

BIOMETRIA APLICADA AO ESTUDO DA DIVERSIDADE GENÉTICA

**Cosme Damião Cruz
Fábio Medeiros Ferreira
Luiz Alberto Pessoni**



**Viçosa - MG
2020**



Universidade Federal de Viçosa - UFV

FAPEMIG Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG

2020

Esta obra trata dos princípios biométricos envolvidos na análise da diversidade genética entre acessos, coleções e populações a partir de informações de diversa natureza. É apresentada ampla abordagem de como analisar e processar dados, para fins de estudo da diversidade, diferenciação, fixação e estruturação da variabilidade fenotípica e genotípica, com os mais diversos propósitos. São apresentadas metodologias de análise aplicáveis a diversas situações relativas ao material genético disponível, aos dados mensurados e aos propósitos da investigação.

**Cosme Damião Cruz
Fábio Medeiros Ferreira
Luiz Alberto Pessoni**

**BIOMETRIA APLICADA AO
ESTUDO DA DIVERSIDADE
GENÉTICA**

**VIÇOSA - MG
2020**

© 2011 by Cosme Damião Cruz
Fábio Ferreira Medeiros
Luiz Alberto Pessoni

1^a. edição : 2011
2^a. edição : 2020

Não é permitida a reprodução total ou parcial deste livro sem a autorização expressa dos autores

Apoio:

Fundação de Amparo à Pesquisa do Estado de Minas Gerais - **FAPEMIG**
Universidade Federal de Viçosa - UFV

Impresso no Brasil

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da
Biblioteca Central da UFV

Cruz, Cosme Damião, 1958-
C957b Biometria aplicada ao estudo da diversidade genética
2011 Cosme Damião Cruz, Fábio Medeiros Ferreira, Luiz Alberto
Pessoni. - Viçosa : UFV, 2020.
xxx. : il.

ISBN: 978-85-60249-70-1

1. Variabilidade (Biologia) - Métodos estatísticos. 2. Biometria.
3. Genética quantitativa. 4. Diferenciação.
CDD 22 ed.576.54

Revisão lingüística: Nelson Coeli

Capa. Fotolito, impressão e acabamento
Suprema Gráfica e Editora

Endereço dos autores

Prof. Cosme Damião Cruz
Departamento de Biologia Geral -UFV
E-mail: cdcruz@ufv.br
Prof. Fábio Medeiros Ferreira
Instituto de Ciências Exatas e Tecnologia em Itacoatiara/UFAM
E-mail: ferreirafmt@ufam.edu.br
Prof. Luiz Alberto Pessoni
Departamento de Biologia - UFRR
E-mail: lualpessoni@ufrr.br

Aos meus pais João S. Cruz e Hilda C. Cruz (*in memoriam*),
À minha esposa Rita de Cássia Rosado Cruz.
Aos meus filhos Patrícia, Rafael e Natália.

Cosme Damião Cruz

Aos meus pais, Laudelino da Silva Ferreira (*in memoriam*) e Maria de Lourdes M. Ferreira,
À minha esposa, Rafaela Fernanda Batista Ferreira,
Aos meus filhos Luiza e Heitor.

Fábio Medeiros Ferreira

Aos meus pais Otílio Pessoni (*in memoriam*) e Romilda Montagnini Pessoni,
Aos meus irmãos, Ricardo, Lindalva, Eliane, Célio, Simone e Flávio

Luiz Alberto Pessoni

AGRADECIMENTOS

À Universidade Federal de Viçosa, pelo apoio.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (**Fapemig**), Conselho Nacional de Desenvolvimento Científico e Tecnológico (**CNPq**) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (**Capes**) pelo apoio financeiro que viabilizaram a publicação deste livro

Aos alunos de pós-graduação, em especial aos orientados e aconselhados, que muito contribuíram para o aprimoramento deste livro.

A todos que, direta ou indiretamente, contribuíram para a concretização desta obra.

Os autores.

PREFÁCIO

Esta obra visa proporcionar aos técnicos, pesquisadores e docentes conhecimentos fundamentais para o estudo da diversidade genética. O leitor encontrará abordagem ampla de como analisar e processar seus dados, para fins de estudo da diversidade genética, com base em informações fenotípicas e genotípicas com os mais diversos propósitos. São apresentadas metodologias de análise aplicáveis a diversas situações relativas ao material genético disponível, aos dados mensurados e aos propósitos da investigação. Também será considerado o interesse de se investigar a diversidade genética entre e dentro de populações, de coleções e de ecótipos. Esses estudos sobre a diversidade têm sido de grande importância para orientar hibridações em programas de melhoramento genético, para avaliação da história e das associações evolutivas entre espécies, para subsidiar trabalhos em banco de germoplasma e para a conservação e manejo de material genético. Neste livro são apresentadas metodologias para o estudo da diversidade genética a partir de informações fenotípicas, que dizem respeito à avaliação de características de distribuições contínuas ou discretas, esta última sendo dos tipos multicategórica ou binária, ou de informações genotípicas, que são obtidas a partir de marcadores moleculares, domintes ou co-dominantes, ou seqüenciamento do DNA.

Preocupou-se com os ensinamentos sobre como analisar e, principalmente, como interpretar parâmetros que certamente irão orientar a utilização de materiais genéticos e recursos humanos, técnicos e financeiros, que normalmente são escassos.

Desde já, agradecemos quaisquer sugestões, críticas e eventuais correções por parte dos leitores, que, certamente, irão contribuir para a melhoria desta obra em novas edições.

Os autores.

SUMÁRIO

CAPÍTULO 1- Importância 1

CAPÍTULO 2 – Diversidade Genética Baseada em Informações Fenotípicas 28

Introdução.....	29
Medidas de Dissimilaridade.....	30
Técnicas de Agrupamento	74
Comparação e propriedades dos métodos de agrupamento SAHN	128
Métodos de agrupamento baseados em dispersão gráfica	144

CAPÍTULO 3 – Discriminação de Populações Baseada em Informações Fenotípicas 186

Introdução.....	187
Análise discriminante baseada em componentes principais.....	188
Análise discriminante linear de Fisher	196
Análise discriminante de Anderson.....	207
Análise discriminante quadrática.....	210
Avaliação de função discriminante.....	211
Dissimilaridade entre as populações	213

CAPÍTULO 4 - Estrutura Genética de Populações 218

Introdução.....	219
Freqüência alélica e genotípica de uma população.....	220
Processos que afetam a freqüência gênica	221
Equilíbrio de Hardy-Weinberg	222
Acasalamentos.....	265
Endogamia	279
Fixação Gênica	293
Índice de Fixação	309
Fluxo Gênico ou Migração.....	316

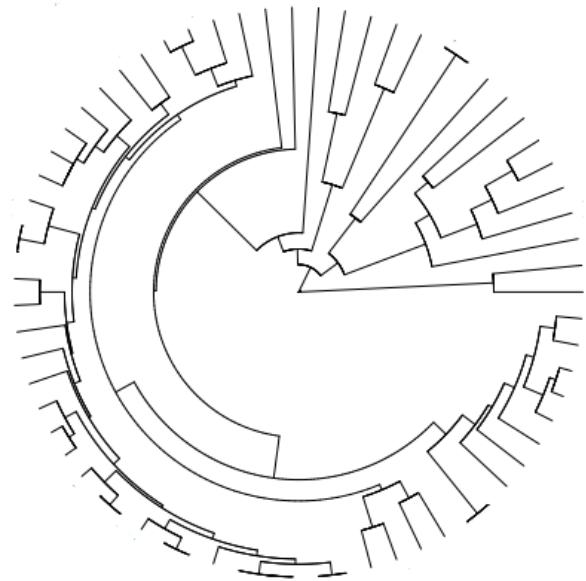
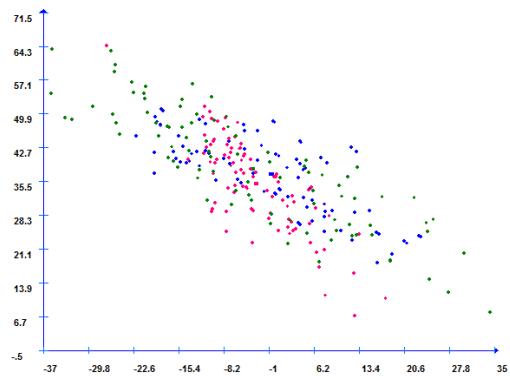
CAPÍTULO 5 - Diversidade Genética Baseada em Informações Moleculares..... 321

Diversidade em Populações e Coleções	322
Diversidade entre Acessos ou Indivíduos	327
Distância Genética entre Populações ou Coleções.....	382
Estatísticas Descritivas da Diversidade Dentro de Populações.....	418
Análise Discriminante Não-paramétrica	422
Importância Relativa de Marcadores Moleculares.....	424

CAPÍTULO 6 – Diferenciação Genética Baseada em Informações Moleculares	429
Introdução	430
Identidade, Heterozigosidade e Diversidade Genética - Estatística GST de Nei	431
Diferenciação baseada na frequencia gênica ou genotípica	444
Diferenciação baseada na análise de variância de uma variável indicadora.....	450
Diferenciação por meio da estatística F de Wright	453
Diferenciação medida pela heterozigosidade – Weir(1966).....	470
Análise molecular de variância - AMOVA.....	486
Emprego das medidas de diversidade	505
Emprego das medidas de distância genética	507
Técnicas de agrupamento a partir de índices de fixação	508
CAPÍTULO 7 - Análise Filogenética Molecular	509
Introdução.....	510
Definições e Terminologias Empregadas na Análise Filogenética	512
Mudanças Evolutivas e Diferenças Entre Seqüências de Nucleotídeos.....	524
Métodos de Reconstrução de Árvores	536
Acurácia e Testes Estatísticos para Árvores Filogenéticas.....	588
Vantagens e Desvantagens dos Diferentes Métodos de Reconstrução	591
Problemas Associados À Reconstrução Filogenética	593
Estratégias Para Minimizar Erros na Análise Filogenética	594
ANEXOS.....	597
LITERATURA CITADA.....	603

Capítulo 1

Diversidade Genética – Importância



Introdução

A conservação de recursos genéticos de espécies vegetais e animais é um dos temas considerados de grande relevância na atualidade, razão pela qual grande número de estudos tem sido realizado na quantificação da diversidade genética e no entendimento de sua magnitude, natureza e distribuição entre e dentro de populações. O sucesso de qualquer programa de pré-melhoramento ou de conservação é dependente do conhecimento da quantidade de variação presente na espécie de interesse.

A avaliação da diversidade genética era, originalmente, realizada a partir de informações fenotípicas relativas às características morfológicas ou de desempenho agronômico ou zootécnico. Contudo, os recentes avanços na biologia molecular abriram novas perspectivas para a pesquisa em conservação de espécies e para os estudos de biologia populacional. Com a utilização de marcadores moleculares, é possível a detecção da variabilidade existente diretamente em nível do DNA.

Percebe-se que grande número de metodologias está, atualmente, disponível para a quantificação e a avaliação da diversidade em estudos populacionais, a partir de informações fenotípicas e genotípicas. No entanto, a abrangência dos estudos, de informações, de métodos e de material biológico tem levado a certa dificuldade em escolher e aplicar corretamente as metodologias disponíveis e interpretar, convenientemente, o significado dos resultados das análises biométricas.

Constata-se também que, após a obtenção de informações a partir das técnicas biométricas, ainda existe certa carência quanto ao melhor aproveitamento dos dados colhidos e à adequada interpretação de seus resultados. Dessa forma, a disponibilidade de um referencial teórico e prático, que oriente a utilização dos recursos biométricos, permitirá melhor aproveitamento destes e também levará às interpretações corretas dos resultados obtidos.

Nesta obra, o leitor encontrará abordagem ampla sobre como analisar e processar seus dados, para fins de estudo da diversidade e discriminação genética, com base em informações fenotípicas (Capítulo 2 e 3), genotípicas ou alélicas (Capítulos 4, 5 e 6) e de nucleotídeos (Capítulo 7), com os mais diversos propósitos. Serão apresentadas metodologias de análise aplicáveis às diversas situações relativas ao material genético disponível, aos dados mensurados e aos propósitos da investigação. Assim, se o interesse do leitor é pelo melhoramento e conservação de bancos de germoplasma os dados disponíveis, provavelmente, serão fenotípicos e os Capítulos 2 e 3 serão indispensáveis em seus estudos. Se o interesse é a caracterização genética, estimar o fluxo gênico e conhecer a estruturação das populações, os dados disponíveis serão genotípicos ou alélicos e o Capítulo 4 proporcionará conhecimentos básicos de genética de populações e os Capítulos 5 e 6 darão suporte às suas análises. Por fim, se o interesse é o estudo evolutivo determinante da diferenciação entre as espécies, os dados deverão ser do DNA e a leitura do capítulo 7 proporcionará valiosas informações.

Também será considerado o interesse de se investigar a diversidade genética entre e dentro de populações, de coleções e de ecótipos. Entende-se por populações um grupo de indivíduos, pertencentes a uma determinada espécie, com sistema de acasalamento supostamente conhecido, de forma que as relações genotípicas e alélicas podem ser estabelecidas e relacionadas entre populações no espaço e no tempo. As unidades de uma população são indivíduos, os quais são mensurados, gerando informações fenotípicas ou genotípicas. Coleções são agrupamentos de acessos pertencentes, ou não, a uma mesma espécie, mas geralmente bastante relacionados, que constituem o somatório da informação genética de uma entidade que se deseja preservar. Os ecótipos são reuniões de táxons (ou unidade taxonômica) cujas relações evolutivas são objetos de estudos na diversidade genética.

Os estudos sobre a diversidade também distinguem-se pelos seus objetivos, podendo estar voltados para o melhoramento genético, para as associações evolutivas e para conservação e manejo de material genético. Em cada caso, é

necessário o uso de metodologia adequada e informações apropriadas. As informações fenotípicas dizem respeito à avaliação de características de distribuições contínuas ou discretas, esta última sendo dos tipos multicategórica ou binária. Informações genotípicas são obtidas a partir de marcadores moleculares ou seqüenciamento do DNA. No caso de marcadores, estes podem ser dominantes ou co-dominantes, do tipo dialélico ou multialélico.

Biodiversidade

A biodiversidade reúne todas as espécies de plantas, animais e microrganismos, assim como os ecossistemas e os processos ecológicos dos quais estas fazem parte. A potencialidade de utilização dos recursos biológicos pode ser observada e manejada por meio dos *recursos genéticos*, os quais possuem agregado um valor econômico atual ou potencial presente nos genes, como ocorre com outros recursos: florestais, minerais, energéticos etc. (QUEROL, 1993). Esses recursos constituem-se em parte essencial da biodiversidade, responsável pelo desenvolvimento sustentável da agricultura e da agroindústria. Essa porção da biodiversidade, constituída por plantas, animais e microrganismos de valor socioeconômico atual ou potencial, para uso em sistemas de produção no setor agrícola, abrangendo os agroecossistemas que os contêm, é denominada hoje de *agrobiodiversidade*.

Os *recursos genéticos* são estabelecidos por acessos que representam a variabilidade genética organizada em um conjunto de materiais diferentes entre si, denominados *germoplasma*. Cada unidade de germoplasma deve constituir uma cópia única do material genético e representativa do organismo vivo de interesse atual ou potencial. Conseqüentemente, o germoplasma é o elemento dos *recursos genéticos* que maneja a variabilidade genética inter e intra-específica, para conservá-la e utilizá-la na pesquisa em geral, especialmente em programas de melhoramento genético. Assim, os recursos genéticos

compreendem a diversidade do material genético contido nas variedades primitivas, obsoletas, tradicionais, modernas, parentes silvestres das espécies alvo, espécies silvestres e linhas primitivas, que podem ser usadas no presente ou no futuro, para a alimentação, agricultura e outros fins.

Na década de 1990 já se estimava que a diversidade global das espécies de plantas superiores girasse em torno de 300.000 a 500.000 espécies, das quais cerca de 250.000 foram identificadas e descritas (FAO, 1996). No entanto, a fragmentação de espécies, a perda da variabilidade genética e mesmo a extinção de espécies são processos nítidos e alarmantes em muitas regiões. Muitas regiões estão sendo dilapidadas pela consolidação das culturas anuais ou pelo avanço de outras atividades humanas no processo de urbanização ou industrialização. Além disso, muitas áreas de vegetação nativa ou faixas de floresta remanescentes sofreram ou sofrem algum grau de degradação pela retirada de produtos (lenha e madeira), pela entrada de outras espécies e por fatores ambientais, como incêndios, urbanização e povoamentos diversos. Dessa forma, espécies que antes mantinham suas populações em tamanho e níveis reprodutivos adequados, hoje se encontram reduzidas, muitas vezes ocorrendo indivíduos isolados que não conseguem trocar pólen (material genético) com outros, o que dificulta ou mesmo inviabiliza a capacidade reprodutiva da espécie naquele local. Dessa forma, a diversidade de espécies também vai progressivamente diminuindo.

Estudos sobre diversidade genética têm sido de grande importância para fins de melhoramento genético e para avaliar o impacto da atividade humana na biodiversidade. São igualmente importantes no entendimento dos mecanismos micro e macroevolutivos que atuam na diversificação das espécies, envolvendo estudos populacionais, bem como na otimização do processo de conservação da diversidade genética. Também são fundamentais no entendimento de como as populações naturais se estruturam no tempo e no espaço e quais os efeitos das atividades antrópicas nessa estruturação e, por consequência, nas suas

chances de sobrevivência e/ou extinção. Essas informações constituem subsídio para constatar as perdas genéticas geradas pelo isolamento das populações e dos indivíduos, o que se refletirá nas futuras gerações, permitindo o estabelecimento das melhores estratégias para incrementar e preservar a diversidade das espécies e a diversidade dentro das espécies.

Perda da Diversidade

Estudos sobre a diversidade genética e o nível de diferenciação genética entre as populações das espécies são essenciais para definir os estoques genéticos e subsidiar políticas de exploração e manejo desses recursos, bem como para traçar estratégias de conservação em escalas regional e geográfica.

A diversidade genética das espécies é uma importante forma de manter a capacidade natural de responder às mudanças climáticas e a todos os tipos de estresses bióticos e abióticos. Na atualidade existe grande preocupação em avaliar a biodiversidade, em razão da perda acentuada da diversidade genética, sobretudo devido à ação do homem, substituindo variedades locais por variedades modernas, híbridos e, mais recentemente, clones, de forma que grandes extensões de área são ocupadas por uma ou poucas variedades ou materiais de base genética estreita.

A perda dessa diversidade provavelmente diminuirá a capacidade dos organismos de responderem às mudanças ambientais e eliminará também informações biológicas potencialmente úteis aos homens, como a diversidade genética de espécies cultivadas e valiosos compostos bioquímicos ainda nem conhecidos.

As causas mais importantes da perda da biodiversidade e dos *recursos genéticos* são as que se seguem:

Destrução dos habitats e comunidades naturais

Tem sido ocasionada por práticas agrícolas não-sustentáveis. A destruição das florestas e dos bosques nativos para o preparo de áreas para cultivos, falta de manejo sustentável dos recursos naturais, pressão da população, conflitos civis, urbanização, degradação do ambiente, pastoreio excessivo, mudanças no sistema de produção e também a legislação e as políticas governamentais que ignoram essas áreas.

Há consenso de que a ocupação e a exploração desordenada da terra é uma das principais causas de extinção de muitas espécies. O desmatamento e a degradação dos ambientes naturais, o avanço da fronteira agrícola, assim como a introdução de espécies exóticas, são fatores que participam de forma efetiva no processo de extinção. À medida que a população cresce e os índices de pobreza aumentam, este processo amplia-se e sua principal e imediata consequência é a fragmentação das paisagens naturais.

Vulnerabilidade genética

A vulnerabilidade ocorre quando um material genético (variedade, linhagem ou população) amplamente cultivado se torna uniforme, suscetível a uma doença, praga ou azar climático, como resultado de sua constituição genética restrita, criando potencial para extensa perda do cultivo (KLOPPENBURG; KLEINMAN, 1987).

Algumas espécies são naturalmente mais frágeis, seja por que são muito procuradas por seu valor econômico, fornecendo frutos (cerejeira), madeira nobre (cedro, grápia, canjerana, ipê-mandioca) ou produtos medicinais (murta, jaborandi, canjerana, erva-mate); ou porque possuem populações pouco densas (cedro); ou porque exigem áreas de vegetação bem preservadas para sobreviver (jaborandi). A conservação dessas espécies viabiliza, além da

manutenção da paisagem nativa, atividades econômicas de forma sustentável em longo prazo.

Uma vez que algumas espécies são mais vulneráveis do que outras à redução da área provocada pela atividade antrópica, é fundamental avaliar os efeitos desse fator, além de outros, na composição das espécies e de seus possíveis fragmentos nos diferentes biomas. Um parâmetro importante a ser avaliado, portanto, é verificar a fração dos componentes intra e interpopulacionais da variabilidade genética total de uma dada espécie. Em espécie com baixa variabilidade genética, o componente de variação interpopulacional pode ser grande, devido à adaptação local ou simplesmente à divergência devida ao baixo fluxo gênico (FRANKHAM et al., 2003). A substituição constante de variedades locais por cultivares modernos homogêneos constitui-se em importante causa da vulnerabilidade genética. Assim, programas de conservação devem proteger os ambos componentes da diversidade genética.

Erosão genética

A erosão constitui na redução da diversidade genética, com perda de genes individuais (NRC, 1972) e de combinações particulares de genes (*gene-pools*), como aquelas manifestadas nas raças locais adaptadas. É relatado que apenas 15 a 30 espécies de plantas superiores sejam responsáveis por 90% da alimentação humana. Essa situação reflete considerável perda da diversidade genética vegetal, o que representa uma erosão genética ao longo do tempo.

A causa principal da erosão genética dos cultivos é também a substituição das variedades locais por espécies exóticas e variedades melhoradas. Assim, programas de melhoramento – responsáveis pelo desenvolvimento de variedades superiores que são a base da agricultura

moderna – são considerados a maior causa de erosão genética. Diante desse fato, deve ser destacado que, na busca de genótipos cada vez mais eficientes, há muitos desafios na área de conservação que necessitam ser enfrentados e superados. O melhorista deve ver a variabilidade genética como fator indispensável à obtenção de ganhos genéticos, e suas técnicas devem ser direcionadas para o desenvolvimento de materiais genéticos superiores, mas com o comprometimento de que a recuperação e manutenção de populações de espécies ameaçadas de extinção sejam também metas prioritárias, para a própria sobrevivência da humanidade. A grande quantidade de variabilidade genética encontrada dentro das populações naturais, a exemplo da seringueira (PAIVA et al., 1994), associada à forma de distribuição dos indivíduos nessas populações, é um forte indicativo de que o cultivo racional de espécies autóctones permitirá novos enfoques e novas estratégias de melhoramento genético. Trabalhos realizados com os mais variados tipos de espécies arbóreas, florestais e/ou frutíferas (pinho caribenho, em ZHENG; ENNOS, 1999; dendê, em BARCELOS et al., 2002; canela-amarela, em MORAES; DERBYSHIRE, 2002; cagaita, em ZUCCHI et al., 2003; pimenta-de-macaco, em GAIA et al., 2004; pequi, em Melo Júnior et al., 2004; pimenta-longa, em WADT; KAGEYAMA, 2004) vêm demonstrando a importância de não apenas conhecer, mas também manter o equilíbrio biológico das populações no ecossistema tropical.

O progresso promovido pela espécie humana tem trazido, além dos benefícios almejados, uma série de problemas paralelos, que vêm preocupando, cada vez mais, as autoridades e o cidadão comum. O conhecimento técnico-científico atual sobre as espécies nativas de nossa fauna e flora ainda é insuficiente, o que dificulta o planejamento de ações de conservação em um contexto de escassez de recursos financeiros. Nossa flora tem sido fortemente afetada pelas atividades antrópicas, que levam, freqüentemente, à fragmentação do habitat. Essa fragmentação, por sua vez, conduz à redução no

tamanho das populações, fator que leva à perda de variabilidade genética e a mudanças na distribuição dessa variabilidade entre as populações (SPIESS, 1989).

A perda de variabilidade genética pode resultar na diminuição da flexibilidade evolutiva, ou seja, na capacidade de resposta das espécies a alterações ambientais e, também, na diminuição do valor adaptativo. Proteger e utilizar racionalmente os recursos de nossa flora exige ações de manejo que demandam conhecimento, técnica, controle e monitoramento. A proteção e o manejo ordenado da flora nativa na busca de sua conservação podem e devem ser feitos pelo governo e pela sociedade de forma integrada, no sentido de defender o que é de todos: o patrimônio natural do país, bem de uso comum de todos os brasileiros e garantia para as futuras gerações. A análise da biodiversidade disponível, associada ao estudo dos processos evolutivos e biogeográficos responsáveis pela geração dessa biodiversidade, torna-se, portanto, fundamental para que possam ser tomadas medidas efetivas para sua preservação.

Reconhecendo que o intenso desenvolvimento agropastoril, aliado, em algumas regiões, à industrialização, tem conduzido à fragmentação e poluição das paisagens naturais, é necessário ampliar as investigações o impacto desses processos sobre a diversidade taxonômica e estrutura genética de algumas espécies da flora e contribuir para o desenvolvimento de estratégias visando sua conservação.

Deriva genética

A deriva genética é um dos fatores que alteram a freqüência gênica de forma dispersiva, ou seja, tem magnitude quantificável, porém o sentido de sua ação é imprevisível, podendo tanto aumentar quanto diminuir a freqüência de

um determinado alelo na população. É um mecanismo que, atuando juntamente com a seleção natural, pode proporcionar grandes mudanças nas características das populações ou espécie ao longo do tempo, e essas alterações induzidas poderão não constituir vantagens adaptativas. A deriva genética e a seleção natural raramente ocorrem independentemente; esses dois fenômenos estão sempre atuando em conjunto numa população, mas o grau em que a freqüência de cada alelo é afetada pode variar em função das circunstâncias. Numa população com um tamanho efetivo elevado, a deriva ocorre muito lentamente, e a seleção que atua sobre um alelo pode, de maneira relativamente rápida, aumentar ou diminuir a sua freqüência (dependendo da viabilidade do alelo). Por outro lado, numa população com tamanho efetivo reduzido, o efeito da deriva genética predomina e, neste caso, o efeito da seleção natural é menos visível, pois o efeito da deriva muitas vezes se sobrepõe.

Trata-se de um processo estocástico, atuante sobre as populações, modificando a freqüência dos alelos e, consequentemente, contribuindo para predominância de certas combinações genotípicas na população. Seus efeitos se manifestam com maior freqüência em populações com tamanhos efetivos reduzidos. Assim, os desvios estatísticos adquirem importância especial, uma vez que, nestas populações de tamanho reduzido, podem até mesmo eliminar determinados genótipos.

Um caso particular da deriva genética, ou oscilação genética, é aquele proporcionado pelo *princípio do fundador*, que se refere ao estabelecimento de uma nova população a partir dos intercruzamentos de poucos indivíduos formadores da população inicial, geralmente migrantes de uma população original ou de referência. Esses indivíduos representarão pequena fração da variação genética da população original, e seus descendentes manifestarão apenas essa variabilidade restrita, ampliada somente quando novos alelos surgirem por mutação. O princípio do fundador é, portanto, determinante da uniformidade genética e fenotípica, fazendo com que essa pequena população

esteja sujeita à ação pronunciada dos efeitos dos vários fatores evolutivos, podendo originar indivíduos diferentemente adaptados. Há elevada probabilidade de ocorrer endogamia, resultando num nível anormal de defeitos, relacionados com a expressão de alelos recessivos em populações estabelecidas com número reduzido de genitores. O estabelecimento de populações pelo princípio do fundador parece ser um dos métodos mais comuns de dispersão de inúmeras espécies de animais e de plantas.

Outro fenômeno a ser considerado em estudos de deriva genética é o *efeito gargalo*, que pode resultar em rápidas e dramáticas alterações nas freqüências alélicas. Esse efeito ocorre de maneira independente da seleção natural, e muitas adaptações benéficas poderão ser eliminadas da população. O efeito de gargalo tem sido definido como um evolutivo evolucionário no qual uma porcentagem significativa da população de uma espécie morre ou é impedida de se reproduzir.

Conservação dos Recursos Genéticos - Banco de Germoplasma

Uma das grandes preocupações da sociedade é o estabelecimento de bancos de germoplama, uma vez que a redução significativa no número de populações de espécies, em especial de vegetais, torna necessário o armazenamento de sementes em coleções dos bancos de germoplasma. Esse tipo de conservação *ex situ* tem sido apontado como uma estratégia importante, pois permite, em caso de redução significativa ou desaparecimento de populações naturais, o reforço ou reintrodução na natureza. Entretanto, para que as sementes constituam uma importante reserva da diversidade da espécie, é necessário que sejam representativas da diversidade alélica e do potencial evolutivo.

Assim, estudos genéticos, com vista ao conhecimento da variabilidade genotípica, são de grande importância, pois permitem determinar quais os

indivíduos/populações que apresentam reduzida diversidade genética. A informação sobre a diversidade genética possibilitará a determinação de indivíduos/populações mais adequados a ações que envolvam reforço ou reintrodução. Os estudos genéticos também permitem a identificação de problemas, como sistemas alélicos auto-incompatíveis, perda de alelos devido à deriva genética e depressão endogâmica, e acumulação de mutações prejudiciais, as quais podem conduzir à extinção de uma espécie.

Uma vez estabelecidos os bancos de germoplamas, as análises de diversidade de diferentes acessos têm permitido estudos precisos da biodiversidade, de forma a orientar a utilização de diferentes genitores e, assim, melhor planejamento de cruzamentos a serem avaliados no programa de melhoramento. Além disso, tais análises têm possibilitado a eliminação de duplicatas e a análise de pureza dos diferentes acessos.

Diversidade genética para fins de melhoramento genético

Embora o cenário tenha apresentado mudanças na concepção de interação de recursos genéticos e programas de melhoramento, ainda existe uma lacuna entre essas atividades (NASS, 2001). O processo de pré-melhoramento serve de elo entre pesquisa de coleta e conservação da variabilidade dos recursos genéticos e exploração da diversidade genética disponível nos programas.

Conceitua-se pré-melhoramento como o conjunto de atividades que visam a identificação de caracteres e/ou genes de interesse, presentes em germoplasmas exóticos e germoplasmas semi-exóticos (NASS e PATERNIANI, 2000),. Programas de pré-melhoramento como os de sorgo, tomate e milho foram evidenciados pela literatura (SMITH; DUVICK, 1989; HOYT, 1992; NASS e PATERNIANI, 2000) como importantes estratégias para a formação dos materiais modernos, atualmente, adaptados a climas diferenciados, resistentes

a moléstias, de melhor qualidade de grãos, tolerantes a estresse abiótico e com melhor estabilidade.

O processo inicial de avaliação da diversidade genética visa a identificação de genitores adequados à obtenção de híbridos com maior efeito heterótico e que proporcionem maior segregação em recombinações, possibilitando o aparecimento de transgressivos. A importância dos estudos sobre a divergência genética para o melhoramento reside no fato de que cruzamentos envolvendo genitores geneticamente diferentes são os mais convenientes para produzir alto efeito heterótico e, também, maior variabilidade genética em gerações segregantes (RAO *et al.*, 1981). Para isso, busca-se população-base para seleção que alie ampla variabilidade genética com alta média para o caráter a ser selecionado (MALUF; FERREIRA, 1983).

A expectativa de que pais divergentes proporcionem bons híbridos decorre do fato de que, segundo Falconer (1987), a heterose manifestada em híbridos é função dos efeitos da dominância dos genes para o caráter em questão e do quadrado da diferença das freqüências gênicas de seus genitores, além de efeitos epistáticos que geralmente são negligenciados. Há duas maneiras básicas de se inferir a diversidade genética, sendo a primeira de natureza quantitativa e a outra de natureza preditiva.

Entre os métodos de natureza quantitativa de avaliação da diversidade, ou da heterose manifestada nos híbridos, citam-se as análises dialélicas. Nesses métodos, é necessária a avaliação de p genitores e de todas (ou amostras de) suas combinações híbridadas, resultando num total de $p(p-1)/2$ híbridos a serem avaliados. Em algumas culturas, a polinização manual é onerosa, difícil de ser executada e com pouca probabilidade de êxito na obtenção da semente híbrida ou em quantidade muito reduzida, para que possam ser feitas avaliações apropriadas.

Entre os métodos preditivos da heterose citam-se aqueles que tomam por base as diferenças morfológicas, fisiológicas ou moleculares, quantificando-as

em alguma medida de dissimilaridade que expressa o grau de diversidade genética entre os genitores. A inferência da diversidade genética com base na diversidade geográfica também é exemplo preditivo da heterose.

Uma medida de divergência genética que possa ser obtida antes que os cruzamentos sejam efetuados permitiria ao melhorista concentrar esforços naquelas combinações que apresentem maiores chances de sucesso (MALUF; FERREIRA, 1983). Moraes et al. (2005) avaliaram a proximidade entre genótipos usados como genitor recorrente e genitor doador de alelos para alto teor de proteína, orientados por marcadores microssatélites, visando a obtenção de linhagens de soja com altos teores de proteína com o mínimo de gerações de retrocruzamento. Os autores definiram grupos de potenciais cruzamentos entre os genótipos recorrentes e genótipos doadores, a partir das menores distâncias genéticas entre recorrente e doador. Essa proposição permite otimizar o processo de retrocruzamento, mantendo assim as características desejáveis do genitor recorrente.

Moll et al. (1965), em estudo sobre divergência em milho, concluíram que deve existir um grau ótimo de divergência para expressão máxima da heterose. Esse ótimo ocorre dentro de uma faixa de divergência que é suficientemente estreita, de forma que barreiras de incompatibilidade, como aquelas causadas por irregularidades citológicas, não se manifestam. Estes autores argumentam, fundamentando-se no trabalho realizado por Paterniani e Lonnquist (1963), que deve existir um nível ótimo de divergência entre os genitores para obtenção de heterose, uma vez que raças com ampla divergência foram cruzadas, porém pouca ou nenhuma diferença na heterose do cruzamento entre raças de mesmo tipo de endosperma e aquelas de tipos de endosperma diferentes foi observada.

Dois genitores não distantes geneticamente entre si tendem a compartilhar muitos genes ou alelos em comum. Quando dois destes genitores são cruzados, há pouca complementaridade e baixo vigor em razão do baixo nível de heterozigosidade alélica no cruzamento. Entretanto, quando dois genitores são mais

distantes geneticamente, é admitido que eles difiram de forma crescente no número de locos nos quais os efeitos da dominância estão evidentes, contribuindo, consequentemente, para a maior manifestação da heterose (Ghaderi et al., 1984). Outra forma de obter heterose para características complexas é pela substituição de genótipos e complementação de características nos híbridos superando as fraquezas individuais de seus genitores (WILLIANS, 1959; GRAFIUS, 1961; SINGH et al., 1981; GHADERI et al., 1984).

Assim, é preciso cautela na avaliação da divergência genética de indivíduos. Conforme Ghaderi et al. (1984), dois genótipos podem ser completamente distantes geneticamente – isto é, o melhor e o pior segregante de um cruzamento – e ainda serem estreitamente relacionados, por serem membros de uma mesma população. Seria mais interessante selecionar, como genitores, dois genótipos que apresentam bom desempenho, mas não relacionados, ou seja, geneticamente distantes entre si, pois, devido à sua não-proximidade, contribuiriam com um arranjo genético diferente e mais proveitoso. Embora teoricamente possível, não é provável que dois pais possam ser geneticamente próximos e ainda produzir heterose, por causa da distribuição contrastante para alelos nos locos, que afeta a característica. A circunstância mais provável é que, se são geneticamente próximos, eles terão arranjos genéticos similares para aquela característica (GHADERI et al., 1984).

No campo do melhoramento genético, os marcadores moleculares também têm dado valiosas contribuições, com destaque para a detecção do parentesco genético entre diferentes germoplasmas em bancos de sementes e programas de melhoramento; a predição da heterose; a busca por grupos heteróticos promissores para constituição de híbridos; a identificação de duplicatas nos bancos de germoplasma; a avaliação do fluxo gênico ao longo do tempo; e a identificação de variedades essencialmente derivadas de variedade de planta protegida.

Diversidade Genética Molecular

O desenvolvimento de novas tecnologias no campo da biologia molecular, especialmente a reação em cadeia da polimerase (PCR, do inglês *polymerase chain reaction*), permitiu a popularização de métodos moleculares em diversas áreas do conhecimento biológico. De acordo com Carlini-Garcia et al. (2001), o surgimento dos marcadores bioquímicos, a exemplo das isoenzimas, e marcadores moleculares proporcionou um salto qualitativo e quantitativo em estudos sobre a estrutura populacional e o sistema reprodutivo das espécies vegetais. Atualmente, a análise de polimorfismos de fragmentos de DNA tem sido uma das principais ferramentas para estudo da biodiversidade, permitindo desde inferir padrões de distribuição espacial da diversidade genética, até testar hipóteses explícitas de biogeografia histórica e definir áreas prioritárias de conservação. A avaliação de polimorfismos moleculares em regiões não-codificadoras tem fornecido informações importantes sobre os vários níveis de diversidade genética, intra x interpopulacional e intra x interespecífica, tanto em animais como em plantas. Nesse sentido, os microssatélites (SSRs), repetições em *tandem* de um a seis nucleotídeos, são ferramentas ideais se distribuírem aleatoriamente no genoma da maioria dos organismos eucariotos, por apresentarem co-dominância e alto grau de polimorfismo e serem passíveis de detecção por PCR (GLOWATZKI-MULLIS et al., 1995; ARROYO et al., 1994).

Alguns estudos (CARDLE et al., 2000; MORGANTE et al., 2002) mostram que os microssatélites são amplamente distribuídos no genoma das plantas superiores e com freqüência de distribuição variada quando se considera a infinidade de classes de microssatélites. Isso engloba todas as combinações de um a seis nucleotídeos, somados à variação do número de unidades repetitivas que caracterizam um (alelo) microssatélite. A identificação e caracterização de SSRs baseados em screening de bibliotecas genômicas têm sido relatadas para várias espécies (SAGHAI-MAROOF et al., 1984; AKKAYA et al., 1992; WU; TANKSLEY, 1993; MORGANTE et al., 1994; BECKER & HEUN, 1995a) e sua

utilidade em plantas tem sido demonstrada em vários estudos genéticos relacionados à construção de mapas genéticos, (BELL; ECKER, 1994; AKKAYA et al., 1995; BECKER; HEUN, 1995b; LIU et al., 1996; TEMNYKH et al., 2000) e associação evolutiva entre espécies correlatas (SANTACRUZ-VARELA et al., 2004).

Diversidade Genética e Geográfica

A utilização da diversidade geográfica como indicadora da diversidade genética tem recebido críticas, pelo fato de que, por esse critério, não se quantifica a diversidade existente entre as populações e de que, em muitos casos, não se verifica relação entre diversidade genética e distância geográfica. Sugere-se que informações ecológicas, de biologia populacional e/ou de biologia reprodutiva devam ser extraídas para que os padrões genético-geográficos sejam melhor entendidos.

Ran e Panwar (1970) relataram que o padrão de agrupamento de 120 variedades de arroz esteve relacionado com a diversidade geográfica do material avaliado. Resultados semelhantes foram encontrados por Joshi e Dhawan (1966) em várias culturas autógamas e por Singh e Singh (1976) em *Capsicum annum* L. Entretanto, Peter e Rai (1976) constataram que não havia nenhuma relação entre a diversidade genética de cultivares de tomate e a diversidade geográfica do material avaliado, pois, em muitos casos, as populações pertencentes a diferentes regiões ficaram agrupadas conjuntamente, enquanto as populações originárias de uma mesma região foram classificadas em grupos diferentes. Esse mesmo fato foi relatado em trabalhos com algodão (AMALRAJ, 1982; SINGH; GILL, 1984).

Murty e Arunachalan (1966) e Upadhyay e Murty (1970) relataram que a deriva genética e a seleção, em diferentes ambientes, podem causar maior diversidade que a distância geográfica. Além disso, nos tempos atuais, há

intensas trocas de sementes ou estruturas propagativas das várias espécies entre pesquisadores e instituições de diferentes regiões. Como consequência, há perdas de individualidade e de ocorrência de tipos particulares em virtude da interferência humana.

O uso de marcadores moleculares no estudo de associações genéticas e geográficas também é dificultado pela complexidade do padrão na distribuição da variação genética exibido pelas espécies (FAHIMA et al., 1999), necessitando do uso de análises multifacetadas. No trabalho de Del Rio e Bamberg (2004), os autores revisaram alguns trabalhos com espécies de batata em que se relacionaram parâmetros geográficos (altitude, temperatura, longitude etc.) e a proximidade geográfica com a predição da variação genética nestas espécies. Em um dos estudos revisados, verificou-se correlação significativa entre as diferenças gerais dos marcadores moleculares e a separação geográfica para populações dentro de espécies da série *Etuberosa* (um grupo com nível de endogamia extremamente alto). Outra investigação da associação genético-geográfica foi avaliada por Del Rio et al. (2001) e Del Rio e Bamberg (2002) com acessos selvagens de batata, diferentes sistemas de acasalamento e nível de ploidia. Numa visão global, os autores afirmaram que não existe uma associação significativa entre as variáveis ecogeográficas testadas e a diversidade genética em batata. Mesmo fisicamente separadas, a diversidade genética entre os acessos não foi predita. Já Del Rio e Bamberg (2004) encontraram tal associação e a justificaram em função da natureza uniforme e endogâmica da espécie *Solanum verrucosum*. Os autores afirmaram que a distinção genético-geográfica de grupos gera informações úteis na “captura” da diversidade genética (ou alélica) para a composição de futuras coleções de germoplasma.

Filogenia Molecular

A análise filogenética refere-se ao estudo das relações de parentesco entre organismos por meio da utilização de dados fenotípicos ou moleculares, como seqüências de DNA e de proteínas, ou, ainda, de outros marcadores moleculares. Este tipo de análise é de fundamental importância para compreender a história evolucionária das espécies ou de quaisquer outras entidades biológicas, possibilitando a reconstrução dos laços genealógicos corretos que as unem e a realização de inferências a respeito do tempo de divergência entre elas (isto é, o tempo desde quando tais entidades compartilharam um ancestral comum), assim como o estabelecimento da cronologia de seqüências de eventos das diferentes linhagens evolutivas.

Trabalhos de filogenia baseados em dados moleculares iniciaram-se há mais de um século, quando testes sorológicos mostraram reação cruzada intensa entre organismos biologicamente mais próximos do que entre aqueles relativamente distantes. As implicações evolutivas desses achados foram utilizadas para inferir a filogenia entre vários grupos de animais, como os mamíferos placentários, primatas e ungulados. A partir destes estudos, também foi possível estabelecer que os grandes macacos (chimpanzé, gorila e orangotango) são os parentes mais próximos da espécie humana, seguidos, em ordem decrescente de parentesco, pelos primatas do velho mundo, primatas do novo mundo e prossímios (GRAUR; LI, 2000).

Várias técnicas de biologia molecular, desenvolvidas a partir do final da década de 1950, passaram a ser amplamente utilizadas na geração informações úteis à análise filogenética, como o seqüenciamento de proteínas, ou outras menos complexas, como a hibridização DNA-DNA, eletroforese de isoenzimas e métodos imunológicos. Entretanto, a filogenia molecular começou a se desenvolver de maneira mais acelerada a partir do final dos anos de 1970, como resultado do acúmulo de informações referentes ao seqüenciamento direto do DNA. Isso porque dados de seqüências de DNA são mais abundantes e mais fáceis de serem analisados de seqüências de proteínas. Além do mais, esse tipo de dado pode ser

empregado tanto na determinação de filogenias, envolvendo populações ou espécies próximas, como as populações humanas, quanto em reconstruções filogenéticas de entidades com origens evolutivas muito antigas, como as origens da mitocôndria e do cloroplasto ou divergência entre filos e reinos.

Considerações mais detalhadas sobre filogenia molecular serão apresentadas em capítulo específico nesta obra.

Filogeografia

A filogeografia (AVISE et al., 1987) trata do estudo dos princípios e processos que governam a distribuição geográfica de linhagens genealógicas, providenciando um meio para elucidar o papel do fluxo gênico histórico na estruturação genética das populações. Possibilita conhecer os processos histórico-evolutivos responsáveis pelos padrões observados de distribuição da variabilidade genética nas populações em escalas regional e continental.

Assim, em uma abordagem sobre filogeográfica são considerados os componentes histórico e filogenético da distribuição espacial de linhagens gênicas. Ou seja, tempo e espaço são os eixos da filogeografia nos quais são mapeadas genealogias de genes de interesse. Por outro lado, abordagens que considerem aspectos filogenéticos da distribuição espacial de qualquer caráter genético – sejam eles morfológico, comportamental ou molecular – também pode ser qualificadas como filogeográficas.

A análise e interpretação da distribuição de linhagens requerem, usualmente, o acúmulo e processamento de grande quantidade de informações da genética molecular, genética de populações, etologia, demografia, filogenia, paleontologia, geologia e geografia histórica. Por isso, a filogeografia exerce um papel integrativo entre disciplinas das áreas de micro e macroevolução.

Como uma área da biogeografia, a filogeografia coloca em um contexto temporal mais amplo as perspectivas tradicionais da ecogeografia, que enfatiza o papel de pressões ecológicas contemporâneas em moldar a distribuição espacial dos caracteres dos organismos. Dessa forma, a filogeografia procura interpretar a

extensão e o modo pelos quais processos históricos, incluindo pressões seletivas e outros fatores evolutivos, deixaram marcas evolutivas na distribuição geográfica de caracteres genéticos dos organismos.

A filogeografia também serve como um escopo conceitual que cobre cenários históricos alternativos que possam explicar o arranjo espacial dos organismos e suas características, como, por exemplo, fazer a distinção entre eventos de dispersão e vicariância.

As raízes históricas da origem da filogeografia estão interligadas ao desenvolvimento dos estudos empíricos do DNA mitocondrial de animais. Pesquisas conduzidas entre as décadas de 1970 e 1980 desvendaram as principais características moleculares e padrões de transmissão hereditária do mtDNA, que ainda hoje continua sendo um dos principais marcadores de filogenia microevolutiva.

Papel do DNA Mitocondrial nos Estudos Filogeográficos

A introdução do DNA mitocondrial na análise de genética populacional no final da década de 1970 promoveu uma mudança revolucionária do ponto de vista histórico e de perspectivas genealógicas na estrutura populacional intra-específica. Pelo fato de as seqüências de mtDNA animal evoluírem rapidamente em nível de seqüência de nucleotídeos, não mostrarem recombinação intramolecular e exibirem herança materna, elas fornecem dados de haplótipos que podem ser ordenados filogeneticamente dentro de espécies, produzindo uma genealogia gênica (ou filogenia intra-específica) interpretável como um componente matriarcal do *pedigree* de um organismo.

A transmissão do mtDNA em animais constitui uma transferência materna análoga à transferência do sobrenome paterno aos filhos, que ocorre em muitas sociedades humanas. Tantos os filhos quanto as filhas herdam o genótipo mitocondrial de suas mães, mas somente as filhas o transmitem à próxima geração. Por isso, as linhagens de mtDNA refletem “nomes de famílias de matrilineagens” relacionadas, porém diferenciadas por mutações pontuais, dentro de uma espécie. Por isso, suas dinâmicas históricas podem ser interpretadas de acordo com os

modelos teóricos utilizados por demógrafos para analisar a distribuição de sobrenomes nas sociedades humanas.

As propriedades especiais do mtDNA animal proporcionaram a emergência de várias perspectivas não-ortodoxas em microevolução:

- a) Introduziu a noção de que organismos individuais poderiam ser tratados como OTUs (Unidade Taxonômica Operacional) em uma análise genética populacional. Essa sugestão veio do fato de que cada animal apresenta um haplótipo de DNA, caracterizável e herdável intacto, sem recombinação intramolecular, através de seus ancestrais maternais.
- b) Introduziu explicitamente o conceito de filogenia na evolução intra-específica. Antes desses estudos havia uma noção cristalizada de que a filogenia não tinha sentido em nível intra-específico porque, em organismo de reprodução sexuada, linhagens co-específicas são anastomóticas em vez de hierarquicamente ramificadas.
- c) O mtDNA possui modo de transmissão assexuado mesmo entre organismos de reprodução sexuada. Por isso, os históricos evolutivos das matrilinhagens de organismos co-específicos podem ser recuperados pela aplicação de algoritmos baseados em perspectivas filogenéticas.

Filogeografia em Vegetais

A detecção de variação intra-específica filogeograficamente informativa é, provavelmente, o maior problema enfrentado pelos biólogos de plantas interessados em estudos com técnicas filogeográficas. Enquanto os estudos de filogeografia em animais recaem majoritariamente em análises do DNA mitocondrial, as seqüências do DNA mitocondrial vegetal mostraram-se inadequadas para esse tipo de investigação, uma vez que mtDNA de plantas exibe baixa taxa de substituições de nucleotídeos, em níveis tais que locos específicos não contêm variação adequada para gerarem filogenias intra-específicas. Outro problema é que o mtDNA vegetal é propenso à recombinação molecular extensiva, resultando em heteroplasmia

(presença de mais de um tipo de mtDNA em uma mesma célula ou organismo) e atrapalhando a avaliação da variação em nível do genoma como um todo.

Por outro lado, as plantas apresentam outro genoma extranuclear de herança materna, que é o DNA dos cloroplastos (cpDNA). Embora apresente níveis de variação intra-específica inferiores aos do mtDNA de animais, o cpDNA tem se mostrado adequado para estudos filogenéticos em plantas. De fato, virtualmente, todos os estudos publicados, envolvendo filogeografia de plantas, recaem sobre o genoma do cpDNA como fonte de variação genética. Similarmente ao mtDNA, o genoma do cloroplasto pode ser considerado uma unidade simples, não recombinante, de herança. Este genoma apresenta taxas de mutações variáveis entre as suas regiões, com a maioria da variação aparentemente ocorrendo dentro de uma grande região de cópia única, e não dentro de regiões repetidas de cópias invertidas.

Avanços Metodológicos

Progressos nos campos da biologia molecular e na análise computacional contribuíram substancialmente para o desenvolvimento de estudos filogeográficos, dentre os quais se destacam:

- a) Introdução e automação de procedimentos de seqüenciamento, que possibilitaram a análise direta de seqüências particulares de DNA, em vez de analisar sítios de restrição apenas, como foi feito inicialmente.
- b) Amplificação *in vitro* do DNA, via reação em cadeia da polimerase (PCR), possibilitando implementar análises a partir de uma quantidade inicial mínima de DNA molde.
- c) Disponibilização de *primers* universais ou não espécie-específicos, tanto para amplificação de seqüências de mtDNA quanto de cpDNA.
- d) Desenvolvimento de algoritmos filogenéticos que possibilitam extrair informações históricas dos dados moleculares associados a aplicativos computacionais mais amigáveis.

- e) Refinamento e extensão da teoria da coalescência, aplicando-a a populações de estruturas e demografias variadas.

Teoria da Coalescência

O termo coalescência refere-se à “junção das partes que estavam separadas”. No contexto de genética de populações, essas “partes separadas” referem-se a cada uma das duas fitas da molécula de DNA. A “junção” dessas fitas se daria pela investigação, retroativa no tempo, do processo de replicação do material genético. Por isso, a teoria da coalescência está fundamentada em três princípios:

- a) Depois de passado tempo suficientemente longo, todos os genes de um loco de uma determinada população devem descender, cada um deles, de uma única e mesma cópia de um determinado gene presente na população ancestral inicial.
- b) Todas as outras linhagens alélicas extinguiram-se.
- c) A probabilidade de a linhagem existente ter se originado a partir de um gene X ancestral é igual à freqüência inicial de X.

Atualmente, teoria da coalescência é o nome aplicado ao tratamento matemático e estatístico formal da genealogia de genes dentro e entre espécies relacionadas. Assim como a sistemática filogenética, a teoria da coalescência relaciona-se com estruturas de ramificação hierárquicas em árvores evolutivas, embora elas tenham se desenvolvido como disciplinas independentes com raízes históricas separadas em análises micro e macroevolutivas, respectivamente.

Análises genealógicas são bastante adequadas para uma visualização ampla dos padrões de variação nucleotídica em diferentes modelos. Esse tipo de abordagem é especialmente informativo em análises de regiões genômicas que não apresentam recombinação, como o mtDNA dos animais e cpDNA dos vegetais, embora, em princípio, possa ser aplicado a partes dos genomas nucleares não recombinantes também. A comparação entre seqüências homólogas de DNA permite estabelecer padrões de similaridade entre elas. De modo geral, as diferenças observadas nessas comparações contêm informações a respeito da história evolutiva dessas seqüências. Dentre essas informações, destacam-se

aquelas que evidenciam quais as seqüências mais relacionadas entre si e aquelas que fornecem dados a respeito de que ponto no tempo as seqüências coalescem, isto é, onde se localiza sua seqüência ancestral comum mais recente (ACMR).

A comparação entre seqüências homólogas de DNA amostradas de espécies diferentes pode fornecer informações a respeito das relações evolutivas dessas espécies, permitindo muitas vezes a construção de árvores filogenéticas que refletem as relações evolutivas dos grupos em estudo. A comparação entre seqüências amostradas de indivíduos diferentes de uma mesma espécie, por sua vez, fornece informações genealógicas que permitem a construção de árvores gênicas ou alélicas, as quais, muitas vezes, evidenciam o ponto temporal em que se encontra a seqüência ACMR.

A teoria da coalescência pode ser visualizada mais facilmente quando se analisa uma nova mutação que, com o passar do tempo, alcança a fixação em uma população ou espécie. Todos os alelos nesse loco são idênticos por descendência ou coalescem quando se remete ao alelo de cópia única de onde surgiu a mutação.

Por outro lado, é importante destacar que a coalescência em um único ancestral de uma genealogia de matrilihagens, por exemplo, não implica que somente uma fêmea vivia na geração coalescente. Apenas indica que quaisquer outras fêmeas que viveram naquela geração não sobreviveram por meio de descendentes que carregassem suas matrilihagens. Também não se pode afirmar que a coalescência de uma árvore matrilinear indica que outras fêmeas fracassaram em contribuir geneticamente para as gerações posteriores. Genes nucleares de outras fêmeas estão, provavelmente, representados nos descendentes por alelos transmitidos por uma multidão de caminhos não-matrilineares.

Como extensão lógica dos princípios de genética mendeliana e demografia populacional, árvores de genes nucleares intra-específicos podem ser obtidas, em teoria, apenas com pequenas modificações das esperanças dos processos de ramificação e coalescência deduzidos para genes mitocondriais ou de cloroplastos. Por outro lado, várias complicações podem surgir nas tentativas empíricas de recuperar genealogias microevolucionárias de locos autossômicos:

- a) Identificação de locos que apresentem taxas evolutivas apropriadas para a natureza da investigação.
- b) Problemas técnicos relativos ao isolamento de haplótipos a partir de tecidos diplóides.
- c) Possibilidade de recombinação intragênica.

O desenvolvimento recente de métodos laboratoriais que permitem a separação física de duas moléculas homólogas de DNA, amplificadas de um loco diplóide heterozigoto (com posterior seqüenciamento de ambas), promete minimizar a segunda das complicações listadas anteriores. As outras duas são de natureza biológica e só poderão ser superadas quando se conseguir alvejar apenas locos de genes que evoluam rapidamente e sejam livres de recombinação intra-alélica.

Por fim, é importante ressaltar que qualquer árvore de genes representa somente uma minúscula fração da história genética (*pedigree*) de uma espécie. Por isso, deve-se tomar cuidado ao delinear conclusões em nível populacional a partir de dados de árvores genéticas baseados em poucos locos gênicos. Por outro lado, genealogias de genes contêm informações históricas que não podem ser recuperadas pela análise de freqüências alélicas apenas.

Capítulo 2

Diversidade Genética Baseada em Informações Fenotípicas



2.1. Introdução

A diversidade genética entre e dentro de populações encontradas em suas condições naturais, em bancos de germoplasma ou desenvolvidas nos programas de melhoramento genético pode ser predita pelas diferenças entre os valores fenotípicos mensurados em suas unidades (indivíduos, famílias etc.). Na caracterização da diversidade genética das espécies vegetais, animais e de microrganismos, os pesquisadores têm o interesse de agrupar genótipos similares, de maneira que as maiores diferenças ocorram entre os grupos formados. Técnicas multivariadas, como análise discriminante, componentes principais, análise de coordenadas e de agrupamento, podem ser aplicadas nesse tipo de estudo. A adoção de uma, entre as técnicas citadas, varia de acordo com o padrão de resultado desejado e com a informação disponível, seja ela característica morfológica, fisiológica, ecológica ou genético-molecular (DINIZ FILHO, 2000).

Em muitas situações, principalmente aquelas voltadas para fins de melhoramento genético, tem sido comum o estudo da diversidade genética com a finalidade de identificar genitores adequados ao cruzamento, tendo em vista a obtenção de híbridos de maiores efeitos heteróticos, que proporcionem maior segregação em recombinações e possibilitem o aparecimento de transgressivos. Embora o volume de informações genéticas provenientes de marcadores moleculares tenha aumentado em grandes proporções para os estudos de diversidade genética, continua-se a dar ênfase ao estudo da diversidade por meio de características fenotípicas, principalmente de natureza quantitativa. Essas características apresentam, geralmente, distribuição contínua, são determinadas por poligenes de pequenos efeitos e influenciadas pelo ambiente. Entretanto, são de grande interesse, tendo em vista a sua importância econômica e a necessidade de grandes esforços para maximizar o êxito na escolha adequada de combinações híbridas, de modo a não comprometer o sucesso das estratégias de seleção.

A definição de como avaliar a diversidade entre os pares de acessos é dependente do conjunto de informações disponíveis, sejam elas fenotípicas,

genotípicas ou geográficas. Neste capítulo será tratada a estimação de medidas de dissimilaridade entre acessos considerando informações fenotípicas originadas de características de distribuição contínua e discreta, podendo, esta última, ser multicategórica ou binária.

Alguns cuidados iniciais devem ser tomados, como a padronização ou não das observações e a transformação dos dados. Outra questão a ser considerada diz respeito à amostragem dos dados, isto é, se as informações são representativas da(s) entidade(s) a ser(em) analisada(s). Com esses cuidados, pode-se diminuir o risco de se fazerem extrapolações incorretas sobre as relações de similaridade.

Após essas definições iniciais, escolhe-se a medida de distância (dissimilaridade) a ser utilizada na formação da matriz de distância entre pares de indivíduos, para que, posteriormente, possa(m) ser aplicado(s) métodos multivariados capazes de produzir uma estrutura de grupos. Em outras situações, o próprio conjunto de dados originais pode ser analisado, geralmente usando técnicas de projeção em eixos ortogonais, de forma que, a partir da dispersão obtida das entidades, são estabelecidos grupos e suas formações são, então, interpretadas. A análise de grupos implica avaliar a capacidade de alocação ou de discriminação de indivíduos, nos seus respectivos centros de referências (populações), com base nas variáveis avaliadas, bem como formular e testar hipóteses sobre as causas dessa aglomeração ou dispersão. A definição das medidas ou coeficientes de (dis)similaridade a serem utilizadas é baseada no tipo de variável sob análise.

2.2. Medidas de Dissimilaridade

O sucesso de um programa de melhoramento reside na existência de variabilidade na população de trabalho. Melhoristas têm recomendado, para a formação de uma população-base (ou população de melhoramento), o intercruzamento entre cultivares de desempenho superiores e divergentes entre si. Essa divergência pode ser avaliada a partir de características agronômicas, morfológicas, moleculares, entre outras. As informações múltiplas de cada cultivar são expressas em medidas de

dissimilaridade, que representam a diversidade existente no conjunto de acessos estudados.

De maneira geral, estudos da diversidade genética têm sido realizados a partir de informações das seguintes medidas:

- i. Medidas de dissimilaridade obtidas de variáveis quantitativas contínuas ou discretas.
- ii. Medidas de dissimilaridade obtidas de variáveis qualitativas binárias.
- iii. Medidas de dissimilaridade obtidas de variáveis qualitativas multicategóricas.

Nesse ponto, deve ser considerado que uma característica (ou variável) é todo atributo mensurável em uma população, gerando para cada elemento (indivíduo ou família) um determinado valor. Seus valores variam de elemento para elemento e podem assumir grandezas numéricas ou não-numéricas. Existem várias maneiras de classificar as variáveis em diferentes tipos; a mais comum considera dois grandes grupos, chamados de variáveis qualitativas e quantitativas.

As variáveis classificadas como quantitativas são aquelas que podem ser medidas em escala real, podendo ser contínuas ou discretas. As variáveis contínuas são as que assumem, dentro de um intervalo finito, uma infinidade de valores, incluindo inteiros e fracionários. As variáveis discretas podem assumir apenas um número finito ou infinito contável de valores e, assim, são expressas por valores inteiros. As variáveis qualitativas (ou categóricas) são as que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias ou classes, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais. As variáveis nominais são aquelas em que não existe ordenação entre as categorias (exemplo: cores variadas), ao contrário do que ocorre com as ordinais (exemplo: tonalidades). Tanto para a variável do tipo nominal quanto para a ordinal pode-se, por razões de conveniência, associar valores numéricos às diferentes categorias. Contudo, é importante lembrar que esses valores numéricos não têm significado como tal, nem mesmo no caso de variáveis

de tipo ordinal. Por exemplo, pode-se associar os valores 1 e 2 às categorias tardio e precoce da variável floração, ou os valores 1, 2 e 3 às categorias amarela, branca e roxa para cor do bulbo da cebola, por exemplo. Todavia, os números 1 e 2, no primeiro caso, e 1, 2 e 3, no segundo, não são nada mais que símbolos para representar as categorias. Outros valores poderiam ser utilizados, de forma que, por exemplo, ao associar valores 1, 2 e 3 às categorias amarela, branca e roxa não significa que o padrão amarelo seja mais distante do roxo que do branco.

Em algumas situações pode existir o interesse em transformar dados quantitativos em qualitativos ou categóricos. Para este fim, o número de classes pode ser estabelecido arbitrariamente ou estimado considerando algumas regras. Assim, pode-se considerar que existem n observação e pretende-se estabelecer k classes. Uma das maneiras de determinar o número de classes é usando a Regra de Sturges que determina k em função de n , por meio da expressão:

$$k = 1 + 3,322\log(n) \quad \text{por exemplo, se } n=100 \text{ então } k=7,644$$

Outra forma é utilizar a expressão:

$$k = \sqrt{n} \quad \text{por exemplo, se } n=100 \text{ então } k=10$$

Após estabelecido o número de classes, pode-se agrupar o conjunto de dados adotando como critério a divisão da amplitude de variação em intervalos de classes ou a distribuição equitativas dos dados no grupo de classes previamente estabelecido.

Para variáveis de distribuição conhecida como, por exemplo, as que seguem distribuição normal, é também possível subdividir o conjunto de dados em intervalos que levem em consideração a média e o desvio padrão (DP). Assim, se há o interesse em seis classes, poder-se-ia adotar o critério:

Classe 1: [mínimo, média - 2DP]

Classe 2: (média - 2DP, média-1P]

Classe 3: (média - DP, media]

Classe 4: (média , media + DP]

Classe 5: (média + DP, media + 2DP]

Classe 6: (média + 2DP, máximo]

Por fim, deve ser lembrado que uma variável originalmente quantitativa pode ser coletada de forma qualitativa, como, por exemplo, o tamanho de frutos, que, depois de medidos, são classificados em grande, médio ou pequeno. Por outro lado, nem sempre uma variável representada por números é quantitativa, como corre, por exemplo, em relação a sexo, raças etc. Muitas vezes o sexo do indivíduo é registrado na planilha de dados como 1, se macho, e 2, se fêmea, e, dessa forma, analisado.

2.2.1. Variáveis quantitativas contínuas e discretas

As medidas de dissimilaridade são de grande importância em estudos de diversidade genética em que se procura identificar genitores a serem utilizados em programas de hibridação. Há expectativa de que genitores de bom desempenho, com certo grau de diversidade, possam apresentar constituição genética complementar que proporcionariam, na F_1 , maior heterose e, nas gerações segregantes, indivíduos transgressivos.

Em outras situações, estudos sobre a diversidade genética têm sido realizados com o intuito de identificar grupos de cultivares com maior similaridade, visando a formação de multilinhas. Em avaliações de banco de germoplasma, os coeficientes de similaridade evidenciam a existência de duplicatas, as quais poderiam ser eliminadas reduzindo os custos e a mão-de-obra necessários para a conservação dos acessos.

Primeiramente, para que uma medida de distância seja considerada métrica é necessário satisfazer algumas propriedades. Considere os indivíduos (genótipos) i e j e um coeficiente d_{ij} capaz de medir a distância entre eles. Desse modo, têm-se as seguintes propriedades:

(i) $d_{ij} \geq 0$

(ii) $d_{ij} = d_{ji}$

(iii) $d_{ij} = 0$ se $i = i'$

(iv) $d_{ij} \leq d_{ik} + d_{jk}$ (inequação triangular).

Essencialmente, o atendimento às duas últimas propriedades possibilita qualificar a distância como sendo métrica. As medidas para caracteres quantitativos descritas a seguir satisfazem essas propriedades, e as mais utilizadas nos estudos genéticos são: a distância euclidiana, o quadrado da distância euclidiana, a distância euclidiana média, a distância ponderada e a distância generalizada de Mahalanobis. Os valores de distâncias são, geralmente, obtidos a partir de informações de g genótipos mensurados em relação a v caracteres, conforme planilha ilustrada a seguir:

Genótipos	Características				
	1	2	3	...	v
1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1v}
2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2v}
...					
i	Y_{i1}	Y_{i2}	Y_{i3}	...	Y_{iv}
i'	$Y_{i'1}$	$Y_{i'2}$	$Y_{i'3}$...	$Y_{i'v}$
...					
g	Y_{g1}	Y_{g2}	Y_{g3}	...	Y_{gv}

Distâncias Euclidianas

Considerando Y_{ij} a observação no i -ésimo genótipo (clone, cultivar, linhagem etc.) para a j -ésima característica, define-se a distância euclidiana entre o par de genótipos i e i' por meio da expressão:

$$d_{ii'} = \sqrt{\sum_j (Y_{ij} - Y_{i'j})^2}$$

Como ilustração, será considerado o exemplo que envolve a avaliação de três genótipos (P_1 , P_2 e P_3), com relação a dois caracteres (Y_1 e Y_2) não-correlacionados, cujos dados são apresentados na Tabela 2.1.

Tabela 2.1 - Médias de três genótipos avaliados em relação a dois caracteres (Y_1 e Y_2)

Genótipos	Y_1	Y_2
P_1	2	3
P_2	4	4
P_3	5	4

Assim: $d_{12} = 2,24$; $d_{13} = 3,16$; e $d_{23} = 1,00$. A representação gráfica dos genitores no plano definido pelos eixos Y_1 e Y_2 é observada na Figura 2.1.

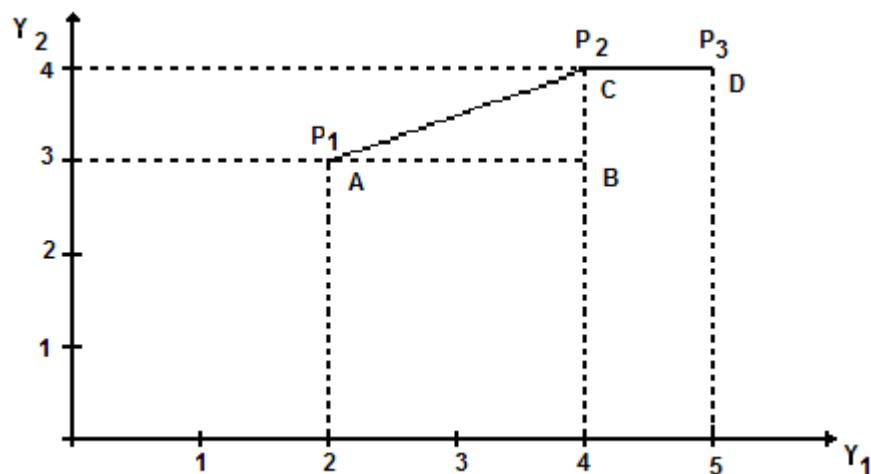


Figura 2.1 - Dispersão dos genótipos P_1 , P_2 e P_3 em relação aos eixos representados pelos caracteres Y_1 e Y_2 .

Logo, a distância euclidiana entre P_1 e P_2 equivale ao comprimento da hipotenusa do triângulo ABC, e entre P_2 e P_3 , ao comprimento do segmento CD.

Distância Euclidiana Média

Como a distância euclidiana sempre aumenta com o acréscimo do número de características consideradas na análise, tem sido usada, de forma alternativa, a distância euclidiana média, dada por:

$$d_{ij} = \sqrt{\frac{1}{v} \sum_j (Y_{ij} - Y_{i'j})^2}$$

sendo v o número de características estudadas.

Ressalta-se que a distância euclidiana foi originalmente proposta para variáveis quantitativas. Portanto, trata-se de uma medida sensível à correlação entre variáveis e, assim, de utilidade restrita a variáveis independentes (DIAS, 1998). Entretanto, como em estudos de melhoramento é praticamente impossível avaliar um conjunto de características não-relacionadas, o uso da distância euclidiana tem sido indiscriminado e mostrado de grande utilidade mesmo nas situações em que a independência entre as características mensuradas não é constatada.

Ilustração

Considere os valores médios da Tabela 2.2, oriundos de sete populações simuladas, com cinco repetições e avaliadas por quatro características quantitativas (Y_1 , Y_2 , Y_3 e Y_4).

Tabela 2.2 - Valores médios originais de sete populações avaliadas por quatro características quantitativas

Populações	Y_1	Y_2	Y_3	Y_4
1	0,21*	196,24	26,44	2889,53
2	0,21	197,09	28,45	3616,27
3	0,21	186,68	26,13	3017,48
4	0,18	176,32	23,95	4392,71
5	0,23	181,81	25,15	3023,84
6	0,22	173,38	26,69	2721,67
7	0,23	172,80	21,60	4167,86

* Arredondamento para duas casas decimais e observações não-padronizadas.

Por exemplo, as distâncias euclidianas médias entre as populações 1 e 2 e 1 e 7 são, respectivamente, as seguintes:

$$\bar{d}_{12} = \sqrt{\frac{1}{4} \sum_{j=1}^4 (Y_{1j} - Y_{2j})^2} = \sqrt{\frac{1}{4} (0,21 - 0,21)^2 + \dots + (2889,53 - 3616,27)^2} = 363,37$$

$$\bar{d}_{17} = \sqrt{\frac{1}{4} \sum_{j=1}^4 (Y_{1j} - Y_{7j})^2} = \sqrt{\frac{1}{4} (0,21 - 0,23)^2 + \dots + (2889,53 - 4167,68)^2} = 639,28$$

O cálculo para as demais distâncias entre os pares de populações segue o mesmo raciocínio. Todos os valores dessas distâncias estarão dispostos numa matriz simétrica, denominada matriz de distância, que será utilizada nas análises de agrupamento.

	1	2	3	4	5	6	7
1	0,00	363,37	64,15	751,66	67,54	84,70	639,28
2	363,37	0,00	299,44	388,37	296,32	447,45	276,09
3	64,15	299,44	0,00	687,64	4,03	148,05	575,24
4	751,66	388,37	687,64	0,00	684,44	835,52	112,44
5	67,54	296,32	4,03	684,44	0,00	151,14	572,03
6	84,70	447,45	148,05	835,52	151,14	0,00	723,10
7	639,28	276,09	575,24	112,44	572,03	723,10	0,00

Quadrado da Distância Euclidiana Média

Outra forma de expressar a dissimilaridade entre dois genótipos ou populações, quando se avaliam v características quantitativas, é por meio do quadrado da distância euclidiana média, fornecido por:

$$d_{ij}^2 = \frac{1}{v} \sum_j (Y_{ij} - \bar{Y}_{ij})^2$$

Esta medida é algumas vezes preferida, por manter relação com a soma de quadrados de desvio, ou seja:

$$SQD_{ii'} = \frac{v}{2} d_{ii'}^2$$

em que:

$$SQD_{ii'} = \sum_{j=1}^v SQD_{j(ii')}$$

sendo $SQD_{j(ii')}$ a soma de quadrados dos desvios, para a j-ésima variável, considerando os acessos i e i' .

Padronização dos dados

Em todas as distâncias até então citadas, a escala afeta o valor obtido. Adicionalmente, elas são quantificadas em diferentes medidas (peso, comprimento, porcentagens etc.), sendo, nesses casos, recomendável o cálculo das distâncias utilizando-se os valores padronizados, feito por meio de:

$$y_j = \frac{Y_j}{\hat{\sigma}_j}$$

em que $\hat{\sigma}_j$ é o desvio-padrão associado à j-ésima característica.

Existem duas razões para a padronização dos dados. A primeira visa evitar que as unidades escolhidas para medir as variáveis afetem arbitrariamente a similaridade entre os indivíduos. A segunda é que a padronização faz com que as variáveis contribuam igualmente na avaliação da similaridade entre indivíduos.

O efeito da escala é ilustrado no exemplo a seguir, em que se consideram duas características (Y_1 e Y_2) mensuradas em três indivíduos. Numa primeira situação considera-se que as unidades de ambas as variáveis são representadas por metros (m). Numa outra situação, a característica Y_2 é mensurada em decímetro (dm). Percebe-se que a modificação de escala muda as conclusões sobre a diversidade, alterando a magnitude das medidas de dissimilaridade e, principalmente, a conclusão a respeito da similaridade entre pares de acessos.

Indivíduos	$Y_1(m)$	$Y_2(m)$	$Y_1(m)$	$Y_2(dm)$
I_1	2	4	2	40
I_2	2	6	2	60
I_3	4	6	4	60

Assim, tem-se com o primeiro conjunto de dados:

$$d_{12} = d_{23} = 2$$

mas, para o segundo conjunto de dados, obtém-se:

$$d_{12} = 20 \quad \text{e} \quad d_{23} = 2$$

O efeito da grandeza e amplitude de variação entre as características também é ilustrado no exemplo a seguir, em que se consideram duas características (Y_1 e Y_2), representativas do peso de frutos e altura da planta, mensuradas em três indivíduos. Assim, a importância da variação entre alturas de plantas para a diversidade genética torna-se mínima diante das variações de peso. Os genótipos 1 e 3, com diferença de 20 gramas, destacam-se como mais divergentes do que os genótipos 1 e 3, cuja variação em altura é enorme. Após a padronização, os resultados demonstram equivalência na diversidade entre esses pares de genótipos.

Indivíduos	Dados originais		Dados padronizados	
	Peso(g)	Altura(m)	Peso	Altura
I_1	200	1,0	20	0,8660
I_2	190	3,0	19	2,5981
I_3	180	1,0	18	0,8660
$\bar{\sigma}_j$	10,0	1,1547	1,0	1,0

Assim, tem-se com o primeiro conjunto de dados:

$$d_{12} = 10,44 \quad \text{e} \quad d_{13} = 20,00$$

mas, para o segundo conjunto de dados, tem-se:

$$d_{12} = 2,00 \quad \text{e} \quad d_{13} = 2,00$$

Uma vez definidas as variáveis, espera-se que a contribuição delas seja equivalente numa análise de agrupamento. Embora uma variável, considerada importante, deva apresentar grande variação e não ser redundante, espera-se que a distância entre dois indivíduos não seja alterada com a adoção de outras unidades de medida para um mesmo conjunto de dados e que todas as variáveis escolhidas apresentem poder discriminatório que não seja baseado na amplitude de seus valores.

Se a amplitude dos valores de uma variável for muito maior do que a outra, então a primeira terá maior peso na análise. Na ilustração com os dados da Tabela 2.2, a contribuição de cada variável no cálculo das distâncias foi bastante discrepante. A variável x_4 apresenta contribuição relativa para a divergência genética entre os genótipos de 99,98%, enquanto as demais variáveis contribuem com apenas 0,02%.

Nos estudos de diversidade genética em que são avaliadas características quantitativas têm sido comuns dois tipos de função de padronização. Pode-se adotar a estratégia de modo que a média seja zero e a variância igual a um [$y_{ij} = (Y_{ij} - \bar{Y}) / \hat{\sigma}_j$], ou apenas para variância unitária ($y_{ij} = Y_{ij} / \hat{\sigma}_j$), em que \bar{Y}_j e $\hat{\sigma}_j$ são a média e o desvio-padrão, respectivamente, da variável j . É necessário ter o cuidado para que a padronização possa diluir as diferenças entre potenciais grupos delineados pelas variáveis originais (EVERITT, 1993).

A Tabela 2.3 apresenta os valores médios padronizados obtidos dos dados originais apresentados na Tabela 2.2. As distâncias euclidianas médias padronizadas encontram-se na Tabela 2.4.

Tabela 2.3 - Valores médios padronizados[#] de sete populações avaliadas por quatro características quantitativas

Populações	y_1	y_2	y_3	y_4
1	11,91*	19,17	12,01	4,36
2	11,58	19,26	12,92	5,46
3	11,69	18,24	11,86	4,56
4	9,89	17,23	10,88	6,64
5	12,70	17,76	11,42	4,57
6	12,36	16,94	12,12	4,11
7	12,81	16,88	9,81	6,30

* Arredondamento para duas casas decimais.

[#] Fórmula para padronização: $y_{ij} = Y_{ij} / \hat{\sigma}_j$.

Outra medida de dissimilaridade pode ser estabelecida a partir do quadrado da distância euclidiana dos valores médios padronizados da Tabela 2.3, originando a matriz D_{7x7} .

$$D_{7x7} = \begin{array}{ccccccccc} & & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{matrix} & \left[\begin{array}{cccccc} 0,00 & & & & & & \\ 2,16 & 0,00 & & & & & \\ 0,98 & 2,99 & 0,00 & & & & \text{Simétrica} \\ 14,32 & 12,53 & 9,55 & 0,00 & & & \\ 3,00 & 6,55 & 1,44 & 12,75 & 0,00 & & \\ 5,25 & 8,45 & 2,41 & 14,12 & 1,49 & 0,00 & \\ 14,66 & 17,56 & 10,33 & 9,91 & 6,37 & 10,34 & 0,00 \end{array} \right] \end{array}$$

Deve-se estar atento na análise de agrupamento, pois diferentes grupos podem ser formados apenas em função da padronização dos dados, devido à mudança das estimativas de distâncias. Por ser um método de descrição, na análise de agrupamento a transformação de dados tem tido menor relevância do que em uma análise de variância multivariada (MANOVA), pois a MANOVA é um método de inferência no qual a validação depende da normalidade do conjunto de dados (ROMESBURG, 1984).

Distância Ponderada

Neste caso, considera-se, no cálculo da distância euclidiana, a diferença de precisão de cada variável mensurada, de forma que aquelas que foram medidas com menor precisão venham contribuir menos para o valor da diversidade entre pares de acessos. Para o cálculo dessa distância, é necessário dispor da informação da variação residual, obtida em análise prévia de variância em modelos estatísticos apropriados, de forma que a distância possa ser calculada da seguinte maneira:

$$d_{ii'}^2 = \frac{d_1^2}{\hat{\sigma}_1^2} + \frac{d_2^2}{\hat{\sigma}_2^2} + \dots + \frac{d_v^2}{\hat{\sigma}_v^2} = \sum_{j=1}^v \frac{d_j^2}{\hat{\sigma}_j^2}$$

ou

$$d_{ii'}^2 = \delta' \Delta^{-1} \delta$$

em que:

$d_{ii'}^2$: distância ponderada entre os genótipos i e i' ;

Δ : matriz diagonal cujos elementos são as variâncias residuais;

$\delta' = [d_1 \ d_2 \ \dots \ d_v]$, sendo $d_j = Y_{ij} - Y_{i'j}$;

$\hat{\sigma}_j^2$: quadrado médio do resíduo associado à j -ésima variável;

e

Y_{ij} : média do i -ésimo genótipo em relação à j -ésima variável.

Distância de Gower

Esta é uma medida que pode ser utilizada para obtenção da dissimilaridade entre indivíduos a partir da análise conjunta de variáveis quantitativas e qualitativas. A expressão geral é:

$$S_{ijk} = \frac{\sum_{k=1}^v w_{ijk} C_{ijk}}{\sum_{k=1}^v w_{ijk}}$$

em que

S_{ijk} : similaridade entre os indivíduos i e j, para a variável k;

w_{ijk} : peso dado à comparação entre os indivíduos i e j, para a variável k, atribuindo valor 1 para as comparações válidas e 0 para as comparações inválidas pela falta de informação em pelo menos um indivíduo do par de comparação;

C_{ijk} : contribuição da variável k na similaridade entre os indivíduos i e j.

Assim, para um conjunto de variáveis quantitativas contínuas, a dissimilaridade pela estatística de Gower é calculada por meio de:

$$d_{ij} = \frac{1}{v} \sum_{j=1}^v \frac{|Y_{ij} - Y_{i'j}|}{R_j}$$

em que:

R_j : amplitude de variação verificada na j-ésima característica.

Esta medida de distância varia de 0 a 1.

Distância Generalizada de Mahalanobis

Uma crítica que se faz à distância euclidiana é o fato de ela não levar em consideração as variâncias e, também, as covariâncias residuais que existem entre as características mensuradas, possíveis de serem quantificadas quando as avaliações são realizadas em genótipos avaliados em delineamentos experimentais.

Considerando novamente a Figura 2.1, verifica-se que, em função de Y_1 e Y_2 serem independentes, os eixos são perpendiculares, e a distância de segmentos de retas ou as hipotenusas de triângulos retângulos correspondem à distância euclidiana. Entretanto, como acontece na maioria das vezes, os caracteres estudados apresentam certo grau de correlação, não sendo totalmente independentes. Dessa forma, os eixos dos gráficos de dispersão são oblíquos, e a estimativa da diversidade genética por meio da distância euclidiana torna-se inadequada.

Se as variáveis Y_1 e Y_2 , apresentadas na Figura 2.1, são correlacionadas, a representação gráfica apropriada é a da Figura 2.2. Nesta situação, as estimativas

das distâncias genotípicas são mais fáceis de serem obtidas em relação aos eixos Z_1 e Z_2 do que em relação aos eixos Y_1 e Y_2 .

Os valores de Z_{i1} e Z_{i2} ($i = 1, 2, \dots, n$; $n =$ número de genitores), representados na Figura 2.2, são obtidos por transformações lineares de Y_{i1} e Y_{i2} . Estabelecendo nesta figura que o eixo representado por Z_1 coincide com Y_1 , a transformação é dada por:

$$Z_{i1} = Y_{i2} \cos \theta + Y_{i1}$$

$$Z_{i2} = Y_{i2} \sin \theta$$

Assim, por exemplo, a coordenada do genótipo P_i , dada por (Y_{i1}, Y_{i2}) , deve ser substituída por (Z_{i1}, Z_{i2}) , que equivale a $(Y_{i2} \cos(\theta) + Y_{i1}, Y_{i2} \sin(\theta))$. Após transformadas as coordenadas de cada genótipo, a distância generalizada de Mahalanobis (MAHALANOBIS, 1936), denominada D^2 , é estimada por:

$$D^2 = \sum_j (Z_{ij} - Z_{i'j})^2$$

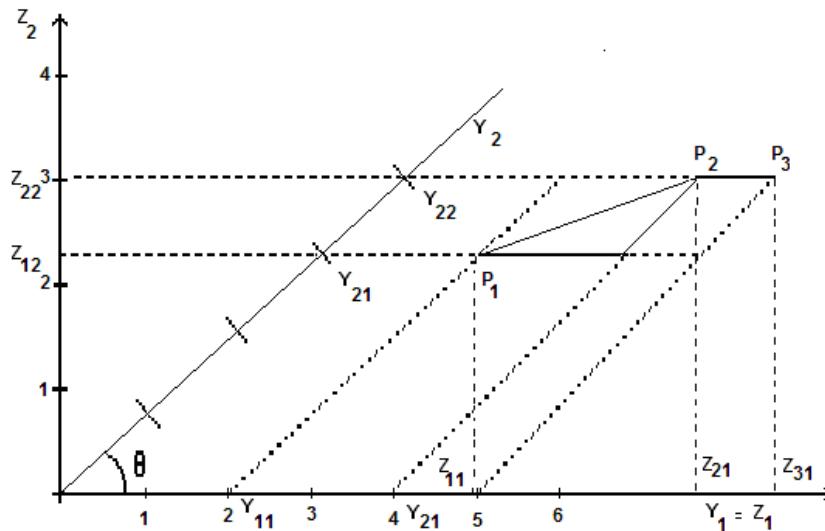


Figura 2.2 - Dispersão dos genótipos P_1 , P_2 e P_3 em relação aos eixos oblíquos representados pelos caracteres Y_1 e Y_2 .

Dessa forma, fica evidenciado que a distância D^2 tem a vantagem, em relação à distância euclidiana, de levar em consideração a correlação entre os caracteres considerados. Quando se dispõe de vários caracteres, o valor de D^2 pode ser, alternativamente, estimado a partir das médias dos dados originais e da matriz de covariâncias residuais (matriz de dispersão), ou a partir dos dados transformados. Essa transformação é equivalente ao processo de rotação de eixos.

Na transformação, são obtidas combinações lineares das variáveis originais, cujos coeficientes de ponderação são derivados de um processo denominado condensação pivotal, que é aplicado na matriz de dispersão. Este processo evita a inversão de matrizes, que, quando de ordem elevada, proporcionam grandes erros numéricos e tornam os cálculos, à semelhança das distâncias euclidianas, mais simples.

Quando se dispõe de informações provenientes de ensaios experimentais é possível obter a matriz de dispersão residual (Ψ) e as médias das características. De posse dessas informações, obtém-se as estimativas das distâncias de Mahalanobis por meio da expressão:

$$D_{ii'}^2 = \delta' \Psi^{-1} \delta$$

em que:

$D_{ii'}^2$: distância de Mahalanobis entre os genótipos i e i' ;

Ψ : matriz de variâncias e covariâncias residuais;

$\delta' = [d_1 \ d_2 \ \dots \ d_v]$, sendo $d_j = Y_{ij} - Y_{i'j}$;

e

Y_{ij} : média do i -ésimo genótipo em relação à j -ésima variável.

Para obter as estimativas dos quadrados médios e dos produtos médios do resíduo, entre dois caracteres Y_1 e Y_2 , recomenda-se fazer as análises de variâncias individuais, segundo um modelo estatístico apropriado, e a análise da soma dos valores de X e Y , de forma que os produtos médios (covariâncias), associados a cada fonte de variação, incluindo a do resíduo, possam ser estimados por meio de:

$$\text{Cov}(Y_1, Y_2) = \frac{V(Y_1+Y_2) - V(Y_1) - V(Y_2)}{2}$$

Considerando os caracteres Y_{1ik} e Y_{2ik} , medidos em g genótipos ou tratamentos ($i = 1, 2, \dots, g$), avaliados em blocos ao acaso com r repetições ($k = 1, 2, \dots, r$), tem-se o esquema da análise de variância apresentado a seguir:

Esquema da análise de variância dos caracteres Y_1 , Y_2 e da soma Y_1+Y_2 , para o experimento em blocos casualizados

FV	GL	QM		
		Y_1	Y_2	Y_1+Y_2
Blocos	r-1			
Tratamentos	g-1	QMT_{Y_1}	QMT_{Y_2}	$QMT_{Y_1+Y_2}$
Resíduo	(r-1)(g-1)	QMR_{Y_1}	QMR_{Y_2}	$QMR_{Y_1+Y_2}$

Os produtos médios associados a tratamentos e resíduo são obtidos por meio das expressões:

$$PMT_{Y_1,Y_2} = (QMT_{Y_1+Y_2} - QMT_{Y_1} - QMT_{Y_2})/2$$

e

$$PMR_{Y_1,Y_2} = (QMR_{Y_1+Y_2} - QMR_{Y_1} - QMR_{Y_2})/2$$

Assim, os elementos da diagonal de ψ são os quadrados médios do resíduo e os elementos fora da diagonal, os produtos médios do resíduo, entre os pares de caracteres mensurados.

Além de possibilitar o estudo da diversidade genética, é possível, por meio das distâncias generalizadas de Mahalanobis, quantificar a contribuição relativa dos caracteres para a divergência genética utilizando o critério proposto por Singh (1981), baseado na estatística $S_{.j}$. Neste caso, considera-se que:

$$D_{ii'}^2 = \delta' \psi^{-1} \delta = \sum_{j=1}^n \sum_{j'=1}^n \omega_{jj'} d_j d_{j'}$$

em que $\omega_{jj'}$ é o elemento da j-ésima linha e j'-ésima coluna da inversa da matriz de variâncias e covariâncias residuais.

No cálculo dos valores da distância generalizada de Mahalanobis, os n^2 valores podem ser arranjados em uma tabela de dupla entrada. Por exemplo, com $n = 3$, os componentes de D^2 podem ser assim arranjados:

$\omega_{11}d_1d_1$	$\omega_{12}d_1d_2$	$\omega_{13}d_1d_3$
$\omega_{21}d_2d_1$	$\omega_{22}d_2d_2$	$\omega_{23}d_2d_3$
$\omega_{31}d_3d_1$	$\omega_{32}d_3d_2$	$\omega_{33}d_3d_3$
$Sm_1 = \sum_{i=1}^3 \omega_{i1}d_id_i$	$Sm_2 = \sum_{i=1}^3 \omega_{i2}d_id_i$	$Sm_3 = \sum_{i=1}^3 \omega_{i3}d_id_i$

Assim, Sm_1 representa a contribuição da variável 1 para o valor de uma determinada distância D_{ij}^2 representada por D_m^2 . O total das distâncias envolvendo todos os pares de genótipos é dado por:

$$\sum_{i < i'} \sum D_{ii'}^2 = \sum_m D_m^2 = \sum S_j$$

Os valores percentuais de $S_{.j}$ constituem a medida da importância relativa da variável j para o estudo da diversidade genética. Deve ser ressaltado que, se a matriz ψ fosse estabelecida apenas pelos seus elementos diagonais, ou seja, as covariâncias residuais fossem todas nulas, o valor de S_{mj} seria diretamente proporcional à estatística F. Assim, nessa situação é verificado que para uma determinada variável (Y) tem-se:

$$SQD = \sum_{i=1}^g Y_i^2 - \frac{1}{g} \left(\sum_{i=1}^g Y_i \right)^2 = \frac{g-1}{g} \sum_{i=1}^g Y_i^2 - \frac{2}{g} \sum_{i < i'} \sum_{i'} Y_i Y_{i'}$$

$$\sum_{i < i'} \sum_{i'} d_{ii'}^2 = (Y_1 - Y_2)^2 + \dots + (Y_{g-1} - Y_g)^2 = (g-1) \sum_{i=1}^g Y_i^2 - 2 \sum_{i < i'} \sum_{i'} Y_i Y_{i'}$$

portanto,

$$SQD = \frac{1}{g} \sum_{i < i'} \sum_{i'} d_{ii'}^2$$

e

$$S_{mj} = \frac{\sum_{i<}^g \sum_{i'}^g d_{ii'}^2}{\hat{\sigma}^2} = \frac{g(g-1)}{r} \frac{QMG_j}{QMR_j} = \frac{g(g-1)}{r} F_j$$

em que:

QMG_j : quadrado médio de genótipos associado à j -ésima característica;

QMR_j : quadrado médio do resíduo associado à j -ésima característica; e

$F_j = QMG_j/QMR_j$: estatística F associada a $g-1$ e $(g-1)(r-1)$ graus de liberdade, em experimentos conduzidos em blocos ao acaso com r repetições.

Ilustração

Considere os dados da Tabela 2.2 para o cálculo da distância de Mahalanobis e a importância relativa das características. Embora as três primeiras características não tenham apresentado diferenças significativas ($P < 0,05$) entre as médias dos genótipos, as distâncias de Mahalanobis serão calculadas a título de ilustração. Para o exemplo em consideração, tem-se a seguinte matriz de variâncias e covariâncias:

$$\Psi = \begin{bmatrix} 0,000954 & 0,124148 & -0,02089 & 5,039563 \\ 0,124148 & 224,3381 & -6,44213 & -1557,19 \\ -0,02089 & -6,44213 & 7,468046 & -264,021 \\ 5,039563 & -1557,19 & -264,021 & 472992,4 \end{bmatrix}$$

As estimativas das distâncias generalizadas de Mahalanobis são obtidas a partir dos valores médios originais (Tabela 2.2) e da matriz de dispersão Ψ . Para os genótipos 1 e 2, tem-se:

$$D_{12}^2 = \delta' \psi^{-1} \delta$$

em que:

$$\delta = \begin{bmatrix} 0,21 - 0,21 \\ 196,24 - 197,09 \\ 26,44 - 28,45 \\ 2889,53 - 3616,27 \end{bmatrix}$$

Assim, obtém-se $D_{12}^2 = 2,22$. As demais estimativas das distâncias de Mahalanobis e distâncias euclidianas médias padronizadas são apresentadas na Tabela 2.4.

Tabela 2.4 - Estimativas das distâncias de Mahalanobis (parte superior) e distâncias Euclidianas médias padronizadas (parte inferior) entre os sete genótipos simulados

Populações	1	2	3	4	5	6	7
1		2,22	0,46	9,20	1,76	3,33	8,39
2	0,74		2,94	7,47	5,70	7,87	10,82
3	0,50	0,86		7,20	0,74	1,86	5,82
4	1,89	1,77	1,54		9,35	12,82	4,63
5	0,86	1,28	0,60	1,79		0,79	4,44
6	1,15	1,45	0,78	1,88	0,61		7,33
7	1,91	2,09	1,61	1,57	1,26	1,61	

A maior distância de Mahalanobis foi verificada entre os genótipos 4 e 6 (12,86), e a menor, entre os genótipos 1 e 3 (0,46). Para as distâncias euclidianas médias, a maior foi entre os genótipos 2 e 7 (2,09), e a menor, entre os genótipos 1 e 3 (0,50). A correlação entre as estimativas de distâncias foi igual a 0,9398.

Recomenda-se a utilização da distância generalizada de Mahalanobis (D^2) em estudos sobre divergência genética, cujas características são avaliadas com grau significativo de correlações entre si e dispõe-se de informações das variâncias e covariâncias residuais,

A importância relativa das características está descrita na Tabela 2.5.

Tabela 2.5 - Importância relativa das características para a diversidade genética estabelecida entre sete populações

Características	S.j	Valor em %
X ₁	20,22*	17,56
X ₂	23,08	20,04
X ₃	29,09	25,26
X ₄	42,76	37,14

* Cálculo feito com médias não-padrонizadas.

Aplicação geral

Será considerada a avaliação de 15 cultivares de milho-pipoca em blocos ao acaso com três repetições. Foram estudados os caracteres altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela). Os dados experimentais encontram-se descritos na Tabela A1 (Anexos).

Pela análise de variância (Tabela 2.6), constata-se que existem diferenças significativas entre as médias dos cultivares para todas as nove características avaliadas. O coeficiente de variação variou de 10,19 a 13,34%, demonstrando existir boa precisão experimental na avaliação de todas as características consideradas.

Tabela 2.6 - Resultado da análise de variância de nove características agronômicas avaliadas em cultivares de milho-pipoca

Características	QMB	QMT	QMR	Média	CV(%)
AP	0,0257	1,0174**	0,0440	2,06	10,19
AE	0,0013	0,1659**	0,0193	1,06	13,10
NE	11,6515	122,9208**	7,5072	24,14	11,35
PROL	0,1122	0,2422**	0,0089	0,88	10,71
CE	10,7978	250,0928**	7,7418	25,46	10,93
QUE	0,0004	0,0267**	0,0007	0,20	13,34
DOEN	0,0002	0,0036**	0,0002	0,11	12,28
PCG	16,1842	203,8536**	8,8635	22,34	13,33
PROD	458531,2889	2467828,1551**	149133,5746	2992,02	12,90
Graus de liberdade	2	14	28		

Na Tabela 2.7 podem ser observadas as diferenças entre as médias obtidas para as características avaliadas, indicando que a variabilidade genética entre esses cultivares é de grande magnitude. Pretende-se, por meio de técnicas multivariadas, quantificar essa dissimilaridade de modo que seja possível indicar possíveis cruzamentos entre os genitores e reconhecer o padrão de similaridade entre eles.

Serão adotadas duas medidas de dissimilaridade: distância euclidiana média padronizada e distância de Mahalanobis. Para o primeiro caso, ignora-se a existência das informações em nível de repetições, tomando-se apenas os valores padronizados das médias, como descritos na Tabela 2.8. Para cálculo das distâncias de Mahalanobis, utilizam-se as médias originais e as estimativas das variâncias e covariâncias residuais entre os caracteres estudados. Os valores das variâncias residuais correspondem àqueles apresentados na Tabela 2.6, porém as covariâncias requerem novas análises, de forma a serem estimados os produtos médios do resíduo (covariâncias residuais). Neste caso,

como mencionado anteriormente, utiliza-se o artifício de considerar somas de pares de caracteres, que são submetidos à análise de variância; neste caso, o produto médio pode ser obtido pela expressão:

$$PM_{Y_1,Y_2} = [QM(Y_1+Y_2) - QMY_1 - QMY_2]/2$$

sendo:

PM_{Y_1,Y_2} : produto médio entre dois caracteres (Y_1 e Y_2); e

$QM(Y_1+Y_2)$, QMY_1 e QMY_2 : quadrados médios obtidos da análise de variância da soma $Y_1 + Y_2$, de Y_1 e de Y_2 , respectivamente.

Tabela 2.7 - Médias de nove características agronômicas avaliadas em 15 cultivares de milho-pipoca

Médias originais									
AP	AE	NE	PROL	CE	QUE	DOEN	PCG	PROD	
1,82 defgh	1,20 abcd	33,84 ab	1,06 bc	14,17 d	0,32 ab	0,19 a	22,01 bc	3339,00 abc	
2,90 ab	1,43 ab	24,46 cde	1,06 bc	14,88 d	0,27 bc	0,15 ab	21,86 c	3950,33 abc	
2,52 bc	1,16 abcd	29,81 bc	1,60 a	13,63 d	0,23 cd	0,15 ab	17,87 c	3512,67 abc	
2,02 cdefg	0,97 cde	24,94 cde	1,08 bc	11,96 d	0,33 ab	0,16 ab	16,40 c	2845,33 cde	
3,48 a	1,53 a	39,37 a	1,18 b	13,71 d	0,38 a	0,10 cd	21,93 c	3022,67 bcd	
1,39 gh	0,83 cde	23,94 cdef	0,73 defg	25,88 c	0,11 ef	0,08 d	14,08 c	2142,00 de	
1,62 fgh	0,67 e	20,64 def	0,94 bcd	35,66 ab	0,12 ef	0,09 cd	14,63 c	1806,33 e	
1,70 efg	0,86 cde	27,32 bcd	0,70 defg	31,26 abc	0,09 f	0,08 d	14,73 c	1981,33 de	
1,35 h	0,87 cde	20,51 def	0,90 bcde	34,62 ab	0,09 f	0,09 cd	14,03 c	1734,67 e	
1,60 fgh	1,11 bcd	24,74 cde	0,84 cdef	28,68 bc	0,10 ef	0,10 cd	13,60 c	1818,33 e	
1,71 efg	1,05 bcde	19,87 def	0,62 efg	28,84 bc	0,20 cd	0,13 bc	36,17 a	4105,67 ab	
2,04 cdef	1,25 abc	18,84 ef	0,63 efg	28,92 bc	0,26 bc	0,08 d	33,63 a	2901,33 cde	
2,25 cde	1,08 bcde	15,76 f	0,60 fg	37,75 a	0,17 def	0,09 cd	31,77 a	4343,67 a	
2,39 bcd	0,81 de	16,80 ef	0,72 defg	30,65 abc	0,21 cd	0,10 cd	31,38 a	3565,33 abc	
2,08 cdef	1,09 bcde	21,27 def	0,53 g	31,29 abc	0,18 de	0,10 cd	31,00 ab	3811,67 abc	
0,5824*	0,2359	6,4005	0,2841	9,1306	0,0937	0,0345	8,2426	906,9794	

(*) Desvio-padrão. Médias seguidas pelas mesmas letras não diferem entre si pelo teste Tukey (5%)

Depois de realizadas as análises que permitem a obtenção das estimativas das covariâncias residuais, obtém-se a matriz de dispersão, cujos elementos da diagonal são os quadrados médios do resíduo, e fora da diagonal, os produtos médios do resíduo, entre cada par de caracteres. Para o exemplo em consideração, tem-se a seguinte matriz:

$$\psi = \begin{bmatrix} 0,0440 & -0,0018 & 0,0271 & 0,0029 & -0,0259 & -0,0011 & 0,0004 & 0,0917 & 25,0480 \\ & 0,0193 & 0,0001 & -0,0033 & -0,0391 & -0,0004 & 0,0001 & -0,0713 & 13,3990 \\ & & 7,5072 & 0,0694 & 0,6516 & 0,0243 & -0,0034 & 1,8767 & -235,4233 \\ & & & 0,0089 & -0,0440 & 0,0001 & 0,0001 & 0,0243 & -6,3108 \\ & & & & 7,7418 & 0,0081 & -0,0030 & -0,3478 & -196,3793 \\ & & & & & 0,0007 & -0,0001 & -0,0003 & -0,7331 \\ & & & & & & 0,0002 & 0,0067 & 1,2645 \\ & & & & & & & 8,8635 & -265,0569 \\ & & & & & & & & 149133,5746 \end{bmatrix}$$

Tabela 2.8. Médias padronizadas de nove características agronômicas avaliadas em 15 cultivares de milho pipoca.

Cultivares	Características (dados padronizados)								
	AP	AE	NE	PROL	CE	QUE	DOEN	PCG	PROD
1	3,125	5,087	5,287	3,731	1,552	3,414	5,501	2,670	3,681
2	4,979	6,062	3,822	3,731	1,630	2,881	4,343	2,652	4,355
3	4,327	4,917	4,657	5,631	1,493	2,454	4,343	2,168	3,873
4	3,468	4,112	3,897	3,801	1,310	3,521	4,633	1,990	3,137
5	5,975	6,486	6,151	4,153	1,502	4,054	2,895	2,661	3,333
6	2,387	3,518	3,740	2,569	2,834	1,174	2,316	1,708	2,362
7	2,782	2,840	3,225	3,308	3,906	1,280	2,606	1,708	1,992
8	2,919	3,646	4,268	2,464	3,424	0,960	2,316	1,787	2,185
9	2,318	3,688	3,204	3,167	3,792	0,960	2,606	1,702	1,913
10	2,747	4,705	3,865	2,956	3,141	1,067	2,895	1,650	2,005
11	2,936	4,451	3,104	2,182	3,159	2,134	3,764	4,388	4,527
12	3,503	5,299	2,944	2,217	3,167	2,774	2,316	4,080	3,199
13	3,863	4,578	2,462	2,112	4,134	1,814	2,606	3,854	4,789
14	4,104	3,434	2,625	2,534	3,357	2,240	2,895	3,807	3,931
15	3,571	4,621	3,323	1,865	3,427	1,920	2,895	3,761	4,203

Para cálculo da distância euclidiana média padronizada, utilizam-se os valores padronizados apresentados na Tabela 2.8. Considerando os cultivares 1 e 2, tem-se a seguinte estimativa:

$$d_{12} = \sqrt{\frac{1}{9} \sum_j (y_{1j} - y_{2j})^2} = \sqrt{\frac{1}{9} [(3,125 - 4,979)^2 + \dots + (3,681 - 4,355)^2]} = 0,98$$

As estimativas das distâncias generalizadas de Mahalanobis são obtidas a partir dos valores médios apresentados na Tabela 2.8 e a matriz de dispersão Ψ . Considerando os cultivares 1 e 2, tem-se a seguinte estimativa:

$$D_{12}^2 = \delta' \Psi^{-1} \delta$$

sendo:

$$\delta = \begin{bmatrix} 1,82 - 2,90 \\ 1,20 - 1,43 \\ 33,84 - 24,46 \\ 1,06 - 1,06 \\ 14,17 - 14,88 \\ 0,32 - 0,27 \\ 0,19 - 0,15 \\ 22,01 - 21,86 \\ 333900 - 395033 \end{bmatrix}$$

Fazendo as operações matriciais indicadas, obtém-se $D_{12}^2 = 66,0$.

As demais estimativas das distâncias de Mahalanobis e euclidiana média padronizada são apresentadas na Tabela 2.9.

Associação entre matrizes de dissimilaridade

A medida de associação entre os valores de duas matrizes de dissimilaridade pode ser obtida por meio do coeficiente de correlação de Pearson, porém deve ser observado que os valores de cada matriz não são independentes e, portanto, um teste de significância apropriado deve ser utilizado. Nessas situações, tem sido recomendado o uso do teste de significância de Mantel, que se baseia num processo de aleatorização pelo qual várias estimativas de correlações são obtidas. A partir das informações de uma das matrizes a serem correlacionadas, promove-se previamente o embaralhamento por critério de permutação aleatória, ou seja, linhas e colunas da matriz são permutadas.

As várias estimativas de correlações, obtidas após cada processo de permutação, são ordenadas e os níveis críticos de significância, a 1 e 5% de

probabilidade, são identificados, considerando os valores-limite das estimativas após excluídos os valores extremos desse conjunto de valores. Assim, como ilustração, considerando 5.000 permutações ($p = 5.000$), para determinação do nível crítico a 1% de probabilidade deverão ser excluídas as 50 estimativas dos extremos (25 maiores e 25 menores estimativas de correlação), de forma que o valor da correlação será significativo quando inferior ao correspondente à ordem 26^a ou superior ao correspondente à ordem 4.975^a desse conjunto de estimativas.

Para as distâncias de Mahalanobis, a maior distância foi verificada entre os acessos 5 e 9 ($D^2 = 391,6$) e a menor entre os acessos 7 e 9 ($D^2 = 5,3$). Quanto às distâncias euclidianas médias padronizadas, o maior valor foi verificado entre os acessos 5 e 7 ou 5 e 9 ($d = 2,33$), e a menor distância foi verificada entre os acessos 6 e 8 ($d = 0,34$). Há grande concordância entre as duas medidas de distância, uma vez que a correlação entre as estimativas obtidas foi de 0,9486, significativa a 1% pelo teste de Mantel, com nível crítico estabelecido após 5.000 permutações, conforme ilustrado na Figura 2.3.

Segundo Arunachalan (1981), em estudos sobre a diversidade genética, avaliam-se vários caracteres com graus significativos de correlação, sendo desaconselhável a quantificação da diversidade genética pela distância euclidiana. Nessas situações, o procedimento recomendável é a utilização da estatística D^2 , que leva em consideração as associações entre as características por meio da matriz de variâncias e covariâncias residuais entre as variáveis.

Deve ser ressaltado que, para o cálculo de D^2 , é pressuposto existir distribuição normal multidimensional e homogeneidade entre as matrizes de covariâncias das unidades amostrais (RAO, 1952), o que restringe, portanto, o seu uso. No entanto, já foi demonstrada considerável robustez para a violação dessas hipóteses, o que faz da distância generalizada de Mahalanobis (D^2) instrumento útil (CRUZ; CARNEIRO, 2003), além da vantagem de proporcionar maior analogia entre as técnicas multivariadas e outras técnicas de agrupamento.

Tabela 2.9 - Medidas de dissimilaridade entre 15 cultivares de milho-pipoca.

Abaixo da diagonal estão apresentadas as estimativas das distâncias de Mahalanobis e, na parte superior, as distâncias euclidianas médias padronizadas

Cult.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		0,98	0,95	0,71	1,43	1,71	1,88	1,75	1,85	1,55	1,43	1,61	1,88	1,67	1,55
2	66,0		0,86	0,98	1,12	1,76	1,91	1,78	1,88	1,55	1,33	1,32	1,49	1,40	1,30
3	77,1	53,7		0,89	1,28	1,70	1,77	1,73	1,77	1,52	1,67	1,72	1,87	1,64	1,68
4	31,3	48,1	74,0		1,53	1,37	1,49	1,48	1,52	1,31	1,39	1,44	1,72	1,37	1,44
5	157,5	68,4	139,7	127,8		2,17	2,33	2,09	2,33	1,96	2,06	1,77	2,15	2,01	1,91
6	187,0	199,6	188,0	179,9	341,6		0,56	0,34	0,46	0,50	1,36	1,25	1,41	1,19	1,16
7	216,5	200,9	174,7	195,9	356,6	29,0		0,57	0,35	0,73	1,48	1,39	1,44	1,17	1,31
8	216,6	219,1	212,0	218,4	345,2	8,9	28,2		0,50	0,48	1,41	1,27	1,40	1,21	1,15
9	229,2	223,3	190,2	219,6	391,6	24,5	5,3	25,2		0,50	1,46	1,34	1,44	1,27	1,28
10	168,9	167,0	152,3	165,0	298,5	15,2	22,4	14,2	15,4		1,38	1,20	1,42	1,29	1,17
11	181,7	164,0	222,8	218,3	354,3	46,2	147,5	165,6	154,0	162,7		0,78	0,67	0,69	0,47
12	199,1	114,1	210,8	182,4	225,8	160,3	149,2	175,7	163,5	157,5	51,2		0,78	0,77	0,57
13	294,0	201,0	271,4	304,0	378,7	172,5	140,1	174,0	153,5	180,4	34,7	50,9		0,59	0,45
14	195,9	117,0	187,2	178,9	262,7	127,9	104,7	138,6	127,8	136,3	34,0	24,4	25,4		0,56
15	193,0	143,6	215,5	210,9	294,9	101,4	105,4	105,9	114,4	114,1	17,5	34,4	21,5	17,9	

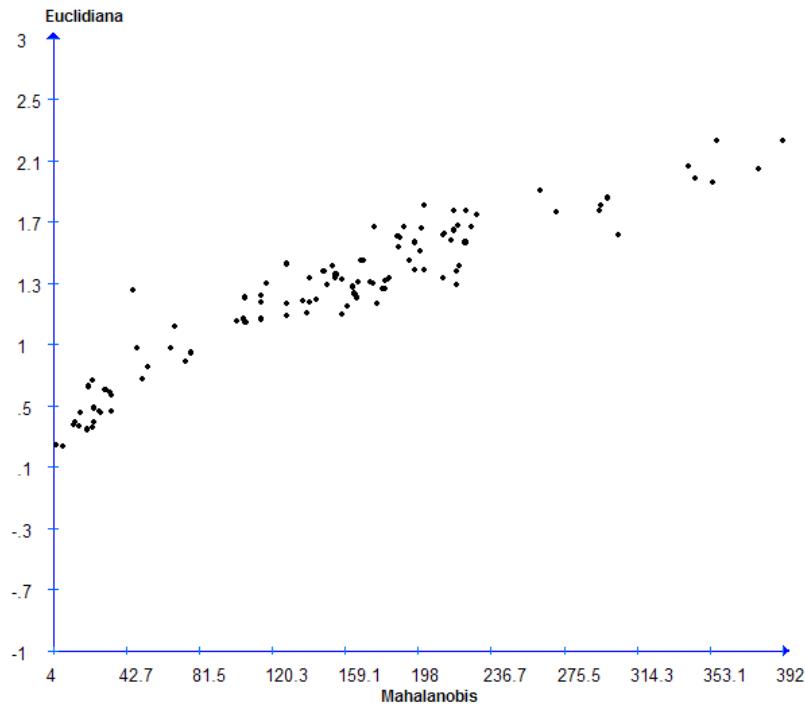


Figura 2.3 - Figura ilustrativa da relação entre distância de Mahalanobis e distância Euclidiana média entre 15 cultivares de milho-pipoca, obtida a partir de nove características agronômicas.

Cruz (1990) realizou estudo da diversidade genética entre cultivares de milho por meio da distância euclidiana média e de Mahalanobis, encontrando correlação entre as estimativas de 0,97; esse autor ressaltou que a concordância nas estimativas é dependente da magnitude das correlações residuais que possam existir entre os caracteres considerados.

A técnica multivariada de quantificação da distância de Mahalanobis permite quantificar a importância relativa de caracteres para a diversidade genética, por meio da avaliação da contribuição destes para os valores de D^2 . Neste estudo, a importância relativa de cada característica está descrita na Tabela 2.10. As características menos importantes foram altura de espigas (AE), número de espigas (NE) e produção de grãos (PROD). Pelo método de Singh (1981), consideram-se de menor importância as características que apresentam menor variabilidade ou que estão representadas por outras. Neste exemplo, as características consideradas de

menor importância foram aquelas de mais baixa herdabilidade e de menor relação entre o coeficiente de variação genético e o ambiental.

Tabela 2.10 - Importância relativa de caracteres agronômicos para estudo da diversidade genética entre 15 cultivares de milho-pipoca e parâmetros genéticos associados a essas características

Características	S. _j	F	Valor de S. _j (%)	Herdabilidade	CVg	CVg/Cve
AP	2008,31	23,12	12,47	95,67	27,67	2,71
AE	819,78	8,60	5,09	88,35	20,83	1,59
NE	950,59	16,37	5,90	93,89	25,69	2,26
PROL	1312,79	27,21	8,15	96,33	31,70	2,96
CE	2410,30	32,30	14,97	96,90	35,30	3,23
QUE	3462,37	38,14	21,50	97,23	45,67	3,42
DOEN	1455,17	18,00	9,04	94,75	30,12	2,45
PCG	2487,19	23,00	15,44	95,65	36,09	2,71
PROD	1196,30	16,55	7,43	93,95	29,38	2,28

AP: altura da planta; AE: altura da espiga; NE: número de espigas; PROL: proliferação; CE: capacidade de expansão; QUE: proporção de plantas quebradas; DOEN: proporção de plantas doentes; PCG: peso de cem grãos (PCG); PROD: produção de grãos.

2.2.2. Variáveis fenotípicas discretas – padrão binário

Os caracteres binários por conveniência são codificados em 0 e 1, e é partir dessas informações que são calculadas as medidas de dissimilaridade entre pares de acessos. Geralmente é adotado o critério de que indivíduos similares são aqueles que possuem maior concordância de manifestação de coincidência de padrão 1 (um) e, algumas vezes, incluindo também o padrão 0 (zero). Segundo Johson e Wichern (1988), embora a distância euclidiana possa ser usada para avaliar a similaridade entre indivíduos a partir das informações de dados binários, ela fornece uma contagem do número de padrões discordantes (1-0 e 0-1) entre dois indivíduos e atribui pesos iguais aos padrões

concordantes (1-1 e 0-0), sendo inadequada quando for necessário atribuir pesos diferentes a cada um desses padrões.

Em certos estudos, tanto em informações moleculares quanto fenotípicas, costuma-se desconsiderar o número de coincidências do tipo 0-0 no cálculo da similaridade ou dissimilaridade entre dois indivíduos. Pelo exposto, a literatura apresenta várias maneiras de avaliar a diversidade genética entre acessos de acordo com as informações obtidas de dados binários, a partir dos quais é avaliada a presença ou ausência de certo padrão, codificadas em zeros e uns. Nessa situação, os coeficientes de similaridade entre pares de genótipos podem ser obtidos considerando os valores:

- a: valor que quantifica o número de coincidência do tipo 1-1 para cada par de genótipos;
- b: valor que quantifica o número de discordância do tipo 1-0 para cada par de genótipos;
- c: valor que quantifica o número de discordância do tipo 0-1 para cada par de genótipos; e
- d: valor que quantifica o número de coincidência do tipo 0-0 para cada par de genótipos.

Com base nesses valores podem ser obtidas as medidas de dissimilaridade, assim como a distância euclidiana média, dada por:

$$d_{ii'} = \sqrt{\frac{a(1-1)^2 + b(1-0)^2 + c(0-1)^2 + d(0-0)^2}{a+b+c+d}}$$

$$d_{ii'} = \sqrt{\frac{b+c}{a+b+c+d}}$$

esta medida é conhecida por distância binária de Sokal.

Outras medidas de similaridade, geralmente recomendadas para estudos de análise de marcadores moleculares, também podem ser recomendadas para variáveis binárias de natureza fenotípica (Sneath e Sokal, 1973). São elas:

Coeficiente de coincidência simples

$$s_{ii'} = \frac{a + d}{a + b + c + d}$$

Coeficiente de Roger e Tanimoto

$$s_{ii'} = \frac{a + d}{a + 2(b + c) + d}$$

Coeficiente de Sokal e Sneath

$$s_{ii'} = \frac{2(a + d)}{2(a + d) + b + c}$$

Coeficiente de Russel e Rao

$$s_{ii'} = \frac{a}{a + b + c + d}$$

Coeficiente de Jaccard

$$s_{ii'} = \frac{a}{a + b + c}$$

Coeficiente de Sorenson ou Dice ou Nei e Li

$$s_{ii'} = \frac{2a}{2a + b + c}$$

Coeficiente de Ochiai

$$s_{ii'} = \frac{a}{\sqrt{(a + b)(a + c)}}$$

Coeficiente de Baroni, Urbani e Buser

$$s_{ii'} = \frac{a + ad}{a + b + c + ad}$$

Coeficiente de Haman

$$s_{ii'} = \frac{(a+d)-(b+c)}{a+b+c+d}$$

Coeficiente de Yule

$$s_{ii'} = \frac{ad-bc}{ad+bc}$$

Coeficiente de Ochiai II

$$s_{ii'} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Outros coeficientes:

$$s_{ii'} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

$$s_{ii'} = 0,5 \left[\frac{a}{a+b} + \frac{a}{a+c} \right]$$

$$s_{ii'} = 0,25 \left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d} \right]$$

Como se trata de medidas de similaridade, é necessário, na análise de agrupamento, fazer uso de medidas de dissimilaridade, expressas pelo complemento aritmético ou pelo inverso do coeficiente obtido, ou seja:

a) O complemento aritmético de um coeficiente de similaridade é dado por:

$$d_{ii'} = 1 - s_{ii'}$$

Muitos índices de similaridade têm intervalo de variação de 0 a 1, e essa transformação tem a vantagem de manter esse mesmo intervalo de variação na medida de dissimilaridade.

b) O inverso do coeficiente, acrescido de uma unidade, com a finalidade de contornar os problemas de indeterminação provocados na condição em que o coeficiente assume valor zero, é formada por:

$$d_{ii'} = \frac{1}{s_{ii'} + 1}$$

Para os índices de similaridade com intervalo de variação de 0 a 1, essa transformação faz com que a variação na medida de dissimilaridade seja apenas de 0,5 a 1. Assim, se a similaridade for total, a dissimilaridade será de 50%.

Em alguns casos recomenda-se o cálculo da raiz quadrada da medida de dissimilaridade, para que a mesma adquira propriedade euclidiana, bastante desejável em análises por dispersão gráfica.

O complemento aritmético do coeficiente de coincidência simples tem a vantagem de ser idêntico ao quadrado da distância euclidiana média, medida bastante utilizada em características quantitativas de distribuição contínua. Entretanto, apresenta a particularidade de considerar como fator de similaridade o número de coincidência do tipo "0-0", que, em alguns casos, pode ser indesejável.

Os índices de Jaccard e de Nei e Li passam a ser interessantes nas situações em que se deve excluir a coincidência do tipo "0-0" como fator de similaridade. Para Jaccard, apenas a coincidência do tipo "1-1" deve ser levada em consideração, na similaridade entre dois acessos. O coeficiente de Nei e Li também tem essa pressuposição, mas admite que, em certos estudos, essa coincidência é pouco esperada e, ao ser encontrada, deve ser ponderada por um peso mais elevado - no caso, igual a 2. Assim, tem sido consenso que, ao trabalhar com materiais exóticos ou envolvendo espécies diferenciadas, a princípio pouco relacionadas, o adequado seria adotar o coeficiente de Nei e Li. Entretanto, se o estudo é feito dentro de uma população, ou espécie, em que coincidências de ocorrência de certos padrões podem ser admitidas como um fenômeno esperado, recomendar-se-ia a utilização do índice de Jaccard. Dentro deste mesmo ponto de vista, e nas situações em que a coincidência "0-0" é considerada causa de similaridade, pode-se optar pelo índice de Sokal e Sneath nas situações em que a coincidência é rara e, quando detectada, deve ser ponderada por peso 2.

Aplicação

Será considerada, como ilustração, a análise da dissimilaridade entre cinco genótipos, considerando as informações relativas a nove características de padrão binário. A presença de uma particularidade foi caracterizada pelo número 1 e a ausência pelo número zero, conforme ilustrado na Tabela 2.11.

Tabela 2.11 - Resistência (valor 1) e suscetibilidade (valor 0) a nove raças de um patógeno, manifestada em cinco genótipos

Genótipos	R1	R2	R3	R4	R5	R6	R7	R8	R9
G1	1	1	1	0	0	1	0	1	1
G2	0	1	0	1	0	0	1	0	0
G3	1	0	1	1	1	1	0	0	1
G4	0	1	1	0	0	1	0	1	0
G5	1	1	0	1	1	0	1	0	1

Com base nas informações da Tabela 2.11, estimam-se, por exemplo, os coeficientes de dissimilaridade entre os genótipos 1 e 2, dados por:

Complemento aritmético do coeficiente de coincidência simples

$$d_{ii'} = 1 - \frac{a+d}{a+b+c+d} = 1 - \frac{2}{9} = \frac{7}{9} = 0,78$$

Complemento aritmético do coeficiente de Jaccard

$$d_{ii'} = 1 - \frac{a}{a+b+c} = 1 - \frac{1}{8} = \frac{7}{8} = 0,875$$

Complemento aritmético do coeficiente de Nei e Li

$$d_{ii'} = 1 - \frac{2a}{2a+b+c} = 1 - \frac{2}{9} = \frac{7}{9} = 0,78$$

As demais estimativas dos coeficientes de dissimilaridade são apresentadas na Tabela 2.12.

Tabela 2.12 - Coeficientes de dissimilaridade entre cinco genótipos, estimados a partir de informações de nove raças de um patógeno

Genótipos	a(1-1)	b(1-0)	c(0-1)	d(0-0)	Coincidência simples (*)	Jaccard (*)	Nei e Li (*)
1 e 2	1	5	2	1	0,78	0,88	0,78
1 e 3	4	2	2	1	0,44	0,50	0,33
1 e 4	4	2	0	3	0,22	0,33	0,20
1 e 5	3	3	3	0	0,67	0,67	0,50
2 e 3	1	2	5	1	0,78	0,88	0,78
2 e 4	1	2	3	3	0,55	0,83	0,71
2 e 5	3	0	3	3	0,33	0,50	0,33
3 e 4	2	4	2	1	0,67	0,75	0,60
3 e 5	4	2	2	1	0,44	0,50	0,33
4 e 5	1	3	5	0	0,89	0,89	0,80

(*) Complemento aritmético do respectivo coeficiente de similaridade.

Os três coeficientes utilizados foram concordantes em apontar como mais dissimilares os genótipos 4 e 5 e, como mais similares, os genótipos 1 e 4.

2.2.3. Variáveis fenotípicas discretas – padrão multicategórico

Caracteres multicategóricos são comumente avaliados no melhoramento vegetal, principalmente relacionados com particularidades morfológicas e estruturais da planta, além de se ter grande interesse em certos atributos que conferem qualidade do produto comercializado, como a forma, a coloração, o sabor do fruto, a consistência da polpa, entre outros. Quando os valores dos caracteres multicategóricos podem ser ordenados, ou seja, eles são de natureza ordinal e descritos em uma escala, há possibilidade de eles serem analisados como variáveis quantitativas discretas (SNEATH; SOKAL, 1973). Entretanto, em muitos casos, os valores de certas variáveis multicategóricas não podem ser ordenados, pois são de natureza nominal, causando dificuldade na avaliação da

dissimilaridade entre os genótipos estudados. Apesar de se atribuírem valores numéricos, como 1, 2, 3 etc., para cada categoria, não é apropriado o uso de medidas de dissimilaridade tradicionais, como a distância euclidiana, uma vez que não é possível afirmar que indivíduos com valores mais discrepantes sejam mais distantes fenotipicamente do que indivíduos com valores mais próximos. Quando se dispõe de características multicategóricas, podem-se utilizar os seguintes índices:

Coincidência simples

Uma maneira de estimar a similaridade considerando um conjunto de variáveis multicategóricas, avaliadas em uma coleção ou população em que cada acesso (família, variedade ou população) apresenta um único padrão fenotípico, é por meio do índice:

$$s_{ii'} = \frac{C}{C + D}$$

em que:

C: concordância de categoria; e

D: discordância de categoria.

Com este índice, um determinado valor expressa a porcentagem de coincidência de similaridade considerando os vários caracteres analisados. Assim, um valor de $s_{ii'}$ igual a 0,40 revela que os dois genótipos (i e i') são similares em 40% das características multicategóricas estudadas. A dissimilaridade é dada por:

$$d_{ii'} = \frac{D}{C + D}$$

Ilustração

Considere a avaliação de cinco genótipos em relação a três características multicategóricas: nível de resistência às raças de um determinado patógeno (1 = raça I; 2 = raça II; 3 = raça I e II; e 4 = nenhuma raça), cor da flor [1 = branca (B); 2 =

amarela (A); e 3 = creme (C)] e formato do fruto [1 = redondo (R); 2 = alongado (A); e 3 = oval (O)]. Os valores de cada genótipo estão descritos a seguir:

Genótipo	Nível de resistência	Cor da flor	Formato do fruto
1	4	3	1
2	2	3	1
3	0	1	2
4	4	2	2
5	3	2	3

Os índices de dissimilaridade entre pares de genótipos são os seguintes:

Genótipos	C	D	$D_{ii'}$
1 e 2	2	1	0,33
1 e 3	0	3	1,00
1 e 4	1	2	0,66
1 e 5	0	3	1,00
2 e 3	0	3	1,00
2 e 4	0	3	1,00
2 e 5	0	3	1,00
3 e 4	1	2	0,66
3 e 5	0	3	1,00
4 e 5	1	2	0,66

Dissimilaridade de Cole-Rodgers et al. (1997)

Este índice deve ser aplicado para estimar a similaridade levando-se em conta um conjunto de variáveis multicategóricas, avaliadas em uma coleção ou população em que cada acesso, representado por vários indivíduos, pode apresentar mais de um padrão fenotípico. Dessa forma, o índice de dissimilaridade deve contemplar a coincidência e discordância de valores nas

várias categorias de uma mesma variável multicategórica. Assim, Cole-Rodgers et al. (1997) apresentaram o seguinte índice de dissimilaridade para essa situação:

$$d_{ij} = \frac{D_1}{C_1 + D_1} + \frac{D_2}{C_2 + D_2} + \dots + \frac{D_v}{C_v + D_v} = \sum_{j=1}^v \frac{D_j}{C_j + D_j}$$

em que:

C_j : número de concordância entre categorias para a j-ésima variável multicategórica; e

D_j : número de discordância entre categorias para a j-ésima variável multicategórica.

Neste índice, o valor máximo de dissimilaridade é igual ao número de característica avaliada e o valor mínimo é zero.

Ilustração

Serão considerados três genótipos em relação às seguintes características multicategóricas: nível de resistência às raças de um determinado patógeno, formato do fruto e cor da flor (conforme descritos anteriormente). Neste caso, têm-se os seguintes valores:

Genótipo	Resistência à raça do patógeno [#]				Cor da flor			Formato do fruto		
	Raça 1	Raça 2	As duas raças	Nenhuma raça	B	A	C	R	A	O
1	1	1	0	1	1	0	1	1	0	0
2	0	1	1	0	1	1	1	1	0	1
3	1	0	0	1	0	0	1	0	1	0

[#] Neste exemplo, os genótipos representam famílias ou populações que podem apresentar material resistente às duas raças ou apenas uma delas.

A dissimilaridade entre os genótipos é dada por:

$$d_{12} = \frac{3}{4} + \frac{1}{3} + \frac{1}{3} = 1,42$$

$$d_{13} = \frac{1}{4} + \frac{1}{3} + \frac{2}{3} = 1,25$$

$$d_{23} = \frac{4}{4} + \frac{2}{3} + \frac{3}{3} = 2,67$$

Dissimilaridade de Cole-Rodgers et al. (1997) modificada

No índice original proposto por Cole-Rodger et al. (1997), o valor da concordância de ausência (0-0) de uma categoria em cada variável contribui para diminuir a dissimilaridade. Portanto, se esse fato não é desejável, deve-se optar pelo índice alternativo:

$$d_{ii} = \frac{D_1}{C_{p1} + D_1} + \frac{D_2}{C_{p2} + D_2} + \dots + \frac{D_v}{C_{pv} + D_v} = \sum_{j=1}^v \frac{D_j}{C_{pj} + D_j}$$

em que C_{pj} é o número de concordância envolvendo a presença de categoria no par de genótipos.

Pelo exemplo anterior, tem-se:

$$d_{12} = \frac{3}{4} + \frac{1}{3} + \frac{1}{2} = 1,58$$

$$d_{13} = \frac{1}{3} + \frac{1}{2} + \frac{2}{2} = 1,83$$

$$d_{23} = \frac{4}{4} + \frac{2}{3} + \frac{3}{3} = 2,67$$

Distância euclidiana média ponderada

Outra crítica que é apresentada aos índices, original e modificado, propostos por Cole-Rodger et al. (1997) refere-se ao fato de que eles não levam em consideração a freqüência de ocorrência de cada categoria na população estudada. Entretanto, ao considerar as freqüências de cada categoria, a medida de dissimilaridade poderia ser qualquer uma adotada para características quantitativas, como, por exemplo, a distância euclidiana média.

A dissimilaridade, expressa pela distância euclidiana média, é dada por:

$$d_{ii'} = \sqrt{\frac{1}{v} \sum_{j=1}^v \sum_{k=1}^{c_j} \frac{(f_{ijk} - f'_{ijk})^2}{c_j}}$$

em que:

f_{ijk} : freqüência observada no genótipo i , para a categoria k da variável j ;

v : número de variáveis analisadas;

c_j : número de categorias dentro da variável j ;

$i = 1, 2, \dots, g$

$j=1, 2, \dots, v$

$k = 1, 2, \dots, c_j$

Ilustração

Serão considerados, a título de ilustração, os seguintes dados:

Genótipo	Resistência a raça do patógeno				Cor da flor			Formato do fruto		
	Raça 1	Raça 2	As duas raças	Nenhuma raça	B	A	C	R	A	O
	1	25 [#]	25	0	50	95	0	5	100	0
2	0	40	60	0	30	35	35	70	0	30
3	90	0	0	10	0	0	100	0	100	0

[#] Valores em porcentagem.

Assim, tem-se:

$$d_{12} = \sqrt{\frac{1}{3} \left[\frac{(0,25 - 0)^2 + \dots + (0,5 - 0)^2}{4} + \dots + \frac{(1 - 0,7)^2 + (0 - 0)^2 + (0 - 0,3)^2}{3} \right]} =$$

0,3853

$d_{13} = 0,6903$

$d_{23} = 0,5981$

2.2.4. Associação de variáveis qualitativas e quantitativas

Quando o estudo da diversidade genética é feito a partir de vários tipos de variáveis, podem ser recomendadas algumas estratégias de análise, descritas a seguir:

a) Numa primeira alternativa, recomenda-se que as características sejam subdivididas em grupos, de forma a ser utilizada, para cada grupo, a medida de dissimilaridade mais apropriada, tendo-se em vista as particularidades das mensurações efetuadas. Com esse procedimento são obtidas várias matrizes de dissimilaridade, que podem ser utilizadas, individualmente, em análises futuras de projeções ou agrupamento ou gerar uma matriz de dissimilaridade conjunta cujos elementos serão dados pela média das dissimilaridades obtidas por meio da cada conjunto de dados. No cálculo desta média, recomenda-se que os valores de cada matriz devam ser previamente padronizados, visto que a amplitude de variação em cada uma delas poderá variar drasticamente. Nesta padronização considera-se que a matriz de dissimilaridade de dimensão $n \times n$ tenha $N = n(n-1)/2$ elementos d_{ij} , com média \bar{d} e variância $\hat{\sigma}_d^2$. A padronização é feita por meio de:

$$d_{pij} = \frac{d_{ij}}{\hat{\sigma}_d}$$

Assim, a matriz de distância média obtida a partir de duas matrizes de dissimilaridade D_1 e D_2 , por exemplo, é obtida por meio da expressão:

$$\bar{D} = \frac{p_1 D_1 + p_2 D_2}{p_1 + p_2}$$

sendo p_1 e p_2 os pesos estabelecidos para D_1 e D_2 , respectivamente. Estes pesos podem ser estabelecidos de forma arbitrária ou adotando algum critério como, por exemplo, um valor proporcional ao número de caracteres utilizados da obtenção das matrizes de dissimilaridade.

b) Uma outra alternativa é adotar uma medida de dissimilaridade única para todos os tipos de variáveis ou diferentes medidas de dissimilaridade, mas que

apresentam mesmo intervalo de definição. De maneira geral, os índices de similaridade adotados para variáveis binárias ou multicategóricas variam de 0 a 1. Assim, uma opção é adotar para as variáveis quantitativas uma medida que também varie dentro deste intervalo.

Gower (1971) propôs um coeficiente geral de similaridade que é aplicável simultaneamente aos três tipos de características, binárias, multicategóricas e quantitativas. É obtido um valor de similaridade, para o par de indivíduos i e i' , para cada característica, denotado por $s_{ii'k}$, de forma que a similaridade entre i e i' seja dada por:

$$s_{ii'} = \frac{\sum_{k=1}^v s_{ii'k} \omega_{ii'k}}{\sum_{k=1}^v \omega_{ii'k}}$$

Sendo $\omega_{ii'k}$ um peso que assume valor 1 quando não há valor perdido para a característica k , quando se comparam os indivíduos i e i' , e 0 em caso contrário.

Para características binárias, a similaridade poderá ser obtida pelo grau de coincidência de padrão “1-1”, assumindo peso ($\omega_{ii'k}$) nulo nas situações em que um padrão fenotípico não se manifesta em ambos os indivíduos, para a k -ésima característica. Assim, quando se analisam unicamente características binárias, esse coeficiente será igual ao proposto por Jaccard.

Quanto a características multicategóricas, esse coeficiente também poderá ser expresso pelo grau de concordância de padrões fenotípicos entre os acessos i e i' , de forma que o valor será 1 se os acessos i e i' apresentarem o mesmo padrão fenotípico para a característica multicategórica analisada e 0, em caso contrário.

Para uma característica quantitativa k , o valor de $s_{ii'k}$ poderá ser obtido por meio de:

$$s_{ii'k} = 1 - \frac{|Y_{ik} - Y_{i'k}|}{R_k}$$

sendo R_k a amplitude de variação da característica k em toda população.

Para v variáveis, calcula-se:

$$s_{ii'} = \frac{1}{v} \sum_{k=1}^v s_{ii'k}$$

De maneira geral, o valor $s_{ii'}$ será igual a 1 quando os acessos i e i' apresentarem o mesmo padrão fenotípico para todas as características analisadas e zero quando eles apresentam valores discordantes para características qualitativas (binárias ou multicategóricas) e extremos para as quantitativas.

c) Por fim, pode-se recomendar a conversão de todas variáveis em um único padrão. Métodos de transformação de variáveis quantitativas em multicategóricas já foram abordados neste livro, de forma que todas possam ser tratadas, por exemplo, como de natureza discreta de padrão binário ou multicategórico.

2.2.5 Utilização da Matriz de Dissimilaridade

As estimativas de dissimilaridade atendem aos objetivos do melhorista por quantificarem e informarem sobre o grau de semelhança ou de diferença apresentado entre dois quaisquer genótipos. Algumas vezes, seus valores, por si só, são de grande utilidade, principalmente quando há interesse em orientar hibridações, concentrando esforços em cruzamentos em que há diversidade genética entre genitores, sendo indicativo de sua complementaridade gênica.

No melhoramento vegetal, a realização de cruzamentos depende de sincronia de florescimento masculino e feminino. Verifica-se que nem sempre há coincidência de períodos em que o pólen está viável e os estigmas receptivos para a fecundação, entre os genitores desejados. A informação sobre quais cruzamentos devem ser evitados e priorizados torna-se, portanto, fundamental. Assim, pode-se ter interesse em apenas identificar os k acessos mais e menos similares a cada um dos genótipos estudados, identificados a partir da matriz de dissimilaridade estimada.

Uma medida útil nessa situação é a média geral das distâncias dos k acessos mais e menos similares em relação a um particular genótipo i de interesse, ou seja:

$$\bar{d}_i = \frac{1}{k} \sum_{j \neq i}^k d_{ij}$$

A seguir são apresentadas as informações relativas aos dois genótipos mais e menos similares, em relação conjunto de sete genótipos avaliados, cuja matriz de dissimilaridade é dada por:

$$D_{7 \times 7} = \begin{array}{c|ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 1 & 0,00 & & & & & & \\ 2 & 2,16 & 0,00 & & & & & \\ 3 & 0,98 & 2,99 & 0,00 & & & & \text{Simétrica} \\ 4 & 14,32 & 12,53 & 9,55 & 0,00 & & & \\ 5 & 3,00 & 6,55 & 1,44 & 12,75 & 0,00 & & \\ 6 & 5,25 & 8,45 & 2,41 & 14,12 & 1,49 & 0,00 & \\ 7 & 14,66 & 17,56 & 10,33 & 9,91 & 6,37 & 10,34 & 0,00 \end{array}$$

Assim, tem-se:

Acessos	Menos similares	\bar{d}_i	Mais similares	\bar{d}_i	\bar{d}_i (Geral)
1	7	14,5	3	1,6	6,7
2	7	14,0	1	2,6	8,0
3	7	9,9	1	1,2	4,6
4	1	14,2	3	9,7	12,2
5	4	9,7	3	1,5	5,3
6	4	12,2	5	2,0	7,0
7	2	15,1	5	8,1	11,2

Constata-se que, em relação ao genótipo 1, deve-se dar preferência aos cruzamentos com os genótipos 7 e 4 e evitar cruzamentos com o 3 e 2, com dissimilaridade muito abaixo da média geral da diversidade do genitor 1.

2.3. Técnicas de Agrupamento

Em muitas situações, o pesquisador está interessado em avaliar padrões de agrupamento, formular e testar hipóteses sobre a similaridade ou diversidade obtida. Contudo, em razão de o número de estimativas de dissimilaridade ser relativamente grande (igual a $n(n-1)/2$, em que n é o número de acessos considerados no estudo), torna-se impraticável o reconhecimento de grupos homogêneos pelo simples exame visual daquelas estimativas. Assim, para realizar essa tarefa, utilizando-se métodos de agrupamento ou de projeções de distâncias em gráficos bi ou tridimensionais, em que cada coordenada é obtida a partir da medida de dissimilaridade escolhida.

A versatilidade e o poder discriminatório dos métodos de agrupamento têm permitido sua aplicação nas mais variadas áreas da ciência (EVERITT, 1993). A literatura apresenta inúmeras técnicas de agrupamento, que se distinguem pelo tipo de resultado a ser fornecido e pelas diferentes formas de definir a proximidade entre um indivíduo e um grupo já formado ou entre dois grupos quaisquer. Em todos os casos não se conhece, *a priori*, o número de grupos a serem estabelecidos, e diferentes métodos proporcionam diferentes resultados. A análise de agrupamento é uma técnica puramente exploratória, que visa a geração de hipóteses sobre o padrão de aglomeração estabelecido e pode ser suplementada ou complementada por outras técnicas de visualização (DIAS, 1998). Além disso, não é necessária qualquer hipótese acerca da distribuição de probabilidade dos dados.

Destacam-se três abordagens particulares de agrupamento. A primeira delas relaciona-se às técnicas que produzem dendrogramas, em que o primeiro passo é calcular as medidas de dissimilaridade entre todos os pares possíveis de acessos e, assim, formar grupos por processos aglomerativos ou divisivos; a segunda envolve partições em um grupo em que os indivíduos possam se mover para fora ou para dentro deste e de outros grupos em diferentes estágios da análise. Por último, têm-

se as técnicas baseadas em dispersão gráfica, em que se consideram as posições relativas de acessos em gráficos bi ou tridimensionais.

Segundo Sneath e Sokal (1973), tais métodos de agrupamento podem ser classificados em diversas categorias contrastantes, da seguinte maneira:

a) Métodos aglomerativos e divisivos

Nos métodos aglomerativos consideram-se n indivíduos, que serão agrupados de forma sucessiva, com base em sua proximidade, em g grupos ($g < n$). No processo divisivo, o procedimento é oposto, ou seja, inicialmente todos os indivíduos estão num mesmo grupo, que se divide em um ou mais subgrupos, os quais se subdividem sucessivamente até o final do processo. Nos processos divisivos, se um indivíduo for alocado de forma inadequada, sua posição não será posteriormente corrigida. As técnicas aglomerativas são freqüentemente empregadas – principalmente quando associadas aos procedimentos seqüenciais, hierárquicos e sem sobreposição – e estão representadas por uma variedade de métodos.

b) Métodos hierárquicos e não-hierárquicos

Um método de agrupamento é considerado hierárquico se ele consiste numa seqüência de $w + 1$ agrupamentos, G_0, G_1, \dots, G_w , em que G_0 é a partição disjunta de todos os n indivíduos e G_w é a partição conjunta. O número de partes k_i na partição G_i , deve obedecer à regra $k_i \geq k_{i+1}$, em que k_{i+1} é o número de partes do grupo G_{i+1} .

Os métodos são considerados não-hierárquicos quando o indivíduo não tem ordem parcial. Exemplos são dados pelas técnicas de projeção bi ou tridimensionais.

Nos métodos hierárquicos, os genótipos são agrupados por um processo que se repete em vários níveis, até que seja estabelecido o dendrograma ou o diagrama de árvore. Nesse caso, não há preocupação com o número ótimo de grupos, uma vez que o interesse maior está na "árvore" e nas ramificações que são obtidas. As delimitações podem ser estabelecidas por um exame visual do dendrograma, em

que se avaliam pontos de alta mudança de nível, tomando-os em geral como delimitadores do número de genótipos para determinado grupo.

O conceito de representação hierárquica dos dados foi desenvolvido primeiramente na biologia, onde em muitos casos, principalmente os de taxonomia e evolução, é pressuposto existir uma estrutura hierárquica (MARDIA et al, 1979). Atualmente, os métodos hierárquicos têm uso generalizado, incluindo aquelas áreas em que a estrutura hierárquica pode não ser mais apropriada e, obviamente, o seu uso necessitaria de avaliação mais criteriosa.

c) Métodos sem sobreposição e métodos com sobreposição

Nos métodos sem sobreposição, os grupos, em dado nível hierárquico, são mutuamente exclusivos. Quando o conceito sem sobreposição é combinado com a hierarquia, surge um novo conceito: a classificação hierárquica aninhada. A classificação hierárquica é dada quando a sobreposição está relacionada com a hierarquia.

d) Métodos seqüenciais e métodos simultâneos

Nos métodos seqüenciais, é aplicada, ao grupo de indivíduos estudados, uma seqüência recorrente. Nos métodos simultâneos, um único procedimento não recorrente é aplicado ao grupo inteiro de indivíduos. A maioria dos métodos de agrupamento é seqüencial.

e) Critérios locais e globais

Em geral, muitos métodos seqüenciais aglomerativos estimam a similaridade entre indivíduos dentro de um grupo de forma confiável, ao passo que, quanto maiores forem os grupos considerados, menor será a confiabilidade. Ao contrário, a análise de componentes principais fornece uma representação confiável das distâncias intergrupo (daqueles mais distantes), porém tem na informação de distâncias dos indivíduos intragrupo menor confiabilidade. No entanto, as confiabilidades locais e globais podem diferir em vários métodos de agrupamento.

f) Soluções diretas e iterativas – métodos de otimização

São considerados métodos de agrupamentos diretos aqueles em que o algoritmo, para construção da classificação, é feito de modo direto e a solução obtida é considerada como ótima em algum sentido. Nos métodos seqüenciais, as soluções buscam a otimização local, isto é, um agrupamento em qualquer nível de fusão é geralmente computado por algum critério de otimização. Uma vez que a estrutura é obtida em certo nível de fusão, não há mudanças nos passos seguintes de agrupamento. Nos métodos de otimização, os grupos são formados pela adequação de algum critério de agrupamento, ou seja, o objetivo é alcançar uma partição dos indivíduos que otimize (maximize ou minimize) alguma medida predefinida. Um dos métodos mais comumente utilizado, na área de melhoramento genético é o proposto por Tcher, citado por Rao (1952).

Os procedimentos iterativos visam a otimização local, global ou ambas.

g) Agrupamentos ponderados e não-ponderados

Os métodos ponderados fornecem pesos aos ramos que se unem durante o processo de agrupamento de algum método aglomerativo seqüencial. A ponderação também é usada quando se considera que algumas dimensões são mais importantes que outras.

Métodos não-ponderados são aqueles que atribuem pesos iguais. Não dão preferência a um agregado de indivíduos em detrimento de outro conjunto de indivíduos, nem atribuem ponderação entre os indivíduos com base no tamanho do seu grupo.

h) Agrupamentos adaptativos e não-adaptativos

Nos métodos de agrupamento não-adaptativos, o algoritmo é aplicado de forma iterativa ou direta, para encontrar uma solução capaz de interagir com todos os pontos no espaço v-dimensional para formar o grupo. A maioria dos métodos é não-adaptativo.

Um método de agrupamento ideal seria o adaptativo, ou seja, inicialmente seriam explorados os tipos de conformações (por exemplo, elipsóide) de pontos (indivíduos) que provavelmente representam o conjunto de dados e, então, se decidiria qual delas é a mais adequada. Modificar-se-ia o algoritmo de agrupamento, para ponderar algumas distâncias entre indivíduos diferentemente.

Dos métodos de agrupamento, os mais utilizados são os de otimização e os hierárquicos.

2.3.1 Método de Agrupamento de Otimização

Método de Tocher

O método requer a obtenção da matriz de dissimilaridade, sobre a qual é identificado o par de indivíduos mais similares. Esses indivíduos formarão o grupo inicial. A partir daí é avaliada a possibilidade de inclusão de novos indivíduos, adotando-se o critério de que a distância média intragrupo deve ser menor que a distância média intergrupo.

A entrada de um indivíduo em um grupo sempre aumenta o valor médio da distância dentro do grupo. Assim, pode-se tomar a decisão de incluir o indivíduo em um grupo por meio da comparação entre o acréscimo no valor médio da distância dentro do grupo e um nível máximo permitido, que pode ser estabelecido arbitrariamente, ou adotado, como tem sido geralmente feito, o valor máximo (θ) da medida de dissimilaridade encontrado no conjunto das menores distâncias envolvendo cada indivíduo.

Neste caso, a distância entre o indivíduo k e o grupo formado pelos indivíduos ij é dada por:

$$d_{(ij)k} = d_{ik} + d_{jk}$$

Durante o processo de agrupamento há necessidade de avaliar o acréscimo no total da diversidade do grupo, no estágio t , pela inclusão de um acesso k de maior similaridade ao grupo. Assim, devem ser observados os seguintes valores:

Estágio	Número de acessos	Número de distâncias	Total das distâncias
t	n	$n(n-1)/2$	$d_{(grupo)}$
t+1	$n' = n+1$	$(n+1)n/2$	$d_{(grupo+k)} = d_{(grupo)} + d_{(grupo)k}$
Acréscimo		n	$d_{(grupo)k}$

A inclusão, ou não, do indivíduo k no grupo é, então, feita considerando que o acréscimo médio promovido pela inclusão de um indivíduo k em um grupo previamente estabelecido seja menor que θ .

Se $\frac{d_{(grupo)k}}{n} \leq \theta$, inclui-se o indivíduo k no grupo;

Se $\frac{d_{(grupo)k}}{n} > \theta$, o indivíduo k não é incluído no grupo; e

sendo n o número de indivíduos que constitui o grupo original.

Deve ser lembrado que ao considerar, por exemplo, um indivíduo m e um grupo com três indivíduos i, j e k, tem-se as seguintes relações de distâncias:

$$- d_{(grupo)} = d_{(ijk)} = d_{ij} + d_{ik} + d_{jk}$$

$$- d_{(grupo)indivíduo} = d_{(ijk)m} = d_{im} + d_{jm} + d_{km}$$

portanto,

$$d_{(grupo + indivíduo)} = d_{(grupo)} + d_{(grupo)indivíduo}$$

ou

$$d_{(grupo + indivíduo)} = d_{(ijk m)} = (d_{ij} + d_{ik} + d_{jk}) + (d_{im} + d_{jm} + d_{km})$$

Ilustração

Considere a matriz de distâncias utilizada nas ilustrações anteriores:

$$D_{7 \times 7} = \begin{array}{c|ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 1 & 0,00 & & & & & & \\ 2 & 2,16 & 0,00 & & & & & \\ 3 & 0,98 & 2,99 & 0,00 & & & & \text{Simétrica} \\ 4 & 14,32 & 12,53 & 9,55 & 0,00 & & & \\ 5 & 3,00 & 6,55 & 1,44 & 12,75 & 0,00 & & \\ 6 & 5,25 & 8,45 & 2,41 & 14,12 & 1,49 & 0,00 & \\ 7 & 14,66 & 17,56 & 10,33 & 9,91 & 6,37 & 10,34 & 0,00 \end{array}$$

Primeiro Passo:

- a) Estimar a maior distância dentre o conjunto de menores distâncias envolvendo cada genótipo.

As menores distâncias em relação a cada um dos genótipos são as seguintes:

Genótipo	1	2	3	4	5	6	7
mínimo d_{ii}	0,98	2,16	0,98	9,55	1,44	1,49	6,37

Assim, é estabelecido $\theta = 9,55$ como limite de acréscimo, na média da distância dentro do grupo, para formação ou inclusão de um novo elemento no grupo, porque esse valor é o maior dentro do conjunto de menores distâncias apresentado anteriormente.

b) Formação do grupo I

Inicialmente, o grupo I é formado pelos genótipos 1 e 3, cuja distância é a menor de todas, sendo menor do que o limite estabelecido.

Logo após, calculam-se as distâncias entre este grupo e os demais genótipos por meio de $d_{(ij)k} = d_{ik} + d_{jk}$. Então, tem-se:

$$d_{(13)2} = d_{12} + d_{23} = 2,16 + 2,99 = 5,15$$

$$d_{(13)4} = d_{14} + d_{34} = 14,32 + 9,55 = 23,87$$

$$d_{(13)5} = d_{15} + d_{35} = 3,00 + 1,44 = 4,44$$

$$d_{(13)6} = d_{16} + d_{36} = 5,25 + 2,41 = 7,66$$

$$d_{(13)7} = d_{17} + d_{27} = 14,66 + 10,33 = 24,99$$

Constata-se que o genótipo 5 é o mais similar. A possibilidade de inclusão deste genótipo no grupo em formação é feita da seguinte maneira:

$$\frac{d_{(13)5}}{2} = 2,22$$

Como $\frac{d_{(13)5}}{2} \leq \theta$, então o genótipo 5 pode ser incluído no grupo I.

Dando continuidade ao processo, serão calculadas as distâncias dos demais indivíduos com o grupo 1,3 e 5.

$$d_{(135)2} = d_{12} + d_{23} + d_{25} = 2,16 + 2,99 + 6,55 = 11,70$$

$$d_{(135)4} = d_{14} + d_{34} + d_{45} = 14,32 + 9,55 + 12,75 = 36,62$$

$$d_{(135)6} = d_{16} + d_{36} + d_{56} = 5,25 + 2,41 + 1,49 = 9,15$$

$$d_{(135)7} = d_{17} + d_{37} + d_{57} = 14,66 + 10,33 + 6,37 = 31,36$$

Constata-se que o genótipo 6 é o mais similar. A possibilidade de inclusão deste genótipo no grupo em formação é feita da seguinte maneira:

$$\frac{d_{(135)6}}{3} = 3,05$$

Como $\frac{d_{(135)6}}{3} \leq \theta$, então o genótipo 6 pode ser incluído no grupo I.

Agora, serão calculadas as distâncias com o grupo 1,3,5 e 6.

$$d_{(1356)2} = d_{12} + d_{23} + d_{25} + d_{26} = 2,16 + 2,99 + 6,55 + 8,45 = 20,15$$

$$d_{(1356)4} = d_{14} + d_{34} + d_{45} + d_{46} = 14,32 + 9,55 + 12,75 + 14,12 = 50,74$$

$$d_{(1356)7} = d_{17} + d_{37} + d_{57} + d_{67} = 14,66 + 10,33 + 6,37 + 10,34 = 41,70$$

O genótipo 2 é o mais similar. A possibilidade de inclusão deste genótipo no grupo em formação é feita da seguinte maneira:

$$\frac{d_{(1356)2}}{4} = 5,04$$

Como $\frac{d_{(1356)2}}{4} \leq \theta$, então o genótipo 2 pode ser incluído no grupo I.

O cálculo das distâncias dos indivíduos com o grupo 1,3,5,6 e 2 é feito da seguinte forma:

$$d_{(13562)4} = d_{14} + d_{34} + d_{45} + d_{46} + d_{24} = 14,32 + 9,55 + 12,75 + 14,12 + 12,53 \\ = 63,27$$

$$d_{(13562)7} = d_{17} + d_{37} + d_{57} + d_{67} + d_{27} = 14,66 + 10,33 + 6,37 + 10,34 + 17,56 \\ = 59,26$$

O genótipo 7 é o mais similar. A possibilidade de inclusão deste genótipo no grupo em formação é feita da seguinte maneira:

$$\frac{d_{(13562)7}}{5} = 11,85$$

Como $\frac{d_{(13562)7}}{5} > \theta$, então o genótipo 7 não pode ser incluído no grupo I.

c) Formação do Grupo II

É avaliada a possibilidade de os genótipos 4 e 7 formarem um novo grupo. A distância entre estes genótipos é $d_{47} = 9,91$. Como este valor é superior a θ , o agrupamento não é estabelecido, ficando cada genótipo em grupos separados.

Assim, no agrupamento final dos genótipos, pelo método de otimização de Tocher, foram estabelecidos os seguintes grupos (Tabela 2.13).

Tabela 2.13 - Grupos de genótipos estabelecidos pelo método de Tocher, com base no quadrado da distância euclidiana

Grupo	Genótipos	Distância média	Distância média
		intragrupo	intergrupo
I	1, 2, 3, 5 e 6	3,47	$D_{I,II} = 11,45$
II	7	-	$D_{I,III} = 12,65$
III	4		$D_{II,III} = 9,91$

$\theta = 9,55$ é o critério global de agrupamento.

Caso haja interesse, é possível investigar o padrão agrupamento dentro de um dos grupos já formados. O processo é idêntico ao descrito anteriormente, devendo-se apenas considerar uma nova matriz de dissimilaridade, que inclua apenas os acessos a serem estudados.

Como ilustração, será considerada a possibilidade de se obter uma partição dentro do grupo I, formado pelos genótipos 1,2,3, 5 e 6. A matriz a ser utilizada será:

$$D_{5 \times 5} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 5 \\ 6 \end{matrix} & \begin{bmatrix} 0,00 & & & & \\ 2,16 & 0,00 & & & \\ 0,98 & 2,99 & 0,00 & & \\ 3,00 & 6,55 & 1,44 & 0,00 & \\ 5,25 & 8,45 & 2,41 & 1,49 & 0,00 \end{bmatrix} \end{matrix} \text{ Simétrica}$$

As menores distâncias em relação a cada um dos genótipos são as seguintes:

Genótipo	1	2	3	5	6
mínimo $d_{ii'}$	0,98	2,16	0,98	1,44	1,49

Dessa forma, é estabelecido como critério local de agrupamento o valor $\theta = 2,16$ como limite de acréscimo, na média da distância dentro do grupo, para formação ou inclusão de um novo elemento no grupo.

Utilizando o mesmo procedimento, é obtido o resultado a seguir:

Grupo	Genótipos	Distância média	Distância média
		intragrupo	intergrupo
I	1, 2, 3, 5 e 6	3,47	$D_{I,II} = 11,45$
Ia	1,3	0,98	$D_{Ia,Ib} = 3,03$
Ib	5,6	1,49	$D_{Ia,Ic} = 2,58$
Ic	2	-	$D_{Ib,Ic} = 7,50$
II	7	-	$D_{I,III} = 12,65$
III	4		$D_{II,III} = 9,91$

$\theta = 9,55$ é o valor adotado como critério global de agrupamento. $\theta = 9,26$ é valor adotado como critério local de agrupamento entre genótipos do grupo I.

Uma outra ilustração de agrupamento pelo método de Tocher pode ser apresentada em relação aos dados dos 15 cultivares de milho, apresentados nas Tabelas A1-Anexo (dados originais) e 2.7 (médias dos cultivares), cujas medidas de dissimilaridade, expressas pela distância generalizada de Mahalanobis ou distância euclidiana média padronizada, são apresentadas na Tabela 2.9. Neste caso, são formados os grupos apresentados a seguir na Tabela 2.14. Pode ser verificado que o padrão de agrupamento foi coincidente, apesar de terem sido utilizadas medidas diferentes para expressar a diversidade genética dos cultivares analisados.

Tabela 2.14. Grupos de genótipos estabelecidos pelo método de Tocher, com base na dissimilaridade expressa pela distância de Mahalanobis ou euclidiana média padronizada

Grupos	<u>Distância de Mahalanobis</u>	Distância Euclidiana Média Padronizada
I	7, 9, 10, 8 e 6	6, 8, 9, 7 e 10
II	11, 15, 14, 13 e 12	13, 15, 11, 14 e 12
III	1, 4, 2, e 3	1, 4, 3 e 2
IV	5	5
Limite intergrupo (θ)	68,44	1,12

Método de Tocher Modificado

Este método, proposto por Vasconcelos et al (2007), difere do original pelo fato de ser adotado critério diferenciado para inclusão de acessos em cada grupo que é formado, de modo que o processo de agrupamento deixa de ser simultâneo e passa a ser seqüencial. O método requer a obtenção da matriz de dissimilaridade, sobre a qual é identificado o par de indivíduos mais similares. Estes indivíduos formarão o grupo inicial. A partir daí, é avaliada a possibilidade de inclusão de novos indivíduos, em que a decisão de incluir o indivíduo em um grupo é tomada por meio da comparação entre o acréscimo no valor médio da distância dentro do grupo e um nível máximo permitido, que é o valor máximo (θ_1) da medida de dissimilaridade encontrado no conjunto das menores distâncias envolvendo cada indivíduo.

Na formação do próximo grupo, o procedimento é similar, diferindo apenas do fato de que é o valor máximo (θ_2) da medida de dissimilaridade encontrado no conjunto das menores distâncias envolvendo cada indivíduo, porém excluindo as informações daqueles anteriormente já agrupados, e assim sucessivamente.

Destaca-se o fato de que, se houver apenas dois últimos acessos a serem investigados no agrupamento, o valor de θ_ℓ será assumido como aquele obtido para a formação do grupo anterior ($\theta_{\ell-1}$).

Vasconcelos et al (2007), para ilustrar e comparar os métodos de Tocher simultâneo (original) e seqüencia, realizaram simulações de coleções de acessos com diferentes características, tanto para avaliação individual quanto para avaliação de experimentos com repetições. O número de grupos formados com o método de Tocher seqüencial foi menor que o número de grupos formados pelo método de Tocher original. No método de Tocher seqüencial, não foi verificada influência dos genótipos já agrupados, no agrupamento dos genótipos mais distantes. O limite de acréscimo na média da distância intragrupo, estimado após a formação de um novo grupo, constituiu uma estimativa da dissimilaridade existente entre os acessos dos grupos. Segundo os autores, o agrupamento dos genótipos com maior

dissimilaridade foi realizado com maior eficácia pelo método de Tocher seqüencial do que pelo método de Tocher original.

2.3.2. Métodos de agrupamento seqüenciais, aglomerativos, hierárquicos, sem sobreposição

Estes métodos, também conhecidos como SAHN (do inglês, *seqüencial, agglomerative, hierarchical, nonoverlapping*), têm a necessidade, em cada estágio do agrupamento, de recalcular o coeficiente de dissimilaridade entre os grupos estabelecidos e os possíveis candidatos a futuras admissões no grupo. Além disso, reconsidera-se o critério de admissão de novos indivíduos aos grupos já estabelecidos (SNEATH; SOKAL, 1973).

Tais métodos se processam por séries sucessivas de fusões. Os n indivíduos são classificados em n-1, n-2,..., até 1 grupo, por sua vez, baseados em subdivisões hierárquicas operadas sobre a matriz de distância, permitindo a geração de uma árvore de classificação bidimensional, chamada dendrograma. Em todos os métodos SAHN, a cada novo estágio, o algoritmo de agrupamento baseia-se na menor distância entre todos os grupos (e/ou pares) de indivíduos como critério para a formação de um novo grupo.

O conceito de representação hierárquica dos dados foi desenvolvido primeiramente na biologia, em que, em muitos casos, principalmente os de taxonomia e evolução, é pressuposto existir uma estrutura hierárquica (MARDIA et al., 1979).

Atualmente, os métodos hierárquicos têm uso generalizado, incluindo aquelas áreas em que a estrutura hierárquica pode não ser a mais apropriada, e, obviamente, o seu uso necessita de avaliação mais criteriosa. No entanto, tais métodos têm sido rotineiramente empregados em estudos de diversidade genética. Geralmente, nessas situações, não há preocupação com o número ótimo de grupos, uma vez que o interesse maior está na “árvore” e nas ramificações que são obtidas. As delimitações podem ser estabelecidas por um exame visual do dendrograma, em que se avaliam pontos de alta mudança de nível, tomando-se em

geral como delimitadores do número de genótipo para determinado grupo, embora, em alguns casos, a alta mudança de nível seja um critério de difícil visualização, impedindo assim a interpretação dos grupos.

Vale ressaltar que a escolha de um método de agrupamento depende do material e dos objetivos em questão, pois métodos de agrupamento diferentes podem conduzir a resultados bem distintos. Não há método considerado o melhor, mas alguns são mais indicados para determinadas situações do que outros (KAUFMAM; ROSSEEUW, 1990). Além do algoritmo usado e do material avaliado, os resultados do agrupamento podem ser influenciados pelo coeficiente de dissimilaridade escolhido (JACKSON et al., 1989).

A seguir, serão apresentados os principais métodos de agrupamento SAHN.

Método da ligação simples (simple linkage) ou do vizinho mais próximo

Este método tem sido amplamente utilizado no melhoramento genético, apresentando como desvantagem o fato de não discernir grupos entre acessos pobemente separados (JOHNSON; WICHERN, 1988). Entretanto, tem sido considerado um dos poucos métodos que pode delinear grupos não-elipsóides, ou seja, neste método evita-se estabelecer grupos únicos. Quando os genótipos se dispõem numa estrutura de filamentos, conhecida por encadeamento, a formação de um único grupo, neste caso, pode ser enganosa, se os indivíduos em extremidades opostas da cadeia são, de fato, completamente dissimilares (SNEATH; SOKAL, 1973, JOHNSON; WICHERN, 1988).

Neste método, o dendrograma é estabelecido pelos genótipos com maior similaridade, sendo a distância entre um indivíduo k e um grupo, formado pelos indivíduos i e j, dada por:

$$d_{(ij)k} = \min \{d_{ik}; d_{jk}\}$$

ou seja, $d_{(ij)k}$ é dada pelo menor elemento do conjunto das distâncias dos pares de indivíduos (i e k) e (j e k). As conexões entre indivíduos e grupos, ou entre grupos, são feitas por ligações simples entre indivíduos, ou seja, a distância entre os grupos

é definida como sendo aquela entre os indivíduos mais parecidos dentre esses grupos.

A distância entre dois grupos é dada por:

$$d_{(ij)(kl)} = \min\{d_{ik}; d_{il}; d_{jk}; d_{jl}\}$$

ou seja, a distância entre dois grupos formados, respectivamente, pelos indivíduos (i e j) e (k e l) é dada pelo menor elemento do conjunto, cujos elementos são as distâncias entre os pares de indivíduos (i e k), (i e l), (j e k) e (j e l).

Ilustração

Considere a matriz de distâncias já apresentada anteriormente:

	1	2	3	4	5	6	7
1	0,00						
2	2,16	0,00					
3	0,98	2,99	0,00				
4	14,32	12,53	9,55	0,00			
5	3,00	6,55	1,44	12,75	0,00		
6	5,25	8,45	2,41	14,12	1,49	0,00	
7	14,66	17,56	10,33	9,91	6,37	10,34	0,00

No primeiro estágio identificam-se as entidades mais similares como sendo os genótipos 1 e 3, com distância de 0,98.

As distâncias entre esse grupo e os outros cinco genótipos são obtidas da seguinte maneira:

$$d_{(13)2} = \min(d_{12}; d_{23}) = \min(2,16; 2,99) = d_{12} = 2,16$$

$$d_{(13)4} = \min(d_{14}; d_{34}) = \min(14,32; 9,55) = d_{34} = 9,55$$

$$d_{(13)5} = \min(d_{15}; d_{35}) = \min(3,00; 1,44) = d_{35} = 1,44$$

$$d_{(13)6} = \min(d_{16}; d_{36}) = \min(5,25; 2,41) = d_{36} = 2,41$$

$$d_{(13)7} = \min(d_{17}; d_{37}) = \min(14,66; 10,33) = d_{37} = 10,33$$

Para qualquer método SAHN, se em qualquer estágio houver caso de empate, escolhe-se ao acaso uma das distâncias empatadas e continua-se o processo normalmente. Todavia, os algoritmos desenvolvidos para computação escolhem, em geral, o primeiro par de distâncias empatado.

Uma nova matriz de dissimilaridade é recalculada, compondo-se de distâncias entre genótipos e grupos:

	(1, 3)	2	4	5	6	7
(1, 3)	0,00					
2	2,16	0,00				
4	9,55	12,53	0,00			Simétrica
5	1,44	6,55	12,75	0,00		
6	2,41	8,45	14,12	1,49	0,00	
7	10,33	17,56	9,91	6,37	10,34	0,00

Estágio 2: Entidades mais similares: grupo (1, 3) e genótipo 5

Distância entre entidades: 1,44

As novas distâncias são estimadas por meio de:

$$d_{(135)2} = \min(d_{12}, d_{23}, d_{25}) = \min(2,16; 2,99; 3,00) = d_{12} = 2,16$$

$$d_{(135)4} = \min(d_{14}, d_{34}, d_{45}) = \min(14,32; 9,55; 12,75) = d_{34} = 9,55$$

$$d_{(135)6} = \min(d_{16}, d_{36}, d_{56}) = \min(5,25; 2,41; 1,49) = d_{56} = 1,49$$

$$d_{(135)7} = \min(d_{17}, d_{37}, d_{57}) = \min(14,66; 10,33; 6,37) = d_{57} = 6,37$$

Assim, é estabelecida a nova matriz de dissimilaridade:

	(1, 3, 5)	2	4	6	7
(1, 3, 5)	0,00				
2	2,16	0,00			
4	9,55	12,53	0,00		Simétrica
6	1,49	8,45	14,12	0,00	
7	6,37	17,56	9,91	10,34	0,00

Estágio 3: Entidades mais similares: grupo (1, 3, 5) e genótipo 6

Distância entre entidades: 1,49

As novas distâncias são estimadas por meio de:

$$d_{(135)2} = \min(d_{12}, d_{23}, d_{25}, d_{26}) = \min(2,16; 2,99; 6,55; 8,45) = d_{12} = 2,16$$

$$d_{(1356)4} = \min(d_{14}; d_{34}; d_{45}; d_{46}) = \min(14,32; 9,55; 12,75; 14,12) = d_{34} = 9,55$$

$$d_{(1356)7} = \min(d_{17}; d_{37}; d_{57}; d_{67}) = \min(14,66; 10,33; 6,37; 10,34) = d_{57} = 6,37$$

A nova matriz de dissimilaridade pode, então, ser reestruturada:

$$D_{4 \times 4} = \begin{array}{c|cccc} & (1,3,5,6) & 2 & 4 & 7 \\ \begin{matrix} (1,3,5,6) \\ 2 \\ 4 \\ 7 \end{matrix} & \left[\begin{array}{ccccc} 0,00 & & & & \\ 2,16 & 0,00 & & & \text{Simétrica} \\ 9,55 & 12,53 & 0,00 & & \\ 6,37 & 17,56 & 9,91 & 0,00 & \end{array} \right] \end{array}$$

Estágio 4: Entidades mais similares: grupo (1, 3, 5, 6) e genótipo 2

Distância entre entidades: 2,16

As novas distâncias são estimadas por meio de:

$$d_{(13562)4} = \min(d_{14}; d_{34}; d_{45}; d_{46}; d_{24}) = \min(14,32; 9,55; 12,75; 14,12; 12,53)$$

$$= d_{34} = 9,55$$

$$d_{(13562)7} = \min(d_{17}; d_{37}; d_{57}; d_{67}; d_{27}) = \min(14,66; 10,33; 6,37; 10,34; 17,56)$$

$$= d_{57} = 6,37$$

A nova matriz de dissimilaridade é, então, dada por:

$$D_{3 \times 3} = \begin{array}{c|cc} (1,3,5,6,2) & 4 & 7 \\ \begin{matrix} (1,3,5,6,2) \\ 4 \\ 7 \end{matrix} & \left[\begin{array}{ccc} 0,00 & & \\ 9,55 & 0,00 & \\ 6,37 & 17,56 & 0,00 \end{array} \right] \end{array}$$

Estágio 5: Entidades mais similares: grupo (1, 3, 5, 6, 2) e genótipo 7

Distância entre entidades: 6,37

A nova medida de distâncias é estabelecida:

$$\begin{aligned} d_{(135627)4} &= \min(d_{14}; d_{34}; d_{54}; d_{64}; d_{24}; d_{74}) \\ &= \min(14,32; 9,55; 12,75; 14,12; 12,53; 17,56) = d_{34} = 9,55 \end{aligned}$$

Dessa forma, considera-se a matriz de dissimilaridade:

$$D_{2x2} = \begin{pmatrix} (1, 3, 5, 6, 2, 7) & 4 \\ (1, 3, 5, 6, 2, 7) & 0,00 \\ 4 & 9,55 & 0,00 \end{pmatrix}$$

Estágio 6: Finalmente, o último grupo é formado por todos os indivíduos (genótipos). A distância entre o grupo (1, 3, 5, 6, 2, 7) e o genótipo 4 é de 9,55. As fusões em cada fase foram:

Estágio	Entidade X	Entidade Y	Níveis de fusões
1	1	3	0,98
2	1, 3	5	1,44
3	1, 3, 5	6	1,49
4	1, 3, 5, 6	2	2,16
5	1, 3, 5, 6, 2	7	6,37
6	1, 3, 5, 6, 2, 7	4	9,55

A correspondente representação gráfica é mostrada na Figura 2.4. Um corte feito à distância 7,1, que corresponde a 75% do valor da distância no último nível de fusão (9,55), possibilita estabelecer dois grupos hierárquicos. Um deles constituído pelo acesso 4 e o outro pelos demais acessos.

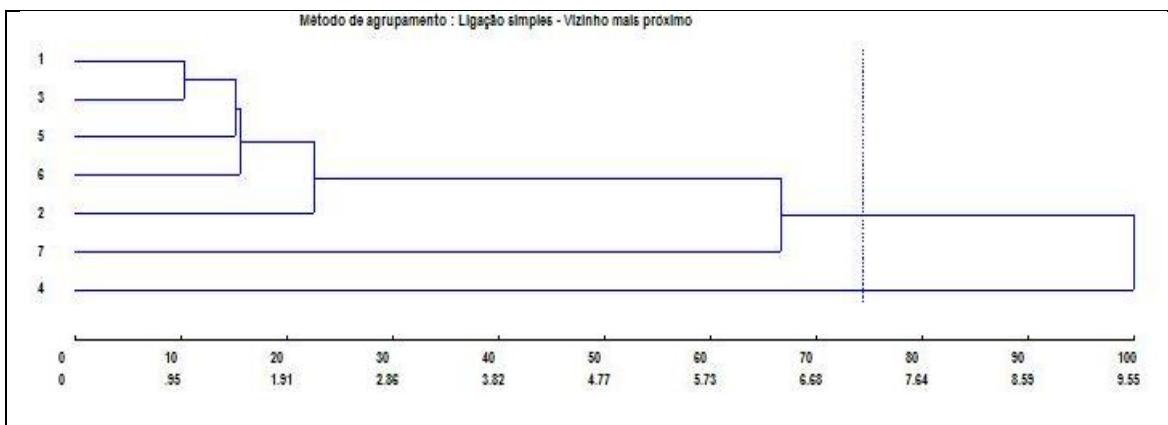


Figura 2.4 - Dendrograma obtido pelo método da ligação simples, baseado no quadrado da distância euclidiana média entre sete acessos. A primeira linha de valores no eixo da abscissa corresponde a valores percentuais em relação à dissimilaridade no último nível de fusão (9,55).

Algumas observações interessantes devem ser feitas:

- No método do vizinho mais próximo, a distância mínima entre acessos está incluída no dendrograma. No exemplo, a menor distância se verifica entre os acessos 1 e 3 e vale 0,98.
- A distância máxima entre os acessos não está representada no dendrograma. O valor da distância no último nível de fusão (9,55) é inferior à maior medida de dissimilaridade estimada, que foi de 17,56, estabelecida entre os acessos 2 e 7.
- A disposição dos acessos na vertical não deve ser considerada como um padrão fixo e, portanto, sem significado biológico. No dendrograma apresentado, o acesso 7 está próximo do 2 e distante dos acessos 1 e 3, mas isso é decorrente apenas de um arranjo gráfico, com vista à melhor estética. Pela matriz de distância, verifica-se que a distância entre os acessos 2 e 7 (17,56) é superior à verificada entre os acessos 7 e 1 (14,66) ou entre os acessos 7 e 3 (10,33).
- O agrupamento explora pouco a diversidade entre os acessos, tendo como referência de dissimilaridade total um pequeno valor da distância no último nível de fusão. Em certos casos, não é capaz de discernir grupos pobemente separados.

v. O padrão de agrupamento geralmente é do formato de encadeamento ou do “tipo escada”. O padrão de agrupamento é contínuo, sem clara distinção de formação de grupos discretos.

Método da ligação completa (complete linkage) ou do vizinho mais distante

O método da ligação completa é uma antítese ao método da ligação simples. No método da ligação completa, a similaridade entre dois grupos é dada pelos indivíduos de cada grupo que menos se parecem. Este método, geralmente, leva a grupos compactos e discretos, tendo os seus valores de similaridade relativamente pequenos.

A construção do dendrograma é similar ao método da ligação simples: a cada estágio são estabelecidos os indivíduos de maior similaridade. Entretanto, a distância entre um indivíduo k e um grupo, formado pelos indivíduos i e j, é fornecida por:

$$d_{(ij)k} = \max(d_{ik}, d_{jk})$$

ou seja, $d_{(ij)k}$ é dada pelo maior elemento do conjunto das distâncias dos pares de indivíduos (i e k) e (j e k).

A distância entre dois grupos é dada por:

$$d_{(ij)(kl)} = \max(d_{ik}, d_{il}; d_{jk}, d_{jl})$$

ou seja, a distância entre dois grupos, formados respectivamente pelos indivíduos (i e j) e (k e l), é determinada pelo maior elemento do conjunto, cujos elementos são distâncias entre pares de indivíduos de grupos (i e k), (i e l), (j e k) e (j e l).

Ilustração

Por ter filosofia semelhante à do método da ligação simples, serão apresentados apenas os três primeiros estágios do método da ligação completa, o que certamente não comprometerá seu entendimento. Aplicando esse método na mesma matriz da Ilustração 7, tem-se:

Estágio 1: Entidades mais similares: genótipos 1 e 3

Distância entre entidades: 0,98

As distâncias entre o grupo (1 e 3) e os demais genótipos são obtidas da seguinte maneira:

$$d_{(13)2} = \max(d_{12}; d_{23}) = \max(2,16; 2,99) = d_{32} = 2,99$$

$$d_{(13)4} = \max(d_{14}; d_{34}) = \max(14,32; 9,55) = d_{14} = 14,32$$

$$d_{(13)5} = \max(d_{15}; d_{35}) = \max(3,00; 1,44) = d_{15} = 3,00$$

$$d_{(13)6} = \max(d_{16}; d_{36}) = \max(5,25; 2,41) = d_{16} = 5,25$$

$$d_{(13)7} = \max(d_{17}; d_{37}) = \max(14,66; 10,33) = d_{17} = 14,66$$

Uma nova matriz de dissimilaridade é recalculada, compondo-se de distâncias entre genótipos e grupos:

$$D_{6 \times 6} = \begin{array}{c|cccccc} & (1, 3) & 2 & 4 & 5 & 6 & 7 \\ \hline (1, 3) & 0,00 & & & & & \\ 2 & 2,99 & 0,00 & & & & \\ 4 & 14,32 & 12,53 & 0,00 & & & \text{Simétrica} \\ 5 & 3,00 & 6,55 & 12,75 & 0,00 & & \\ 6 & 5,25 & 8,45 & 14,12 & 1,49 & 0,00 & \\ 7 & 14,66 & 17,56 & 9,91 & 6,37 & 10,34 & 0,00 \end{array}$$

Estágio 2: Entidades mais similares: genótipo 5 e genótipo 6

Distância entre entidades: 1,49

As novas distâncias são, então, estimadas:

$$d_{(56)(13)} = \max(d_{15}; d_{35}; d_{16}; d_{36}) = \max(3,00; 1,44; 5,25; 2,41) = d_{16} = 5,25$$

$$d_{(56)2} = \max(d_{25}; d_{26}) = \max(6,55; 8,45) = d_{26} = 8,45$$

$$d_{(56)4} = \max(d_{45}; d_{46}) = \max(12,75; 14,12) = d_{46} = 14,12$$

$$d_{(56)7} = \max(d_{57}; d_{67}) = \max(6,37; 10,34) = d_{67} = 10,34$$

A nova matriz de dissimilaridade é, portanto, dada por:

$$D_{5 \times 5} = \begin{array}{c} (1,3) \\ 2 \\ 4 \\ (5, 6) \\ 7 \end{array} \left[\begin{array}{ccccc} (1,3) & 2 & 4 & (5, 6) & 7 \\ 0,00 & & & & \\ 2,99 & 0,00 & & & \text{Simétrica} \\ 14,32 & 12,53 & 0,00 & & \\ 5,25 & 8,45 & 14,12 & 0,00 & \\ 14,66 & 17,56 & 9,91 & 10,34 & 0,00 \end{array} \right]$$

No terceiro estágio, as entidades mais similares foram o grupo (1, 3) e o genótipo 2. O processo continua até que seja formado um único grupo com todos os indivíduos. As demais fusões em cada estágio estão apresentadas a seguir:

Estágio	Entidade X	Entidade Y	Níveis de fusões
1	1	3	0,98
2	5	6	1,49
3	1,3	2	2,99
4	1, 3, 2	5, 6	8,45
5	4	7	9,91
6	1, 3, 2, 5, 6	4, 7	17,56

A correspondente representação gráfica é mostrada na Figura 2.5. Um corte feito à distância 13,17, que corresponde a 75% do valor da distância no último nível de fusão (17,56), possibilita estabelecer dois grupos hierárquicos: um deles constituído pelos acessos 4 e 7 e o outro pelos demais acessos.

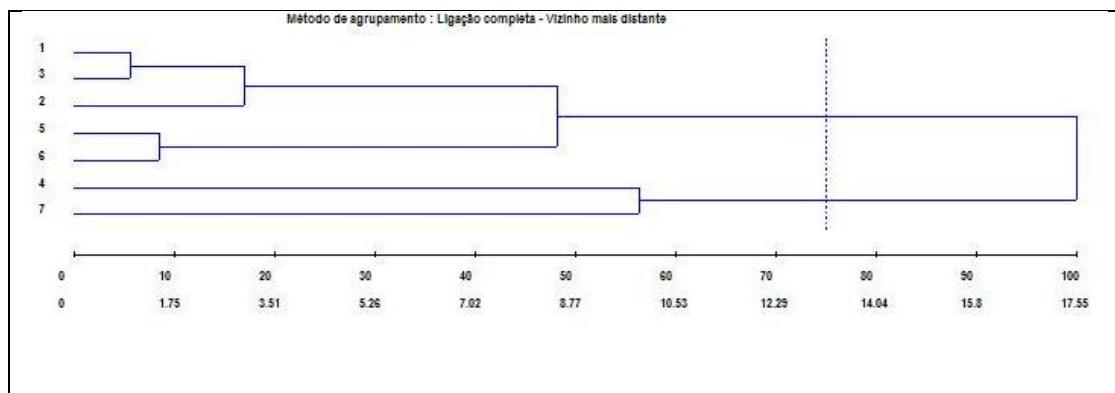


Figura 2.5 - Dendrograma obtido pelo método da ligação completa, baseado no quadrado da distância euclidiana média entre sete acessos. A primeira linha de valores no eixo da abscissa corresponde a valores percentuais em relação à dissimilaridade no último nível de fusão (17,56).

Algumas observações interessantes devem ser feitas:

- No método do vizinho mais distante, a distância mínima entre acessos está incluída no dendrograma. No exemplo, a menor distância se verifica entre os acessos 1 e 3 e vale 0,98.
- A distância máxima entre os acessos também está explicitamente representada no dendrograma. O valor da distância no último nível de fusão (17,56) é exatamente à maior medida de dissimilaridade estimada, que foi entre os acessos 2 e 7.
- A disposição dos acessos na vertical não deve ser considerada um padrão fixo e, portanto, sem significado biológico. Nenhuma método de agrupamento hierárquico tem por objetivo estabelecer projeção linear dos acessos quanto à suas dissimilaridades. A disposição dos acessos é decorrente de um arranjo gráfico apropriado, com vista à melhor estética.
- O agrupamento explora adequadamente a diversidade entre os acessos, tendo como referência de dissimilaridade total, expressa no último nível de fusão, o valor máximo da distância entre acessos. Há formação de grupos compactos e discretos.

Método da ligação média entre grupos ou UPGMA (*Unweighted pair-group method using arithmetic averages*)

O método da ligação média não-ponderada entre grupos, mais conhecido como UPGMA, tem sido utilizado com maior freqüência em ecologia e sistemática (JAMES; McCULLOCH, 1990) e em taxionomia numérica (SNEATH; SOKAL, 1973). Trata-se de uma técnica de agrupamento que utiliza as médias aritméticas (não-ponderadas) das medidas de dissimilaridade, evitando assim caracterizar a dissimilaridade por valores extremos (mínimo e máximo) entre os genótipos considerados.

Como regra geral, a construção do dendrograma é estabelecida pelo genótipo de maior similaridade. Entretanto, a distância entre um indivíduo k e um grupo, formado pelos indivíduos i e j, é fornecida por:

$$d_{(ij)k} = \text{média } (d_{ik}; d_{jk}) = \frac{d_{ik} + d_{jk}}{2}$$

ou seja, $d_{(ij)k}$ é dada pela média do conjunto das distâncias dos pares de indivíduos (i e k) e (j e k).

A distância entre dois grupos é dada por:

$$d_{(ij)(klm)} = \text{média } (d_{ik}; d_{il}; d_{im}; d_{jk}; d_{jl}; d_{jm}) = \frac{d_{ik} + d_{il} + d_{im} + d_{jk} + d_{jl} + d_{jm}}{6}$$

ou seja, a distância entre dois grupos, formados respectivamente pelos indivíduos (i e j) e (k, l e m), é determinada pela média entre os elementos do conjunto, cujos elementos são distâncias entre pares de indivíduos de grupos (i e k), (i e l), (i e m), (j e k), (j e l) e (j e m).

Uma expressão geral para a média não-ponderada entre grupos pode ser apresentada da seguinte maneira:

$$d_{(ij)k} = \frac{n_i}{n_i + n_j} d_{ik} + \frac{n_j}{n_i + n_j} d_{jk}$$

em que $d_{(ij)k}$ é definida como a distância entre o grupo (ij), com tamanho interno n_i e n_j , respectivamente, e o grupo k. Nesta expressão, caracterizam-se os indexadores

i, j e k como indivíduos ou grupos. Essa interpretação deverá ser a mesma para os métodos subseqüentes.

Assim, para o cálculo da distância $d_{(12)3}$, em que se considera o grupo formado pelos acessos 1 ($n_i = 1$) e 2 ($n_j = 1$), tem-se:

$$d_{(12)3} = \frac{1}{1+1} d_{13} + \frac{1}{1+1} d_{23} = \frac{d_{13} + d_{23}}{2}$$

Para o cálculo da distância $d_{(12,3)4}$, em que se considera o grupo formado pelos acessos 1 e 2 ($n_i = 2$) e 3 ($n_j=1$), tem-se:

$$d_{(12,3)4} = \frac{2}{2+1} d_{(12)4} + \frac{1}{2+1} d_{34} = \frac{2}{3} \left(\frac{1}{1+1} d_{14} + \frac{1}{1+1} d_{24} \right) + \frac{d_{34}}{3} = \frac{d_{14} + d_{24} + d_{34}}{3}$$

Ilustração

Será considerada a mesma matriz do exemplo anterior, de forma que se tenha:

Estágio 1: Entidades mais similares: genótipos 1 e 3

Distância entre entidades: 0,98

As distâncias entre o grupo 1 e 3 formado e os demais genótipos são obtidas da seguinte maneira:

$$d_{(13)2} = \text{média } (d_{12}; d_{23}) = (d_{12} + d_{23})/2 = (2,16 + 2,99)/2 = 2,575$$

Pela expressão geral, tem-se:

$$d_{(13)2} = \frac{1}{1+1} d_{12} + \frac{1}{1+1} d_{23} = (d_{12} + d_{23})/2 = 2,575$$

As demais distâncias são obtidas por:

$$d_{(13)4} = \text{média } (d_{14}; d_{34}) = (d_{14} + d_{34})/2 = (14,32 + 9,55)/2 = 11,935$$

$$d_{(13)5} = \text{média } (d_{15}; d_{35}) = (d_{15} + d_{35})/2 = (3,00 + 1,44)/2 = 2,22$$

$$d_{(13)6} = \text{média } (d_{16}; d_{36}) = (d_{16} + d_{36})/2 = (5,25 + 2,41)/2 = 3,83$$

$$d_{(13)7} = \text{média } (d_{17}; d_{37}) = (d_{17} + d_{37})/2 = (14,66 + 10,33)/2 = 12,495$$

A nova matriz de dissimilaridade é a seguinte:

$$D_{6 \times 6} = \begin{array}{c|cccccc} & (1, 3) & 2 & 4 & 5 & 6 & 7 \\ \hline (1, 3) & 0,00 & & & & & \\ 2 & 2,58 & 0,00 & & & & \\ 4 & 11,94 & 12,53 & 0,00 & & & \text{Simétrica} \\ 5 & 2,22 & 6,55 & 12,75 & 0,00 & & \\ 6 & 3,83 & 8,45 & 14,12 & 1,49 & 0,00 & \\ 7 & 12,50 & 17,56 & 9,91 & 6,37 & 10,34 & 0,00 \end{array}$$

Estágio 2: Entidades mais similares: genótipos 5 e 6

Distância entre entidades: 1,49

As novas distâncias são dadas por:

$$d_{(56)(13)} = \text{média } (d_{15}; d_{35}; d_{16}; d_{36}) = (3,00 + 1,44 + 5,25 + 2,41)/4 = 3,025$$

Pela expressão geral, tem-se:

$$d_{(ij)(k)} = d_{(56)(13)} = \frac{1}{1+1} d_{(5)(13)} + \frac{1}{1+1} d_{(6)(13)}$$

sendo:

$$\frac{1}{2} d_{(13)(5)} = \frac{1}{2} \left[\frac{1}{1+1} d_{15} + \frac{1}{1+1} d_{35} \right] = \frac{1}{2} \left[\frac{1}{2} (d_{15} + d_{35}) \right] = \frac{1}{4} (d_{15} + d_{35})$$

e

$$\frac{1}{2} d_{(13)(6)} = \frac{1}{2} \left[\frac{1}{1+1} d_{16} + \frac{1}{1+1} d_{36} \right] = \frac{1}{2} \left[\frac{1}{2} (d_{16} + d_{36}) \right] = \frac{1}{4} (d_{16} + d_{36})$$

logo:

$$d_{(56)(13)} = \frac{1}{4} (d_{15} + d_{35} + d_{16} + d_{36}) = 3,025$$

As demais distâncias são obtidas por:

$$d_{(56)2} = (d_{25} + d_{26})/2 = (6,55 + 8,45)/2 = 7,5$$

$$d_{(56)4} = (d_{45}; d_{46})/2 = (12,75 + 14,12)/2 = 13,435$$

$$d_{(56)7} = (d_{57}; d_{67})/2 = (6,37 + 10,34)/2 = 8,355$$

A nova matriz de dissimilaridade é, então, dada por:

$$D_{5 \times 5} = \begin{array}{ccccc} & (1,3) & 2 & 4 & (5,6) & 7 \\ (1,3) & 0,00 & & & & \\ 2 & 2,58 & 0,00 & & & \text{Simétrica} \\ 4 & 11,94 & 12,53 & 0,00 & & \\ (5,6) & 3,03 & 7,5 & 13,44 & 0,00 & \\ 7 & 12,50 & 17,56 & 9,91 & 8,36 & 0,00 \end{array}$$

Estágio 3: Entidades mais similares: grupo (1,3) e genótipo 2

Distância entre entidades: 2,58

As novas distâncias, como, por exemplo, $d_{(13,2)4}$, podem ser obtidas por meio de:

$$d_{(132)4} = \text{média } (d_{14}; d_{34}; d_{24}) = (14,32 + 9,55 + 12,53)/3 = 12,133$$

ou pela expressão geral:

$$\begin{aligned} d_{(i,j)k} = d_{(13,2)4} &= \frac{2}{2+1} d_{(13)4} + \frac{1}{1+2} d_{(2)4} = \frac{2}{3} \left[\frac{1}{2} (d_{14} + d_{34}) \right] + \frac{1}{3} d_{24} \\ &= \frac{1}{3} (d_{14} + d_{34} + d_{24}) = 12,133 \end{aligned}$$

As demais distâncias são obtidas da seguinte maneira:

$$\begin{aligned} d_{(132)(56)} &= \text{média } (d_{15} + d_{16} + d_{35} + d_{36} + d_{25} + d_{26}) \\ &= (3,00 + 5,25 + 1,44 + 2,41 + 6,55 + 8,45)/6 = 4,5167 \end{aligned}$$

$$d_{(132)7} = (14,66 + 10,33 + 17,56)/3 = 14,1833$$

A nova matriz de distância é, portanto, dada por:

$$D_{4 \times 4} = \begin{array}{ccccc} & (1,3,2) & 4 & (5,6) & 7 \\ (1,3,2) & 0,00 & & & \\ 4 & 12,13 & 0,00 & & \text{Simétrica} \\ (5,6) & 4,52 & 13,44 & 0,00 & \\ 7 & 14,18 & 9,91 & 8,36 & 0,00 \end{array}$$

Estágio 4: Entidades mais similares: grupos (1,3,2) e (5,6)

Distância entre entidades: 4,52

As novas distâncias são dadas por:

$$\begin{aligned}d_{(13256)4} &= \text{média } (d_{14}; d_{34}; d_{24}; d_{45}; d_{46}) \\&= (14,32 + 9,55 + 12,53 + 12,75 + 14,12)/5 = 12,654\end{aligned}$$

ou pela expressão geral, tendo-se:

$$d_{(132,56)4} = \frac{3}{5}d_{(132)4} + \frac{2}{5}d_{(56)4}$$

em que:

$$\frac{3}{5}d_{(132)4} = \frac{3}{5}\left(\frac{2}{3}d_{(13)4} + \frac{1}{3}d_{24}\right) = \frac{3}{5}\left[\frac{2}{3}\left(\frac{1}{2}d_{14} + \frac{1}{2}d_{34}\right) + \frac{1}{3}d_{24}\right] = \frac{1}{5}(d_{14} + d_{34} + d_{24})$$

e

$$\frac{2}{5}d_{(56)4} = \frac{2}{5}\left(\frac{1}{2}d_{45} + \frac{1}{2}d_{46}\right) = \frac{1}{5}(d_{45} + d_{46})$$

logo:

$$d_{(132,56)4} = \frac{1}{5}(d_{14} + d_{34} + d_{24} + d_{45} + d_{46}) = 12,654$$

Também pode ser obtido:

$$\begin{aligned}d_{(13256)7} &= \text{média } (d_{17}; d_{37}; d_{27}; d_{57}; d_{67}) = (14,66 + 10,33 + 17,56 + 6,37 + 10,34)/5 \\&= 11,852\end{aligned}$$

A nova matriz de dissimilaridade é dada por:

$$D_{3 \times 3} = \begin{array}{c|cc|c} & (1,3,2,5,6) & 4 & 7 \\ \begin{matrix} (1,3,2,5,6) \\ 4 \\ 7 \end{matrix} & \left[\begin{matrix} 0,00 & & \\ 12,65 & 0,00 & \\ 11,85 & 9,91 & 0,00 \end{matrix} \right] & & \text{Simétrica} \end{array}$$

Estágio 5: Entidades mais similares: genótipos 4 e 7

Distância entre entidades: 9,91

As novas distâncias podem ser obtidas pela média:

$$\begin{aligned} d_{(13256)47} &= \text{média } (d_{14}; d_{34}; d_{24}; d_{45}; d_{46}; d_{17}; d_{37}; d_{27}; d_{57}; d_{67}) \\ &= (14,32 + 9,55 + 12,53 + \dots + 6,37 + 10,34)/10 = 12,253 \end{aligned}$$

ou pela expressão geral, em que se tem:

$$d_{(132,56)47} = \frac{3}{5}d_{(132)47} + \frac{2}{5}d_{(56)47}$$

sendo:

$$\begin{aligned} \frac{3}{5}d_{(132)47} &= \frac{3}{5}\left\{\frac{2}{3}d_{(13)47} + \frac{1}{3}d_{(2)47}\right\} = \frac{3}{5}\left\{\frac{2}{3}\left[\frac{1}{2}(d_{(1)47} + d_{(3)47})\right] + \frac{1}{3}\left[\frac{1}{2}(d_{24} + d_{27})\right]\right\} \\ &= \frac{3}{5}\left\{\frac{2}{3}\left[\frac{1}{2}\left(\frac{1}{2}d_{14} + \frac{1}{2}d_{17}\right) + \frac{1}{2}\left(\frac{1}{2}d_{34} + \frac{1}{2}d_{37}\right)\right] + \frac{1}{3}\left[\frac{1}{2}(d_{24} + d_{27})\right]\right\} \\ &= \frac{1}{10}(d_{14} + d_{17} + d_{34} + d_{37} + d_{24} + d_{27}) \end{aligned}$$

$$\begin{aligned} \frac{2}{5}d_{(56)47} &= \frac{2}{5}\left\{\frac{1}{2}[d_{(5)47} + d_{(6)47}]\right\} = \frac{2}{5}\left\{\frac{1}{2}\left[\frac{1}{2}(d_{45} + d_{57}) + \frac{1}{2}(d_{46} + d_{67})\right]\right\} \\ &= \frac{1}{10}(d_{45} + d_{57} + d_{46} + d_{67}) \end{aligned}$$

logo:

$$d_{(132,56)47} = \frac{1}{10}(d_{14} + d_{17} + d_{34} + d_{37} + d_{24} + d_{27}) + \frac{1}{10}(d_{45} + d_{57} + d_{46} + d_{67}) = 12,253$$

A nova matriz de dissimilaridade é, então, dada por:

$$D_{2 \times 2} = \begin{pmatrix} (1,3,2,5,6) & (4,7) \\ (1,3,2,5,6) & 0,00 \\ (4,7) & 12,25 \\ & 0,00 \end{pmatrix}$$

Estágio 6: O último grupo é formado pela união do grupo (1,3,2,5,6) ao grupo (4,7), com distância 12,25. As fusões produzidas em cada estágio foram as seguintes:

Estágio	Entidade X	Entidade Y	Nível de fusões
1	1	3	0,98
2	5	6	1,49
3	1, 3	2	2,58
4	1, 3, 2	5, 6	4,52
5	4	7	9,91
6	1, 3, 2, 5, 6	4, 7	12,25

A correspondente representação gráfica é mostrada na Figura 2.6. Um corte feito à distância 9,18, que corresponde a 75% do valor da distância no último nível de fusão (12,25), possibilita estabelecer três grupos hierárquicos: primeiro é constituído pelo acesso 4; outro pelo acesso 7; e o último, pelos demais acessos.

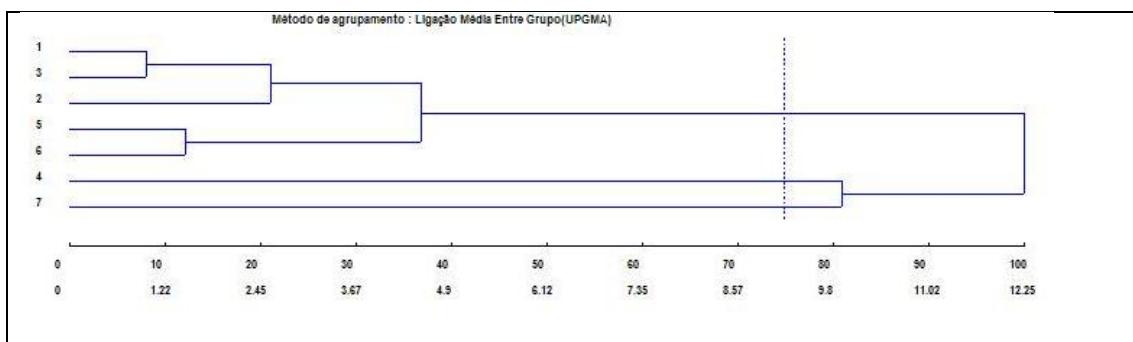


Figura 2.6 - Dendrograma obtido pelo método UPGMA, baseado no quadrado da distância euclidiana média entre sete acessos. A primeira linha de valores no eixo da abscissa corresponde a valores percentuais em relação à dissimilaridade no último nível de fusão (12,25).

Algumas observações interessantes devem ser feitas:

- No método UPGMA, a distância mínima entre acessos está incluída no dendrograma. No exemplo, a menor distância se verifica entre os acessos 1 e 3 e vale 0,98.
- A distância máxima entre os acessos também não está representada no dendrograma. O valor da distância no último nível de fusão (12,25) está abaixo da maior medida de dissimilaridade estimada, que foi entre os acessos 2 e 7, e acima

do nível de fusão estabelecido pelo método do vizinho mais próximo (9,55). Assim, nesta metodologia evita-se caracterizar a dissimilaridade pelos valores extremos (máximo ou mínimo) de dissimilaridade entre acessos.

iii. A disposição dos acessos na vertical não deve ser considerada um padrão fixo e, portanto, sem significado biológico. A disposição dos acessos é decorrente de um arranjo gráfico apropriado, com vista à melhor estética.

iv. O agrupamento explora adequadamente a diversidade entre os acessos, porém adotar, como referência de dissimilaridade total, um valor inferior à distância máxima.

Método da ligação média ponderada entre grupos ou WPGMA (*weighted pair-group method using arithmetic averages*)

Neste método, a cada passo, obtém-se uma média ponderada dos novos coeficientes de distância, que serão recalculados para compor a “nova” matriz de distância.

A expressão para o cálculo da média ponderada é dada por:

$$d_{(ij)k} = \frac{d_{ik} + d_{jk}}{2}$$

Assim, para o cálculo da distância $d_{(12)3}$ tem-se:

$$d_{(12)3} = \frac{d_{13} + d_{23}}{2}$$

Para o cálculo da distância $d_{(12,3)4}$, em que se considera o grupo formado pelos acessos 1 e 2 com a junção do 3, tem-se:

$$d_{(12,3)4} = \frac{d_{(12)4} + d_{34}}{2} = \frac{\frac{d_{14} + d_{24}}{2} + d_{34}}{2} = \frac{d_{14} + d_{24} + 2d_{34}}{4}$$

Ilustração

Novamente considerando os dados da matriz anterior, obtém-se:

Estágio 1: Entidades mais similares: genótipos 1 e 3

Distância entre entidades: 0,98

As distâncias entre o grupo 1 e 3 formado e os demais genótipos são obtidas da seguinte maneira:

$$d_{(1,3)2} = (d_{12} + d_{23})/2 = (2,16 + 2,99)/2 = 2,575$$

$$d_{(1,3)4} = (d_{14} + d_{34})/2 = (14,32 + 9,55)/2 = 11,935$$

$$d_{(1,3)5} = (d_{15} + d_{35})/2 = (3,00 + 1,44)/2 = 2,22$$

$$d_{(1,3)6} = (d_{16} + d_{36})/2 = (5,25 + 2,41)/2 = 3,83$$

$$d_{(1,3)7} = (d_{17} + d_{37})/2 = (14,66 + 10,33)/2 = 12,495$$

A nova matriz de dissimilaridade é a seguinte:

$$D_{6 \times 6} = \begin{array}{c|cccccc} & (1, 3) & 2 & 4 & 5 & 6 & 7 \\ \hline (1, 3) & 0,00 & & & & & \\ 2 & 2,58 & 0,00 & & & & \\ 4 & 11,94 & 12,53 & 0,00 & & & \text{Simétrica} \\ 5 & 2,22 & 6,55 & 12,75 & 0,00 & & \\ 6 & 3,83 & 8,45 & 14,12 & 1,49 & 0,00 & \\ 7 & 12,50 & 17,56 & 9,91 & 6,37 & 10,34 & 0,00 \end{array}$$

Estágio 2: Entidades mais similares: genótipos 5 e 6

Distância entre entidades: 1,49

As novas distâncias podem ser calculadas por meio de:

$$d_{(5,6)(13)} = \frac{d_{5(13)} + d_{6(13)}}{2} = (2,22 + 3,83)/2 = 3,025$$

$$d_{(5,6)2} = (d_{25} + d_{26})/2 = (6,55 + 8,45)/2 = 7,5$$

$$d_{(5,6)4} = (d_{45}; d_{46})/2 = (12,75 + 14,12)/2 = 13,435$$

$$d_{(5,6)7} = (d_{57}; d_{67})/2 = (6,37 + 10,34)/2 = 8,355$$

A nova matriz de dissimilaridade é, portanto, dada por:

$$D_{5 \times 5} = \begin{array}{ccccc} & (1,3) & 2 & 4 & (5, 6) & 7 \\ \begin{matrix} (1,3) \\ 2 \\ 4 \\ (5, 6) \\ 7 \end{matrix} & \left[\begin{array}{ccccc} 0,00 & & & & \\ 2,58 & 0,00 & & & \text{Simétrica} \\ 11,94 & 12,53 & 0,00 & & \\ 3,03 & 7,5 & 13,44 & 0,00 & \\ 12,50 & 17,56 & 9,91 & 8,36 & 0,00 \end{array} \right] \end{array}$$

Estágio 3: Entidades mais similares: grupo (1,3) e genótipo 2

Distância entre entidades: 2,58

As novas distâncias são facilmente calculadas por:

$$d_{(13,2)4} = \frac{d_{(13)4} + d_{24}}{2} = (11,935 + 12,53)/2 = 12,2325$$

$$d_{(13,2)(56)} = \frac{d_{(13)(56)} + d_{2(56)}}{2} = (3,025 + 7,5)/2 = 5,2625$$

$$d_{(13,2)7} = \frac{d_{(13)7} + d_{27}}{2} = (12,495 + 17,56)/2 = 15,0275$$

A nova matriz de distância é fornecida por:

$$D_{4 \times 4} = \begin{array}{ccccc} & (1,3,2) & 4 & (5,6) & 7 \\ \begin{matrix} (1,3,2) \\ 4 \\ (5,6) \\ 7 \end{matrix} & \left[\begin{array}{ccccc} 0,00 & & & & \\ 12,13 & 0,00 & & & \text{Simétrica} \\ 5,26 & 13,44 & 0,00 & & \\ 14,18 & 9,91 & 8,36 & 0,00 & \end{array} \right] \end{array}$$

Estágio 4: Entidades mais similares: grupos (1,3,2) e (5,6)

Distância entre entidades: 5,26

Tem-se, agora, as novas distâncias, estimadas por:

$$d_{(132,56)4} = \frac{d_{(132)4} + d_{(56)4}}{2} = (12,2325 + 13,435)/2 = 12,8338$$

$$d_{(132,56)7} = \frac{d_{(132)7} + d_{(56)7}}{2} = (15,0275 + 8,355)/2 = 11,6913$$

A nova matriz de dissimilaridade é, então, estabelecida:

$$D_{3 \times 3} = \begin{matrix} & (1,3,2,5,6) & 4 & 7 \\ (1,3,2,5,6) & \left[\begin{matrix} 0,00 & & \\ 12,83 & 0,00 & \\ 11,69 & 9,91 & 0,00 \end{matrix} \right] \\ 4 & & & \text{Simétrica} \\ 7 & & & \end{matrix}$$

Estágio 5: Entidades mais similares: genótipos 4 e 7

Distância entre entidades: 9,91

Por fim, a nova distância será:

$$d_{(47)(13256)} = \frac{d_{4(13256)} + d_{7(13256)}}{2} = (11,6913 + 12,8338)/2 = 12,2626$$

A matriz final de dissimilaridade é:

$$D_{2 \times 2} = \begin{matrix} & (1,3,2,5,6) & (4,7) \\ (1,3,2,5,6) & \left[\begin{matrix} 0,00 & \\ 12,26 & 0,00 \end{matrix} \right] \\ (4,7) & & \end{matrix}$$

Estágio 6: O último grupo é formado pela união do grupo (1, 3, 2, 5, 6) ao grupo (4,7), com distância 12,26. As fusões produzidas em cada estágio foram as seguintes:

Estágio	Entidade X	Entidade Y	Níveis de fusões
1	1	3	0,98
2	5	6	1,49
3	1, 3	2	2,58
4	1, 3, 2	5, 6	5,26
5	4	7	9,91
6	1, 3, 2, 5, 6	4, 7	12,26

A correspondente representação gráfica é mostrada na Figura 2.7. Um corte feito à distância 9,19, que corresponde a 75% do valor da distância no último nível de fusão (12,26), possibilita estabelecer três grupos hierárquicos: um deles é constituído pelo acesso 4; outro, pelo acesso 7; e o último pelos demais acessos.

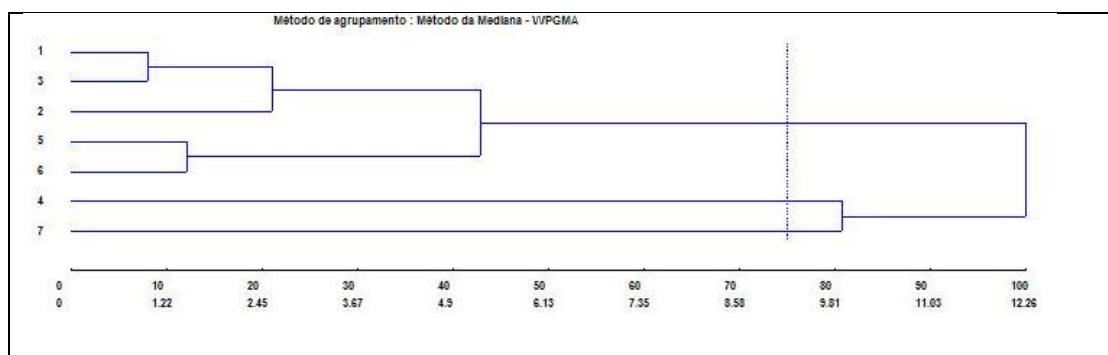


Figura 2.7 - Dendrograma obtido pelo método WPGMA, baseado no quadrado da distância euclidiana média entre sete acessos. A primeira linha de valores no eixo da abscissa corresponde a valores percentuais em relação à dissimilaridade no último nível de fusão (12,26).

Algumas observações interessantes devem ser feitas:

i. Os métodos UPGMA e WPGMA apresentam resultados bastante próximos.

Veja, no exemplo, as coincidências na evolução dos agrupamentos e a concordância nos valores dos níveis de fusão:

Estágio	Entidade X	Entidade Y	Níveis de fusões	Níveis de fusões
			WPGMA	UPGMA
1	1	3	0,98	0,98
2	5	6	1,49	1,49
3	1,3	2	2,58	2,58
4	1,3,2	5,6	5,26	4,52
5	4	7	9,91	9,91
6	1,3,2,5,6	4,7	12,26	12,25

ii. No método WPGMA, a distância mínima entre acessos está incluída no dendrograma. No exemplo, a menor distância se verifica entre os acessos 1 e 3 e vale 0,98.

iii. A distância máxima entre os acessos também não está representada no dendrograma. O valor da distância no último nível de fusão (12,26) está abaixo da maior medida de dissimilaridade estimada, que foi entre os acessos 2 e 7, e acima do nível de fusão estabelecido pelo método do vizinho mais próximo (9,55). Assim, neste método evita-se caracterizar a dissimilaridade pelos valores extremos (máximo ou mínimo) de dissimilaridade entre acessos.

iv. A disposição dos acessos na vertical não deve ser considerada um padrão fixo e, portanto, sem significado biológico. A disposição dos acessos é decorrente de um arranjo gráfico apropriado, com vista à melhor estética.

v. O agrupamento explora adequadamente a diversidade entre os acessos, porém adote, como referência de dissimilaridade total um valor inferior à distância máxima.

Método do centróide ou UPGMC (*Unweighted pair-group centroid method*)

No método do centróide, também conhecido como método do centróide não-ponderado, o objetivo é representar as distâncias entre grupos através do centróide dos indivíduos pertencentes aos seus respectivos grupos. No entanto, não produz resultados monotônicos, ou seja, os níveis de fusões não variam numa mesma proporção. Esse aspecto é constatado no exemplo a seguir.

A expressão para o cálculo do centróide é a seguinte:

$$d_{(ij)k} = \left[\frac{n_i}{n_i + n_j} \right] d_{ik} + \left[\frac{n_j}{n_j + n_i} \right] d_{jk} - \left[\frac{n_i \cdot n_j}{(n_i + n_j)^2} \right] d_{ij}$$

Ilustração

Considere a mesma matriz das ilustrações anteriores.

No estágio 1, as entidades mais similares foram os genótipos 1 e 3 com distância de 0,98

As distâncias entre o grupo 1 e 3 e os demais genótipos são:

$$d_{(13)2} = \frac{1}{2} d_{12} + \frac{1}{2} d_{23} - \frac{1}{(1+1)^2} d_{13} = \frac{2,16 + 2,99}{2} - \frac{0,98}{4} = 2,33$$

$$d_{(13)4} = \frac{1}{2} d_{14} + \frac{1}{2} d_{34} - \frac{1}{(1+1)^2} d_{13} = \frac{14,32 + 9,55}{2} - \frac{0,98}{4} = 11,69$$

$$d_{(13)5} = \frac{1}{2} d_{15} + \frac{1}{2} d_{35} - \frac{1}{(1+1)^2} d_{13} = \frac{3,00 + 1,44}{2} - \frac{0,98}{4} = 1,975$$

$$d_{(13)6} = \frac{1}{2} d_{16} + \frac{1}{2} d_{36} - \frac{1}{(1+1)^2} d_{13} = \frac{5,25 + 2,41}{2} - \frac{0,98}{4} = 3,585$$

$$d_{(13)7} = \frac{1}{2} d_{17} + \frac{1}{2} d_{37} - \frac{1}{(1+1)^2} d_{13} = \frac{14,66 + 10,33}{2} - \frac{0,98}{4} = 12,25$$

A nova matriz de dissimilaridade é a seguinte:

$$D_{6 \times 6} = \begin{array}{c|cccccc} & (1, 3) & 2 & 4 & 5 & 6 & 7 \\ \hline (1, 3) & 0,00 & & & & & \\ 2 & 2,33 & 0,00 & & & & \\ 4 & 11,69 & 12,53 & 0,00 & & & \text{Simétrica} \\ 5 & 1,98 & 6,55 & 12,75 & 0,00 & & \\ 6 & 3,59 & 8,45 & 14,12 & 1,49 & 0,00 & \\ 7 & 12,25 & 17,56 & 9,91 & 6,37 & 10,34 & 0,00 \end{array}$$

Estágio 2: Entidades mais similares: genótipos 5 e 6

Distância entre entidades: 1,49

As novas distâncias são obtidas por meio de:

$$d_{(5,6)(13)} = \frac{1}{2} d_{5(13)} + \frac{1}{2} d_{6(13)} - \frac{1}{(1+1)^2} d_{13} = \frac{1,975 + 3,585}{2} - \frac{1,49}{4} = 2,4075$$

$$d_{(5,6)2} = \frac{1}{2} d_{25} + \frac{1}{2} d_{26} - \frac{1}{(1+1)^2} d_{56} = \frac{6,55 + 8,45}{2} - \frac{1,49}{4} = 7,1275$$

$$d_{(5,6)4} = \frac{1}{2} d_{45} + \frac{1}{2} d_{46} - \frac{1}{(1+1)^2} d_{56} = \frac{12,75 + 14,12}{2} - \frac{1,49}{4} = 13,0625$$

$$d_{(5,6)7} = \frac{1}{2} d_{57} + \frac{1}{2} d_{67} - \frac{1}{(1+1)^2} d_{56} = \frac{6,37 + 10,34}{2} - \frac{1,49}{4} = 7,9825$$

A nova matriz de dissimilaridade é, portanto, dada por:

$$D_{5 \times 5} = \begin{array}{c|ccccc} & (1, 3) & 2 & 4 & (5, 6) & 7 \\ \hline (1, 3) & 0,00 & & & & \\ 2 & 2,33 & 0,00 & & & \text{Simétrica} \\ 4 & 11,69 & 12,53 & 0,00 & & \\ (5, 6) & 2,41 & 7,13 & 13,06 & 0,00 & \\ 7 & 12,25 & 17,56 & 9,91 & 7,98 & 0,00 \end{array}$$

Estágio 3: Entidades mais similares: grupo (1,3) e genótipo 2

Distância entre entidades: 2,33

As novas distâncias serão:

$$d_{(13,2)4} = \frac{2}{3}d_{(13)4} + \frac{1}{3}d_{24} - \frac{2 \times 1}{(2+1)^2} d_{13,2} = \frac{2(11,69) + (12,53)}{3} - \frac{2}{9} \cdot 2,33 \\ = 11,4522$$

$$d_{(13,2)(56)} = \frac{2}{3}d_{(13)56} + \frac{1}{3}d_{2(56)} - \frac{2 \times 1}{(2+1)^2} d_{13,2} = \frac{2(2,4075) + 7,1275}{3} - \frac{2}{9} \cdot 2,33 \\ = 3,4631$$

$$d_{(13,2)7} = \frac{2}{3}d_{(13)7} + \frac{1}{3}d_{27} - \frac{2 \times 1}{(2+1)^2} d_{13,2} = \frac{2(12,25) + (17,56)}{3} - \frac{2}{9} \cdot 2,33 \\ = 13,5022$$

A nova matriz de distância é, então, reestruturada:

$$D_{4 \times 4} = \begin{array}{ccccc} & (1,3,2) & 4 & (5,6) & 7 \\ \begin{matrix} (1,3,2) \\ 4 \\ (5,6) \\ 7 \end{matrix} & \left[\begin{array}{cccc} 0,00 & & & \\ 11,45 & 0,00 & \text{Simétrica} & \\ 3,46 & 13,06 & 0,00 & \\ 13,50 & 9,91 & 7,98 & 0,00 \end{array} \right] \end{array}$$

Estágio 4: Entidades mais similares: grupos (1,3,2) e (5,6)

Distância entre entidades: 3,4631

As novas distâncias são dadas por:

$$d_{(132,56)4} = \frac{3}{5}d_{(132)4} + \frac{2}{5}d_{(56)4} - \frac{3 \cdot 2}{(3+2)^2} d_{132,56} = \\ = \frac{3(11,4522) + 2(13,0625)}{5} - \frac{6}{25} \cdot 3,4631 = 11,2652$$

$$d_{(132,56)7} = \frac{3}{5}d_{(132)7} + \frac{2}{5}d_{(56)7} - \frac{3 \cdot 2}{(3+2)^2} d_{132,56} = \\ = \frac{3(13,5022) + 2(7,9825)}{5} - \frac{6}{25} \cdot 3,4631 = 10,4632$$

A nova matriz de dissimilaridade é, então, dada por:

$$D_{3 \times 3} = \begin{matrix} & (1,3,2,5,6) & 4 & 7 \\ (1,3,2,5,6) & \left[\begin{matrix} 0,00 & & \\ 11,27 & 0,00 & \\ 10,46 & 9,91 & 0,00 \end{matrix} \right] \\ 4 & & & \text{Simétrica} \\ 7 & & & \end{matrix}$$

Estágio 5: Entidades mais similares: genótipos 4 e 7

Distância entre entidades: 9,91

Por fim, a nova distância é dada por:

$$\begin{aligned} d_{(4,7)(13256)} &= \frac{1}{2}d_{4(13256)} + \frac{1}{2}d_{7(13256)} - \frac{1}{(1+1)^2}d_7 \\ &= \frac{11,2652 + 10,4632}{2} - \frac{1}{4}9,91 = 8,3867 \end{aligned}$$

Nova matriz de dissimilaridade:

$$D_{2 \times 2} = \begin{matrix} & (1,3,2,5,6) & (4,7) \\ (1,3,2,5,6) & \left[\begin{matrix} 0,00 & \\ 8,39 & 0,00 \end{matrix} \right] \\ (4,7) & & \end{matrix}$$

Estágio 6: O último grupo é formado pela união do grupo (1, 3, 2, 5, 6) ao grupo (4, 7), com distância 8,39. As fusões produzidas em cada estágio estão descritas a seguir:

Estágio	Entidade X	Entidade Y	Níveis de fusões
1	1	3	0,98
2	5	6	1,49
3	1, 3	2	2,33
4	1, 3, 2	5, 6	3,46
5	4	7	9,91
6	1, 3, 2, 5, 6	4, 7	8,39

A correspondente representação gráfica é mostrada na Figura 2.8.

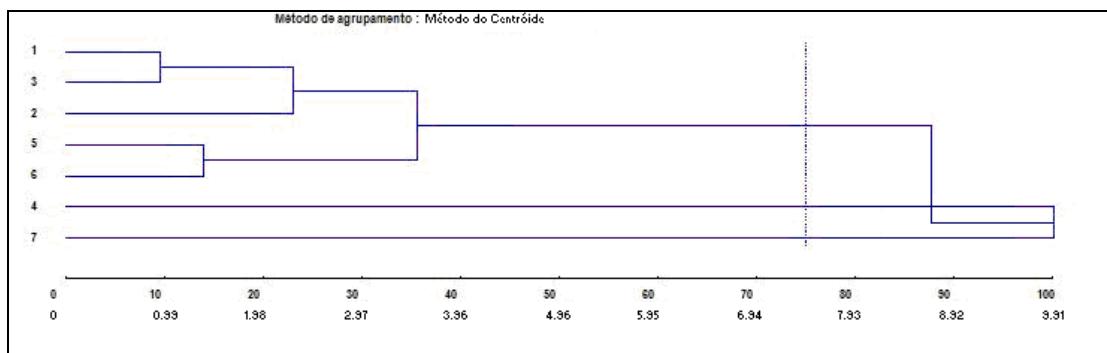


Figura 2.8 - Dendrograma obtido pelo método do centróide, baseado no quadrado da distância euclidiana média entre sete acessos.

Algumas observações interessantes devem ser feitas:

- i. Os métodos UPGMC, WPGMA e do centróide apresentam resultados bastante próximos. Veja, para o exemplo, as coincidências na evolução dos agrupamentos e a concordância nos valores dos níveis de fusão:

Estágio	Níveis de fusão				
	Entidade X	Entidade Y	Centróide	WPGMA	UPGMA
1	1	3	0,98	0,98	0,98
2	5	6	1,49	1,49	1,49
3	1,3	2	2,33	2,58	2,58
4	1,3,2	5,6	3,46	5,26	4,52
5	4	7	9,91	9,91	9,91
6	1,3,2,5,6	4,7	8,39	12,26	12,25

- ii. No método do centróide, a distância mínima entre acessos está incluída no dendrograma. No exemplo, a menor distância se verifica entre os acessos 1 e 3 e vale 0,98.

- iii. A distância máxima entre os acessos também não está representada no dendrograma. O valor máximo da distância entre os níveis de fusão (9,91) está abaixo da maior medida de dissimilaridade estimada, que foi entre os acessos 2 e 7, e até mesmo inferior ao nível de fusão máximo estabelecido pelo método do vizinho

mais próximo (9,55). Há possibilidade de o valor da distância do nível de fusão de um estágio de agrupamento anterior ser superior ao de um valor no nível posterior, como ocorre no presente exemplo, em relação aos estágios 5 e 6. Essa particularidade causa uma conformação diferenciada no dendrograma, conforme ilustrado pela Figura 2.8.

- iv. A disposição dos acessos na vertical não deve ser considerada um padrão fixo e, portanto, sem significado biológico. A disposição dos acessos é decorrente de um arranjo gráfico apropriado, visando melhor estética.
- v. Apesar de a diversidade máxima ter valor pequeno, o agrupamento possibilita a formação de grupos compactos e discretos.

Método da mediana ou WPGMC (weighted pair-group centroid method)

O método do centróide ponderado, também chamado de método da mediana, foi proposto por Gower (1967) com o intuito de evitar a contribuição desigual dos centróides de grupos diferentes na constituição de um novo centróide (grupo) candidato. Khattree e Naik (2000) relatam que, se um grupo for considerado muito pequeno, em termos de números de indivíduos, a sua contribuição para a formação de um grupo com novo centróide não será muito diferente da média (centróide) do grupo com maior número de indivíduos.

A mediana pode ser obtida pela seguinte expressão:

$$d_{(i,j)k} = \frac{d_{ik} + d_{jk}}{2} - \frac{1}{4} d_{ij}$$

Ilustração

Inicialmente, as entidades mais similares são os genótipos 1 e 3

Distância entre entidades: 0,98

As demais distâncias dos indivíduos com o grupo 1 e 3 são as seguintes:

$$d_{(13)2} = \frac{d_{12} + d_{23}}{2} - \frac{1}{4} d_{13} = \frac{2,16 + 2,99}{2} - \frac{0,98}{4} = 2,33$$

$$d_{(13)4} = \frac{d_{14} + d_{34}}{2} - \frac{1}{4}d_{13} = \frac{14,32 + 9,55}{2} - \frac{0,98}{4} = 11,69$$

$$d_{(13)5} = \frac{d_{15} + d_{35}}{2} - \frac{1}{4}d_{13} = \frac{3,00 + 1,44}{2} - \frac{0,98}{4} = 1,975$$

$$d_{(13)6} = \frac{d_{16} + d_{36}}{2} - \frac{1}{4}d_{13} = \frac{5,25 + 2,41}{2} - \frac{0,98}{4} = 3,585$$

$$d_{(13)7} = \frac{d_{17} + d_{37}}{2} - \frac{1}{4}d_{13} = \frac{14,66 + 10,33}{2} - \frac{0,98}{4} = 12,25$$

A nova matriz de dissimilaridade é a seguinte:

$$D_{6 \times 6} = \begin{array}{c|cccccc} & (1, 3) & 2 & 4 & 5 & 6 & 7 \\ \hline (1, 3) & 0,00 & & & & & \\ 2 & 2,33 & 0,00 & & & & \\ 4 & 11,69 & 12,53 & 0,00 & & & \text{Simétrica} \\ 5 & 1,98 & 6,55 & 12,75 & 0,00 & & \\ 6 & 3,59 & 8,45 & 14,12 & 1,49 & 0,00 & \\ 7 & 12,25 & 17,56 & 9,91 & 6,37 & 10,34 & 0,00 \end{array}$$

Estágio 2: Entidades mais similares: genótipos 5 e 6

Distância entre entidades: 1,49

As novas distâncias são dadas por:

$$d_{(5,6)(13)} = \frac{1}{2}d_{5(13)} + \frac{1}{2}d_{6(13)} - \frac{1}{4}d_{56} = \frac{1,975 + 3,585}{2} - \frac{1,49}{4} = 2,4075$$

$$d_{(5,6)2} = \frac{1}{2}d_{25} + \frac{1}{2}d_{26} - \frac{1}{4}d_{56} = \frac{6,55 + 8,45}{2} - \frac{1,49}{4} = 7,1275$$

$$d_{(5,6)4} = \frac{1}{2}d_{45} + \frac{1}{2}d_{46} - \frac{1}{4}d_{56} = \frac{12,75 + 14,12}{2} - \frac{1,49}{4} = 13,0625$$

$$d_{(5,6)7} = \frac{1}{2}d_{57} + \frac{1}{2}d_{67} - \frac{1}{4}d_{56} = \frac{6,37 + 10,34}{2} - \frac{1,49}{4} = 7,9825$$

A nova matriz de dissimilaridade é, portanto, dada por:

$$D_{5 \times 5} = \begin{array}{ccccc} & (1,3) & 2 & 4 & (5, 6) & 7 \\ \begin{matrix} (1,3) \\ 2 \\ 4 \\ (5, 6) \\ 7 \end{matrix} & \left[\begin{array}{ccccc} 0,00 & & & & \\ 2,33 & 0,00 & & & \text{Simétrica} \\ 11,69 & 12,53 & 0,00 & & \\ 2,41 & 7,13 & 13,06 & 0,00 & \\ 12,25 & 17,56 & 9,91 & 7,98 & 0,00 \end{array} \right] \end{array}$$

Estágio 3: Entidades mais similares: grupo (1,3) e genótipo 2

Distância entre entidades: 2,33

Têm-se as estimativas das novas distâncias, dadas por:

$$d_{(13,2)4} = \frac{1}{2}d_{(13)4} + \frac{1}{2}d_{24} - \frac{1}{4}d_{13,2} = \frac{(11,69) + (12,53)}{2} - \frac{1}{4} \cdot 2,33 = 11,5275$$

$$d_{(13,2)(56)} = \frac{1}{2}d_{(13)56} + \frac{1}{2}d_{2(56)} - \frac{1}{4}d_{13,2} = \frac{2,4075 + 7,1275}{2} - \frac{1}{4} \cdot 2,33 = 4,185$$

$$d_{(13,2)7} = \frac{1}{2}d_{(13)7} + \frac{1}{2}d_{27} - \frac{1}{4}d_{13,2} = \frac{(12,25) + (17,56)}{2} - \frac{1}{4} \cdot 2,33 = 14,3225$$

A partir desses valores, a matriz de distância é novamente estabelecida:

$$D_{4 \times 4} = \begin{array}{ccccc} & (1,3,2) & 4 & (5,6) & 7 \\ \begin{matrix} (1,3,2) \\ 4 \\ (5,6) \\ 7 \end{matrix} & \left[\begin{array}{ccccc} 0,00 & & & & \\ 11,53 & 0,00 & & & \text{Simétrica} \\ 4,19 & 13,06 & 0,00 & & \\ 14,32 & 9,91 & 7,98 & 0,00 & \end{array} \right] \end{array}$$

Estágio 4: Entidades mais similares: grupos (1, 3, 2) e (5, 6)

Distância entre entidades: 4,19

As estimativas das novas distâncias são apresentadas a seguir:

$$\begin{aligned} d_{(132,56)4} &= \frac{1}{2}d_{(132)4} + \frac{1}{2}d_{(56)4} - \frac{1}{4}d_{132,56} = \frac{(11,5275) + (13,0625)}{2} - \frac{1}{4} \cdot 4,185 = \\ &= 11,2486 \end{aligned}$$

$$d_{(132,56)7} = \frac{1}{2}d_{(132)7} + \frac{1}{2}d_{(56)7} - \frac{1}{4}d_{132,56} = \frac{(14,3225) + (7,9825)}{2} - \frac{1}{4} \cdot 3,4631 =$$

$$= 10,1063$$

Portanto, a nova matriz de dissimilaridade será:

$$D_{3 \times 3} = \begin{matrix} & (1,3,2,5,6) & 4 & 7 \\ (1,3,2,5,6) & \left[\begin{matrix} 0,00 & & \\ 11,25 & 0,00 & \\ 10,11 & 9,91 & 0,00 \end{matrix} \right] \\ 4 & & & \text{Simétrica} \\ 7 & & & \end{matrix}$$

Estágio 5: Entidades mais similares: genótipos 4 e 7

Distância entre entidades: 9,91

Por fim, calcula-se a distância:

$$\begin{aligned} d_{(4,7)(13256)} &= \frac{1}{2}d_{4(13256)} + \frac{1}{2}d_{7(13256)} - \frac{1}{4}d_7 = \frac{11,2486 + 10,1063}{2} - \frac{1}{4}9,91 \\ &= 8,1999 \end{aligned}$$

Assim, a matriz final para agrupamento será:

$$D_{2 \times 2} = \begin{matrix} & (1,3,2,5,6) & (4,7) \\ (1,3,2,5,6) & \left[\begin{matrix} 0,00 & \\ 8,20 & 0,00 \end{matrix} \right] \\ (4,7) & & \end{matrix}$$

Estágio 6: O último grupo é formado pela união do grupo (1,3,2,5,6) ao grupo (4,7), com distância 8,20. As fusões produzidas em cada estágio estão descritas a seguir:

Estágio	Entidade X	Entidade Y	Níveis de fusões
1	1	3	0,98
2	5	6	1,49
3	1, 3	2	2,33
4	1, 3, 2	5, 6	4,19
5	4	7	9,91
6	1, 3, 2, 5, 6	4, 7	8,20

A correspondente representação gráfica é mostrada na Figura 2.9.

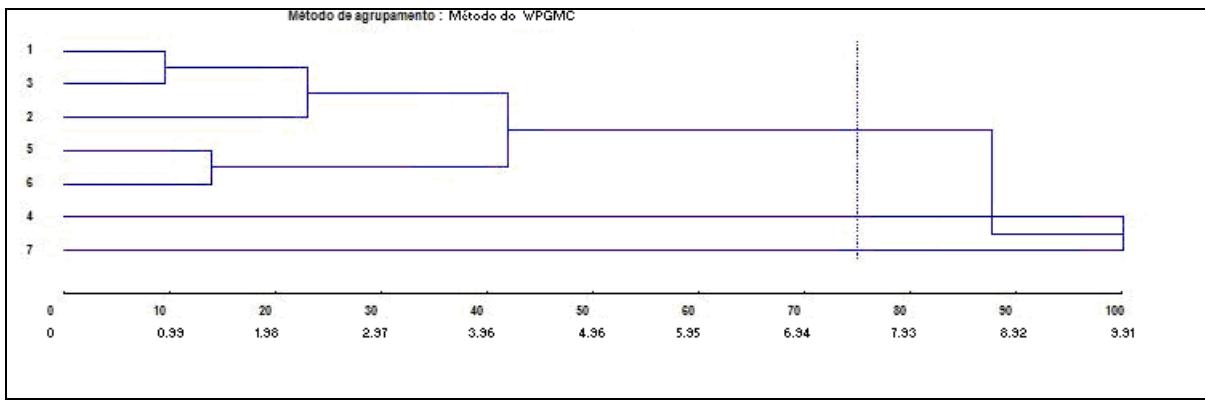


Figura 2.9 - Dendrograma obtido pelo método da mediana, baseado no quadrado da distância euclidiana média entre sete acessos.

Os métodos UPGMA, WPGMA, do centróide e da mediana apresentam resultados bastante próximos. As observações feitas para o método do centróide são também aplicáveis ao método da mediana.

Método da variância mínima de Ward

Neste método, consideram-se, para a formação inicial do grupo, aqueles indivíduos que proporcionam a menor soma de quadrados dos desvios. Admite-se que, em qualquer estágio, há perda de informações em razão do agrupamento realizado, o qual pode ser quantificado pela razão entre a soma de quadrados dos desvios dentro do grupo em formação e a soma de quadrados total dos desvios. A soma de quadrados dos desvios dentro é calculada considerando apenas os acessos dentro do grupo em formação, e a soma de quadrados dos desvios total é calculada considerando todos os indivíduos disponíveis para análise de agrupamento.

O agrupamento é feito a partir das somas de quadrados dos desvios entre acessos ou, alternativamente, a partir do quadrado da distância euclidiana, uma vez que se verifica a relação:

$$SQD_{ii'} = \frac{1}{2} d_{ii'}^2$$

em que:

$$SQD_{ii'} = \sum_{j=1}^v SQD_{j(ii')}$$

sendo $SQD_{j(ii')}$ a soma de quadrados dos desvios, para a j-ésima variável, considerando os acessos i e i' ; e

e

$$d_{ii'}^2 = \sum_{j=1}^v (X_{ij} - X_{i'j})^2$$

em que:

$d_{ii'}^2$: quadrado da distância euclidiana entre os genótipos i e i' ;

v: número de caracteres avaliados; e

X_{ij} : valor do caráter j para o genótipo i.

A soma de quadrados dos desvios total é dada por:

$$SQDT_{\text{Total}} = \frac{1}{g} \sum_{i < i'}^g \sum_{j=1}^g d_{ii'}^2$$

sendo g o número de acessos a serem agrupados.

Nesta análise de agrupamento, identifica-se na matriz D (cujos elementos são os quadrados das distâncias euclidianas - $d_{ii'}^2$) ou na matriz S (cujos elementos são as somas dos quadrados dos desvios - $SQD_{ii'}$) o par de acessos que proporciona menor soma de quadrados dos desvios. Com estes acessos agrupados, uma nova matriz de dissimilaridade, de dimensão inferior, é recalculada, considerando que:

$$SQD_{(ijk)} = \frac{1}{\alpha} d_{(ijk)}^2 \quad (\alpha \text{ é o número de acessos no grupo, que, neste caso, é igual a } 3).$$

$$d_{(ijk)}^2 = d_{(ij)}^2 + d_{(ij)k}^2 = d_{ij}^2 + d_{ik}^2 + d_{jk}^2$$

e ainda que:

$$SQD_{(ijkm)} = \frac{1}{\alpha} d_{(ijkm)}^2 \quad (\alpha \text{ é o número de acessos no grupo, que, neste caso, é igual a } 4).$$

$$d_{(ijklm)}^2 = d_{ij}^2 + d_{ik}^2 + d_{jk}^2 + d_{im}^2 + d_{jm}^2 + d_{km}^2$$

e assim sucessivamente.

No procedimento, realiza-se a análise de agrupamento fornecendo os g-1 passos de agrupamento para que seja formado o dendrograma.

Ilustração

Será considerado o agrupamento de cinco genótipos avaliados em relação a duas características (X_1 e X_2), conforme Tabela 2.15.

Tabela 2.15 - Valores para as características hipotéticas X_1 e X_2 , avaliadas em cinco genótipos

Genótipos	X_1	X_2
1	1	2
2	2	4
3	7	3
4	9	5
5	12	7
Total	31	21

A soma de quadrados de desvios, para cada característica, é dada por:

$$SQD_1 = \sum_{i=1}^g X_{i1}^2 - \frac{\left(\sum_{i=1}^g X_{i1}\right)^2}{g} = (1^2 + \dots + 12^2) - \frac{31^2}{5} = 86,80$$

$$SQD_2 = \sum_{i=1}^g X_{i2}^2 - \frac{\left(\sum_{i=1}^g X_{i2}\right)^2}{g} = (2^2 + \dots + 7^2) - \frac{21^2}{5} = 14,80$$

As dissimilaridades, expressas pelos quadrados das distâncias euclidianas, são dadas por:

Dissimilaridade (d_{ii}^2)	Apenas X_1	Apenas X_2	Total
12	$(1-2)^2 = 1$	$(2-4)^2 = 4$	5
13	$(1-7)^2 = 36$	$(2-3)^2 = 1$	37
14	$(1-9)^2 = 64$	$(2-5)^2 = 9$	73
15	$(1-12)^2 = 121$	$(2-7)^2 = 25$	146
23	$(2-7)^2 = 25$	$(4-3)^2 = 1$	26
24	$(2-9)^2 = 49$	$(4-5)^2 = 1$	50
25	$(2-12)^2 = 100$	$(4-7)^2 = 9$	109
34	$(7-9)^2 = 4$	$(3-5)^2 = 4$	8
35	$(7-12)^2 = 25$	$(3-7)^2 = 16$	41
45	$(9-12)^2 = 9$	$(5-7)^2 = 4$	13
Total	434	74	508

Verifica-se que:

a) Para a característica X_1

$$SQD_1 = \frac{1}{g} \sum_{i < j} \sum_{i'j'} d_{ii'}^2(1) = \frac{434}{5} = 86,80$$

b) Para a característica X_2

$$SQD_2 = \frac{1}{g} \sum_{i < j} \sum_{i'j'} d_{ii'}^2(2) = \frac{74}{5} = 14,80$$

c) Para o conjunto de características X_1 e X_2

$$SQD = \sum_{j=1}^n SQD_j, \text{ para o exemplo } SQD = SQD_1 + SQD_2 = 101,60$$

$$SDist = \sum_i^g \sum_{i' < i} \left[\sum_{j=1}^n (X_{ij} - X_{i'j})^2 \right] = 508$$

$$\text{sendo } SQD = \frac{SDist}{g} = \frac{508}{5} = 101,60$$

Assim, a dissimilaridade entre os acessos, dada pela soma de quadrados dos desvios, pode ser apresentada de duas formas:

$$S = \begin{bmatrix} 0 & & & & \\ SQD_{12} & 0 & & & \\ SQD_{13} & SQD_{23} & 0 & & \\ \dots & \dots & \dots & \dots & \\ SQD_{1g} & SQD_{2g} & SQD_{3g} & \dots & 0 \end{bmatrix} \quad \text{ou} \quad S = \frac{1}{2} D = \frac{1}{2} \begin{bmatrix} 0 & & & & \\ d_{12}^2 & 0 & & & \\ d_{13}^2 & d_{23}^2 & 0 & & \\ \dots & \dots & \dots & \dots & \\ d_{1g}^2 & d_{2g}^2 & d_{3g}^2 & \dots & 0 \end{bmatrix}$$

Para o exemplo em consideração, tem-se:

$$S = \begin{bmatrix} 0,00 \\ 2,50 & 0,00 \\ 18,50 & 13,00 & 0,00 \\ 36,50 & 25,00 & 4,00 & 0,00 \\ 73,00 & 54,50 & 20,50 & 6,50 & 0,00 \end{bmatrix} \quad e$$

$$D = \begin{bmatrix} 0,00 \\ 5,00 & 0,00 \\ 37,00 & 26,00 & 0,00 \\ 73,00 & 50,00 & 8,00 & 0,00 \\ 146,00 & 109,00 & 41,00 & 13,00 & 0,00 \end{bmatrix}$$

Para realizar o agrupamento, consideram-se os seguintes passos:

Passo 1: Identifica-se na matriz D ou S o elemento de menor valor. O par de genótipos que proporciona este menor valor é agrupado. Neste exemplo, os genótipos mais similares são o 1 e o 2. Tem-se, então:

$$\text{Soma de quadrados dos desvios: } SQD_{12} = \frac{1}{2}d_{12}^2 = 2,50$$

$$\text{Soma de quadrados dos desvios acumulada: } SQDa = SQD_{12} = 2,50$$

Perda de informação com o grupo:

$$P = R^2 = \frac{100SQDa}{SQTotal} = \frac{100(2,50)}{101,60} = 2,46\%$$

Passo 2: Deve ser estimada a soma de quadrados dos desvios envolvendo os genótipos do grupo em formação, com cada um dos demais não agrupados.

Dessa forma, tem-se:

$$SQD_{(12)3} = \frac{1}{3}d_{(12)3}^2 = \frac{1}{3}[d_{12}^2 + d_{(12)3}^2] = \frac{1}{3}(d_{12}^2 + d_{13}^2 + d_{23}^2) = \frac{1}{3}(5 + 37 + 26) = 22,67$$

$$SQD_{(12)4} = \frac{1}{3}d_{(12)4}^2 = \frac{1}{3}[d_{12}^2 + d_{(12)4}^2] = \frac{1}{3}(d_{12}^2 + d_{14}^2 + d_{24}^2) = \frac{1}{3}(5 + 73 + 50) = 42,67$$

$$SQD_{(12)5} = \frac{1}{3}d_{(12)5}^2 = \frac{1}{3}[d_{12}^2 + d_{(12)5}^2] = \frac{1}{3}(d_{12}^2 + d_{15}^2 + d_{25}^2) = \frac{1}{3}(5 + 146 + 109) = 86,67$$

Deve ser ressaltado que estas somas de quadrados poderiam ser calculadas da forma convencional. Assim, por exemplo, tem-se:

$$SQD_{(123)} = \left[(1^2 + 2^2 + 7^2) - \frac{10^2}{3} \right] + \left[(2^2 + 4^2 + 3^2) - \frac{9^2}{3} \right] = 20,67 + 2,00 = 22,67$$

Tem-se uma nova matriz de dissimilaridade, dada por:

$$S = \begin{matrix} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0,00 & & & \\ 3 & 22,67 & 0,00 & & \\ 4 & 42,67 & 4,00 & 0,00 & \\ 5 & 86,67 & 20,50 & 6,50 & 0,00 \end{matrix}$$

Verifica-se agora que os genótipos mais similares são o 3 e 4, tendo-se:

$$\text{Soma de quadrados dos desvios: } SQD_{34} = \frac{1}{2}d_{34}^2 = 4,00$$

$$\begin{aligned} \text{Soma de quadrados dos desvios acumulada: } SQDa &= SQD_{12} + SQD_{34} \\ &= 6,50 \end{aligned}$$

Perda de informação com o grupo:

$$P = R^2 = \frac{100SQDa}{SQTotal} = \frac{100(6,50)}{101,60} = 6,39\%$$

Passo 3: Deve ser estimada a soma de quadrados do desvio envolvendo os genótipos e grupos ou entre dois grupos formados. Assim, tem-se:

$$SQD_{(12,34)} = \frac{1}{4}d_{(12/34)}^2 = \frac{1}{4}(d_{12}^2 + d_{13}^2 + d_{14}^2 + d_{23}^2 + d_{24}^2 + d_{34}^2) = \frac{1}{4}(5 + 37 + 73 + 26 + 50 + 8) = 49,75$$

$$SQD_{(34,5)} = \frac{1}{3}d_{(34/5)}^2 = \frac{1}{3}[d_{34}^2 + d_{(34)5}^2] = \frac{1}{3}(d_{34}^2 + d_{35}^2 + d_{45}^2) = \frac{1}{3}(8 + 41 + 13) = 20,67$$

Tem-se uma nova matriz de dissimilaridade, dada por:

$$S = \begin{matrix} & (12) & (34) & 5 \\ (12) & 0,00 & & \\ (34) & 49,75 & 0,00 & \\ 5 & 86,67 & 20,67 & 0,00 \end{matrix}$$

Verifica-se agora que os mais similares são aqueles do grupo formado pelos genótipos 3, 4 e 5, tendo-se:

Soma de quadrados dos desvios: $SQD_{345} = 20,67$

Soma de quadrados dos desvios acumulada: $SQDa = SQD_{12} + SQD_{345}$
 $= 23,17$

Perda de informação com o grupo:

$$P = R^2 = \frac{100SQDa}{SQT\text{otal}} = \frac{100(23,17)}{101,60} = 22,80\%$$

Passo 4: Por fim, calcula-se a soma de quadrados dos desvios entre os dois grupos restantes, ou seja, o grupo estabelecido pelos genótipos 1 e 2 e o grupo formado pelos genótipos 3, 4 e 5. Assim, tem-se:

$$SQD_{(12\ 345)} = 101,60$$

As fusões produzidas em cada estágio estão descritas a seguir:

Estágio	Entidade X	Entidade Y	SQD	R ² (%)	100- R ² (%)
1	1	2	2,50	2,46	97,54
2	3	4	4,00	6,39	93,61
3	(3, 4)	5	20,67	22,80	76,20
4	(3, 4, 5)	(1, 2)	101,60	100,00	0,00

A correspondente representação gráfica é mostrada na Figura 2.10.

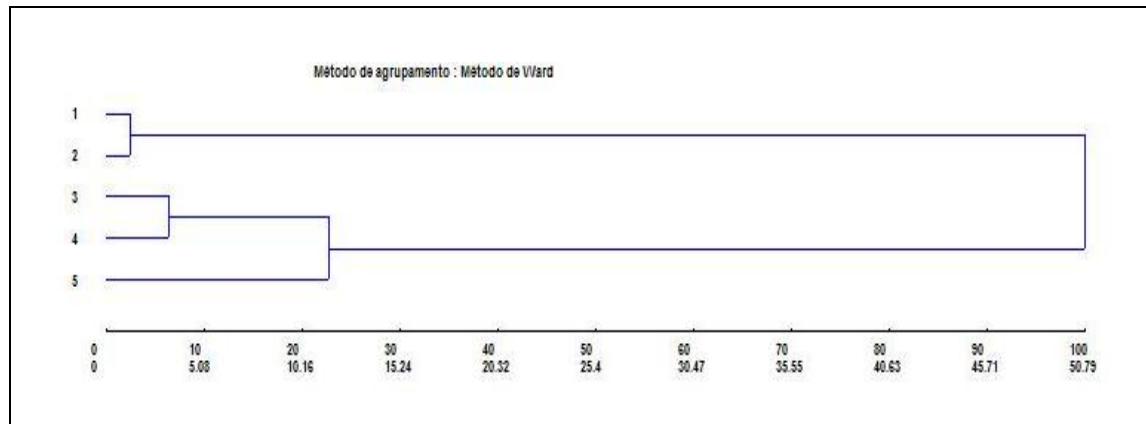


Figura 2.10 - Dendrograma obtido pelo método de Ward, a partir das medidas de dissimilaridade entre cinco genótipos. A primeira linha de valores no eixo da abscissa corresponde a valores percentuais em relação à dissimilaridade no último nível de fusão (50,80).

Para o exemplo relativo à avaliação de sete acessos, cuja medida de dissimilaridade é o quadrado da distância euclidiana, têm-se as seguintes informações relativas ao agrupamento:

Estágio	Entidade X	Entidade Y	Níveis de fusões	$R^2(\%)$
1	1	3	0,4900	2,0519
2	5	6	1,2350	5,1717
3	1,3	2	2,7883	11,6764
4	1,3,2	5,6	6,9440	29,0787
5	4	7	11,8990	49,8283
6	1,3,2,5,6	4,7	23,8800	100,0000

A correspondente representação gráfica é mostrada na Figura 2.11. Um corte feito à distância de 17,90, que corresponde a 75% do valor da distância no último nível de fusão (23,88), possibilita estabelecer dois grupos hierárquicos: um deles é constituído pelos acessos 4 e 7; e o outro, pelos demais acessos.

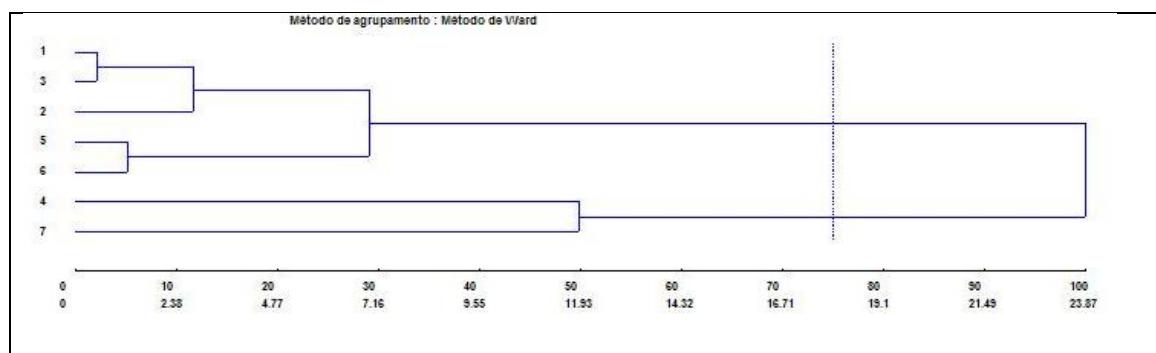


Figura 2.11 - Dendrograma obtido pelo método de Ward, baseado no quadrado da distância euclidiana média entre sete acessos. A primeira linha de valores no eixo da abscissa corresponde a valores percentuais em relação à dissimilaridade no último nível de fusão (23,88).

A título de comparação, percebe-se, que ao ser definido o mesmo valor percentual de corte (75%) para os diferentes algoritmos de agrupamento SAHN, é possível formar diferentes números de grupos e/ou com diferentes composições. Esse fato reforça o que foi dito anteriormente, que, para qualquer caso não se conhece, *a priori*, o número de grupos a serem estabelecidos, e diferentes métodos podem proporcionar diferentes resultados.

O método de Ward é atraente por se basear numa medida com forte apelo estatístico e por gerar grupos que, assim como o do método vizinho mais distante, possue alta homogeneidade interna (Barroso e Artes, 2003). Romesburg (1984) cita as seguintes características desse método: Apresenta bons resultados tanto para distâncias Mahalanobis quanto para distâncias eucliadianas (especialmente o quadrado da distância euclidiana); pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é praticamente igual; tem tendência a combinar grupos com poucos elementos; e é bastante sensível à presença de acessos muito diferenciados (pontos outliers entre o conjunto de dados).

Representação gráfica dos dendrogramas

Usualmente, os trabalhos de diversidade genética têm apresentado graficamente as distâncias nos dendrogramas através dos níveis de fusões de cada estágio. Deve-se estar atento ao método do centróide e da mediana, cuja representação gráfica das distâncias pode não ser bem entendida pelo usuário, uma vez que estes métodos fornecem distâncias não-monotônicas. A representação gráfica destes métodos pode ser dada pela ordem dos níveis de fusões.

Outra maneira de representar os valores de distâncias em cada estágio de agrupamento é por meio do percentual da distância máxima entre os níveis de fusão, ou seja:

$$d(\text{relativa}) = \left(\frac{\text{distância}_{\text{estágio}}}{\text{distânciamáxima}} \times 100 \right)$$

Pode-se também converter as distâncias em percentual de similaridade, pela expressão:

$$s(\text{relativa}) = \left[1 - \left(\frac{\text{distância}_{\text{estágio}}}{\text{distânciamáxima}} \right) \right] \times 100$$

2.4. Comparação e propriedades dos métodos de agrupamento SAHN

Utilização dos métodos SAHN

A literatura tem fornecido boas discussões acerca das propriedades e dos problemas inerentes aos métodos de agrupamento SAHN. O comum a todos estes métodos é que eles operam sobre a matriz de distância, dispensando recorrer à matriz de dados originais. Por outro lado, Romesburg (1984) e Khattree e Naik (2000) apresentam toda a teoria do método da variância mínima de Ward a partir do conjunto de dados originais.

Os métodos hierárquicos impõem uma estrutura hierárquica aos dados, que pode não existir. De acordo com Bussab et al. (1990), o conceito intuitivo de agrupamento natural, geometricamente, é subentendido como sendo regiões de alta densidade de pontos, delimitadas por regiões de baixa densidade.

No método da ligação simples – e no método da mediana, segundo Everitt (1993) – há uma tendência de gerar agrupamentos em cadeia. Esse fenômeno, conhecido como encadeamento (*chaining*), descreve uma situação em que o dendrograma apresenta níveis de “escadas” entre os grupos formados. Essa é uma propriedade inerente aos métodos, às vezes tida como um defeito. A formação de um único grupo pode ser enganosa se os indivíduos em extremidades opostas da cadeia são, de fato, completamente dissimilares (SNEATH; SOKAL, 1973; JOHNSON; WICHERN, 1988).

Segundo Romesburg (1984), a ocorrência ou não do encadeamento vai depender também do conjunto de dados avaliado e do coeficiente utilizado. É possível que uma matriz de distâncias venha a produzir encadeamento com os demais métodos. Naturalmente, que a informação proveniente do dendrograma em cadeia não deve ser descartada. O encadeamento parcial – cadeias que ocupam ramos do dendrograma – também poderá informar sobre a similaridade dos indivíduos. Em alguns casos, o método da ligação simples pode ser mais acurado em relação à estrutura dos dados do que os outros (JARDINE; SIBSON, 1968).

Uma das propriedades matemáticas favoráveis aos métodos da ligação simples e ligação completa é a notável estabilidade do agrupamento sob transformações monotônicas da matriz de distância. Isso significa que tais métodos produzirão os mesmos resultados sobre outras matrizes de distância, cujos elementos constituintes são da mesma ordem de posto dos elementos da matriz original. Jardine e Sibson (1971), em uma abordagem matemática, verificaram a capacidade dos métodos SAHN em transformar um coeficiente de dissimilaridade em um dendrograma hierárquico, de maneira que a inequação ultramétrica – $d_{xy} \leq \max(d_{xz}; d_{yz})$ – fosse imposta ao coeficiente de distância, o qual originalmente satisfaz apenas a inequação triangular. Neste estudo, o método da ligação simples satisfez às condições de continuidade, distorção mínima, entre outros aspectos.

Hughes (1979) utilizou os métodos da ligação simples e ligação completa para investigar a estabilidade da análise de agrupamento. Ela argumenta que, se os métodos são filosoficamente opostos, mas se apresentarem a mesma árvore para uma determinada matriz de dados, então os grupos são bem definidos em relação ao seu espaço v-dimensional (espaço amostral das variáveis) e o agrupamento pode ser dito como “real”. Contrariamente, se os dois métodos fornecem dendrogramas completamente diferentes, os grupos são mal definidos ou talvez seja um artefato do método de agrupamento.

Romesburg (1984) relata que o método da ligação simples tende a produzir árvores mais compactas. Já no método da ligação completa, as árvores são mais extensas e, no método UPGMA, elas são intermediárias entre esses dois extremos.

Uma das razões teóricas que favorecem o método UPGMA é que ele tende a gerar valores mais altos da matriz de correlação cofenética (SOKAL; ROHLF, 1962). Isso significa dizer que ele produz menor distorção quanto à representação das similaridades entre indivíduos em um dendrograma, embora a estimativa da correlação cofenética não seja uma verdade absoluta.

Dudley (1994), após revisar de literatura, também concluiu que o método UPGMA é superior aos métodos da ligação simples e ligação completa, quando comparados com informação de *pedigree*. O método UPGMA tem sido um dos preferidos pelos pesquisadores em várias áreas da ciência. Este método tem sido utilizado com maior freqüência em estudos de ecologia, sistemática e taxonomia numérica. O seu uso também tem sido comum nos estudos de diversidade genética. Já o método da ligação média ponderada (WPGMA) foi criado por biólogos taxonomistas, porém, caiu em desuso, por fazer classificações sem caráter biológico e sem base evolutiva.

O método UPGMC (centróide) é conceitualmente atrativo, por computar o centróide dos indivíduos que se unem para formar grupos. Assim, as distâncias são computadas entre esses centróides. Por não seguir a inequação ultramétrica, os processos de agrupamento pelo centróide e pela mediana não produzem resultados monotônicos, como visto nas Figuras 2.8 e 2.9 por meio dos “reversos” na união dos grupos 4 e 7 com o grupo 1, 3, 2, 5, 6. O “reverso” ocorre quando um indivíduo (ou grupo) une-se a um grupo depois de ele já ter sido formado, porém em um nível de similaridade mais alto (ou em um nível de dissimilaridade mais baixo) do que o do grupo já formado (SNEATH; SOKAL, 1973).

A preferência pelo método de Ward tem sido em função do efeito gráfico gerado pelo dendrograma, possibilitando a visualização de grupos bem definidos (ROMESBURG, 1984). Contudo, isso não é de real significância, pois a soma dos quadrados dos desvios é função do quadrado da diferença, tendendo a aumentar de forma não-linear ao longo dos estágios do agrupamento. Esse autor sugere a construção do dendrograma em função da raiz quadrada dos valores das somas de quadrados dos desvios. Com isso, a árvore pode não apresentar grupos tão bem

definidos como anteriormente, porém o dendrograma passa a ter melhor conotação matemática. O método de Ward restringe as opções quanto à utilização de outras medidas de dissimilaridade, uma vez que o quadrado da distância euclidiana possui relação com a soma de quadrados dos desvios.

Kuiper e Fisher (1975) concluíram que o método de Ward agrupa tão bem quanto a análise discriminante de Fisher quando existe igual número de indivíduos nos grupos e seguindo distribuição normal multivariada. Já com tamanho desigual de indivíduos nos grupos, os métodos do centróide, UPGMA e ligação completa apresentaram resultados melhores.

Não é possível definir claramente que um único método é superior aos demais para qualquer conjunto de dados. Por exemplo, Milligan (1980), na presença de observações discrepantes, verificou que os agrupamentos nos métodos de ligação simples, centróide e mediana não foram afetados, ao contrário dos métodos de Ward e UPGMA. Em contraste, quando os dados apresentavam indivíduos intermediários (“ruídos”) entre os grupos, os métodos de ligação simples, centróide e mediana não constituíram um agrupamento satisfatório, enquanto os métodos de Ward e UPGMA tiveram um desempenho melhor.

Hands e Everitt (1987), ao utilizarem dados binários, encontraram melhor desempenho na formação de grupos pelo método de Ward quando estes apresentavam tamanhos (número de indivíduos) aproximadamente iguais. Contudo, quando o tamanho dos grupos foi desigual, a técnica do centróide mostrou resultados mais satisfatórios.

Pereira (1999) submeteu 49 cultivares de arroz (tipo moderno e tradicional, dos grupos I e II) à análise multivariada, utilizando os métodos de agrupamento de ligação simples, UPGMA e Ward, além do método de otimização de Tocher, com base nas distâncias euclidianas e de Mahalanobis. Segundo o autor, o padrão de agrupamento diferiu tanto em relação aos métodos utilizados quanto às medidas de distância. A análise discriminante aplicada aos grupos formados pela análise de agrupamento mostrou não existir um padrão de agrupamento consistente com todos os métodos. No entanto, tais métodos multivariados, associados ao conhecimento

biológico por parte do pesquisador, permitiram distinguir os cultivares dos tipos tradicional e moderno.

Para que a formação e classificação de grupos sejam inerentes à técnica ou algoritmo utilizado, é preciso estar atento a alguns fatores, citados anteriormente, que podem interferir no processo de agrupamento. A necessidade de padronização dos dados é fundamental quando se trabalha com informações quantitativas. O pesquisador não pode obter resultados sob influência da unidade de medida tampouco da escala adotada para as variáveis. A escolha das medidas de distância para a formação da matriz de dissimilaridade deve ser cuidadosa, pois diferentes índices geram distâncias diferenciadas. Em alguns casos, as medidas guardam propriedades matemáticas e biológicas capazes de representar os fenômenos de maneira mais adequada.

Determinação do número de grupos

O processo, em todas as técnicas de agrupamento hierárquico aglomerativo, continua até que todos os indivíduos estejam alocados em um único grupo. Contudo, como medir a qualidade do agrupamento e decidir o número “verdadeiro” de grupos formados?

Essa tomada de decisão tem sido subjetiva em muitos estudos. De maneira geral, pode-se lançar mão de critério subjetivo, com base na inspeção visual das ramificações do dendrograma, ou recorrer a algum procedimento estatístico. Assim, devem ser considerados os seguintes aspectos:

Definição a priori do número de grupos

O número de grupos pode ser definido a priori, através de algum conhecimento que se tenha sobre os dados, pela conveniência do pesquisador ou por simplicidade (Barroso e Artes, 2003),

O pesquisador pode também, ter razões práticas para estabelecer o número de agrupamentos, com base no uso que pretende fazer dele. Qualquer que seja a abordagem empregada, geralmente é aconselhável observar o padrão total de

agrupamentos. Isto pode proporcionar uma medida da qualidade do processo de agrupamento e do número de grupos que se manifesta nos vários níveis do critério de agrupamento. Geralmente há expectativa de se ter mais de um nível de agrupamento (Aaker et al., 2001).

Análise visual das ramificações do dendrograma

Em teoria, os pontos de corte, para estabelecimento de grupos, seriam aqueles a partir dos quais se dá a mudança abrupta da ramificação presente no dendrograma. Este critério é comumente usado e pode ser útil, mas ele carrega consigo a possível influência de uma expectativa *a priori*. Geralmente, tem-se adotado o critério com base no conhecimento prévio do pesquisador sobre a relação de similaridade entre os indivíduos de seu estudo. Todavia, essa separação pode não estar muito clara visualmente, principalmente quando o número de indivíduos a serem agrupados é muito grande ou o método utilizado não lhe permite bom discernimento sobre os grupos formados.

A dificuldade maior surge quando não há uma expectativa prévia da conformação dos grupos e o número de indivíduos sob análise é elevado. Cabe ao pesquisador conhecer um pouco sobre as propriedades (matemáticas) dos métodos de agrupamento (SAHN e Tocher) e do seu conjunto de dados, para que o objetivo seja atingido. Diferentes métodos podem gerar agrupamentos divergentes, e a definição dos grupos pode estar mais evidente em uma técnica (dendrograma) do que em outra.

Critério estatístico

A literatura tem apresentado algumas maneiras de se determinar o número de grupos.

a) Método de Mojena (1977)

Mojena (1977) sugeriu um procedimento baseado no tamanho relativo dos níveis de fusões (distâncias) no dendrograma. A proposta é selecionar o número de grupos no estágio j que, primeiramente, satisfizer a seguinte inequação:

$$\alpha_j > \theta_k$$

em que:

α_j é o valor de distâncias dos níveis de fusão correspondentes ao estágio j ($j = 1, 2, \dots, n$); e θ_k é o valor referencial de corte, dado por:

$$\theta_k = \bar{\alpha} + k\hat{\sigma}_\alpha$$

sendo $\bar{\alpha}$ e $\hat{\sigma}_\alpha$ a média e o desvio-padrão não-viesado dos valores de α , respectivamente; e k é uma constante. A adoção de valores de k em torno de 2,75 e 3,50 é sugerida por Mojena (1977). No entanto, Milligan e Cooper (1985) sugeriram o valor de $k = 1,25$ como regra de parada na definição do número de grupos.

Como ilustração, serão considerados os resultados obtidos pelo método UPGMA, dados por:

Estágio	Entidade X	Entidade Y	Níveis de fusões (α_j)
1	1	3	0,98
2	5	6	1,49
3	1,3	2	2,58
4	1,3,2	5,6	4,52
5	4	7	9,91
6	1,3,2,5,6	4,7	12,25

Assim, tem-se:

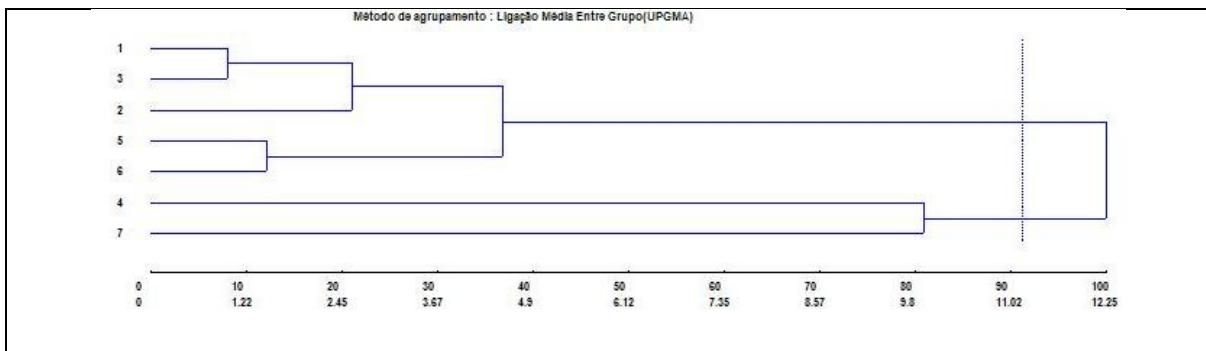
$$\bar{\alpha} = \frac{1}{g-1} \sum_{j=1}^{g-1} \alpha_j = 5,2883$$

$$\hat{\sigma}_\alpha = \sqrt{\left(\sum_{j=1}^{g-1} \alpha_j^2 - \frac{1}{g-1} \left(\sum_{j=1}^{g-1} \alpha_j \right)^2 \right) / (g-2)} = 4,7061$$

Para $k = 1,25$ tem-se:

$$\theta_k = 5,2883 + 1,25(4,7061) = 11,17$$

Assim, o valor α_j que supera θ_j ocorre no estágio 6, de modo que poderia ser estabelecido um grupo formado pelos acessos 1,3,2,5 e 6 e outro pelos acessos 4 e 7. A representação gráfica seria, portanto, dada por:

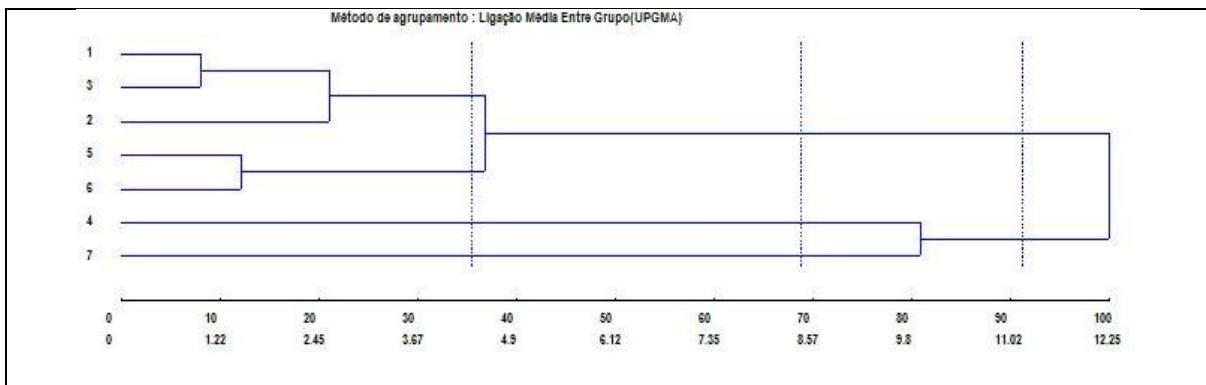


A análise poderia também ser feita para cada estágio do processo de agrupamento, adotando o mesmo procedimento. Dessa forma, tem-se:

Estágio	Nível de fusões (α_j)	$\bar{\alpha}$	$\hat{\sigma}_\alpha$	$\theta_k = \bar{\alpha} + k\hat{\sigma}_\alpha$
1	0,98	-	-	
2	1,49	1,2250	0,3606	1,6858
3	2,58	1,6833	0,8173	2,7050
4	4,52	2,3925	1,5675	4,3518*
5	9,91	3,8960	3,6256	8,4281*
6	12,25	5,2883	4,7061	11,1700*

$k = 1,25$

Assim, pode-se considerar que seria possível efetuar três cortes em diferentes estágios de agrupamento, e o dendrograma poderia ser representado da seguinte maneira:



b) Khattree e Naik (2000)

Khattree e Naik (2000) apresentam outras três estatísticas, que, aliadas aos níveis de fusão (no caso do método de Ward), podem ajudar a definir o número de grupos. Esses critérios devem ser aplicados sobre o conjunto de dados originais ou padronizados, para qualquer técnica de agrupamento SAHN.

Um deles é o desvio-padrão médio (DP_w), ou seja, a raiz quadrada da variância do novo grupo formado sobre todas as variáveis. Especificamente, se o novo grupo formado (G_w) tem as respectivas variâncias $s_{w1}^2, s_{w2}^2, \dots, s_{wv}^2$ para v variáveis, então:

$$DP_w = \sqrt{\frac{1}{v} \sum_{j=1}^v s_{wj}^2}$$

Evidentemente que, quanto mais similares os dois grupos (ou indivíduos) unidos para formar G_w , mais homogêneo o grupo G_w será e, consequentemente, ter-se-á o menor valor de DP_w .

Outro critério é o coeficiente de determinação (R^2) para o grupo G_w , definido como:

$$R_w^2 = 1 - \frac{SQD_w}{SQT}$$

em que SQT é a soma de quadrados total corrigida de todos os indivíduos somada sobre todas as variáveis; e SQD_w é a soma de quadrados comum (corrigida) dentro

do grupo, somada sobre todas as variáveis, no estágio em que o grupo G_w foi formado. Naturalmente que SQT é um valor constante para um determinado conjunto de dados e SQD_w irá mudar de acordo com o grupo. É desejável encontrar pequenos valores de SQD_w , isto é, altos valores de R_w^2 , variando de 0 a 1.

Outro critério é o R^2 semiparcial, que mede a perda de homogeneidade pela fusão de dois grupos (indivíduos). Essa estatística é definida como a redução no correspondente valor de R^2 , quando dois grupos (indivíduos) são unidos. Quando o novo grupo formado é relativamente homogêneo, como desejado, essa perda na homogeneidade deve ser pequena. Conseqüentemente, buscam-se pequenos valores de R^2 parcial.

A visualização gráfica dessas três últimas estatísticas contra o número de grupos pode ajudar na interpretação e identificação de pontos cujas mudanças de um valor para outro são menores, isto é, tende-se à estabilidade. O programa SAS (SAS, 1989) fornece essas alternativas de análise.

Como ilustração, serão considerados os resultados obtidos pelo método UPGMA, fornecidos por:

Estágio	Entidade X	Entidade Y	Grupos	SQ_w	DP_w	R²
1	1	3	(1,3) 2,4,5,6,7	1,080	0,5196	0,9548
2	5	6	(1,3)(5,6)2,4,7	0,745	0,4316	0,9688
3	1,3	2	(1,3,2)(5,6),4,7	6,130	1,2379	0,7432
4	1,3,2	5,6	(1,3,2,5,6),4,7	6,944	1,3176	0,7092
5	4	7	(1,3,2,5,6), (4,7)	4,955	1,1130	0,7925
6	1,3,2,5,6	4,7	(1,3,2,5,6,4,7)	23,888	2,4433	0,0000

Consistência do Agrupamento (ou do dendrograma)

A literatura fornece algumas maneiras de se avaliar a consistência do padrão de agrupamento ou de validá-lo. Serão descritas algumas das alternativas mais utilizadas para esse propósito.

Métodos de otimização

A consistência do padrão de agrupamento obtido pelo método de otimização pode ser avaliada aplicando-se a análise discriminante, geralmente baseada no método proposto por Anderson (1958), aos dados obtidos ou por meio da análise discriminante não-paramétrica (KHATTREE; NAIK, 2000).

Nesse caso, os dados originais, sem informações prévias sobre grupos, darão origem à matriz de distância, que, por sua vez, é utilizada para promover o agrupamento. Retorna-se aos dados originais, classificando-os conforme resultado obtido pela análise de otimização efetuada. Realiza-se, então, a estimativa de funções discriminantes, que permitem classificar novos genótipos não incluídos na análise e ratificar a alocação daqueles em estudo nos grupos considerados. Neste último caso, obtém-se a taxa de erro aparente, que quantifica o número de genótipos classificados de forma diferente da prevista pela análise da técnica de otimização, medindo a adequação do agrupamento anteriormente realizado.

Métodos hierárquicos

Com a formação do dendrograma, pode haver distorções sobre o padrão de dissimilaridade entre os indivíduos estudados e ocorrer considerável simplificação das informações originais (EVERITT, 1993; CRUZ; CARNEIRO, 2003). Então, é necessário julgar a adequação da classificação. A consistência do agrupamento é checada após a obtenção do dendrograma.

i. Correlação entre medidas de dissimilaridade originais e gráficas

A consistência do agrupamento é feita depois de obtido o dendrograma, em que se tem uma nova leitura da dissimilaridade ou similaridade entre os genótipos avaliados. Os novos coeficientes de semelhança indicados nos eixos são

estabelecidos de acordo com o método de agrupamento escolhido e podem ser utilizados no estabelecimento de uma nova matriz de dissimilaridade, denominada de matriz de coeficientes de semelhança cofenéticos. Deve ser ressaltado que na leitura do coeficiente de similaridade no dendrograma a informação entre pares de genótipos é, muitas vezes, perdida, dando lugar a novas referências relativas aos grupos ou níveis dentro da árvore de similaridade.

O coeficiente de correlação cofenética, segundo Sokal e Rohlf (1962), é um coeficiente de correlação momento-produto calculado entre os elementos acima da diagonal da matriz de dissimilaridade e os valores da matriz cofenética. Quanto maior o valor da correlação, menor será a distorção provocada pelo agrupamento.

Rohlf e Fisher (1968) estudaram a distribuição do coeficiente de correlação cofenética sob a hipótese de que os indivíduos são escolhidos aleatoriamente a partir de uma única distribuição normal multivariada. Eles observaram que, em média, os coeficientes de correlação cofenética tendem a diminuir com o aumento do número de indivíduos e que são quase sempre independentes do número de variáveis. Os mesmos autores sugeriram que uma correlação cofenética acima de 0,8 é suficiente. Por sua vez, Rohlf (1982) alerta que uma correlação cofenética próxima de 0,9 não garante que o dendrograma tenha sido capaz de sintetizar a relação fenética dos indivíduos – relação esta definida por Sneath e Sokal (1973) como a similaridade entre os indivíduos baseada em um conjunto de características fenotípicas.

Ilustração

Será considerado o agrupamento do método da ligação simples (ou vizinho mais próximo). Os elementos c_{ij} (matriz C) podem ser obtidos considerando os passos do agrupamento; desse modo, tem-se:

$$D_{7 \times 7}^* = \begin{array}{ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{matrix} & \left[\begin{matrix} 0,00 \\ 2,16 & 0,00 \\ 0,98 & 2,99 & 0,00 \\ 14,32 & 12,53 & 9,55 & 0,00 \\ 3,00 & 6,55 & 1,44 & 12,75 & 0,00 \\ 5,25 & 8,45 & 2,41 & 14,12 & 1,49 & 0,00 \\ 14,66 & 17,56 & 10,33 & 9,91 & 6,37 & 10,34 & 0,00 \end{matrix} \right] \end{array}$$

* Matriz de distância

$$C_{7 \times 7}^* = \begin{array}{ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{matrix} & \left[\begin{matrix} 0,00 \\ 2,16 & 0,00 \\ 0,98 & 2,16 & 0,00 \\ 9,55 & 9,55 & 9,55 & 0,00 \\ 1,44 & 2,16 & 1,44 & 9,55 & 0,00 \\ 1,49 & 2,16 & 1,49 & 9,55 & 1,49 & 0,00 \\ 6,37 & 6,37 & 6,37 & 9,91 & 6,37 & 6,37 & 0,00 \end{matrix} \right] \end{array}$$

* Matriz cofenética

A correlação cofenética entre os elementos de D e C é de 0,8186. Este valor foi significativo a 1% de probabilidade pelo teste de Mantel, com base em 1.000 reamostragens. Resultados da aplicação de outros métodos de agrupamento são apresentados a seguir:

Agrupamento	Correlação cofenética
Vizinho mais próximo	0,8186
Vizinho mais distante	0,8502
UPGMA	0,8724
Centróide	0,8186
Mediana	0,8704
Ward	0,8615

ii. Comparação de resultados entre diferentes técnicas de agrupamento do tipo SAHN

Bussab et al. (1990) recomendam a aplicação de mais de um algoritmo SAHN ao conjunto de dados e que o resultado mais concordante seja aceito como o mais adequado. Isso evita que a classificação seja um mero artefato da técnica utilizada, já que cada técnica impõe determinada estrutura aos dados. Nesse contexto, para validar os agrupamentos, com base na concordância e comparação entre dendrogramas compostos por métodos diferentes ou por um mesmo método, mas estabelecidos a partir de diferentes medidas de dissimilaridade, Rohlf (1982) propôs o índice consenso (CI), que estima o nível de similaridade entre dois dendrogramas. Este índice é baseado no número de passos (estágios) em comum observado na construção de um dendrograma. Esse valor é obtido pela divisão do número de estágios em comum pelo número máximo possível de estágios no agrupamento, excluindo o último estágio, em que todos os indivíduos são alocados em um único grupo. O denominador é igual a $n-2$, sendo n o número de indivíduos avaliados. O valor de CI varia de 0 (nenhum consenso) a 1 (dendrogramas idênticos).

Evidentemente que a análise e comparação apenas entre os métodos hierárquicos podem não retratar o verdadeiro caráter do agrupamento. Outros métodos estatísticos de análise auxiliarão na tomada de decisão e crítica a respeito dos métodos. A importância das técnicas de agrupamento é justamente devido à capacidade de resumir a informação contida na matriz de distância, facilitando a identificação de grupos homogêneos pela simples visualização do dendrograma. Todavia, o agrupamento conduz à perda de informações em nível de indivíduos, restando somente informações em nível de grupos de indivíduos similares. Por esse motivo, o estudo da diversidade genética recebe suporte, simultaneamente, de outras técnicas multivariadas.

iii. Comparação de resultados entre diferentes técnicas de agrupamento

Uma alternativa para verificação da consistência do agrupamento pelos métodos SAHN é estabelecer comparações dos grupos obtidos com outras técnicas multivariadas (análise de escala multidimensional – MSD, análise de coordenadas, análise discriminante, análise de componentes principais etc.). A concordância dos resultados pode indicar a consistência dos grupos formados.

Estabilidade do Agrupamento

Para avaliar a estabilidade do agrupamento, Everitt (1993) sugeriu dividir ao acaso o conjunto de dados originais em dois subconjuntos e aplicar o algoritmo escolhido a cada um deles. Se a alocação dos indivíduos nos subconjuntos for a mesma do conjunto completo, diz-se que o agrupamento foi estável.

Quando se realizam experimentos em diferentes ambientes há o interesse em se ter um padrão de agrupamento em cada um deles e inferir quanto à estabilidade das associações entre os acessos. Entretanto, grandes diferenças podem ocorrer em razão da própria resposta biológica dos acessos quando submetidos a diferentes ambientes, não sendo, necessariamente, provocadas pelo uso da metodologia estatística inapropriada. De maneira geral, acessos exibem maior variabilidade em condições ambientais favoráveis que possibilitam a expressão máxima do potencial genético dos indivíduos avaliados. Murty e Arunachalan (1966) e Upadhyay e Murty (1970) relataram que a deriva genética e a seleção, em diferentes ambientes, podem causar grande diversidade genética de forma que os resultados de agrupamento podem diferir drasticamente de um ambiente para outro.

Sugere-se que informações ecológicas, da biologia populacional e, ou, da biologia reprodutiva dos acessos devam ser levadas em consideração quando houver interesse na comparação e no entendimento dos padrões de agrupamento genético-ambiental obtidos .

Robustez do Agrupamento

Outro aspecto, conhecido como condição de continuidade e robustez, traduz o fato de que pequenas mudanças no conjunto de dados, como deleção ou inserção de variáveis e/ou indivíduos, seriam acompanhadas por pequenas mudanças no dendrograma.

Consistência das Ramificações do Dendrograma

A reamostragem com reposição (*bootstrapping*) das variáveis no conjunto de dados originais pode ser uma alternativa para se checar a repetibilidade dos nós formados nos dendrogramas. Para exemplificar, considere a Figura 2.12, com 10 genótipos e nove locos (variáveis). O processo de reamostragem é realizado nos locos com 10.000 simulações, e em cada uma delas podem ou não ser obtidos ramos diferenciados. Após todas as simulações definiu-se o dendrograma consenso, cujos nós foram os mais freqüentes em todas as 10.000 reamostragens. Evidentemente que o número de reamostragem pode variar, ficando a critério do pesquisador ou limitado pelo *software* utilizado. Contudo, um número de reamostragem igual ou superior a 1.000 é tido como satisfatório.

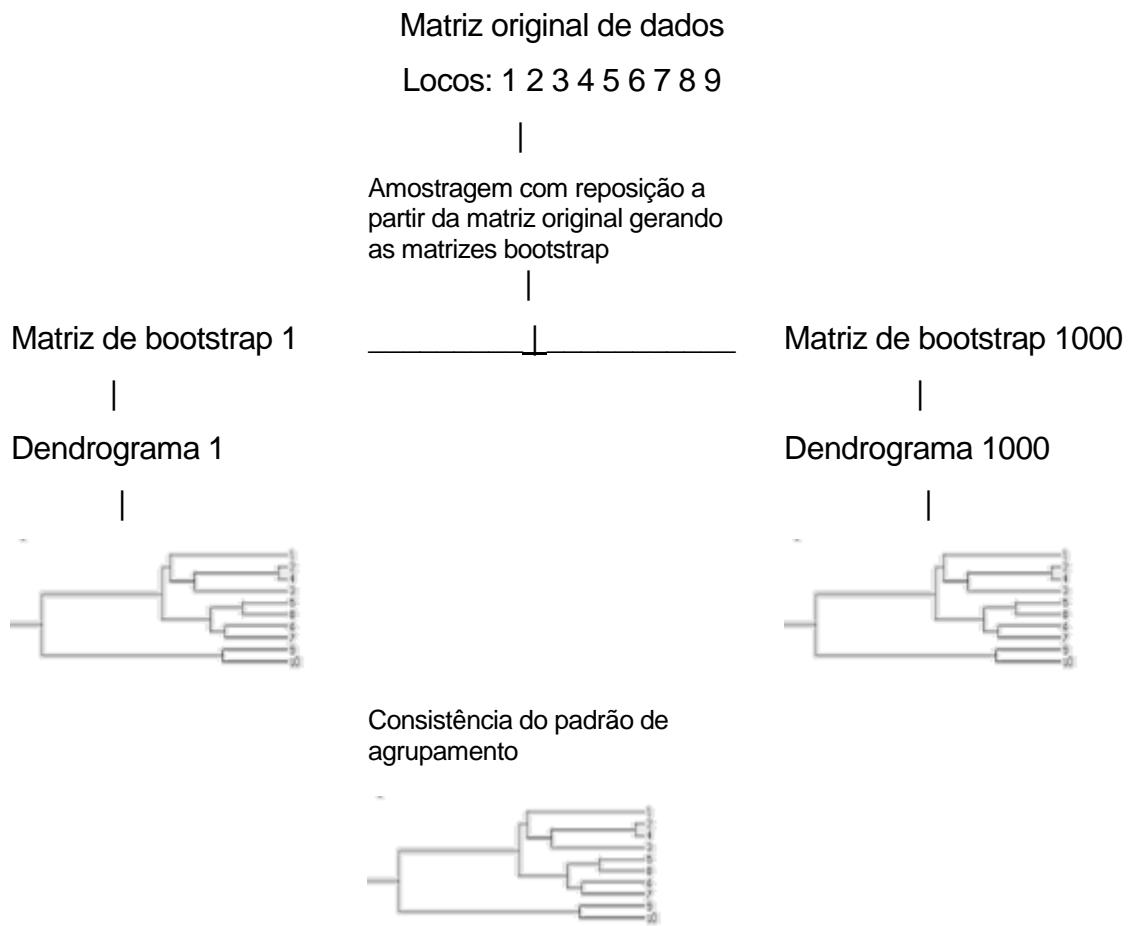


Figura 2.12 - Esquema ilustrativo da realização de *bootstrap* na matriz de dados originais.

2.5. Métodos de agrupamento baseados em dispersão gráfica

Como já relatado, os métodos de agrupamento podem envolver matrizes de dissimilaridade de ordem elevada, e o agrupamento realizado pode provocar perda de informações do grau de dissimilaridade, principalmente dos indivíduos pertencentes a um mesmo grupo. Uma maneira alternativa aos métodos de agrupamento, para avaliação da diversidade entre genótipos, é a análise da dispersão gráfica, normalmente utilizando o espaço bi ou tridimensional.

O fato questionável nas análises baseadas em dispersão gráfica é a dificuldade do estabelecimento de grupos de similaridade de forma menos subjetiva, com base na simples inspeção visual da dispersão. Há casos em que a visualização do grau de similaridade entre os genótipos não é muito clara. Em outros casos, a análise não consegue resumir o complexo de informações das variáveis originais com apenas informações representadas por dois ou três eixos na análise gráfica.

Assim, pelas vantagens e desvantagens apresentadas, a utilização conjugada de métodos de dispersão gráfica com os de agrupamento tem sido a alternativa mais adequada em estudos de diversidade genética. Recomenda-se utilizar a análise gráfica quando puderem ser estimadas variáveis – a serem usadas em eixos cartesianos – que envolvem o máximo da variação disponível nos dados originais. Neste gráfico, quando o grau de distorção entre as distâncias originais e as representadas for mínimo, poder-se-á inferir a dissimilaridade entre todos os genótipos considerados, com relativa facilidade. Entretanto, toma-se o cuidado de delinear os grupos com base na análise de agrupamento, evitando-se a subjetividade, uma vez que é admitido um critério de otimização.

Há várias técnicas que possibilitam a análise da diversidade genética por meio de dispersão gráfica. As técnicas de componentes principais e variáveis canônicas são utilizadas quando se dispõe dos dados originais e guardam correspondência com o agrupamento baseado nas dissimilaridades expressas pela distância euclidiana e de Mahalanobis, respectivamente. Por outro lado, a projeção é recomendada quando o interesse é o de representar graficamente distâncias obtidas por qualquer coeficiente, escolhidos pelo pesquisador por descreverem o fenômeno estudado dentro de princípios biológicos de interesse. A possibilidade de representar graficamente qualquer medida entre pares de genótipos torna esta metodologia mais flexível e de grande interesse, principalmente em dados discretos binários ou multicategóricos, em que são adotados diferentes coeficientes de similaridade.

Para variáveis quantitativas contínuas, a diversidade genética pode ser convenientemente estudada por meio das técnicas de componentes principais ou das variáveis canônicas, cujas particularidades serão apresentadas a seguir. Também pode ser estudada por meio da projeção de uma determinada medida de dissimilaridade no plano ou espaço tridimensional. Os procedimentos que utilizam combinações lineares de variáveis, como as análises canônicas e de componentes principais, são apropriados quando se objetiva interpretar combinações lineares ótimas de variáveis. Contudo, se as variáveis não são linearmente relacionadas, a matriz de variâncias e covariâncias, indispensáveis nos procedimentos anteriormente citados, é pobre sumário estatístico. Assim, para variáveis qualitativas, em especial as binárias, essas técnicas são menos apropriadas e devem ser substituídas pela técnica da projeção de medidas de dissimilaridade (JAMES; MCCULLOCH, 1990)

2.5.1. Projeção de medidas de dissimilaridade

Projeção 2D

Neste procedimento as medidas de dissimilaridade são convertidas em escores relativos a duas variáveis (X e Y), que, quando representadas em gráficos de dispersão, irão refletir, no espaço bidimensional, as distâncias originalmente obtidas a partir do espaço v-dimensional ($v =$ número de caracteres utilizados para obtenção das distâncias).

A viabilidade do uso desta técnica é avaliada pela correlação entre as distâncias originais e as que serão representadas no gráfico de dispersão, ou pelo grau de distorção ($1 - \alpha$), considerando que:

$$\alpha = \frac{\sum_{i < i'} \sum d_{gi}^2}{\sum_{i < i'} \sum d_{oi}^2}$$

em que d_{gi}^2 e d_{oi}^2 são as distâncias gráficas (espaço bidimensional) e originais (espaço v-dimensional), respectivamente, de todos os pares de indivíduos i e i' .

Também tem sido utilizado como medida da adequação da representação gráfica o valor do coeficiente de estresse (s), dado por:

$$s = 100 \sqrt{\frac{\sum_{i<} \sum_{i'} (d_{gi'i'} - d_{oi'i'})^2}{\sum_{i<} \sum_{i'} d_{oi'i'}^2}}$$

De maneira geral, admite-se ser satisfatória a representação gráfica quando o valor da correlação entre as medidas de distâncias originais e gráficas é superior a 0,9 e os valores de distorção e estresse são inferiores a 20%.

Para fazer a representação das medidas de similaridade em gráficos, calcula-se a coordenada das medidas mais divergentes e, a seguir, daquelas que demonstram, em ordem decrescente, as maiores diversidades com os genótipos já considerados, conforme descrito por Cruz e Viana (1994). Sendo i e j os genótipos mais divergentes, considera-se, arbitrariamente, que a coordenada de i seja igual a $(0,0)$, e a de j , igual a $(d_{ij}, 0)$.

O próximo genótipo k a ser considerado para representação gráfica será aquele de maior valor $d_{(ij)k}$, fornecido por:

$$d_{(ij)k} = d_{ik} + d_{jk}$$

Ressalta-se que estes valores ($d_{(ij)k}$) são utilizados apenas para ordenação dos genótipos, em relação à diversidade com aqueles cujas coordenadas já foram estabelecidas. Entretanto, a coordenada deste terceiro genótipo, dada por (X_k, Y_k) , é estabelecida matematicamente, considerando as propriedades de um triângulo. Assim, tem-se:

$$X_k = \frac{d_{jk}^2 - d_{ik}^2 - d_{ij}^2}{-2d_{ij}}$$

e

$$Y_k = \sqrt{d_{ik}^2 - X_k^2}$$

O mesmo critério é usado para a próxima unidade ℓ , ou seja, escolhe-se ℓ tal que o valor $d_{(ijk)\ell}$ seja o maior entre todos. Assim, tem-se:

$$d_{(ijk)\ell} = d_{i\ell} + d_{j\ell} + d_{k\ell}$$

A coordenada das demais unidades é estimada estatisticamente, visando minimizar a distorção entre a distância original e a distância gráfica. Assim, a coordenada da unidade ℓ é estimada considerando que:

- a) O genótipo i apresenta coordenada $(0, 0)$.
- b) O genótipo j apresenta coordenada (X_j, Y_j) , em que $X_j = d_{ij}$ e $Y_j = 0$.
- c) O genótipo k apresenta coordenada (X_k, Y_k) sendo X_k e Y_k estimados conforme expressão matemática dada anteriormente.
- d) O genótipo ℓ apresenta coordenada (X_ℓ, Y_ℓ) estimada pelo sistema de equações:

$$d_{\ell i}^2 = X_\ell^2 + Y_\ell^2$$

$$d_{\ell j}^2 = (X_j - X_\ell)^2 + (Y_j - Y_\ell)^2 = X_j^2 - 2X_j X_\ell + X_\ell^2 + Y_j^2 - 2Y_j Y_\ell + Y_\ell^2$$

$$d_{\ell k}^2 = (X_k - X_\ell)^2 + (Y_k - Y_\ell)^2 = X_k^2 - 2X_k X_\ell + X_\ell^2 + Y_k^2 - 2Y_k Y_\ell + Y_\ell^2$$

O qual pode ser colocado sob notação matricial $Y = X\beta + \varepsilon$, obtendo-se:

$$Y = \begin{bmatrix} d_{\ell j}^2 - d_{\ell i}^2 - d_{ij}^2 \\ d_{\ell k}^2 - d_{\ell i}^2 - d_{ik}^2 \end{bmatrix}; \quad X = -2 \begin{bmatrix} X_j & Y_j \\ X_k & Y_k \end{bmatrix}; \quad \beta = \begin{bmatrix} X_\ell \\ Y_\ell \end{bmatrix} \text{ e } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

Para as demais coordenadas, são acrescentadas linhas no vetor Y e na matriz X , as quais passam a ter as dimensões $(m - 2) \times 1$ e $(m - 2) \times 2$, respectivamente, sendo m o número de genótipos até então estudados.

A solução do sistema é obtida por $X'X\hat{\beta} = X'Y$, de forma que a coordenada estimada para o genótipo ℓ apresente a menor distorção de distância com os demais, cujas coordenadas já foram estabelecidas. Assim, considera-se:

$$X'X = 4 \begin{bmatrix} \sum_{n=1}^{m-2} X_n^2 & \sum_{n=1}^{m-2} X_n Y_n \\ \sum_{n=1}^{m-2} X_n Y_n & \sum_{n=1}^{m-2} Y_n^2 \end{bmatrix} \quad \text{e} \quad X'Y = -2 \begin{bmatrix} \sum_{n=1}^{m-2} \Delta_{\ell n} X_n \\ \sum_{n=1}^{m-2} \Delta_{\ell n} Y_n \end{bmatrix}$$

sendo:

$$\Delta_{\ell j} = d_{\ell j}^2 - d_{\ell i}^2 - d_{ij}^2$$

$$\Delta_{\ell k} = d_{\ell k}^2 - d_{\ell i}^2 - d_{ik}^2$$

...

$$\Delta_{\ell n} = d_{\ell n}^2 - d_{\ell i}^2 - d_{in}^2$$

Ressalta-se o fato de que n é um indexador que assume os valores correspondentes aos pontos (genótipos) cujas coordenadas já foram calculadas. Assim, $n = i, j, k, l$. A minimização das distorções permite obter a seguinte solução:

$$X_{\ell} = \frac{(\sum X_n Y_n)(\sum Y_n \Delta_{\ell n}) - (\sum Y_n^2)(\sum X_n \Delta_{\ell n})}{2 \left[\sum X_n^2 \sum Y_n^2 - (\sum X_n Y_n)^2 \right]}$$

e

$$Y_{\ell} = \frac{(\sum X_n Y_n)(\sum X_n \Delta_{\ell n}) - (\sum X_n^2)(\sum Y_n \Delta_{\ell n})}{2 \left[\sum X_n^2 \sum Y_n^2 - (\sum X_n Y_n)^2 \right]}$$

Ilustração

Será considerada, como ilustração, a projeção das medidas de dissimilaridade entre cinco genótipos expressos pelo complemento aritmético do índice de Jaccard, dado por:

$$D = \begin{bmatrix} 0 & 0,88 & 0,50 & 0,33 & 0,67 \\ & 0 & 0,88 & 0,83 & 0,50 \\ & & 0 & 0,75 & 0,50 \\ & & & 0 & 0,89 \\ & & & & 0 \end{bmatrix}$$

Para fazer a projeção destas medidas de dissimilaridade no espaço bidimensional, devem ser adotados os seguintes passos:

Passo 1. Identificação na matriz de dissimilaridade dos genótipos mais divergentes, que, no caso considerado, são o 4 e o 5, cuja distância é igual a 0,89. Ficam estabelecidas as seguintes coordenadas:

- a) Genótipo 5: coordenada (0,0)
- b) Genótipo 4: coordenada $(d_{45}; 0) = (0,89 ; 0)$

Passo 2. Estabelecimento da ordem de dissimilaridade em relação aos genótipos 4 e 5, por meio de:

$$d_{(45)1} = d_{14} + d_{15} = 0,33 + 0,67 = 1,00$$

$$d_{(45)2} = d_{24} + d_{25} = 0,83 + 0,50 = 1,33$$

$$d_{(45)3} = d_{34} + d_{35} = 0,75 + 0,50 = 1,25$$

Identifica-se, portanto, o genótipo 2 como o de maior distância ($d = 1,33$) em relação aos demais, cujas coordenadas já foram estabelecidas. Obtém-se a coordenada deste genótipo por meio das equações:

$$X_k = X_2 = \frac{d_{jk}^2 - d_{ik}^2 - d_{ij}^2}{-2d_{ij}} = \frac{d_{24}^2 - d_{25}^2 - d_{45}^2}{-2d_{45}} = \frac{0,83^2 - 0,50^2 - 0,89^2}{-2(0,89)} = 0,1984$$

e

$$Y_k = Y_2 = \sqrt{d_{ik}^2 - X_k^2} = \sqrt{d_{25}^2 - X_2^2} = \sqrt{0,50^2 - 0,1984^2} = 0,4589$$

Passo 3. Estabelecimento da ordem de dissimilaridade em relação aos genótipos 2, 4 e 5, por meio de:

$$d_{(245)1} = d_{12} + d_{14} + d_{15} = 0,88 + 0,33 + 0,67 = 1,88$$

$$d_{(245)3} = d_{23} + d_{34} + d_{35} = 0,88 + 0,75 + 0,50 = 2,13$$

Identifica-se, portanto, o genótipo 3 como o de maior distância ($d = 2,13$) em relação aos demais, cujas coordenadas já foram estabelecidas. Obtém-se a coordenada deste genótipo por meio das informações:

n	X _n	Y _n	X _n ²	Y _n ²	X _n Y _n	d _{ℓn} ²	d _{in} ²	Δ _{ℓn}	X _n Δ _{ℓn}	Y _n Δ _{ℓn}
4	0,8900	0,0	0,7921	0,0	0,0	0,5625	0,7921	-0,4796	-0,4268	0,0
2	0,1984	0,4589	0,0394	0,2106	0,0910	0,7744	0,2500	0,2744	0,0532	0,1259
Total			0,8315	0,2106	0,0910				-0,3736	0,1259

Neste caso, $i = 5$ (i é o genótipo de coordenada $(0, 0)$) e $d_{\ell i}^2 = d_{35}^2 = 0,50^2 = 0,25$.

Assim:

$$X_{\ell} = \frac{(\sum X_n Y_n)(\sum Y_n \Delta_{\ell n}) - (\sum Y_n^2)(\sum X_n \Delta_{\ell n})}{2 \left[\sum X_n^2 \sum Y_n^2 - (\sum X_n Y_n)^2 \right]}$$

$$X_3 = \frac{(0,0910)(0,1259) - (0,2106)(-0,3736)}{2[(0,8315)(0,2106) - (0,0910)^2]} = 0,2702$$

e

$$Y_{\ell} = \frac{(\sum X_n Y_n)(\sum X_n \Delta_{\ell n}) - (\sum X_n^2)(\sum Y_n \Delta_{\ell n})}{2 \left[\sum X_n^2 \sum Y_n^2 - (\sum X_n Y_n)^2 \right]}$$

$$Y_3 = \frac{(0,0910)(-0,3736) - (0,8315)(0,1259)}{2[(0,8315)(0,2106) - (0,0910)^2]} = -0,4157$$

Passo 4. Finalmente, calcula-se a coordenada do genótipo 1, que ainda não foi considerado. Obtém-se a coordenada deste genótipo por meio das informações:

n	X _n	Y _n	X _n ²	Y _n ²	X _n Y _n	d _{ℓn} ²	d _{ln} ²	Δ _{ℓn}	X _n Δ _{ℓn}	Y _n Δ _{ℓn}
4	0,8900	0,0	0,7921	0,0	0,0	0,1089	0,7921	-0,1321	-1,0076	0,0
2	0,1940	0,4589	0,0394	0,2106	0,0910	0,7744	0,2500	0,0755	0,0146	0,0346
3	0,2702	-0,4157	0,0730	0,1728	-0,1123	0,2500	0,2500	-0,4489	-0,1213	0,1866
Total			0,9045	0,3834	-0,0213				-1,1143	0,2212

Neste caso, i = 5 (i é o genótipo de coordenada (0, 0)) e $d_{\ell i}^2 = d_{15}^2 = 0,67^2 = 0,4489$.

Assim:

$$X_{\ell} = \frac{(\sum X_n Y_n)(\sum Y_n \Delta_{\ell n}) - (\sum Y_n^2)(\sum X_n \Delta_{\ell n})}{2 \left[\sum X_n^2 \sum Y_n^2 - (\sum X_n Y_n)^2 \right]}$$

$$X_1 = \frac{(-0,0213)(0,2212) - (0,3834)(-1,1143)}{2[(0,9045)(0,3834) - (-0,0213)^2]} = 0,6099$$

e

$$Y_{\ell} = \frac{(\sum X_n Y_n)(\sum X_n \Delta_{\ell n}) - (\sum X_n^2)(\sum Y_n \Delta_{\ell n})}{2 \left[\sum X_n^2 \sum Y_n^2 - (\sum X_n Y_n)^2 \right]}$$

$$Y_1 = \frac{(-0,0213)(-1,1143) - (0,9045)(0,2212)}{2[(0,9045)(0,3834) - (-0,0213)^2]} = -0,2545$$

Para avaliação da adequação da projeção realizada, obtém-se as distâncias gráficas a partir das seguintes coordenadas:

Genótipo	1	2	3	4	5
X _i	0,6099	0,1984	0,2702	0,89	0,0
Y _i	-0,2545	0,4589	-0,4157	0,00	0,0

Dessa forma, tem-se a matriz de distâncias originais, dada pelo complemento aritmético do índice de Jaccard (matriz D), e a nova matriz obtida pelas distâncias gráficas (matriz Dg), ou seja:

$$D = \begin{bmatrix} 0 & 0,88 & 0,50 & 0,33 & 0,67 \\ 0 & 0 & 0,88 & 0,83 & 0,50 \\ 0 & 0 & 0 & 0,75 & 0,50 \\ 0 & 0 & 0 & 0 & 0,89 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad Dg = \begin{bmatrix} 0 & 0,8551 & 0,4310 & 0,3364 & 0,7166 \\ 0 & 0 & 0,8775 & 0,8300 & 0,500 \\ 0 & 0 & 0 & 0,7463 & 0,4958 \\ 0 & 0 & 0 & 0 & 0,8900 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Veja que as estimativas de distâncias entre os três primeiros acessos (4, 5 e 2) considerados no processo de projeção são reproduzidas exatamente como as distâncias na matriz original. Para o cálculo da distorção, considera-se:

$$\alpha = \frac{\sum_{i<} \sum_{i'} d_{gii'}^2}{\sum_{i<} \sum_{i'} d_{oii'}^2} = 97,77\%$$

sendo, portanto, a estimativa do grau de distorção dada por:

$$\text{grau distorção} = 1 - \alpha = 100 - 97,77 = 2,23\%$$

Neste exemplo, a correlação entre distâncias originais e as estimadas foi de 0,9750, sendo significativo a 1% de probabilidade pelo teste t feito com g(g - 1)/2 graus de liberdade, em que g é a dimensão da matriz de dissimilaridade.

A estimativa do coeficiente de estresse é dada por:

$$s = 100 \sqrt{\frac{\sum_{i<} \sum_{i'} (d_{oij} - d_{gij})^2}{\sum_{i<} \sum_{i'} d_{oii'}^2}} = 6,54\%$$

O processo de dispersão das medidas de dissimilaridade no plano pode ser considerado satisfatório quando os coeficientes que expressam o grau de distorção e o estresse são inferiores a 20%. O resultado é apresentado graficamente, como ilustrado na Figura 2.13.

Outra ilustração da projeção de medidas de dissimilaridade no plano pode ser considerada tomando-se a matriz de dissimilaridade entre 15 genótipos, por meio da distância euclidiana média padronizada, conforme apresentado na Tabela 2.9. Para este caso, é obtida a dispersão gráfica

ilustrada na Figura 2.14. O gráfico apresenta coeficientes de distorção de 18,40%; correlação entre as distâncias originais e estimadas de 0,96 (significativo a 1% de probabilidade); e coeficiente de estresse de 20,79%. Os resultados apresentados são coerentes com aqueles obtidos.

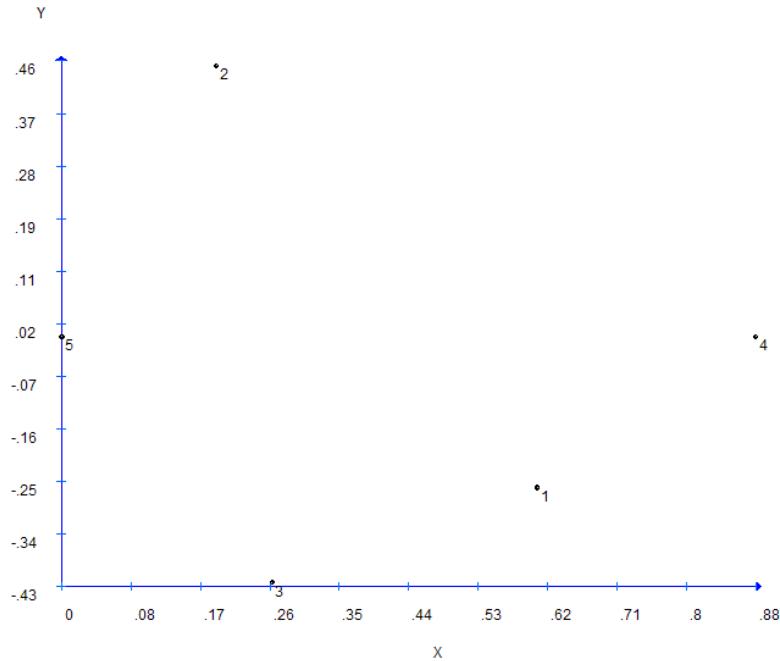


Figura 2.13 - Projeção da dissimilaridade entre cinco genótipos, expressa pelo complemento aritmético do índice de Jaccard, no espaço bidimensional.

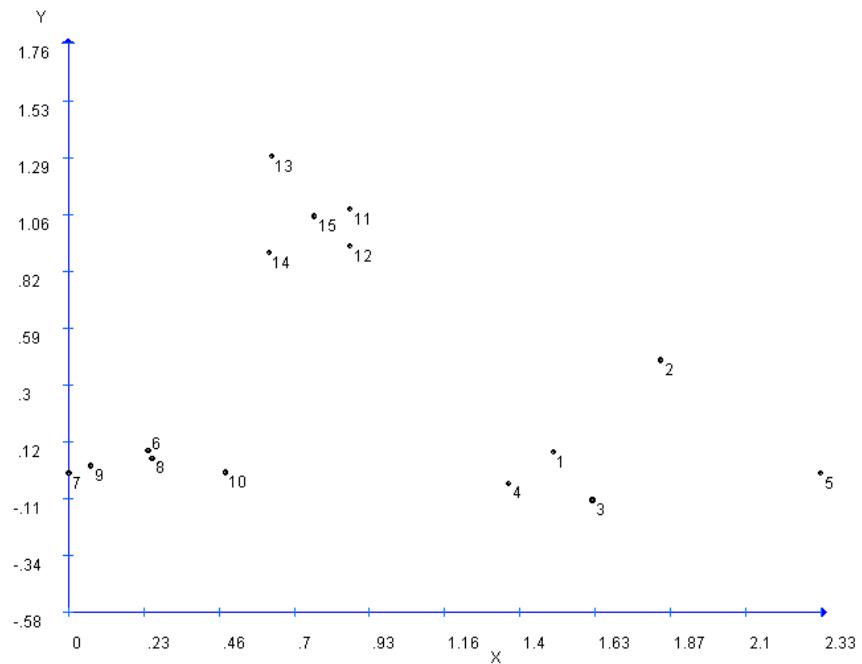


Figura 2.14 - Projeção da dissimilaridade entre 15 genótipos, expressa pela distância euclidiana média padronizada, no espaço bidimensional.

pelas técnicas de agrupamento que apontavam a formação de quatro grupos. O genótipo 5, definido como grupo isolado nos métodos de agrupamentos, está mais próximo do grupo III, formado pelos genótipos 1,2,3 e 4.

Projeção 3D

Neste tipo de projeção, as medidas de dissimilaridade são convertidas em escores relativos a três variáveis (X, Y e Z), que, quando representadas em gráficos de dispersão, irão refletir, no espaço tridimensional (3D), as distâncias originalmente obtidas a partir do espaço n-dimensional ($n =$ número de caracteres utilizados para obtenção das distâncias).

O procedimento consiste em calcular as coordenadas das medidas mais divergentes e, a seguir, daquelas que demonstram, em ordem decrescente, maior diversidade com os pontos (genótipos) já considerados.

Sendo i e j as unidades mais divergentes, a próxima unidade k a ser considerada será aquela de maior valor $d(ij)k$, dado por:

$$d_{(ij)k} = d_{ik} + d_{jk}$$

O mesmo critério é usado para a próxima unidade ℓ , ou seja, escolhe-se ℓ , tal que o valor $d_{(ijk)\ell}$ seja o maior entre todos. Dessa forma, tem-se:

$$d_{(ijk)\ell} = d_{i\ell} + d_{j\ell} + d_{k\ell}$$

A coordenada das duas primeiras unidades é estabelecida arbitrariamente. Considerando duas unidades i e j , tem-se que a coordenada de i é igual a $(0,0,0)$ e a de j igual a $(d_{ij}, 0, 0)$. A coordenada da terceira unidade, dada por $(X_k, Y_k, 0)$, é estabelecida matematicamente, levando-se em conta as propriedades de um triângulo, obtendo-se:

$$X_k = \frac{d_{jk}^2 - d_{ik}^2 - d_{ij}^2}{-2d_{ij}}$$

e

$$Y_k = \sqrt{d_{ik}^2 - X_k^2}$$

A coordenada da quarta unidade, fornecida por (X_l, Y_l, Z_l) , é também estabelecida matematicamente por meio da expressão:

$$X_l = \frac{d_{jl}^2 - d_{il}^2 - d_{ij}^2}{-2d_{ij}}$$

$$Y_l = \frac{d_{kl}^2 - d_{il}^2 - d_{ik}^2 + 2X_l X_k}{-2d_{ij}}$$

$$Z_l = \sqrt{d_{il}^2 - X_l^2 - Y_l^2}$$

A coordenada das demais unidades é estimada estatisticamente, visando minimizar a distorção entre a distância original e a distância gráfica. Assim, a coordenada da unidade ℓ é estimada considerando que:

- e) O genótipo i apresenta coordenada $(0, 0, 0)$.

- f) O genótipo j apresenta coordenada (X_j, Y_j, Z_j) , em que $X_j = d_{ij}$, $Y_j = 0$ e $Z_j = 0$.
- g) O genótipo k apresenta coordenada (X_k, Y_k, Z_k) , sendo $Z_k = 0$ e X_k e Y_k estimados conforme expressão matemática dada anteriormente.
- h) O genótipo l apresenta coordenada (X_l, Y_l, Z_l) estimada conforme descrito anteriormente.
- i) O genótipo m apresenta coordenada (X_m, Y_m, Z_m) estimada pelo sistema de equações:

$$d_{mi}^2 = X_m^2 + Y_m^2$$

$$d_{mj}^2 = (X_m - X_j)^2 + (Y_m - Y_j)^2 + (Z_m - Z_j)^2 = d_{ij}^2 + d_{im}^2 - 2X_j X_m - 2Y_j Y_m - 2Z_j Z_m$$

$$d_{mk}^2 = (X_m - X_k)^2 + (Y_m - Y_k)^2 + (Z_m - Z_k)^2 = d_{ik}^2 + d_{im}^2 - 2X_k X_m - 2Y_k Y_m - 2Z_k Z_m$$

$$d_{ml}^2 = (X_m - X_l)^2 + (Y_m - Y_l)^2 + (Z_m - Z_l)^2 = d_{il}^2 + d_{im}^2 - 2X_l X_m - 2Y_l Y_m - 2Z_l Z_m$$

Este sistema de equações pode ser colocado sob notação matricial $Y = X\beta + \varepsilon$, obtendo-se:

$$Y = \begin{bmatrix} d_{mj}^2 - d_{ij}^2 - d_{im}^2 \\ d_{mk}^2 - d_{ik}^2 - d_{im}^2 \\ d_{ml}^2 - d_{il}^2 - d_{im}^2 \end{bmatrix} \quad X = -2 \begin{bmatrix} X_j & Y_j & Z_j \\ X_k & Y_k & Z_k \\ X_l & Y_l & Z_l \end{bmatrix} \quad \beta = \begin{bmatrix} X_m \\ Y_m \\ Z_m \end{bmatrix} \quad \varepsilon = \begin{bmatrix} e_m \\ e_m \\ e_m \end{bmatrix}$$

Para as demais coordenadas, são acrescentadas linhas no vetor Y e na matriz X, as quais passam a ter as dimensões $(t - 3) \times 1$ e $(t - 3) \times 3$, respectivamente, sendo t o número de genótipos até então estudados.

A solução do sistema é obtida por $X'X\hat{\beta} = X'Y$, de forma que a coordenada estimada para o genótipo m apresente a menor distorção de distância com os demais, cujas coordenadas já foram estabelecidas. Assim, considera-se:

$$X' X = 4 \begin{bmatrix} \sum_{n=1}^t X_n^2 & \sum_{n=1}^t X_n Y_n & \sum_{n=1}^t X_n Z_n \\ \sum_{n=1}^t X_n Y_n & \sum_{n=1}^t Y_n^2 & \sum_{n=1}^t Y_n Z_n \\ \sum_{n=1}^t X_n Z_n & \sum_{n=1}^t Y_n Z_n & \sum_{n=1}^t Z_n^2 \end{bmatrix} \quad \text{e} \quad X' Y = -2 \begin{bmatrix} \sum_{n=1}^t \Delta_{nm} X_n \\ \sum_{n=1}^t \Delta_{nm} Y_n \\ \sum_{n=1}^t \Delta_{nm} Z_n \end{bmatrix}$$

sendo:

$$\Delta_{jm} = d_{jm}^2 - d_{im}^2 - d_{ij}^2$$

$$\Delta_{km} = d_{km}^2 - d_{im}^2 - d_{ik}^2$$

...

$$\Delta_{lm} = d_{lm}^2 - d_{im}^2 - d_{il}^2$$

Ressalta-se o fato de que n é um indexador que assume os valores correspondentes às unidades cujas coordenadas já foram calculadas. Assim, n = i, j, k.

Após obtenção das coordenadas de cada acesso, calcula-se a eficiência da projeção gráfica, comparando-se as estimativas das distâncias originais e das que serão apresentadas no gráfico de dispersão. Como citado anteriormente, as estatísticas utilizadas para medir a eficiência da projeção são: a correlação entre as distâncias originais e as que serão representadas no gráfico de dispersão; o grau de distorção ($1-\alpha$); e o valor do estresse (s).

Novamente será tomada como ilustração a projeção das medidas de dissimilaridade entre 15 genótipos, expressas por meio da distância euclidiana média padronizada, conforme apresentado na Tabela 2.9. Para este caso, é obtida a dispersão gráfica ilustrada na Figura 2.15. O gráfico apresenta coeficientes de distorção de 11,5%; correlação entre as distâncias originais e estimadas de 0,98 (significativo a 1% de probabilidade), e coeficiente de estresse de 13,55%. Os resultados apresentados são coerentes com aqueles obtidos anteriormente.

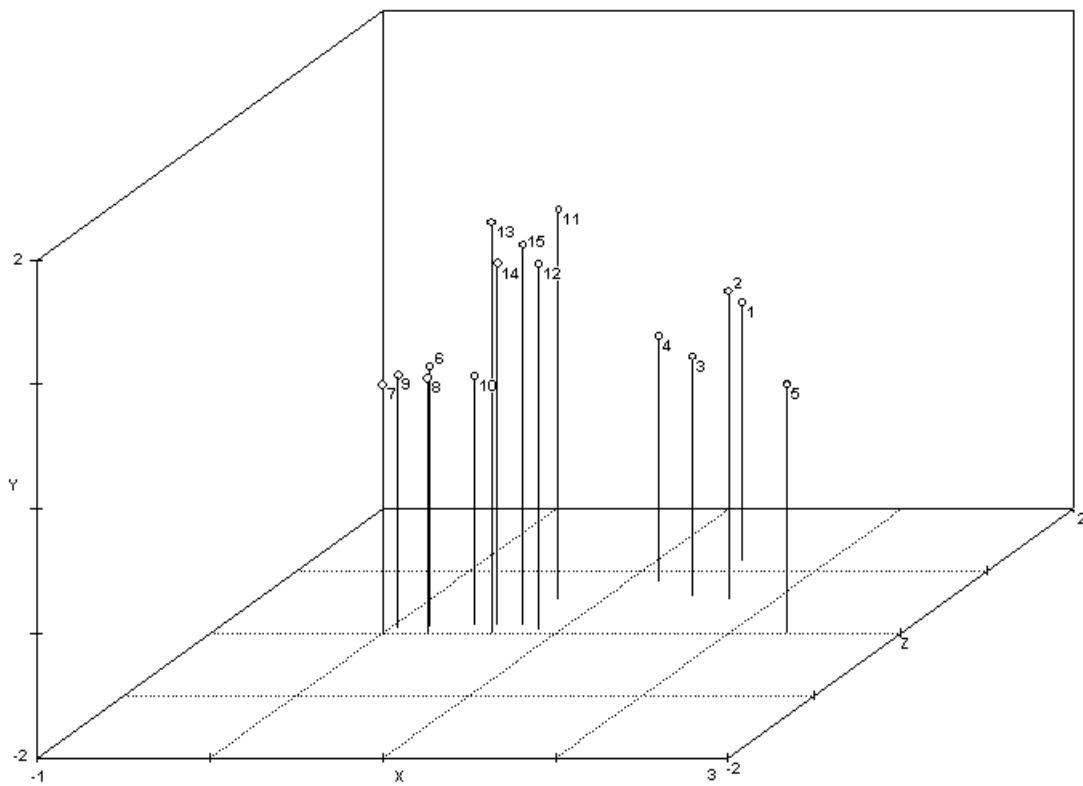


Figura 2.15 - Projeção da dissimilaridade entre 15 genótipos, expressa pela distância euclidiana média padronizada, no espaço tridimensional.

2.5.2. Componentes Principais

A técnica de componentes principais foi originalmente descrita por Pearson (1901) e posteriormente aplicada por Hotelling (1933, 1936) em diversas áreas da ciência. A análise por componentes principais consiste em transformar um conjunto original de variáveis (por exemplo, altura, produção etc.) em outro conjunto de dimensão equivalente, mas com propriedades importantes, que são de grande interesse em certos estudos de melhoramento. Cada componente principal é uma combinação linear das variáveis originais. Além disso, são independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo da informação, em termos de variação total, contida nos dados iniciais.

A viabilidade de utilização dos componentes principais em estudos sobre divergência genética dependerá da possibilidade de resumir o conjunto de variáveis originais em poucos componentes, o que significará ter boa aproximação do comportamento dos indivíduos (genótipos), oriundo de um espaço v-dimensional ($v =$ número de caracteres estudados) em um espaço bi ou tridimensional. Quando este requisito for satisfeito, a referida técnica proporcionará uma simplificação considerável nos cálculos estatísticos e na interpretação dos resultados com relação aos demais métodos alternativos, principalmente quando o número de genótipos avaliado for relativamente grande.

Em geral, os primeiros componentes principais em estudos da divergência genética têm sido utilizados quando estes envolvem pelo menos 80% da variação total. Nos casos em que este limite não é atingido nos dois primeiros componentes, a análise é complementada com a dispersão gráfica em relação ao terceiro e quarto componentes.

Além de possibilitar o estudo da diversidade genética de um grupo de acessos, a técnica dos componentes principais tem a vantagem de possibilitar a avaliação da importância de cada caráter estudado sobre a variação total disponível entre os genótipos avaliados. O interesse nesta avaliação reside na possibilidade de se descartarem caracteres que contribuem pouco para a discriminação do genótipo avaliado, reduzindo, dessa forma, mão-de-obra, tempo e custo despendidos na experimentação agrícola.

A técnica de componentes principais baseia-se apenas nas informações individuais de cada acesso, sem a necessidade de dados com repetições.

Para a realização da análise, geralmente feita com dados padronizados, considera-se que x_{ij} é a média padronizada do j-ésimo caráter ($j = 1, 2, \dots, v$) avaliado no i-ésimo genótipo ($i = 1, 2, \dots, g$) e R a matriz de covariâncias ou de correlação entre esses caracteres (ou matriz de correlação fenotípica entre os caracteres baseada nos dados originais). A técnica dos componentes principais consiste em transformar o conjunto de v variáveis ($x_{i1}, x_{i2}, \dots, x_{iv}$) em um novo

conjunto ($Y_{i1}, Y_{i2}, \dots, Y_{iv}$), que são funções lineares dos x_j 's e independentes entre si.

As seguintes propriedades são verificadas:

- a) Se Y_{i1} é um componente principal, então Y_{i1} é uma combinação linear das variáveis x_j 's, como descrito a seguir:

$$Y_{i1} = a_1 x_{i1} + a_2 x_{i2} + \dots + a_v x_{iv}$$

- b) Se Y_{i2} é outro componente principal, então Y_{i2} é outra combinação linear das variáveis x_j 's, ou seja:

$$Y_{i2} = b_1 x_{i1} + b_2 x_{i2} + \dots + b_v x_{iv}$$

- c) Entre todos os componentes, Y_{i1} apresenta a maior variância, Y_{i2} a segunda maior, e assim sucessivamente.

Também são consideradas as restrições:

$$\sum_j a_j^2 = \sum_j b_j^2 = 1$$

$$\sum_j a_j b_j = 0, \text{ ou seja, os componentes são não-correlacionados.}$$

Com base na propriedade c, objetiva-se, em estudos sobre divergência genética por meio dos componentes principais, avaliar a possibilidade de estudar a dispersão dos genótipos em sistemas de eixos cartesianos nos quais o aproveitamento da variabilidade disponível seja maximizada. O problema estatístico consiste fundamentalmente em estimar os coeficientes de ponderação dos caracteres em cada componente e a variância a eles associada.

Sendo Y_{i1} (ou simplesmente Y_1) o primeiro componente principal, sua variância é dada por:

$$V(Y_{i1}) = V(Y_1) = \sum_j a_{j1}^2 r_{jj} + \sum_{j \neq j'} \sum_j a_j a_{j'} r_{jj'} = \sum_j \sum_{j'} a_j a_{j'} r_{jj'}$$

em que r_{jj} é o elemento da j-ésima linha e da j'-ésima coluna de R, lembrando que $r_{jj} = 1$.

Sob forma matricial, tem-se:

$$V(Y_1) = a'Ra$$

em que a' é um vetor $1 \times v$ de elementos a_j ($j = 1, 2, \dots, v$).

Objetiva-se obter o vetor a de forma que a variância de Y_1 seja maximizada, impondo-se a restrição no conjunto de soluções de a por meio de $a'a = 1$. Expressando a variância de Y_1 pela função ω_1 e incorporando a restrição pelo multiplicador λ_1 de Lagrange, tem-se:

$$\omega_1 = a'Ra + \lambda_1(1 - a'a)$$

Por diferenciação, encontra-se:

$$\delta\omega_1 = 2\delta a'Ra - 2\lambda_1\delta a'a, \text{ ou}$$

$$\frac{\delta\omega_1}{\delta a} = 2(R - \lambda_1I)a$$

Fazendo $\delta\omega/\delta a = \phi$, tem-se:

$$(R - \lambda_1I)a = \phi \quad (1)$$

A solução do sistema deve ser tal que $a \neq \phi$. Assim, é necessário que o determinante de $R - \lambda_1I$ seja nulo ($|R - \lambda_1I| = 0$) para que o sistema se torne indeterminado e a solução possa ser escolhida entre aquelas que satisfaçam a condição $a'a = 1$.

Sendo λ_1 os valores que satisfazem $|R - \lambda_1I| = 0$, então, por definição, λ_1 são as raízes características (ou autovalores) de R e a o vetor característico (ou autovetor) associado. Para o primeiro componente principal, o valor de λ_1 deve ser o maior dos v autovalores estimados, pois pré-multiplicando (1) por a' , verifica-se que:

$$a'Ra - a'a\lambda_1 = 0$$

logo:

$$\lambda_1 = \mathbf{a}' \mathbf{R} \mathbf{a} = V(Y_1)$$

Como o vetor a foi escolhido para maximizar $V(Y_1)$, tem-se que λ_1 assume, nesta condição, o valor máximo entre os elementos do conjunto de autovalores de R .

A variância do segundo componente principal é dada por:

$$V(Y_{i_2}) = V(Y_2) = \mathbf{b}' \mathbf{R} \mathbf{b}$$

Na obtenção do vetor b , cujos elementos são os coeficientes b_j ($j = 1, 2, \dots, v$), devem-se considerar as restrições $b'b = 1$ e $b'a = a'b = 0$, as quais são incorporadas na função de maximização por meio dos multiplicadores λ_2 e θ de Lagrange. Assim, é estabelecido que:

$$\omega_2 = b' \mathbf{R} b + \lambda_2(1 - b'b) + \theta a'b.$$

A restrição $b'b = 1$ é necessária para garantir a unicidade de b , ao passo que $a'b = 0$ garante que Y_1 e Y_2 sejam ortogonais.

A solução que maximiza ω_2 é obtida pela derivação de ω_2 em relação a b , dada por:

$$\frac{\delta \omega_2}{\delta b} = 2(\mathbf{R} - \lambda_2 I)b + \theta a$$

Fazendo $\delta \omega_2 / \delta b = \phi$, tem-se:

$$2(\mathbf{R} - \lambda_2 I)b + \theta a = \phi \quad (2)$$

Pré-multiplicando (2) por a' , obtém-se:

$$2a' \mathbf{R} b + \theta a = 0 \quad (3)$$

Pré-multiplicando (1) por b' , obtém-se:

$$b' Ra = a' Rb = 0. \quad (4)$$

Substituindo (4) em (3), conclui-se que $\theta = 0$. Assim, (2) pode ser simplificado para:

$$(R - \lambda_2 I) b = 0 \quad \text{e} \quad |R - \lambda_2 I| = 0$$

em que se conclui que λ_2 é o segundo maior autovalor de R e b o seu autovetor associado. Os demais componentes principais são estimados de maneira análoga à descrita para os dois primeiros.

Importância Relativa de um Componente

A importância relativa de um componente é avaliada pela porcentagem da variância total que ele explica. A importância relativa (IR) de cada componente, útil na avaliação da capacidade de discriminação do componente estimado, é calculada por meio da expressão:

$$IR_j = \frac{\lambda_j}{\text{Traço}(R)}$$

Em estudos sobre a divergência genética entre genótipos, é desejável que a variância acumulada nos dois primeiros componentes principais exceda 80%. Neste caso, a distorção das coordenadas de cada cultivar, no gráfico de dispersão cujos eixos são os componentes principais, será considerada aceitável e as inferências no estudo da diversidade genética, satisfatórias.

Relação entre Componentes Principais e a Análise de Agrupamento

Um dos objetivos do uso dos componentes principais em estudo sobre a divergência genética é avaliar a dissimilaridade dos genótipos em gráficos de dispersão que têm os primeiros componentes como eixos de referência. Como nesta técnica é feita uma simplificação do espaço n-dimensional para o bi ou tridimensional, há certas distorções nas distâncias gráficas. Entretanto, há entre as estimativas das distâncias gráficas, ou distâncias euclidianas baseadas nos escores

dos primeiros componentes principais, e a distância euclidiana baseada nos dados originais n-dimensional uma relação matemática dada por:

$$\alpha = \frac{\sum_{i < i'} \sum d^2 cp_{ii'}}{v \sum_{i < i'} d^2_{ii'}}$$

em que:

$d^2 cp_{ii'}$: quadrado da distância euclidiana estimado a partir dos escores de v_1 componentes principais; e

$d^2_{ii'}$: quadrado da distância euclidiana média estimado a partir das v variáveis originais.

A estatística $1 - \alpha$ mede o grau de distorção proporcionado pela técnica dos componentes principais ao passar do espaço n-dimensional para o n_1 -dimensional ($v_1 < v$). Deduz-se ainda que α é a porcentagem acumulada da variância, explicada pelos n_1 componentes principais.

Importância Relativa dos Caracteres na Divergência Genética

Outra informação interessante, em estudos de diversidade genética, é a identificação de variáveis que menos contribuem para a diferenciação dos genótipos, sendo possível o seu descarte em estudos futuros.

Para fazer o descarte de variáveis, procura-se identificar os componentes cuja variância seja nula ou bem próxima de zero, os quais representarão combinações lineares que resultam em valores de escores dados por uma constante. Pelo menos uma das variáveis que determina esta combinação linear perfeita, ou quase perfeita, poderá ser eliminada em estudos futuros, por já está sendo representada pelas demais da relação, sem prejuízo de informação relativa à diversidade entre os acessos. Com base nesse princípio, pela técnica de componentes principais, as variáveis de maiores pesos nos últimos autovetores são consideradas de menor importância para estudo da diversidade genética. Normalmente consideram-se os últimos autovetores associados a autovalores da

matriz de correlação aquele(s) nulo(s) ou, de forma mais flexível, aqueles cuja magnitude seja inferior a 0,7. Por outro lado, as variáveis de maiores pesos nos primeiros autovetores são consideradas de maior importância para o estudo da diversidade genética quando o autovalor explica uma fração considerável da variação disponível, normalmente limitado em valor mínimo de 80%.

Em estudos com caracteres padronizados, nos quais os autovetores são obtidos a partir da matriz de correlação, tem sido comum descartar o caráter de maior coeficiente (em valor absoluto) a partir do último componente até aquele cujo autovalor não excede 0,70, segundo recomendações de Jolliffe (1972, 1973) e Mardia et al. (1979). Quando em um componente de menor variância o maior coeficiente de ponderação está associado a um caráter já previamente descartado, tem-se optado por não fazer nenhum outro descarte com base nos coeficientes daquele componente, mas prosseguir a identificação da importância relativa dos caracteres no outro componente de variância imediatamente superior.

A importância dos caracteres pode, alternativamente, ser avaliada no conjunto de cargas totais associadas às variáveis, considerando que:

$$C_j = \alpha R$$

em que C_j é o vetor linha de ordem j , cujos elementos são C_{kj} .

A carga da variável k no j -ésimo componente principal é dada por:

$$\text{Carga} = \frac{C_{kj}}{\sqrt{\lambda_j}}$$

Neste caso, considera-se que a correlação entre a k -ésima variável e o j -ésimo componente é dada por:

$$r_{kj} = \frac{\text{Cov}(x_k, y_j)}{\sqrt{V(x_k)V(y_j)}} = \frac{\alpha_{1j}r_{1k} + \alpha_{2j}r_{2k} + \dots + \alpha_{kj} + \dots + \alpha_{vj}r_{vk}}{\sqrt{\lambda_j}}$$

sendo α_{ij} o i -ésimo elemento do j -ésimo autovetor.

Ilustração

Será considerada, como ilustração, a obtenção dos componentes principais a partir das médias padronizadas de 15 genótipos de milho-pipoca, avaliados em relação aos caracteres altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela) (Tabela 2.8).

Obteve-se a matriz de covariâncias entre os valores padronizados, que corresponde à matriz de correlação dos dados expressos pelas médias originais (Tabela 2.8), a partir da qual se estimaram os autovalores e autovetores apresentados na Tabela 2.16.

Na Tabela 2.16, verifica-se serem necessários três componentes para explicarem um mínimo de 80% da variação originalmente disponível nos dados. Neste caso, recomenda-se fazer a dispersão gráfica bidimensional considerando os componentes 1 e 2 e, posteriormente, os componentes 1 e 3. Nesses gráficos, os genótipos serão considerados similares se consistentemente situarem-se próximos e serão considerados dissimilares se consistentemente situarem-se graficamente distantes. Não havendo consistência, conclui-se que a distorção gráfica impossibilita inferir o padrão de similaridade dos genótipos analisados. Outra opção é a representação tridimensional, em que se consideram simultaneamente os três primeiros componentes principais.

A Tabela 2.16 também permite avaliar a importância relativa das características sobre a diversidade genética. Características menos importantes são aquelas correlacionadas com outras consideradas no estudo. Fazendo a análise dos elementos dos seis últimos autovetores, ou seja, do último até aquele em que o valor do autovetor

Tabela 2.16 - Estimativas de autovalores obtidos da matriz de correlação entre os caracteres altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela), avaliados em cultivares de milho-pipoca

λ_j	λ_j (%) acumulada	Elementos dos autovetores associados a								
		AP	AE	NE	PROL	CE	QUE	DOEN	PCG	PROD
4,5799	50,89	-0,358	-0,370	-0,336	-0,326	0,427	-0,417	-0,337	-0,030	-0,214
2,4252	77,84	0,195	0,168	-0,322	-0,354	0,178	0,140	-0,059	0,620	0,515
0,9087	87,93	0,455	0,374	0,268	-0,092	0,152	-0,017	-0,695	-0,054	-0,252
0,4289	92,70	0,490	-0,302	-0,391	0,605	0,121	-0,206	-0,120	-0,134	0,245
0,2708	95,71	-0,153	0,588	0,079	0,096	0,150	-0,670	0,186	-0,138	0,230
0,1804	97,72	0,088	-0,480	0,734	-0,051	0,192	-0,193	-0,004	0,149	0,352
0,1179	99,02	0,326	-0,130	-0,116	-0,587	-0,384	-0,134	0,043	-0,534	0,260
0,0641	99,74	0,305	0,062	0,010	-0,161	0,698	0,229	0,489	-0,260	-0,175
0,0237	100,00	0,396	-0,080	-0,013	-0,108	-0,233	-0,455	0,332	0,447	-0,503

obtido da matriz de correlação é inferior a 0,7, identificam-se, no estudo, os caracteres PROD, CE, PROL, NE, QUE e, novamente, PROL como passíveis de descarte. Uma maneira alternativa de avaliar a importância dos caracteres é por meio das cargas canônicas, como apresentado na Tabela 2.17. Entretanto, deve ser ressaltado que as cargas canônicas medem a associação total de uma variável com o componente, incluindo seus efeitos indiretos sobre as demais variáveis consideradas no estudo.

Nas Figuras 2.16 e 2.17 são apresentadas as dispersões de 15 cultivares em relação a eixos estabelecidos por componentes principais. No primeiro gráfico é nítida a formação de três grupos, tal como já estabelecido pelos métodos de agrupamento.

As informações da Figura 2.16 devem ser utilizadas apenas como complementares às informações da Figura 2.15, uma vez que apresenta grande distorção. O terceiro componente, por reter menor proporção da variação disponível, dispõe os genótipos com maior proximidade. Neste gráfico, os

grupos continuam bem estabelecidos, apesar de o grupo I (estabelecido pelo método de Tocher, Tabela 2.14) localizar-se mais próximo do grupo II. O genótipo 5 aparece ainda mais distanciado do grupo III, mostrando a sua

Tabela 2.17 - Estimativas das cargas canônicas associadas aos caracteres altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela), avaliados em cultivares de milho-pipoca

AP	AE	NE	PROL	CE	QUE	DOEN	PCG	PROD
-0,7661	-0,7924	-0,7192	-0,6972	0,9129	-0,8918	-0,7206	-0064	-0,4581
0,3038	0,2609	-0,5014	-0,5508	0,2769	0,2179	-0,0912	0,9663	0,8026
0,4337	0,357	0,2554	-0,0879	0,1452	-0,0162	-0,6624	-0,0515	-0,2399
0,3207	-0,1977	-0,2561	0,3966	0,0796	-0,1347	-0,0785	-0,0876	0,1607
-0,0799	0,306	0,0413	0,0499	0,0779	-0,3489	0,0966	-0,0717	0,1559
0,0373	-0,2039	0,3118	-0,0216	0,0816	-0,0819	-0,0017	0,0634	0,1497
0,1118	-0,0447	-0,0398	-0,2014	-0,1317	-0,0461	0,0147	-0,1832	0,0893
0,0771	0,0157	0,0025	-0,0408	0,1767	0,058	0,1238	-0,0658	-0,0442
0,061	-0,0124	-0,0021	-0,0167	-0,036	-0,070	0,0512	0,0689	-0,0775

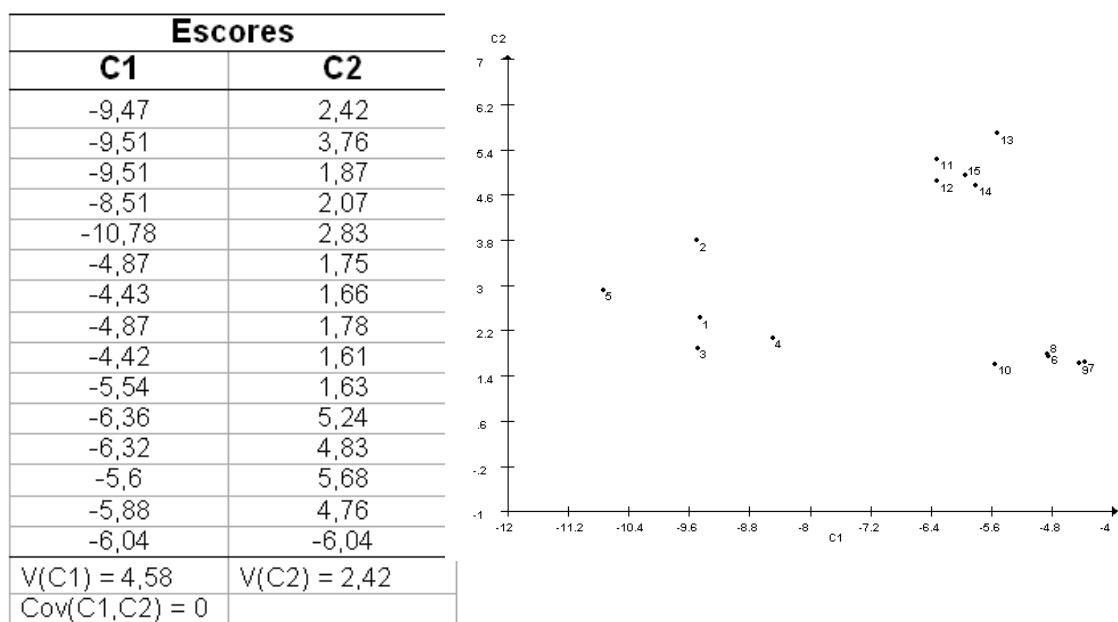


Figura 2.16 - Dispersão gráfica de 15 cultivares de milho-pipoca, em relação ao primeiro e ao segundo componente principal, estabelecido pela combinação linear de nove características agronômicas.

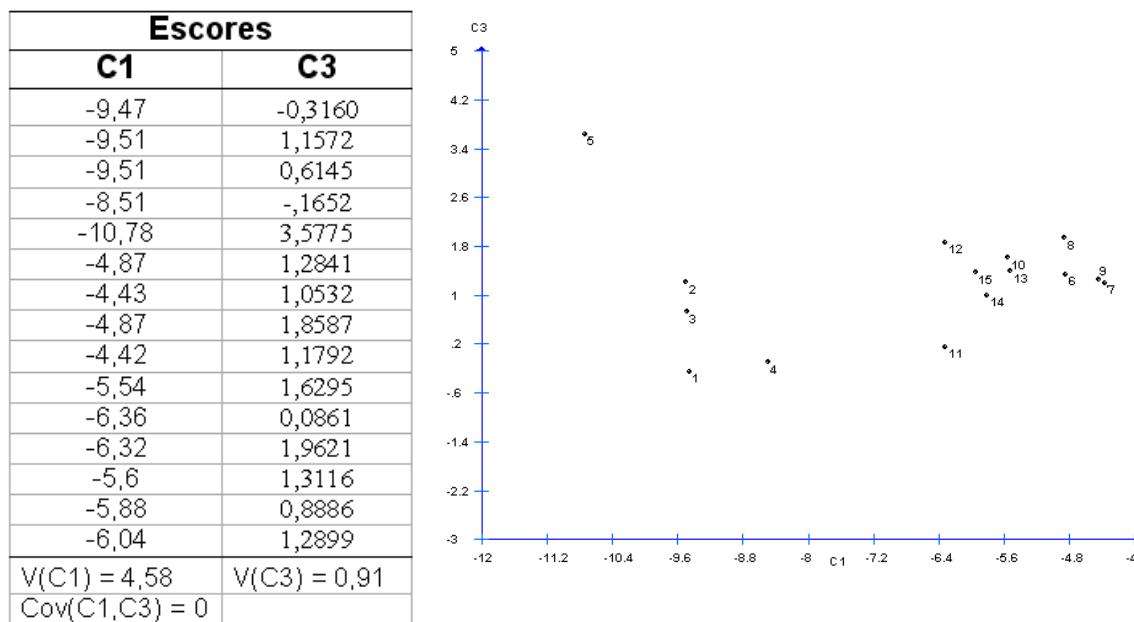


Figura 2.17 - Dispersão gráfica de 15 cultivares de milho-pipoca, em relação ao primeiro e ao terceiro componente principal, estabelecido pela combinação linear de nove características agronômicas.

diversidade em relação aos genótipos integrantes deste grupo. Na Figura 2.18 é apresentada a dispersão numa figura tridimensional, em que são considerados simultaneamente os eixos representativos dos três primeiros componentes principais. Também nesta figura constata-se a existência de três grupos (incluindo o genótipo 5 no grupo III), tal como já constatado nos métodos anteriores. É necessário que o programa utilizado para plotar os três eixos simultaneamente permita a rotação destes para que se possa escolher uma posição de maior dispersão dos genótipos e, consequentemente, maior visualização da diversidade existente entre os genótipos avaliados.

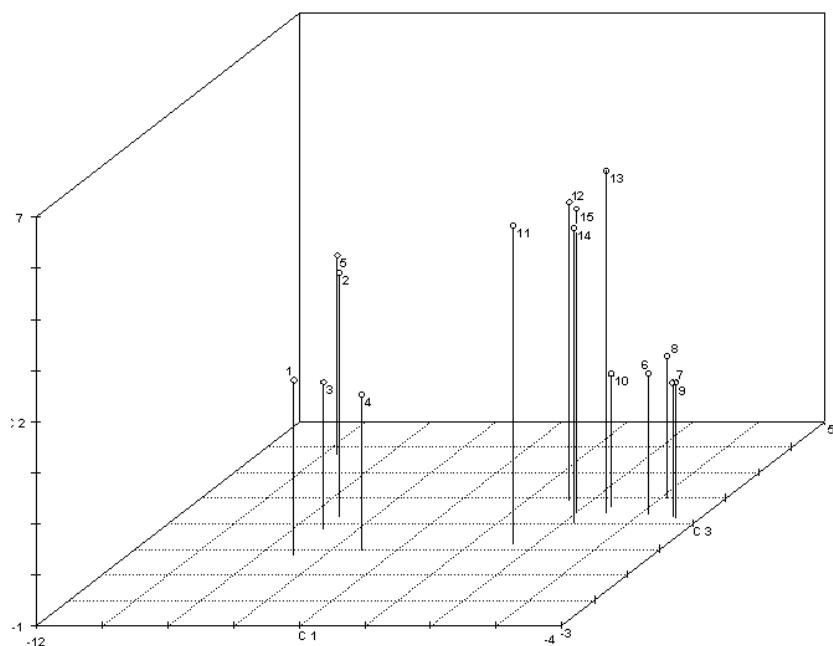


Figura 2.18 - Dispersão gráfica de 15 cultivares de milho-pipoca, em relação aos três primeiros componentes principais, estabelecidos pela combinação linear de nove características agronômicas.

2.5.3. Variáveis canônicas

A análise multivariada, com base em variáveis canônicas, foi relatada por Rao (1952). Trata-se de um processo alternativo para avaliação do grau de similaridade genética entre genótipos que leva em consideração tanto a matriz de covariância residual quanto a de covariância fenotípica entre os caracteres avaliados.

A técnica de variáveis canônicas é similar à de componentes principais, pois permite a simplificação no conjunto de dados, resumindo as informações, originalmente contidas em um grupo de v variáveis, em poucas variáveis, que apresentam as propriedades de reterem o máximo da variação originalmente disponível e serem independentes entre si. Entretanto, esta técnica baseia-se nas informações *entre e dentro* de genótipos (ou entre indivíduos de cada genótipo), havendo, portanto, necessidade de dados, em nível de acessos, com repetições.

Considera-se, de forma geral, que a análise de variáveis canônicas constitui-se em um procedimento alternativo à análise de componentes principais nas situações em que se dispõe de dados experimentais com informações de repetições, de modo que seja possível obter as médias e a matriz de dispersão (matriz de variâncias e covariâncias) residual entre os dados.

A análise por variáveis canônicas, quando utilizada em estudos de divergência genética, tem como propósito possibilitar a identificação de genótipos similares em gráficos de dispersão bi ou tridimensionais, à semelhança dos componentes principais. Esta técnica apresenta a vantagem adicional de manter o princípio do processo de agrupamento com base na distância D^2 , de Mahalanobis, o qual leva em conta as correlações residuais existentes entre as médias dos genótipos.

A viabilidade do uso das variáveis canônicas em estudos sobre divergência genética, em gráficos de dispersão, também está restrita à concentração da variabilidade disponível entre as primeiras variáveis. Sua estimativa, ao contrário dos componentes principais, requer o conhecimento da matriz residual, que, em muitas condições experimentais, como aquelas inerentes à avaliação de bancos de

germoplasma ou coleta de dados em condições naturais, não estão disponíveis ou são de difícil estimação.

Obtenção das Variáveis Canônicas

Seja X_{ij} a média do j-ésimo caráter ($j = 1, 2, \dots, v$) avaliada no i-ésimo genótipo ($i = 1, 2, \dots, g$), T a matriz de covariâncias entre médias de genótipos e E a matriz de covariâncias residuais. A técnica de variáveis canônicas, à semelhança dos componentes principais, consiste em transformar o conjunto de n variáveis originais em um novo conjunto de variáveis que são funções lineares dos X_i 's.

As seguintes propriedades são verificadas:

a) Se Y_{ij} é uma variável canônica, então:

$$Y_{ij} = a_1 X_{i1} + a_2 X_{i2} + \dots + a_v X_{iv}$$

b) Se $Y_{ij'}$ é outra variável canônica, então:

$$Y_{ij'} = b_1 X_{i1} + b_2 X_{i2} + \dots + b_v X_{iv}$$

e

$$\sum_j \sum_{j'} a_j a_{j'} \sigma_{jj'} = \sum_j \sum_{j'} b_j b_{j'} \sigma_{jj'} = 1$$

$$\sum_j \sum_{j'} a_j b_{j'} \sigma_{jj'} = 0$$

$\sigma_{jj'}$ é a covariância residual entre os caracteres j e j'.

c) Dentre todas as variáveis canônicas, Y_{i1} apresenta a maior variância, Y_{i2} a segunda maior, e assim sucessivamente.

A propriedade (b) garante a ponderação da influência das variâncias e covariâncias residuais sobre as estimativas dos coeficientes de cada caráter, bem como a independência entre essas variáveis.

Nesse caso, o problema estatístico consiste em estimar os coeficientes de ponderação dos caracteres em cada variável canônica e as suas respectivas variâncias.

Sendo Y_{11} (ou simplesmente Y_1) a primeira variável canônica, então sua variância é dada por:

$$V(Y_{11}) = V(Y_1) = \sum_j a_{jj}^2 t_{jj} + \sum_{j \neq j'} a_j a_{j'} t_{jj'} = \sum_j a_j a_{j'} t_{jj'}$$

em que $t_{jj'}$ é o elemento da j -ésima linha e j' -ésima coluna de T .

Sob forma matricial, tem-se:

$$V(Y_1) = a' T a$$

em que a' é um vetor $1 \times v$ de elementos a_j ($j = 1, 2, \dots, v$).

Objetiva-se obter o vetor a de forma que a variância de Y_1 seja maximizada e que os elementos deste vetor sejam estimados, a fim de que as influências das variâncias e covariâncias residuais sejam eliminadas. Assim, obtém-se a impondo-se a restrição $a'Ea = 1$, que é incorporada à função da variância de Y_1 , representada por ω_1 , pelo multiplicador λ_1 de Lagrange. Assim, tem-se:

$$\omega_1 = a' T a + \lambda_1 (1 - a'Ea)$$

Por diferenciação, encontra-se:

$$\delta\omega_1 = 2\delta a' T a - 2\lambda_1 \delta a'Ea$$

ou

$$\frac{\delta\omega_1}{\delta a} = 2(T - \lambda_1 E)a$$

Fazendo $\delta\omega_1/\delta a = \phi$, tem-se:

$$(T - \lambda_1 E)a = \phi \quad (1)$$

Pré-multiplicando a equação (1) por E^{-1} , resulta:

$$(E^{-1}T - \lambda_1 I)a = \phi$$

logo, à semelhança do relatado anteriormente, λ_1 são as raízes características (ou autovalores) de $E^{-1}T$ e a o respectivo vetor característico (ou autovetor) associado.

Deve ser ressaltados alguns aspectos importantes ao se fazer análise de variáveis canônicas. O primeiro diz respeito à necessidade de investigar problemas relativos à existência de multicolinearidade entre as variáveis analisadas, que poderia levar à singularidade da matriz E. Procedimentos para diagnósticos de multicolinearidade são descritos em Cruz e Carneiro (2006) e devem ser aplicado à matriz E ou, de forma alternativa, à matriz R_e cujos elementos são as correlações residuais entre pares de características estudadas. Outro aspecto, de natureza computacional, diz respeito à necessidade de se dispor de aplicativos capazes de obter autovalores e autovetores de matriz assimétrica, tal como $E^{-1}T$, que apesar de ser produto de duas matrizes simétricas apresenta assimetria. Por fim, deve-se estar atento à dimensionalidade da matriz $E^{-1}T$, identificando-se o número provável de autovalores não nulos em razão do número de variáveis e de genótipos usados nas análises.

Para a primeira variável canônica, o valor de λ_1 deve ser o maior dos n autovalores estimados, pois pré-multiplicando (1) por a' , verifica-se que:

$$a' T a - \lambda_1 a' E a = 0$$

logo:

$$\lambda_1 = a' T a = V(Y_1)$$

Como o vetor a foi escolhido para maximizar $V(Y_1)$, λ_1 assume, nesta condição, o valor máximo entre suas possíveis estimativas.

A variância da segunda variável canônica é dada por:

$$V(Y_{i2}) = V(Y_2) = b' T b$$

O vetor b , cujos elementos são os coeficientes b_j ($j = 1, 2, \dots, v$), deve ser estimado de forma que a $V(Y_2)$ seja maximizada, as influências das variâncias e covariâncias residuais sejam levadas em consideração e que as variáveis canônicas sejam não-correlacionadas. Estas duas últimas condições são incorporadas na função de maximização por meio dos multiplicadores λ_2 e θ de Lagrange. Dessa forma, é estabelecido que:

$$\omega_2 = b' Tb + \lambda_2(1 - b' Eb) + \theta b' Ea.$$

A solução que maximiza ω_2 é obtida pela derivação de ω_2 em relação a b , dada por:

$$\frac{\delta \omega_2}{\delta b} = 2(T - \lambda_2 E)b + \theta Ea$$

Fazendo $\delta \omega_2 / \delta b = \phi$, tem-se:

$$2(T - \lambda_2 E)b + \theta Ea = \phi \quad (2)$$

Pré-multiplicando (2) por a' , obtém-se:

$$2a' Tb + \theta = 0 \quad (3)$$

Pré-multiplicando (1) por b' , obtém-se:

$$b' Ta = a' Tb = 0. \quad (4)$$

Substituindo (4) em (3), conclui-se que $\theta = 0$. Assim, (2) pode ser simplificado para:

$$(T - \lambda_2 E)b = \phi$$

em que se conclui que λ_2 é o segundo maior autovalor de $E^{-1}T$ e b o seu autovetor associado. As demais variáveis canônicas são estimadas de maneira análoga à anteriormente descrita para as duas primeiras variáveis canônicas.

Quando se utiliza esse procedimento, é comum a transformação das variáveis originais em variáveis padronizadas e não-correlacionadas, de modo que a matriz de dispersão se iguala à identidade, ou seja, covariâncias residuais nulas e variâncias residuais iguais a um. Para essa transformação, tem sido utilizado o processo de condensação pivotal. Uma vez realizada a transformação, o processo de estimação equivale ao descrito para componentes principais.

Feita a transformação, os parâmetros são estimados da mesma maneira descrita para os componentes principais. Admitindo as matrizes de covariâncias entre médias, matriz T , e a de dispersão residual, matriz E , verifica-se que, após a condensação pivotal, as variáveis transformadas apresentam matriz de covariância

entre médias dada por T^* e matriz de covariâncias residuais igual à identidade ($E^* = I$).

A transformação é obtida por meio de $Z' = VX$, em que:

Z : matriz $g \times v$ de médias transformadas de g genótipos em relação aos v caracteres;

X : matriz $g \times v$ de médias originais; e

V : matriz $v \times v$ de transformação, obtida pelo processo de condensação pivotal.

As estimativas dos autovalores, que medem a variância de cada variável canônica, são obtidas por meio de:

$$\det(T^* - I\lambda) = 0, \text{ que equivale aos autovalores obtidos de } \det(E^{-1}T - I\lambda) = 0$$

As estimativas dos autovetores associados às variáveis transformadas por condensação pivotal são obtidas por meio de:

$$(T^* - I\lambda)\alpha = \phi$$

Neste caso, α representa o autovetor cujos elementos são coeficientes de ponderação das variáveis obtidas por condensação pivotal. É interessante estimar os coeficientes de ponderação associados às variáveis originais, para que seja avaliada a contribuição de cada característica para uma determinada variável canônica. Esses coeficientes constituem o autovetor a , que pode ser obtido de α ou a partir do sistema:

$$(E^{-1}T - I\lambda)a = \phi$$

Para a dispersão gráfica é indiferente considerar uma combinação linear de variáveis transformadas (por condensação pivotal) ou a combinação linear das características originais, pois os escores obtidos serão os mesmos. Ressalta-se que a análise gráfica, para estudo do padrão de similaridade entre os genótipos, deve ser considerada quando for possível resumir em poucas variáveis mais de 80% da variação total disponível. Assim, considera-se que:

$$VC_1 = \alpha_{11}Z_1 + \alpha_{12}Z_2 + \dots + \alpha_{1v}Z_v = a_{11}X_1 + a_{12}X_2 + \dots + a_{1v}X_v$$

...

$$VC_n = \alpha_{n1}Z_1 + \alpha_{n2}Z_2 + \dots + \alpha_wZ_v = a_{n1}X_1 + a_{n2}X_2 + \dots + a_wX_v$$

Em VC_1, VC_2, \dots, VC_n , tem-se:

$$\sum_j \alpha_{jj}^2 = 1, \text{ para cada } j=1,2,\dots,n; \text{ e}$$

$$\sum_j \alpha_{jj} \alpha_{jj'} = 0, \text{ para qualquer par } j' \neq j \text{ de variáveis canônicas estimadas por}$$

um processo idêntico ao descrito para os componentes principais.

Uma vez estimados os coeficientes $\alpha_{jj'}$, os coeficientes $a_{jj'}$, associados às variáveis originais, podem ser calculados por meio de:

$$[a_{j1} \ a_{j2} \ \dots \ a_{jv}] = [\alpha_{j1} \ \alpha_{j2} \ \dots \ \alpha_{jv}] V$$

Importância Relativa de uma Variável Canônica

A importância relativa de cada variável canônica é também dada pela razão entre a variância por ela explicada e o total da variância disponível. Uma vez que há, nas primeiras variáveis, a concentração de grande proporção da variância total, em geral referenciada como acima de 80%, é viável o estudo da divergência genética por meio das distâncias geométricas entre genótipos em gráficos de dispersão, cujas coordenadas são escores relativos às primeiras variáveis canônicas.

Relação entre Variáveis Canônicas e a Análise de Agrupamento

A utilização das variáveis canônicas tem por objetivo básico proporcionar uma simplificação estrutural dos dados, de modo que a divergência genética, influenciada a princípio por um conjunto v -dimensional ($v =$ número de caracteres considerados no estudo), possa ser avaliada por um complexo no espaço bi ou tridimensional de fácil interpretação geométrica. A eficácia de sua utilidade depende do grau de distorção provocado nas distâncias entre os progenitores quando se passa do espaço n -dimensional para o v_1 -dimensional ($v_1 < v$).

Como as distâncias gráficas, em relação a eixos que representam as primeiras variáveis canônicas, são influenciadas pelas variações entre (variâncias e covariâncias fenotípicas) e dentro (variâncias e covariâncias residuais), pode-se

quantificar o grau de distorção dessas distâncias comparando o seu total com o total das distâncias generalizadas de Mahalanobis, ou seja:

$$\text{Grau de distorção} = 1 - \alpha$$

sendo:

$$\alpha = \frac{\sum_{i < i'} \sum d^2 v c_{ii'}}{\sum_{i < i'} D_{ii'}^2}$$

em que:

$d^2 v c_{ii'}$ = quadrado da distância euclidiana estimada a partir dos escores de n variáveis canônicas; e

$D_{ii'}^2$ = distância generalizada de Mahalanobis estimada a partir de n variáveis originais.

Demonstra-se que α é também a porcentagem acumulada da variância explicada pelas v_1 variáveis canônicas.

Importância Relativa dos Caracteres na Divergência Genética

Identificam-se os caracteres de menor importância para divergência genética entre os acessos avaliados como sendo aqueles cujos coeficientes de ponderação, obtidos com a padronização das variáveis, são os de maior magnitude, em valor absoluto, nas últimas variáveis canônicas.

Assim, é fundamental que, ao se fazerem inferências sobre as características analisadas, os efeitos de escala de mensuração sejam eliminados. Os coeficientes a_j 's devem ser multiplicados pelo desvio-padrão do erro experimental ($\hat{\sigma}_j$), de modo que:

$$\theta_j x_j = a_j \hat{\sigma}_j \left(\frac{X_j}{\hat{\sigma}_j} \right)$$

logo:

$$\theta_j = a_j \hat{\sigma}_j$$

portanto, os valores θ_j medem a importância relativa de uma característica em cada variável canônica.

Se Y_v é a variável canônica de menor importância relativa, dada por $Y_v = \theta_{v1}x_1 + \theta_{v2}x_2 + \dots + \theta_{vn}x_n$, em que x_1, x_2, \dots, x_n são as variáveis originais padronizadas, então identifica-se o caráter de menor importância como aquele associado ao maior dos elementos $\theta_{v1}, \theta_{v2}, \dots, \theta_{vn}$. A segunda variável de menor importância é identificada, com o mesmo critério, pelos coeficientes da variável canônica Y_{v-1} , e assim sucessivamente.

Quando em uma variável canônica de menor variância o maior coeficiente de ponderação está associado a um caráter já previamente descartado, é recomendado não fazer nenhum outro descarte com base nos coeficientes daquela variável canônica, mas prosseguir a identificação da importância relativa dos caracteres na outra variável de variância imediatamente superior.

Ilustração

Será considerada, como ilustração, a obtenção dos componentes principais a partir das médias padronizadas de 15 genótipos de milho-pipoca, avaliados em relação aos caracteres altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela). Para este caso, a matriz E, de variâncias e covariâncias residuais, é dada por:

$$E = \begin{bmatrix} 0,0440 & -0,0018 & 0,0271 & 0,0029 & -0,0258 & -0,0011 & 0,0004 & 0,0916 & 25,0479 \\ & 0,0193 & 0,0001 & -0,0033 & -0,0391 & -0,0004 & 0,0001 & -0,0713 & 13,3990 \\ & & 7,5072 & 0,0694 & 0,6516 & 0,0243 & -0,0034 & 1,8667 & -235,4233 \\ & & & 0,0089 & -0,0440 & 0,0001 & 0,0001 & 0,0242 & -6,3108 \\ & & & & 7,7418 & 0,0081 & -0,0030 & -0,3478 & -196,3793 \\ & & & & & 0,0007 & -0,0001 & -0,0003 & -0,73331 \\ & & & & & & 0,0002 & 0,0067 & 1,2645 \\ & & & & & & & 8,8635 & -265,0569 \\ & & & & & & & & 1491335746 \end{bmatrix}$$

e a matriz T, de variâncias e covariâncias fenotípicas entre médias dos genótipos, é fornecida por:

$$T = \begin{bmatrix} 0,3392 & 0,1018 & 1,6073 & 0,0702 & -2,8353 & 0,03925 & 0,0044 & 1,3388 & 283,7651 \\ & 0,0556 & 0,8029 & 0,0219 & -1,2895 & 0,0154 & 0,0029 & 0,5312 & 100,2161 \\ & & 40,9667 & 1,1947 & -43,4981 & 0,3148 & 0,0924 & -21,3900 & -730,1541 \\ & & & 0,0807 & -1,9411 & 0,0118 & 0,0056 & -1,1344 & -12,4416 \\ & & & & 83,3684 & -0,6691 & -0,2397 & 15,3778 & -174,20661 \\ & & & & & 0,0088 & 0,0020 & 0,2317 & 42,3104 \\ & & & & & & 0,0012 & -0,0044 & 12,8376 \\ & & & & & & & 67,9410 & 585,16164 \\ & & & & & & & & 822,6616049 \end{bmatrix}$$

As variâncias e os coeficientes de ponderação são obtidos pelos autovalores e autovetores, respectivamente, de $E^{-1}T$ ou pelas matrizes associadas às variáveis obtidas por condensação pivotal, por meio de $Z' = VX$, sendo:

$$V = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0,0402 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0,6192 & -0,0594 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0,0531 & 0,1665 & -0,0091 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0,3398 & 3,3060 & -0,1554 & 7,2907 & 1 & 0 & 0 & 0 & 0 \\ 0,02811 & 0,0226 & -0,0034 & 0,0092 & -0,0005 & 1 & 0 & 0 & 0 \\ -0,0079 & -0,0074 & 0,0004 & -0,0102 & 0,0002 & 0,0446 & 1 & 0 & 0 \\ -1,2785 & 4,52431 & -0,3146 & 2,4054 & 0,08159 & 6,8581 & -38,5417 & 1 & 0 \\ -690,4023 & -465,5900 & 20,2429 & 700,3220 & 23,9897 & -1803,1059 & -6044,3725 & 32,5299 & 1 \end{bmatrix}$$

Após a transformação, padronizam-se as variáveis considerando as variâncias:

$$\text{Var}(Z_1) = 0,0440$$

$$\text{Var}(Z_4) = 0,0076$$

$$\text{Var}(Z_7) = 0,0002$$

$$\text{Var}(Z_2) = 0,0192$$

$$\text{Var}(Z_5) = 7,1815$$

$$\text{Var}(Z_8) = 7,6043$$

$$\text{Var}(Z_3) = 7,4904$$

$$\text{Var}(Z_6) = 0,0006$$

$$\text{Var}(Z_9) = 96762,1306$$

Assim, a matriz de variâncias e covariâncias entre médias de genótipos, utilizando as variáveis transformadas, é dada por:

$$T^* = \begin{bmatrix} 7,7092 & 3,9659 & 2,4234 & 2,9965 & -3,7736 & 9,1862 & 0,7561 & 1,5517 & -0,5806 \\ 3,3430 & 2,0866 & 1,7181 & -3,0144 & 5,7040 & 1,2536 & 0,9868 & -0,5776 & \\ 5,2091 & 3,5207 & -4,9874 & 3,5655 & 2,2696 & -4,6847 & -3,9800 & \\ 8,2287 & -4,9730 & 4,5867 & 3,6638 & -5,3338 & -3,7100 & \\ 8,6323 & -8,6056 & -5,3064 & 3,8967 & 3,8733 & \\ 17,7568 & 5,4821 & 3,3491 & -0,8858 & \\ 6,4269 & -1,6509 & -0,4726 & \\ 11,7172 & 7,7506 & \\ 7,6568 & \end{bmatrix}$$

Na Tabela 2.18 são apresentados os autovalores e os autovetores, cujos elementos são os coeficientes de ponderação das características avaliadas. Verifica-se que as duas primeiras variáveis canônicas explicam próximo de 80% da variação disponível nos dados, sendo satisfatória a análise da diversidade genética por meio da dispersão gráfica, em relação a eixos representados por

essas duas variáveis canônicas. Cada variável canônica é uma combinação linear de caracteres, e os pesos atribuídos são descritos na Tabela 2.18.

Na Tabela 2.19 Encontram-se os coeficientes associados às variáveis analisadas, que expressam as suas importâncias relativas no estudo da diversidade genética. Variáveis com pequena variabilidade ou que estão correlacionadas com outras consideradas no estudo apresentarão coeficientes de grande magnitude nos últimos autovetores. Assim, considerando as últimas variáveis canônicas, que representam menos de 10% da variação total, constata-se que as características em que seria recomendado o descarte, em estudos futuros, seriam PROD, DOEN, CE, AE e NE.

Tabela 2.18 - Estimativas de autovalores e coeficientes de ponderação das características altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela), avaliados em cultivares de milho-pipoca

λ_j	λ_j (%) acumulada	AP	AE	NE	PROL	CE	QUE	DOEN	PCG	PROD
37,579	49,01	2,306	2,064	0,241	2,028	-0,192	25,143	25,915	-0,080	-0,001
23,254	79,33	0,690	1,593	-0,165	-0,609	0,076	16,881	-18,756	0,287	0,001
6,891	88,32	-3,715	-2,314	-0,001	0,824	-0,043	0,071	46,110	0,056	0,001
4,109	93,68	0,060	1,164	-0,153	10,897	0,145	-7,142	-6,976	0,021	0,001
2,331	97,72	-0,596	1,051	0,330	-0,289	0,006	-20,182	-4,353	0,045	0,002
1,028	98,06	1,905	-6,142	0,041	-3,247	-0,062	-1,813	-0,237	-0,107	0,001
0,696	98,97	0,078	-1,713	0,135	1,123	0,239	12,694	16,671	-0,010	-0,001
0,635	99,80	1,910	2,562	-0,067	-3,198	,1300	-1,711	45,249	-0,118	-0,001
0,154	100,00	1,672	-1,165	0,006	-0,614	-0,055	-11,772	24,193	0,149	-0,001

Tabela 2.19 - Estimativas dos coeficientes que expressam a importância relativa dos caracteres altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela), para estudo da diversidade entre cultivares de milho-pipoca

AP	AE	NE	PROL	CE	QUE	DOEN	PCG	PROD
0,484	0,287	0,066	0,191	-0,535	0,684	0,356	-0,237	-0,257
0,145	0,221	-0,451	-0,057	0,211	0,459	-0,258	0,856	0,521
-0,779	-0,322	-0,001	0,078	-0,120	0,002	0,634	0,166	0,440
0,012	0,162	-0,419	1,027	0,404	-0,194	-0,096	0,063	0,351
-0,125	0,146	0,905	-0,027	0,017	-0,549	-0,060	0,135	0,625
0,400	-0,854	0,113	-0,306	-0,172	-0,049	-0,003	-0,320	0,394
0,016	-0,238	0,369	0,106	0,665	0,345	0,229	-0,031	-0,041
0,401	0,356	-0,184	-0,301	0,361	-0,046	0,622	-0,351	-0,193
0,351	-0,162	0,017	-0,058	-0,153	-0,320	0,333	0,444	-0,549

Na Figura 2.19 é apresentada a dispersão dos 15 cultivares em relação aos eixos estabelecidos pelas duas primeiras variáveis canônicas. Novamente é

percebida a formação de três grupos de similaridade, tal como já ressaltado pelos demais métodos de agrupamento ou de dispersão gráfica apresentados neste capítulo.

Escores	
VC1	VC2
15,91	10,70
15,82	14,10
15,45	9,69
16,39	10,21
20,27	13,61
4,29	7,16
3,64	7,91
3,60	6,92
2,89	7,23
5,57	6,65
5,08	18,28
7,45	18,59
2,93	19,37
6,15	17,57
4,72	16,95
$V(VC1) = 37,58$	
$V(VC2) = 23,25$	
$Cov(VC1, VC2) = 0$	

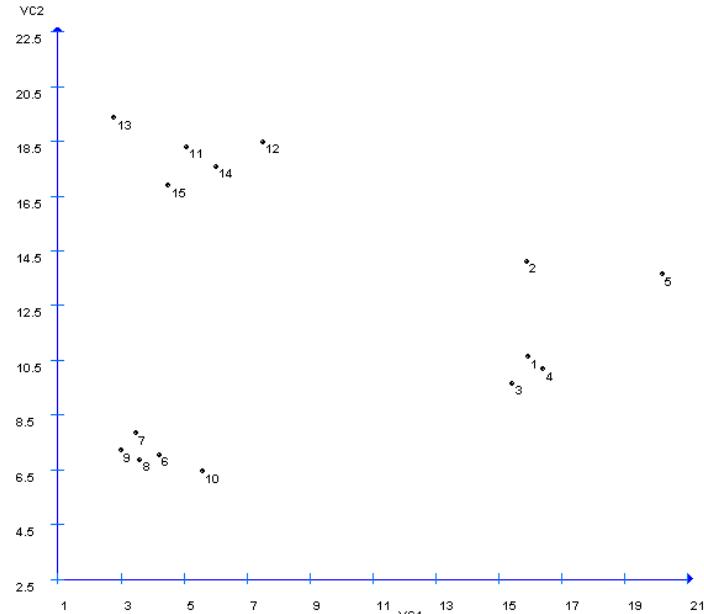
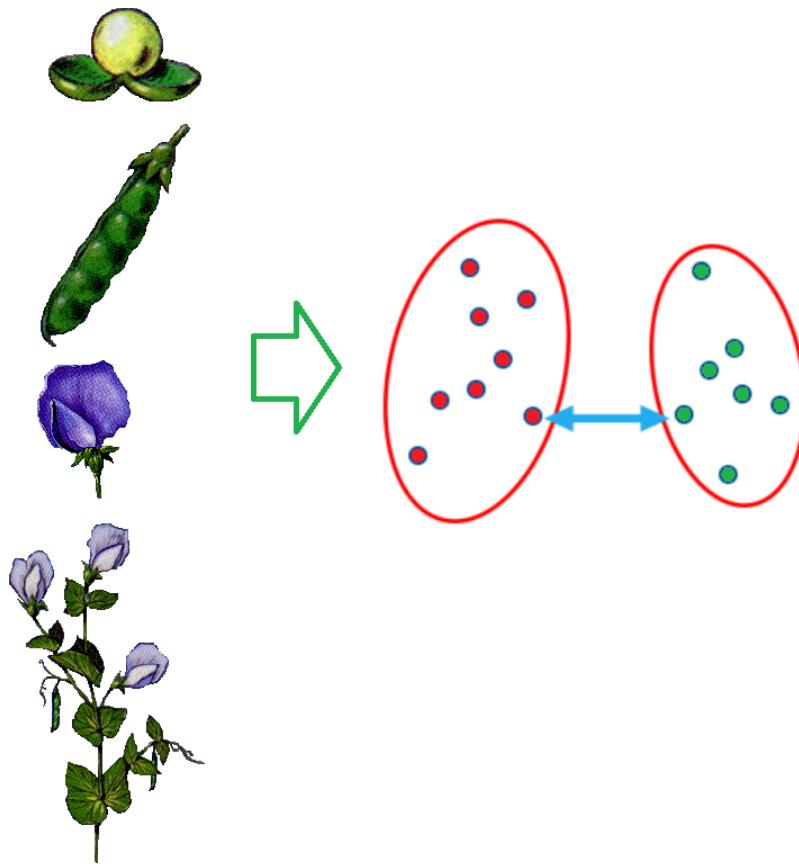


Figura 2.19 - Dispersão gráfica de 15 cultivares de milho, em relação às duas primeiras variáveis canônicas, estabelecidas pela combinação linear de nove características agronômicas.

Capítulo 3

Discriminação de Populações Baseada em Informações Fenotípicas



3.1 Introdução

Nas análises de discriminação, ou de classificação, procura-se obter funções ou algoritmos que permitam classificar um indivíduo, a partir das informações de um conjunto de características mensuradas, em uma dentre várias populações conhecidas, buscando minimizar a probabilidade de má classificação. Assim, deve-se obter um critério, expresso por funções ou algoritmo, que permitam alocar um indivíduo na população à qual ele realmente pertence.

Nessas análises, os estudos são realizados preliminarmente, a partir das informações de populações previamente conhecidas. Constatada a eficácia da discriminação, as funções podem ser utilizadas para alocar novos indivíduos, dos quais se desconhece a origem.

Exemplos de estudos de classificação podem ser encontrados com funções discriminantes em diversas áreas da ciência. No melhoramento genético, cita-se como ilustração o estudo realizado por Ferreira et al. (1995), em que o objetivo era o de discriminar cultivares de arroz tolerantes à toxidez de alumínio, em condições de cultivo hidropônico. Para estabelecimento das funções que permitissem a discriminação entre cultivares tolerantes e sensíveis, foram utilizados genótipos cujo comportamento, em nível de cultivo de campo, já era conhecido. Dessa forma, genótipos sabidamente tolerantes foram reunidos para formar a população T, e os sabidamente sensíveis formaram a população S. Constatou-se ser possível, em condições de cultivo hidropônico, distinguir indivíduos tolerantes e sensíveis, considerando os caracteres comprimento de raiz, peso da matéria seca da raiz, da parte aérea e total, e altura da planta, após dez dias de cultivo em solução nutritiva. Outro estudo, baseado em função discriminante, foi realizado por Pereira (1999), visando diferenciar cultivares de arroz considerados de padrão moderno daqueles tradicionais.

A classificação pode ser estabelecida a partir de critérios subjetivos. Assim, por exemplo, sabe-se que o número de dias para florescimento e

maturidade são caracteres que refletem a precocidade em genótipos de soja. Estes são classificados em grupos de maturidade, sendo o ciclo precoce de até 100 dias, semiprecoce entre 101 a 110 dias, médio de 111 a 125 dias, semitardio entre 125 e 145 dias, e tardio maior que 145 dias (NEPOMUCENO et al., 2008). Não obstante, cabe destacar que a classificação de cultivares de soja quanto ao ciclo, deve-se levar em consideração a faixa de latitude, em razão da sensibilidade de genótipos de soja ao fotoperíodo.

3.2. Análises discriminantes baseada em componentes principais

A análise discriminante por componentes principais visa estabelecer funções que permitam discriminar um conjunto de genótipos distribuídos em grupos (ou populações) previamente conhecidos, denominadas funções discriminantes. O número de funções discriminantes necessário para caracterizar as diferenças entre os grupos é igual ao mínimo entre ($p - 1$, v), em que p representa o número de grupos (ou populações) e v o número de caracteres. As funções discriminantes são obtidas em ordem decrescente de importância. A primeira função explica o máximo da variância entre os grupos, e a segunda, ortogonal à primeira, o máximo da variância remanescente, e assim sucessivamente, até se esgotar a variância na matriz que representa a variação entre os centróides das populações.

Para esta análise, consideram-se as informações de um grupo de indivíduos pertencentes a diferentes populações. A partir da matriz de variâncias e covariâncias entre populações são obtidos componentes principais, independentes entre si e capazes de reterem, em ordem de estimativação, o máximo da variância originalmente existente entre as populações. Obtidas as funções que melhor discriminam as populações, são estimados os escores a partir dos valores dos indivíduos que pertencem a cada população, de forma que se possa, por meio de uma análise de dispersão gráfica, avaliar as dissimilaridades entre estes indivíduos e as populações a que pertencem.

Para análise discriminante com base na técnica dos componentes principais, em que se ignora o delineamento experimental ou não há condições de prover estimativas da matriz de dispersão residual, considera-se o modelo:

$$Y_{ijk} = \mu_k + P_{ik} + I_{ijk}$$

em que:

Y_{ijk} : valor, para uma variável k , do j -ésimo indivíduo dentro do i -ésima população, ou grupo; ($k=1,2,\dots,v$);

μ_k : média geral da k -ésima variável;

P_{ik} : efeito da i -ésima população ($i = 1, 2, \dots, p$), considerando a variável k ; e

I_{ijk} : efeito do j -ésimo indivíduo dentro do i -ésima população ($j=1,2,\dots,n_i$), considerando a variável k .

A partir dessas informações são obtidas as médias dentro de cada grupo, dadas por:

$$X_{ik} = \frac{\sum_{j=1}^{n_i} Y_{ijk}}{n_i}$$

em que n_i é o número de indivíduos dentro da população i .

Os valores a serem utilizado em análises gráficas são resultantes da padronização, feita por meio de:

$$x_{ik} = \frac{X_{ik}}{\hat{\sigma}_{xk}}$$

Obtém-se a matriz de variâncias e covariâncias (B), com base nas médias destes grupos, ou, quando os dados são padronizados, a matriz de correlação (R). A partir da matriz R (ou B) obtêm-se os autovalores (λ_k) e os autovetores correspondentes (α_k), como descrito a seguir:

$$\det(R - I\lambda_k) = 0$$

e

$$(R - I\lambda_k)\alpha_k = \phi$$

O número de autovalores não-nulos da matriz R é dado por:

$$\eta = \min\{v, f\}$$

sendo v o número de variáveis consideradas na análise e f ($f = p-1$) os graus de liberdade associados às variâncias e covariâncias que deram origem a esta matriz.

Nesta análise, a importância relativa (IR) de cada componente é calculada por meio da expressão:

$$IR_k = \frac{\lambda_k}{\text{Traço}(R)}$$

São obtidos os escores considerando as médias dos grupos como referência e também os valores dos acessos. Admitindo que existam η componentes principais, considera-se que cada um seja estabelecido por:

$$CP_1 = a_{11}x_1 + a_{21}x_2 + \dots + a_{v1}x_v$$

...

$$CP_\eta = a_{1\eta}x_1 + a_{2\eta}x_2 + \dots + a_{v\eta}x_v$$

Assim, utilizando o primeiro componente principal como exemplo, serão obtidos os escores a partir da expressão:

$$CP_{1i} = a_{11}x_{i1} + a_{21}x_{i2} + \dots + a_{v1}x_{iv}$$

em que x_{ik} representa a média padronizada da população i , para a variável k .

Neste caso, verifica-se que:

$$V(CP_\ell) = \lambda_\ell, \text{ para } \ell = 1, 2, \dots, \eta.$$

Também são calculados os escores para cada indivíduo. Utilizando o primeiro componente principal como exemplo, estes escores são obtidos por:

$$CP_{1ij} = a_{11}y_{ij1} + a_{21}y_{ij2} + \dots + a_{v1}y_{ijv}$$

Os valores de y são obtidos por meio de:

$$y_{ijk} = \frac{Y_{ijk}}{\hat{\sigma}_{xk}}$$

Verifica-se, portanto, que não se trata de uma padronização desses valores, mas uma mudança de escala, tornando-os apropriados para a dispersão gráfica. Nesta análise também deve ser identificada a ordem de variáveis

associadas aos maiores elementos identificados no último até o primeiro autovetor. As variáveis de maiores pesos nos últimos autovetores são consideradas de menor importância para estudo da diversidade genética entre as populações. Normalmente, consideram-se os últimos autovetores associados a autovalores da matriz de correlação inferior a 0,7. As variáveis de maiores pesos nos primeiros autovetores são consideradas de maior importância para o estudo da diversidade genética quando o autovalor explica uma fração considerável da variação disponível, normalmente limitado em valor mínimo de 80%.

O conjunto de cargas totais associado às variáveis é obtido considerando que:

$$C_j = \alpha R$$

em que C_j é o vetor linha de ordem j , cujos elementos, denotados C_{kj} , são utilizados para obtenção da estimativa da carga da variável k no j -ésimo componente principal, por meio de:

$$\text{Carga} = \frac{C_{kj}}{\sqrt{\lambda_j}}$$

Ilustração

Será considerada, como ilustração, a obtenção de funções discriminantes, baseadas em componentes principais, a partir das médias padronizadas de 15 genótipos de milho-pipoca, avaliados em relação aos caracteres altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela). Considerase, neste caso, o estudo de três populações, nas quais a população 1 é representada pelos cultivares de 1 a 5; a população 2, pelos cultivares de 6 a 10; e a população 3, pelos cultivares de 11 a 15, conforme especificado na Tabela 3.1.

As médias das populações estão representadas na Tabela 3.2. São também apresentados os valores padronizados a serem utilizados na obtenção das funções discriminantes.

A partir dos valores padronizados, é obtida a matriz de correlação mostrada a seguir:

$$R = \begin{bmatrix} 1 & 0,9957 & 0,5324 & 0,5879 & -0,8244 & 0,9982 & 0,8966 & 0,3660 & 0,7823 \\ 0,9957 & 1 & 0,6084 & 0,6602 & -0,8732 & 0,9995 & 0,9337 & 0,2783 & 0,7213 \\ 0,5324 & 0,6084 & 1 & 0,9978 & -0,9180 & 0,5819 & 0,8522 & -0,5929 & -0,1108 \\ 0,5879 & 0,6602 & 0,9978 & 1 & -0,9425 & 0,6351 & 0,8853 & -0,5377 & -0,0440 \\ -0,8244 & -0,8732 & -0,9180 & -0,9425 & 1 & -0,8567 & -0,9898 & 0,2251 & -0,2923 \\ 0,9982 & 0,9995 & 0,5819 & 0,6351 & -0,8567 & 1 & 0,9214 & 0,3099 & 0,7438 \\ 0,8966 & 0,9337 & 0,8522 & 0,8853 & -0,9898 & 0,9214 & 1 & -0,0839 & 0,4256 \\ 0,3660 & 0,2783 & -0,5929 & -0,5377 & 0,2251 & 0,3099 & -0,0839 & 1 & 0,8660 \\ 0,7823 & 0,7213 & -0,1108 & -0,0440 & -0,2923 & 0,7438 & 0,4256 & 0,8660 & 1 \end{bmatrix}$$

Para a análise considerada, têm-se nove características ($v = 9$), porém os graus de liberdade associados às variâncias é $p - 1 = 2$ (p é o número de populações avaliadas, que, no exemplo, é igual a 3). Assim, são estimadas duas funções discriminantes, cujas

Tabela 3.1 - Média de 15 genótipos de milho-pipoca, pertencentes a três populações, para nove características agronômicas

Populações	AP	AE	NE	PROL	CE	QUE	DOEN	PCG	PROD
1	1,82	1,20	33,84	1,06	14,17	0,32	0,19	22,01	3339,00
1	2,90	1,43	24,46	1,06	14,88	0,27	0,15	21,86	3950,33
1	2,52	1,16	29,81	1,60	13,63	0,23	0,15	17,87	3512,67
1	2,02	0,97	24,94	1,08	11,96	0,33	0,16	16,40	2845,33
1	3,48	1,53	39,37	1,18	13,71	0,38	0,10	21,93	3022,67
2	1,39	0,83	23,94	0,73	25,88	0,11	0,08	14,08	2142,00
2	1,62	0,67	20,64	0,94	35,66	0,12	0,09	14,63	1806,33
2	1,70	0,86	27,32	0,70	31,26	0,09	0,08	14,73	1981,33
2	1,35	0,87	20,51	0,90	34,62	0,09	0,09	14,03	1734,67
2	1,60	1,11	24,74	0,84	28,68	0,10	0,10	13,60	1818,33
3	1,71	1,05	19,87	0,62	28,84	0,20	0,13	36,17	4105,67
3	2,04	1,25	18,84	0,63	28,92	0,26	0,08	33,63	2901,33
3	2,25	1,08	15,76	0,60	37,75	0,17	0,09	31,77	4343,67
3	2,39	0,81	16,80	0,72	30,65	0,21	0,10	31,38	3565,33
3	2,08	1,09	21,27	0,53	31,29	0,18	0,10	31,00	3811,67
DP	0,5824	0,2359	6,4005	0,2841	9,1306	0,0937	0,0345	8,2426	906,9794

Tabela 3.2 - Médias originais e padronizadas de populações de milho-pipoca, para nove características agronômicas

Populações	AP	AE	NE	PROL	CE	QUE	DOEN	PCG	PROD
1	2,547	1,260	30,483	1,196	13,67	0,306	0,151	20,011	3334,429
2	1,531	0,868	23,426	0,822	31,22	0,101	0,087	14,215	1897,126
3	2,096	1,055	18,507	0,620	31,49	0,204	0,096	32,791	3746,213
DP	0,5091	0,1959	6,0197	0,2922	0,2113	0,1025	0,0346	9,5042	970,7839
Médias padronizadas									
1	5,001	6,423	5,064	4,089	1,339	2,98	4,392	2,105	3,435
2	3,005	4,423	3,891	2,811	3,057	,98	2,545	1,496	1,954
3	4,115	5,375	3,074	2,118	3,084	1,989	2,804	3,45	3,859

variâncias são representadas pelos autovalores de R, e os coeficientes de ponderação, representados pelos elementos dos autovetores associados a estes autovalores, mostrados na Tabela 3.3. Neste exemplo, em que foram

consideradas apenas 3 populações, verifica-se, como esperado, que os dois primeiros componentes explicam 100% da variação disponível. Também constata-se que, à exceção de PCG no primeiro componente, todas as variáveis têm peso com magnitude relativamente próxima, e alguns apresentam sinais contrários. A importância de PCG é maior no segundo componente, indicando que, ao utilizar os dois componentes, todas as variáveis terão pesos consideráveis na discriminação das populações em estudo. A análise das cargas fatoriais também permite as mesmas conclusões obtidas pelos elementos dos autovetores.

Na Figura 3.1 está representada a dispersão dos centróides das populações estudadas. Estes centróides são estabelecidos pelos escores dos dois primeiros componentes principais. Percebe-se que eles se encontram bem dispersos, indicando ser possível discriminá-las a partir das nove características agronômicas mensuradas.

Na Figura 3.2 é apresentada a dispersão dos genótipos em torno dos centróides, cujas coordenadas foram estabelecidas pelas funções discriminantes, obtidas com base em componentes principais, e estimadas com o objetivo de maximizar as diferenças entre as populações analisadas. As populações 2 e 3 mostram ser mais homogêneas, pois seus cultivares estão bem próximos aos seus respectivos centróides. A população 1 é mais heterogênea, porém mantém características que a tornam possível de ser distinguida das outras duas consideradas na análise.

Tabela 3.3.- Estimativas de autovalores e coeficientes de ponderação das características altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela), avaliados em cultivares de milho-pipoca

λ_j	λ_j (%) acumulada	AP	AE	NE	PROL	CE	QUE	DOEN	PCG	PROD
Autovetores										
6,247	69,41	0,3764	0,3874	0,3151	0,3309	-0,3870	0,3838	0,3975	0,0117	0,2101
2,753	100,00	0,2041	0,1508	-0,3714	-0,3387	0,1527	0,1700	-0,0681	0,6025	0,5129
Cargas fatoriais										
-	-	0,9409	0,9682	0,7876	0,8271	-0,9674	0,9594	0,9936	0,0292	0,5251
-	-	0,3387	0,2502	-0,6162	-0,562	0,2534	0,2820	-0,1129	0,9996	0,8510

Escores	
C1	C2
10,44	2,16
5,63	1,25
6,84	4,47
V(C1)=6,24	V(C2)=2,75
Cov(C1,C2) =0	

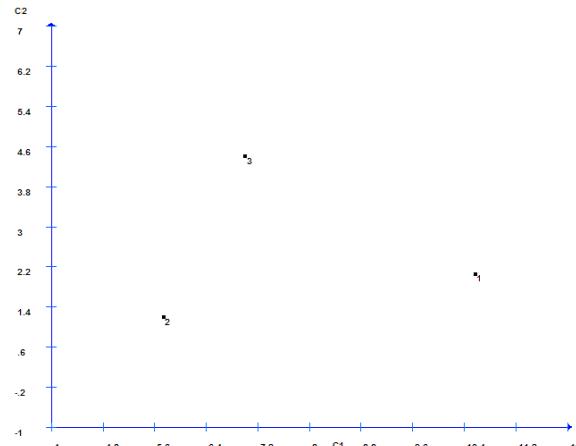


Figura 3.1- Dispersão gráfica dos centróides de três populações de milho-pipoca, em relação a funções discriminantes, estimadas com base em componentes principais, estabelecidos pela combinação linear entre nove características agronômicas.

Escores	
C1	C2
10,34	1,86
10,55	3,37
10,41	1,48
9,20	1,71
11,69	2,40
5,58	1,31
5,17	1,22
5,65	1,30
5,27	1,21
6,49	1,20
7,14	4,53
7,15	4,25
6,43	5,06
6,63	4,20
6,83	4,29

$$(C1)=4,79 \quad V(C2)=2,16$$

$$\text{Cov}(C1,C2) = 0$$

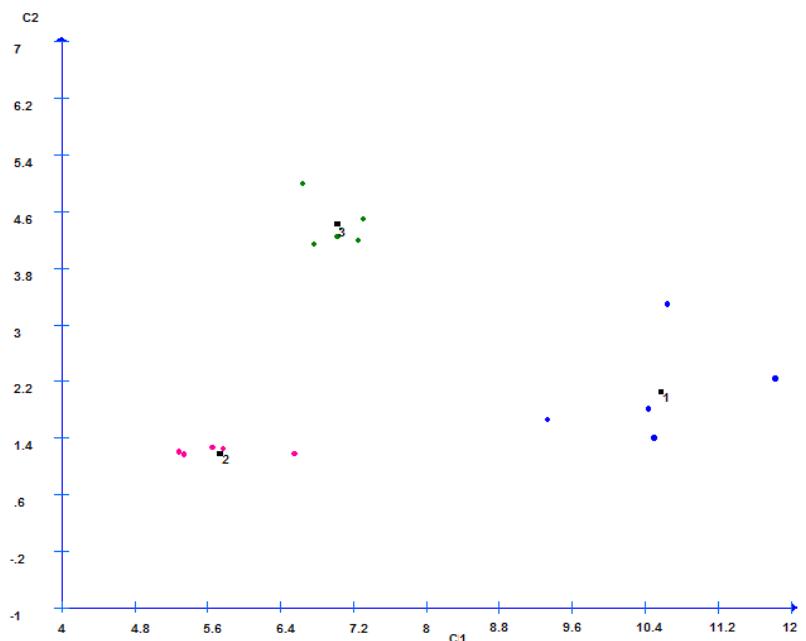


Figura 3.2 - Dispersão gráfica de genótipos em torno dos centróides de três populações de milho-pipoca, em relação a funções discriminantes, estimadas com base em componentes principais, estabelecidos pela combinação linear entre nove características agronômicas.

3.3 Análise discriminante linear de Fisher

3.2.1 Abordagem univariada

Inicialmente, será admitido o caso mais simples em que se dispõe apenas das informações relativas a uma determinada característica x mensurada em indivíduos pertencentes a duas populações distintas. Será suposto que a variável avaliada na população 1, denotada por π_1 , tem distribuição normal univariada com média μ_1 e variância σ_1^2 . Assim, para a

população π_1 , tem-se $x \sim N(\mu_1, \sigma_1^2)$. Por outro lado, é admitido que a variável avaliada na população 2, denotada por π_2 , tem distribuição normal univariada com média μ_2 e variância σ_2^2 . Assim, para a população π_2 tem-se $x \sim N(\mu_2, \sigma_2^2)$. Tal situação é ilustrada por Mardia et al. (1979), usando a Figura 3.3.

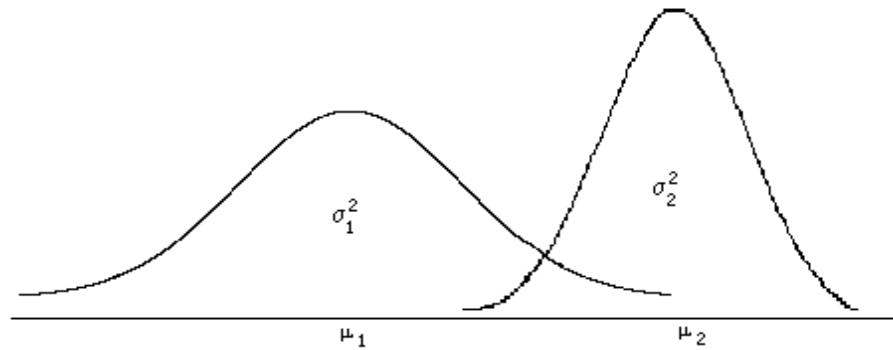


Figura 3.3 - Curvas representativas de duas populações com médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2 .

Na Figura 3.3, observa-se que estão sendo consideradas duas populações em que há diferenças entre as médias ($\mu_1 < \mu_2$) e entre as variâncias ($\sigma_1^2 > \sigma_2^2$). Procura-se estabelecer funções discriminantes de máxima verossimilhança que sejam capazes de alocar convenientemente um indivíduo, com valor de observação x , em uma das duas populações consideradas (π_1 ou π_2).

Se a variável tem distribuição normal, então a função densidade de probabilidade é dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

A expressão de máxima verossimilhança que expressa a probabilidade de x pertencer a cada uma das duas populações é dada por:

$$f_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}$$

e

$$f_2(x) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}$$

Assim, x é alocado em π_1 ao invés de π_2 quando:

$$\frac{f_1(x)}{f_2(x)} > 1$$

que equivale a:

$$\frac{\sigma_2}{\sigma_1} e^{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]} > 1$$

Aplicando logaritmo, tem-se:

$$\ln\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right] > 0$$

ou

$$\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2 < 2\ln\left(\frac{\sigma_2}{\sigma_1}\right)$$

Uma situação interessante se verifica quando $\sigma_1^2 = \sigma_2^2$, de forma que, para se ter $\frac{f_1(x)}{f_2(x)} > 1$, é necessário:

$$(x - \mu_1)^2 - (x - \mu_2)^2 < 0$$

$$-2x\mu_1 + \mu_1^2 + 2x\mu_2 - \mu_2^2 < 0$$

sendo $\mu_2 > \mu_1$, é conveniente adotar:

$$2x(\mu_2 - \mu_1) < \mu_2^2 - \mu_1^2$$

ou ainda:

$$x < \frac{\mu_2^2 - \mu_1^2}{2(\mu_2 - \mu_1)}$$

de onde se conclui que:

$$\frac{f_1(x)}{f_2(x)} > 1 \text{ quando } x < \frac{\mu_2 + \mu_1}{2}$$

Assim, se $\mu_1 < \mu_2$, a função discriminante de máxima verossimilhança aloca x em π_1 se $x < \frac{\mu_2 + \mu_1}{2}$, e em π_2 em caso contrário.

3.2.2 Abordagem multivariada

Será considerado agora que se dispõe de diversas variáveis para caracterizar as populações (π_i , com $i = 1, 2, \dots, g$), as quais apresentam matriz de variâncias e covariâncias homogêneas, denotada por Σ . Admite-se, portanto, que em cada população o vetor de observações \tilde{x} tem distribuição $N_v(\mu_i, \Sigma)$. Assim, \tilde{x} , em uma determinada população π_i , apresenta a seguinte função densidade de probabilidade:

$$f_i(\tilde{x}) = |2\pi\Sigma|^{-1/2} e^{-\frac{1}{2}[(\tilde{x} - \mu_i)\Sigma^{-1}(\tilde{x} - \mu_i)]}$$

Tomando um par de populações π_1 e π_2 , aloca-se um indivíduo, com vetor de observações \tilde{x} , em π_1 se:

$$\frac{f_1(\tilde{x})}{f_2(\tilde{x})} > 1$$

que, por analogia ao descrito anteriormente, significa que:

$$(\tilde{x} - \mu_1)' \Sigma^{-1} (\tilde{x} - \mu_1) < (\tilde{x} - \mu_2)' \Sigma^{-1} (\tilde{x} - \mu_2)$$

ou

$$D_1^2 < D_2^2$$

sendo D_i^2 a distância generalizada de Mahalanobis, calculada tomando as informações do indivíduo a ser classificado e a média da i -ésima população. Assim, a análise discriminante consiste, portanto, em alocar um indivíduo com um conjunto de observações \tilde{x} numa população π_i , adotando-se o critério de que a distância de Mahalanobis seja mínima.

Será considerada novamente a inequação:

$$(\tilde{x} - \mu_1)' \Sigma^{-1} (\tilde{x} - \mu_1) - (\tilde{x} - \mu_2)' \Sigma^{-1} (\tilde{x} - \mu_2) < 0$$

Expandindo-a, tem-se:

$$(\tilde{x}' \Sigma^{-1} \tilde{x} - \tilde{x}' \Sigma^{-1} \mu_1 - \mu_1' \Sigma^{-1} \tilde{x} + \mu_1' \Sigma^{-1} \mu_1) - (\tilde{x}' \Sigma^{-1} \tilde{x} - \tilde{x}' \Sigma^{-1} \mu_2 - \mu_2' \Sigma^{-1} \tilde{x} + \mu_2' \Sigma^{-1} \mu_2) < 0$$

tendo-se, para os escalares, a seguinte igualdade:

$$\tilde{x}' \Sigma^{-1} \mu_1 = \mu_1' \Sigma^{-1} \tilde{x}$$

$$\tilde{x}' \Sigma^{-1} \mu_2 = \mu_2' \Sigma^{-1} \tilde{x}$$

logo:

$$-2\tilde{x}' \Sigma^{-1} \mu_1 + \mu_1' \Sigma^{-1} \mu_1 + 2\tilde{x}' \Sigma^{-1} \mu_2 + \mu_2' \Sigma^{-1} \mu_2 < 0$$

Arranjando os elementos, somando e subtraindo os escalares $\mu_1' \Sigma^{-1} \mu_2 = \mu_2' \Sigma^{-1} \mu_1$ na expressão anterior, tem-se:

$$-2\tilde{x}' \Sigma^{-1} (\mu_1 - \mu_2) + (\mu_1' \Sigma^{-1} \mu_1 + \mu_1' \Sigma^{-1} \mu_2) - (\mu_2' \Sigma^{-1} \mu_2 + \mu_2' \Sigma^{-1} \mu_1) < 0$$

$$-2\tilde{x}'\Sigma^{-1}(\mu_1 - \mu_2) + \mu_1'\Sigma^{-1}(\mu_1 + \mu_2) - \mu_2'\Sigma^{-1}(\mu_1 + \mu_2) < 0$$

$$-2\tilde{x}'\Sigma^{-1}(\mu_1 - \mu_2) + (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) < 0$$

e, finalmente:

$$\tilde{x}'\Sigma^{-1}(\mu_1 - \mu_2) - (\mu_1 - \mu_2)'\Sigma^{-1}\frac{1}{2}(\mu_1 + \mu_2) > 0$$

Mardia et al. (1979) relatam que a decisão de alocar um indivíduo com vetor de observações \tilde{x} em π_1 ao invés de π_2 ocorre quando:

$$\alpha'(\tilde{x} - u) > 0$$

em que:

$$\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$$

e

$$u = \frac{1}{2}(\mu_1 + \mu_2)$$

Considerando duas populações com vetor de médias v-variado μ_i e $\mu_{i'}$ e matriz de variâncias e covariâncias comuns Σ , de ordem v, define-se a função discriminante linear de Fisher pela expressão:

$$D_{ii'}(\tilde{x}) = \alpha'\tilde{x} = (\mu_i - \mu_{i'})'\Sigma^{-1}\tilde{x}$$

Assim, a função discriminante $D_{ii'}(\tilde{x})$ é uma combinação linear do conjunto de caracteres que possibilita alocar um determinado indivíduo, com vetor de observações \tilde{x} , em uma população i, ou i' , com máxima probabilidade de acerto.

$$D_{ii'}(\tilde{x}) = \alpha'\tilde{x} = a_1x_1 + a_2x_2 + \dots + a_vx_v$$

Define-se também o ponto médio entre duas populações i e i' pelo valor m, expresso por:

$$m_{ii'} = \frac{1}{2}(\mu_i - \mu_{i'})'\Sigma^{-1}(\mu_i + \mu_{i'}) = \alpha'u = \frac{1}{2}(\alpha'\mu_1 + \alpha'\mu_2)$$

ou

$$m_{ii} = \frac{1}{2} [D(\mu_1) + D(\mu_2)]$$

d. Tomada de decisão

Com a função discriminante estimada, adota-se a regra de classificação:

- Aloca-se \tilde{x} em π_i se:

$$D_{ii}(\tilde{x}) = \alpha' \tilde{x} = (\mu_1 - \mu_2)' \Sigma^{-1} \tilde{x} \geq m_{ii}$$

- Aloca-se \tilde{x} em π_j se:

$$D_{jj}(\tilde{x}) = \alpha' \tilde{x} = (\mu_1 - \mu_2)' \Sigma^{-1} \tilde{x} < m_{jj}$$

Em termos práticos, é difícil dispor de informações paramétricas das populações em consideração de forma que o vetor de médias μ_i ($i = 1, 2, \dots, g$) e a matriz de variâncias e covariâncias comum, Σ , são desconhecidos. Para contornar esses problemas, utilizam-se os estimadores de μ_i e Σ , obtidos a partir de observações efetuadas em indivíduos que sabidamente pertencem às populações em consideração.

Assim, novamente admitindo que se dispõe de duas populações em que as variáveis mensuradas têm distribuição normal multivariada, pode-se, a partir de n_i ($i = 1, 2$) observações (indivíduos), definir:

$X_1 = [\tilde{x}_{11} \quad \tilde{x}_{21} \quad \dots \quad \tilde{x}_{p1}]$ (matriz $n_1 \times p$ de observações na população π_1);

$$X_1 = \begin{bmatrix} x_{11(1)} & x_{12(1)} & \dots & x_{1p(1)} \\ x_{21(1)} & x_{22(1)} & \dots & x_{2p(1)} \\ \dots & \dots & \dots & \dots \\ x_{n1,1(1)} & x_{n1,2(1)} & \dots & x_{n1,p(1)} \end{bmatrix}$$

e

$X_2 = [\tilde{x}_{12} \quad \tilde{x}_{22} \quad \dots \quad \tilde{x}_{p2}]$ (matriz $n_2 \times p$ de observações na população π_2)

$$X_2 = \begin{bmatrix} x_{11(2)} & x_{12(2)} & \dots & x_{1p(2)} \\ x_{21(2)} & x_{22(2)} & \dots & x_{2p(2)} \\ \dots & \dots & \dots & \dots \\ x_{n2,1(2)} & x_{n2,2(2)} & \dots & x_{n2,p(2)} \end{bmatrix}$$

em que $x_{ki(j)}$ é a observação no k-ésimo indivíduo, em relação à i-ésima variável, avaliado na j-ésima população.

A partir das matrizes de observações X_1 e X_2 estimam-se os vetores de médias amostrais dados por:

$$\hat{\mu}_j = [\bar{x}_{1j} \quad \bar{x}_{2j} \quad \dots \quad \bar{x}_{pj}]$$

sendo:

$$\bar{x}_{ij} = \frac{1}{n_j} \sum_{k=1}^{n_j} x_{kj}, \text{ para } i=1,2,\dots,p \text{ variáveis e } j=1,2 \text{ populações, nestas considerações.}$$

Também são estimadas as matrizes de variâncias e covariâncias amostrais. Para a população π_1 tem-se a matriz:

$$S_1 = \begin{bmatrix} \hat{\sigma}_{1(1)}^2 & \hat{\sigma}_{12(1)} & \dots & \hat{\sigma}_{1p(1)} \\ & \hat{\sigma}_{2(1)}^2 & & \hat{\sigma}_{2p(1)} \\ & & \dots & \\ \text{sim} & & & \hat{\sigma}_{p(1)}^2 \end{bmatrix}$$

Para a população π_2 tem-se:

$$S_2 = \begin{bmatrix} \hat{\sigma}_{1(2)}^2 & \hat{\sigma}_{12(2)} & \dots & \hat{\sigma}_{1p(2)} \\ & \hat{\sigma}_{2(2)}^2 & & \hat{\sigma}_{2p(2)} \\ & & \dots & \\ \text{sim} & & & \hat{\sigma}_{p(2)}^2 \end{bmatrix}$$

Na análise discriminante é admitido que as populações apresentam matriz de variâncias e covariâncias comum, denominada de Σ . Assim, apesar de serem

estimados S_1 e S_2 , utiliza-se a matriz de variâncias e covariâncias amostral comum, dada por:

$$S_c = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

Destaca-se o fato de que se deve ter $n_1 + n_2 - 2 > p$; caso contrário, S_c é singular e não apresenta inversa comum. Quando se dispõe de várias populações (π_j , com $j=1,2,\dots,g$), cada uma com n_j indivíduos a partir dos quais são estimadas as matrizes de variâncias e covariância S_j , obtém S_c por meio de:

$$S_c = \frac{\sum_{j=1}^g (n_j - 1)S_j}{\sum_{j=1}^g n_j - g}$$

A partir dos vetores e matrizes amostrais são estabelecidas as funções discriminantes entre duas populações 1 e 2:

$$D_{12}(\tilde{x}) = \hat{\alpha}' \tilde{x} = (\hat{\mu}_1 - \hat{\mu}_2)' S_c^{-1} \tilde{x}$$

E o ponto médio entre duas amostras é dado por:

$$m_{12} = \frac{1}{2} (\hat{\mu}_1 - \hat{\mu}_2)' S_c^{-1} (\hat{\mu}_1 + \hat{\mu}_2) = \hat{\alpha}' u = \frac{1}{2} (\hat{\alpha}' \hat{\mu}_1 + \hat{\alpha}' \hat{\mu}_2)$$

ou

$$m_{12} = \frac{1}{2} [D(\hat{\mu}_1) + D(\hat{\mu}_2)]$$

sendo $\hat{\alpha} = S_c^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$

A regra de decisão é análoga à apresentada anteriormente.

Ilustração

Será considerada, como ilustração, a obtenção das funções discriminantes a partir das médias de 15 genótipos de milho-pipoca, avaliados em relação aos caracteres altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela). Considera-se, nesse caso, a avaliação de três populações, nas quais a população 1 é representada pelos cultivares de 1 a 5; a população 2, pelos cultivares de 6 a 10; e a população 3, pelos cultivares de 11 a 15, conforme especificado na Tabela A1 (Anexo).

A matriz de variâncias e covariâncias entre as médias das populações (estas médias estão descritas na Tabela 2.7) é obtida como apresentado a seguir:

$$\omega = \begin{bmatrix} 0,3387 & 0,1018 & 1,5989 & 0,0700 & -2,8286 & 0,0393 & 0,0043 & 1,3493 & 285,1822 \\ & 0,0556 & 0,8139 & 0,0222 & -1,3066 & 0,0157 & 0,0031 & 0,5252 & 99,9740 \\ & & 40,9792 & 1,1936 & -43,5004 & 0,3120 & 0,1000 & -21,3904 & -729,4514 \\ & & & 0,0812 & -1,9462 & 0,0118 & 0,0061 & -1,1397 & -13,1292 \\ & & & & 83,3717 & -0,6714 & -0,2554 & 15,3992 & -174,13165 \\ & & & & & 0,0089 & 0,0021 & 0,2388 & 43,2991 \\ & & & & & & 0,0013 & -0,0169 & 11,8725 \\ & & & & & & & 67,9434 & 585,12827 \\ & & & & & & & & 822597,3411 \end{bmatrix}$$

As seguintes funções $D_{ii'}$ e pontos médios $m_{ii'}$ foram obtidos:

Caracteres	D ₁₂	D ₁₃	D ₂₃
AP	0,0888	0,2924	0,2036
AE	0,4046	1,2041	0,7995
NE	-0,0259	0,0293	0,0552
PROL	1,4999	0,7271	-0,7728
CE	-0,0227	0,0016	0,0243
QUE	13,5391	7,2532	-6,2859
DOEN	6,9166	15,6984	8,7818
PCG	-0,0346	-0,2369	-0,2023
PROD	0,0011	0,0004	-0,0007
Ponto médio (m _{ij'})	6,6859	2,3193	-4,3666

A classificação de qualquer indivíduo I, com vetor de observações \tilde{x} , poderá ser feita com apenas p-1 ($p=3$, neste exemplo) funções discriminantes. Assim, para o primeiro genótipo, com vetor de médias (Tabela 2.7) dado por:

$$\tilde{x} = [1,82 \quad 1,20 \quad 33,84 \quad 1,06 \quad 14,17 \quad 0,32 \quad 0,19 \quad 22,01 \quad 3339,00]$$

são obtidos os valores:

$D_{12} = 9,5969$. Este valor é superior ao ponto médio m_{12} indicando que a melhor alocação genótipo 1 é na população 1 do que na 2.

$D_{13} = 5,1871$. Este valor é superior ao ponto médio m_{13} indicando que a melhor alocação genótipo 1 é na população 1 do que na 3.

$D_{23} = -4,4098$. A comparação deste valor é desnecessária, uma vez que as comparações anteriores são suficientes para alocar o indivíduo 1 na população 1.

3.4 Análise discriminante de Anderson

Na análise discriminante proposta por Anderson (1958), consideram-se as informações de indivíduos sabidamente pertencentes a diferentes populações. A partir dessas informações são geradas funções, que são combinações lineares das características avaliadas e têm por finalidade promover a melhor discriminação entre os indivíduos, alocando-os em suas devidas populações. Estas funções, uma vez estimadas, passam a ser de grande utilidade por permitirem classificar novos materiais genéticos, de comportamento desconhecido, nas populações já conhecidas. A eficácia das variáveis utilizadas em promover a discriminação também é avaliada, permitindo conhecer a adequação da função estimada.

Para o estabelecimento da função discriminante de Anderson, novamente se considera que, para uma população π_j ($j=1,2\dots g$), o vetor de observações \tilde{x} tem distribuição $N_v(\mu_j, \Sigma)$, com a seguinte função densidade de probabilidade:

$$f_j(\tilde{x}) = 2\pi\Sigma^{-1/2} e^{-\frac{1}{2}[(\tilde{x}-\mu_j)'\Sigma^{-1}(\tilde{x}-\mu_j)]}$$

Anderson (1958) relata que é fundamental considerar na classificação de observações a probabilidade a priori inerente às várias populações avaliadas em um determinado estudo. Há casos em que a probabilidade de um determinado indivíduo, com vetor de observações \tilde{x} , pertencer a uma determinada população já é, de certa forma, conhecida pelo pesquisador, a partir de sua experiência e do conhecimento que se tem sobre as populações estudadas e do indivíduo a ser classificado.

Como ilustração, pode-se imaginar o estabelecimento de funções discriminantes entre duas populações π_1 e π_2 a partir de um conjunto de variáveis mensuráveis. Entretanto, pode-se supor que, além dos caracteres avaliados e da função discriminante estimada, o pesquisador tem conhecimento de que a população π_1 é típica de uma região de baixada, enquanto a π_2 é típica de uma região de encosta. Assim, se é desejado classificar um indivíduo desconhecido a

partir de um vetor de observações, mas se sabe que este indivíduo foi coletado numa região de baixada, é de se pressupor que a sua maior probabilidade é a de pertencer à população π_1 . Dessa forma, a probabilidade a priori deve ser computada, fazendo com que a discriminação seja feita de forma mais eficaz.

Será admitido que a probabilidade de uma observação pertencer a uma determinada população é p_j ($\sum_{j=1}^g p_j = 1$), conhecida a priori. Assim, pode-se estabelecer a função discriminante, dada pela probabilidade de \tilde{x} pertencer a π_j , por meio do logaritmo da função densidade de probabilidade de \tilde{x} , de forma que se tenha:

$$D_j(\tilde{x}) = -\frac{1}{2} [\ln(2\pi) + \ln|\Sigma_j|] - \frac{1}{2} [(\tilde{x} - \mu_j)' \Sigma_j^{-1} (\tilde{x} - \mu_j)] + \ln(p_j)$$

Supondo a homogeneidade das matrizes de variâncias e covariâncias ($\Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$) e retirando-se os componentes constantes da função $D_j(\tilde{x})$ que são dispensáveis na discriminação entre duas populações, tem-se:

$$D_j(\tilde{x}) = \ln(p_j) - \frac{1}{2} (\tilde{x} - \mu_j)' \Sigma^{-1} (\tilde{x} - \mu_j)$$

ou

$$D_j(\tilde{x}) = \ln(p_j) - \frac{1}{2} (\tilde{x}' \Sigma^{-1} \tilde{x} - \tilde{x}' \Sigma^{-1} \mu_j - \mu_j' \Sigma^{-1} \tilde{x} + \mu_j' \Sigma^{-1} \mu_j)$$

Sendo $\tilde{x}' \Sigma^{-1} \tilde{x}$ uma constante, ela também deve ser excluída da função, por ser dispensável na discriminação. Além disso, verifica-se na expressão anterior que:

$$\tilde{x}' \Sigma^{-1} \mu_j = \mu_j' \Sigma^{-1} \tilde{x}$$

logo:

$$D_j(\tilde{x}) = \ln(p_j) - \frac{1}{2} (-2\tilde{x}' \Sigma^{-1} \mu_j + \mu_j' \Sigma^{-1} \mu_j)$$

de forma que se tenha:

$$D_j(\tilde{x}) = \ln(p_j) + \left(\tilde{x} - \frac{1}{2}\mu_j \right)' \Sigma^{-1} \mu_j$$

que é a função discriminante de Anderson (1958). Esta técnica tem por finalidade classificar novos materiais genéticos, de comportamento desconhecido, em populações já conhecidas. Como exemplo, serão consideradas três populações, em que:

π_1 , π_2 e π_3 : populações 1, 2 e 3, respectivamente;

μ_1 , μ_2 e μ_3 : vetor de médias dos p caracteres avaliados em π_1 , π_2 e π_3 , respectivamente;

Σ_1 , Σ_2 e Σ_3 : matriz de variâncias e covariâncias entre os caracteres avaliados em π_1 , π_2 e π_3 , respectivamente. A matriz de variâncias e covariâncias comum é denotada por Σ ;

p_1 , p_2 e p_3 : probabilidades, *a priori*, de os indivíduos pertencerem a π_1 , π_2 e π_3 , respectivamente; e

\tilde{x} : vetor de variáveis representativas dos caracteres envolvidos na análise.

As funções discriminantes são obtidas por meio de:

$$D_1(x) = \ln(p_1) + (x - \frac{1}{2}\mu_1)' \Sigma^{-1} \mu_1$$

$$D_2(x) = \ln(p_2) + (x - \frac{1}{2}\mu_2)' \Sigma^{-1} \mu_2$$

e

$$D_3(x) = \ln(p_3) + (x - \frac{1}{2}\mu_3)' \Sigma^{-1} \mu_3$$

De forma genérica, tem-se:

$$D_j(\tilde{x}) = \ln(p_j) - \frac{1}{2} \left(-2\tilde{x}' \Sigma^{-1} \mu_j + \mu_j' \Sigma^{-1} \mu_j \right) = \kappa_j + \tilde{x}' \Sigma^{-1} \mu_j$$

ou ainda:

$$D_j(\tilde{x}) = \kappa_j + \alpha_{j1}x_1 + \alpha_{j2}x_2 + \dots + \alpha_{jv}x_v$$

em que:

$$\kappa_j = \ln(p_j) - \frac{1}{2} \mu_j' \Sigma^{-1} \mu_j$$

: constante associada à função discriminante;

α_{jk} : coeficiente de ponderação da k-ésima variável ($k=1,2,\dots,v$) na j-ésima função discriminante; e

x_k : valor representativo do escore da k-ésima variável da observação que se deseja classificar em uma das populações em estudo.

Classifica-se o t-ésimo material genético, com vetor de média \tilde{x}_i , na população π_j se e somente se $D_j(\tilde{x}_i)$ for o maior entre os elementos do conjunto $\{D_1(\tilde{x}_i), D_2(\tilde{x}_i), D_3(\tilde{x}_i)\}$. Utilizando as funções discriminantes e os dados das próprias populações π_1, π_2 e π_3 , estima-se a taxa de erro aparente que mede a eficiência da função discriminante em classificar os genótipos, corretamente, nas populações previamente estabelecidas.

3.5 Análise discriminante quadrática

Há situações em que não é possível estabelecer uma matriz de variâncias e covariâncias comum (Σ), pois as v matrizes Σ_i são heterogêneas (heterocedasticidade entre elas). Então, recorre-se à análise discriminante quadrática, própria para este tipo de situação. O critério de classificação é o mesmo da função discriminante linear de Anderson. Entretanto, na função discriminante

$$D_j(\tilde{x}) = \ln(p_j) - \frac{1}{2} (-2\tilde{x}' \Sigma^{-1} \mu_j + \mu_j' \Sigma^{-1} \mu_j)$$

, a matriz de variância e covariâncias comum (Σ) é substituída pela matriz de variância e covariâncias da população i (Σ_i), de modo que a função fica definida por:

$$D_j(\tilde{x}) = \ln(p_j) - \frac{1}{2} (-2\tilde{x}' \Sigma_i^{-1} \mu_j + \mu_j' \Sigma_i^{-1} \mu_j) = \kappa_j + \tilde{x}' \Sigma_i^{-1} \mu_j$$

Embora mais complexa, as funções discriminantes quadráticas também visam minimizar a probabilidade de má classificação total.

3.6 Avaliação da função discriminante

a. Taxa de erro aparente

Uma vez obtidas as funções discriminantes, é fundamental avaliar a sua eficácia, que é dependente do grau de dissimilaridade entre as populações analisadas e, principalmente, da quantidade e qualidade das variáveis consideradas na discriminação. Como as funções discriminantes são obtidas a partir de análises prévias de observações que se supõe serem, de fato, pertencentes às populações consideradas, pode-se calcular a probabilidade de má classificação, reclassificando toda observação até então disponível. A classificação de uma observação pertencente a uma população π_j em outra é indicativo de menor eficiência da função discriminante estimada, contribuindo para o acréscimo na taxa de erro aparente. Considerando que foram tomadas n_j observações em cada população π_j , deve-se construir o seguinte quadro de classificação:

População	Classificado em				Total de observações	Acertos	Erros
	1	2	...	g			
1	k_{11}	k_{12}	...	k_{1g}	n_1	k_{11}	$m_1 = n_1 - k_{11}$
2	k_{21}	k_{22}	...	k_{2g}	n_2	k_{22}	$m_2 = n_2 - k_{22}$
...
g	k_{g1}	k_{g2}	...	k_{gg}	n_g	k_{gg}	$m_g = n_g - k_{gg}$
Total					$N = \sum_{j=1}^g n_j$	$\sum_{j=1}^g m_j$	

A probabilidade de má classificação, para cada população, é dada por:

$$\hat{p}_j = \frac{m_j}{n_j} \quad (j=1,2,\dots,g)$$

sendo m_j o número de observações retiradas da população π_j que foram, por meio das funções discriminantes obtidas, classificadas em outra população $\pi_{j'}$ sendo $j' \neq j$.

A soma de todos os casos desfavoráveis encontrados em cada população fornece a taxa de erro aparente (TEA), dada por:

$$TEA = \frac{1}{N} \sum_{j=1}^g m_j$$

Esta taxa é, de fato, subestimada, uma vez que se utiliza o artifício de empregar os mesmos dados para obtenção da função discriminante e da probabilidade de má classificação. Quando a taxa de erro aparente é alta deve-se concluir que:

- As populações analisadas não são suficientemente diferenciadas para que possam ser distinguidas por meio das funções discriminantes.
- As populações analisadas, apesar de serem diferenciadas, não puderam ser distinguidas em razão da quantidade e qualidade das variáveis consideradas na discriminação.

b. Validação cruzada

Um dos caminhos para resolver o problema de subestimação é dividir o conjunto de dados em duas partes: uma amostra para obtenção das funções discriminantes e outra para avaliá-las. Entretanto, essa sugestão necessita de uma amostra grande da população (ou grupo) para obter estimativas satisfatórias de má classificação. Há ainda situações em que a função discriminante construída com parte dos dados disponíveis pode estimar coeficientes inadequados à outra amostra.

Um método alternativo para estimar a taxa de erro é a validação cruzada. Neste método, omite-se o primeiro indivíduo da análise e gera-se a(s) função(ões)

discriminante(s) usando os $(\sum_{i=1}^g n_i - 1)$ indivíduos restantes. Classifica-se então o

indivíduo omitido. Este procedimento é repetido para todos os indivíduos $(\sum_{i=1}^g n_i)$.

Finalmente, o número de classificações erradas para cada população é contabilizado e a taxa de erro de cada população é computada. A taxa de erro aparente (global) pode ser obtida com as médias ponderadas dessas proporções, considerando as probabilidades *a priori* como pesos. Com este procedimento, tem-

se um menor viés. O interessante é que, nesta técnica, a cada iteração, uma função discriminante diferente está sendo avaliada; contudo, com a eliminação de apenas um indivíduo, espera-se que essa alteração não afete significativamente a função discriminante, a menos que este indivíduo seja uma informação muito discrepante da amostra sob análise.

3.7 Dissimilaridade entre as populações

A dissimilaridade entre as duas populações expressa pela distância generalizada de Mahalanobis é indicativo da eficácia da discriminação das observações. Nas análises discriminantes considera-se uma observação \bar{x} pertencente a uma população π_j , ao invés de $\pi_{j'}$, quando:

$$(\bar{x} - \hat{\mu}_j)' S_c^{-1} (\bar{x} - \hat{\mu}_j) < (\bar{x} - \hat{\mu}_{j'})' S_c^{-1} (\bar{x} - \hat{\mu}_{j'})$$

Por outro lado, a distância entre π_j e $\pi_{j'}$ é dada por:

$$D_{jj'}^2 = (\hat{\mu}_j - \hat{\mu}_{j'})' S_c^{-1} (\hat{\mu}_j - \hat{\mu}_{j'})$$

A verificação da dissimilaridade entre π_j e $\pi_{j'}$ corresponde a avaliar a hipótese $H_o : \mu_j = \mu_{j'}$ versus $H_a : \mu_j \neq \mu_{j'}$, ou seja, verificar se os dois vetores de média das duas populações diferem estatisticamente. Considerando que as observações nas populações seguem distribuição normal multivariada, com matriz de variâncias e covariâncias comum, Σ , o teste é efetuado por meio de:

$$F_o = \frac{n_j + n_{j'} - p - 1}{p(n_j + n_{j'} - 2)} \frac{n_j n_{j'}}{n_j + n_{j'}} D_{jj'}^2$$

sendo

n_j e $n_{j'}$: número de indivíduos pertencentes às populações π_j e $\pi_{j'}$, respectivamente; e

p : número de variáveis consideradas nas funções discriminantes.

O valor de F_o obtido tem distribuição F com g_1 e g_2 graus de liberdade, dados por:

$$g_1 = p$$

$$g_2 = n_1 + n_2 - p - 1$$

Valores significativos indicam que as populações são bastante distintas; portanto, a classificação de um novo indivíduo em um dos grupos se fará com maior probabilidade de acerto. Esse teste pode ser realizado até mesmo antes do estabelecimento das funções discriminantes, uma vez que populações ou grupos semelhantes terão elevadas taxas de classificação errada.

Ilustração

Novamente será considerada, como ilustração, a obtenção das funções discriminantes a partir das médias de 15 genótipos de milho-pipoca descritas no exemplo anterior. Para o exemplo em consideração, foi admitido que as probabilidades de classificação fossem todas iguais a 1/3. Assim, com base nas médias de cada população e na matriz ω , obtiveram-se as funções discriminantes apresentadas na Tabela 3.4. Cada função é uma combinação linear das nove características agronômicas, existindo tantas funções quanto for o número de populações avaliadas.

Tabela 3.4 - Funções discriminantes de populações de milho-pipoca, obtidas em função de nove características agronômicas

Descrição	D ₁ (x)	D ₂ (x)	D ₃ (x)
Constante	-206,6206	-199,9775	-204,2153
AP	5,188932	5,0895	4,9920
AE	35,1771	34,9193	34,0730
NE	2,4987	2,5194	2,4642
PROL	39,5887	38,1416	38,8615
CE	6,6392	6,6605	6,6300
QUE	177,8310	164,1034	169,4405
DOEN	615,6343	609,6025	602,4976
PCG	-0,8172	-0,7799	-0,5689
PROD	0,0024	0,0014	0,0019

Com base nas funções discriminantes obtidas, estimou-se, para cada genótipo, o valor discriminante, conforme apresentado na Tabela 3.5. Assim, para o genótipo 1, os valores discriminantes para as populações 1, 2 e 3 foram, respectivamente, 229,69, 226,94 e 226,99. Dentre esses valores, o máximo é aquele correspondente à população 1, devendo este genótipo ser alocado nesta população. Como se trata da classificação de genótipos em que se conhece previamente a população a que pertence, é possível avaliar a adequação das funções discriminantes obtidas.

Tabela 3.5 - Valor discriminante e classificação de cultivares de milho-pipoca, a partir de funções discriminantes estabelecidas pela combinação linear de nove características agronômicas

Genótipo	D1(x)	D2(x)	D3(x)	Classificação
1	229,6965	226,9401	226,9884	Pop -1
2	192,7700	189,9187	190,5178	Pop -1
3	202,8230	200,2624	199,9392	Pop -1
4	173,2098	170,6077	170,1894	Pop -1
5	219,9865	216,9745	217,1289	Pop -1
6	152,8793	155,3459	152,9667	Pop -2
7	220,1073	222,6194	220,2842	Pop -2
8	194,0388	197,1702	194,2966	Pop -2
9	211,9085	214,8970	212,0990	Pop -2
10	198,8963	201,5407	198,3617	Pop -2
11	200,9484	200,5977	204,0269	Pop -3
12	187,0649	187,2999	190,0023	Pop -3
13	227,1178	227,0670	229,6983	Pop -3
14	190,2401	190,1714	192,8537	Pop -3
15	201,9561	202,3549	204,3356	Pop -3

Neste exemplo, a taxa de erro aparente é nula, como pode ser verificado no quadro a seguir:

População	Classificado em			Total de observações	Acertos	Erros
	1	2	3			
1	5	0	0	n ₁ = 5	k ₁₁ = 5	m ₁ = 0
2	0	5	0	n ₂ = 5	k ₂₂ = 5	m ₂ = 0
3	0	0	5	n ₃ = 5	k ₃₃ = 5	m _g = 0
Total				N = 15		0

logo:

$$TEA = \frac{1}{N} \sum_{j=1}^g m_j = 0$$

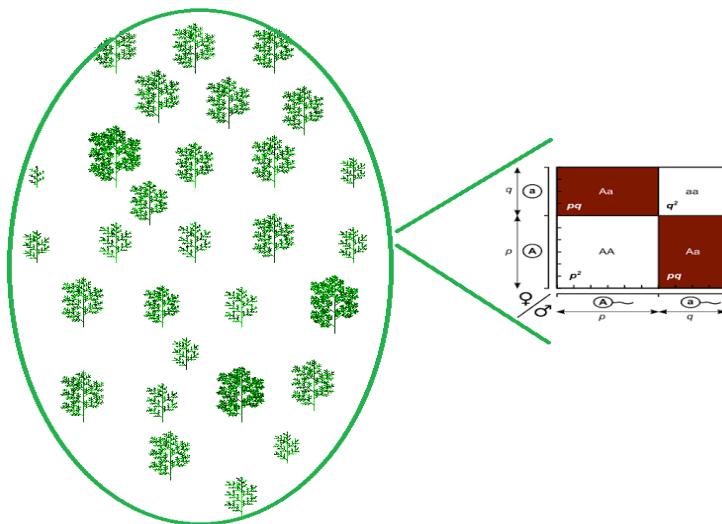
Assim, verifica-se que a taxa de erro aparente dada pela relação entre o número de classificações erradas e o número total de classificações é igual a

zero, indicando que as funções discriminantes estimadas são eficazes em alocar os genótipos em suas devidas populações a partir das nove características agronômicas estudadas.

Outros cultivares poderiam ser avaliados e classificados em relação às funções obtidas, tornando o processo mais eficaz e de menor subjetividade.

Capítulo 4

Estrutura Genética de Populações



4.1. Introdução

Na Genética de Populações, o termo população refere-se à reunião de indivíduos com diferentes genótipos, pertencentes a uma mesma espécie e vivendo dentro de uma área suficientemente restrita, com sistema de acasalamento definido e que possibilite a formação de descendentes em freqüência proporcional à contribuição gamética de seus genitores. Com esse conceito, o termo assume um componente genético (indivíduos pertencentes à mesma espécie) e um componente espacial (convívio dentro de uma mesma área).

Conhecer a estrutura genética de uma população permite presumir, senão desvendar, quais os fenômenos ecológicos e genéticos atuantes nela. A investigação sobre a estrutura de populações permite detectar modos de reprodução e estrutura familiar, estimar níveis de migração e dispersão, ajudar no manejo de espécies ameaçadas de extinção e na manutenção de bancos de germoplasma, além de auxiliar no diagnóstico do histórico evolutivo de um conjunto de táxons. Ainda, para o melhoramento genético, informações da estrutura populacional permitem ao melhorista realizar, ou predizer, mudanças em magnitude e sentido desejados.

A teoria da genética populacional está preocupada, principalmente, em entender duas variáveis bastante conectadas: a freqüência gênica e a freqüência genotípica. Elas são os descritores básicos, embora ênfase possa ser dada também à heterozigosidade (diversidade gênica).

Modelos teóricos desenvolvidos na Genética de Populações auxiliam na busca por respostas sobre o comportamento das populações ou espécies. Cita-se, como exemplo, o modelo reprodutivo de acasalamento ao acaso, que possui uma função importante em outros modelos da genética populacional porque, freqüentemente, serve como ponto de partida para a consideração de situações mais realísticas.

4.2. Freqüências alélicas e genotípicas de uma população

A estrutura de uma população é definida pela freqüência dos alelos que compõem os diferentes genótipos dos diferentes indivíduos integrantes da população. Considerando apenas o gene A/a, define-se uma população de tamanho **N** como sendo aquela constituída pelo agrupamento de **N₁₁** indivíduos AA, **N₁₂** Aa e **N₂₂** aa, tal como ilustrado na Tabela 4.1.

Tabela 4.1 - Freqüência genotípica numa população

Genótipos	Nº de indivíduos	Freqüência
AA	N ₁₁	D = N ₁₁ /N
Aa	N ₁₂	H = N ₁₂ /N
aa	N ₂₂	R = N ₂₂ /N
Total	N	1

em que:

$$N = N_{11} + N_{12} + N_{22} = N_{AA} + N_{Aa} + N_{aa}$$

$$D + H + R = 1,0$$

A freqüência do alelo k (no caso, A e a) na população pode ser obtida por meio da expressão:

$$f(\text{Alelo } k) = \frac{\text{número de alelos } k}{\text{número total de alelos}}$$

Assim, lembrando que numa espécie diplóide cada indivíduo apresenta dois alelos, tem-se:

$$f(A) = p = \frac{n_A}{n} = \frac{2N_{AA} + N_{Aa}}{2N} = D + \frac{1}{2}H$$

e

$$f(a) = q = \frac{n_a}{n} = \frac{2N_{aa} + N_{Aa}}{2N} = R + \frac{1}{2}H$$

sendo:

$$p + q = 1 \text{ e } n_A + n_a = n = 2N$$

Será considerada, como ilustração, uma população com 1.000 indivíduos, conforme descrito a seguir:

Genótipo	Número observado
AA	200
Aa	400
aa	400
Total	1000

Com os valores observados, estimam-se as freqüências genotípicas, tal como apresentado a seguir:

$$D = 200/1000 = 0,20$$

$$H = 400/1000 = 0,40$$

$$R = 400/1000 = 0,40$$

A partir destes valores são obtidas as freqüências dos alelos A e a, dadas por:

$$p = f(A) = D + \frac{1}{2} H = 0,20 + \frac{1}{2} (0,40) = 0,4$$

$$q = f(a) = R + \frac{1}{2} H = 0,40 + \frac{1}{2} (0,40) = 0,6$$

4.3. Processos que afetam a freqüência gênica

Os seguintes processos afetam a freqüência gênica de uma população:

a. Processos Sistemáticos

São aqueles cuja alteração na freqüência gênica é conhecida tanto em magnitude quanto em direção. Consideram-se como processos sistemáticos a seleção, a migração e a mutação.

b. Processos Dispersivos

São aqueles em que é possível conhecer apenas a magnitude da alteração da freqüência, mas não a direção em que ela foi alterada. Como processo dispersivo é considerada a oscilação (ou deriva) genética ou a amostragem.

4.4. Equilíbrio de Hardy-Weinberg

4.4.1. Equilíbrio com relação a um loco

O equilíbrio de Hardy-Weinberg (também princípio de Hardy-Weinberg, ou lei de Hardy-Weinberg) é a base da genética de populações. Foi demonstrado independentemente por Godfrey Harold Hardy na Inglaterra e por Wilhelm Weinberg, na Alemanha, em 1908. Pelos pressupostos de Hardy e Weinberg considera-se que, em uma população suficientemente grande e na ausência de seleção, migração e mutação, o equilíbrio é atingido após uma geração de acasalamento ao acaso (“aaa”), de maneira que a relação genotípica torna-se igual ao quadrado da freqüência gênica e, com as sucessivas gerações de acasalamento ao acaso, permanece inalterada. Para ilustrar esse fato, será considerada uma população inicial com genótipos AA, Aa e aa, nas freqüências D, H e R, respectivamente. As freqüências alélicas são p e q, para A e a, respectivamente. Considerando que ocorre acasalamento ao acaso entre os indivíduos desta população, pode-se predizer a descendência, conforme ilustrado na Tabela 4.2.

Tabela 4.2 - Freqüência genotípica e alélica numa população antes e após acasalamento ao acaso (aaa)

População Inicial (P_0)		População após o “aaa” (P_1)
$AA = D$ $Aa = H$ $aa = R$ $f(A) = p = D + H/2$ $f(a) = q = R + H/2$	\Rightarrow	$AA = D_1 = p^2$ $Aa = H_1 = 2pq$ $aa = R_1 = q^2$ $f(A) = p_1 = p$ $f(a) = q_1 = q$

Assim, pode-se conhecer as freqüências genotípica e alélica que ocorrerão numa geração futura, derivada de sucessivos acasalamentos ao acaso numa população inicial, a partir da sua freqüência alélica (p e q) original. Esse conhecimento preditivo permite aos pesquisadores estabelecer estratégias de melhoramento e manipulação de população, bem como reconhecer a dinâmica evolutiva da espécie em determinadas regiões.

O exposto pode ser facilmente demonstrado se forem considerados todos os possíveis tipos de cruzamento, na população original, ilustrados na Tabela 4.3.

Tabela 4.3 - Relação dos possíveis acasalamentos numa população (P_0) e previsão da descendência (P_1) resultante do acasalamento ao acaso

Cruzamentos			Descendência em P_1		
em P_0		Freqüência	AA	Aa	aa
AA	x	AA	D^2	D^2	-
AA	x	Aa	$2DH$	DH	DH
AA	x	aa	$2DR$	-	$2DR$
Aa	x	Aa	H^2	$\frac{1}{4} H^2$	$\frac{1}{4} H^2$
Aa	x	aa	$2HR$	-	HR
aa	x	aa	R^2	-	R^2
Total		1	$D_1 = (D + \frac{1}{2} H)^2$ p^2	$H_1 = 2(D + \frac{1}{2} H)(R + \frac{1}{2} H)$ $2pq$	$R_1 = (R + \frac{1}{2} H)^2$ q^2

Assim, demonstra-se que a freqüência genotípica da descendência pode ser predita por meio do conhecimento da freqüência alélica na população genitora. Com o acasalamento ao acaso, a freqüência alélica não se altera, ou seja:

$$f(A \text{ em } P_1) = p_1 = D_1 + \frac{1}{2} H_1 = p^2 + \frac{1}{2} 2pq = p$$

$$f(a \text{ em } P_1) = q_1 = R_1 + \frac{1}{2} H_1 = q^2 + \frac{1}{2} 2pq = q$$

A relação genotípica da descendência é, portanto, dada por $(p + q)^2$.

O mais importante no estudo das condições de equilíbrio de Hardy-Weinberg são as suposições implícitas no modelo genético, que impõem várias restrições como:

a) Ausência de migração, ou de fatores relacionados à incorporação de novos indivíduos em uma população que difira em freqüência alélica.

b) Ausência de mutação, ou seja, ausência de fatores que promovam mudança na estrutura dos alelos existente na população, gerando variação em suas freqüências.

c) Ausência de seleção que promova a perpetuação diferencial e não aleatória de diferentes genótipos.

d) Acasalamento ao acaso.

e) População grande - os genes transmitidos de uma geração para outra são sempre amostras dos genes existentes na geração paterna e, portanto, as freqüências alélicas podem estar sujeitas à variação de amostragem que ocorre nas gerações que se sucedem. Quanto menor o número de genitores maior será a variação causada pela amostragem. Estas variações podem modificar a constituição genética da população.

Além dessas suposições são assumidas várias condições para a população, quais sejam:

- a) organismos diplóides;
- b) genes autossônicos;
- c) dois alelos por loco;
- d) um único gene.

4.4.2. Descendência do acasalamento ao acaso

Utilizando o princípio de Hardy-Weinberg é possível predizer a descendência resultante do acasalamento ao acaso considerando a população como um todo, em vez de particularizar os cruzamentos individuais, como normalmente é tratado na Genética Mendeliana. Assim, pensando em indivíduos particulares, a freqüência do alelo A, por exemplo, será igual a 1, $\frac{1}{2}$ ou 0 para genótipos iguais a AA, Aa ou aa, respectivamente. Entretanto, considerando todos os indivíduos da população, a freqüência de A será p ($0 \leq p \leq 1$) e a descendência, admitindo todos os acasalamentos possíveis, poderá ser previda a partir deste valor p .

Como ilustração, será considerada uma população constituída por 200 indivíduos AA, 80 Aa e 74 aa. Para esta população, a freqüência dos indivíduos derivados do acasalamento ao acaso será de 18,49, 49,02 e 32,49%, de AA, Aa e aa, respectivamente. Esses valores correspondem ao que seria esperado no equilíbrio, pois se têm, originalmente, os valores:

$$D = 46/200 = 0,23$$

$$H = 80/200 = 0,40$$

$$R = 74/200 = 0,37$$

A partir destes valores, obtém-se:

$$p = f(A) = D + \frac{1}{2} H = 0,23 + \frac{1}{2} (0,40) = 0,43$$

e

$$q = f(a) = R + \frac{1}{2} H = 0,37 + \frac{1}{2} (0,40) = 0,57$$

A freqüência esperada na descendência será, portanto, a seguinte:

$$f(AA) = p^2 = 0,43^2 = 18,49\%$$

$$f(Aa) = 2pq = 2 \times 0,43 \times 0,57 = 49,02\%$$

$$f(aa) = q^2 = 0,57^2 = 32,49\%$$

O leitor obterá esses mesmos valores se considerar todos os cruzamentos possíveis, conforme ilustrado na Tabela 4.3.

4.4.3. Variância e covariância da freqüência alélica

Análise de dois alelos por loco

Será considerada uma população constituída pelos genótipos A_1A_1 , A_1A_2 e A_2A_2 , em que o alelo A_1 ocorre n_1 vezes na população e o genótipo A_iA_j ocorre N_{ij} vezes. Assim, tem-se:

Genótipo	Ocorrência	Alelos	Ocorrência
A_1A_1	N_{11}	A_1	n_1
A_1A_2	N_{12}	A_2	n_2
A_2A_2	N_{22}		
Total	N		$n = 2N$

sendo: $n_1 = 2N_{11} + N_{12}$

$$n_2 = 2N_{22} + N_{12}$$

$$n = n_1 + n_2 = 2N$$

A quantidade de ocorrência do alelo A_1 , medida por n_1 , é uma variável que segue distribuição binomial e, portanto, se verifica:

$$E(n_1) = np$$

$$V(n_1) = npq$$

em que $p = f(A_1)$ e $q = 1 - p = f(A_2)$

A partir das quantidades de alelos ou de genótipos podem ser estimadas as freqüências alélicas, por meio de:

$$\hat{p} = f(A_1) = \frac{n_1}{n} = \frac{N_{11}}{N} + \frac{1}{2} \frac{N_{12}}{N}$$

portanto:

$$E(\hat{p}) = E\left(\frac{n_1}{n}\right) = \frac{1}{n} np = p$$

e

$$V(\hat{p}) = V\left(\frac{n_1}{n}\right) = \frac{1}{n^2} npq = \frac{1}{n} pq$$

Para se ter variância na freqüência gênica é necessário que os alelos não estejam fixados. É interessante observar que:

$$V(\hat{p}) = E(\hat{p}^2) - [E(\hat{p})]^2$$

logo:

$$E(\hat{p}^2) = p^2 + \frac{1}{n} pq$$

A probabilidade de se ter conjuntamente $n_1 n_2$ ocorrência dos alelos A_1 e A_2 pode também ser obtida utilizando princípios da distribuição multinomial, de forma que se deduz que:

$$E(n_1 n_2) = n(n-1)pq$$

A partir desta expressão calculam-se as covariâncias:

$$\text{Cov}(n_1, n_2) = -npq$$

$$\text{Cov}(\hat{p}, \hat{q}) = -\frac{1}{n} pq$$

Para o caso de apenas dois alelos por loco, a correlação entre as freqüências gênicas será igual a -1, pois:

$$r_{\hat{p}\hat{q}} = \frac{\text{Cov}(\hat{p}, \hat{q})}{\sqrt{V(\hat{p})V(\hat{q})}} = -1$$

Análise de alelos múltiplos

Neste caso, considera-se a existência de s alelos para um determinado loco e que a população é constituída por genótipos $A_i A_j$ cujo número de ocorrência é dado por N_{ij} . Também se considera que cada alelo A_i ocorre n_i vezes na população. Assim:

Genótipos	Ocorrência	Alelos	Ocorrência
$A_i A_i$	N_{ii}	A_i	n_i
$A_i A_j$	N_{ij}	A_j	n_j
Total	$N = \sum_i N_{ii} + \sum_{i < j} N_{ij}$		

As freqüências genotípicas amostrais são denotadas por \hat{P}_{ii} , para os genótipos $A_i A_i$, e \hat{P}_{ij} para os genótipos $A_i A_j$, tendo-se os seguintes estimadores:

$$\hat{P}_{ii} = \frac{N_{ii}}{N} \quad \text{e} \quad \hat{P}_{ij} = \frac{N_{ij}}{N}$$

Também, a partir da amostra, é estimado:

$$\hat{p}_i = \frac{n_i}{n} = \frac{n_i}{2N}$$

sendo: $n_i = 2N_{ii} + \sum_{i \neq j} N_{ij}$

$$n = \sum_{i=1}^s n_i = 2N$$

Em termos populacionais, podem ser definidos:

P_{ii} : freqüência de $A_i A_i$

P_{ij} : freqüência de $A_i A_j$

$$p_i = P_{ii} + \frac{1}{2} \sum_{j=1}^s P_{ij}$$

de forma que possas ser obtidas as esperanças matemáticas:

$$E(n_i) = 2NP_{ii} + N \sum_{j=1}^s P_{ij}$$

e:

$$E(\hat{p}_i) = E\left(\frac{n_i}{n}\right) = \frac{E(n_i)}{2N} = P_{ii} + \frac{1}{2} \sum_{j=1}^s P_{ij}$$

concluindo-se que:

$$E(\hat{p}_i) = E\left(\frac{n_i}{n}\right) = p_i$$

A variância da quantidade de um determinado alelo é dada por:

$$V(n_i) = V(2N_{ii} + \sum_{j=1}^s N_{ij})$$

sabe-se que:

$$V(N_{ii}) = NP_{ii}(1 - P_{ii})$$

$$\text{Cov}(N_{ij}, N_{i'j'}) = \frac{N-1}{N} P_{ij} P_{i'j'}$$

então:

$$V(n_i) = 2N(p_i + P_{ii} - 2p_i^2)$$

e, portanto:

$$V(\hat{p}_i) = \frac{1}{2N}(p_i + P_{ii} - 2p_i^2)$$

Como

$$V(\hat{p}) = E(\hat{p}^2) - [E(\hat{p}_i)]^2$$

então:

$$E(\hat{p}_i^2) = V(\hat{p}) + [E(\hat{p}_i)]^2 = p_i^2 + \frac{1}{2N}(p_i + P_{ii} - 2p_i^2)$$

Assim, a variância da freqüência de um alelo depende de sua segregação na população e será maior à medida que ele se concentra na forma homozigota. Veja que, para o caso particular de dois alelos por loco e equilíbrio de Hardy-Weinberg, tem-se:

$$P_{ii} = p_i^2$$

logo: $V(p_i) = (1/n)p_i(1-p_i)$

De maneira análoga é obtida a covariância entre as freqüências de pares de alelos, tendo-se:

$$\text{Cov}(\hat{p}_i, \hat{p}_j) = -\frac{1}{2N}p_i p_j$$

Uso de variável indicadora no cálculo da freqüência gênica

Muitas expressões associadas às freqüências alélicas ou genotípicas podem facilmente ser obtidas usando variáveis indicadoras. Em estudos de população com N indivíduos portadores de dois alelos, pode-se utilizar a variável x_{ij} ($i=1,2,\dots,N$ e $j=1,2$), com as seguintes particularidades:

$x_{ij} = 1$, se o alelo j no indivíduo i é A.

$x_{ij} = 0$ nos casos contrários.

Como ilustração, pode ser considerada uma população constituída por 10 indivíduos, dada a seguir:

Indivíduo	1	2	3	4	5	6	7	8	9	10
Genótipo	AA									

Nesta situação, tem-se:

Genótipo	Ocorrência	Alelo	Ocorrência
AA	$N_{11}=3$	A	$n_1=11$
Aa	$N_{12}=5$	a	$n_2=9$
aa	$N_{22}=2$		
Total	$N=10$		$n=20=2N$

A análise da freqüência alélica pode ser feita alternativamente, por meio dos valores de x_{ij} apresentados a seguir:

Indivíduo	Alelo	x_{ij} (A)	X^*_{ij} (a)	Indivíduo	Alelo	x_{ij} (A)	X^*_{ij} (a)
1 AA	1	1	0	6 Aa	1	1	0
	2	1	0		2	0	1
2 AA	1	1	0	7 Aa	1	1	0
	2	1	0		2	0	1
3 AA	1	1	0	8 Aa	1	1	0
	2	1	0		2	0	1
4 Aa	1	1	0	9 aa	1	0	1
	2	0	1		2	0	1
5 Aa	1	1	0	10 aa	1	0	1
	2	0	1		2	0	1

Algumas propriedades interessantes podem ser estabelecidas em relação a x_{ij} :

a) $E(x_{ij})$

É dada por:

$$E(x_{ij}) = 1P(x_{ij} = 1) + 0P(x_{ij} = 0) = 1p_i + 0(1-p_i) = p_i$$

No exemplo:

$$E(x_{ij}) = 1\left(\frac{11}{20}\right) + 0\left(\frac{9}{20}\right) = 0,55$$

b) $E(x_{ij}^2) = p_i$, pois $x_{ij}^2 = x_{ij}$

c) $E(x_{ij}, x_{ij'})$

É fornecida por:

$$E(x_{ij}, x_{ij'}) = \frac{N_{ii}(1x1) + N_{ij}(1x0) + N_{jj'}(0x0)}{N} = \frac{N_{ii}}{N} = P_{ii}$$

Sendo P_{ii} a freqüência de AA na população.

d) $E(x_{ij}, x_{ij'})$

Numa população em que os indivíduos não são parentados, é dado por:

$$E(x_{ij}, x_{ij'}) = E(x_{ij})E(x_{ij'}) = p_i^2$$

Em algumas situações deve-se imaginar que diferentes indivíduos podem não ser independentemente distribuídos. Nesse caso, seria conveniente admitir que:

$$E(x_{ij}, x_{i'j'}) = P_{ii'}$$

sendo $P_{ii'}$ a probabilidade de encontrar mesmo alelo i em dois indivíduos tomados ao acaso da população.

e) $V(x_{ij})$

É dada por:

$$V(x_{ij}) = E(x_{ij}^2) - [E(x_{ij})]^2 = p_i - p_i^2 = p_i(1-p_i)$$

f) $Cov(x_{ij}, x_{ij'})$

É fornecida por:

$$Cov(x_{ij}, x_{ij'}) = P_{ii'} - p_i^2$$

Em uma população sob endogamia, e admitindo apenas dois alelos (A e a), tem-se a seguinte estrutura genotípica:

Genótipo	Freqüência ($F=0$)	Freqüência ($F \neq 0$)
AA	p^2	$p^2 + pqF$
Aa	$2pq$	$2pq(1-F)$
aa	q^2	$q^2 + pqF$

sendo F igual ao coeficiente de endogamia f de Cockerhan (1969) ou F_{IS} de Wright (1951).

Assim, existindo endogamia, tem-se:

$$Cov(x_{ij}, x_{ij'}) = pqF$$

g) $Cov(x_{ij}, x_{i'j'})$

É dada por:

$$Cov(x_{ij}, x_{i'j'}) = E(x_{ij}x_{i'j'}) - E(x_{ij})E(x_{i'j'}) = p_i^2 - p_i p_i = 0, \quad \text{admitindo } x_{ij} \text{ e } x_{i'j'} \text{ independentes.}$$

ou

$$Cov(x_{ij}, x_{i'j'}) = E(x_{ij}x_{i'j'}) - E(x_{ij})E(x_{i'j'}) = P_{ii'} - p_i p_i = P_{ii'} - p_i^2$$

Se $P_{ii'} = p_i^2$, então a covariância é nula. Entretanto, é possível admitir que:

$$P_{ii'} = p_i^2 + p_i(1-p_i)F$$

e

$$P_{i/i} = p_i^2 + p_i(1-p_i)\theta$$

de forma que se tenha:

$$\text{Cov}(x_{ij}, x_{i'j'}) = p_i(1-p_i)\theta$$

sendo θ o coeficiente de co-ancestralidade.

Nesse ponto, é necessário apresentar uma nova expressão da variância da freqüência gênica. Nas considerações anteriores foi visto que:

$$V(\hat{p}_i) = \frac{1}{2N} (p_i + P_{ii} - 2p_i^2)$$

agora,

$$V(\hat{p}_i) = V\left(\frac{\sum_i \sum_j x_{ij}}{2N}\right) = \frac{1}{4N^2} V(\sum_i \sum_j x_{ij})$$

$$V(\hat{p}_i) = \frac{1}{4N^2} [2NV(x_{ij}) + 2NCov(x_{ij}, x_{ij'}) + (4N^2 - 4N)]Cov(x_{ij}, x_{i'j'})$$

$$V(\hat{p}_i) = (P_{i/i} - p_i^2) + \frac{1}{2N} (p_i + P_{ii} - 2P_{i/i})$$

de forma que:

$$V(\hat{p}_i) = p_i(1-p_i) \left(\theta + \frac{F-\theta}{n} + \frac{1-F}{2n} \right)$$

Sob acasalamento envolvendo indivíduos não aparentados ($\theta=0$) tem-se que $V(\hat{p}_i) = p_i(1-p_i)F/2N$, para uma população infinitamente grande.

4.4.4. Equilíbrio com relação a uma série de alelos múltiplos

Mesmo quando mais de dois alelos são considerados em um loco (alelos múltiplos) o equilíbrio é estabelecido após uma única geração de acasalamento ao acaso. Também neste caso a relação genotípica da geração em equilíbrio é dada pelo quadrado da freqüência dos alelos da geração original. Assim, considerando k alelos (sendo $k = 1, 2, \dots, s$; e A_k com freqüência $f(A_k)$), tem-se no equilíbrio a

seguinte propriedade:

$$\text{Relação genotípica no equilíbrio} = [f(A_1) + f(A_2) + \dots + f(A_s)]^2.$$

Será considerada, a título de exemplo, uma série constituída por apenas três alelos: A_1 , A_2 e A_3 , com freqüência p , q e r , respectivamente. Os possíveis genótipos e as respectivas freqüências genotípicas são dados na Tabela 4.4, apresentada a seguir:

Tabela 4.4 - Freqüência genotípica numa população, considerando um loco gênico com três alelos (A_1 , A_2 e A_3)

Genótipo	Nº de indivíduos	Freqüência Genotípica
A_1A_1	N_{11}	$P_{11} = N_{11} / N$
A_1A_2	N_{12}	$P_{12} = N_{12} / N$
A_1A_3	N_{13}	$P_{13} = N_{13} / N$
A_2A_2	N_{22}	$P_{22} = N_{22} / N$
A_2A_3	N_{23}	$P_{23} = N_{23} / N$
A_3A_3	N_{33}	$P_{33} = N_{33} / N$
Total	N	1,0

As freqüências alélicas podem ser obtidas por meio das expressões apresentadas a seguir:

$$f(A_1) = p = \frac{2N_{11} + N_{12} + N_{13}}{2N} = P_{11} + \frac{(P_{12} + P_{13})}{2}$$

$$f(A_2) = q = \frac{2N_{22} + N_{12} + N_{23}}{2N} = P_{22} + \frac{(P_{12} + P_{23})}{2}$$

$$f(A_3) = r = \frac{2N_{33} + N_{13} + N_{23}}{2N} = P_{33} + \frac{(P_{13} + P_{23})}{2}$$

Após uma geração de acasalamento ao acaso, têm-se as freqüências genotípicas descritas na Tabela 4.5.

Tabela 4.5 - Freqüência genotípica numa população em equilíbrio de Hardy-Weinberg, considerando um loco gênico com três alelos (A_1 , A_2 e A_3)

Genótipo	Freqüência
A_1A_1	p^2
A_1A_2	$2pq$
A_1A_3	$2pr$
A_2A_2	q^2
A_2A_3	$2qr$
A_3A_3	r^2
Total	1,0

4.4.5. Verificação das condições de equilíbrio

Há grande importância em avaliar se uma determinada população, ou melhor, cada loco independente, encontra-se, ou não, em equilíbrio de Hardy-Weinberg. Caso isso ocorra, há indicativo de que ela não está sujeita à pressão de seleção e que o fluxo de migração e a taxa de mutação são fatores desprezíveis. Havendo informações sobre a ocorrência dos genótipos, pode-se verificar as condições de equilíbrio, por meio do clássico teste de qui-quadrado, teste de razão de verossimilhança e, ainda, teste exato de Fisher. A hipótese de nulidade a ser testada é a de que os genótipos observados são produtos de uma união aleatória de gametas masculinos e femininos.

Teste de Qui-quadrado (χ^2)

Neste caso, para avaliar se uma população encontra-se em equilíbrio, devem-se calcular as freqüências gênicas, predizer os valores esperados para as freqüências genotípicas e avaliar, por meio do teste de qui-quadrado, se os valores observados e esperados não são discrepantes, sendo, neste caso,

indicativos de que a população se encontra em equilíbrio. Para o caso da análise de um único loco, com s alelos, as seguintes classes genotípicas e suas quantidades observadas e esperadas devem ser consideradas:

Genótipos	Núm. Observado(O_{ij})	Núm. Esperado(E_{ij})	Desvio($=O_{ij}-E_{ij}$)
A_1A_1	N_{11}	p_1^2N	d_{11}
A_1A_2	N_{12}	$2p_1p_2N$	d_{12}
A_1A_3	N_{13}	$2p_1p_3N$	d_{13}
...
$A_{s-1}A_s$	$N_{s-1,s}$	$2p_{s-1}p_sN$	$d_{s-1,s}$
A_sA_s	$N_{s,s}$	p_s^2N	$d_{s,s}$
Total	N	N	0

A estatística qui-quadrado é obtida por meio de:

$$\chi^2 = \sum_{i=1}^s \sum_{j \geq i}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}} , \text{ que está associada a } n_c - r \text{ graus de liberdade, em}$$

que:

n_c : é o número de classe genotípicas avaliadas. Para s alelos, tem-se que:

$$n_c = \frac{s(s+1)}{2}$$

r: é o número de restrições ou o número de informações necessárias para ser estabelecidos os valores esperados no teste de qui-quadrado. Portanto, nesta caso, é necessário informações sobre o número total de indivíduos estudados (N) e os valores das freqüência de s-1 alelos. Assim, tem-se que:

$$r = 1 + (s-1) = s$$

Assim, os graus de liberdade associado ao valor de χ^2 será igual a $s(s-1)/2$.

Dois exemplos são ilustrados a seguir, em que se pretende verificar se as populações P_1 e P_2 se encontram em equilíbrio, considerando as freqüências genotípicas dadas na Tabela 4.6.

Tabela 4.6 - Valores observados e freqüências genotípicas das populações P_1 e P_2 para um loco com dois alelos (A e a)

Genótipo	População 1 (P_1)		População 2 (P_2)	
	Ocorrência	Freqüência	Ocorrência	Freqüência
AA	15	0,375	14	0,350
Aa	12	0,300	16	0,400
aa	13	0,325	10	0,250
Total	40	1,000	40	1,000

Com os dados disponíveis estimam-se as freqüências alélicas, como descrito a seguir:

a) Para a população P_1

$$p = f(A) = D + \frac{1}{2}H = 0,375 + \frac{1}{2}0,300 = 0,525$$

e

$$q = f(a) = R + \frac{1}{2}H = 0,325 + \frac{1}{2}0,350 = 0,475$$

No equilíbrio, espera-se que a freqüência seja igual a p^2 para AA, $2pq$ para Aa e q^2 para aa, que corresponde a 0,2756 AA, 0,4987 Aa e 0,2256 aa. Assim, considerando os 40 indivíduos, podem-se comparar os valores esperados com os observados, como descrito a seguir:

Genótipo	Observado (O)	Esperado no equilíbrio(E)	Desvio(O-E)
AA	15	11,025	3,975
Aa	12	19,950	-7,950
aa	13	9,025	3,975

Como se dispõe de três classes fenotípicas, com valores esperados obtidos por meio das estimativas de p (ou de q) e de N (tamanho amostral), estima-se a estatística χ^2 , associada a 1 grau de liberdade. Assim, para os dados considerados, tem-se:

$$\chi^2 = \frac{(15 - 11,025)^2}{11,025} + \frac{(12 - 19,950)^2}{19,950} + \frac{(13 - 9,025)^2}{9,025} = 6,352$$

O valor de probabilidade associado é $\alpha = 0,01171$ (1,171%). Daí, conclui-se que os dados não se ajustam ao esperado, sendo, portanto, indicativo de que a população P_1 não se encontra em equilíbrio.

a) Para a população P_2

$$p = f(A) = D + \frac{1}{2}H = 0,35 + \frac{1}{2}0,40 = 0,55$$

e

$$q = f(a) = R + \frac{1}{2}H = 0,25 + \frac{1}{2}0,40 = 0,45$$

No equilíbrio, espera-se que a freqüência seja igual a 0,3025 (p^2) para AA, 0,4950 (2pq) para Aa e 0,2025 (q^2) para aa. Assim, considerando os 40 indivíduos, podem-se comparar os valores esperados com os observados, como descrito a seguir:

Genótipo	Observado (O)	Esperado no equilíbrio(E)	Desvio(O-E)
AA	14	12,1	1,9
Aa	16	19,8	-3,8
aa	10	8,1	1,9

Estima-se a estatística χ^2 , associada a 1 grau de liberdade, por meio de:

$$\chi^2 = \frac{(14 - 12,1)^2}{12,1} + \frac{(16 - 19,8)^2}{19,8} + \frac{(10 - 8,1)^2}{8,1} = 1,4733$$

O valor de probabilidade associado é $\alpha = 0,2248$ (22,48%). Daí, tomando um valor referencial de 5% como erro mínimo para se rejeitar H_0 , conclui-se que os dados se ajustam ao esperado, sendo, portanto, indicativo de que a população P_2 se encontra em equilíbrio.

Cabe ressaltar que o teste de qui-quadrado é sensível ao tamanho amostral. Assim, a estimativa da estatística χ^2 não deve ser obtida com dados de freqüência das classes genotípicas.

Teste da Razão de Verossimilhança (Teste G ou G²)

A razão de verossimilhança oferece um caminho sistemático para se testar o equilíbrio quando existem mais de dois alelos por loco. O teste é baseado na razão entre duas funções de máxima verossimilhança: uma definida pelas freqüências genotípicas esperadas no equilíbrio (L_0) e a outra definida pelas freqüências genotípicas observadas (L_1). Admitindo que o número de ocorrência das classes genotípicas segue distribuição multinomial, as funções L_0 e L_1 , para quaisquer k alelos autossômicos de um loco, são definidas por:

$$L_0 = \lambda \prod_{k=1}^s \left[\left(\frac{n_k}{2N} \right)^2 \right]^{N_{kk}} \prod_{\substack{k=1 \\ k'>k}}^s \left(2 \frac{n_k n_{k'}}{2N 2N} \right)^{N_{kk'}}$$

e

$$L_1 = \lambda \prod_{k=1}^s \left(\frac{N_{kk}}{N} \right)^{N_{kk}} \prod_{\substack{k=1 \\ k'>k}}^s \left(\frac{N_{kk'}}{N} \right)^{N_{kk'}}$$

em que:

$$\lambda = \frac{N!}{\prod_{k=1}^s N_{kk'}!}$$

N: tamanho amostral;

n_k : número de alelos k da população amostrada; e

N_{kk} e $N_{kk'}$: números de homozigotos e de heterozigotos amostrados, respectivamente.

O primeiro produtório de cada função corresponde a todos os homozigotos e o segundo a todos os heterozigotos, em relação ao loco sob estudo. Para que as funções L_0 e L_1 tornem-se mais tratáveis para o cálculo, adota-se a função suporte de L_0 e L_1 , que nada mais é do que o logaritmo neperiano destas, de modo que:

$$\ln L_0 = \ln(\lambda) + \sum_{k=1}^s N_{kk} \ln \left(\frac{n_k}{2N} \right)^2 + \sum_{\substack{k=1 \\ k'>k}}^s N_{kk'} \ln \left(2 \frac{n_k n_{k'}}{2N 2N} \right)$$

e

$$\ln L_1 = \ln(\lambda) + \sum_{k=1}^s N_{kk} \ln\left(\frac{N_{kk}}{N}\right) + \sum_{\substack{k=1 \\ k' > k}}^s N_{kk'} \ln\left(\frac{N_{kk'}}{N}\right)$$

O teste para o equilíbrio é dado pela seguinte razão de verossimilhança:

$$G^2 = -2 \ln\left(\frac{L_0}{L_1}\right) = -2 (\ln L_0 - \ln L_1)$$

O número de informações necessário para obter L_1 é dado por $g - 1$, sendo g o número de classes genotípicas avaliadas, ou seja: $g = s(s+1)/2 - 1$. Entretanto, o número de informações necessário para obter L_0 é dado pela freqüência de $s-1$ alelos. Assim, sob equilíbrio, a quantidade $-2 (\ln L_0 - \ln L_1)$ tem distribuição aproximada de qui-quadrado, com $s(s-1)/2$ graus de liberdade (WEIR, 1996).

A utilização do teste da razão de verossimilhança será ilustrada com as populações P_1 e P_2 , cujas informações das classes genotípicas foram descritas na Tabela 4.6. Considerando o número de alelos A e a , com freqüências alélicas p e q , respectivamente, obtém-se as freqüências genotípicas observadas e esperadas, conforme a Tabela 4.7.

Tabela 4.7 - Freqüências genotípicas observadas e esperadas nas populações P_1 e P_2 , descritas na Tabela 4.6

Genótipo	População P_1				População P_2			
	Num. Ind.	Freq. Obs.	Freq. Esp.		Num. Ind.	Freq. Obs.	Freq. Esp.	
A_1A_1	15	0,375	0,275625		14	0,35	0,3025	
A_1A_2	12	0,300	0,498750		16	0,40	0,4950	
A_2A_2	13	0,325	0,225625		10	0,25	0,2025	
A_1	42				44			
A_2	38				36			

A partir das informações da Tabela 4.7 é possível construir as funções suporte de L_0 e L_1 , de modo que:

Para a população P_1

$$\begin{aligned}\ln L_0 &= \ln(\lambda) + N_{11} \ln\left(\frac{n_1}{2N}\right)^2 + N_{22} \ln\left(\frac{n_2}{2N}\right)^2 + N_{12} \ln\left(2 \frac{n_1 n_2}{2N 2N}\right)^2 \\ &= \ln(\lambda) + 15 \ln(0,275625) + 13 \ln(0,225625) + 2 \ln(0,49875) \\ &= \ln(\lambda) - 47,0339\end{aligned}$$

e

$$\begin{aligned}\ln L_1 &= \ln(\lambda) + N_{11} \ln\left(\frac{N_{11}}{N}\right) + N_{22} \ln\left(\frac{N_{22}}{N}\right) + N_{12} \ln\left(\frac{N_{12}}{N}\right) = \\ &= \ln(\lambda) + 15 \ln(0,375) + 13 \ln(0,325) + 2 \ln(0,300) = \ln(\lambda) - 43,7712\end{aligned}$$

A razão de verossimilhança é dada por:

$$G^2 = -2 \left(\frac{L_0}{L_1} \right) = -2 (\ln L_0 - \ln L_1) = -2(-47,0339 + 43,7712) = 6,5254$$

O valor de G^2 está associado a $2(2 - 1)/2 = 1$ grau de liberdade, e o valor de probabilidade associado é $\alpha = 0,01063$ (1,063%). Admitindo um nível crítico de probabilidade igual a 5%, conclui-se que os dados não se ajustam ao esperado, sendo, portanto, indicativo de que a população P_1 não se encontra em equilíbrio, estando de acordo com o resultado anteriormente obtido pelo teste de qui-quadrado.

Para a população P_2

$$\ln L_0 = \ln(\lambda) + 14 \ln(0,3025) + 10 \ln(0,2025) + 16 \ln(0,4950) = \ln(\lambda) - 43,9607$$

e

$$\ln L_1 = \ln(\lambda) + 14 \ln(0,35) + 10 \ln(0,25) + 16 \ln(0,40) = \ln(\lambda) - 43,2211$$

A razão de verossimilhança é dada por:

$$G^2 = -2 \left(\frac{L_0}{L_1} \right) = -2 (\ln L_0 - \ln L_1) = -2(-43,9607 + 43,2211) = 1,4742$$

O valor de G^2 está associado a $2(2 - 1)/2 = 1$ grau de liberdade, e o valor de probabilidade associado é $\alpha = 0,2238$ (22,38%). Admitindo um nível crítico de probabilidade igual a 5%, conclui-se que os dados observados nesta população se ajustam ao esperado para as condições de equilíbrio. Assim, os resultados obtidos constituem indicativo de que a população P_2 se encontra em equilíbrio.

Devem-se considerar alguns aspectos para o teste de qui-quadrado (χ^2) e a razão de verossimilhança (teste G^2). O primeiro deles é que são testes sensíveis a pequenos valores esperados nas classes genotípicas. O que se tem feito no teste de qui-quadrado é usar a correção de Yates (YATES, 1934). Contudo, em situações de alelos múltiplos, em que alguns possuem freqüências bem pequenas na população, a proposta de correção ou até mesmo uso do χ^2 clássico pode levar a resultados inadequados (GUO; THOMPSON, 1992). Outra solução alternativa seria agrupar classes genotípicas de modo a aumentar o número esperado de uma classe; todavia, esta é uma solução pobre do ponto de vista estatístico, pois, obviamente, há perda de informação de classes genotípicas. Diante desses impedimentos, recorre-se ao teste exato de Fisher.

Teste Exato de Fisher

O teste exato (FISHER, 1935; LEVENE, 1949; HALDANE, 1954) é baseado na probabilidade assumida por um possível arranjo genotípico (conjunto de classes genotípicas) condicionado às freqüências alélicas amostradas na população. A probabilidade condicional de cada arranjo genotípico é estimada assumindo que as classes genotípicas seguem as proporções do equilíbrio de Hardy-Weinberg. Ocorre que alguns ou vários arranjos genotípicos podem ser obtidos para um particular conjunto de alelos observados, o que depende da quantidade de alelos e de tamanho da amostra (N). Assim, algumas ou várias probabilidades condicionais podem ser estimadas. Cada arranjo genotípico pode ser delineado em uma tabela de contingência com dimensão $s \times s$, em que s é o número de alelos detectados no

loco amostrado. A hipótese de nulidade (H_0 : união ao acaso dos gametas) é rejeitada quando a soma de uma parte ou de todas as probabilidades (condicionais) ordenadas, de forma crescente, for menor do que um nível de significância α preestabelecido. Ressalta-se que a probabilidade condicional proveniente do arranjo genotípico da amostra original também deve ser inserida no ordenamento.

Para dois alelos (A e a), a probabilidade do arranjo genotípico, baseado no número de indivíduos AA, Aa e aa (N_{AA} , N_{Aa} , N_{aa}), é calculada considerando que a população está em equilíbrio logo, segue distribuição multinomial, descrita por:

$$P(N_{AA}, N_{Aa}, N_{aa}) = \frac{N!}{N_{AA}! N_{Aa}! N_{aa}!} (p^2)^{N_{AA}} (2pq)^{N_{Aa}} (q^2)^{N_{aa}}$$

em que p e q são as freqüências alélicas de A e a, respectivamente.

A probabilidade apresentada anteriormente estará condicionada à probabilidade das quantidades alélicas n_A e n_a observadas, definida por:

$$P(n_A, n_a) = \frac{(2N)!}{n_A! n_a!} (p)^{n_A} (q)^{n_a}$$

Com algumas operações algébricas, constrói-se a seguinte probabilidade condicional:

$$P(N_{AA}, N_{Aa}, N_{aa} | n_A, n_a) = \frac{P(N_{AA}, N_{Aa}, N_{aa})}{P(n_A, n_a)} = \frac{N! n_A! n_a! 2^{N_{Aa}}}{N_{AA}! N_{Aa}! N_{aa}! (2N)!}$$

Observa-se que a inferência sobre o equilíbrio não é influenciada pelas freqüências alélicas da população.

As probabilidades condicionais ainda podem ser calculadas em termos de um único alelo (A ou a) e o número de heterozigotos (N_{Aa}), conforme propôs Haldane (1954), da seguinte forma:

$$P(N_{Aa} | n_A) = \frac{N! n_A! (2N - n_A)! 2^{N_{Aa}}}{[(n_A - N_{Aa})/2]! N_{Aa}! [2N - (n_A + N_{Aa})/2]! (2N)!}$$

Assim todas as probabilidades são estimadas para todos os possíveis valores de N_{Aa} , da amostra de tamanho N e quantidade alélica n_A (ou n_a). Os valores de N_{Aa} são ordenados de acordo com suas probabilidades condicionais.

Para ilustrar o teste exato, serão considerados novamente os dados das populações P_1 e P_2 com 40 indivíduos de genótipos AA, Aa e aa, nas freqüências D, H e R, respectivamente.

Para o cálculo das $P(N_{Aa}|n_A)$ devem-se considerar as seguintes informações:

População	N	n_A
P_1	40	42
P_2	40	44

Tabela 4.8 - Arranjos genotípicos para a população P_1 , com 40 indivíduos, com resultados do teste exato de Fisher ($P(N_{Aa}|n_A)$ e $P(Acum)$), do teste de qui-quadrado (χ^2 e $P(\chi^2)$) e do teste do desvio da freqüência de homozigotos (D_A e z)

AA	Aa	aa	$P(N_{Aa} n_A)$	$P(Acum)$	χ^2	$P(\chi^2)$	D_A	z
21	0	19	0,0	0,0*	40,0000	0,00000*	0,2494	6,3245*
20	2	18	0,0	0,0*	32,3800	0,00000*	0,2244	5,6905*
2	38	0	0,0	0,0*	32,7437	0,00000*	-0,2256	-5,7222*
19	4	17	0,0	0,0*	25,5679	0,00001*	0,1994	5,0564*
3	36	1	0,0	0,0*	25,8895	0,00001*	-0,2006	-5,0881*
18	6	16	0,00001	0,00001*	19,5579	0,00001*	0,1744	4,4224*
4	34	2	0,00001	0,00002*	19,8393	0,00001*	-0,1756	-4,4541*
17	8	15	0,00011	0,00013*	14,3519	0,00015*	0,1494	3,7883*
5	32	3	0,00019	0,00032*	14,5931	0,00013*	-0,1506	-3,8200*
16	10	14	0,00130	0,00162*	9,94994	0,00161*	0,1244	3,1543*
6	30	4	0,00196	0,00358*	10,1509	0,00144*	-0,1256	-3,1860*
15	12	13	0,00880	0,01238*	6,3519	0,01173*	0,0994	2,5203*
7	28	5	0,01220	0,02457*	6,5127	0,01071*	-0,1006	-2,5520*
14	14	12	0,03770	0,06228	3,5580	0,05926	0,0744	1,8862
8	26	6	0,04802	0,11030	3,6786	0,05511	-0,0756	-1,9179
13	16	11	0,10557	0,21587	1,5680	0,21049	0,0494	1,2522
9	24	7	0,12386	0,33973	1,6484	0,19917	-0,0506	-1,2839
12	18	10	0,19734	0,53707	0,3821	0,53645	0,0244	0,61819
10	22	8	0,21366	0,75073	0,4223	0,51576	-0,0256	-0,64989
11	20	9	0,24927	1,00000	0,0002	0,98735	-0,0006	-0,01585

*Rejeita-se a H_0 a 5% de significância.

Tabela 4.9 - Arranjos genotípicos para a população P_2 , com 40 indivíduos, com resultados do teste exato de Fisher ($P(N_{Aa}|nA)$ e $P(\text{Acum})$), do teste de qui-quadrado (χ^2 e $P(\chi^2)$) e do teste do desvio da freqüência de homozigotos (D_A e z)

AA	Aa	aa	$P(N_{Aa} nA)$	$P(\text{Acum})$	χ^2	$P(\chi^2)$	D_A	z
22	0	18	0,0	0,0*	40,0000	0,0*	0,2475	6,3245*
21	2	17	0,0	0,0*	32,3273	0,0*	0,2225	5,6857*
4	36	0	0,0	0,0*	26,7769	0,0*	-0,2025	-5,1746*
20	4	16	0,0	0,0*	25,4709	0,0*	0,1975	5,0468*
5	34	1	0,0	0,0*	20,5734	0,00001*	-0,1775	-4,5357*
19	6	15	0,00001	0,00001*	19,4309	0,00001*	0,1725	4,4080*
6	32	2	0,00013	0,00014*	15,1862	0,00010*	-0,1525	-3,8969*
18	8	14	0,00013	0,00027*	14,2067	0,00016*	0,1475	3,7691*
17	10	13	0,00144	0,00170*	9,7999	0,00175*	0,1225	3,1303*
7	30	3	0,00151	0,00321*	10,6152	0,00112*	-0,1275	-3,2581*
16	12	12	0,00962	0,01283*	6,2075	0,01272*	0,0975	2,4914*
8	28	4	0,01026	0,02309*	6,8605	0,00881*	-0,1025	-2,6192 *
15	14	11	0,04059	0,06368	3,4323	0,06393	0,0725	1,8526
9	26	5	0,04308	0,10676	3,9220	0,04766*	-0,0775	-1,9804 *
14	16	10	0,11162	0,21838	1,4733	0,22482	0,0475	1,2138
10	24	6	0,11667	0,33505	1,7998	0,17973	-0,0525	-1,3415
13	18	9	0,20427	0,53932	0,3305	0,56532	0,0225	0,5749
11	22	7	0,2091	0,74842	0,4938	0,48223	-0,0275	-0,7027
12	20	8	0,25158	1,00000	0,0041	0,94906	-0,0025	-0,0638

*Rejeita-se a H_0 a 5% de significância.

Pelas Tabelas 4.8 e 4.9, verifica-se que o arranjo genotípico da população P_1 encontra-se entre aqueles com probabilidade acumulada inferior a 5%, de forma que se deve considerá-lo incompatível com a situação esperada de equilíbrio, prevista a partir do número de alelos A da população e da quantidade de heterozigotos observada. O nível crítico associado à hipótese é de 0,01238 (1,238%). Para a população P_2 , constata-se que o arranjo genotípico está entre aqueles esperados para uma população em equilíbrio de Hardy-Weinberg, com base na sua quantidade de alelo A e de heterozigotos observados, e seu nível crítico de probabilidade atinge valor de 0,2183 (21,83%), acima de 5%.

Generalizando para s alelos, a probabilidade de um arranjo com a quantidade genotípica $N_{kk'}$ de heterozigotos (para $k < k'$), condicionado às quantidades alélicas observadas (n_k 's) provenientes de s alelos, é expressa por:

$$P[(N_{kk'})|n_k] = \frac{N! 2^H \prod_k n_k!}{(2N)! \prod_{k \leq k'} (N_{kk'})!}$$

em que $H = \sum_k \sum_{k' \neq k} N_{kk'}$ é o número de indivíduos heterozigotos na população.

Na situação em que o tamanho amostral da população é grande e o loco investigado possui vários alelos, tem-se elevado número de arranjos genotípicos possíveis (ou tabelas de contingência) e as probabilidades condicionais são muito pequenas, embora a quantidade relevante seja a soma das probabilidades, a qual representa o valor de α . Para contornar esse problema, Guo e Thompson (1992) propuseram uma versão permutada para se obter o valor de α .

O processo baseia-se na desestruturação de todos os N genótipos (indivíduos) da população amostrada, de maneira que pares de alelos são tomados aleatoriamente até que os indivíduos sejam reconstituídos novamente. Esse processo é repetido inúmeras vezes. Sob equilíbrio de Hardy-Weinberg, os alelos estariam distribuídos independentemente nos genótipos; assim, um arranjo genotípico encontrado pelo processo de permutação (embaralhamento dos alelos) corresponderia a um dos possíveis arranjos a serem encontrados sob equilíbrio. O valor de α (área de rejeição) é dado pela proporção de arranjos genotípicos (ou tabelas de contingência) que tiveram as probabilidades condicionais menores ou iguais à probabilidade condicional do arranjo genotípico observado.

Os aplicativos computacionais têm estabelecido variações para esse processo de permutação, com destaque para os métodos MCMC (cadeia de Markov e Monte Carlo, do inglês, *Markov Chain Monte Carlo*). Em vez de enumerar todos os possíveis arranjos genotípicos, usa-se um processo de “caminhada” aleatória (cadeia de Markov) capaz de explorar eficientemente o espaço de todas as

tabelas de contingência, mantendo-se constante o número de alelos observados, ou seja, as marginais das tabelas de contingência.

Para iniciar o processo MCMC, os programas costumam pedir que seja definido o número de “dememorização”. Esse número correspondente a uma quantidade de passos que permite à cadeia de Markov executar o processo de construção de novas tabelas de contingência de maneira independente do estado inicial observado, mantendo-se, evidentemente, as quantidades alélicas originais. Raramente é necessário exceder 1.000 passos de “dememorização”. Outro procedimento a ser definido pelos usuários é o processo de *batching* (“loteamento”), que consiste em subdivisões da quantidade total de permutações a serem executadas. Esse processo permite atingir um ponto de convergência a ser checado automaticamente. Basicamente, o valor de p é calculado para cada *batch* (“lote”) e o desvio-padrão (ou coeficiente de variação) desses valores de p é comparado com o critério de convergência (β). O processo, como um todo, pára quando o desvio-padrão se encontra menor do que o valor de β especificado pelo usuário, ou seja, quando o critério de convergência foi atingido.

O teste exato apresenta algumas particularidades interessantes, como: (i) não fazer uso de distribuição assintótica; (ii) trata-se de uma distribuição de probabilidade independente de parâmetros sobre a hipótese de nulidade, um importante requerimento, que junto com (i) leva ao uso de distribuições condicionais particulares, como a apresentada por Haldane (1954), que independe das freqüências (alélicas e genotípicas) paramétricas; e (iii) usa as probabilidades de uma particular configuração (arranjos genotípicos) como um teste estatístico (ROUSSET; RAYMOND, 1995).

Levene (1949) também mostra que probabilidades condicionais podem ser obtidas por uma razão de verossimilhança entre a função de verossimilhança de uma particular amostra S e a soma das funções de verossimilhança de todas as possíveis amostras. Assim, a distribuição condicional de qualquer estatística sobre a hipótese de nulidade pode ser computada. Diferentes testes estatísticos definem diferentes ordenamentos das

possíveis amostras, porém o valor de α é definido igualmente como a soma de probabilidades exatas de amostras de ordem mais extrema, de modo que todos os testes são exatos (ROUSSET; RAYMOND, 1995).

Teste z

Pode-se considerar que, se não há equilíbrio de Hardy-Weinberg, alguma força seletiva atua contribuindo para o acréscimo de homozigotos ou de heterozigotos. Para dois locos, seriam esperadas as freqüências genotípicas listadas na Tabela 4.10.

Tabela 4.10 - Freqüência genotípica numa população em desequilíbrio

Genótipo	Freqüência observada	Freqüência esperada
AA	$\hat{P}_{AA} = N_{AA}/N$	$p^2 + D_A$
Aa	$\hat{P}_{Aa} = N_{Aa}/N$	$2pq - 2D_A$
aa	$\hat{P}_{aa} = N_{aa}/N$	$q^2 + D_A$

O valor de taxa de desequilíbrio (D_A) é dado por:

$$\hat{D}_A = \hat{P}_{AA} - \hat{p}^2$$

Sabendo que:

$$E(\hat{P}_{AA}) = P_{AA}$$

$$E(\hat{p}^2) = p^2 + \frac{1}{2N}(p + P_{AA} - 2p^2)$$

tem-se que:

$$E(\hat{D}_A) = D_A - \frac{1}{2N}[p(1-p) + D_A]$$

logo, \hat{D}_A é um estimador viesado de D_A , cujo viés decresce com o tamanho da amostra. Pode-se deduzir que:

$$\hat{V}(\hat{D}_A) = \frac{\hat{p}^2(1-\hat{p})^2}{2N}$$

É possível realizar o teste de hipótese $H_0: D_A = 0$ vs $H_a: D_A \neq 0$ por meio da estatística z, dada por:

$$z = \frac{\hat{D}_A}{\sqrt{V(\hat{D}_A)}}$$

Se o valor absoluto de z excede 1,96, rejeita-se a hipótese de que a proporção de homozigotos observada está em conformidade com o esperado sob hipótese de equilíbrio de Hardy-Weinberg. O teste apresentado é bilateral, de forma que, ao rejeitar H_0 , não é possível afirmar que exista excesso ou falta de homozigotos na população. Hipóteses alternativas à $H_0: D_A = 0$ podem ser $H_a: D_A > 0$, de forma que a rejeição de H_0 significará assumir excesso de homozigotos, ou $H_a: D_A < 0$, cuja rejeição de H_0 significará falta de homozigotos. Ambas serão significativas, a 5% de probabilidade, em teste unilateral quando o valor de z for maior que 1,64.

4.4.6. Equilíbrio com relação a genes ligados ao sexo

Quando se consideram genes ligados ao sexo, constata-se que o equilíbrio é alcançado na população quando as freqüências dos alelos nos diferentes sexos são iguais, o que não é atingido, normalmente, em uma única geração de acasalamento ao acaso.

Como ilustração, será admitida uma população na geração G_0 , numa espécie em que o macho é heterogamético (XY) e a fêmea é homogamética (XX). Se o gene A/a é ligado ao sexo, então são observados os seguintes genótipos:

Fêmeas	Obs.	Freq.	Exemplo		Machos	Obs.	Freq.	Exemplo	
X ^A X ^A	N _{AA}	D=N _{AA} /N	277	D=0,819	X ^A Y	n _A	n _A /N	311	0,881
X ^A X ^a	N _{Aa}	H=N _{Aa} /N	54	H=0,160	X ^a Y	n _a	n _a /N	42	0,119
X ^a X ^a	N _{aa}	R=N _{aa} /N	7	R=0,021					
Total		1	338					353	

Assim, tem-se:

a) Nas fêmeas

$$f(A) = p_f = D + \frac{1}{2} H = 0,819 + \frac{1}{2} 0,160 = 0,899$$

$$f(a) = q_f = R + \frac{1}{2} H = 0,021 + \frac{1}{2} 0,160 = 0,101$$

b) Nos machos

$$f(A) = p_m = n_A/N = 0,881$$

$$f(a) = q_m = n_a/N = 0,119$$

Considerando a população como um todo, têm-se as freqüências médias dadas por:

$$f(A) = \bar{p} = \frac{1}{3} p_m + \frac{2}{3} p_f = 0,893$$

$$f(a) = \bar{q} = \frac{1}{3} q_m + \frac{2}{3} q_f = 0,107$$

A diferença de freqüência alélica obtida entre os machos e entre as fêmeas é dada por:

$$d_0 = p_f - p_m$$

que, para o exemplo em consideração, é $d_0 = 0,899 - 0,881 = 0,018$.

A população, na geração G₁, resultante do acasalamento ao acaso será formada a partir dos encontros gaméticos ilustrados a seguir:

		Fêmeas	
Machos		$X^A (p_m)$	$X^a (q_m)$
$X^A (p_f)$		$X^A X^A$	$X^A X^a$
$X^a (q_f)$		$X^A X^a$	$X^a X^a$
Y		$X^A Y$	$X^a Y$

De forma que se tenha:

Fêmeas	Freqüência	Machos	Freqüência
$X^A X^A$	$D = p_m p_f$	$X^A Y$	p_f
$X^A X^a$	$H = p_m q_f + p_f q_m$	$X^a Y$	q_f
$X^a X^a$	$R = q_m q_f$		

Agora as freqüências alélicas serão dadas por:

a) Nas fêmeas

$$f(A) = p_f = D + \frac{1}{2}H = \frac{p_m + p_f}{2} = \frac{0,899 + 0,881}{2} = 0,890$$

$$f(a) = q_f = R + \frac{1}{2}H = \frac{q_m + q_f}{2} = \frac{0,101 + 0,119}{2} = 0,110$$

b) Nos machos

$$f(A) = p_m = p_f = 0,899$$

$$f(a) = q_m = q_f = 0,101$$

Novamente, considerando a população como um todo, têm-se as freqüências médias dadas por:

$$f(A) = \bar{p} = \frac{1}{3}p_m + \frac{2}{3}p_f = 0,893$$

$$f(a) = \bar{q} = \frac{1}{3}q_m + \frac{2}{3}q_f = 0,107$$

Constata-se que, na geração G_1 , a diferença entre as freqüências alélicas nos machos e nas fêmeas é reduzida, sendo dada por:

$$d_1 = p_f - p_m = \frac{p_m + p_f}{2} - p_m = -\frac{p_f - p_m}{2} = -\frac{1}{2}d_0$$

ou seja:

$$d_1 = 0,890 - 0,899 = -0,009$$

Nas gerações seguintes haverá maior aproximação para a condição de equilíbrio, de forma que o valor de d na n-ésima geração tenderá para zero e as freqüências p_m e p_f tornar-se-ão iguais, verificando-se as relações genotípicas descritas na Tabela 4.11.

Tabela 4.11 - Freqüência genotípica numa população em equilíbrio de Hardy-Weinberg para um gene ligado ao sexo

Machos	$X^A Y$	$X^a Y$
Freqüências	p	q
Fêmeas	$X^A X^A$	$X^A X^a$
Freqüências	p^2	$2pq$
		$X^a X^a$
		q^2

Como ilustração, será considerada uma população em que a freqüência do alelo a, entre as fêmeas, dada por q_f , é igual a 1, e entre os machos, dada por q_m , é igual a zero. Assim, verificam-se as seguintes mudanças nas freqüências alélicas após sucessivas gerações de acasalamento ao acaso:

Geração	q_m	q_f	$d = p_f - p_m = q_m - q_f$	$\bar{q} = \frac{1}{3} q_m + \frac{2}{3} q_f$
0	0	1	-1	0,667
1	1	0,5	0,5	0,667
2	0,5	0,75	-0,25	0,667
3	0,75	0,625	0,125	0,667
4	0,625	0,6875	-0,625	0,667
5	0,6875	0,65625	0,03125	0,667
6	0,65625	0,671875	-0,0015625	0,667

Considerando um gene deletério dominante ligado ao sexo (A), em que $f(A)$ é igual a p , espera-se observar maior freqüência de defeito entre as fêmeas ($p^2 + 2pq$

> p). Já para o caso de um gene deletério recessivo (b) ligado ao sexo, espera-se maior freqüência de defeitos entre os machos ($q > q^2$).

4.4.7 Equilíbrio com relação a mais de um gene

O estabelecimento das condições de equilíbrio depois de uma geração de acasalamento ao acaso é verdadeiro para todos os locos quando considerados isoladamente, mas não o é para genótipos referentes a dois ou mais locos considerados conjuntamente, o que conduz a um outro conceito importante na genética de populações, denominado de desequilíbrio gamético. O termo desequilíbrio gamético também tem sido referido como desequilíbrio de ligação, desequilíbrio de fase gamética e associação alélica (FLINT-GARCIA et al., 2003). Trata-se da associação não-aleatória de alelos de diferentes locos nos gametas. Em uma população panmítica, com locos segregando independentemente, na ausência de forças evolutivas, locos polimórficos estarão em equilíbrio gamético (FALCONER; MACKAY, 1996). O conhecimento do desequilíbrio permite elucidar fenômenos genéticos e evolutivos ocorridos ao longo de gerações nas populações ou espécies.

Como ilustração, considera-se uma população e as informações em relação a dois genes A/a e B/b. Os gametas produzidos pela população, na geração 1 tomada como referência, são dados por:

Gameta	Freqüência
AB	$P_{11(1)}$
Ab	$P_{10(1)}$
aB	$P_{01(1)}$
Ab	$P_{00(1)}$

O desequilíbrio de fase gamética é, nesta população, quantificado por meio de:

$$\Delta = P_{11(1)}P_{00(1)} - P_{10(1)}P_{00(1)}$$

Estimação do desequilíbrio

A freqüência gamética em qualquer geração pode ser estimada conhecendo-se as freqüências alélicas e o desequilíbrio de ligação em qualquer geração (Δ_n) pode ser dado por:

$$\Delta_n = P_{11(n)} P_{00(n)} - P_{01(n)} P_{10(n)}$$

Como:

$$P_{11(n)} + P_{00(n)} + P_{01(n)} + P_{10(n)} = 1$$

$$p_A = P_{11} + P_{10}$$

e

$$p_B = P_{11} + P_{01}$$

tem-se:

$$\Delta_n = P_{11(n)} [1 - P_{11(n)} - P_{01(n)} - P_{10(n)}] - P_{01(n)} P_{10(n)}$$

$$\Delta_n = P_{11(n)} - p_A p_B$$

Conseqüentemente:

$$P_{11(n)} = p_A p_B + \Delta_n$$

$$P_{10(n)} = p_A q_B - \Delta_n$$

$$P_{01(n)} = q_A p_B - \Delta_n$$

$$P_{00(n)} = q_A q_B + \Delta_n$$

Assim, se forem considerados dois alelos para cada um de dois genes codominantes ($f(A) = p_A$; $f(a) = q_A$; $f(B) = p_B$; $f(b) = q_B$), ocorrem as seguintes expectativas de freqüências genotípicas:

Genótipo	Freqüência	Esperado	Obs.
AABB	P_{11}^2	$(p_A p_B + \Delta)^2$	n_1
AABb	$2P_{11} P_{10}$	$2(p_A p_B + \Delta)(p_A q_b - \Delta)$	n_2
AAbb	P_{10}^2	$(p_A q_b - \Delta)^2$	n_3
AaBB	$2P_{11} P_{01}$	$2(p_A p_B + \Delta)(q_a p_B - \Delta)$	n_6
AaBb	$2(P_{10} P_{01} + P_{11} P_{00})$	$2[(p_A q_b - \Delta)(q_a p_B - \Delta) + (p_A p_B + \Delta)(q_a q_b + \Delta)]$	n_5
Aabb	$2P_{10} P_{00}$	$2(p_A q_b - \Delta)(q_a q_b + \Delta)$	n_6
aaBB	P_{01}^2	$(q_a p_B - \Delta)^2$	n_7
aaBb	$2P_{01} P_{00}$	$2(q_a p_B - \Delta)(q_a q_b + \Delta)$	n_8
aabb	P_{00}^2	$(q_a q_b + \Delta)^2$	n_9

Uma vez que os valores de p_A , q_a , p_B e q_b , são facilmente obtidos em uma população, torna-se possível estimar o valor de Δ por método de máxima verossimilhança, admitindo que o número de ocorrência das classes genotípicas segue distribuição multinomial, dada por:

$$L(p_A, q_a, p_B, q_b, \Delta; n_i) = \frac{N!}{n_1! n_2! \dots n_9!} p_1^{n_1} p_2^{n_2} \dots p_9^{n_9}$$

em que p_1, p_2, \dots, p_9 são as freqüências observadas das classes genotípicas.

O estimador de verossimilhança de Δ é dado pelo diferencial da função $L(p_A, q_a, p_B, q_b, \Delta; n_i)$, sendo:

$$\frac{\partial L(p_A, q_a, p_B, q_b, \Delta; n_i)}{\partial \Delta}$$

A hipótese de nulidade é testada por uma razão de verossimilhança, definida por:

$$LOD = \log_{10} \left(\frac{L_1}{L_0} \right)$$

em que:

L_0 : função de verossimilhança considerando a hipótese de nulidade $\Delta = 0$; e

L_1 : função de verossimilhança considerando a hipótese alternativa $\Delta \neq 0$, ou seja, levando-se em conta o valor de Δ estimado.

Se $L_1 > L_0$, o LOD escore é positivo. Conclui-se que os locos estão em desequilíbrio quando o LOD é maior que 3, ou seja, probabilidade de 1.000 para 1.

Como exemplo, consideram-se dois locos A/a e B/b e os seguintes valores observados das classes genotípicas:

Genótipo	Esperado	Num. obs.
AABB	$(p_A p_B + \Delta)^2$	$n_1 = 13$
AABb	$2(p_A p_B + \Delta)(p_A q_b - \Delta)$	$n_2 = 4$
AAAb	$(p_A q_b - \Delta)^2$	$n_3 = 2$
AaBB	$2(p_A p_B + \Delta)(q_a p_B - \Delta)$	$n_6 = 12$
AaBb	$2[(p_A q_b - \Delta)(q_a p_B - \Delta) + (p_A p_B + \Delta)(q_a q_b + \Delta)]$	$n_5 = 8$
Ab/ab	$2(p_A q_b - \Delta)(q_a q_b + \Delta)$	$n_6 = 3$
aB/aB	$(q_a p_B - \Delta)^2$	$n_7 = 4$
aaBb	$2(q_a p_B - \Delta)(q_a q_b + \Delta)$	$n_8 = 3$
aabb	$(q_a q_b + \Delta)^2$	$n_9 = 1$

Para esse par de locos, o valor máximo do coeficiente de desequilíbrio será de 16,5% e o valor estimado de Δ é 2,5%, conforme ilustrado graficamente na Figura 4.1.

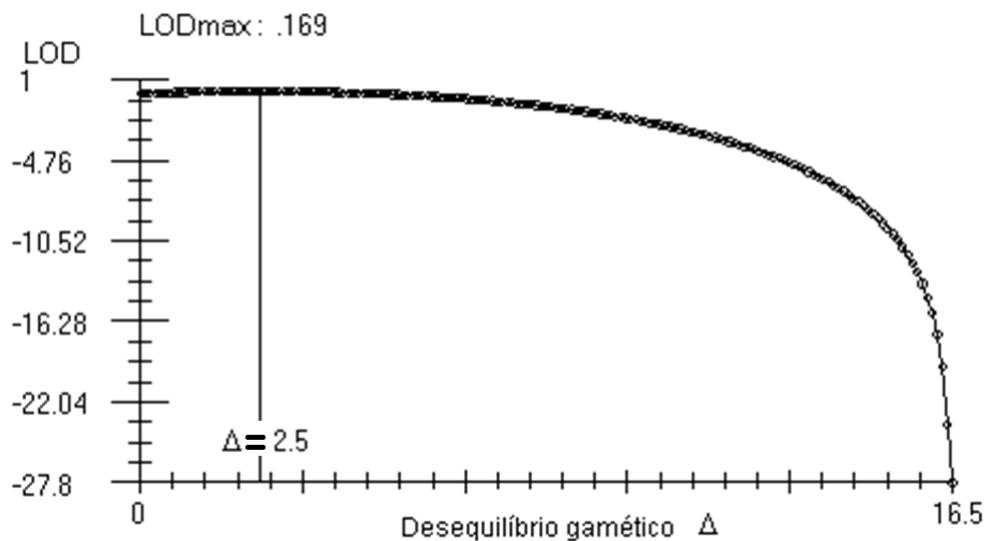


Figura 4.1 - Representação gráfica do comportamento do coeficiente de desequilíbrio gamético (Δ) e os respectivos valores de LOD escore, do exemplo anterior. Observe que o valor de LOD máximo ($LOD_{\max} = 0,169$) é obtido quando se tem $\Delta = 2,5 \%$.

Os valores do desequilíbrio estão compreendidos nos seguintes intervalos:

$$\Delta_{\min} : \text{maior } [(-p_A p_B); (-q_a q_b)]$$

$$\Delta_{\max} : \text{menor } [(p_A q_b); (q_a p_B)]$$

Para o exemplo em consideração, tem-se:

$$p_A = 0,61 \text{ e } q_a = 0,39$$

$$p_B = 0,73 \text{ e } q_b = 0,27$$

logo, tem-se:

$$\begin{aligned} \Delta_{\max} &= \text{menor } [(p_A q_b); (q_a p_B)] = \text{menor}[(0,61)(0,27);(0,39)(0,73)] \\ &= \text{menor}(0,1647 ; 0,2847) = 0,1647 \end{aligned}$$

Se a causa do desequilíbrio gamético fosse unicamente a ligação fatorial, a distância entre os locos poderia ser estimada por meio da expressão:

$$d = 50 \left(1 - \frac{\Delta}{\Delta_{\max}} \right)$$

Para o exemplo, tem-se:

$$d = 50 \left(1 - \frac{2,5}{16,5} \right) = 42,42 \text{ cMorgans}$$

Uma forma aproximada de obter a medida de desequilíbrio é por meio de:

$$\Delta = \frac{[(n_{AABB} + n_{aabb}) - (n_{AAbb} + n_{aaBB})]/N - [(p^2r^2 + q^2s^2) - (p^2s^2 + q^2r^2)]}{2}$$

$$\text{Sendo } p = p_A \quad q = p_a \quad r = p_B \quad s = q_b$$

Assim, no exemplo tem-se:

$$\Delta = \frac{[(13+1) - (2+4)]/50 - [(0,2094) - (0,1081)]}{2} = 0,0293 = 2,93\%$$

Outras medidas relativas de desequilíbrio apresentadas na literatura são:

Estatística r^2

Uma medida do desequilíbrio, denotada r^2 , é fornecida por:

$$r^2 = \frac{\Delta^2}{p_A q_a p_B q_b}$$

de forma que, no exemplo, tem-se:

$$r^2 = \frac{\Delta^2}{p_A q_a p_B q_b} = \frac{(0,025)^2}{0,61 \times 0,39 \times 0,73 \times 0,27} = 1,33\%$$

O r^2 representa o quadrado da correlação entre valores gaméticos considerando a informação de dois genes A/a e B/b, conforme ilustrado a seguir:

A/a	B/b	Valor gamético (X)	Valor gamético (Y)	Frequência
A	B	1	1	$p_A p_B + \Delta$
A	b	1	0	$p_A q_b - \Delta$
a	B	0	1	$q_a p_B - \Delta$
a	b	0	0	$q_a q_b + \Delta$

Assim, tem-se

$$V(X) = p_A q_a$$

$$V(Y) = p_B q_b$$

$$\text{Cov}(X, Y) = \Delta$$

logo,

$$r^2 = \frac{\Delta^2}{p_A q_a p_B q_b}$$

Estatística D'

A estatística D' é calculada como:

$$|D'| = \frac{\Delta}{\Delta_{\text{máximo}}}$$

Ou seja,

$$|D'| = \frac{\Delta}{\text{menor } [(p_A q_b); (q_a p_B)]} \quad \text{para } \Delta > 0$$

$$|D'| = \frac{\Delta}{\text{maior } [(-p_A p_B); (-q_a q_B)]} \quad \text{para } \Delta < 0$$

de forma que, no exemplo, tem-se:

$$|D'| = \frac{\Delta}{\text{menor } [(p_A q_b); (q_a p_B)]} = \frac{0,25}{\text{menor } [(0,61 \times 0,27); (0,39 \times 0,73)]}$$

$$|D'| = \frac{0,25}{\text{menor } [0,1647; 0,2847]} = \frac{0,25}{0,1647} = 1,52\%$$

As estatísticas r^2 e D' refletem diferentes aspectos do desequilíbrio de ligação e comportam-se diferentemente sob condições variadas.

Numa população (experimental) F_2 , em que todas as classes genotípicas estão representadas, o desequilíbrio será máximo e expresso por:

$$\Delta = \frac{1 - 2d}{4}$$

Se $d = 0$, o valor de Δ será de 25%, e se $d = 0,5$ (50 cMorgans), o desequilíbrio de fase gamética provocado pela ligação fatorial será nulo.

Desequilíbrio após gerações de acasalamento ao acaso

O equilíbrio de fase gamética não é obtido em apenas uma geração de acasalamento ao acaso, mas sua intensidade é reduzida a cada geração, conforme será verificado a seguir. Assim, pode-se considerar uma determinada geração em que os genótipos encontrados e suas respectivas freqüências são dados a seguir. Neste exemplo, consideraram-se dois genes ligados cuja porcentagem de recombinação é igual a d.

Genótipos	Freqüência
AB/AB	$P_{11(1)}^2$
AB/Ab	$2 P_{11(1)} P_{10(1)}$
AB/aB	$2 P_{11(1)} P_{01(1)}$
AB/ab	$2 P_{11(1)} P_{00(1)}$
Ab/Ab	$P_{10(1)}^2$
Ab/aB	$2 P_{10(1)} P_{01(1)}$
Ab/ab	$2 P_{10(1)} P_{00(1)}$
aB/aB	$P_{01(1)}^2$
aB/ab	$2 P_{01(1)} P_{00(1)}$
ab/ab	$P_{00(1)}^2$

Os gametas produzidos pela população são mostrados a seguir, omitindo-se o indexador que caracteriza a geração analisada:

Genótipo	Freqüência	Gametas			
		AB	Ab	aB	ab
AB/AB	P_{11}^2	P_{11}^2			
AB/Ab	$2 P_{11} P_{10}$	$P_{11} P_{10}$	$P_{11} P_{10}$		
AB/aB	$2 P_{11} P_{01}$	$P_{11} P_{01}$		$P_{11} P_{01}$	
AB/ab	$2 P_{11} P_{00}$	$(1-d)P_{11} P_{00}$	$dP_{11} P_{00}$	$dP_{11} P_{00}$	$(1-d)P_{11} P_{00}$
Ab/Ab	P_{10}^2		P_{10}^2		
Ab/aB	$2 P_{10} P_{01}$	$dP_{10} P_{01}$	$(1-d)P_{10} P_{01}$	$(1-d)P_{10} P_{01}$	$dP_{10} P_{01}$
Ab/ab	$2 P_{10} P_{00}$		$P_{10} P_{00}$		$P_{10} P_{00}$
aB/aB	P_{01}^2			P_{01}^2	
aB/ab	$2 P_{01} P_{00}$			$P_{01} P_{00}$	$P_{01} P_{00}$
ab/ab	P_{00}^2				P_{00}^2

As freqüências dos gametas produzidos seriam:

$$\begin{aligned}
f(AB) &= P_{11(2)} = P_{11}^2 + P_{11}P_{10} + P_{11}P_{01} + (1-d)P_{11}P_{00} + dP_{10}P_{01} \\
&= P_{11}(P_{11} + P_{10} + P_{10} + P_{00}) - d(P_{11}P_{00} - P_{10}P_{01}) \\
&= P_{11(1)} - d\Delta_1
\end{aligned}$$

de maneira análoga, tem-se:

$$f(AB) = P_{11(2)} = P_{11(1)} - d\Delta_1$$

$$f(Ab) = P_{10(2)} = P_{10(1)} + d\Delta_1$$

$$f(aB) = P_{01(2)} = P_{01(1)} + d\Delta_1$$

$$f(ab) = P_{00(2)} = P_{00(1)} - d\Delta_1$$

Com as freqüências gaméticas conhecidas, obtém-se as freqüências genotípicas da descendência obtida por acasalamento ao acaso, descritas a seguir:

Genótipo	Freqüência
AB/AB	$P_{11(2)}^2$
AB/Ab	$2P_{11(2)}P_{10(2)}$
AB/aB	$2P_{11(2)}P_{01(2)}$
AB/ab	$2P_{11(2)}P_{00(2)}$
Ab/Ab	$P_{10(2)}^2$
Ab/aB	$2P_{10(2)}P_{01(2)}$
Ab/ab	$2P_{10(2)}P_{00(2)}$
aB/aB	$P_{01(2)}^2$
aB/ab	$2P_{01(2)}P_{00(2)}$
ab/ab	$P_{00(2)}^2$

Com o acasalamento ao acaso as freqüências alélicas permanecem, mas após uma geração de acasalamento ao acaso houve maior quebra do bloco gênico, de forma que a freqüência dos gametas paternais diminuiu com o aumento relativo das formas recombinantes.

Nesse exemplo, pode ser verificado que o desequilíbrio da fase gamética é dado por:

$$\Delta_1 = P_{11(1)}P_{00(1)} - P_{10(1)}P_{01(1)}$$

e

$$\begin{aligned}\Delta_2 &= P_{11(2)}P_{00(2)} - P_{10(2)}P_{01(2)} \\ &= (P_{11(1)} - d\Delta_1)(P_{00(1)} - d\Delta_1) - (P_{10(1)} - d\Delta_1)(P_{01(1)} - d\Delta_1) \\ &= (1 - d)\Delta_1\end{aligned}$$

Pela expressão anterior, constata-se que o valor de d reduz-se à metade a cada geração de acasalamento ao acaso para dois genes independentes. Contudo, se os genes estiverem ligados, a velocidade de decréscimo do valor de d será reduzida e o equilíbrio só será atingido com número maior de gerações de acasalamento ao acaso.

De forma generalizada, tem-se:

$$\Delta_n = (1 - d)^{n-1} \Delta_1$$

O desequilíbrio é dependente da taxa de recombinação (d) e quanto maior a taxa de recombinação, mais rápida será a aproximação do equilíbrio entre os dois locos. Se os genes são independentes, tem-se $d = 0,5$ e, portanto:

$$\Delta_n = (1/2)^{n-1} \Delta_1$$

Se o valor de n é suficientemente grande, tem-se:

$$\Delta_n = (1 - d)^{n-1} \Delta_1 = 0$$

E, nesta situação, atinge-se o equilíbrio de ligação, de forma que se tenha: $P_{11(n)}$

$$= p_A p_B + \Delta_n = p_A p_B$$

$$P_{10(n)} = p_A q_B - \Delta_n = p_A q_B$$

$$P_{01(n)} = q_A p_B - \Delta_n = q_A p_B$$

$$P_{00(n)} = q_A q_B + \Delta_n = q_A q_B$$

Mecanismos que promovem desequilíbrio gamético são a seleção, a ligação factorial e deriva genética, aleatória e não-aleatória (RIDLEY, 2006). Se a seleção favorece indivíduos com combinações particulares de alelos, então ela produz desequilíbrio gamético. Em se tratando de locos ligados, um número maior de gerações é necessário para que a recombinação realize a sua função de tornar as

associações genéticas aleatórias. Locos fracamente ligados não irão apresentar desequilíbrio de ligação por muito tempo. Na ausência de seleção e em uma população infinita e de cruzamentos aleatórios, a quantidade de desequilíbrio de ligação sofre uma queda exponencial a uma taxa igual à de recombinação entre dois locos.

Processos aleatórios possuem a propriedade interessante de serem capazes de causar desequilíbrio de ligação persistente, não apenas transitório. Se a deriva genética produz excesso de um determinado gameta (haplótipo) em uma geração, o desequilíbrio gamético terá aparecido. Acrescenta-se que isso pode ser verdadeiro considerando todos os possíveis haplótipos: a amostragem aleatória que produz excesso de qualquer tipo gamético irá perturbar o estado de equilíbrio (RIDLEY, 2006).

Essas associações persistem por mais tempo em locos fortemente ligados, de modo que, quanto mais elevada for a taxa de recombinação, mais rápida será a destruição da associação. Entretanto, como a taxa de recombinação entre dois locos diminui, o tempo em que os alelos podem estar associados entre si de forma não-aleatória aumenta (RIDLEY, 2006).

Outro fator gerador de desequilíbrio gamético diz respeito aos cruzamentos não-aleatórios ou preferenciais. Geralmente o desequilíbrio reduz-se mais rapidamente em espécies alógamas, quando comparado com espécies de autofecundação (NORDBORG, 2000). Cruzamentos não-aleatórios proporcionam aumentos (ou diminuição) de certos haplótipos, fazendo com que haja freqüência em excesso (ou deficiência) em relação a situações de cruzamentos aleatórios. A alta homozigosidade dos genes, em espécies autógamas, implica que a recombinação raramente resultará em novos haplótipos que ainda não estão presentes nos parentais. A predominância de autofecundação tende a retardar a proximidade do equilíbrio gamético, porque para atingi-lo são necessárias recombinações entre duplos heterozigotos, que são raros nas populações autógamas (HARTL; CLARK, 1997).

Outro fator, como fluxo gênico entre indivíduos de populações geneticamente distintas seguido por intercruzamentos, resulta na introdução de diferentes informações de genéticas de ancestrais e diferentes freqüências alélicas. Freqüentemente, o resultado do desequilíbrio se estende a sítios não ligados, mesmo em diferentes cromossomos, mas que são quebrados rapidamente com o processo de acasalamento ao acaso (PITCHARD; ROSENBERG, 1999).

4.5. Acasalamentos

Conhecer o sistema de acasalamento de uma população ou espécie permite delinear estratégias de conservação, melhoramento e manejo sustentado. De acordo com Liu (1997), as populações obtidas por cruzamentos controlados entre genitores selecionados podem ser consideradas como autênticas populações de melhoramento ou experimentais. Nelas, a estrutura e variabilidade genética são previsíveis, se não conhecidas. Assim, o melhorista tem a possibilidade de predizer mudanças em magnitude e sentido desejado e formular hipóteses acerca do seu comportamento genético. Enquadram-se aí as populações de híbridos F_1 , gerações avançadas F_n (ou S_n), retrocruzamentos, entre outras. Estas populações, ao serem obtidas, passam por seleções e avaliações, até que se tornem de uso comercial.

As populações naturais são aquelas geradas por acasalamentos ocorridos naturalmente, ou seja, sem controle artificial. Trata-se de unidades sobre as quais incide o manejo para a conservação e utilização dos recursos naturais, além de fonte de germoplasma para os programas de melhoramento genético (ROBINSON, 1998). Algumas espécies podem ser geradas por acasalamentos preferenciais ou mesmo por completa autofecundação em vez de unicamente pelo acasalamento ao acaso. Exemplos de populações naturais são dados por famílias de meios-irmãos, populações mistas cujos indivíduos são derivados de polinização cruzada e autopolinização e populações com sobreposição de gerações (LIU, 1997).

Os cruzamentos aqui considerados são:

- a) Cruzamentos direcionados.

- b) Autofecundações.
- c) Acasalamento ao acaso.

Como ilustração serão consideradas duas populações P_1 e P_2 , com a seguinte constituição genotípica:

Genótipo	P_1	P_2
AA	20	50
Aa	30	50
aa	50	0

Os seguintes cruzamentos podem ser ilustrados:

- a) Hibridação entre P_1 e P_2

A hibridação, ou cruzamento direcionado, tem sido utilizada rotineiramente no melhoramento genético com a intenção de reunir genes favoráveis presentes em ambos os genitores. Preocupa-se em cruzar materiais genéticos superiores e que exibam diversidade genética de forma que o híbrido obtido manifeste vigor, ou heterose, e que as populações segregantes avançadas manifestem ampla variabilidade a ser explorada por técnica seletiva. Em muitos casos são encontrados na população F2, e em outras segregantes, indivíduos transgressivos cujo valor fenotípico está além, ou aquém, dos limites estabelecidos pelos genitores.

Considera-se, inicialmente, a freqüência genotípica em cada população, dada a seguir:

Freqüência	P_1	P_2
$D = f(AA)$	0,20	0,50
$H = f(Aa)$	0,30	0,50
$R = f(aa)$	0,50	0,0

Neste caso, são envolvidos os seguintes acasalamentos:

$P_1 \times P_2$	Probabilidade	Descendência		
		AA	Aa	aa
AA x AA	0,20 x 0,50	0,10		
AA x Aa	0,20 x 0,50	0,05	0,05	
Aa x AA	0,30 x 0,50	0,075	0,075	
Aa x Aa	0,30 x 0,50	0,0375	0,075	0,0375
aa x AA	0,50 x 0,50		0,25	
aa x Aa	0,50 x 0,50		0,125	0,125
Total	1	0,2625	0,575	0,1625

Assim, a população híbrida terá a seguinte constituição:

Genótipo	Freqüência
AA	0,2625
Aa	0,5750
aa	0,1625

Com conhecimentos adquiridos em genética de populações, a freqüência dos genótipos na população híbrida poderá ser facilmente estimada considerando as expressões:

$$f(A) = p = D + \frac{1}{2}H$$

e

$$f(a) = q = R + \frac{1}{2}H$$

Assim, para as populações P_1 e P_2 , tem-se:

$$\text{Para } P_1: f(A) = p_1 = 0,35 \quad \text{e} \quad f(a) = q_1 = 0,65$$

$$\text{Para } P_2: f(A) = p_2 = 0,75 \quad \text{e} \quad f(a) = q_2 = 0,25$$

logo:

Genótipos em $F_1 = P_1 \times P_2$	Esperado	Freqüência
AA	$p_1 p_2$	$D' = 0,2625$
Aa	$p_1 q_2 + p_2 q_1$	$H' = 0,5750$
aa	$q_1 q_2$	$R' = 0,1625$

Algumas informações importantes relativas à população híbrida (F_1) são:

- *Freqüência alélica*

É interessante observar que a freqüência alélica da população híbrida é igual à media das freqüências alélicas das populações genitoras, ou seja:

$$f(A) = p_h = D' + \frac{1}{2}H' = p_1 p_2 + \frac{1}{2}(p_1 q_2 + p_2 q_1) = \frac{p_1 + p_2}{2}$$

e

$$f(a) = q_h = R' + \frac{1}{2}H' = q_1 q_2 + \frac{1}{2}(p_1 q_2 + p_2 q_1) = \frac{q_1 + q_2}{2}$$

No exemplo, tem-se:

$$f(A) = p_h = 0,55 \quad e \quad f(a) = q_h = 0,45$$

- *Frequênciade heterozigotos e Endogamia*

A frequência de heterozigotos observada é dada por:

$$H_{obs} = p_1 q_2 + p_2 q_1$$

sendo $\Delta = p_2 - p_1$ tem-se:

$$H_{obs} = p_1 q_2 + p_2 q_1 = p_1(q_1 - \Delta) + q_1(p_1 - \Delta) = 2p_1 q_1 + \Delta(q_1 - p_1)$$

Se a população estivesse em equilíbrio de Hardy-Weinberg, a frequência esperada seria dada por:

$$H_{esp.EHW} = 2p_h q_h = 2\left(\frac{p_1 + p_2}{2}\right)\left(\frac{q_1 + q_2}{2}\right) = 2p_1 q_1 + \Delta(q_1 - p_1) - (\frac{1}{2})\Delta^2$$

Assim, o coeficiente de endogamia (F) esperado pode ser calculado por meio da relação entre as frequências de heterozigotos, tendo-se:

$$F = \frac{H_{esp.EHW} - H_{obs}}{H_{esp.EHW}} = \frac{-(1/2)\Delta^2}{H_{esp.EHW}}$$

Observa-se que este coeficiente representa a fixação e a diversidade entre as populações genitores. Este valor será negativo e alto quando as populações envolvidas na hibridação apresentar grande diferença entre as frequências alélicas dos locos considerados.

- *Média de uma população híbrida*

O cruzamento entre duas populações em equilíbrio de Hardy-Weinberg produzirá uma população híbrida, cujo valor esperado da média será:

$$\mu_H = \frac{\mu_{P1} + \mu_{P2}}{2} + h$$

sendo h o valor da heterose manifestada no cruzamento, dada por:

$$h = \mu_H - \frac{\mu_{P1} + \mu_{P2}}{2} = d(p_1 - p_2)^2$$

Pela expressão anterior, fica evidente que a heterose manifestada em populações híbridas é função direta do valor genotípico do heterozigoto (d), expresso pelo grau médio de dominância, e da diferença de freqüência gênica, ou da diversidade genética, entre as populações intercruzadas. Também é compreensível que a heterose será máxima quando um alelo for fixo em uma população e o outro na outra população; se as populações não diferirem em freqüência gênica, não haverá heterose. Com base nessa expressão, têm sido fundamentados muitos estudos de avaliação da diversidade genética entre populações, à procura daqueles de bons desempenhos e que exibam diversidade, recomendando-se seus cruzamentos com a expectativa de que a população híbrida manifeste heterose e que as populações dela derivada apresentem variabilidade e indivíduos transgressivos. Segregantes transgressivos são aqueles manifestados

em populações segregantes, cujos valores fenotípicos superam os limites estabelecidos pelas populações genitoras.

- Variabilidade em uma população híbrida

Na Figura 4.2 são apresentadas curvas que descrevem o comportamento da variância genotípica, aditiva e devida aos desvios da dominância em função da freqüência genotípica em três diferentes situações de dominância (d/a , em que d e a são os valores genotípicos do heterozigoto Aa e homozigoto AA , respectivamente).

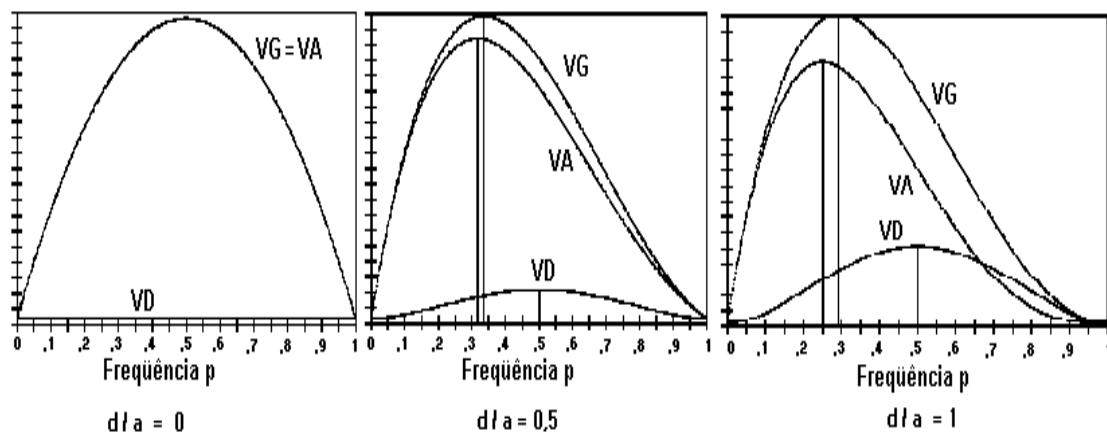


Figura 4.2 - Comportamento da variância genotípica (VG), aditiva (VA) e devida aos desvios da dominância (VD) em função da freqüência gênica p , em três diferentes situações de relação de dominância (d/a).

Fica evidente a relação quadrática entre os valores de variância e os da freqüência gênica da população. Assim, como a freqüência gênica da população híbrida é intermediária, tem-se que a variabilidade nesta população também será intermediária à variância destas populações genitoras quando ambas apresentarem freqüência gênica acima ou abaixo do máximo esperado. Entretanto, o cruzamento entre populações genitoras com grande diferença de freqüência gênica poderá proporcionar população híbrida cuja variabilidade poderá ser bem maior que a apresentada em qualquer uma de suas populações genitoras.

b) Autofecundação em P1

A autofecundação é um processo de propagação sexuada, que se verifica naturalmente em muitas espécies vegetais autógamas, que contam com os aparelhos reprodutores, masculino e feminino, numa mesma planta. Também é utilizado em programas de melhoramento, com vistas à obtenção de linhagens homozigóticas, para a produção de híbridos heteróticos a partir de seus intercruzamentos.

Uma maneira de visualizar as consequências da autofecundação em sucessivas gerações é apresentada na Figura 4.3. Nesse esquema, considera-se que cada indivíduo autofecundado deixa quatro descendentes para a próxima geração. O estabelecimento um número fixo de descendentes torna possível quantificar o total de heterozigotos e homozigotos em cada geração.

Geração	Genótipos			Freqüência	
	AA	Aa	aa	Heterozigotos	Homozigotos
0		1		100	0
1	1	2	1	$1/2 = 50$	$1/2 = 50$
2	6	4	6	$1/4 = 25$	$3/4 = 75$
3	28	8	28	$1/8 = 12,5$	$7/8 = 87,5$
4	120	16	120	$1/16 = 6,25$	$15/16 = 93,75$
5	496	32	496	$1/32 = 3,125$	$31/32 = 96,875$
6	2016	64	2016	$1/64 = 1,5625$	$63/64 = 98,4375$
7	8128	128	8128	$1/128 = 0,78125$	$127/128 = 99,21875$

Figura 4.3. Freqüência de homozigotos e heterozigotos em sucessivas autofecundações realizadas em populações derivadas da F_1 com 100% de heterozigotos.

No esquema apresentado, constata-se que há aumento da freqüência de homozigotos a cada geração, enquanto a freqüência de heterozigotos reduz-se à metade. Após sete gerações de autofecundações sucessivas em uma população originalmente constituída por apenas heterozigotos, têm-se mais de 99% de homozigotos.

A freqüência de heterozigotos a cada geração de autofecundação pode ser estimada por meio da seguinte expressão:

$$f(H_t) = \left(\frac{1}{2}\right)^t f(H_0)$$

em que:

$f(H_t)$: freqüência de heterozigotos após t gerações de autofecundações; e

$f(H_0)$: freqüência inicial de heterozigotos.

A freqüência de homozigotos ($f(D)$ e $f(R)$ para os homozigotos dominantes e recessivos, respectivamente) na t-ésima geração de autofecundação também pode ser estimada por meio de:

$$f(D_t) = f(D_0) + \frac{1}{2} \Delta$$

e

$$f(R_n) = f(R_0) + \frac{1}{2} \Delta$$

sendo $\Delta = f(H_0) - f(H_t)$ a perda na freqüência de heterozigotos na t-ésima geração.

A velocidade do acréscimo da freqüência de homozigotos e do decréscimo de heterozigotos, com sucessivas gerações de autofecundações, é ilustrada na Figura 4.4. Novamente se verifica nesse gráfico que após sete gerações de autofecundações há praticamente 100% de homozigotos (e freqüência nula de heterozigotos) na população.

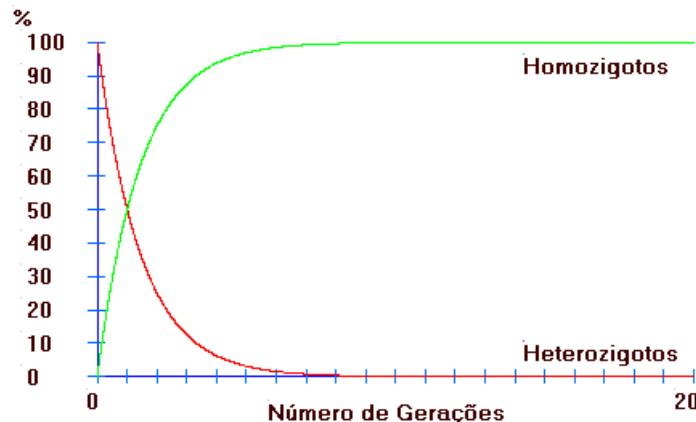


Figura 4.4 - Alteração na freqüência de heterozigotos e homozigotos após sucessivas gerações de autofecundação.

No exemplo, considerando a população P_1 , são envolvidas as seguintes autofecundações:

Genótipos de P_1	Probabilidade	Descendência		
		AA	Aa	aa
AA	0,20	0,20		
Aa	0,30	0,075	0,15	0,075
aa	0,50			0,50
Total	1	0,275	0,15	0,575

Essa população descendente é facilmente predita sabendo-se que a cada geração de autofecundação a freqüência de heterozigotos reduz-se à metade. Assim, a freqüência que originalmente é 0,30 passa para 0,15. O restante (0,15) é distribuído equitativamente entre os homozigotos. Logo, a freqüência de AA torna-se $0,20 + \frac{1}{2}(0,15) = 0,275$, e a de aa, $0,50 + \frac{1}{2}(0,15) = 0,575$.

Assim, a população autofecundada terá a seguinte constituição:

Genótipo	Freqüência
AA	0,275
Aa	0,150
aa	0,575

Se uma população de heterozigotos (100% Aa) é autofecundada, a descendência (F_1) será formada de 50% de heterozigotos e 50% de homozigotos (AA + aa). Assim, em apenas uma geração de autofecundação a freqüência de heterozigotos reduz-se à metade. Esse fato se verifica nas gerações seguintes, ou seja, se a F_1 for novamente autofecundada, ter-se-á:

Genótipos da F_1 que se autofecundam	Probabilidade da autofecundação	Descendência		
		AA	Aa	aa
AA	0,25	0,25		
Aa	0,50	0,125	0,25	0,125
aa	0,25			0,25
Total	1	0,375	0,25	0,375

Verifica-se que, agora, a freqüência de heterozigotos, reduzida à metade, é de 25%. Os 25% restantes são distribuídos eqüitativamente para os respectivos homozigotos AA e aa.

Algumas informações importantes relativas à população obtida por autofecundação:

- *Freqüência alélica*

As freqüências genotípicas em uma população antes e após sucessivas gerações de autofecundação são apresentadas na Tabela 4.12.

Tabela 4.12 - Freqüências genotípicas em população antes e após n gerações de autofecundação

Genótipo	Inicial	Após autofecundações
AA	D_0	$D_n = D_0 + \frac{1}{2}\Delta$
Aa	H_0	$H_n = (\frac{1}{2})^n H_0$
aa	R_0	$R_n = R_0 + \frac{1}{2}\Delta$

$$\Delta = H_0 - H_n$$

Assim, constata-se que, originalmente, a freqüência alélica é dada por:

$$f(A) = p_0 = D_0 + \frac{1}{2}H_0 \quad \text{e} \quad f(a) = p_0 = R_0 + \frac{1}{2}H_0$$

Após as autofecundações, obtém-se:

$$f(A) = p_n = D_n + \frac{1}{2}H_n = \left(D_0 + \frac{1}{2}\Delta \right) + \frac{1}{2}(H_0 - \Delta) = p_0$$

e

$$f(a) = q_n = R_n + \frac{1}{2}H_n = \left(R_0 + \frac{1}{2}\Delta \right) + \frac{1}{2}(H_0 - \Delta) = q_0$$

ou seja, a autofecundação não altera a freqüência gênica da população original.

- Média de uma população resultante de autofecundação

Considerando ainda os valores apresentados na Tabela 4.12 e admitindo um loco em que os valores genotípicos são u+a, u+d e u-a para AA, Aa e aa, respectivamente, os valores das médias das populações podem ser obtidos por meio de:

$$\mu_0 = u + a(D_0 - R_0) + dH_0$$

e

$$\mu_n = u + a(D_n - R_n) + dH_n = \mu_0 - d\Delta$$

Pela expressão anterior, fica evidente que a média de uma população é afetada pela perda de heterozigoto e pelo valor genotípico apresentado por ele. Geralmente o valor de d é positivo e, consequentemente, a maioria dos caracteres apresentará redução em seus valores médios, sendo tal fato atribuído à depressão por endogamia. Entretanto, em algumas situações o valor de d será negativo quando houver dominância do fenótipo de menor grandeza, e, neste caso, a média terá acréscimo após processo de autofecundação. O valor da perda de heterozigoto poderá ser expresso pelo coeficiente de endogamia da população, denotado por F.

- *Variabilidade em uma população resultante da autofecundação*

A variância genética numa população não-endogâmica e em equilíbrio de Hardy-Weinberg é dada por:

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$$

em que:

σ_A^2 : variância aditiva; e

σ_D^2 : variância atribuída aos desvios da dominância.

Com a autofecundação, cujo efeito é quantificado pelo coeficiente de endogamia F, a variância genética da população aumenta, sendo dada por:

$$\sigma_G^2 = (1+F) \sigma_A^2 + (1-F^2) \sigma_D^2$$

Se a perda de heterozigoto é total é há fixação de homozigotos, de forma que o valor de F atinja o máximo igual a 1, a variância aditiva dobrará seu valor e a de dominância será nula, ou seja:

$$\sigma_G^2 = 2 \sigma_A^2$$

c) Acasalamento ao acaso em P₂

Um dos modelos mais importantes na genética de populações é o acasalamento ao acaso, o que significa dizer que cada indivíduo tem igual possibilidade de se acasalar com qualquer outro indivíduo da população. Em outras palavras, diz-se que os casais têm as mesmas chances de se acasalarem, como se fossem gerados pela união aleatória entre indivíduos. A chance de um organismo, de genótipo definido, se acasalar com outro é igual à freqüência deste genótipo na população. O aspecto importante é que não haverá tendências para que o acasalamento ocorra entre indivíduos com genótipos semelhantes ou entre aqueles relacionados por ascendência (FALCONER, 1987).

Nesse caso, são envolvidos os seguintes acasalamentos:

População P ₂	Probabilidade	Descendência		
		AA	Aa	aa
AA x AA	0,50 x 0,50	0,25		
AA x Aa(*)	2x0,50 x 0,50	0,25	0,25	
Aa x Aa	0,50 x 0,50	0,0625	0,125	0,0625
Total	1	0,5625	0,375	0,0625

(*) Inclui também o cruzamento Aa x AA

Com conhecimentos adquiridos em genética de populações, poder-se-á estimar com facilidade as freqüências gênicas ou alélicas da população, pelas expressões:

$$f(A) = p = D + \frac{1}{2}H$$

e

$$f(a) = q = R + \frac{1}{2}H$$

Assim, para população P₂, tem-se:

$$f(A) = p = 0,75 \quad e \quad f(a) = q = 0,25$$

Utilizando o princípio do equilíbrio de Hardy-Weinberg, aplicado a populações derivadas de acasalamento ao acaso, também pode ser obtido:

Genótipo	Esperado	Freqüência
AA	p^2	0,5625
Aa	$2pq$	0,3750
aa	q^2	0,0625

Sistemas de acasalamento preferenciais também ocorrem naturalmente, quando a formação dos casais é baseada no fenótipo. A maioria dos acasalamentos preferenciais são positivos, o que significa dizer que os casais formados têm, em média, fenótipos mais similares do que o esperado com acasalamento ao acaso.

O período de florescimento de uma espécie é um exemplo de acasalamento preferencial positivo. Geralmente, o período de florescimento de uma planta é relativamente menor do que a duração total da estação de florescimento. Plantas que florescem mais cedo na estação são polinizadas preferencialmente por outras que também florescem mais cedo, e aquelas que florescem tarde são preferencialmente polinizadas por outras de florescimento tardio (HARTL; CLARK, 1997).

Modelos de acasalamento preferenciais tendem a ser complexos, pois a maioria das características (fenotípicas) é poligênica. Mas, todavia, considerando que fenótipos semelhantes tendem a se acasalar, acasalamentos preferenciais tendem a aumentar a frequência de genótipos homozigotos e diminuir genótipos heterozigotos na população; logo, a variância fenotípica populacional também é aumentada.

4.6. Endogamia

4.6.1 Introdução

A endogamia é o fenômeno que ocorre em decorrência do acasalamento entre indivíduos aparentados. Pode ter consequências sobre a média de uma população e afeta a similaridade das linhas derivadas. O coeficiente de endogamia refere-se à probabilidade de que os alelos de um loco de um indivíduo sejam idênticos por ascendência. Estes alelos são idênticos quando derivam ou são cópias de um alelo comum, encontrado nos ancestrais daquele indivíduo.

Numa população, podem-se encontrar homozigotos com alelos idênticos por ascendência ou idênticos apenas pela função que exercem. Assim, para um indivíduo I de genótipo A_pA_m , define-se o coeficiente de endogamia por meio de:

$$F = P(A_p \equiv A_m)$$

em que:

\equiv : “símbolo que significa idêntico por ascendência”.

4.6.2. Conceitos de endogamia

Endogamia como consequência de gametas que se unem

Segundo Wright (1951), endogamia ocorre em consequência da identidade dos gametas masculinos e femininos que se unem e é expressa pela correlação entre os valores gaméticos que formam a progênie derivada de uma população. Em uma população panmítica, onde ocorre o acasalamento ao acaso, essa correlação é nula, porém em uma população endogâmica a correlação não é nula, sendo proporcional ao coeficiente de endogamia.

Será considerada, inicialmente, uma população em que ocorre o acasalamento ao acaso e em que são formados gametas A e a, com freqüência p e q, respectivamente, para formar a nova descendência. Será também admitido que os valores dos gametas A e a são, respectivamente, 1 e 0. Assim, tem-se:

Gametas (♀)	Gametas (♂)		Valor do Gameta (Y_i)	Freqüência $f(Y_i)$
	A	a		
A	AA (p^2)	Aa (pq)	1	p
a	Aa (pq)	aa (q^2)	0	q
Valor do Gameta (X_i)		1	0	
Freqüência $f(X_i)$		p	q	

Verifica-se que:

a) Variância entre gametas ♀

$$V(f) = \sum Y_i^2 f(Y_i) - [\sum Y_i f(Y_i)]^2 = (1^2 p + 0^2 q) - (1p + 0q)^2 = pq$$

b) Variância entre gametas ♂

$$V(m) = \sum X_i^2 f(X_i) - [\sum X_i f(X_i)]^2 = (1^2 p + 0^2 q) - (1p + 0q)^2 = pq$$

c) Covariância entre gametas ♀ e ♂

Para o cálculo da covariância, deve-se considerar a distribuição conjunta dos valores de X e Y. Assim, tem-se:

Valores de X_i	Valores de Y_i	$X_i Y_i$	Freqüência
1	1	1	p^2
1	0	0	pq
0	1	0	
0	0	0	q^2

logo:

$$\text{Cov}(m, f) = \sum X_i Y_i f(X_i Y_i) - [\sum X_i f(X_i)][\sum Y_i f(Y_i)]$$

$$\text{Cov}(m, f) = [1p^2 + 0(2pq + q^2)] - (1p + 0q)(1p + 0q) = 0$$

Conclui-se que:

$$F = r_{mf} = \frac{\text{Cov}(m, f)}{\sqrt{V(m)V(f)}} = 0$$

ou seja, numa população em que há o acasalamento ao acaso o coeficiente de endogamia é nulo.

Será considerada, de outra forma, outra população com a particularidade de que o acasalamento não ocorre ao acaso, mas sim entre indivíduos aparentados, tendo como consequência aumento na freqüência de homozigotos em detrimento da freqüência de heterozigotos. Será também admitido que o acréscimo na freqüência de homozigotos, em relação à população panmítica, seja dado por ε . Assim, tem-se:

Gametas (\textcircled{f})	Gametas (\textcircled{m})		Valor do Gameta (Y_i)	Freqüência $f(Y_i)$
	A	a		
A	AA ($p^2 + \varepsilon$)	Aa ($pq - \varepsilon$)	1	0
a	Aa ($pq - \varepsilon$)	aa ($q^2 + \varepsilon$)	0	q
Valor do Gameta (X_i)		1	0	
Freqüência $f(X_i)$		p	q	

Verifica-se que:

a) Variância entre gametas \textcircled{f}

$$V(f) = \sum Y_i^2 f(Y_i) - [\sum Y_i f(Y_i)]^2 = (1^2 p + 0^2 q) - (1p + 0q)^2 = pq$$

b) Variância entre gametas \textcircled{m}

$$V(m) = \sum X_i^2 f(X_i) - [\sum X_i f(X_i)]^2 = (1^2 p + 0^2 q) - (1p + 0q)^2 = pq$$

c) Covariância entre gametas \textcircled{f} e \textcircled{m}

Para o cálculo da covariância, deve-se considerar a distribuição conjunta dos valores de X e Y. Assim, tem-se:

Valores de X_i	Valores de Y_i	$X_i Y_i$	Freqüência
1	1	1	$p^2 + \varepsilon$
1	0	0	$pq - \varepsilon$
0	1	0	
0	0	0	$q^2 + \varepsilon$

logo:

$$\text{Cov}(m, f) = \sum X_i Y_i f(X_i Y_i) - [\sum X_i f(X_i)][\sum Y_i f(Y_i)] = [1(p^2 + \varepsilon) + 0(2pq + q^2 - \varepsilon)] - (1p + 0q)(1p + 0q) = \varepsilon$$

Conclui-se que:

$$F = r_{mf} = \frac{\text{Cov}(m, f)}{\sqrt{V(m)V(f)}} = \frac{\varepsilon}{\sqrt{(pq)(pq)}} = \frac{\varepsilon}{pq}$$

A correlação não é nula, produzindo progênie com coeficiente de endogamia F. O efeito que a endogamia proporciona sobre as freqüências de heterozigotos e homozigotos é dado por $\varepsilon = pqF$.

Endogamia resultante do acasalamento entre parentados

O coeficiente de endogamia é expresso pela probabilidade de que os dois alelos que o indivíduo possui, para um determinado loco, sejam idênticos por ascendência. Assim, considerando um indivíduo X, de constituição genotípica ab ($X_{(ab)}$), tem-se:

$$F_x = P(a \equiv b)$$

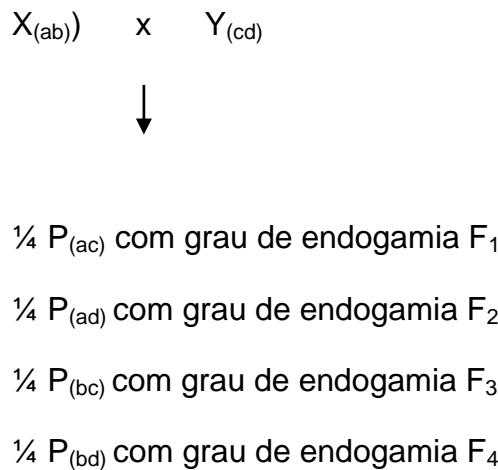
em que:

\equiv : símbolo que significa ser idêntico por ascendência.

Por outro lado, o coeficiente de parentesco é a probabilidade de que dois indivíduos tenham alelos idênticos por ascendência. Assim, considerando dois indivíduos X e Y, de constituição genotípica ab e cd ($X_{(ab)}$ e $Y_{(cd)}$), respectivamente, tem-se:

$$r_{XY} = \frac{1}{4} [P(a \equiv c) + P(a \equiv d) + P(b \equiv c) + P(b \equiv d)]$$

Com base nessas definições, constata-se que os coeficientes de endogamia e de parentesco estão intimamente relacionados. Assim, verifica-se que o coeficiente de endogamia médio de uma progêniese igual ao coeficiente de parentesco de seus genitores. Tomando os indivíduos $X_{(ab)}$ e $Y_{(cd)}$, obtém-se a progênie:



O coeficiente de endogamia médio da progênie F_{ij} é dado por:

$$F_P = \frac{1}{4} (F_1 + F_2 + F_3 + F_4)$$

Como: $F_1 = P(a \equiv c)$, $F_2 = P(a \equiv d)$, $F_3 = P(b \equiv c)$ e $F_4 = P(b \equiv d)$, então:

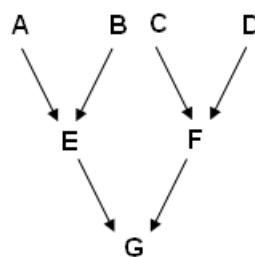
$$F_P = r_{XY}$$

Ou seja, a endogamia média na progênie é dada pelo parentesco entre seus genitores.

A seguir é apresentado o coeficiente de parentesco entre vários indivíduos e sua relação com o coeficiente de endogamia:

Relação	Simbologia	Genealogia	Valor
Indivíduo com ele próprio	$r_{X\otimes}$	$X \longleftrightarrow X$	$\frac{1}{2}(1+F_X)$
Irmãos completos	$r_{IC} = r_{AB}$	$\begin{array}{ccccc} X & & x & & Y \\ \downarrow & & \diagup & & \downarrow \\ & & \times & & \\ & & \diagdown & & \\ A & & & & B \end{array}$	$\frac{1}{4}\left(1 + \frac{F_X + F_Y}{2}\right)$
Meios-irmãos	$r_M = r_{AB}$	$\begin{array}{ccccc} X & & x & & W & & x & & Y \\ & & \downarrow & & & & \downarrow & & \\ & & & & W & & & & Y \\ & & & & \downarrow & & & & \downarrow \\ A & & & & & & & & B \end{array}$	$\frac{1}{8}(1+F_W)$
Pai e filho	$r_{PF} = r_{XA}$	$\begin{array}{ccccc} X & & x & & Y \\ & & \downarrow & & \\ & & A & & \end{array}$	$\frac{1}{4}(1+F_X)$

Dada a genealogia:



Verificam-se as propriedades:

a) $r_{EF} = r_{E,CD} = \frac{1}{2}(r_{EC} + r_{ED})$

b) $r_{EF} = r_{AB,F} = \frac{1}{2}(r_{AF} + r_{BF})$

c) $r_{EF} = r_{AB,CD} = \frac{1}{4}(r_{AC} + r_{AD} + r_{BC} + r_{BD})$

d) $r_{GG} = \frac{1}{2}(1+F_G)$ e $F_G = r_{EF}$ sendo F_G o coeficiente de endogamia de G.

4.6.3 Sistemas de Acasalamento e Estimação do Coeficiente de Endogamia

a) Por meio do coeficiente de parentesco

Será considerada, inicialmente, a estimação do coeficiente de endogamia em gerações sucessivas de autofecundação numa população original F_2 , derivada do cruzamento entre linhagens contrastantes, de forma que as freqüências gênicas sejam dadas por $p = q = 0,5$. Para a situação em que há sucessivas gerações de autofecundação, pode-se considerar a genealogia:

$$F_2 \xrightarrow{\otimes} F_3 \xrightarrow{\otimes} F_4 \xrightarrow{\otimes} \dots \xrightarrow{\otimes} F_n$$

Assim:

$$F_{F3} = r_{F2,F2} = \frac{1}{2}(1+F_{F2}) = \frac{1}{2}$$

$$F_{F4} = r_{F3,F3} = \frac{1}{2}(1+F_{F3}) = \frac{3}{4}$$

$$F_{F5} = r_{F4,F4} = \frac{1}{2}(1+F_{F4}) = \frac{7}{8}$$

...

$$F_{Ft} = r_{Ft-1,Ft-1} = \frac{1}{2}(1+F_{Ft-1})$$

por conseqüência, tem-se:

$$F_t = \frac{1}{2} + \left(\frac{1}{2}\right)^2 (1+F_{t-2}) = \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 (1+F_{t-3})$$

$$F_t = \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \dots + \left(\frac{1}{2}\right)^{t-1} + \left(\frac{1}{2}\right)^t (1+F_0)$$

sendo:

$$F_0 = 0$$

então:

$$F_t = \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \dots + \left(\frac{1}{2}\right)^t$$

A soma dos n termos de uma progressão geométrica é dada por:

$$S_n = \frac{a_1(1-q^n)}{1-q}$$

sendo a_1 o primeiro termo de progressão e q a razão. Logo:

$$F_t = \frac{\left(\frac{1}{2}\right) \left[1 - \left(\frac{1}{2}\right)^t \right]}{1 - \frac{1}{2}}$$

Apartir desta expressão, conclui-se que:

$$F_t = 1 - \left(\frac{1}{2}\right)^t \quad (t \geq 0) \tag{1}$$

ou, alternativamente, tem-se:

$$F_t = 1 - \left(\frac{1}{2}\right)^{t-1} \quad (t \geq 1) \tag{2}$$

Por esta última expressão, obtém-se:

Geração	1	2	3	4	5	... ∞
F	0	1/2	3/4	7/8	15/16	... 1

Por outro lado, considerando que $F_{F_1} = F_{F_2} = 0$, a expressão apropriada seria:

$$F_t = 1 - \left(\frac{1}{2}\right)^{t-2} \quad (t \geq 2) \quad (3)$$

Por esta expressão, obtém-se:

Geração	2	3	4	5	... ∞
F	0	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{7}{8}$...

De forma generalizada, tem-se:

$$F_t = 1 - \left(\frac{1}{2}\right)^{t-g}$$

sendo g a geração em que se verifica o coeficiente de endogamia nulo ($F_g = 0$)

b) Por meio da redução na freqüência de heterozigotos

Para ilustrar o efeito da endogamia sobre a freqüência genotípica de uma população, será considerada, inicialmente, uma população em equilíbrio de Hardy-Weinberg (não-endogâmica) submetida a sucessivos ciclos de autofecundação, conforme apresentado na Tabela 4.13.

Tabela 4.13 - Freqüência genotípica, média e coeficiente de endogamia de populações submetidas a sucessivas autofecundações

Geração	<u>Genótipos</u>			Coef. de Endogamia
	AA	Aa	aa	
0	p^2	$2pq$	q^2	0
1	$p^2 + \frac{1}{2}pq$	pq	$q^2 + \frac{1}{2}pq$	$\frac{1}{2}$
2	$p^2 + \frac{3}{4}pq$	$\frac{1}{2}pq$	$q^2 + \frac{3}{4}pq$	$\frac{3}{4}$
3	$p^2 + \frac{7}{8}pq$	$\frac{1}{4}pq$	$q^2 + \frac{7}{8}pq$	$\frac{7}{8}$
...				
∞	p	0	q	1

A freqüência do heterozigoto na t-ésima geração de autofecundação é dada por:

$$H_{Ft} = \left(\frac{1}{2}\right)^t 2pq$$

ou seja, a cada geração de autofecundação a freqüência de heterozigotos reduz-se à metade. Sendo H_0 a freqüência de heterozigoto na geração inicial, tem-se:

$$H_0 = 2pq \quad \text{e} \quad F_t = 1 - \left(\frac{1}{2}\right)^t$$

Assim:

$$F_t = \frac{H_0 - H_{Ft}}{H_0} = \frac{\Delta}{H_0}$$

ou seja:

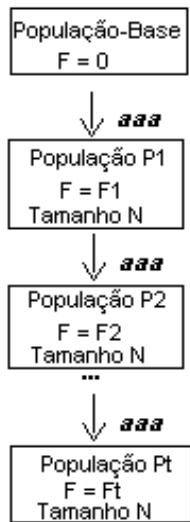
$$H_{Ft} = H_0 (1 - F_t)$$

Verifica-se, portanto, que a freqüência de heterozigotos, em relação à freqüência original, é reduzida na proporção $H_0 F_t = 2pq F_t$ na t-ésima geração.

c) Gerações sucessivas de acasalamento ao acaso numa população de tamanho finito

Já foi ressaltado que após uma geração de acasalamento ao acaso a endogamia desaparece. Essa afirmativa é válida para uma população suficientemente grande, de forma que a probabilidade do encontro gamético de alelos idênticos por ascendência é nula. Para uma população de tamanho finito (N), onde os indivíduos são monóicos e ocorre o acasalamento ao acaso, surge a cada geração um efeito da endogamia. Esse efeito é resultante da probabilidade de que alelos idênticos por ascendência se unem com probabilidade que não é desprezível e diretamente proporcional ao valor de N .

Para efeito de ilustração, será considerado o esquema:



Será admitido que para a formação da população P1 foram produzidos $2N$ gametas a partir de N indivíduos da população-base. Assim, cada indivíduo produziu dois tipos de gametas. As seguintes probabilidades podem ser calculadas:

$$P(\text{ocorrência de um tipo de gameta}) = \frac{1}{2N}$$

$$P(\text{ocorrência de 2 gametas do mesmo tipo}) = \left(\frac{1}{2N}\right)^2$$

$$P(\text{alelos idênticos por ascendência em P1}) = \left(\frac{1}{2N}\right)^2 2N = \frac{1}{2N}$$

Assim, conclui-se que o coeficiente de endogamia da população P1 é igual a $\frac{1}{2N}$. Se o valor de N é infinitamente grande, esta endogamia será desprezível, ou seja:

$$F_1 = \frac{1}{2N}$$

Essa abordagem será exemplificada, considerando uma população de tamanho $N = 3$ com indivíduos de genótipo aa' , bb' e cc' . Os gametas e os indivíduos formados estão ilustrados a seguir:

Freq.:	1/2N	1/2N	1/2N	1/2N	1/2N	1/2N
Gametas	a	a'	b	b'	c	c'
a	aa	aa'	ab	ab'	ac	ac'
a'	a'a	a'a'	a'b	a'b'	a'c	a'c'
b	ba	ba'	bb	bb'	bc	bc'
b'	b'a	b'a'	b'b	b'b'	b'c	b'c'
c	ca	ca'	cb	cb'	cc	cc'
c'	c'a	c'a'	c'b	c'b'	c'c	c'c'

Assim, se g e g' são dois tipos de gametas, tem-se:

$$F_1 = P(g \equiv g) + \left(1 - \frac{1}{2N}\right)P(g \equiv g') = \frac{1}{2N}$$

Para a formação da população P2, deve-se considerar a existência da endogamia em P1, ou seja, existe a proporção $F_1 = \frac{1}{2N}$ indivíduos com alelos idênticos por ascendência. A população P2 provém de N indivíduos, gerando uma endogamia igual a $\frac{1}{2N}$ como visto anteriormente. Contudo, de todos os gametas produzidos pela população P1, uma fração $1/(2N)$ se une e carrega cópias de alelos, e uma fração $1 - 1/(2N)$ que se une carrega cópias de outros alelos. Dessa forma, considera-se:

$$F_2 = \text{endogamia nova} + \text{endogamia remanescente}$$

ou

$$F_2 = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)F_1$$

Por analogia ao exemplo anterior, pode-se considerar:

$$F_2 = \frac{1}{2N}P(g \equiv g) + \left(1 - \frac{1}{2N}\right)P(g \equiv g')$$

De maneira análoga podem-se estimar os demais coeficientes de endogamia, ou seja:

$$F_3 = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)F_2$$

e

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)F_{t-1}$$

Fazendo:

$$\Delta_F = \frac{1}{2N} : \text{incremento da endogamia a cada geração}$$

e

$$P = 1 - F: \text{índice de panmixia}$$

tem-se:

$$1 - P_t = \Delta_F + (1 - \Delta_F)(1 - P_{t-1})$$

ou

$$P_t = (1 - \Delta_F)P_{t-1}$$

Por extensão:

$$P_{t-1} = (1 - \Delta_F)P_{t-2}$$

logo:

$$P_t = (1 - \Delta_F)^2 P_{t-2} = \dots = (1 - \Delta_F)^t P_0$$

Considerando que o índice de panmixia na população inicial (P_0) é 1 (e $F_0 = 0$), tem-se:

$$P_t = (1 - \Delta_F)^t$$

e

$$F_t = 1 - (1 - \Delta_F)^t = 1 - \left(1 - \frac{1}{2N}\right)^t$$

É interessante observar que, se $N = 1$, então tem-se situação idêntica à da autofecundação, caracterizada por:

$$F_t = 1 - \left(\frac{1}{2}\right)^t$$

4.6.4 Sistema misto de acasalamento

Em certas populações pode ocorrer simultaneamente o acasalamento ao acaso e autofecundações em taxas diferenciadas. Assim, se numa população de acasalamento ao acaso – que apresenta as freqüências de AA, Aa e aa iguais a p^2 , $2pq$ e q^2 , respectivamente – ocorrer taxa de autofecundação igual a w e de acasalamento ao acaso igual a $1-w$, esperar-se-á que a próxima geração tenha a seguinte constituição genotípica:

Genótipo	Acasalamento ao acaso (taxa = $1-w$)	Autofecundação (taxa = w)	Total
AA	p^2	$p^2 + \frac{1}{2} pq$	$p^2 + \frac{1}{2} pqw$
Aa	$2pq$	pq	$2pq(1 - \frac{1}{2}w)$
aa	q^2	$q^2 + \frac{1}{2} pq$	$q^2 + \frac{1}{2} pqw$

Se o processo de cruzamento misto, envolvendo autofecundações e acasalamento ao acaso, é continuado, nas mesmas taxas, a redução na freqüência de heterozigotos da primeira até a t -ésima geração poderá ser previda por:

$$H_1 = 2pq\left(1 - \frac{1}{2}w\right)$$

$$H_2 = 2pq\left(1 - w\right) + pq\left[1 - \left(\frac{1}{2}w\right)\right]w = 2pq\left[1 - \frac{1}{2}w - \left(\frac{1}{2}w\right)^2\right]$$

...

$$H_t = 2pq\left[1 - \frac{1}{2}w - \left(\frac{1}{2}w\right)^2 - \left(\frac{1}{2}w\right)^3 - \dots - \left(\frac{1}{2}w\right)^t\right] = 2pq\left(1 - \frac{w}{2-w}\right)$$

Pode-se, portanto, prever que o coeficiente de endogamia no equilíbrio será dado por:

$$F = 1 - \frac{H_t}{H_0} = \frac{w}{2-w}$$

Veja alguns valores de w e F:

w	0	0,25	0,5	0,75	1
F	0	0,14	0,33	0,60	1

4.7. Fixação Gênica

4.7.1 Em populações derivadas de autofecundações

Como verificado anteriormente, em cada geração a redução na freqüência dos heterozigotos é de $2pq F$. Assim, a freqüência dos homozigotos é dada por:

Genótipo	Original	No equilíbrio	Freqüência após n	
			gerações de autofecundações	após n
AA	D_0	p^2	D_n	$p^2 + pqF_t$
Aa	H_0	$2pq$	H_n	$2pq - 2pqF_t$
aa	R_0	q^2	R_n	$q^2 + pqF_t$

Por este quadro, observa-se que:

- Com infinitas gerações de autofecundação, tem-se:

$$F_t = F_\infty = 1$$

e então:

$$f(Aa) = 2pq(1 - F_\infty) = 0$$

$$f(AA) = p^2 + pqF_\infty = p^2 + pq = p(p+q) = p$$

$$f(aa) = q^2 + pqF_\infty = q^2 + pq = q(p+q) = q$$

- As sucessivas autofecundações fazem com que ocorra a fixação gênica, ou seja, a eliminação de locos em heterozigose, de modo que, ao final do processo de autofecundação, todos os genes, teoricamente, terão segregação na proporção p:q (AA:aa).

- Se após a endogamia houver uma geração de acasalamento ao acaso, a população resultante do cruzamento entre os indivíduos da população endogâmica apresentará a seguinte relação genotípica:

Genótipo	Freqüência
AA	p^{*2}
Aa	$2pq^*$
aa	q^{*2}

Pela lei do equilíbrio de Hardy-Weinberg, as freqüências alélicas podem ser estimadas a partir da freqüência genotípica da população genitora, que, no caso, se trata da t-ésima geração de autofecundação da população original, que tinha freqüências alélicas iguais a p e q.

Assim, tem-se que:

$$p^* = f(A) = D_t + \frac{1}{2}H_t = p^2 + pqF_t + \frac{1}{2}(2pq - 2pqF_t) = p$$

$$q^* = f(a) = R_t + \frac{1}{2}H_t = q^2 + pqF_t + \frac{1}{2}(2pq - 2pqF_t) = q$$

Verifica-se, portanto, que a população descendente do acasalamento ao acaso terá a mesma freqüência gênica da população-base submetida ao processo de autofecundação.

4.7.2. Deriva genética como causa de fixação

Considerações iniciais

A deriva genética é o processo decorrente da amostragem de uma população responsável pela alteração dispersiva na freqüência alélica. Se uma amostra é adequada, espera-se que a freqüência alélica estimada (\hat{p}), tendo-se por base o

número de ocorrência de cada classe genotípica na amostra, reflita a freqüência alélica da população original(p), de forma que se obtenha:

$$E(\hat{p}) = p$$

$$E(\hat{q}) = E(1 - \hat{p}) = 1 - E(\hat{p}) = 1 - p = q$$

Entretanto, as várias amostras de uma população apresentarão diferentes valores de \hat{p} , visto que a ocorrência do alelo A poderá ser de diferente magnitude nas várias amostras.

Será admitido que na população original, de tamanho η , encontram-se η_{11} genótipos AA, η_{12} Aa e η_{22} aa, de forma que:

$$f(A) = p = \frac{2\eta_{11} + \eta_{12}}{2\eta}$$

e

$$f(a) = q = \frac{2\eta_{22} + \eta_{12}}{2\eta}$$

Ao ser tomada, desta população original, uma amostra de tamanho N ($N < \eta$), deve ser admitido que ela foi originada do encontro entre $2N$ gametas, em que N originou-se dos genitores femininos e os outros N , dos genitores masculinos. A relação genotípica esperada (RGe) na amostra será:

$$RGe = [p + q]^{2N}$$

Nesta amostra, que tem N indivíduos e, portanto, $2N$ alelos, a ocorrência do alelo A poderá ser quantificada pela variável x , de distribuição binomial, que assumirá os seguintes valores: 0, 1, 2, ..., $2N$. Tendo x distribuição binomial, verificam-se as propriedades:

$$\bar{x} = E(x) = 2Np$$

$$V(x) = 2Npq$$

Para uma particular amostra, a freqüência do alelo A será quantificada por meio de:

$$\hat{p} = \frac{x}{2N} \quad (0 \leq \hat{p} \leq 1)$$

portanto:

$$E(\hat{p}) = E\left(\frac{x}{2N}\right) = \frac{2Np}{2N} = p$$

e

$$V(\hat{p}) = V\left(\frac{x}{2N}\right) = \frac{V(x)}{4N^2} = \frac{pq}{2N}$$

A probabilidade de que \hat{p} assuma um particular valor, por exemplo, igual a $k/2N$, poderá ser calculada por meio de:

$$P\left(\hat{p} = \frac{k}{2N}\right) = \frac{(2N)!}{k!(2N-k)!} \hat{p}^k (1-\hat{p})^{2N-k}$$

A variação esperada na freqüência gênica da população pelo processo de amostragem pode, então, ser quantificada por:

$$\Delta p = p - \hat{p}$$

e

$$V(\Delta p) = V(p - \hat{p}) = \frac{pq}{2N}$$

Deve ser lembrado que N é o número de indivíduos da população original que efetivamente contribuíram para a formação da amostra. Muitas vezes o que se procura é exatamente conhecer o valor do tamanho efetivo da amostra.

Oscilação na freqüência gênica

A oscilação na freqüência gênica poderá ser computada admitindo que são retiradas s amostras de tamanho N e que os valores \hat{p}_j ($j=1,2,\dots,s$) seguem distribuição aproximadamente normal, para s suficientemente grande. Considera-se que:

$$E(\hat{p}_j \pm \hat{\sigma}_{pj}) = p \pm \sqrt{\frac{pq}{2N}}$$

e que

$$P(p_j - \sigma_{pj} \leq \hat{p}_j \leq p_j + \sigma_{pj}) = 68,2\%$$

Assim, se for admitida a existência de 50 amostras de tamanho N igual a 50, derivadas de uma população original cuja freqüência gênica é $p=q=0,5$, que apenas 68,2% das amostras irão apresentar freqüência entre 0,45 e 0,55, sendo, portanto, difícil de manter a identidade gênica com a população original.

Nesse ponto, cabe esclarecer que, estando apenas o loco sob discussão, não é possível compreender propriamente o que pode ocorrer em uma amostra, exceto quando considerada como sendo uma entre várias amostras. No entanto, o que ocorre aos alelos de um loco, em um número s de amostras, também ocorre aos alelos de vários locos em uma única amostra, desde que todos eles iniciem com a mesma freqüência gênica. Para ilustrar esse fato, pode-se admitir uma primeira situação em que haja uma única amostra e investiga-se a freqüência de alelos em 100 locos, com freqüências iniciais $p=q=0,5$. Assim, em uma amostra de 50 indivíduos espera-se que:

- 2,1 locos terão freqüência entre 0,35 e 0,40.
- 13,6 locos terão freqüência entre 0,40 e 0,45.
- 34,1 locos terão freqüência entre 0,45 e 0,50.

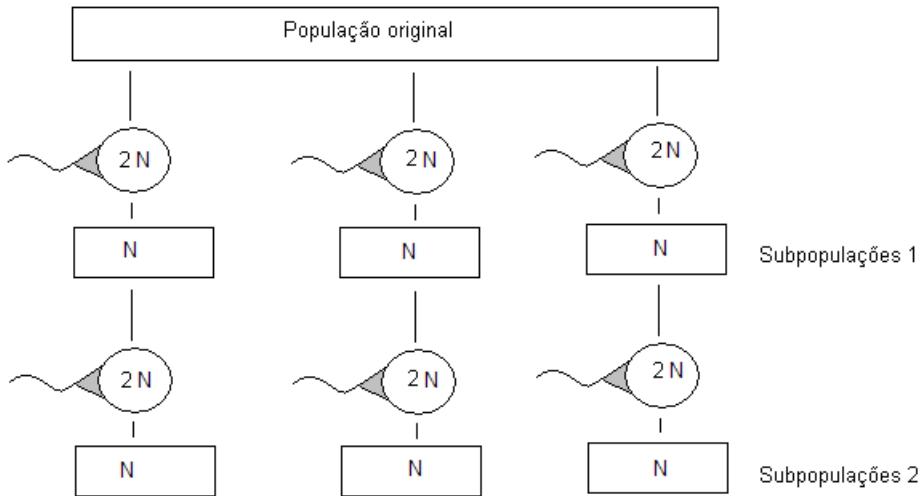
Em uma segunda situação, considera-se uma população a partir da qual se retiram 100 amostras de tamanho 50 e avalia-se a oscilação gênica em um alelo que se encontra na freqüência original de $p=q=0,5$. Espera-se que:

- 2,1 amostras terão freqüência do alelo entre 0,35 e 0,40.
- 13,6 amostras terão freqüência do alelo entre 0,40 e 0,45.
- 34,1 amostras terão freqüência do alelo entre 0,45 e 0,50.

Efeito da deriva genética ao longo do tempo - fixação

Para estudar os efeitos da deriva genética proporcionada pela subdivisão de uma população original em subpopulações, cada uma com s amostras, ao longo de um tempo t de acasalamentos ao acaso, será admitida uma população ideal, constituída por indivíduos diplóides e monóicos. Admite-se que o acasalamento ocorra unicamente entre os indivíduos da mesma amostra, que as gerações são distintas e que não haja seleção, migração e mutação. Também se considera que o

número de indivíduos que se acasalam para dar origem a próxima geração é constante e igual a N em todas as amostras e em todas as subpopulações. Esquematicamente, tem-se:



No diagrama, considera-se que a população original é dividida em várias amostras (ou linhas) ao longo do tempo. Em cada geração, $2N$ gametas de uma amostra se unem para gerar os N indivíduos da próxima subpopulação. Admite-se que a freqüência alélica na população original seja p_0 e q_0 e que a freqüência do alelo A na amostra (ou linha) j na subpopulação t seja dada por \hat{p}_{jt} ou simplesmente \hat{p}_t .

As seguintes análises poderão ser feitas:

Análise da subpopulação 1

Com já visto anteriormente, tem-se:

$$E(\hat{p}_{j1}) = E(\hat{p}_1) = p_0$$

$$V(\hat{p}_{j1}) = V(\hat{p}_1) = \frac{p_0 q_0}{2N}$$

Análise da subpopulação 2

A esperança matemática da freqüência gênica da subpopulação 2, em relação à subpopulação, é dada por:

$$E_2(\hat{p}_{j2}) = E_2(\hat{p}_2) = \hat{p}_1$$

Entretanto, é interessante que a referência seja a população original e não a antecessora. Assim, deve-se quantificar a esperança matemática da freqüência gênica da subpopulação 2, em relação à população original, denotada por:

$$E_1 E_2(\hat{p}_{j2}) = E_1 E_2(\hat{p}_2) = E_1(\hat{p}_1) = p_0$$

A variância da freqüência gênica na subpopulação 2 é dada por:

$$V(\hat{p}_{j2}) = V(\hat{p}_2) = E_2[\hat{p}_2 - E_2(\hat{p}_2)]^2 = E_2(\hat{p}_2 - p_0)^2$$

Em relação à população original, é esperado que:

$$\begin{aligned} V(\hat{p}_2) &= E_1 E_2[(\hat{p}_2 - p_0)^2] = E_1 E_2[(\hat{p}_2 - \hat{p}_1) + (\hat{p}_1 - p_0)]^2 \\ &= E_1 E_2[(\hat{p}_2 - \hat{p}_1)]^2 + 2E_1 E_2[(\hat{p}_2 - \hat{p}_1)(\hat{p}_1 - p_0)] + E_1 E_2[(\hat{p}_1 - p_0)]^2 \end{aligned}$$

Para obter $V(\hat{p}_2)$, deve-se investigar cada termo da expressão anterior.

Assim, tem-se:

- $E_1 E_2[(\hat{p}_2 - \hat{p}_1)]^2$

$$\begin{aligned} E_1 E_2[(\hat{p}_2 - \hat{p}_1)]^2 &= E_1 \left\{ E_2[(\hat{p}_2 - E_2(\hat{p}_2))]^2 \right\} = E_1[V(\hat{p}_2)] = E_1 \left[\frac{\hat{p}_1(1-\hat{p}_1)}{2n} \right] \\ &= E_1 \left[\frac{(\hat{p}_1 - p_0^2) - (\hat{p}_1^2 - p_0^2)}{2N} \right] = \frac{1}{2N} [E(\hat{p}_1) - E(p_0^2) - E_1(\hat{p}_1^2 - p_0^2)] \\ &= \frac{1}{2N} [p_0 - p_0^2 - V(\hat{p}_1)] = \frac{1}{2N} \left[p_0 - p_0^2 - \frac{p_0(1-p_0)}{2N} \right] = \frac{p_0(1-p_0)}{2N} \left(1 - \frac{1}{2N} \right) \end{aligned}$$

- $2E_1 E_2[(\hat{p}_2 - \hat{p}_1)(\hat{p}_1 - p_0)]$

$$2E_1 E_2[(\hat{p}_2 - \hat{p}_1)(\hat{p}_1 - p_0)] = 2E_1(\hat{p}_1 - p_0)[E_2(\hat{p}_2 - \hat{p}_1)] = 2E_1(\hat{p}_1 - p_0)[E_2(\hat{p}_2) - \hat{p}_1] = 0$$

Dado que $E_2(\hat{p}_2) = \hat{p}_1$ e $E_2(\hat{p}_1) = \hat{p}_1$

- $E_1 E_2[(\hat{p}_1 - p_0)]^2$

$$E_1 E_2 [(\hat{p}_1 - p_0)]^2 = E_1 [(\hat{p}_1 - p_0)]^2 = V(\hat{p}_1) = \frac{p_0(1-p_0)}{2N}$$

Voltando à expressão original, pode-se concluir que:

$$V(\hat{p}_2) = \frac{p_0(1-p_0)}{2N} \left[\left(1 - \frac{1}{2N} \right) + 1 \right]$$

Análise da subpopulação t

A esperança matemática da freqüência gênica da subpopulação t, em relação à população original, é dada por:

$$E_1 \dots E_t (\hat{p}_{jt}) = p_0$$

Pode-se deduzir que:

$$V(\hat{p}_t) = \frac{p_0(1-p_0)}{2N} \left[\left(1 - \frac{1}{2N} \right)^{t-1} + \left(1 - \frac{1}{2N} \right)^{t-2} + \dots + \left(1 - \frac{1}{2N} \right)^1 + 1 \right]$$

Sabendo que a soma dos termos de uma progressão geométrica (S) de razão r, com t termos, cujo primeiro elemento é a_1 , é dada por:

$$S = \frac{a_1(1-r^t)}{1-r}$$

tem-se:

$$V(\hat{p}_t) = \frac{p_0(1-p_0)}{2n} \left[\frac{1 - (1/2N)^t}{1/2N} \right] = p_0(1-p_0) \left[1 - \left(1 - \frac{1}{2N} \right)^t \right]$$

Sabendo que:

$$\lim_{t \rightarrow \infty} \left(1 - \frac{1}{2N} \right)^t = 0$$

obtém-se:

$$V(\hat{p}_{t \rightarrow \infty}) = p_0(1-p_0)$$

ou seja, a variância da freqüência gênica aumenta 2N vezes em relação à existente na subpopulação inicial. Também se constata que o valor $p_0(1-p_0)$ corresponde

exatamente à variância da freqüência gênica em um conjunto de amostras em que o alelo A foi fixado em p_0 linhas (ou amostras) e o alelo a, em q_0 linhas.

Tamanho Efetivo

Para melhor entendimento de seu efeito sobre a freqüência gênica, deve-se definir tamanho ideal e tamanho efetivo de população. A importância desses conceitos torna-se mais clara à medida que se leva em consideração que é uma amostra de genes de uma determinada população que será transmitida à próxima. Assim, a freqüência gênica na progênie será influenciada pela variação amostral, que será tanto maior quanto menor for o número de pais.

Tamanho efetivo de população representa o número de indivíduos que contribuem efetivamente para a variância de amostragem, ou taxa de endogamia, desde que acasalados de acordo com as premissas da população ideal, que podem ser, segundo Falconer (1987), resumidas simplesmente no seguinte: "é aquela na qual a variação de amostragem é tão pequena que pode ser desconsiderada".

O número ou tamanho efetivo de uma população corresponde ao número de reprodutores na população ideal que proporcionaria a mesma taxa de endogamia numa população em estudo. Para estudos do tamanho efetivo, considera-se que a população ideal tem N indivíduos e que apenas N_e se intercruzam de forma a proporcionar a taxa de endogamia estimada na população em estudo.

As relações entre os números observados (N) em populações (ou amostras) em estudo e o tamanho efetivo (N_e), nas situações mais comumente encontradas, são dadas a seguir.

Caso 1. Organismos hermafroditas, mas sem autofecundações

Nesta situação, considera-se que o número de machos e o de fêmeas que contribuem para a população em estudo sejam iguais, ou seja:

$$N_m = N_f \quad \text{e, portanto,} \quad p_m = p_f$$

sendo p_m e p_f as freqüências do alelo A, entre reprodutores masculino e feminino.

Se a população em estudo tem tamanho N, então sua taxa de endogamia será dada por:

$$\Delta F' = \frac{1}{2N+1}$$

Na população ideal, deve-se ter um número de reprodutores N_e que proporcione a mesma taxa de endogamia, ou seja:

$$\Delta F = \frac{1}{2N_e} = \frac{1}{2N+1}$$

Portanto, apesar da população em estudo ter N indivíduos, seu tamanho efetivo será:

$$N_e = N + 1/2$$

Nesta situação, a exclusão da autofecundação influí pouco na taxa de endogamia, exceto nos casos em que N é muito pequeno.

Caso 1. Organismos dióicos

Nesta situação, considera-se que o número de machos e o de fêmeas que contribuem para a população em estudo possam ser diferentes. A freqüência gênica esperada na progênie obtida do intercruzamento desses reprodutores será:

$$p_1 = \frac{p_m + p_f}{2}$$

sendo p_m e p_f as freqüências do alelo A, entre reprodutores masculino e feminino.

A variância da freqüência gênica proporcionada pela amostragem diferencial nos dois sexos poderá ser estimada como se segue:

$$V(p_1) = V\left(\frac{p_m + p_f}{2}\right) = \frac{1}{4}[V(p_m) + V(p_f)] = \frac{1}{4}\left(\frac{p_m q_m}{2N_m} + \frac{p_f q_f}{2N_f}\right)$$

Não se espera que haja diferença entre as freqüências gênicas nos dois sexos, de forma que:

$$p_m = p_f = p_0$$

Assim:

$$V(p_1) = \frac{p_0 q_0}{8} \left(\frac{1}{N_m} + \frac{1}{N_f} \right)$$

Para que $V(p_0)$, da população ideal, seja igual a $V(p_1)$ da população em estudo, deve-se ter:

$$V(p_0) = \frac{p_0 q_0}{2N_e} = V(p_1)$$

portanto:

$$\frac{p_0 q_0}{2N_e} = \frac{p_0 q_0}{8} \left(\frac{1}{N_m} + \frac{1}{N_f} \right)$$

logo:

$$N_e = \frac{4N_m N_f}{N_m + N_f}$$

Se a população em estudo é proveniente do intercruzamento entre N_m machos e N_f fêmeas, então sua taxa de endogamia será dada por:

$$\Delta F = \frac{1}{2N_e} = \frac{1}{8} \left(\frac{1}{N_m} + \frac{1}{N_f} \right)$$

Como ilustração, considera-se uma população derivada do acasalamento entre cinco genitores masculinos e 95 femininos. Nesta situação, tem-se:

$$N_e = \frac{4N_m N_f}{N_m + N_f} = \frac{4(5)(95)}{5 + 95} = 19$$

e

$$\Delta F = \frac{1}{2N_e} = \frac{1}{38} = 0,0263$$

A população constituída a partir destes 100 indivíduos proporciona valor de $V(p_1)$ ou de ΔF equivalente ao fornecido pelo intercruzamento entre 19 indivíduos da população ideal. Nota-se que a taxa de endogamia depende, principalmente, do menor número de indivíduos entre os dois sexos.

Se uma população for mantida com um número grande de fêmeas, mas apenas um único macho em cada geração, o número efetivo será:

$$N_e = \frac{4N_m N_f}{N_m + N_f} = \frac{4(1)N_f}{1+N_f} \cong 4$$

Assim uma família de meio-irmãos corresponde a apenas 4 indivíduos de uma população ideal.

Outra situação a ser analisada é quando o número de machos e fêmeas que se intercruzam é igual. Assim:

$$N_m = N_f = N$$

e

$$N_e = \frac{4N_m N_f}{N_m + N_f} = \frac{4N^2}{2N} = 2N$$

Assim uma família de irmãos completos (situação em que $N_m = N_f = 1$) corresponde a apenas 2 indivíduos de uma população ideal.

Para um dado número de reprodutores, quanto maior a desproporcionalidade entre a contribuição dos dois sexos, maior será a taxa de endogamia. Se $N_m = N_f$, a taxa de endogamia é minimizada.

Caso 3. Variação do número de indivíduos de geração em geração

O número de indivíduos de uma população, submetida ao acasalamento ao acaso, poderá variar de geração para geração. Assim, pode-se considerar que a população, no tempo t , apresenta n_j indivíduos, de forma que a variância média da freqüência gênica seria dada por:

$$V(p) = \frac{1}{t} \sum_{j=1}^t \frac{p_0 q_0}{2N_j} = \frac{p_0 q_0}{2t} \left(\frac{1}{N_1} + \frac{1}{N_2} + \dots + \frac{1}{N_t} \right)$$

O tamanho efetivo poderá, então, ser quantificado por meio de:

$$V(p) = \frac{p_0 q_0}{2N_e} = \frac{p_0 q_0}{2t} \left(\frac{1}{N_1} + \frac{1}{N_2} + \dots + \frac{1}{N_t} \right)$$

logo:

$$\frac{1}{N_e} = \frac{1}{t} \left(\frac{1}{N_1} + \frac{1}{N_2} + \dots + \frac{1}{N_t} \right)$$

Considerando t gerações, o número efetivo é dado pela média harmônica dos números de indivíduos de cada geração. O valor de N_e está bastante influenciado pelo menor valor de N . Assim, se, por exemplo, tem-se:

$$N_1 = 100 \quad N_2 = 10 \quad N_3 = 100 \quad e \quad N_4 = 100$$

então:

$$\frac{1}{N_e} = \frac{1}{4} \left(\frac{1}{100} + \frac{1}{10} + \frac{1}{100} + \frac{1}{100} \right) = \frac{13}{400} \quad \text{logo } N_e = 30,76$$

As gerações contendo menores números de indivíduos influenciam o tamanho efetivo e, consequentemente, a endogamia da população. A expansão do tamanho da população pouco reduz a endogamia já estabelecida.

4.7.3. Fixação por isolamento de subpopulações - Princípio Wahlund

Para fazer abordagem sobre o princípio de fixação gênica descrito por Wahlund, será considerada uma população subdividida em várias unidades de melhoramento ou subpopulações, mesmo que não completamente isoladas, sobre as quais interessa o estudo da variação entre e dentro das subpopulações.

Será considerado que a população original foi subdividida em s subpopulações, cada uma das quais em equilíbrio de Hardy-Weinberg. Seja p_j a freqüência do alelo A_1 na j -ésima subpopulação, de forma que a freqüência genotípica de A_1A_1 , A_1A_2 e A_2A_2 possa ser expressa por p_j^2 , $2p_j(1-p_j)$ e $(1-p_j)^2$. Também será considerado que as subpopulações apresentam diferentes tamanhos, de forma que o tamanho relativo da j -ésima população seja denotado por θ_j , tendo-

se $\sum_{j=1}^s \theta_j = 1$. Esquematicamente, tem-se:

Subpopulação	A_1A_1	A_1A_2	A_2A_2	$p_j=f(A_1)$	θ_j
1	p_1^2	$2p_1(1-p_1)$	$(1-p_1)^2$	p_1	θ_1
2	p_2^2	$2p_2(1-p_2)$	$(1-p_2)^2$	p_2	θ_2
s	p_s^2	$2p_s(1-p_s)$	$(1-p_s)^2$	p_s	θ_s
Total (O_i)	$\sum_{j=1}^s \theta_j p_j^2$	$2 \sum_{j=1}^s \theta_j p_j (1-p_j)$	$\sum_{j=1}^s \theta_j (1-p_j)^2$		1

A população original, determinada pelo agrupamento de todas as s subpopulações, terá a seguintes características:

i. Freqüência genotípica observada

A freqüência genotípica de A_1A_1 , A_1A_2 e A_2A_2 a ser observada na população original é a média ponderada das freqüências genotípicas das subpopulações, ou seja:

$$f(A_1A_1) = \theta_1 p_1^2 + \theta_2 p_2^2 + \dots + \theta_s p_s^2 = \sum_{j=1}^s \theta_j p_j^2$$

$$f(A_1A_2) = \theta_1 [2p_1(1-p_1)] + \theta_2 [2p_2(1-p_2)] + \dots + \theta_s [2p_s(1-p_s)] = 2 \sum_{j=1}^s \theta_j p_j (1-p_j)$$

$$f(A_2A_2) = \theta_1 (1-p_1)^2 + \theta_2 (1-p_2)^2 + \dots + \theta_s (1-p_s)^2 = \sum_{j=1}^s \theta_j (1-p_j)^2$$

Contudo, a freqüência de cada classe genotípica pode ser denotada de outra maneira, a partir dos parâmetros:

$$\bar{p} = \sum_{j=1}^s \theta_j p_j$$

e

$$\sigma^2 = \sum_{j=1}^s \theta_j (p_j - \bar{p})^2 = \sum_{j=1}^s \theta_j p_j^2 - \bar{p}^2$$

Assim, tem-se:

$$f(A_1A_1) = \sum_{j=1}^s \theta_j p_j^2 = \bar{p}^2 + \sigma^2$$

$$f(A_1A_2) = 2 \sum_{j=1}^s \theta_j p_j (1 - p_j) = 2\bar{p}(1 - \bar{p}) - 2\sigma^2$$

$$f(A_2A_2) = \sum_{j=1}^s \theta_j (1 - p_j)^2 = (1 - \bar{p})^2 + \sigma^2$$

ii. Freqüência genotípica esperada

A freqüência genotípica de A_1A_1 , A_1A_2 e A_2A_2 esperada na população original é aquela correspondente à de uma população em equilíbrio de Hardy-Weinberg com freqüência alélica dada pela esperança matemática das freqüências nas subpopulações. Assim, tem-se:

$$f(A_1A_1) = \bar{p}^2$$

$$f(A_1A_2) = 2\bar{p}(1 - \bar{p})$$

$$f(A_2A_2) = (1 - \bar{p})^2$$

A seguir, podem-se comparar as freqüências esperadas e observadas numa população resultante do agrupamento de várias subpopulações em equilíbrio de Hardy-Weinberg.

Genótipo	Observado	Esperado
A_1A_1	$\sum_{j=1}^s \theta_j p_j^2 = \bar{p}^2 + \sigma^2$	\bar{p}^2
A_1A_2	$2 \sum_{j=1}^s \theta_j p_j (1 - p_j) = 2\bar{p}(1 - \bar{p}) - 2\sigma^2$	$2\bar{p}(1 - \bar{p})$
A_2A_2	$\sum_{j=1}^s \theta_j (1 - p_j)^2 = (1 - \bar{p})^2 + \sigma^2$	$(1 - \bar{p})^2$

\bar{p} : média das freqüências do alelo A_1 nas s subpopulações

Isso indica que, se uma população é subdividida em várias subpopulações que se acasalam ao acaso, a freqüência de homozigotos desta população tende a ser

maior do que a esperada com base no equilíbrio de Hardy-Weinberg estabelecido a partir das freqüências médias das subpopulações. Na prática, a comparação entre os valores esperados e observados provê estimativa de um importante parâmetro, σ^2 , que mede o grau de diferenciação das subpopulações em relação à população original, cuja freqüência alélica é esperada ser \bar{x} .

Aplicação – a seguir são apresentados valores de σ^2 , considerando várias situações de subdivisão de uma determinada população em quatro isolados (ou amostras)

A- Os isolados apresentam tamanhos iguais ($\theta_j = 1/s$)

Isolado	θ_j	p_j	$1 - p_j$	X_{11}	X_{12}	X_{22}
1	0,25	0,9	0,1	0,81	0,18	0,01
2	0,25	0,7	0,3	0,49	0,42	0,09
3	0,25	0,5	0,5	0,25	0,50	0,25
4	0,25	0,3	0,7	0,09	0,42	0,49
Observado		0,6	0,4	0,41	0,38	0,21
Esperado		0,6	0,4	0,36	0,48	0,16
Diferença				0,05	-0,10	0,05
				(σ^2)	$(-2\sigma^2)$	(σ^2)

Nesta situação, a população foi dividida em quatro isolados com mesmo tamanho relativo. O valor de heterozigoto observado é inferior ao esperado, e a magnitude de σ^2 , que mede o grau de diferenciação, é de 0,05.

B – O isolados apresentam tamanhos diferentes e freqüências alélicas equivalentes às de A

Isolado	θ_j	p_j	$1 - p_j$	X_{11}	X_{12}	X_{22}
1	0,3	0,9	0,1	0,81	0,18	0,01
2	0,2	0,7	0,3	0,49	0,42	0,09
3	0,2	0,5	0,5	0,25	0,50	0,25
4	0,3	0,3	0,7	0,09	0,42	0,49
Observado		0,6	0,4	0,418	0,364	0,218
Esperado		0,6	0,4	0,360	0,480	0,160
Diferença				0,058	-0,116	0,058
				(σ^2)	$(-2\sigma^2)$	(σ^2)

Na situação B, a população também foi dividida em quatro isolados com tamanho relativo diferenciado. O valor de heterozigoto observado é inferior ao esperado, e a magnitude de σ^2 , que mede o grau de diferenciação, é maior que a verificada na situação anterior, apesar de as freqüências alélicas nos isolados serem idênticas.

C- Isolados de tamanhos iguais e freqüências alélicas mais similares do que em A e

B						
Isolado	θ_j	p _j	1- p _j	X ₁₁	X ₁₂	X ₂₂
1	0,25	0,8	0,2	0,64	0,32	0,04
2	0,25	0,7	0,3	0,49	0,42	0,09
3	0,25	0,5	0,5	0,25	0,50	0,25
4	0,25	0,4	0,6	0,16	0,48	0,36
Observado		0,6	0,4	0,385	0,430	0,185
Esperado		0,6	0,4	0,360	0,480	0,160
Diferença				0,025	-0,050	0,025
				(σ^2)	(-2 σ^2)	(σ^2)

Na situação C, a população também foi dividida em quatro isolados com o mesmo tamanho relativo, mas com menor variação nos valores das freqüências gênicas. O valor de heterozigoto observado é inferior ao esperado, e a magnitude de σ^2 , que mede o grau de diferenciação, é menor que os valores verificados nas situações A e B, descritas anteriormente.

4.8. Índice de Fixação

4.8.1. Relação entre freqüências de heterozigotos

Quando há dois alelos para um determinado loco, qualquer desvio da condição de equilíbrio de Hardy-Weinberg pode ser medido pelo parâmetro F, denominado de índice de fixação (WRIGHT, 1951, 1965). Este parâmetro mede a probabilidade de um indivíduo da população ser homozigoto por autozigose de um gene qualquer presente em seus ancestrais. Assim, considerando uma população constituída por indivíduos A₁A₁, A₁A₂ e A₂A₂, é possível admitir que entre A₁A₁ (ou para A₂A₂) devem ser encontrados autozigotos, em que os alelos A₁ são idênticos

por ascendência, e alozigotos, em que os alelos são idênticos por funcionalidade, mas não por ancestralidade. Indivíduos heterozigotos (A_1A_2) são alozigotos.

Dessa forma, se numa população os alelos A_1 e A_2 , ocorrem na freqüência p e $(1-p)$ e há uma taxa F de acasalamento entre indivíduos parentados, podem-se calcular as seguintes freqüências genotípicas:

$$\begin{aligned}
 f(A_1A_1) &= f(A_1A_1 \text{ autozigotos}) + f(A_1A_1 \text{ alozigotos}) \\
 &= F f(A_1) + (1-F) [f(A_1)/\text{macho } f(A_1)/\text{fêmea}] = pF + (1-F) p^2 \\
 f(A_2A_2) &= f(A_2A_2 \text{ autozigotos}) + f(A_2A_2 \text{ alozigotos}) \\
 &= F f(A_2) + (1-F) [f(A_2)/\text{macho } f(A_2)/\text{fêmea}] = (1-p)F + (1-F) (1-p)^2 \\
 f(A_1A_2) &= f(A_1A_2 \text{ alozigotos}) \\
 &= (1-F) [f(A_1)/\text{macho } f(A_2)/\text{fêmea}] + (1-F) [f(A_2)/\text{macho } f(A_1)/\text{fêmea}] \\
 &= 2(1-F)p(1-p)
 \end{aligned}$$

Assim, tem-se a seguinte freqüência genotípica:

Genótipo	Freqüência	Panmítico	Fixação	Total	$F=0$	$F=1$
O						
A_1A_1	X_{11}	$(1-F)p^2$	Fp	$(1-F)p^2 + Fp$	$(1-F)p^2$	p
A_1A_2	X_{12}	$2(1-F)pq$		$2(1-F)pq$	$2pq$	0
A_2A_2	X_{22}	$(1-F)q^2$	$F(1-p)$	$(1-F)q^2 + Fq$	q^2	q

$$q = 1 - p$$

O acréscimo na freqüência de homozigotos decorrente de acasalamento entre parentados é dado por:

$$\varepsilon = p(1-p)F$$

De maneira prática, pode-se considerar que as populações naturais apresentam diferentes sistemas reprodutivos, podendo-se classificá-los como alogamia ou panmixia, em que F é igual a zero; autogamia, em que F é igual a 1; e sistema misto, em que $0 < F < 1$.

A seguir é apresentada a freqüência genotípica esperada em uma população, considerando os diferentes tipos de acasalamento e admitindo que as freqüências alélicas são iguais a 0,5 ($p = q = 0,5$):

Genótipo	Freqüência ($F=0$)	Panmixia ($F=1$)	Autogamia ($F=0,02$)	Misto ($F=0,20$)
A_1A_1	$p^2 + \varepsilon$	0,25	0,50	0,255
A_1A_2	$2pq - 2\varepsilon$	0,50	0	0,490
A_2A_2	$q^2 + \varepsilon$	0,25	0,50	0,255

$$q = 1 - p \quad \text{e} \quad \varepsilon = pqF$$

O índice de fixação (F) pode ser estimado a partir da diferença entre as freqüências esperadas (h_e) e observadas (h_o) de heterozigotos, ou seja:

$$h_o = X_{12} = 2pq - 2\varepsilon$$

$$h_e = 2pq$$

Assim:

$$\varepsilon = \frac{h_e - h_o}{2} = pqF$$

logo:

$$F = \frac{h_e - h_o}{2pq} = \frac{h_e - h_o}{h_e}$$

Este valor de índice de fixação pode, em alguns casos, ser negativo.

4.8.2. Diferenciação entre subpopulações

O parâmetro F pode agora ser definido como um índice de fixação total capaz de computar o acréscimo na freqüência de homozigotos por acasalamento entre parentados e por efeito da subdivisão da população. Assim, têm-se as seguintes freqüências genotípicas esperadas:

Genótipo	Freqüência Observada	Freqüência	Freqüência esperada
		Fixação	Princípio Wahlund
A ₁ A ₁	X ₁₁	(1-F) $\bar{p}^2 + F \bar{p}$	$\sum_{j=1}^s \theta_j p_j^2 = \bar{p}^2 + \sigma^2$
A ₁ A ₂	X ₁₂	2(1-F) $\bar{p}\bar{q}$	$2\sum_{j=1}^s \theta_j p_j (1-p_j) = 2\bar{p}\bar{q} - 2\sigma^2$
A ₂ A ₂	X ₂₂	(1-F) $\bar{q}^2 + F \bar{q}$	$\sum_{j=1}^s \theta_j (1-p_j)^2 = \bar{q}^2 + \sigma^2$

Comparando os dois valores esperados, pode-se estimar F por meio de:

$$F = \frac{\sigma^2}{\bar{p} (1 - \bar{q})}$$

O valor de F é definido por: $0 \leq F \leq 1$. Assim, F será zero quando a freqüência p_j ($j=1,2,\dots,s$) é a mesma para todas as subpopulações, de forma que σ^2 é nulo. E F será igual a 1 quando uma fração t das subpopulações estiver fixada para o alelo A₁, e uma fração s-t, para A₂. Assim, tem-se:

Subpopulação fixada com	Freqüência	f(A1)
A ₁	t/s = m	1
A ₂	(s-t)/s = 1-m	0

Neste caso, tem-se:

$$\bar{x} = \frac{t(1) + (s-t)0}{s} = m$$

$$\sigma^2 = m1^2 + (1-m)0^2 - m^2 = m(1-m) = \bar{p}(1-\bar{p})$$

logo:

$$F = \frac{\sigma^2}{\bar{p} (1 - \bar{p})} = \frac{\bar{p} (1 - \bar{p})}{\bar{p} (1 - \bar{p})} = 1$$

Nas expressões anteriores, é assumido que o valor θ_j é conhecido. Entretanto, na prática há disponibilidade apenas da informação do tamanho da amostra estudada, mas raramente é conhecido o valor real do tamanho da subpopulação. Para contornar essa situação, geralmente é admitido que:

$$\theta_1 = \theta_2 = \dots = \theta_s = \frac{1}{s}$$

Essa pressuposição é razoável, uma vez que o tamanho da população (ou subpopulação) é transitório e, de fato, o que realmente interessa é caracterizar a diferenciação genética entre as subpopulações, independentemente de seu tamanho naquele determinado espaço temporal ou geográfico.

Pelo exposto até então, admitindo a avaliação de populações submetidas a diferentes sistemas de acasalamento, pode-se inferir que o grau de fixação pode ser quantificado por meio das seguintes estatísticas:

- a) Grau de fixação de populações submetidas a sucessivas gerações de autofecundação

A fixação pode ser estimada pelo coeficiente de endogamia F ou pela freqüência de heterozigotos (H) dados por:

$$F_t = 1 - \left(\frac{1}{2}\right)^t \quad \text{ou} \quad H_t = \left(\frac{1}{2}\right)^t H_0$$

Em que:

t é o número de gerações de autofecundação

H_0 é a freqüência de heterozigotos na população panmítica.

- b) Grau de fixação de populações derivadas de acasalamento ao acaso a partir do intercruzamento de um número N finito de genitores.

Neste caso, a fixação é medida por meio de:

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t$$

- c) Grau de fixação em populações derivadas de amostras, de tamanho N , sucessivas de uma população panmítica

Quantifica-se a fixação por meio da variância de freqüência alélica, dada por:

$$V(p_t) = p_0(1-p_0) \left[1 - \left(1 - \frac{1}{2N} \right)^t \right]$$

- d) Grau de fixação de um conjunto de amostras resultante da subdivisão de uma população.

A fixação poderá ser estimada por meio de:

$$F = \frac{\sigma^2}{\bar{p}(1-\bar{p})}$$

sendo σ^2 a variância da freqüência alélica entre as amostras.

Fixação gênica – casos especiais

Alelos múltiplos

As expressões apresentadas podem ser facilmente generalizadas nas situações em que há alelos múltiplos, fato comumente encontrado em estudos de diversidade genética de populações a partir de marcadores microssatélites. Nessa situação, serão utilizadas as seguintes simbologias:

- Para as subpopulações

p_{ij} : freqüência do alelo i ($i=1,2,\dots,a$) na subpopulação j ($j=1,2,\dots,s$)

G_{ij} : freqüência observada do genótipo A_iA_i na j -ésima subpopulação

G_{ikj} : freqüência observada do genótipo A_iA_k na j -ésima subpopulação

- Para a população geral

X_{ii} : freqüência observada do genótipo A_iA_i na população total

X_{ik} : freqüência observada do genótipo A_iA_k na população total

sendo:

$$G_{ij} = p_{ij}^2$$

$$G_{ikj} = 2p_{ij}p_{kj}$$

e

$$X_{ii} = \sum_{j=1}^s \theta_j p_{ij}^2 = \bar{p}_i^2 + \sigma_i^2$$

$$X_{ij} = 2 \sum_{j=1}^s \theta_j p_{ij} p_{kj} = 2\bar{p}_i \bar{p}_k + \sigma_{ik}$$

em que:

$$\bar{p}_i = \sum_{j=1}^s \theta_j p_{ij}$$

$$\sigma_i^2 = \sum_{j=1}^s \theta_j (p_{ij} - \bar{p}_i)^2$$

$$\sigma_{ij} = \sum_{j=1}^s \theta_j (p_{ij} - \bar{p}_i)(p_{kj} - \bar{p}_k)$$

De maneira análoga, podem-se expressar as freqüências genotípicas dos homozigotos e heterozigotos a partir de índices de fixação (F), de forma que se tenha:

$$X_{ii} = (1 - F_{ii})\bar{p}_i^2 + F_{ii}\bar{p}_i$$

$$X_{ik} = 2(1 - F_{ik})\bar{p}_i \bar{p}_k$$

de onde se conclui que:

$$F_{ii} = \frac{\sigma_i^2}{\bar{p}_i(1 - \bar{p}_i)}$$

$$F_{ik} = \frac{-\sigma_{ik}}{\bar{p}_i \bar{p}_k}$$

Se a diferenciação entre as subpopulações ocorrem ao acaso, é esperado que:

$$F_{ii} = F_{ik} = F$$

Assim, pelos estudos baseados em locos com vários alelos, é possível testar a hipótese de diferenciação casual ou não das subpopulações. No entanto, deve-se ter em mente que, se há grande número de alelos por locos e o tamanho da

amostra (ou subpopulação) é reduzido, os valores de F (F_{ii} ou F_{ik}) estarão sujeitos a grandes erros de amostragem, e a avaliação da hipótese poderá estar comprometida.

Análise de vários locos

Nas expressões apresentadas foi considerado apenas um único loco, mas no estudo da variação genética entre subpopulações ou populações é indispensável avaliar grande número de locos, representativos do genoma. Por outro lado, dado o tamanho do genoma, é quase impossível caracterizar as populações a partir de todos os locos da espécie; por isso, na prática, utiliza-se uma amostra adequada de locos.

Quando muitos locos são estudados considera-se a diversidade gênica média, ou seja, os valores médios são obtidos a partir das informações de cada loco.

4.9. Fluxo Gênico ou Migração

É a transferência de material genético de uma espécie (daninha ou cultivada) para outras plantas (da mesma espécie ou de espécies diferentes), que podem apresentar, após esse processo, novas expressões fenotípicas, que poderão conferir à população maior ou menor adaptação (ARIAS; RIESEBERG, 1994). O fluxo gênico é também conhecido como escape gênico e está associado ao termo poluição gênica, que seria uma dispersão descontrolada de genes. A migração pode resultar em importantes mudanças nas freqüências dos genes, de forma que a imigração pode provocar adição de material genético novo ao conjunto gênico estabelecido de uma espécie em particular, enquanto a emigração pode resultar na remoção de material genético.

Vários fatores afetam a taxa de fluxo gênico entre diferentes populações. Um dos mais significativos é a mobilidade, e animais tendem a ser mais móveis que plantas. A maior mobilidade de um indivíduo tende a lhe dar maior potencial migratório. Em vegetais, o fluxo gênico ocorre via pólen, sendo dependente de

vários fatores: sincronismo de florescimento, elevada compatibilidade, abundância de vetores e métodos de difusão de pólen, distância de movimentação do pólen e condições ambientais apropriadas para polinização cruzada (CARPENTER et al., 2002).

A estrutura genética de uma espécie reflete o número de alelos intercambiados entre populações, e esses fluxos gênicos tornam homogêneas as freqüências de alelos entre essas populações, determinando os efeitos relativos da seleção e da deriva genética e, consequentemente, a composição genética dos indivíduos. Um fluxo gênico alto evita a adaptação local, reduzindo a fixação de alelos que são favorecidos sob condições locais e impedindo o processo de especiação. Por outro lado, o fluxo gênico gera novos polimorfismos nas populações e aumenta o tamanho efetivo da população local e sua habilidade de resistir a mudanças aleatórias nas freqüências gênicas, opondo-se à deriva genética e gerando novas combinações de genes, nas quais a seleção natural pode atuar (BALLOUX et al., 2002).

As estimativas do fluxo gênico entre as populações foram obtidas segundo a equação proposta por Crow e Aoki (1984):

$$\eta_m = \frac{1}{4k} \left(\frac{1}{F_{ST}} - 1 \right)$$

em que:

F_{ST} é uma medida de diferenciação entre populações que será abordada no próximo capítulo.

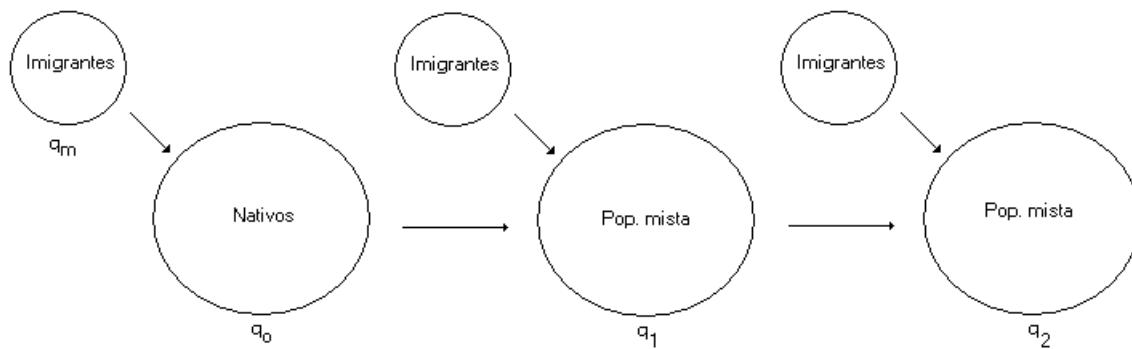
$$k = [p / (p - 1)]^2,$$

sendo η_m o número de migrantes e p o número de populações.

O tamanho da vizinhança é estimado por:

$$\tau_v = 2p\eta_m$$

O esquema a seguir ilustra o feito da imigração sobre uma população. Neste esquema, considera-se q uma população mista formada a partir de uma proporção m de migrantes e 1-m de nativos.



A alteração na freqüência gênica pode ser observada, considerando as gerações:

a) Primeira geração de migração

O valor da freqüência da população que inclui nativos e migrantes é dado por:

$$q_1 = mq_m + (1-m)q_0 = m(q_m - q_0) + q_0$$

Podem ser obtidas duas medidas de desequilíbrio: Uma em relação à população de nativos originais, denotada por Δ_q , e outra em relação ao conjunto de migrantes, denotada por Δ'_q . Esses valores são dados por:

$$\Delta q_1 = (q_1 - q_0) = m(q_m - q_0)$$

e

$$\Delta'_q = (q_1 - q_m) = (1-m)(q_0 - q_m)$$

b) Segunda geração de migração

Agora, o valor da freqüência da população, que inclui a população mista e novos migrantes, é dado por:

$$q_2 = mq_m + (1-m)q_1 = m(q_m - q_1) + q_1$$

As taxas de desequilíbrio são dadas por:

$$\Delta q_2 = (q_2 - q_1) = m(q_m - q_1) = m(1-m)(q_m - q_0)$$

e

$$\Delta q'_2 = (q_2 - q_m) = (1-m)(q_1 - q_m) = (1-m)^2(q_0 - q_m)$$

c) Demais gerações

De forma análoga é obtido:

$$q_3 = mq_m + (1-m)q_2 = m(q_m - q_2) + q_3$$

$$\Delta q_3 = (q_3 - q_2) = m(q_m - q_2) = m(1-m)^2(q_m - q_0)$$

e

$$\Delta q'_3 = (q_3 - q_m) = (1-m)(q_2 - q_m) = (1-m)^3(q_0 - q_m)$$

Assim, na t -ésima geração, tem-se:

$$\Delta q_t = (q_t - q_{t-1}) = m(1-m)^t(q_m - q_0)$$

e

$$\Delta q'_t = (q_t - q_m) = (1-m)(q_{t-1} - q_m) = (1-m)^t(q_0 - q_m)$$

Após sucessivas gerações, espera-se que $\Delta q_t = \Delta q'_t = 0$, de forma que as freqüências gênicas se estabilizem em valor igual a q_m .

De maneira geral, devem ser considerados dois tipos de fluxos gênicos:

a) Fluxo gênico vertical

Ocorre entre plantas e variedades da mesma espécie ou entre espécies aparentadas. Esse fenômeno vem ocorrendo há milhares de anos, com efeitos no geral benéficos. Os cruzamentos que ocorrem entre as plantas de diferentes

populações são características da espécie, e todo e qualquer fluxo gênico altera a constituição e estrutura genética da população receptora.

b) Fluxo gênico horizontal

Corresponde à transferência de genes entre espécies diferentes. Esse fato se dá em razão de a compatibilidade sexual ser encontrada entre muitas espécies vegetais. A probabilidade de fluxo gênico horizontal depende de muitos fatores, como a dinâmica das populações envolvidas, os mecanismos de polinização e dispersão das sementes e o ambiente da liberação. A imposição de mecanismos de isolamento, espaciais ou temporais, é fundamental, sendo porém importante o monitoramento de possíveis situações de escape.

Uma grande parte das plantas cultivadas superiores teve origem em cruzamentos tanto intra como interespecíficos. O trigo, por exemplo, é o produto da combinação de três espécies diferentes. O triticale – espécie artificialmente obtida combinando os genomas do trigo e do centeio – combina a produtividade do trigo com a rusticidade do centeio.

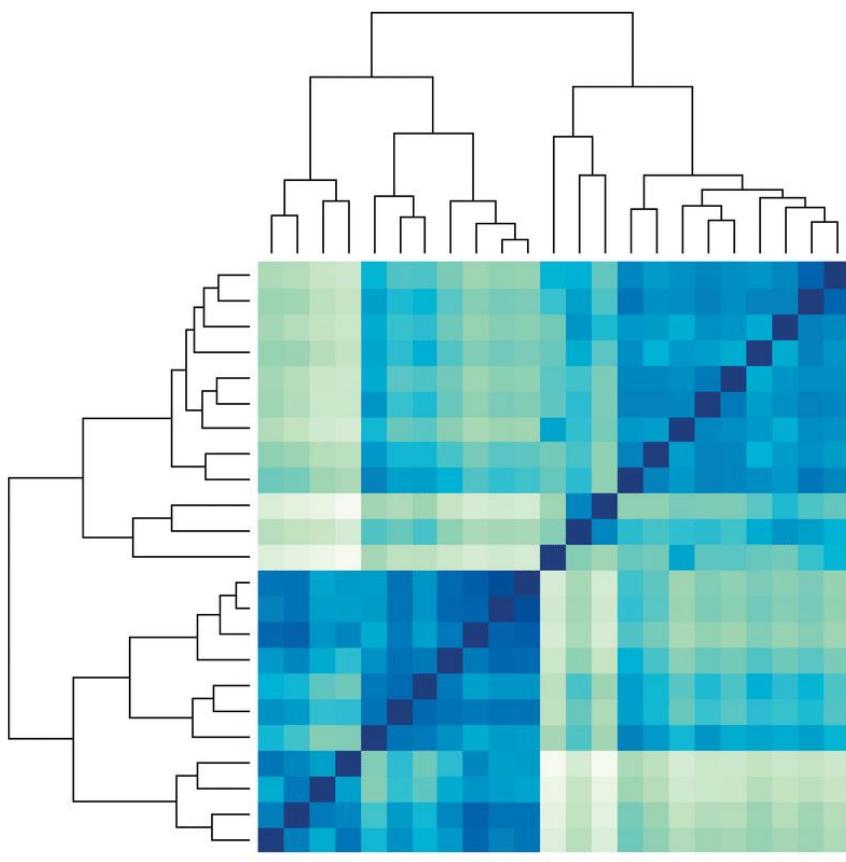
Barreiras ao fluxo gênico

Na natureza, o cruzamento entre espécies diferentes é normalmente impedido pelos mecanismos de isolamento reprodutivo, podendo, entretanto, ocorrer entre espécies evolutivamente mais próximas. As barreiras geográficas interrompem o fluxo gênico, permitindo que as duas populações separadas tomem caminhos evolutivos diferentes, uma vez que os agentes seletivos atuantes são diferentes em ambientes diferentes.

Contudo, em muitas situações são necessárias, para manutenção de variedades livres de contaminação gênica, medidas específicas de isolamento, conforme a natureza das espécies envolvidas. Essas medidas são rotineiramente empregadas pelos pesquisadores, bem como por agricultores interessados em preservar seus materiais genéticos.

Capítulo 5

Diversidade Genética Baseada em Informações Moleculares



5.1. Diversidade em Populações e Coleções

5.1.1 Introdução

Um dos principais objetivos, na área de genética de populações, é entender as causas da diferenciação das populações ao longo da distribuição temporal e geográfica da(s) espécie(s). Nesse sentido, podem-se estudar as forças evolutivas em ação (seleção natural, deriva genética, migração etc.) e perceber os padrões filogenéticos e biogeográficos entre os organismos, através do espaço e do tempo (SILVA; RUSSO, 2000). A maneira como os indivíduos, pertencentes a uma espécie, se distribuem no espaço físico depende principalmente: dos limites estabelecidos pelas variáveis ecológicas; do modo de reprodução e dos mecanismos de dispersão da espécie; dos eventos estocásticos que resultam na formação e extinção de populações ou em variações em seu tamanho efetivo; e das variáveis ambientais, que impõem diferentes coeficientes de seleção a cada genótipo (ROBINSON, 1998).

Populações naturais são unidades sobre as quais incide o manejo para a conservação e utilização dos recursos naturais, bem como a fonte de germoplasma para os programas de melhoramento genético (ROBINSON, 1998). O desenvolvimento de marcadores bioquímicos e moleculares viabilizou estudos fundamentais da estrutura populacional e de seu sistema reprodutivo, indispensáveis para o melhoramento e a conservação de espécies. Várias informações populacionais, como grau de endogamia, sistema reprodutivo predominante, entre outros, podem ser obtidas e são de grande importância na determinação de estratégias de conservação da variabilidade genética na natureza (CARLINI-GARCIA et al., 2001), assim como para sua melhor utilização em programas de melhoramento genético das espécies domesticadas.

Em um trabalho de revisão, Silva e Russo (2000) agruparam as técnicas moleculares, aplicadas à biologia populacional, em quatro categorias principais:

- 1) O primeiro agrupamento comprehende problemas relacionados com a análise da variação genética dentro de indivíduos, abrangendo áreas como heteroplasmia (variação avaliada em mutações no DNA mitocondrial), evolução de famílias multigênicas, assim como problemas relacionados à medicina forense.
- 2) No segundo grupo foram incluídos problemas que envolvem a variação dentro de populações, como efeito da endogamia e do afunilamento genético na variação hereditária, consangüinidade, nepotismo, estrutura social e sucesso reprodutivo.
- 3) Na terceira categoria foram agrupados os estudos diretamente relacionados com a variação genética entre populações, envolvendo problemas como bioinvasão e fluxo gênico, estruturação de estoques naturais em populações de exploração extrativista, bem como problemas relacionados com o *status* taxonômico de morfótipos ou ecótipos.
- 4) Finalmente, na quarta categoria foram agrupados os problemas que envolvem a variação genética acima do nível de espécie, incluindo os efeitos dos diferentes ciclos de vida no processo de diferenciação e isolamento das espécies, na hibridização e estabelecimento de zonas híbridas, bem como na reconstrução filogenética dos táxons.

Para as mais diversas aplicações práticas, o interesse em marcadores moleculares se concentra em quantificar a variabilidade genética, descrever como esta se distribui entre e dentro de populações e como pode ser manipulada (ROBINSON, 1998). Diversas técnicas moleculares estão disponíveis, e todas elas, em certo sentido, representam uma maneira indireta de obter informações a respeito de diferenças nas seqüências genômicas, uma vez que o seqüenciamento completo (que representa a informação genética em um nível de detalhamento maximizado) ainda é muito trabalhoso e caro e, por isso, indisponível para este tipo de estudo (SILVA; RUSSO, 2000). Convém ressaltar também que o material genético utilizado nos trabalhos com marcadores moleculares é, em última análise, uma amostra da população original. Conseqüentemente, a variabilidade contida na

amostra dependerá do polimorfismo e da estrutura genética existente na população e do modo como foi feita a amostragem (ROBINSON, 1998). O polimorfismo acessado, por sua vez, depende em grande parte da técnica molecular adotada. Em qualquer situação, por outro lado, a pressuposição básica é a de que os marcadores moleculares analisados sejam herdáveis, reproduzíveis e independentes (SILVA; RUSSO, 2000).

5.1.2. Diversidade Genética Dentro de Uma População ou Coleção

O desenvolvimento das técnicas de eletroforese de isoenzimas revolucionou os estudos genéticos durante a década de 1950. Mais recentemente, com o advento das técnicas modernas de biologia molecular, surgiram diversos métodos de detecção de polimorfismo em nível de DNA, a exemplo dos marcadores RFLP (*restriction fragment length polymorphism*), RAPD (*random amplified polymorphic DNA*), microssatélites (ou SSR – *single sequence repeat*) e AFLP (*amplified fragment length polymorphism*), que permitiram ampla cobertura do genoma, proporcionando um salto qualitativo e quantitativo em estudos sobre a estrutura das populações e sistema reprodutivo de diversas espécies.

Entretanto, sabe-se muito pouco sobre os mecanismos de especiação e evolução de reprodução de espécies isoladas, apesar de ser relativamente fácil avaliar a extensão da variação genética entre e dentro de populações em nível molecular (NEI; KUMAR, 2000).

No melhoramento genético, quantificar o grau de dissimilaridade entre linhagens e variedades também tem sido primordial. Nesse sentido, os marcadores moleculares têm contribuído, com destaque, na detecção do parentesco genético entre diferentes germoplasmas em bancos de sementes e programas de melhoramento; na predição da heterose; na busca por grupos heteróticos promissores para constituição de híbridos; na identificação de duplicatas nos bancos

de germoplasma; na avaliação do fluxo gênico ao longo do tempo; e na identificação de variedades protegidas.

A maioria dos estudos de diversidade genética se baseia em informações de locos amostrados aleatoriamente em populações não estruturadas hierarquicamente. Dessa maneira, diversas medidas de dissimilaridade (distância) têm sido propostas para verificar o grau de similaridade e a variação genética em amostras de populações, com aplicações variando em nível individual, intrapopulacional e interpopulacional.

As medidas de dissimilaridade podem ser consideradas como estatísticas multivariadas de redução de dados ou informações, ou apenas uma maneira de comparar pares de populações, ou ainda a base para a construção do histórico evolucionário das populações (WEIR, 1996). As estimativas de distância obtidas par a par entre as populações geram uma matriz capaz de proporcionar uma classificação objetiva e estável, tanto quanto possível, das populações estudadas (DIAS, 1998). Algumas dessas medidas compilam as informações de forma binária, cujos dados são codificados em uns e zeros, representando, respectivamente, presença e ausência de um determinado alelo ou marca. Essas medidas têm sido mais utilizadas para informações oriundas de marcadores dominantes, uma vez que não é possível distinguir o genótipo homozigoto dominante do heterozigoto e o cálculo das freqüências alélicas para esse tipo de marcador somente é possível sob condições especiais, ou seja, quando as populações estiverem em equilíbrio em relação ao sistema reprodutivo e for conhecida a taxa natural de fertilização cruzada (ROBINSON, 1998). Já outras medidas de dissimilaridade se baseiam nas freqüências alélicas ou genotípicas, cujo cálculo é realizado com base em locos individuais, sendo a distância final representada pela média das distâncias de cada loco.

Atualmente, há grande quantidade de índices e coeficientes de dissimilaridade disponíveis na literatura e, talvez, a maior dificuldade encontrada pelos pesquisadores seja justamente decidir qual a melhor medida a ser usada de

acordo com seus objetivos. Nenhuma das medidas de dissimilaridade deve ser usada em situações para as quais não são destinadas (WEIR, 1996).

As medidas de dissimilaridade podem ser agrupadas com base na natureza dos dados moleculares da seguinte forma:

- i. Medidas de dissimilaridade para marcadores dominantes, cuja informação alélica é codificada de forma binária em presença e ausência da banda, comumente conhecidas como coeficientes de (dis)similaridade
- ii. Medidas de dissimilaridade para marcadores co-dominantes, cuja informação de freqüência alélica é possível de ser acessada, comumente conhecidas como distâncias genéticas (e/ou geométricas) e distâncias genotípicas.

Em muitas situações, há o interesse em caracterizar a diversidade genética dentro de uma população a partir das informações moleculares de um conjunto de indivíduos que a constituem. Uma situação similar é aquela em que há também o interesse de avaliar a diversidade genética de um conjunto de acessos que constituem um grupo de interesse, tal como se verifica em bancos de germoplasma. Neste último caso, a unidade de inferência denomina-se acesso, ao contrário do primeiro caso, em que se tem o indivíduo como menor unidade representativa da população.

Quando se tem uma população, a diversidade avaliada reflete a variabilidade genotípica, muitas vezes expressa em termos de freqüência gênica, cujos fatores que a alteram são objetos de estudo para entendimento dos mecanismos evolutivos que sobre ela incide.

Para estudar a diversidade da população, considera-se ser possível dispor de informações moleculares de natureza dominante; assim, cada genótipo pode ser caracterizado por um sistema binário, ou de informações co-dominantes e multialélicas, como ocorre em algumas situações em que se dispõe de informações de marcadores microssatélites. Para ilustração e considerações metodológicas, serão considerados dois conjuntos de dados, relativos a duas populações com 50 acessos (ou indivíduos) cada. Na população 1, as informações são relativas a

marcadores dominantes descritos por código binário, sendo utilizado código 1 para representar a ocorrência de bandas e código 0 para caracterizar a ausência destas. Para a população 2, a caracterização dos indivíduos foi realizada por marcadores co-dominantes multialélicos, de forma que a descrição genotípica foi feita por códigos numéricos informativos dos alelos que possuem. Assim, considerando a existência de três alelos (A1, A2 e A3), têm-se os genótipos homozigotos descritos por 11, 22 e 33 e genótipos heterozigotos descritos por 12, 13 e 23. Dados ilustrativos são apresentados nas Tabelas A2 e A3.

5.2. Diversidade Entre Acessos ou Indivíduos

5.2.1. Dados de marcadores dominantes

Codificação da informação de marcadores co-dominantes em populações

Uma característica inerente aos marcadores dominantes (como RAPD e AFLP) é a de não apresentar sensibilidade quantitativa suficiente para discriminar o genótipo homozigoto “dominante” (AA) do heterozigoto (Aa), razão pela qual tais genótipos são alocados na mesma classe fenotípica, isto é apresentam a banda no gel de agarose, enquanto o genótipo homozigoto “recessivo” (aa) é identificado pela ausência da banda no gel (fenótipo nulo). As informações são descritas por código binário, sendo utilizado código 1 para representar a ocorrência de bandas e código 0 para caracterizar a ausência destas

Estudo da diversidade entre indivíduos dentro da população

Para o estudo da diversidade dentro da população ou entre acessos, a partir de informações de marcadores dominantes, pode-se adotar técnicas de agrupamento ou de projeção de medidas de dissimilaridade. Em ambos os casos, torna-se necessária a matriz de dissimilaridade entre pares de indivíduos por meio de índices apropriado.

a. Coeficientes de similaridade e de dissimilaridade

Os coeficientes de similaridade (S), originalmente propostos para estudos de taxonomia numérica, utilizam esse tipo de informação molecular como sendo variáveis binárias, codificadas como 0 na ausência da banda e 1 na presença. Nessa situação, os coeficientes de similaridade entre pares de populações (sejam elas materiais exóticos, linhagens, variedades, espécies etc.) são obtidos levando-se em conta:

- a: número de coincidências do tipo 1-1 para cada par de populações;
- b: número de discordâncias do tipo 1-0 para cada par de populações;
- c: número de discordâncias do tipo 0-1 para cada par de populações; e
- d: número de coincidências do tipo 0-0 para cada par de populações.

A partir da Tabela 5.1 são apresentados os principais coeficientes de similaridade descritos na literatura (SNEATH; SOKAL, 1973; ROMESBURG, 1984).

Tabela 5.1 - Expressões de alguns dos coeficientes de similaridade (S)

Coeficientes	Expressão	Intervalo
Coincidência simples	$S_{CS} = \frac{a+d}{a+b+c+d}$	0, 1*
Sokal e Sneath	$S_{SS} = \frac{2(a+d)}{2(a+d)+b+c}$	0, 1
Jaccard	$S_J = \frac{a}{a+b+c}$	0, 1
Sorensen-Dice ou Nei e Li	$S_{SD} = \frac{2a}{2a+b+c}$	0, 1
Ochiai I	$S_O = \frac{a}{\sqrt{(a+b)(a+c)}}$	0, 1
Ochiai II	$S_{OII} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	0, 1
Andenberg	$S_A = \frac{a}{a+2(b+c)}$	0, 1
Rogers e Tanimoto	$S_{RT} = \frac{a+d}{a+2(b+c)+d}$	0, 1
Russel e Rao	$S_{RR} = \frac{a}{a+b+c}$	0, 1
Kulczynski	$S_K = \frac{1}{2} \left[\frac{a}{a+b} + \frac{a}{a+c} \right]$	0, 1
Yule	$S_Y = \frac{ad-bc}{ad+bc}$	-1, 1
Hamman	$S_H = \frac{(a+d)-(b+c)}{a+b+c+d}$	-1, 1
Phi	$S_P = \frac{ad-bc}{\sqrt{(a+b)+(a+c)+(b+d)+(c+d)}}$	-1, 1

* Corresponde à máxima similaridade entre dois indivíduos.

Como se trata de medidas de similaridade, é recomendável, em análises de agrupamento, fazer uso de medidas de dissimilaridade, definidas por:

- a) $D = 1 - S$.
- b) $D = 1 / (k + S)$. Nesta expressão, a inclusão da constante k (geralmente $k=1$) visa contornar os problemas de indeterminação ocasionados quando o valor de S é nulo. Nesta situação, se o valor de S varia de 0 a 1, o valor de D variará de 1 a 0,5, para k igual a 1.
- c) $D = \sqrt{1-S}$ ou $\sqrt{2(1-S)}$. Esta expressão atribui a algumas medidas de dissimilaridade propriedades euclidianas, com exceção do coeficiente de Yule, que não é métrico (GOWER; LEGENDRE, 1986).

Outra medida de dissimilaridade é a própria distância euclidiana, dada por:

$$D_{EU} = \sqrt{\frac{b+c}{a+b+c+d}}$$

Esta medida é conhecida como a distância binária de Sokal. Naturalmente que todos os coeficientes euclidianos são métricos, embora nem todo coeficiente métrico tenha propriedades euclidianas (JACKSON et al., 1989).

Na prática, os coeficientes de similaridade e dissimilaridade mais utilizados nos trabalhos de diversidade genética têm sido o de coincidência simples (SNEATH; SOKAL, 1973), o de Jaccard (JACCARD, 1908) e o de Nei e Li (NEI; LI, 1979). O de coincidência simples tem, no seu complemento aritmético, a vantagem de ser idêntico ao quadrado da distância euclidiana média (D_{ii}^2) e, uma vez aplicada a raiz quadrada($\sqrt{D_{ii}^2}$), torna-se a própria distância binária de Sokal.

Em contraste com o coeficiente de coincidência simples, os coeficientes de Jaccard e Nei e Li apresentam a vantagem de não considerarem coincidências do tipo 0-0. Por exemplo, se existir alta probabilidade de não-amplificação de bandas e a ausência delas, em ambas as populações, não

puder ser interpretada como uma característica comum, ou seja, não significa necessariamente regiões do DNA idênticas, é mais adequado aplicar coeficientes que excluem a co-ocorrência negativa. Alguns autores têm destacado esse aspecto (DUARTE et al., 1999; EMYGDIO et al., 2003; ARIEL et al., 2004). Nesse contexto, a escolha de qualquer um dos coeficientes seja – ele o de Jaccard, Nei e Li, Ochiai I, Kulczynski e Andenberg – é recomendada, com destaque para o índice de Jaccard, com interpretação facilitada por representar a razão entre o número de coincidências e o número total de bandas, excluindo a coincidência negativa (MEYER et al. 2004). No entanto, não se devem descartar as medidas de similaridade que levam em consideração a concordância negativa, pois em determinados situações, como é caso de populações F_2 , a ausência da marca pode ser tratada como um fator de similaridade entre as duas populações a serem comparadas.

Alguns trabalhos compararam a eficiência dos coeficientes de similaridade em expressar o grau de divergência genética em espécies vegetais com marcadores RAPD e AFLP (DUARTE et al., 1999; EMYGDIO, 2003; MEYER et al., 2003; ARRIEL et al., 2004). A conclusão geral obtida pelos autores foi de que o número de grupos formados altera conforme o coeficiente de similaridade utilizado, e aqueles que consideram a co-ocorrência negativa (0-0) produzem resultados bastante discrepantes dos demais, não sendo sugerido seu uso na avaliação da divergência genética.

Tem sido consenso a idéia de que, ao se trabalhar com materiais exóticos ou envolvendo espécies diferenciadas, a princípio pouco relacionadas, o adequado seria adotar o coeficiente de Nei e Li. Contudo, se o estudo for feito dentro de uma população, ou espécie, em que coincidências de ocorrência de bandas podem ser admitidas como um fenômeno esperado, recomenda-se a utilização do coeficiente de Jaccard.

Ilustração 1

Será apresentada, como exemplo, a análise da dissimilaridade entre cinco acessos e vinte marcadores moleculares. A partir da informação extraída no gel de agarose dos marcadores dominantes é estabelecida uma tabela de dupla entrada, cuja presença da marca foi caracterizada pelo número 1 e a ausência pelo número 0, de acordo com a Tabela 5.2.

Tabela 5.2 - Presença e ausência de 20 marcadores moleculares, avaliadas em cinco acessos

Acessos	Marcadores																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	0	1	0	1	0	0	0	1	1	0	0	1	1	1	0	1	0	1	0
2	1	0	1	1	0	1	1	1	0	0	0	1	0	1	1	1	1	1	1	0
3	0	1	1	0	1	0	0	0	1	1	1	0	0	1	1	1	0	1	0	0
4	1	1	1	0	0	1	0	0	0	1	1	1	0	0	1	0	1	0	1	0
5	1	1	1	0	1	0	1	0	0	1	0	0	0	1	1	0	1	1	1	0

Conforme as informações da Tabela 5.2, foram obtidos os coeficientes de similaridade de coincidência simples, Jaccard e Nei e Li, entre os acessos 1 e 2, calculados por:

Complemento aritmético do coeficiente de coincidência simples

$$d_{12} = 1 - \frac{a+d}{a+b+c+d} = 1 - \frac{7+4}{7+3+6+4} = 0,45$$

Complemento aritmético do coeficiente de Jaccard

$$d_{12} = 1 - \frac{a}{a+b+c} = 1 - \frac{7}{7+3+6} = 0,56$$

Complemento aritmético do coeficiente de Nei e Li

$$d_{12} = 1 - \frac{2a}{2a+b+c} = 1 - \frac{2.7}{2.7+3+6} = 0,39$$

As demais estimativas dos coeficientes de dissimilaridade encontram-se na Tabela 5.3.

Tabela 5.3 - Medidas de dissimilaridade entre cinco acessos, estimadas a partir de 20 marcas moleculares

Acessos	a (1-1)	b (1-0)	c (0-1)	D (0-0)	Coincidência		
					simples*	Jaccard*	Nei e Li*
1 e 2	7	3	6	4	0,45	0,56	0,39
1 e 3	6	4	4	6	0,40	0,57	0,40
1 e 4	6	4	4	6	0,40	0,57	0,40
1 e 5	8	2	3	7	0,25	0,38	0,24
2 e 3	5	8	5	2	0,65	0,72	0,57
2 e 4	6	7	4	3	0,55	0,65	0,48
2 e 5	7	6	4	3	0,50	0,59	0,42
3 e 4	5	5	5	5	0,50	0,67	0,50
3 e 5	7	3	4	6	0,35	0,50	0,33
4 e 5	7	3	4	6	0,35	0,50	0,33

*Complemento aritmético do respectivo coeficiente de similaridade.

Os três coeficientes designaram os acessos 1 e 5 como os mais similares e os acessos 2 e 3 como os mais divergentes. A correlação de Spearman entre as medidas de dissimilaridade, dadas pelo complemento de Jaccard e de Nei e Li, foi igual a 1 e entre as medidas destes complementos com a obtida pelo complemento da coincidência simples foi de 0,9272.

Aplicação

Para o exemplo de avaliação da população 1 (Tabela A2), podem-se ilustrar os valores da dissimilaridade entre os 10 primeiros acessos, expressos pelo complemento aritmético do coeficiente de Jaccard, conforme apresentados na Tabela 5.4. Constatase que os acessos 4 e 9 apresentam o mesmo padrão de ocorrência de bandas nos dez locos avaliados. Os acessos 3 e 10 foram os mais dissimilares, com valor igual a 0,778.

Tabela 5.4 - Medidas de dissimilaridade (complemento aritmético do índice de Jaccard) entre dez acessos, estimadas a partir de dez marcadores moleculares (locos) (dados originais disponíveis na Tabela A2)

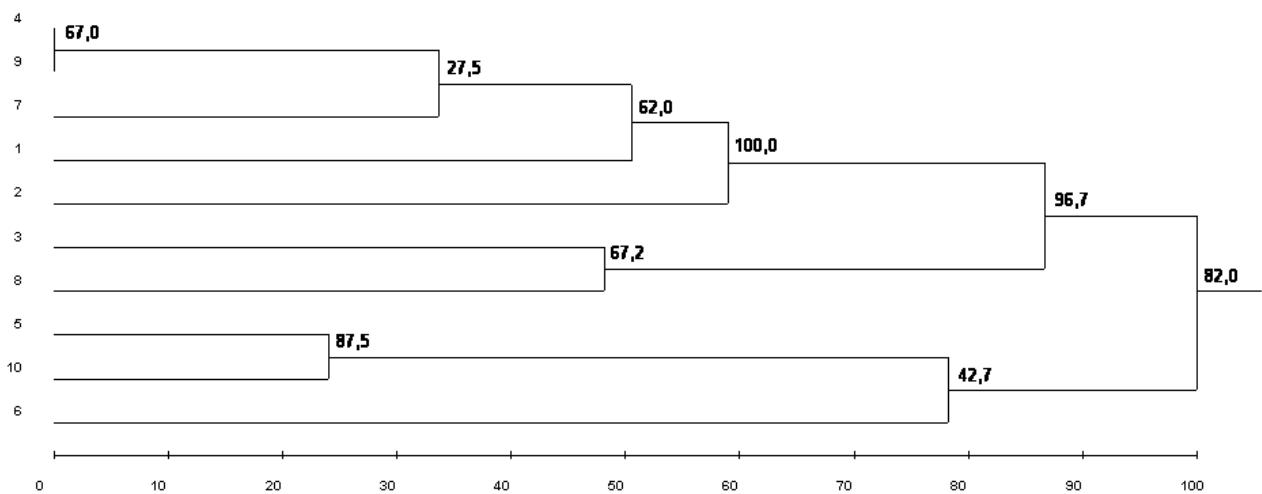
Acessos	A	b	c	d	$D_{ij''}$	Acessos	a	b	c	d	$D_{ij''}$
1 2	2	1	1	6	0,500	4 5	3	1	4	2	0,625
1 3	3	0	2	5	0,400	4 6	3	1	2	4	0,500
1 4	3	0	1	6	0,250	4 7	4	0	1	5	0,200
1 5	2	1	5	2	0,750	4 8	4	0	3	3	0,429
1 6	3	0	2	5	0,400	4 9	4	0	0	6	0,000
1 7	3	0	2	5	0,400	4 10	3	1	3	3	0,571
1 8	3	0	4	3	0,571	5 6	4	3	1	2	0,500
1 9	3	0	1	6	0,250	5 7	4	3	1	2	0,500
1 10	2	1	4	3	0,714	5 8	4	3	3	0	0,600
2 3	2	1	3	4	0,667	5 9	3	4	1	2	0,625
2 4	3	0	1	6	0,250	5 10	6	1	0	3	0,143
2 5	2	1	5	2	0,750	6 7	4	1	1	4	0,333
2 6	2	1	3	4	0,667	6 8	4	1	3	2	0,500
2 7	3	0	2	5	0,400	6 9	3	2	1	4	0,500
2 8	3	0	4	3	0,571	6 10	4	1	2	3	0,429
2 9	3	0	1	6	0,250	7 8	4	1	3	2	0,500
2 10	2	1	4	3	0,714	7 9	4	1	0	5	0,200
3 4	3	2	1	4	0,500	7 10	4	1	2	3	0,429
3 5	2	3	5	0	0,800	8 9	4	3	0	3	0,429
3 6	3	2	2	3	0,571	8 10	4	3	2	1	0,556
3 7	3	2	2	3	0,571	9 10	3	1	3	3	0,571
3 8	5	0	2	3	0,286						
3 9	3	2	1	4	0,500						
3 10	2	3	4	1	0,778						

b. Análise de agrupamento

Uma matriz de distância envolvendo todos os acessos poderá ser estabelecida e submetida análise de agrupamento por alguma técnica específica. Vários métodos de agrupamento, hierárquicos e de otimização, e de dispersão gráfica poderão ser utilizados, como já descrito no capítulo 2, para características fenotípicas. A base teórica desses métodos já foi apresentada e poderá ser revisada pelo leitor. Assim, as metodologias de agrupamento serão aplicadas aos dados moleculares a título de ilustração, de forma que podem ser considerados os resultados mostrados a seguir.

i. Análise de agrupamento por métodos hierárquicos

Neste tipo de análise interessa estudar a diversidade por meio de um dendrograma no qual as relações entre acessos poderão ser avaliadas por meio das ramificações da árvore obtida. Como ilustração, será considerado o agrupamento dos 10 acessos pertencentes à população 1 (Tabela A2) estudados pela técnica de agrupamento UPGMA, cujo resultado é descrito a seguir:



Constata-se no dendrograma obtido a existência de acessos com 100% de similaridade (4 e 9), sendo indicativo de se tratar de réplicas dentro da população ou indivíduos bem próximos, apesar de a análise ter sido feita apenas com base em 10 locos.

Para melhor interpretação dos resultados obtidos pela análise de agrupamento hierárquica torna-se importante destacar as seguintes informações:

- i. Valor da dissimilaridade obtida no último nível de fusão. Neste exemplo, a dissimilaridade no último nível de fusão foi igual a 0,5931 inferior à dissimilaridade máxima verificada entre os acessos 3 e 10 (igual a 0,778).
- ii. Identificação dos pontos de cortes obtidos por meio de algum critério ótimo de partição. Utilizando o método de Mojema (1977), com valor de $k = 1,25$, são recomendados corte a 79% e 87% de dissimilaridade.

- iii. A consistência dos nós e bifurcações. Neste tipo de estudo é desejável que a consistência do padrão de agrupamento seja avaliada, sendo recomendado, para este fim, o uso da técnica de *bootstrap*. Para este exemplo, consideraram-se 5.000 simulações de forma que a consistência do agrupamento, de cada nó da árvore apresentada, é ilustrada na figura analisada.
- iv. Valor do coeficiente de correlação cofenético. Para este exemplo, a correlação entre as medidas de dissimilaridades originais e aquelas geradas graficamente a partir do dendrograma apresentado é de 0,8266.

ii. Análise de agrupamento por métodos de otimização

Outra alternativa para estudar a existência de grupos homogêneos é por meio de métodos de otimização, sendo um dos mais utilizados o descrito por Tocher, citado por Rao (1952). Este método, quando aplicado à matriz de dissimilaridade expressa pelo complemento aritmético do índice de Jaccard, proporciona o resultado apresentado na Tabela 5.5.

Tabela 5.5 - Medidas de dissimilaridade entre 10 acessos, estimadas a partir de 10 marcadores moleculares (locos)

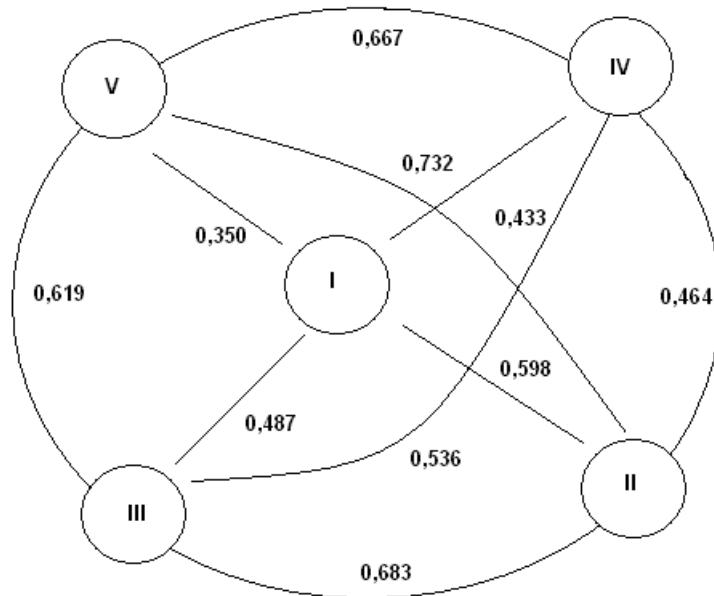
Grupos	Acessos	Soma das distâncias	Média das distâncias
		distâncias	
I	4,9,7 e 1	10,300	0,217
II	5 e 10	0,143	0,143
III	3 e 8	0,286	0,286
IV	6	-	-
V	2	-	-

O resultado obtido pelo método de Tocher é idêntico ao obtido pela técnica hierárquica UPGMA se for admitido ponto de corte no valor de 50% da

dissimilaridade estabelecida no último nível de fusão por esta técnica de agrupamento.

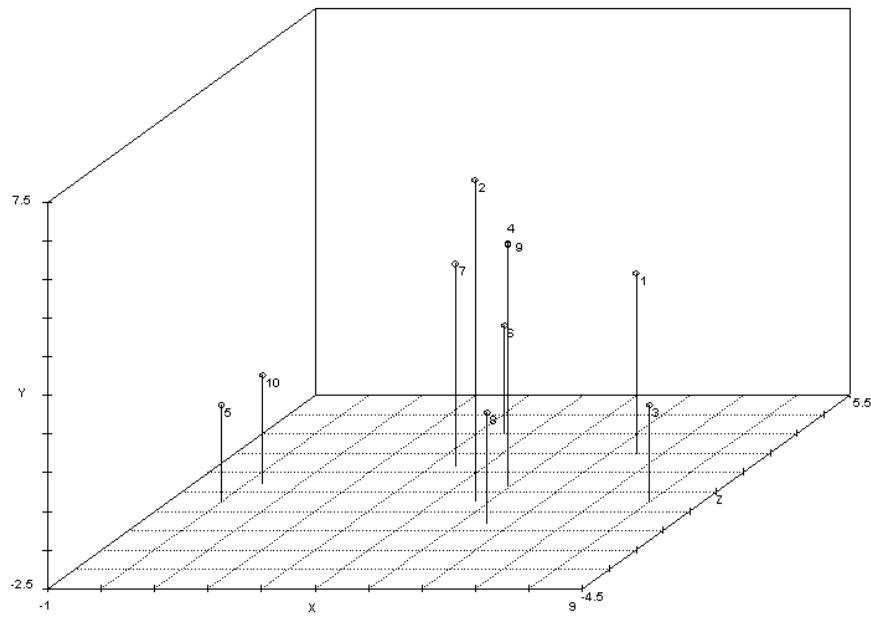
Quando se utiliza a metodologia de agrupamento de Tocher é importante destacar:

- i. O valor da estatística utilizada como critério global para inclusão de genótipos dentro de um grupo. Neste exemplo, o valor obtido foi de 0,3333.
- ii. Os valores das distâncias intergrupos, que devem ser inferiores às distâncias intragrupo. Neste exemplo, os valores das distâncias são representados graficamente por meio de:



iii. Análise por dispersão gráfica

Também é possível apresentar a dissimilaridade entre acessos por meio da dispersão gráfica. Neste caso, recomenda-se o uso da projeção de distâncias em gráficos 2D ou 3D. Para o exemplo considerado, a projeção 3D é:



Para a projeção gráfica, constatam-se as seguintes medidas da qualidade da dispersão:

Projeção	2D	3D
Distorção	14,34%	3,78%
Correlação entre distâncias originais e gráficas	0,910	0,986
Estresse	20,63%	6,94%

A projeção 3D apresenta qualidade de representação gráfica superior à 2D e, portanto, deve ser preferida. Neste exemplo, há considerável decréscimo no grau de distorção e estresse e elevação da correlação, entre as distâncias originais e gráfica, para 0,986. Assim, considera-se que o posicionamento dos escores dos genótipos no gráfico possibilita inferir apropriadamente sobre a diversidade genética e seu padrão de agrupamento.

c. Freqüência alélica na população

Quando se dispõe de informações relativas a marcadores dominantes, a freqüência alélica poderá ser estimada se houver evidência de que os indivíduos avaliados pertencem a uma população que se encontra em equilíbrio de Hardy-

Weinberg. Nesta situação, espera-se que a freqüência de indivíduos com fenótipo dominante se $p^2 + 2pq$ e de fenótipos recessivos dada por q^2 . A pressuposição de equilíbrio poderá ter por base o conhecimento prévio sobre o sistema de acasalamento, tamanho populacional, fatores seletivos, dentre outros.

Assim, para o exemplo em consideração referente à população 1 (Tabela A2), envolvendo todos os 50 indivíduos avaliados, têm-se as seguintes informações:

Marcador	A-	aa	N	$p=f(A)$	$q=f(a)$	$H=2pq$
1	8	42	50	0,084	0,917	0,153
2	13	37	50	0,140	0,860	0,241
3	32	18	50	0,400	0,600	0,480
4	46	4	50	0,717	0,283	0,406
5	15	35	50	0,163	0,837	0,273
6	28	22	50	0,337	0,663	0,447
7	7	43	50	0,073	0,927	0,135
8	49	1	50	0,859	0,141	0,243
9	21	29	50	0,238	0,762	0,363
10	22	28	50	0,252	0,748	0,377

Para o marcador 1, o número de indivíduos com a presença de banda é de 8, e com ausência, igual a 42. Assim, denominando a freqüência de fenótipos dominantes por $D+H$ e de recessivos por R , obtém-se as seguintes relações, pressupondo-se que a população encontra-se em equilíbrio de Hardy-Weinberg:

$$D+H = p^2 + 2pq$$

$$R = q^2$$

Logo, para o marcador 1, tem-se:

$$D+H = 8/50 = 0,16$$

$$R = 42/50 = 0,84$$

portanto:

$$q = \sqrt{R} = \sqrt{0,84} = 0,917 \quad \text{e } p = 1 - q = 0,153$$

A freqüência esperada de heterozigotos em uma população é $2pq$ e, portanto, será de 0,153 para o marcador 1. A freqüência máxima de heterozigotos numa população em equilíbrio de Hardy-Weinberg, considerando dois alelos por loco, é de 0,50, que se verifica quando $p=q = 0,5$. Assim, neste exemplo, constata-se que os locos 4, 6 e, principalmente, o 3 apresentam grau de herozigosidade elevado.

d. Número ótimo de marcadores

Em estudos genéticos, a quantificação da diversidade genética numa população é fundamental para vários propósitos, seja relacionado ao melhoramento, à evolução ou à conservação de germoplasma. Certamente que, quanto maior o número de marcadores avaliados, maior será a possibilidade de quantificar apropriadamente a diversidade entre os acessos. Contudo, na maioria das vezes, os recursos, técnicos e financeiros, são escassos e há necessidade de caracterizar vários acessos e, muitas vezes, várias populações; assim, o estabelecimento de um número ótimo de marcadores, para descrever a diversidade genética, pode ser desejável.

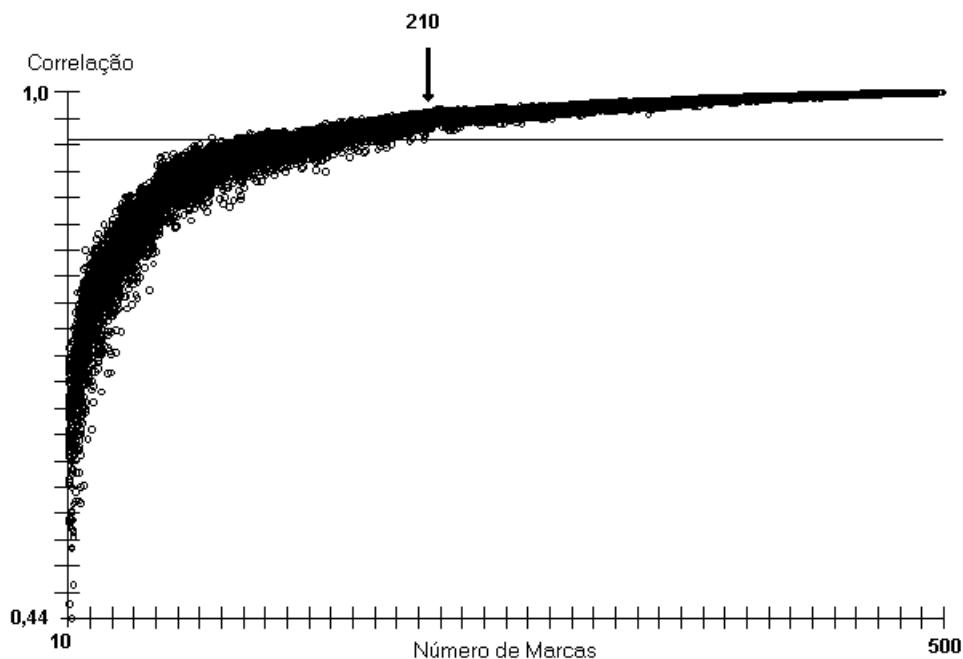
A fim de estabelecer o número adequado de marcadores para representar a diversidade genética de uma população, consideram-se as informações preliminares, em que é avaliada, originalmente, uma amostra adequada da população e quantificada a informação genética de um número elevado de marcadores que amostre adequadamente o genoma da população de interesse. A partir daí, procura-se identificar um número ótimo de marcadores capaz de representar tão bem a diversidade da população quanto aquele inicialmente utilizado.

Como ilustração, foi considerada a análise de uma população constituída por 50 indivíduos, que foram avaliados em relação a 500 marcadores. Para fins de simulação e estabelecimento do número ótimo de marcadores, é estabelecido inicialmente o número de amostras por simulação, que se refere ao número de réplicas (r) de uma determinada simulação, em que se estabelece a adequabilidade

de uma amostra, contendo m' marcadores, representar a diversidade da população original, originalmente caracterizada no experimento por m marcadores ($m' \leq m$).

O processo de simulação consiste em estabelecer um número inicial de marcadores (no exemplo, foi considerado igual a 10). Para cada tamanho, são estabelecidas r réplicas de amostras submetidas à análise da diversidade genética, utilizando-se um índice apropriado de dissimilaridade (no exemplo, foi considerado o complemento aritmético do índice de Jaccard). Também é estabelecido um incremento (Δ) a ser dado aos sucessivos tamanhos, de forma a ser possível estabelecer amostras que variam de um tamanho inicial m'_i até m , numa progressão aritmética de razão Δ .

A análise é feita com amostras de um tamanho inicial (m'_i) até um tamanho final (m), com incremento Δ . Assim, por exemplo, tendo-se uma população, caracterizada originalmente por 500 marcadores, pode-se avaliar a possibilidade de se ter 50 ($r = 50$) amostras de tamanho 10,11,12...499 marcadores ($m'_i = 10$, $m = 499$ e $\Delta = 1$), de forma que serão realizadas 24.500 análises de dados. Obtidas as matrizes de dissimilaridade, com o índice desejado pelo pesquisador, elas são comparadas com a matriz original (obtida com m marcas e n indivíduos). Os valores de cada matriz gerada são correlacionados com os da matriz original, obtendo-se o seguinte gráfico:



Essa análise permite constatar que com 210 marcas, em todas as 50 amostragens realizadas, é possível representar a diversidade genética entre os acessos com grau de concordância satisfatório ao obtido com as 500 marcas originais, visto que a correlação entre as medidas de dissimilaridade obtida com 500 marcas foi acima de 0,95 para todas as 50 amostras de 210 marcas tomadas entre as 500 disponíveis. Portanto, o número 210 é considerado o tamanho ótimo, para substituir as 500 originalmente propostas, para estudo da diversidade.

e. Fingerprint

Fingerprint é um termo utilizado para referenciar uma técnica que permite identificar padrões únicos de cada um de um conjunto de acessos por meio de um número mínimo e ótimo de informações.

Para se realizar a análise, considera-se um conjunto de m informações moleculares relativas a g acessos de um conjunto de interesse. A partir desta informação obtém-se uma matriz de distância D , de dimensão $g \times g$, em que cada elemento é dado por:

$$d_{ii'} = \sum_{j=1}^m \alpha_j$$

em que α_j assume valor 1 se a informação molecular do indivíduo i for diferente da apresentada pelo indivíduo i' e 0 se for igual.

É feita uma investigação preliminar sobre os elementos da matriz D . Se existir algum valor, fora da diagonal de D , nulo conclui-se que há réplicas entre o conjunto de acessos e que a quantidade e, ou a variabilidade, de marcadores utilizados é insuficiente para gerar um padrão específico de identificação para cada acesso. Se não houver réplicas poderemos investigar a possibilidade de estabelecer um número mínimo ótimo de marcadores para estabelecer padrões (*fingerprint*) de cada acesso, conforme descrito a seguir.

- a. Obter a soma das distâncias originais (matriz D) considerando a informação de $g(g-1)/2$ pares de acessos.
- b. Obter m matrizes de distâncias de dimensão gxg (matriz D'_j sendo $j=1,2\dots,m$) pela análise de todos, exceto o j -ésimo, marcadores. Verificar qual marcador que, ao ser excluído, proporciona menor redução na soma das distâncias.
- c. Verificar se existe na matriz D'_k (sendo k o índice identificador do marcador excluído da análise) algum elemento nulo fora da diagonal. Caso não exista, repetir o procedimento descrito no item anterior.
- d. O procedimento finalizará na análise anterior àquela que conduz a uma matriz D'_j contendo pelo menos um valor nulo fora de sua diagonal.

Veja que este procedimento conduzirá a um conjunto $m' < m$ de marcadores capazes de discriminar os acessos, preservando o máximo da diversidade e garantindo que se tenham padrões únicos diferenciáveis. O procedimento fornecerá quantos e quais marcadores serão úteis para o padrão de identificação bem como o próprio padrão de cada acesso, ou seja seu *fingerprint*.

5.2.2. Dados de marcadores co-dominantes

Codificação da informação de marcadores co-dominantes em populações

Com os marcadores co-dominantes, a exemplo das isoenzimas, RFLP e microssatélites, em cada loco estudado é possível identificar genótipos heterozigotos e homozigotos, gerando mais informação em nível genético. Nessa situação, a informação de freqüência alélica se dá em nível populacional, de modo que as freqüências dos alelos são quantificadas a partir dos genótipos amostrados de cada população. Por exemplo, considerando duas populações (1 e 2), amostradas por 10 indivíduos e quatro locos (A, B, C e D), tem-se, inicialmente, a seguinte informação oriunda do gel(Tabela 5.6).

Tabela 5.6 - Planilha de dados de quatro marcas (locos) co-dominantes avaliadas em duas populações (1 e 2), amostradas cada uma por 10 indivíduos

População	Indivíduo	Locos			
		A	B	C*	D
I	1	A ₁ A ₁	B ₃ B ₃	C ₁ C ₁	D ₁ D ₂
	2	A ₁ A ₁	B ₁ B ₃	C ₁ C ₁	D ₂ D ₂
	3	A ₁ A ₁	B ₂ B ₂	C ₁ C ₁	D ₁ D ₁
	4	A ₁ A ₁	B ₂ B ₂	C ₁ C ₁	D ₁ D ₁
	5	A ₁ A ₁	B ₂ B ₃	C ₁ C ₁	D ₁ D ₂
	6	A ₁ A ₁	B ₁ B ₂	C ₁ C ₁	D ₁ D ₂
	7	A ₁ A ₁	B ₁ B ₂	C ₁ C ₁	D ₂ D ₂
	8	A ₁ A ₁	B ₁ B ₂	C ₁ C ₁	D ₁ D ₂
	9	A ₁ A ₁	B ₁ B ₃	C ₁ C ₁	D ₁ D ₂
	10	A ₁ A ₁	B ₁ B ₁	C ₁ C ₁	D ₂ D ₂
II	1	A ₃ A ₃	B ₁ B ₁	C ₁ C ₁	D ₁ D ₁
	2	A ₃ A ₃	B ₃ B ₃	C ₁ C ₁	D ₁ D ₂
	3	A ₁ A ₃	B ₁ B ₃	C ₁ C ₁	D ₁ D ₁
	4	A ₁ A ₃	B ₂ B ₂	C ₁ C ₁	D ₁ D ₂
	5	A ₂ A ₃	B ₁ B ₃	C ₁ C ₁	D ₂ D ₂
	6	A ₁ A ₃	B ₁ B ₁	C ₁ C ₁	D ₁ D ₁
	7	A ₁ A ₃	B ₁ B ₁	C ₁ C ₁	D ₁ D ₂
	8	A ₁ A ₂	B ₁ B ₂	C ₁ C ₁	D ₁ D ₁
	9	A ₁ A ₃	B ₁ B ₃	C ₁ C ₁	D ₁ D ₂
	10	A ₁ A ₃	B ₁ B ₃	C ₁ C ₁	D ₁ D ₂

* O loco C é monomórfico para o alelo C₁ nas duas populações amostradas

A Figura 5.1 ilustra uma auto-radiografia de RFLP cujas amostras com uma banda são indivíduos homozigotos, e as de duas bandas, de indivíduos heterozigotos.



Figura 5.1 - Exemplo de uma auto-radiografia de RFLP em população de milho.

De posse das informações da Tabela 5.6, é possível construir uma tabela de dupla entrada com as freqüências alélicas de cada população, de acordo com a Tabela 5.7.

Tabela 5.7 - Freqüências alélicas de quatro locos (A, B, C e D) obtidas das populações I e II

Loco/alelo	População I			População II		
	1 [#]	2 [#]	3 [#]	1 [#]	2 [#]	3 [#]
A	1,00			0,30	0,10	0,60
B	0,35	0,40	0,25	0,55	0,15	0,30
C	1,00			1,00		
D	0,45	0,55		0,65	0,35	-

[#]Sendo 1, 2 e 3 os alelos (A₁, B₁, C₁ e D₁), (A₂, B₂, C₂ e D₂) e (A₃, B₃, C₃ e D₃), respectivamente.

Veja que, para cada loco, o número máximo de genótipos a partir de **a** alelos será:

- Número de genótipos homozigotos é igual a: **a**
- Número de genótipos heterozigotos é igual a: $\frac{a(a - 1)}{2}$
- Número total de genótipos é igual a: $\frac{a(a + 1)}{2}$

Estudo da diversidade entre indivíduos dentro da população

Para o estudo da diversidade dentro da população, é possível adotar técnicas de agrupamento ou de projeção de medidas de dissimilaridade. Em ambos os casos, torna-se necessária a matriz de dissimilaridade entre pares de indivíduos por meio de um índice apropriado.

a. Coeficientes de similaridade e de dissimilaridade

As informações de marcadores moleculares co-dominantes são descritas, para cada acesso avaliado, na forma de representação genotípica, dadas em função do número de alelos por cada loco. Assim, se o loco apresenta três alelos, tem-se a representação 11, 22 e 33 para as formas homozigotas e 12, 13 e 23 para as heterozigotas. De maneira geral, dois genótipos i e i' apresentarão genótipos G_{ii} e $G_{i'i}$ ($j = 1, 2, \dots, l$) conforme ilustrado a seguir:

		Locos			
Genótipo	1	2	...	L	Total
i	G_{i1}	G_{i2}		G_{iL}	
i'	$G_{i'1}$	$G_{i'2}$		$G_{i'L}$	
Num.alelos	a_1	a_2		a_L	A

A similaridade entre pares de acessos pode ser calculada pelos seguintes índices:

a) Índice não-ponderado

É dado por:

$$S_{ii'} = \frac{1}{2L} \sum_{j=1}^L c_j$$

em que:

L: número total de locos estudiados; e

c_{ij} : número de alelos comuns entre os pares de acessos i e j .

O número de alelos comuns entre dois acessos de genótipos será:

- Dois para: $A_i A_i$ e $A_j A_i$ (ou $A_i A_j$ e $A_i A_j$)
 - Um para: $A_i A_i$ e $A_i A_j$ (ou $A_i A_j$ e $A_i A_k$)
 - Zero par: $A_i A_i$ e $A_k A_l$

O índice especificado varia de 0 a 1.

b) *Índice ponderado*

É fornecido por:

$$S_{ii'} = \frac{1}{2} \sum_{j=1}^L p_j c_j$$

em que:

$$p_j = \frac{a_j}{A} : \text{peso associado ao loco } j, \text{ determinado por:}$$

a_j : número total de alelos do loco j

A : número total de alelos estudados

$$\text{sendo } \sum_{j=1}^L p_j = 1$$

c_j : número de alelos comuns entre os pares de acessos i e i' .

Para os casos em que $a_1 = a_2 = \dots = a_L = a$, tem-se $A = aL$, e o índice ponderado torna-se igual ao não-ponderado.

Como se trata de medidas de similaridade, é recomendável, em análises de agrupamento, utilizar-se medidas de dissimilaridade, definidas por:

i. $D = 1 - S$.

ii. $D = 1 / (S + k)$. A inclusão da constante k , geralmente igual a 1, visa contornar os problemas de indeterminações ocasionados quando o valor de S é nulo.

iii. $D = \sqrt{1 - S}$ ou $\sqrt{2(1 - S)}$. Esta expressão atribui, a algumas medidas de dissimilaridade, propriedades euclidianas.

c) *Índice d² de Smouse e Peakall*

Dado por:

$$d_{ii'}^2 = \frac{1}{2L} \sum_{j=1}^L \sum_{k=1}^{a_j} (y_{ijk} - y_{i'jk})^2$$

Em que y_{ijk} refere-se quantidade de alelos k no loco j que o indivíduo i apresenta.

Se for considerado apenas um loco, seriam possíveis obter as seguintes distâncias:

Genótipos	Alelos				d^2
	i	j	k	l	
A _i A _i e A _i A _i	0				0
A _i A _i e A _j A _j	2	-2			4
A _i A _i e A _i A _j	1	-1			1
A _i A _i e A _j A _k	2	-1	-1		3
A _i A _j e A _i A _i	-1	1			1
A _i A _j e A _j A _j	1	-1			1
A _i A _j e A _i A _j	0	0			0
A _i A _j e A _i A _k	0	1	-1		1
A _i A _j e A _j A _k	1	0	-1		1
A _i A _j e A _k A _i	1	1	-1	-1	2

Por este índice, dois genótipos homozigotos e dois genótipos heterozigotos que não compartilham alelos comuns terão dissimilaridade igual a 4 e 2, respectivamente. como ilustrado a seguir:

HomxHom	A ₁	A ₂	A ₃	A ₄	HetxHet	A ₁	A ₂	A ₃	A ₄
A ₁ A ₁	2	0	0	0	A ₁ A ₂	1	1	0	0
A ₂ A ₂	0	2	0	0	A ₃ A ₄	0	0	1	1
(y _i -y _{i'}) ²	4	4	0	0	(y _i -y _{i'}) ²	1	1	1	1
$d^2=(4+4+0+0)/2=4$					$d^2=(1+1+1+1)/2=2$				

Esta dissimilaridade será diferente (e igual a 3), se dois genótipos não compartilharem alelos comuns e um deles for homozigoto e o outro heterozigoto, como representado a seguir:

HomxHet	A ₁	A ₂	A ₃	A ₄
A ₁ A ₁	2	0	0	0
A ₂ A ₃	0	1	1	0
(y _i -ȳ) ²	4	1	1	0
$d^2 = (4+1+1+0)/2=3$				

Para o exemplo referente à população 2 (Tabela A3), pode-se ilustrar o cálculo da similaridade entre os acessos 1 e 2 considerando as informações:

Loco	Genótipo Aces. 1	Genótipo Aces. 2	Num. Alelos comuns (c _j)	Num. total de alelos (a) [*]	c _j a _j
1	11	11	2	5	10
2	11	11	2	3	6
3	24	34	1	4	4
4	13	11	1	3	3
5	35	23	1	5	5
6	11	12	1	4	4
7	22	22	2	2	4
8	23	33	1	3	3
9	11	11	2	2	4
10	22	22	2	2	4
Total			15	33	47

(*) Informação obtida considerando todos os indivíduos das populações estudadas

Neste exemplo, tem-se:

L = 10: número total de locos

$$A = \sum_{j=1}^L a_j = 5 + 3 + \dots + 2 = 33: \text{número total de alelos}$$

O índice não-ponderado que expressa a similaridade entre os acessos 1 e 2 é dado por:

$$S_{12} = \frac{1}{2L} \sum_{j=1}^L c_j = \frac{1}{20} (2+2+\dots+2) = \frac{15}{20} = 0,75$$

Por sua vez, o índice ponderado é dado por:

$$S_{12} = \frac{1}{2} \sum_{j=1}^L p_j c_j = \frac{1}{2} \left(\frac{5}{33} 2 + \frac{3}{33} 2 + \frac{4}{33} 1 + \dots + \frac{2}{33} 2 \right) = \frac{47}{2(33)} = 0,7121$$

O valor de d^2 seria dado por:

$$d_{12}^2 = \frac{1}{2L} \left\{ \left[(y_{11,1} - y_{11,1})^2 \right] + \left[(y_{11,1} - y_{11,1})^2 \right] + \left[(y_{24,2} - y_{34,2})^2 + (y_{24,3} - y_{34,3})^2 + (y_{24,4} - y_{34,4})^2 \right] + \dots + \left[(y_{22,2} - y_{22,2})^2 \right] \right\} = 0,5$$

Logo,

$$d_{12} = 0,7071$$

Para análise dos dados da população 2 (Tabela A3), podem-se ilustrar os valores da dissimilaridade entre os 10 primeiros acessos, expresso pelo complemento aritmético do índice de similaridade ponderado, por meio dos elementos da matriz D, fornecida a seguir:

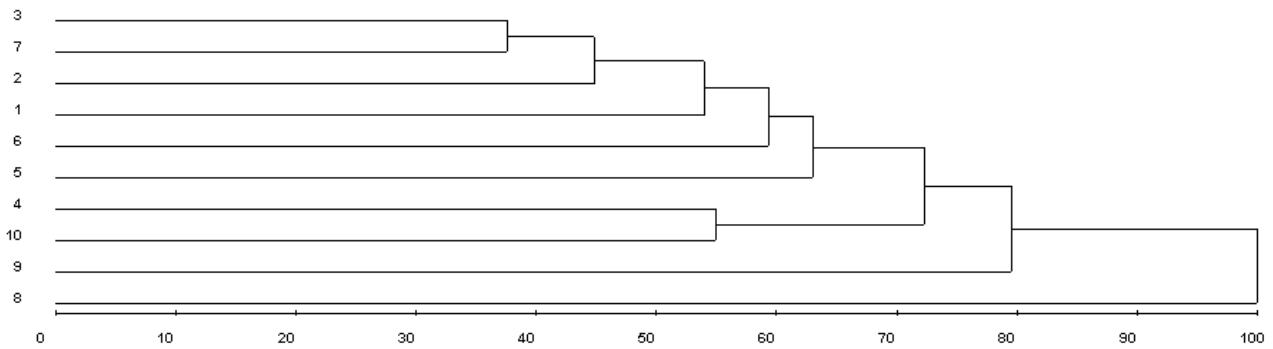
	1	2	3	4	5	6	7	8	9	10
1	0	0,288	0,348	0,424	0,379	0,364	0,212	0,500	0,379	0,348
2	0,288	0	0,273	0,424	0,394	0,348	0,197	0,439	0,409	0,348
3	0,348	0,273	0	0,242	0,242	0,227	0,197	0,636	0,439	0,379
4	0,424	0,424	0,242	0	0,333	0,288	0,439	0,455	0,545	0,288
5	0,379	0,394	0,242	0,333	0	0,318	0,318	0,667	0,379	0,470
6	0,364	0,348	0,227	0,288	0,318	0	0,303	0,409	0,333	0,424
7	0,212	0,197	0,197	0,439	0,318	0,303	0	0,576	0,379	0,424
8	0,500	0,439	0,636	0,455	0,667	0,409	0,576	0	0,485	0,545
9	0,379	0,409	0,439	0,545	0,379	0,333	0,379	0,485	0	0,470
10	0,348	0,348	0,379	0,288	0,470	0,424	0,424	0,545	0,470	0

b. Análise de agrupamento

Uma vez obtida a matriz de dissimilaridade entre os acessos, ela poderá ser submetida à análise de agrupamento pelas técnicas já descritas neste livro (capítulo 2). A base teórica desses procedimentos já foi apresentada e, agora, podem ser considerados os seguintes resultados, como ilustração do emprego de tais técnicas:

- i. Análise de agrupamento por métodos hierárquicos

Como ilustração será considerado o agrupamento dos 10 primeiros acessos estudado pela técnica de agrupamento UPGMA, a partir dos valores da matriz D, apresentada anteriormente, cujo resultado é descrito a seguir:



Algumas informações indispensáveis para melhor interpretação dos resultados obtidos pela análise de agrupamento hierárquica são:

- i. Valor da dissimilaridade obtida no último nível de fusão. Neste exemplo, a dissimilaridade no último nível de fusão foi igual a 0,5236 inferior à dissimilaridade máxima verificada entre os acessos 5 e 8 (igual a 0,667).
- ii. Identificação dos pontos de cortes obtidos por meio de algum critério ótimo de partição. Utilizando o método de Mojema (1977), com valor de $k = 1,25$, são recomendados corte a 73% e 80% de dissimilaridade. Estabelecendo corte a 73% de dissimilaridade identificam-se três grupos de similaridade: um formado pelo acesso 8, outro pelo acesso 9 e o terceiro pelos demais acessos.
- iii. A consistência dos nós e bifurcações, avaliada pelo uso da técnica de *bootstrap*.
- iv. Valor do coeficiente de correlação cofenético. Para este exemplo, a correlação entre as medidas de dissimilaridades originais e aquelas geradas graficamente a partir do dendrograma apresentado é de 0,8057.

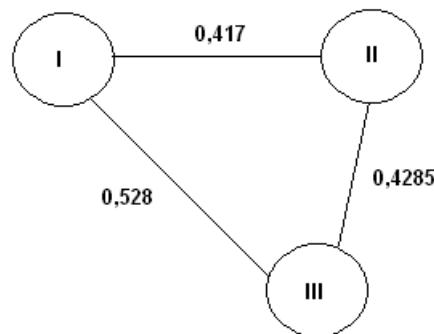
ii. Análise de agrupamento por métodos de otimização

Outra alternativa para estudar a existência de grupos homogêneos entre acessos é usar o método de otimização por Tocher, que ao ser aplicado à matriz de dissimilaridade D, entre os 10 acessos, proporciona o seguinte resultado:

Grupos	Acessos	Soma das distâncias	Média das distâncias
I	2 7 3 1 6 5 4	9,239	0,330
	10		
II	9	-	-
III	8	-	-

Alguns pontos importantes a serem destacados para melhor interpretação do resultado do agrupamento de Tocher são:

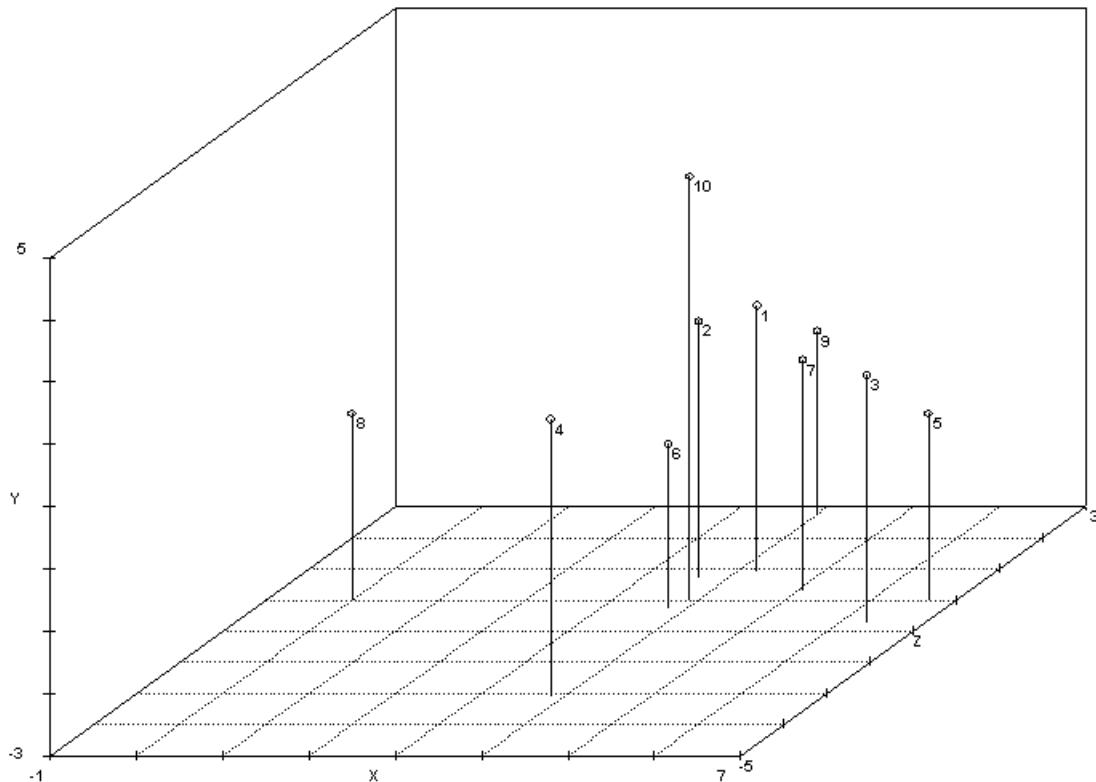
- i. O valor da estatística utilizada como critério global para inclusão de genótipos dentro de um grupo. Neste exemplo, o valor obtido foi de 0,409.
- ii. Os valores das distâncias intergrupos, que devem ser inferiores às distâncias intragrupo. Neste exemplo, os valores das distâncias são representados graficamente por meio de:



iii. Análise por dispersão gráfica

Outra alternativa para avaliação do padrão de agrupamento, com base em informações contidas numa matriz de dissimilaridade, é apresentar a diversidade

entre acessos por meio da dispersão gráfica, obtida pela projeção de distâncias em gráficos 2D ou 3D. Para o exemplo considerado, optou-se por representar graficamente os valores das distâncias apresentadas na matriz D em gráfico 3D, conforme ilustrado a seguir:



Para avaliar a qualidade da dispersão pela projeção gráfica 2D ou 3D apresentada, têm-se as seguintes medidas:

Projeção	2D	3D
Distorção	29,54%	10,56%
Correlação entre as distâncias originais e as gráficas	0,710	0,838
Estresse	38,97	21,64

Constata-se que a análise em gráfico 3D apresenta qualidade superior à 2D, com correlação entre distâncias acima de 0,80 e estresse próximo a 20%. O valor de distorção também reduziu consideravelmente, passando de 29,54% para 10,56%.

Freqüência alélica na população

A freqüência alélica poderá ser estimada a partir do número de ocorrência das diferentes classes genotípicas. Assim, denominando a ocorrência de homozigotos na população de N_{ii} e de heterozigotos de N_{ij} , tem-se:

$$f(A_i) = \frac{2N_{ii} + \sum_{j=1, j \neq i}^a N_{ij}}{2N}$$

sendo a o número de alelos apresentados por o loco estudado e N é o número total de indivíduos da população

Para a análise da população 2 (Tabela A3), em que se tem uma população com 50 indivíduos, o loco 1 apresenta cinco alelos. Dessa forma, há expectativa do aparecimento de $a(a+1)/2$ (igual a 15) diferentes genótipos na população. Estes genótipos e sua ocorrência estão relacionados a seguir:

Genótipo	Observados	Esperados no equilíbrio
11	$N_{11} = 19$	Np_1^2
12	$N_{12} = 10$	$N2p_1p_2$
13	$N_{13} = 11$	$N2p_1p_3$
14	$N_{14} = 2$	$N2p_1p_4$
15	$N_{15} = 0$	$N2p_1p_5$
22	$N_{22} = 0$	Np_2^2
23	$N_{23} = 3$	$N2p_2p_3$
24	$N_{24} = 0$	$N2p_2p_4$
25	$N_{25} = 0$	$N2p_2p_5$
33	$N_{33} = 2$	Np_3^2
34	$N_{34} = 2$	$N2p_3p_4$
35	$N_{35} = 0$	$N2p_3p_5$
44	$N_{44} = 0$	Np_4^2
45	$N_{45} = 1$	$N2p_4p_5$
55	$N_{55} = 0$	Np_5^2
Total	$N = 50$	N

Assim, as freqüências alélicas são estimadas por meio das expressões:

$$f(A_1) = \frac{2N_{11} + N_{12} + N_{13} + N_{14} + N_{15}}{2N} = \frac{38 + 10 + 11 + 2 + 0}{100} = 0,61$$

$$f(A_2) = \frac{2N_{22} + N_{12} + N_{23} + N_{24} + N_{25}}{2N} = \frac{0 + 10 + 3 + 0 + 0}{100} = 0,13$$

$$f(A_3) = \frac{2N_{33} + N_{13} + N_{23} + N_{34} + N_{35}}{2N} = \frac{4 + 11 + 3 + 2 + 0}{100} = 0,20$$

$$f(A_4) = \frac{2N_{44} + N_{14} + N_{24} + N_{34} + N_{35}}{2N} = \frac{0 + 2 + 0 + 2 + 1}{100} = 0,05$$

e

$$f(A_5) = \frac{2N_{55} + N_{15} + N_{25} + N_{35} + N_{45}}{2N} = \frac{0 + 0 + 0 + 0 + 1}{100} = 0,01$$

De maneira análoga são obtidas as freqüências alélicas para os demais locos, tendo-se para a população 2 (Tabela A3), os seguintes resultados:

Loco	Número de alelos	$f(A_1)$	$f(A_2)$	$f(A_3)$	$f(A_4)$	$f(A_5)$
1	5	0,61	0,13	0,20	0,05	0,01
2	3	0,73	0,26	0,01		
3	4	0,33	0,05	0,29	0,33	
4	3	0,52	0,03	0,45		
5	5	0,26	0,26	0,40	0,06	0,02
6	4	0,73	0,19	0,01	0,07	
7	2	0,03	0,97			
8	3	0,07	0,34	0,59		
9	2	0,64	0,36			
10	2	0,10	0,90			

É interessante dispor de locos polimórficos que apresentam grande quantidade de alelos para que se possa inferir a diversidade de uma população em relação a outras ou com ela própria, ao longo do tempo, principalmente quando submetida a forças evolutivas que promovam diferenciações.

Equilíbrio de Hardy-Weinberg

Populações grandes que se reproduzem por acasalamentos ao acaso, ou aleatório, e que não estão sujeitas a forças evolutivas, como seleção, mutação e migração, encontram-se em equilíbrio. Quando os pressupostos de Hardy-Weinberg não são satisfeitos ocorrem desvios das expectativas, mas, dependendo de qual pressuposto não é satisfeito, esses desvios podem ou não ser estatisticamente detectáveis. Desvios podem ser causados por efeito Wahlund, fluxo gênico, mutações, acasalamento não-aleatório, seleção ou deriva genética. Acasalamento não-aleatório mudará apenas as freqüências genotípicas relativas aos genes especificamente envolvidos. Deriva genética é particularmente efetiva em populações pequenas. Desvios causados por seleção, fluxo gênico e mutações de qualquer natureza muitas vezes precisam de valores significativamente altos para poderem ser detectados, o que torna o teste de desvios das proporções Hardy-Weinberg fraco para avaliação das mudanças numa população. Entretanto, quando se consideram as informações de vários locos, é possível obter informações de maior credibilidade sobre as reais condições da população sob análise.

A seguir são apresentadas técnicas biométricas para avaliação das condições de equilíbrio, tendo como base os dados da população 2 (Anexo A3).

- Teste de qui-quadrado

Com base nas freqüências alélicas estimadas, é possível verificar se a população 2 (Tabela 5.2) encontra-se em equilíbrio de Hardy-Weinberg, comparando-se valores observados com os esperados. Assim, tem-se para o primeiro loco estudado:

Genótipo	Observado	Esperado no equilíbrio	$d = (O - E)^2/E$
11	$N_{11} = 19$	$Np_1^2 = 18,605$	0,008
12	$N_{12} = 10$	$N2p_1p_2 = 7,930$	0,540
13	$N_{13} = 11$	$N2p_1p_3 = 12,200$	0,118
14	$N_{14} = 2$	$N2p_1p_4 = 3,050$	0,361
15	$N_{15} = 0$	$N2p_1p_5 = 0,610$	0,610
22	$N_{22} = 0$	$Np_2^2 = 0,845$	0,845
23	$N_{23} = 3$	$N2p_2p_3 = 2,600$	0,062
24	$N_{24} = 0$	$N2p_2p_4 = 0,650$	0,650
25	$N_{25} = 0$	$N2p_2p_5 = 0,130$	0,130
33	$N_{33} = 2$	$Np_3^2 = 2,000$	0,000
34	$N_{34} = 2$	$N2p_3p_4 = 1,000$	1,000
35	$N_{35} = 0$	$N2p_3p_5 = 0,200$	0,200
44	$N_{44} = 0$	$Np_4^2 = 0,125$	0,125
45	$N_{45} = 1$	$N2p_4p_5 = 0,050$	18,050
55	$N_{55} = 0$	$Np_5^2 = 0,005$	0,005
Total	$N = 50$	50	22,705

Os graus de liberdade associados são igual a $c - a$, em que:

c é o número de classes genotípicas esperadas, dado por $a(a+1)/2$; e

a é o número de alelos que o loco apresenta.

Dessa forma, têm-se os resultados dos testes de qui-quadrado realizados para todos os locos estudados:

Loco	GL	χ^2	Probabilidade	
1	10	22,705	0,012	*
2	3	3,911	0,271	ns
3	6	7,148	0,307	ns
4	3	5,427	0,143	ns
5	10	6,924	0,733	ns
6	6	3,290	0,772	ns
7	1	0,048	0,827	ns
8	3	0,507	0,917	ns
9	1	0,871	0,351	ns
10	1	0,617	0,432	ns

Assim, para os locos considerados, apenas o primeiro apresenta desequilíbrio em relação ao esperado pelos princípios postulados por Hardy-Weinberg. A significância, neste exemplo, deve ser vista com certa reserva, uma vez que o tamanho da amostra, com apenas 50 indivíduos, pode ter sido insuficiente para quantificar a verdadeira ocorrência de certas classes genotípicas, em particular daquelas portadoras do alelo A_5 , cuja freqüência estimada foi relativamente baixa.

- Teste da razão de verossimilhança (teste G ou G^2)

Como visto no capítulo 3, há possibilidade de avaliar o equilíbrio por meio do teste da razão entre duas funções de máxima verossimilhança: uma definida pelas freqüências genotípicas esperadas no equilíbrio (L_0) e a outra pelas freqüências genotípicas observadas (L_1), ou seja:

$$G^2 = -2 \ln \left(\frac{L_0}{L_1} \right) = -2 (\ln L_0 - \ln L_1)$$

sendo:

$$\ln(L_0) = \ln(\lambda) + \sum_{k=1}^a N_{kk} \ln \left(\frac{n_k}{2N} \right)^2 + \sum_{\substack{k=1 \\ k'>k}}^a N_{kk'} \ln \left(2 \frac{n_k n_{k'}}{2N 2N} \right)$$

$$\ln(L_1) = \ln(\lambda) + \sum_{k=1}^a N_{kk} \ln \left(\frac{N_{kk}}{N} \right) + \sum_{\substack{k=1 \\ k'>k}}^a N_{kk'} \ln \left(\frac{N_{kk'}}{N} \right)$$

em que:

$$\lambda = \frac{N!}{\prod_{k=1}^s N_{kk}!}$$

a: número de alelos apresentado pelo loco estudado;

N: tamanho amostral;

n_k : número de alelos k da população amostrada; e

N_{kk} e $N_{kk'}$: respectivos números de homozigotos e de heterozigotos amostrados.

Sob equilíbrio, a quantidade $-2(\ln L_0 - \ln L_1)$ tem distribuição aproximada de qui-quadrado, com $a(a-1)/2$ graus de liberdade (WEIR, 1996). Assim, têm-se os resultados dos testes feitos para todos os locos estudados:

Loco	GL	G^2	Probabilidade	
1	10	11,0937	0,3503	ns
2	3	3,9071	0,2717	ns
3	6	8,8318	0,1833	ns
4	3	6,5673	0,0870	ns
5	10	5,5974	0,8479	ns
6	6	3,2632	0,7752	ns
7	1	0,0928	0,7607	ns
8	3	0,7437	0,8629	ns
9	1	0,8617	0,3533	ns
10	1	0,5079	0,4760	ns

Constata-se que, por este teste, todos os locos apresentam freqüência genotípica de acordo com o esperado, segundo princípio de equilíbrio proposto por Hardy-Weinberg.

- Teste exato de Fisher

Este teste é baseado na probabilidade assumida por um possível arranjo genotípico (conjunto de classes genotípicas) condicionado às freqüências alélicas amostradas na população. Para dois alelos, as probabilidades condicionais ainda podem ser calculadas em termos de um alelo (A), e o número de heterozigotos (N_{Aa}), conforme propôs Haldane (1954), da seguinte forma:

$$P(N_{Aa} | n_A) = \frac{N! n_A!(2N - n_A)! 2^{N_{Aa}}}{[(n_A - N_{Aa})/2]! N_{Aa}! [N - (n_A + N_{Aa})/2]! (2N)!}$$

Dessa forma, todas as probabilidades são estimadas para todos os possíveis valores de N_{Aa} , da amostra de tamanho N e quantidade alélica n_A (ou n_a). Os valores de N_{Aa} são ordenados de acordo com suas probabilidades condicionais. Para vários

alelos, é possível realizar o teste de acordo com a proposta de Guo e Thompson (1992), ou adotar um processo simplificado, em que se consideram apenas dois alelos (A e \bar{A}), tomando-se A como o alelo mais freqüente da população e \bar{A} representativo dos demais alelos.

Loco	Alelo (*)	AA	A \bar{A}	$\bar{A}\bar{A}$	P($N_{\bar{A}\bar{A}} n_A$)	P.Acum	\hat{D}_A	z
1	A_1	19	23	8	0,2232	0,7738	0,0079	0,2348
2	A_1	29	15	6	0,0624	0,1436	0,0471	1,6897
3	A_1	4	25	21	0,1821	0,5241	-0,0289	-0,9243
4	A_1	10	32	8	0,0364	0,0871	-0,0704	-1,9944*
5	A_3	6	28	16	0,1293	0,3758	-0,0400	-1,1785
6	A_1	28	17	5	0,1598	0,3034	0,0271	0,9722
7	A_1	0	3	47	0,9697	1,0000	-0,0009	-0,2187
8	A_3	0	7	43	0,7987	1,0000	-0,0049	-0,5322
9	A_1	22	20	8	0,1463	0,3643	0,0304	0,9330
10	A_2	41	8	1	0,3334	0,3923	0,0100	0,7857

(*) Alelo mais freqüente em cada loco.

A análise revela que, para todos os marcadores, as freqüências genotípicas estão de acordo com o esperado, segundo equilíbrio de Hardy-Weinberg.

- Teste z

Pode-se considerar que, se não há equilíbrio de Hardy-Weinberg, alguma força seletiva atua contribuindo para o acréscimo de homozigotos ou de heterozigotos. Para dois locos, o valor de taxa de desequilíbrio (D_A) é dado por:

$$\hat{D}_A = \hat{P}_{AA} - \hat{p}^2$$

sendo \hat{P}_{AA} a freqüência observada de genótipos AA.

É possível realizar o teste de hipótese $H_0: D_A = 0$ vs $H_a: D_A \neq 0$ por meio da estatística z, dada por:

$$z = \frac{\hat{D}_A}{\sqrt{V(\hat{D}_A)}}$$

Definições de \hat{D}_A e $V(\hat{D}_A)$ são as mesmas descritas no item 3.4.5.

Se o valor absoluto de z excede 1,96, rejeita-se a hipótese de que a proporção de homozigotos observada está em conformidade com o esperado sob hipótese de equilíbrio de Hardy-Weinberg. Por este teste, constata-se que apenas o loco 4 apresenta freqüência de homozigoto aquém do esperado.

Endogamia, Heterozigose e PIC

Uma população poderá estar sujeita aos efeitos da endogamia quando ocorrer cruzamentos preferenciais entre parentados, que tem por consequência o efeito sobre a freqüência de homozigotos e heterozigotos. No exemplo em consideração, tomando o primeiro loco como referência, tem-se:

Genótipo	Observados	Freqüência Observada
Homozigotos	21	0,42
Heterozigotos	29	$H_e = 0,58$
Total	50	1,00

A freqüência esperada, pressupondo que a população esteja em equilíbrio de Hardy-Weinberg (EHW), de homozigotos para este loco é obtida a partir das freqüências alélicas, ou seja:

$$f(\text{Homozigoto}_{\text{sob EHW}}) = \sum_{i=1}^a p_i^2 = 0,61^2 + \dots + 0,01^2 = 0,4316$$

e

$$f(\text{Heterozigoto}_{\text{sob EHW}}) = H_0 = 1 - \sum_{i=1}^a p_i^2 = 1 - 0,4316 = 0,5684$$

Estima-se o coeficiente médio de endogamia da população por meio de:

$$F = 1 - \frac{H_e}{H_0} = 1 - \frac{0,58}{0,5684} = -0,0204$$

Lembrando que:

H_0 : freqüência de heterozigotos numa população supostamente em equilíbrio de Hardy-Weinberg (dada por $2pq$, quando se consideram apenas dois alelos por loco); e

H_ε : freqüência de heterozigotos numa população sujeita ao acasalamento entre aparentados (dada por $2pq - 2\varepsilon$, quando se consideram apenas dois alelos por loco e $\varepsilon = pqF$).

O valor negativo deve ser interpretado como estimativa de endogamia nula, de forma que na população não há acasalamentos entre aparentados, como revelado por este loco.

De maneira análoga são obtidos os valores para os demais locos, conforme ilustrado a seguir:

Loco	$f(A_1)$	$f(A_2)$	$f(A_3)$	$f(A_4)$	$f(A_5)$	H_ε	H_0	F
1	0,61	0,13	0,20	0,05	0,01	0,580	0,5684	-0,0204
2	0,73	0,26	0,01			0,300	0,3994	0,2489
3	0,33	0,05	0,29	0,33		0,660	0,6956	0,0512
4	0,52	0,03	0,45			0,640	0,5262	-0,2163
5	0,26	0,26	0,40	0,06	0,02	0,740	0,7008	-0,0559
6	0,73	0,19	0,01	0,07		0,400	0,4260	0,0610
7	0,03	0,97				0,060	0,0582	-0,0309
8	0,07	0,34	0,59			0,540	0,5314	-0,0162
9	0,64	0,36				0,400	0,4608	0,1319
10	0,10	0,90				0,160	0,1800	0,1111

O coeficiente médio de endogamia, computado os 10 locos, é de 0,0264.

Segundo Liu (1997), um estimador não-viesado da heterozigosidade é dado por:

$$H_o = \frac{N}{N-1} \left(1 - \sum_{i=1}^a p_i^2 \right)$$

sendo N o tamanho da amostra. A variância aproximada deste estimador é dada por:

$$V_{H_0} = \frac{N}{(N-1)^2} \left[\sum_{j=1}^a p_j^3 - \left(\sum_{j=1}^a p_j^2 \right)^2 \right]$$

Outra informação de importância para caracterizar a diversidade da população é o conteúdo médio de informação polimórfica (PIC) (BOTSTEIN et al., 1980), expresso por:

$$PIC = 1 - \sum_{i=1}^a p_i^2 - \sum_{i,j=1}^a p_i^2 p_j^2$$

em que p_i é a freqüência do i -ésimo alelo do loco estudado.

O valor de PIC fornece uma estimativa do poder discriminatório do marcador, por considerar não somente o número de alelos por loco, mas também a freqüência relativa desses alelos. De maneira geral, os valores de PIC são inferiores aos obtidos para a heterozigosidade.

Segundo Ott (1992), um loco será considerado polimórfico quando $H_0 \geq 0,1$ (que corresponde, aproximadamente, à situação em que o alelo mais freqüente tem freqüência menor que 0,95) e altamente polimórfico quando $H_0 \geq 0,7$.

Quando se considera um loco com a alelos, em que a freqüência é igual para todos eles, tem-se:

$$H_0 = 1 - \sum_{j=1}^a p_j^2 = 1 - \frac{1}{a}, \text{ se } a = 2, \text{ então } H_0 = 0,5$$

e

$$PIC = 1 - \sum_{i=1}^a p_i^2 - \sum_{i,j=1}^a p_i^2 p_j^2 = 1 - \frac{1}{a} - \frac{1}{a^2} + \frac{1}{a^3}, \text{ se } a = 2 \text{ então } PIC = 0,375$$

Para um gene, com dois alelos A e a com frequência p e q , respectivamente, tem-se:

$$H_0 = 1 - \sum_{j=1}^a p_j^2 = 1 - p^2 - q^2 = 2pq$$

$$PIC = 1 - \sum_{i=1}^a p_i^2 - \sum_{i,j=1}^a p_i^2 p_j^2 = 1 - (p^2 + q^2) - (p^2 q^2 + q^2 p^2) = 2pq - 2p^2 q^2$$

A seguir são apresentadas estimativas de H e de PIC para as situações de vários alelos em um loco, com freqüências iguais. Assim, tem-se:

Número de alelos	H	PIC
2	0,5000000	0,3750000
3	0,6666666	0,5925925
4	0,7500000	0,7031250
5	0,8000000	0,7680000
6	0,8333333	0,8101851
7	0,8571429	0,8396502
8	0,8750000	0,8613281
9	0,8888889	0,8779150
10	0,9000000	0,8910000
11	0,9090909	0,9015777
12	0,9166667	0,9103010
13	0,9230769	0,9176149
14	0,9285714	0,9238338
15	0,9333333	0,9291852
16	0,9375000	0,9338379
17	0,9411765	0,9379198
18	0,9444444	0,9415295
19	0,9473684	0,9447442
20	0,9500000	0,9476250

Para o primeiro marcador, o valor do PIC poderá ser obtido a partir das seguintes informações:

$$\sum_{i=1}^a p_i^2 = 0,61^2 + 0,13^2 + 0,20^2 + 0,05^2 + 0,01^2 = 0,4316$$

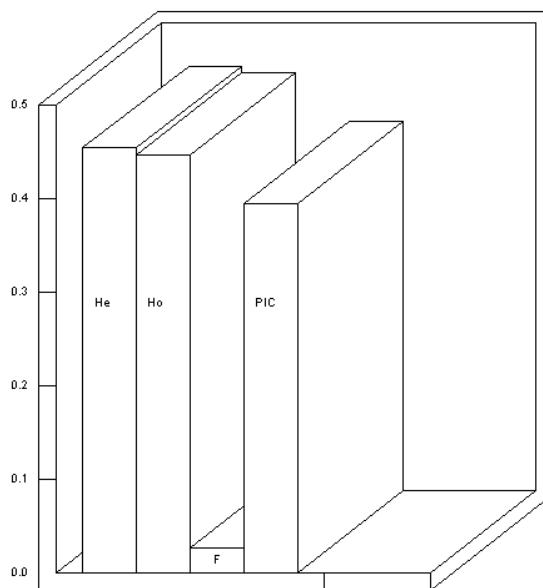
$$\begin{aligned} & \sum_{i,j=1}^a p_i^2 p_j^2 \\ &= \{[0,61^2(0,13^2+0,20^2+\dots+0,01^2)]+[0,13^2(0,61^2+0,20^2+\dots+0,01^2)]+\dots+ \\ & \quad [0,01^2(0,61^2+0,13^2+0,20^2+\dots+0,05^2)]\} = 0,0459 \end{aligned}$$

$$PIC = 1 - \sum_{i=1}^a p_i^2 - \sum_{i,j=1}^a p_i^2 p_j^2 = 1 - (0,4316 + 0,0459) = 0,5225$$

Para os demais locos, tem-se:

Loco	$f(A_1)$	$f(A_2)$	$f(A_3)$	$f(A_4)$	$f(A_5)$	PIC
1	0,61	0,13	0,20	0,05	0,01	0,5225
2	0,73	0,26	0,01			0,3272
3	0,33	0,05	0,29	0,33		0,6337
4	0,52	0,03	0,45			0,4158
5	0,26	0,26	0,40	0,06	0,02	0,6460
6	0,73	0,19	0,01	0,07		0,3818
7	0,03	0,97				0,0565
8	0,07	0,34	0,59			0,4464
9	0,64	0,36				0,3546
10	0,10	0,90				0,1638

O número médio de alelos é 3,3, e o valor médio do PIC é igual a 0,3948. Graficamente, têm-se as magnitudes relativas das medidas de diversidade dentro da população:



Medidas Adicionais de Diversidade

Além das medidas baseadas nas estatísticas mencionadas anteriormente, trabalhos envolvendo estudos de estrutura populacional e diversidade genética, geralmente, apresentam sempre algumas das variáveis relacionadas a seguir.

Proporção de locos polimórficos (P)

É fornecida por:

$$P = (\text{número de locos polimórficos}) / (\text{número total de locos analisados})$$

Três critérios são comumente utilizados para classificar um loco como polimórfico (COLE, 2003):

- loco exibindo polimorfismo em pelo menos um indivíduo da amostra;
- loco em que o alelo mais comum tem freqüência menor que 99%; e
- loco em que o alelo mais comum tem freqüência menor que 95%.

Para o exemplo em consideração (população 2 – Anexo A3), tem-se:

Loco	$f(A_1)$	$f(A_2)$	$f(A_3)$	$f(A_4)$	$f(A_5)$	Máxima
1	0,61	0,13	0,20	0,05	0,01	0,61
2	0,73	0,26	0,01			0,73
3	0,33	0,05	0,29	0,33		0,33
4	0,52	0,03	0,45			0,52
5	0,26	0,26	0,40	0,06	0,02	0,40
6	0,73	0,19	0,01	0,07		0,73
7	0,03	0,97				0,97
8	0,07	0,34	0,59			0,59
9	0,64	0,36				0,36
10	0,10	0,90				0,90

Assim, pode-se inferir que:

- Pelo critério 1: $P = 100\%$
- Pelo critério 2: $P = 100\%$
- Pelo critério 3: $P = (9/10) \times 100 = 90\%$

Número total (A) e médio de alelos por loco (N_m)

É dado pela razão entre o número total de alelos e o número de locos analisados.

$$N_m = \frac{A}{L}$$

No exemplo, tem-se:

$$A = 33 \quad \text{e} \quad L = 10$$

$$N_m = \frac{33}{10} = 3,3$$

Número efetivo de alelos por loco polimórfico (N_e)

É dado pela razão entre o número total de alelos dos locos polimórficos e o número total de locos polimórficos. Assim:

- Pelo critério 1: $N_e = 33/10 = 3,3$
- Pelo critério 2: $N_e = 33/10 = 3,3$
- Pelo critério 3: $N_e = 31/9 = 3,44$

Proporção de alelos contidos em cada população (P_A)

$P_A = (\text{número de alelos da população}) / (\text{número total de alelos na espécie})$

Número de alelos raros (N_r)

N_r = contagem de alelos com freqüência menor que 0,05 em cada subpopulação ou amostra. No exemplo, há seis alelos raros.

Desequilíbrio de fase gamética

O teste de desequilíbrio é de grande importância, pois, em muitos casos, apesar de a população ser panmítica, pode persistir uma estratificação dentro da população, de forma que ainda permanece um conjunto gênico preferencial, derivado dos ancestrais, que subdivide a população, que só gradualmente irá homogeneizar-se. Essa homogeneização dependerá, principalmente, da taxa de recombinação entre os dois locos, podendo ser extremamente lenta, sobretudo entre os locos estreitamente ligados. Assim, o equilíbrio, quando constatado, será indicativo de que a população já passou por sucessivos ciclos de acasalamentos ao acaso e encontra-se livre de forças evolutivas, como deve ter ocorrido no exemplo considerado.

Se o desequilíbrio de ligação é detectado, apesar de ser conhecida a existência de acasalamentos ao acaso, postula-se a hipótese de que a introdução ("migração") de indivíduos de outras populações, a seleção, a mutação e a deriva devem ter atuado e constituem elementos perturbadores em relação ao modelo de

Hardy-Weinberg. Mesmo que a quebra do isolamento ou processos seletivos abranjam apenas uma geração (o equilíbrio nas novas frequências em cada loco é atingido logo na geração seguinte), gera novos desequilíbrios gaméticos, que podem, se entre locos próximos, persistir durante muitas gerações; por isso, o grande interesse na avaliação dos desequilíbrios de ligação reside no fato de eles assinalarem retrospectivamente eventos de introdução de alelos e alteração nas suas freqüências no passado da população.

Várias metodologias para estimar o desequilíbrio de ligação têm sido amplamente descritas em revisões (FLINT-GARCIA et al., 2003; GAUT; LONG, 2003; GUPTA et al., 2005). Em algumas revisões são descritos os métodos disponíveis, a estatística utilizada para se testar a significância das medidas obtidas e as estimativas obtidas envolvendo locos multialélicos e condições multilocos (GUPTA et al., 2005; JORDE, 2000; LIANG et al., 2001; GORELICK; LAUBICHLER, 2004).

a) Estimador de máxima verossimilhança

Em uma população em equilíbrio de Hardy-Weinberg, os gametas produzidos, em relação a dois genes A/a e B/b, são dados por:

Gameta	Freqüência
AB	$P_{11(1)}$
Ab	$P_{10(1)}$
aB	$P_{01(1)}$
ab	$P_{00(1)}$

O desequilíbrio de fase gamética é quantificado por meio de:

$$\Delta = P_{11(1)}P_{00(1)} - P_{10(1)}P_{01(1)}$$

A freqüência gamética em qualquer geração de acasalamento ao acaso pode ser estimada conhecendo-se as freqüências alélicas e o valor de desequilíbrio inicial, ou seja:

$$P_{11(n)} = p_A p_B + \Delta_n$$

$$P_{10(n)} = p_A q_B - \Delta_n$$

$$P_{01(n)} = q_A p_B - \Delta_n$$

$$P_{00(n)} = q_A q_B + \Delta_n$$

Para dois alelos para cada um de dois genes co-dominantes ($f(A) = p_A$; $f(a) = q_a$; $f(B) = p_B$; $f(b) = q_b$), ocorrem as seguintes expectativas de freqüências genotípicas:

Genótipo	Esperado	Observado
AABB	$(p_A p_B + \Delta)^2$	n_1
AABb	$2(p_A p_B + \Delta)(p_A q_B - \Delta)$	n_2
AAbb	$(p_A q_B - \Delta)^2$	n_3
AaBB	$2(p_A p_B + \Delta)(q_A p_B - \Delta)$	n_6
AaBb	$2[(p_A q_B - \Delta)(q_A p_B - \Delta) + (p_A p_B + \Delta)(q_A q_B + \Delta)]$	n_5
Aabb	$2(p_A q_B - \Delta)(q_A q_B + \Delta)$	n_6
aaBB	$(q_A p_B - \Delta)^2$	n_7
aaBb	$2(q_A p_B - \Delta)(q_A q_B + \Delta)$	n_8
aabb	$(q_A q_B + \Delta)^2$	n_9

Com relatado no capítulo 3, o valor de Δ é estimado por método de máxima verossimilhança, admitindo que o número de ocorrência das classes genotípicas segue distribuição multinomial, dada por:

$$L(p, q, r, s, \Delta; n_i) = \frac{N!}{n_1! n_2! \dots n_9!} p_1^{n_1} p_2^{n_2} \dots p_9^{n_9}$$

Para o exemplo considerado, em que se analisam os dados da população 2 (Anexo – Tabela A3), foi admitido que, para todos os locos estudados, havia apenas dois alelos, sendo um o mais freqüente; o outro representa o conjunto dos demais alelos. O novo conjunto de dados foi, então, representado por:

Loco 1	Loco 2	Loco 3	Loco 4	Loco 5	Loco 6	Loco 7	Loco 8	Loco 9	Loco 10
11	11	22	12	12	11	11	12	11	11
11	11	22	11	12	12	11	11	11	11
12	11	12	11	12	11	11	12	12	11
12	12	12	12	12	11	11	11	22	11
12	11	22	22	12	11	11	12	22	22
22	11	12	12	12	11	11	11	12	11
11	11	12	11	12	11	11	12	11	12
22	12	22	12	12	22	11	11	11	11
12	11	22	22	22	11	11	12	11	11
11	11	22	12	22	11	11	11	12	11
22	11	11	11	12	11	11	22	11	11
11	11	12	12	22	12	11	12	12	12
11	11	22	12	11	12	11	11	11	11
12	12	22	12	11	11	11	12	11	11
12	11	22	12	12	22	11	12	11	12
11	11	22	12	11	11	11	12	12	11
11	11	22	12	11	11	11	12	12	11
11	12	12	11	22	12	11	11	12	11
12	22	12	12	12	12	11	22	22	11
12	12	12	12	12	22	11	12	11	11
12	11	22	12	22	11	11	12	22	12
22	12	12	22	12	12	11	12	12	11
22	22	12	12	12	11	12	11	11	11
11	12	11	11	12	12	11	11	22	11
11	11	12	12	12	11	11	12	11	11
11	11	12	12	12	12	11	12	11	11
11	12	22	12	22	22	12	12	12	11
22	11	22	12	22	12	11	11	12	12
12	12	22	22	12	11	11	12	11	11
12	11	22	12	22	12	11	12	11	11
12	11	11	22	11	12	11	12	12	11
12	11	12	12	11	11	11	11	12	11
12	11	12	22	12	11	11	12	11	11
11	11	12	12	22	22	11	12	11	11
11	11	12	12	12	12	11	22	11	12
12	11	12	11	12	11	11	12	11	11
11	12	22	12	22	12	11	22	11	11
22	12	12	22	22	11	11	12	22	11
11	11	12	12	12	11	11	22	12	11
11	22	22	11	12	11	11	12	12	11
22	11	12	12	12	12	11	11	12	12
11	22	22	12	11	11	11	11	12	11
12	22	22	12	11	11	11	11	12	11
12	22	12	12	22	11	12	22	12	11
12	22	12	12	22	11	12	22	12	11
12	11	22	12	22	11	11	12	12	11
12	12	12	12	12	11	11	11	11	11
11	11	11	22	22	11	11	11	11	11
12	12	12	12	12	11	11	22	12	11
12	11	12	12	22	11	11	11	11	11
12	11	12	12	12	11	11	11	11	11
12	11	12	12	22	11	11	11	11	11
12	11	12	12	22	11	11	11	11	12

Para este novo conjunto de dados, tem-se:

Loco	Num. indivíduos	GL	χ^2	Probabilidade	f(A)	f(a)
1	50	1	0,055	0,814 ns	0,61	0,39
2	50	1	20,855	0,091 ns	0,73	0,27
3	50	1	0,854	0,355 ns	0,33	0,67
4	50	1	30,978	0,046 *	0,52	0,48
5	50	1	10,389	0,239 ns	0,40	0,60
6	50	1	0,945	0,331 ns	0,73	0,27
7	50	1	0,048	0,827 ns	0,97	0,03
8	50	1	0,056	0,813 ns	0,59	0,41
9	50	1	0,871	0,351 ns	0,64	0,36
10	50	1	0,617	0,432 ns	0,90	0,10

Assim, no estudo de avaliação de desequilíbrio gamético, em se tratando do locos 1 e 2, devem ser considerados os valores:

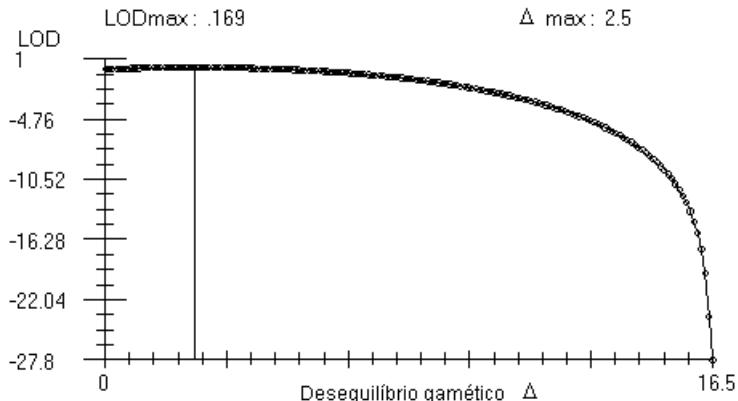
Genótipo	Esperado	Obs ¹	Obs ²
AABB	$(p_A p_B + \Delta)^2$	$n_1 = 13$	$n_1 = 2$
AABb	$2(p_A p_B + \Delta)(p_A q_B - \Delta)$	$n_2 = 4$	$n_2 = 4$
AAAb	$(p_A q_B - \Delta)^2$	$n_3 = 2$	$n_3 = 13$
AaBB	$2(p_A p_B + \Delta)(q_A p_B - \Delta)$	$n_6 = 12$	$n_6 = 3$
AaBb	$2[(p_A q_B - \Delta)(q_A p_B - \Delta) + (p_A p_B + \Delta)(q_A q_B + \Delta)]$	$n_5 = 8$	$n_5 = 8$
Ab/ab	$2(p_A q_B - \Delta)(q_A q_B + \Delta)$	$n_6 = 3$	$n_6 = 12$
aB/aB	$(q_A p_B - \Delta)^2$	$n_7 = 4$	$n_7 = 1$
aaBb	$2(q_A p_B - \Delta)(q_A q_B + \Delta)$	$n_8 = 3$	$n_8 = 3$
aabb	$(q_A q_B + \Delta)^2$	$n_9 = 1$	$n_9 = 4$

¹ e ²: Situações admitindo ser o alelo B aquele codificado por 1 ou 2, respectivamente.

Assim, são considerados os dois conjuntos de valores observados sendo obtidos os seguintes resultados:

Caso 1:

Para este par de locos, o valor máximo de desequilíbrio será de 16,5%, e o valor estimado de Δ é 2,5%, conforme ilustrado graficamente:



$$f(A) = 0,61 \quad f(a) = 0,39 \quad f(B) = 0,73 \quad f(b) = 0,27$$

Caso 2:

Nesta situação, o valor máximo de desequilíbrio será de 10,6%, e o valor estimado de Δ é 0,1%, com LOD associado igual a -0,014. Assim, admite-se que a função de verossimilhança que melhor explica os resultados é aquela em que se considera o primeiro conjunto de valores observados.

Se a causa do desequilíbrio gamético fosse unicamente a ligação fatorial, a distância entre os locos poderia ser estimada por meio da expressão:

$$d = 50 \left(1 - \frac{\Delta}{\Delta_{\max}} \right)$$

Para o exemplo, tem-se:

$$d = 560 \left(1 - \frac{2,5}{16,5} \right) = 42,42 \text{ cMorgans}$$

Numa população F_2 o desequilíbrio máximo é expresso por:

$$\Delta = \frac{1 - 2d}{4}$$

Se $d = 0$, o valor de Δ será de 25%, e se $d = 0,5$ (50 cMorgans), o desequilíbrio de fase gamética provocado pela ligação fatorial será nulo.

Quanto aos demais pares de locos, tem-se:

Loco i	Loco j	p _A	q _a	p _B	q _b	Δ (%)	Δmax(%)	LOD	r ²	D'
1	2	0,61	0,39	0,27	0,73	2,5	16,5	0,1692	0,0133	0,0059
1	3	0,61	0,39	0,67	0,33	2,5	20,2	0,1188	0,0119	0,0031
1	4	0,61	0,39	0,48	0,52	5,2	20,3	0,4499	0,0455	0,0144
1	5	0,61	0,39	0,60	0,40	0,2	15,6	0,0009	0,0001	0,0000
1	6	0,61	0,39	0,27	0,73	0,8	10,6	0,0193	0,0014	0,0006
1	7	0,61	0,39	0,03	0,97	0,5	1,9	0,0505	0,0036	0,0021
1	8	0,61	0,39	0,41	0,59	2,5	16,0	0,0952	0,0109	0,0039
1	9	0,61	0,39	0,36	0,64	3,0	22,0	0,2353	0,0164	0,0064
1	10	0,61	0,39	0,10	0,90	0,9	6,1	0,0544	0,0038	0,0021
2	3	0,73	0,27	0,67	0,33	0,7	9,0	0,0142	0,0011	0,0003
2	4	0,73	0,27	0,48	0,52	1,6	13,0	0,0688	0,0052	0,0020
2	5	0,73	0,27	0,60	0,40	0,4	16,2	0,0039	0,0003	0,0001
2	6	0,73	0,27	0,27	0,73	0,3	19,8	0,0036	0,0002	0,0001
2	7	0,73	0,27	0,03	0,97	2,1	2,2	1,425	0,0769	0,0544
2	8	0,73	0,27	0,41	0,59	2,2	16,0	0,1443	0,0102	0,0044
2	9	0,73	0,27	0,36	0,64	5,0	17,3	0,8949	0,0551	0,0257
2	10	0,73	0,27	0,10	0,90	2,6	2,7	1,4103	0,0381	0,0250
3	4	0,33	0,67	0,48	0,52	0,9	17,2	0,0054	0,0015	0,0005
3	5	0,33	0,67	0,60	0,40	1,4	19,8	0,0178	0,0037	0,0015
3	6	0,33	0,67	0,27	0,73	0,9	9,0	0,0177	0,0019	0,0005
3	7	0,33	0,67	0,03	0,97	0,1	2,1	-0,0001	0,0002	0,0001
3	8	0,33	0,67	0,41	0,59	2,0	19,5	0,0773	0,0075	0,0021
3	9	0,33	0,67	0,36	0,64	1,3	21,2	0,0349	0,0033	0,0008
3	10	0,33	0,67	0,10	0,90	3,2	3,3	0,3449	0,0515	0,0153
4	5	0,52	0,48	0,60	0,40	1,9	19,2	0,0474	0,0060	0,0017
4	6	0,52	0,48	0,27	0,73	0,1	14,1	0,0001	0,0000	0,0000
4	7	0,52	0,48	0,03	0,97	1,5	1,6	0,0716	0,0310	0,0156
4	8	0,52	0,48	0,41	0,59	1,8	21,4	0,0522	0,0054	0,0017
4	9	0,52	0,48	0,36	0,64	0,5	18,8	0,0042	0,0004	0,0001
4	10	0,52	0,48	0,10	0,90	2,7	5,2	0,2015	0,0325	0,0152
5	6	0,40	0,60	0,27	0,73	2,5	10,8	0,1526	0,0132	0,0039
5	7	0,40	0,60	0,03	0,97	1,1	1,2	0,3546	0,0173	0,0067
5	8	0,40	0,60	0,41	0,59	3,5	16,4	0,1859	0,0211	0,0052
5	9	0,40	0,60	0,36	0,64	1,5	14,4	0,0541	0,0041	0,0010
5	10	0,40	0,60	0,10	0,90	2,0	4,0	0,2103	0,0185	0,0067
6	7	0,73	0,27	0,03	0,97	0,2	2,2	0,0131	0,0007	0,0005
6	8	0,73	0,27	0,41	0,59	0,7	16	0,0132	0,0010	0,0004
6	9	0,73	0,27	0,36	0,64	2,0	9,8	0,1196	0,0088	0,0041
6	10	0,73	0,27	0,10	0,90	0,6	7,3	0,0229	0,0020	0,0013
7	8	0,97	0,03	0,41	0,59	0,4	1,8	0,0336	0,0023	0,0013
7	9	0,97	0,03	0,36	0,64	0,4	1,1	0,0103	0,0024	0,0015
7	10	0,97	0,03	0,10	0,90	0,2	0,3	0,0951	0,0015	0,0013
8	9	0,59	0,41	0,36	0,64	1,2	21,3	0,0355	0,0026	0,0010
8	10	0,59	0,41	0,10	0,90	0,2	4,1	0,0019	0,0002	0,0001
9	10	0,64	0,36	0,10	0,90	1,4	6,4	0,1245	0,0095	0,0054

b) Estatística r^2

De maneira geral, consideram-se dois locos, com dois alelos cada, ou seja, A e a, e B e b, com freqüências alélicas p_A, q_a, p_B e q_b (muitas vezes referenciadas na literatura como sendo π_A, π_a, π_B e π_b), respectivamente, resultando nas freqüências gaméticas $\pi_{AB}, \pi_{Ab}, \pi_{aB}$ e π_{ab} para cada possibilidade. O componente básico para o cálculo de desequilíbrio é a diferença entre a freqüência esperada e a observada dos gametas, dada por:

$$\Delta = (\pi_{AB} - p_A p_B)$$

Uma medida do desequilíbrio, denotada r^2 , é fornecida por:

$$r^2 = \frac{(\Delta)^2}{p_A q_a p_B q_b}$$

É conveniente considerar r^2 como o quadrado do coeficiente de correlação entre dois locos. Entretanto, ao menos que os dois locos tenham freqüências alélicas idênticas, o valor 1 não é possível.

c) Estatística D'

Alternativamente, a estatística D' é calculada como:

$$|D'| = \frac{|\Delta|}{\min(p_A p_B, q_a q_b)} \text{ para } \Delta < 0$$

$$|D'| = \frac{|\Delta|}{\min(p_A q_b, q_a p_B)} \text{ para } \Delta > 0$$

O valor de D' é baseado nas freqüências alélicas observadas e irá variar entre 0 e 1, se as freqüências alélicas diferirem entre os locos. D' poderá ser menor do que 1 apenas se todos os quatro possíveis gametas forem observados, assumindo, consequentemente, que eventos de recombinação ocorreram entre os locos.

As estatísticas r^2 e D' refletem diferentes aspectos do desequilíbrio de ligação e comportam-se diferentemente sob condições variadas. No exemplo considerado, todas estatísticas estimadas foram de pequena magnitude indicando que os pares de locos na população não se encontram em desequilíbrio.

Vários fatores influenciam o desequilíbrio de ligação. Alguns deles são responsáveis pelo aumento do desequilíbrio, incluindo as autofecundações, pequenos tamanhos de populações, isolamento genético entre linhagens, subdivisão populacional, baixa taxa de recombinação, mistura populacional, seleção artificial e natural, entre outros. Alguns outros fatores são responsáveis pela queda ou quebra do desequilíbrio, incluindo acasalamento ao acaso, elevadas taxas de recombinação, elevadas taxas de mutações, entre outros. Existem fatores que podem aumentar ou quebrar o desequilíbrio, ou podem aumentar o desequilíbrio entre determinado par de alelos e diminuí-lo entre outros pares de alelos. Por exemplo, a mutação pode romper o desequilíbrio entre pares de alelos envolvendo alelos selvagens e promover desequilíbrio entre os pares de alelos dos mutantes envolvidos. Outros fatores que afetam o desequilíbrio, incluindo estrutura populacional, epistasia e conversão gênica, não têm recebido a atenção desejada nas revisões realizadas (GUPTA et al., 2005).

De forma resumida, os fatores explicados por Flint-Garcia et al. (2003) que afetam o desequilíbrio são apresentados a seguir.

A freqüência alélica e a recombinação entre sítios afetam o desequilíbrio, e a maior parte dos processos observados na genética de populações afeta o desequilíbrio. O tamanho populacional também tem um importante papel. Em pequenas populações, os efeitos da deriva genética resultam em perda consistente de combinações alélicas raras, o que aumenta os níveis de desequilíbrio.

Os sistemas de acasalamento e misturas também podem influenciar fortemente o desequilíbrio; geralmente, ele diminui mais rapidamente em espécies alógamas, comparada com espécies autógamas. Isso porque a recombinação é menos efetiva em espécies que se autofecundam, onde os indivíduos são mais semelhantes por serem homozigotos, do que em espécies de fecundação cruzada.

Mistura é a circulação gênica entre indivíduos geneticamente distintos, seguidas por intercruzamentos. A mistura resulta na introdução de genes de diferentes ancestrais e freqüências alélicas, porém os valores de desequilíbrio caem rapidamente com os cruzamentos aleatórios.

O desequilíbrio pode também ser criado em populações que tiveram recentemente seu tamanho populacional reduzido devido ao afunilamento genético. Durante o afunilamento, apenas poucos alelos combinados são passados para gerações futuras, o que pode gerar um desequilíbrio considerável.

A seleção, que causa afunilamento genético para locos específicos, também pode criar desequilíbrio entre o alelo selecionado e o loco ligado. Além disso, a seleção a favor ou contra um fenótipo controlado por dois genes não ligados (epistasia) pode criar desequilíbrio, embora os locos não estejam fisicamente ligados. Finalmente, o fluxo gênico entre indivíduos de diferentes populações pode introduzir novas combinações cromossômicas e diferentes freqüências alélicas, resultando em desequilíbrio.

Riqueza genotípica

A função de Shannon-Wiener proposta para medir a diversidade ou riqueza de espécies em estudos de ecologia também pode ser empregada para medir a diversidade fenotípica ou genotípica dentro de uma população. Define-se a estatística de Shannon-Wiener como:

$$H' = -\sum_{i=1}^c P_i \ln (P_i)$$

em que:

c: número de classes genotípicas para um dado loco; e

P_i : freqüência do i-ésimo genótipo;

Assim, temos valores de H' para algumas populações:

	Pop1	Pop2	Pop3	Pop4
AA	1	0,5	0,33	0,25
Aa	0	0,5	0,33	0,50
aa	0	0	0,33	0,25
H'	0	0,69	1,09	1,04

Pode ser demonstrado que, para um dado número c de genótipos, o valor de H' poderá atingir o máximo dado por $\ln(c)$, que ocorre quando todos os genótipos estão presentes e com a mesma freqüência. Assim, tem-se:

c	2	3	4	5	6	7	8	9	10
H'_{\max}	0,693	1,098	1,386	1,609	1,791	1,945	2,079	2,197	2,302
	1	6	3	4	7	9	4	2	5

É possível calcular a variância da estatística H' pela seguinte fórmula:

$$V_{H'} = \frac{\sum_{i=1}^c P_i \ln^2(P_i) - \left[\sum_{i=1}^c P_i \ln(P_i) \right]^2}{N} + \frac{c-1}{2N^2},$$

sendo N o tamanho da amostra.

Desse modo, o teste t (de Student) pode ser usado para testar a significância da diferença entre os valores H' de duas populações.

Para o exemplo considerado relativos aos dados da população 2, pode-se relacionar os seguintes genótipos para cada um dos locos estudados e suas respectivas freqüências:

Genótipo	Loco 1	Loco 2	Loco 3	Loco 4	Loco 5	Loco 6	Loco 7	Loco 8	Loco 9	Loco 10
11	0,38	0,58	0,08	0,20	0,06	0,56			0,44	0,02
12	0,20	0,28	0,04	0,06	0,12	0,24	0,06	0,04	0,40	0,16
13	0,22	0,02	0,24	0,58	0,26	0,02			0,10	
14	0,04		0,22		0,02	0,08				
15										
22		0,12			0,06	0,04	0,94	0,12	0,16	0,82
23	0,06				0,24			0,40		
24			0,06		0,02	0,06				
25					0,02					
33	0,04		0,12	0,16	0,12			0,34		
34	0,04		0,10		0,04					
35					0,02					
44			0,14		0,02					
45	0,02									
55										

Assim, para o primeiro loco, a estatística H' é calculada por meio de:

$$H' = -\sum_{i=1}^c P_i \ln(P_i) = -[0,38\ln(0,38) + 0,20\ln(0,20) + \dots + 0,02\ln(0,02)] = 1,6559$$

Pode-se também calcular:

$$\sum_{i=1}^c P_i \ln^2(P_i) = [0,38\ln^2(0,38) + 0,20\ln^2(0,20) + \dots + 0,02\ln^2(0,02)] = 3,4025$$

$$V_H = \frac{\sum_{i=1}^c P_i \ln^2(P_i) - \left[\sum_{i=1}^c P_i \ln(P_i) \right]^2}{N} + \frac{c-1}{2N^2} = \frac{3,4025 - (1,6559)^2}{50} + \frac{8-1}{5000} = 0,0146$$

A estatística t que possibilita testar a hipótese $H_0: H' = 0$ é dada por:

$$t = \frac{H'}{\sqrt{V_{H'}}} = \frac{1,6559}{\sqrt{0,0146}} = 14,0354 \text{ associada a } N-1 \text{ graus de liberdade.}$$

Assim, para $N-1 = 49$ tem-se $\alpha = 0,0001$, ou seja, rejeita-se a hipótese H_0 , concluindo haver diversidade significativa para o loco em consideração.

O intervalo de confiança também pode ser estabelecido a partir da expressão:

$$IC_{1-\alpha}(H') = H' \pm t_{\alpha/2} \sqrt{V_{H'}}$$

O intervalo estabelecido desta maneira terá $1 - \alpha$ de probabilidade de confiança de que conterá o verdadeiro valor de H' .

Quanto aos demais locos, são obtidos os resultados:

Loco	H'	V(H')	t	$\alpha(\%)$	IC _{0,95} (H')	
					LI	LS
1	1,6560	0,0146	14,0354	0,0	1,3962	1,9158
2	1,0050	0,0098	10,1395	0,0	0,7919	1,2182
3	1,9352	0,0066	23,8642	0,0	1,7608	2,1095
4	1,0998	0,0105	10,7565	0,0	0,8800	1,3197
5	2,0592	0,0161	16,2388	0,0	1,7865	2,3318
6	1,2451	0,0177	9,3709	0,0	0,9594	1,5307
7	0,2270	0,0087	2,4278	1,8045	0,0260	0,428
8	1,3468	0,0088	14,3259	0,0	1,1446	1,5489
9	1,0210	0,0029	18,8054	0,0	0,9042	1,1377
10	0,5342	0,0122	4,8349	0,0024	0,2966	0,7717

Comparando os intervalos de confiança, constata-se que a diversidade apresentada pelo loco 1 é estatisticamente superior à apresentada pelo loco 2. A maior variabilidade ou riqueza, em termos de variação genotípica, é encontrada no loco 5, cujo valor de H' é estimado em 2,0592. O menor valor encontrado é relativo ao loco 7, que apresentou apenas dois genótipos na população. Os locos 9 e 10 apresentam três genótipos cada, porém a variabilidade do loco 9 é significativamente superior à do loco 10, dada a melhor distribuição de freqüências destes genótipos neste loco. O valor máximo de H' quando se têm três genótipos é

igual a 1,0986, de forma que o loco 9 pode ser considerado bem próximo da condição ótima de diversidade genotípica.

Riqueza alélica

Em algumas situações o termo riqueza alélica tem sido utilizado para se referir ao número médio ou total de alelos por loco. Segundo Nei (1987), esta estatística é muito influenciada pelo tamanho da amostragem genética (número de locos e de indivíduos), não sendo, por isso, uma boa medida de variabilidade genética para fins de comparação entre amostras de diferentes tamanhos. Entretanto, PETIT et al., 1998 descrevem a riqueza alélica como uma medida que se baseia no princípio da rarefação, utilizado em estudos de ecologia, relacionado à riqueza de espécies em amostras de tamanhos desiguais. Nesse sentido, a técnica consiste na avaliação do número esperado de diferentes alelos entre amostras de tamanhos iguais, retiradas de populações diferentes. A contribuição de cada população para a diversidade total ainda pode ser decomposta em dois componentes: um relacionado ao seu nível intrínseco de diversidade e o outro à sua divergência em relação às demais populações.

5.3. Distância Genética Entre Populações ou Coleções

Diversos métodos para estimar a variação genética e conhecer a estrutura populacional já foram desenvolvidos, com aplicações variadas em níveis individual, intrapopulacional e interpopulacional.

A utilização das freqüências alélicas, ou genotípicas, provenientes de diferentes marcas moleculares (locos) tem sido comum para quantificar distâncias genéticas a serem utilizadas na construção de árvores filogenéticas para um grupo de populações ou espécies de parentesco próximo. Além de distâncias, outras medidas, como heterozigosidade, grau de fixação e correlação intergenotípica, têm sido úteis para avaliar o grau de diversidade que expressa a similaridade e o

distanciamento genético entre populações, provocado por fatores genéticos e ambientais. De acordo com Robinson (1998), dentre as medidas mais utilizadas em trabalhos com marcadores moleculares, destacam-se aquelas baseadas nas estatísticas H de Nei e F de Wright. Outra medida importante é a heterozigosidade. Será considerado o método proposto por Nei (1978) sobre as estimativas não-viesadas de diversidade genética (H_s). Também serão consideradas as metodologias propostas por Lynch e Milligan (1994) e Zhivotovsky (1999) adequadas a marcadores dominantes.

As diferenças genéticas entre as populações são pequenas e freqüentemente quantificadas em termos de diferenças das freqüências alélicas e, em estudos filogenéticos, pelas diferenças nucleotídicas ou de aminoácidos. Naturalmente que qualquer diferença alélica é causada por diferenças na seqüência de nucleotídeos entre alelos.

Não menos importante, no melhoramento genético a utilização de informações de freqüências alélicas tem direcionado estudos de diversidade genética entre os acessos que compõem os bancos de germoplasma e de materiais elites, como linhagens endogâmicas, cujo conhecimento da similaridade genética (ou distância) entre elas é de fundamental importância para o estabelecimento de grupos heteróticos e o desenvolvimento de híbridos promissores.

Usualmente, as medidas de distância têm sido utilizadas para estimar o tempo decorrido de divergência entre as populações a partir de uma população ancestral e a construção de árvores filogenéticas entre as populações. Conseqüentemente, é requerido um modelo genético apropriado para especificar o processo de mutação ou de deriva genética que causou a diferenciação entre as populações. Inúmeros autores têm proposto diferentes medidas de distância, sob vários pontos de vista.

Goodman (1972) ressalta que valores iguais de certas medidas de distâncias geométricas podem não indicar uma equivalência de divergência biológica, e uma similar falta de correspondência entre distância e grau de divergência pode ocorrer se distâncias baseadas em mutação são aplicadas a situações de deriva genética.

Medidas de distância genética com base em propriedades geométricas

Uma distância D_{XY} no espaço euclidiano será métrica caso siga as seguintes condições:

- $D_{XY} \geq 0$, as distâncias não são negativas.
- $D_{XX} = 0$ identifica o ponto X.
- $D_{XY} = D_{YX}$, as distâncias são simétricas.
- $D_{XY} = 0$ se $X = Y$ (conclusividade da função de distância).
- Para qualquer três pontos X, Y e Z, $D_{XZ} + D_{YZ} \geq D_{XY}$; assim, as distâncias satisfazem a inequação triangular.

Essencialmente, o atendimento às duas últimas propriedades possibilita qualificar a distância como sendo métrica (DIAS, 1998). A caracterização de uma distância no espaço euclidiano é importante porque esta é uma pressuposição explícita ou implícita em várias técnicas de análise multivariada, como na análise de coordenadas principais, agrupamento hierárquico e dispersão gráfica. A propriedade euclidiana é um aspecto desejável, mas o principal critério para a escolha de uma medida de distância são suas propriedades genéticas, que vão ao encontro dos objetivos da investigação acerca das populações. Segundo Weir (1996), distâncias geométricas devem satisfazer critérios diferentes em relação às distâncias genéticas, o que não impede que os modelos evolucionários sejam apropriados para as distâncias geométricas.

Ilustração

Como ilustração, serão considerados os dados da Tabela A4 (anexos) relativos à avaliação de cinco populações e dois marcadores (ou locos gênicos). Verifica-se que as populações encontram-se em equilíbrio de Hardy-Weinberg para todos os dois locos estudados, de acordo com os resultados obtidos e descritos a seguir:

População	Marcador	GL	χ^2	Probabilidade	f(A ₁)	f(A ₂)	f(A ₃)	f(A ₄)	f(A ₅)
1	1	1	0,006	0,9381	ns	0,35	0,65		
	2	3	1,7136	0,6339	ns	0,43	0,42	0,15	
2	1	6	0,8397	0,991	ns	0,68	0,16	0,15	0,01
	2	1	0,2833	0,5946	ns	0,31	0,69		
3	1	3	1,0766	0,7827	ns	0,42	0,12	0,46	
	2	3	1,4564	0,6924	ns	0,23	0,33	0,44	
4	1	1	0,1198	0,7292	ns	0,78	0,22		
	2	1	0,2142	0,6435	ns	0,27	0,73		
5	1	6	3,9999	0,6767	ns	0,88	0,06	0,0	0,06
	2	10	3,8207	0,9551	ns	0,01	0,15	0,35	0,33
									0,16

As populações diferem quanto às suas freqüências genotípicas e alélicas. A partir dessas informações, é possível calcular diferentes medidas de distância entre as populações e inferir quais aquelas de maior diferenciação.

No cálculo das distâncias genéticas entre populações será considerada a simbologia:

X_{ijt} : expressa o valor do indivíduo t, referente ao loco j na população i sendo:

i = 1,2,...g, sendo g o número total de populações avaliadas;

j = 1,2,...L, sendo L o número total de locos estudados; e

t = 1,2,...n_i, sendo n_i o número de indivíduos avaliado na população i

$$N = \sum_{i=1}^g n_i$$

No exemplo considerado, tem-se g=5, L=2 e n₁ = n₂ = ... n₅ = 50.

Também deve ser considerado o valor p_{ijk}, que expressa a freqüência do alelo k, do loco j na população i, sendo:

k = 1,2,...a_j, sendo a_j o número de alelos do loco j

$$A = \sum_{ij=1}^L a_j$$

Em algumas medidas de distância utiliza-se a informação $P_{ijmm'}$ que expressa a freqüência do genótipo mm' , do loco j na população i .

Distância Euclidiana

Considere as populações i e i' no plano bidimensional, caracterizado por dois alelos de um único loco, conforme a Figura 5.2. Para a população i , e um particular loco j , têm-se as freqüências alélicas p_{i1} e $p_{i2} = 1 - p_{i1}$, e para a população i' , $p_{i'1}$ e $p_{i'2}=1-p_{i'1}$. A diferença em projeções sobre a abscissa ($p_{i1} - p_{i'1}$) e sobre a ordenada ($p_{i2} - p_{i'2}$) são os catetos de um triângulo retângulo de hipotenusa igual à distância euclidiana,

$D_{Eii'} = \sqrt{(p_{i1} - p_{i'1})^2 + (p_{i2} - p_{i'2})^2}$, medida que se baseia no teorema de Pitágoras, porém aplicada a inúmeros eixos ortogonais.

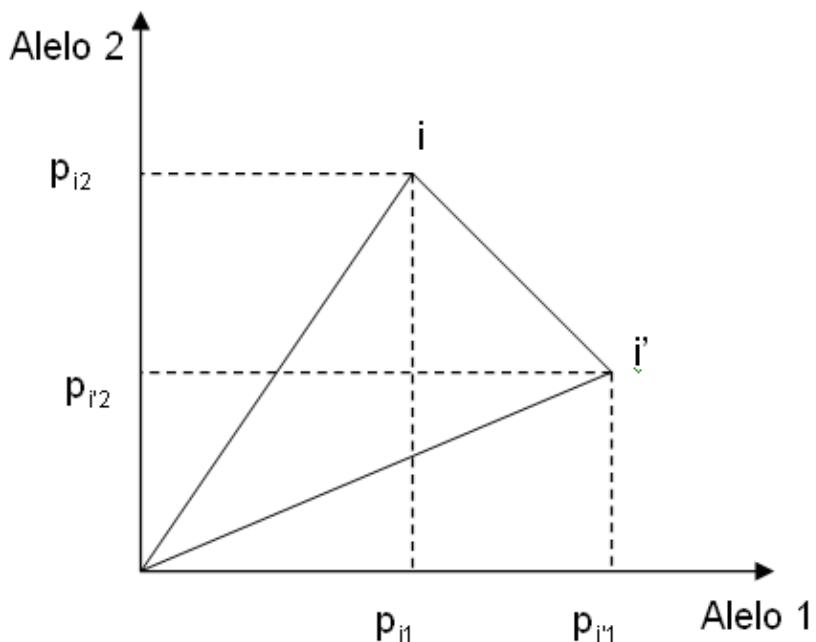


Figura 5.2 - Representação da distância as populações i e i' com base nas informações de freqüências alélicas p_{i1} e p_{i2} .

Considere p_{ijk} e $p_{i'jk}$ a freqüência do k-ésimo alelo para o j-ésimo loco das respectivas populações i e i' ; a_j , o número de alelos do j-ésimo loco; e L, o número de locos analisados.

De forma generalizada, a distância euclidiana pode ser expressa da seguinte forma:

$$D_{Eii'} = \sqrt{\sum_{j=1}^L \sum_{k=1}^{a_j} (p_{ijk} - p_{i'jk})^2}$$

Cada população é representada como um ponto no espaço **A** ($= \sum_{j=1}^L a_j$)

dimensional, com base nas freqüências alélicas dos **A** alelos. Para que duas populações sejam consideradas similares, elas devem ocorrer na mesma região do espaço multidimensional, com pequena distância entre si (DIAS, 1998).

Para L locos, a distância euclidiana assume valores entre 0 e $\sqrt{2L}$, cujos limites são obtidos quando as duas populações têm freqüências alélicas idênticas ou estão fixadas para diferentes alelos, respectivamente. Naturalmente, a distância euclidiana é mais apropriada quando as informações alélicas estão disponíveis e a divergência entre as populações é investigada por técnicas multivariadas que apresentam uma caracterização euclidiana. Como referência, consideram-se as informações:

População	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇	A ₈
1	1	0	0	0	0	0	0	0
2	0	1/7	1/7	1/7	1/7	1/7	1/7	1/7
3	0	0	1/6	1/6	1/6	1/6	1/6	1/6
4	0	0	0	0	0	0	0	1
5	1/4	1/4	1/4	1/4	0	0	0	0
6	0	0	0	0	1/4	1/4	1/4	1/4

As medidas de dissimilaridade baseadas na distância euclidiana são apresentadas a seguir:

Populações	1	2	3	4	5	6
1	0,0	1,0690	1,0801	1,4142	0,8660	1,118
2	1,0690	0,0	0,1543	0,9258	0,4226	0,3273
3	1,0801	0,1543	0,0	0,9129	0,5000	0,2887
4	1,4142	0,9258	0,9129	0,0	1,1180	0,8660
5	0,8660	0,4226	0,5000	1,1180	0,0	0,7071
6	1,1180	0,3273	0,2887	0,8660	0,7071	0,0

Apesar de muitos pares de populações não compartilharem alelos comuns, a dissimilaridade máxima ocorrerá entre as populações 1 e 4 com valor $\sqrt{2}$, em que, além de não haver compartilhamento, há fixação de alelos em ambas as populações.

Para o exemplo cujos dados encontram-se na Tabela A4 (Anexos), podem ser obtidas as seguintes medidas de distâncias:

Populações	Loco 1 (DE_1)	Loco 2 (DE_2)	Locos 1 e 2 (DE_{12})*
1 e 2	0,6096	0,3314	0,6938
1 e 3	0,7053	0,3636	0,7935
1 e 4	0,6081	0,3797	0,7169
1 e 5	0,7954	0,6510	1,0278
2 e 3	0,4067	0,5741	0,7036
2 e 4	0,1903	0,0566	0,1985
2 e 5	0,2739	0,7991	0,8447
3 e 4	0,5926	0,5960	0,8405
3 e 5	0,6560	0,4727	0,8086
4 e 5	0,1980	0,8130	0,8368

$$(*) DE_{12} = \sqrt{DE_1^2 + DE_2^2}$$

Outras formas de expressar a distância entre populações são:

- a) Quadrado da distância euclidiana
- b) Distância euclidiana média
- c) Quadrado da distância euclidiana média

A partir da matriz de distância, é possível estabelecer árvores (Figura 5.3), grupos e projeções (Figura 5.4). Todas as técnicas de agrupamentos realçam a similaridade entre as populações 2 e 4 e o distanciamento da população 5 em relação às demais.

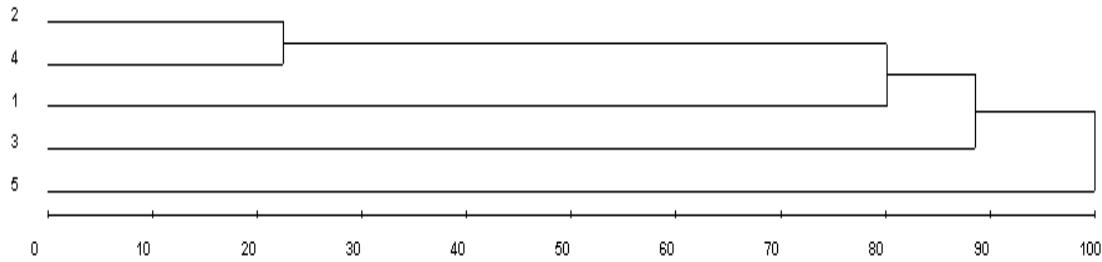


Figura 5.3 - Dendrograma estabelecido pelo método UPGMA a partir de distâncias euclidianas entre cinco populações (dados originais na Tabela A4 – Anexos).

Pelo agrupamento de Tocher, foi verificada a formação de dois grupos, em que a população 5 ficou alocada em um grupo e as demais em outro. Esse distanciamento relativo da população 5 também pode ser constatado na projeção 3D, que será apresentada a seguir. As razões desse padrão de agrupamento deverão ser objeto de investigação de pesquisadores, que, diante de evidências históricas sobre a população e informações biológicas, como isolamento reprodutivo, sistema de acasalamento e fluxo gênico, formularão hipóteses sobre a sua ocorrência e estabelecerão estratégias para melhor uso e conservação dessas populações.

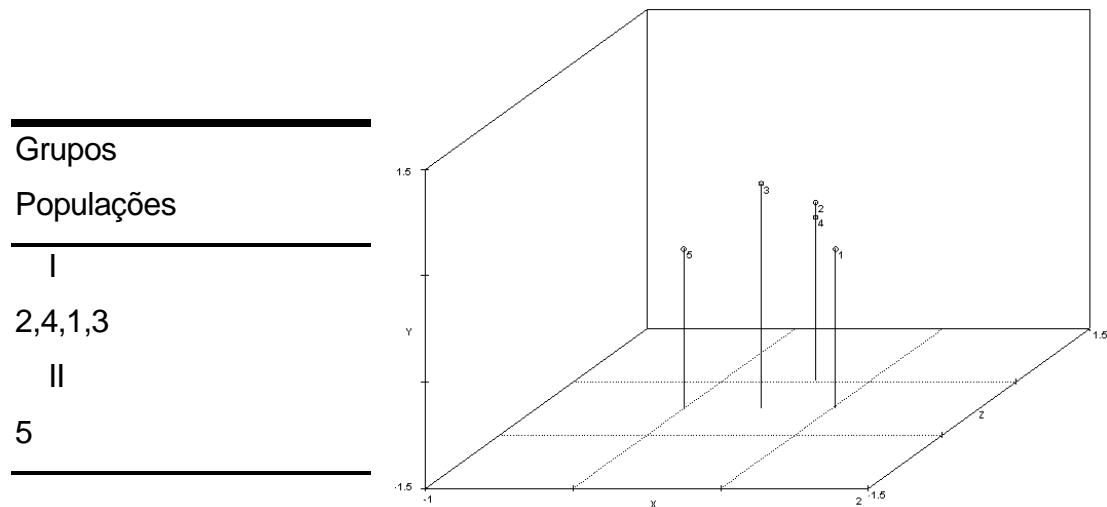


Figura 5.4 - Agrupamento de otimização, pelo método de Tocher, e projeção 3D (distorção igual a 0,61%, estresse de 1,17% e correlação entre distâncias originais e gráficas de 0,9998) a partir de distâncias euclidianas entre cinco populações.

Distância de Rogers

Rogers (1972) propôs uma modificação na distância euclidiana, de modo que a distância sugerida corresponderia à distância euclidiana média em relação a todos os locos, justamente para contornar o problema do número desigual de locos avaliados em diferentes estudos e tornar duas estimativas de distância euclidiana comparáveis. Além disso, outra questão relaciona-se ao limite superior ($\sqrt{2L}$) possível de ser atingido pela distância euclidiana quando duas populações não compartilham alelos comuns, o que é pouco desejável para caracterizar a dissimilaridade entre duas populações.

Assim, Rogers (1972) padronizou a distância pelo fator $\sqrt{1/2}$, limitando os valores de distância ao intervalo [0, 1]. Assim, tem-se:

$$D_{Rij} = \frac{1}{L} \sum_{j=1}^L \sqrt{\frac{1}{2} \sum_{k=1}^{n_j} (p_{ijk} - p'_{ijk})^2}$$

Essa distância (D_{Rii}) foi desenvolvida assumindo-se não existir qualquer conhecimento sobre as forças evolucionárias capaz de promover a divergência entre duas populações sob análise.

No entanto, a distância D_R também apresenta a propriedade de caracterizar como máxima diversidade as populações fixadas que não compartilham alelos comuns. Por exemplo, considere um loco com oito alelos com as freqüências descritas a seguir:

População	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
1	1	0	0	0	0	0	0	0
2	0	1/7	1/7	1/7	1/7	1/7	1/7	1/7
3	0	0	1/6	1/6	1/6	1/6	1/6	1/6
4	0	0	0	0	0	0	0	1
5	1/4	1/4	1/4	1/4	0	0	0	0
6	0	0	0	0	1/4	1/4	1/4	1/4

Para os dados apresentados anteriormente, seriam obtidas as seguintes medidas de distâncias entre os pares das seis populações estudadas:

Populações	1	2	3	4	5	6
1	0,0000	0,7559	0,7638	1,000	0,6124	0,7906
2	0,7559	0,0000	0,1091	0,6547	0,2988	0,2315
3	0,7638	0,1091	0,0000	0,6455	0,3536	0,2041
4	1,0000	0,6547	0,6455	0,0000	0,7906	0,6124
5	0,6124	0,2988	0,3536	0,7906	0,0000	0,5000
6	0,7906	0,2315	0,2041	0,6124	0,5000	0,0000

Assim, por exemplo, considerando as populações 5 e 6, com freqüências iguais a $\frac{1}{4}$, sendo que quatro alelos não são compartilhados entre as duas populações, a distância de Rogers calculada foi igual a apenas 0,5.

A deficiência dessa medida surge quando duas populações são polimórficas e não compartilham alelos comuns, pois o valor de D_R pode ser muito inferior a 1, mesmo quando as populações apresentam seus conjuntos gênicos completamente

diferentes. Ela será igual a 1 somente se duas populações estiverem fixadas para diferentes alelos, como visto nas populações 1 e 4.

A distância de Rogers apresenta propriedades métricas que satisfazem a pressuposição de alguns métodos de construção de árvores filogenéticas (Nei et al., 1983).

Para os dados da Tabela A4 (Anexos), seriam obtidas as seguintes medidas de distâncias entre os pares de populações estudadas:

Populações	Loco 1 (DR_1)	Loco 2 (DR_2)	Locos 1 e 2 (DR_{12})*
1 e 2	0,4310	0,2343	0,3327
1 e 3	0,4987	0,2571	0,3779
1 e 4	0,4300	0,2685	0,3493
1 e 5	0,5624	0,4603	0,5114
2 e 3	0,2876	0,4060	0,3468
2 e 4	0,1345	0,0400	0,0873
2 e 5	0,1936	0,5651	0,3794
3 e 4	0,4190	0,4214	0,4202
3 e 5	0,4639	0,3342	0,3991
4 e 5	0,1400	0,5749	0,3574

$$(*)DR_{12} = (DR_1 + DR_2)/2$$

A partir da matriz de distância é possível estabelecer árvores (Figura 5.5), grupos e projeções (Figura 5.6). As técnicas de agrupamentos realçam também a similaridade entre as populações 2 e 4 e o distanciamento da população 5 em relação às demais. O agrupamento, pelo método de Tocher, evidenciou a formação de três grupos.

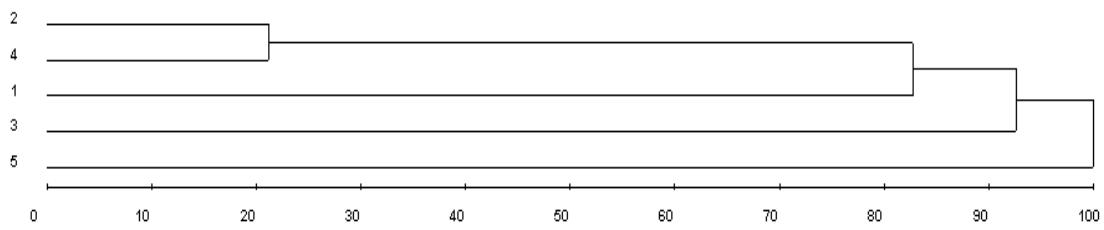


Figura 5.5 - Dendrograma estabelecido pelo método UPGMA a partir de distância de Rogers entre cinco populações.

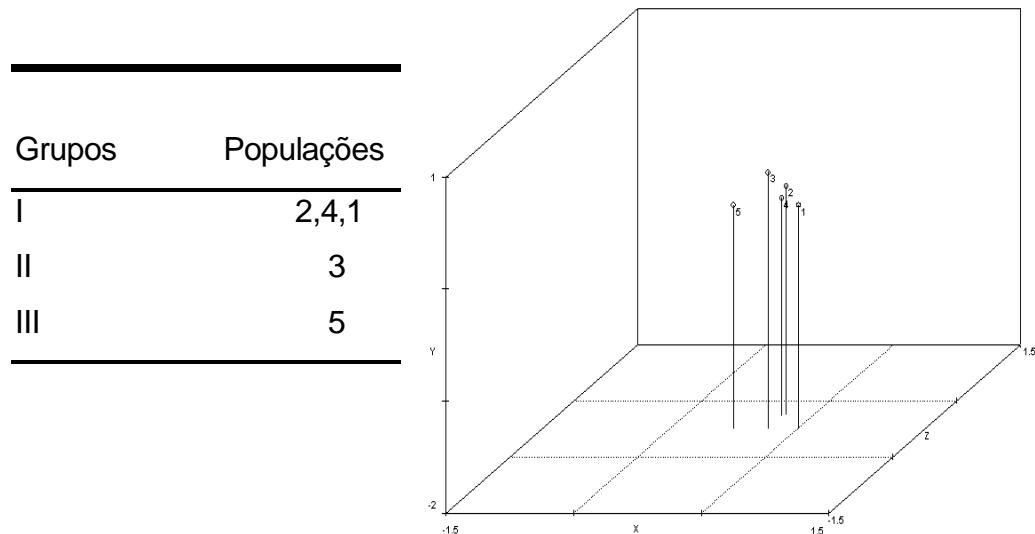


Figura 5.6 - Agrupamento de otimização, pelo método de Tocher, e projeção 3D (distorção de 0,80%, estresse de 1,50% e correlação entre distâncias originais e gráficas de 0,9997) a partir de distâncias de Rogers entre cinco populações.

Distância modificada de Rogers

Goodman e Stuber (1983) propuseram mudança na distância de Rogers no sentido de designar a cada alelo uma dimensão no espaço. Para isso, a distância modificada de Rogers é calculada por meio da expressão:

$$D_{GS\ ii'} = \frac{1}{\sqrt{2L}} \sqrt{\sum_{j=1}^L \sum_{k=1}^{n_j} (p_{ijk} - p_{i'jk})^2}$$

A distância modificada de Rogers varia no intervalo [0, 1]. Assim como D_R , o valor de D_{GS} não será igual a um, no caso de múltiplos alelos, mesmo quando as duas populações não compartilharem alelos em comum.

$D_{GS14} = \frac{1}{\sqrt{2}} \sqrt{[(1-0)^2 + \dots + (0-1)^2]} = 1$ (máxima distância, considerando um único loco)

$$D_{GS56} = \frac{1}{\sqrt{2}} \sqrt{[8(1/4 - 0)^2]} = \frac{1}{2}$$

Também nesse caso, apesar de as populações 5 e 6 não compartilharem alelos comum, a distância entre elas não é máxima.

Para o exemplo cujos dados se encontram na Tabela A4 (Anexos), podem ser obtidas as seguintes medidas de distâncias:

Populações	Rogers			Rogers modificada		
	Loco 1	Loco 2	Média	Loco 1	Loco 2	Média
1 e 2	0,4310	0,2343	0,3327	0,3716	0,1098	0,3469
1 e 3	0,4987	0,2571	0,3779	0,4974	0,1322	0,3967
1 e 4	0,4300	0,2685	0,3493	0,3698	0,1442	0,3585
1 e 5	0,5624	0,4603	0,5114	0,6326	0,4238	0,5139
2 e 3	0,2876	0,4060	0,3468	0,1654	0,3296	0,3518
2 e 4	0,1345	0,0400	0,0873	0,0362	0,0032	0,0992
2 e 5	0,1936	0,5651	0,3794	0,075	0,6386	0,4224
3 e 4	0,4190	0,4214	0,4202	0,3512	0,3552	0,4202
3 e 5	0,4639	0,3342	0,3991	0,4304	0,2234	0,4043
4 e 5	0,1400	0,5749	0,3574	0,0392	0,661	0,4184

Há boa concordância entre as duas medidas de dissimilaridade; ambas apontaram as populações 2 e 4 como as mais similares e a 1 e 5 como as mais divergentes.

Distância angular

Outra forma de representação da distância geométrica entre duas populações é considerá-las numa hiperesfera n-dimensional, cuja medida de distância é dada pelo ângulo (ψ) formado entre as linhas de projeção que unem a origem aos respectivos pontos das populações (CAVALLI-SFORZA; EDWARDS, 1967; EDWARDS 1971). Esse ângulo corresponde ao comprimento da corda ($\sqrt{2D_{CE}^2}$), ligando as duas populações.

Um melhor entendimento se dá quando se considera apenas um loco, com dois alelos para cada uma das populações (i e i'), situadas em um quarto de círculo de raio 1, cujas coordenadas são as raízes quadradas de suas freqüências alélicas p_{ik} e $p_{i'k}$, conforme a Figura 5.7.

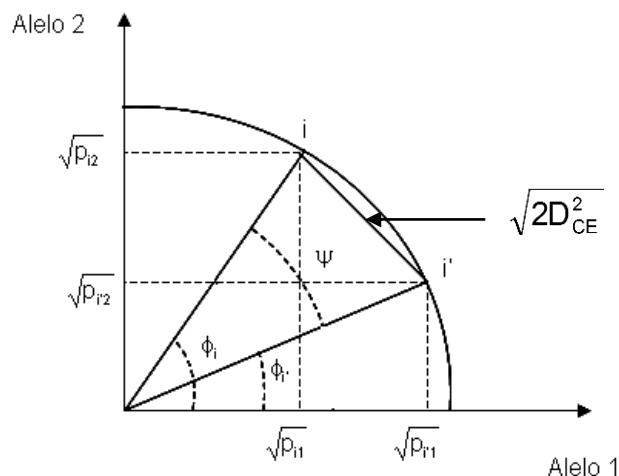


Figura 5.7 - Representação de duas populações (i e i') de coordenadas $\sqrt{p_{ik}}$ e $\sqrt{p_{i'k}}$, cujos pontos situam-se em um quarto de círculo centrado na origem.

O raio da circunferência pode ser obtido por meio de:

$$r^2 = (\sqrt{p_{i1}})^2 + (\sqrt{p_{i2}})^2 = p_{i1} + p_{i2} = 1$$

logo, $r=1$.

Distância medida pelo ângulo Ψ

Sendo ϕ_i e $\phi_{i'}$ os ângulos formados entre o alelo 1 (eixo 1) e as duas linhas de projeção para i e i' , respectivamente, o ângulo ψ entre estas duas linhas é $(\phi_i - \phi_{i'})$, e o seu cosseno é dado por:

$$\cos(\psi) = \cos(\phi_i - \phi_{i'}) = \cos(\phi_i)\cos(\phi_{i'}) + \sin(\phi_i)\sin(\phi_{i'})$$

sendo:

$$\text{Cos}(\phi_i) = \frac{1}{r} \sqrt{p_{i1}} = \sqrt{p_{i1}}$$

$$\text{Cos}(\phi_{i'}) = \frac{1}{r} \sqrt{p_{i'1}} = \sqrt{p_{i'1}}$$

$$\text{Sen}(\phi_i) = \frac{1}{r} \sqrt{p_{i2}} = \sqrt{p_{i2}}$$

$$\text{Sen}(\phi_{i'}) = \frac{1}{r} \sqrt{p_{i'2}} = \sqrt{p_{i'2}}$$

logo:

$$\cos(\psi) = \sqrt{p_{i1}p_{i'1}} + \sqrt{p_{i2}p_{i'2}} \quad (\text{para um loco com dois alelos})$$

Para a₁ alelo por loco, tem-se

$$\cos(\psi) = \sum_{k=1}^{a_1} \sqrt{p_{ik}p_{i'k}}$$

Se as populações i e i' não apresentam alelos em comum, tem-se $\cos(\psi) = 0$. Caso as freqüências alélicas sejam as mesmas ($i = i'$), então $\cos(\psi) = 1$. Assim, é adequado caracterizar o $\cos(\psi)$ como uma medida de similaridade. A medida de distância pode ser representada pelo seu complemento aritmético, ou seja:

$$D_{CE\ ii'}^2 = 1 - \cos(\psi) = 1 - \sum_{k=1}^{a_1} \sqrt{p_{ik}p_{i'k}}$$

e

$$D_{CEii'} = \sqrt{1 - \left(\sum_{k=1}^{n_1} \sqrt{p_{ik} p_{i'k}} \right)}$$

Generalizando para L locos, tem-se:

$$D_{CE ii'} = \sqrt{\frac{1}{L} \sum_{j=1}^L \left(1 - \sum_{k=1}^{a_j} \sqrt{p_{ik} p_{i'k}} \right)},$$

Distância medida pelo comprimento da corda

A corda de uma circunferência é um segmento de reta cujas extremidades pertencem à circunferência. O comprimento da corda é dado por:

$$cc = 2[1 - \cos(\psi)]$$

Se $\Psi = 0$, então $cc = 0$, e se $\Psi = 90^\circ$, tem-se $cc = \sqrt{2}$.

Assim, pode-se expressar:

$$D_{CE ii'}^2 = cc$$

de forma que:

$$D_{CE ii'} = \sqrt{cc} = \sqrt{2[1 - \cos(\Psi)]} = \sqrt{2[1 - \sum_{k=1}^{a_1} p_{ik} p_{i'k}]} \text{ para um loco com } a_1 \text{ alelos}$$

Os valores possíveis da distância angular estão compreendidos entre 0 e $\sqrt{2}$. Particularmente, para duas populações que não compartilham alelos comuns, a distância angular é igual a um mesmo nas situações de múltiplos alelos, sendo uma vantagem sobre D_R e D_{GS} .

A expressão geral para o comprimento da corda é dada por:

$$D_{CE ii'} = \frac{1}{L} \sum_{j=1}^L \sqrt{2 - 2 \left(\sum_{k=1}^{a_j} \sqrt{p_{ik} p_{i'k}} \right)}, \text{ variando de } 0 \text{ a } \sqrt{2}.$$

Cavalli-Sforza e Edwards (1967) e Edwards (1971) definiram ainda a distância (genética) de corda como:

$$D_{CEii'} = \kappa \sqrt{1 - \cos(\psi)}$$

em que $\cos(\psi) = \frac{1}{L} \sum_{j=1}^L \left(\sum_{k=1}^{a_j} \sqrt{p_{ijk} p_{i'jk}} \right)$ e o valor de κ corresponde à função-suporte de distâncias de corda.

Para $\kappa = 1$, D_{CE} varia de 0 a 1.

Para $\kappa = \sqrt{2}$, D_{CE} varia de 0 a $\sqrt{2}$.

Para $\kappa = \frac{2}{\pi} \sqrt{2}$, D_C varia de 0 a $\frac{2}{\pi} \sqrt{2}$.

Nei et al. (1983) sugeriram uma distância mais apropriada a populações filogeneticamente mais próximas, sendo menos influenciada pela presença de alelos raros na amostra. Quando $k = 1$, tem-se:

$$D_A = 1 - \frac{1}{L} \sum_{j=1}^L \left(\sum_{i=1}^{a_j} \sqrt{p_{ijk} p_{i'jk}} \right)$$

Para o exemplo em consideração (seis populações e oito alelos), são obtidas as medidas de dissimilaridade:

Populações	1	2	3	4	5	6
1	0,0000	1,0000	1,0000	1,0000	0,7071	1,0000
2	1,0000	0,0000	0,2724	0,7887	0,6581	0,494
3	1,0000	0,2724	0,0000	0,7693	0,7693	0,4284
4	1,0000	0,7887	0,7693	0,0000	1,0000	0,7071
5	0,7071	0,6581	0,7693	1,0000	0,0000	1,0000
6	1,0000	0,4940	0,4284	0,7071	1,0000	0,0000

Veja que, agora, tanto as populações fixadas (1 e 4) e quanto as que não compartilham alelos comuns (5 e 6) apresentam diversidade máxima igual a 1,0.

Distância medida pelo comprimento da corda (modificado)

De acordo com Nei e Kumar (2000), um ângulo (ψ) = $\frac{\pi}{2} = 90^\circ$ corresponde a uma completa substituição gênica; na prática, usa-se:

$$D_{CE} = \frac{2}{\pi L} \sum_{j=1}^L \sqrt{2(1 - \sum_{k=1}^{a_j} \sqrt{p_{ik} p_{ik}})} ,$$

Como os dados são representados num espaço euclidiano, a escala é dada por uma unidade de distância por substituição gênica.

A distância angular baseia-se no modelo de deriva genética (seletiva) de Kimura (1954), assumindo que a taxa de mutação é pequena e a variação da pressão de seleção é rápida e desordenada (sem direção constante na mudança da freqüência alélica). Há dúvidas se acessos de bancos de germoplasma e materiais melhorados tenham evoluído de acordo com esse modelo, pois nestas situações a pressão de seleção é direcionada, em vez de rápida e desordenada. Contudo, se as informações de freqüências alélicas estão disponíveis e pode-se pressupor o modelo de Kimura (1954), então D_{CE} é um coeficiente apropriado para examinar as relações filogenéticas entre as populações.

Para o exemplo cujos dados se encontram na Tabela A4 (Anexos), podem ser obtidas as seguintes medidas de distâncias:

Populações	Complemento do cosseno			Comprimento da Corda			
	Loco 1	Loco 2	Locos 1 e 2	Loco 1	Loco 2	Média	Modificado
1 e 2	0,1897	0,0966	0,3783	0,6159	0,4395	0,5277	0,3359
1 e 3	0,3373	0,0563	0,4436	0,8214	0,3356	0,5785	0,3683
1 e 4	0,0994	0,1056	0,3201	0,4458	0,4595	0,4526	0,2881
1 e 5	0,2475	0,4543	0,5924	0,7036	0,9532	0,8284	0,5274
2 e 3	0,0643	0,2558	0,4001	0,3587	0,7153	0,5370	0,3419
2 e 4	0,0841	0,0010	0,2062	0,4101	0,0441	0,2271	0,1446
2 e 5	0,1040	0,6226	0,6027	0,4560	1,1159	0,7859	0,5003
3 e 4	0,2652	0,2600	0,5124	0,7282	0,7211	0,7247	0,4613
3 e 5	0,3072	0,3371	0,5676	0,7838	0,8211	0,8025	0,5109
4 e 5	0,0566	0,6171	0,5804	0,3365	1,1110	0,7237	0,4607

Os dendrogramas obtidos utilizando as medidas de distâncias especificadas são apresentados na Figura 5.8

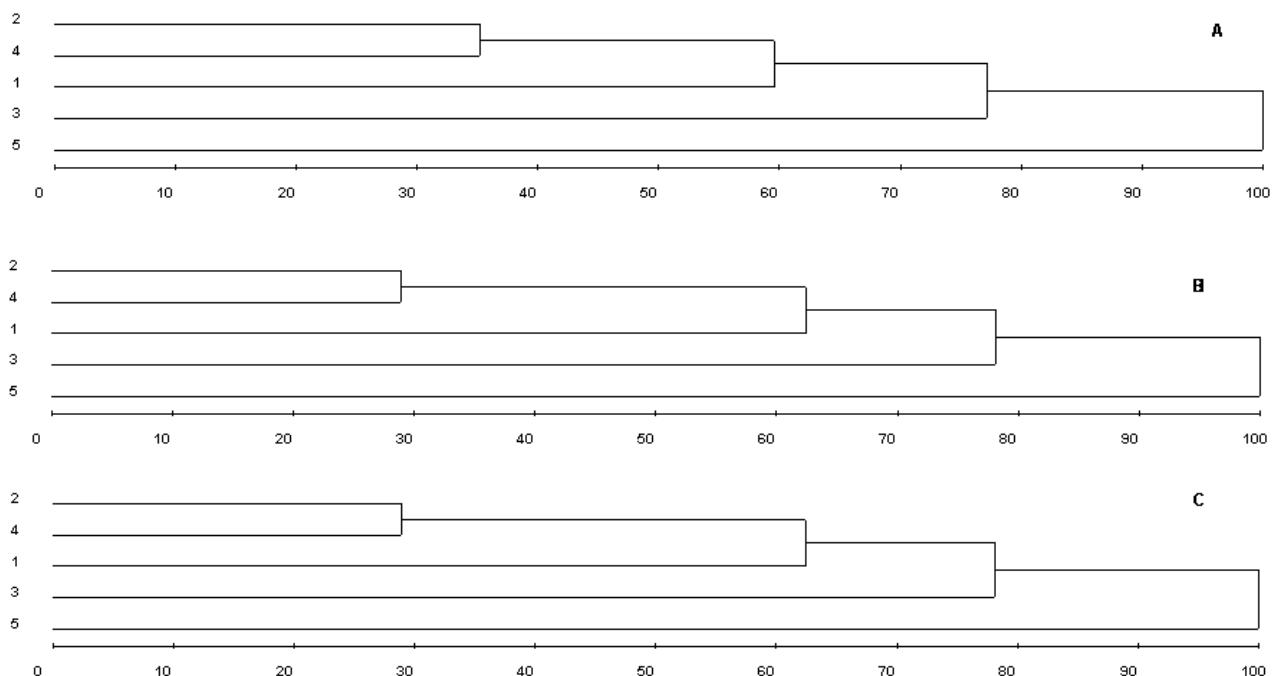


Figura 5.8 - Dendrograma estabelecido pelo método UPGMA a partir de distâncias estabelecidas pelo complemento do cosseno, corda e corda modificada, entre cinco populações avaliadas em relação a dois marcadores co-dominantes multialélicos (Tabela A4 – Anexos).

O resultado do método de agrupamento de Tocher, para as três medidas de dissimilaridade, revelou existirem dois grupos: o primeiro, formado pelas populações 2,4, 1 e 3; e o último, pela população 5.

Nei et al. (1983), em estudo de simulação, sugeriram uma medida de dissimilaridade que se mostrou bastante eficiente para o restabelecimento da verdadeira topologia da árvore evolucionária quando reconstruída a partir de dados de freqüências alélicas (NEI; KUMAR, 2000), dada por:

$$D_{N83} = \frac{1}{L} \sum_{j=1}^L \left(1 - \sum_{k=1}^{n_j} \sqrt{p_{ijk} p_{i'jk}} \right)$$

Para linhagens endogâmicas homozigotas, $D_{N83} = D_R$, essa distância é igual a D_{CE}^2 e assume valores entre zero e um, sendo o último valor obtido quando as duas populações não compartilham alelos em comum. Uma vez que o valor máximo de D_{N83} é igual a um, essa distância não assume relação linear com o número de substituições alélicas. No entanto, quando D_{N83} é pequena, ela aumenta a relação linear com o tempo de evolução (NEI; KUMAR, 2000)..

A Medida D_{N83} não foi desenvolvida sob um modelo genético específico nem é caracterizada como distância métrica e euclidiana. Com isso, Rief et al. (2005) questionaram a aplicabilidade dessa distância em estudos filogenéticos

Distância medida pelo arco

Quando a distância genética é função do arco-cosseno, tem-se:

$$\theta_{ii'} = \frac{1}{L} \sum_{j=1}^{a_i} \left[\frac{2}{\pi} \arccos \left(\sum_{k=1}^{a_j} \sqrt{p_{ijk} p_{i'jk}} \right) \right]$$

Se o ângulo ψ é zero, o valor de $\theta_{ii'}$ será nulo. Se o ângulo ψ for de 90 graus, o arco-cosseno será $\pi/2$, de forma que $\theta_{ii'}$ atinge seu valor máximo igual a 1.

Outras medidas de distância citadas na literatura são apresentadas a seguir:

Distância absoluta

$$D_{ABii'} = \frac{1}{L} \sum_{j=1}^L \sum_{k=1}^{a_j} |p_{ijk} - p_{i'jk}|$$

Similarmente, Provesti et al. (1975) definiram a distância genética como:

$$C_{pii'} = \frac{1}{2L} \sum_{j=1}^L \sum_{k=1}^{a_j} |p_{ijk} - p_{i'jk}|$$

Distância de qui-quadrado

$$\chi^2_{ii'} = \frac{2}{L} \sum_{j=1}^L \sum_{k=1}^{a_j} \frac{(p_{ijk} - p_{i'jk})^2}{(p_{ijk} + p_{i'jk})}$$

Neste caso, consideram-se apenas as comparações em que $p_{ijk} + p_{i'jk}$ seja diferente de zero.

Distância de Reynolds, Weir e Cockerham (ignorando os termos envolvendo tamanho n amostral)

$$\theta_{RWC} = \frac{\sum_{j=1}^L \sum_{k=1}^{a_j} (p_{ijk} - p_{i'jk})^2}{2 \sum_{j=1}^L \left(1 - \sum_{i=1}^{a_j} p_{ijk} p_{i'jk} \right)}$$

Medidas de distância genética com base em modelos evolucionários (evolução a longo prazo)

Ao se assumir que a população ancestral, bem como as populações derivadas dela, diverge em função do balanço ou equilíbrio existente entre mutação e deriva genética, implica dizer que tais populações divergem devido ao aparecimento de novos mutantes dentro das populações. Então, as distâncias podem ser usadas do ponto de vista filogenético, como estimadores do tempo

decorrido de divergência. Em genética de populações são adotados dois modelos de mutação alélica: o modelo de alelos infinitos e o modelo de mutação *stepwise*.

Modelo de (mutação) alelos infinitos (IAM)

No modelo de alelos infinitos, assume-se que a cada nova mutação um novo alelo é criado (KIMURA; CROW, 1964). Devido ao grande número de variação que um gene teoricamente deve exibir, o número de possíveis novos mutantes é esperado ser muito grande (ou infinito). Este modelo aplica-se aos alelos obtidos em nível de seqüência de nucleotídeos ou aminoácidos, bem como o polimorfismo em nível de proteínas detectado por eletroforese.

Modelo de mutação *stepwise* (SMM)

O uso de marcadores de microssatélites implica considerar o modelo de mutação *stepwise*. Os locos de microssatélites são segmentos de repetições de DNA de pequeno comprimento, freqüentemente formados de um a seis nucleotídeos. Por exemplo, para um loco com repetição GT, é possível encontrar uma seqüência GTGTGTGTGTGT, em que o dinucleotídeo GT é repetido seis vezes. Por isso, os locos microssatélites são chamados de repetições (curtas) em *tandem* (do inglês, *short tandem repeat*, STR). Acredita-se que este tipo de marcador está sujeito à mudança mutacional do tipo duplicação ou deleção de unidades repetidas, sendo altamente polimórfico. Portanto, poderão existir alelos com sete, oito e nove repetições de GT. Neste modelo, os alelos representados pelo número de repetições nucleotídicas (tamanho do alelo) e se supõe que o aumento ou diminuição do tamanho do alelo (estado alélico) é dado por uma unidade, como na Figura 5.9.

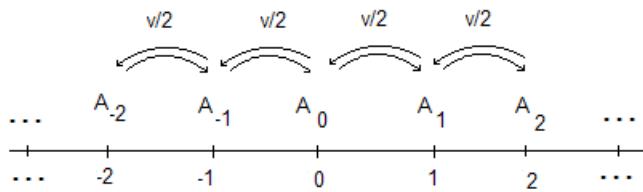


Figura 5.9 - Representação do modelo de mutação *stepwise*.

Na verdade, a situação real encontra-se entre o modelo *stepwise* e o modelo de alelos infinitos (NEI; KUMAR, 2000).

Distância genética de Goldstein

A utilização de marcadores de microssatélites em estudos de divergência genética e evolução levou Goldstein et al. (1995) a proporem uma nova medida de distância para ser utilizada com este tipo de marcador.

Os marcadores de microssatélites apresentam elevado grau de polimorfismo por loco. Cada alelo neste tipo de loco é representado por um número de repetições em tandem, que pode crescer ou decrescer por eventos de mutação que seguem, aproximadamente, o modelo de mutação *stepwise* (SMM). Este modelo assume que um alelo no estado i (um alelo com i repetições) muta para o estado $i+1$ ou $i-1$, com igual probabilidade (TAKEZAKI; NEI, 1996). Com base nessas pressuposições, Goldstein (1995) propôs a medida de distância dada pela seguinte expressão:

$$(\delta\mu_{ii})^2 = \frac{1}{L} \sum_{j=1}^L (\mu_{ij} - \mu_{i'j})^2$$

Para cálculo deste índice, são consideradas as informações:

		Alelos (Loco j)			
Indivíduo		A ₁	A ₂	...	A _k
i	Tamanho	α_{j1}	α_{j2}	...	α_{jk}
	Freqüência	p_{ij1}	p_{ij2}	...	p_{ijk}
i'	Tamanho	α_{j1}	α_{j2}	...	α_{jk}
	Freqüência	$p_{ij'1}$	$p_{ij'2}$...	$p_{ij'k}$

sendo:

$$\mu_{ij} = \sum_{k=1}^{a_j} \alpha_{jk} p_{ijk}$$

e

$$\mu_{i'j} = \sum_{k=1}^{a_j} \alpha_{jk} p_{i'jk}$$

em que μ_{ij} e $\mu_{i'j}$ são os números médios de repetições encontradas nos alelos do j-ésimo loco; e p_{ijk} e $p_{i'jk}$ são as freqüências alélicas do alelo com k repetições (tamanho do alelo) no j-ésimo loco, das populações i e i', respectivamente.

A partir da esperança de $(\delta\mu)^2$ dada por $E(\delta\mu)^2 = 2vt$, pode ser estimado o tempo (t) de divergência entre as populações por $(\delta\mu)^2 / (2v)$.

Na prática, existem alguns problemas com essa distância. Primeiramente, não se sabe a real taxa de mutação para locos microssatélites. Além disso, o coeficiente de variação de $(\delta\mu)^2$ é muito alto quando comparado a outras medidas de distância, como D_{CE} e D_{N83} (TAKEZAKI; NEI, 1996). Desse modo, um grande número de locos deve ser usado para obter uma estimativa adequada do tempo decorrido de divergência mesmo quando a taxa de mutação é conhecida. Há evidências de que o padrão atual de mutação para locos microssatélites é irregular e desvia consideravelmente do modelo stepwise de mutação. Por último, alguns locos

microssatélites são altamente polimórficos para determinadas populações e monomórficos para outras (NEI; KUMAR, 2000).

Distância ASD (Slatkin , 1995)

Esta medida de distância é relacionada a $(\delta\mu_{ii'})^2$ e representada pela diferença das médias ao quadrado (ASD), fornecida por:

$$ASD_{ii'} = \frac{1}{L} \sum_{j=1}^L \left[\sum_{k=1}^{a_j} \sum_{k'=1}^{a_j} (\alpha_{jk} - \alpha_{jk'})^2 p_{ijk} p_{i'jk'} \right]$$

Distância de Shriver et al. (1995)

É dada por:

$$D_{SW} = W_{ii'} - \frac{(W_i + W_{i'})}{2}$$

em que:

$$W_i = \frac{1}{L} \sum_{j=1}^L \left(\sum_{k>}^{a_j} \sum_{k'}^{a_j} |\alpha_{jk} - \alpha_{jk'}| p_{ijk} p_{ijk'} \right)$$

$$W_{i'} = \frac{1}{L} \sum_{j=1}^L \left(\sum_{k>}^{a_j} \sum_{k'}^{a_j} |\alpha_{jk} - \alpha_{jk'}| p_{i'jk} p_{i'jk'} \right)$$

$$W_{ii'} = \frac{1}{L} \sum_{j=1}^L \left(\sum_k^{a_j} \sum_{k'}^{a_j} |\alpha_{jk} - \alpha_{jk'}| p_{ijk} p_{i'jk'} \right)$$

Distância e identidade genética padronizada de Nei

Dentre as várias medidas de distância genética, destaca-se a distância genética padronizada, também conhecida como estatística D de Nei (1972), utilizada quando informações de freqüências alélicas estão acessíveis.

Inicialmente, Nei (1972) caracterizou a *identidade genética* (I) como a probabilidade de que um determinado alelo de um loco, amostrado aleatoriamente em duas populações distintas, fosse idêntico em relação à probabilidade de que dois alelos do mesmo loco, amostrados aleatoriamente em cada população, também fossem idênticos. Logo, a probabilidade de que dois alelos, de um dado

loco j , provenientes da população i e da população i' , sejam idênticos é $j_{ii'} = p_{ijk}p_{i'jk}$.

Partindo-se de princípio similar, a probabilidade de que dois alelos amostrados ao acaso sejam idênticos dentro de cada população i e i' é dada por $j_i = p_{ijk}$ e $j_{i'} = p_{i'jk}$, respectivamente. Desse modo, define-se a identidade genética padronizada para um loco como:

$$I_j = \frac{j_{ii'}}{\sqrt{j_i j_{i'}}},$$

Considerando vários locos, a identidade genética (I) é dada por:

$$I_{ii'} = \frac{J_{ii'}}{\sqrt{J_i J_{i'}}}$$

em que:

$$J_{ii'} = \frac{1}{L} \sum_{j=1}^L \sum_{k=1}^{a_j} p_{ijk}p_{i'jk}$$

$$J_i = \frac{1}{L} \sum_{j=1}^L \sum_{k=1}^{a_j} p_{ijk}^2$$

e

$$J_{i'} = \frac{1}{L} \sum_{j=1}^L \sum_{k=1}^{a_{i'}} p_{i'jk}^2$$

Com base nessas informações, calcula-se a distância genética D de Nei por meio da seguinte expressão:

$$D_{N72\ ii'} = -\ln (I_{ii'}) = -\ln \left(\frac{J_{ii'}}{\sqrt{J_i J_{i'}}} \right) = -\ln \frac{\sum_{j=1}^L \sum_{k=1}^{a_j} p_{ijk}p_{i'jk}}{\sqrt{\sum_{j=1}^L \sum_{k=1}^{a_j} p_{ijk}^2 \sum_{j=1}^L \sum_{k=1}^{a_{i'}} p_{i'jk}^2}}$$

Se duas populações quaisquer exibem os mesmos alelos em freqüências idênticas, $J_i = J_{i'} = J_{ii'} = 1$. Caso elas não exibam alelos comuns, tanto $J_{ii'}$ quanto a identidade genética (I) serão iguais a zero.

Há casos em que D_{N72} é não definida, pois pode ocorrer que a identidade seja nula e, portanto, a distância indeterminada. Teoricamente, é possível computar

a média aritmética dos valores de I_j , estimativa esta que se aproxima do valor estimado da identidade genética, porém com uma interpretação genética complexa (NEI, 1972). Como referência, consideram-se as informações:

População	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
1	1	0	0	0	0	0	0	0
2	0	1/7	1/7	1/7	1/7	1/7	1/7	1/7
3	0	0	1/6	1/6	1/6	1/6	1/6	1/6
4	0	0	0	0	0	0	0	1
5	1/4	1/4	1/4	1/4	0	0	0	0
6	0	0	0	0	1/4	1/4	1/4	1/4

Para este conjunto de dados, seriam obtidos para as populações 1 e 4, por exemplo, os seguintes valores:

$$J_1 = \sum_{k=1}^8 p_{1jk}^2 = 1$$

$$J_4 = \sum_{k=1}^8 p_{4jk}^2 = 1$$

$$J_{14} = \sum_{k=1}^8 p_{1jk} p_{4jk} = 0$$

portanto:

$$I_{14}=0 \quad \text{e } D_{N72\ 14} \text{ é indeterminado}$$

Também para as populações 5 e 6, tem-se:

$$J_5 = \sum_{k=1}^8 p_{5jk}^2 = \frac{1}{4}$$

$$J_6 = \sum_{k=1}^8 p_{6jk}^2 = \frac{1}{4}$$

$$J_{56} = \sum_{k=1}^8 p_{5jk} p_{6jk} = 0$$

logo:

$I_{56}=0$ e $D_{N72\ 56}$ é também indeterminado

Para os demais pares de populações, com base no exemplo em consideração, têm-se as seguintes medidas de similaridade:

Populações	1	2	3	4	5	6
1	1,0000	0,0000	0,0000	0,0000	0,5000	0,0000
2	0,0000	1,0000	0,9258	0,3780	0,5669	0,7559
3	0,0000	0,9258	1,0000	0,4082	0,4082	0,8165
4	0,0000	0,3780	0,4082	1,0000	0,0000	0,5000
5	0,5000	0,5669	0,4082	0,0000	1,0000	0,0000
6	0,0000	0,7559	0,8165	0,5000	0,0000	1,0000

Os fatores que influenciam D_{N72} são a quantidade de divergência genética entre as populações, os tamanhos das amostras e o número de locos estudados (DIAS, 1998). Para grande número de locos, a esperança de D é $E(D) = 2vt$, em que v é a taxa de mutação e t o tempo de divergência entre as populações. O valor D_{N72} assume relação linear com o tempo decorrido da divergência e à taxa de substituição gênica por loco e por geração, assumindo a taxa constante com o tempo.

A distância padronizada de Nei mede o número médio de diferenças (substituições) alélicas acumuladas por loco e pode ser considerada uma medida geral de distância genética para qualquer par de organismos. Para o cálculo da distância padronizada de Nei, também são incluídos os locos monomórficos.

A distância genética padronizada de Nei não se caracteriza como uma distância geométrica (REIF et al., 2005). Por outro lado, NEI et al. (1983) relatam que, à medida que o número de locos aumenta, D_{N72} vai se tornando gradualmente métrica, o que a torna uma medida métrica assintótica.

Essa medida foi desenvolvida com base no modelo de alelos infinitos (KIMURA; CROW, 1964), supondo que uma população ancestral em equilíbrio é dividida em várias subpopulações, as quais divergiram em virtude da deriva genética e da mutação, negligenciado o processo de seleção, sendo, portanto, apropriada para casos de processos evolucionários longos.

Para os dados apresentados na Tabela A4 (Anexos) referentes à avaliação de cinco populações quanto a dois locos co-dominantes multialélicos, têm-se os seguintes valores de identidade e de distância:

População	Loco	J_i	$J_{i'}$	$J_{ii'}$	$I = \frac{J_{ii'}}{\sqrt{J_i \cdot J_{i'}}}$	$D_{ii'} = -\ln(I)$
1 e 2	1	0,545	0,5106	0,342	0,6483	
	2	0,3838	0,5722	0,4231	0,9029	
	Média	0,4644	0,5414	0,3826	0,7629	0,2706
1 e 3	1	0,545	0,4024	0,225	0,4805	
	2	0,3838	0,3554	0,3035	0,8218	
	Média	0,4644	0,3789	0,2642	0,63	0,4621
1 e 4	1	0,545	0,6568	0,416	0,6953	
	2	0,3838	0,6058	0,4227	0,8766	
	Média	0,4644	0,6313	0,4193	0,7745	0,2556
1 e 5	1	0,545	0,7816	0,347	0,5317	
	2	0,3838	0,2796	0,1198	0,3657	
	Média	0,4644	0,5306	0,2334	0,4702	0,7546
2 e 3	1	0,5106	0,4024	0,3738	0,8247	
	2	0,5722	0,3554	0,299	0,663	
	Média	0,5414	0,3789	0,3364	0,7427	0,2974
2 e 4	1	0,5106	0,6568	0,5656	0,9767	
	2	0,5722	0,6058	0,5874	0,9977	
	Média	0,5414	0,6313	0,5765	0,9861	0,014
2 e 5	1	0,5106	0,7816	0,6086	0,9634	
	2	0,5722	0,2796	0,1066	0,2665	
	Média	0,5414	0,5306	0,3576	0,6672	0,4047
3 e 4	1	0,4024	0,6568	0,354	0,6886	
	2	0,3554	0,6058	0,303	0,653	
	Média	0,3789	0,6313	0,3285	0,6717	0,398
3 e 5	1	0,4024	0,7816	0,3768	0,6719	
	2	0,3554	0,2796	0,2058	0,6529	
	Média	0,3789	0,5306	0,2913	0,6497	0,4313
4 e 5	1	0,6568	0,7816	0,6996	0,9764	
	2	0,6058	0,2796	0,1122	0,2726	
	Média	0,6313	0,5306	0,4059	0,7013	0,3548

A partir da matriz de distância é possível estabelecer árvores (Figura 5.10a), grupos e projeções (Figura 5.10b). As técnicas de agrupamentos realçam também a similaridade entre as populações 2 e 4 e o distanciamento da população 5 em

relação às demais. O agrupamento, pelo método de Tocher, evidenciou a formação de três grupos.

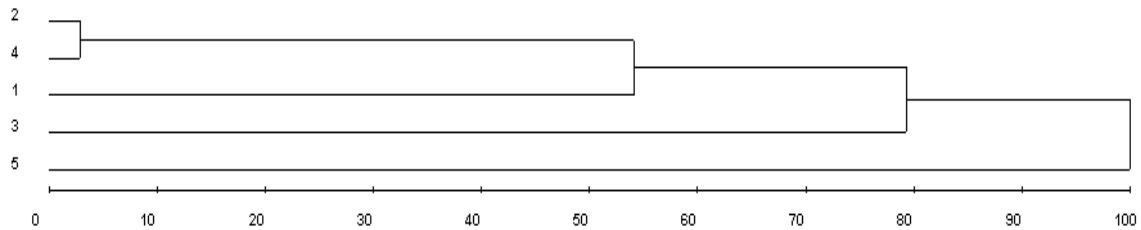


Figura 5.10a - Dendrograma estabelecido pelo método UPGMA a partir de distância genética de Nei (1972).

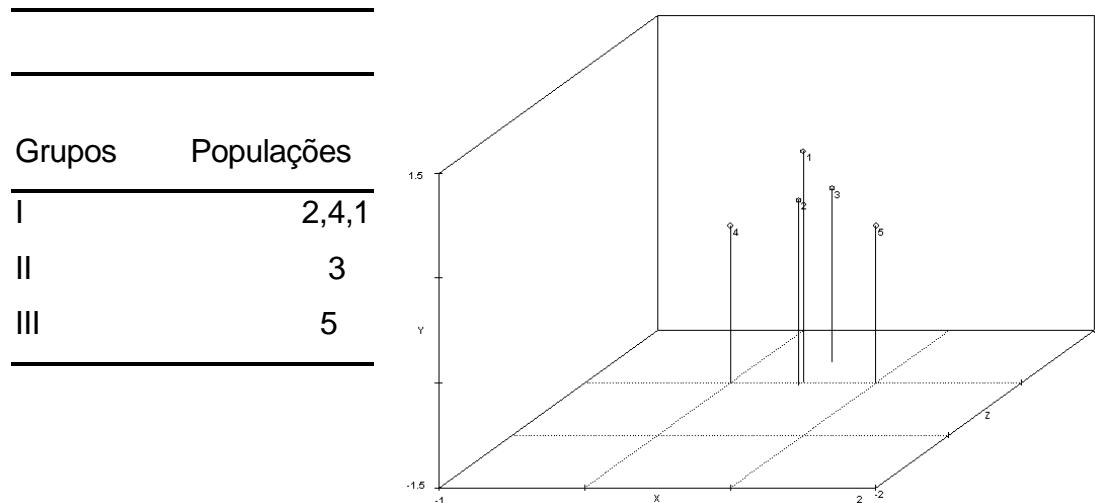


Figura 5.10b - Agrupamento de otimização, pelo método de Tocher, e projeção 3D (distorção de 15,18%, estresse de 37,74% e correlação entre distâncias originais e gráficas de 0,34) a partir de distância dada pela raiz quadrada de D_{ij} cinco populações.

Distância não-viesada de Nei

A distância genética D de Nei foi modificada pelo mesmo autor (NEI, 1978) para situações cujo número de indivíduos amostrados é considerado pequeno. Assim, apresenta-se a seguinte expressão:

$$\hat{D}_{ii'} = -\ln \left(\frac{\hat{G}_{ii'}}{\sqrt{\hat{G}_i \cdot \hat{G}_{i'}}} \right)$$

em que:

$$\hat{G}_i = \frac{2n_i J_i - 1}{2n_i - 1}$$

$$\hat{G}_{i'} = \frac{2n_{i'} J_{i'} - 1}{2n_{i'} - 1}$$

$$\hat{G}_{ii'} = J_{ii'}$$

O número de indivíduos amostrados na população i e i' é n_i e $n_{i'}$.

Distância e Identidade de Latter (1972 e 1973)

Para este autor, a identidade genética entre duas populações i e i' deve ser obtida por meio de:

$$\phi_{ii'} = \frac{(J_i + J_{i'}) - J_{ii'}}{1 - J_{ii'}}$$

e a medida de distância expressa por meio de:

$$D_{Li'i'} = -\ln(1 - \phi_{ii'})$$

Para o exemplo em consideração, tem-se:

$$\phi_{12} = \frac{(1+1)-0}{1-0} = 2$$

e

$$\phi_{34} = \frac{\left(\frac{1}{16} + \frac{1}{16}\right) - 0}{1-0} = \frac{1}{8}$$

Se $J_i + J_{i'} = 1$, então $\phi_{ii'}$ será igual a 1 e o valor de $D_{Li'i'}$ será indeterminado.

Medida de distância genética com base em informações genotípicas

Considerando três populações – $P_1(100AA \text{ e } 100aa)$, $P_2(50AA, 100Aa \text{ e } 50aa)$ e $P_3(200Aa)$, – constata-se que, pelas freqüências alélicas, estas populações não se distanciam geneticamente, uma vez que em todas elas se verifica $f(A) = f(a) = 0,5$. Entretanto, suas constituições genotípicas são extremamente diferentes. Assim, uma maneira alternativa de quantificar a dissimilaridade entre populações é por meio de estatísticas fundamentadas nas freqüências genotípicas em vez da freqüência alélica. A seguir são descritas algumas medidas de distâncias que se baseiam na informação da freqüência genotípica das populações estudadas.

Distância Genotípica de Hedrick

Hedrick (1971) também justificou a comparação entre freqüências genotípicas observadas ao invés de freqüências alélicas, admitindo que a seleção atua, principalmente, nos organismos em seu estado diplóide.

Assim, considerando apenas um loco, define-se a medida de identidade genotípica (I_H) por:

$$I_{Hii'} = \frac{\sum_{g \geq g'}^{a_i} \sum_{g' \geq g}^{a_{i'}} P_{ijgg'} P_{i'j'gg'}}{2 \left(\sum_{g \geq g'}^{a_i} \sum_{g' \geq g}^{a_j} P_{ijgg'}^2 + \sum_{g \geq g'}^{a_{i'}} \sum_{g' \geq g}^{a_j} P_{i'j'gg'}^2 \right)}$$

em que $P_{ijgg'}$ e $P_{i'j'gg'}$ são as freqüências do gg' -ésimo genótipo nas populações i e i' , respectivamente em relação ao loco j ; e a_j é o número de alelos para o loco em consideração.

O numerador expressa a probabilidade de se tomar ao acaso genótipos idênticos provenientes da população i e i' , e o denominador é a probabilidade média de extrair genótipos idênticos provenientes de uma mesma população em duas sucessivas retiradas independentes.

Generalizando para vários locos, a probabilidade média de identidade genotípica entre duas populações ($\bar{I}_{ii'}$) é dada pela média dos valores de $I_{ii'}$.

encontrados para cada loco. Para L locos, a distância genotípica ($D_{Hii'}$) é dada pelo complemento aritmético da identidade genotípica, expressa por:

$$D_{Hii'} = 1 - \bar{I}_{ii'}$$

A presença e a ausência de um alelo podem ser tão significativas do ponto de vista evolucionário quanto a diferença em termos de freqüência gênica e genotípica. Uma extensão lógica da probabilidade de identidade genotípica é a informação sobre os genótipos específicos (únicos) de uma determinada população. Quando se comparam duas populações, a probabilidade de um genótipo único, em uma delas, é calculada por:

$$U_{ii'} = \sum_{g=1}^{a_i} p_{ijgg}, \text{ na condição } p_{i'jgg} = 0$$

$\bar{U}_{ii'}$ e $\bar{U}_{i'i}$ são as probabilidades médias de um genótipo específico para a população i e i' , respectivamente.

Tomando como exemplo os dados de cinco populações, conforme apresentados na Tabela A4 (Anexos), pode ser estabelecida a seguinte relação genotípica com as suas respectivas freqüências:

Genótipo	Loco 1					Loco 2				
	P ₁	P ₂	P ₃	P ₄	P ₅	P ₁	P ₂	P ₃	P ₄	P ₅
11	0,12	0,44	0,16	0,60	0,78	0,18	0,08	0,04	0,06	
12	0,46	0,24	0,12	0,36	0,10	0,36	0,46	0,20	0,42	
13		0,22	0,40			0,14		0,18		
14		0,02			0,10					0,02
15										
22	0,42	0,02		0,04		0,16	0,46	0,08	0,52	0,02
23		0,04	0,12			0,16		0,30		0,08
24				0,02						0,12
25										0,06
33		0,02	0,20				0,20			0,14
34										0,20
35										0,14
44										0,12
45										0,08
55										0,02

Tendo em vista os valores descritos na tabela apresentada anteriormente, pode-se obter a distância entre cada par das cinco populações estudadas, conforme apresentado a seguir:

População	Loco	p_i^2	$p_{i'}^2$	$p_{ii'}$	$I_{Hi'}$	$D_{ii'}$
1 e 2	1	0,4024	0,3024	0,1716	0,4869	0,5131
	2	0,2328	0,4296	0,2536	0,7657	0,2343
	Média	0,3176	0,366	0,2126	0,6263	0,3737
1 e 3	1	0,4024	0,2544	0,0744	0,2266	0,7734
	2	0,2328	0,2104	0,1652	0,7455	0,2545
	Média	0,3176	0,2324	0,1198	0,4860	0,5140
1 e 4	1	0,4024	0,4912	0,2544	0,5694	0,4306
	2	0,2328	0,4504	0,2452	0,7178	0,2822
	Média	0,3176	0,4708	0,2498	0,6436	0,3564
1 e 5	1	0,4024	0,6288	0,1396	0,2708	0,7292
	2	0,2328	0,1256	0,016	0,0893	0,9107
	Média	0,3176	0,3772	0,0778	0,1800	0,8200
2 e 3	1	0,3024	0,2544	0,196	0,7040	0,2960
	2	0,4296	0,2104	0,132	0,4125	0,5875
	Média	0,3660	0,2324	0,164	0,5583	0,4417
2 e 4	1	0,3024	0,4912	0,3512	0,8851	0,1149
	2	0,4296	0,4504	0,4372	0,9936	0,0064
	Média	0,3660	0,4708	0,3942	0,9394	0,0606
2 e 5	1	0,3024	0,6288	0,3692	0,7930	0,2070
	2	0,4296	0,1256	0,0092	0,0331	0,9669
	Média	0,3660	0,3772	0,1892	0,4130	0,5870
3 e 4	1	0,2544	0,4912	0,1392	0,3734	0,6266
	2	0,2104	0,4504	0,128	0,3874	0,6126
	Média	0,2324	0,4708	0,1336	0,3804	0,6196
3 e 5	1	0,2544	0,6288	0,1368	0,3098	0,6902
	2	0,2104	0,1256	0,0536	0,3190	0,6810
	Média	0,2324	0,3772	0,0952	0,3144	0,6856
4 e 5	1	0,4912	0,6288	0,5049	0,9000	0,1000
	2	0,4504	0,1256	0,0104	0,0361	0,9639
	Média	0,4708	0,3772	0,2572	0,4681	0,5319

A partir da nova matriz de distância, é possível estabelecer árvores (Figura 5.11a), grupos e projeções (Figura 5.11b). As técnicas de agrupamentos realçam também a similaridade entre as populações 2 e 4 e o distanciamento da população 5 em relação às demais. O agrupamento pelo método de Tocher evidenciou a formação de dois grupos. Houve considerável melhoria na projeção 3D após transformar as medidas de dissimilaridade pela raiz quadrada dos valores obtidos.

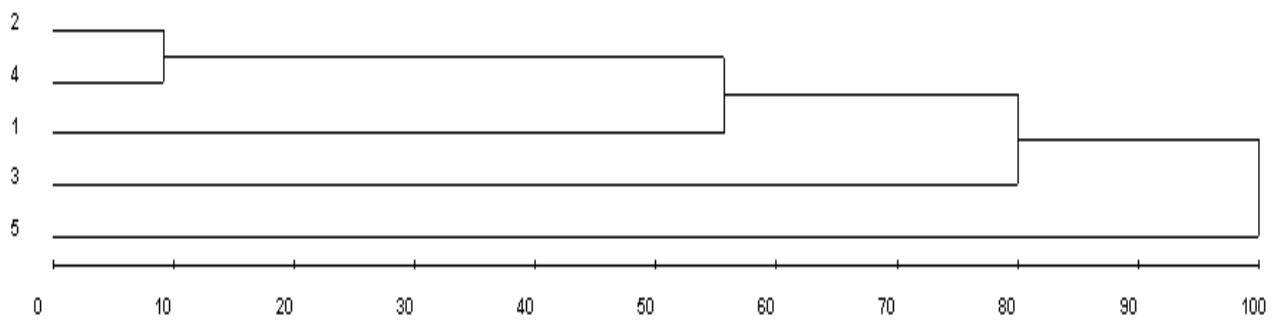


Figura 5.11a - Dendrograma estabelecido pelo método UPGMA a partir de distância genotípica de Hedrick entre cinco populações.

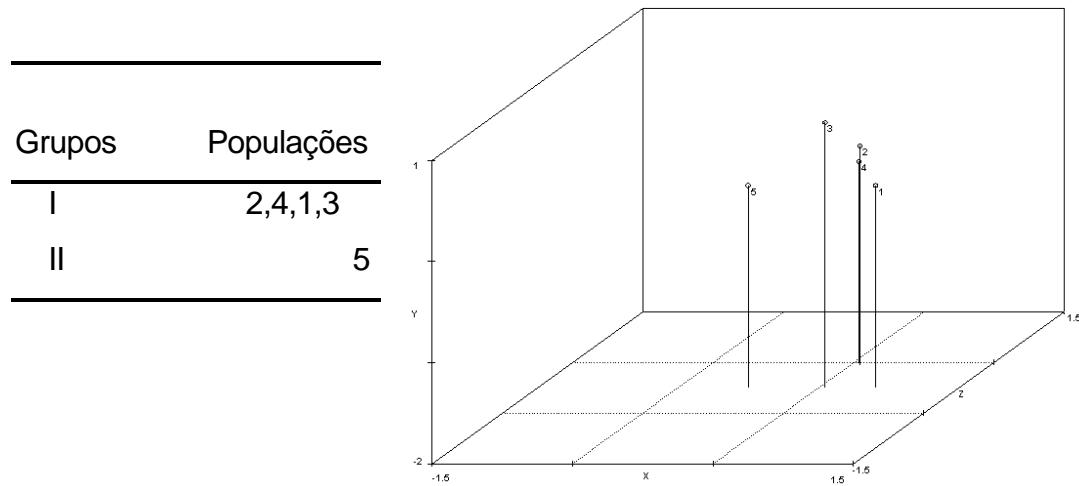


Figura 5.11b. Agrupamento de otimização, pelo método de Tocher, e projeção 3D (distorção de 2,09%, estresse de 3,75% e correlação entre distâncias originais e gráficas de 0,9982) a partir de distância dada pela raiz quadrada de D_{ij} de cinco populações.

Distância genotípica Powell

Expressa a distância entre duas populações a partir das informações sobre as freqüências genotípicas, utilizando a seguinte expressão:

$$DP_{ii'} = \frac{1}{L} \sum_{j=1}^L \left[\frac{\sum_{g \geq g'}^{a_j} |P_{ijgg'} - P_{i'jgg'}|}{2} \right]$$

5.4. Estatísticas Descritivas da Diversidade Dentro de Populações

Além das medidas de distâncias, várias outras informações sobre a população e a variabilidade manifestada nos marcadores utilizados são de grande importância. A comparação de resultados de diferentes populações permite estabelecer sua história evolutiva de convergência ou isolamento. A seguir serão descritas algumas importantes estatísticas auxiliares úteis para descrever a dissimilaridade genética entre e dentro das populações estudadas. Como ilustração, serão consideradas as informações moleculares de cinco locos (L_1, L_2, L_3, L_4 e L_5) genotipados em três populações, cujas informações são descritas a seguir:

Pop	L1	L2	L3	L4	L5	Pop	L1	L2	L3	L4	L5	Pop	L1	L2	L3	L4	L5
1	11	14	11	12	12	2	13	11	11	14	13	3	33	11	12	11	11
1	11	14	13	22	13	2	11	12	12	12	13	3	33	11	13	11	12
1	11	11	22	12	11	2	13	11	22	11	12	3	33	15	11	11	11
1	11	11	12	12	11	2	11	11	11	11	11	3	23	11	12	12	11
1	11	15	22	22	11	2	14	12	12	11	22	3	33	25	24	11	11
1	11	14	22	12	11	2	13	13	12	12	12	3	33	11	12	11	12
1	11	34	12	11	11	2	34	11	12	11	22	3	33	11	24	11	11
1	11	11	22	11	13	2	14	11	22	11	11	3	33	11	22	12	11
1	11	12	22	12	11	2	14	11	12	11	12	3	33	15	22	12	12
1	11	14	12	12	34	2	11	13	12	23	11	3	33	22	25	22	11
1	12	13	22	11	11	2	14	12	22	11	22	3	33	11	22	11	11
1	11	14	22	13	11	2	11	13	22	11	12	3	33	11	12	11	11
1	11	11	12	22	23	2	34	11	12	11	22	3	33	11	13	11	11
1	11	11	22	12	12	2	14	11	22	14	12	3	33	12	23	12	11
1	11	34	12	11	12	2	14	12	12	11	12	3	33	25	12	11	11
1	11	11	12	12	11	2	11	12	22	11	12	3	33	11	44	11	12
1	11	13	12	12	14	2	11	12	22	11	11	3	33	12	23	11	11
1	11	11	13	11	11	2	33	11	11	11	22	3	33	11	12	11	11
1	11	11	23	11	14	2	11	12	12	11	23	3	33	12	44	11	11
1	11	11	12	12	13	2	24	12	22	11	22	3	33	12	24	12	11

A partir das informações sobre os genótipos que ocorrem em cada população, são obtidas as freqüências alélicas, dadas por:

Freqüência alélica e Heterozigose - População: P₁

Loco	A1	A2	A3	A4	A5	PIC	Num. alelo	Máx
1	0,975	0,025				0,0487	2	0,975*
2	0,675	0,025	0,100	0,175	0,0250	0,5025	5	0,675
3	0,300	0,625	0,075			0,5137	3	0,625
4	0,575	0,400	0,025		0,5087		3	0,575
5	0,700	0,100	0,125	0,075	0,4787		4	0,700

Freqüência alélica e Heterozigose - População: P₂

Loco	A1	A2	A3	A4	A5	PIC	Num.alelo	Máx
1	0,575	0,025	0,175	0,225		0,5875	4	0,575
2	0,725	0,200	0,075			0,4287	3	0,725
3	0,375	0,625				0,4688	2	0,625
4	0,850	0,075	0,025	0,05		0,2687	4	0,850
5	0,425	0,500	0,075			0,5637	3	0,500

Freqüência alélica e Heterozigose - População: P₃

Loco	A1	A2	A3	A4	A5	PIC	Num.alelo	Máx
1		0,025	0,975			0,0487	2	0,975*
2	0,700	0,200			0,100	0,4600	3	0,700
3	0,250	0,450	0,100	0,175	0,025	0,6938	5	0,450
4	0,825	0,175				0,2888	2	0,825
5	0,900	0,100				0,1800	2	0,900

De posse das informações alélicas e genotípicas, em cada população, podem ser extraídas as seguintes informações adicionais para avaliação da diversidade genética:

a - Proporção de locos polimórficos (P)

É dada por:

$$P = \frac{\text{número de locos polimórficos}}{\text{número total de locos polimórficos}}$$

Três critérios são comumente utilizados para classificar um loco como polimórfico (COLE, 2003):

- i. loco exibindo polimorfismo em pelo menos um indivíduo da amostra;
- ii. loco em que o alelo mais comum tem freqüência menor que 99%; e
- iii. loco em que o alelo mais comum tem freqüência menor que 95%.

Para os dados analisados, tem-se:

População	Critério 1	Critério 2	Critério 3
P ₁	100%	100%	80%
P ₂	100%	100%	100%
P ₃	100%	100%	80%

Assim, pelo critério 1 todas as populações apresentam polimorfismo, uma vez que não há fixação de alelos em nenhum dos locos avaliados. Pelo critério 2, também todas as populações apresentam 100% de polimorfismo, visto que em nenhuma delas há um alelo com freqüência superior a 0,99. Pelo critério 3, percebe-se que as populações 1 e 3 apresentam um loco (L1) com freqüência alélica acima de 0,95; assim, em cinco locos analisados, apenas quatro são polimórficos, dando grau de polimorfismo, a estas populações, igual a 80%.

b - Número efetivo de alelos por loco polimórfico (n_e)

O número efetivo de alelos é dado por:

$$n_e = \frac{\text{númerototal de alelospolimórficos}}{\text{número de locospolimórficos}}$$

População	Número médio	Número total	Número efetivo de alelos (n_e)		
			Critério 1	Critério 2	Critério 3
P ₁	3,4	7	3,4	3,4	4,25
P ₂	3,2	6	3,2	3,2	3,2
P ₃	2,8	4	2,8	2,8	3,5

c – Proporção de alelos contidos em cada população (P_a)

É formada por:

$$P_a = \frac{\text{númerode alelosda população}}{\text{númerototal de alelosda espécie}}$$

d - Número de alelos raros (N_r)

Expressa o número de alelos com freqüênci menor que 0,05 em cada subpopulação ou amostra.

População	Pa	Nr	Nr (%)
P ₁	0,7727	4	23,5294
P ₂	0,7273	2	12,5000
P ₃	0,6364	2	14,2857

e- Número de alelos privados ou exclusivos

Refere-se à contagem dos alelos presentes em apenas uma das subpopulações amostradas. Para o exemplo considerado, tem-se:

População 1: alelo 4 do loco 2 e o alelo 4 do loco 5

População 2: alelo 4 do loco 1 e o alelo 4 do loco 4

População 3: alelo 4 do loco 3 e o alelo 5 do loco 3

f – Conteúdo médio de informação polimórfica (PIC)

É dado por:

$$PIC_{ij} = 1 - \sum_{k=1}^{a_j} p_{ijk}^2 - \sum_{k=1}^{a_j} \sum_{k'=1}^{a_j} p_{ijk}^2 p_{ijk'}^2$$

Para o exemplo em consideração, a população 3 foi a que apresentou maior valor de PIC para o loco 3.

g – Índice da Marca (MI)

É fornecido pela expressão:

MI = PIC X (proporção de bandas polimórficas) X (número de loco por unidade de ensaio.)

Por exemplo, para RFLPs, uma unidade de ensaio corresponde a uma combinação sonda-enzima; para AFLPs, a uma combinação de *primer*, e para os SSRs, a uma combinação de par de *primer*.

h – Índice de eficiência por ensaio (A_i)

É medido por:

$$A_i = \frac{n_e}{E}$$

em que $N_e = \sum n_e$ é o número total de alelos efetivos detectados; e E é o número total de ensaios realizados para identificação dos alelos.

5.5. Análise Discriminante Molecular (Não-paramétrica)

Na análise discriminante é comum as populações estarem assumindo alguma probabilidade específica de distribuição, sendo a multinormalidade a mais comum para o estudo das populações. Contudo, em outros métodos paramétricos de análise podem ser assumidas pressuposições para outros tipos de distribuição

de probabilidade. Também existem técnicas de análise discriminantes não-paramétricas que podem ser utilizadas sem pressuposições sob a distribuição.

Uma das técnicas de análise discriminante que vêm sendo utilizadas com sucesso nos estudos de genética com base em informações de marcadores moleculares é o método dos k-vizinhos mais próximos.

A técnica de análise discriminante não-paramétrica consiste em, inicialmente, estimar a medida de dissimilaridade entre cada par de indivíduos estudado, considerando um índice apropriado. É necessário também estabelecer a probabilidade *priori* inerente às várias populações avaliadas em um determinado estudo. Caso não haja informações prévias para classificação dos indivíduos, pode-se pressupor que as probabilidades sejam iguais para todas as populações analisadas. Outra opção é admitir que as probabilidades *a priori* sejam proporcionais ao tamanho de cada população.

Podem ser realizadas as seguintes análises:

Análise discriminante pelo vizinho médio

Este método de análise discriminante é baseado na alocação de um indivíduo em uma população de acordo com a classe de seus vizinhos mais próximos, seguindo critérios específicos (algoritmo). Suponha que N_i seja o número de indivíduos da i -ésima população estudada e a dissimilaridade média de um indivíduo à cada população seja:

$$\bar{D}_i = \frac{\sum_{j=1}^{N_i} d_{kj}}{N_i}$$

Sendo d_{kj} : medida de dissimilaridade do indivíduo k , a ser classificado, em relação ao j -ésimo indivíduo da i -ésima população.

Com base na média de todas as distâncias genéticas possíveis de serem estimadas entre os indivíduos, com exceção do indivíduo com ele mesmo,

independentemente da população ao qual o indivíduo pertence, define-se a população a que ele esteja mais próximo alocando-a nesta.

Análise discriminante pelos k vizinhos mais próximos

O método do k-vizinho mais próximo é baseado na alocação de um indivíduo em uma população de acordo com a classe de seus vizinhos mais próximos, seguindo um critério específico. Suponha que k seja o número máximo de acessos mais próximos que se possa obter e, consequentemente, o número que definirá a população a que o acesso pertencerá. Considere ainda que, entre esses k acessos, k_i são provenientes da população i , cuja probabilidade *a priori* é π_i . Então, a probabilidade de um acesso pertencer à i -ésima população é estimada por:

$$\hat{P}_{x_i} = \frac{\pi_i \hat{f}_i(x)}{\sum_{j=1}^g \pi_j \hat{f}_j(x)} = \frac{\pi_i (k_i / N_i)}{\sum_{j=1}^g \pi_j (k_j / N_j)}$$

em que $i = 1, 2, \dots, g$, e N_i é o número de acessos de cada população. Um indivíduo qualquer é alocado na j -ésima população se \hat{P}_x for a maior probabilidade entre as g populações avaliadas. Após a designação dos indivíduos nas populações, pode-se estimar a taxa de erro aparente ou utilizar outros métodos para avaliação da análise discriminante.

Se a probabilidade *a priori* π_i for estabelecida como sendo proporcional ao tamanho de cada população ($\pi_i = N_i / N$), então:

$$\hat{P}_{x_i} = \frac{k_i}{k}$$

Neste tipo de análise deve-se também calcular a taxa de erro aparente, que é dada pela relação entre o número de classificações erradas e o número total de classificações e o número de classificações corretas e incorretas de cada grupo.

5.6. Importância Relativa de Marcadores Moleculares

Uma questão importante na análise discriminante é saber se todos os m marcadores utilizados contêm informações úteis e se alguns deles são suficientes para a discriminação. Para a inclusão ou descarte de um marcador da lista de marcadores importantes, pode-se adotar um critério com base em um nível de significância de um teste F. Para os demais marcadores utiliza-se um critério fundamento em análise multivariadas.

De maneira geral, para identificar os marcadores mais importantes no estudo da diversidade genética são realizadas as seguintes análises estatísticas:

Método Forward

- Identificação da marcador mais importante utilizando a estatística F

Inicialmente é feita análise de variância univariada, considerando um experimento inteiramente casualizado, conforme esquema a seguir:

FV	GL	SQ	QM	F
Entre populações	g-1	SQE	QME	QME/QMD
Dentro de populações	N-g	SQD	QMD	
Total	N-1	SQT		

$$k = \frac{N - \left(\frac{1}{N} \sum_{i=1}^g N_i \right)}{g-1} \quad \text{e} \quad N = \sum_{i=1}^g N_i$$

sendo N_i o número de indivíduo dentro da i-ésima população.

Identifica-se o marcador mais importante para o estudo da diversidade genética como sendo aquele cujo valor da estatística F na análise de variância, entre as m realizadas, foi o de maior magnitude.

- Identificação da importância dos demais marcadores para o estudo da diversidade genética.

Uma vez identificado um marcador mais importante para estudo da diversidade genética, a análise prossegue, estudando-se agora a importância conjunta deste marcador com um outro do conjunto disponível. Assim, nesta primeira fase é feita a análise de variância multivariada (Manova), considerando pares de marcadores, ou seja, aquele já anteriormente identificado juntamente com um outro. Assim, são feitas $m-1$ manovas, de forma que se possa identificar o par mais importante, tomando como critério de avaliação da significância da variação entre aquele estabelecido pelo pesquisador.

Tendo-se identificado os dois marcadores mais importantes, a análise prossegue, agora com três marcadores, sendo dois os anteriormente identificados. Assim, são feitas $m-2$ manovas, de forma a identificar o marcador adicional mais importante e que promova variação entre populações significativa. O processamento prossegue até que todas os marcadores sejam incluídas na manova ou que o valor da estatística utilizada seja não-significativo.

São utilizados os seguintes testes:

- Teste de Wilks
- Teste de Pillai
- Teste de Hotelling-Lowley
- Teste de Roy

Para aplicação destes testes, considera-se:

- E: Matriz de Somas de Quadrados e Produtos do Resíduo;
- H: Matriz de Somas de Quadrados e Produtos de Tratamentos (populações);
- p: número de variáveis (marcadores) de resposta;
- q: número de graus de liberdade associado a H;
- n_e : número de graus de liberdade associado a E, ou seja, é o número de grau de liberdade do resíduo.

Define-se:

$$s = \min(p, q) = 2$$

$$m' = \frac{1}{2}(|p - q| - 1) = 2$$

$$n' = \frac{1}{2}(n_e - p - 1) = 2,5$$

Teste de Wilks

A hipótese de nulidade a ser testada, considerando g tratamentos (populações) e p marcadores, é a de que os vetores de médias das populações são iguais, isto é:

$$H_0: \underset{\sim}{\mu_1} = \underset{\sim}{\mu_2} = \dots = \underset{\sim}{\mu_g}$$

ou seja:

$$H_0: \begin{bmatrix} \mu_{1x} \\ \mu_{1y} \end{bmatrix} = \begin{bmatrix} \mu_{2x} \\ \mu_{2y} \end{bmatrix} = \dots = \begin{bmatrix} \mu_{lx} \\ \mu_{ly} \end{bmatrix}$$

O teste de Wilks é um teste de significância muito usado na análise da variância multivariada, e a estatística do teste é indicada pela letra grega Λ (lambda maiúsculo), assim definida:

$$\Lambda = \frac{\det(E)}{\det(H+E)} = \frac{|E|}{|H+E|}$$

Na presença de diferenças sistemáticas entre populações, refere-se obter $\Lambda < 1$, e tanto mais significativo quanto menor o seu valor.

Para avaliar a significância do valor de Λ obtido, pode-se usar a tabela própria para o teste de Wilks ou, o que é mais comum, transformar o valor de Λ num valor correspondente de F e usar as tabelas de F já conhecidas.

O valor de Λ obtido na tabela para o teste de Wilks é função de α , p, q e n_e .

Rejeita-se H_0 em nível de significância α se $\Lambda_{\text{Calculado}} < \Lambda_{\text{Tabelado}}$. Caso contrário, não se rejeita H_0 .

Teste de Pillai

Define-se a estatística:

$$V = \text{traço}[H(H+E)^{-1}]$$

V se relaciona com a distribuição F, aproximadamente, por:

$$\frac{2n'+s+1}{2m'+s+1} \cdot \frac{V}{s-V} \sim F(n_1, n_2)$$

em que:

$$n_1 = s(2m'+s+1) \text{ e } n_2 = s(2n'+s+1)$$

Rejeita-se H_0 em nível de significância α se $F_0 > F_{[\alpha, n_1, n_2]}$,

em que:

$$F_0 = \frac{2n'+s+1}{2m'+s+1} \cdot \frac{V}{s-V}$$

Teste de Hotelling-Lawley

Define-se a estatística:

$$U = \text{traço}(E^{-1}H)$$

U se relaciona com a distribuição F, aproximadamente, por:

$$\frac{2(sn'+1)}{s^2(2m'+s+1)} \cdot U \sim F[s(2m'+s+1), 2(sn'+1)].$$

Rejeita-se H_0 em nível de significância α se $F_0 > F[\alpha, s(2m'+s+1), 2(sn'+1)]$

em que:

$$F_0 = \frac{2(sn'+1)}{s^2(2m'+s+1)} \cdot U$$

Teste de Roy

Consiste em determinar as raízes características (autovalores) da equação:

$$|E^{-1}H - \lambda I| = 0 \quad \text{ou} \quad |H - \lambda E| = 0$$

No caso de $p=2$, a equação acima fornece duas raízes (λ_1 e λ_2), em que $\lambda_{\max} = \max(\lambda_1, \lambda_2)$.

Sendo λ_{\max} o maior autovalor de $E^{-1}H$, define-se a estatística:

$$\theta_0 = \frac{\lambda_{\max}}{1 + \lambda_{\max}}.$$

Rejeita-se H_0 em nível de significância α se $\theta_0 > \theta_{[\alpha, s, m', n']}$

Método Backward

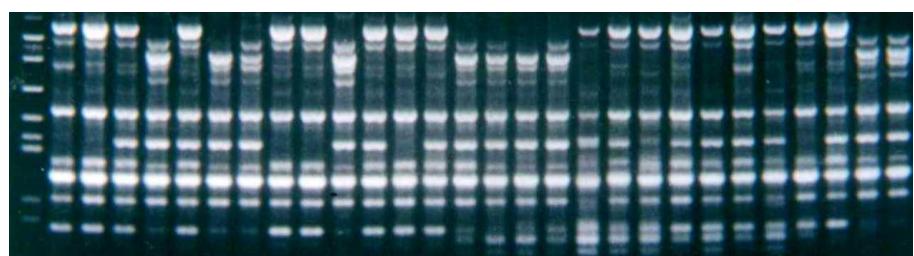
Para o procedimento *backward* é necessário calcular a estatística F para todas os marcadores considerados no modelo MANOVA, descartando em cada estágio aquele de menor poder discriminatório. Então, $\Lambda_B(k)$ é a estatística Wilk's *lambda* com base em todos os k marcadores, e $\Lambda_B(k - 1)$ é baseada em todos, exceto um, marcadores , sendo para este eleiminado avaliado a pertinência de sua eliminação. A estatística F é dada por:

$$F = \frac{(N-g-k+1)}{g-1} \cdot \left(\frac{\Lambda_B(k-1)}{\Lambda_B(k)} - 1 \right)$$

seguindo distribuição F aproximada, com $(g - 1, N - g - k + 1)$ graus de liberdade, para grandes amostras. Identifica-se o marcador que apresente F mínimo em relação às demais; consequentemente, neste estágio, será eliminado do modelo.

Capítulo 6

Diferenciação Genética Baseada em Informações Moleculares



6.1 Introdução

Diversos métodos para estimar variação genética e de estrutura populacional já foram desenvolvidos, com aplicações variadas em níveis individuais, intrapopulacional e interpopulacional. De acordo com Robinson (1998), dentre as medidas mais utilizadas em trabalhos com marcadores moleculares, destacam-se aquelas baseadas nas estatísticas H de Nei e F de Wright. Outra medida importante é a heterozigosidade.

Nesses estudos devem ser considerados os fatores responsáveis pela estruturação da variação genética em nível populacional, quais sejam a deriva genética, o fluxo gênico, a mutação e a seleção. Também devem ser relembrados os modelos de estrutura geográfica das populações, dados por:

- a) Modelo em Ilhas (WRIGHT, 1931) – Todas as populações apresentam o mesmo tamanho efetivo (N_e). Os indivíduos migram de uma subpopulação para a outra com a mesma taxa m . As distâncias entre subpopulações não são levadas em consideração.
- b) Modelo de Alpondras (ou Stepping – Stone) (MALECOT, 1950; KIMURA, 1953) – Todas as subpopulações apresentam o mesmo tamanho efetivo (N_e). A taxa de migração é constante e define a taxa de intercâmbio entre populações vizinhas apenas.
- c) Modelo de Isolamento por Distância (Wright, 1943) – A população forma um *continuum* geográfico, com fluxo gênico ocorrendo localmente entre indivíduos vizinhos. O modelo não permite estabelecer populações locais estáveis no tempo.

O modelo de ilhas é o convencionalmente empregado na análise de populações geograficamente estruturadas, baseada em freqüências gênicas.

6.2. Identidade, Heterozigosidade e Diversidade Genética - Estatística G_{ST} de Nei

A medida da diversidade genética elaborada por Nei (1973, 1975, 1978) é baseada na heterozigosidade (H) gênica ou alélica. Sua partição hierárquica permite estimar um componente de diversidade entre e outro componente de diversidade dentro das unidades experimentais (populações, subpopulações, demes etc.).

De acordo com Nei (1973, 1975), considerando uma população subdividida em g subpopulações, com p_{ijk} sendo a freqüência do k -ésimo alelo, do j -ésimo loco, na i -ésima população podem-se obter, para estas subpopulações, medidas de identidade gênica (igual ao complemento aritmético da diversidade gênica), da heterozigosidade e da diversidade genética. Segundo Robinson (1998), as expressões para cálculo da diversidade apresentadas por Nei podem ser aplicadas a diferentes níveis da classificação hierárquica, como populações dentro de uma região geográfica, subpopulações dentro de populações, indivíduos dentro de subpopulações etc., utilizando-se, em cada caso, as freqüências alélicas correspondentes da categoria em questão.

a) Medidas de identidade genética

A identidade genética na i -ésima subpopulação, para um particular loco, é dada por:

$$J_i = \sum_{k=1}^{a_j} p_{ijk}^2$$

A identidade genética entre a i -ésima e a i' -ésima subpopulações é fornecida por:

$$J_{ii'} = \sum_{k=1}^{a_j} p_{ijk} p_{i'jk}$$

A partir desses valores, podem ser estabelecidas as seguintes estatísticas:

Identidade dentro de subpopulações

A identidade genética total dentro das subpopulações, para um loco, é dada por:

$$J_{IS} = \sum_{i=1}^g J_i = \sum_{i=1}^g \sum_{k=1}^{a_i} p_{ijk}^2$$

e o valor médio é denotado por:

$$\bar{J}_{IS} = \frac{J_{IS}}{g}$$

Identidade entre subpopulações

A identidade genética total entre todos os pares de subpopulações, para um loco, é fornecida por:

$$J_{ST} = \sum_{i=1}^g \sum_{i' < i}^g J_{ii'} = \frac{1}{2} \sum_{i=1}^g \sum_{i'=i}^g J_{ii'}$$

e o valor médio é denotado por:

$$\bar{J}_{ST} = \frac{2J_{ST}}{g(g-1)}$$

Identidade total das subpopulações

É dada por:

$$J_{IT} = \sum_i J_i + \sum_{i \neq 1}^g \sum_{i'}^g J_{ii'} = J_{IS} + 2J_{ST}$$

e o valor médio é fornecido por:

$$\bar{J}_{IT} = \frac{J_{IT}}{g^2} = \frac{1}{g} \bar{J}_{IS} + \frac{g-1}{g} \bar{J}_{ST}$$

Como ilustração, consideram-se três subpopulações e um loco que apresenta três alelos. Os valores da identidade gênica entre e dentro das populações é apresentado a seguir:

Subpopulação	$f(A)$	$f(a)$	J_i	Pares de subpopulações	$J_{ii'}$
1	0,8	0,2	0,68	1 e 2	0,50
2	0,5	0,5	0,50	1 e 3	0,44
3	0,4	0,6	0,52	2 e 3	0,50
Total			1,70		1,44

De maneira geral, podem-se representar os valores de identidade entre e dentro das subpopulações por meio da operação matricial:

$$FF' = \begin{bmatrix} 0,8 & 0,2 \\ 0,5 & 0,5 \\ 0,4 & 0,6 \end{bmatrix} \begin{bmatrix} 0,8 & 0,5 & 0,4 \\ 0,2 & 0,5 & 0,6 \end{bmatrix} = \begin{bmatrix} 0,68 & 0,50 & 0,44 \\ 0,50 & 0,50 & 0,50 \\ 0,44 & 0,50 & 0,52 \end{bmatrix} = \begin{bmatrix} J_1 & J_{12} & J_{13} \\ J_2 & sim & J_{23} \\ J_3 & & \end{bmatrix}$$

em que F : matriz de dimensão $g \times a_j$ cujos elementos são as freqüências dos alelos, para um dado loco, obtidas em cada subpopulação.

Assim:

$J_{IS} = 1,70$ é o traço da matriz FF' .

$J_{ST} = 1,44$ é a soma dos elementos acima da diagonal.

$J_{IT} = 4,58$ é a soma de todos os elementos da matriz FF' .

b) Medida de Heterozigosidade

Nei (1978) apresenta as expressões para o cálculo da diversidade (comumente denominada de heterozigosidade esperada) de uma população. Para um único loco, de uma população de acasalamento ao acaso, a heterozigosidade é definida como:

$$H_i = 1 - \left(\sum_{k=1}^{a_j} p_{ijk}^2 \right)$$

em que p_{ijk} é freqüência do k -ésimo alelo, para o j -ésimo loco que apresenta a_j alelos.

Para múltiplos locos, calcula-se a heterozigosidade média de uma população i (H), que é a média aritmética dos valores de h sobre todos os locos, ou seja:

$$\bar{H}_i = \frac{1}{L} \sum_{j=1}^L H_{i(j)}$$

em que $H_{i(j)}$ (ou simplesmente H_i) é a estimativa de heterozigosidade no j -ésimo loco e L é o número de locos amostrados.

Em uma população em equilíbrio de Hardy-Weinberg, a estimativa h é igual à proporção esperada de heterozigotos. Contudo, nos organismos e, ou, nas situações em que a proporção de heterozigotos não possa ser definida conforme as suposições de equilíbrio, a medida não deve ser chamada de heterozigosidade, devendo ser compreendida apenas como uma medida de variabilidade genética e chamada de índice de heterogeneidade ou índice de diversidade genética (TORGGLER et al., 1995).

Para uma subpopulação, tem-se:

$$H_i = 1 - J_i$$

$$H_{ii'} = 1 - J_{ii'}$$

A partir desses valores, podem ser estabelecidas as seguintes estatísticas:

Heterozigose dentro de subpopulações

A heterozigose genética total dentro das subpopulações, para um loco, é dada por:

$$H_{IS} = \sum_{i=1}^g H_i = g - J_{IS}$$

e o valor médio é denotado por:

$$\bar{H}_{IS} = \frac{H_{IS}}{g} = 1 - \bar{J}_{IS}$$

Heterozigose entre subpopulações

A heterozigose genética total entre todos os pares de subpopulações, para um loco, é dada por:

$$H_{ST} = \sum_{i=1}^g \sum_{i' < i}^g H_{ii'} = \frac{g(g-1)}{2} - J_{ST}$$

e o valor médio é denotado por:

$$\bar{H}_{ST} = \frac{2H_{ST}}{g(g-1)} = 1 - \bar{J}_{ST}$$

Heterozigose total das subpopulações

A heterozigose genética total é fornecida por:

$$H_{IT} = \sum_{i=1}^g H_i + \sum_{i=1}^g \sum_{i' \neq i}^g H_{ii'} = H_{IS} + 2H_{ST}$$

Define-se a heterozigose média das subpopulações pela expressão:

$$\bar{H}_{IT} = \frac{H_{IT}}{g^2} = \frac{1}{g} \bar{H}_{IS} + \frac{g-1}{g} \bar{H}_{ST}$$

ou

$$\bar{H}_{IT} = \frac{1}{g}(1 - \bar{J}_{IS}) + \frac{g-1}{g}(1 - \bar{J}_{ST}) = 1 - \bar{J}_{IT}$$

c) Medida de diversidade

A diversidade genética ($D_{ii'}$) entre a i -ésima e a i' -ésima subpopulações é estimada por:

$$D_{ii'} = H_{ii'} - \frac{H_i + H_{i'}}{2}$$

ou

$$D_{ii'} = \frac{J_i + J_{i'}}{2} - J_{ii'}$$

Como trata-se de diversidade, tem-se:

$$D_{ii} = 0 \quad \text{e} \quad D_{ii'} = D_{i'i}$$

Diversidade dentro de subpopulações

Como $D_{ii}=0$, tem-se:

$$D_{IS} = \sum_i D_{ii} = 0$$

Diversidade entre subpopulações

$$D_{ST} = \sum_{i>i'} \sum_{i'<i''} D_{ii'} = \sum_{i>i'} \sum_{i'} \left(\frac{J_i + J_{i'}}{2} - J_{ii'} \right) = \frac{g-1}{2} \sum_{i=1}^g J_i - \sum_{i=1}^g \sum_{i>i'} J_{ii'}$$

$$D_{ST} = \frac{g-1}{2} J_{IS} - J_{ST}$$

e

$$\bar{D}_{ST} = \frac{2}{g(g-1)} D_{ST} = \bar{J}_{IS} - \bar{J}_{ST}$$

Diversidade total das subpopulações

A diversidade total, computadas todas as populações, é denotada por D_{IT} e calculada por meio de:

$$D_{IT} = \sum_i D_{ii} + \sum_{i \neq i'} \sum_{i'<i''} D_{ii'} = D_{IS} + 2D_{ST} = 2D_{ST}$$

A diversidade média, envolvendo todas as subpopulações, é dada por:

$$\bar{D}_{IT} = \frac{D_{IT}}{g^2} = \frac{g-1}{g} \bar{J}_{IS} - \frac{g-1}{g} \bar{J}_{ST} = \bar{J}_{IS} - \left(\frac{1}{g} \bar{J}_{IS} - \frac{g-1}{g} \bar{J}_{ST} \right)$$

$$\bar{D}_{IT} = \frac{D_{IT}}{g^2} = \bar{J}_{IS} - \bar{J}_{IT}$$

de outra forma:

$$\bar{D}_{IT} = \frac{2D_{ST}}{g^2} = \frac{g-1}{g} \bar{D}_{ST}$$

Deve-se lembrar que \bar{J}_{IS} é identidade genética média dentro de subpopulações, \bar{J}_{IT} é a identidade média geral e \bar{D}_{IT} é diversidade genética total das subpopulações, incluindo comparações de subpopulações com elas próprias (apesar de $D_{ii'}$ ser nulo).

De maneira resumida, podem-se representar os valores encontrados para as estatísticas descritas da seguinte maneira:

FV	Identidade	Heterozigosidade	Diversidade
Dentro	\bar{J}_{IS}	$\bar{H}_{IS} = 1 - \bar{J}_{IS}$	$D_{IS} = 0$
Entre	\bar{J}_{ST}	$\bar{H}_{IS} = 1 - \bar{J}_{ST}$	$\bar{D}_{ST} = \bar{J}_{IS} - \bar{J}_{ST}$ $\bar{D}_{ST} = \bar{H}_{ST} - \bar{H}_{IS}$
Total	$\bar{J}_{IT} = \frac{1}{g} \bar{J}_{IS} + \frac{g-1}{g} \bar{J}_{ST}$	$\bar{H}_{IT} = \frac{1}{g} (1 - \bar{J}_{IS}) + \frac{g-1}{g} (1 - \bar{J}_{ST})$ $\bar{H}_{IT} = 1 - \bar{J}_{IT}$	$\bar{D}_{IT} = \bar{J}_{IS} - \bar{J}_{IT}$ $\bar{D}_{IT} = \bar{H}_{IT} - \bar{H}_{IS}$

Decomposição das heterozigosidade total \bar{H}_{IT} e média entre populações \bar{H}_{ST}

A heterozigosidade média da população \bar{H}_{IT} pode ser decomposta considerando que:

$$\bar{D}_{IT} = \bar{J}_{IS} - \bar{J}_{IT} = (1 - \bar{H}_{IS}) - (1 - \bar{H}_{IT})$$

Logo:

$$\bar{D}_{IT} = \bar{H}_{IT} - \bar{H}_{IS}$$

de forma que:

$$\bar{H}_{IT} = \bar{D}_{IT} + \bar{H}_{IS}$$

Assim, a heterozigosidade da população total (conjunto das subpopulações analisadas) pode ser decomposta em diversidade genética (\bar{D}_{IT}) e heterozigosidade genética média dentro de subpopulações (\bar{H}_{IS}).

Sabe-se que:

$$\bar{H}_{IT} = \frac{1}{g} \bar{H}_{IS} + \frac{g-1}{g} \bar{H}_{ST} \quad \text{e} \quad \bar{D}_{IT} = \frac{g-1}{g} \bar{D}_{ST}$$

logo:

$$\bar{H}_{IT} = \frac{1}{g} \bar{H}_{IS} + \frac{g-1}{g} \bar{H}_{ST} = \frac{g-1}{g} \bar{D}_{ST} + \bar{H}_{IS}$$

então:

$$\bar{H}_{ST} = \bar{D}_{ST} + \bar{H}_{IS}$$

Diferenciação entre subpopulações

A magnitude relativa da diferenciação entre subpopulações pode ser medida por:

$$G = \frac{100\bar{D}_{ST}}{\bar{H}_{ST}} \quad \text{ou} \quad G_{ST} = \frac{100\bar{D}_{IT}}{\bar{H}_{IT}}$$

Para $g=2$, tem-se

$$\bar{H}_{IT} = \frac{1}{g} \bar{H}_{IS} + \frac{g-1}{g} \bar{H}_{ST} = \frac{\bar{H}_{IS} + \bar{H}_{ST}}{2}$$

$$\bar{D}_{IT} = \frac{g-1}{g} \bar{D}_{ST} = \frac{\bar{D}_{ST}}{2}$$

então

$$G_{ST} = \frac{\bar{D}_{ST}}{\bar{H}_{IS} + \bar{H}_{ST}}$$

A medida G_{ST} é denominada de coeficiente de diversidade relativa entre grupos. Ela varia entre 0 e 1 e expressa a proporção da diversidade total explicada por diferenças entre os grupos. Matematicamente, esta medida equivale à estatística F_{ST} de Wright.

Nei (1978) propõe o uso de estimativas não-viesadas de H_i e H quando o número (n) de indivíduos amostrados for pequeno. Nesta situação deve ser utilizado:

$$\hat{h} = \frac{2n}{2n-1} \left[1 - \left(\sum_{k=1}^{a_j} p_{ijk}^2 \right) \right] = \frac{2n}{2n-1} h$$

e

$$\hat{H} = \sum_{j=1}^L \hat{h}_j / L$$

sendo n o número de indivíduos amostrados por loco, podendo ser variável de loco para loco e de população para população. Entretanto, se ocorrerem desvios significativos das proporções esperadas para o equilíbrio de Hardy-Weinberg, como em populações de autofecundação, as estimativas de \hat{h} devem ser obtidas pela expressão:

$$\hat{h} = \frac{n}{n-1} \left[1 - \left(\sum_{k=1}^{a_j} p_{ijk}^2 \right) \right]$$

As expressões para o cálculo da heterozigosidade total (H_T) e do coeficiente de diversidade relativa entre grupos (G_{ST}), corrigidas para o número de grupos avaliados, são:

$$\bar{H}'_{IT} = \bar{H}_{IS} + \bar{D}'_{IT}$$

e

$$G'_{ST} = \frac{100 \bar{D}'_{IT}}{\bar{H}'_{IT}},$$

sendo:

$$\bar{H}_{IS} = \bar{n} \left[1 - \sum_i \bar{p}_i^2 - H_o / 2\bar{n} \right] / (\bar{n} - 1)$$

e

$$D'_{IT} = [n/(np - 1)] D_{IT} \text{ (NEI, 1987, citado por MOTA, 2003)}$$

Aplicação

Para os procedimentos descritos a seguir, serão tomadas como ilustração as informações relativas a oito populações, cujas ocorrência genotípica e freqüência alélica são descritas a seguir:

População	Genótipos			p _{i11}	1- p _{i11}
	A ₁ A ₁	A ₁ A ₂	A ₂ A ₂		
1	14	3	3	0,7750	0,2250
2	15	2	3	0,8000	0,2000
3	13	0	0	1,0000	0,0000
4	23	5	2	0,8500	0,1500
5	23	3	4	0,8167	0,1833
6	29	3	1	0,9242	0,0758
7	5	0	0	1,0000	0,0000
8	0	1	0	0,5000	0,5000

Como medidas auxiliares e descritivas dessas populações, podem ser consideradas as estimativas da heterozigosidade e do PIC, dadas a seguir:

População	H	PIC
1	0,3487	0,2879
2	0,3200	0,2688
3	0,0000	0,0000
4	0,2550	0,2225
5	0,2994	0,2546
6	0,1400	0,1302
7	0,0000	0,0000
8	0,5000	0,375

A identidade genética para o par de populações 1 e 2 pode ser calculada por meio de:

$$J_1 = 0,775^2 + 0,225^2 = 0,65125$$

$$J_2 = 0,8^2 + 0,2^2 = 0,68$$

$$J_{12} = (0,775)(0,8) + (0,225)(0,2) = 0,665$$

Assim, tem-se:

Identidade dentro

$$J_{IS} = J_1 + J_2 = 1,33125$$

$$\bar{J}_{IS} = 0,665625$$

Identidade entre

$$J_{ST} = J_{12} = 0,665$$

$$\bar{J}_{ST} = 0,665$$

Identidade total

$$J_{IT} = J_{IS} + 2J_{ST} = 2,66125$$

$$\bar{J}_{IT} = \frac{1}{2}0,665625 + \frac{1}{2}0,665 = 0,6653125$$

A heterozigosidade para o par de populações 1 e 2 pode ser calculada por meio de:

$$H_1 = 1 - J_1 = 1 - 0,65125 = 0,34875$$

$$H_2 = 1 - J_2 = 1 - 0,68 = 0,32$$

$$H_{12} = 1 - J_{12} = 1 - 0,665 = 0,335$$

Dessa forma:

Heterozigosidade dentro

$$H_{IS} = H_1 + H_2 = 0,34875 + 0,32 = 0,66875$$

$$H_{IS} = g - J_{IS} = 2 - 1,33125 = 0,66875$$

$$\bar{H}_{IS} = 0,334375$$

Heterozigosidade entre

$$H_{ST} = 0,335$$

$$H_{ST} = 1 - J_{ST} = 1 - 0,665 = 0,335$$

$$\bar{H}_{ST} = 0,335$$

Heterozigosidade total

$$H_{IT} = H_{IS} + 2H_{ST} = 0,66875 + 2(0,335) = 1,33875$$

$$\bar{H}_{IT} = 1 - \bar{J}_{IT} = 0,3346975$$

A diversidade genética entre as populações 1 e 2 pode, portanto, ser calculada por meio de:

$$D_{11} = 0$$

$$D_{22} = 0$$

$$D_{12} = H_{12} - \frac{H_1 + H_2}{2} = 0,335 - \frac{0,34875 + 0,32}{2} = 0,000625$$

ou

$$D_{12} = \frac{J_1 + J_2}{2} - J_{12} = \frac{0,65125 + 0,68}{2} - 0,665 = 0,000625$$

Diversidade dentro

$$D_{IS} = 0$$

Diversidade entre

$$D_{ST} = 0,000625$$

$$\bar{D}_{ST} = 0,000625$$

Diversidade total

$$D_{IT} = 2(0,000625) = 0,00125$$

$$\bar{D}_{IT} = \frac{0,00125}{4} = 0,0003125$$

A magnitude relativa da diferenciação entre as populações 1 e 2 é medida por:

$$G_{ST} = \frac{100\bar{D}_{IT}}{\bar{H}_{IT}} = \frac{100(0,0003125)}{0,3346875} = 0,093371\%$$

$$\text{e } G = \frac{100\bar{D}_{ST}}{\bar{H}_{ST}} = \frac{100(0,000625)}{0,335} = 0,1865\%$$

Os índices de diversidade para cada par de populações são apresentados a seguir:

Populações	\bar{H}_{IT}	\bar{H}_{IS}	\bar{D}_{IT}	G_{ST}
1 2	0,3347	0,3344	0 ,0003	0,0934
1 3	0,2355	0,2114	0,0242	10,2632
1 4	0,2952	0,2925	0,0027	0,9146
1 5	0,3200	0,3192	0,0008	0,2604
1 6	0,2293	0,2188	0,0105	4,5653
1 7	0,2952	0,2790	0,0162	5,4878
1 8	0,3628	0,3560	0,0069	1,8906
2 3	0,2130	0,1939	0,0191	8,9655
2 4	0,2822	0,2810	0,0012	0,4252
2 5	0,3078	0,3077	0,0001	0,0433
2 6	0,2152	0,2079	0,0073	3,3707
2 7	0,2688	0,2560	0,0128	4,7619
2 8	0,3367	0,3286	0,0082	2,4243
3 4	0,1874	0,1779	0,0095	5,0649
3 5	0,2231	0,2089	0,0142	6,3555
3 6	0,1028	0,1005	0,0023	2,2640
3 7	0,0000	0,0000	0,0000	0,0000
3 8	0,0689	0,0357	0,0332	48,1481
4 5	0,2778	0,2772	0,0006	0,2000
4 6	0,1975	0,1948	0,0027	1,3921
4 7	0,2241	0,2186	0,0055	2,4590
4 8	0,2706	0,2629	0,0076	2,8269
5 6	0,2217	0,2159	0,0058	2,6038
5 7	0,2649	0,2567	0,0082	3,1073
5 8	0,3122	0,3059	0,0063	2,0055
6 7	0,1229	0,1216	0,0013	1,0670
6 8	0,1609	0,1506	0,0103	6,3865
	0,1528	0,0833	0,0694	45,4545

6.3. Diferenciação baseada na freqüência gênica ou genotípica

Uma possibilidade de avaliar o grau de diferenciação entre populações (ou subpopulações) é por meio da comparação da diferença da quantidade de indivíduos com diferentes genótipos ou da quantidade de diferentes alelos que cada população possui. Considerando um conjunto fixo de populações, diferentes populações de uma mesma espécie poderão ser comparadas simplesmente pelos valores das freqüências alélicas, quando as condições de equilíbrio de Hardy-Weinberg puderem ser adotadas; caso contrário, recomenda-se utilizar as próprias freqüências genotípicas. Assim, são possíveis os seguintes testes:

a) Estatística F_{ST}

Outra maneira de comparar e quantificar as diferenças entre populações, a partir de suas freqüências gênicas, é por meio da estatística F_{ST} . Assim, para um conjunto de g populações e um determinado alelo A, em que as freqüências alélicas sejam representadas por p_i ($i = 1, 2, \dots, g$), é obtido:

$$F_{ST} = \frac{\frac{1}{g-1} \sum_{i=1}^g (p_i - \bar{p})^2}{\bar{p}(1-\bar{p})} = \frac{S^2}{\bar{p}(1-\bar{p})}$$

em que:

$\bar{p} = \frac{1}{g} \sum_{i=1}^g p_i$: média da freqüência de A considerando todas as populações; e

S^2 : variância da freqüência gênica.

Se as populações têm tamanho desigual, então se recomenda o uso da expressão:

$$F_{ST} = \frac{\frac{1}{n(g-1)} \sum_{i=1}^g n_i (p_i - \bar{p})^2}{\bar{p}(1-\bar{p})}$$

sendo:

$$\bar{p} = \frac{\sum_{i=1}^g n_i p_i}{\sum_{i=1}^g n_i} \quad e \quad \bar{n} = \frac{\sum_{i=1}^g n_i}{g}$$

em que n_i é a quantidade de alelos na i -ésima população, ou seja, $n_i = 2N_i$, sendo N_i a quantidade de indivíduos mensurados.

Para o exemplo em consideração, tomando por base as populações 1 e 2, pode-se obter:

$$p_1 = 31/40 = 0,775$$

$$p_2 = 32/40 = 0,800$$

$$\bar{p} = \frac{\sum_{i=1}^g n_i p_i}{\sum_{i=1}^g n_i} = \frac{31(0,775) + 32(0,800)}{80} = 0,7875$$

ou $\bar{p} = \sum_{i=1}^g w_i p_i = 0,7875$, sendo $w_1 = w_2 = \frac{1}{2}$ (w_i é a medida da proporcionalidade do número de indivíduos de cada população)

$$\sum_{i=1}^g w_i p_i^2 = \frac{1}{2}(0,775)^2 + \frac{1}{2}(0,800)^2 = 0,6203125$$

$$S^2 = \sum_{i=1}^g w_i p_i^2 - \left(\sum_{i=1}^g w_i p_i \right)^2 = 0,6203125 - (0,7875)^2 = 1,5625 \cdot 10^{-4}$$

logo:

$$F_{ST} = \frac{1,5625 \cdot 10^{-4}}{0,7875(1 - 0,7875)} = 9,337068 \cdot 10^{-4} = 0,09337068\%$$

Deve ser lembrado que o valor de F_{ST} obtido equivale à medida G_{ST} proposta por Nei (173), dada por:

$$G_{ST} = \frac{100 \bar{D}_{IT}}{\bar{H}_{IT}} = \frac{100(0,0003125)}{0,3346875} = 0,093371\%$$

O valor de F_{ST} aumenta à medida que a diversidade, medida pela diferença da freqüência alélica, também aumenta. Entretanto, é difícil estabelecer um valor de F_{ST} que possa ser considerado significativamente alto. No caso de um conjunto fixo de populações, é possível relacionar o valor de F_{ST} com a estatística qui-quadrado obtida de uma tabela de contingência. Assim, pode ser considerado um alelo A, com freqüência p_i , e seu(s) alelo(s) alternativo(s) \bar{A} , avaliado em uma população com quantidade total de alelos igual a n_i . O valor de qui-quadrado que compara a quantidade de alelos em todas as populações é dado por:

População	Observado			Esperado		
	A	\bar{A}	Total	A	\bar{A}	Total
1	$n_1 p_1$	$n_1(1-p_1)$	n_1	$n_1 \bar{p}$	$n_1(1-\bar{p})$	n_1
2	$n_2 p_2$	$n_2(1-p_2)$	n_2	$n_2 \bar{p}$	$n_2(1-\bar{p})$	n_2
...	
g	$n_g p_g$	$n_g(1-p_g)$	n_g	$n_g \bar{p}$	$n_g(1-\bar{p})$	n_g
Total	$\sum_{i=1}^g n_i p_i$	$\sum_{i=1}^g n_i(1-p_i)$	$n = \sum_{i=1}^g n_i$	$\sum_{i=1}^g n_i \bar{p}$	$\sum_{i=1}^g n_i(1-\bar{p})$	$n = \sum_{i=1}^g n_i$

A partir das informações da tabela anterior, é obtida a estimativa da estatística qui-quadrado, conforme apresentado a seguir:

$$\chi^2 = \sum_{i=1}^g \frac{(n_i p_i - n_i \bar{p})^2}{n_i \bar{p}} + \sum_{i=1}^g \frac{[n_i(1-p_i) - n_i(1-\bar{p})]^2}{n_i(1-\bar{p})} = \frac{1}{\bar{p}(1-\bar{p})} \sum_{i=1}^g n_i (p_i - \bar{p})^2$$

Como:

$S^2 = \frac{1}{ng} \sum_{i=1}^g n_i (p_i - \bar{p})^2$ (algumas vezes utiliza-se no denominador $g-1$ ao invés de g)
 logo:

$$\chi^2 = \bar{n} g F_{ST}$$

Para o exemplo em consideração, tem-se:

$$g = 2$$

$$\bar{n} = \frac{1}{g} \sum_{i=1}^g n_i = 40$$

$$F_{ST} = 9,337068 \cdot 10^{-4}$$

portanto:

$$\chi^2 = (2)40(9,337068 \cdot 10^{-4}) = 0,074696544$$

b) Teste de igualdade das quantidades genotípicas ou alélicas

Pode-se avaliar se as populações apresentam a mesma quantidade genotípica ou alélica por meio de teste de qui-quadrado aplicando-se uma tabela de contingência. Existindo a alelos em um determinado loco, haverá $a(a+1)/2$ diferentes genótipos numa população, que poderão ser arranjados em uma tabela de contingência $[a(a+1)/2] \times g$ (sendo g o número de populações). Esses valores poderão ser comparados pela estatística qui-quadrado associado a $[a(a+1)/2 - 1] \times (g - 1)$ graus de liberdade.

Como ilustração, pode-se considerar o teste de igualdade genotípica para as populações 1 e 2. Assim, tem-se a tabela:

População	Observado			Total	Esperado			Total
	A ₁ A ₁	A ₁ A ₂	A ₂ A ₂		A ₁ A ₁	A ₁ A ₂	A ₂ A ₂	
1	O ₁₁	O ₂₁	O ₃₁	T ₁	E ₁₁	E ₂₁	E ₃₁	T ₁
2	O ₁₂	O ₂₂	O ₃₂	T ₂	E ₁₂	E ₂₂	E ₃₂	T ₂
Total	G ₁	G ₂	G ₃	T	G ₁	G ₂	G ₃	T

Os valores esperados do i -ésimo genótipo na j -ésima população (E_{ij}) são obtidos por meio de:

$$E_{ij} = \frac{T_j G_i}{T}$$

em que:

T_j : total observado de genótipos na j -ésima população;

G_i : total observado do i -ésimo genótipo em todas as populações; e

T : total geral de genótipos.

O valor de qui-quadrado é obtido por meio de:

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^g \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

que está associado a $[a(a+1)/2 - 1] \times (p - 1)$ graus de liberdade.

Para as populações 1 e 2, cujos dados são apresentados no exemplo anterior, pode-se construir a tabela:

População	Observado			Total	Esperado			Total
	A_1A_1	A_1A_2	A_2A_2		A_1A_1	A_1A_2	A_2A_2	
1	14	3	3	20	14,5	2,5	3,0	20
2	15	2	3	20	14,5	2,5	3,0	20
Total	29	5	6	40	29	5	6	40

Podem-se obter os seguintes valores:

$$\chi^2 = \frac{(14 - 14,5)^2}{14,5} + \dots + \frac{(3 - 3)^2}{3} = 0,2345$$

$$GL = [a(a+1)/2 - 1] \times (g - 1) = 2$$

$$P = 0,8894$$

Com base nos resultados, não se rejeita a hipótese de que não há diferença entre as populações quanto às suas quantidades genotípicas.

O uso de classes genotípicas pode apresentar problemas de ordem prática, em razão do elevado número de classe e da possibilidade de que, em algumas delas, a quantidade esperada seja relativamente baixa, comprometendo a estatística qui-quadrado.

Uma outra alternativa para o estudo da diversidade entre populações se fundamenta na comparação de seus conteúdos alélicos. Para tal, é suficiente e apropriado comparar as quantidades alélicas das populações por meio de uma

tabela de contingência de dimensão $a \times g$, a partir da qual o teste de qui-quadrado é realizado com $(a - 1)(g - 1)$ graus de liberdade.

Para o exemplo em consideração, tomando apenas as informações das populações 1 e 2, poderia ser realizado o seguinte teste:

População	Observado			Esperado		
	A ₁	A ₂	Total	A ₁	A ₂	Total
1	31	9	40	31,5	8,5	40
2	32	8	40	31,5	8,5	40
Total	63	17	80	63	17	80

Assim, podem ser obtidos os seguintes valores:

$$\chi^2 = \frac{(31 - 31,5)^2}{31,5} + \dots + \frac{(8 - 8,5)^2}{8,5} = 0,0747$$

$$GL = (a - 1)(g - 1) = 1$$

$$P = 0,7846$$

Com base nos resultados, não se rejeita a hipótese de que não há diferença entre as populações quanto às suas quantidades alélicas, havendo, portanto, indicativo de que as duas populações conservam estrutura genética similar, não sendo sujeita a fatores que poderiam promover as suas diferenciações.

Os testes apresentados são apenas de interpretação qualitativa. Se houver diferenças entre as populações, seria interessante quantificar a magnitude dessas diferenças por meio de estatísticas apropriadas.

6.4. Diferenciação baseada na análise de variância de uma variável indicadora

Quando se avaliam g populações com N_i indivíduos dentro de cada população, é possível avaliar a existência de diferenças entre elas por meio de uma análise de variância convencional, em que se consideram, como fontes de variação, os efeitos entre e dentro de população. Assim, para uma dada variável, pode-se adotar o modelo estatístico:

$$x_{ij} = \mu + P_i + \varepsilon_{ij}$$

em que:

μ : efeito da média geral;

P_i : efeito da i -ésima população, que possibilita avaliar a variação entre populações; e

ε_{ij} : efeito do erro aleatório, que possibilita avaliar a variação dentro das populações.

A questão que surge é o tipo de variável a ser considerado na análise de variância. Weir (1996) recomenda analisar os valores obtidos por variáveis indicadoras fornecida por:

$$Z_{ijk} = 1, \text{ se o alelo é } A$$

$$Z_{ijk} = 0, \text{ se o alelo não é } A$$

A análise poderá ser feita a partir da média dos valores de cada loco, de forma que se tenha:

$$x_{ij} = \frac{1}{2} \sum_{k=1}^2 Z_{ijk}$$

Na prática, significa apenas adotar os valores:

$$x_{ij} = 1, \text{ se o genótipo é } AA$$

$$x_{ij} = \frac{1}{2}, \text{ se o genótipo é } AA_k$$

$$x_{ij} = 0, \text{ se o genótipo é } A_kA_k$$

Apesar das restrições deste modelo quanto às pressuposições de análise de variância, deve-se ter em mente que uma possível estatística F possibilitaria avaliar

a existência de diferenças entre médias das populações. A média dos valores de x_{ij} representa a freqüência do alelo **A** na i -ésima população, ou seja:

$$p_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij} \quad \text{e} \quad \bar{f} = \frac{\sum_{i=1}^g N_i p_i}{\sum_{i=1}^g N_i}$$

Portanto, a análise de variância permite apontar a existência de diferença entre as freqüências do alelo A nas populações. O quadro da análise de variância é apresentado a seguir:

FV	GL	SQ	QM	E(QM)
Entre populações	$g-1$	SQE	QME	$\frac{1}{g-1} \sum_{i=1}^g k_{1i} p_i (1-p_i) + \frac{1}{g-1} \sum_{i=1}^g N_i (p_i - \bar{p})^2$
Dentro de populações	$\sum_{i=1}^g N_i - g$	SQD	QMD	$\frac{1}{\sum_{i=1}^g N_i - g} \sum_{i=1}^g k_{2i} p_i (1-p_i)$

$$N = \sum_{i=1}^g N_i$$

sendo:

$$k_{1i} = 1 - \frac{N_i}{\sum_{i=1}^g N_i} \quad \text{e} \quad k_{2i} = N_i - 1$$

Se $N_1 = N_2 = \dots = N_g = \eta$, tem-se:

$$k_{1i} = \frac{g-1}{g}, \text{ para todo } i$$

$$k_{2i} = \eta - 1, \text{ para todo } i$$

$$E(QME) = \frac{1}{g} \sum_{i=1}^g p_i (1-p_i) + \frac{\eta}{g-1} \sum_{i=1}^g (p_i - \bar{p})^2$$

$$E(QMD) = \frac{1}{g} \sum_{i=1}^g p_i (1-p_i)$$

e a estatística $F = QME/QMD$ testa a hipótese de que as populações não diferem quanto às suas freqüências gênicas.

As somas de quadrados são obtidas pelas expressões:

$$SQE = \sum_{i=1}^g \frac{x_{i\cdot}^2}{N_i} - \frac{x_{..}^2}{N}$$

$$SQD = \sum_{i=1}^g \sum_{j=1}^{N_i} x_{ij}^2 - \sum_{i=1}^g \frac{x_{i\cdot}^2}{N_i}$$

Sabendo que:

$$x_{i\cdot} = N_i p_i \quad \text{e} \quad x_{..} = N \bar{p}_g$$

então:

$$SQE = \sum_{i=1}^g N_i (p_i - \bar{p}_g)^2$$

$$SQD = \sum_{i=1}^g N_i p_i (1 - p_i)$$

Para o exemplo em consideração, a análise de variância seria realizada a partir dos seguintes valores de x_{ij} :

Pop	Indivíduos																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	½	½	½	0	0	0
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	½	½	0	0	0	0

Para este exemplo, tem-se:

$$x_{..} = \sum_{i=1}^g \sum_{j=1}^{N_i} x_{ij} = 31,5$$

$$x_{1\cdot} = 15,5 \quad N_1 = 20$$

$$x_{2\cdot} = 16,0 \quad N_2 = 20$$

$$C = \frac{x_{..}^2}{N} = \frac{31,5^2}{40} = 24,8065$$

$$\sum_{i=1}^g \sum_{j=1}^{N_i} x_{ij}^2 = 30,25$$

$$SQE = \sum_{i=1}^g \frac{x_{i..}^2}{N_i} - \frac{x_{...}^2}{N} = \frac{15,5^2}{20} + \frac{16,0^2}{20} - C = 24,8125 - C = 6,25 \cdot 10^{-3}$$

$$SQD = \sum_{i=1}^g \sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^g \frac{x_{i..}^2}{N_i} = 30,25 - 24,8125 = 5,4375$$

Assim, tem-se:

FV	GL	SQ	QM	F
Entre populações	1	$6,25 \cdot 10^{-3}$	$6,25 \cdot 10^{-3}$	0,04367 ns
Dentro de populações	38	5,4375	0,1430	

Pelos resultados da análise de variância, é possível concluir que as populações não diferem quanto às suas freqüências alélicas.

6.5. Diferenciação por meio da estatística F de Wright

6.5.1. Abordagem fundamentada em índices de fixação

Introdução

Em estudo de diferenciação genética de populações Sewall Wright (1951, 1978) introduziu um método para a partição do índice de fixação de uma população subdividida (F_{IT}) em componentes devido ao acasalamento não casual dentro de populações locais (F_{IS}) e à subdivisão da população (F_{ST}). Segundo Robinson (1998), Wright descreveu três coeficientes de fixação aplicáveis a uma população com um nível hierárquico de subdivisão:

F_{IT} (ou F) – mede o desvio das freqüências genotípicas da população em relação ao equilíbrio de Hardy-Weinberg. Esses desvios resultam de

cruzamentos não ao acaso dentro da população (incluindo a endogamia em todos os níveis). F também é uma medida de correlação entre duas unidades gaméticas que formam um zigoto na população (ou do conjunto de populações amostradas).

F_{ST} (ou θ) – é o coeficiente de ancestria, representando a probabilidade de que dois indivíduos, pertencentes a subpopulações diferentes, possuam um alelo idêntico por descendência. Assim, θ é uma medida da correlação de gametas entre as subpopulações.

F_{IS} (ou f) - mede a fixação em nível de indivíduos, ou seja, mede a probabilidade de que os dois alelos de um loco presentes no mesmo indivíduo sejam idênticos por descendência. Assim, f é também uma medida de correlação de gametas devido à endogamia dentro das subpopulações.

Relação entre índices de fixação

A partir desses índices, Wright também estabeleceu a seguinte relação:

$$(1-F_{IT}) = (1 - F_{IS})(1 - F_{ST}).$$

Os índices de fixação são úteis para se entender a estrutura genética e de melhoramento das subpopulações bem como o padrão de seleção associado com o polimorfismo alélico.

Estas estatísticas F, que expressam índices de fixação, foram redefinidas por Cockerham (1973) e relacionadas com o coeficiente de endogamia e de co-ancestralidade, como descrito a seguir:

$$F_{IS} = \frac{F_{IT} - F_{ST}}{1 - F_{ST}} = \frac{F - \theta}{1 - \theta}$$

$$F_{ST} = \frac{F_{ST} - F_{IS}}{1 - F_{IS}} = \frac{\theta - f}{1 - f}$$

$$F_{IT} = \frac{F_{IT} - F_{IS}}{1 - F_{IS}} = \frac{F - f}{1 - f}$$

sendo:

$$F = F_{IT},$$

$$f = F_{IS}$$

e

$$\theta = F_{ST}$$

em que os indexadores I,S eT representam indivíduos, subpopulações e população total, respectivamente.

Pressuposições de análise da decomposição dos índices de fixação

Na decomposição do índice de fixação são consideradas algumas pressuposições básicas, evidenciadas a seguir:

- i. Assume-se que a população e as subpopulações sob investigação são derivadas de um ancestral comum ao mesmo tempo.
- ii. Todas as subpopulações são igualmente relacionadas uma com as outras ou não há migração entre elas.
- iii. As populações são consideradas amostras ao acaso de um conjunto infinito de subpopulações igualmente relacionadas.

Estimação dos índices de fixação

a. Por meio dos coeficientes de heterozigosidade

Nei (1977) mostra que as estatísticas F de Wright podem ser definidas como razões entre estatísticas H (definidas por NEI, 1973, 1975), em vez de correlações entre unidades gaméticas. Essas definições, segundo Nei (1977), são independentes do número de alelos envolvidos e da atuação da seleção natural. Os valores da heterozigosidade observada, esperada e total são obtidos a partir das seguintes informações:

População	A ₁	A ₂	...	A _a	Observado	Esperado
1	p _{1j1}	p _{1j2}	...	p _{1ja}	H _{o1}	H _{e1} = 1 - $\sum_{k=1}^a p_{1jk}^2$
2	p _{2j1}	p _{2j2}	...	p _{2ja}	H _{o2}	H _{e2} = 1 - $\sum_{k=1}^a p_{2jk}^2$
...
G	p _{gj1}	p _{gj2}	...	p _{gja}	H _{og}	H _{eg} = 1 - $\sum_{k=1}^a p_{gjk}^2$
Média	$\bar{p}_{.j1}$	$\bar{p}_{.j2}$...	$\bar{p}_{.jg}$		H _T = 1 - $\sum_{k=1}^a \bar{p}_{.jk}^2$

Neste caso, tem-se:

$$F_{IS} = \frac{H_S - H_0}{H_S}$$

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

$$F_{IT} = \frac{H_T - H_0}{H_T}$$

em que H_T corresponde à diversidade genética (ou heterozigosidade) na população total, dada por:

$$H_T = 1 - \sum_{k=1}^{a_j} \bar{p}_{.jk}^2 \quad (\text{para o loco } j)$$

$$\text{e} \quad \bar{p}_{.jk} = \sum_{i=1}^g w_i p_{ijk}$$

sendo p_{ijk} a freqüência do alelo k, do loco j, na subpopulação i; e w_i o tamanho relativo da i-ésima população. Quando se comparam populações com o mesmo

número de indivíduos, ou se desconhece a proporcionalidade de indivíduos entre as populações, considera-se $w_i = 1/g$ para todo i .

H_S : corresponde à diversidade genética (ou heterozigosidade) esperada entre subpopulações, dada por:

$$H_S = 1 - \sum_{i=1}^g w_i \left(\sum_{k=1}^{a_j} p_{ijk}^2 \right)$$

H_o : corresponde à diversidade genética (ou heterozigosidade) observada dentro de subpopulações.

$$H_o = \sum_{i=1}^g w_i P_{ijkk'} = \sum_{i=1}^g w_i H_{oi}$$

em que:

$P_{ijkk'}$ é a freqüência do heterozigoto $A_k A_{k'}$ em relação ao loco j na população i .

Com dois alelos possíveis por loco e um número infinito de subpopulações, Wright mostrou que:

$$F_{ST} = \frac{S^2}{\bar{p}(1-\bar{p})}$$

em que:

\bar{p} é a freqüência média de um dos alelos na população, dada por:

$$\bar{p} = \sum_{k=1}^{a_j} w_i p_{ijk} , \text{ para um dado loco } j$$

S^2 é a variância na freqüência de um dos alelos entre as demes.

Para obter estimadores não viesados das heterozigosidades, H_S e H_T devem ser corrigidos para erros de amostragem. Assumindo que:

- N_i é o número de indivíduos diplóides amostrados ao acaso, na i -ésima subpopulação; e

$-p_{ijk} e P_{ijkk'}$ são as freqüências do alelo A_k e genótipo $A_k A_k$, respectivamente, na amostra e

$$- w_i = 1/g$$

tem-se:

$$\tilde{H}_0 = H_0 = 1 - \frac{\sum_{i=1}^g P_{ijkk'}}{g}$$

$$\tilde{H}_s = \frac{\tilde{N}}{\tilde{N}-1} \left[1 - \sum_{j=1}^L \sum_{k=1}^{a_j} p_{.jk}^2 - \frac{\tilde{H}_0}{2\tilde{N}} \right]$$

em que:

$$p_{.jk}^2 = \frac{\sum_i p_{ijk}^2}{g}$$

\tilde{N} = média harmônica de N_i

$$\frac{g}{\tilde{N}} = \frac{1}{N_1} + \frac{1}{N_2} + \dots + \frac{1}{N_g}$$

$$\tilde{H}_T = 1 - \sum_{j=1}^L \sum_{k=1}^{a_j} \bar{p}_{.jk}^2 + \frac{\tilde{H}_s}{\tilde{N}g} - \frac{\tilde{H}_0}{2\tilde{N}g}$$

em que:

$$\bar{p}_{.jk} = \frac{\sum_i p_{ijk}}{g}$$

Se $\tilde{N} > 30$, seu efeito nas estimativas de \tilde{H}_s e \hat{h}_T será negligenciável; logo, a correção para tamanho da amostra é necessária apenas quando \tilde{N} for pequeno.

No modelo de estrutura populacional de infinitas ilhas, Wright mostrou também que:

$$F_{ST} \approx \frac{1}{1 + 4Nm}$$

em que N é o tamanho da população e m , a taxa de migração.

Se F_{ST} for obtido desta maneira, essa relação permite estimar o fluxo gênico entre subpopulações por meio da expressão:

$$Nm = \frac{1}{4} \left[\frac{1}{F_{ST}} - 1 \right].$$

Para um número arbitrário de alelos por loco, Slatkin e Barton (1989) mostraram que F_{ST} pode ser definido em termos de probabilidades de identidade por descendência:

$$F_{ST} = \frac{f_0 - \bar{f}}{1 - \bar{f}}$$

em que:

f_0 : probabilidade de identidade por descendência de dois alelos escolhidos ao acaso de uma única deme; e

\bar{f} : probabilidade de identidade de dois alelos escolhidos ao acaso da população inteira. Esse definição está implícita na definição de G_{ST} de Nei (1973).

Como ilustração, serão novamente considerados os dados genotípicos de duas populações das oito originalmente estudadas. Esses dados são reproduzidos a seguir:

Populações	Genótipos			p _{i11}	1 - p _{i11}
	A ₁ A ₁	A ₁ A ₂	A ₂ A ₂		
1	14	3	3	0,7750	0,2250
2	15	2	3	0,8000	0,2000

Informações auxiliares sobre essas duas populações são dadas a seguir:

Populações	Heterozigotos (obs)	p = f(A)	Heterozigotos (esp)
1	3/20 = 0,15	0,775	0,3487
2	2/20 = 0,10	0,800	0,3200

Assim, podem ser obtidas as seguintes estimativas:

$$\bar{p} = (0,775 + 0,800) / 2 = 0,7875$$

$$H_T = 1 - (0,7875^2 + 0,2125^2) = 0,3347$$

$$H_S = \frac{0,3487 + 0,3200}{2} = 0,3344$$

$$H_0 = \frac{0,15 + 0,10}{2} = 0,125$$

Portanto, os índices de fixação terão os seguintes valores:

$$F_{IS} = \frac{0,3344 - 0,125}{0,3344} = 0,6261$$

$$F_{ST} = \frac{0,3347 - 0,3344}{0,3347} = 0,0009$$

$$F_{IT} = \frac{0,3347 - 0,125}{0,3347} = 0,6265$$

Uma alternativa para se obter F_{ST} é por meio da variação da freqüência alélica entre as populações. Assim, tem-se:

$$\sum_i w_i p_i^2 = 0,6203125$$

$$\left(\sum_i w_i p_i \right)^2 = 0,62015625$$

$$S^2 = \sum_i w_i p_i^2 - \left(\sum_i w_i p_i \right)^2 = 1,5625 \cdot 10^{-4}$$

$$F_{ST}^* = \frac{1,5625 \cdot 10^{-4}}{(0,7875)(0,2125)} = 0,0009$$

Considerando agora as informações de todas as oito populações originalmente disponíveis, tem-se:

$$\bar{p} = 0,8332$$

São observados também os seguintes valores da heterozigose observada e esperada:

População	f(A)	w _i	Heterozigose (obs)	Heterozigose (esp)
1	0,7750	1/8	3/20=0,15	1-0,775 ² -0,225 ² =0,3487
2	0,8000	1/8	2/20=0,10	0,3200
3	1,0000	1/8	0/13=0,0	0,0000
4	0,8500	1/8	5/30=0,1667	0,2550
5	0,8167	1/8	3/30=0,10	0,2994
6	0,9242	1/8	3/33=0,0909	0,1400
7	1,0000	1/8	0/5=0,0	0,0000
8	0,5000	1/8	1/1=1,0	0,5000
Média	0,8332		0,20095	0,2329

Por meio dos valores descritos anteriormente, estima-se:

$$H_T = 1 - (0,8332^2 + 0,1668^2) = 0,2779$$

$$H_S = 0,2329$$

$$H_0 = 0,20095$$

logo, são obtidos os valores dos índices de fixação:

$$F_{IS} = \frac{H_S - H_0}{H_S} = \frac{0,2329 - 0,2009}{0,2329} = 0,1374$$

$$F_{ST} = \frac{H_T - H_S}{H_T} = \frac{0,2779 - 0,2329}{0,2779} = 0,1619$$

$$F_{IT} = \frac{H_T - H_0}{H_T} = \frac{0,2779 - 0,2009}{0,2779} = 0,2771$$

A variância da freqüência gênica é dada por:

$$S^2 = 0,0225$$

Assim, pode-se, alternativamente, calcular F_{ST} por meio de:

$$F_{ST} = \frac{\sigma^2}{\bar{p}(1-\bar{p})} = \frac{0,0225}{(0,8332)(0,1668)} = 0,1619$$

Os valores corrigidos para os tamanhos da amostra são obtidos considerando o número médio de indivíduos, estimado por meio de:

$$\frac{g}{\tilde{n}} = \frac{1}{n_1} + \dots + \frac{1}{n_8}$$

Ou

$$\frac{8}{\tilde{n}} = \frac{1}{20} + \dots + \frac{1}{1} \text{ logo } \tilde{n} = 5,43$$

Os valores obtidos, com e sem correção, são apresentados a seguir:

Método	F_{IS}	F_{ST}	F_{IT}	Nm
Direto	0,1374	0,1619	0,2771	1,29
Corrigido p/ \tilde{n}	0,2166	0,0866	0,2846	2,64

Em outra ilustração, é considerado que uma população original, constituída por 30AA, 60Aa e 30aa, é subdividida em três subpopulações de diferentes maneiras. Os seguintes valores de índices de fixação podem ser obtidos:

Caso 1:

População	AA	Aa	aa	p	$h(\text{obs})$	$H(\text{esp})$	Índices
P_1	10	20	10	0,5	0,5	0,5	$F_{IS}=0,0$
P_2	10	20	10	0,5	0,5	0,5	$F_{ST}=0,0$
P_3	10	20	10	0,5	0,5	0,5	$F_{IT}=0,0$
		$\bar{p}=0,5$		$H_O=0,5$		$H_S=0,5$	$H_T=0,5$

Caso 2:

População	AA	Aa	aa	p	h(obs)	H(esp)	Índices
P ₁	30	0	0	1	0	0	F _{IS} =1
P ₂	0	60	0	0,5	1	0,5	F _{ST} =0,667
P ₃	0	0	30	0	0	0	F _{IT} =0,333
		$\bar{p} = 0,5$		$H_O=0,333$		$H_S=0,167$	$H_T = 0,5$

Caso 3:

População	AA	Aa	aa	p	h(obs)	H(esp)	Índices
P ₁	10	10	5	0,6	0,4	0,48	F _{IS} =0,020
P ₂	10	20	15	0,444	0,444	0,4938	F _{ST} =0,016
P ₃	10	30	10	0,5	0,6	0,5	F _{IT} =0,036
		\bar{p} =0,5148		$H_O=0,4815$		$H_S=0,4913$	$H_T = 0,4996$

Neste exemplo, é possível verificar as situações em que a diferenciação entre as subpopulações é mais pronunciada. No segundo caso o valor de F_{ST} foi elevado, atingindo estimativa de 0,667.

6.5.2. Abordagem fundamentada na análise de variância

Outro método de obter F_{ST}, apresentado por Slatkin e Barton (1989), é o estimador θ de Weir e Cockerham (1984). Embora seja de aplicação direta, o método é algebraicamente complicado. Considerando apenas o caso em que igual número de indivíduos é amostrado de cada localidade, e assumindo que há união ao acaso de gametas nos grupos de acasalamento (isto é, não há endogamia

dentro das subpopulações, a não ser aquela devido à subdivisão da população), o estimador de F_{ST} é θ .

O valor de θ pode ser estimado por meio da análise de variância realizada a partir do seguinte modelo:

$$X_{ijk} = \mu + P_i + I/P_{ij} + G_{ijk}$$

em que:

X_{ijk} : variável que identifica a presença de determinado alelo A_k no genótipo do j -ésimo indivíduo da i -ésima população;

μ : freqüência média do alelo A_k nas populações estudadas;

P_i : efeito da i -ésima população ($i=1,2,\dots,g$) com $P_i \sim (0, \sigma_p^2)$;

I/P_{ij} : efeito do j -ésimo indivíduo dentro da i -ésima população ($j = 1,2,\dots,N_i$) com $I/P_{ij} \sim (0, \sigma_i^2)$; e

G_{ijk} : efeito da presença ou ausência do k -ésimo alelo ($k = 1,2$) com $G_{ijk} \sim (0, \sigma_g^2)$.

O esquema de análise de variância é apresentado a seguir:

FV	GL	SQ	QM	E(QM)
Entre populações	$g-1$	SQP	QMP	$\sigma_g^2 + 2\sigma_i^2 + 2k\sigma_p^2$
Indivíduos/Populações	$\sum_{i=1}^g N_i - g$	SQI	QMI	$\sigma_g^2 + 2\sigma_i^2$
Alelos/Indivíduos	$\sum_{i=1}^g N_i$	SQG	QMG	σ_g^2
Total		$2N-1$		
$k = \frac{1}{g-1} \left(\sum_{i=1}^g N_i - \frac{\sum_{i=1}^g N_i^2}{\sum_{i=1}^g N_i} \right)$		$N = \sum_{i=1}^g N_i$		

As somas de quadrados são obtidas por meio de:

$$C = \frac{X^2}{2N}$$

$$SQP = \sum \frac{x_{i..}^2}{2N_i} - C$$

$$SQI = \frac{1}{2} \sum_{i=1}^g \sum_{j=1}^{N_i} x_{ij..}^2 - \sum \frac{x_{i..}^2}{2N_i}$$

$$SQG = \sum_{i=1}^g \sum_{j=1}^{N_i} \sum_{k=1}^2 x_{ijk}^2 - \frac{1}{2} \sum_{i=1}^g \sum_{j=1}^{N_i} x_{ij..}^2$$

Os estimadores dos componentes de variância são dados por:

$$\hat{\sigma}_g^2 = QMG$$

$$\hat{\sigma}_i^2 = \frac{QMI - QMG}{2}$$

$$\hat{\sigma}_p^2 = \frac{QMP - QMI}{2k}$$

Para entendimento do significado dos componentes de variância, algumas propriedades interessantes sobre a variável x_{ijk} devem ser apresentadas, ou seja:

a) $E(x_{ijk})$

É dada por:

$$E(x_{ijk}) = 1P(x_{ijk} = 1) + 0P(x_{ijk} = 0) = 1p_k + 0(1-p_k) = p_k$$

b) $E(x_{ijk}^2) = p_k$, pois $x_{ijk}^2 = x_{ijk}$

c) $E(x_{ijk}, x_{ijk'})$

É fornecida por:

$$E(x_{ijk}, x_{ijk'}) = \frac{N_{kk'}(1x1) + N_{kk'}(1x0) + N_{kk'}(0x0)}{N} = \frac{N_{kk'}}{N} = P_{kk'}$$

$N_{kk'}$ é o número de genótipos kk' na i -ésima população.

d) $E(x_{ijk}, x_{ij'k'})$

Numa população i em que os indivíduos não são parentados, é dada por:

$$E(x_{ijk}, x_{ij'k'}) = E(x_{ijk})E(x_{ij'k'}) = p_k^2$$

Em algumas situações, deve-se imaginar que diferentes indivíduos podem não ser independentemente distribuídos. Nesse caso, seria conveniente admitir que:

$$E(x_{ijk}, x_{ij'k'}) = P_{kk}$$

Sendo P_{kk} a probabilidade de encontrar mesmo alelo k em dois indivíduos tomados ao acaso da população.

e) $V(x_{ijk})$

É dada por:

$$V(x_{ijk}) = E(x_{ijk}^2) - [E(x_{ijk})]^2 = p_k - p_k^2 = p_k(1-p_k)$$

f) $Cov(x_{ijk}, x_{ij'k'})$

É dada por:

$$Cov(x_{ijk}, x_{ij'k'}) = P_{kk} - p_k^2$$

g) $Cov(x_{ijk}, x_{ij'k'})$

É fornecida por:

$$Cov(x_{ijk}, x_{ij'k'}) = E(x_{ijk}x_{ij'k'}) - E(x_{ijk})E(x_{ij'k'}) = p_k^2 - p_k p_k = 0, \text{ admitindo } x_{ijk} \text{ e } x_{ij'k'} \text{ independentes.}$$

ou,

$$Cov(x_{ijk}, x_{ij'k'}) = E(x_{ijk}x_{ij'k'}) - E(x_{ijk})E(x_{ij'k'}) = P_{kk} - p_k p_k = P_{kk} - p_k^2$$

Se $P_{kk} = p_k^2$, então a covariância é nula. Entretanto, é possível admitir que:

$$P_{kk} = p_k^2 + p_k(1-p_k)\theta$$

e

$$P_{kk} = p_k^2 + p_k(1-p_k)\theta$$

de forma que se tenha:

$$Cov(x_{ijk}, x_{ij'k'}) = p_k(1-p_k)\theta$$

sendo θ o coeficiente de co-ancestralidade.

Assim, tem-se:

$$E(QMP) = p(1-p)[(1-F)+2(F-\theta)+2k\theta]$$

$$E(QMI) = p(1-p)[(1-F)+2(F-\theta)]$$

$$E(QMG) = p(1-p)(1-F)$$

logo, podem ser obtidos:

$$\hat{F} = \frac{\hat{\sigma}_p^2 + \hat{\sigma}_i^2}{\hat{\sigma}_p^2 + \hat{\sigma}_i^2 + \hat{\sigma}_g^2}$$

$$\hat{\theta} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_i^2 + \hat{\sigma}_g^2}$$

$$\hat{f} = \frac{\hat{F} - \hat{\theta}}{1 - \hat{\theta}}$$

Slatkin e Barton (1989) verificaram, com dados de simulação, que F_{ST} pode ser igualmente bem estimado pelos métodos de Nei (1973) e Weir e Cockerham (1984) sob condições moderadas de fluxo gênico, de amostras relativamente grandes, com igual número de observações. Por outro lado, sob condições de elevado fluxo gênico, os dois métodos produziram estimativas viesadas.

Neigel (2002) afirma que várias definições de F_{ST} podem ser encontradas na literatura, porém em muitos casos elas utilizam princípios e parâmetros que são incompatíveis com a definição original de Wright, baseada no coeficiente de endogamia. O autor também discute a importância e pertinência de se utilizar F_{ST} como uma medida indireta do fluxo gênico (aplicação proposta desde Wright) e conclui que, embora existam métodos mais poderosos, F_{ST} permanece como medida útil da estimativa do fluxo gênico médio.

Como ilustração, serão novamente considerados os dados genotípicos de duas populações das oito originalmente estudadas. Esses dados são reproduzidos a seguir:

Populações	Genótipos				p_{i11}	$1 - p_{i11}$
	A_1A_1	A_1A_2	A_2A_2	Total		
1	14	3	3	20	0,7750	0,2250
2	15	2	3	20	0,8000	0,2000

O resultado da análise de variância é:

FV	GL	SQ	QM	E(QM)
Entre populações	1	0,0125	0,0125	$\sigma_g^2 + 2\sigma_i^2 + 2k\sigma_p^2$
Indivíduos/Populações	38	10,8750	0,2862	$\sigma_g^2 + 2\sigma_i^2$
Alelos/Indivíduos	40	2,5000	0,0625	σ_g^2
Total	79			

$$k = 20$$

Os componentes de variância são dados por

$$\hat{\sigma}_g^2 = QMG = 0,0625$$

$$\hat{\sigma}_i^2 = \frac{QMI - QMG}{2} = 0,1118$$

$$\hat{\sigma}_p^2 = \frac{QMP - QMI}{2k} = -0,0068$$

e os índices de fixação estimados por meio de:

$$\hat{F} = \frac{\hat{\sigma}_p^2 + \hat{\sigma}_i^2}{\hat{\sigma}_p^2 + \hat{\sigma}_i^2 + \hat{\sigma}_g^2} = 0,6268$$

$$\hat{\theta} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_i^2 + \hat{\sigma}_g^2} = -0,0408$$

$$\hat{f} = \frac{\hat{F} - \hat{\theta}}{1 - \hat{\theta}} = 0,6415$$

Verifica-se que os resultados são bem próximos aos obtidos pela metodologia anteriormente descrita.

Erros Associados às Estimativas da Diversidade Entre Populações

Muitos trabalhos não apresentam os erros associados às estimativas de parâmetros populacionais fornecidas pelas estatísticas H e F . Segundo Carlini-Garcia et al. (2001), essa situação resulta da pouca informação sobre a natureza das distribuições empíricas das variáveis, bem como da dificuldade de obter expressões explícitas das estimativas das variâncias de F (ou F_{IT}), f (ou F_{IS}) e θ (ou F_{ST}), pois estes estimadores são razões entre variáveis aleatórias com distribuição desconhecida.

Nei e Chakravarti (1977) apresentaram expressões que permitem calcular as variâncias aproximadas das estatísticas F_{ST} e G_{ST} , mas os próprios autores mostraram que os resultados não são satisfatórios quando o número de populações independentes for pequeno e, ou, ocorrer fixação de um dos alelos em todas as subpopulações. Weir e Cockerham (1984) apresentaram estimadores baseados no método dos momentos, bem como suas variâncias, para F , f e θ . Entretanto, os autores consideraram apenas o caso de um loco.

Uma forma alternativa de medir a acurácia das estimativas desses parâmetros, que vem sendo muito utilizada em diversos trabalhos de diversidade genética, é o emprego do método de reamostragem por *bootstrap*.

Carlini-Garcia et al. (2001) avaliaram o método de reamostragem por *bootstrep* aplicado sobre locos, indivíduos, populações e indivíduos e populações concomitantemente, utilizando seis conjuntos diferentes de dados de populações naturais de plantas. Para os parâmetros F (ou F_{IT}), f (ou F_{IS}) e θ (ou F_{ST}) e taxa de cruzamento aparente (ta), foram obtidos os erros associados às suas estimativas, a distribuição empírica dessas estimativas e os intervalos de confiança para os parâmetros. Também foram feitas reamostragens com tamanho variável de amostra *bootstrap*, visando obter o número necessário de

locos, indivíduos e populações para atingir um dado nível de precisão na estimação de F , f e θ . De acordo com os autores, os tamanhos amostrais utilizados nas pesquisas com populações naturais foram suficientes apenas para estimar θ , com a precisão estabelecida. A fonte de variação “locos” foi a responsável pelos maiores erros associados a F e f . Aumentar o número de locos e de populações amostradas é a estratégia recomendada em pesquisas dessa natureza.

Carlini-Garcia et al. (2003), trabalhando com o mesmo conjunto de dados referidos anteriormente, concluíram que existe aditividade das variâncias das estimativas dos parâmetros F , f e θ , obtidas de indivíduos (i), populações (p) e indivíduos e populações concomitantemente ($i+p$). O método mostrou ser possível estimar as variâncias das estimativas de F , f e θ , para cada fonte de variação ou unidade amostral, para depois somá-las e obter a variância total. Esse procedimento facilita o uso do método *bootstrep* em dados de estrutura hierárquica.

6.6. Diferenciação medida pela heterozigosidade – Weir (1996)

6.6.1 Análise baseada em médias dos valores da heterozigosidade observada

Introdução

Uma medida simples da variação em uma população é a quantidade observada de heterozigotos para um único loco e para todos os locos, considerada em média. Portanto, quando se estudam várias populações em relação às informações genéticas de vários locos, podem-se formular várias questões a respeito da diversidade entre estas populações a partir da quantidade de heterozigotos observada em cada uma delas. Assim, deve-se avaliar se a heterozigose observada varia entre populações; em caso afirmativo, haverá indicativo de que elas estão passando por algum tipo de diferenciação ou variação

no sistema de acasalamento. Também pode-se questionar se esta heterozigosidade varia de loco para loco – em caso afirmativo, entende-se que, em certos locos, os alelos estão sujeitos à pressão de seleção, resultando em variações alélicas ou genotípicas. Adicionalmente, a análise simultânea destes dois fatores, locos e populações, possibilita estudos de interação de grande interesse no contexto genético.

Para fazer análise da heterozigose, adotam-se duas estratégias básicas. A primeira relaciona-se à obtenção dos valores médios da heterozigose para locos e populações, suas respectivas variâncias e testes de hipóteses ou intervalos de confiança. A segunda se fundamenta na análise de variância com obtenção de componentes de variância com significado biológico, que quantificam a variação da heterozigose entre populações, entre locos e da interação locos x populações.

Considerações gerais

Para fins de simplificação de notação, serão adotadas as seguintes terminologias e indexação:

g : número de populações analisadas – indexado por $i=1,2,\dots,g$.

N : número de indivíduos dentro da população – indexado por $j = 1,2,\dots, N$ (ou N_i para experimentos desbalanceados).

L : número de locos analisados – indexados por $l = 1,2,\dots,L$.

A_u, A_v : alelos de um determinado loco, sendo A_uA_v a forma heterozigota e A_uA_u e A_vA_v as formas homozigotas.

a_l : quantidade de alelos do loco l .

α_{uv} : quantidade de genótipos A_uA_v na população i , em relação ao loco l .

x_{ijl} : situação do indivíduo j em relação ao loco l numa particular população i . Admite-se $x_{ijl} = 1$ se o genótipo for heterozigoto e 0, caso contrário.

Na descrição da metodologia, será considerado que há disponibilidade de informações sobre os genótipos de indivíduos de várias populações, tendo-se o seguinte conjunto de dados:

População	Indivíduo	Locos			
		1	2	...	L
1	1	X_{111}	X_{112}	...	X_{11L}
	2	X_{121}	X_{122}	...	X_{12L}
	...				
	N	X_{1N1}	X_{1N2}	...	X_{1NL}
	1	X_{211}	X_{212}	...	X_{21L}
2	2	X_{221}	X_{222}	...	X_{22L}
	...				
	N	X_{2N1}	X_{2N2}	...	X_{2NL}
	1	X_{g11}	X_{g12}	...	X_{g1L}
	2	X_{g21}	X_{g22}	...	X_{g2L}
g	...				
	N	X_{gN1}	X_{gN2}	...	X_{gL}
	1				
	2				
	...				

i. Análise de um loco e uma população

Deve-se, inicialmente, apresentar o conceito de heterozigosidade aplicável à metodologia como sendo a freqüência de heterozigotos no loco I. Dessa forma, se numa população, para um particular loco, há, por exemplo, três alelos, pode-se estabelecer o seguinte conjunto de informações:

Genótipo	Quantidade	Valor (x_{ijl})
A_1A_1	α_{11}	0
A_1A_2	α_{12}	1
A_1A_3	α_{13}	1
A_2A_2	α_{22}	0
A_2A_3	α_{23}	1
A_3A_3	α_{33}	0

$$N = \sum_u \sum_v \alpha_{uv}$$

Assim, por definição, tem-se:

$$\hat{H}_l = \frac{\sum_{u>v}^{a_l} \sum_{v}^{a_l} \alpha_{uv,l}}{N}$$

A heterozigosidade média, considerando a existência de L locos, é dada por:

$$\bar{H} = \frac{1}{L} \sum_{l=1}^L \hat{H}_l$$

O valor α_{uv} representa a quantidade de genótipos $A_u A_v$ na população i, em relação ao loco l, com distribuição multinomial com média Nx e variância Nxy , em que x é a freqüência de heterozigotos e $y = 1-x$

A esperança matemática de \hat{H}_l é função da esperança que se tem da quantidade de heterozigotos numa população, porém expressa em valor de freqüência. Assim, pode-se definir:

$$E(\hat{H}_l) = \frac{1}{N} E(\alpha_{uv,l}) = \frac{1}{N} Nx = x = H_l$$

Por sua vez, a esperança da variância de \hat{H}_l é função da variância esperada da quantidade de heterozigotos, tendo-se:

$$E[V(\hat{H}_l)] = E\left[\frac{1}{N^2} V(\alpha_{uv})\right] = \frac{1}{N^2} Nxy = \frac{1}{N} H_l(1-H_l)$$

Na prática, não se conhece a real quantidade de cada um dos genótipos de um determinado loco. Geralmente é tomado um número N de indivíduos, os quais são genotipados, possibilitando obter a estimativa da heterozigosidade amostral. A análise dos N indivíduos de uma população em relação a L locos possibilita gerar as seguintes informações:

Indivíduo	Locos			
	1	2	...	L
1	x_{11}	x_{12}	...	x_{1L}
2	x_{21}	x_{22}	...	x_{2L}
3	x_{31}	x_{32}	...	x_{3L}
...				
N	x_{N1}	x_{N2}	...	x_{NL}

(*) omitindo de x_{ij} o índice i, por se tratar da análise preliminar de uma única população

A condição de heterozigose do indivíduo j para o loco l é dada por:

$x_{jl} = 1$, se o indivíduo é heterozigoto para o loco l; e

$x_{jl} = 0$, caso contrário.

De posse das informações de x_{jl} pode-se estimar o valor da heterozigosidade, para o loco j, por meio da seguinte expressão:

$$\hat{H}_l = \frac{1}{N} \sum_{j=1}^N x_{jl}$$

A questão que surge é se este estimador é não-viesado e qual seria a variância associada ao valor da estimativa obtida. Para resolver esses problemas, é necessário apresentar algumas propriedades interessantes em relação à variável x_{jl} , considerando que existem, para um determinado loco l, N_l heterozigotos (valores de x_{jl} iguais a 1) e $N - N_l$ homozigotos (valores de x_{jl} iguais a 0). Assim, tem-se:

a) $E(x_{jl})$. É dada por:

$$E(x_{jl}) = \frac{N_l(1) + (N - N_l)(0)}{N} = \frac{N_l}{N} = H_l$$

b) $E(x_{jl}^2)$. É fornecida por:

$$E(x_{jl}^2) = \frac{N_l(1^2) + (N - N_l)(0^2)}{N} = \frac{N_l}{N} = H_l$$

- c) $E(x_{jl} - x_{j'l})$. Haverá $N(N-1)/2$ diferenças, cuja esperança destes valores é dada a partir da tabela:

Número de indivíduos	N_l	$N - N_l$
N_l	0	1
$N - N_l$	1	0

$$E(x_{jl} - x_{j'l}) = \frac{2N_l(N - N_l)(1)}{N} = 2H_l(1 - H_l)$$

- d) $E(x_{jl} - x_{j'l})^2$. De forma análoga, tem-se:

$$E(x_{jl} - x_{j'l})^2 = \frac{2N_l(N - N_l)(1^2)}{N} = 2H_l(1 - H_l)$$

- e) $E(x_{jl} x_{j'l})$. É dada por:

$$E(x_{jl} x_{j'l}) = \frac{1}{2} [E(x_{jl}^2) + E(x_{j'l}^2) - E(x_{jl} - x_{j'l})^2] = H_l^2$$

Pode-se agora demonstrar que:

- a) Média esperada de \hat{H}_l

$$\text{Se } \hat{H}_l = \frac{1}{N} \sum_{j=1}^N x_{jl} \text{ então:}$$

$$E(\hat{H}_l) = \frac{1}{N} \sum_{j=1}^N E(x_{jl}) = H_l$$

- b) Variância esperada de \hat{H}_l

Pode ser obtida por meio de:

$$V(\hat{H}_l) = \frac{1}{N^2} V\left(\sum_{j=1}^N x_{jl}\right) = \frac{1}{N^2} [NV(x_{jl}) + N(N-1)\text{Cov}(x_{jl}, x_{j'l})]$$

Sabe-se que:

$$V(x_{jl}) = E(x_{jl}^2) - [E(x_{jl})]^2 = H_l(1 - H_l)$$

$$\text{Cov}(x_{jl}, x_{j'l}) = E(x_{jl} x_{j'l}) - E(x_{jl})E(x_{j'l})$$

logo:

$$\text{Cov}(x_{ji}, x_{j'i}) = H_i^2 - H_i H_{i'} = 0$$

Assim:

$$E[V(\hat{H}_i)] = E\left[V\left(\frac{1}{N} \sum_{j=1}^N x_{ij}\right)\right] = \frac{1}{N^2} N E[V(x_{ij})] = \frac{1}{N} H_i (1 - H_i)$$

O fato de a $\text{Cov}(x_{ji}, x_{j'i})$ ser nula reflete a independência dos valores x_{ji} entre os vários indivíduos dentro de cada loco.

O conhecimento desses estimadores possibilita não só quantificar a heterozigose de uma população, para um particular loco, mas também testar sua significância e estabelecer intervalos de confiança.

ii. Análise de vários locos e uma população

Além da heterozigose manifestada em um loco, pode-se ter o interesse em obter o valor médio para o conjunto de locos estudados. A heterozigosidade média é dada pela média da heterozigosidade obtida para todos os m locos estudados, ou seja:

$$\bar{H} = \frac{1}{L} \sum_{i=1}^L \hat{H}_i = \frac{1}{LN} \sum_{j=1}^N \sum_{i=1}^L x_{ji}$$

Este estimador é não-viesado. Assim, verifica-se que a esperança matemática da heterozigosidade média amostral é a heterozigosidade média da população, ou seja, a heterozigosidade média da amostra é estimador não-tendencioso da heterozigosidade da população. Dessa forma, tem-se:

$$E(\bar{H}) = \frac{1}{L} \sum_{i=1}^L E(\hat{H}_i) = \frac{1}{L} \sum_{i=1}^L H_i = \bar{H}$$

Uma questão interessante é como obter um estimador não-tendencioso da heterozigosidade para uma população, considerando os vários locos, de forma que se possa aplicar testes estatísticos apropriados. Deve ser ressaltado que, de posse

dos valores de \hat{H}_l ($l=1,2,\dots,L$), é possível obter um estimador viesado da variação de H por meio de:

$$\tilde{S}_H^2 = \frac{1}{L-1} \sum_{l=1}^L (\hat{H}_l - \bar{H})^2$$

Entretanto, uma maneira apropriada de se obter a variância de \hat{H}_l é por meio de outro estimador, descrito a seguir:

$$V(\bar{H}) = V\left(\frac{1}{m} \sum_{l=1}^m \hat{H}_l\right) = \frac{1}{L^2} \left[\sum_{l=1}^L V(\hat{H}_l) + \sum_{l \neq l'} \sum_{l'} \text{cov}(\hat{H}_l, \hat{H}_{l'}) \right]$$

de outra forma:

$$V(\bar{H}) = V\left(\frac{1}{LN} \sum_j \sum_{l=1}^L x_{jl}\right) = \frac{1}{L^2 N^2} \left[LNV(x_{jl}) + L[N(N-1)]\text{Cov}(x_{jl}, x_{j'l'}) + N[L(L-1)]\text{Cov}(x_{jl}, x_{j'l'}) \right]$$

Deve-se destacar que a $\text{cov}(\hat{H}_l, \hat{H}_{l'})$, apresentada na expressão anterior, não é nula, pois ela depende da freqüência de duplos heterozigotos, ou seja, depende da probabilidade de surgir um determinado genótipo para um loco, dada a heterozigosidade já estabelecida para outro loco. Essa freqüência de duplos heterozigotos denomina-se $H_{ll'}$ e pode ser entendida como uma medida da $E(x_{jl}, x_{j'l'})$.

Para fins de cálculo da $E(x_{jl}, x_{j'l'})$, devem-se considerar as seguintes informações:

Loco l	Loco l'	Número de indivíduos
1	1	$n_{11,ll'}$
1	0	$n_{10,ll'}$
0	1	$n_{01,ll'}$
0	0	$n_{00,ll'}$

Dessa forma, tem-se:

$$E(x_{jl}, x_{j'l'}) = \frac{n_{11,ll'}(1) + n_{10,ll'}(0) + n_{01,ll'}(0) + n_{00,ll'}(0)}{N} = \frac{n_{11,ll'}}{N} = H_{ll'}$$

portanto:

$$\text{Cov}(x_{jl}, x_{j'l'}) = \frac{1}{N-1} \left[\sum_{j=1}^N x_{jl} x_{j'l'} - \frac{\sum_{j=1}^N x_{jl} \sum_{j=1}^N x_{j'l'}}{N} \right] = \hat{H}_{ll'} - \hat{H}_l \hat{H}_{l'}$$

e

$$\text{Cov}(x_{jl}, x_{jl}) = 0$$

De forma que:

$$\text{Cov}(\hat{H}_l, \hat{H}_{l'}) = \text{Cov}\left(\frac{1}{n} \sum_{j=1}^n x_{jl}, \frac{1}{n} \sum_{j=1}^n x_{j'l'}\right) = \frac{1}{N^2} [NCov(x_{jl}, x_{j'l'}) + (N^2 - N)\text{Cov}(x_{jl}, x_{jl})]$$

ou

$$\text{Cov}(\hat{H}_l, \hat{H}_{l'}) = \frac{1}{N} (\hat{H}_{ll'} - \hat{H}_l \hat{H}_{l'})$$

Voltando à expressão da $V(\hat{H})$ tem-se:

$$V(\hat{H}) = \frac{1}{L^2} \left[\sum_{l=1}^L V(\hat{H}_l) + \sum_{l \neq l'} \sum_{l'}^L \text{cov}(\hat{H}_l, \hat{H}_{l'}) \right]$$

conclui-se que:

$$V(\bar{H}) = \frac{1}{L^2} \left[\frac{1}{N} \sum_{l=1}^L [\hat{H}_l(1 - \hat{H}_l)] + \frac{1}{N} \sum_{l \neq l'} \sum_{l'}^L (\hat{H}_{ll'} - \hat{H}_l \hat{H}_{l'}) \right]$$

lembrando que:

$$\hat{H}_l = \frac{1}{N} \sum_{j=1}^N x_{jl} \quad \text{e} \quad \hat{H}_{ll'} = \frac{1}{N} \sum_{j=1}^N x_{jl} x_{j'l'}$$

Trata-se de um estimador não-tendencioso da variância de \bar{H} .

iii. Análise de vários locos e populações

Para uma análise conjunta das informações contidas em todos os locos e todas as populações, deve-se ter em mente os valores das esperanças matemáticas envolvendo indivíduos, locos e populações diferentes. Assim, tem-se:

Para análise de uma única população

- a) $E(x_{ijl}) = H_{il}$
- b) $E(x_{ijl}^2) = H_{il}$
- c) $E(x_{ijl}, x_{ijl}) = H_{il}^2$
- d) $E(x_{ijl}, x_{ijl}) = H_{i,il}$

Para análise de mais de uma população

a) $E(x_{ijl}, x_{ij'l}) = M_l$

De fato, se for considerado que a covariância entre dois indivíduos diferentes é nula, tem-se:

$$\text{Cov}(x_{ijl}, x_{ij'l}) = 0$$

logo:

$$\text{Cov}(x_{ijl}, x_{ij'l}) = E(x_{ijl}, x_{ij'l}) - E(x_{ijl})E(x_{ij'l}) = E(x_{ijl}, x_{ij'l}) - H_{il}H_{il} = 0$$

então:

$$E(x_{ijl}, x_{ij'l}) = H_{il}H_{il} = M_l$$

Para simplificação de simbologia, será adotado:

$$E(x_{ijl}, x_{ij'l}) = E(x_{ijl}, x_{ijl}) = M_l$$

se $i = i'$, então $M_l = H_{il}^2$; caso contrário, $M_l = H_{il}H_{il}$

b) $E(x_{ijl}, x_{ij'l}) = M_{ll}$

Sabe-se que a covariância entre dois locos diferentes para um mesmo indivíduo não é nula, sendo dada por:

$$\text{Cov}(x_{ijl}, x_{ij'l}) = H_{il}^2 - H_{il}H_{il}$$

Por outro lado, a covariância entre dois locos diferentes, para dois indivíduos diferentes (pertencentes ou não a uma mesma população), pode ser expressa por:

$$\text{Cov}(x_{ijl}, x_{ij'l'}) = E(x_{ijl}, x_{ij'l'}) - E(x_{ijl})E(x_{ij'l'}) = E(x_{ijl}, x_{ij'l'}) - H_{il}H_{il'}$$

assim, denotando $E(x_{ijl}, x_{ij'l'}) = M_{ll'}$ ter-se-á:

$$\text{Cov}(x_{ijl}, x_{ij'l'}) = M_{ll'} - H_{il}H_{il'}$$

Para simplificação de simbologia, será adotado:

$$E(x_{ijl}, x_{ij'l'}) = E(x_{ijl}, x_{ij'l}) = M_{ll'}$$

se $i = i'$, e $j = j'$ então $M_{ll'} = H_{ll'}$

De maneira geral, considera-se que M_l é a probabilidade de que dois indivíduos sejam heterozigotos para o loco l , de forma que $E(x_{ijl}, x_{ij'l}) = E(x_{ijl}, x_{ij'l}) = M_l$. $M_{ll'}$ é a probabilidade de que dois indivíduos ao acaso sejam duplos heterozigotos, um heterozigoto para o loco l e o outro para o loco l' , de forma que $E(x_{ijl}, x_{ij'l'}) = E(x_{ijl}, x_{ij'l}) = M_{ll'}$.

6.6.2 Análise baseada em componentes de variâncias associados à heterozigosidade de populações, locos e interação locos x populações

A análise das hipóteses de existência de heterozigosidade diferencial entre populações ou dentro de populações pode ser feita por meio do modelo estatístico:

$$X_{ijl} = \alpha_i + \beta_{ij} + \gamma_l + \alpha\gamma_{il} + \beta\gamma_{ijl}$$

em que:

α_i : efeito da i -ésima população, com $\alpha_i \sim ID(0, \sigma_p^2)$;

β_{ij} : efeito do j -ésimo indivíduo dentro da i -ésima população, com $\beta_{ij} \sim ID(0, \sigma_{i/p}^2)$;

γ_l : efeito fixo do l -ésimo loco. $E(\gamma_l) = E(x_{ijk}) = H_l$ e $E(\gamma_l^2) = H_l^2$;

$\alpha\gamma_{il}$: efeito da interação população x loco, com $\alpha\beta_{il} \sim ID(0, \sigma_{pl}^2)$; e

$\beta\gamma_{ijl}$: efeito da interação indivíduo/população x loco, com $\beta\gamma_{ijl} \sim ID(0, \sigma_{li/p}^2)$.

Devem ser consideradas as seguintes esperanças matemáticas:

a) $E(x_{ijl}^2)$

$$E(x_{ijl}^2) = E(\alpha_i + \beta_{ij} + \gamma_l + \alpha\gamma_{il} + \beta\gamma_{ijl})^2 = E(\alpha_i^2 + \beta_{ij}^2 + \gamma_l^2 + \alpha\gamma_{il}^2 + \beta\gamma_{ijl}^2 + DP)$$

em que DP se refere aos duplos produtos entre efeitos, considerados independentes e, portanto, com esperança nula.

$$E(x_{ijl}^2) = \sigma_p^2 + \sigma_{i/p}^2 + H_l + \sigma_{pl}^2 + \sigma_{li/p}^2$$

b) $E(x_{ij.}^2)$

$$\begin{aligned} E(x_{ij.}^2) &= E\left(\sum_{l=1}^L \alpha_i + \beta_{ij} + \gamma_l + \alpha\gamma_{il} + \beta\gamma_{ijl}\right)^2 = E\left(L\alpha_i + L\beta_{ij} + \sum_{l=1}^L \gamma_l + \sum_{l=1}^L \alpha\gamma_{il} + \sum_{l=1}^L \beta\gamma_{ijl}\right)^2 \\ &= E\left[L^2\alpha_i^2 + L^2\beta_{ij}^2 + \left(\sum_{l=1}^L \gamma_l\right)^2 + \left(\sum_{l=1}^L \alpha\gamma_{il}\right)^2 + \left(\sum_{l=1}^L \beta\gamma_{ijl}\right)^2 + DP\right] \\ &= L^2\sigma_p^2 + L^2\sigma_{i/p}^2 + \left(\sum_{l=1}^L H_l\right)^2 + L\sigma_{pl}^2 + L\sigma_{li/p}^2 \end{aligned}$$

c) $E(x_{i..}^2)$

$$\begin{aligned} E(x_{i..}^2) &= E\left(\sum_{j=1}^N \alpha_i + \beta_{ij} + \gamma_l + \alpha\gamma_{il} + \beta\gamma_{ijl}\right)^2 = E\left(N\alpha_i + \sum_{j=1}^N \beta_{ij} + N\gamma_l + N\alpha\gamma_{il} + \sum_{j=1}^N \beta\gamma_{ijl}\right)^2 \\ &= E\left[N^2\alpha_i^2 + \left(\sum_{j=1}^N \beta_{ij}\right)^2 + N^2\gamma_l^2 + N^2(\alpha\gamma_{il})^2 + \left(\sum_{j=1}^N \beta\gamma_{ijl}\right)^2 + DP\right] \\ &= N^2\sigma_p^2 + N\sigma_{i/p}^2 + N^2H_l^2 + N^2\sigma_{pl}^2 + N\sigma_{li/p}^2 \end{aligned}$$

d) $E(x_{...}^2)$

$$\begin{aligned} E(x_{...}^2) &= E\left(\sum_{j=1}^N \sum_{l=1}^L \alpha_i + \beta_{ij} + \gamma_l + \alpha\gamma_{il} + \beta\gamma_{ijl}\right)^2 \\ &= E\left(NL\alpha_i + L\sum_{j=1}^N \beta_{ij} + N\sum_{l=1}^L \gamma_l + N\sum_{l=1}^L \alpha\gamma_{il} + \sum_{j=1}^N \sum_{l=1}^L \beta\gamma_{ijl}\right)^2 \\ &= E\left[N^2L^2\alpha_i^2 + L^2\left(\sum_{j=1}^N \beta_{ij}\right)^2 + N^2\left(\sum_{l=1}^L \gamma_l\right)^2 + N^2\left(\sum_{l=1}^L \alpha\gamma_{il}\right)^2 + \left(\sum_{j=1}^N \sum_{l=1}^L \beta\gamma_{ijl}\right)^2 + DP\right] \\ &= N^2L^2\sigma_p^2 + NL^2\sigma_{i/p}^2 + N^2\left(\sum_{l=1}^L H_l\right)^2 + N^2L\sigma_{pl}^2 + NL\sigma_{li/p}^2 \end{aligned}$$

e) $E(x_{..l}^2)$

$$\begin{aligned}
E(x_{..l}^2) &= E\left(\sum_{i=1}^g \sum_{j=1}^N \alpha_i + \beta_{ij} + \gamma_l + \alpha\gamma_{il} + \beta\gamma_{ijl}\right)^2 \\
&= E\left(N \sum_{i=1}^g \alpha_i + \sum_{i=1}^g \sum_{j=1}^N \beta_{ij} + Ng\gamma_l + N \sum_{i=1}^g \alpha\gamma_{il} + \sum_{i=1}^g \sum_{j=1}^N \beta\gamma_{ijl}\right)^2 \\
&= E\left[N^2 \left(\sum_{i=1}^g \alpha_i\right)^2 + \left(\sum_{i=1}^g \sum_{j=1}^N \beta_{ij}\right)^2 + N^2 g^2 \gamma_l^2 + N^2 \left(\sum_{i=1}^g \alpha\gamma_{il}\right)^2 + \left(\sum_{i=1}^g \sum_{j=1}^N \beta\gamma_{ijl}\right)^2 + DP\right] \\
&= N^2 g \sigma_p^2 + Ng\sigma_{i/p}^2 + N^2 g^2 H_l^2 + N^2 g \sigma_{pl}^2 + Ng\sigma_{il/p}^2
\end{aligned}$$

f) $E(x_{...}^2)$

$$\begin{aligned}
E(x_{...}^2) &= E\left(\sum_{i=1}^g \sum_{j=1}^N \sum_{l=1}^L \alpha_i + \beta_{ij} + \gamma_l + \alpha\gamma_{il} + \beta\gamma_{ijl}\right)^2 \\
&= E\left(NL \sum_{i=1}^g \alpha_i + L \sum_{i=1}^g \sum_{j=1}^N \beta_{ij} + Ng \sum_{l=1}^L \gamma_l + N \sum_{i=1}^g \sum_{l=1}^L \alpha\gamma_{il} + \sum_{i=1}^g \sum_{j=1}^N \sum_{l=1}^L \beta\gamma_{ijl}\right)^2 \\
&= E\left[N^2 L^2 \left(\sum_{i=1}^g \alpha_i\right)^2 + L^2 \left(\sum_{i=1}^g \sum_{j=1}^N \beta_{ij}\right)^2 + N^2 g^2 \left(\sum_{l=1}^L \gamma_l\right)^2 + N^2 \left(\sum_{i=1}^g \sum_{l=1}^L \alpha\gamma_{il}\right)^2 + \left(\sum_{i=1}^g \sum_{j=1}^N \sum_{l=1}^L \beta\gamma_{ijl}\right)^2 + DP\right] \\
&= N^2 L^2 g \sigma_p^2 + NL^2 g \sigma_{i/p}^2 + N^2 g^2 \left(\sum_{l=1}^L H_l\right)^2 + N^2 L g \sigma_{pl}^2 + NL g \sigma_{il/p}^2
\end{aligned}$$

Para análise de variância dos valores que representam a heterozigose da população, são obtidas as seguintes somas de quadrados:

a) Soma de quadrados de populações

É dada por:

$$SQP = \frac{1}{LN} \sum_{i=1}^g x_{i..}^2 - C$$

$$\text{sendo } C = \frac{x_{...}^2}{NLg}$$

b) Soma de quadrados de indivíduos/populações

É fornecida por:

$$SQI/P = \frac{1}{L} \sum_{i=1}^g \sum_{j=1}^N x_{ij..}^2 - \frac{1}{NL} \sum_{i=1}^g x_{i..}^2$$

c) Soma de quadrados de locos

É dada por:

$$SQL = \frac{1}{Ng} \sum_{l=1}^L x_{..l}^2 - C$$

d) Soma de quadrados de locos x populações

É fornecida por:

$$SQLP = \frac{1}{N} \sum_{i=1}^g \sum_{l=1}^L x_{i..l}^2 - \frac{1}{LN} \sum_{i=1}^g x_{i..}^2 - \frac{1}{Ng} \sum_{l=1}^L x_{..l}^2 + C$$

e) Soma de quadrados de locos x indivíduos/populações

É dada por:

$$SQL(I/P) = \sum_{i=1}^g \sum_{j=1}^N \sum_{l=1}^L x_{ijl}^2 - \frac{1}{L} \sum_{i=1}^g \sum_{j=1}^N x_{ij..}^2 - \frac{1}{N} \sum_{i=1}^g \sum_{l=1}^L x_{i..l}^2 - \frac{1}{LN} \sum_{i=1}^g x_{i..}^2$$

O quadro de análise de variância é apresentado a seguir:

FV	GL	QM	E(QM)
Populações	g-1	QMP	$\sigma_{ii/p}^2 + L\sigma_{i/p}^2 + N\sigma_{pl}^2 + LN\sigma_p^2$
Indivíduos/Pop.	$g(N-1)$	QMI	$\sigma_{ii/p}^2 + L\sigma_{i/p}^2$
Locos	L-1	QML	$\sigma_{ii/p}^2 + N\sigma_{pl}^2 + \Psi$
Locos x Popul.	$(g-1)(L-1)$	QMLP	$\sigma_{ii/p}^2 + N\sigma_{pl}^2$
Locos x Ind/Pop	$g(N-1)(L-1)$	QMR	$\sigma_{ii/p}^2$

$$\Psi = \frac{1}{L-1} \sum_{l=1}^L (H_l - \bar{H})^2$$

Por fim, resta estabelecer o significado de cada componente de variância do modelo. Assim, tem-se:

a) σ_p^2

$$\sigma_p^2 = \text{Cov}(x_{ijl}, x_{ijl'}) = E(x_{ijl}x_{ijl'}) - E(x_{ijl})E(x_{ijl'}) = M_{ll'} - H_l H_{l'}$$

Considerando todos os locos, tem-se:

$$\sigma_p^2 = \frac{1}{L(L-1)} \sum_{l \neq l'}^L (M_{ll'} - H_l H_{l'})$$

b) $\sigma_{i/p}^2$

$$\sigma_{i/p}^2 = Cov(x_{ij}, x_{ij'}) - Cov(x_{ij}, x_{ij''})$$

mas:

$$Cov(x_{ij}, x_{ij'}) = E(x_{ij}x_{ij'}) - E(x_{ij})E(x_{ij'}) = H_{ii'} - H_i H_{i'}$$

logo:

$$\sigma_{i/p}^2 = Cov(x_{ij}, x_{ij'}) - Cov(x_{ij}, x_{ij''}) = (H_{ii'} - H_i H_{i'}) - (M_{ii'} - H_i H_{i'}) = H_{ii'} - M_{ii'}$$

Considerando todos os locos, tem-se:

$$\sigma_{i/p}^2 = \frac{1}{L(L-1)} \sum_{l \neq l'}^L (H_{ll'} - M_{ll'})$$

c) σ_{pl}^2

$$\sigma_{pl}^2 = Cov(x_{ij}, x_{ij'}) - Cov(x_{ij}, x_{ij''})$$

mas:

$$Cov(x_{ij}, x_{ij'}) = E(x_{ij}x_{ij'}) - E(x_{ij})E(x_{ij'}) = M_i - H_i^2$$

logo:

$$\sigma_{pl}^2 = Cov(x_{ij}, x_{ij'}) - Cov(x_{ij}, x_{ij''}) = (M_i - H_i^2) - (M_{ii'} - H_i H_{i'})$$

considerando todos os locos, tem-se:

$$\sigma_{i/p}^2 = \frac{1}{L} \sum_{l=1}^L (M_l - H_l^2) - \frac{1}{L(L-1)} \sum_{l \neq l'}^L (M_{ll'} - H_l H_{l'})$$

d) $\sigma_{li/p}^2$

$$\sigma_{li/p}^2 = Cov(x_{ij}, x_{ij''}) - (\sigma_p^2 + \sigma_{i/p}^2 + \sigma_{pl}^2)$$

mas:

$$Cov(x_{ij}, x_{ij''}) = E(x_{ij}x_{ij''}) - E(x_{ij})E(x_{ij''}) = H_i - H_i^2$$

Considerando todos os locos, tem-se:

$$\sigma_{i/p}^2 = \frac{1}{L} \sum_{l=1}^L (H_l - M_l) - \frac{1}{L(L-1)} \sum_{l \neq l'}^L \sum_{l'}^L (H_{ll'} - M_{ll'})$$

Aplicação

O mesmo exemplo referente a oito populações será considerado como ilustração.

Deve-se, entretanto, lembrar que se trata do estudo de apenas um loco e, portanto, ter-se-á interesse no componente de variação associado às populações. Assim, são obtidos os resultados:

FV	GL	SQ	QM	F
Populações	7	1,1547	0,1650	1,70ns
Indivíduos/Pop	144	13,9439	0,0968	
Total	151	15,0987		

$$\hat{\sigma}_p^2 = 0,0038$$

$$\hat{\sigma}_{i/p}^2 = 0,0968$$

A quantidade de variação entre populações atribuída unicamente a σ_p^2 é dada por:

$$Var_{ST} = \frac{100\hat{\sigma}_p^2}{QMP/\bar{N}} = 41,3\%$$

$$\text{sendo } \bar{N} = \frac{N - \left(\sum_{i=1}^g N_i^2 \right)/N}{g-1} = 18,0639$$

6.7 Análise molecular de variância - AMOVA

6.7.1 Introdução

A distribuição da variabilidade genética entre e dentro das populações também pode ser avaliada pela metodologia AMOVA (análise de variância molecular), descrita por Excoffier et al. (1992). Nesse método, a matriz de distâncias entre todos os pares de haplótipos (genótipos) é utilizada em um esquema de análise de variância hierarquizada, produzindo estimativas de componentes de variância análogas às estatísticas F de Wright. O termo haplótipo, empregado nesse contexto, refere-se à combinação de sítios de amplificação presente em alguma área definida do genoma. No caso de marcadores moleculares dominantes, o haplótipo de um indivíduo é geralmente representado por um vetor p , com valores 1, se a banda homóloga estiver presente, e 0, em caso contrário. Diferentes métodos para obtenção da matriz distância fenotípica podem ser utilizados.

A caracterização da diversidade entre populações (ou subpopulações) por meio do índice de fixação F de Wright é baseada nas comparações das freqüências de gene entre subpopulações. Contudo, informações moleculares podem ser tanto úteis em estimar a freqüência alélica quanto quantificar as diferenças mutacionais entre diferentes genes. Assim, técnicas usadas para calcular diferenciação de população analisando diferenças entre seqüências moleculares são igualmente úteis.

A análise de Variância Molecular (AMOVA) é uma metodologia capaz de estudar a diversidade entre populações a partir de dados moleculares e também de testar hipóteses a respeito de tal diferenciação. Uma variedade de dados moleculares, como informações de marcadores dominantes ou co-dominantes e dados de seqüência, pode ser analisada usando este método (Excoffier et al., 1992).

A AMOVA estuda qualquer tipo de dados moleculares estruturados em um vetor p_j de variável booliana, ou seja, uma matriz $1 \times n$ constituída por uns ou zeros, em que 1 que indica a presença de um marcador e 0 sua ausência. Um marcador

poderia ser uma base de nucleotídeo, uma seqüência básica, um fragmento de restrição ou um evento de mutacional (EXCOFFIER, et al., 1992).

Para dois haplótipos j e k , os vetores contendo as informações de quatro marcadores moleculares poderiam ser representados por:

$$p_j = [1 \ 1 \ 0 \ 1]$$

$$p_k = [0 \ 1 \ 1 \ 0]$$

Para realizar a AMOVA é necessário, inicialmente, calcular a distância euclidiana entre pares de vetores (ou pares de indivíduos), subtraindo o vetor booliano de um haplótipo de outro, de acordo com a fórmula $(p_j - p_k)$. A distância euclidiana é um escalar que expressa a menor distância menor entre esses dois indivíduos. Normalmente, calculam-se os quadrados das distâncias euclidianas por meio da expressão:

$$d_{jk}^2 = (p_j - p_k)' W (p_j - p_k)$$

em que W é uma matriz de pesos atribuídos a cada informação molecular. Geralmente estes pesos são estabelecidos a partir de informações sobre mudanças moleculares localizadas em diferentes posições numa seqüência ou numa árvore filogenética (EXCOFFIER, et al., 1992). Na falta de informações sobre esses pesos, utiliza-se uma matriz de identidade.

Para uso desta metodologia, é interessante ressaltar a associação entre a soma de quadrado de distância e a soma de quadrados de desvio para uma (ou várias) variável. Assim, supondo uma variável X , com valores X_1, X_2, \dots, X_n , pode ser obtida a estatística:

$$SQ = \sum_{i=1}^n X_i^2 - \frac{1}{n} \left(\sum_{i=1}^n X_i \right)^2$$

Para cada par de indivíduos, é também obtido o quadrado da distância que pode ser colocados na matriz T , dada por:

$$\tilde{T} = \begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 & \dots & d_{1n}^2 \\ d_{21}^2 & 0 & d_{23}^2 & \dots & d_{2n}^2 \\ d_{31}^2 & d_{32}^2 & 0 & \dots & d_{3n}^2 \\ \dots & \dots & \dots & \dots & \dots \\ d_{n1}^2 & d_{n2}^2 & d_{n3}^2 & \dots & 0 \end{bmatrix}$$

Verifica-se que existe relação entre SQ e o total dos n^2 elementos de \tilde{T} (dados por T), expressa por:

$$SQ = \frac{1}{2n} T$$

Assim, por exemplo, se X assume os valores 2, 4 e 6, tem-se:

$$\sum x_i = 12 \quad \sum x_i^2 = 144 \quad \text{e} \quad SQ = 56 - (144/3) = 8$$

$$\tilde{T} = \begin{bmatrix} 0 & 4 & 16 \\ 4 & 0 & 4 \\ 16 & 4 & 0 \end{bmatrix} \quad \text{logo, } T = 48$$

portanto:

$$SQ = \frac{1}{2n} T = \frac{1}{6} 48 = 8$$

6.7.2 AMOVA com dois níveis hierárquicos

Para análise da diversidade entre e dentro de populações pela metodologia proposta por Excoffier et al. (1992), deve-se, inicialmente, calcular a matriz de dissimilaridade, cujos elementos são os quadrados das distâncias euclidianas entre pares de acessos, denotada por D^2 . Esta matriz é, então, particionada, agrupando as informações de pares de haplótipos dentro de cada subpopulação. Assim, a matriz original D^2 pode ser apresentada das seguintes formas:

$$D^2 = [\tilde{T}]$$

ou

$$D^2 = \begin{bmatrix} [\tilde{D}_1] & & & \\ & [\tilde{D}_2] & & \\ & & \dots & \\ & & & [\tilde{D}_g] \end{bmatrix}$$

Tendo-se as seguintes informações:

Subpopulações	Matriz	Total elementos	dos indivíduos	de
1	\tilde{D}_1	D_1	N_1	
2	\tilde{D}_2	D_2	N_2	
...	
p	\tilde{D}_g	D_g	N_g	
Total	\tilde{T}	$T \neq \sum_{i=1}^p D_i$	$N = \sum_{i=1}^g N_i$	

Obtenção de somas de quadrados

Soma de quadrados das distâncias – total

É dada por

$$SQT = \frac{1}{2N} T$$

Soma de quadrados das distâncias – dentro de subpopulações

É fornecida por:

$$SQD = \sum_{i=1}^g \frac{D_i}{2N_i}$$

Soma de quadrados das distâncias – entre subpopulações

É dada por:

$$SQE = SQT - SQD$$

Esquema da análise de variância

De posse dos valores de somas de quadrados, é estabelecido o quadro de análise de variância molecular (AMOVA), conforme descrito a seguir. Os valores das esperanças dos quadrados médios são obtidos considerando um modelo de dois fatores de classificação, em que as medidas de dissimilaridade são estabelecidas a partir das informações das variações entre e dentro de subpopulações, ou seja:

$$Y_{ij} = u + P_i + D_{ij}$$

em que:

Y_{ij} : medida da dissimilaridade entre pares de indivíduos (j) numa subpopulação i;

u: constante

P_i : efeito da subpopulação i; e

D_{ij} : efeito da dissimilaridade entre pares de indivíduos j na subpopulação i.

O esquema da análise de variância molecular (AMOVA), com dados agrupados em dois níveis hierárquicos, é apresentado a seguir:

Fonte de Variação	GL	SQ	QM	E(QM)
Entre Populações	g-1	SQE	QME	$\sigma_i^2 + \tilde{N}\sigma_p^2$
Dentro Populações	N-g	SQD	QMD	σ_i^2
Total	N-1	SQT	-	σ_T^2

em que g representa o número médio de haplótipos amostrados por população.

Com amostras de tamanhos desiguais, \tilde{N} é obtido por:

$$\tilde{N} = \frac{N - \sum_i \frac{N_i^2}{N}}{g-1}, \text{ sendo } N_i \text{ o número de indivíduos ou haplótipos da } i\text{-ésima subpopulação.}$$

Estimação de componentes de variância

Os estimadores dos componentes de variância que expressam as diferenciações entre e dentro de subpopulações são obtidos por meio de:

$$\hat{\sigma}_i^2 = \text{QMD}$$

e

$$\hat{\sigma}_p^2 = \frac{\text{QME} - \text{QMD}}{\tilde{N}}$$

Associação entre componente de variância e estatísticas Φ

As estatísticas obtidas pela AMOVA, denominadas de “estatísticas Φ ”, refletem a correlação da diversidade de haplótipos em diferentes níveis da subdivisão hierárquica, para o caso em que se consideram dois níveis hierárquicos (populações e indivíduos ou haplótipos avaliados).

Assim, o coeficiente de correlação entre haplótipos, amostrados aleatoriamente entre populações, é dado por:

$$r = \Phi_{ST} = \frac{\text{Cov}(Y_{ij}, Y_{ij'})}{\sqrt{V(Y_{ij})V(Y_{ij'})}}$$

Logo, $\Phi_{ST} = \frac{\sigma_p^2}{\sigma_T^2}$, que é uma medida da diversidade relativa entre as populações

avaliadas.

Testes estatísticos dos componentes de variância e das estatísticas Φ

A significância dos componentes σ_p^2 e Φ_{ST} pode ser testada pela permutação de haplótipos entre populações, por meio dos procedimentos fundamentados em permutação dos dados. Dois tipos de permutação são realizados. No primeiro, para estabelecimento da nulidade dos componentes σ_i^2 e σ_p^2 , todos os dados de cada informação molecular são permutados, independentemente das populações a que pertencem. No segundo teste, admite-se ser verdadeira a divisão da população em subpopulações, e testa-se apenas a

existência de variabilidade dentro das subpopulações. Neste caso, a permutação dos dados deve ocorrer apenas dentro de subpopulações.

O uso da AMOVA poderá ser estendido a outros modelos hierarquizados e servirá para estimar outros parâmetros populacionais.

6.7.3 AMOVA com três níveis hierárquicos

Neste caso, também são calculados os quadrados das distâncias euclidianas para todos os arranjos de pares de vetores booliano, os quais são dispostos e organizados em uma matriz. São feitas as partições nesta matriz, gerando-se submatrizes, que corresponderão às subdivisões dentro da população (EXCOFFIER, et al., 1992). Como ilustração, pode ser considerada a avaliação de uma população em relação a 14 indivíduos, caracterizada por dois níveis de subdivisão. O primeiro refere-se a regiões, tendo-se no exemplo cinco indivíduos na região 1 e nove na região 2. Uma segunda divisão ocorre dentro de cada região. Para o exemplo em consideração, há dentro da região 1, duas subpopulações com dois e três indivíduos cada. Dentro da região 2, têm-se três subpopulações com dois, três e quatro indivíduos cada. A matriz D^2 pode ser caracterizada por:

$$\left[\begin{array}{c} \left[\begin{array}{cc} 0 & d_{12}^2 \\ 0 & \end{array} \right] \quad d_{13} \quad d_{14} \quad d_{15} \\ \left[\begin{array}{ccc} d_{23} & d_{24} & d_{25} \\ 0 & d_{34}^2 & d_{35}^2 \\ 0 & d_{45}^2 \\ 0 \end{array} \right] \end{array} \right] \quad \left[\begin{array}{ccc} d_{1,6}^2 & \dots & d_{1,14}^2 \\ d_{2,6}^2 & \dots & d_{2,14}^2 \\ d_{3,6}^2 & \dots & d_{3,14}^2 \\ d_{4,6}^2 & \dots & d_{4,14}^2 \\ d_{5,6}^2 & \dots & d_{5,14}^2 \end{array} \right]$$

$$\left[\begin{array}{c} \left[\begin{array}{c} 0 \quad d_{67}^2 \\ 0 \end{array} \right] \\ \left[\begin{array}{ccc} 0 & d_{89}^2 & d_{8,10}^2 \\ 0 & d_{9,10}^2 \\ 0 \end{array} \right] \end{array} \right] \quad \left[\begin{array}{c} \left[\begin{array}{cccc} 0 & d_{11,12}^2 & d_{11,13}^2 & d_{11,14}^2 \\ 0 & d_{12,13}^2 & d_{12,14}^2 \\ 0 & d_{13,14}^2 \\ 0 \end{array} \right] \end{array} \right]$$

Assim, podem ser caracterizadas as subdivisões em D^2 :

$D^2 = [\tilde{T}]_{14 \times 14}$. A soma de todos os elementos de \tilde{T} é representada por T .

$$D^2 = \begin{bmatrix} [\tilde{R}_1]_{5 \times 5} & [\tilde{R}_{12}]_{5 \times 9} \\ [\tilde{R}_{21}]_{9 \times 5} & [\tilde{R}_2]_{9 \times 9} \end{bmatrix}. \text{ A soma dos elementos de } \tilde{R}_i \text{ é representada por } R_i.$$

$$D^2 = \begin{bmatrix} [\tilde{S}_{1.1}]_{2 \times 2} & [\tilde{S}_{12.1}]_{2 \times 3} & & & \\ [\tilde{S}_{21.1}]_{3 \times 2} & [\tilde{S}_{2.1}]_{3 \times 3} & & & \dots \\ & & \dots & & \\ & & & [\tilde{S}_{1.2}]_{2 \times 2} & [\tilde{S}_{12.2}]_{2 \times 3} & [\tilde{S}_{13.2}]_{2 \times 4} \\ & & & [\tilde{S}_{21.2}]_{3 \times 2} & [\tilde{S}_{2.2}]_{3 \times 3} & [\tilde{S}_{23.2}]_{3 \times 4} \\ & & & & [\tilde{S}_{31.2}]_{4 \times 2} & [\tilde{S}_{32.2}]_{4 \times 3} & [\tilde{S}_{3.2}]_{4 \times 4} \end{bmatrix}$$

A soma dos elementos de \tilde{S}_{jk} é S_{jk} . Assim, têm-se as seguintes informações:

Regiões	Subpopulações	Total de d_{ij}^2	Número de indivíduos N_{ij}
1	1	S_{11}	N_{11}
	2	S_{21}	N_{21}
	...		
	s_1	$S_{s1.1}$	$N_{s1.1}$
Total de região		R_1	$N_{.1}$
1			
2	1	S_{12}	N_{12}
	2	S_{22}	N_{22}
	...		
	s_2	$S_{s2.2}$	$N_{s2.2}$
Total de região		R_2	$N_{.2}$
2			
...			
R	1	S_{1r}	N_{1r}
	2	S_{2r}	N_{2r}
	...		
	s_r	$S_{sr.r}$	$N_{sr.r}$
Total de região		R_r	$N_{.r}$
r			
Total geral		T	N

Veja que a matriz é arranjada e particionada de tal modo que as submatrizes, dispostas na diagonal da matriz original, contenham as informações dos pares de indivíduos de uma mesma população, enquanto aqueles fora da diagonal representam pares de indivíduos de populações diferentes. As somas dos elementos das matrizes diagonais proporcionam os totais das diversidades para os vários níveis hierárquicos da população.

Essas diversidades podem ser analisadas dentro de um esquema hierárquico, permitindo realizar testes de hipótese para avaliação da diversidade entre e dentro de grupos (EXCOFFIER, et al., 1992; EXCOFFIER, 2001), conforme o esquema:

FV	GL	SQ	QM	E(QM)
Regiões (R)	r-1	SQR	QMR	$\sigma_i^2 + k_2 \sigma_p^2 + k_3 \sigma_r^2$
Populações/R	$\sum_{i=1}^r s_i - r$	SQE	QME	$\sigma_i^2 + k_1 \sigma_p^2$
Indivíduos/P/R	$N - \sum_{i=1}^r s_i$	SQD	QMD	σ_i^2
Total	N-1	SQT		

Estimação de somas de quadrados

- a) Soma de quadrados total

$$SQT = \frac{1}{2N} T$$

Neste exemplo, tem-se:

$$SQT = \frac{1}{28} [(d_{12}^2 + \dots + d_{1.14}^2) + (d_{22}^2 + \dots + d_{2.14}^2) + \dots + (d_{14.1}^2 + \dots + d_{14.13}^2)]$$

- b) Soma de quadrados dentro de populações ou entre indivíduos dentro de populações
e regiões

$$SQD = \sum_{i=1}^r \sum_{j=1}^{s_i} \frac{S_{ij}}{2N_{ij}}$$

Neste caso, tem-se:

$$r : \text{número de regiões } (r = 2)$$

s_r : número de subpopulações por região ($s_1 = 2$ e $s_2 = 3$)

N_{ij} : número de indivíduos na subpopulação i da região j

em que:

$$N_{11} = 2 \quad N_{21} = 3 \quad N_{12} = 2 \quad N_{22} = 3 \quad N_{32} = 4$$

$$N = 14$$

logo,

$$SQD = \frac{d_{12}^2 + d_{21}^2}{2(2)} + \frac{d_{34}^2 + d_{35}^2 + d_{43}^2 + d_{45}^2 + d_{53}^2 + d_{54}^2}{2(3)} + \dots + \frac{d_{11.12}^2 + d_{14.13}^2}{2(4)}$$

c) Soma de quadrados entre populações/regiões

$$SQE = \sum_{i=1}^r \frac{R_i}{2N_i} - SQD$$

Neste exemplo, tem-se:

$$SQE = \frac{d_{12}^2 + d_{13}^2 + \dots + d_{53}^2 + d_{54}^2}{2(6)} + \dots + \frac{d_{67}^2 + d_{68}^2 + \dots + d_{14.12}^2 + d_{14.13}^2}{2(9)}$$

d) Soma de quadrados entre regiões

É dada pela diferença:

$$SQR = SQT - (SQE + SQD)$$

Estimação dos componentes de variância

Para estimação dos componentes de variância, é necessário estabelecer os valores dos coeficientes k_1 , k_2 e k_3 , dados por:

O coeficiente n_1 é fornecido por:

$$k_1 = \frac{N - \alpha_1}{\sum_{i=1}^r s_i}$$

sendo:

$$\alpha_1 = \sum_{i=1}^r \left(\frac{\sum_{j=1}^{s_i} N_{ij}^2}{\sum_{j=1}^{s_i} N_{ij}} \right)$$

No exemplo, tem-se:

Região	População	N_{ij}	N_{ij}^2
1	1	2	4
	2	3	9
Total		5	13
2	1	2	4
	2	3	9
	3	4	16
Total		9	29
Total geral		14	42

logo:

$$\alpha_1 = \frac{13}{5} + \frac{29}{9} = 5,8222$$

então:

$$k_1 = \frac{N - \alpha_1}{r} = \frac{14 - 5,8222}{5}$$

O coeficiente k_2 é dado por:

$$k_2 = \frac{\alpha_1 - \alpha_2}{r - 1}$$

sendo,

$$\alpha_2 = \frac{\sum_{i=1}^r \sum_{j=1}^{s_i} N_{ij}^2}{N}$$

No exemplo, tem-se:

$$\alpha_2 = \frac{42}{14} = 3$$

$$\text{e } k_2 = \frac{\alpha_1 - \alpha_2}{r - 1} = \frac{5,8222 - 3}{1} = 2,8222$$

O coeficiente n_3 é dado por:

$$k_3 = \frac{N - \alpha_3}{r - 1}$$

sendo:

$$\alpha_3 = \frac{\sum_{i=1}^r \left(\sum_{j=1}^{s_i} N_{ij} \right)^2}{N}$$

No exemplo, tem-se:

$$\alpha_3 = \frac{5^2 + 9^2}{14} = 7,5714$$

e

$$k_3 = \frac{N - \alpha_3}{r - 1} = \frac{14 - 7,5714}{1} = 6,4286$$

É interessante observar que, se o estudo fosse realizado com r regiões, cada uma com g populações, todas de tamanho igual a n , ter-se-ia:

$$k_1 = \frac{N(g-1)}{g}$$

$$k_2 = N$$

e

$$k_3 = Ng$$

Assim, os componentes de variância podem ser obtidos por meio das expressões:

$$\hat{\sigma}_i^2 = QMD$$

$$\hat{\sigma}_p^2 = \frac{QME - QMD}{k_1}$$

e

$$\hat{\sigma}_r^2 = \frac{k_1 QMR - k_2 QME - (k_1 - k_2) QMD}{k_1 k_3}$$

Associação com estatísticas Φ

Os componentes de variância podem ser usados para calcular uma série de estatísticas – chamadas de estatísticas phi (Φ) – que resumem o grau de diferenciação entre divisões de população e são análogas às estatísticas F. As estatísticas Φ podem ser obtidas como segue (EXCOFFIER, et al., 1992; EXCOFFIER, 2001):

Nível hierárquico da população	Estatística Φ
Entre subpopulações dentro da região	$\Phi_{SG} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_i^2 + \hat{\sigma}_p^2}$
Entre regiões	$\Phi_{GT} = \frac{\hat{\sigma}_r^2}{\hat{\sigma}_T^2}$
Entre subpopulações independente da região	$\Phi_{ST} = \frac{\hat{\sigma}_r^2 + \hat{\sigma}_p^2}{\hat{\sigma}_T^2}$

$$\hat{\sigma}_T^2 = \hat{\sigma}_r^2 + \hat{\sigma}_p^2 + \hat{\sigma}_i^2$$

As estatísticas Φ são originadas de correlações intraclass, considerando o modelo:

$$Y_{ijk} = \mu + R_i + S/R_{ij} + I/S/R_{ijk}$$

em que:

Y_{ijk} : diversidade entre pares de indivíduos k, na subpopulação j e região k;

μ : constante;

R_i : diversidade entre regiões;

S/R_{ij} : diversidade entre subpopulação de uma região; e

$I/S/R_{ijk}$: diversidade entre pares de indivíduos k, na subpopulação j da região i.

Assim:

$$\Phi_{GT} = \frac{Cov(Y_{ijk}, Y_{ijk'})}{\sqrt{V(Y_{ijk})V(Y_{ijk'})}} = \frac{\hat{\sigma}_r^2}{\hat{\sigma}_T^2}$$

$$\Phi_{ST} = \frac{Cov(Y_{ijk}, Y_{ijk'})}{\sqrt{V(Y_{ijk})V(Y_{ijk'})}} = \frac{\hat{\sigma}_r^2 + \hat{\sigma}_p^2}{\hat{\sigma}_T^2}$$

e

$$\Phi_{SG} = \frac{Cov(y_{jk}, y_{jk'})}{\sqrt{V(y_{jk})V(y_{jk'})}} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_i^2 + \hat{\sigma}_p^2}$$

sendo y_{jk} os valores de dissimilaridade dentro de uma particular região x, ou seja:

$$y_{jk} = Y_{xijk} = \mu + S_j + I/S_{jk}$$

Deve ser ressaltado que os dados sob análise consistem em distâncias euclidianas derivadas de vetores de uns e zeros, sendo, portanto, improvável seguirem uma distribuição normal. Uma distribuição nula é computada então por reamostragem dos dados (EXCOFFIER, et al., 1992). Em cada permutação, cada indivíduo é alocado aleatoriamente dentro da população, mantendo-a de tamanho constante. Essas permutações são repetidas muitas vezes, permitindo a construção de uma distribuição nula, bem como comparar o grau de significância da estatística obtida.

Considerando que as distribuições nulas são obtidas através de reamostragem, não é preciso assumir que a distância euclidiana entre haplótipos seja normalmente distribuída ou que apresente homogeneidade de variância entre as subpopulações.

Devem ser feitas certas suposições sobre a natureza da população (EXCOFFIER, et al., 1992), por exemplo, que o acasalamento é completamente ao acaso e que não ocorre endogamia. Caso contrário, o nível de heterozigose será inferior e a variabilidade poderá não ser convenientemente estimada. Em razão da deriva genética, fica difícil assumir que qualquer um haplótipo seja representativo da variação completa do genoma. Também é importante que os dados sejam derivados de um número adequado de marcadores ou pares de base.

Os efeitos de seleção também não são completamente considerados nesta metodologia de análise da diversidade. Se as subpopulações estão sob diferentes pressões seletivas, a seleção poderá ter efeitos muito diferentes sobre os alelos e sobre as combinações genotípicas.

A AMOVA parece ser uma metodologia robusta para examinar as distâncias entre haplótipos. Excoffier et al. (1992) examinaram os resultados desta metodologia aplicáveis às matrizes de distâncias obtidas por várias medidas de dissimilaridade e concluíram a respeito da consistência das informações obtidas.

Teste de Hipótese

Uma estatística Φ pode ser tratada como instrumento de avaliação de uma hipótese sobre o nível de diferenciação de uma população; por exemplo, a

estatística Φ associada a ST pode ser tratada como ferramenta para testar a hipótese sobre diferenciação da população em seus componentes de subpopulações. As hipóteses que usam a distribuição nula dos componentes de variação podem ser testadas; se a variação das subpopulações não difere significativamente da distribuição nula da variação da população, a hipótese de que essas subpopulações são diferenciadas da população original seria rejeitada.

A permutação é feita diretamente na matriz de distância (D^2), alterando-se as posições dos elementos de linhas e das colunas correspondentes. A permutação pode se dar em toda a matriz, sem distinção de subespaço de regiões ou de subpopulações, ou admitir que a subdivisão em regiões e, ou, em subpopulações é verdadeira. Nesses casos, a permutação deve ocorrer dentro dos subespaços apropriados da matriz.

Aplicação

Como ilustração, será considerada a avaliação de 14 indivíduos em estrutura caracterizada por dois níveis de subdivisão, em relação a 10 marcadores moleculares. O primeiro refere-se a regiões, tendo-se no exemplo cinco indivíduos na região 1 e nove na região 2. Uma segunda divisão ocorre dentro de cada região. Para o exemplo em consideração, tem-se, dentro da região 1, duas subpopulações com 2 e 3 indivíduos cada. Dentro da região 2, há três subpopulações com dois, três e quatro indivíduos cada. Os dados encontram-se descritos a seguir:

Reg	Pop	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
1	1	0	1	0	1	0	1	1	0	1	0
1	1	1	0	1	1	0	0	1	1	1	1
1	2	1	1	1	0	1	1	0	0	1	1
1	2	1	1	0	0	1	0	1	1	0	0
1	2	1	0	1	0	1	1	0	1	1	1
2	3	0	0	0	0	0	0	0	1	1	0
2	3	1	1	0	0	1	0	1	0	0	0
2	4	0	0	0	1	0	0	0	1	0	0
2	4	0	1	0	0	1	0	0	0	1	0
2	4	0	0	0	1	1	1	1	0	1	0
2	5	1	0	0	0	0	0	0	1	0	1
2	5	1	1	0	1	0	0	1	1	0	1
2	5	1	1	0	0	1	0	1	0	0	0
2	5	1	0	0	0	1	0	0	0	1	1

O resultado da AMOVA é apresentado a seguir:

FV	GL	SQ	QM	E(QM)
Regiões (R)	1	3,0063	3,0063	$\sigma_i^2 + k_2\sigma_p^2 + k_3\sigma_r^2$
Populações/R	3	7,8389	2,6130	$\sigma_i^2 + k_1\sigma_p^2$
Indivíduos/P/R	9	21,5833	2,3981	σ_i^2
Total	13	32,4286	2,4945	

Os valores das constantes que multiplicam os componentes de variância são:

$$k_1 = 1,6356$$

$$k_2 = 2,8222$$

$$k_3 = 6,4286$$

As estimativas dos componentes de variância são, então, dadas por:

$$\hat{\sigma}_i^2 = QMD = 2,3981 \text{ (representa } 1,44\% \text{ da variação total)}$$

$$\hat{\sigma}_p^2 = \frac{QME - QMD}{k_1} = 0,1313 \text{ (representa } 5,12\% \text{ da variação total)}$$

$$\hat{\sigma}_r^2 = \frac{k_1 QMR - k_2 QME - (k_1 - k_2) QMD}{k_1 k_3} = 0,0369 \text{ (representa } 93,44\% \text{ da variação total)}$$

$$\hat{\sigma}_T^2 = \hat{\sigma}_i^2 + \hat{\sigma}_p^2 + \hat{\sigma}_r^2 = 2,5663$$

Por fim:

$$\Phi_{GT} = \frac{\hat{\sigma}_r^2}{\hat{\sigma}_T^2} = 0,0144$$

$$\Phi_{ST} = \frac{\hat{\sigma}_r^2 + \hat{\sigma}_p^2}{\hat{\sigma}_T^2} = 0,0656$$

$$\Phi_{SG} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_i^2 + \hat{\sigma}_p^2} = 0,0519$$

Com os testes de permutações, são obtidos os limites de significância a 10, 5% e 1%, conforme descrito a seguir:

	LS(10%)	LI(10%)	LS(5%)	LI(5%)	LS(1%)	LI(1%)	Sig
$\hat{\sigma}_r^2$	-0,3376	0,8632	-0,4186	0,9596	-0,6120	1,1035	ns
$\hat{\sigma}_p^2$	-2,4117	-0,0476	-2,7151	0,1200	-3,1046	0,5593	*
$\hat{\sigma}_i^2$	2,3472	3,4907	2,2083	3,6250	2,0833	3,787	ns
Φ_{ST}	-1,2430	-0,0416	-1,4906	0,0377	-1,8048	0,1600	*
Φ_{SG}	-2,3835	-0,0200	-3,4487	0,0477	-5,5893	0,2041	*
Φ_{GT}	-0,1707	0,4564	-0,2406	0,5369	-0,3637	0,6428	ns

6.8 Emprego das Medidas de Diversidade

Grande parte dos trabalhos envolvendo estudos de estrutura populacional diversidade genética de populações naturais, baseados na análise de marcadores moleculares, utilizam metodologias baseadas nas estatísticas H e F. Alguns exemplos ilustrativos dos resultados obtidos serão comentados a seguir.

Mota (2003) fez um estudo de caracterização, por meio da análise de 30 locos de microsatélites, de 172 clones de cacau (*Theobroma cacao* L.), provenientes de coletas de populações espontâneas de 16 bacias hidrográficas da Amazônia brasileira. O autor obteve 29 locos polimórficos, considerados seletivamente neutros apesar de estar sob seleção natural, com um total de 187 alelos, sendo 68 considerados raros, dentre os quais 23 foram exclusivos. A diversidade genética total (H_T) foi de 0,658, contra uma heterozigosidade observada (H_o) de 0,295 e coeficiente de endogamia de 0,408. O número médio de alelos efetivos por loco foi igual a 3,4, e a riqueza alélica variou de 2,273 a 3,595 alelos por subpopulação, para uma amostra padronizada de 16 genes (= ao número de indivíduos da menor amostra). As subpopulações apresentaram-se fortemente estruturadas, com significativa diferenciação ($G_{ST} = 0,292$ e $F_{ST} = 0,290$). A análise das amostras agrupadas, por quatro regiões geográficas principais, mostrou que 66,6% da variabilidade genética se encontra dentro das subpopulações, 14,86% entre subpopulações dentro dos quatro grupos e 18,5% entre os quatro grupos.

As estatísticas de diversidade genética de Nei foram utilizadas por Mengistu et al. (2000) na avaliação da diversidade da gramínea invasora *Poa annua* L. no oeste do estado do Oregon. Os autores utilizaram 18 marcadores de RAPD e avaliaram 1.357 indivíduos de 47 populações, coletadas em 16 sítios diferentes e em três estações do ano (outono, inverno e primavera). Pelo fato de os marcadores utilizados serem dominantes, esses autores assumiram, nas análises de estrutura populacional, que todas as populações se encontravam em equilíbrio de Hardy-Weinberg e utilizaram um coeficiente de endogamia, de dados da literatura sobre a

espécie, igual a $F_{IS}=0,64$. Os resultados indicaram que as populações de *P. annua* apresentam elevada diversidade, com uma média do índice h de Nei igual a 0,241 e uma diversidade total (H_T) de 0,245. Por outro lado, grande proporção da diversidade total se deveu mais à diversidade dentro das populações ($H_S = 0,209$) do que à diversidade entre populações ($G_{ST} = 0,146$); essa condição foi ainda mais acentuada quando as populações foram agrupadas por localidade ou estações do ano. Populações coletadas em sítios com histórico de elevada pressão de seleção pelo emprego de herbicidas mostraram menor diferenciação entre as datas de coleta, bem como menor diversidade, com G_{ST} tão baixo quanto 0,016 e $h = 0,155$, respectivamente, ao passo que aquelas coletadas em campos de baixa pressão de seleção por herbicidas apresentaram valores de G_{ST} tão elevados quanto 0,125 e de h iguais a 0,155.

Sebbenn et al. (2000) analisaram os efeitos do manejo na variabilidade genética intrapopulacional de *Tabebuia cassinoides* (Bignoniaceae), através do estudo de 13 locos enzimáticos, em uma população manejada e uma população natural não manejada. Esses autores observaram que o índice médio de fixação de alelos dentro das populações (F_{IS}) e para o conjunto de populações (F_{IT}) apresentou valores positivos, altos e significativos (0,259 e 0,282), sugerindo alta endogamia. A divergência entre populações (F_{ST}) foi baixa para a média dos locos (0,031), revelando que a maior parte da variabilidade genética (97%) é de natureza intrapopulacional. A análise dentro das populações mostrou que o manejo florestal utilizado levou a perda de alelos raros, redução na heterozigosidade (H_o), diversidade genética (H_e) e porcentagem de locos polimórficos, aumento no índice de fixação (f) e tendência de aumento da taxa de fecundação.

A estrutura genética e o fluxo gênico em 10 populações de cagaita (*Eugenia dysenterica* DC) foram avaliados por Zucchi et al. (2003), através do uso de 10 pares de *primers* de marcadores microssatélites e uma amostra total de 116 indivíduos. Sete locos foram polimórficos, com uma média de 10,4 alelos por loco, heterozigosidade média observada (H_o) de 0,458 e esperada (H_e) de 0,442. O índice médio de fixação (F_{IS}) foi igual a -0,037, indicando que a espécie possui taxa de

polinização cruzada compatível com a alogamia. A divergência genética entre populações (F_{ST}) foi igual a 0,250. Esse alto grau de divergência interpopulacional associado a um grande número de alelos exclusivos é, segundo os autores, indicativo de redução no fluxo gênico entre as populações, com possível conseqüência negativa para a estrutura de metapopulação.

6.9. - Emprego das Medidas de Distância Genética

Takezaki e Nei (1996), por meio de simulações, avaliaram diferentes medidas de distâncias genéticas, aplicadas a dados de locos de microssatélites, quanto à probabilidade de cada uma obter a topologia correta (T_c) da árvore filogenética. Foram considerados os dois modelos de mutação: o de alelos infinitos (IAM) e o de mutações em etapa ou *stepwise* (SMM). Os resultados mostraram que, em ambos os modelos (IAM e SMM), os métodos da “corda” de Cavalli-Sforza e a distância de Nei et al. (1983), geralmente, apresentaram valores mais elevados de T_c que outras medidas de distância, com ou sem efeito de afunilamento genético. Por outro lado, a distância D de Nei (1972) e a de Goldstein et al. $[(\delta\mu)^2]$ foram mais apropriadas para estimar o tempo de divergência evolutiva.

Mota (2003) utilizou a metodologia de Nei (1972) para estimar as distâncias genéticas entre 172 clones de cacau (*Theobroma cacao L.*), provenientes de coletas de populações espontâneas de 16 bacias hidrográficas da Amazônia brasileira, por meio da análise de 30 locos de microssatélites, com um total de 187 alelos. Segundo o autor, o dendrograma construído com as distâncias genéticas D de Nei apresentou topologia mais robusta que os obtidos por metodologias alternativas (distância de Reynolds e distância de Goldstein).

Zucchi et al. (2003), trabalhando com sete locos microssatélites, empregaram a distância de Nei na avaliação da distância genética e da correlação entre distância genética e espacial, de 10 populações de *Eugenia dysenterica* do cerrado goiano. A consistência do agrupamento, produzido pelo método UPGMA, foi avaliada pelo método *bootstrap*, com 10.000 reamostragens, e o padrão de variação espacial foi

analisado através do coeficiente de correlação de Pearson. Os autores obtiveram um agrupamento com elevado coeficiente de correlação co-fenética (0,943), assim como um coeficiente de correlação elevado e positivo ($r=0,872$ $p<0,001$) entre distância genética e geográfica. Esses resultados indicaram que o padrão de variabilidade genética entre populações é estruturado no espaço, confirmando dados obtidos anteriormente, para as mesmas populações, a partir de marcadores isoenzimáticos.

A metodologia de Nei (1972) também foi utilizada por Rosseto et al. (1999) na análise da distância genética de 500 indivíduos de 40 populações da planta de chá *Melaleuca alternifolia*, de ocorrência nativa da Austrália. Os autores analisaram cinco locos de microssatélite, com um total de 98 alelos úteis no estudo de genética de populações. Foram testadas quatro metodologias de agrupamento, sendo concordantes entre si os métodos das distâncias D de Nei, o coeficiente de co-ancestralidade de Reynolds e a “corda” de Cavalli-Sforza. Os filogramas produzidos por estes métodos também foram coerentes com a distribuição geográfica das populações.

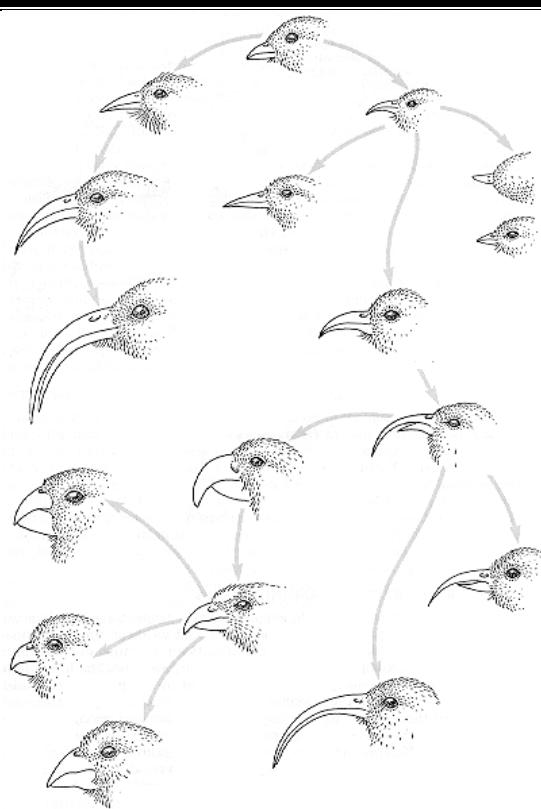
6.10. Técnicas de agrupamento a partir de índices de fixação

O uso de métodos de agrupamento visa facilitar o reconhecimento de grupos homogêneos pelo simples exame visual das estimativas de dissimilaridade. Todas as técnicas de análise de agrupamento, já abordadas neste livro, podem ser utilizadas para estudar o padrão de agrupamento das populações. A matriz de dissimilaridade pode ser estabelecida pelas diferentes estatísticas de distâncias baseadas em freqüências alélicas ou genotípicas. Também podem ser utilizados coeficientes que expressam a diferenciação entre as populações, como G_{ST} , de Nei(1977), F_{ST} de Wright ou Φ_{ST} de Excoffier et al (1992).

Além das análises de agrupamentos, é possível fazer as projeções das medidas de distância ou de diferenciação das populações no plano ou espaço 3D pelos processos também já descritos neste livro.

Capítulo 7

Análise Filogenética Molecular



7.1 Introdução

A filogenética tem por objetivo o estudo das relações de ancestralidade entre organismos por meio da utilização de dados fenotípicos ou moleculares, como seqüências de DNA e de proteínas, ou ainda de outros marcadores moleculares. A análise é de fundamental importância para compreender a história evolucionária das espécies ou de quaisquer outras entidades biológicas, possibilitando a reconstrução dos laços genealógicos corretos que as unem e a realização de inferências a respeito do tempo de divergência entre elas (isto é, o tempo desde quando tais entidades compartilharam um ancestral comum), assim como o estabelecimento da cronologia de seqüências de eventos das diferentes linhagens evolutivas.

O rápido avanço da biologia molecular, no campo da sistemática filogenética, foi encarado com grande dose de ceticismo por parte dos biólogos sistematas, com treinamento em morfologia comparada e biologia comportamental. Ao mesmo tempo, muitos dos primeiros biólogos moleculares eram mal informados em relação a áreas como ecologia e evolução. Esse quadro proporcionou boa dose de antagonismo entre os dois grupos, com persistência até em anos recentes. Por outro lado, nos últimos anos, uma atitude mais compreensiva emergiu com o reconhecimento de que dados moleculares e de organismos completos podem ser reciprocamente informativos e, de fato, requer os conhecimentos um do outro (GRAUR e LI, 2000; AVISE, 2004).

De acordo com Avise (2004), os avanços no campo dos marcadores moleculares se processam na forma de uma série de “ondas”, cada qual iniciada pelo desenvolvimento de um novo método laboratorial. Esse padrão é caracterizado pela introdução de uma técnica para acessar as informações contidas no DNA ou nas proteínas; a essa introdução se segue um sem número de ensaios de avaliação. Métodos ineficientes (seja por dificuldades técnicas, baixa reproduzibilidade ou ambigüidades na interpretação genética) são abandonados, enquanto aqueles que sobrevivem ao processo de avaliação inicial têm seus resultados confrontados com os paradigmas da evolução molecular. Por exemplo,

qual seria o papel da seleção natural na manutenção da variabilidade molecular revelada? Simultaneamente, esses novos marcadores genéticos são amplamente empregados na solução de problemas de história natural ou de evolução, nas situações em que parecem apropriados. Depois de alguns anos, entretanto, o interesse por tais marcadores decresce e uma nova onda de prestígio é direcionada a um procedimento metodológico recém-introduzido. Usualmente, os métodos antigos bem sucedidos não são abandonados, mas meramente incorporados ao conjunto crescente de técnicas moleculares que continuam a ser aplicadas em estudos da biologia de organismo, história natural e evolução.

Existem várias razões pelas quais dados moleculares, particularmente seqüências de DNA e aminoácidos, são mais apropriados para estudos evolutivos do que os dados morfológicos e fisiológicos, com destaque para:

- a) Seqüências de DNA e proteínas são entidades estritamente herdáveis, diferentemente de dados morfológicos, que são muito influenciados pelo ambiente.
- b) A descrição dos caracteres moleculares e dos respectivos estados de caráter não são ambíguas.
- c) Caracteres moleculares evoluem de maneira muito mais regular do que caracteres morfológicos e fisiológicos.
- d) Caracteres moleculares são muito mais apropriados para tratamentos quantitativos que dados morfológicos. De fato, teorias estatísticas e matemáticas sofisticadas foram desenvolvidas para a análise quantitativa de dados de seqüências de DNA.
- e) o estabelecimento de homologias é mais fácil com dados moleculares do que entre caracteres morfológicos;
- f) Alguns dados moleculares podem ser utilizados para acessar relações evolutivas entre organismos filogeneticamente muito distantes.
- g) Dados moleculares são muito mais abundantes que dados morfológicos. Essa abundância é especialmente útil quando se trabalha com organismos como algas, fungos e bactérias, que possuem limitado número de caracteres morfológicos que podem ser utilizados em estudos filogenéticos.

7.2 Definições e Terminologias Empregadas na Análise Filogenética

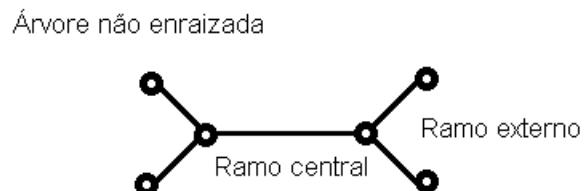
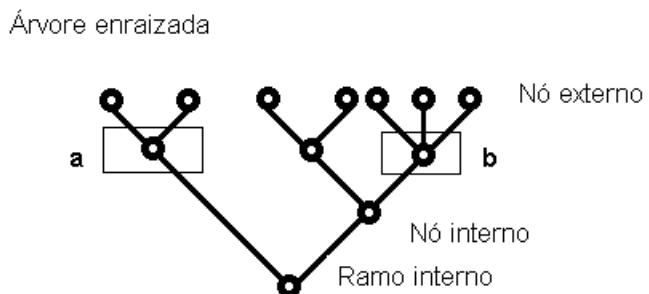
A seguir são apresentadas definições de alguns termos mais rotineiramente utilizados em estudos filogenéticos.

Relativas à representação gráfica ou dendrograma

Uma **árvore filogenética** (ou dendrograma) é a representação gráfica das inter-relações filogenéticas entre um grupo de organismos ou unidades taxonômicas. Esse gráfico é composto por nós e ramos, de modo que somente um ramo conecta dois nós adjacentes. Os **nós** representam as unidades taxonômicas, que podem ser referentes a espécies (ou táxons de níveis superiores), populações, indivíduos ou genes, enquanto os **ramos** representam relações de parentesco entre elas, em termos de ancestrais e descendentes. O padrão de ramificação do dendrograma é chamado de **topologia**.

Os nós de uma árvore são distinguidos em internos e terminais. Os **nós terminais** representam as unidades taxonômicas operacionais (**OTUs**), ou seja, as unidades reais que estão sendo comparadas, enquanto os **nós internos** representam as unidades taxonômicas hipotéticas (**HTUs**), ou seja, unidades taxonômicas ancestrais inferidas. Os nós internos geralmente são **bifurcantes**, sendo conectado por um ramo ancestral e dois derivados, assumindo que o processo de especiação é binário, isto é, cada evento de especiação resulta na formação de não mais que duas espécies a partir de um único estoque ancestral. Entretanto, **nós multifurcantes**, ou seja, com mais de duas linhagens descendentes, também podem ocorrer. A multiforquilha **ou politomia** de árvore pode ser explicada uma das seguintes razões: ela representa uma seqüência verdadeira de eventos, na qual um táxon ancestral originou simultaneamente mais de dois táxons descendentes, ou ela representa uma situação na qual a ordem exata de dois ou mais processos binários de especiação não puderam ser estabelecidos sem ambigüidades.

Ilustração de alguns componentes de uma árvore filogenética é apresentada a seguir:



a: Dicotomia (três ramos e um nó interno)

b: Politomia (mais de três ramos e um nó interno)

As relações de parentesco entre os táxons ou OTUs, expressas pela árvore filogenética, permitem definir **grupos monofiléticos** ou **clados**, ou seja, grupos incluindo um ancestral e todos os seus descendentes. Um dos principais objetivos da análise filogenética é a identificação de clados naturais.

As árvores filogenéticas tanto podem ser enraizadas ou não-enraizadas, em escala ou sem escala. **Árvore enraizada** é aquela que apresenta um nó particular, a raiz, do qual um único caminho leva a qualquer outro nó. A direção de cada caminho corresponde ao tempo evolutivo, e a raiz, ao ancestral comum mais recente de todas as unidades taxonômicas sob investigação. Já uma **árvore não-enraizada** é aquela que especifica somente o grau de relacionamento entre as unidades taxonômicas, porém não define o caminho evolutivo. Por isso, estritamente falando, uma árvore não-enraizada não pode ser considerada uma árvore filogenética, uma vez que o ancestral comum não é especificado. A **árvore em escala** apresenta comprimentos de ramos proporcionais ao número de mudanças evolutivas (como, por exemplo, ao número de substituições de nucleotídeos) que ocorreram ao longo de cada ramo. Já na **árvore sem escala** ou

ultramétrica, os comprimentos dos ramos não são proporcionais ao número de mudanças evolutivas. Árvores com essa simetria são geradas sob o pressuposto de taxas evolutivas constantes para as diferentes OTUs.

O número de topologias possíveis, estabelecidas a partir de m táxons, é dado por:

$$n_t = 1 \times 3 \times 5 \times \dots \times (2m-3), \text{ para árvores enraizadas; e}$$

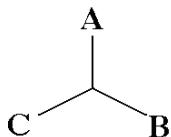
$$n_t = 1 \times 3 \times 5 \times \dots \times (2m-5), \text{ para árvores não-enraizadas.}$$

Assim, tem-se:

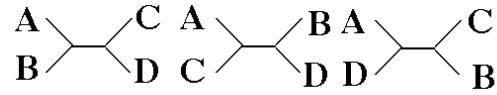
Número de táxons (m)	Número de topologias (n_t)	
	Nº de árvores enraizadas	Nº de árvores não-enraizadas
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
...
10	34.459.425	2.027.025

Ilustrações de árvores não-enraizadas com três, quatro e cinco taxas são apresentadas a seguir:

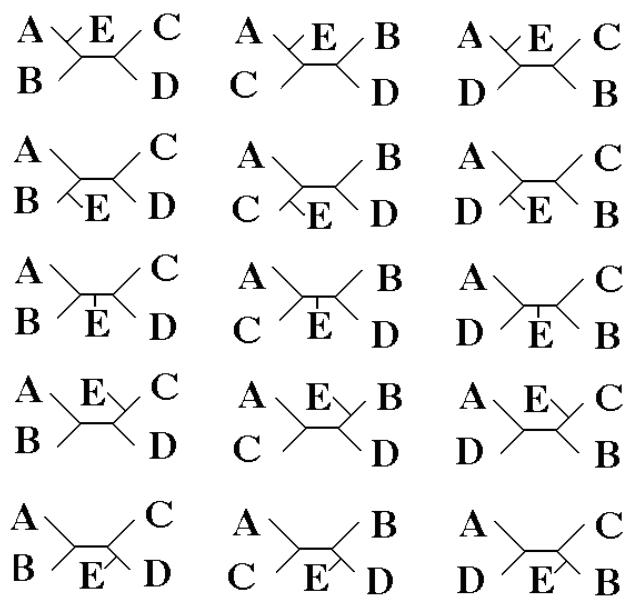
Três taxas:



Quatro taxas:

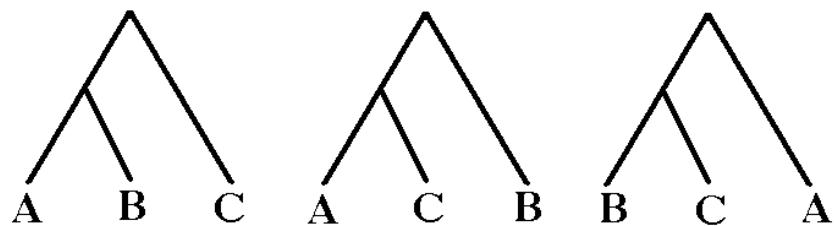


Cinco taxas:

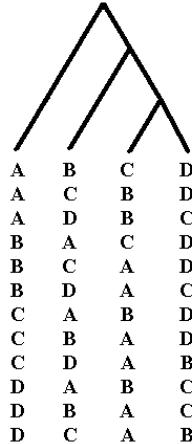
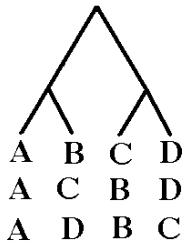


Ilustrações de árvores enraizadas com três e quatro taxas são apresentadas a seguir:

Três taxas



Quatro taxas



Para muitos estudos, entretanto, a maioria das topologias poderá ser excluída de análises em razão de serem estabelecidas a partir de um padrão pouco provável evolutivamente ou por qualquer razão ou importância biológica adicional. Por outro lado, mesmo excluindo alguns padrões, o número de topologias a ser considerado nas análises ainda é relativamente grande, quando o número de táxons for elevado.

O número de ramos e de nós também difere entre árvores com ou sem raiz.

Para árvores sem raízes, têm-se:

Número de ramos: $2m - 3$

Número de ramos externos: m

Número de ramos internos: $m - 3$

Número de nós internos: $m - 2$

Para árvores com raízes, têm-se:

Número de ramos: $2m - 2$

Número de ramos externos: m

Número de ramos internos: $m - 2$

Número de nós internos: $m - 1$

Um aspecto importante na construção de árvores em escala refere-se à distinção entre árvores de espécies e árvores de genes. Em uma **árvore de**

espécies, uma bifurcação representa o momento em que houve especiação, isto é, o momento a partir do qual duas espécies se tornaram distintas e, reprodutivamente, isoladas entre si. Uma **árvore de genes** pode diferir de uma árvore de espécies em dois aspectos: primeiro, porque a divergência de dois genes amostrados de duas espécies distintas pode ser anterior à divergência entre as duas espécies – isso resultará em uma superestimação do comprimento dos ramos; e, segundo, porque o padrão de ramificação da árvore de genes pode ter topologia diferente do padrão de ramificação da árvore de espécies. A razão dessa diferença é o polimorfismo genético da espécie ancestral (NEI; KUMAR, 2000).

É importante também fazer a distinção entre árvore verdadeira e árvore inferida. A **árvore verdadeira** é aquela que representa a verdadeira história evolutiva de um dado grupo de OTUs ou clado. Uma vez que os eventos de especiação que levam à formação de qualquer grupo de OTUs são, historicamente, únicos, pode-se dizer que existe apenas uma árvore verdadeira para este grupo. A **árvore inferida** é aquela obtida a partir da análise de determinado grupo de dados, por meio de determinado método de reconstrução de árvore. Esta árvore pode ou não ser idêntica à árvore verdadeira (NEI; KUMAR, 2000).

Relativas aos tipos de dados utilizados

Dados moleculares empregados em reconstruções filogenéticas pertencem a uma de duas categorias: caracteres ou distâncias. Um caráter fornece uma informação a respeito de uma OTU individual. Uma distância representa um valor quantitativo de dissimilaridade entre duas OTUs.

Um **caráter** é uma característica bem definida que, em uma OTU, assume um de dois ou mais estados de caráter mutuamente exclusivos. Em outras palavras, um caráter é uma variável independente, podendo ser tanto qualitativa quanto quantitativa. Um **estado de caráter** de uma variável quantitativa é usualmente contínuo, sendo medido por uma escala de intervalos. O estado de caráter de uma variável qualitativa é discreto. Um caráter será **binário** se assumir somente dois estados de caráter. Um caráter molecular pode ser denominado de **não ordenado**

quando a mudança de um estado a outro ocorre em um único passo – por exemplo, a mudança de um nucleotídeo em uma seqüência de DNA. Um caráter **parcialmente ordenado** é aquele em que número de passos de mudanças de estado varia para diferentes pares de combinação de estados, mas não há relação definida entre o número de passos e o número de estado de caráter – por exemplo, mudanças nas seqüências de aminoácidos. Um aminoácido não pode mudar para qualquer um dos demais aminoácidos em um único passo. Já um **caráter ordenado** é aquele em que o número de passos de um estado a outro é igual, em valor absoluto, entre seus respectivos número de estados – por exemplo, a mudança de estado de 1 para 3 é assumida ocorrer através de dois passos, ou seja, de 1 para 2 e de 2 para 3. A mudança recíproca também só poderá ocorrer através desses mesmos dois passos, na ordem e sentido inversos.

Os estados de caracteres homólogos podem também ser ordenados em termos de antiguidade temporal. Assim, um **caráter plesiomórfico** representa o estado mais primitivo ou ancestral do caráter entre as OTUs consideradas, enquanto um **caráter apomórfico** refere-se ao estado de caráter derivado, representando uma novidade evolutiva. Um estado primitivo compartilhado por vários táxons é denominado de **simplesiomórfico**, e um estado de caráter derivado, compartilhado por um grupo de táxons, é denominado de **sinapomórfico**.

Dados de distâncias envolvem pares de OTUs. Nesse caso, as distâncias evolutivas são computadas entre todos os pares de táxons, e uma árvore filogenética é construída levando-se em conta as interrelações entre estes valores de distância. Dados de distância não podem ser convertidos em dados de caracteres. Por outro lado, dados de caracteres podem ser convertidos em dados de distância. Três razões básicas justificam esse tipo de conversão:

- a) Uma longa lista de estados de caracteres, como seqüências de DNA, não possui significado intrínseco em um contexto evolutivo. Por outro lado, se esta lista for convertida em medidas de similaridades (distâncias) entre as OTUs, podem-se invocar inter-relações evolutivas entre elas.

- b) As múltiplas substituições, que ocorrem em um dado sítio, podem ser estimadas a partir de pressuposições razoáveis a respeito do processo evolutivo. Essas múltiplas substituições podem ser consideradas no estabelecimento de uma medida de distância, mas não na análise dos estados de caráter dessas seqüências.
- c) Existem numerosos métodos para inferir árvores evolutivas a partir de dados de distância. A maioria deles é rápida e eficiente, mesmo quando o número de OTUs é elevado.

Os dados de distância podem ser aditivos, ultramétricos ou nenhum dos dois. As distâncias são aditivas quando a divergência entre quaisquer duas OTUs é igual à soma dos comprimentos de todos os ramos que as conectam. No caso de dados ultramétricos, todas as OTUs são eqüidistantes da raiz.

Relativas à natureza dos caracteres

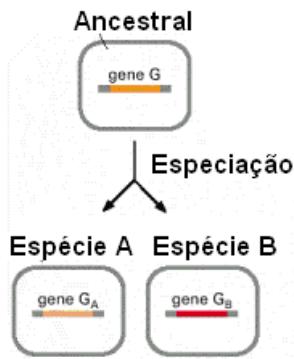
A sistemática molecular baseia-se no conceito de homologia de caracteres para descrever a história evolutiva dos táxons em questão. Diz-se que estruturas ou órgãos são homólogos se eles têm uma origem embrionária comum, não necessariamente apresentando a mesma função. A homologia entre estruturas de dois organismos diferentes sugere que eles se originaram de um grupo ancestral comum, embora não indique um grau de proximidade e, quando partem de várias linhas evolutivas que originaram várias espécies diferentes, fala-se que ocorrem **irradiação adaptativa**. Por exemplo, as patas dianteiras do jacaré, as asas dos morcegos, os braços humanos e as asas das aves exercem funções diferentes, mas apresentam uma mesma origem embrionária.

Estruturas que não apresentam uma origem embrionária comum, embora possuam semelhanças morfológicas e funcionais, são chamadas análogas, como as asas dos insetos e as asas dos morcegos. O fenômeno resultante do conjunto de processos (convergência, paralelismo e reversão) que leva à formação de duas estruturas análogas é denominado homoplasia.

As estruturas análogas não refletem por si sós qualquer grau de parentesco. Elas fornecem indícios da adaptação de estruturas de diferentes organismos a uma mesma variável ecológica. Quando organismos não intimamente aparentados apresentam estruturas semelhantes, exercendo a mesma função, diz-se que eles sofreram **evolução convergente**.

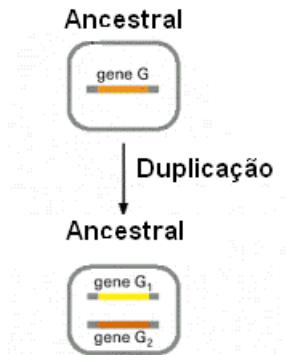
Embora esses termos tenham sido definidos inicialmente para caracteres morfológicos, eles também se aplicam a genes. Quando se está tratando de genes, o conceito de homologia torna-se um pouco mais complexo. Assim, quatro conceitos podem ser apresentados em relação a genes homólogos:

- a) Genes ortólogos são aqueles em que a homologia ocorreu em função da especiação e, portanto, apresentam um ancestral comum. Os genes mitocondriais, por exemplo, são considerados ortólogos, pois acredita-se que a origem da mitocôndria nos seres vivos tenha ocorrido em um estádio inicial da história da vida, antes da divergência entre os seres vivos atuais. Exemplo de genes ortólogos (G_A e G_B) é apresentado a seguir:

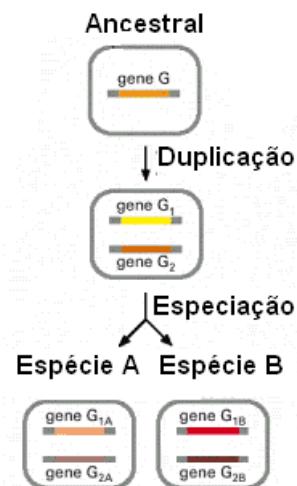


- b) Genes parálogos são aqueles em que a homologia deve-se à duplicação de um gene ancestral em uma mesma espécie. Em alguns casos, poucas modificações nucleotídicas são suficientes para conferir ao gene uma função diferente. A utilização de genes pertencentes a famílias gênicas (paralogia) dificulta a obtenção de seqüências de uma mesma região do genoma, uma vez que as cópias podem possuir diferentes pressões de seleção, ter diferentes histórias evolutivas e até

mesmo ocupar cromossomos diferentes. Exemplo de genes parálogo (G_1 e G_2) é apresentado a seguir:



De maneira geral, genes parálogos e ortólogos ocorrem simultaneamente, como ilustra a figura a seguir:



- c) Genes xenólogos ocorrem quando um gene de uma espécie é introduzido em outra espécie (transferência horizontal). Isso pode ocorrer, por exemplo, através de retrovírus, ou através da formação de híbridos férteis que cruzam com indivíduos de uma das espécies parentais.
- d) Genes plerólogos ocorrem por meio da interação entre exons e íntrons de um mesmo gene, como, por exemplo, através de embaralhamento de exons (“exon shuffling”) ou de evolução em concerto (“concerted evolution”). Genes parálogos e

plerólogos podem ser úteis para mostrar eventos de surgimento de novos genes ou famílias gênicas, mas não refletem a história evolutiva entre unidades taxonômicas.

Relativas aos grupos taxonômicos

Outro ponto importante da reconstrução filogenética é o conceito de monofilia. Quando se deseja inferir a relação de um dado grupo taxonômico, parte-se do pressuposto de que ele seja monofilético, ou seja, as OTUs analisadas apresentam um ancestral comum. Grupos monofiléticos por definição incluem aqueles táxons que apresentam sinapomorfias, isto é, caracteres derivados de um mesmo ancestral comum.

Grupos não-monofiléticos podem ser denominados parafiléticos ou polifiléticos (HENNIG, 1966). Parafilia ocorre quando o grupo de OTUs possui caracteres diagnósticos que são na verdade simplesiomorfias, isto é, é um conjunto de caracteres primitivos, e nem todos os descendentes do ancestral comum deste grupo estão incluídos no grupo interno.

Uma maneira de se testar a monofilia de um grupo de OTUs é por meio da inclusão de um grupo externo (MADDISON et al., 1984), isto é, um táxon sabidamente próximo, porém não pertencente ao grupo de análise – o grupo interno.

O uso do grupo externo surgiu na escola cladística, diferenciando-a da escola fenética. Esta por sua vez, agrupa os táxons simplesmente por semelhanças compartilhadas. Para a escola cladista, a escolha do grupo externo está baseada em usar um ou mais táxons que apresentem caracteres primitivos em relação aos táxons do grupo interno. O grupo externo será responsável pela polarização das modificações das características morfológicas. Na escola molecular, um táxon não pertencente ao grupo interno pode ser usado para enraizar a árvore filogenética, sem uma hipótese *a priori* de polarização de mudança de nucleotídeos. Isso é relevante nesse caso porque a reversão de um nucleotídeo para o seu estado ancestral pode ocorrer, e a mudança de um nucleotídeo para outro não pode ser

polarizada. Afinal, uma mudança de um nucleotídeo A para G, por exemplo, não requer a passagem pelos outros estados (C ou T).

Se vários táxons são usados como grupos externos, a monofilia do grupo interno pode ser constatada se nenhum dos táxons internos for separado dos demais por um ou mais táxons do grupo externo. Um procedimento que dispensa o uso de grupos externos é a construção de árvores não-enraizadas, o que consiste em demonstrar apenas as relações entre as OTUs sem, no entanto, fazer relações temporais de divergência e sem definir relações de ancestrais e descendentes. Dessa maneira, também não ocorre a polarização dos caracteres *a priori*.

De modo geral, a análise e reconhecimento de grupos monofiléticos em árvores não enraizadas é mais difícil do que em árvores enraizadas.

Alinhamento de Seqüências Moleculares

O alinhamento de seqüências biológicas é uma tarefa de extrema importância em Biologia Molecular, permitindo explorar o grau de similaridade entre estruturas de DNA, RNA ou proteínas. Um grau alto de similaridade indica alta probabilidade de as funções executadas pelas moléculas comparadas serem semelhantes. Os resultados obtidos permitem a descoberta, por exemplo, de funções de novas proteínas ou a identificação de possíveis mutações genéticas. O alinhamento, em estudos filogenéticos, é utilizado para medir a distância evolutiva entre duas ou mais espécies, com base na homologia das seqüências comparadas.

A inferência da árvore evolutiva de um grupo de táxons a partir da comparação entre as seqüências de genes homólogos requer que estas seqüências estejam alinhadas. O alinhamento possibilita testar a hipótese de homologia e, como tal, consiste em trabalhar com seqüências homólogas e definir posições homólogas ao longo dessas seqüências.

Ao fazer o alinhamento, é necessário ter em mente o conceito de homologia posicional, tendo em vista que, para uma dada seqüência, cada posição ou sítio é tratada como um caráter, e pressupõe-se que essas posições sejam homólogas.

A homologia posicional em seqüências ortólogas permite reconhecer substituições e eventos de inserção e deleção ("indels") de bases em uma ou mais seqüências. O termo "indel" é útil nesse caso porque nem sempre é possível determinar *a priori* em um alinhamento se houve um evento de inserção de nucleotídeos em um grupo de seqüências ou se ocorreu a perda de alguns nucleotídeos no outro grupo. Indels que ocorrem durante o processo evolutivo podem ser mantidos, especialmente se não afetarem o valor adaptativo do portador. É importante saber se eles representam realmente eventos de inserção/deleção que ocorreram ao longo da evolução, ou se eles simplesmente representam lacunas que se originaram durante o alinhamento das seqüências, pela ocupação de uma mesma região por bases diferentes. As regiões onde os indels ocorrem apresentam lacunas de alinhamento. Os métodos de inferência filogenética podem lidar com as lacunas como um quinto estado do caráter ou então ignorar na análise os sítios que as apresentam. O problema de desconsiderar essas regiões com lacunas é que se pode perder informação evolutiva se elas representarem verdadeiros eventos indels (MORRISON; ELLIS, 1997).

Freqüentemente, o conceito de homologia é confundido com de similaridade. Se dois genes são sabidamente homólogos e as posições homólogas foram definidas, pode-se estabelecer a quantidade de sítios iguais entre as seqüências desses dois genes. Assim, se as seqüências deles apresentam 100 sítios, dos quais 92 são idênticos e 8 deles são variáveis, eles ainda continuam sendo homólogos e apresentam 92% de similaridade entre si, e não 92% de homologia.

7.3 Mudanças Evolutivas e Diferenças entre Seqüências de Nucleotídeos

Introdução

Ney e Kumar (2000) destacam que a análise das mudanças evolutivas nas seqüências de DNA são mais complicadas do que nas seqüências de proteínas, porque existem vários tipos de regiões na molécula de DNA, como regiões codificadoras de proteínas, regiões não-codificadoras, exons, introns, regiões

flanqueadoras, seqüências de DNA repetitivo e seqüências de inserção. Por isso, é importante conhecer o tipo e a função da região de DNA sob investigação, uma vez que as taxas de mutação variam amplamente entre as diferentes regiões. Mesmo se forem consideradas apenas as regiões codificadoras de proteínas, os padrões de substituições de nucleotídeos na primeira, segunda e terceira posições dos códons não são iguais. Além do mais, algumas regiões estão mais sujeitas à seleção natural que outras, e isso também contribui para a variação no padrão evolutivo entre as diferentes regiões do DNA.

Medidas de distâncias genéticas entre seqüências

a) Distância *p*

Quando duas seqüências de DNA são derivadas de uma seqüência ancestral comum, elas gradualmente divergem por substituição de nucleotídeos. Uma medida simples dessa divergência é a proporção *p* de nucleotídeos, em que as duas seqüências são diferentes. Por exemplo, considerando as seqüências X e Y, suas diferenças são dadas por:

$$\hat{p} = \frac{n_d}{n}$$

sendo:

n_d: número de nucleotídeos diferentes entre as duas seqüências; e

n: número total de nucleotídeos examinados nas seqüências X ou Y.

Essa medida é, rotineiramente, denominada de **distância *p*** para seqüências de nucleotídeos.

Uma vez que a quantidade de substituição *n_d* tem distribuição binomial, pode-se afirmar que:

$$V(\hat{p}) = \frac{p(1-p)}{n}$$

O valor de *p* é a medida da diferença total entre as seqüências X e Y, porém há maior interesse em particularizar o tipo de substituição de nucleotídeo a

nucleotídeo, caracterizando-os em transições e transversões. Uma vez que há quatro tipos de nucleotídeos (A, T, C, e G) em cada seqüência, ocorrerão 16 tipos de pares substituições de nucleotídeos, conforme mostra a Figur a seguir:

Classes	Pares de nucleotídeos				Freqüência
Idênticos	AA	TT	CC	GG	
Freqüência	O ₁	O ₂	O ₃	O ₄	O
Transição	AG	GA	TC	CT	
Freqüência	P ₁₂	P ₂₁	P ₄₃	P ₃₄	P
Transversão	AT	TA	AC	CA	
Freqüência	Q ₁₃	Q ₃₁	Q ₁₄	Q ₄₁	Q
Transversão	TG	GT	CG	GC	
Freqüência	Q ₃₄	Q ₄₃	Q ₄₂	Q ₂₄	Q

Constata-se, na tabela anterior, que quatro pares são de nucleotídeos idênticos; quatro, de transições (P); e oito, de transversões (Q).

Quando as freqüências de nucleotídeos estão em equilíbrio, tem-se:

$$P_{12} = P_{21} \quad \text{e} \quad P_{34} = P_{43}$$

$$Q_{13} = Q_{31} \quad Q_{14} = Q_{41} \quad Q_{34} = Q_{43} \quad \text{e} \quad Q_{42} = Q_{24}$$

A partir das informações da tabela anterior, pode-se estabelecer:

$$O + P + Q = 1 \quad \text{e} \quad \hat{p} = \frac{n_d}{n} = P + Q$$

endo P a freqüência de substituições do tipo transição e Q a freqüência de substituição do tipo transversão.

A razão entre a freqüência de substituições do tipo transição e transversão é dada por meio de:

$$\hat{R} = \frac{\hat{P}}{\hat{Q}}$$

em que \hat{P} e \hat{Q} são os valores obtidos a partir de dados observados e constituem estimativas de P e Q . Se a substituição ocorre ao acaso entre os quatro nucleotídeos, espera-se que:

$$Q = 2P$$

e, portanto:

$$R = \frac{P}{Q} = \frac{P}{2P} = 0,5$$

Na prática, tem sido verificado que a transição ocorre com maior freqüência que a transversão, sendo comum encontrar valor de R maior que 1. Para genes nucleares, o valor de \hat{R} tem apresentado estimativas entre 0,5 e 2,0, mas para DNA mitocondrial a razão pode atingir valores iguais a 15.

b) Distâncias evolucionárias

A distância p fornece uma estimativa do número de substituições de nucleotídeos por sítio quando as duas seqüências são proximamente relacionadas. Por outro lado, quando p é grande, ele subestima este valor porque não leva em consideração a ocorrência de mutações reversas ou paralelas. Por isso, para estimar o número de substituições de nucleotídeos, de forma mais acurada, é necessário usar um determinado modelo matemático. Diversos modelos foram desenvolvidos por diferentes autores, conforme citam Nei e Kumar (2000), e alguns deles serão descritos a seguir.

b1) Método de Jukes e Cantor

Um modelo simples de substituição de nucleotídeos é aquele que assume que a substituição ocorre, em qualquer sítio, com igual freqüência e que, em cada sítio, um nucleotídeo muta para qualquer um dos três nucleotídeos remanescentes, com igual probabilidade (α) por ano (ou qualquer outra unidade de tempo). Disso decorrem as seguintes possibilidades, e respectivas probabilidades, de mutação:

	A	T	C	G
A	-	α	α	α
T	α	-	α	α
C	α	α	-	α
G	α	α	α	-

Neste modelo, tem-se:

$$P = 4\alpha$$

$$Q = 8\alpha$$

em que α é a taxa de substituição de um nucleotídeo por outro, por unidade de tempo.

A probabilidade de substituição de um nucleotídeo é, portanto, dada por:

$$r = P(\text{substituição de um nucleotídeo})$$

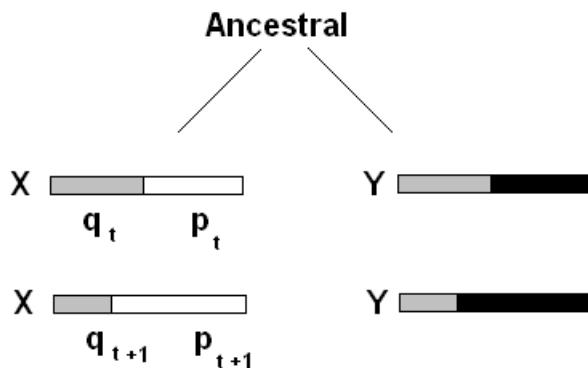
Assim, para um particular nucleotídeo – por exemplo, o nucleotídeo A – tem-se:

$$r = P(A \rightarrow G) + P(A \rightarrow C) + P(A \rightarrow T) = 3\alpha$$

Considerando duas seqüências de nucleotídeos X e Y, as quais divergem de uma seqüência ancestral comum a t anos atrás, e admitindo também que:

q_t : proporção de nucleotídeos idênticos entre X e Y e

$p_t = 1 - q_t$: proporção de nucleotídeos diferentes entre X e Y.



Assim, tem-se que:

a) A proporção de nucleotídeos idênticos no tempo $t+1$ é dada por:

$$p_i = (1-r)^2$$

b) A proporção de nucleotídeos idênticos no tempo t , que permanecerão idênticos em $t+1$, será:

$$p_{im} = q_t(1-r)^2 \approx q_t(1-2r)$$

c) A proporção de nucleotídeos diferentes no tempo t , que permanecerão idênticos em $t+1$, depende de assumir certas situações de forma que a probabilidade possa ser expressa por:

$$P(A \text{ e } T \text{ idênticos em } t+1) = [P(A \rightarrow T / X) P(T=T/Y)] + [P(A=A/X)P(T \rightarrow A/Y)]$$

$$= \alpha(1-r) + \alpha(1-r) = 2\alpha(1-r) = \frac{2}{3}r(1-r) \approx \frac{2}{3}r$$

Desse modo, a probabilidade de os nucleotídeos serem idênticos em $t+1$ é dada pela soma das probabilidades de serem idênticos em t , e permanecerem idênticos em $t+1$, e de serem diferentes em t , porém tornando-se idênticos em $t+1$. Ou seja:

$$q_{t+1} = P(\text{idênticos em } t, \text{idênticos em } t+1) + P(\text{diferentes em } t, \text{idênticos em } t+1) \text{ ou}$$

$$q_{t+1} = (1-2r)q_t + \frac{2}{3}r(1-q_t)$$

logo:

$$\Delta_q = q_{t+1} - q_t = \frac{2}{3}r - \frac{8r}{3}q_t$$

Deve ser lembrado que uma equação diferencial é uma expressão que descreve uma relação entre uma função de uma variável e sua derivada. Ela tem uma solução geral que descreve a trajetória de q no tempo t contínuo ($\Delta t \rightarrow 0$). Assim, a expressão anterior pode ser escrita na forma:

$$\frac{\partial q}{\partial t} = \frac{2r}{3} - \frac{8r}{3}q_t$$

que é uma equação diferencial do tipo:

$$\frac{\partial q}{\partial t} = w - zq$$

e, neste caso, a solução geral é dada por:

$$q(t) = \frac{w}{z} + ce^{-zt}$$

em que c é uma constante que pode ser calculada com base na condição inicial.

Pela equação original, tem-se que $w = (2/3)r$ e $z = (8/3)r$. Então, quando $q = 1$ e $t = 0$ (condição inicial); tem-se $c = 3/4$.

A solução particular (dada a condição inicial) é dada por:

$$q(t) = \frac{1}{4} + \frac{3}{4}e^{-\frac{8r}{3}t}$$

que também pode ser escrita como:

$$q = 1 - \frac{3}{4}(1 - e^{-\frac{8rt}{3}})$$

Neste modelo, o número esperado de substituição por sítio (d) para duas seqüências é dado por:

$$d = 2rt$$

logo:

$$q = 1 - \frac{3}{4}(1 - e^{-\frac{4d}{3}})$$

de forma que:

$$d = -\frac{3}{4} \ln \left[1 - \frac{4}{3}(1 - q) \right]$$

$$e \quad V(d) = \frac{9p(1-p)}{(3-4p)^2 r}$$

No modelo anterior, é assumido que a taxa de substituição de nucleotídeos é a mesma para todos os pares; logo, as freqüências esperadas de A, T, C e G, eventualmente, serão iguais a 0,25. Por outro lado, uma vez que não se faz

pressuposições sobre as freqüências iniciais, a equação para o cálculo de d é válida.

b2) Modelo de dois parâmetros de Kimura

A taxa de substituição de nucleotídeos é freqüentemente maior para transições do que para transversões. Kimura (1980) propõe um método para estimar o número de substituições de nucleotídeos por sítio, assumindo que a taxa de transição por sítio e por unidade de tempo (α) é diferente da taxa de transversão(2β). Assim, têm-se:

	A	T	C	G
A	-	β	β	α
T	β	-	α	β
C	β	α	-	β
G	α	β	β	-

A taxa total de substituição por sítio e por unidade de tempo é, portanto, dada por:

$$r = \alpha + 2\beta$$

O valor que se procura é o de d, que expressa a taxa de substituição em unidade de tempo, de um nucleotídeo para outro, dada por:

$$d = 2rt = 2\alpha t + 4\beta t$$

Kimura demonstra que:

$$P = \frac{1}{4}(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t}) \quad \text{e}$$

$$Q = \frac{1}{2}(1 - e^{-8\beta t}),$$

em que t é o tempo de divergência entre duas seqüências (X e Y). Por isso, o número esperado de substituições de nucleotídeos por sítio entre X e Y é dado por:

$$d = 2rt = 2\alpha t + 4\beta t \quad \text{ou}$$

$$d = -\frac{1}{2} \ln(1-2P-Q) - \frac{1}{4} \ln(1-2Q).$$

sendo

$$V(d) = \frac{1}{n} [c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2]$$

em que:

$$c_1 = \frac{1}{1-2P-Q} \quad c_2 = \frac{1}{1-2Q} \quad \text{e} \quad c_3 = \frac{c_1 + c_2}{2}$$

No presente modelo é possível estimar o número de transições ($s = 2\alpha t$) e transversões $v = 4\beta t$

No modelo de Kimura, a freqüência de equilíbrio de cada nucleotídeo é 0,25. Por outro lado, as equações anteriores são aplicáveis independentemente das freqüências dos diferentes nucleotídeos, e, com relação a isso, o modelo é similar ao de Jukes-Cantor. Essa propriedade torna os dois modelos aplicáveis a um maior espectro de situações do que os demais modelos.

Como ilustração, é considerado um exemplo em que foi calculado o valor da distância p entre duas seqüências, considerando informações do 1º, 2º e 3º. nucleotídeo para cada códon. As medidas transformadas de d , segundo diferentes modelos, são apresentadas a seguir:

Posição no códon	$p = 1-q$	Jukes-Cantor	Kimura
1º.	0,155	0,173	0,178
2º.	0,085	0,091	0,098
3º.	0,368	0,506	0,523

Verifica-se que os valores obtidos pelo modelo de Jukes-Cantor e Kimura são relativamente próximos e que a transformação apresenta variações de maior magnitude para valores de p relativamente alto.

b3) Distância de Tajima e Nei

Sabe-se que, além da variação na taxa de substituição para diferentes posições do códon, o conteúdo relativo de nucleotídeos G e C (denominado de conteúdo GC) na terceira posição varia segundo a espécie considerada. Entretanto, havendo variação no conteúdo GC entre as seqüências, pode haver distorção nas estimativas de distâncias, já que a probabilidade de substituição entre as bases também não é igual. Portanto, duas seqüências com proporção de bases convergentemente semelhante parecerão mais próximas do que realmente o são. Por isso, Tajima e Nei desenvolveram um modelo para contornar esse problema. Nesse caso, o conteúdo GC também é estimado diretamente dos dados.

A obtenção de d é feita por meio de:

$$d = -b \ln \left(1 - \frac{p}{b} \right)$$

sendo:

$$b = \frac{1}{2} \left[1 - \sum_{i=1}^4 g_i^2 + \frac{p^2}{c} \right] \text{ e } c = \sum_{i=1}^3 \sum_{j=i+1}^4 \frac{x_{ij}^2}{2g_i g_j}$$

em que:

x_{ij} : a freqüência relativa do par de nucleotídeos i e j nas duas seqüências comparadas; e

g_i : a freqüência do i-ésimo nucleotídeo nas duas seqüências comparadas.

Tem-se ainda que:

$$V(d) = \frac{b^2 p (1-p)}{(b-p)^2 n}$$

b4) Distância de Tamura com 3-parâmetros

A distância de 3 parâmetros de Tamura agrupa todos os parâmetros previamente mencionados, ou seja, ela considera substituições múltiplas, conteúdo de GC e taxa de transição/transversão no cálculo da distância entre duas seqüências. À primeira vista, pode parecer que esta seja a mais adequada de todas

as medidas de distâncias, por levar em consideração mais parâmetros. Isso porque, quanto maior o número de parâmetros incluídos no cálculo da distância, mais próxima seria a distância média estimada da distância real. Entretanto, o aumento do número de parâmetros também aumenta a variância associada a essa medida.

Portanto, a distância de três parâmetros de Tamura é recomendada somente para casos em que o conteúdo de GC e a razão transição/transversão são bem diferentes do esperado (0,5).

Neste caso, tem-se:

$$d = -h \ln \left(1 - \frac{P}{h} - Q \right) - \frac{1}{2}(1-h) \ln(1-2Q)$$

em que:

$$h = 2\theta(1-\theta)$$

sendo θ a freqüência de GC.

b5) Distância de Tamura e Nei

O tipo de distância que leva em consideração o maior número de parâmetros é da distância de Tamura e Nei. Nesse caso, considera-se que as taxas de transição entre as purinas (A e G) sejam diferentes das taxas de transição entre as pirimidinas (C e T); por isso, elas são computadas separadamente na análise dos dados. Esta distância foi desenvolvida especificamente para ser utilizada na região controladora do DNA mitocondrial.

b6) Distância Gamma-Poisson

Todas as distâncias discutidas anteriormente são baseadas no pressuposto de que todos os sítios evoluem a uma mesma taxa, o que raramente acontece. Por exemplo, em regiões codificadoras, a terceira posição do códon evolui mais rapidamente que as outras duas posições. Quando cada sítio evolui de acordo com a distribuição de Poisson, mas a taxa (λ) varia de sítio para sítio, essas taxas seguem uma distribuição gama. A distribuição gama é descrita por dois parâmetros,

α e β . O primeiro parâmetro é a variação das taxas de acordo com o sítio ($\alpha = 0$ para taxas completamente diferentes e $\alpha = \infty$ para taxas idênticas). O segundo parâmetro é a calibração de acordo com a média. Esses dois parâmetros são relacionados pela equação: $\frac{\alpha}{\beta} = \mu$, em que $\alpha = \frac{(\mu^2)}{c}$; μ é a média e c é a variância das taxas para todos os sítios.

Vários métodos foram desenvolvidos para estimar o número de substituições por sítio, a média e a variância, ao longo dos sítios.

A correção gama apenas adiciona o parâmetro de variação de taxas no cálculo do número de substituições. Assim, pode-se incorporá-la a todos os modelos de evolução de nucleotídeos apresentados anteriormente, como Jukes-Cantor, Kimura 2-parâmetros, Tamura 3-parâmetros etc., ou ao modelo de Poisson, no caso de seqüências de aminoácidos. Dessa forma, assume-se que todos os sítios evoluem de acordo com uma determinada distribuição e as taxas de evolução diferem entre os sítios.

A correção para a heterogeneidade entre taxas de mutação de diferentes sítios deve ser, em princípio, mais importante para seqüências de nucleotídeos do que para seqüências de aminoácidos. No entanto, é importante confirmar se o valor de α é baixo, ou seja, menor que 0,65. Nesse caso, a correção gama torna-se desnecessária.

c) Comparação Entre os Métodos

A comparação teórica entre as medidas pode ser feita assumindo $n = \infty$. Nesse caso, diferentes medidas de distância fornecem resultados substancialmente diferentes quando $d \geq 0,6$. A distância de Tamura é virtualmente idêntica à distância de Tamura-Nei para valores de d acima de 0,5, enquanto as distâncias de Tamura, Kimura e Jukes-Cantor são essencialmente iguais à distância de Tamura-Nei quando $d \geq 0,25$. Mesmo a distância p torna-se muito similar `de outras medidas de

distâncias quando $p \leq 0,1$. Por isso, ao investigar seqüências proximamente relacionadas, não há necessidade de utilizar medidas complexas de distâncias. Nesses casos, é melhor utilizar a mais simples, porque ela terá menor variância (NEI; KUMAR, 2000).

Também dever ser destacado que, na construção de árvores filogenéticas a partir de medidas de distâncias, a utilização de modelos mais sofisticados, no cálculo das medidas de distâncias, não é mais eficiente para obter a topologia correta do que o emprego de um modelo mais simples. Por sua vez, no cálculo dos comprimentos dos ramos de uma árvore, um método de medida de distâncias que se ajuste melhor aos dados, geralmente, produz resultados mais confiáveis (NEI; KUMAR, 2000).

7.4 Métodos de Reconstrução de Árvores

A reconstrução de árvores filogenéticas por meio da utilização de métodos estatísticos foi iniciada, independentemente, na taxonomia numérica para caracteres morfológicos (SOKAL; SNEATH, 1963) e na genética de populações para dados de freqüências gênicas (CAVALLI-SFORZA; EDWARDS, 1964). Alguns dos métodos elaborados naquela época ainda são empregados na análise de dados moleculares, porém nos últimos anos muitos métodos novos foram desenvolvidos.

Inferir uma filogenia é um processo de estimação no qual uma ‘melhor estimativa’ da história evolutiva é feita com base em informações incompletas, uma vez que usualmente não se dispõe de informações do passado, mas apenas de seqüências de organismos contemporâneos.

Pelo fato de que muitas filogenias diferentes podem ser geradas com um mesmo conjunto de dados, devem-se especificar critérios para selecionar as árvores que melhor representem a melhor estimativa da verdadeira história evolutiva do grupo sob investigação. Por isso, uma reconstrução filogenética consiste em dois passos:

- 1) Definição de um critério de otimização ou de uma função objetiva, isto é, o valor que é associado a uma árvore e subsequentemente utilizado para compará-la com outras.
- 2) Delineamento de um algoritmo específico para computar o valor da função objetiva e identificar a árvore (ou grupo de árvores) que possui o melhor valor, de acordo com este critério.

Por outro lado, deve ficar claro que uma árvore inferida é tão boa quanto as pressuposições nas quais o método de reconstrução filogenética é baseado. Geralmente se está interessado em árvore filogenética que represente a história evolutiva de um grupo de espécies ou populações. Em uma árvore de espécies, o tempo de divergência entre duas OTUs (espécies ou populações) refere-se ao tempo no qual estas duas OTUs se tornaram reprodutivamente isoladas. Por outro lado, quando uma árvore é construída a partir de um gene amostrado de cada uma das OTUs, a árvore obtida não concorda, necessariamente, com a árvore de espécies. Isso se dá pelo fato de que, na presença de alelos polimórficos no loco amostrado, espera-se que o tempo de divergência do gene em questão seja anterior ao tempo de divergência das espécies.

Na teoria de inferência filogenética assume-se que as seqüências de DNA ou proteínas que serão estudadas são infinitamente longas e que grande número de nucleotídeos ou aminoácidos que representem estas longas seqüências será amostrado para execução de qualquer estudo. Uma árvore construída por meio da utilização das informações contidas nas seqüências infinitamente longas é denominada de **árvore esperada**, enquanto uma árvore construída com base no número total de substituições de seqüências inteiras, mas reais, é denominada de **árvore realizada**. Já uma árvore reconstruída a partir das seqüências amostradas (observadas) é denominada de **árvore inferida**. Esses três tipos de árvores são, geralmente, diferentes devido a erros estocásticos, decorrentes da taxa de substituições de nucleotídeos, assim como dos processos de amostragem dos dados.

Pode-se argumentar que o interesse geral é, de fato, conhecer a árvore verdadeira ou árvore esperada e não a árvore realizada. Na prática, porém, é mais fácil reconstruir uma árvore realizada que uma árvore esperada, porque os dados disponíveis se referem à árvore realizada. Além do mais, a topologia da árvore realizada é a mesma da árvore esperada, a menos que a primeira se torne multifurcante por causa de erros estocásticos. Entretanto, à medida que o número de nucleotídeos examinados aumenta, a árvore realizada se aproxima da árvore esperada.

Existem muitos métodos estatísticos que podem ser utilizados na reconstrução de árvores filogenéticas a partir de dados moleculares. Os métodos comumente utilizados são classificados em três categorias:

- a) Métodos baseados distâncias – que constroem um dendrograma a partir de uma matriz de distâncias entre as OTUs. Nesta categoria, as medidas de distâncias são obtidas de informações como taxas de substituições de nucleotídeos ou aminoácidos. Dentre eles, destacam-se o UPGMA, o método das distâncias transformadas, o método *neighbor-relation* ou da “relação entre vizinhos” e o *neighbor-joinnig*.
- b) Métodos baseados em parcimônia – que consideram a distribuição dos “estados” de caracteres entre as OTUs, como, por exemplo, o tipo de nucleotídeo ou aminoácido presente em uma determinada posição, ou a presença/ausência de uma deleção/inserção em um dado loco. Nessa categoria é buscada a topologia que pode ser explicada pelo menor número de mudanças evolutivas ou passos. O método da máxima parcimônia é um representante típico dessa estratégia.
- c) Métodos de máxima verossimilhança – que buscam a topologia que apresenta a maior probabilidade de ter gerado os dados observados, assumindo determinado modelo evolutivo. Nessa estratégia, os dados podem ser tanto de distâncias quanto de estados de caráter.

Atualmente, considera-se que a reconstrução de uma árvore filogenética equivale a uma inferência estatística da árvore verdadeira, que é desconhecida. Dois processos estão envolvidos nessa inferência:

- estimação da topologia da árvore; e
- estimação dos comprimentos dos ramos para uma dada topologia.

Quando a topologia é conhecida, a estimativa dos comprimentos dos ramos é relativamente simples, e existem vários métodos estatísticos que podem ser utilizados. O problema é a estimativa ou reconstrução da topologia, uma vez que o número de topologias possíveis é muito grande, crescendo exponencialmente com o número de OTUs investigadas; por isso, é muito difícil encontrar a verdadeira topologia entre todas as topologias possíveis.

Normalmente, a topologia mais provável é buscada por meio da otimização de um princípio como a máxima verossimilhança ou o princípio da evolução mínima. A base teórica desse procedimento não está bem entendida, porém simulações de computadores têm mostrado que os princípios de otimização utilizados atualmente funcionam muito bem se o número de nucleotídeos ou aminoácidos amostrados for grande. Quando esse número é pequeno e a quantidade de seqüências analisadas é grande, o princípio da otimização tende a fornecer topologias incorretas.

As metodologias de reconstrução filogenética ainda geram bastante controvérsia. Além de preferências pessoais, existem três razões básicas para essa controvérsia:

- a) Formação acadêmica dos pesquisadores. Alguns foram formados originalmente como sistematas, utilizando caracteres morfométricos, por isso preferem métodos baseados em parcimônia, que requerem um número mínimo de pressuposições. Outros foram treinados no campo da genética e biologia molecular e tendem a preferir métodos analíticos, mas tendem a não depositar confiança em modelos matemáticos altamente sofisticados. Um terceiro grupo teve um treinamento majoritariamente matemático ou estatístico e tende a ver o problema de reconstrução filogenética como um problema matemático mais do que um problema prático.

- b) Interesse focado em problemas microevolutivos, circunscritos dentro e entre espécies próximas, ou macroevolutivos, abrangendo famílias, ordens ou níveis hierárquicos superiores.
- c) Na análise filogenética, a árvore verdadeira é quase sempre desconhecida, sendo difícil testar a acurácia de árvores obtidas por métodos de reconstrução diferentes, a não ser através de um número simplificado de pressuposições.

É importante enfatizar que a filogenia molecular é área relativamente nova da ciência e quaisquer de seus métodos estatísticos são apenas aproximações da realidade, com pontos fortes e fracos, assim como com graus de adequação variáveis, conforme a situação envolvida.

Outro aspecto a ser considerado é o foco de interesse entre cladistas e feneticistas. **Cladistas** estão interessados em definir as “rotas” ou passos evolutivos; dessa forma, um **cladograma** é uma árvore enraizada que reflete as inter-relações ancestral-descendente. Por outro lado, os **feneticistas** estudam os grupos de organismos procurando estabelecer os graus de similaridade entre eles, sejam elas morfológicas, anatômicas ou moleculares. Uma árvore expressando relações fenéticas é denominada de **fenograma**. Se houver uma relação linear entre tempo evolutivo e grau de divergência genética, os dois tipos de árvore serão idênticos.

O método da máxima parcimônia é um típico representante da abordagem cladística, enquanto o UPGMA é um tipo representante da abordagem feneticista. Por outro lado, outros métodos de reconstrução filogenética não podem ser associados tão especificamente a nenhuma dessas abordagens.

7.4.1. Métodos Baseados em Distâncias

Para o estabelecimento de árvores filogenéticas por métodos baseados em distância é necessário, inicialmente, obter uma matriz cujos elementos expressam a medida de distância evolucionária entre cada par de táxons. Geralmente são obtidas

seqüências, as quais são comparadas e obtido, preliminarmente, o valor da estatística p . Este valor é, então, transformado em uma medida de distância que reflete as mudanças evolucionárias entre os taxas comparados.

Segundo Felsenstein (2004), a melhor maneira de compreender os métodos baseados em matriz de distâncias é considerar as distâncias como estimativas de comprimentos de ramos que separam cada par de OTUs, de modo que cada distância consista na inferência da melhor árvore não-enraizada entre elas. Assim, tem-se um grande número de árvores estimadas, e o objetivo é encontrar uma árvore com n -OTUs que atenda a esse critério. A dificuldade em fazer isso resulta do de fato que as distâncias individuais entre duas OTUs não podem ser “acomodadas” com exatidão em uma árvore envolvendo n -OTUs. Por isso, é necessário encontrar a árvore que represente a melhor aproximação das diversas árvores envolvendo pares de OTUs.

Há várias metodologias para o estabelecimento de uma determinada topologia. Serão abordadas, a seguir, algumas que se mostraram mais úteis para uma maior amplitude de situações. Essas metodologias são capazes de estabelecer uma topologia apropriada e prover estimativas dos comprimentos dos ramos internos e externos da árvore.

Método UPGMA

O UPGMA (*unweighted pair-group method using arithmetic averages*) é uma metodologia amplamente utilizada no agrupamento de acesso ou de espécies em estudos de taxonomia numérica. Seu emprego, na construção de árvores filogenéticas, tem sido recomendado nos casos em que as taxas de substituição gênica entre táxons são mais ou menos constantes. Particularmente, quando dados de freqüências gênicas são utilizados na reconstrução filogenética, este método produz árvores razoavelmente confiáveis, quando comparado com outros métodos de distâncias.

Por outro lado, se a taxa de substituição gênica não for constante ou o número de genes amostrados for pequeno, a utilização do UPGMA resulta,

frequêntemente, em erros topológicos. Isso porque o método gera um dendrograma em que os ramos que partem do mesmo nó possuem, necessariamente, o mesmo comprimento, ou seja, a árvore gerada pelo UPGMA é sempre ultramétrica, porque o pressuposto básico é de taxas constantes de evolução.

O algoritmo do UPGMA opera, a partir dos dados de uma matriz de distância, através dos passos:

- a) O par de OTUs mais semelhante é unido. Esse par fica separado por uma distância idêntica ao valor que os separa na matriz, com um nó posicionado no ponto médio entre os dois.
- b) O par de OTUs recé-unidos passa a ser uma nova OTU ou OTU composta. Uma nova matriz de distâncias é calculada em relação a esta OTU composta, considerando a média aritmética entre cada uma das OTUs da unidade composta com as demais OTUs. A menor distância da nova matriz determina quais táxons deverão ser unidos. Novamente, o nó desta união é posicionado no ponto médio da distância que separa as OTUs.
- c) O procedimento anterior é repetido até a incorporação de todas as OTUs no dendrograma.

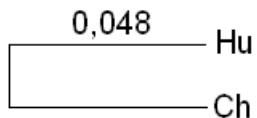
Uma árvore obtida pelo UPGMA geralmente é apresentada com uma raiz, porque é fácil inferir a raiz de uma árvore sob a pressuposição de taxa constante de evolução. Por outro lado, a comparação de métodos e a aplicação de testes de confiabilidade geralmente são factíveis para árvores não-enraizadas. Por isso, não se deve enraizar uma árvore resultante de UPGMA.

Como ilustração, será considerada a seguinte matriz de distância, transcrita de Nei e Kumar (2000). Os valores foram obtidos da comparação de seqüências contendo 896 nucleotídeos de DNA mitocondrial de humanos (Hu), chimpanzés (Ch), gorila (Go), orangotango (Or) e gibão (Gi):

	Hu	Ch	Go	Or	Gi
Hu	-	0,095	0,113	0,183	0,212
Ch		-	0,118	0,201	0,225
Go			-	0,195	0,225
Or				-	0,222
Gi					-

Primeiro Passo

Consiste em identificar, na matriz de distância, os dois táxons mais próximos. Neste caso, trata-se dos táxons Hu e Ch, com valor de distância de $d = 0,095$. A árvore filogenética é então inicialmente construída admitindo-se igual comprimento dos braços em cada ramificação, dado por $d/2$. Assim, tem-se:



Segundo Passo

Consiste em estabelecer uma nova matriz de distância, de dimensão inferior à anterior, em uma unidade e que apresente as distâncias entre pares de táxons e entre um táxon e um determinado grupo. Considera-se que a distância entre o táxon k e o grupo estabelecido pelos táxons ij seja dada por:

$$d_{(ij)k} = \frac{d_{ik} + d_{jk}}{2}$$

Assim, tem-se:

$$d_{(Hu,Ch)Go} = (0,113 + 0,118) / 2 = 0,1155$$

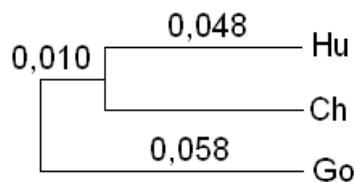
$$d_{(Hu,Ch)Or} = (0,183 + 0,201) / 2 = 0,1920$$

$$d_{(Hu,Ch)Gi} = (0,212 + 0,225) / 2 = 0,2185$$

A nova matriz de distâncias, que expressa a dissimilaridade evolucionária entre táxons, é fornecida por:

	Hu,Ch	Go	Or	Gi
Hu,Ch	–	0,1155	0,1920	0,2185
D = Go		–	0,1950	0,2250
Or			–	0,2220
Gi				–

Agora, constata-se que os elementos mais próximos são Go e o agrupamento (Hu,Ch), com valor de distância igual a 0,1155. O valor do comprimento do braço da árvore é de $d/2$, igual a 0,058. Dessa forma, tem-se:



Terceiro Passo

Novamente é estabelecida a matriz de dissimilaridade com dimensão menor, contendo, agora, as distâncias entre o agrupamento [Hu, Ch, Go] e os demais táxons. Assim, tem-se:

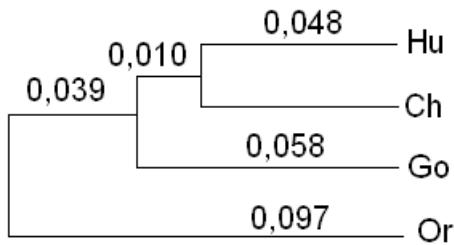
$$d_{(Hu,Ch,Go)Or} = (0,183 + 0,201 + 0,195) / 3 = 0,193$$

$$d_{(Hu,Ch,Go)Gi} = (0,212 + 0,225 + 0,225) / 3 = 0,221$$

logo:

	Hu, Ch, Go	Or	Gi
Hu,Ch,Go	–	0,193	0,221
Or	–	–	0,222
Gi			–

Os mais próximos são [Hu, Ch, Go] e Or, com distância $d = 0,193$ e comprimento de braço igual a 0,097. Assim, tem-se a árvore:



Quarto Passo

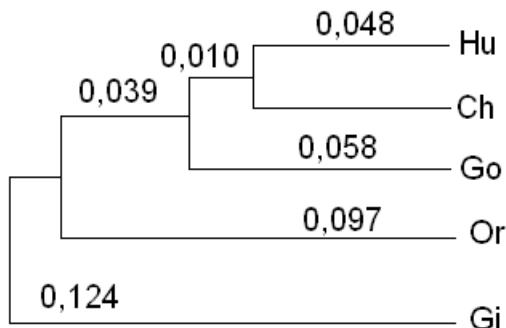
Neste último passo, calcula-se a distância do táxon ainda não agrupado (G_i) com o grupo formado pelos demais táxons, ou seja:

$$d_{(Hu,Ch,Go,Or)G_i} = (0,212 + 0,225 + 0,225 + 0,222) / 4 = 0,221$$

O tamanho do braço é obtido por diferença, ou seja:

$$\text{braço} = 0,221 - 0,097 = 0,124$$

A árvore filogenética, pelo método UPGMA, fica então estabelecida da seguinte maneira:

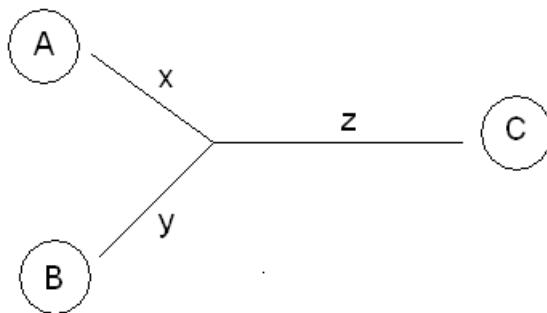


Método ou Algoritmo de Fitch-Margoliash

Fitch e Margoliash (1967) apresentam uma maneira de estimar o comprimento de ramos baseada no princípio de uma árvore com três táxons, cujas estimativas de comprimento dos três ramos apresentam solução única. Não é propriamente um método de reconstrução de árvores, mas um algoritmo possível de ser adotado em

diferentes situações em que o objetivo seja o de obter estimativas de comprimento de ramos. Entretanto, pode-se aplicá-lo, por exemplo, em conjunto com a metodologia UPGMA, obtendo-se um resultado em que não haja necessidade de pressupor taxas de substituição gênica entre táxons pouco variantes e a topologia obtida deixa de ser ultramétrica.

Portanto, admite-se que o comprimento de um ramo da topologia pode ser estimado a partir das informações da seguinte estrutura de árvore:



Assim, podem ser estabelecidas as equações:

$$d_{AB} = x + y$$

$$d_{AC} = x + z$$

$$d_{BC} = y + z$$

As três incógnitas x , y e z podem ser estimadas a partir das informações das três distâncias d_{AB} , d_{AC} e d_{BC} , de forma que se tenha:

$$x = \frac{d_{AB} + d_{AC} - d_{BC}}{2}$$

$$y = \frac{d_{AB} + d_{BC} - d_{AC}}{2}$$

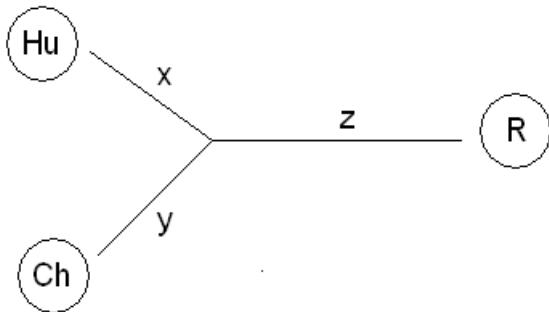
$$z = \frac{d_{AC} + d_{BC} - d_{AB}}{2}$$

Como exemplo, também será considerada a matriz de distâncias dada por:

$$D = \begin{bmatrix} & Hu & Ch & Go & Or & Gi \\ Hu & - & 0,095 & 0,113 & 0,183 & 0,212 \\ Ch & & - & 0,118 & 0,201 & 0,225 \\ Go & & & - & 0,195 & 0,225 \\ Or & & & & - & 0,222 \\ Gi & & & & & - \end{bmatrix}$$

Primeiro Passo

Consiste em identificar na matriz $D_{5 \times 5}$ os táxons mais próximos – no caso, Hu e Ch. A questão a ser resolvida agora é conhecer a estimativa do comprimento dos braços da ramificação que une estes táxons. Dessa forma, tem-se:



A letra R representa todos os demais táxons não considerados, que, no exemplo, são Go, Or e Gi. São estabelecidas as equações:

$$d_{Hu,Ch} = x + y = 0,095$$

$$d_{Hu,R} = x + z = (0,113 + 0,183 + 0,212)/3 = 0,1693$$

$$d_{Ch,R} = y + z = (0,118 + 0,201 + 0,225)/3 = 0,1813$$

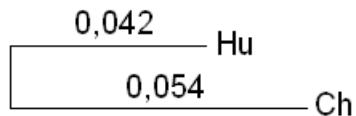
logo:

$$x = \frac{d_{AB} + d_{AC} - d_{BC}}{2} = \frac{0,095 + 0,1693 - 0,1813}{2} = 0,042$$

$$y = \frac{d_{AB} + d_{BC} - d_{AC}}{2} = \frac{0,095 + 0,1813 - 0,1693}{2} = 0,054$$

$$z = \frac{d_{AC} + d_{BC} - d_{AB}}{2} = \frac{0,1693 + 0,1813 - 0,095}{2} = 0,128$$

Dessa forma, fica estabelecida a seguinte topologia inicial:

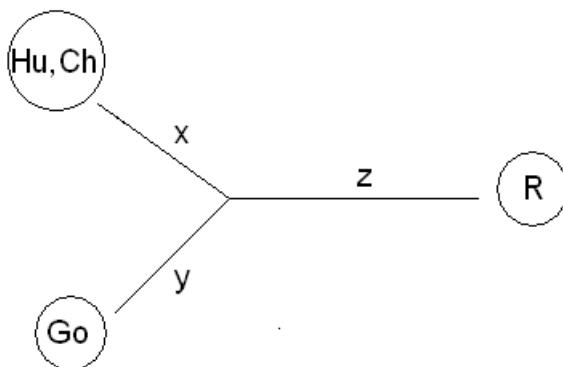


Segundo Passo

A matriz de distâncias de dimensão inferior ($D_{4 \times 4}$) é obtida de forma idêntica à descrita para a metodologia UPGMA, dada por:

$$D = \begin{matrix} & \text{Hu,Ch} & \text{Go} & \text{Or} & \text{Gi} \\ \text{Hu,Ch} & - & 0,1155 & 0,1920 & 0,2185 \\ \text{Go} & & - & 0,1950 & 0,2250 \\ \text{Or} & & & - & 0,2220 \\ \text{Gi} & & & & - \end{matrix}$$

Novamente, identificam-se as OTUs mais similares na matriz de distâncias. Neste exemplo, trata-se do grupo Hu e Ch e do táxon Go, cuja distância do braço na ramificação da árvore filogenética é estabelecida por meio de:



As seguintes medidas podem ser determinadas:

$$d_{(Hu,Ch)R} = x + z = d_{(Hu,Ch)(Or,Gi)} = \frac{0,183 + 0,212 + 0,201 + 0,225}{4} = 0,2055$$

$$d_{Go,R} = y + z = d_{Go(Or,Gi)} = \frac{0,195 + 0,225}{2} = 0,210$$

$$d_{(Hu,Ch)Go} = x + z = 0,1155$$

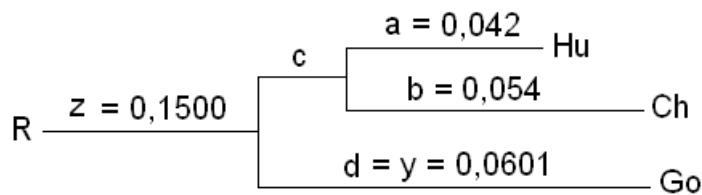
logo:

$$x = \frac{0,2053 + 0,1155 - 0,210}{2} = 0,0554$$

$$y = \frac{0,210 + 0,1155 - 0,2053}{2} = 0,0601$$

$$z = \frac{0,2053 + 0,210 - 0,1155}{2} = 0,1500$$

Assim, tem-se:



Com base na topologia supracitada, são calculados os valores c e d, por meio de:

$$d_{(Hu,Ch)R} = \frac{a+b}{2} + c + z = 0,2053$$

$$d_{Go,R} = z + d = 0,210$$

$$d_{(Hu,Ch)Go} = \frac{a+b}{2} + c + d = 0,1155$$

Sabendo que $(a+b)/2 = 0,048$ e $z = 0,150$, tem-se:

$$c = 0,2053 - 0,048 - 0,150 = 0,008$$

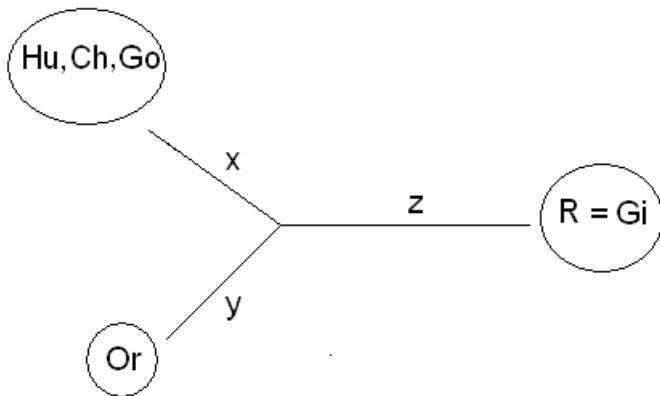
$$d = 0,210 - 0,150 = 0,06$$

Terceiro Passo

A matriz de dissimilaridade com dimensão $D_{3 \times 3}$ é dada por:

$$D = \begin{matrix} & \text{Hu,Ch,Go} & \text{Or} & \text{Gi} \\ \text{Hu,Ch,Go} & - & 0,193 & 0,221 \\ \text{Or} & - & - & 0,222 \\ \text{Gi} & - & - & - \end{matrix}$$

Os táxons mais próximos são Hu, Ch, Go e Or. As distâncias indispensáveis para cálculo do tamanho do braço da nova ramificação são dadas por:



Assim, tem-se:

$$d_{(Hu,Ch,Go)R} = x + z = 0,221$$

$$d_{Or,R} = d_{OR,Gi} = y + z = 0,222$$

$$d_{(Hu,Ch,Go)Or} = x + y = 0,193$$

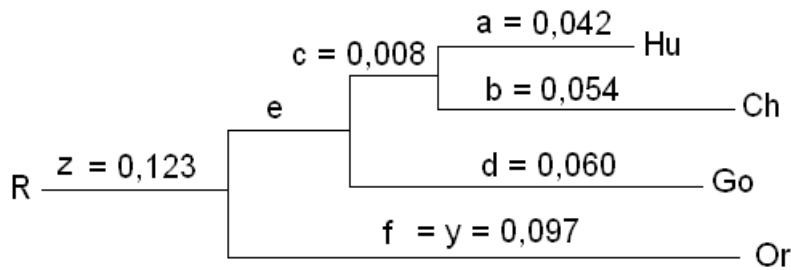
de forma que:

$$x = \frac{0,221 + 0,193 - 0,222}{2} = 0,0965$$

$$y = \frac{0,222 + 0,193 - 0,221}{2} = 0,097$$

$$z = \frac{0,221 + 0,222 - 0,193}{2} = 0,123$$

Com base nessas informações, estimam-se os comprimentos dos braços na árvore:



Dessa forma:

$$d_{(Hu,Ch,Go)R} = z + e + \frac{\frac{a+b}{2} + c + d}{2} = 0,221$$

$$d_{Or,R} = z + f = 0,222$$

$$d_{(Hu,Ch,Go)Or} = \frac{\frac{a+b}{2} + c + d}{2} + e + f$$

Assim, é possível estimar:

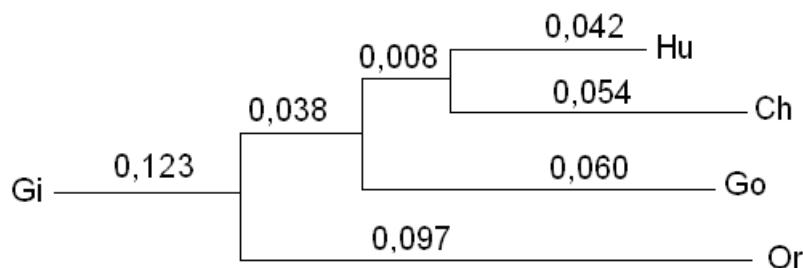
$$e = 0,221 - 0,125 - 0,0575 = 0,03835$$

$$f = 0,222 - 0,125 = 0,097$$

Quarto Passo

Neste último passo, estima-se a distância entre o táxon Gi e os demais já agrupados, da seguinte maneira:

$$d_{(Hu,Ch,Go,Or)Gi} = z = 0,123$$



Método do Vizinho Próximo

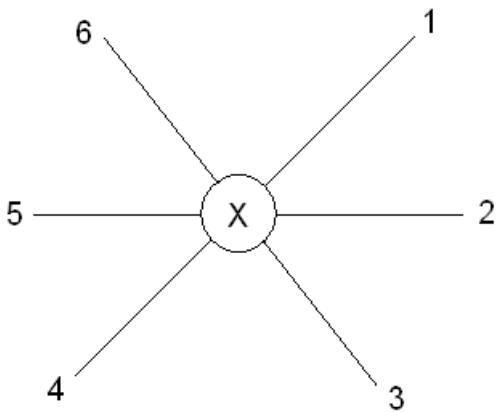
O método do vizinho próximo ou *neighbor joining* (NJ) é considerado uma versão simplificada do método da evolução mínima (ME), uma vez ele não requer o exame de todas as topologias possíveis, porém a cada estágio de agrupamento das OTUs o princípio da evolução mínima é utilizado.

Um conceito importante do método NJ é o de **vizinhos** ou **neighbors**, que são definidos como dois táxons conectados por um único nó em uma árvore não-enaizada. O conceito também é estendido à combinação de táxons, que passa a ser considerado um novo táxon (composto), que é conectado, também por um nó, a um terceiro táxon, seja ele simples ou composto.

Uma vantagem do NJ em relação ao UPGMA é que ele não requer o pressuposto de taxas de evolução constantes. Por isso, a topologia final da árvore apresenta comprimentos diferentes dos ramos, que representam variações nas taxas evolutivas. Por outro lado, por ser um método heurístico de estimar uma árvore curta, é possível que o método nem sempre encontre a árvore mais curta de todas para um particular conjunto de dados.

O algoritmo do NJ é mais complexo que do que o de UPGMA, mas em linhas gerais opera por meio da adição sucessiva de táxons a um núcleo inicial, sempre incorporando novos táxons, de modo a minimizar o aumento do tamanho da árvore. Busca-se obter uma topologia final cuja soma de ramos seja mínima.

O algoritmo NJ considera, inicialmente, uma árvore ‘estrela’ ou nula que é produzida sob a pressuposição de ausência de agrupamento dos táxons analisados. Como ilustração, será considerada a construção de uma árvore filogenética a partir de seis táxons; assim, tem-se a seguinte árvore nula:



Nesta árvore, a soma dos braços é dada por:

$$S_0 = L_{1X} + L_{2X} + \dots + L_{6X} = \sum_{i=1}^m L_{ix}$$

sendo m o número de táxons

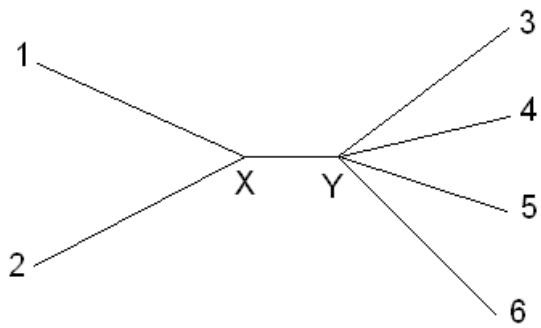
No entanto, sabe-se que:

$$\begin{aligned}
 L_{1X} + L_{2X} &= d_{12} \\
 L_{1X} + L_{3X} &= d_{13} \\
 &\dots \\
 L_{5X} + L_{6X} &= d_{56} \\
 \hline
 (m-1) \sum_{i=1}^m L_{ix} &= T = \sum_{i<j}^m d_{ij}
 \end{aligned}$$

logo:

$$S_0 = \sum_{i=1}^m L_{ix} = \frac{T}{m-1}$$

Outra árvore a ser considerada é aquela em que se estabelece com a associação de um par de táxons potencialmente mais próximos – por exemplo, 1 e 2 – e avalia-se a distância destes em relação aos demais e, posteriormente, calculase a soma de braços da árvore estabelecida. Assim, tem-se:



A soma dos braços desta árvore, denotada por S_{12} , é dada por:

$$S_{12} = L_{1X} + L_{2X} + L_{XY} + \sum_{i=3}^m L_{iY}$$

Verifica-se que:

- $L_{1X} + L_{2X} = d_{12}$
- O valor de L_{XY} é obtido considerando:

$$d_{13} = L_{1X} + L_{XY} + L_{Y3}$$

$$d_{14} = L_{1X} + L_{XY} + L_{Y4}$$

$$d_{15} = L_{1X} + L_{XY} + L_{Y5}$$

$$d_{16} = L_{1X} + L_{XY} + L_{Y6}$$

$$\text{Definindo } R_1 = \sum_{j=1}^m d_{1j}$$

obtém-se:

$$R_1 - d_{12} = (m - 2)(L_{1X} + L_{XY}) + \sum_{i=3}^m L_{iY} \quad (1)$$

Também tem-se:

$$d_{23} = L_{2X} + L_{XY} + L_{Y3}$$

$$d_{24} = L_{2X} + L_{XY} + L_{Y4}$$

$$d_{25} = L_{2X} + L_{XY} + L_{Y5}$$

$$d_{26} = L_{2X} + L_{XY} + L_{Y6}$$

Definindo $R_2 = \sum_{j=1}^m d_{2j}$

tem-se:

$$R_2 - d_{12} = (m - 2)(L_{2X} + L_{XY}) + \sum_{i=3}^m L_{iY} \quad (2)$$

Somando as equações (1) e (2), obtém-se:

$$L_{XY} = \frac{1}{2(m-2)} \left(R_1 + R_2 - md_{12} - 2 \sum_{i=3}^m L_{iY} \right)$$

e

$$\sum_{i=3}^m L_{iY} = \frac{T - R_1 - R_2 + d_{12}}{m-3}$$

Assim, conclui-se que:

$$S_{12} = \frac{2T - R_1 - R_2}{2(m-2)} + \frac{d_{12}}{2}$$

Como não se sabe *a priori* quais pares de táxons são vizinhos de fato, o procedimento descrito anteriormente é adotado para todas as combinações de pares de táxons, definindo-se como vizinhos o par i e j que produza o menor valor de S_{ij} .

Aplicação

Será considerado o estabelecimento de uma árvore filogenética a partir das medidas evolucionárias de distância entre seis táxons. Os valores de distância entre esses táxons são apresentados a seguir:

Táxon	1	2	3	4	5	6	R_i
1	0	9	12	15	20	16	72
2	9	0	7	10	15	11	52
3	12	7	0	5	10	6	40
4	15	10	5	0	11	7	48
5	20	15	10	11	0	8	64
6	16	11	6	7	8	0	48
Total							324

$$\sum_{i=1}^6 R_i = 324 \quad T = (1/2) \sum_{i=1}^6 R_i = 162$$

Para estabelecer a árvore filogenética, são necessários os seguintes passos:

Passo 1

Inicialmente, calcula-se a soma de braços de cada árvore topológica para todos os pares de táxons agrupados, conforme descrito a seguir:

$$S_{12} = (324 - 72 - 52)/8 + (9/2) = 29,5$$

$$S_{13} = (324 - 72 - 40)/8 + (12/2) = 32,5$$

...

$$S_{56} = (324 - 64 - 48)/8 + (8/2) = 30,5$$

A matriz auxiliar, de dimensão 6x6, contendo as informações da soma dos tamanhos dos braços das árvores filogenéticas, tomando cada par de táxons como vizinhos mais próximos, é dada por:

$$S = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & 29,5 & 32,5 & 33,0 & 33,5 & 33,5 \\ 2 & & 0 & 32,5 & 33,0 & 33,5 & 33,5 \\ 3 & & & 0 & 32,0 & 32,5 & 32,5 \\ 4 & & & & 0 & 32,0 & 32,0 \\ 5 & & & & & 0 & 30,5 \\ 6 & & & & & & 0 \end{bmatrix}$$

Nesta matriz S, constata-se que os táxons mais próximos são o 1 e o 2, cujo valor S_{12} (29,5) é o menor de todos.

Passo 2

É agora estabelecida uma nova matriz de distâncias $D_{5 \times 5}$, entre os táxons remanescentes e o grupo estabelecido. Essa distância é obtida pela expressão:

$$d_{(ij)k} = \frac{d_{ik} + d_{jk} - d_{ij}}{2}$$

Assim:

$$d_{(12)3} = d_{A3} = \frac{d_{13} + d_{23} - d_{12}}{2} = \frac{12 + 7 - 9}{2} = 5$$

$$d_{(12)4} = d_{A4} = \frac{d_{14} + d_{24} - d_{12}}{2} = \frac{15 + 10 - 9}{2} = 8$$

...

$$d_{(12)6} = d_{A6} = \frac{d_{16} + d_{26} - d_{12}}{2} = \frac{16 + 11 - 9}{2} = 9$$

de forma que:

	A	3	4	5	6	R_i
A	0	5	8	13	9	35
3	5	0	5	10	6	26
D = 4	8	5	0	11	7	31
5	13	10	11	0	8	42
6	9	6	7	8	0	30

Nesta matriz, tem-se:

$$\sum_{i=1}^6 R_i = 164 \quad \text{e} \quad T = (1/2) \sum_{i=1}^6 R_i = 82$$

A matriz auxiliar contendo a informação das somas de braços para cada par de vizinhos mais próximos é dada por:

$$S_{A3} = (164 - 35 - 26)/6 + (5/2) = 19,67$$

$$S_{A4} = (164 - 35 - 31)/6 + (8/2) = 20,30$$

...

$$S_{56} = (164 - 42 - 30)/6 + (8/2) = 19,30$$

Dessa forma, S fica estabelecida por:

$$S = \begin{matrix} & A & 3 & 4 & 5 & 6 \\ A & \left[\begin{matrix} 0 & 19,7 & 20,3 & 21,0 & 21,0 \\ 3 & 0 & 20,3 & 21,0 & 21,0 \\ 4 & & 0 & 20,7 & 20,7 \\ 5 & & & 0 & 19,3 \\ 6 & & & & 0 \end{matrix} \right] \end{matrix}$$

Constata-se, portanto, que os acessos mais próximos são o 5 e 6, com soma de braços dada por 19,3. Esses táxons formam, então, o grupo B.

Passo 3

Uma nova matriz de distância é construída, a partir das informações:

$$d_{BA} = d_{(56)A} = \frac{d_{5A} + d_{6A} - d_{56}}{2} = \frac{13 + 9 - 8}{2} = 7$$

$$d_{B3} = d_{(56)3} = \frac{d_{53} + d_{63} - d_{56}}{2} = \frac{10 + 6 - 8}{2} = 4$$

$$d_{B4} = d_{(56)4} = \frac{d_{54} + d_{64} - d_{56}}{2} = \frac{11 + 7 - 8}{2} = 5$$

A matriz de distância é, então, fornecida por:

$$D = \begin{matrix} & A & 3 & 4 & B & R_i \\ A & \left[\begin{matrix} 0 & 5 & 8 & 7 & 20 \\ 3 & 5 & 0 & 5 & 14 \\ 4 & 8 & 5 & 0 & 18 \\ B & 7 & 4 & 5 & 0 & 16 \end{matrix} \right] \end{matrix}$$

Nesta matriz, tem-se:

$$\sum_{i=1}^6 R_i = 68 \text{ e} \quad T = (1/2) \sum_{i=1}^6 R_i = 34$$

A partir desta matriz, estimam-se os valores dos elementos da matriz auxiliar S, dados por:

$$S_{A3} = (68 - 20 - 14)/4 + (5/2) = 11,0$$

$$S_{A4} = (68 - 20 - 18)/4 + (8/2) = 11,5$$

...

$$S_{4B} = (68 - 18 - 16)/4 + (5/2) = 11,0$$

A matriz auxiliar é, então, dada por:

$$S = \begin{matrix} & \begin{matrix} A & 3 & 4 & B \end{matrix} \\ \begin{matrix} A \\ 3 \\ 4 \\ B \end{matrix} & \begin{bmatrix} 0 & 11,0 & 11,5 & 11,5 \\ 0 & 0 & 11,5 & 11,5 \\ 0 & 0 & 0 & 11,0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

Os elementos mais próximos são o táxon 3 e o grupo A, formando-se, portanto, o grupo C.

Passo 4

Novamente, obtém-se a matriz de distância, por meio das informações:

$$d_{C4} = d_{(A3)4} = \frac{d_{A4} + d_{34} - d_{A3}}{2} = \frac{8 + 5 - 5}{2} = 4$$

$$d_{CB} = d_{(A3)B} = \frac{d_{AB} + d_{3B} - d_{A3}}{2} = \frac{7 + 4 - 5}{2} = 3$$

A matriz de distância é, então, fornecida por:

$$D = \begin{matrix} & \begin{matrix} C & 4 & B \end{matrix} & R_i \\ \begin{matrix} C \\ 4 \\ B \end{matrix} & \begin{bmatrix} 0 & 4 & 3 \\ 4 & 0 & 5 \\ 3 & 5 & 0 \end{bmatrix} & \begin{matrix} 7 \\ 9 \\ 8 \end{matrix} \end{matrix}$$

Nesta matriz, tem-se:

$$\sum_{i=1}^6 R_i = 24 \text{ e } T = (1/2) \sum_{i=1}^6 R_i = 12$$

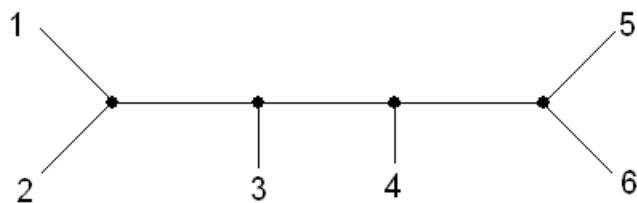
A matriz S_{3x3} é, então, dada pelos elementos:

$$S_{C4} = (24 - 7 - 9)/2 + (4/2) = 6,0$$

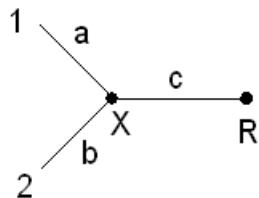
$$S_{CB} = (24 - 7 - 8)/2 + (3/2) = 6,0$$

$$S_{4B} = (24 - 9 - 8)/2 + (5/2) = 6,0$$

A confecção da árvore poderia ser estabelecida unicamente com as informações dadas até o passo 3, em que se obteriam dados:



Um problema adicional é a estimativa dos comprimentos dos braços. Para isso, admite-se a ramificação:



As seguintes equações são estabelecidas:

$$d_{ij} = a + b$$

$$d_{iR} = a + c = \frac{R_i - d_{12}}{m - 2}$$

$$d_{jR} = b + c = \frac{R_j - d_{12}}{m - 2}$$

de forma que:

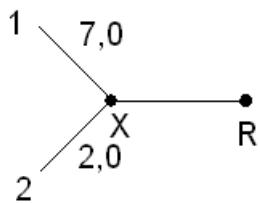
$$a = \frac{1}{2(m-2)} [(m-2)d_{ij} - R_i - R_j]$$

$$b = \frac{1}{2(m-2)} [(m-2)d_{ij} - R_j - R_i]$$

No passo 1, tem-se:

$$a = \frac{1}{8} [4(9) + 72 - 52] = 7,0$$

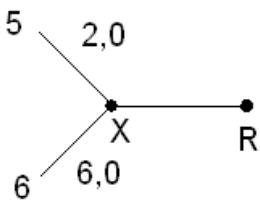
$$b = \frac{1}{8} [4(9) + 52 - 72] = 2,0$$



No passo 2:

$$a = \frac{1}{6} [3(8) - 42 + 30] = 2,0$$

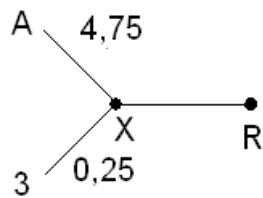
$$b = \frac{1}{6} [3(8) - 30 + 42] = 6,0$$



No passo 3:

$$a = \frac{1}{4} [2(5) - 26 + 35] = 4,75$$

$$b = \frac{1}{4} [2(5) + 26 - 34] = 0,25$$



Uma crítica que se faz ao método NJ é de que, embora ele seja baseado no princípio da evolução mínima, apenas uma topologia final com estimativas de comprimento de ramos é obtida. Por isso, alguns autores argumentam que o método da evolução mínima (ME) é preferível ao NJ. Por outro lado, o ME é muito mais trabalhoso e só produz resultados confiáveis se o número de nucleotídeos investigado for suficientemente elevado e um método não-viesado de estimar a taxa de substituição de nucleotídeos for utilizado como medida de distância. Em resumo, o NJ é um método rápido de construção de árvores filogenéticas, sendo apropriado para analisar um grande conjunto de dados, que pode ser testado por meio de *bootstrap*.

Método dos Quadrados Mínimos

A metodologia dos mínimos quadrados (MQ) busca, entre todas as topologias possíveis, aquela que minimiza a soma de quadrados residuais (R_s), a qual resulta da seguinte equação:

$$R_s = \sum_{i < j} w_{ij} (d_{ij} - \phi_{ij})^2$$

em que:

d_{ij} : distância observada entre os táxons i e j;

ϕ_{ij} : distância predita ou distância patrística entre os táxons i e j; e

w_{ij} : fator de ponderação específico do método MQ empregado.

A distância patrística entre i e j é a soma das estimativas dos comprimentos de todos os ramos que conectam os dois táxons na árvore. O cálculo de R_s preconizado por Cavalli-Sforza e Edwards (1967) assume que $w_{ij} = 1$, enquanto

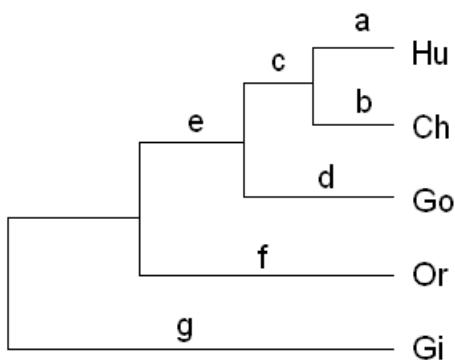
Fitch e Margoliash (1967) sugerem que $w_{ij} = \frac{1}{d_{ij}^2}$; assim, a equação para o cálculo de

R_s pode ser reescrita da seguinte forma:

$$R_s^* = \sum_{i < j} \left[\frac{(d_{ij} - \phi_{ij})^2}{d_{ij}} \right]$$

Na prática, entretanto, os resultados obtidos com e sem ponderação produzem árvores com topologias muito similares.

O problema básico dos métodos MQ é a estimativa dos comprimentos dos ramos de uma determinada topologia para, em seguida, computar a soma de quadrados residuais (R_s) desta topologia. Uma maneira de estimar os comprimentos dos ramos é através da solução de um sistema de equações lineares, que pode ser obtida tanto por método exato quanto iterativo. No caso do MQ originalmente proposto por Cavalli-Sforza e Edwards, cujo fator de ponderação é $w_{ij} = 1$, o sistema de equações é relativamente simples. Assim, como ilustração, pode-se considerar a árvore genealógica:



Para estimar os valores de a, b,..., f e g, adota-se o sistema de equações:

$$d = A\beta + \varepsilon$$

em que:

d: vetor de distâncias, de dimensão $[m(m-1)/2] \times 1$;

A: matriz de incidência, de dimensão $[m(m-1)/2] \times (2m-3)$;

β : vetor de parâmetros a serem estimados $(2m-3) \times 1$; e

ε : vetor de erros aleatórios, de dimensão $[m(m-1)/2] \times 1$.

Assim, tem-se:

Taxas	d_{ij}	a	b	c	d	e	f	g
Hu Ch	0,095	1	1	0	0	0	0	0
Hu Go	0,113	1	0	1	1	0	0	0
Hu Or	0,183	1	0	1	0	1	1	0
Hu Gi	0,212	1	0	1	0	1	0	1
Ch Go	0,118	0	1	1	1	0	0	0
Ch Or	0,201	0	1	1	0	1	1	0
Ch Gi	0,225	0	1	1	0	1	0	1
Go Or	0,195	0	0	0	1	1	1	0
Go Gi	0,225	0	0	0	1	1	0	1
Or Gi	0,222	0	0	0	0	0	1	1

E os parâmetros são estimados a partir de:

$$\hat{\beta} = (A'A)^{-1} A'd$$

De forma que:

Taxas	d_{ij} observada	ϕ_{ij} distância patrística
Hu Ch	0,095	0,096
Hu Go	0,113	0,110
Hu Or	0,183	0,186
Hu Gi	0,212	0,214
Ch Go	0,118	0,122
Ch Or	0,201	0,198
Ch Gi	0,225	0,226
Go Or	0,195	0,204
Go Gi	0,225	0,232
Or Gi	0,222	0,222

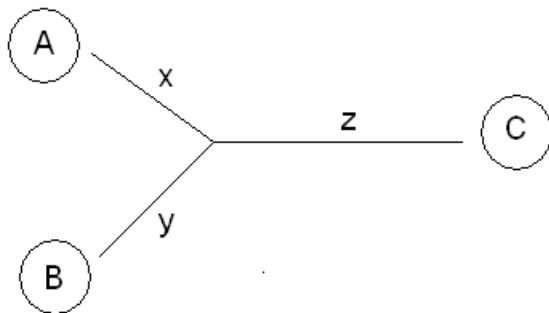
Obtendo-se:

$$R_s = 0,000047$$

Na prática, entretanto, a estimativa de comprimentos dos ramos por quadrados mínimos através de álgebra de matrizes pode não ser simples, requerendo grande quantidade de tempo computacional quando o número de seqüências a serem analisadas se torna elevado. Por isso, alguns algoritmos alternativos são empregados no processo de estimativa dos comprimentos dos ramos para as diversas árvores que precisam ser analisadas.

Algoritmo Fitch e Margoliash

Como visto anteriormente, Fitch e Margoliash (1967) apresentam uma maneira de estimar os comprimentos dos ramos, a partir das informações da seguinte estrutura de árvore:



As três incógnitas x , y e z podem ser estimadas a partir das informações das três distâncias d_{AB} , d_{AC} e d_{BC} , de forma que se tenha:

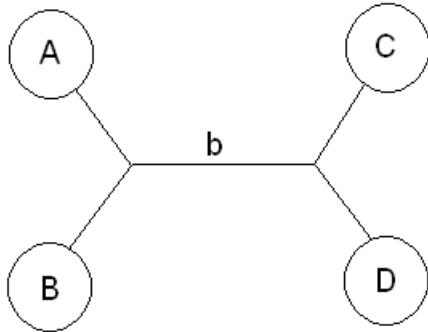
$$x = \frac{d_{AB} + d_{AC} - d_{BC}}{2}$$

$$y = \frac{d_{AB} + d_{BC} - d_{AC}}{2}$$

$$z = \frac{d_{AC} + d_{BC} - d_{AB}}{2}$$

Algoritmo Rzhetsky-Nei

Outra maneira de estimar os comprimentos dos ramos sem recorrer à álgebra de matrizes foi proposta por Rzhetsky e Nei (1993). No esquema a seguir, o comprimento do ramo assinalado por b pode ser estimado pela equação:



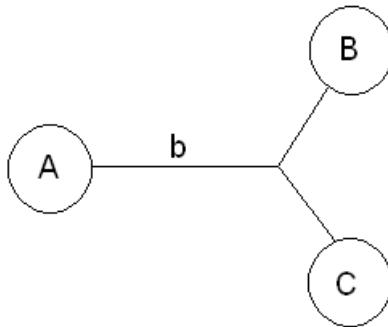
$$\hat{b} = \frac{1}{2} [\gamma(\bar{d}_{AC} + \bar{d}_{BC}) + (1-\gamma)(\bar{d}_{BC} + \bar{d}_{AD}) - (\bar{d}_{AB} + \bar{d}_{CD})]$$

sendo:

$$\gamma = \frac{m_B m_C + m_A m_D}{(m_A + m_B)(m_C + m_D)}$$

Aqui, m_A , m_B , m_C e m_D referem-se ao número de seqüências nos táxons ou agrupamentos A, B, C, e D, respectivamente.

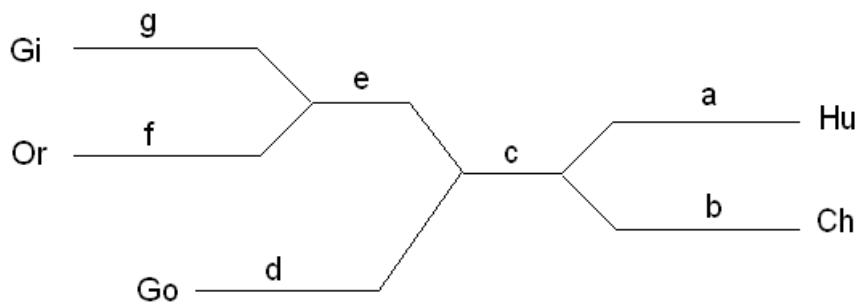
Para o seguinte esquema, tem-se:



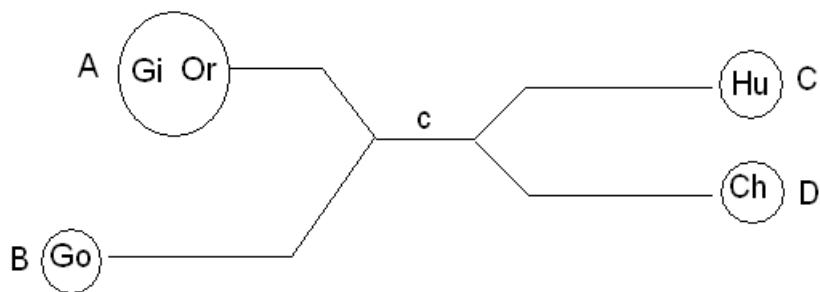
Assim:

$$\hat{b} = \frac{1}{2} \left(\frac{d_{AB}}{m_B} + \frac{d_{AC}}{m_C} - \frac{d_{BC}}{m_B m_C} \right)$$

Como exemplo, pode-se considerar:



Para estimação de c, por exemplo, tem-se:



Assim:

$$d_{AB} = d_{(Gi,Or)Go} = \frac{0,195 + 0,225}{2} = 0,210$$

$$d_{AC} = d_{(Gi,Or)Hu} = \frac{0,212 + 0,183}{2} = 0,1975$$

$$d_{AD} = d_{(Gi,Or)Ch} = \frac{0,225 + 0,201}{2} = 0,213$$

$$d_{BC} = d_{Go,Hu} = 0,113$$

$$d_{BD} = d_{Go,Ch} = 0,118$$

$$d_{CD} = d_{Hu,Ch} = 0,095$$

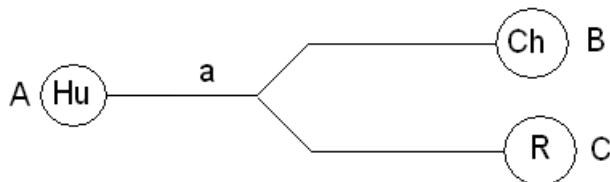
logo:

$$\gamma = \frac{1x1 + 2x1}{(3)(2)} = 0,5$$

então:

$$c = \frac{1}{2} \left[0,5 \left(\frac{0,1975}{2} + \frac{0,118}{1} \right) + 0,5 \left(\frac{0,113}{1} + \frac{0,213}{2} \right) - (0,210 + 0,095) \right] = 0,009065$$

Para obtenção da estimativa de a , considera-se:



$$d_{AB} = 0,095$$

$$d_{AC} = d_{Hu(Go,Or,Gi)} = \frac{0,113 + 0,183 + 0,212}{3} = 0,1693$$

$$d_{BC} = d_{Ch(Go,Or,Gi)} = \frac{0,118 + 0,201 + 0,225}{3} = 0,1813$$

logo

$$a = \frac{1}{2} \left(\frac{0,095}{1} + \frac{0,1693}{3} - \frac{0,1813}{(1)(3)} \right) = 0,045$$

O MQ é um método estatístico de estimação de parâmetros muito bem estabelecido. Quando as variáveis são de distribuição normal, ele é tão eficiente quanto o método de máxima verossimilhança. No caso em questão, se o número de nucleotídeos ou aminoácidos for grande, os comprimentos de ramos (b_i) devem seguir distribuição normal. Assim, espera-se que o método MQ forneça estimativas confiáveis dos comprimentos de ramos.

Por outro lado, deve ser lembrado que o interesse primário é determinar a topologia da árvore; se a topologia estiver incorreta, as estimativas de comprimentos dos ramos não terão significado biológico. Entretanto, as equações utilizadas no método MQ não permitem estimar uma topologia, por isso não há como definir diretamente qual é a topologia correta. Na prática, o que se faz é assumir que aquela topologia cujas estimativas de comprimentos de ramos são mais próximas

do observado deve ser uma boa topologia. Adicionalmente, se estimativas não-viesadas de distâncias evolutivas são utilizadas, e o número de nucleotídeos ou aminoácidos torna-se infinitamente grande, o valor de R_s deverá ser zero para a topologia correta.

Método da Evolução Mínima (ME)

O algoritmo da evolução mínima procura a árvore com a menor soma total dos comprimentos de ramos. Por isso, a soma (S) das estimativas de comprimentos de todos os ramos é computada para todas as topologias ou todas as topologias plausíveis; a topologia que apresentar o menor valor de S é escolhida como a melhor árvore. Embora o método ME possua boas propriedades estatísticas, assim como o método de mínimos quadrados, ele pressupõe o exame de todas as topologias possíveis, resultando em consumo de substancial tempo computacional quando o número de OTUs é grande. Por isso, uma alternativa é construir, inicialmente, uma árvore pelo método NJ e, então, examinar um grupo de topologias próximas a esta para encontrar a árvore com o menor valor de S . As bases teóricas dessa estratégia é que a árvore de ME geralmente é idêntica próxima à árvore obtida por NJ, quando o número de OTUs é pequeno. Assim, uma árvore NJ pode ser utilizada como árvore inicial na busca pela árvore ME, quando o número de OTUs for grande.

Na estratégia de busca exaustiva, para cada topologia é obtida a soma de todos os comprimentos dos braços, de forma que se tenha:

$$S_k = \sum_{i=1}^{n_b} \hat{b}_i$$

sendo:

$$k=1,2,\dots, N_i;$$

$$n_b = \text{número de braços} = 2m - 3; \text{ e}$$

$$m = \text{número de taxas.}$$

Admitindo que o comprimento dos braços seja obtido pelo método dos quadrados mínimos, tem-se:

$$\hat{\beta} = (A' A)^{-1} A' d = Ld$$

e a soma de braços é dada por:

$$S_{\text{braço}} = U\hat{\beta} = ULd$$

sendo U um vetor linha de uns.

Sejam duas topologias A e B , então:

S_A : soma dos braços da topologia A ; e

S_B : soma de braços da topologia B .

A diferença da soma de braços das topologias A e B é dada por:

$$\delta = S_B - S_A = UL_A d - UL_B d = U(L_A - L_B)d$$

Se $\delta > 0$, então a topologia B é melhor que A , e, caso contrário, se $\delta < 0$.

7.4.2. Métodos de Reconstrução Baseados em Máxima Parcimônia

Os métodos de máxima parcimônia (MP) foram originalmente desenvolvidos para aplicação em caracteres morfológicos. Posteriormente, algoritmos MP mais rigorosos foram desenvolvidos para aplicação em dados moleculares (nucleotídeos ou aminoácidos). Nesta classe de métodos MP, quatro ou mais seqüências alinhadas ($m \geq 4$) são consideradas, e os nucleotídeos (ou aminoácidos) de um táxon ancestral são inferidos separadamente para cada sítio de uma dada topologia, sob a pressuposição de que as mutações ocorrem em todas as direções entre os quatro nucleotídeos. Dessa forma, o menor número de substituições que explique o processo evolutivo para a topologia em consideração é computado. Essa computação é feita para todas as topologias potencialmente corretas, e aquelas que requerem o menor número de substituições são escolhidas como as melhores árvores. A base teórica desse procedimento é sustentada pela idéia filosófica de que a melhor hipótese para explicar um processo é aquela baseada em um menor número de pressuposições. Uma vez que qualquer modelo matemático, atualmente utilizado na reconstrução de filogenias, é uma aproximação grosseira da realidade, modelos livres de pressuposições podem produzir árvores mais confiáveis.

Quando não haver substituições reversas nem paralelas (homoplasias) em cada um dos sítios e o número de nucleotídeos examinados (n) for grande, espera-se que os métodos MP produzam árvores (realizadas) corretas. Na prática, entretanto, seqüências de nucleotídeos estão, freqüentemente, sujeitas a substituições reversas e paralelas, associadas à utilização de um valor de n relativamente pequeno nas análises. Nestes casos, os métodos MP tendem a produzir topologias incorretas. Outro problema com os métodos MP é o fenômeno denominado de atração de ramos longos ou de ramos curtos, em que os ramos mais longos (ou os ramos mais curtos) tendem a se juntar na árvore reconstruída, mesmo que isso não reflita a topologia correta.

Por outro lado, os métodos MP apresentam algumas vantagens em relação a outros métodos de reconstrução de árvores:

- a) Eles são relativamente livres de várias pressuposições requeridas na aplicação de métodos baseados em distâncias ou de máxima verossimilhança.
- b) Análises baseadas em máxima parcimônia são muito úteis para alguns tipos de dados moleculares, como seqüências de inserção/deleção.

De acordo com Nei e Kumar (2000), os métodos de MP desenvolvidos para dados moleculares podem ser divididos em duas categorias:

- a) MP não-ponderados, que assumem que as taxas de substituições de nucleotídeos são iguais ou muito próximas em todas as direções.
- b) MP ponderados, que assumem que alguns tipos de substituições (transições) ocorrem com maior freqüência do que outras (transversões).

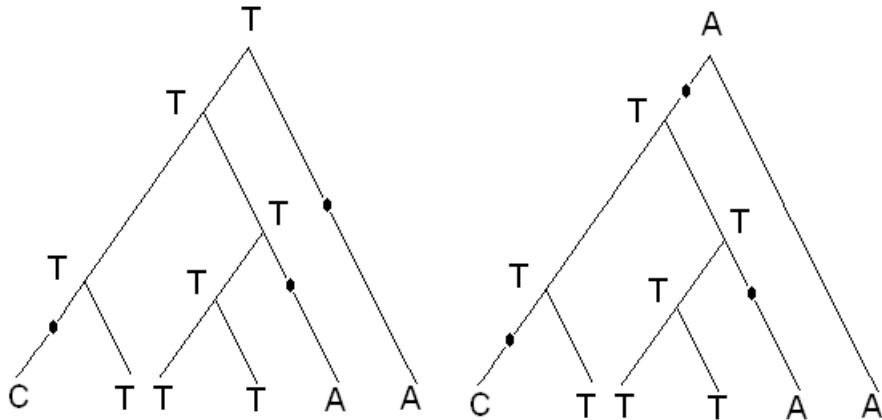
Neste último caso, é razoável atribuir diferentes pesos para os diferentes tipos de substituições quando o número mínimo de substituições é computado para uma dada topologia.

Máxima Parcimônia não-Ponderada

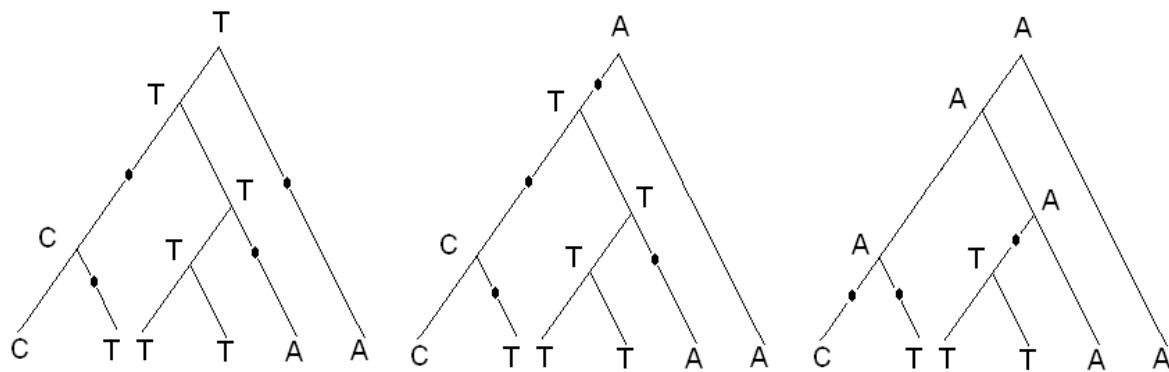
Para encontrar uma árvore de máxima parcimônia, é preciso determinar o número mínimo de substituições envolvidas, estabelecer o comprimento da árvore, distinguir os sítios informativos de sítios não-informativos e estimar o grau de homoplásia das seqüências analisadas.

a) Número mínimo de substituições

Como ilustração, será considerada a análise de um sítio em relação a seis táxons, cada um apresentando, respectivamente, as bases C, T, T, T, A e A. Será admitida a existência de cinco ancestrais, conforme figura a seguir:



Outras topologias também podem ser consideradas, como, por exemplo:

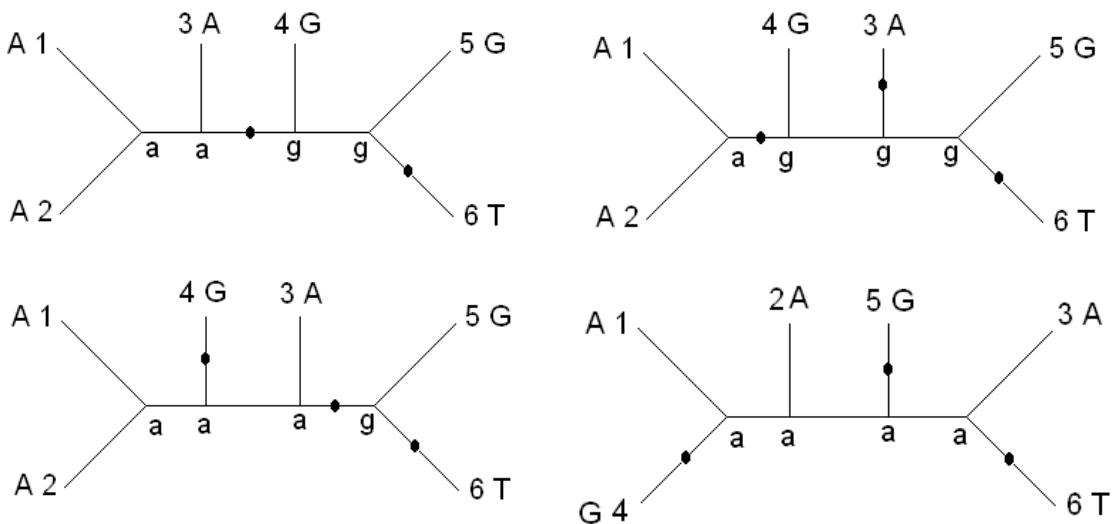


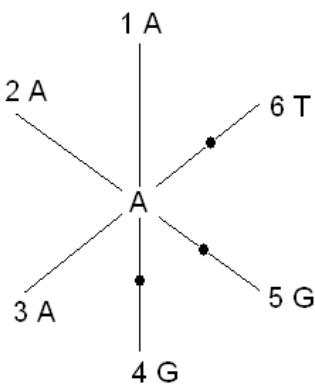
Veja que há várias possibilidades de estabelecer topologias com o mesmo número de substituições, porém com variação no nucleotídeo do ancestral. Apesar disso, é possível, em qualquer situação, contar o número de substituições gênicas.

b) Comprimento da árvore

No exemplo anterior, foi considerada apenas uma topologia, mas na prática devem-se considerar todas elas, ou pelo menos as potencialmente corretas, e determinar qual delas requer o menor número de substituições.

Será admitido agora, como exemplo, a análise de seis táxons, com a seguinte configuração em um sítio: A, A, A, G, G e T. Como se dispõe de seis táxons, o número total de topologias a serem analisadas é 105. A seguir são apresentadas cinco delas.





Apenas a primeira topologia apresentada requer duas substituições, enquanto as outras quatro requerem três substituições. É preciso computar todas as substituições para cada sítio de cada uma das 105 topologias possíveis, obtendo-se o comprimento de árvore (TL), dado por:

$$TL_k = \sum_{j=1}^n s_{kj}$$

sendo:

$$k = 1, 2, \dots, N_t; e$$

n: número de sítios.

A árvore de máxima parcimônia é aquela que apresenta o menor comprimento. Na prática, é possível que duas ou mais topologias tenham o mesmo número mínimo total de substituições; nesse caso, não será possível determinar uma única topologia. Por isso, admite-se que todas elas sejam árvores filogenéticas potencialmente corretas.

c) Sítios informativos

Quando se analisa um conjunto de táxons (m), em relação a vários sítios (n), devem ser considerados os sítios variáveis, informativos e não-informativos, e os sítios invariáveis, todos não-informativos. Apenas os sítios variáveis informativos são utilizados na análise MP.

O sítio variável não-informativo (*singleton*) é aquele cuja forma mutante aparece uma única vez. Como o número de substituições em qualquer topologia considerada será o mesmo, este sítio não permite identificar a melhor topologia. Deve ser ressaltado, entretanto, que o tipo de sítio poderá ser útil em outras metodologias. Por isso, seria mais apropriado denominá-los de sítios variáveis **parcimonialmente não informativos**.

Para que um **sítio seja informativo**, na construção de uma árvore MP, é necessário que ele tenha, no mínimo, dois tipos de nucleotídeos, cada um com pelo menos duas cópias. Como somente os sítios informativos contribuem para busca da árvore MP, é importante que haja muitos desses sítios para que a árvore seja confiável.

d) Homoplasia

A homoplasia refere-se ao fato de dois táxons serem geneticamente semelhantes, porém evolutivamente independentes, sendo originados em decorrência de retromutações ou de substituições paralelas. Por isso, quando há extensa ocorrência de homoplasias nas seqüências analisadas, as árvores MP não serão confiáveis, mesmo com a utilização de muitos sítios informativos. Com o intuito de contornar essas limitações, foram concebidos alguns índices para avaliar o grau de homoplasia de um grupo de seqüências, conforme descrito a seguir.

- Índice de consistência

É uma medida relativa, que leva em consideração o número mínimo de substituições que ocorrem em um j-ésimo sítio. É expressa por:

$$c_j = \frac{m_j}{s_j}$$

em que:

m_j : número mínimo possível de substituições no j-ésimo sítio, considerando todas as topologias; e

s_j : número mínimo de substituições para a topologia em análise.

Quando c_j é igual a 1, considera-se que a topologia, para este sítio, atende ao princípio da MP.

Quando se consideram todos os sítios, tem-se:

$$C = \frac{\sum_{j=1}^n m_j}{\sum_{j=1}^n s_j}$$

- Índice de Homoplasia

É definido por:

$$H = 1 - C$$

Quando não há retromutações nem substituições paralelas, o valor de C é igual a 1, e o de H, igual a zero. Neste caso, a topologia de MP será única.

- Índice de Retenção

Este índice varia de 0 a 1 e leva em consideração tanto o número mínimo quanto o máximo de substituições em cada sítio. É expresso por:

$$r_j = \frac{g_j - s_j}{g_j - m_j}$$

sendo g_j é o número máximo de substituições no j-ésimo sítio e pode ser estimado a partir da topologia estrela (topologia E), em que o nucleotídeo mais freqüente é colocado no ponto central da árvore filogenética.

O índice de retenção será igual a zero quando o sítio for o menos informativo para a construção da árvore, adotando-se o critério MP.

Quando se consideram todos os sítios, tem-se:

$$R = \frac{\sum_{j=1}^n g_j - \sum_{j=1}^n s_j}{\sum_{j=1}^n g_j - \sum_{j=1}^n m_j}$$

Em sistemática, $HI = 1 - CI$ é denominado de índice de homoplasia. Quando não há substituições reversas ou paralelas, $CI=1$ e $HI=0$, a topologia tem determinação única.

- Índice de consistência corrigido

É dado pelo produto entre o índice de consistência e o índice de retenção, ou seja:

$$rc_j = r_j c_j$$

Quando são considerados todos os sítios, tem-se:

$$RC = RxC$$

A seguir, são apresentados os valores encontrados para os índices relatados, considerando quatro topologias (A, B, C e D).

Topologia	s_j	c_j	r_j	rc_j
A	2	1	1	1
B	3	2/3	0	0
C	3	2/3	0	0
D	3	2/3	0	0

$$m_j = 2 \text{ e } g_j = 3$$

Estratégias de Busca de Árvores MP

Quando o número de seqüências ou táxons (m) é pequeno, isto é, $m < 10$, podem-se computar os comprimentos de todas as árvores possíveis e determinar qual é a de máxima parcimônia (busca exaustiva). Por outro lado, à medida que m aumenta ($m > 10$), uma busca exaustiva se torna impraticável. Por isso, foram desenvolvidos algoritmos que possibilitam encontrar árvores MP sem a necessidade de busca exaustiva.

Busca por ramo-e-ligação (branch-and-bound)

Neste método, árvores que tenham comprimentos maiores que uma outra, previamente examinada, são ignoradas, e a árvore MP é determinada pelo exame de árvores que, potencialmente, apresentem comprimentos menores. O método garante o encontro de todas as árvores MP, embora não seja um método exaustivo.

Existem várias versões de algoritmos para o método de ramo-e-ligação. Uma das mais comuns é a que se inicia com uma árvore nuclear não-enraizada de três táxons. Os táxons remanescentes são adicionados a esta árvore nuclear, um a um, de acordo com certa ordem. O comprimento da nova árvore é computado a cada estágio de adição de um novo táxon. Se a adição de um táxon, a um ramo particular da árvore nuclear, resultar em um comprimento maior do que um limite superior predeterminado (L_U), esta e todas as topologias, subsequentemente derivadas, são ignoradas nas investigações mais detalhadas.

Este algoritmo economiza muito tempo computacional, porque muitas árvores não precisam ser examinadas. No entanto, ele se torna muito demorado quando $m \geq 20$, sendo necessária a utilização de um algoritmo heurístico.

Busca Heurística

Neste método, somente uma pequena porção de todas as árvores possíveis é examinada, não havendo garantia de que a árvore MP será encontrada. Por outro lado, é possível elevar a probabilidade de obtenção da árvore MP pela utilização de vários algoritmos.

De modo geral, os algoritmos heurísticos, para busca de árvores MP, estão baseados no mesmo princípio. Inicialmente, uma árvore MP provisória é construída através de um procedimento denominado de algoritmo de adição *stepwise* (*stepwise addition algorithm*). Posteriormente, esta árvore provisória é submetida a algum tipo de troca de ramos, com a finalidade de se encontrar uma árvore mais parcimoniosa.

Nos algoritmos *stepwise*, uma árvore nuclear de três táxons é formada de acordo com certa regra (geralmente pelo triplete de menor comprimento), e cada táxon remanescente é adicionado a um dos três ramos da árvore nuclear. Os

comprimentos das árvores resultantes são computados, selecionando-se a árvore de quatro táxons com o menor comprimento. Esta árvore será utilizada na etapa posterior de adição de mais um táxon. O processo continua até a incorporação de todos os táxons, gerando a árvore provisória.

A troca de ramos geralmente é feita por um dos seguintes algoritmos:

- a) Troca entre vizinhos mais próximos (NNI) – que equivale a examinar todas as árvores que são diferentes da árvore provisória por uma distância topológica de $d_T=2$.
- b) Poda de subárvore e reenxerto (SPR) – neste algoritmo, um ramo da árvore provisória é cortado gerando uma subárvore e uma árvore residual. A subárvore podada é então reenxertada em cada ramo da árvore residual, gerando novas topologias. Aquela topologia de menor comprimento será escolhida como a árvore MP.
- c) Bissecção e reconexão de árvore (TBR) – neste procedimento uma árvore provisória é cortada em duas subárvores e reconectadas através da união de dois ramos, sendo consideradas todas as combinações de ramos entre as duas subárvores. Este procedimento possibilita o exame de maior número de topologias que os demais.

Árvore Consenso

Pelo fato de os métodos baseados em MP frequentemente produzirem várias árvores igualmente parcimoniosas, foram propostos métodos para combiná-las de modo a produzir uma única árvore consenso.

Uma árvore de **consenso estrito** é produzida através da fusão, em nós multifurcantes, de quaisquer padrões de ramificações conflitantes entre árvores MP rivais. Já em uma árvore de **consenso majoritário** o padrão de ramificação prevalente entre todas as árvores MP, segundo um limite preestabelecido, é adotado na representação da árvore consenso.

Estimação dos Comprimentos dos Ramos

Em geral, os métodos MP utilizados na construção de topologias de árvores não calculam os comprimentos dos ramos. Entretanto, é possível estimar tais valores assumindo certos pressupostos na reconstrução de uma árvore.

Para estimar os comprimentos dos ramos de uma árvore MP, são consideradas todas as rotas evolutivas de cada sítio variável e computados os números médios de substituições para cada ramo – interno e externo.

Método da Máxima Parcimônia Ponderada

A expectativa é de que métodos MP produzam árvores mais confiáveis quando a freqüência de homoplasias é baixa. Por isso, se um grupo de seqüências utilizado na análise filogenética inclui sítios de evolução lenta e sítios de evolução rápida, é esperado que os primeiros sejam mais úteis na reconstrução filogenética por MP do que os últimos, quando se analisam seqüências distivamente relacionadas. Por isso, se for atribuído um peso maior aos sítios de evolução lenta, pode-se obter uma árvore mais confiável do que se as duas classes de sítios fossem ponderadas igualmente.

De modo geral, as partes funcionalmente menos importantes de um gene evoluem mais rapidamente do que as mais importantes. Por isso, é plausível atribuir maior peso para substituições ocorridas em regiões mais importantes do que aquelas ocorridas em regiões menos importantes, quando seqüências de táxons distivamente relacionados são estudados.

O método MP ponderado também permite que diferentes pesos sejam atribuídos a diferentes tipos de substituições que ocorrem em um sítio. Por exemplo, é possível atribuir maior peso às transversões (que são mais raras) do que às transições (que são mais freqüentes).

Por outro lado, existem alguns problemas com a parcimônia ponderada. O principal é que, geralmente, não se sabe quais são os pesos apropriados a serem atribuídos aos dados sob investigação, embora, em alguns casos, seja possível utilizar informações de estudos prévios. Outro problema é que sítios que apresentem baixa taxa de substituição são informativos apenas quando se compararam seqüências distivamente relacionadas. Quando seqüências mais próximas são utilizadas, os sítios de elevada taxa de substituição são mais informativos. Entretanto, geralmente um conjunto de dados apresenta tanto seqüências próximas quanto distivamente relacionadas, e, neste caso, não está claro o quanto útil é o método MP ponderado.

Dados de Seqüências de Proteínas *versus* de Seqüências de Nucleotídeos na MP

Dados de seqüências de aminoácidos foram os primeiros utilizados na reconstrução filogenética por MP. Os autores do método consideraram os 20 aminoácidos como estados de caráter e assumiram que as mudanças evolutivas poderiam ocorrer em todas as direções entre eles. Teoricamente, essa estratégia é aproximada, porque a mudança de alguns aminoácidos requerer duas ou três substituições de nucleotídeos, enquanto outros necessitam de apenas uma substituição. Além do mais, alguns aminoácidos são mais similares bioquimicamente entre si; por isso, as substituições ocorrem mais frequentemente dentro de cada grupo de similaridade do que entre grupos.

Quando dados de seqüências de DNA se tornaram disponíveis nos anos de 1980, a maior parte dos trabalhos de filogenia foi, progressivamente, sendo direcionada para utilização deste tipo de dado. Por outro lado, gradualmente, tornou-se claro que o padrão de evolução do DNA é tão complexo que tais seqüências não são, necessariamente, melhores que seqüências de proteínas na inferência filogenética. Além da variação na taxa de substituição para diferentes posições do códon, o conteúdo de GC na terceira posição varia segundo a espécie considerada.

7.4.3. Método da Máxima Verossimilhança

Nos métodos de máxima verossimilhança (MV), a probabilidade de um particular grupo de seqüências observadas é maximizada, segundo um modelo evolutivo definido *a priori*, para cada uma das topologias possíveis; a topologia com máxima probabilidade de ter gerado os dados é escolhida como a árvore final. Os parâmetros a serem considerados não são as topologias, mas os comprimentos de ramos de cada uma delas, uma vez que a probabilidade é maximizada para estimar os comprimentos de ramos. Assim, o valor de verossimilhança para uma topologia é obtido pelo cálculo da probabilidade de ter ocorrido uma combinação de eventos evolutivos naquela topologia, capaz de resultar na distribuição de dados observados.

A construção de uma árvore por MV é muito demorada porque é necessário considerar todos os possíveis nucleotídeos em cada nó interno, assim como todas as topologias. Dessa forma, o número de combinações de nucleotídeos a ser examinado é dado por $4^{(m-2)}$, para cada árvore não-enzaizada, ou $4^{(m-1)}$, para árvores enraizadas, em que m é o número de táxons e $m-2$ (ou $m-1$) é o número de nós internos. Considerando uma árvore não-enzaizada com 10 táxons, o número de diferentes combinações de nucleotídeos a serem examinadas é de 65.536, e o número de topologias, de 2.027.025. Por isso, não é muito fácil obter uma árvore de MV global quando m é grande. Entretanto, existem algoritmos para facilitar essa operação.

Vários algoritmos heurísticos de busca foram desenvolvidos para métodos de reconstrução baseados em MV. Muitos desses algoritmos são similares àqueles utilizados para obtenção de árvores pelos métodos da evolução mínima (ME) e de máxima parcimônia (MP), como os algoritmos NNI (*nearest neighbor interchanges*) e o TBR (*tree-bisection-reconnection*). Por sua vez, a eficiência destes algoritmos em obter a topologia correta é, necessariamente, a mesma para os métodos ME, MP e MV. Assim como nos métodos ME e MP, os métodos MV tendem a produzir

topologias incorretas quando m é grande e n é pequeno, independentemente do algoritmo empregado. Por isso, não vale a pena dispensar um tempo computacional excessivo na busca pela árvore MV. O importante é encontrar a árvore verdadeira ou uma árvore que seja próxima a ela, em vez da árvore de máxima verossimilhança.

O método de máxima verossimilhança é um método estatístico bem estabelecido de estimação de parâmetros. Ele é conhecido por produzir a menor variância possível para o parâmetro estimado quando o tamanho amostral é grande. Por isso, este método é rotineiramente utilizado quando se deseja formular uma função de verossimilhança envolvendo parâmetros desconhecidos em um dado espaço probabilístico. A estimativa de valores de recombinação na genética clássica e a de freqüências alélicas na genética de populações são exemplos familiares da utilização de MV. A estimativa dos comprimentos dos ramos por métodos MV para uma dada topologia (verdadeira) pode ser também justificada, contando que o modelo correto de substituição seja utilizado. A estratégia empregada pelos algoritmos que implementam a análise MV é modificar repetidamente os valores dos comprimentos de ramos de uma árvore, até que se obtenha um valor que não pode ser melhorado. Dessa forma, são estimados os comprimentos dos ramos que maximizam a probabilidade de obter os dados observados.

Um dos principais problemas na utilização do método MV é a reconstrução da topologia da árvore. Nos métodos correntes de MV, aplicados à inferência filogenética, a função de verossimilhança não inclui parâmetros relativos às topologias. Por isso, não se estima uma topologia pela maximização da verossimilhança. Simplesmente se escolhe uma topologia com o maior valor de MV, sob a pressuposição de que a topologia com boas estimativas de comprimentos de ramos é, provavelmente, a da árvore verdadeira. Essa pressuposição não é, necessariamente, verdadeira. De fato, quando a taxa de substituição varia muito de ramo para ramo, topologias incorretas podem ser escolhidas mais freqüentemente que topologias verdadeiras, mesmo quando o número de nucleotídeos (n) examinados for grande. Também é possível encontrar exemplos nos quais os

métodos MV mostram menor probabilidade de obtenção da topologia correta do que os métodos MP e alguns métodos baseados em distâncias.

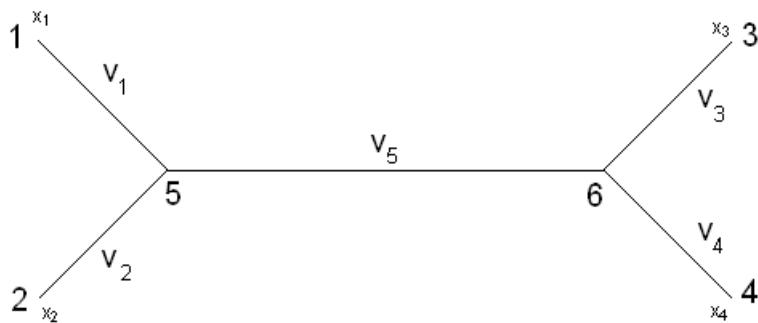
Embora um modelo sofisticado de substituição não implique, necessariamente, a melhoria da confiabilidade da topologia da árvore inferida, espera-se que ele melhore a acurácia das estimativas dos comprimentos dos ramos, medida pelo número de substituições de nucleotídeos ou de aminoácidos. Por essa razão, diversos autores desenvolveram métodos estatísticos para estimar os parâmetros envolvidos nos vários modelos de substituições. Quando todos os sítios evoluem independentemente segundo um mesmo modelo de substituição, uma maneira simples de estimar os parâmetros de substituição é inferir as seqüências ancestrais por métodos de máxima parcimônia e contar o número dos diferentes tipos de substituição.

Por outro lado, as estimativas obtidas por máxima parcimônia podem ser viesadas, pelo fato de as múltiplas substituições ocorridas em um sítio não serem levadas em conta. Teoricamente, uma maneira melhor é utilizar um método MV para uma topologia que seja, provavelmente, a correta. A função geral de verossimilhança é dada pela expressão $L=f(x; \theta)$, em que θ é o grupo de parâmetros a ser estimado. Por exemplo, o modelo de Kimura inclui os parâmetros de substituição α e β , assim como 2-3 parâmetros, para comprimentos de ramos. Desse modo, pode-se estimar o grupo de parâmetros θ pela maximização de L .

Outro ponto importante é que o padrão de substituição de nucleotídeos é muito complexo. Por isso, imagina-se que um modelo matemático com muitos parâmetros seja melhor que um com menor número de parâmetros, na construção de árvores filogenéticas. Na prática isso não é sempre o caso. Um modelo com muitos parâmetros ajusta-se melhor aos dados que um modelo mais simples, porém a predição estatística ou a estimativa de topologias baseada em um modelo com muitos parâmetros é mais sujeita a erros. Dessa forma, é aconselhável utilizar um modelo simples, desde que este modelo represente o padrão de substituição razoavelmente bem.

Determinação da Função de Verossimilhança

O processo consiste em estabelecer uma determinada topologia e avaliar o ajuste de um modelo a ela. Como exemplo, será considerada a análise de quatro táxons, denotados por 1, 2, 3 e 4, que apresentam para um particular sítio os nucleotídeos x_1 , x_2 , x_3 e x_4 . Assim, tem-se:



Nesta topologia não se conhece o nucleotídeo nos nós 5 e 6, que são assumidos serem x_5 e x_6 , podendo, obviamente, ser qualquer um dos quatro nucleotídeos: A,T, C e G.

Será considerado que o valor $P_{ij(t)}$ mede a probabilidade do nucleotídeo i , no tempo zero, tornar-se o nucleotídeo j , no tempo t , para um dado sítio. No método MV, assume-se que a taxa de substituição de um nucleotídeo, denotada por r , varia de ramo para ramo, sendo conveniente medir o tempo evolucionário em termos de valor esperado de substituição, ou seja:

$$d = v = rt$$

Na topologia especificada, é admitido que:

$$v_i = r_i t_i$$

sendo:

r_i : taxa de substituição; e

t_i : tempo evolucionário para o ramo i .

No método MV, v_i é o parâmetro a ser estimado e x_1, x_2, \dots, x_n são as observações disponíveis. A função de verossimilhança é dada por $L(v;x)$. Assim, para um sítio k, tem-se:

$$L_k(v;x) = \sum_{x_5} \sum_{x_6} g_{x_5} P_{x_5x_1}(v_1) P_{x_5x_2}(v_2) P_{x_5x_6}(v_5) P_{x_6x_3}(v_3) P_{x_6x_4}(v_4)$$

Por isso, para a topologia considerada, pode-se admitir, por exemplo, que $x_1 = A$, $x_2 = A$, $x_3 = A$ e $x_4 = A$. Assim, haverá possibilidade de $x_5 = A, G, C$ ou T e, também, de $x_6 = A, G, C$ ou T . O valor de $L(v;x)$ seria dado por:

$$L(v;x) = W_1 + W_2 + \dots + W_{16}$$

Os valores de W_i são apresentados a seguir, para as várias possibilidades de x_5 e x_6 :

x_5	x_6	
A	A	$W_1 = g_A P_{AA}(v_1) P_{AA}(v_2) P_{AA}(v_5) P_{AA}(v_3) P_{AA}(v_4)$
	G	$W_2 = g_A P_{AA}(v_1) P_{AA}(v_2) P_{AG}(v_5) P_{GA}(v_3) P_{GA}(v_4)$
	C	$W_3 = g_A P_{AA}(v_1) P_{AA}(v_2) P_{AC}(v_5) P_{CA}(v_3) P_{CA}(v_4)$
	T	$W_4 = g_A P_{AA}(v_1) P_{AA}(v_2) P_{AT}(v_5) P_{TA}(v_3) P_{TA}(v_4)$
G	A	$W_5 = g_G P_{GA}(v_1) P_{GA}(v_2) P_{GA}(v_5) P_{AA}(v_3) P_{AA}(v_4)$
	G	$W_6 = g_G P_{GA}(v_1) P_{GA}(v_2) P_{GG}(v_5) P_{GA}(v_3) P_{GA}(v_4)$
	C	$W_7 = g_G P_{GA}(v_1) P_{GA}(v_2) P_{GC}(v_5) P_{CA}(v_3) P_{CA}(v_4)$
	T	$W_8 = g_G P_{GA}(v_1) P_{GA}(v_2) P_{GT}(v_5) P_{TA}(v_3) P_{TA}(v_4)$
C	A	$W_9 = g_C P_{CA}(v_1) P_{CA}(v_2) P_{CA}(v_5) P_{AA}(v_3) P_{AA}(v_4)$
	G	$W_{10} = g_C P_{CA}(v_1) P_{CA}(v_2) P_{CG}(v_5) P_{GA}(v_3) P_{GA}(v_4)$
	C	$W_{11} = g_C P_{CA}(v_1) P_{CA}(v_2) P_{CC}(v_5) P_{CA}(v_3) P_{CA}(v_4)$
	T	$W_{12} = g_C P_{CA}(v_1) P_{CA}(v_2) P_{CT}(v_5) P_{TA}(v_3) P_{TA}(v_4)$
T	A	$W_{13} = g_T P_{TA}(v_1) P_{TA}(v_2) P_{TA}(v_5) P_{AA}(v_3) P_{AA}(v_4)$
	G	$W_{14} = g_T P_{TA}(v_1) P_{TA}(v_2) P_{TG}(v_5) P_{GA}(v_3) P_{GA}(v_4)$
	C	$W_{15} = g_T P_{TA}(v_1) P_{TA}(v_2) P_{TC}(v_5) P_{CA}(v_3) P_{CA}(v_4)$
	T	$W_{16} = g_T P_{TA}(v_1) P_{TA}(v_2) P_{TT}(v_5) P_{TA}(v_3) P_{TA}(v_4)$

Na expressão da função MV, apresentada anteriormente observa-se que:

- g_{x_5} é a probabilidade *a priori* de que o nó 5 apresente o nucleotídeo x_5 . Esse valor é freqüentemente estimado a partir da freqüência do próprio nucleotídeo, considerando todo o conjunto de seqüências estudadas, mas pode também ser estimado por procedimento MV.

$P_{ij(v)}$ é a probabilidade de substituição do nucleotídeo i pelo nucleotídeo j , que é função da taxa de substituição do nucleotídeo (r_i) e do tempo evolucionário t_i .

Para conhecer P_{ij} , deve-se adotar um modelo específico de substituição. Considerando o modelo de “equal-input”, tem-se:

$$P_{ii}(v) = g_i + (1-g_i)e^{-v}$$

e

$$P_{ij}(v) = g_j(1-g_i)e^{-v}$$

A função suporte, para um sítio, é dada por:

$$\ell_k(v;x) = \ln[L_k(v;x)]$$

e a função suporte considerando a análise conjunta de todos os n sítios pode ser expressa por:

$$\ell_k(v;x) = \sum_{k=1}^n \ln[L_k(v;x)]$$

O parâmetro v é estimado a partir da maximização da função suporte $\ell_k(v;x)$, por meio de procedimentos computacionais. Além de estimar os valores dos elementos do vetor v , é obtido o valor máximo da verossimilhança para cada uma das possíveis topologias.

7.5 Acurácia e Testes Estatísticos para Árvores Filogenéticas

Aspectos Gerais

Quando se constrói uma árvore filogenética, é importante conhecer sua confiabilidade. Existem dois tipos de erros em uma filogenia: os de topologia e os de comprimentos de ramos. Os primeiros se referem às diferenças nos padrões de ramificação entre a árvore inferida e a árvore verdadeira, e os últimos, aos desvios das estimativas dos comprimentos de ramos em relação aos verdadeiros comprimentos de ramos (realizados ou esperados). Na prática, a topologia verdadeira de uma árvore geralmente não é conhecida. Por isso, a confiabilidade da topologia obtida é usualmente testada pelo exame de várias partes da topologia (padrões de ramificação).

Mesmo que a topologia de uma árvore se mostre incorreta, deve-se testar a confiabilidade dos comprimentos de seus ramos. Isso porque os testes estatísticos dos comprimentos de ramos são importantes tanto no exame da acurácia da topologia quanto das estimativas dos próprios comprimentos de ramos. A confiabilidade dos comprimentos de ramos pode ser testada por métodos analíticos ou *bootstrap*.

Os princípios de maximização ou minimização utilizados nos métodos de reconstrução por máxima parcimônia (MP), evolução mínima (ME) e máxima verossimilhança (MV) tendem a produzir topologias incorretas quando o número de nucleotídeos ou aminoácidos investigados (n) é pequeno. Isso é causado por erros estocásticos de substituição e torna-se sério quando o número de seqüências é grande. Por outro lado, métodos baseados no princípio da otimização funcionam bem quando n é grande, a menos que surjam problemas de inconsistência. Erros topológicos também ocorrem quando alguns dos ramos internos são curtos e sujeitos a erros estocásticos. Esses ramos, entretanto, podem ser identificados por *bootstrap* ou por outro teste estatístico.

Se um ramo interno não é bem estabelecido, o padrão de ramificação associado a ele deveria ser registrado como não-resolvido. Caso tal estratégia seja adotada, ainda se pode continuar utilizando métodos baseados no princípio da otimização. Neste caso, entretanto, não é necessário empregar algoritmos de busca exaustivos ou heurísticos para encontrar a árvore ótima. Algoritmos de busca rápida parecem ser suficientes para a inferência de árvores filogenéticas, uma vez que as árvores inferidas estão sujeitas a testes estatísticos.

Testes Bootstrap

Um dos testes mais utilizados para averiguar a confiabilidade de uma árvore inferida é o *bootstrap*. A base do método consiste de reamostragens aleatórias, com reposição, dos dados. Em cada reamostragem, o número total de sítios amostrados é constante e igual ao número da amostra original que gerou a árvore inferida. No entanto, a cada evento de retirada de uma observação amostral todos os sítios apresentam a mesma probabilidade de serem amostrados. Conseqüentemente, alguns sítios podem ser escolhidos mais de uma vez, enquanto outros podem não ser escolhidos nenhuma vez na confecção de uma replicação amostral. Cada amostra gerada por *bootstrap* é empregada na construção de uma árvore réplica através do mesmo método de reconstrução utilizado na árvore inferida. A topologia desta árvore é comparada com a da árvore original. Atribui-se valor 1 aos ramos internos que apresentam partição equivalente à dos ramos da árvore original, e zero, em caso contrário. Esse processo é repetido centenas a milhares de vezes, computando-se as porcentagens que cada ramo interno da árvore original recebeu escore 1. Esses valores são denominados de valores de confiança de *bootstrap* ou valores de *bootstrap* (P_B). Em geral, se P_B é igual a 95% ou maior, o ramo interno é considerado como significativamente positivo.

Em uma outra versão do teste, uma árvore consenso das árvores réplicas é obtida. Nesse caso, se um grande número de sítios é utilizado e todos eles evoluem independentemente, por um mesmo modelo de substituição, esta árvore consenso

pode ser mais próxima da árvore esperada do que da árvore inferida a partir dos dados originais.

O teste *bootstrap* necessita construir uma árvore para cada conjunto de dados reamostrados. Por isso, o tempo computacional dessa operação torna-se um fator importante do *bootstrap*. Por essa razão, este teste é utilizado rotineiramente em árvores NJ, mas é muito demorado para árvores MP e MV, a menos que o número de seqüências seja pequeno.

As propriedades estatísticas do *bootstrap* são complicadas e não muito bem entendidas. Por outro lado, quando o teste é aplicado a uma árvore NJ, a interpretação dos resultados é relativamente simples. Se (1) cada sitio das seqüências de DNA evolui independentemente, (2) a medida de distância utilizada é um estimador não-viesado do número de substituições de nucleotídeos e (3) o número de seqüências (m) e o de sítios (n) são suficientemente grandes, a hipótese de nulidade do teste é de que o comprimento de cada ramo interno é zero. Por isso, a suposição testada é a de que o P_B para um ramo meça a probabilidade (P_C) de seu comprimento ser maior que zero.

O *bootstrap* é considerado um teste muito conservador, mas apropriado para ser aplicado em análises filogenéticas. Isso porque o padrão de substituição de nucleotídeos é muito complexo, mudando, com freqüência, entre sítios e períodos evolutivos.

Árvores Condensadas

Quando uma árvore filogenética possui baixos valores de P_C e P_B , para vários ramos internos, é útil produzir uma árvore multifurcada, assumindo que tais ramos apresentam comprimentos iguais a zero. A eliminação dos ramos internos de baixa significância produz uma árvore condensada, que enfatiza as porções confiáveis do padrão de ramificação. Por outro lado, não há métodos gerais de estimar os comprimentos de ramos de uma árvore condensada. Assim, geralmente se apresenta apenas a topologia de tais árvores.

Deve-se notar também que **árvore condensada** é diferente de **árvore consenso**. Uma árvore consenso é produzida a partir de muitas árvores igualmente parcimoniosas, enquanto uma árvore condensada é meramente uma apresentação simplificada de uma árvore, que pode ter sido produzida por qualquer tipo de método de reconstrução filogenética.

7.6 Vantagens e Desvantagens dos Diferentes Métodos de Reconstrução Filogenética

Há vários critérios para comparar diferentes métodos de reconstrução. Os mais importantes são:

- velocidade computacional;
- consistência como estimador para uma dada topologia;
- possuir propriedades que permitam a aplicação de testes estatísticos;
- probabilidade de recuperar a verdadeira topologia; e
- confiabilidade nas estimativas dos comprimentos dos ramos.

A velocidade computacional é uma medida relativamente simples, embora dependa do algoritmo e do equipamento utilizado. Por esse critério, UPGMA e NJ são superiores aos demais métodos atualmente utilizados. Estes dois métodos podem manipular um grande número de seqüências ($m > 500$), e testes *bootstrap* podem ser facilmente aplicados.

Um método de reconstrução é considerado um “estimador consistente” se ele tende a produzir a topologia correta, quando o número de nucleotídeos (n) amostrados tende para o infinito. Os métodos NJ, ME e MQ são todos estimadores consistentes, se estimativas não-viesadas de substituição de nucleotídeos são utilizadas como medidas de distâncias. No caso de MV, o método será consistente quando o modelo de substituição de nucleotídeos for correto. Por sua vez, o modelo MP é muitas vezes inconsistente, além de ser sujeito ao fenômeno da “atração dos ramos”.

Na prática, entretanto, n geralmente não é grande (variando de centenas a milhares de sítios apenas), e, neste caso, NJ, ME, MQ, MV e MP podem fracassar

na recuperação da árvore verdadeira. Por isso, a consistência não é um critério muito útil de comparação de métodos.

Atualmente, os métodos estatísticos de avaliação dos métodos NJ e ME estão bem estabelecidos, assim como para métodos MQ. Por outro lado, há muita dificuldade de empregar testes estatísticos de avaliação aplicáveis aos métodos MP e MV. O melhor teste para estes métodos ainda é o *bootstrap*.

A probabilidade de obtenção da topologia verdadeira é um dos critérios mais importantes na comparação de métodos de reconstrução, porém é o problema mais difícil a ser estudado. Isso porque é necessário conhecer a topologia verdadeira – conhecimento este raramente disponível, a não ser por simulação.

Outro critério importante para comparação de diferentes métodos é a confiabilidade das estimativas dos comprimentos dos ramos. Uma vez que a topologia correta tenha sido estabelecida, esse problema pode ser avaliado com relativa facilidade. Teoricamente, espera-se que MV, MQ, NJ e ME produzam estimativas de ramos mais confiáveis que MP. Atualmente, árvores obtidas por MP são apresentadas sem estimativas dos comprimentos de ramos, provavelmente, porque o método tende a produzir valores subestimados.

Nos últimos anos houve muito debate a respeito da eficiência relativa dos diferentes métodos de reconstrução de árvores, especialmente NJ, MP e MV. Está claro que não existe um método que seja superior a outro em todas as condições. Sabe-se que as bases teóricas da reconstrução de filogenia ainda não estão bem estabelecidas e que alguns métodos funcionam melhor que outros sob certas condições, sendo, porém, muito ruins em outras condições. Nas análises com dados reais, em que a extensão das divergências entre seqüências não é muito elevada e um substancial número delas é utilizados, os métodos NJ, MP e MV geralmente produzem topologias iguais ou muitos similares. Quando há diferenças, geralmente causadas por interarranjos de ramos internos, elas podem facilmente ser identificadas por testes *bootstrap*.

As informações supramencionadas sugerem que uma árvore construída com valores de *bootstrap*, em qualquer um destes três métodos (NJ, MV e MP), produz,

essencialmente, as mesmas conclusões a respeito das relações filogenéticas dos organismos ou genes sob investigação.

Pesquisadores são, com freqüência, ansiosos para prover a validade de uma árvore obtida. Entretanto, é recomendável que qualquer árvore filogenética seja submetida vários testes estatísticos e seja aceita somente quando nenhum destes testes a rejeitar, considerando que tal árvore é uma hipótese científica sujeita à equívocos.

Por outro lado, isso não significa que uma árvore seja sem valor, a menos que todos os ramos internos tenham elevados índices de *bootstrap*. De fato, toda árvore filogenética construída é a melhor árvore obtida sob o princípio da reconstrução utilizada. Assim, mesmo que muitos dos ramos internos de uma árvore não sejam bem sustentados pelo *bootstrap*, ela não deveria ser descartada. Ela é uma árvore hipotética, mas pode ser a árvore correta. Simulações de computador têm mostrado que muitos padrões de ramos de uma árvore inferida estão corretos, mesmo não sendo sustentados por elevados valores de *bootstrap*.

7.7 Problemas Associados À Reconstrução Filogenética

Nenhum método de reconstrução filogenética pode ser qualificado como o melhor sob quaisquer condições. Cada um deles possui vantagens e desvantagens e pode ser bem sucedido ou falho, dependendo da natureza do processo evolutivo, que é em grande parte desconhecido. Por isso, em resumo, tem-se:

- a) O método UPGMA funciona bem somente sob taxas evolutivas constantes, mas possui a vantagem de ser computacionalmente rápido.
- b) Métodos aditivos, incluindo método das distâncias transformadas, *neighbor-relation* e *neighbor-joining*, são livres de erros sistemáticos se as distâncias satisfazem a “condição dos quatro pontos”. Os desempenhos destes métodos dependem também do método que transforma estados de caráter em medidas de distâncias. A principal vantagem destes métodos é que o

tempo computacional é muito curto, tornando-os apropriados quando se deseja analisar um grande número de OTUs.

- c) Métodos de máxima parcimônia não requerem nenhuma pressuposição explícita, exceto que a árvore que requer menor número de substituições é melhor que outra que exige maior número. Por outro lado, uma árvore que minimiza o número de substituições também minimiza o numero de homoplasias, isto é, paralelismos, convergências, substituições reversas, etc. Dessa forma, estes métodos possuem baixo desempenho quando alguns dos ramos da árvore são muito mais longos que os demais, porque eles tendem a posicionar junto esses ramos maiores (atração dos ramos longos). Além do mais, quando o número de OTUs é grande, é impraticável realizar uma busca exaustiva pelo método da parcimônia, sendo necessário o emprego de um algoritmo heurístico. Neste caso, não há garantia de se obter a árvore com máxima parcimônia.
- d) Método de máxima verossimilhança utiliza as informações de estados de caráter de todos os sítios; por outro lado, ele requer pressuposições explícitas quanto à taxa e ao padrão de substituição dos nucleotídeos. Sua principal desvantagem, entretanto, é o longo e tedioso tempo computacional consumido.

7.8 Estratégias para Minimizar Erros na Análise Filogenética

A melhor maneira de minimizar erros aleatórios é usar grande quantidade de dados e somente seqüências que evoluem em uma taxa apropriada para a questão filogenética sob investigação. Também se recomenda que a análise conjunta de dados de diferentes seqüências só deve ser realizada se elas apresentam comportamento evolutivo similar. Deve-se evitar, sempre que possível, a análise de OTUs com grandes distâncias filogenéticas.

Para que um caráter seja útil do ponto de vista filogenético, é preciso que ele seja informativo e confiável; sempre se deve estar ciente que de qualquer árvore inferida frequentemente contém erros, independentemente das precauções que são tomadas.

De maneira geral, a acurácia de uma árvore inferida depende de dois fatores no mínimo: da relação linear entre a medida de distância utilizada com o número de substituições ocorridas; e do erro-padrão ou do coeficiente de variação da medida de distância estimada.

Embora não haja, atualmente, nenhum método estatístico geral que possa ser aplicado na escolha de uma medida de distância (ou modelo matemático) apropriada para construção de topologias de árvores, algumas considerações gerais, obtidas a partir de dados empíricos e simulações de computador, podem servir de orientação.

- 1) Quando a estimativa do número de substituições por sítio é de, aproximadamente, 0,05 ou menor ($d \leq 0,05$), deve-se utilizar a distância Jukes-Cantor (p), tendo ou não viés na razão transição/transversão, assim como a variação nas taxas de substituições entre os sítios. Nessas circunstâncias, medidas mais complicadas fornecem os mesmos valores que a medida p, porém com maiores variâncias.
- 2) Quando o número de nucleotídeos examinados é grande e $0,05 < d < 1,0$, deve-se utilizar a distância p, a menos que a razão transição/transversão (R) seja alta, isto é, $R > 5$. Quando R é elevado e o número de nucleotídeos examinados (n) for grande, deve-se utilizar a distância de Kimura ou gama. Por outro lado, quando o número de seqüências examinadas for grande, mas n for pequeno, a distância p fornecerá melhores resultados, a menos que a taxa de evolução entre os sítios dentro das linhagens seja muito variável.
- 3) Quando $d > 1$ para muitos pares de seqüências, a árvore produzida não é confiável, por uma série de razões (p. ex. variâncias elevadas dos \hat{d} 's, erros nos alinhamentos). Por isso, sugere-se desconsiderar esses dados, ou evitar utilizá-los o máximo possível.

- 4) Como regra geral, se duas medidas de distância fornecem valores similares para um particular conjunto de dados, deve-se utilizar a mais simples, porque ela possui variância menor. Há tantos fatores desconhecidos na análise de dados reais, que qualquer árvore produzida dever ser interpretada com o devido cuidado.

As recomendações anteriores são indicadas para a construção de árvores. Em se tratando de estimativas dos comprimentos de ramos ou de extensão de tempos evolutivos, estimadores não-viesados produzem melhores resultados do que estimadores viesados.

ANEXOS

Tabela A1 - Dados experimentais de 15 cultivares de milho-pipoca, avaliados em blocos ao acaso, com três repetições, relativos a altura da planta (AP, em metros), da espiga (AE, em metros), número de espigas (NE), prolificidade (PROL), capacidade de expansão (CE), proporção de plantas quebradas (QUE), proporção de plantas doentes (DOEN), peso de cem grãos (PCG) e produção de grãos (PROD, em kg por parcela)

T	R	AP	AE	NE	PROL	CE	QUE	DOEN	PCG	PROD
1	1	1,58	1,29	35,82	1,14	15,74	0,28	0,19	20,94	3184
1	2	1,79	1,15	33,18	1,02	12,96	0,37	0,19	23,11	3724
1	3	2,08	1,17	32,53	1,01	13,81	0,31	0,19	21,97	3109
2	1	3,00	1,41	27,98	1,04	15,00	0,27	0,13	24,60	3441
2	2	2,69	1,47	21,85	0,98	13,44	0,26	0,18	18,45	4189
2	3	3,02	1,41	23,54	1,16	16,19	0,29	0,14	22,52	4221
3	1	2,58	0,91	28,09	1,65	12,70	0,23	0,14	14,84	3014
3	2	2,72	1,24	27,82	1,51	13,84	0,19	0,15	18,63	4142
3	3	2,27	1,33	33,51	1,64	14,36	0,27	0,16	20,13	3382
4	1	1,92	1,05	23,73	0,99	11,69	0,30	0,13	10,42	2469
4	2	1,98	0,97	22,55	1,02	14,08	0,34	0,17	18,51	2736
4	3	2,15	0,90	28,54	1,24	10,11	0,35	0,18	20,26	3331
5	1	3,42	1,49	38,16	1,15	13,94	0,42	0,08	19,96	2840
5	2	3,42	1,58	43,02	1,12	13,81	0,40	0,11	24,64	2758
5	3	3,59	1,52	36,94	1,27	13,38	0,32	0,11	21,19	3470
6	1	1,03	0,78	23,22	0,64	23,37	0,11	0,06	14,74	2024
6	2	1,76	0,82	26,08	0,92	24,35	0,12	0,10	15,57	2238
6	3	1,38	0,90	22,53	0,64	29,92	0,09	0,08	11,94	2164
7	1	1,37	0,78	19,94	0,93	35,61	0,11	0,09	14,13	1515
7	2	1,71	0,50	21,27	0,98	32,63	0,12	0,08	15,32	1743
7	3	1,77	0,72	20,7	0,92	38,75	0,12	0,09	14,45	2161
8	1	1,99	0,81	29,68	0,67	34,11	0,08	0,08	15,16	1762
8	2	1,48	0,95	26,49	0,61	25,98	0,09	0,09	16,67	1928
8	3	1,64	0,83	25,78	0,81	33,69	0,10	0,07	12,36	2254
9	1	1,31	0,81	15,83	0,70	34,74	0,08	0,10	14,52	1618
9	2	1,4	0,94	21,1	1,01	35,84	0,11	0,08	14,01	1570
9	3	1,35	0,87	24,59	0,99	33,27	0,09	0,09	13,55	2016
10	1	1,56	0,99	20,11	0,86	27,15	0,10	0,11	11,97	1763
11	2	1,68	1,15	26,91	0,86	26,72	0,09	0,11	13,25	1728

0										
1	3	1,55	1,18	27,19	0,81	32,16	0,10	0,09	15,57	1964
0										
1	1	1,58	1,07	19,5	0,55	32,29	0,24	0,12	33,77	4222
1										
1	2	1,63	0,79	22,05	0,67	31,09	0,20	0,13	34,78	3651
1										
1	3	1,92	1,30	18,06	0,65	23,14	0,17	0,13	39,96	4444
1										
1	1	2,34	1,26	16,12	0,55	31,25	0,23	0,09	34,91	3305
2										
1	2	1,7	1,35	18,23	0,67	27,04	0,27	0,08	32,37	2393
2										
1	3	2,09	1,13	22,16	0,68	28,47	0,28	0,08	33,61	3006
2										
1	1	2,41	1,11	16,72	0,64	35,73	0,17	0,09	31,13	3957
3										
1	2	2,08	1,09	13,63	0,57	35,53	0,18	0,08	26,41	4850
3										
1	3	2,27	1,03	16,94	0,58	41,99	0,16	0,09	37,78	4224
3										
1	1	2,47	1,11	15,55	0,67	31,18	0,21	0,10	25,52	3731
4										
1	2	2,17	0,59	15,15	0,81	31,15	0,20	0,09	36,46	2655
4										
1	3	2,52	0,74	19,71	0,67	29,61	0,22	0,10	32,15	4310
4										
1	1	2,15	1,01	18,47	0,60	28,25	0,15	0,10	31,49	3882
5										
1	2	2,14	1,21	22,69	0,46	30,30	0,19	0,08	30,24	3808
5										
1	3	1,96	1,04	22,64	0,53	35,32	0,19	0,11	31,28	3745
5										

Tabela A2 - Descrição genotípica de 50 indivíduos em relação a 10 locos - população 1

Ind.	Loco 1	Loco 2	Loco 3	Loco 4	Loco 5	Loco 6	Loco 7	Loco 8	Loco 9	Loco 10
1	0	0	1	1	0	0	0	1	0	0
2	0	0	0	1	0	1	0	1	0	0
3	0	1	1	1	0	0	1	1	0	0
4	0	0	1	1	0	1	0	1	0	0
5	1	0	1	0	1	1	0	1	1	1
6	0	0	1	1	1	0	0	1	0	1
7	0	0	1	1	0	1	0	1	0	1

8	0	1	1	1	1	1	1	1	1	0	0
9	0	0	1	1	0	1	1	0	1	0	0
10	0	0	0	1	1	1	1	0	1	1	1
11	0	0	1	1	1	1	0	0	1	0	1
12	0	0	0	0	1	1	1	0	1	1	1
13	1	1	1	1	1	1	1	0	0	0	0
14	0	1	1	1	1	1	0	0	1	0	1
15	0	1	1	0	0	1	0	0	1	0	1
16	0	1	0	1	1	0	0	0	1	0	0
17	0	0	1	1	0	0	0	0	1	1	1
18	0	0	1	1	0	1	0	0	1	0	0
19	0	0	1	1	0	1	0	0	1	1	0
20	0	0	0	0	0	0	0	0	1	0	1
21	1	1	1	1	1	0	0	0	1	1	0
22	0	0	1	1	0	0	1	0	1	0	0
23	1	0	1	1	0	0	0	1	1	0	1
24	0	1	1	1	0	0	1	0	1	0	0
25	0	0	1	1	0	0	1	0	1	1	1
26	0	0	1	1	0	0	0	0	1	0	1
27	0	0	1	1	0	0	0	0	1	0	1
28	0	0	1	1	0	0	0	0	1	1	1
29	0	0	1	1	0	0	0	0	1	1	1
30	0	1	0	1	0	0	0	1	1	1	0
31	0	0	1	1	1	1	0	0	1	0	0
32	1	0	1	1	0	0	1	0	1	1	0
33	0	0	0	1	0	0	1	0	1	1	1
34	0	0	1	1	0	0	1	0	1	0	0
35	0	0	0	1	0	0	0	0	1	1	0
36	0	0	0	1	1	1	1	0	1	0	1
37	1	0	0	1	0	0	1	0	1	0	0
38	1	1	1	1	1	1	1	1	1	1	0
39	0	1	0	1	0	0	0	0	1	0	0
40	0	0	0	1	0	0	1	0	1	1	0
41	0	0	0	1	0	0	1	0	1	0	0
42	1	0	0	1	0	0	0	0	1	0	1
43	0	0	0	1	1	1	0	0	1	1	1
44	0	0	1	1	0	0	0	0	1	1	0
45	0	1	0	1	1	0	0	0	1	0	1
46	0	0	0	1	0	0	1	1	1	1	0
47	0	0	0	1	0	0	0	0	1	0	0
48	0	0	1	1	0	0	1	1	1	1	0
49	0	0	1	1	0	0	1	0	1	1	0
50	0	0	0	1	0	0	0	0	1	1	1

Tabela A3 - Descrição genotípica de 50 indivíduos em relação a 20 locos - população 2

Ind.	Loco 1	Loco 2	Loco 3	Loco 4	Loco 5	Loco 6	Loco 7	Loco 8	Loco 9	Loco 10
1	11	11	24	13	35	11	22	23	11	22
2	11	11	34	11	23	12	22	33	11	22
3	12	11	13	11	23	11	22	23	12	22
4	12	12	13	13	13	11	22	33	22	22
5	12	11	34	33	23	11	22	23	22	11
6	23	11	14	13	23	11	22	33	12	22
7	11	11	14	11	23	11	22	23	11	12
8	33	12	44	13	13	24	22	33	11	22
9	13	11	44	33	22	11	22	13	11	22
10	11	11	33	13	11	11	22	33	12	22
11	23	11	11	11	13	11	22	22	11	22
12	11	11	13	13	22	12	22	23	12	12
13	11	11	33	13	33	12	22	33	11	22
14	14	12	44	12	33	11	22	23	11	22
15	13	11	34	13	23	24	22	23	11	12
16	11	11	34	12	33	11	22	23	12	22
17	11	13	13	11	12	12	22	33	12	22
18	13	22	14	12	13	12	22	12	22	22
19	13	12	14	13	13	22	22	23	11	22
20	12	11	34	13	25	11	22	13	22	12
21	34	12	14	33	23	12	22	23	12	22
22	45	22	13	13	13	11	12	33	11	22
23	11	12	11	11	13	12	22	33	22	22
24	11	11	14	13	13	11	22	23	11	22
25	11	11	12	13	23	14	22	13	11	22
26	11	12	44	13	12	22	12	23	12	22
27	34	11	33	13	12	13	22	33	12	12
28	12	12	33	33	23	11	22	13	11	22
29	13	11	44	13	14	12	22	23	11	22
30	12	11	11	33	33	14	22	23	12	22
31	12	11	14	13	33	11	22	33	12	22
32	13	11	13	33	13	14	22	23	12	22
33	13	12	24	11	23	12	22	13	22	22
34	13	22	14	13	12	12	22	22	22	22
35	11	11	13	13	12	24	22	23	11	22
36	11	11	13	13	23	12	22	22	11	12
37	13	11	14	11	13	11	22	23	11	22
38	11	12	24	13	24	14	22	22	11	22
39	33	12	13	33	11	11	22	23	22	22
40	11	11	12	13	13	11	22	22	12	22

41	11	22	33	11	13	11	22	23	12	22
42	23	11	14	13	23	12	22	33	12	12
43	11	22	44	13	33	11	22	33	12	22
44	12	12	44	11	34	11	22	33	12	22
45	12	22	14	13	12	11	12	22	12	22
46	14	11	33	13	44	11	22	23	12	22
47	13	12	13	13	13	11	22	33	11	22
48	11	11	11	33	11	11	22	33	11	22
49	12	12	13	13	34	11	22	12	12	22
50	13	11	13	13	22	11	22	33	11	12

Tabela A4 - Descrição genotípica de cinco populações, com 50 indivíduos, em relação a dois locos codominantes multialélicos

População 1			População 2		População 3		População 4		População 5	
Ind	Loco 1	Loco 2	Loco 1	Loco 2	Loco 1	Loco 2	Loco 1	Loco 2	Loco 1	Loco 2
1	22	11	11	22	12	13	11	12	11	45
2	22	11	11	12	12	23	11	22	14	34
3	12	13	13	22	23	23	11	22	14	35
4	12	12	13	12	13	13	11	12	11	23
5	12	23	13	12	13	13	11	12	11	44
6	12	22	12	22	13	23	12	22	11	34
7	12	22	12	22	13	23	11	12	12	34
8	12	23	12	12	23	33	12	22	11	44
9	12	11	12	22	33	12	11	22	11	33
10	12	12	12	12	33	13	11	12	11	25
11	22	12	11	22	12	11	12	22	11	55
12	12	23	12	12	33	12	11	12	11	33
13	12	11	13	12	13	33	12	11	12	34
14	22	13	11	22	33	22	11	22	11	24
15	11	12	11	22	13	22	12	12	11	24
16	22	11	13	22	33	23	11	22	11	34
17	22	22	12	22	12	11	11	22	11	35
18	22	12	33	12	23	33	11	12	11	24
19	12	11	13	12	23	13	12	12	11	33
20	22	22	13	11	13	23	11	22	12	34
21	11	13	13	12	11	23	11	22	11	24
22	22	23	12	12	13	33	12	22	11	44
23	22	12	12	22	33	23	22	12	11	44
24	22	12	23	12	13	13	12	12	11	25
25	11	12	14	12	11	23	11	11	11	44
26	22	23	11	12	13	22	12	12	11	23
27	22	23	11	22	23	12	11	22	14	33
28	12	12	11	22	11	23	11	22	11	34
29	12	12	23	22	11	33	11	22	14	24
30	12	13	11	22	13	33	12	22	11	35
31	22	11	11	11	13	23	11	11	11	34
32	12	12	11	12	33	13	11	22	11	34
33	12	13	11	22	13	12	12	12	11	14
34	12	22	11	12	13	12	11	22	11	35
35	22	22	13	22	13	12	11	22	11	35
36	22	13	11	12	13	12	12	22	14	45
37	22	12	11	12	13	33	11	22	24	23
38	22	12	11	22	33	23	12	12	12	25
39	12	12	22	12	13	13	11	22	11	22

40	12	12	11	12	13	12	11	12	11	33
41	12	12	11	12	23	33	12	12	11	44
42	22	12	13	22	12	23	11	22	11	33
43	12	13	11	12	33	23	11	22	11	33
44	22	11	12	22	11	33	12	22	11	34
45	12	23	11	22	12	22	11	22	11	23
46	11	12	12	12	11	13	22	12	11	45
47	22	22	12	22	13	33	12	12	11	24
48	11	22	11	11	11	12	12	12	11	35
49	12	11	13	11	33	23	12	12	11	35
50	11	23	11	22	11	12	11	12	12	45

LITERATURA CITADA

- AAKER, D. A; KUMAR, V; DAY, G. S. **Pesquisa de marketing**. São Paulo: Atlas, 2001. 745 p.
- AKKAYA, M. S.; BHAGWAT, A. A.; CREGAN., E. P. B. Length polymorphisms of simple sequence repeat DNA in soybean. **Genetics**. v.132, p. 1131-1139, 1992.
- AKKAYA, M. S.; et al. Integration of simple sequence repeat DNA markers into a soybean linkage map. **Crop Science**. v. 35, p. 1439-1445, 1995.
- AMALRAJ, S. F. A. Genetic divergence in *Gossypium hirsutum* L. **Genetic Agriculture**, Bangalore, v. 36, p. 23-30, 1982.
- ANDERSON, T. W. **An introduction to multivariate statistical analysis**. New York, John Wiley & Sons. 242p. 1958.
- ARIAS, D. M.; RIESEBERG, L. H. Gene flow between cultivated and wild sunflowers. **Theor Appl Genet**, Berlin, v. 89, p. 655-660, 1994.,
- ARRIEL, N. H. C.; et al. Comparison of similarity coefficients in sesame cultivars clustering using RAPD markers. **Crop Breeding and Applied Biotechnology**. v. 4, p. 192-199, 2004.
- ARROYO, M. T. K., et al. Convergence in the mediterranean floras in central Chile and California: insights from comparative biogeography. In: ARROYO, M. T. K.; ZEDLER, P. H.; FOX, M. D. (Eds.), *Ecology and biogeography of mediterranean ecosystems in Chile, California, and Australia*. New York: Springer-Verlag, 1994. p. 43-88.
- ARUNACHALAM, V. Genetic distance in plant breeding. **The Indian Journal of Genetics and Plant Breeding**, v.41, n.2, p.226-236. 1981.
- AVISE, J.C.; et al. Intraespecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. **Annual Review of Ecology and Systematics**, v. 18, p. 489-522, 1987.
- AVISE, J. C. **Molecular Markers, Natural History, and Evolution**. 2ed, Sinauer Associates, Sunderland, MA. USA, 2004
- BALLOUX, F.; LUGON-MOULIN, N. The estimation of population differentiation with microsatellite markers. **Molecular Ecology**. v. 11, p.155-165, 2002.
- BARCELOS, E.; et al. Genetic diversity and relationship in American and African oil palm as revealed by RFLP and AFLP molecular markers. **Pesquisa Agropecuária Brasileira**. v. 37, n. 8, p. 1105-1114, 2002.
- BARROSO, L. P.; ARTES, R. **Análise multivariada**. Lavras: UFLA, 2003. 151 p.
- BECKER J.; HEUN, M. Barley microsatellites, allele variation and mapping. **Plant Mol. Biol.** v. 27, p. 835-845, 1995a.
- BECKER J.; HEUN, M. Mapping of digested and undigested random amplified microsatellite polymorphisms in barley. **Genome**. v. 38, p. 991-998. 1995b.
- BELL, C. J.; ECKER, J. R. Assignment of 30 microsatellite loci to the linkage map of *Arabidopsis*. **Genomics**. v. 19, p. 137-144, 1994.

BOTSTEIN, D.; et al. Construction of a genetic linkage map in man using restriction fragment lenght polymorphisn. **American Journal of Human Genetics**. v.32, p. 314-331, 1980.

BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. F. **Introdução à análise de agrupamento**. São Paulo, IME, USP, 1990. 105 p.

CARDLE, L.; et al. Computational and experimental characterization of physically clustered simple sequence repeats in plants. **Genetics**. v. 156, p. 847–854, 2000.

CARLINI-GARCIA, L. A.; VENCOVSKY, R.; COELHO, A.S.G. Factorial analysis of bootstrap variances of population genetic parameter estimates. **Genetics and Molecular Biology**. v. 29, n. 2, p. 308-313, 2006.

CARLINI-GARCIA, L. A.; VENCOVSKY, R.; COELHO, A. S. G. Métodos bootstrap aplicados em níveis de reamostragem na estimação de parâmetros genéticos populacionais. **Scientia Agrícola**. v. 58, n. 4, p. 785-793, 2001.

CARLINI-GARCIA, L. A.; VENCOVSKY, R.; COELHO, A.S.G. Variance additivity of genetic populational parameter estimates obtained through bootstrapping. **Scientia Agricola**. v. 60, n. 1, p. 97-103, 2003.

CARPENTER, J. et al. **Comparative Environmental Impacts of Biotechnology-derived and traditional soybeans, corn, and cotton crops**. Council for Agricultural Science and Technology. Ames: Iowa, 2002. 190 p.

CAVALLI-SFORZA, L. L., EDWARDS, A. W. F. Analysis of human evolution. IN: **Proc. 11th Intl. Cong. Genet.**, New York, Pergamon, p. 923-933. 1964.

CAVALLI-SFORZA L.; EDWARDS A. W. F. Phylogenetic analysis models and estimation procedure. **Evolution**. v. 21, p. 550-570, 1967.

COCKERHAM, C.C. Analysis of gene frequencies. **Genetics**. n.74, p. 679-700, 1973.

COLE, C.T. Genetic variation in rare and common plants. **Annual Reviews Ecology Systems**. v. 34, p. 213-237, 2003.

COLE-RODGERS, P.; SMITH, D.W.; BOSLAND, P.W. A novel statistical approach to analyze genetic resource evaluations using capsicum as an example. **Crop Science**, v.37. p.1000-1002, 1997.

CROW, J. F.; AOKI, K. Group selection for polygenic behavioral trait: estimating the degree of population subdivision. **Proceedings of the National Academy of Sciences of the United States of America**. v. 81, p. 6073-6077, 1984

CRUZ, C. D. **Aplicação de algumas técnicas multivariadas no melhoramento de plantas**. Piracicaba: ESALQ/USP, 188 p. 1990. Tese Doutorado.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. vol.2. Editora UFV, Viçosa, 2003. 585 p.

CRUZ, C. D.; VIANA, J. M. S. A methodology of genetic divergence analysis based on sample uni projection on two-dimensional space. **Revista Brasileira de Genética**, Ribeirão Preto, v.17, n.1, p.69-73, 1994.

del RIO, A. H.; BAMBERG, J. B. Geographical parameters and proximity to related species predict genetic variation in the inbred potato species *Solanum verrucosum* Sclechtd. **Crop Science**. v. 44, p.1170-1177, 2004.

del RIO, A. H., BAMBERG, J. B.; HUAMAN, Z.; SALAS, A.; VEJA, S.E. Association of ecogeographical variables and RAPD marker variation in wild potato populations of the USA. **Crop Sci.** v. 41, p. 870–878, 2001.

del RIO, A. H.; BAMBERG, J. B. Lack of association between genetic and geographic origin characteristics for the wild potato *Solanum sucrense* Hawkes. **Am. J. Potato Res.** v. 79, p. 335-338, 2002.

DIAS, L. A. S. Análises multidimensionais. In: Alfenas, A.C. (ed) Eletroforese de isoenzimas e proteínas afins. Ed. UFV, Viçosa-MG,1998, p. 405-475.

DINIZ FILHO, J. A. **Métodos filogenéticos comparativos**. Ribeirão Preto: Holos., 2000. 120 p.

DUARTE, M. C.; SANTOS J. B.; MELO, L. C. Comparison of similarity coefficients based on RAPD markers in the common bean. **Genetics and Molecular Biology**. v. 22, p. 427-432, 1999.

DUDLEY, J. W. Comparison of genetic distance estimators using molecular data. In: SYMPOSIUM ANALYSIS OF MOLECULAR DATA, Corvallis, Oregon, 1994. **Proceedings** ... Cornavallis, Oregon, American Society for Horticultural Science, Crop Science Society of American, 1994. p 3-7.

EDWARDS, A. W. F. Distances between populations on the basis of gene frequencies. **Biometrics**. v. 27, p. 873–881, 1971.

EMYGDIO, B. M.; ANTUNES, I. F.; CHOEL, E.; NEDEL, J. L. Eficiência de coeficientes de similaridade em genótipos de feijão mediante marcadores RAPD. **Pesq. agropec. bras.**. v. 38, p. 243-250, 2003.

EVERITT, B. S. **Cluster analysis**. Cambridge, Edward Arnold, University Press, 1993, 170 p.

EXCOFFIER, L. Analysis of population subdivision, IN: BALDING, D.J., BISHOP, M., CANNINGS, C. (eds). **Handbook of statistical genetics**. Chichester (UK): John Wiley & Sons. 2001, p. 271-307.

EXCOFFIER, L.; SMOUSE, P. E.; QUATTRO, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. **Genetics**. v. 131, p. 479-491, 1992.

FAHIMA, T.; et al. RAPD polymorphism of wild emmer populations, *Triticum dicoccoides*, in Israel. **Theor. Appl. Genet.** v. 98, p. 434-447, 1999.

FALCONER, D. S. **Introdução à genética quantitativa**. Viçosa: UFV, 1987. 279 p.

FALCONER, D. S.; MACKAY T. F. C. **Introduction to quantitative genetics**. 4th ed. New York: Longman, 1996. 464 p.

FAO. **First state of the world's plant genetic resources**: erosion of biodiversity and loss of genes continues; many gene banks threatened. FAO, Rome, Italy. 1996.

FELSENSTEIN, J. **Inferring Phylogenies**. Sinauer, Sunderland. 2004, 664p.

FERREIRA, R. de P.; CRUZ, C. D.; SEDIYAMA, C. S.; FAGERIA, N. K. Identificação de cultivares de arroz tolerantes à toxidez de alumínio por técnica multivariada. **Pesquisa Agropecuária Brasileira**. Brasília: v. 30, n. 6, p. 789-795, 1995.

FISHER, R. A. The logic of inductive inference. **J. Roy. Stat. Soc.** v.98, p. 39-54, 1935.

FITCH, W. M.; MARGOLIASH, E. Construction of phylogenetic trees. **Science**. v. 155, p. 279 284, 1967.

FLINT-GARCIA, S. A.; THORNSBERRY, J . M.; BUCKLER IV, E. S. Structure of linkage disequilibrium in plants. **Annual Reviews of Plant Biology**. v. 54, p. 357-74, 2003.

FRANKHAM, R.; BALLOU, J. D.; BRISCOE, D. A. **Introduction to conservation genetics**. Cambridge University Press, Cambridge, UK: 2003.

GAIA, J. M. D.; MOTA, M. G. C.; CONCEIÇÃO, C. C. C. Similaridade genética de populações naturais de pimenta-de-macaco por análise de RAPD. **Horticultura Brasileira**. v. 22, p. 686-689, 2004.

GAUT, B. S.; LONG, A. D. The Lowdown on Linkage Disequilibrium. **The Plant Cell**, v. 15, p. 1502-1506, July 2003.

GHADERI, A.; ADAMS, M. W.; NASSIB, A. M. Relationship between genetic distance and heterosis for yield and morphological traits in dry edible bean and fava bean. **Crop Science**. Madison, v.14, n.1, p.24-27, 1984.

GLOWATZKI-MULLIS, M-L.; et al. Microsatellite-based parentage control in cattle. **Anim. Genet.** v. 26, p. 7-12, 1995.

GOLDSTEIN, D. B.; LINARES, A. R.; CAVALLI-SFORZA, L. L.; FELDMAN, M. W. An evaluation of genetic distances for use with microsatellite loci. **Genetics**. v. 139, p. 463–471, 1995.

GOODMAN, M. M. Distance analysis in biology. **Systematic Zoology**. v. 21, p. 174-286, 1972.

GOODMAN, M. M; STUBER, C. W. Races of maize: VI. Isozyme variation among races of maize in Bolivia. **Maydica**. v.28, p.169-187, 1983.

GORELICK, R.; LAUBICHLER, M. D. Decomposing Multi-locus Linkage Disequilibrium. **Genetics**, v. 166, p. 1581-1583, 2004.

GOWER, J. C. A comparison of some methods of cluster analysis. **Biometrics**. v. 23, p. 623-637, 1967.

GOWER, J. C. A general coefficient of similarity and some of its properties. **Biometrics**. v.27, p. 857-871, 1971.

GOWER, J. C.; LEGENDRE, P. Metric and euclidian properties of dissimilarity coefficients. **J. Classif.** v.3, p.5-48, 1986.

GRAFIUS, J. E. The complex trait as a geometric construct. **Heredity**, v.16. n.1, p.225-228, 1961.

GRAUR, D.; LI, W. H. **Fundamentals of molecular evolution**. 2a ed. Massachusetts: Sinauer Associates, 2000. 481p.

- GUO, S. W, THOMPSON, E. A. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. **Biometrics**. v. 48, p. 361-372, 1992.
- GUPTA, P. K.; RUSTGI, S.; KULWAL, P. L. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. **Plant Mol Biol**, v. 57, p. 461-485, 2005.
- HALDANE, J. B. S. An exact test for randomness of mating. **Journal of Genetics**. v. 52, p. 631-635, 1954.
- HANDS, S.; EVERITT, B. S. A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. **Multivariate Behavior Research**. v. 22, p. 235-243, 1987.
- HARTL, D. L.; CLARK, A. G. **Principles of Population Genetics**. 3rd edition. Sunderland (MA): Sinauer Associates, 1997. 542 p.
- HEDRICK, P. W. A new approach to measuring genetic similarity. **Evolution**. v. 25, p. 276-280, 1971.
- HENNIG, W. **Phylogenetic systematics**. Urbana: University of Illinois Press. 1996.
- HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**. v. 24, p. 417-441, 1933.
- HOTELLING, H. Simplified calculation of principal components. **Psychometrika**. v.1, p. 27-35, 1936.
- HOYT, E. **Conservação dos parentes silvestres das plantas cultivadas**. Tradução por Coradin, L. Delaware: Addison-Wesley Iberoamericana, 1992. 52 p. (Apoio IBPGR, IUCN, WWF e Embrapa/Cenargen).
- HUGHES, M. M. Exploration and play re-visited: a hierarchical analysis. **International Journal of Behavioral development**. v. 2, p. 225-233, 1979.
- JACCARD, P. Nouvelles recherches sur la distribution florale. **Bull. Soc. Vaud. Sci. Nat.**, v. 44, p. 223-270, 1908.
- JACKSON, A. A.; SOMERS, K. M.; HARVEY, H. H. Similarity coefficients: measures for cooccurrence and association or simply measures of occurrence? **American Naturalist**. v. 133, 436-453, 1989.
- JAMES, F. C.; McCULLOCH, C. E. Multivariate analysis in ecology and systematics: Panacea of pandora's box? **Annual Review Ecology Systematic**, v.21, p.129-166, 1990.
- JARDINE, N. E SIBSON, R. The construction of hierachic and non-hierachic classification. **Computer Journal**, v. 11, p. 117-184, 1968.
- JARDINE, N. E SIBSON, R. **Mathematical Taxonomy**. New York: John Wiley and Sons, Chichester, 1971.
- JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. New Jersey: Prentice-Hall, 1988. 607p.
- JOLLIFFE, I. T. Discarding variables in a principal component analysis. I. Artificial data. **Appl. Stat.** v. 21, n. 2, p. 160-173, 1972.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. II. Real data. **Appl. Stat.**, v. 21, n. 1, p. 21-31, 1973.

JORDE, L.B. Linkage disequilibrium and the search for complex disease genes. **Genome Res.** v. 10, p. 1435–1444, 2000.

JOSHI, A. B.; DHAWAN, N.L. Genetic improvement of yield with special reference to self fertilizing crops. **Indian Journal Genetics.** v. 26, n.1, p.101-113, 1966.

KAUFMAN, L. E.; ROUSSEEUW, P. J. **Fiding Groups in Data**, Wiley & Sons, New York, 1990.

KHATTREE, R.; NAIK, D. N. **Multivariate data reduction and discrimination with SAS Software**. Cary, NC: SAS Institute Inc, 2000, 338p.

KIMURA, M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. **Journal of Molecular Evolution.** v. 16, p. 111-120, 1980.

KIMURA, M; CROW, J. F. The number of alleles that can be maintained in a finite population. **Genetics.** v. 49, p.725-738,1964.

KIMURA, M. Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. **Genetics.** v. 39, p. 280-295, 1954.

KIMURA, M. “Stepping stone” model of population. **Ann. Rept. Nat. Inst. Genetics.** v. 3, p. 62-63, 1953.

KLOPPENBURG J. R.; KLEINMAN, D. L. Plant germplasm controversy – analyzing empirically the distribution of the world's plant genetic resources. **Bioscience.** v. 37, p. 190-198, 1987.

KUIPER, F. K.; FISHER, L. A Monte Carlo comparison of six clustering procedures. **Biometrics**, v. 31, p. 777-783, 1975.

LATTER, B. D. H. Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. **Genetics.** v. 70, p. 475-490, 1972.

LATTER, B. D. H. The island model of differentiation: a general solution. **Genetics.** v. 73, p. 147-157, 1973.

LEVENE, H. On a matching problem arising in genetics. **Annals of mathematical statistics.** v. 21, p. 91–94, 1949.

LIANG, K. Y.; CHIU, Y.F.; BEATY, T. H. A robust identity-by-descent procedure using affected sib pairs: multipoint mapping for complex diseases. **Hum Hered.** v.51, p. 64-78, 2001.

LIU, B. H. **Statistical genomics: linkage, mapping, and QTL analysis**. Boca Raton: CRC Press, 1997. 611 p.

LIU, Z-W; et al. Development of simple sequence repeat DNA markers and their integration into a barley linkage map. **Theoretical and Applied Genetics.** v. 93: p. 869-876, 1996.

LYNCH, M.; MILLIGAN, B. G. Analysis of population genetic structure with RAPD marks. **Molecular Ecology.** v. 3, p. 91-99, 1994.

- MADDISON, W.P.; DONOGHUE M. J.; MADDISON, R. Outgroup analysis and parsimony. **Systematic Zoology**. v. 33, p. 83-183, 1984.
- MAHALANOBIS, P. C. On the generalized distance in statistics. **Proceedings of the National Institute of Sciences of India**, New Delhi, v. 2, p. 49-55, 1936.
- MALÉCOT, G. Quelques schemas probabilistes sur la variabilité des populations. **Ann. Univ. Lyon Sci.** v. 13, p. 37-60, 1950.
- MALUF, W.R.; FERREIRA, P.E.; MIRANDA, J.E.C. Genetic divergence in tomatoes and its relationship with heterosis for yield in F₁ hybrids. **Revista Brasileira de Genética**. Ribeirão Preto: v.3, p.453-460, 1983.
- MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. London: Academic Press, 521p. 1979.
- MELO JÚNIOR, A. F.; CARVALHO, D. de; PÓVOA, J. S. R. Estrutura genética de populações naturais de pequi (Caryocar brasiliense Camb.). **Scientia Florestalis**. n. 66, p. 56-65, 2004.
- MENGISTU, L. W.; MUELLER-WARRANT, G. W.; BARKER, R. E. Genetic diversity of *Poa annua* in western Oregon grass seed crops. **Theor. Appl Genet**. v. 101, p. 70-79. 2000.
- MEYER, A. S.; GARCIA, A. A. F.; SOUZA, A P.; SOUZA JÚNIOR, C. L. Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (*Zea mays* L.). **Genetics and Molecular Biology**. v. 27, p. 83-91, 2004.
- MILLIGAN, G. W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. **Psychometrika**. v. 45, p. 325-342, 1980.
- MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**. v. 50, p. 159-179, 1985.
- MOJENA, R. Hierarchical grouping method and stopping rules: an evaluation. **Computer Journal**. v. 20, p. 359-363, 1977.
- MOLL, R. H.; et al. The relationship of heterosis and genetic divergence in maize. **Genetics**. v.52, n.1, p.139-144, 1965.
- MORAES, P. L. R.; DERBYSHIRE, M. T. V. C. Estrutura genética de populações naturais de *Cryptocarya Aschersoniana* Mez (lauraceae) através de marcadores isoenzimáticos. **Biota Neotropica**. v. 2, n. 2, p. 1-10, 2002. Disponível em: <<http://www.biota-neotropica.org.br/v2n2/pt/abstract?article+BN02402022002>> Acesso em: 22 set. 2006.
- MORAES, R. M. A.; et al. Genetic divergence in soybean parents for backcrossing programs. **Crop Breeding and Applied Biotechnology**. v. 5, n. 3, p. 339-346, 2005.
- MORGANTE, M.; et al. Genetic mapping and variability of seven soybean simple sequence repeat loci. **Genome**. v. 37, p. 763-769, 1994.
- MORGANTE, M.; HANAFAY, M.; POWELL, W. Microsatellites are preferentially associated with non-repetitive DNA in plant genomes. **Nature Genetics**. v. 30, p. 194-200, 2002.

MORRISON, D. A.; ELLIS, J. T. Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa. **Molecular Biology and Evolution**, v.14, p. 428-441, 1997.

MOTA, J. W. da S. e. **Análise da diversidade genética de germoplasma de Theobroma cacao L. da Amazônia brasileira por microssatélites**. 2003. 97p. Tese (Doutorado) – Universidade Federal de Viçosa, Viçosa-MG.

MURTY, B.R.; ARUNACHALAM, V. The nature of divergence in relation to breeding system in some crop plants. **The Indian Journal of Genetics and Plant Breeding**. v.26, n.2, p.188-198, 1966.

NASS, L. L.; PATERNIANI, E. Pre-breeding: a link between genetic resources and maize breeding. **Scientia Agricola**. v.57, p.581-587, 2000.

NASS, L. L. Utilização de recursos genéticos vegetais no melhoramento. In: NASS, L. L.; et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 2, p. 29-55.

NATIONAL RESEARCH COUNCIL - NRC. Atmospheric Contaminants in Manned Spacecraft. Washington, D.C.: **National Academy of Sciences**. 1972.

NEI, M. Analysis of gene diversity in subdivided populations. **Proceedings of the National Academy of Sciences of the United States of America**. Washington, v. 70, p. 3321-3323, 1973.

NEI, M.; CHAKRAVARTI, A. Drift variances of FST and GST statistics obtained from a finite number of isolated populations. **Theoretical Population Biology**, v. 11, n. 3, 1977.

NEI, M. Estimation of average heterozygosity and genetic distance from small number of individuals. **Genetics**. v. 89, p. 583-590, 1978.

NEI, M. F-statistics and analysis of gene diversity in subdivided populations. **Annual Human Genetics**. v. 41, p.225-233, 1977.

NEI, M. Genetic distance between populations. **American Naturalist**. Chicago, v. 106, p. 238-292, 1972.

NEI, M. KUMAR, S. **Molecular evolution and phylogenetics**. New York: Oxford University Press, 2000. 333p.

NEI, M.; LI, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. **Proc. Natl. Acad. Sci. USA**. v. 76, n. 10, p. 5269-5273. 1979.

NEI, M. **Molecular Evolutionary Genetics**. Columbia University Press, New York, 1987.

NEI, M. **Molecular Population Genetics and Evolution**. North-Holland, Amsterdam and New York, 1975, 290 p.

NEI, M.; TAJIMA, F.; TATENO, Y. Accuracy of estimated phylogenetic trees from molecular data. **Journal of Molecular Evolution**. v.19, p. 153-170, 1983.

NEIGEL, J.E. Is F_{ST} obsolete? **Conservation Genetics**, n. 3, p. 167-173, 2002.

- NORDBORG M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial selffertilization. **Genetics**. v. 154, p. 923–929, 2000.
- OTT, J. Strategies for characterizing highly polymorphic markers in human gene mapping. **American Journal of Human Genetics**. v.51, p. 283-290, 1992.
- PAIVA, J. R.; KAGEYAMA, P. Y.; VENCOVSKY, R. Genetics of rubber tree (*Hevea brasiliensis* (Willd. ex A. Juss.) Muell. Arg. 2. Mating system. **Silvae Genetica**. v. 34, p. 373-376, 1994.
- PATERNANI, E.; LONNQUIST, J. H. Heterosis in interracial crosses of corn (*Zea mays* L.). **Crop Science**. v.3, n.1, p. 504-507, 1963.
- PEARSON, K. On lines and planes of closet fit to systems of points in space. **Philosophical Magazine**. v. 2, p. 559- 572, 1901.
- PEREIRA, J. J. **Análise de agrupamento e discriminante no melhoramento genético – aplicação na cultura do arroz (*Oryza sativa* L.)**. 1999. 191 p. Tese (Doutorado).
- PETER, K. V.; RAI, B. Genetic divergence in tomato. **The Indian Journal of Genetics and Plant Breeding**. v.36, n.3, p.379-383, 1976.
- PETTIT, R.J.; EL MOUSADIK, A.; PONS, O. Identifying Populations for conservation on the basis of genetics markers. **Conservation Biology**. v. 12, n.4, p. 844-855, 1998.
- POWELL, J R.; LEVENE, H; DOBZHANSKY, T. Chromosomal Polymorphism in *Drosophila pseudoobscura* Used for Diagnosis of Geographic Origin. **Evolution**, v. 26, n. 4 , p. 553-559, 1972.
- PREVOSTI, A.; OCANA, J.; ALONZO, G. Distances between populations for *Drosophila Subobscura* based on chromosome arrangement frequencies. **Theor. Appl. Genet..** v. 45, p. 231-241, 1975.
- PRITCHARD J. K.; ROSENBERG N. A. Use of unlinked genetic markers to detect population stratification in association studies. **American Journal of Human Genetic**. v. 65, p. 220–228, 1999.
- QUEROL, D. **Recursos genéticos, nosso tesouro esquecido:** abordagem técnica e sócio-econômica. Rio de Janeiro: AS-PTA, 1993. 206 p.
- RAM, J., PANWAR, D.V.S. Interspecific divergence in Rice. **Indian J. Genet.** V.30, p. 1-10, 1970.
- RAO, N. K. S.; SWAMY, R. D.; CHACO, E. K. Differentiation of plantlets in hybrid embryo callus of pineapple. **Scientia Horticulturae**. Amsterdam, v. 15, p. 235-238, 1981.
- RAO, R. C. **Advanced statistical methods in biometric research**. New York: John Wiley and Sons, 390p. 1952.
- REIF, J. C.; MELCHINGER, A. M.; FRISCH, M. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. **Crop Science**. v.45, p.1-7, 2005.
- REYNOLDS, J.; WEIR, B. S.; COCKERHAM, C. C. Estimation of the coancestry coefficient: basis for a short-term genetic distance. **Genetics**. v.105, p. 767-779, 1983.

RIDLEY, M. **Evolução**. 3. ed. Tradução por Ferreira, H. B. Porto Alegre: Artmed, 2006. 752 p.

ROBINSON, I. P. Aloenzimas na Genética de Populações de Plantas. In: Alfenas, A. C. (Ed.) **Eletroforese de isoenzimas e proteínas afins:** fundamentos e aplicações em plantas e microrganismos. Viçosa: UFV, 1998, cap. 7, p. 329-380.

ROGERS, J. S. Measures of genetic similarity and genetic distance. In: **Studies in genetics**. VII. Austin, University of Texas, 1972. p. 145-153.

ROHLF, F. J. Consensus indices for comparing classifications. **Mathematical Bioscience**. v. 59, p. 131-144, 1982.

ROHLF, F. J., FISHER, D. R. Test for hierarchical structure in random data sets. **Systematic Zoology**. v.17, p. 407-412, 1968.

ROMESBURG, H. C. **Cluster analysis for researchers**. California, Lifetime Learning, 1984, 334 p.

ROSSETTO, M.; SLADE,R.W.; BAVERSTOCK, P.R.; HENRY,R.J.; LEE, L.S. Microsatellite variation and assessment of genetic struture in tea tree (*Melaleuca alternifolia* – Myrtaceae). **Molecular Ecology**. n. 8, p. 633-643, 1999.

ROUSSET, F.; RAYMOND, M. Testing heterozygote excess and deficiency. **Genetics**. v. 140, p. 1413-1419, 1995.

RZHETSKY, A.; NEI, M. Theoretical foundation of the minimum-evolution method of phylogenetic inference. **Mol. Biol. Evol.** v. 10, p. 1073-1095, 1993.

SAGHAI-MAROOF, M. A.; et al. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. **Proceedings of the National Academy of Sciences of the United States of America**, v.81, p.8014-8018, 1984.

SANTACRUZ-VARELA, A.; et al. Phylogenetic relationships among North American popcorns and their evolutionary links to Mexican and South American popcorns. **Crop Science**. v.4, p.1456-1467, 2004.

SAS. **SAS/STAT user´s guide. Version 6.4.** , v.2, NC, Cary, SAS Institute, 1989.

SEBBENN, A. M.; SEOANE, C. E. S.; KAGEYAMA, P. Y.; VENCOVSKY, R. Efeitos do Manejo na estrutura genética de populações de caixeta (*Tabebuia cassinoides*). **Scientia Forestalis**, n. 58, p. 127-143, 2000.

SHRIVER, M., et al. A novel measure of genetic distance for highly polymorphic tandem repeat loci. **Mol. Biol. Evol..** v. 12, p. 914-920, 1995.

SILVA, E. P.; RUSSO, .C. A. M. Techniques and statistical data analysis in molecular population genetics. **Hydrobiologia**, n. 420, p. 119-135, 2000.

SINGH, S. P.; SINGH, R. B. Triple test cross analysis in two wheat crosses. **Heredity**. v. 37, p. 173-177, 1976.

SINGH, T. H.; GILL, S. S. Genetic diversity in upland cotton under different environments. **Indian Journal and Plant Breeding**. New Delhi, v. 44, p. 506-513, 1984.

SINGH, Y. P.; KUMAR, A.; CHAUHAN, B. P. S. Genetic divergence in pearl millet. **The Indian J. of Genet. and Plant Breed.** v.41, n.1, p.186-190, 1981.

SLATKIN, M. A measure of population subdivision based on microsatellite allele frequencies. **Genetics**, 139: 457-462, 1995.

SLATKIN, M.; BARTON, N. H. A comparison of three indirect methods for estimating average levels of gene flow. **Evolution**. v. 43, n. 7, p. 1349-1368, 1989.

SMITH, J. S. C.; DUVICK, D. N. Germplasm collections and the private plant breeder. In: BROWN, A. D. H.; et al. (Eds.) **The Use of Plant Genetic Resources**. Cambridge: Cambridge University Press. 1989. p. 17-31.

SNEATH, P. H.; SOKAL, R. R. **Numerical taxonomy**: the principles and practice of numerical classification. San Francisco: W.H. Freeman, 573 p. 1973.

SOKAL, R. R.; ROHLF, F. J. The comparison of dendograms by objective methods. **TAXON**. v.11, p. 30-40, 1962.

SOKAL, R. R.; SNEATH, P. H. A. **Principles of Numerical Taxonomy**. San Francisco: W. H. Freeman and Company, 1963. 359p.

SPIESS, E. L. Genes in Population. 2nd ed. John Wiley & Sons, Inc., *Bulletin of Science Technology Society*. v. 12, p. 53-54, 1989.

TAKEZAKI, N.; NEI, M. Genetic Distances and reconstruction of trees from microsatellite DNA. **Genetics**. v. 144, p. 389-399, 1996.

TEMNYKH, S.; et al. Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). **Theoretical Applied Genetics**. v. 100, p. 697-712, 2000.

TORGGLER, M. G. F.; CONTEL, E. P. B.; TORGGLER, S. P. **Isoenzimas: variabilidade genética em plantas**. Sociedade Brasileira de Genética, Ribeirão Preto, 1995, 186p.

UPADHYUAY, M. K.; MURTY, B. R. Genetic divergence in relation to geographical distribution in pearl millet. **The Indian Journal of Genetics and Plant Breeding**, v.30, n.3, p.704-715, 1970.

VASCONCELOS, E. D.; CRUZ, C. D.; BHERING, L. L.; RESENDE JUNIOR, M. F. R. Método alternativo para análise de agrupamento. **Pesq. agropec. bras.** v. 42, p. 10, 2007.

WADT, L. H.; KAGEYAMA, P. Y. Estrutura genética e sistema de acasalamento de *Piper hispidinervum*. **Pesquisa Agropecuária Brasileira**. v. 39, n.2, p. 151-157, 2004.

WEIR, B. S.; COCKERHAM, C. C; Estimating F-statistics for the analysis of population structure. **Evolution**. v. 38, p. 1358-1370, 1984.

WEIR, B. S. **Genetic Data Analysis II**. Sinauer Associates, Inc., MA, USA, 1996. 445p.

WILLIAMS, W. Heterosis and the genetics of complex characters. **Nature**. v.184, p. 527–530, 1959.

WRIGHT, S. Evolution in Mendelian populations. **Genetics**. v. 16, p. 97–159, 1931.

WRIGHT, S. Isolation by distance. **Genetics**. v. 28, p.114-138, 1943.

WRIGHT, S. The interpretation of population structure by F-statistics with special regards to systems of mating. **Evolution**. v. 19, p. 395-420, 1965.

WRIGHT, S. **Variability within and among natural populations**. Chicago: The University of Chicago Press, 1951. 580 p.

WRIGHT, S. **Variability within and among natural populations**. Vol. 4, The University of Chicago Press, Chicago, 1978, 580p.

WU K-S.; TANKSLEY S. D. Abundance, polymorphism and genetic mapping of microsatellites in rice. **Mol. Gen. Genet.** v. 241, p. 225-235., 1993

YATES, F. Contingency tables involving small numbers and the χ^2 test. **Journal of the Royal Statistical Society Supplement**. v. 1, p. 217-235, 1934.

ZHENG, Y. Q.; ENNOS, R. A. Genetic variability and structure of natural and domesticated populations of Caribbean pine (*Pinus caribaea* Morelet). **Theoretical and Applied Genetics**. v. 98, p. 765-771, 1999.

ZHIVOTOVSKY, L. A. Estimating population structure in diploids with multilocus dominant DNA markers. **Molecular Ecology**. v. 8, p. 907-913, 1999.

ZUCCHI, M.I.; et al. Genetic struture and gene flow in *Eugenia dysenterica* DC in Brazilian Cerrado utilizing SSR markers. **Genetics and Molecular Biology**. v. 4, n. 26, p. 449-457, 2003.