



Tópicos Especiais: Filogenômica e Evolução Molecular

Módulo 1 – Fundamentos de Filogenética e Evolução Molecular

Dr. Tiago Belintani
<https://tiagobelintani.github.io/>

Laboratório de Aracnologia de Rio Claro, Departamento de Biodiversidade,
Instituto de Biociências,
Universidade Estadual Paulista, Rio Claro, Brazil

Rio Claro
2026

Módulo 1 - Fundamentos de Filogenética e Evolução Molecular

Objetivo do módulo

Introduzir os fundamentos conceituais da filogenética e da evolução molecular, estabelecendo a base teórica necessária para compreender métodos modernos de inferência filogenética.

Foco deste módulo

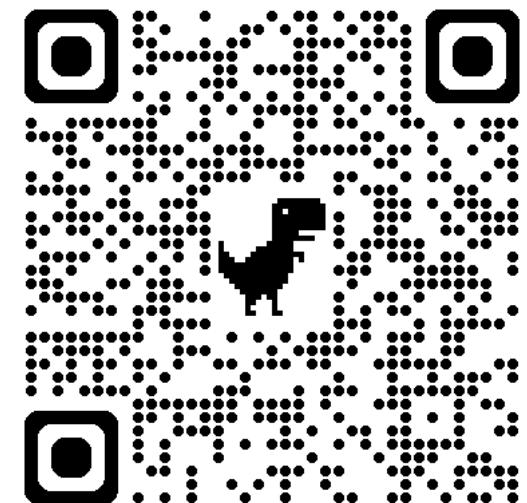
O que é filogenia

Como são gerados os dados genômicos

Origem do sinal filogenético

Modelos de substituição molecular

Homologia, ortologia e tipos de caracteres moleculares



Disciplina-ECO001011-UNESP-
Campus-de-Rio-Claro

O que é Filogenética?

é o campo da biologia que estuda as relações evolutivas entre organismos, ou seja, como as espécies (ou outros grupos taxonômicos) estão relacionadas umas com as outras através da ancestralidade comum.

Sistemática Clássica

- ✓ Baseada em caracteres morfológicos
- ✓ Classificação hierárquica
- ✓ Limitada por convergência

Era Molecular

- ✓ Análise de sequências de DNA
- ✓ Primeiros árvores filogenéticas
- ✓ Descoberta da variabilidade molecular

Era Genômica

- ✓ Análise de genomas completos
- ✓ Múltiplas sequências de referência
- ✓ Comparações em larga escala

Filogenômica Moderna

- ✓ Análises abrangentes
- ✓ Dados filogenômicos massivos
- ✓ Integração de múltiplas fontes



Filogenia como Hipótese Histórica

Conceito Central

- ✓ Filogenética hoje é uma ciência baseada em dados
- ✓ Hipóteses filogenéticas são **testáveis e revisáveis**
- ✓ Mudança paradigmática em relação à sistemática clássica

Propriedades das Hipóteses Filogenéticas:

 **Testabilidade**

 **Revisabilidade**

Ciclo de Hipóteses Filogenéticas



"A filogenia é uma hipótese sobre a história evolutiva que pode ser testada e refinada com dados"

Filogenética (ômica) na Biologia Evolutiva

- Base para sistemática
- Integração com biogeografia
- Relação com genética e ecologia



Filogenética vs Filogenômica

Filogenética Tradicional

- 📊 Análise de dados em escala reduzida (genes ou marcadores específicos)
- 📌 Abordagem baseada em caracteres morfológicos ou genéticos limitados
- ⌚ Modelos evolutivos simplificados com menos parâmetros
- 🕒 Árvores filogenéticas baseadas em menor número de sequências

VS

Filogenômica

- 🧩 Análise de dados em escala genômica (e.g. transcritomas ou genomas completos)
- 🧩 Uso de milhares de loci ou todos os genes disponíveis
- ➡️ Modelos evolutivos mais complexos que lidam com heterogeneidade
- 🧩 Novos desafios analíticos: conflito entre genes, saturação

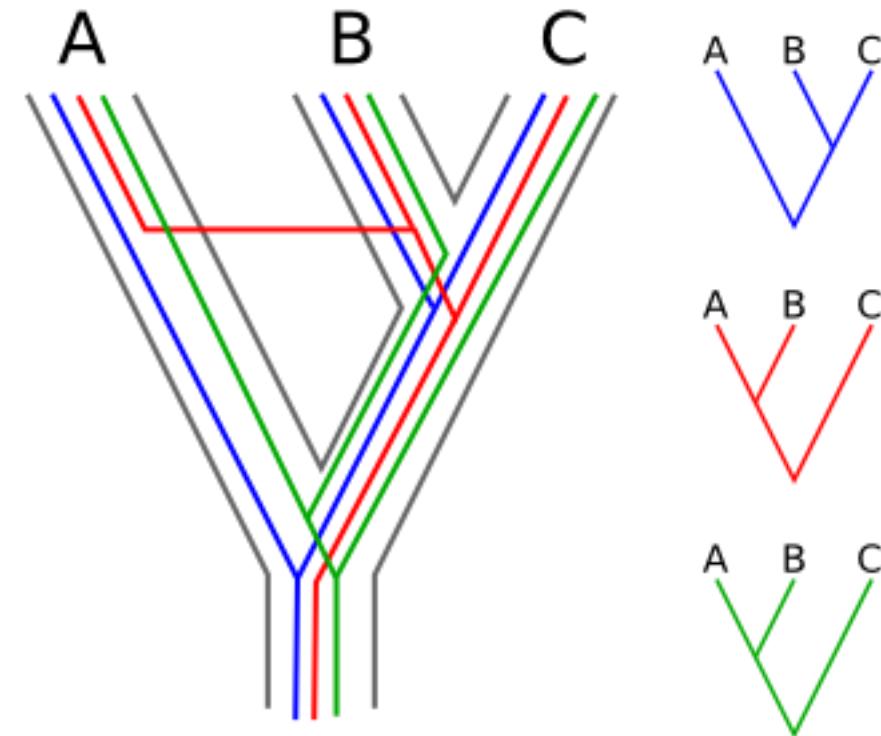
Marcadores genéticos



Genomas ou transcriptomas

O que muda com dados genômicos

- Mais sinal, mas também mais ruído
- Conflito entre genes
- Novos desafios conceituais
- Expansão do conhecimento





Como são gerados os dados genômicos?

O que é Sequenciamento



Técnica laboratorial para determinar a ordem das bases do DNA



Identificação sequencial de nucleotídeos ao longo da molécula



Base da genômica e da filogenômica

Genoma e Número de Genes

-  Relação complexa entre tamanho genômico e número de genes
-  Correlação forte em procariotos
-  Variação extrema em organismos eucariotos

| Organism | # of protein-coding genes | # of genes naïve estimate: (genome size /1000) | BNID |
|--------------------------------|---------------------------|--|-------------------|
| HIV 1 | 9 | 10 | 105769 |
| <i>Influenza A virus</i> | 10-11 | 14 | 105767 |
| Bacteriophage λ | 66 | 49 | 105770 |
| Epstein Barr virus | 80 | 170 | 103246 |
| <i>Buchnera sp.</i> | 610 | 640 | 105757 |
| <i>T. maritima</i> | 1,900 | 1,900 | 105766 |
| <i>S. aureus</i> | 2,700 | 2,900 | 105500 |
| <i>V. cholerae</i> | 3,900 | 4,000 | 105760 |
| <i>B. subtilis</i> | 4,400 | 4,200 | 111448 |
| <i>E. coli</i> | 4,300 | 4,600 | 105443 |
| <i>S. cerevisiae</i> | 6,600 | 12,000 | 105444 |
| <i>C. elegans</i> | 20,000 | 100,000 | 101364 |
| <i>A. thaliana</i> | 27,000 | 140,000 | 111380 |
| <i>D. melanogaster</i> | 14,000 | 140,000 | 111379 |
| <i>F. rubripes</i> | 19,000 | 400,000 | 111375 |
| <i>Z. mays</i> | 33,000 | 2,300,000 | 110565 |
| <i>M. musculus</i> | 20,000 | 2,800,000 | 100308 |
| <i>H. sapiens</i> | 21,000 | 3,200,000 | 100399, 111378 |
| <i>T. aestivum</i> (hexaploid) | 95,000 | 16,800,000 | 105448, 102713 |

Gerações de Tecnologias de Sequenciamento



Primeira Geração

- ✓ Método de Sanger
- ✓ Terminação de cadeia
- ✓ Alto impacto científico



Segunda Geração

- ✓ Leituras curtas
- ✓ Alto rendimento
- ✓ Democratização do sequenciamento



Terceira Geração

- ✓ Leituras longas
- ✓ Sequenciamento de molécula única
- ✓ Redução de vieses

Método de Sanger



Desenvolvido por Frederick Sanger no final da década de 1970



Baseado na terminação de cadeia de DNA



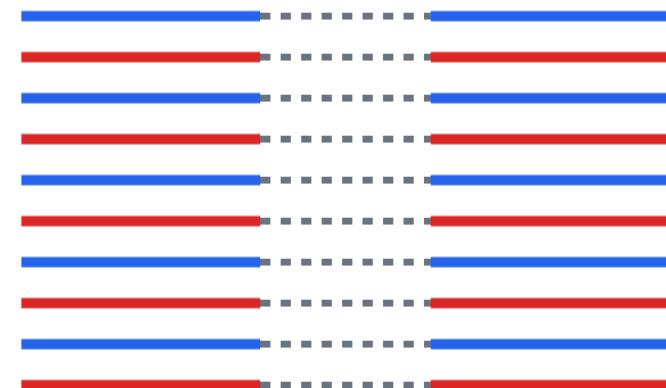
Técnica com alto impacto científico que revolucionou a biologia molecular



Iniciou a era genômica moderna



Alto impacto científico



Características do Sanger

Tecnologia com leituras longas de alta precisão mas baixo rendimento para aplicações genômicas



Leituras Longas

500 a 1.000 pares de bases



Alta Precisão

Baixa taxa de erro



Baixo Rendimento

Processamento limitado por amostra

Transição para o NGS



Necessidade de alto rendimento



Impulsionou avanços tecnológicos



Explosão da genômica

Alto Rendimento

Necessidade de processar milhões de sequências simultaneamente para estudar genomas complexos

Inovação Tecnológica

Desenvolvimento de novas abordagens de sequenciamento baseadas em amplificação e detecção em paralelo

Expansão Genômica

Capacidade de sequenciar genomas completos de organismos não-modelo e estudar diversidade genética

Origem do NGS



Primeiras plataformas comerciais em 2005



Introdução do conceito de shotgun em larga escala



Revolução metodológica

Sequenciamento de Nova Geração (NGS)



Alto Rendimento

Tecnologia de alto rendimento capaz de sequenciar milhões de fragmentos simultaneamente



Processamento em Paralelo

Capacidade de processamento massivo em paralelo, aumentando dramaticamente a velocidade



Redução de Custos

Redução drástica de custos por base sequenciada, tornando o genoma acessível

Características do NGS



Processamento Simultâneo

Milhões de fragmentos de DNA sequenciados em paralelo



Leituras Curtas

Fragmentos de 50 a 500 pares de bases, dependentes da plataforma



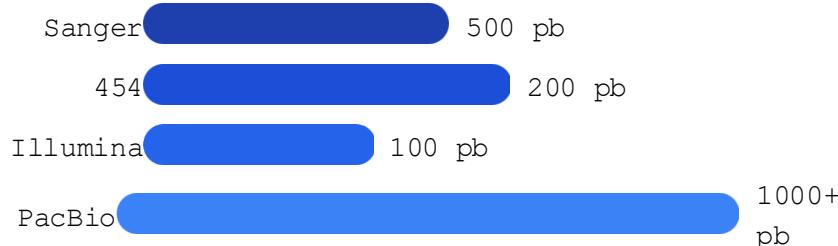
Alta Cobertura

Profundidade de sequenciamento elevada, permitindo montagem consistente

Leituras no NGS



Comprimento típico: 50 a 500 pb



Dependência da plataforma de sequenciamento

- 454: Leituras médias de 200-400 pb
- Illumina: Leituras de 300-500 pb
- Ion Torrent: Leituras de 200-400 pb



Impacto na montagem

- Leituras mais longas: Montagens melhores
- Leituras curtas: Mais complexo
- Comprimento: Fator crucial

Exemplos de Plataformas de 2^a Geração



454

- Tecnologia de sequenciamento por síntese
- Leituras de comprimento médio de 400-700 pb
- Análise de expressão gênica



Solexa / Illumina

- Tecnologia de sequenciamento por síntese
- Leituras de 50-300 pb
- Alta densidade de dados



Ion Torrent

- Tecnologia de sequenciamento sem necessidade de leitura
- Leituras de 100-400 pb
- Rápido e de baixo custo

Impacto da Segunda Geração



Democratização do sequenciamento



Aumento exponencial de dados



Base da filogenômica moderna



Terceira Geração de Sequenciamento



Sequenciamento de molécula única

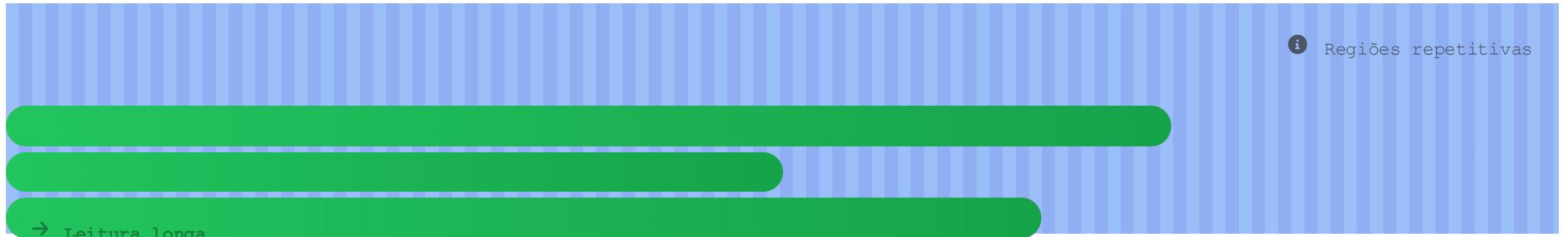


Leituras longas



Redução significativa de vieses

Leituras Longas



-  Fragmentos de dezenas de milhares de pares de bases
-  Capacidade de atravessar regiões repetitivas
-  Avanço estrutural em relação às leituras curtas



Aplicações das Leituras Longas



Montagem de Genomas Complexos

Permite a montagem de genomas complexos com alta cobertura e resolução de regiões repetitivas.



Resolução de Regiões Repetitivas

Capacidade de atravessar regiões repetitivas que dificultam a montagem com leituras curtas.



Fechamento de Genomas

Permite o fechamento de genomas completos, incluindo regiões que eram resistentes a abordagens anteriores.

Genomas Repetitivos



Alta Proporção

DNA repetitivo em altas proporções



Desafios

Desafios computacionais



Solução

Leituras longas resolvem desafios



Leituras longas atravessam regiões repetitivas

Exemplo de Genoma Complexo



Genoma Humano

Alta proporção de DNA repetitivo, regiões complexas e desafios computacionais para a montagem



Organismos Não-Modelo

Variedade de genomas complexos em plantas, animais e microrganismos não cultivados em laboratório



Grandes Genomas Eucariotos

Características comuns: tamanho variável, conteúdo repetitivo e estrutura organizacional desafiadora

Custo por Megabase



Redução histórica dramática de custos ao longo do tempo

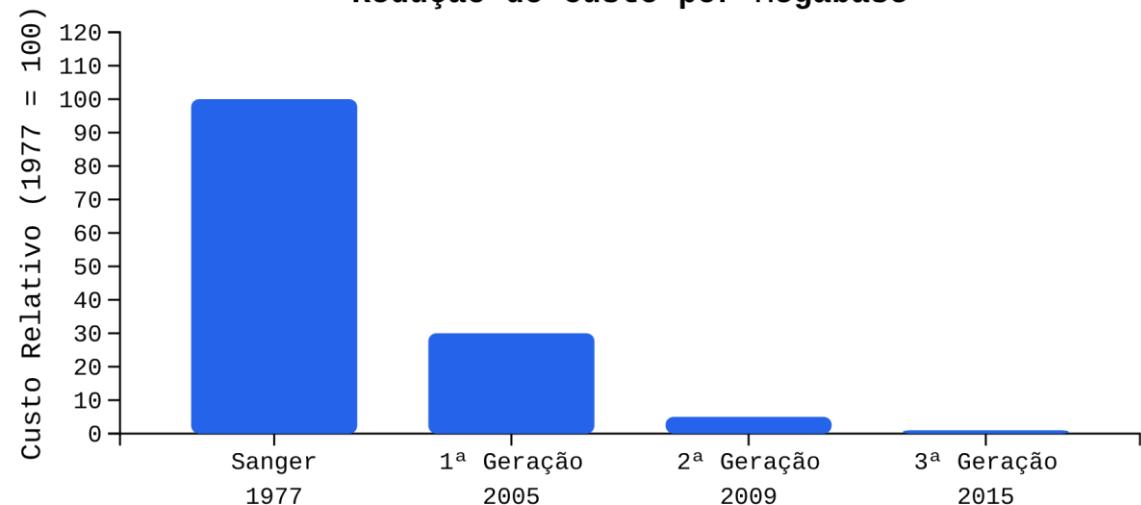


Comparação entre tecnologias de sequenciamento



Impacto científico das mudanças de custo

Redução do Custo por Megabase



Custo por Genoma



Sequenciamento de genomas completos atingiu viabilidade econômica



Permite escala populacional



Transformou a biologia evolutiva e filogenômica



Técnicas de Redução de Complexidade



UCEs

- ✓ Regiões ultraconservadas
- ✓ Flancos informativos



AHE

- ✓ Hibridização adaptativa
- ✓ Enriquecimento direcionado



Subamostragem

- ✓ Redução aleatória
- ✓ Conservação de loci

Etapas Gerais do NGS



1. Preparação de
Amostras



2. Sequenciamento



3. Análise de Dados

i Cada etapa é crucial para a qualidade do resultado final e requer controle rigoroso de qualidade.

Fluxo Geral do Sequenciamento



Extração de DNA

Isolamento e purificação do DNA da amostra



Construção de Bibliotecas

Fragmentação, padronização e preparação para sequenciamento



Geração de Dados

Leitura e análise das sequências geradas

- Cada etapa é critical para a qualidade dos dados finais e deve ser otimizada conforme o objetivo do estudo.

Construção de Bibliotecas

Preparação padronizada do DNA através de fragmentação controlada para sequenciamento



1. Preparação do DNA



2. Fragmentação Controlada

Divisão do DNA em fragmentos de tamanho específico



3. Padronização

Ajuste do tamanho e qualidade dos fragmentos

Fragmentação do DNA

Métodos para fragmentação controlada do material genético



Mecânica

- Sonicação
- Impacto mecânico
- Fragments aleatórios



Enzimática

- Enzimas de restrição
- Digestão controlada
- Padrões específicos



Direcionada

- Amplificação PCR
- Captura de regiões específicas
- Enriquecimento alvo

Tamanho dos Fragmentos



Compatibilidade com Plataforma

Cada tecnologia de sequenciamento possui faixas de tamanho específicas para otimizar o desempenho.



Impacto no Sequenciamento

O tamanho dos fragmentos afeta diretamente a cobertura, a qualidade das leituras e a montagem do genoma.



Planejamento Experimental

A escolha do tamanho dos fragmentos deve ser feita com base nos objetivos do projeto e nas características da plataforma.

Inserção de Adaptadores



Sequências Conhecidas

Adaptadores com sequências especificadas permitem amplificação e sequenciamento



Primers e Barcodes

Primers para amplificação, barcodes para identificação das amostras

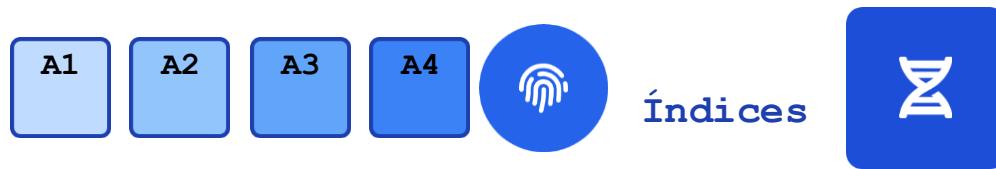


Identificação das Amostras

Adaptadores únicos permitem multiplexação e rastreabilidade



Multiplexação



Múltiplas Amostras

Sequenciamento simultâneo de múltiplas amostras em um único fluxo

Índices Específicos

Adição de sequências únicas para identificação pós-sequenciamento

Otimização de Custos

Redução significativa do custo por amostra e melhor aproveitamento do throughput

Dados Brutos



Reads de qualidade variável requerem filtragem adequada



Necessidade de controle de qualidade rigoroso



Impacto na análise subsequente

Qualidade dos Dados



Erros de Leitura

- Taxas de erro variáveis
- Erro de qualidade



Viés de Plataforma

- Viés de GC
- Viés estrutural

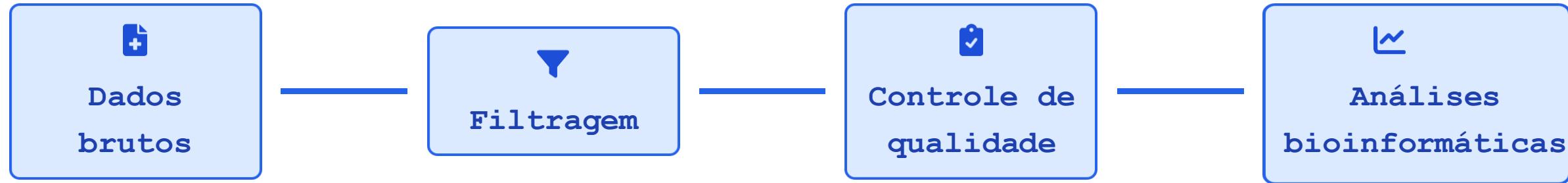


Avaliação Crítica

- Análise detalhada
- Interpretação cuidadosa

 Necessidade de avaliação crítica para interpretação adequada dos dados

Pós-processamento



Filtragem

Remoção de reads de baixa qualidade, contaminação e regiões com alta frequência de erros.

Métricas

Avaliação de qualidade baseada em phred scores, comprimento, GC content e distribuição de erros.

Preparação

Transformação dos dados para formatos compatíveis com algoritmos de montagem e análise filogenética.

Armazenamento de Dados



Grandes Volumes

Dados brutos do NGS requerem armazenamento em escala massiva



Infraestrutura Robusta

Sistemas de armazenamento em camadas para otimização



Gestão Adequada

Organização e versionamento de dados para análise



A gestão eficiente de dados é fundamental para a integridade da pesquisa filogenômica.

Impacto Computacional



Demanda por computação de alto desempenho



Necessidade de algoritmos eficientes



Impulsiona desenvolvimento bioinformático



NGS e Filogenômica



Geração de grandes matrizes de dados



Base para inferência evolutiva



Integração com métodos tradicionais



Limitações do NGS



Erros Sistemáticos

Viéses e erros constantes que podem distorcer os resultados



Dependência de Pipeline

Processamento requer pipelines específicos que impactam resultados



Interpretação Cuidadosa

Resultados requerem análise crítica para evitar erros conceituais

Tendências Futuras



Leituras Mais Longas

Desenvolvimento de tecnologias para leituras progressivamente mais longas, superando as limitações atuais do NGS



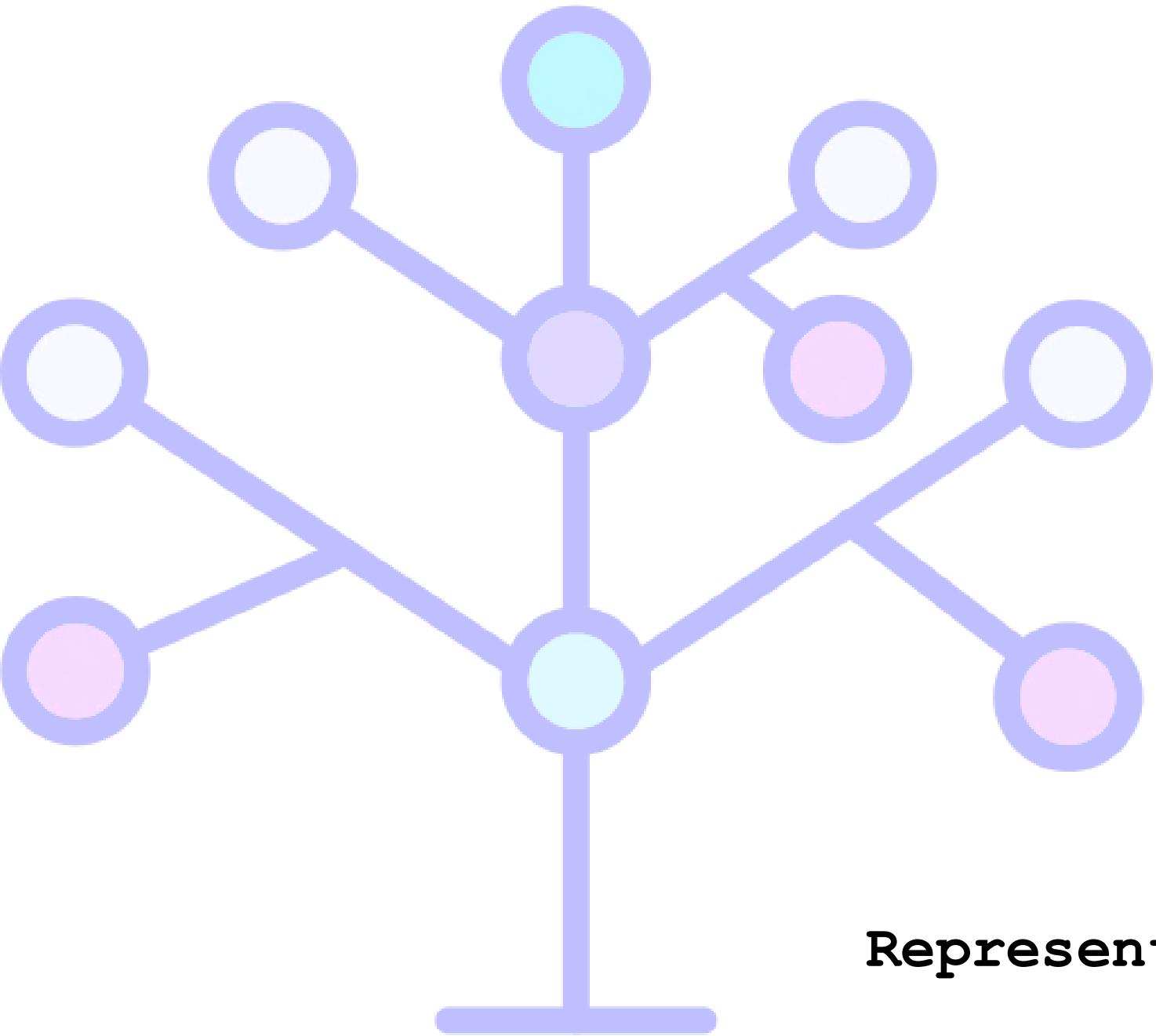
Custos Reduzidos

Continuação da tendência de queda exponencial de custos, tornando o sequenciamento acessível para aplicações em larga escala



Integração Multiômica

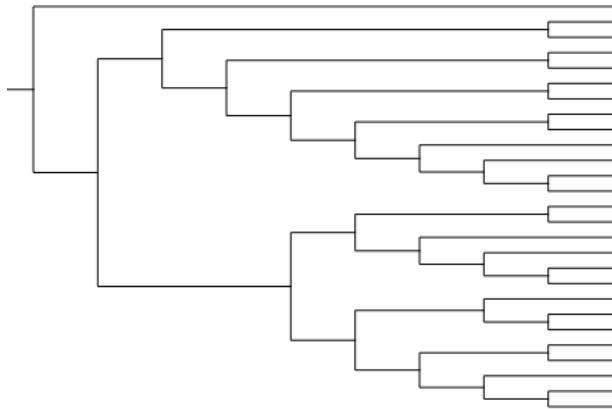
Avanço na integração de dados genômicos com outras camadas de informação biológica para uma visão mais completa da biologia evolutiva



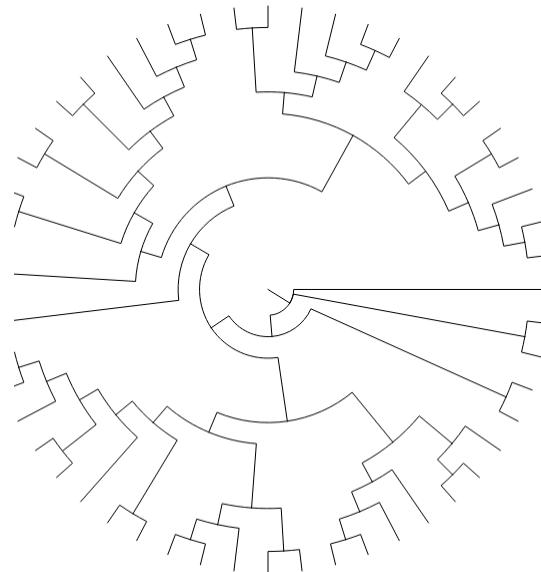
Representando filogenias

Arvores filogenéticas

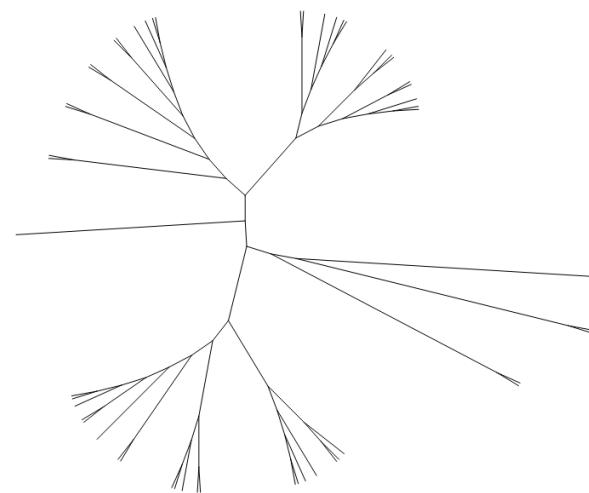
- As árvores filogenéticas (filogenias) = representação gráfica
 - Apresentadas de 3 formas principais:



Retangular

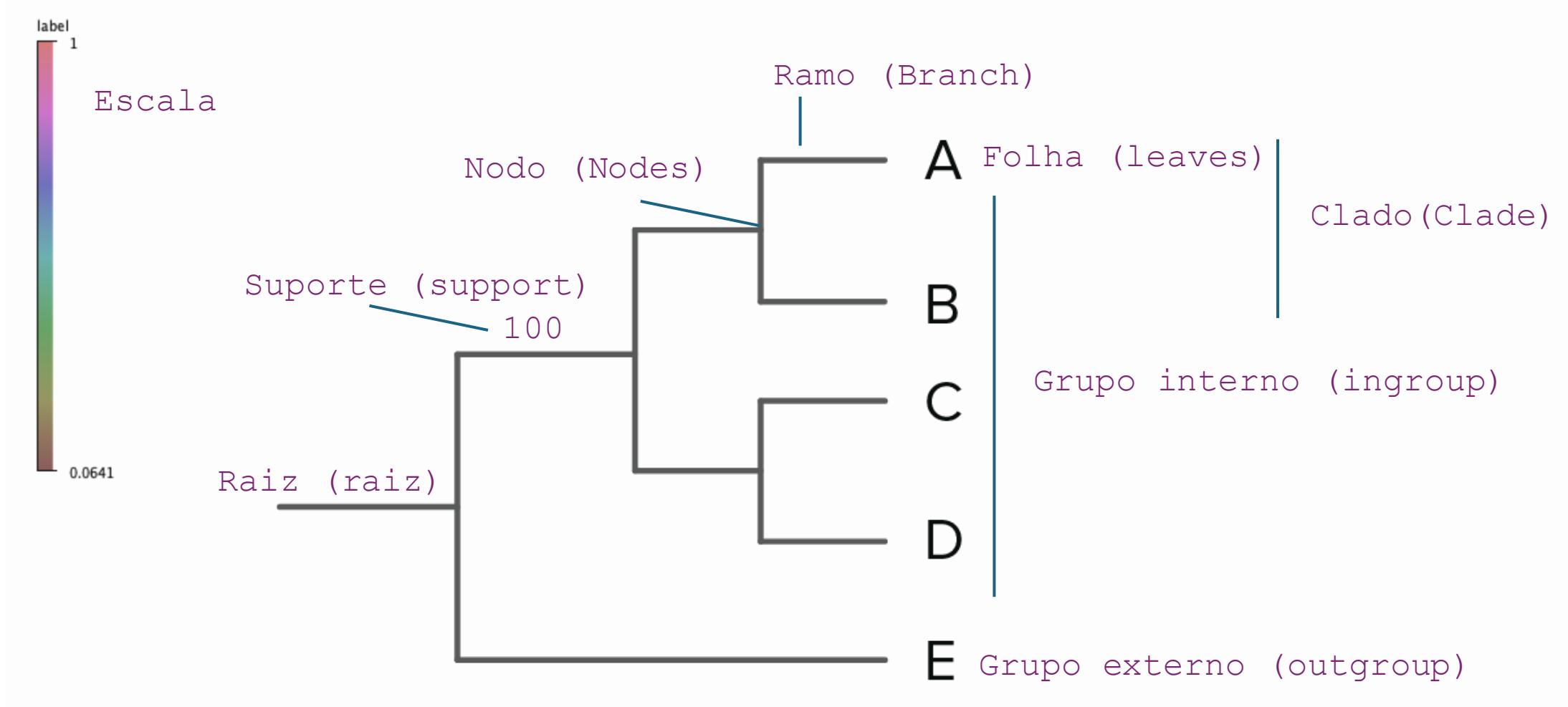


polar



radial

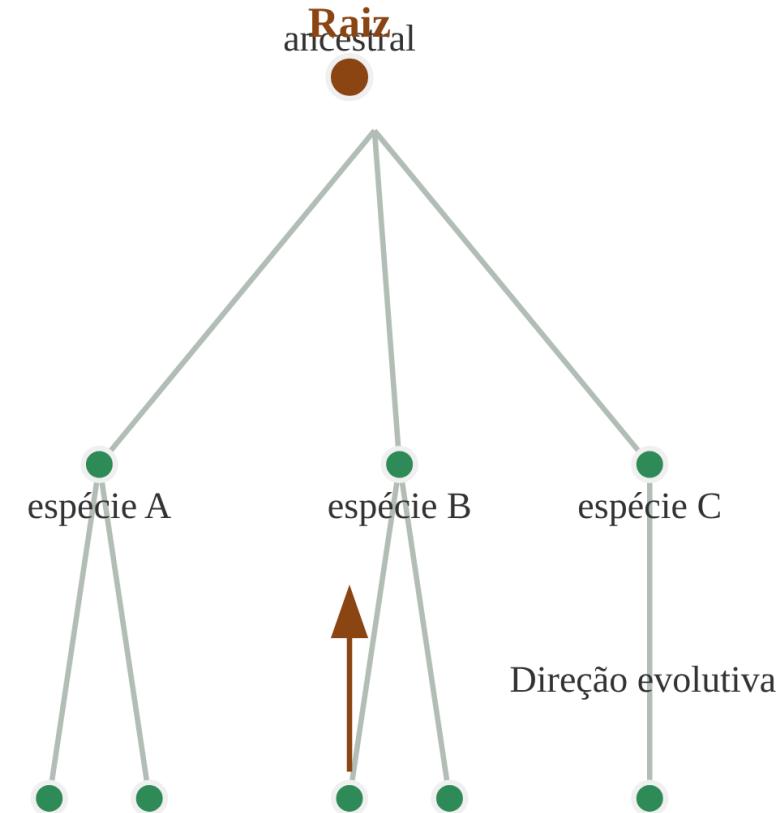
Estrutura básica de um filogenia



Árvores Enraizadas

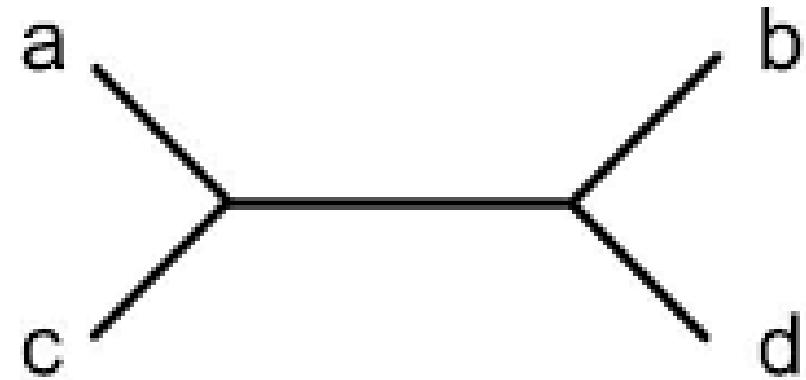
↗ **Direção evolutiva definida:** Possui uma raiz que indica o sentido do processo evolutivo

👉 **Identificação do ancestral:** Representa explicitamente o ancestral comum a todos os táxons



Árvores Não Enraizadas

- ☛ Representam relações evolutivas relativas sem definição de direção temporal
- 🚫 Ausência de ancestral externo (outgroup) definido
- 💡 Uso analítico: foco nas relações
- ⌚ Equivalentes a árvores enraizadas em termos de informação filogenética



Enraizamento com Outgroup

✓ Critérios para Seleção

- Distância evolutiva adequada (não muito próxima nem muito distante)
- Representatividade do grupo ancestral
- Disponibilidade de dados morfológicos ou moleculares

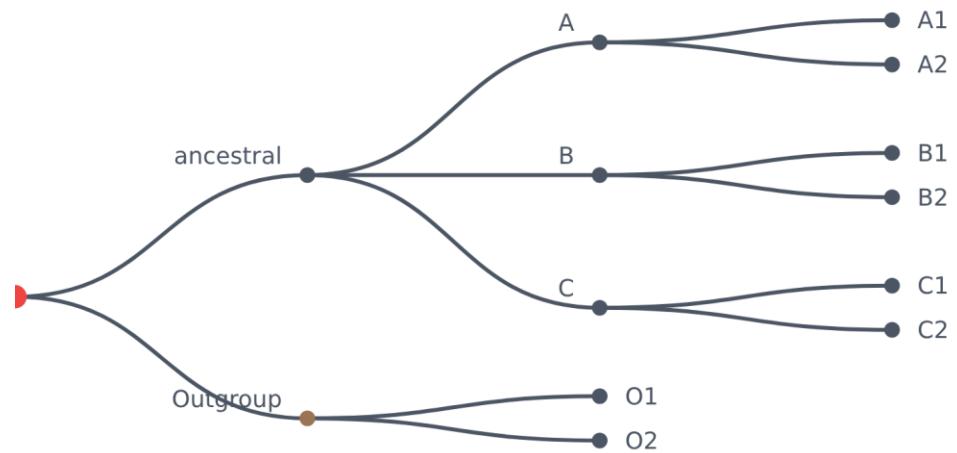
⚠ Riscos Methodológicos

- Outgroup muito próximo pode levar a enraizamento incorreto
- Escolha inadequada pode distorcer a topologia filogenética

🧪 Exemplos Práticos

- Enraizamento de árvores de primatas usando loris (grupo mais distante)
- Enraizamento de árvores de mamíferos usando répteis (outgroup externo)
- Enraizamento de árvores de aves usando dinossauros (outgroup histórico)

Ilustração do Conceito de Outgroup



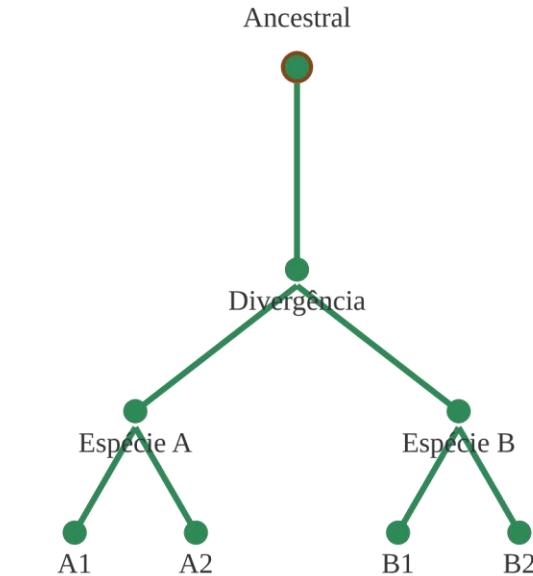
Espécies do grupo em estudo
Outgroup
Nodo de enraizamento

Nós Internos

 **Ancestrais comuns hipotéticos:** Representam espécies que não foram preservadas no registro fóssil mas que servem como pontos de divergência na árvore filogenética

 **Eventos de divergência:** Marcam momentos em que linhagens evolutivas se separaram, resultando em ramificações na árvore filogenética

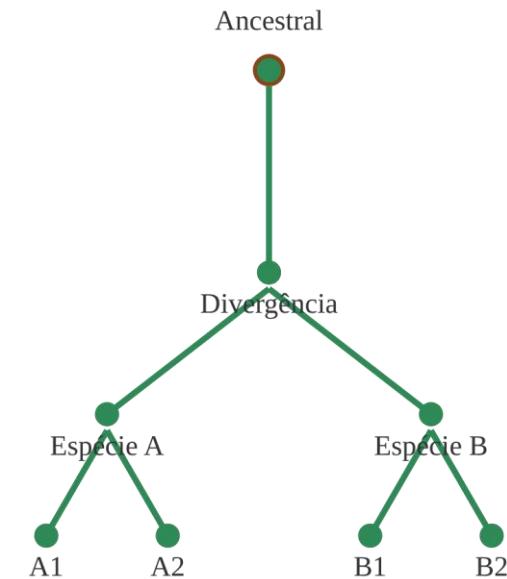
 **Natureza inferencial:** Os nós internos representam hipóteses sobre a história evolutiva que devem ser testadas com base nos dados disponíveis



 **Importante:** A posição dos nós internos influencia diretamente a topologia da árvore e a interpretação das relações evolutivas

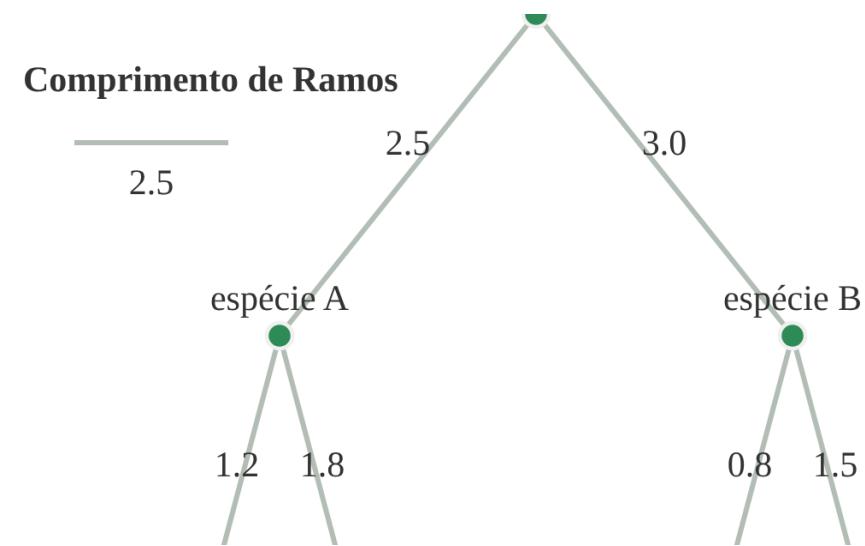
Ramos

- /Branch icon/ Conexões entre linhagens que representam continuidade evolutiva
- /Timeline icon/ Representação visual da evolução ao longo do tempo
- /Dna icon/ Significado biológico: linhagens descendentes e mudanças genéticas



Comprimento de Ramos

- Número de substituições:** Medida do total de mudanças nucleotídicas ou aminoácidas ao longo do ramo
- Tempo evolutivo:** Quando baseado em taxas de mutação, representa o intervalo de tempo entre eventos evolutivos
- Informação quantitativa:** Permite comparações entre ramos e testes de hipóteses sobre processos evolutivos



Topologia Filogenética

💡 O que é Topologia?

A topologia filogenética refere-se à organização estrutural das relações evolutivas entre os táxons, representada através da posição dos nós e ramos em uma árvore.

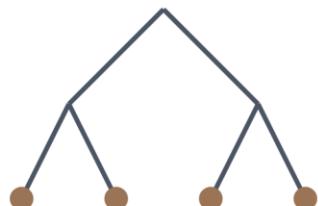
➡️ Independência da Métrica

As topologias são conceitos independentes da métrica temporal. A mesma topologia pode ser representada com diferentes comprimentos de ramos.

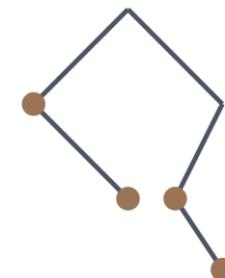
📊 Importância

A topologia filogenética é fundamental para a inferência evolutiva, fornecendo a base para a compreensão das relações de ancestralidade.

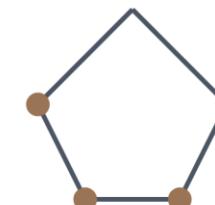
Topologia A



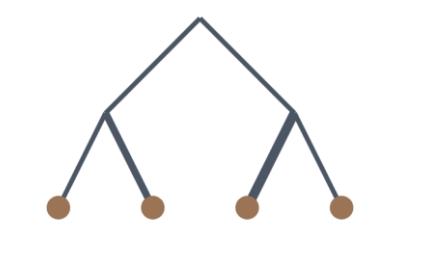
Topologia B



Topologia C

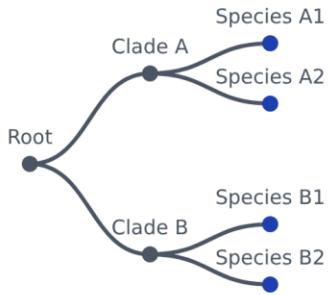


Mesma Topologia,
Diferentes Métricas



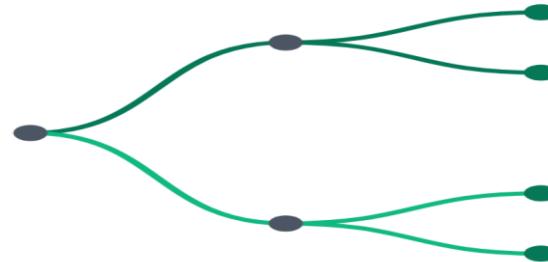
Tipos de Árvores

Cladograma



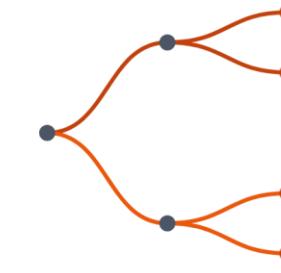
- Mostra apenas a topologia (estrutura de ramificação)
- Não representa tempo ou distâncias evolutivas
- Usado para ilustrar relações taxonômicas

Filograma



- Mostra topologia e distâncias evolutivas
- Comprimento dos ramos proporcional à quantidade de mudança
- Permite comparações de diversidade evolutiva

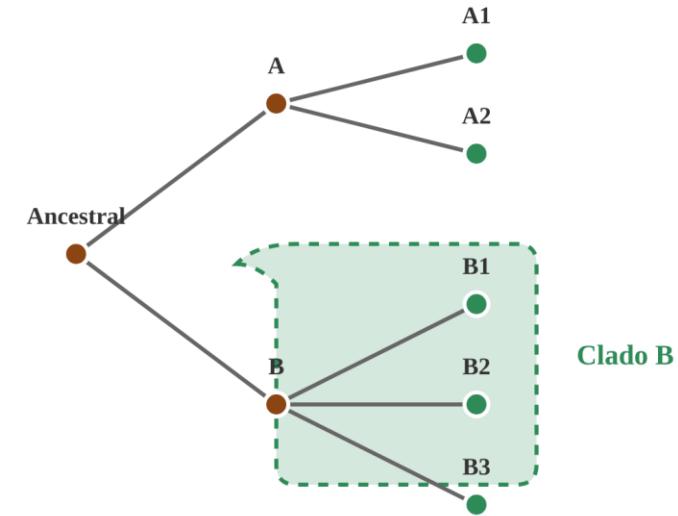
Cronograma



- Mostra topologia, distâncias e tempo evolutivo
- Nós enraizados representam eventos evolutivos no tempo
- Requer calibragem para estimativas de tempo

Monofilia

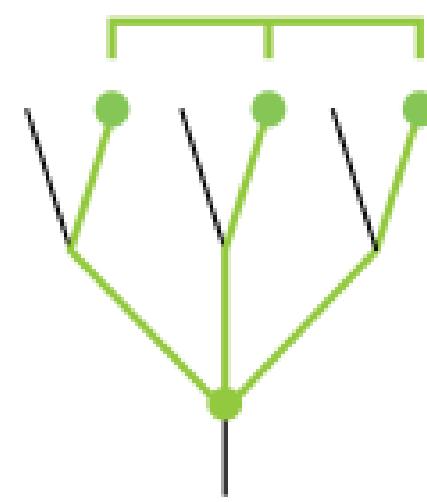
- 人群图标 **Grupos monofiléticos** que incluem um ancestral comum e todos os seus descendentes
- 连接图标 **Unidade fundamental** da classificação filogenética moderna
- 勾图标 **Grupos naturais** que refletem a história evolutiva real
- 天平图标 **Base da sistemática** para uma taxonomia evolutivamente informativa



Área destacada representa um clado: ancestral comum + todos descendentes

Parafilia

- **Exclusão de descendentes:** Grupos que incluem um ancestral e alguns, mas não todos, dos seus descendentes.
- ⚠ **Problemas conceituais:** Contradiz o princípio de inclusão completa dos descendentes em grupos taxonômicos.
- **Implicações taxonômicas:** Pode levar a classificações redundantes e dificultam relações evolutivas claras.

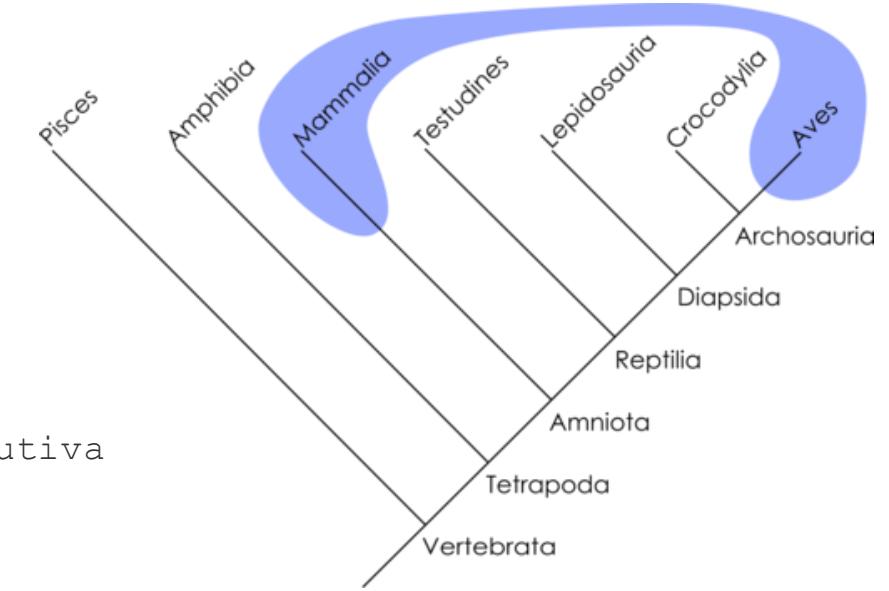


Polifilia

⚠️ **Agrupamentos artificiais** que não refletem história comum

⟳ **Convergência evolutiva** leva a características similares por origem independente

⚠️ **Erros de classificação** que distorcem a verdadeira história evolutiva



"Grupos polifílicos são resultantes de convergência, não representando ancestralidade comum."

Homologia

Conceito Fundamental

Homologia refere-se à ancestralidade comum entre caracteres, ou seja, características compartilhadas por espécies devido a herança de um ancestral comum.

- Diferente de **similaridade** – que pode surgir por convergência
- Base para inferência de relações evolutivas
- Fundamental para construção de árvores filogenéticas

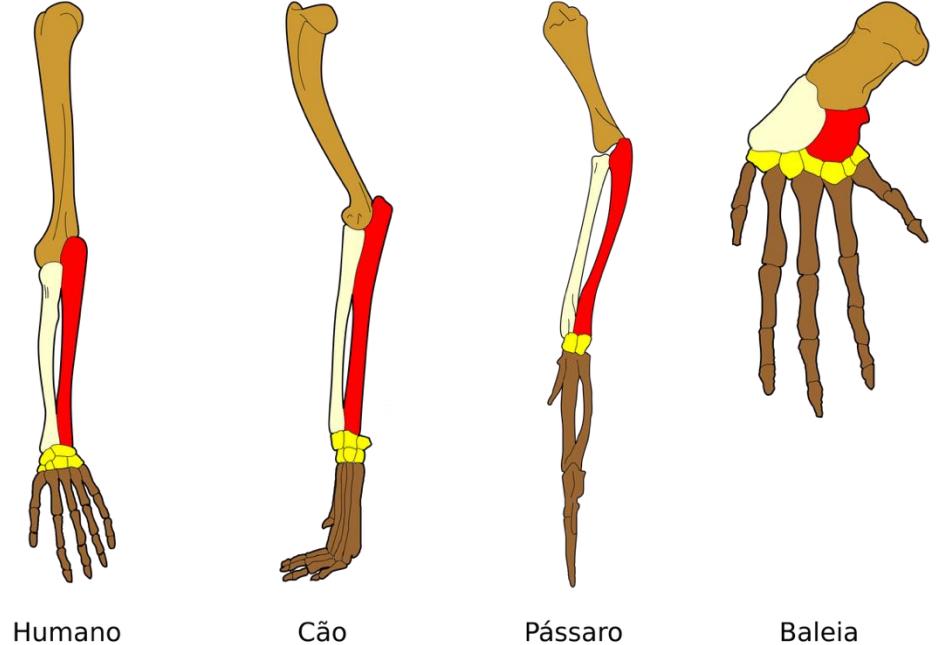
💡 Importante distinguir:

Homologia

Herança de ancestral comum

Similaridade

Pode surgir por convergência



Humano

Cão

Pássaro

Baleia

Os caracteres homólogos são herdados do ancestral comum

Homoplasia

Homoplasia refere-se a semelhanças entre espécies que não são resultado de ancestralidade comum, mas sim de processos evolutivos independentes.



Convergência

Evolução de características semelhantes em diferentes linhagens devido a pressões seletivas semelhantes.

- Mesmo ambiente => mesmos resultados
- Ex: asas de aves e morcegos



Paralelismo

Evolução independente de caracteres semelhantes em linhagens proximamente relacionadas, partindo de estados ancestrais similares.

- Mesma mutação => mesmos resultados
- Ex: duplicações gênicas independentes



Reversão

Restauração de um estado ancestral por meio de mutação ou conversão gênica.

- Voltar ao estado ancestral
- Ex: perda de capacidade de voo em aves



Importante distinguir homoplasia de homologia para construir árvores filogenéticas.

Homologia Molecular

Definição

Homologia molecular refere-se à similaridade entre sequências de DNA, RNA ou proteínas que se originam de um ancestral comum. É a base para inferir relações filogenéticas.

Pontos-Chave

- Sequências como caracteres: cada posição em uma sequência pode ser tratada como um caractere para análise filogenética.
- Alinhamento como hipótese: o alinhamento múltiplo representa uma hipótese sobre a homologia entre sequências.
- A homologia molecular é fundamental para a filogenômica, permitindo a comparação de sequências para inferir relações evolutivas.

Alinhamento como Hipótese de Homologia

Espécie A

A T G C A T G C

Espécie B

A T G C A T G C

Espécie C

A T G C A T G T

Homologia Substituição

Homologia vs. Similaridade

A similaridade molecular é importante, mas a homologia (origem comum) é o que importa para a filogenética.

→ Similaridade ≡ Homologia ←

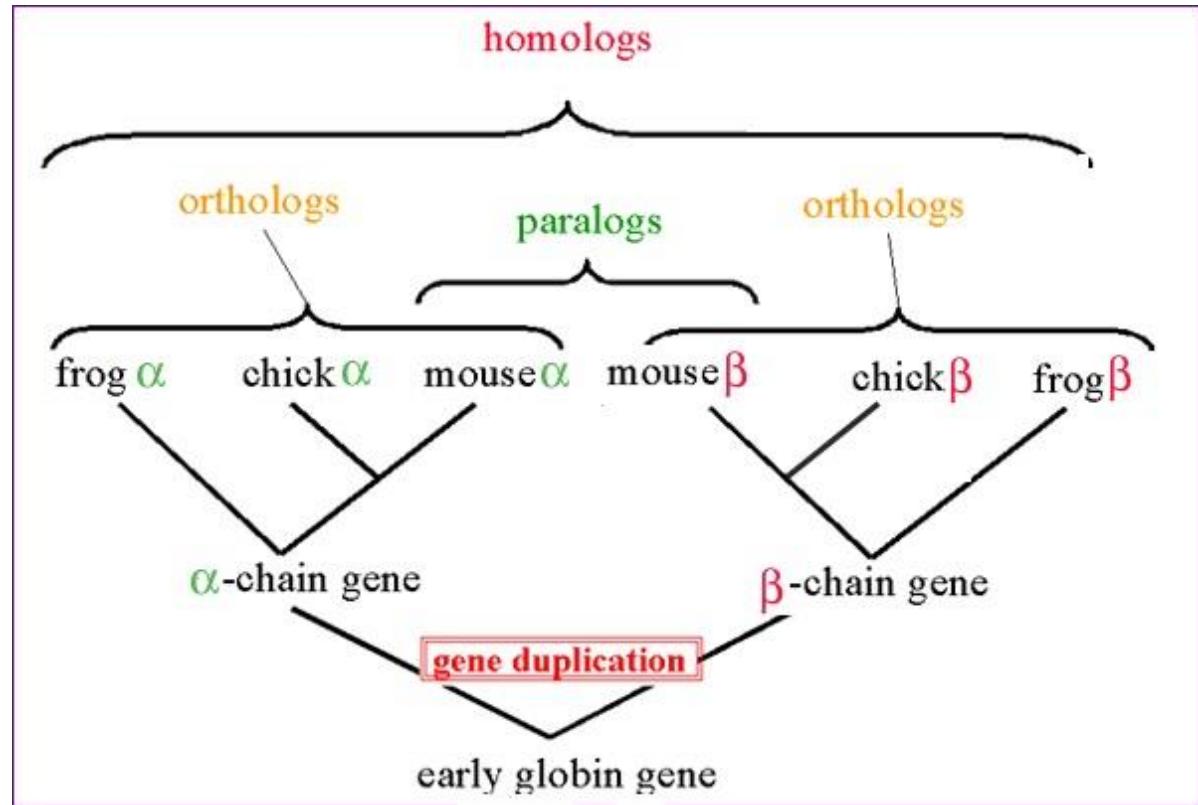
Ortologia

💡 O que são ortólogos?

- Genes que se originam através de **especiação**
- Divididos por linhagens evolutivas distintas
- Centralidade em filogenômica

💡 Por que são importantes?

Ortólogos são fundamentais para inferir relações evolutivas entre espécies e construir árvores filogenéticas robustas.



★ Centralidade em Filogenômica

Ortólogos são fundamentais para inferir relações evolutivas entre espécies e construir árvores filogenéticas robustas.

Paralogia

O que é Paralogia?

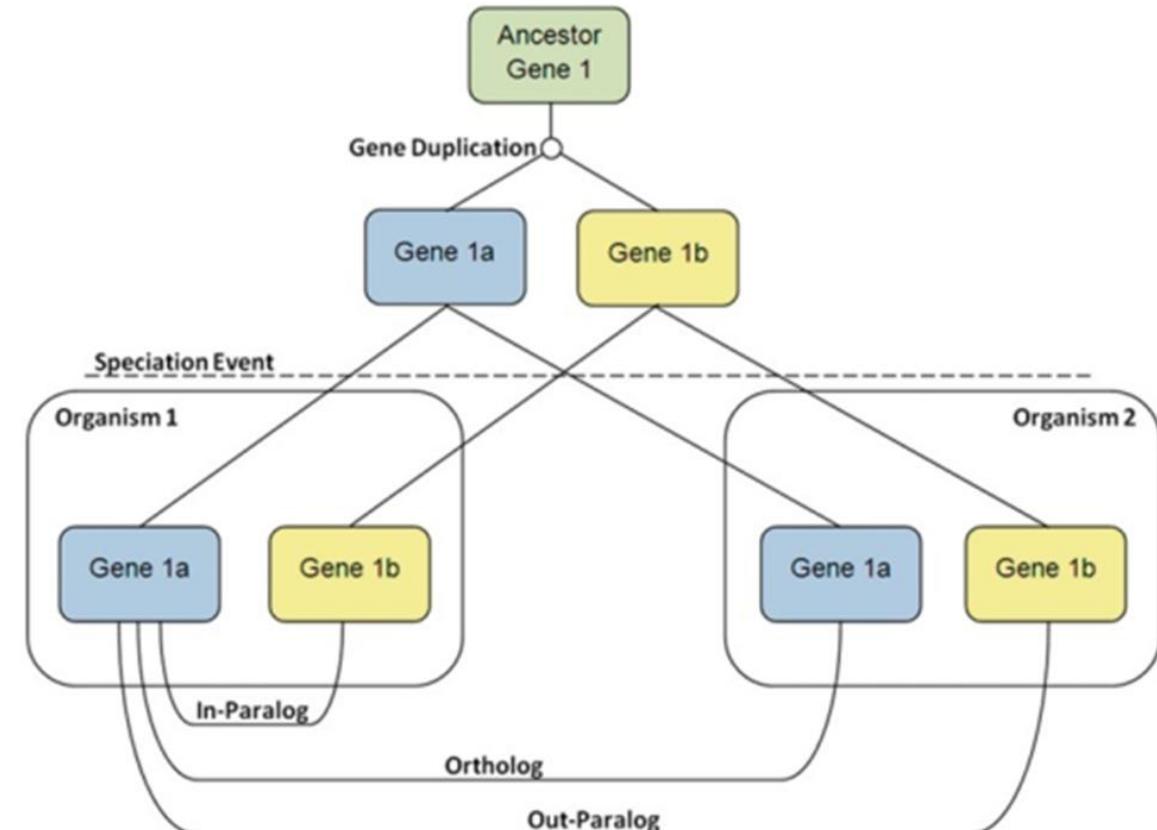
Duplicações gênicas que ocorrem durante evolução, resultando em múltiplas cópias de um gene.

Impacto na Topologia Filogenética

- Complica a interpretação filogenética
- Pode levar a conclusões errôneas
- Desafia métodos tradicionais

Consequências

- Dificuldade em identificar ortologias
- Risco de inferir relações incorretas
- Necessidade de abordagens mais sofisticadas



Xenologia

Transferência Horizontal de Genes

Xenologia refere-se à transferência de genes entre espécies distintas, geralmente através de processos

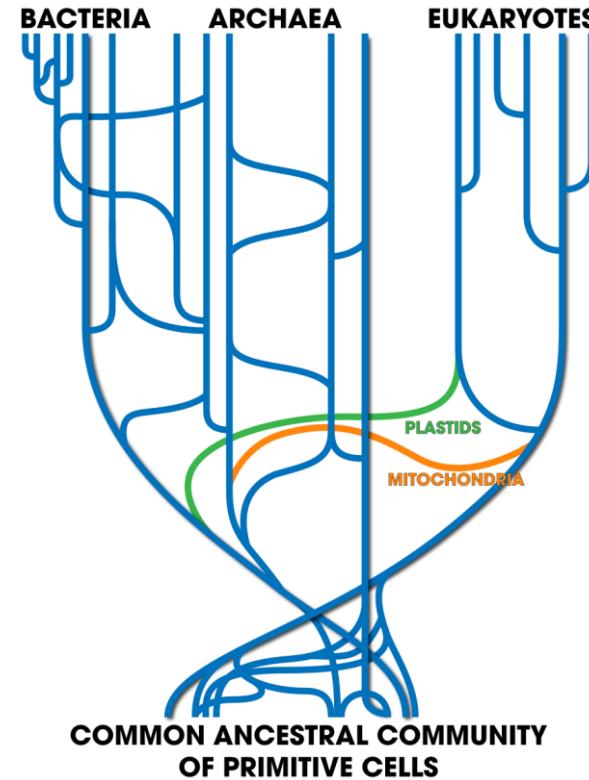
como:

- Transferência lateral entre organismos
- Processos de endosimbiose
- Viral transdução

Casos Clássicos Documentados

Exemplos destacados de transferência horizontal de genes:

- Genes mitocondriais transferidos para o núcleo
- Transferência de genes entre bactérias e eucariotos
- Inserções de genes virais no genoma hospedeiro



i A transferência horizontal de genes é um importante mecanismo de evolução e pode ser detectada através de análises filogenéticas.

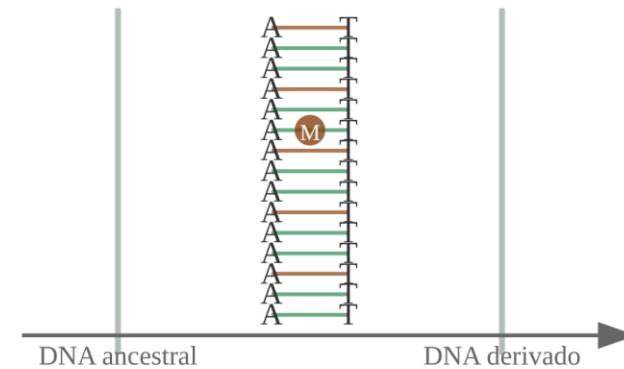
Escalas Evolutivas

A filogenética opera em diferentes escalas evolutivas, cada uma com seu próprio nível de complexidade e informações evolutivas.



Introdução à Evolução Molecular

-  Mudanças evolutivas em nível de DNA
 -  Acúmulo de alterações genéticas ao longo do tempo
 -  Origem do sinal filogenético observado
 -  Base para inferências evolutivas



Sequência ancestral:

Sequência derivada com mutações:

ACGTACGGT**T**ACGTACGTACGTACGTACGTACGTACGTACGT

Tipos de Substituição

↪ Substituições Sinônimas

- Mudanças que não alteram o aminoácido codificado
- Freqüência maior em regiões codificantes
- Efeito neutro ou benéfico para a função

➡ Substituições Não Sinônimas

- Mudanças que alteram o aminoácido
- Pode impactar a função proteica
- Classificadas por pressão seletiva

Consequências Funcionais

- Silenciosa: sem efeito na proteína
- Mássima: alteração radical
- Neo: nova função parcial

Substituição Sinônima

$\text{CAA} \rightarrow \text{GAA}$
Glutamina → Glutamato

Substituição Não Sinônima

$\text{CAA} \rightarrow \text{CAG}$
Glutamina → Glutamina (mudança)

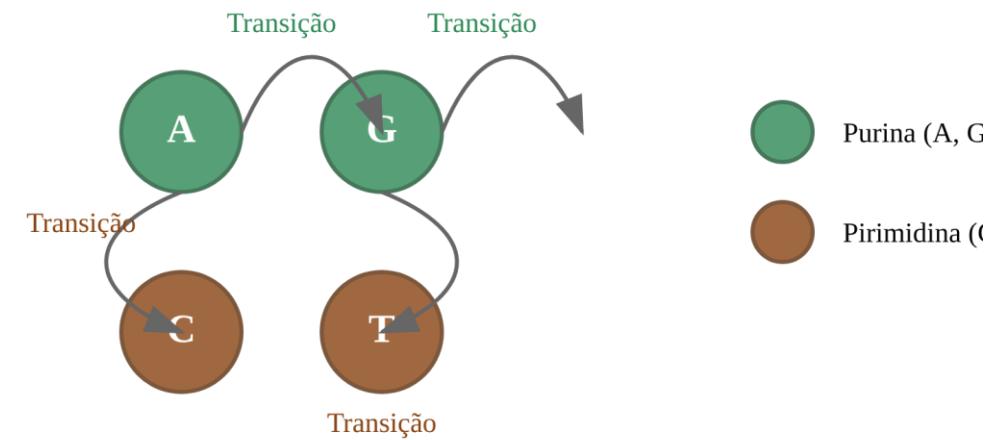


Transições

- ☒ Substituições entre bases quimicamente similares
- ☒ Maior frequência de ocorrência na evolução molecular
- ⚙️ Importância nos modelos de substituição

i Exemplos

- : A ↔ G (purinas)
- C ↔ T (pirimidinas)



Transversões

Transversões são mudanças entre bases quimicamente distintas:

- ➡ Troca entre purinas (A, G) e pirimidinas (C, T)
- ➡ Menor frequência evolutiva
- ➡ Alto valor informativo filogenético

Importante:

Transversões são fundamentais para detectar divergência evolutiva, especialmente em sequências com alta similaridade.

Mecanismo de Transversão



Inserções e Deleções (Indels)

- Ganho ou perda de nucleotídeos:** Mutações que alteram o comprimento da sequência
- Desafios de alinhamento:** Complicações para alinhamento múltiplo e análise filogenética
- Informação evolutiva:** Pode fornecer evidências de eventos evolutivos
- Risco de ruído:** Pode introduzir homoplasia e distorcer resultados filogenéticos

Exemplo de Alinhamento com Indels

Seq1: ATCGATCGATCG
Seq2: ATCG-TCGATCG
Seq3: ATCGAT--TCG
Seq4: AT-TCGATCG

Indels de 1, 2 e 3 nucleotídeos apresentam diferentes implicações

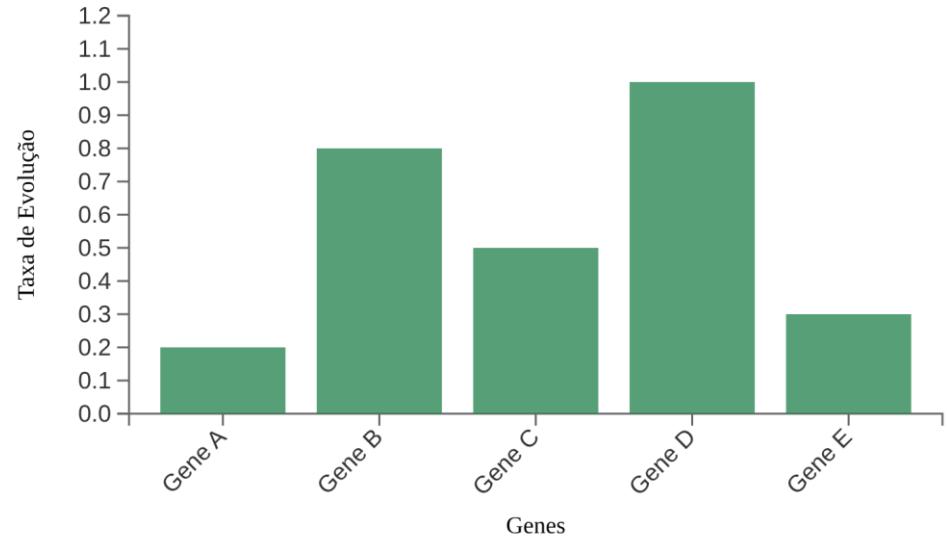
Importante: Ajustes específicos de modelo são necessários para lidar com indels em filogenética.

Taxas de Evolução

- DNA** **Variação entre genes:** Diferentes genes evoluem a taxas distintas devido à pressões seletivas, tamanho populacional e eficiência de reprodução.
- lhagens** **Variação entre linhagens:** Em algumas linhagens, a evolução ocorre mais rapidamente, resultando em ramos longos nas árvores filogenéticas.
- Impacto na inferência:** A heterogeneidade nas taxas de evolução desafia modelos simples e pode levar a resultados filogenéticos distorcidos se não for adequadamente modelada.

Exemplo de Taxas de Evolução

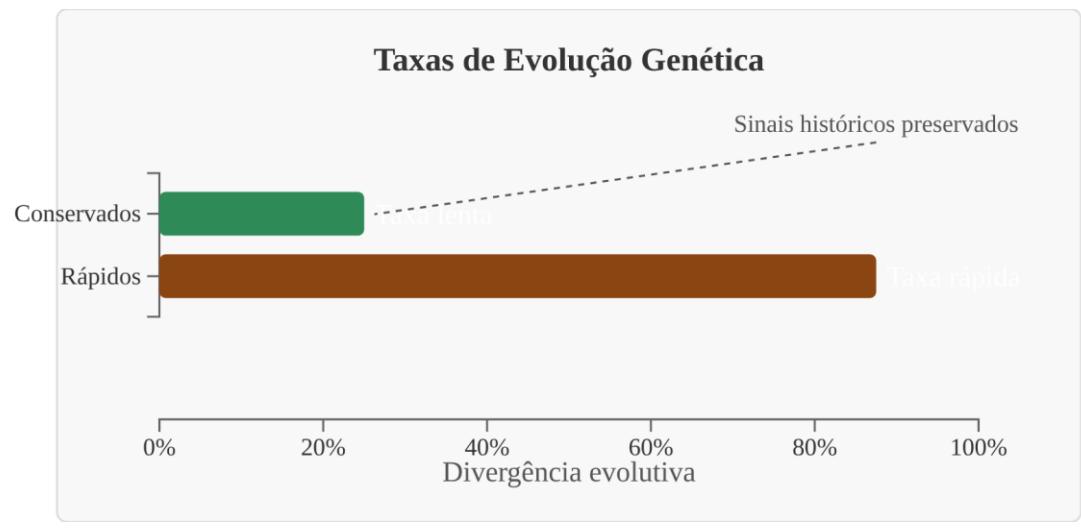
Variação de Taxas entre Genes (Conceptual)



A heterogeneidade nas taxas de evolução é um desafio constante para a filogenética

Heterogeneidade Entre Genes

- ☒ Genes conservados evoluem a taxas mais lentas e preservam sinais históricos
- ⚡ Genes rápidos acumulam mudanças e podem obscurecer relações evolutivas antigas
- 🕒 Escalas temporais diferentes requerem marcadores genéticos adequados
- 💡 Escolha de marcadores baseada em taxas evolutivas e objetivos analíticos



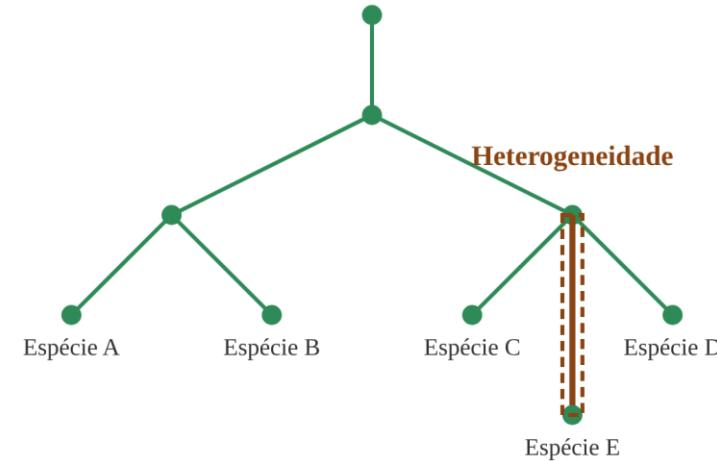
Heterogeneidade Entre Linhagens

- ⌚ **Relógio molecular imperfeito:** taxas de evolução variam entre linhagens, violando a suposição do relógio molecular
- 🕒 **Ramos longos:** linhagens que evoluíram rapidamente geram ramos longos que podem criar artefatos filogenéticos
- ⚠ **Artefatos filogenéticos:** resultados distorcidos devido à heterogeneidade evolutiva

Consequências:

- Distância evolutiva exagerada
- Topologias incorretas
- Erros na inferência de ancestralidade

Exemplo: Ramos longos podem levar a agrupamentos artificiais



Saturação

 **Múltiplas substituições:** Acúmulo de mudanças em sítios evolutionariamente conservados

 **Perda de sinal histórico:** Dificuldade em reconstruir a verdadeira relação evolutiva

 **Limites dos dados:** Atingimos um ponto onde mais mutações não podem ser detectadas

Principalmente quando:

- taxas de evolução são altas
- ramos são longos
- o tempo de divergência é grande
- regiões do DNA evoluem rápido (ex.: 3^a posição de códons)

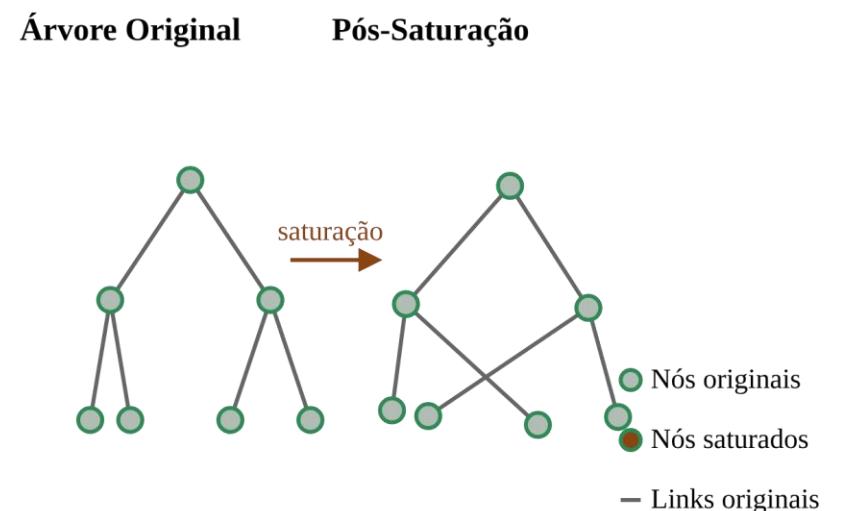


Consequências da Saturação

👉 **Homoplasia:** Similaridade convergente que pode ser interpretada como evidência de parentesco

👉 **Distorção de distâncias:** Substituição de sítios originais por múltiplas substituições

👉 **Erros topológicos:** Alteração da estrutura da árvore filogenética



Introdução aos Modelos de Substituição

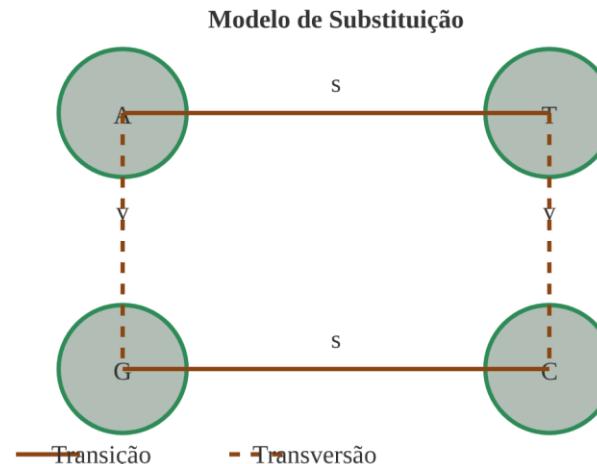
Modelos evolutivos são regras matemáticas que descrevem como o DNA muda ao longo do tempo.

Correção de mudanças ocultas: Modelos que corrigem para substituições que não são observáveis diretamente

Base estatística: Fundamentos probabilísticos que permitem estimativa de parâmetros evolutivos

Papel na inferência: Fundamentam métodos como máxima verossimilhança e bayesianos

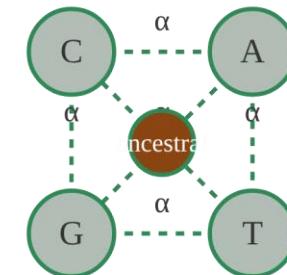
Representação: Funções matemáticas descrevendo taxas de mudança entre estados



Modelos de substituição são ferramentas essenciais para inferência filogenética rigorosa

Modelo Jukes-Cantor

- Modelo evolutivo mais simples com taxas iguais para todas as substituições
- Apenas um parâmetro (α) para descrever todas as taxas de substituição
- Assume que todas as posições nucleotídicas são equivalentes
- Uso didático e base para modelos mais complexos



💡 O modelo Jukes-Cantor corrige para mudanças ocultas e serve como baseline para modelos mais realistas.

Modelo Kimura 2-Parâmetros

Distingue transições de transversões

Tipos de substituições:

Transições (π_1)

Purina \leftrightarrow Purina ($A \rightleftharpoons G$)

Pirimidina \leftrightarrow Pirimidina ($C \rightleftharpoons T$)

Transversões (π_2)

Purina \leftrightarrow Pirimidina ($A \rightleftharpoons C, A \rightleftharpoons T, G \rightleftharpoons C, G \rightleftharpoons T$)

Vantagens:

Modela melhor o processo de substituição real

Mais preciso que modelos com taxa única (ex: Jukes-Cantor)

Limitações:

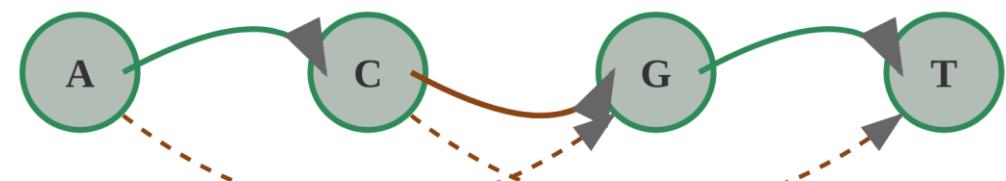
Ainda assume homogeneidade entre sítios

Não considera variação de taxas ao longo do genoma

Modelo HKY Hasegawa-Kishino-Yano

Modelo de substituição nucleotídica que estabeleceu padrões para análises filogenéticas avançadas.

- ✓ Incorpora frequências desiguais de bases (π_A , π_C , π_G , π_T)
- ↔ Taxas diferenciadas de substituição (k_1 para transições, k_2 para transversões)
- ✓ Maior realismo evolutivo que modelos anteriores
- ➡ Base para implementações em análise filogenética moderna

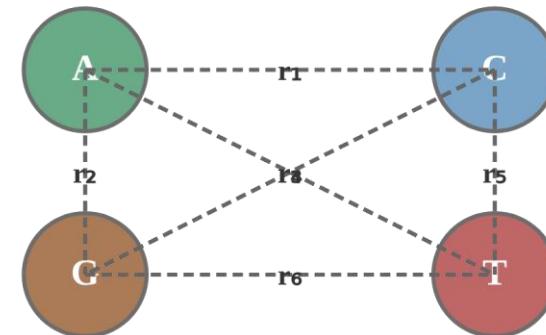


Modelo HKY generaliza o modelo K2P incluindo frequências nucleotídicas.

Modelo GTR - General Time-Reversible

- ☒ **General Time-Reversible:** Modelo mais complexo com taxas independentes para cada tipo de substituição
- ☒ **Máxima flexibilidade:** Adequa-se a diferentes padrões de evolução nucleotídica
- ☒ **Parâmetros:** 6 taxas de substituição, frequências bases, e parâmetro de reversibilidade temporal
- ☒ **Custo computacional:** Maior exigência de processamento
- ✓ **Vantagem:** Menos restrições sobre padrões de substituição

Modelo GTR - Taxas Independentes



Distribuição Gamma

- ↳ **Modelagem da variação de taxas:** Ferramenta estatística para descrever como as taxas de substituição variam entre sítios em uma sequência evolutiva.
- ↳ **Parâmetro alfa (α):** Controla a forma da distribuição. Valores baixos ($\alpha < 1$) indicam alta heterogeneidade de taxas, enquanto valores altos ($\alpha > 1$) aproximam a distribuição à uma curva delta.
- ↳ **Uso em filogenética:** Corrigir para heterogeneidade de taxas entre sítios, melhorando a precisão da inferência filogenética ao considerar que alguns sítios evoluem mais rapidamente do que outros.
- ↳ **Implicações:** Distribuições com α menores resultam em maior correção para mudanças ocultas, potencialmente aumentando a veracidade das árvores filogenéticas.

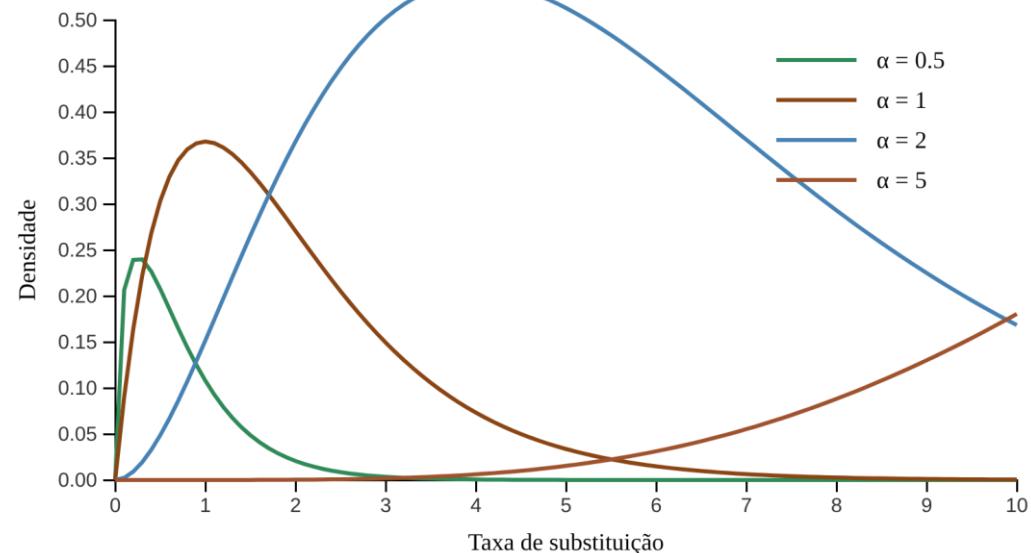


Figura: Distribuição Gamma para diferentes valores de alfa (α)

Sítios Invariantes

-  **Posições altamente conservadas** que não apresentam variação entre sequências analisadas
-  **Informação estrutural** importante refletindo restrições funcionais e evolutivas
-  **Complemento ao gamma** na modelagem da heterogeneidade de taxas entre sítios
-  **Modelagem estatística** melhora a estimativa de parâmetros evolutivos

Exemplo Conceitual

Sequência 1: ATCGGTACGGCATGC

Sequência 2: AGCAGTCCGGCATGT

Sequência 3: ATCTGTACGGCATGA

Sítios Invariantes

Posições conservadas em todas as sequências

Modelos como Aproximações



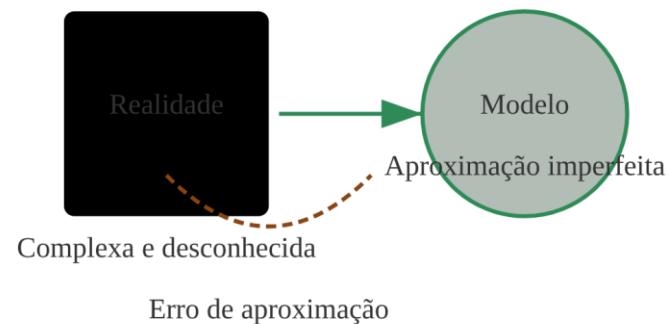
Nenhum modelo evolutivo é perfeito ou completamente verdadeiro



Modelos são simplificações que buscam capturar os aspectos essenciais da evolução



Escolha crítica baseada em propósitos analíticos e dados disponíveis



Seleção de Modelos

⚖️ Critérios de Informação

Os critérios AIC (Akaike) e BIC (Bayesiano) são ferramentas usadas para comparar modelos e selecionar o mais apropriado aos dados.

AIC (Akaike)

Critério que balança ajuste dos dados com complexidade do modelo.

$$\square \text{ AIC} = -2\ln(L) + 2k$$

- Preferência por modelos que ajustam bem
- Penaliza complexidade com k parâmetros

BIC (Bayesiano)

Critério que incorpora uma penalidade mais forte por complexidade.

$$\square \text{ BIC} = -2\ln(L) + k\ln(n)$$

- Penalidade aumenta com tamanho da amostra
- Mais severo com modelos complexos

Seleção de Modelo

JC69

| | |
|-------------|-----|
| AIC: | 100 |
| BIC: | 95 |

K80

| | |
|-------------|----|
| AIC: | 92 |
| BIC: | 87 |

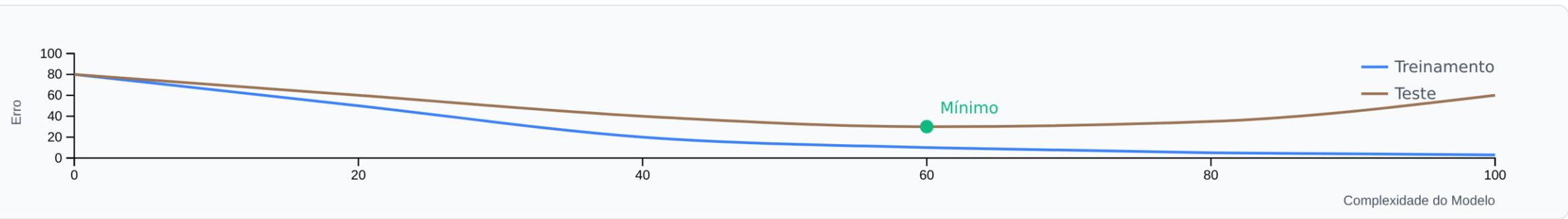
HKY

| | |
|-------------|----|
| AIC: | 88 |
| BIC: | 82 |

GTR

| | |
|-------------|----|
| AIC: | 85 |
| BIC: | 79 |

Overfitting vs Underfitting



⚠️ Overfitting

Modelo muito complexo

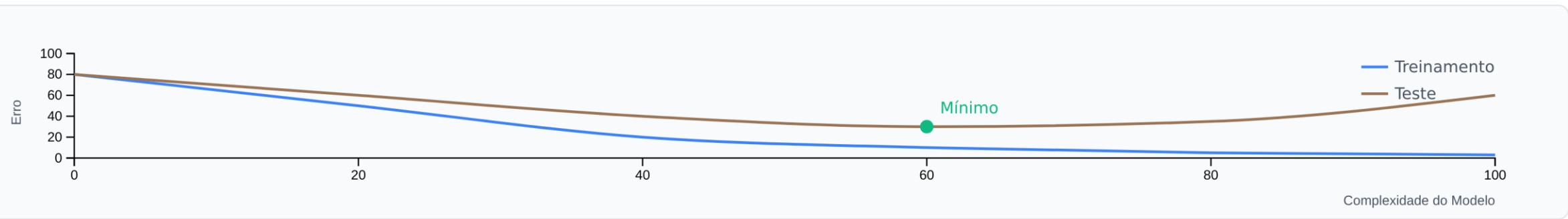
- Exemplo: Modelo GTR com muitos parâmetros
- Consequências:
 - Sensibilidade excessiva aos dados
 - Boa ajuste aos dados de treinamento
 - Pior performance em dados novos

ℹ️ Underfitting

Modelo muito simples

- Exemplo: Modelo JC69 para dados complexos
- Consequências:
 - Adequação insuficiente aos dados
 - Perda de informação evolutiva
 - Incapacidade de detectar relações

Overfitting vs Underfitting



⚠ Overfitting

Modelo muito complexo

- Exemplo: Modelo GTR com muitos parâmetros
- Consequências:
 - Sensibilidade excessiva aos dados
 - Boa ajuste aos dados de treinamento
 - Pior performance em dados novos

ℹ Underfitting

Modelo muito simples

- Exemplo: Modelo JC69 para dados complexos
- Consequências:
 - Adequação insuficiente aos dados
 - Perda de informação evolutiva
 - Incapacidade de detectar relações

*O melhor resultado filogenético não depende apenas de mais dados, mas de **dados adequados, modelos apropriados e interpretação crítica.***