Filogenética molecular aplicada

2. Modelos de evolução de sequências, máxima verossimilhança e bootstrapping

Matheus E. Bianconi
Extensão PUC-PR Nov/2023

Revisão aula 1

- Importância das árvores filogenéticas
- Ancestralidade comum e o conceito de homologia
- Terminologia e propriedades das árvores filogenéticas
- Inferindo filogenias, parte 1: métodos baseados em distância (NJ)
- Inferindo filogenias, parte 2: métodos baseados em caracteres
  parte 2.1: máxima parcimônia
- Limitações dos métodos basedos em distância e máxima parcimônia
- Prática: familiarização com o kit de ferramentas para análise filogenética

Plano de aula

- Modelos de evolução de sequências
- Métodos baseados em caracteres II: máxima verossimilhança
- Qual o melhor modelo?
- Medidas de suporte de uma árvore

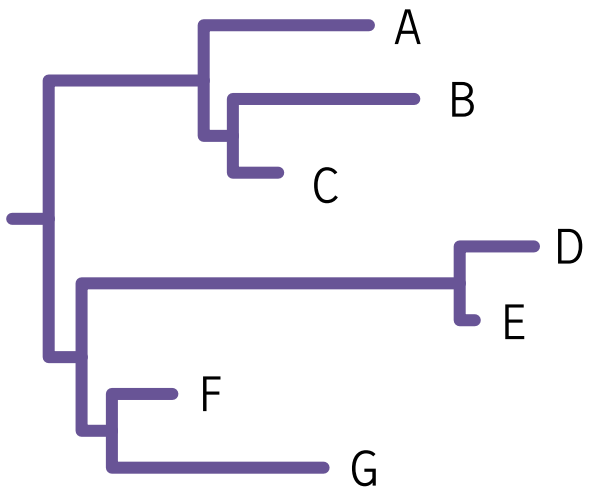
Prática: Inferência filogenética com máxima verossimilhança e bootstrapping, seleção de modelos de substituição

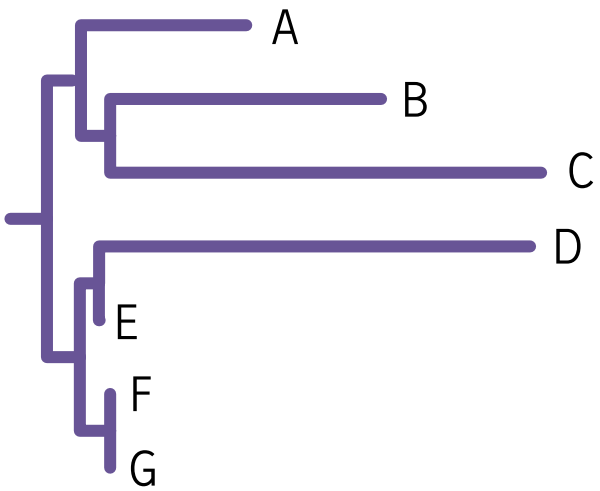
Modelos de evolução de sequências

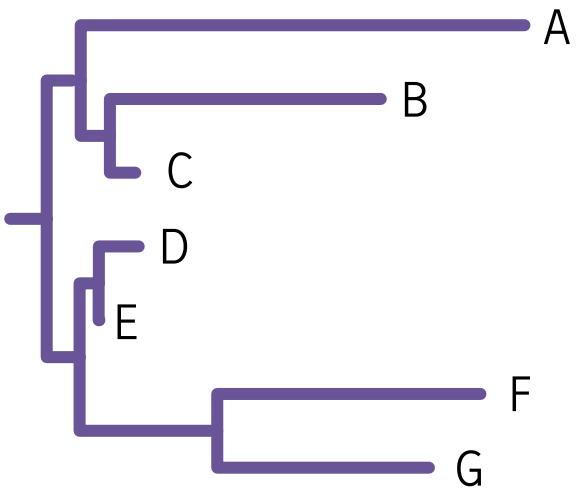
- Por que métodos de distância e parcimônia não são suficientes?
- Modelos de evolução de sequências

Por que não usar métodos mais simples?

- Métodos baseados em distância e parcimônia funcionam, mas sob condições restritas
  Em alguns casos, é quase garantido que eles retornarão uma árvore incorreta (e.g. atração de ramos longos –'long branch attraction')







Um método deve ser robusto o suficiente para lidar com as inúmeras histórias evolutivas possíveis

Por que não usar métodos mais simples?

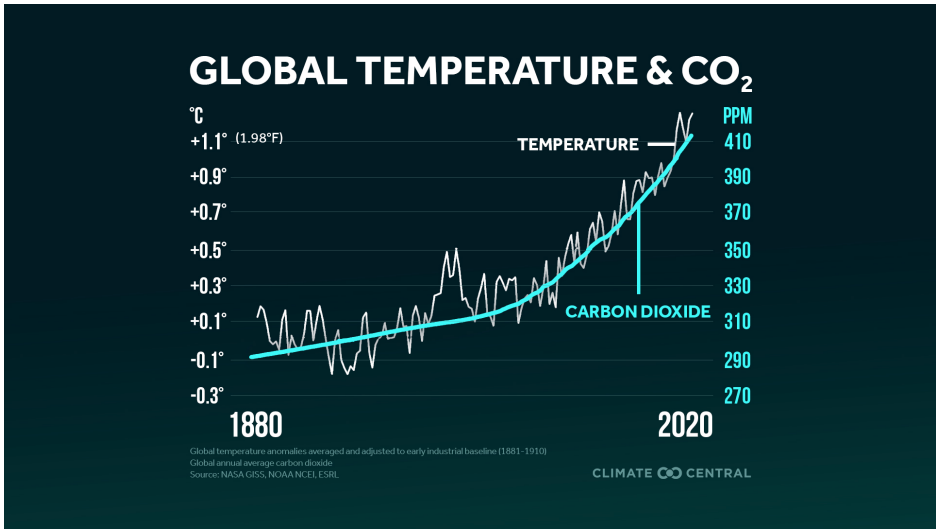
- Tais métodos não oferecem boas estimativas da distância genética entre sequências, e por isso podem produzir agrupamentos que são artefatos
- Como então produzir boas estimativas para a distância genética?
  Através de modelos → descrição matemática de como as sequências evoluem

O que são modelos?

- Descrição explícita de um fenômeno ou um processo
- Simplifica a realidade a fim de permitir predições quantitativas

  Exemplo:

  Como a temperatura do planeta varia de acordo com a quantidade de CO_2 emitida na atmosfera?

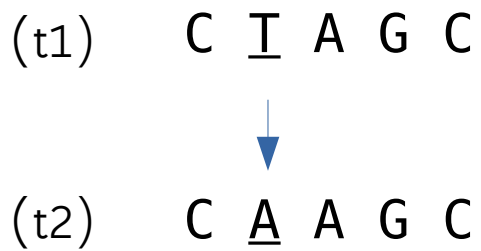


Um possível modelo:

<u>Variável resposta</u> = diferença na temperatura média do planeta

<u>Variáveis preditoras</u> = concentração atual de CO_2, quantidade de CO_2 emitida por ano, quantidade de CO_2 fixada por processos biológicos e químicos, fatores físico–químicos da atmosfera, etc...

Como modelar a evolução de sequências?



Objetivo do modelo: Calcular a probabilidade de uma base **i** ser substituída por uma base **j** entre os tempos **t1** e **t2**?

Possíveis premissas:

- Transições (A <–> G e C <–> T) são mais comuns que transversões
- A frequência das bases (A, G, C e T) é constante e igual a 25% cada
- A taxa evolutiva não é constante ao longo da sequência (i.e. algumas regiões mudam mais rapidamente que outras)

Cada uma dessas premissas pode ser expressa como um parâmetro no modelo

Modelos de evolução de sequências

- Para gerar os valores de probabilidade, é necessário um modelo com parâmetros que descrevem como as sequências evoluem
  Modelos de substituição/modelos de evolução de sequência de DNA
  (substitution models/models of DNA sequence evolution)
- Modelos descrevem a probabilidade uma base *i* ser substituída por uma base *j* em uma determinada posição da sequência (matriz de transição)

	A	C	G	T
A	–	a	b	c
C	a	–	d	e
G	b	d	–	f
T	c	e	f	–

$P(A \rightarrow C) = a$

$P(A \rightarrow G) = b$

$P(A \rightarrow T) = c$

...

Modelos de evolução de sequências

- Modelo mais simples: Jukes–Cantor (1969) 'JC69'

  A taxa de substituição é a mesma para todas as bases

  A frequência de todas as bases é a mesma

Parâmetros do modelo JC69:

Taxa de substituição: a= b= c= d= e= f
(uma única taxa é estimada junto com a árvore)

Frequência das bases: fixa = $0,25$

	A	C	G	T
A	–	a	b	c
C	a	–	d	e
G	b	d	–	f
T	c	e	f	–

Modelos de evolução de sequências

- Outros modelos
  - Felsenstein (1981) 'F81'

    A taxa de substituição é a mesma para todas as bases

    A frequência das bases pode variar

Parâmetros do modelo F81:

Taxa de substituição: a= b= c= d= e= f
(uma única taxa é estimada junto com a árvore)

Frequência das bases: estimada
(a frequência de cada base é estimada junto com a árvore)

	A	C	G	T
A	–	a	b	c
C	a	–	d	e
G	b	d	–	f
T	c	e	f	–

Modelos de evolução de sequências

- Outros modelos
  - Hasegawa–Kishino–Yano (1985) 'HKY85'

    Há duas taxas de substituição, uma para transições e outra para transversões

    A frequência de todas as bases pode variar

Parâmetros do modelo HKY85:

Taxa de substituição: transições diferem de transversões (b = e; a = c = d = f)

(duas taxas – uma para transições e outra para tranversões – são estimadas junto com a árvore)

Frequência das bases: estimada

(a frequência de cada base é estimada junto com a árvore)

	A	C	G	T
A	–	a	b	c
C	a	–	d	e
G	b	d	–	f
T	c	e	f	–

Modelos de evolução de sequências

- Modelos mais complexos (maior número de parâmetros estimados)
  - General Time Reversible (1986) 'GTR'

    A taxa de substituição difere para todas as bases

    A frequência de todas as bases pode variar

Parâmetros do modelo GTR:

Taxa de substituição: a != b != c != d != f
(as seis são taxas estimadas junto com a árvore)

Frequência das bases: estimada
(a frequência de cada base é estimada junto com a árvore)

	A	C	G	T
A	–	a	b	c
C	a	–	d	e
G	b	d	–	f
T	c	e	f	–

Modelos de evolução de sequências

- Modelos mais complexos (maior número de parâmetros estimados)
  - General Time Reversible (1986) 'GTR'

    A taxa de substituição difere para todas as bases

    A frequência de todas as bases pode variar

    Observação:
    - Todos os modelos que tem apenas as duas classes de parâmetros (taxa de substituição e frequência de bases) são casos específicos da família de modelos GTR (modelos aninhados → ao constringir um dos parâmetros, você chega a outro modelo)

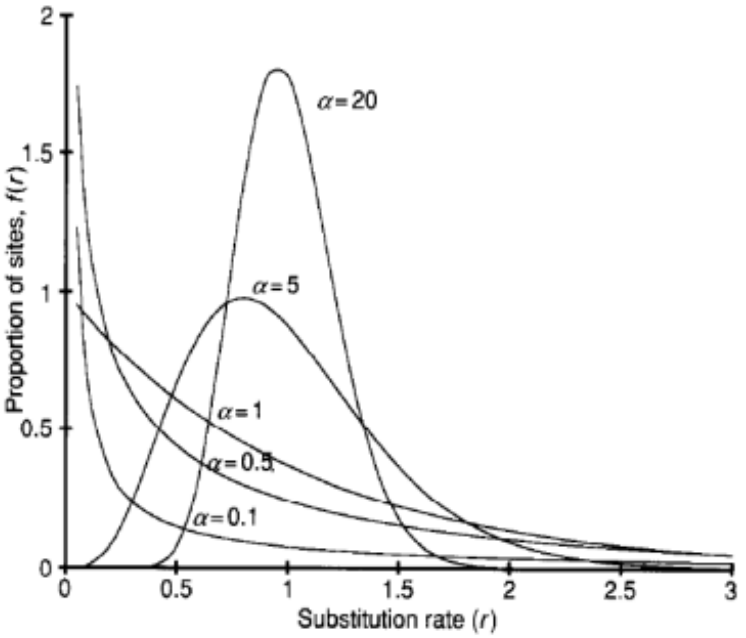
Modelos de evolução de sequências

Outros parâmetros que podem ser incluídos para adicionar mais 'realismo' ao modelo:

- Variação da taxa de substituição entre posições da sequência (parâmetro da forma da distribuição gama, *gamma shape parameter*):

Assume–se que a taxa de substituição varia ao longo da sequência, e segue uma distribuição gama, que tem um parâmetro que determina sua forma ('shape parameter', chamado de alfa)

- alfa < 1 = significativa variação da taxa evolutiva entre posições



Modelos de evolução de sequências

Outros parâmetros que podem ser incluídos para adicionar mais 'realismo' ao modelo:

- Variação da taxa de substituição entre posições da sequência
  (parâmetro da forma da distribuição gama, *gamma shape parameter*)
- Proporção de posições que são invariáveis
  (proportion of invariant sites)

  Assume-se que uma proporção das posições da sequência não variam (essa proporção pode ser fixa ou estimada)

Modelos de evolução de sequências

Modelos para sequências de proteínas

- Matriz de substituição 20 x 20, 189 parâmetros a serem estimados a partir dos dados
  → Necessita de grandes conjuntos de dados para estimativas acuradas
- Uso de modelos empíricos: taxas de substituição pré–calculadas a partir de grandes conjuntos de dados
- Diversas matrizes disponíveis (Dayhoff, JTT, WAG, LG, outras para grupos específicos e organelas)

Métodos baseados em caracteres II: máxima verossimilhança

- O que é máxima verossimilhança
- Inferência filogenética usando máxima verossimilhança
- Seleção de modelos de evolução de sequências

Incorporando modelos de evolução na análise

- Para que um modelo seja usado para estimar as distâncias genéticas entre sequências, é preciso que haja uma maneira de estimar os melhores valores para os parâmetros do modelo

Máxima verossimilhança
(*maximum likelihood*)

Verossimilhança

- Verossimilhança: Probabilidade de um determinado conjunto de dados ter sido gerado seguindo um determinado modelo

  $$\text{Verossimilhança}\,(\text{Modelo}) = P(\text{Dados} \mid \text{Modelo})$$

  Exemplo:
  Paciente com um conjunto de sintomas; cinco doenças prováveis. Para cada doença provável (modelo/hipótese), qual a chance de que aquele conjunto de sintomas (dados) tenha sido gerado por ela?

Máxima verossimilhança

- Método para estimar os parâmetros de um modelo

Máxima verossimilhança

- Dado um modelo e um conjunto de dados (que assumimos que foi gerado pelo processo descrito naquele modelo), quais os valores dos parâmetros que tornam mais provável a hipótese de que aqueles dados foram gerados por aquele modelo?

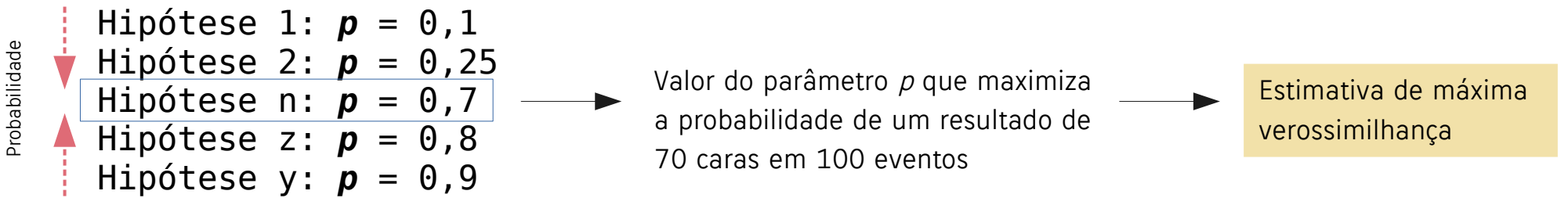
  Exemplo:
  Dados: Resultado de 100 lançamentos de moeda: 70 caras e 30 coroas
  Modelo: $P(\text{cara}) = p$; $P(\text{coroa}) = 1-p$
  Qual o valor mais provável do parâmetro p ? (= estimativa de máxima verossimilhança de p)

Máxima verossimilhança

- Dado um modelo e um conjunto de dados (que assumimos que foi gerado pelo processo descrito naquele modelo), quais os valores dos parâmetros que tornam mais provável a hipótese de que aqueles dados foram gerados por aquele modelo?

  Exemplo:
  Dados: Resultado de 100 lançamentos de moeda: 70 caras e 30 coroas
  Modelo: P(cara) = p ; P(coroa) = $1-p$
  Qual o valor mais provável do parâmetro p ? (= estimativa de máxima verossimilhança de p)

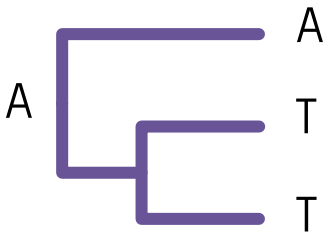
Lógica da estimativa por máxima verossimilhança:
Encontrar o valor de p que maximiza a probabilidade do resultado de 70 caras (x) em 100 lançamentos.



Inferência filogenética por máxima verossimilhança

Na inferência filogenética:

- Dados = Alinhamento de sequências
- Modelo = Um modelo de como uma sequência ancestral evoluiu para dar origem às sequências observadas no alinhamento
- Parâmetros do modelo:
  - A topologia da árvore e o tamanho dos ramos
  - Parâmetros relacionados ao modelo de evolução escolhido (frequência das bases, taxa de substituição entre bases, etc)

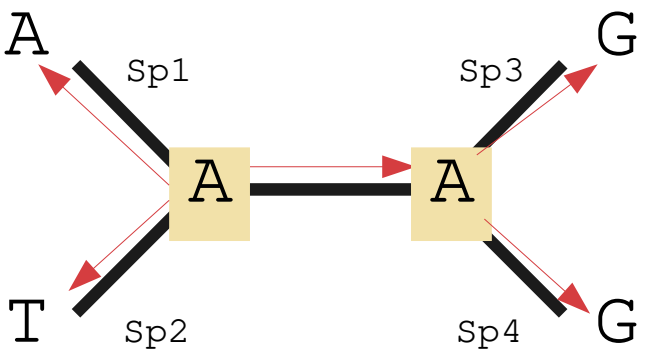


Máxima verossimilhança

Como funciona a análise por máxima verossimilhança?

- A função de verossimilhança de cada posição do alinhamento é calculada para todas as possíveis combinações de sequências ancestrais

	1	2	3	4	5
Sp1	A	A	C	G	A
Sp2	A	T	C	G	A
Sp3	A	G	C	T	A
Sp4	A	G	C	T	A



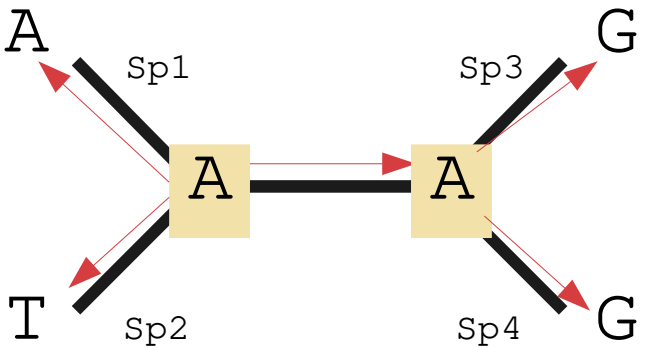
Pr = função das probs. de substituição entre cada nó da árvore e das frequências de cada nucleotídeo

Máxima verossimilhança

Como funciona a análise por máxima verossimilhança?

- A função de verossimilhança de cada posição do alinhamento é calculada para todas as possíveis combinações de sequências ancestrais
- Faz-se o mesmo para todas as posições
- Somam-se as probabilidades de todas as posições = verossimilhança total daquele conjunto de parâmetros

	1	2	3	4	5
Sp1	A	A	C	G	A
Sp2	A	T	C	G	A
Sp3	A	G	C	T	A
Sp4	A	G	C	T	A



Pr = função das probs. de substituição entre cada nó da árvore e das frequências de cada nucleotídeo

Máxima verossimilhança

Como funciona a análise por máxima verossimilhança?

- A função de verossimilhança de cada posição do alinhamento é calculada para todas as possíveis combinações de sequências ancestrais
- Faz-se o mesmo para todas as posições
- Somam-se as probabilidades de todas as posições = verossimilhança total daquele conjunto de parâmetros
- Mudam-se os parâmetros (topologia, tamanho dos ramos, parâmetros do modelo) e repete-se o procedimento
- Assim, cada conjunto de parâmetros avaliado recebe um valor de verossimilhança
- A árvore e o conjunto de parâmetros que produz a verossimilhança máxima é a árvore resultante da análise (estimativa de máxima verossimilhança)

Como saber se o modelo de evolução é o correto?

- 'Todos os modelos estão errados', porém uns são melhores que outros
- A inferência por máxima verossimilhança inicia–se com a escolha de um modelo de evolução
- A escolha do modelo é fundamental para que se obtenham as melhores estimativas, e portanto a melhor árvore.
- Dentre os modelos existentes, como saber qual o melhor?

JC69 x F81 x HKY x GTR ... ?

Como saber se o modelo de evolução é o correto?

- O uso da máxima verossimilhança permite comparar vários modelos
- Modelos com mais parâmetros são mais realistas… Mas nem sempre a inclusão de parâmetros adicionais melhora significativamente o resultado

  → Diferentes modelos produzem diferentes valores de máxima verossimilhança; modelos com mais parâmetros produzem valores maiores (= melhor ajuste)
- Métodos de seleção de modelos:
  - Teste da razão de verossimilhança (likelihood ratio test, LRT)
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)

Teste da razão de verossimilhança

- Modelos devem ser aninhados (um modelo é um caso especial do outro)

  Exemplo:

  Duas árvores são estimadas usando máxima verossimilhança, uma usando o modelo JC69 e outra o modelo F81

  Qual a melhor árvore?

  - Parâmetros JC69 = 1 taxa de substituição (u) (1 parâmetro)
  - Parâmetros F81 = 1 taxa de substituição (u), 3 frequências de bases (+ 3 parâmetros)

Teste da razão de verossimilhança

- Modelos devem ser aninhados (um modelo é um caso especial do outro)

  Exemplo:

  Duas árvores são estimadas usando máxima verossimilhança, uma usando o modelo JC69 e outra o modelo F81

  Qual a melhor árvore?

  - Parâmetros JC69 = 1 taxa de substituição (u) (1 parâmetro)
  - Parâmetros F81 = 1 taxa de substituição (u), 3 frequências de bases (+ 3 parâmetros)

Likelihood árvore JC69 = $-\ln(L1) = 1787.1$

Likelihood árvore F81 = $-\ln(L2) = 1784.8$

Likelihood ratio test (LRT) = $2 \times [\ \ln(L1) - \ln(L2)\ ]$

$= 2 \times (1787.1 - 1784.8) = 4.6$

Estatística LRT tem distribuição qui-quadrado com D graus de liberdade

D = número de parâmetros modelo 2 – número de parâmetros modelo 1

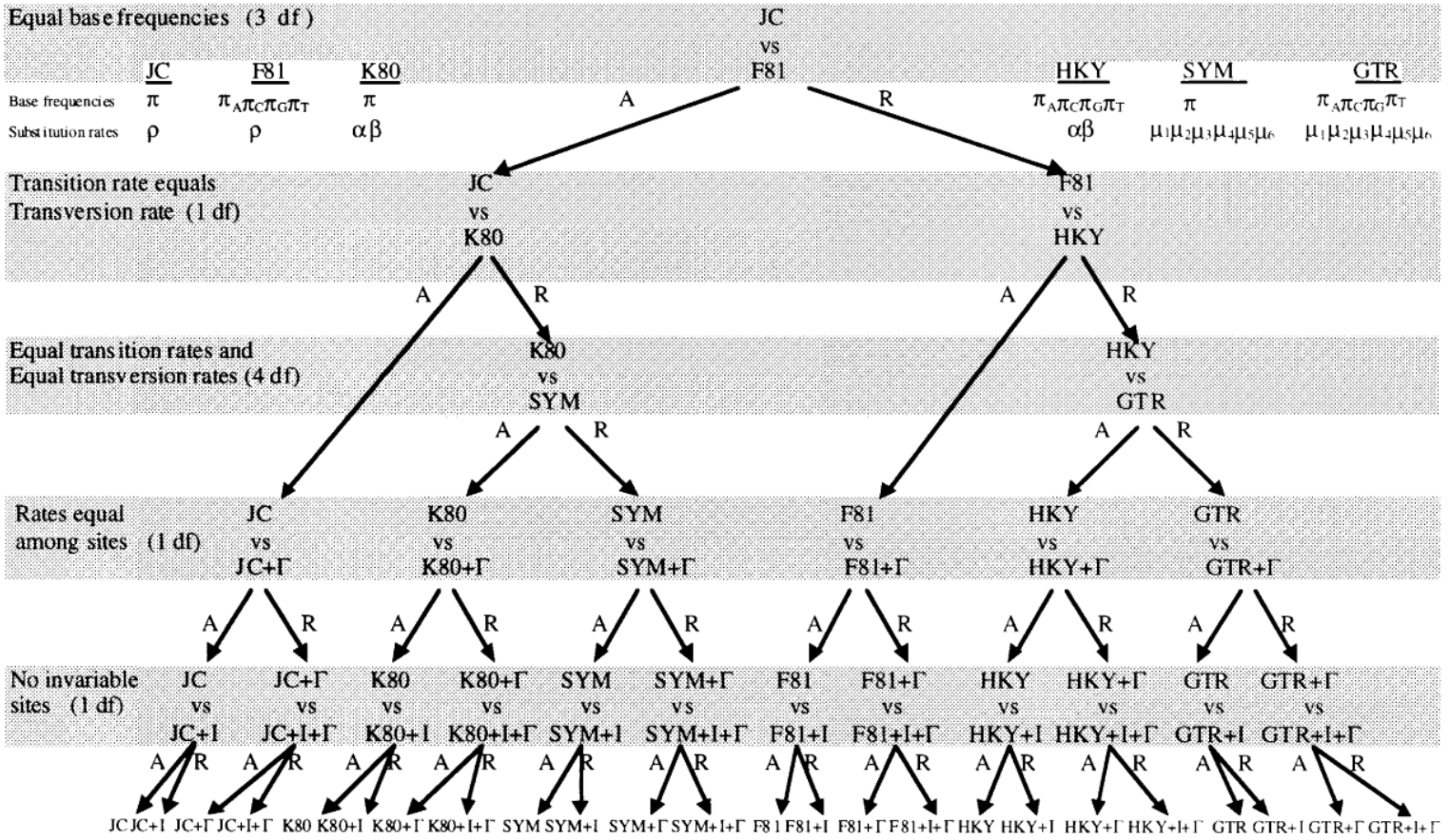
= 3

Valor crítico (P = 0.05) para 3 graus de liberdade (tabela qui-quadrado) = 7.81

Hipótese nula não é rejeitada (ou seja, nesse caso, F81 não produz uma árvore melhor que JC69)

Teste da razão de verossimilhança

- Modelos devem ser aninhados (um modelo é um caso especial do outro)



Posada and Crandall (1998)

Seleção de modelos na prática

- Montar scripts para automatizar a testagem de modelos

Seleção de modelos na prática

- Montar scripts para automatizar a testagem de modelos
- Usar programas que fazem a seleção de modelos: MODELTEST, jmodeltest

Seleção de modelos na prática

- Montar scripts para automatizar a testagem de modelos
- Usar programas que fazem a seleção de modelos: MODELTEST, jmodeltest
- Usar programas que estimam árvores usando vários modelos e escolhem automaticamente o melhor modelo → exemplo: IQ–TREE

Seleção de modelos na prática

- Montar scripts para automatizar a testagem de modelos
- Usar programas que fazem a seleção de modelos: MODELTEST, jmodeltest
- Usar programas que estimam árvores usando vários modelos e escolhem automaticamente o melhor modelo → exemplo: IQ–TREE
- Usar o modelo com maior número de parâmetros (GTR) + Gamma shape parameter

  Vantagens:
  - na prática, é na maioria das vezes o melhor modelo para conjuntos de dados médios/grandes
  - alguns programas para análise em larga escala só oferecem o modelo GTR

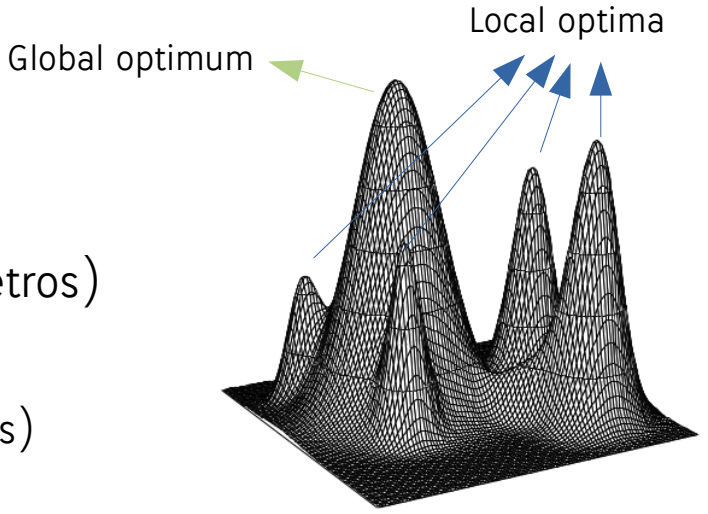
  Problemas:
  - é mais lento (mais parâmetros para serem estimados)
  - os dados podem não ser suficientes para os parâmetros do modelo serem estimados com acurácia

Resumo

- Máxima verossimilhança é o 'padrão ouro' na análise filogenética usando sequências de DNA e proteínas
- Permite a testagem de modelos evolutivos diferentes
- Com o avanço da computação e dos algoritmos de busca de árvores, atualmente a falta de poder computacional não é mais uma limitação (exceto para superárvores)

Críticas:

- Os modelos são simplificações do processo evolutivo e podem estar errados
- Há o risco da busca pela melhor árvore (e conjunto de parâmetros) atingir um ótimo local (local optima) ao invés do ótimo global (global optimum) (problema geral da busca por métodos heurísticos)

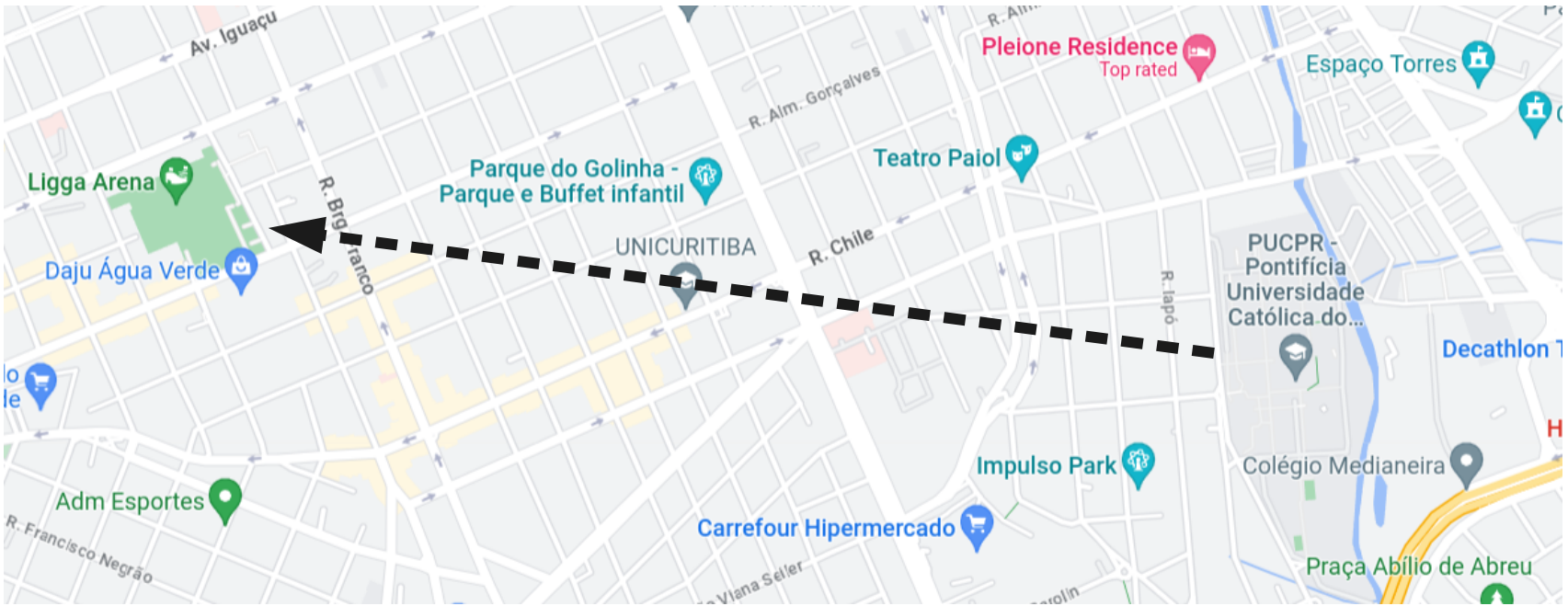


Resumo

Analogia:

Qual o caminho mais rápido da PUC até a arena da baixada?

Parcimônia: a menor distância entre dois pontos é uma linha reta

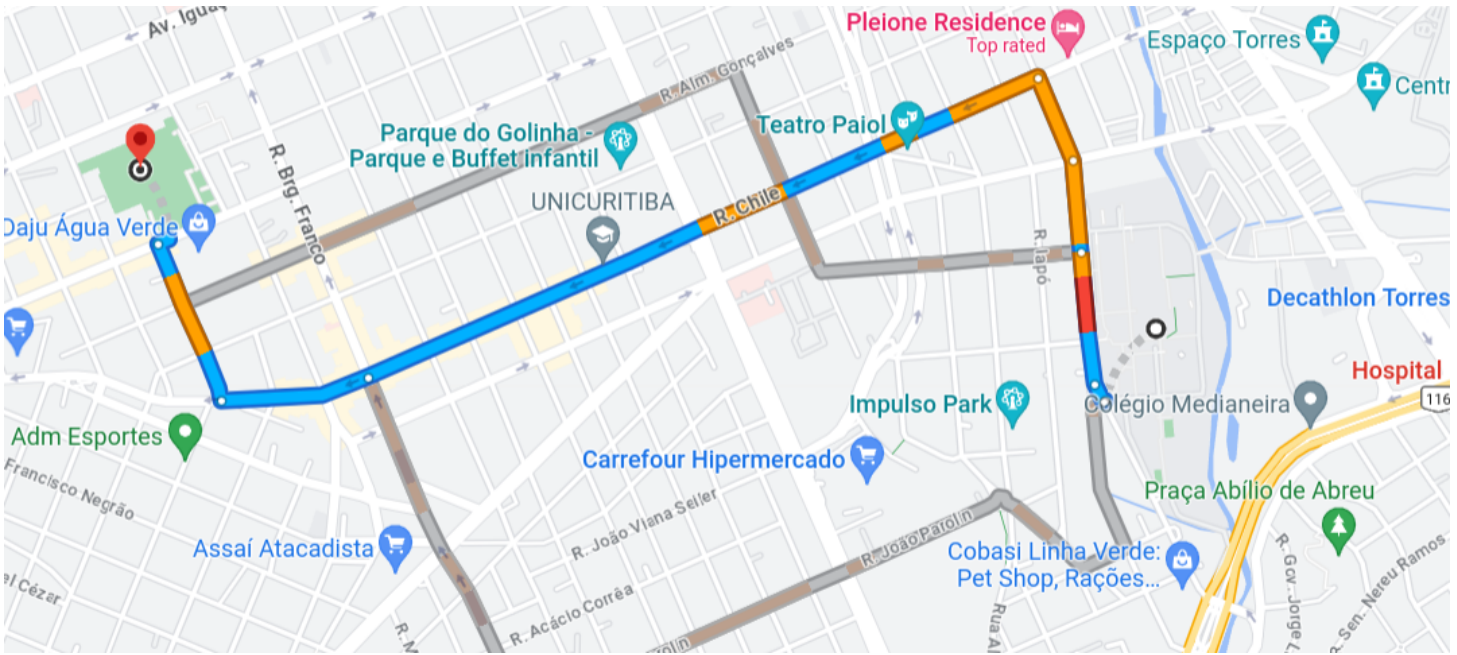


Resumo

Analogia:

Qual o caminho mais rápido da PUC até a arena da baixada?

Métodos baseados em modelos: existem ruas, semáforos, em alguns horários o fluxo em algumas ruas é maior, ... (n parâmetros) → mais realista



Medidas de suporte estatístico para as árvores inferidas

- O quanto podemos confiar nas árvores inferidas?
- Bootstrapping na análise filogenética

O quanto podemos confiar nas árvores inferidas?

Principais fontes de erro/incerteza na inferência filogenética:

1. Dados ruins
(e.g. uso de sequências não homólogas, falha no alinhamento)

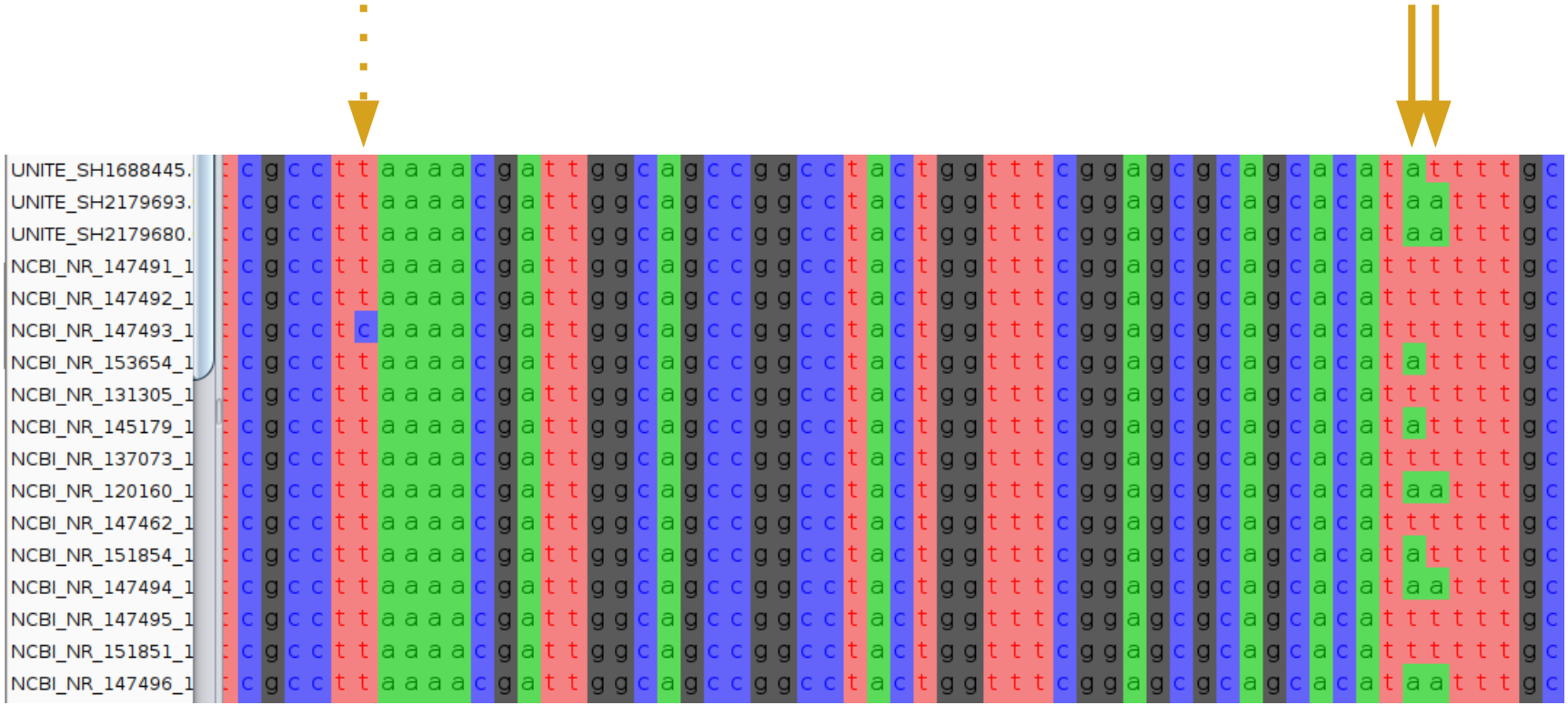
2. Erros sistemáticos devido ao método utilizado
(e.g. atração de ramos longos usando parcimônia, distâncias não corrigidas usando neighbour joining)

3. Modelos de substituição sub–ótimos
(falha na seleção de modelos usando máxima verossimilhança)

4. Erros aleatórios
(dados insuficientes → erro de amostragem)

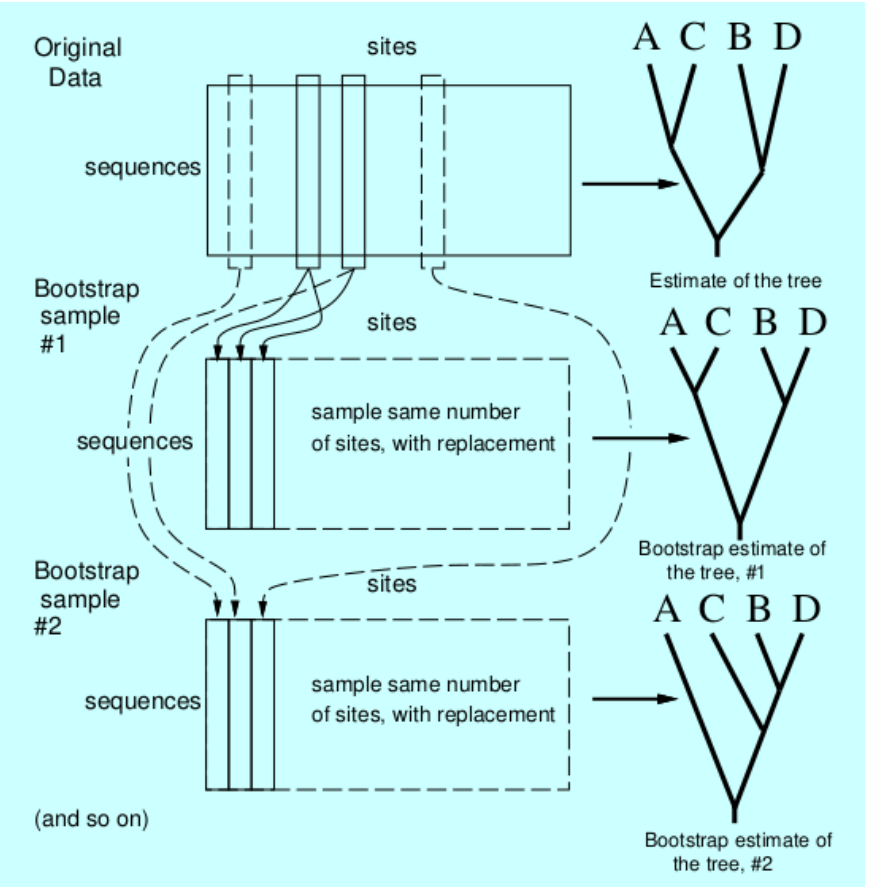
Meus dados são suficientes?

Alinhamentos muito curtos (e/ou com poucas posições variáveis)



Medida de suporte dos ramos: Bootstrapping

- Técnica de reamostragem
- Objetivo: Avaliar o suporte de cada agrupamento da árvore
- Procedimento:
  - Crie um novo alinhamento, do tamanho do original, a partir da amostragem das colunas do alinhamento original (= pseudoreplicatas)
  - Refaça a análise filogenética
  - Repita o procedimento X vezes

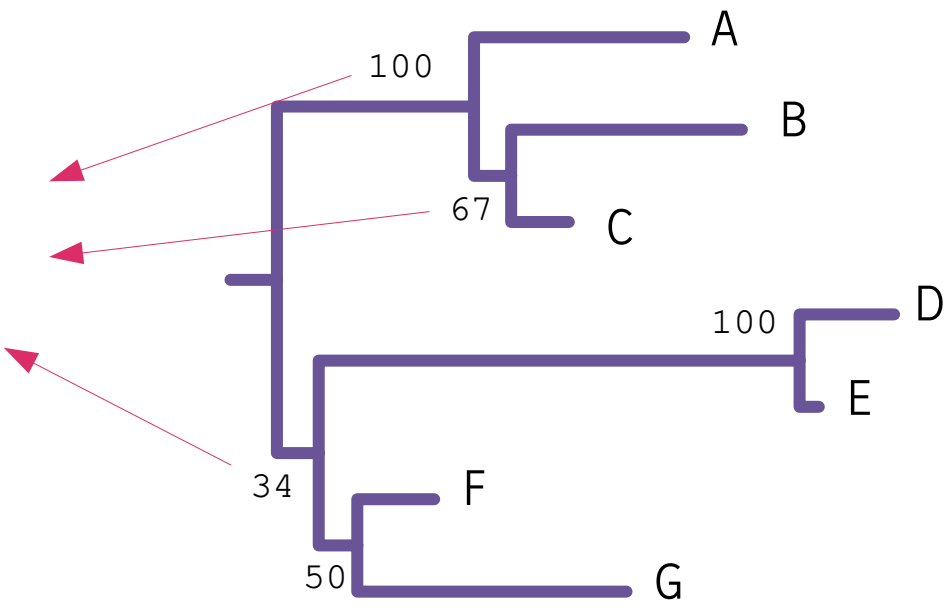


Retirado de Joe Felsenstein (2016)

Medida de suporte dos ramos: Bootstrapping

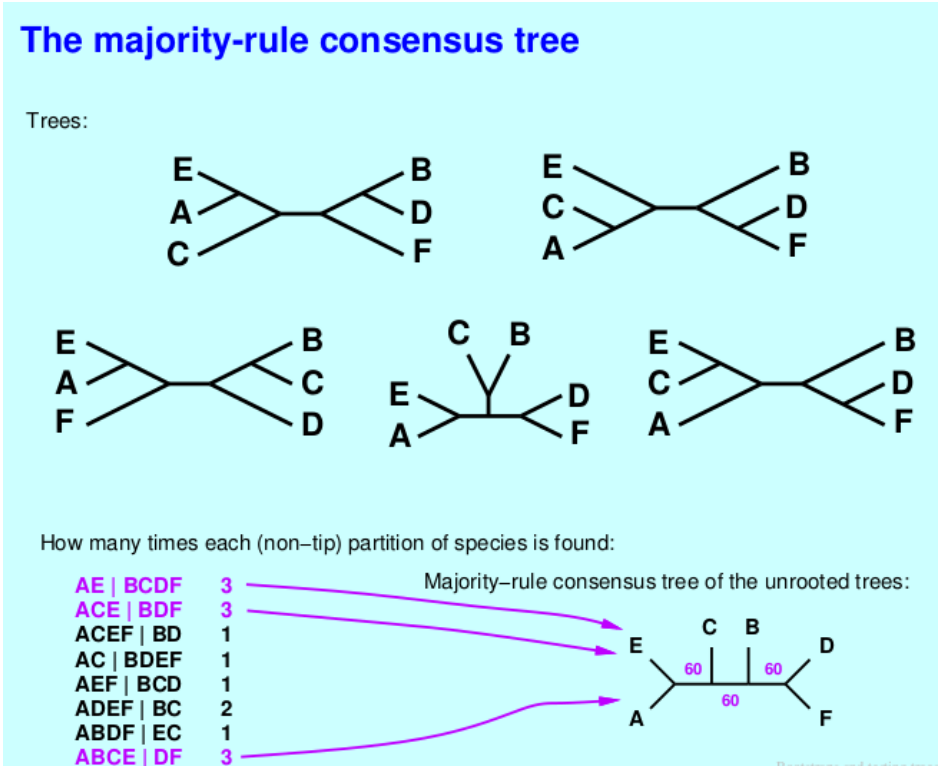
- Após x pseudoreplicatas (geralmente 100), quão frequentemente cada ramo da árvore original aparece nas árvores produzidas por bootstrap? = Valor de bootstrap

Significado: Este agrupamento apareceu X vezes nas análises das pseudoreplicatas do alinhamentos



Medida de suporte dos ramos: Bootstrapping

- Árvore consenso de maioria = montada a partir do conjunto de partições (agrupamentos) mais comuns encontrados nas árvores de boostrap



Retirado de Joe Felsenstein (2016)

Suporte bootstrap

- Não é uma medida de probabilidade da topologia estar correta

Suporte bootstrap

- Não é uma medida de probabilidade da topologia estar correta
- Pode ser usado para qualquer tipo de método (exceto Inferência Bayesiana, que tem suas próprias medidas de suporte da topologia)

Suporte bootstrap

- Não é uma medida de probabilidade da topologia estar correta
- Pode ser usado para qualquer tipo de método (exceto Inferência Bayesiana, que tem suas próprias medidas de suporte da topologia)
- É útil para indicar se a quantidade de dados usada na análise é suficiente ou não para suportar cada agrupamento mostrado na árvore

Suporte bootstrap

- Não é uma medida de probabilidade da topologia estar correta
- Pode ser usado para qualquer tipo de método (exceto Inferência Bayesiana, que tem suas próprias medidas de suporte da topologia)
- É útil para indicar se a quantidade de dados usada na análise é suficiente ou não para suportar cada agrupamento mostrado na árvore
- O usuário escolhe se quer indicar os valores de bootstrap (i.e. quantidade de vezes que cada agrupamento aparece nas pseudoreplicatas) na árvore gerada, ou se quer gerar uma nova árvore consenso de maioria a partir dos agrupamentos mais frequentemente observados nas pseudoreplicatas.

Literatura sugerida

Livros:

- M Nei & S Kumar. 2000. Molecular Evolution and Phylogenetics.

Artigos

- Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39: 783–791.
- Kapli P, Yang Z, Telford MJ. 2020. Phylogenetic tree building in the genomic age. Nature Reviews Genetics 21: 428–444.
- Kelchner SA, Thomas MA. 2007. Model use in phylogenetics: nine key questions. Trends in Ecology and Evolution 22: 87–94.