Tópicos em Filogenômica e Evolução Molecular

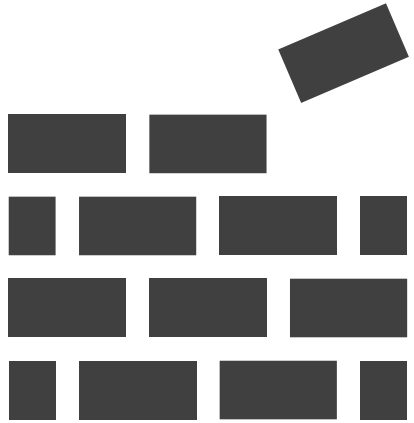
Módulo 2 – Métodos de Inferência em Escala Genômica

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

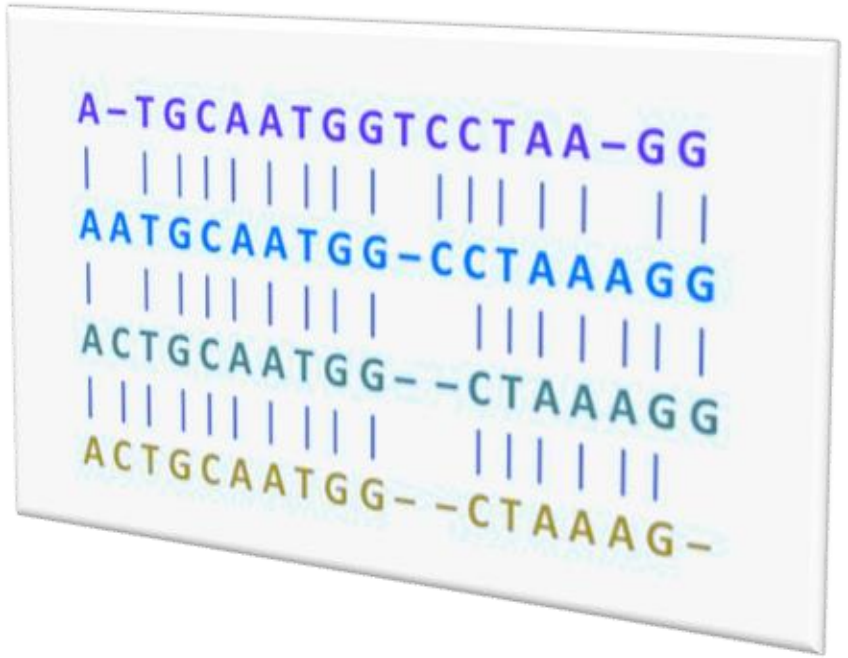
Estrutura da aula

- **Panorama da aula**
  - Como passamos de sequências alinhadas a hipóteses filogenéticas
  - Onde surgem erros e incertezas ao longo do pipeline
- **Bloco 1 — Alinhamento em larga escala**
  - Alinhamento múltiplo de milhares de sequências
  - Avaliação da qualidade do alinhamento
  - Impacto do alinhamento na inferência filogenética
- **Bloco 2 — Suporte e inferência filogenética**
  - Princípios de avaliação de suporte filogenético
  - Métodos de inferência em escala genômica
  - Parcimônia, verossimilhança e abordagem bayesiana
- **Bloco 3 — Inferência moderna em filogenômica**
  - Métodos baseados em distância e buscas heurísticas
  - Coalescência multiespécies e inferência multi-locus
  - Softwares e pipelines filogenômicos



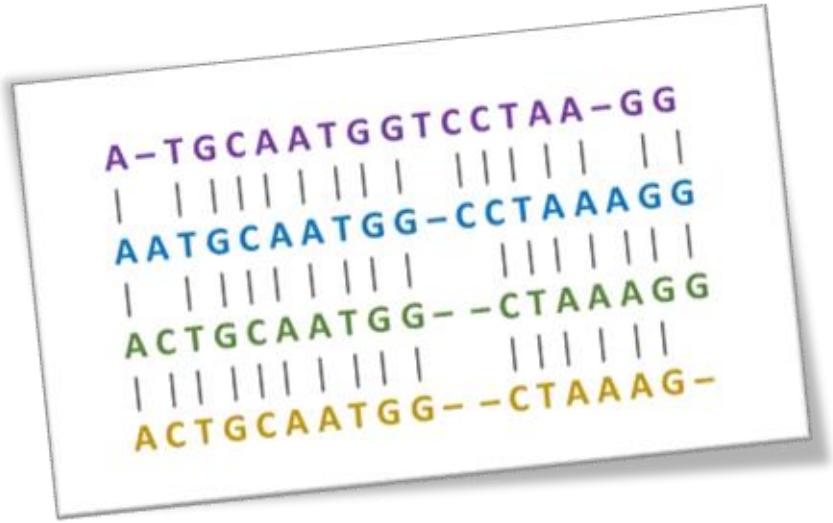
Alinhamento de sequências biológicas

O alinhamento de sequências é o procedimento analítico que estabelece correspondências posicionais entre sequências de nucleotídeos ou proteínas, permitindo a comparação sistemática de seus sítios homólogos.



Alinhamento não é?

- **Não é uma etapa neutra**
  - Envolve decisões analíticas explícitas
  - Introduz pressupostos evolutivos implícitos
- **Não é apenas uma etapa técnica**
  - Define hipóteses de homologia posicional
  - Afeta diretamente a topologia filogenética
- **Não é independente da inferência**
  - Erros de alinhamento propagam-se para a árvore
  - Inferência estatística não corrige alinhamentos ruins
- **Não garante homologia verdadeira**
  - Similaridade ≠ ancestralidade comum
  - Convergência e saturação podem enganar
- **Não escala sem consequências**
  - Milhares de sequências exigem heurísticas
  - Heurísticas implicam compromissos entre precisão e viabilidade

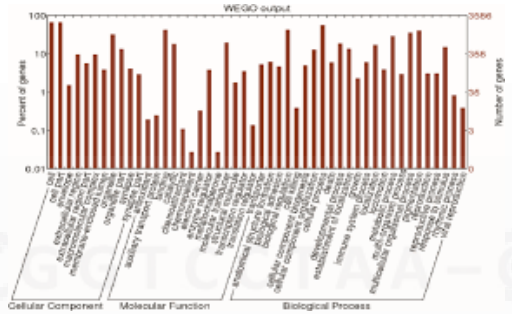


Por que alinhar sequências?

- **Para viabilizar análises genômicas comparativas**
  - Comparação entre genes, genomas e espécies
  - Identificação de regiões conservadas e divergentes
- **Para apoiar montagem e anotação genômica**
  - Validação de *contigs* e *scaffolds*
  - Identificação de ortólogos e parálogos
  - Transferência de anotação funcional
- **Para definir caracteres em filogenia**
  - Cada coluna do alinhamento representa um caráter
  - Base direta para inferência filogenética
- **Para reconstruir filogenias**
  - Permite inferir relações evolutivas entre táxons
  - Fundamenta análises concatenadas e multi-locus
- **Para integrar dados em filogenômica**
  - Comparação de centenas a milhares de loci
  - Avaliação de concordância e conflito entre genes

Conceitos importantes

- **Identidade** = quantitativo
- **Similaridade** = quantitativo
- **Homologia** = qualitativo



Material sobre ortólogos

Como avaliar um alinhamento

Alinhamentos são avaliados por funções de pontuação

Cada alinhamento recebe um valor numérico
A pontuação reflete:

- correspondências
- discrepâncias
- Lacunas

O "melhor" alinhamento é aquele com **maior_pontuação**, dado um conjunto de regras.

Critérios de pontuação

Lacunas (gaps): penalidade por cada lacuna

Correspondêcias (Match): pontuação por correspondência

Discripâncias (mismatches): penalidade por cada discrepância

Esquema de pontuação (exemplo)

Parâmetros utilizados
Correspondência (*match*): **+2**
Discrepância (*mismatch*): **−4**
Lacuna (*gap*): **−1**

Observação
Esses valores são arbitrários
Diferentes parâmetros → diferentes alinhamentos ótimos

Alinhamento avaliado

```
Sequência α′:  CAGGATGCTAGCAAAAACCATCGCGGGCGATAA
Sequência β′:  ——GGCATGTAGCACACACGACGCTGGGAGAAT—
```

O alinhamento converte sequências biológicas em uma matriz de caracteres homólogos, na qual cada coluna representa um caráter evolutivo comparável entre os táxons.

Cálculo de pontuação

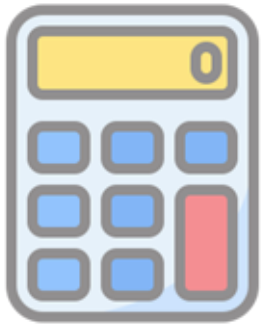
Componentes
Lacunas:
3 × (−1) = **−3 -**
Correspondências: |
16 × (+2) = **+32**
Discrepâncias: *
14 × (−4) = **−56**

```
Sequência α′:  CAGGATGCTAGCAAAAACCATCGCGGGCGATAA
Sequência β′:  ——GGCATGTAGCACACACGACGCTGGGAGAAT—
```

Pontuação final do alinhamento

$$-3 + 32 - 56 = \mathbf{-27}$$

Interpretação
Um score negativo não significa um alinhamento “errado”, apenas que, sob esses parâmetros, há mais penalidades do que recompensas.



Por que isso importa?

Algoritmos escolhem alinhamentos com base na pontuação
Parâmetros influenciam diretamente:

- alinhamento final
- regiões consideradas homólogas
- inferência filogenética

Tipos de alinhamentos

Tipo de Alinhamento	Descrição	Principais Usos	Programas / Algoritmos
Alinhamento Global	Alinha duas sequências completas, maximizando a pontuação ao longo de todo o comprimento do alinhamento. Indicado quando as sequências são semelhantes e de tamanho comparável.	Comparação de genes ou proteínas homólogas; análise de sequências completas.	Needleman–Wunsch (algoritmo exato); *EMBOSS Needle*
Alinhamento Semi-Global	Similar ao alinhamento global, mas não penaliza lacunas nas extremidades, permitindo alinhar prefixos e sufixos de sequências de comprimentos diferentes.	Montagem de genomas ou fragmentos; alinhamento de sequências parcialmente sobrepostas.	Variações do Needleman–Wunsch; *EMBOSS Stretcher*
Alinhamento Local	Identifica a melhor correspondência entre regiões (subsequências) de duas sequências, ignorando regiões externas ao alinhamento ótimo. Ideal para detectar similaridade local.	Identificação de domínios conservados; busca de homólogos distantes em grandes bases de dados.	Smith–Waterman (algoritmo exato); BLAST (heurístico)

Abordagens Algorítmicas em Alinhamentos

Abordagem	Princípio	Tipo de Alinhamento	Exemplos de Algoritmos / Programas	Vantagens	Limitações	Principais Aplicações
Programação Dinâmica	Resolve o alinhamento de forma ótima por decomposição recursiva, maximizando uma função de pontuação definida explicitamente.	Par-a-par (global ou local)	Needleman–Wunsch (global); Smith–Waterman (local)	Garante alinhamento ótimo; controle explícito de penalidades	Alto custo computacional; não escala para muitas sequências	Comparação precisa entre duas sequências; validação de alinhamentos
Métodos Heurísticos	Utilizam aproximações para reduzir o espaço de busca e acelerar o alinhamento.	Par-a-par e múltiplo	CLUSTALW; MUSCLE; MAFFT	Rápidos; escaláveis; viáveis para centenas ou milhares de sequências	Não garantem ótimo global do problema; dependem de heurísticas	Filogenômica; genômica comparativa; alinhamento em larga escala

Necessidade de Heurísticas

Crescimento combinatorial: O número de árvores possíveis cresce exponencialmente com o número de espécies, tornando métodos exatos impraticáveis para conjuntos de dados genômicos.

Limites computacionais: Análises exatas requerem tempo e memória que excedem capacidades atuais de computação para grandes conjuntos de dados.

Heurísticas eficientes: Métodos como estratégias progressivas e iterativas permitem explorar o espaço de árvores de maneira prática, embora não garantam otimalidade global.

Heurísticas representam um acordo entre precisão e viabilidade computacional

"Em filogenômica, heurísticas não são alternativas à precisão, mas condições necessárias para obter resultados em tempo razoável."

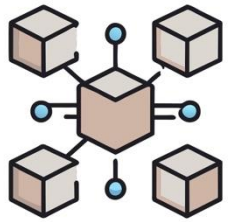
Algoritmos Heurísticos Modernos

Estratégias Progressivas

- **Abordagem top-down:** Constrói alinhamento múltiplo através de alinhamentos progressivos de pares de sequências.
- **Árvore filogenética:** Utiliza uma árvore inicial para orientar o processo de alinhamento.
- **Divide e conquista:** Quebra o problema em subproblemas menores e resolvê-los recursivamente.
- **Vantagem:** Complexidade computacional gerenciável, adequada para grandes conjuntos de dados.

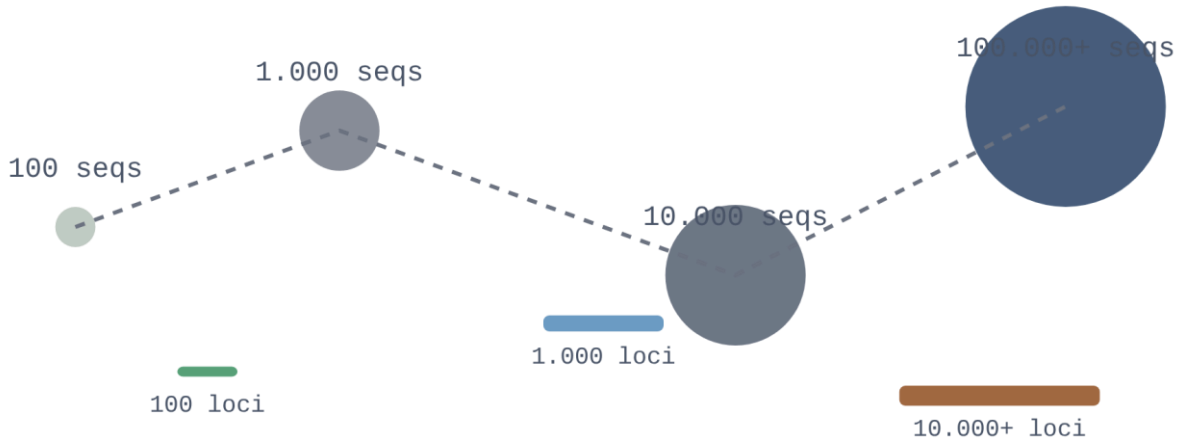
Estratégias Iterativas

- **Refinamento contínuo:** Parte de um alinhamento inicial e aplica melhorias iterativas.
- **Busca local:** Explora vizinhanças para encontrar alinhamentos de melhor qualidade.
- **Atualizações alternadas:** Alterna entre árvore e alinhamento até convergência.
- **Vantagem:** Potencial para encontrar alinhamentos de maior qualidade que algoritmos gulosos.



Escala do Alinhamento Moderno

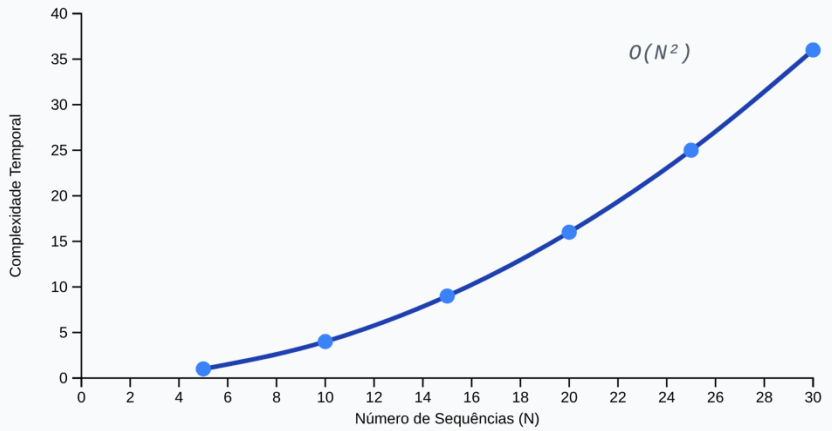
- **Número de sequências (táxons)**
- **Número de loci (marcadores)**
- **Duas dimensões da escala**
- **Consequência analítica**



A escala do alinhamento moderno representa um dos principais desafios e oportunidades da filogenômica contemporânea.

Limitações do Alinhamento Ótimo

- Custo computacional alto
- Escala limitada
- Crescimento rápido do problema
- Inviabilidade



à medida que o número de sequências aumenta, o custo computacional do alinhamento ótimo cresce rapidamente

"A impraticabilidade do alinhamento ótimo leva à necessidade de abordagens heurísticas na filogenômica moderna."

Desafios do Alinhamento em Larga Escala

Principais obstáculos para alinhamento correto em filogenômica

Ambiguidade Posicional

- Regiões onde a homologia posicional é incerta
- Desafios para algoritmos de alinhamento
- Requer avaliação criteriosa

Saturação

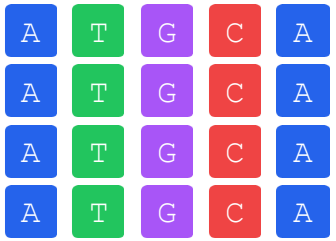
- Múltiplas substituições ocultas
- Perda de sinal evolutivo
- Distorção de distâncias evolutivas

Regiões Hipervariáveis

- Alta taxa de variação
- Risco elevado de erros de alinhamento
- Desafios para métodos de inferência

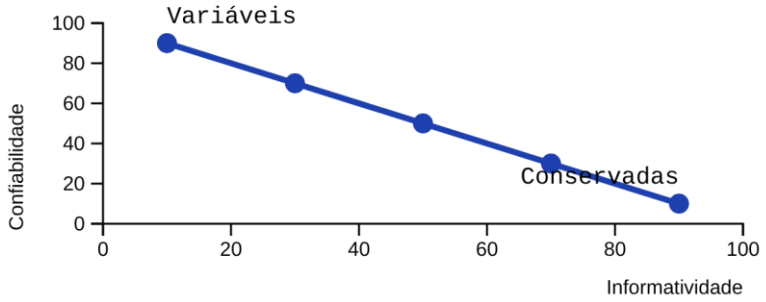
Regiões Conservadas

Representação de Região Conservada



Alinhamento de sequências conservadas

Relação Confiabilidade vs. Informatividade



- **Alta confiabilidade:** Sequências conservadas têm maior probabilidade de representar corretamente a homologia posicional devido à pressão seletiva.
- **Baixa informatividade:** Pouca variação entre sequências, resultando em pouco sinal filogenético para distinguir relações evolutivas.
- **Características:** Altamente conservadas, menores taxas de substituição, pressão seletiva constante.
- **Implicações:** Múltiplas substituições saturadas podem mascarar relações evolutivas, mesmo com alinhamento correto.

"Regiões conservadas fornecem alinhamentos confiáveis, mas não necessariamente árvores filogenéticas informativas."

Regiões Variáveis

Representação de Regiões Variáveis

Exemplo de sequências variáveis:

```
ATCGATCGATCGATCG
```

```
ATCGGTCGATCGCGAT
```

- Variações nucleotídicas
- Substituições sinônimas e não sinônimas
- Inserções e deleções (indels)

--- **Alto potencial filogenético:** Regiões variáveis acumulam mudanças que refletem eventos evolutivos recentes, proporcionando bons sinais para discriminar linhagens próximas.

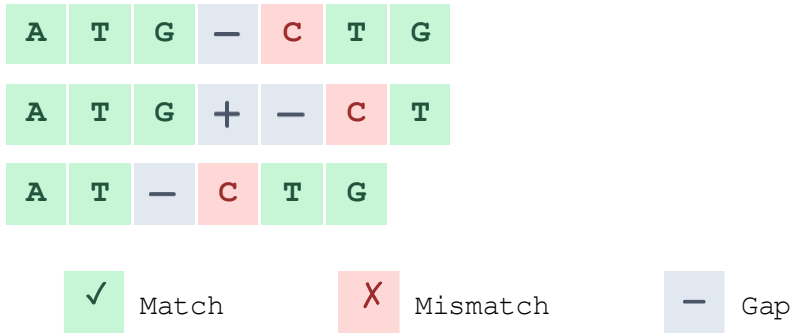
-- **Maior risco de erro analítico:** Variações excessivas podem levar a erros de alinhamento, interpretação distorcida da evolução e suporte filogenético enganoso.

····· **Desafios analíticos:** Risco de saturação de substituições, convergência evolutiva e distorção de distâncias filogenéticas em regiões hipervariáveis.

-· **Abordagens recomendadas:** Análise criteriosa com modelos evolutivos apropriados, trimming seletivo e validação cruzada dos resultados.

Gaps e Indels

Representação de Indels

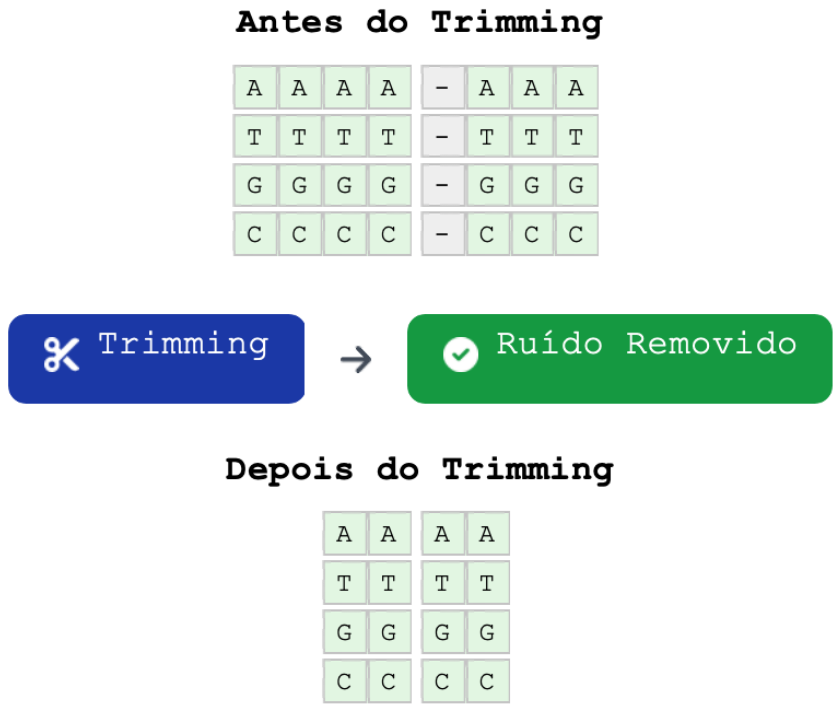


- **Gaps e indels:** Representam ganho ou perda de nucleotídeos ou aminoácidos durante a evolução, não simplesmente erros de alinhamento.
- **Interpretação evolutiva:** Indels precisam ser analisados no contexto de sua origem e manutenção evolutiva.
- **Desafios filogenéticos:** Indels podem causar distorções topológicas se mal interpretados ou tratados de forma inadequada.
- **Critérios de avaliação:** Qualidade do alinhamento deve ser avaliada não apenas por pontuações, mas por consistência evolutiva.

Indels refletem eventos históricos de inserção ou deleção, cuja interpretação depende do alinhamento e do modelo evolutivo adotado.

Trimming de Alinhamentos

- **Objetivo:** reduzir ruído sem perder sinal filogenético
- Trimming afeta diretamente a inferência filogenética
- Não existe estratégia universalmente ótima



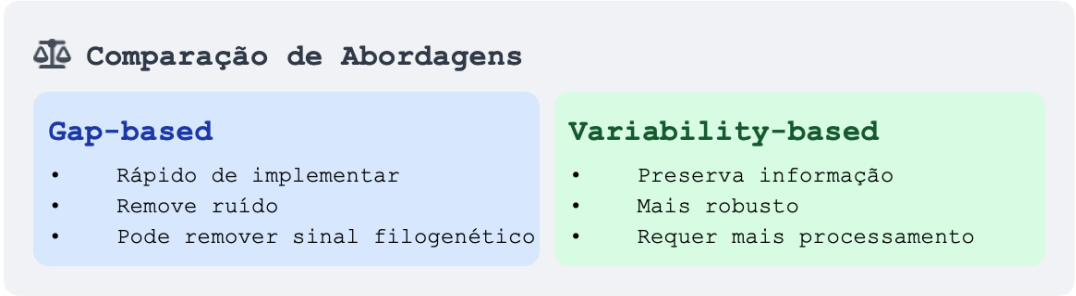
Estratégias de Trimming

Abordagem baseada em gaps

- Remove colunas com muitos gaps
- Rápida e simples
- Pode remover sinal filogenético

Abordagem baseada em variabilidade

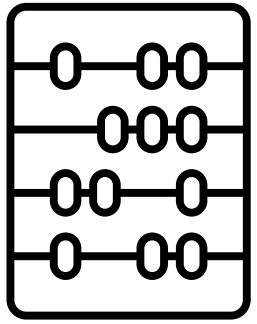
- Prioriza colunas informativas
- Mais robusta
- Maior custo computacional



Estratégias de Trimming

Trade-off Fundamental do Trimming

- Menos dados ≠ melhor inferência
- Mais dados ≠ mais sinal filogenético
- O equilíbrio depende do tipo de dado e da pergunta



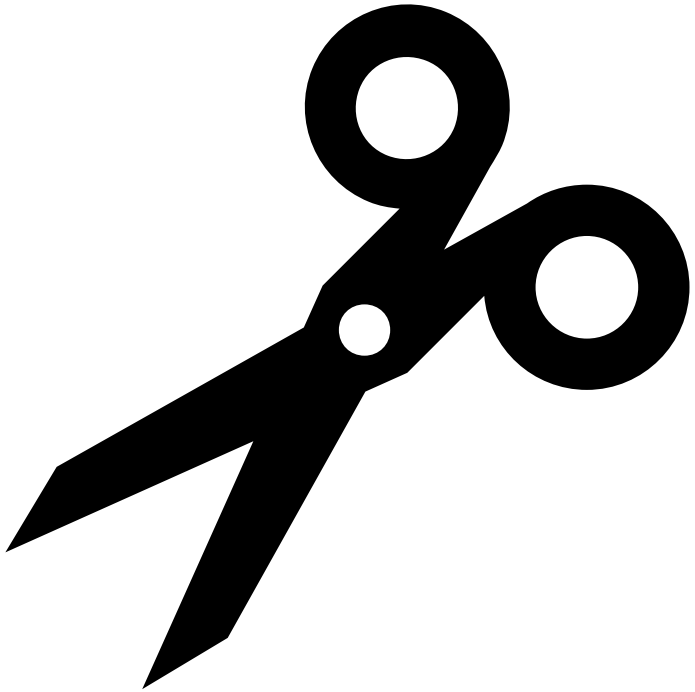
Trimming como Decisão Analítica

- Trimming não é etapa neutra
- Diferentes critérios produzem alinhamentos diferentes
- Impacta topologia, suporte e comprimento de ramos

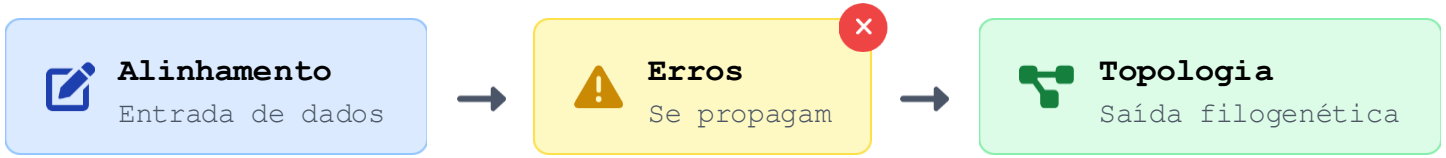
Estratégias de Trimming

- **Quando Ser Conservador?**
- Dados com poucos táxons
- Loci altamente conservados
- Perguntas filogenéticas profundas
-

- **Quando Ser Mais Agressivo?**
- Muitos táxons
- Regiões altamente divergentes
- Dados com alinhamento incerto



Relação Alinhamento × Topologia



Propagação de Erros

Erros de alinhamento não se limitam à etapa de preprocessamento, mas se propagam através do processo de inferência, afetando diretamente a topologia resultante.

Impacto Cumulativo

A influência dos erros de alinhamento se acumula durante a inferência filogenética, podendo levar a conclusões erradas sobre relações evolutivas.

Consequências Topológicas

Erros seletivos em regiões hipervariáveis ou mal alinhadas podem distorcer as relações entre taxons, alterando significativamente a topologia inferida.

Compromisso Qualidade

A qualidade do alinhamento condiciona a qualidade da árvore filogenética resultante, exigindo cuidados analíticos rigorosos.

"A qualidade do alinhamento condiciona a qualidade da árvore filogenética resultante."

Refinamento: Automático vs. Manual

Aspecto	Refinamento Automático	Refinamento Manual
	Ajuste do alinhamento realizado por algoritmos computacionais, com base em critérios estatísticos e heurísticos.	Ajuste realizado pelo pesquisador a partir da inspeção visual e conhecimento biológico.
Objetivo principal	Melhorar a consistência e a pontuação do alinhamento de forma reprodutível e escalável.	Corrigir erros evidentes não resolvidos automaticamente.
Principais ferramentas	MAFFT (L-INS-i, FFT-NS-2); Clustal Omega; T-Coffee	Jalview, AliView
Abordagem	Heurística e automatizada; pode ser iterativa e baseada em múltiplas estratégias de alinhamento.	Interativa; baseada em inspeção visual e julgamento do pesquisador.
Vantagens	Rápido; reprodutível; viável para grandes conjuntos de dados; adequado para filogenômica.	Permite correção fina de regiões problemáticas; incorpora conhecimento biológico específico.
Limitações	Pode manter erros sistemáticos; depende fortemente dos parâmetros escolhidos.	Subjetivo; pouco reprodutível; inviável para grandes matrizes filogenômicas.
Quando usar	Etapa principal do pipeline; alinhamentos iniciais e refinamentos globais.	Casos pontuais; poucos loci; análises exploratórias ou validação visual.
Uso em filogenômica	Essencial e dominante.	Uso restrito e cauteloso.

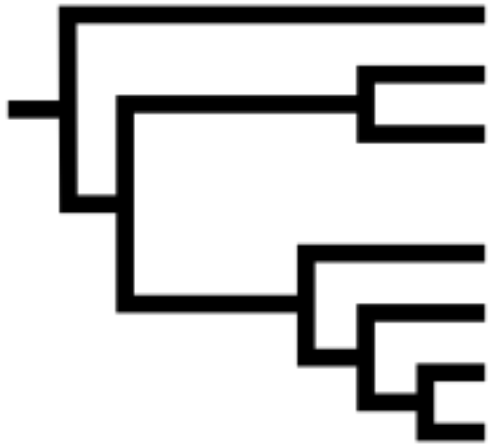
Perda de Sinal Filogenético

Alinhamento conservador: Remove variações sutis que podem ser informativas para a filogenia.

Sinal filogenético: Informação que reflete a história evolutiva verdadeira entre os taxons.

Compromisso analítico: Precisamos balancear ruído (que deve ser removido) com sinal (que deve ser preservado).

Consequências: Perda de sinal resulta em árvores filogenéticas com menor suporte e maior risco de erro.

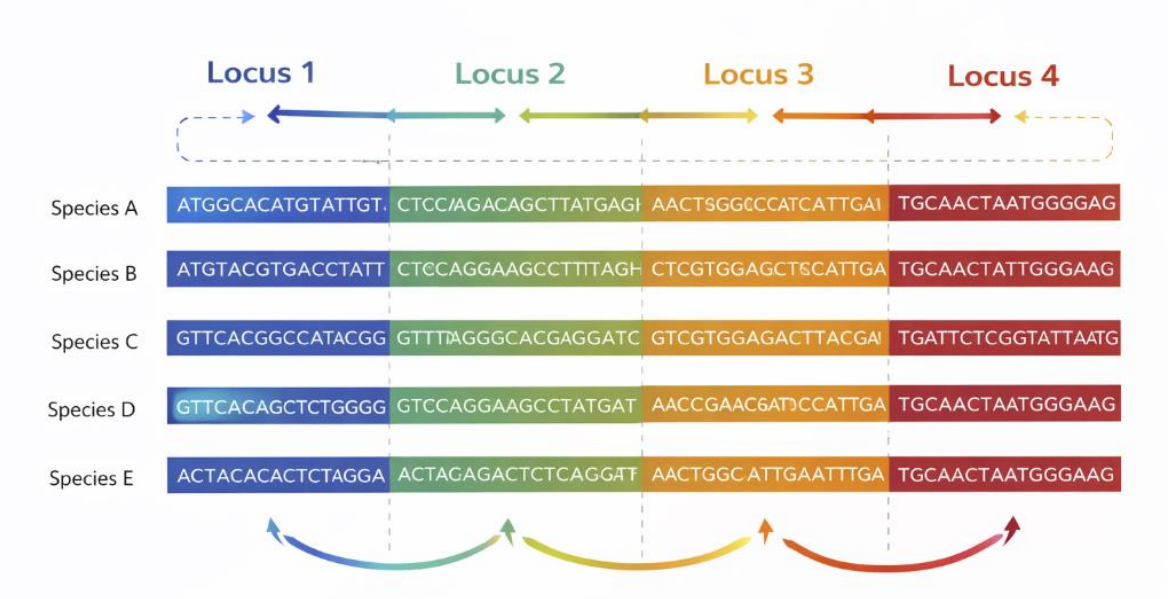


Cuidado com alinhamentos excessivamente conservadores!

"O alinhamento deve ser suficientemente flexível para preservar informação evolutiva relevante."

Alinhamento e Escala Genômica

Heterogeneidade entre loci



- Dados modernos envolvem muitos táxons e muitos loci
- Diferentes loci apresentam qualidades e histórias evolutivas distintas
- A heterogeneidade entre loci aumenta a complexidade da inferência

Alinhamento como Parte da Inferência

Não é uma etapa preliminar neutra

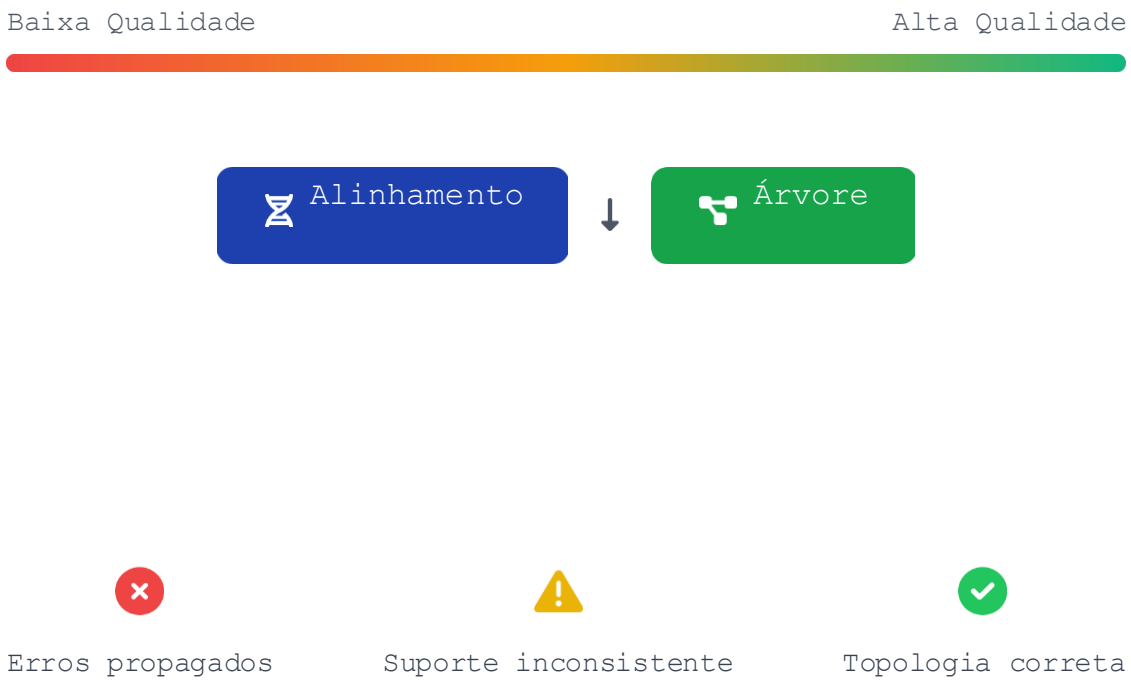
- **Decisões críticas:** Escolhas de alinhamento representam decisões que impactam resultados finais.
- **Interpretação evolutiva:** Alinhamento como hipótese sobre homologia posicional.

Componente integral da análise

- **Integração contínua:** Alinhamento e inferência ocorrem em paralelo durante a análise.
- **Ciclo iterativo:** Resultados de inferência informam melhorias no alinhamento.

Síntese do Bloco

- **Alinhamento como hipótese:** Representa decisões analíticas baseadas em critérios evolutivos, não etapas automáticas.
- **Homologia posicional:** Escolha de padrões de similaridade que representam ancestralidade.
- **Métricas de qualidade:** Avaliação visual e automática para garantir confiabilidade dos dados.
- **Relação direta:** Qualidade do alinhamento condiciona diretamente a qualidade da árvore filogenética.



Programa	Abordagem	Critério principal	Vantagens	Limitações	Uso típico em filogenômica
Gblocks	Baseado em blocos conservados	Remove regiões com muitos gaps e alta variabilidade	Conservador; fácil de interpretar	Pode remover sinal filogenético útil	Filogenias profundas; dados ruidosos
TrimAl	Estatística e heurística	Gaps, entropia e variabilidade	Flexível; vários modos automáticos	Parâmetros podem ser pouco intuitivos	Filogenômica multi-locus
ClipKit	Baseado em informatividade	Retém sítios informativos (PI sites)	Mantém sinal filogenético; moderno	Menos conservador que Gblocks	UCEs, AHE, exons
BMGE	Baseado em entropia	Entropia e substituições	Bom controle estatístico	Mais complexo de parametrizar	Dados heterogêneos
No trimming	Nenhuma remoção	—	Preserva todo o sinal	Mantém ruído e erros	Testes comparativos

→

O que é Suporte Filogenético

Medida de Confiança
Não verdade absoluta

Incerteza Inerente
Sempre presente

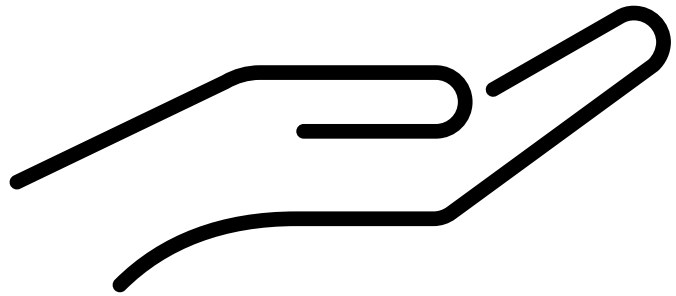
- **Medida de confiança:** Quantifica o quanto os dados sustentam uma relação evolutiva.
- **Incerteza analítica:** Expressa incerteza estatística, não verdade absoluta.
- **Dependência do método:** Valores variam conforme dados e abordagem inferencial.

Mensagem-chave: suporte ≠ verdade evolutiva.

"O suporte filogenético é uma ferramenta de avaliação, não uma prova definitiva."

Por que Avaliar Suporte

- **Comparação objetiva:** Permite contrastar hipóteses filogenéticas alternativa.
- **Distinção sinal × artefato:** Ajuda a identificar padrões espúrios.
- **Tomada de decisão:** Fundamenta inferências evolutivas mais robustas.



Bootstrap

Reamostragem de caracteres: Técnica que envolve gerar múltiplas amostras com substituição dos caracteres originais para avaliar a estabilidade das ramificações filogenéticas.

Avaliação de estabilidade: Medida de confiança baseada na frequência com que um ramo é recuperado em árvores bootstrap, expressa como porcentagem (200-1000 repetições).

Interpretação específica: Valores de bootstrap devem ser interpretados no contexto da hipótese filogenética e não representam verdade absoluta sobre a corretude do ramo.

Cuidados na interpretação: Valores altos não garantem corretude, e valores baixos não invalidam necessariamente o ramo, especialmente em presença de ruído filogenético.

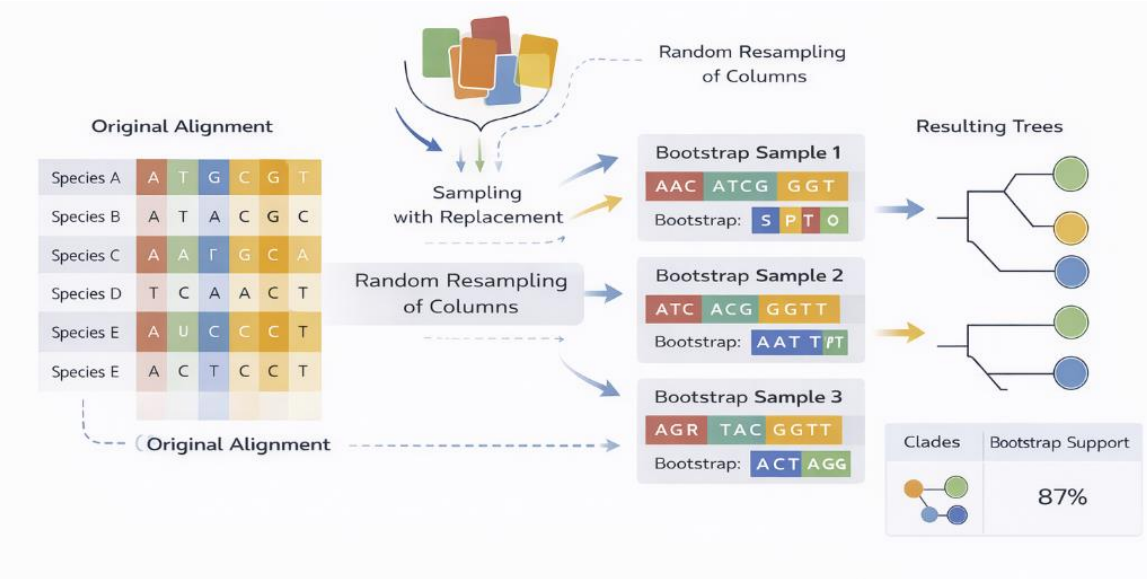


Ilustração conceitual do processo de bootstrap

"O bootstrap avalia a consistência do suporte dado pelos dados à topologia filogenética, não a verdadeira corretude do ramo."

Limitações do Bootstrap

- **Dependência do alinhamento:** Bootstrap propaga erros de alinhamento, amplificando incertezas posicionais em toda a análise.
- **Comportamento em escala genômica:** Valores de bootstrap podem ser inflacionados pela quantidade massiva de dados, mascarando conflitos reais.
- **Heterogeneidade entre loci:** Bootstrap não captura adequadamente conflitos entre genes individuais em análises concatenadas.
- **Saturação de sinal:** Em dados altamente informativos, bootstrap pode sugerir confiança excessiva mesmo com problemas subjacentes.
- **Interpretação inadequada:** Valores altos não garantem correção biológica, apenas estabilidade estatística dos dados.

Bootstrap em filogenômica requer interpretação crítica: alta confiança estatística não implica correção biológica.

Suporte em Máxima Verossimilhança

SH-aLRT Shimodaira–Hasegawa approximate Likelihood Ratio Test

- Aproximação analítica rápida baseada em qui-quadrado.
- Computacionalmente mais eficiente que bootstrap
- Utiliza a distribuição qui-quadrado para calcular valores de suporte

UFBoot Ultrafast Bootstrap

- Versão otimizada do bootstrap para máxima verossimilhança
- Reduz a variabilidade dos valores de suporte
- Proporciona resultados mais estáveis em comparação ao bootstrap clássico

Método	Base Estatística	Eficiência
Bootstrap Clássico	Reamostragem	Baixa
SH-aLRT	Qui-quadrado	Alta
UFBoot	Reamostragem	Média-Alta

Vantagens

- Menor custo computacional
- Valores de suporte mais estáveis
- Facilidade de implementação em pipelines filogenômicos

"SH-aLRT e UFBoot representam alternativas eficientes ao bootstrap clássico, especialmente em escala genômica."

Suporte Bayesiano

Probabilidade Posterior

- Medida de confiança baseada em verossimilhança e priors
- Representa incerteza a priori e dados observados
- Interpretação direta de probabilidade

Conceitos Fundamentais

- Baseada em inferência bayesiana
- Utiliza distribuições a priori
- Avança além do bootstrap

VS

Bootstrap

- Medida de estabilidade baseada em reamostragem
- Não representa verdadeira probabilidade
- Baseada em frequência relativa

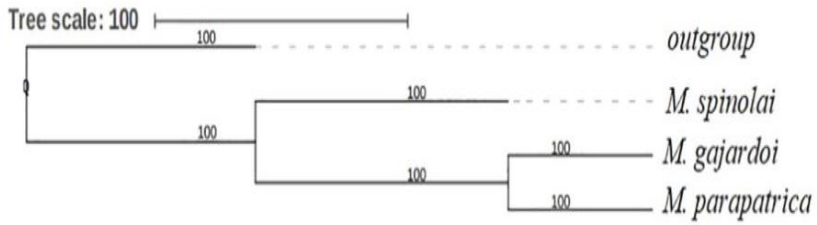
Diferenças Conceptuais

- Bootstrap não incorpora informação a priori
- Posterior probability tem interpretação mais rica
- Bootstrap pode ser visto como aproximação

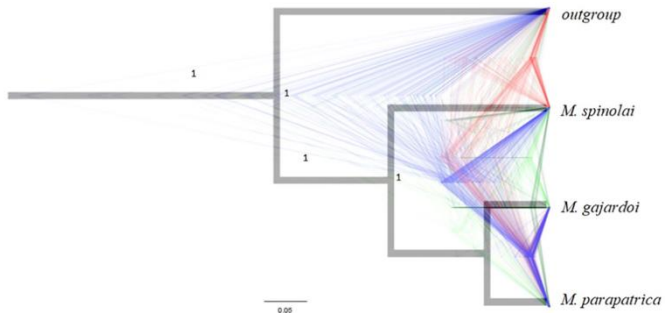
"A probabilidade posterior apresenta fundamentos conceituais diferentes do bootstrap, baseando-se em verossimilhança e informação a priori."

Suporte e Conflito Gênico

Concordância Genética



Discordância Genética



- **Concordância:** Árvores gênicas individuais apresentam padrões de relacionamento que se alinham com a árvore de espécies, indicando suporte robusto.
- **Discordância:** Árvores gênicas diferentes apresentam relacionamentos contraditórios, refletindo processos evolutivos complexos como o incompleto sorteio de linhagens.
- **Fontes de conflito:** Processos como o incompleto sorteio de linhagens, seleção natural, deriva genética e migração genealógica podem causar discordância.
- **Análise crítica:** Métodos filogenômicos devem integrar e interpretar o suporte concordante e discordante entre diferentes loci para inferir a história evolutiva.

"A compreensão do conflito gênico é fundamental para a interpretação correta do suporte filogenético em análises genômicas."

Interpretação Crítica de Suporte

Contexto Biológico

- Estrutura evolutiva
- História natural
- Conflito gênico

Tipo de Dado

- Genes codificantes
- Regiões não codificantes
- SNPs

Interpretação

- Contexto biológico
- Tipos de suporte
- Cautela interpretativa

Fatores Críticos para Interpretação de Suporte

- **Suporte ≠ Verdade:** Valores devem ser interpretados em contexto.
- **Equilíbrio:** Compromisso entre rigor e flexibilidade.
- **Consistência:** Comparação entre métodos e dados.
- **Prudência:** Cautela ao interpretar suporte alto em contextos de conflito.

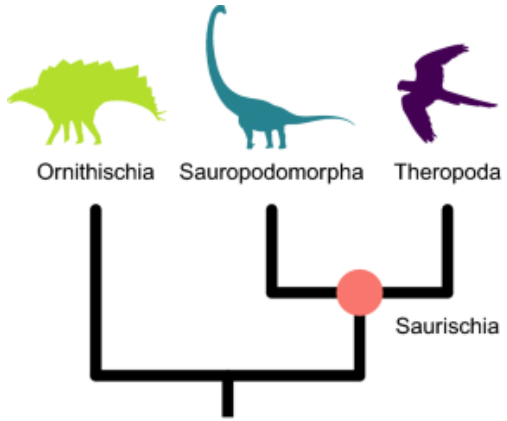
Inferência Filogenética

Reconstrução histórica
Processo de inferir hipóteses sobre relações evolutivas entre táxons a partir de dados moleculares, utilizando modelos explícitos e métodos computacionais.

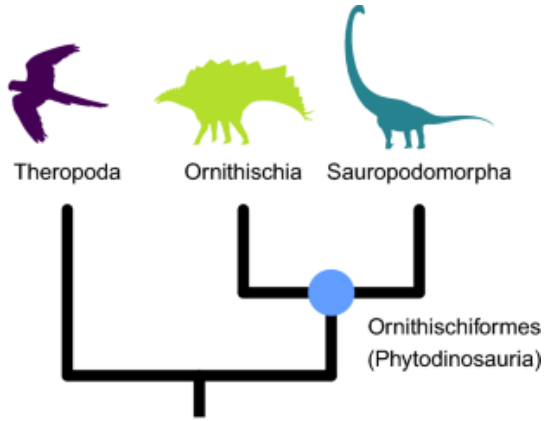
Base quantitativa
modelos matemáticos e estatísticos
Hipóteses testáveis
hipóteses científicas que podem ser comparadas

Abordagens metodológicas
Diferentes critérios inferenciais refletem distintas formas de modelar a evolução:

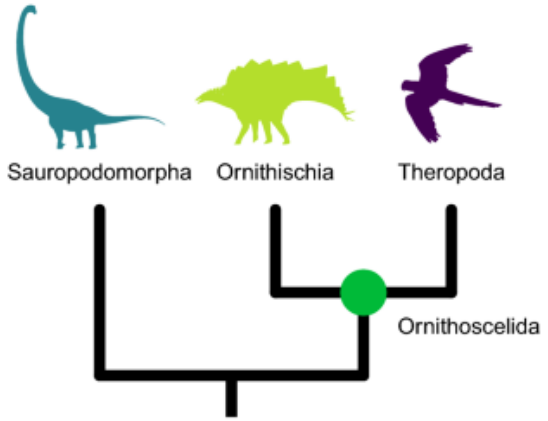
- Parcimônia
- Máxima verossimilhança
- Inferência bayesiana
- Métodos baseados em distância



After Seeley (1887)



After Cooper (1985)



After Baron et al. (2017)

Métodos Baseados em Caracteres

Abordagens fundamentais para inferência filogenética baseadas na análise de caracteres moleculares.

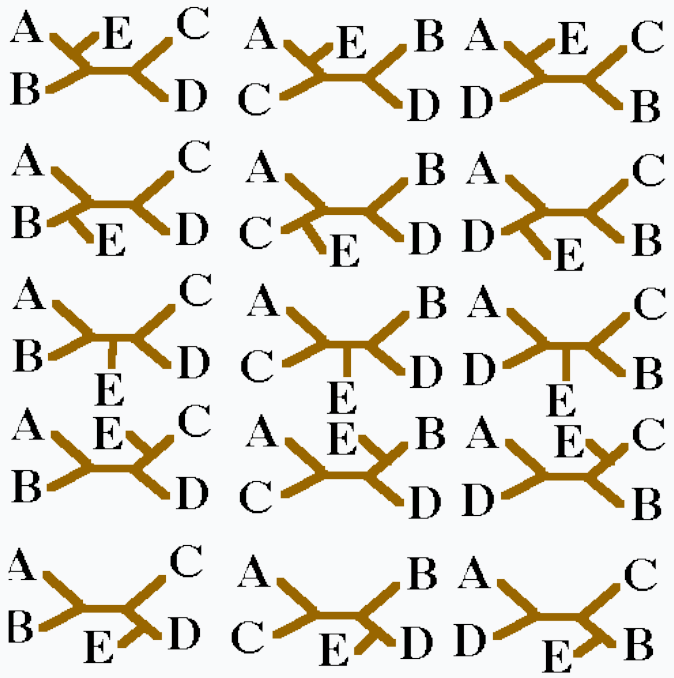
Parcimônia

- **Princípio:** Número mínimo de mudanças evolutivas para explicar os dados.
- **Características:** Não requer modelo evolutivo, baseia-se em critérios de otimização.
- **Vantagens:** Intuitiva, computacionalmente eficiente para árvores pequenas.
- **Limitações:** Pode levar a resultados distorcidos por homoplasia.

Verossimilhança

- **Princípio:** Probabilidade dos dados dado o modelo evolutivo.
- **Características:** Requer modelo evolutivo, baseia-se em maximização de verossimilhança.
- **Vantagens:** Mais robusta, permite testes estatísticos e avaliação de suporte.
- **Limitações:** Computationalmente intensiva, requer mais parâmetros.

Máxima Parcimônia

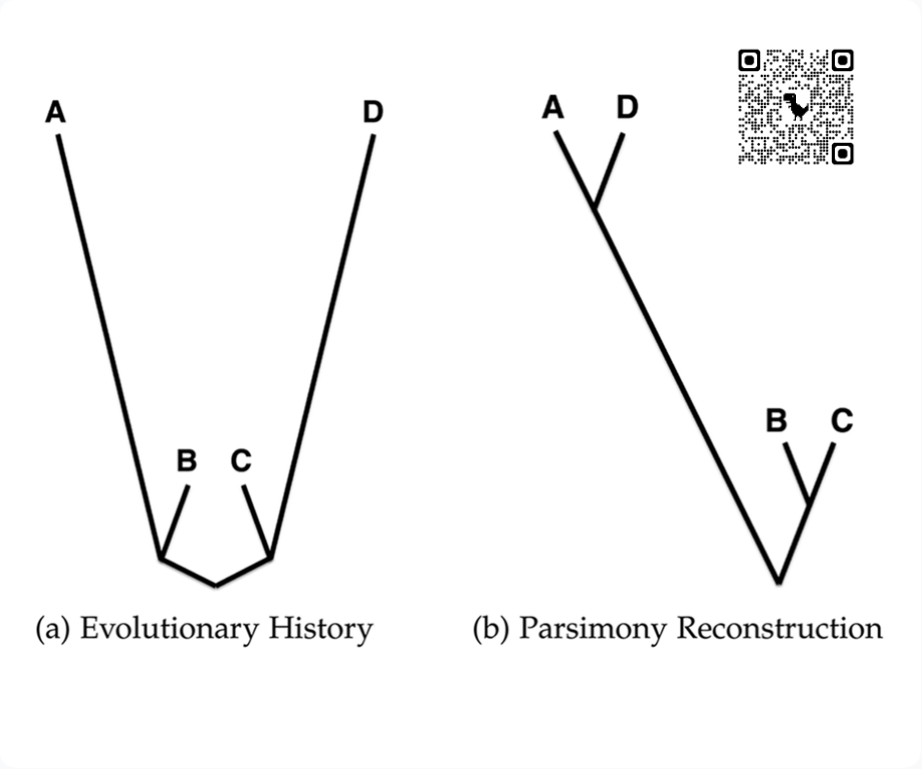


- **Princípio da parcimônia:** Busca a árvore que requer o menor número de eventos evolutivos para explicar as diferenças observadas nos dados.
- **Contagem de mudanças:** Cada mudança de estado no caractere é contada como um evento evolutivo, independentemente do número de passos intermediários.
- **Aplicação prática:** Para cada caractere, determinam-se os estados no ancestral e contam-se as mudanças ao longo das ramificações da árvore.
- **Limitações:** Pode levar a resultados enganosos em casos de alta saturação de mudanças ou convergência evolutiva.

Princípio da Mínima Mudança

A árvore que requer o menor número total de mudanças evolutivas para explicar os dados observados é a mais parcimoniosa.

Limitações da Parcimônia



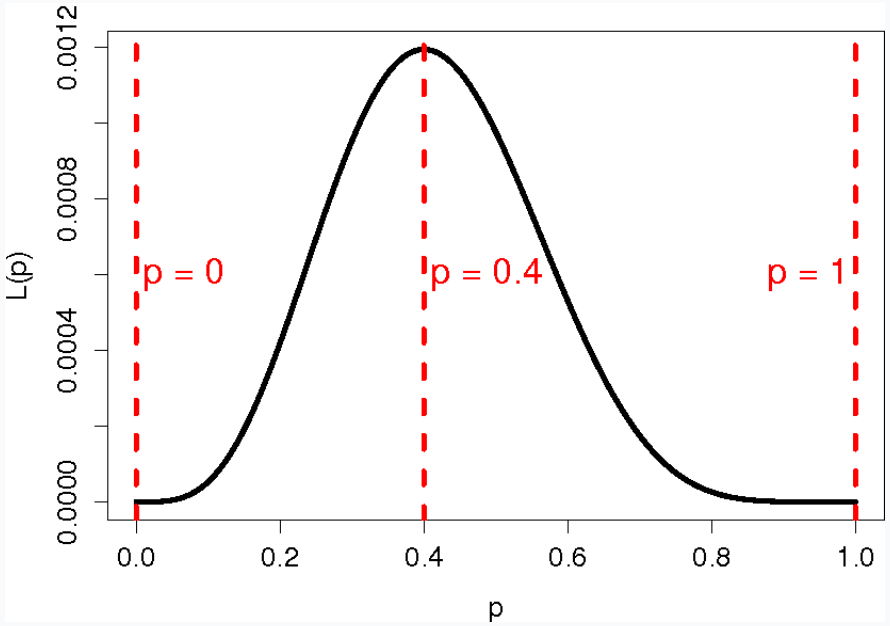
Long branch attraction como artefato sistemático

- **Long branch attraction:** Fenômeno onde ramos evolutivos longos se atraem erroneamente, levando a topologias filogenéticas incorretas.
- **Causa:** Parcimônia assume o menor número de mudanças possível, mas pode ser enganada por mudanças convergentes ou homoplásicas.
- **Condições favoráveis:** Maior em dados com alta taxa de evolução, longos ramos e baixa densidade de caracteres.
- **Soluções:** Uso de métodos alternativos (verossimilhança, bayesiana), análise de suporte e validação cruzada.

"A parcimônia, embora intuitiva, pode levar a inferências enganosas em certas condições analíticas."

Máxima Verossimilhança

- **Probabilidade dos dados** observados dado um modelo evolutivo específico.
- **Princípio:** Encontrar a árvore que maximiza a verossimilhança dos dados observados.
- **Modelos evolutivos:** Incorporam taxas de mutação, frequências de bases e outras características evolutivas.
- **Critério de seleção:** Compara múltiplas árvores para encontrar a que melhor se ajusta aos dados.



Máxima verossimilhança é um dos métodos mais robustos para inferência filogenética.

Permite avaliação de **suporte filogenético** através de testes likehood ratio.

"A máxima verossimilhança busca a árvore que torna os dados observados mais prováveis."

Inferência Bayesiana

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

- **Priors:** Representam conhecimento prévio sobre os parâmetros filogenéticos, incorporando informações externas à sequência.
- **Teorema de Bayes:** Atualiza a probabilidade a priori com base nos dados observados para calcular a probabilidade a posteriori.
- **Espaço de hipóteses:** Examina todo o espaço de árvores possíveis, ponderado pela verossimilhança e priors.
- **Probabilidade posterior:** Distribuição final que representa a incerteza sobre as hipóteses filogenéticas.
- **Vantagem:** Fornece inferência completa com medidas de incerteza, mas requer especificação de priors adequados.

"A inferência bayesiana combina evidência dos dados com conhecimento prévio para atualizar hipóteses filogenéticas."

Comparação entre MP, ML e Bayes

	Máxima Parcimônia	Máxima Verossimilhança	Inferência Bayesiana
Pressupostos	• Princípio do menor número de mudanças • Similaridade por ancestralidade • Modelo evolutivo implícito	• Probabilidade dos dados dado o modelo • Modelo evolutivo explicitado • Assunções sobre distribuição de caracteres	• Probabilidade posterior • Priors sobre parâmetros • Modelo evolutivo completo com incerteza
Escalabilidade	• Boa para árvores pequenas a médias • Algoritmos eficientes (ex: TNT) • Limitado por espaço exponencial	• Intermediária para conjuntos médios • Depende da complexidade do modelo • Implementações otimizadas	• Geralmente a mais lenta • MCMC pode ser muito demorado • Depende do número de gerações
Características	• Simplidade interpretável • Pode ser viciada por long branch attraction • Menos sensível a modelos incorretos	• Maior poder de detecção • Resposta a modelos inadequados • Estimaadores consistentes	• Trata incerteza completa • Resultados probabilísticos • Mais robusto a modelos evolutivos

Métodos Baseados em Distâncias

- **Construção de matrizes de distância:** Métricas de distância entre sequências são calculadas primeiro, antes da inferência da árvore.
- **Matrizes de distância:** Tabelas simétricas onde cada elemento representa a distância evolutiva entre pares de espécies.
- **Inferência simplificada:** Ignora a informação detalhada do alinhamento, focando apenas nas distâncias.
- **Modelos de substituição:** Distâncias são baseadas em modelos evolutivos simplificados.

Exemplo de Matriz de Distância

	A	B	C	D
A	0.00	0.10	0.30	0.25
B	0.10	0.00	0.28	0.26
C	0.30	0.28	0.00	0.12
D	0.25	0.26	0.12	0.00

Métodos conhecidos: UPGMA e Neighbor-Joining são os principais métodos baseados em distâncias.

"Métodos baseados em distâncias simplificam a inferência filogenética ao ignorar a informação detalhada do alinhamento."

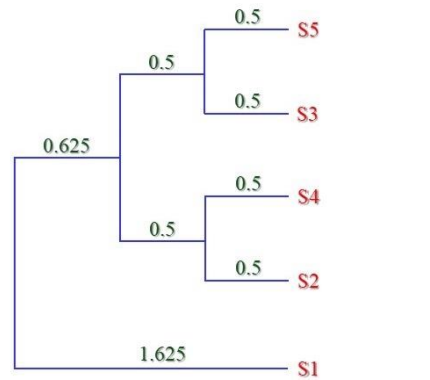
UPGMA

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) é um método de agrupamento hierárquico que constrói árvores filogenéticas baseadas em matrizes de distância.

- **Pressuposto de relógio molecular estrito:** todas as linhagens evoluem à mesma taxa constante, resultando em distâncias evolutivas proporcionais ao tempo.
- **Método de distância:** trabalha diretamente com matrizes de distância entre sequências, não utilizando os dados brutos.
- **Agrupamento hierárquico:** constrói uma árvore bifurcada passo a passo, unindo os grupos mais similares.
- **Cálculo:** a distância entre grupos é a média das distâncias entre todos os pares de elementos entre os grupos.

Exemplo de UPGMA

D	S1	S2	S3	S4	S5
S1	-	3	2	4	4
S2	3	-	2	1	4
S3	2	2	-	1	1
S4	4	1	1	-	1
S5	4	4	1	1	-



(((S5:0.5,S3:0.5):0.5,(S4:0.5,S2:0.5):0.5):0.625,S1:1.625);

Propriedades e Limitações

Vantagens

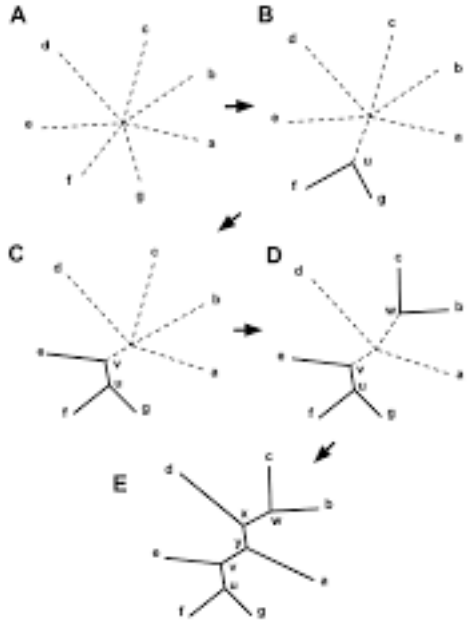
- Fácil implementação
- Rápido e eficiente
- Base matemática sólida

Limitações

- Assume relógio molecular
- Sensível a distâncias mal estimadas
- Não tolera bem dados heterogêneos

"UPGMA é inadequado para dados que violam o relógio molecular, como em grupos com taxas evolutivas divergentes."

Neighbor-Joining



Neighbor-Joining é um método de construção de árvores filogenéticas baseado em distâncias que resolve a limitação do UPGMA ao não pressupor relógio molecular.

- **Flexibilidade:** Não pressupõe relógio molecular, permitindo árvores com ramos de diferentes comprimentos que refletem diferentes taxas evolutivas.
- **Escala adequada:** Complexidade computational $O(n^2)$ torna viável para conjuntos de dados genômicos de grande porte.
- **Busca heurística:** Utiliza estratégias como o "best improvement" para encontrar a melhor topologia de árvore.
- **Vantagem sobre UPGMA:** Permite desviations do relógio molecular, resultando em árvores mais realistas para dados evolutivos complexos.

"Neighbor-Joining se mostrou robusto para dados genômicos e amplamente implementado em softwares filogenômicos."

Algoritmos de Busca Heurística

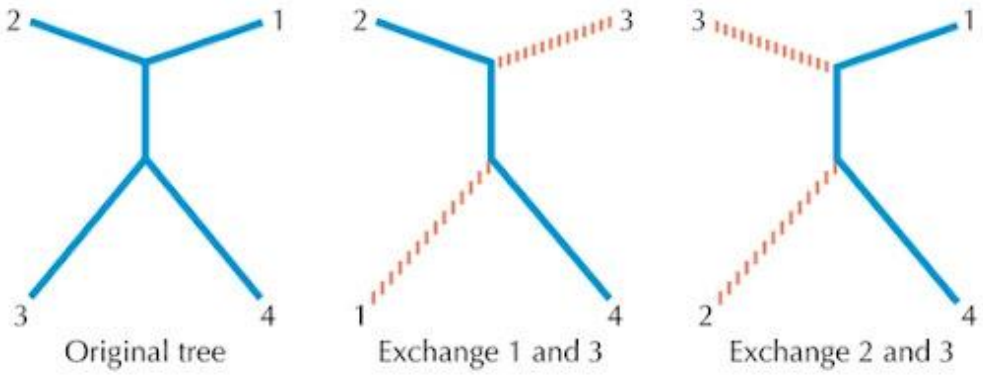
Espaço de árvores é grande demais para busca exaustiva

- Heurísticas exploram apenas parte do espaço
- Buscam boas soluções, não o ótimo garantido
- Essenciais em parcimônia, ML e Bayesiano

.

Como a Busca Heurística Funciona

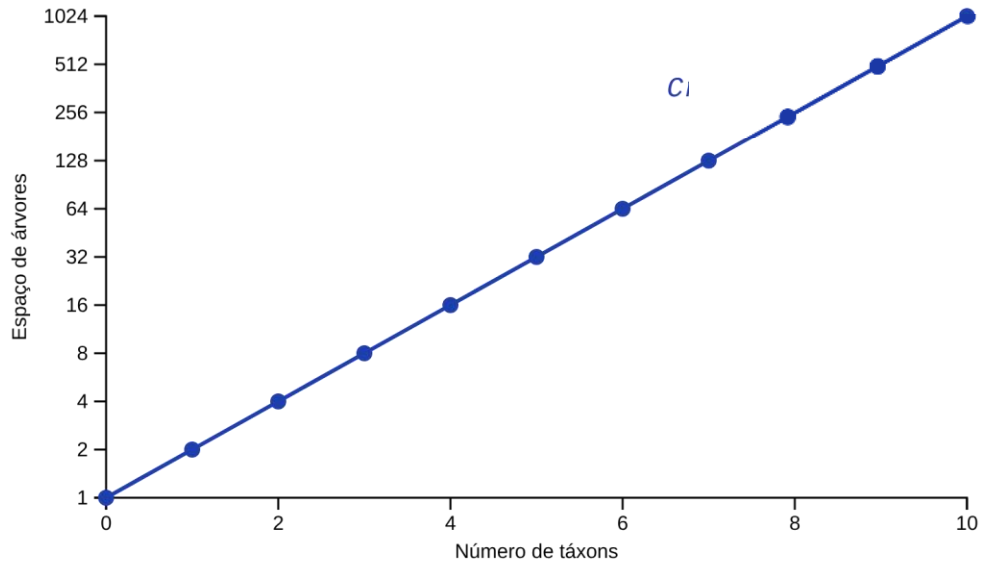
- Começa com uma árvore inicial
- Faz pequenas modificações na topologia
- Avalia se a nova árvore é melhor ou pior
- Mantém árvores melhores e descarta piores



- Pode parar em um **ótimo local**
- Resultados dependem da **árvore inicial**
- Diferentes buscas podem gerar **topologias diferentes**
- Repetições ajudam a avaliar **robustez**

Inferência em Larga Escala

- **Complexidade computacional:** O espaço de árvores filogenéticas cresce exponencialmente com o número de táxons, tornando a busca exaustiva impraticável para conjuntos de dados grandes.
- **Limites de tempo:** dias ou semanas para convergir em conjuntos de dados genômicos
- **Consumo de memória:** Métodos baseados em verossimilhança requerem matrizes de distâncias que crescem quadraticamente com o número de sequências.
- **Impacto na inferência:** Limitações computacionais impõem restrições práticas que influenciam escolhas metodológicas e interpretações filogenéticas.



Busca Heurística em Filogenética

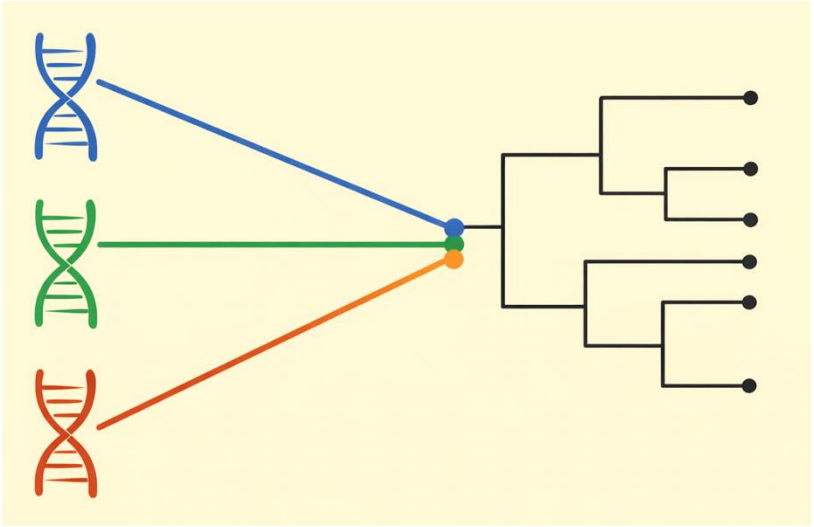
Aspecto	Resumo Essencial
Problema central	Número de árvores cresce exponencialmente → busca exaustiva se torna inviável.
Busca exaustiva	Avalia todas as árvores; precisa demais; só viável para poucos táxons.
Busca heurística	Explora parte do espaço; viabiliza inferência, mas sem garantir solução ideal.
Princípio heurístico	Começa com uma árvore e aplica mudanças que melhoram um critério.
NNI	Trocas pequenas e rápidas entre ramos vizinhos.
SPR	Corta e reinsere subárvores; mudanças intermediárias.
TBR	Divide e reconecta; explora mudanças mais profundas, com maior custo.
Dependência da busca	Resultado pode variar conforme o ponto de partida e estratégia usada.
Relação com suporte	Suporte alto ≠ melhor árvore; diferentes árvores podem ter suportes variados.
Parcimônia	Minimiza o número total de mudanças evolutivas.
Máx. Verossimilhança	Maximiza a probabilidade dos dados segundo um modelo evolutivo.
Inf. Bayesiana	Usa MCMC para amostrar árvores com base na probabilidade posterior.
Limitação fundamental	Heurísticas podem parar em soluções locais; não cobrem todo o espaço.
Boas práticas	Fazer várias buscas, comparar árvores e testar estabilidade.
Mensagem final	Filogenia é uma otimização com incerteza estatística e computacional.

Integração com Dados Genômicos

Por que é desafiador?

•Genes diferentes evoluem de formas diferentes, por causa de seleção natural, tamanho populacional, recombinação etc.

•Métodos antigos assumem que todos os genes seguem o mesmo padrão, o que nem sempre é verdade.



Desafios principais

•Loci (regiões do DNA) evoluem em velocidades diferentes.

•Pode haver conflito entre genes (ex: um gene indica uma árvore, outro indica outra).

•Genes podem ter histórias diferentes no tempo.

Soluções modernas

•Métodos como inferência multi-locus e summary methods foram criados para lidar com essas diferenças.

💡 O que isso muda?

Modelo de Coalescência Multiespécies

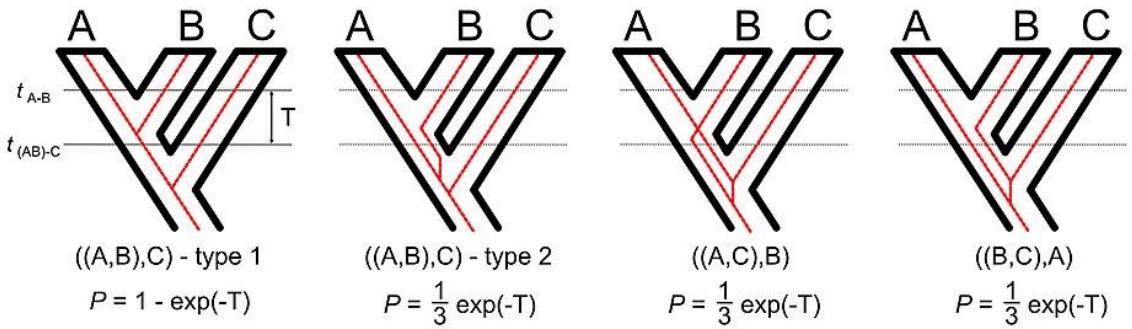
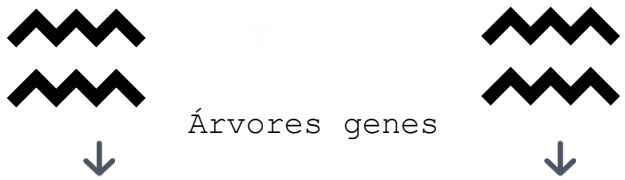


Ilustração da coalescência multiespécie mostrando a relação entre a *árvore de espécies* (contorno preto) e *as árvores gênicas* (linhas vermelhas tracejadas inseridas na árvore de espécies).

- **Extensão do modelo clássico:** O modelo de coalescência multiespécies (MSC) estabelecce a distribuição a posteriori das árvores gênicas condicionada à história evolutiva das espécies.
- **Unificação de perspectivas:** Conecta abordagens bayesianas e de máxima verossimilhança para inferência de árvores de espécies.
- **Árvores gênicas explícitas:** Modela a discordância entre árvores gênicas e a árvore de espécies como resultado de processos de coalescência.
- **Probabilidades:** Calcula a probabilidade de uma árvore gênica dada uma árvore de espécies, fundamental para métodos bayesianos fully Bayesian.

Inferência Concatenada



Árvores genes

Árvore de espécies

Pressupostos Restritivos

- **Unidade evolutiva:** Assume que todos os genes compartilham a mesma história evolutiva.

- **Relógio molecular:** Impõe taxa constante de evolução para todos os genes e linhagens.

Limitações Conhecidas

- **Discordância genética:** Falha quando há discordância entre árvores gênicas e a árvore de espécies.

- **Suposição de concatenação:** Pode levar a resultados distorcidos quando há heterogeneidade evolutiva entre loci.

"A inferência concatenada é limitada pela complexidade evolutiva real dos sistemas biológicos.”]

Vantagens e Limitações

Vantagens

- **Robustez estatística:** Métodos de summary apresentam propriedades estatísticas bem definidas sob modelos de coalescência.
- **Integração de informação:** Utilizam toda a informação disponível nos quartetos de genes para inferir a árvore de espécies.
- **Balanceamento de pressupostos:** Reduzem impacto de viéses em modelos individuais de evolução.

Limitações

- **Dependência crítica:** Resultados dependem fortemente da qualidade das árvores gênicas inferidas.
- **Complexidade computacional:** Múltipla etapa de processamento requer tempo e recursos significativos.
- **Saturação de convergência:** Risco de erro em scenarios de alta taxa de evolução.

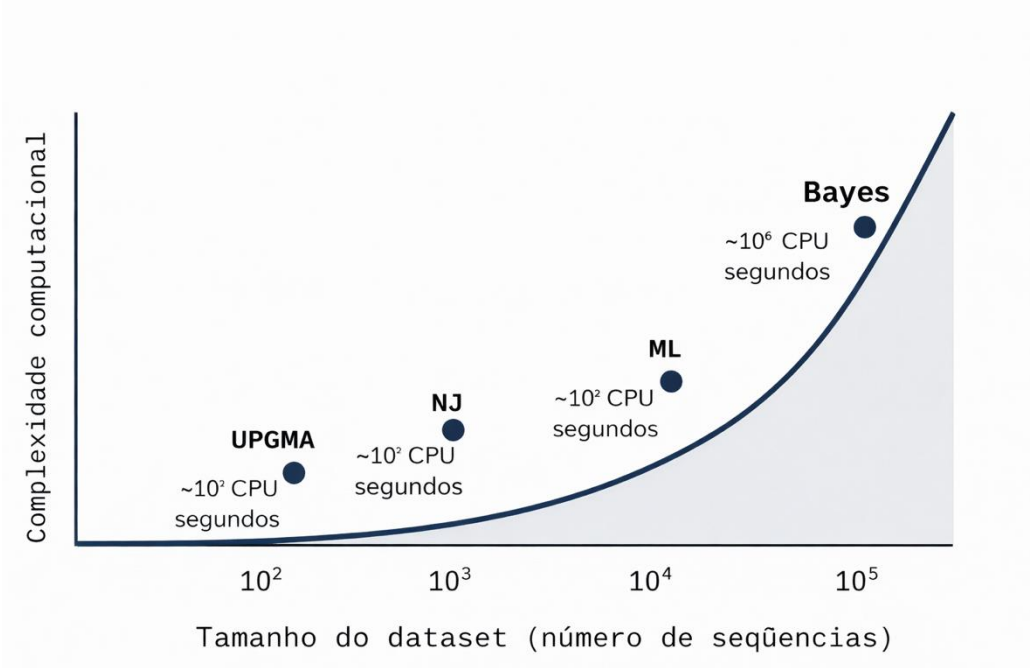
A qualidade do alinhamento e dos modelos evolutivos influencia diretamente a robustez

---

"Escolha do método deve ser guiada por qualidade dos dados e perguntas biológicas específicas."

Escala Computational

- **Limites de memória:**
- **Tempo de processamento**
- **Complexidade algorítmica**
- **Compromisso qualidade × custo**

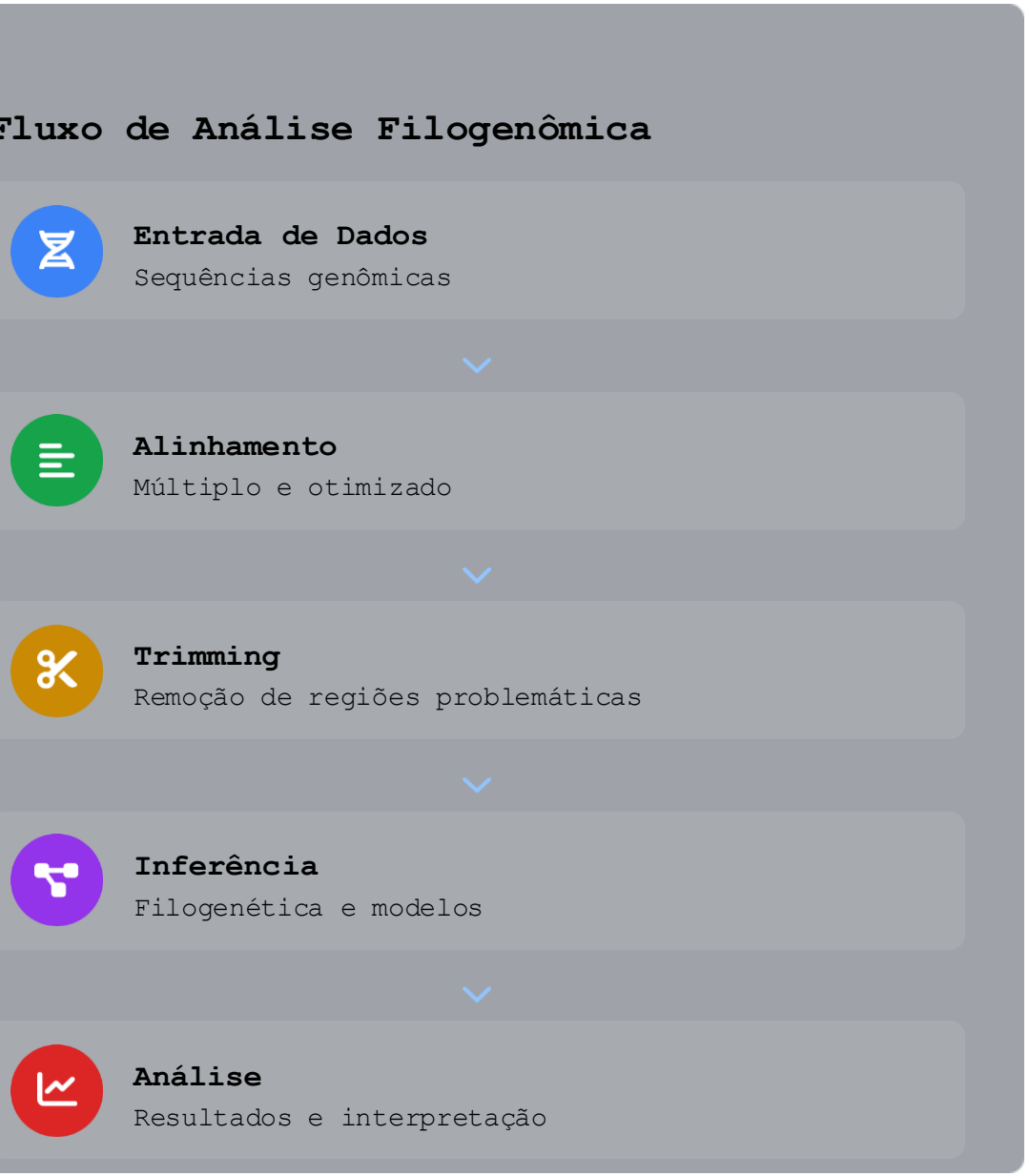


⚠ Consequências práticas:

- Limitações em análises de grande escala
- Restrições no número de sequências e loci
- Desafios para métodos bayesianos completos
- Necessidade de abordagens approximadas

Pipelines Filogenômicos

Fluxo de Análise Filogenômica



Automação e Reprodutibilidade

Automação: Processos repetitivos são automatizados, reduzindo erros humanos e aumentando a consistência.

Reprodutibilidade: Mesmos resultados para mesmas entradas, fundamental para validação científica.

Versionamento: Controle de versões para código e parâmetros.

Escalabilidade: Processamento de grandes volumes de dados.

Benefícios

- Padronização dos processos
- Documentação implícita
- Facilita a colaboração
- Permite atualizações globais

Conexão com a Prática

Alinhamento

- Qualidade do alinhamento condiciona a árvore
- Análise crítica das decisões analíticas
- Tratamento criterioso de regiões problemáticas

Inferência

- Métodos baseados em caracteres ou distâncias
- Avaliação de suporte filogenético
- Integração com modelos evolutivos

Interpretação

- Análise em contexto biológico
- Relação entre dados e hipóteses evolutivas
- Conexão entre resultados e perguntas biológicas