Algoritmos para Análise de Sequências Biológicas

Análise Filogenética

Sumário

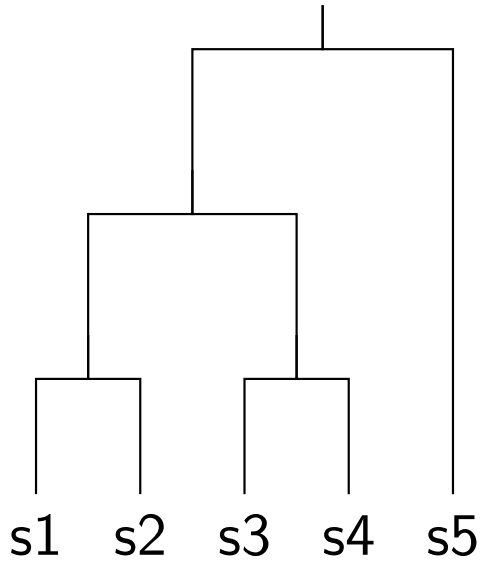
- Análise filogenética
- Algoritmo UPGMA

Definição

- **Análise filogenética** de um conjunto de sequências (DNA, RNA, proteínas) é a determinação de como cada sequência pode ter sido derivada ao longo do processo de **evolução** natural.
- Relações evolutivas são visualizadas colocando as sequências como folhas de uma **árvore evolucionária**, onde os nós de ramificação representam eventos de mutação (substituição, inserção, remoção).

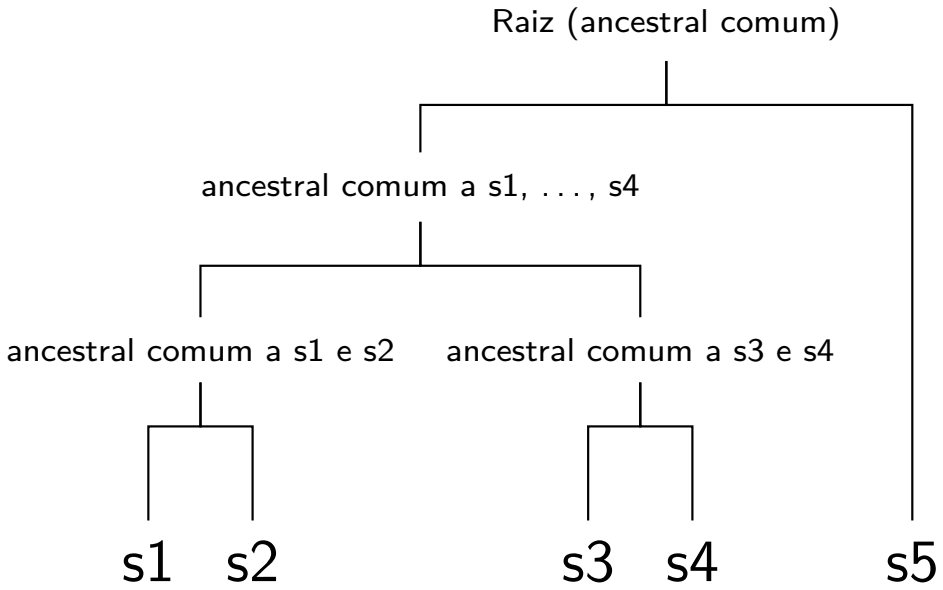
Análise Filogenética

Uma árvore filogenética sugere as relações de proximidade entre sequências;
Proximidade na árvore sugere proximidade evolutiva



Análise Filogenética

Os nós da árvore indicam ancestrais comuns



Aplicações

- Determinar a **árvore da vida** - evolução das diferentes espécies - complementando os métodos tradicionais baseados na morfologia; permitem estabelecer relações taxonómicas entre espécies ou ancestralidade entre indivíduos ou populações;
- Ajuda na determinação da **função** de sequências de DNA/ proteínas - determinação de ramos com domínios específicos que podem ter consequências funcionais;
- Análise de espécies com mutações rápidas (e.g. virus) – pode ajudar na **epidemiologia**; permite hierarquizar mutações numa árvore - antigas vs. recentes;
- Primeiro passo para alguns algoritmos de Alinhamento Múltiplo (progressivos).

Árvore de gene/sequência vs. espécie

- A evolução de um gene na maioria dos casos segue a evolução observada da espécie
- A reconstrução filogenética de um gene humano terá preponderância a agrupar o gene humano com o chimpanzé, e ambos com o gorila

Exceções

- Nem sempre a filogenia pode estar correcta; não se pode fazer inferência a partir de um só gene.
- A relação entre espécies (taxonomia) pode estar incorrecta
- Transferência horizontal
  - Típica das bactérias
  - Gene é incorporado no genoma de uma fonte exterior
  - Não seguiu a história evolutiva da espécie onde se inseriu

Árvores Evolucionárias

- Indicam o sentido da passagem do tempo
- Pode assumir-se a hipótese do relógio molecular – taxas de mutação uniformes
- Árvores podem ser representadas pelos clusters que se obtêm juntando taxa (folhas) presentes abaixo de cada nó interno (sub-árvores)
- Nº de árvores aumenta muito rapidamente com o aumento do nº de sequências.

Algoritmos de Análise Filogenética

- Objectivo: a partir de um **conjunto de sequências** (DNA ou proteínas), determinar a **árvore evolucionária** que melhor explique a sua evolução.
- Problema de **otimização**: de entre todas as árvores possíveis, escolher a que maximiza uma dada função objetivo.
- Espaço de procura tipicamente bastante grande – problema muito complexo.

Complexidade do problema

# seqs	# pares de seqs	# arvores	# ramos/árvore
3	3	3	4
4	6	15	6
5	10	105	8
6	15	945	10
10	45	34459425	18
30	435	4.95×10^{38}	58
N	$\frac{N(N-1)}{2}$	$\frac{(2N-3)!}{2^{N-2}(N-2)!}$	$2N-2$

Algoritmos de previsão filogenética

Baseados na distância Baseia-se na distância (alterações) entre pares de sequências: Neighbor Joining, UPGMA

Máxima parcimónia (ou mínima evolução) Retornam a árvore que minimiza nº de mutações necessárias para explicar a variação das sequências

Máxima verosimilhança Emprega modelos probabilísticos

Métodos baseados na distância

Baseiam-se na **distância** (inverso da similaridade) entre os diversos **pares de sequências** considerados.

Objectivo: tentar identificar sequências a colocar como **vizinhas** e determinar **comprimentos dos ramos** da árvore filogenética que representem, o mais fielmente possível, as distâncias entre os pares de sequências.

São usados como primeiro passo dos **métodos progressivos de AM** (e.g. ClustalW).

Métodos baseados na distância

Pretende-se encontrar a árvore T que minimiza

$$\text{SQE } \sum_{ij}(d_{ij}(T) - Dij)^2$$

Esta é a soma do quadrado dos erros entre a distância na árvore e a distância nas sequências dos vários taxa

O problema de estimar a árvore que minimiza SQE é um problema **NP-difícil**

Cálculo da distância

- Tipicamente, distância medida pelo **nº de carateres distintos** entre as duas sequências (edit distance)
- Métodos mais complexos podem fazer uso de matrizes de substituição (e.g. PAM, BLOSUM).
- Pode usar-se a função de mérito dos alinhamentos normalizada entre 0 e 1 (distância será 1 – mérito normalizado).

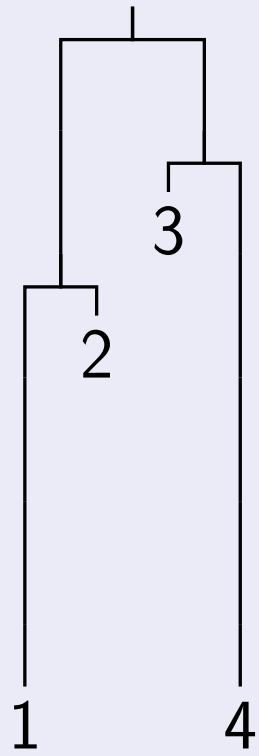
Algoritmo Unweighted Pair Group Method Using Arithmetic Averages (UPGMA)

- Algoritmo heurístico (não dá garantias de soluções óptimas mas é eficiente)
- Começa pelo par de sequências mais próximo e vai agrupando as sequências usando sempre a distância menor como critério
- Usa um algoritmo clássico de clustering: **clustering hierárquico**

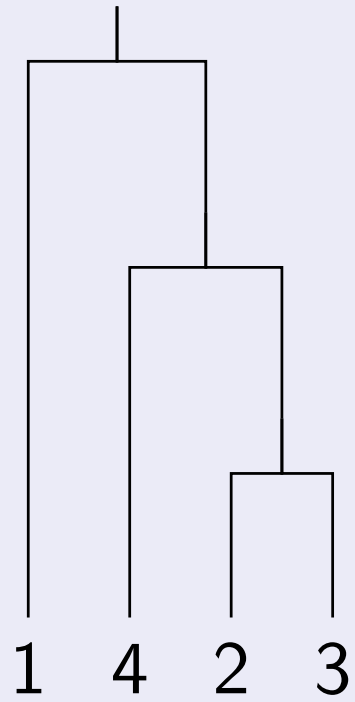
Algoritmo UPGMA

- Assume taxas de mutação uniformes em todos os ramos, logo árvores criadas são **ultramétricas**

Correta



UPGMA



UPGMA

Cada sequência é agrupada num cluster