

Exercício 6.30 (Papadimitriou)

Alice Duarte Scarpa, Bruno Lucian Costa

2015-06-23

1 Enunciado

Reconstruindo árvores filogenéticas pelo método da máxima parcimônia

Uma árvore filogenética é uma árvore em que as folhas são espécies diferentes, cuja raiz é o ancestral comum de tais espécies e cujos galhos representam eventos de especiação.

Queremos achar:

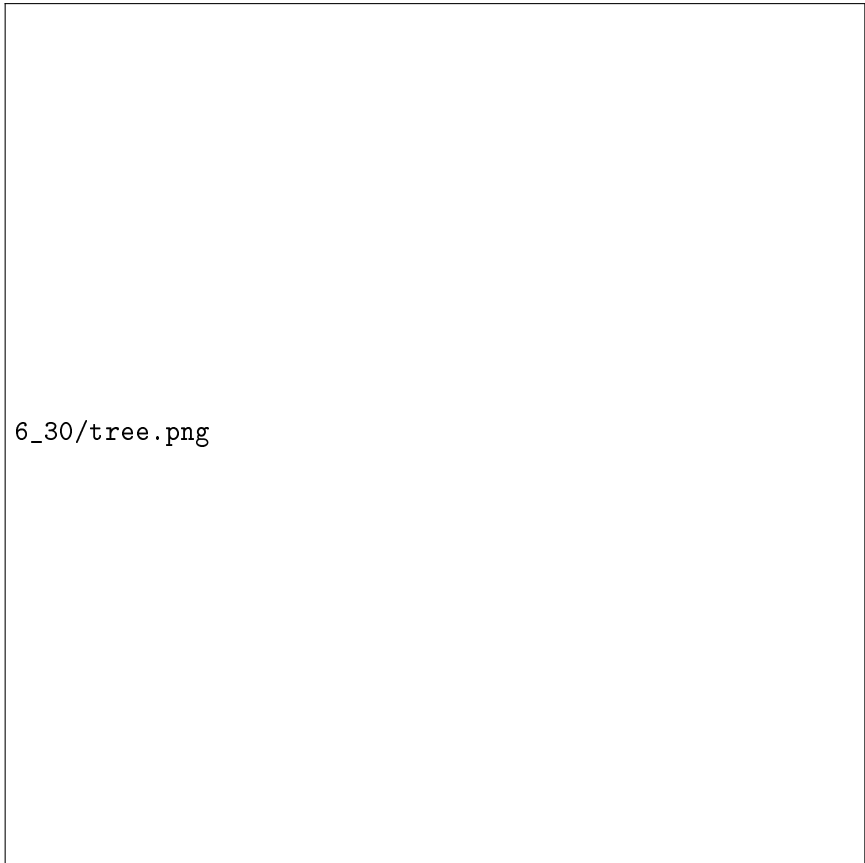
- Uma árvore (binária) evolucionária com as espécies dadas
- Para cada nó interno uma string de comprimento k com a sequência genética daquele ancestral.

Dada uma árvore acompanhada de uma string $s(u) \in \{A, C, G, T\}^k$ para cada nó $u \in V(T)$, podemos atribuir uma nota usando o método da máxima parcimônia, que diz que menos mutações são mais prováveis:

$$\text{nota}(T) = \sum_{(u,v) \in E(T)} (\text{número de posições em que } s(u) \text{ e } s(v) \text{ diferem}).$$

Achar a árvore com nota mais baixa é um problema difícil. Aqui vamos considerar um problema menor: Dada a estrutura da árvore, achar as sequências genéticas $s(u)$ para os nós internos que dêem a nota mais baixa.

Um exemplo com $k = 4$ e $n = 5$:



6_30/tree.png

1. Ache uma reconstrução para o exemplo seguindo o método da máxima parcimônia.
2. Dê um algoritmo eficiente para essa tarefa.

2 Solução

Como o valor só depende TODO. Vamos calcular a resposta para cada letra independentemente e depois concatenar as respostas para obter a árvore final.

Nós vamos usar um algoritmo de programação dinâmica para encontrar o valor das folhas intermediárias em uma árvore P em que cada folha tem valor A, G, T ou C

TODO: estrutura de dados para representar a árvore.

Colocar aqui uma estrutura de dados para a árvore

TODO: achar um nome melhor para *ans*

Vamos computar $ans[v][\ell]$ como a melhor maneira de preencher os nós da sub-árvore enraizada em v , dado que o pai de v tem valor ℓ .

TODO: justificar a inicializacao

`valor = {}`

`ans = {}`

Vamos computar *ans* de baixo para cima. Então, o caso base para esse algoritmo é a resposta para as folhas, isso é, $ans[folha][\ell]$.

Uma sub-árvore que contém apenas uma folha e seu pai vai ter nota = 0 se a folha e o pai tiverem ambos o mesmo valor (A, G, T ou C) ou nota = 1, se os dois tiverem valores diferentes:

$$ans[folha][\ell] = \begin{cases} 0 & \text{se } valor[folha] = \ell \\ 1 & \text{caso contrário} \end{cases}$$

Podemos então preencher as folhas:

#TODO: preencher as folhas

Tendo o caso base, podemos computar $ans[v][\ell]$ assumindo que $ans[w][\ell]$ já foi computado para todo w filho de v e $\ell \in \{A, G, T, C\}$.

TODO: explicar em algum lugar que a raiz é especial

A nota da sub-árvore quando o valor de v é igual a m é:

$$[\ell \neq m] + \sum_{w \text{ filho de } v} ans[w][m]$$

Onde

$$[\ell \neq m] = \begin{cases} 0 & \text{se } m = \ell \\ 1 & \text{caso contrário} \end{cases}$$

Queremos escolher um valor $m \in \{A, G, T, C\}$ para v que minimize a nota final da sub-árvore. Então:

$$ans[v][\ell] = \min_{m \in \{A, G, T, C\}} \left([\ell \neq m] + \sum_{w \text{ filho de } v} ans[w][m] \right)$$

TODO: preencher os outros vértices (exceto a raiz)

Após computarmos $ans[v][\ell]$ para todos os vértices exceto a raiz podemos encontrar a nota da árvore como o mínimo entre os possíveis valores para a raiz:

$$\min_{\ell \in \{A, G, T, C\}} \sum_{v \text{ filho da raiz}} ans[v][\ell]$$

3 Dados reais

3.1 Formato Newick

Um formato muito usado para árvores em bioinformática é o formato Newick. Assim como LISP, ele usa o fato de que parenteses podem ser usados para especificar uma árvore.

TODO: especificar o formato, referência do formato

3.1.1 Parseando o formato Newick

3.2 Rosalind

Obtemos os dados do Rosalind, TODO: explicar o Rosalind.

Rosalind MULT, GLOB, EDTA, PERM, EDIT, LCSQ, CSTR, CTBL, NWCK, SSET, MRNA, KMP, PROB SSEQ, SPLC, LCSM

3.3 Rodando o algoritmo com dados reais

4 Extensões

Ao fazer esse exercício, notamos que a árvore já é uma entrada do problema. Como é possível obter a árvore de menor valor a partir das espécies

Esse problema é NP-completo [TODO: colocar referência] e o melhor algoritmo conhecido é [TODO]