MAC–IME–USP

Carlos Eduardo Ferreira
Sala 297A Tel.: 818 6140
e-mail cef@ime.usp.br

MAC 323 – Estruturas de Dados

Primeiro semestre de 1999

Exercício-Programa 3 – Entrega: **31 de maio de 1999**

Árvores filogenéticas

Biologia computacional é a área da Ciência da Computação que surge das aplicações em ciências biológicas. Alguns dos exemplos mais importantes destas aplicações são montagens de fragmentos de DNA, inferência da estrutura secundária de moléculas, etc.

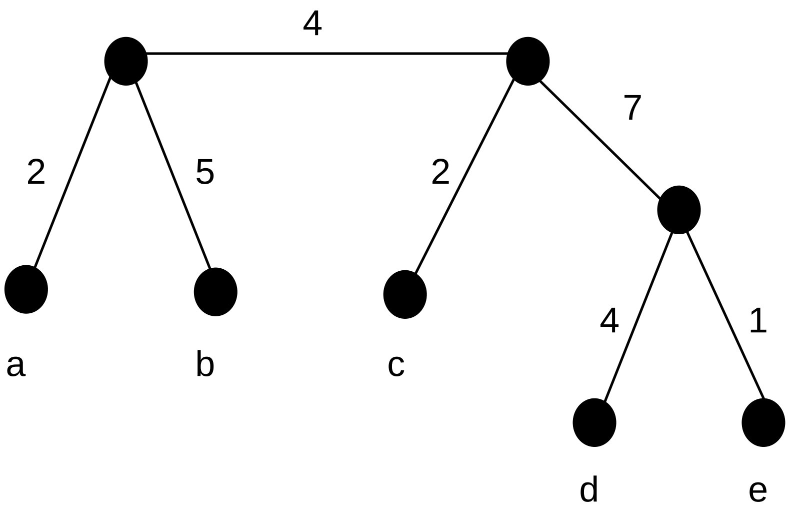
Uma aplicação muito importante é a construção de **árvores filogenéticas**. Estas árvores são explicações plausíveis de como foi o desenvolvimento das espécies até chegar às existentes hoje. A construção dessas árvores envolve muitos aspectos teóricos, e é objeto de estudo de diversos pesquisadores da área. Os interessados em Biologia Computacional podem ler mais a respeito em:

[1] Introduction to Molecular Biology
J. Meidanis, J.C. Setubal
PWS Publishing 1997.

Neste exercício-programa estudaremos a construção de um certo tipo de árvore filogenética. São dadas n espécies e uma matriz simétrica $A_{n\times n}$, em que na posição a_{ij} é dada a distância evolutiva entre as espécies i e j. O objetivo é encontrar uma árvore T, em que todos as folhas correspondam às n espécies e os nós internos têm todos grau 3. Os nós internos correspondem a possíveis “espécies” extintas das quais as espécies atuais seriam descendentes. Na árvore T cada aresta deve ter um peso de tal forma que ao calcular a distância entre duas folhas i e j em T devemos obter a_{ij}.

Exemplo: Dada a matriz abaixo, referente à distância evolutiva entre cinco espécies $\{a, b,$

$c, d, e\}$, a árvore da figura ao lado satisfaz a propriedade requerida.

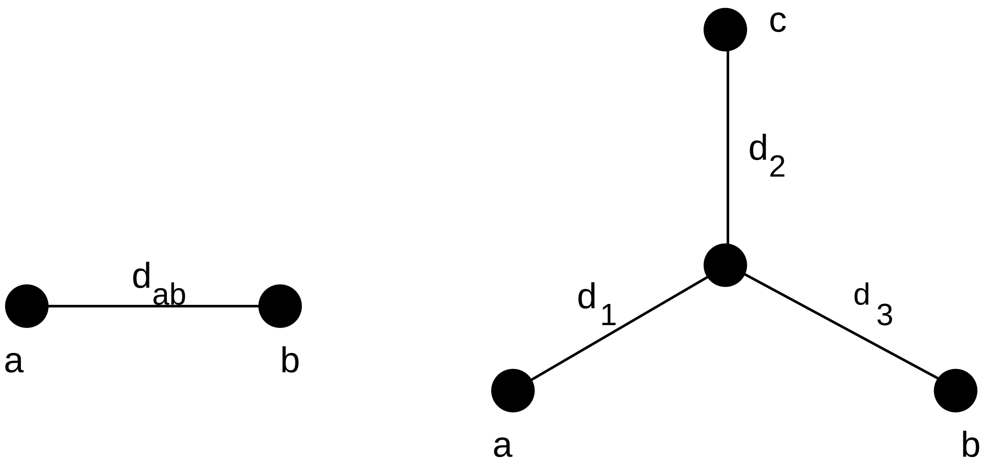


$$\begin{pmatrix} 0 & 7 & 8 & 17 & 14 \\ 7 & 0 & 11 & 20 & 17 \\ 8 & 11 & 0 & 13 & 10 \\ 17 & 20 & 13 & 0 & 5 \\ 14 & 17 & 10 & 5 & 0 \end{pmatrix}$$

No exemplo acima, a distância na árvore entre a e d é 17, a entrada correspondente na matriz.

Pode-se mostrar que, para que exista uma tal árvore, a matriz A, além de simétrica, deve respeitar a propriedade triangular, ou seja, para todo $i, j, k \in \{1, \ldots, n\}$, $a_{ij} \leq a_{ik} + a_{kj}$.

A idéia para resolver o problema é iterativa. Para duas espécies, basta construir uma árvore com apenas uma aresta, com o peso dado. Para inserir uma terceira espécie, basta encontrar o ponto em que esta aresta deverá ser dividida, dando origem a um vértice interno, a que a nova espécie será ligada. Veja a figura abaixo.



Os valores d_1, d_2 e d_3 na figura deverão satisfazer as equações:

$$\begin{array}{llll} d_1 & & +d_3 & = d_{ab} \\ d_1 & +d_2 & & = d_{ac} \\ & d_2 & +d_3 & = d_{bc} \end{array}$$

Que é um sistema de três equações e três incógnitas, facilmente resolvível. O procedimento para inserir uma nova espécie segue então a mesma idéia. A idéia é tomar um par de espécies, e tentar inserir a nova espécie no caminho que as liga. Para fazer isso devemos resolver um sistema semelhante ao exposto acima. Há duas possibilidades para a solução deste sistema:

- a solução dá origem a um novo nó interno da árvore, e a espécie nova será ligada diretamente a este vértice.
- a solução coincide com um vértice já existente. Neste caso não podemos ligar a nova espécie a este vértice, pois estaríamos criando um nó de grau 4. Então, devemos tentar um novo par de vértices para incluir a nova espécie.

Se, ao tentarmos todos os possíveis pares de vértices não pudermos incluir a nova espécie, seu programa deverá dizer que é impossível construir uma árvore filogenética para a matriz dada.

Com isso, a entrada para este exercício-programa será uma matriz A como descrito acima, e a saída deverá ser uma árvore filogenética satisfazendo as condições descritas, caso exista. Pense bem que operações você está executando sobre estas árvores, que estruturas de dados são mais eficientes para implementá-las, etc.

O algoritmo para resolver este problema está descrito no livro [1], seção 6.5.1.