INF 02798 Algoritmos para Biologia Computacional - 2004/2 16/01/05

Aula 6: Algoritmo de Aproximação para o Problema do Alinhamento Múltiplo

Instrutor: Berilhes Borges Garcia Escriba: Daniel Guimarães

DRAFT

1 Introdução

A seguir nós apresentaremos um algoritmo de tempo polinomial que computa um alinhamento múltiplo com escore SP de no máximo duas vezes o valor ótimo (Gusfield 93).

O método de Gusfield, o qual é utilizado por algumas heurísticas de uso prático, baseia-se em estudar o alinhamento uma string de cada vez.

2 O Método de Gusfield

Idéia Chave: Quando alinhando um conjunto de strings $S = \{S_1, ..., S_k\}$, concentre-se em otimizar k-1 distâncias de edição dadas na forma de uma árvore.

T é uma árvore que tem strings $S_1, ..., S_k$ como seus nós. Um alinhamento múltiplo **M** de $\{S_1, ..., S_k\}$ é consistente com **T** se o escore do alinhamento dois a dois induzido de S_1 e S_j é igual a D(S_i, S_j) quando (S_i, S_j) é uma aresta de **T**. Isto é, strings que são adjacentes na árvore são alinhadas dois a dois de uma forma ótima (enquanto que os outros podem não ser).

Example: Exemplo de um alinhamento consistente com uma árvore.
Considere o alinhamento de strings:
S_1 = A X Z A
S_2 = A X Z B
S_3 = A X X Z A
S_4 = A Y Z A
S_5 = A Y X X Z A

Considere a árvore **T**, feita das strings $S_1, ..., S_k$, mostrada na Figura 1.

Um alinhamento consistente com **T** é:
S_1': A X – – Z A
S_2': A – X – Z B
S_3': A X X – Z A
S_4': A Y – – Z A
S_5': A Y X X Z A
⊠

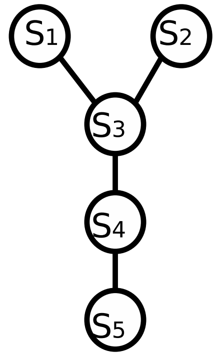

Figure 1: Exemplo de um árvore feita das strings $S_1, ..., S_k$

Teorema 1. *Dado um conjunto de strings $S_1, ..., S_k$ e uma árvore* ***T*** *feita das strings de S, nós podemos eficientemente computar um alinhamento* ***M*** *de S que é consistente com T.*

Proof. Primeiro compute um alinhamento ótimo (com distância D(S_i, S_j) para qualquer pares de strings S_i e S_j adjacentes na árvore.

Até que todas strings tenham sido alinhadas, selecione duas strings, $\overline{S}_i$ em **M** (possivelmete com espaços inseridos) e S' ainda não alinhado, tal que (S_i, S') é uma aresta de **M**; alinhe $\overline{S}_i$ e S' atribuindo peso zero para espaços inseridos em S' contra espaços em $\overline{S}_i$, este alinhamento tem escore D(S_i, S').

Então adicione $\overline{S}$' (possivelmente com espaços adicionados) ao alinhamento; se novos espaços são inseridos em $\overline{S}_i$, adicione-os nas colunas correspondentes das outras linhas de **M**, também. Note que o escore do alinhamento dois a dois induzido de **M** não muda.

□

Complexidade: Se **l** é o tamanho final da linha de **M**, todos os k-1 alinhamentos dois a dois pode ser computado em tempo total O(k$\mathbf{l}^2$).

Alinhamento baseado em árvore é utilizado para aproximar um alinhamento otimal, utilizando uma árvore estrela. Esta árvore possui uma string central, convenientemente escolhida, que é conectada diretamente às outras strings.

Definição: Erro consesual de uma string S_k relativa a conjunto de strings $\mathcal{S}$ é

$$E(S_k) = \sum_{S_i \in \mathcal{S}} D(S_k, S_i)$$

Uma string $S_c \in \mathcal{S}$ é uma string central se $E(S_c) \leqslant E(S_i)$ para todo $S_i \in \mathcal{S}$.

Uma estrela central é uma árvore (livre) que contém uma aresta conectando a string central a cada uma das outras strings de $\mathcal{S}$.

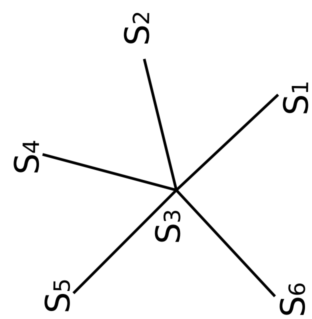

Figure 2: Estrela Central das strings $S_1, ..., S_k$

Uma string central pode ser encontrada em tempo polinomial. (Exercício).

$\mathbf{M}_c$ é um alinhamento múltiplo de $\mathcal{S}$ que é consistente com uma estrela central de $\mathcal{S}$.

$\mathbf{M}_c$ pode ser utilizado para aproximar um alinhamento otimal se o esquema de pontuação satisfaz a desigualdade triangular: $s(x, y) \leqslant s(x, z) + s(z, y)$.

Note que nem todas as matrizes de pontuação em biologia computacional satisfazem a desigualdade triangular.

Denote por $d(S_i, S_j)$ o escore do alinhamento dois a dois das strings S_i e S_j induzidos por $\mathbf{M}_c$. De modo a simplificar a apresentação chame $d(S_i, S_j)$ a distância de S_i e S_j induzidos pelo alinhamento $\mathbf{M}_c$.

Lema 1. *Se a matriz de pontuação satizfaz a desigualdade triangular, então*

$$d(S_i, S_j) \leqslant d(S_i, S_c) + d(S_c, S_j) \tag{1}$$

$$d(S_i, S_j) = D(S_i, S_c) + D(S_c, S_j) \tag{2}$$

Proof. (1) é verdade porque a desigualdade triangular é verificada em cada coluna do alinhamento de S_i, S_c e S_j.
(2) é verdade porque o alinhamento $\mathbf{M}_c$ é consistente com a estrela central. □

Assuma que $\mathbf{M}^*$ é um alinhamento múltiplo otimal de $\mathcal{S} = S_1, ..., S_k$.

Denote por $d^*(S_i, S_j)$ o escore do alinhamento dois a dois de S_j, S_j induzidos por $\mathbf{M}^*$.

Denote o escore SP de um alinhamento $\mathbf{M}$ por $d(\mathbf{M})$, nós temos que

$$d(\mathbf{M}^*) = \sum_{i<j} d^*(S_i, S_j)$$

e

$$d(\mathbf{M}_c) = \sum_{i<j} d(S_i, S_j)$$

Nós podemos estabelecer o resultado principal:

Teorema 2. $\dfrac{d(\boldsymbol{M}_c)}{d(\boldsymbol{M}^*)} \leqslant (2 - \dfrac{2}{k})$.

Proof. Considere a razão $\dfrac{d(\mathbf{M}_c)}{d(\mathbf{M}^*)}$ em termos das expressões que contam as distâncias induzidas duas vezes:

$$v(\mathbf{M}_c) = \sum_{i \neq j} d(S_i, S_j)$$

e

$$v(\mathbf{M}^*) = \sum_{i \neq j} d^*(S_i, S_j)$$

ou seja, $v(\mathbf{M}_c) = 2d(\mathbf{M}_c)$ e $v(\mathbf{M}^*) = 2d(\mathbf{M}^*)$, dessa forma

$$\frac{d(\mathbf{M}_c)}{d(\mathbf{M}^*)} = \frac{v(\mathbf{M}_c)}{v(\mathbf{M}^*)}$$

Nós agora podemos estimar $v(\mathbf{M}_c)$

$$v(\mathbf{M}_c) \leqslant \sum_{i \neq j} [D(S_i, S_c) + D(S_c, S_j)]$$

$$v(\mathbf{M}_c) = 2(k-1) \sum_{j} D(S_c, S_j)$$

$$v(\mathbf{M}_c) = 2(k-1) E(S_c)$$

Por outro lado, $v(\mathbf{M}^*)$ pode ser estimado da seguinte forma:

$$v(\mathbf{M}^*) = \sum_{(i,j)} d^*(S_i, S_j)$$

$$v(\mathbf{M}^*) \geq \sum_{(i,j)} D(S_i, S_j) = \sum_{i} \sum_{j} D(S_i, S_j)$$

$$v(\mathbf{M}^*) \geq k \sum_{j} D(S_i, S_j) = kE(S_c)$$

Logo,

$$\frac{d(\mathbf{M}_c)}{d(\mathbf{M}^*)} = \frac{v(\mathbf{M}_c)}{v(\mathbf{M}^*)}$$

$$\frac{d(\mathbf{M}_c)}{d(\mathbf{M}^*)} \leq \frac{2(k-1)E(S_c)}{kE(S_c)} = 2 - \frac{2}{k}$$

□

Observações sobre a Aproximação Estrela Central
Na prática, o escore SP de um alinhamento estrela central pode ser menor que duas vezes o escore otimal. Alguns testes mostraram que o alinhamento estrela central desvia-se do ótimo de 2% *à* 16% somente. Note que um PTAS (esquema de aproximação de tempo polinomial) também foi desenvolvido para o problema do alinhamento SP. Este método produz um alinhamento múltiplo de K strings com escore $SP \leq 2 - \frac{q}{k}$ vezes o ótimo. A precisão da aproximação pode ser melhorado, mas isto também aumenta o tempo de execução (como uma função de q).