Biologia Computacional - 2004/2 11/30/04

Aula 4: Alinhamento Múltiplo de Strings

Instrutor: Berilhes Borges Garcia Escriba: Letícia Rosetti Margoto

DRAFT

1 Definição

Um alinhamento múltiplo global de $k > 2$ strings é uma generalização direta de alinhamento de duas strings. Insira espaços dentro das strings de modo que elas tenham o mesmo tamanho (digamos l caracters), arrange-as em k linhas e l colunas, de modo que só exista um caracter ou espaço por coluna.
Exemplo: Alinhamento múltiplo das strings {abca, ababa, accb, cbbc}.

a	b	c	_	a
a	b	a	b	a
a	c	c	b	_
c	b	_	b	c

Também existem alinhamento múltiplos locais.
Um alinhamento múltiplo local das strings S_1,..., S_k consiste em selecionar uma substring S_i' de cada S_i e alinhar estas k substrings globalmente.

2 Motivação

Alinhamento múltiplo é uma formalização de comparação múltipla de strings. Esta é uma das metodologias mais importantes e uma área de pesquisa bastante ativa em análise de bio-sequências. Esta é utilizada para:

- extrair e representar similaridades biologicamente importantes de um conjunto de strings. Similaridades estas que podem passar despercebidas se nós somente compararmos duas strings.
- inferir história evolutiva de sequências de DNA ou proteínas.

Encontrar partes comuns é utilizado para caracterizar (e entender) famílias de proteínas.

- família: um conjunto de proteínas relacionadas pela estrutura, função ou história evolutiva, por exemplo, globinas e imunoglobulinas.

A representação de familías de proteínas é útil, pois estima-se que aproximadamente 30.000 proteínas humanas poderiam ser agrupadas em algumas centenas de famílias.
Uma nova sequência pode ser testada para inclusão em uma família comparando-a com a representação da família.
Formas normalmente utilizadas para representar um família são: perfis, sequências de consenso e assinaturas. Todas essas representações são derivados da comparação múltipla de strings.

3 Perfis como representação de famílias

Um perfil de um alinhamento múltiplo M com tamanho de linha l uma matriz p de dimensão $|\sum^{'}|$ x l, onde p(y, j) é a frequência de ocorrência do caracter y na coluna j do alinhamento M.
Exemplo: O perfil do alinhamento múltiplo mostrado anteriormente é:

p	1	2	3	4	5
a	0,75	0,00	0,25	0,00	0,50
b	0,00	0,75	0,00	0,75	0,00
c	0,25	0,25	0,50	0,00	0,50
_	0,00	0,00	0,25	0,25	0,25

Como comparar uma string e um perfil?

4 Alinhando uma string a um perfil

Um perfil p é uma sequência de colunas. Nós podemos alinhar uma string S com p, inserindo-se espaços no perfil e na string.
Exemplo:

$p^{'}$	1	_	2	3	4	5
a	0,75		0,00	0,25	0,00	0,50
b	0,00		0,75	0,00	0,75	0,00
c	0,25		0,25	0,50	0,00	0,50
_	0,00	1,00	0,00	0,25	0,25	0,25

S'	a	a	b	_	b	c

Como atribuir um escore a um alinhamento perfil/string?
Um abordagem comum é:

1. O escore S(x, j) de um caracter x alinhado com uma coluna j é a média dos escores dos pares de caracter com relação à x e qualquer caracter na coluna j.

$$S(x,j) = \sum_{y \in \sum^{'}} [s(x,y)p(y,j)]$$

2. A pontuação do alinhamento total é igual a soma dos escores das colunas.

Exemplo: Assuma o seguinte escore s(a, a)=2, s(a, b)= s(a, _)=-1 e s(a, c)=-3. A primeira coluna do alinhamento anterior adiciona S(a, 1)= 0,75. 2 + 0,25.(-3) ao escore total e a segunda coluna S(a, _)=1,0.(-1).

Um alinhamento string/perfil pode ser computado como uma extensão direta do alinhamento ótimo do prefixo $S[1...i]$ com colunas i,...,j do perfil p.
Recorrências:
Caso Base:

$$V(0,j) = \sum_{k=1}^{j} S(\_, k)$$

("_" contra as j primeiras colunas de p)

$$V(i,0) = \sum_{k=1}^{i} s(S_1[k], \_)$$

($S_1[1...i]$ contra espaços)

Caso indutivo para i, $j > 0$:

$$V(i,j) = max \begin{cases} V(i-1, j-1) + S(S_1[i], j) \\ V(i-1, j) + s(S_1[i], \_) \\ V(i, j-1) + s(\_, j) \end{cases} \quad (1)$$

Com estas recorrências um alinhamento string/perfil ótimo pode ser computado, similarmente ao alinhamento string/string, em tempo O($|\sum|$nm).
O fator $|\sum|$ vem de considerar todos os caracteres da coluna j para computar o escore de alinhamento de um caracter na coluna j.