Biologia Computacional - 2004/2 16/11/04

Aula 2: Distância de edição com peso nas operações, similaridade de *strings*, casamento de padrões aproximado

Instrutor: Berilhes Borges Garcia Escriba: Idilio Drago

DRAFT

1 Distância de edição com peso nas operações

Example: Um alinhamento com:

- substituição $r = 2$
- espaço $d = 4$
- casamento $e = 1$

```
v i n t n e r _
w r i t _ e r s
```

$peso = 2 + 2 + 2 + 1 + 4 + 1 + 1 + 4 = 17$ ⊠

Peso nas operações introduz modificações diretas nas recorrências.

- Caso base com peso nas operações:

$$D(i, 0) = i \times d \tag{1}$$

$$D(0, j) = j \times d \tag{2}$$

(← Quantos caracteres são removidos ou inseridos)

- Caso indutivo, para $i, j > 0$, com peso nas operações:

$$D(i,j) = min \begin{cases} D(i-1,j) + d & \text{(D)} \\ D(i,j-1) + d & \text{(I)} \\ D(i-1,j-1) + t(i,j) & \text{(M,R)} \end{cases} \tag{3}$$

com

$$t(i,j) = \begin{cases} e & \text{se } S_1[i] = S_2[j] \\ r & \text{caso contrário} \end{cases} \tag{4}$$

Com essas modificações, a distância de edição com peso nas operações e a correspondente transcrição/alinhamento podem ser computadas em tempo $O(nm)$, exatamente como antes.

2 Outra generalização

O custo das operações de edição pode também depender de quais caracteres do alfabeto estão envolvidos (distância de edição com alfabeto ponderado).

O que é normalmente utilizado, peso nas operações ou peso nos alfabetos?

- Proteínas são freqüentemente comparadas utilizando peso no alfabeto, neste caso o alfabeto dos amino-ácidos.

  Note que não existe nenhuma concordância sob quais pesos utilizar. As matrizes de peso PAM e BLOSUM são as duas matrizes de peso mais freqüentemente utilizadas.

- Cadeias de DNA são mais freqüentemente comparadas utilizando-se pesos nas operações ou pesos iguais. O programa de pesquisa em banco de dados BLAST utiliza $c = +5$ e $r = -4$ Note que estas aplicações maximizam a similaridade (e não minimizam a distância de edição).

Uma formalização alternativa para a proximidade de *strings* é sua similaridade (ao invés de distância), que é utilizada em muitas bio-aplicações.

Assuma que Σ é o alfabeto de duas *string* S_1 e S_2, e Σ' e Σ mais o caractere "_" denotando um espaço.

O escore do alinhamento dos caracteres x e y de Σ' é denotado por $s(x, y)$. $S'_1[1...l]$ e $S'_2[1...l]$ são duas versões das *strings* S_1 e S_2 que são complementadas com espaços para permitir um alinhamento A.

O valor de alinhamento (ou escore total) A é:

$$\sum_{i=1}^{l} s(S'_1[i], S'_2[i]) \tag{5}$$

Considere a seguinte matriz de escore para $\Sigma' = \{a, b, c, d, \_\}$

s	a	b	c	d	_
a	1	-1	-2	0	-1
b		3	-2	-1	0
c			0	-4	-2
d				3	-1
_					0

Freqüentemente $s(x, y) >= 0$ se e somente se $x = y$.

Valor de um alinhamento $S_1 = cacdbd$ e $S_2 = cabbdb$:
S'_1 : $c\ a\ c\ \_\ d\ b\ d$
S'_2 : $c\ a\ b\ b\ d\ b\ \_$
A : 0 1 -2 0 3 3 -1 = 4

A similaridade das strings S_1 e S_2 é o valor de um alinhamento de S_1 e S_2 que maximiza o escore total, dando o valor do alinhamento ótimo.

A similaridade de strings e a distância de edição com peso no alfabeto estão relacionadas, mas não são equivalentes. Nós veremos que similaridade é mais conveniente para computar alinhamentos locais.

A similaridade das strings S_1 e S_2 pode ser computada com uma modificação direta da distância de edição. Para isso, vamos definir $V(i, j)$ como sendo o valor do alinhamento ótimo dos préfixos $S_1[1...i]$ e $S_2[1...j]$.

- A condição base é óbvia:

$$V(0, j) = \sum_{k=1}^{j} s(\text{"\_"}, S_2[k]) \tag{6}$$

$$V(i, 0) = \sum_{k=1}^{i} s(S_1[k], \text{"\_"}) \tag{7}$$

- O caso indutivo para $i, j > 0$ considera os escore expecíficos de caracteres:

$$V(i, j) = max \begin{cases} V(i-1, j) + s(S_1[i], \text{"\_"}) \\ V(i, j-1) + s(\text{"\_"}, S_2[j]) \\ V(i-1, j-1) + s(S_1[i], S_2[j]) \end{cases} \tag{8}$$

Aplicando estas recorrências, a similaridade pode ser computada de forma semelhante à distância de edição, obtendo $V(n, m)$ no canto inferior direito da tabela, dessa forma a similaridade e o alinhamento ótimo de $S_1[...n]$ e $S_2[...m]$ podem ser computados em tempo $O(nm)$.

3 Casamento de padrões aproximado

Uma generalização importante do casamento exato consiste em determinar as ocorrências similares de um padrão (não apenas cópias exatas).

Uma *substring* T' de T é uma ocorrência aproximada de P se e somente se a similaridade de P e T' é pelo menos δ (para um determinado δ).

Ocorrências aproximadas do padrão P no texto T podem ser computadas como uma pequena variação do alinhamento global. Esta variação consiste em aplicar a recorrência anterior (com P no lugar de S_1, e T no lugar de S_2), mas o caso base muda de modo que:

$$V(0, j) = 0 \tag{9}$$

Dessa forma espaços no começo são "livres".

A tabela $V(i, j)$ pode ser preenchida, como antes, em tempo $O(mn)$.

Teorema 1 (Gusfield). *$T[k...j]$ é uma ocorrência aproximada de $P[1...n]$ em T se e somente se $V(n, j) >= \delta$, e existe um caminho da célula (n, j) para $(0, k)$.*

Podem existir ocorrências aproximadas múltiplas de P (de tamanhos diferentes) terminando na mesma posição J de T. A mais curta pode ser localizada da seguinte forma:

1. Encontre cada coluna j na linha n onde $V(n, j) >= \delta$
2. Para cada uma delas, trilhe os ponteiros de (n, j) para uma coluna k da linha 0, preferindo ponteiros $\uparrow$ à $\nwarrow$ e $\nwarrow$ à $\leftarrow$.

4 Alinhamento local

Ao invés de computar a similaridade global de *strings*, é algumas vezes mais importante localizar as regiões similares das *strings*.

Alinhamento local (ou similaridade local): Dados duas *strings* S_1 e S_2, encontre *substrings* α de S_1 e β de S_2 de similaridade máxima.

Note que *substrings* de distância de edição mínima seriam *substrings* casando perfeitamente (possivelmente com um único caractere). Maximizar similaridade é assim mais útil quando procurando áreas de alta similaridade.