Coletânea de Exercícios

26 de setembro de 2016

1 [SM97, adap. 3-2] Score alignment below according to the following system: $p(a, b) = 1$ if $a = b$, $p(a, b) = 0$ if $a \neq b$, and $g = -1$.

```
GA-CGGATTAG
GATCGGAATAG
```

2 [SM97, adap. 3-3] Find all optimal alignments between AAAG and ACG, with the scoring system $p(a, b) = 1$ if $a = b$, $p(a, b) = -1$ if $a \neq b$, and $g = -2$.

3 Encontre um alinhamento global ótimo entre `TACTGTTAGT` e `TCTAT` com pontuação $M = 1$, $m = -1$ e $g = -2$.

4 Encontre um alinhamento local ótimo entre `TACTGTTAGT` e `TCTAT` com pontuação $M = 1$, $m = -1$ e $g = -2$.

5 [Adi12, 8] Dadas as seqüências $s = ABRACADABRA$ e $t = CABECADECABRA$, encontre todos os alinhamentos ótimos globais, locais e semi-globais dessas seqüências.

6 [Adi12, 9] Escreva um algoritmo que receba uma matriz de alinhamento de duas seqüências s e t, as seqüências s e t e devolva todos os alinhamentos ótimos globais dessas duas seqüências.

7 [JP04, 6.20] Consider the sequences v = TACGGGTAT and w = GGACGTACG. Assume that the match premium is +1 and that the mismatch and indel penalties are -1.

a. Fill out the dynamic programming table for a global alignment between v and w. Draw arrows in the cells to store the backtrack information. What is the score of the optimal global alignment and what alignment does this score correspond to?

b. Fill out the dynamic programming table for a local alignment between v and w. Draw arrows in the cells to store the backtrack information. What is the score of the optimal local alignment in this case and what alignment achieves this score?

c. Suppose we use an affine gap penalty where it costs -20 to open a gap, and -1 to extend it. Scores of matches and mismatches are unchanged. What is the optimal global alignment in this case and what score does it achieve?

8 [SM97, 3-8] Given two sequences, which value is larger: their local similarity or their global similarity? Why? How does their semiglobal similarity compare with the other two values?

9 [SM97, adap. 3-4] Um alinhamento é chamado de upmost se ele for produzido dando preferência para a célula de cima quando há empates na pontuação e é chamado de downmost se ele for

produzido dando preferência para a célula da esquerda quando há empates na pontuação. Os alinhamentos upmost e downmost normalmente delimitam uma região da matriz que contém todos os alinhamentos ótimos entre duas cadeias. Mostre um exemplo em que isso não acontece.

10 A distância de Levenshtein (também chamada de distância de edição) é o número de operações de inserção, remoção e substituição de caracteres para transformar uma cadeia em outra. Mostre como calcular a distância de edição entre duas cadeias s e t.

11 Uma subseqüência de uma cadeia s é um subconjunto de símbolos de s que preserva a ordem relativa. Por exemplo, se $s = elefante$ então eft é subseqüência de s mas ftn não é. Mostre como encontrar a subseqüência comum mais longa entre duas cadeias s e t.

12 Dado um parâmetro δ, uma subcadeia t' de t é uma ocorrência aproximada de p em t se e somente se o alinhamento global entre p e t' tem pontuação mior ou igual a δ. O *problema do casamento aproximado de cadeias* consiste em encontrar todas as ocorrências aproximadas de p em t. Mostre como resolver esse problema.

13 [Gus97, 11-11] Since the traceback paths in a dynamic programming table correspond one-to-one with the optimal alignments, the number of distinct cooptimal alignments can be obtained by computing the number of distinct traceback paths. Give an algorithm to compute this number in $O(nm)$ time. Hint: Use dynamic programming.

14 [Gus97, 11-12] As discussed in the previous problem, the cooptimal alignments can be found by enumerating all the traceback paths in the dynamic programming table. Give a backtracking method to find each path, and each cooptimal alignment, in $O(n + m)$ time per path.

15 [Gus97, 11-13] In a dynamic programming table for edit distance, must the entries along a row be nondecreasing? What about down a column or down a diagonal of the table? Now discuss the same questions for optimal global alignment.

16 [Gus97, 11-30] Give a simple algorithm to solve the local alignment problem in $O(nm)$ time if no spaces are allowed in the local alignment.

17 [Gus97, 11-31] Local alignment between two different strings finds pairs of substrings from the two strings that have high similarity. It is also important to find substrings of a single string that have high similarity. Those substrings represent inexact repeated substrings. This suggests that to find inexact repeats in a single string one should locally align of a string against itself. But there is a problem with this approach. If we do local alignment of a string against itself, the best substring will be the entire string. Even using all the values in the table, the best path to a cell (i, j) for $i \neq j$ may be strongly influenced by the main diagonal. There is a simple fix to this problem. Find it. Can your method produce two substrings that overlap? Is that desirable?

18 [JP04, 6.32] A string x is called a supersequence of a string v if v is a subsequence of x. For example, ABLUE is a supersequence for BLUE and ABLE. Given strings v and w, devise an algorithm to find the shortest supersequence for both v and w.

19 Calcule a pontuação SP para o alinhamento abaixo usando a matriz BLOSUM62.

```
M-QPIALLLG
M-LR-ALL-G
M-K-IALLLG
MWPPVA--LG
```

20 Uma outra forma de calcular a pontuação de uma coluna em um alinhamento múltiplo é calculando a entropia. A entropia para uma coluna é

$$\sum_{x \in \mathcal{A}[i]} f_x \log_2 f_x,$$

onde f_x é a freqüência do símbolo x. A entropia é igual a 0 se todos os símbolos na coluna forem iguais e será maior se a diversidade de símbolos for maior. Calcule a pontuação por entropia do alinhamento abaixo.

```
M-QPIALLLG
M-LR-ALL-G
M-K-IALLLG
MWPPVA--LG
```

21 [Dia12, 1] Mostre como calcular o valor de soma de pares para uma coluna de um alinhamento múltiplo em tempo $O(k + |\mathcal{A}|)$ para um esquema de pontuação tal que $p(x, x) = M$ e $p(x, -) = g$, para todo símbolo x e $p(x, y) = m$, para todo símbolo $x \neq y$.

22 [SM97, adap. 3-14] How much space does algorithm for multiple sequence alignment below need apart from the pool of relevant cells?

MultiAlign$(s_1, \ldots, s_k, L)$

```
1  for every x and y, 1 ≤ x < y ≤ k
2       Compute C_xy, the total score array for s_x and s_y
3  for every x and y, 1 ≤ x < y ≤ k
4       L_xy = L − Σ_{i<j,(x,y)≠(i,j)} sim(s_i, s_j)
5  pool = {0}
6  while pool ≠ ∅
7       i = remove the lexicographically smallest cell in pool
8       if C_xy[i_x, i_y] ≥ L_xy, ∀x, y, 1 ≤ x < y ≤ k
9            for every j dependent on i
10                if j ∉ pool
11                     pool = pool ∪ j
12                     a[j] = a[i] + SP(col(s, i, j − i))
13                else
14                     a[j] = max(a[j], a[i] + SP(col(s, i, j − i))
15 return a[n_1, ..., n_k]
```

23 Construa o alinhamento estrela para as cadeias abaixo. Escolha o centro como a cadeia que maximiza o somatório das pontuações dos alinhamentos par-a-par com todas as demais.

```
AXXZ AXYXZ AXZ AYZ AYZYZ
```

24 [Gus97, 14-7] In the center-star method, the multiple alignment is constructed by successively aligning each new string to the center string S_c. However, the order that strings are aligned to S_c was not specified. Show that the same multiple alignment is built, no matter what order is used.

25 Faça o alinhamento global dos alinhamentos abaixo usando pontuação SP.

```
M-QPIL
M-LR-L

M--AL
MPPAL
```

26 Descreva os passos típicos de um alinhador múltiplo de seqüências progressivo.

27 Para as cadeias $s_1 = CGC, s_2 = GGATGC, s_3 = GCAGGT, s_4 = AGGG, s_5 = ATT$, construa um alinhamento progressivo usando a árvore $((s_1, s_3), ((s_2, s_4), s_5))$ como guia.

28 [Dia12, 2] Mostre um alinhamento estrela com pontuação pelo menos 1/3 maior que o melhor alinhamento múltiplo possível. Considere distância de edição para o alinhamento entre as seqüências e soma de pares (coluna a coluna) para a pontuação do alinhamento múltiplo.

29 [Dia12, 3] Considere um alinhamento múltiplo $\mathcal{A}$ de k seqüências, todas de tamanho n. Seja p o valor do alinhamento, considerando pontuação por soma de pares (coluna a coluna) e distância de edição para comparação entre os caracteres. E possível modificar o alinhamento múltiplo em tempo polinomial (em termos de k e n) e obter um alinhamento $\mathcal{A}'$, com pontuação p', tal que $p < p'$ (caso $\mathcal{A}$ não seja um alinhamento ótimo). Justifique sua resposta.

Referências

[Adi12] Said S. Adi. Lista de exercícios 1, 2012.

[Dia12] Zanoni Dias. Lista de exercícios 3, 2012.

[Gus97] D. Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology.* Cambridge University Press, 1997.

[JP04] N. C. Jones and P. A. Pevzner. *An Introduction To Bioinformatics Algorithms.* MIT Press, 2004.

[SM97] J. C. Setubal and J. Meidanis. *Introduction to computational molecular biology.* PWS Publishing Co., 1997.