Exercício Programa 1
Alinhamento Múltiplo de Sequências Biológicas

MAC0465

5 de outubro de 2018

1 Introdução

Primeiramente, vamos rever alguns conceitos de biologia, que servirão de base para a formulação do exercício proposto. O exercício propriamente dito está proposto nas próximas seções.

Existem três tipos principais de sequências biológicas: as sequências de DNA, RNA e as proteínas. Cada elemento da sequência de DNA é um dos quatro nucleotídeos: `A`, `C`, `T` ou `G`, enquanto que um elemento de uma sequência de RNA é um dos nucleotídeos `A`, `C`, `U` ou `G`. No casos das proteínas, são 20 os possíveis aminoácidos que aparecem nas sequências a formarem as proteínas.

Um alinhamento de sequências de DNA, de RNA ou proteínas de diferentes espécies de seres vivos descreve uma hipótese de evolução entre os nucleotídeos ou aminoácidos que constituem as sequências biológicas dos indivíduos de cada espécie. Os alinhamentos são importantes para inferir a história evolutiva das sequências. Um alinhamento múltiplo, envolvendo múltiplas sequências, é utilizado para extrair e representar similaridades biologicamente importantes de um conjunto de sequências, que por sua vez se tornam bem mais evidentes quando várias sequências são comparadas. Outra aplicação importante é a identificação de repetições, como é o caso de repetições Alu de aproximadamente 300 nucleotídeos que aparecem cerca de 600000 mil vezes no DNA humano.

Um alinhamento de k sequências biológicas é uma disposição matricial destas k sequências. Cada linha da matriz é formada pelos símbolos da sequência em questão, eventualmente com símbolos - (representando *espaços*, ou *gaps*) acrescentados entre símbolos consecutivos da sequência original. Colunas somente com *espaços* não são permitidas, o alinhamento com maior número de colunas possível é um em que em cada coluna há $k-1$ *espaços* e um único símbolo que relativo a uma das k sequências. Em particular, o número de colunas de um alinhamento não excede $\sum_i^k n_i$. Na Figura 1 vemos um exemplo de alinhamento de três proteínas (sequências de letras num alfabeto de 20 aminoácidos). Na primeira coluna do alinhamento, vemos alinhados os aminoácidos `I`, `I` e `V`, que descreve uma substituição do símbolo `I` na primeira e segunda sequências pelo símbolo `V` na terceira. Diversos alinhamentos são possíveis para um mesmo conjunto de sequências e a próxima seção descreve

```
IVNGEEAvpgSWPWQVSLQDktgF---HFCGGSLINENWVVTAAHCgvttSDVVVA-GEFdqgsssekiQKLKIA
IVGGYTCganTVPYQVSLN--SgY---HFCGGSLINSQWVVSAAHCyksGIQVRL--GEDninvvegneQFISAS
VVGGTEAqrnSWPSQISLQYrsgsswAHTCGGTLIRQNWVMTAAHCvdrelTFRVVVGEHnlnqnngteQYVGVQ

KVFKNSKYNSltinn--DITLLKLSTAASFSqTVSAVCLPsasddfaagTTCVTTGWGLTRytnantPDRLQQAS
KSIVHPSYNSntlnn--DIMLIKLKSAASLNsRVASISLPtsca--sagTQCLISGWGNTKssgtsyPDVLKCLK
KIVVHPYWNTddvaagyDIALLRLAQSVTLNsYVQLGVLPragtilannSPCYITGWGLTR-tngqlAQTLQQAY

LPLLSNtnckk--ywgtkikdaMICAG-asgV-SSCMGDSGGPLVCKkngaWTLVGIVSWGss-tcs-tstPGVY
APILSDsscks--aypgqitsnMFCAGyleggKDSCQGDSGGPVVC--SGK--LQGIVSWGs--gcaqknkPGVY
LPTVDYaicssssywgstvknsMVCAG-gngvRSGCQGDSGGPLHCLvngqYAVHGVTSFVsrlgcnvtrkPTVF

ARVTALVNWVQQTLAAN
TKVCNYVSWIKQTIASN
TRVSAYISWINNVIASN
```

Figura 1: Alinhamento das proteínas: chymotripsina bovina, tripsina bovina, e elastase suína. As 20 letras usadas representam os 20 aminoácidos possíveis. Letras maiúsculas foram usadas em colunas com maior pontuação. Hífens representam *espaços*, por exemplo devidos à remoção histórica de algum símbolo.

estratégias usadas para escolher um que seja melhor. Além disso, descreveremos implicitamente um modo alternativo para representar um alinhamento através de um conjunto de arestas ligando pontos de coordenadas inteiras, que serão melhor definidos a seguir.

2 Pontuação de um alinhamento

Vários alinhamentos (hipóteses evolutivas) podem ser propostos para um mesmo conjunto de sequências. Um problema que naturalmente se coloca é o de qual é o melhor alinhamento. O que vem a ser este melhor alinhamento A, está associado a uma pontuação que idealmente poderia ser $-\log p(A)$, onde $p(A)$ é a probabilidade de que o alinhamento A seja aquele que reflete as evoluções do ancestral comum às sequências em questão.

Para tanto, dá-se uma pontuação adequada a cada alinhamento múltiplo e procura-se obter um alinhamento de pontuação máxima. Em geral, se deseja e se calcula as pontuações de tal forma a associar uma pontuação maior a um alinhamento mais provável evolutivamente. São feitos levantamentos estatísticos para se definir um bom esquema de pontuação.

Costumeiramente são feitas várias hipóteses simplificadoras nas elaborações dos modelos matemáticos usados em biologia computacional. Por exemplo, comumente adota-se a hipótese de que a pontuação adequada de um alinhamento é a soma de pontuações das colunas e as pontuações das colunas de um alinhamento independem completamente da pontuação das demais colunas. Cada coluna c de um alinhamento é composta de k símbolos, sendo que na i-ésima linha desta coluna, temos um símbolo da i-ésima sequência sendo alinhada ou um *espaço*. Comumente adota-se o esquema de pontuação conhecido como *soma de pares*, em que se define a pontuação da coluna como a soma das pontuações dos $\frac{k(k-1)}{2}$ pares de símbolos tomados dentre os k símbolos da coluna. Para $k = 3$, temos três pares, correspondentes: às linhas 1 e 2; às linhas 2 e 3; e às linhas 1 e 3.

Adotamos a seguinte pontuação para cada par de símbolos:

1. o parâmetro r (*reward*) premia *símbolos idênticos diferentes de espaço* (`-`);
2. o parâmetro q pontua *símbolos diferentes entre si e diferentes de espaço*;
3. o parâmetro g, se exatamente um deles é *espaço/gap*;
4. 0, caso sejam *dois símbolos espaço* (`-`).

Por exemplo, dadas as sequências de nucleotídeos `ATC`, `CGGA` e `CACT`, um possível alinhamento múltiplo é o destacado na Figura 2. A cada coluna do alinhamento corresponde

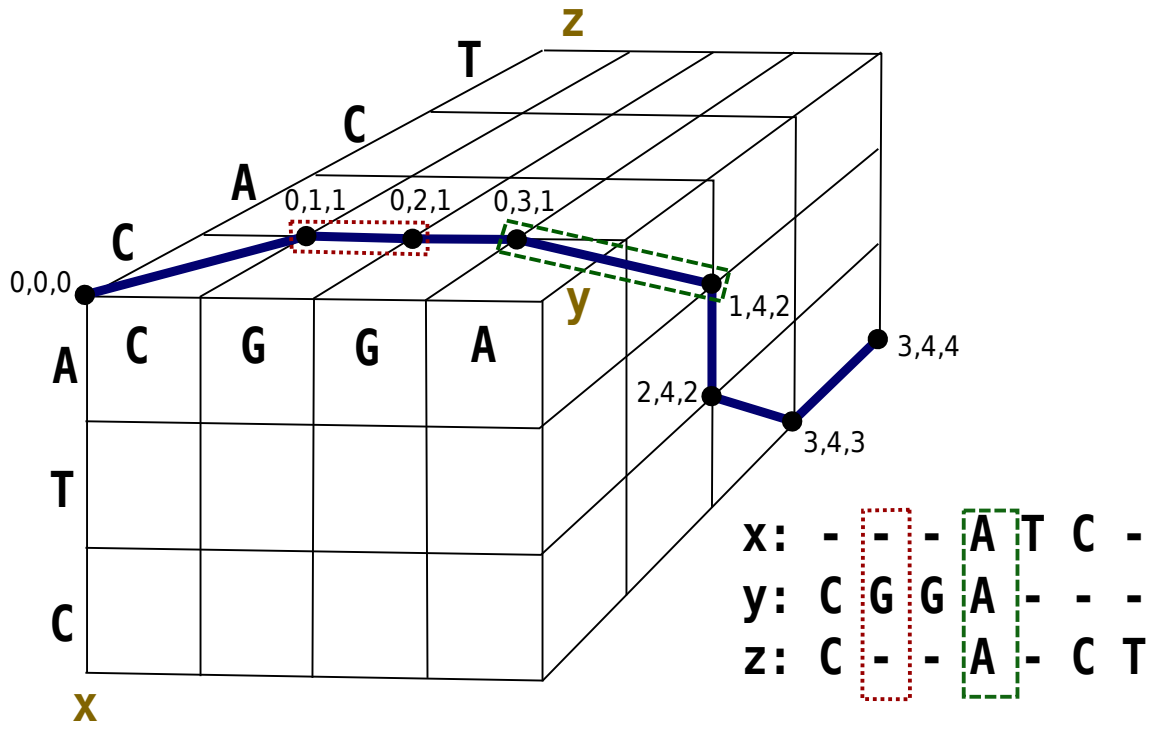

Figura 2: Um alinhamento múltiplo das sequências `ATC`, `CGGA` e `CACT` com pontuação $5r+12g$. Este alinhamento é ótimo para $r = 1$ e $q = 0 = g$.

uma aresta do caminho azul do vértice de coordenadas $(0, 0, 0)$ até o vértice de coordenadas $(3, 4, 4)$. (Observe que 3, 4 e 4 são os respectivos comprimentos das sequências x, y e z.) A aresta do caminho destacada por linha pontilhada vermelha, por exemplo, corresponde à coluna do alinhamento destacada por linha pontilhada vermelha. Nesta coluna estão presentes os símbolos `-`, `G` e `-`. Equivale a dizer que não é alinhado nenhum nucleotídeo das sequências `x` e `z`, denotado pelos *espaços* `-`, enquanto que é alinhado o nucleotídeo `G` da sequência `y`. Observe que somente a coordenada da sequência `y` sofreu alteração já que a aresta em questão parte do ponto $(0, 1, 1)$ para chegar ao ponto $(0, 2, 1)$. De fato, os símbolos presentes são muito bem definidos pelas coordenadas dos pontos de partida e chegada da aresta em questão: um *espaço* `-` é usado se e só se a coordenada em questão não muda.

De uma maneira geral, seja uma aresta que parte do ponto P e chega no ponto $Q = (Q_1, \ldots, Q_k)$. Associamos à aresta em questão uma coluna de um alinhamento da sequinte forma. Seja $1 \leq i \leq k$ e seja $j_i = Q_i$ a i-ésima coordenada do destino Q da aresta. O símbolo associado à sequência S_i na coluna do alinhamento é:

- o j_i-ésimo símbolo de S_i, se a i-ésima coordenada da origem P for $j_i - 1$;
- um espaço -, se a i-ésima coordenada do ponto de origem P for j_i.

A cada diferença possível $Q - P$ chamamos de *deslocamento*. Observe que o conjunto dos deslocamentos são os vértices no cubo k-dimensional de coordenadas 0 ou 1, exceto $(0, 0, \ldots, 0)$, como os cubos na Figura 3. Definimos o conjunto $\Delta = \{0, 1\}^k \setminus \{0\}^k$ como sendo este conjunto de diferenças possíveis. Observe que Δ tem $2^k - 1$ deslocamentos e $\Delta = \{(0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 0, 1), (1, 1, 0), (1, 1, 1)\}$ no caso em que $k = 3$. Na

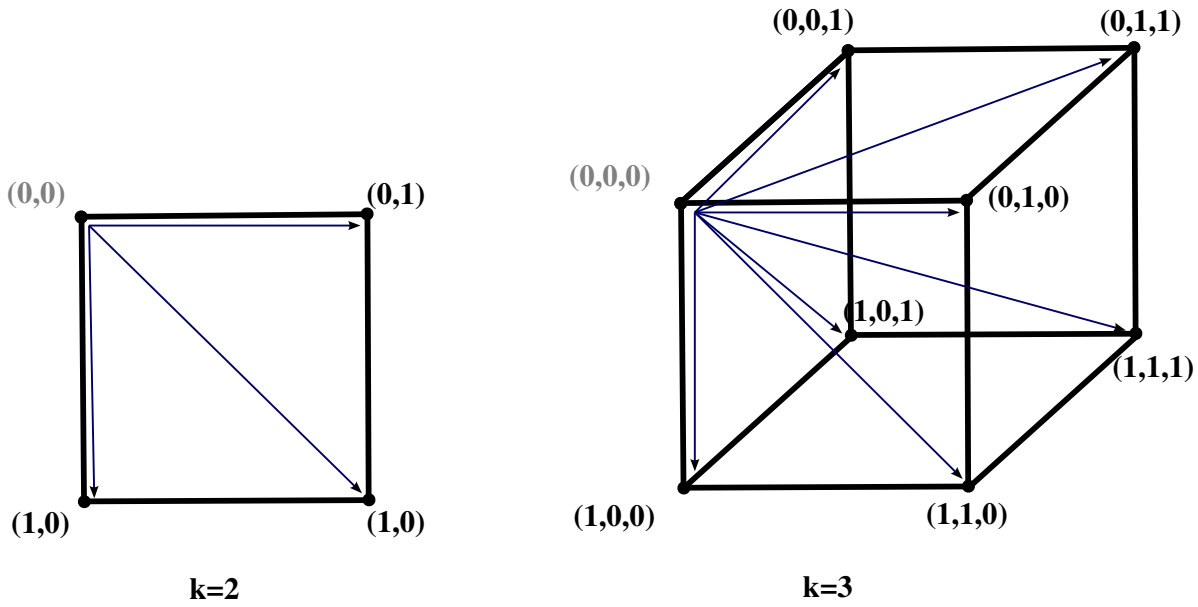

Figura 3: Cubos k-dimensionais correspondentes às diferenças entre pontos consecutivos associados a colunas do alinhamento para $k = 2$ e $k = 3$. Quando $k = 3$, temos $2^k - 1 = 7$ diferenças possíveis a formar o conjunto de diferenças possíveis Δ.

Figura 3, vemos exemplos para $k = 2$ sequências e para $k = 3$ sequências. A pontuação numa tal coluna associada a uma aresta do ponto P até o ponto Q é denotada por

$$W(P, Q).$$

Voltando ao exemplo da Figura 2, a pontuação da coluna pontilhada vermelha é a soma das pontuações associadas aos pares (-,G), (-,-) e (G,-), que dá $g + 0 + g = 2g$. Podemos fazer o mesmo para a aresta e coluna destacadas por linhas tracejadas azuis, que corresponde

ao alinhamento de três `A`'s. Neste caso todas as coordenadas são incrementadas de 1 quando se considera a origem $(0, 3, 1)$ e o destino $(1, 4, 2)$ da aresta. A pontuação da coluna é $3r$. Outro exemplo de alinhamento múltiplo para as mesmas sequências pode ser visto na Figura 4, onde as três colunas finais do alinhamento são diferentes, apesar da pontuação dos dois alinhamentos ser a mesma: $5r + 12g$.

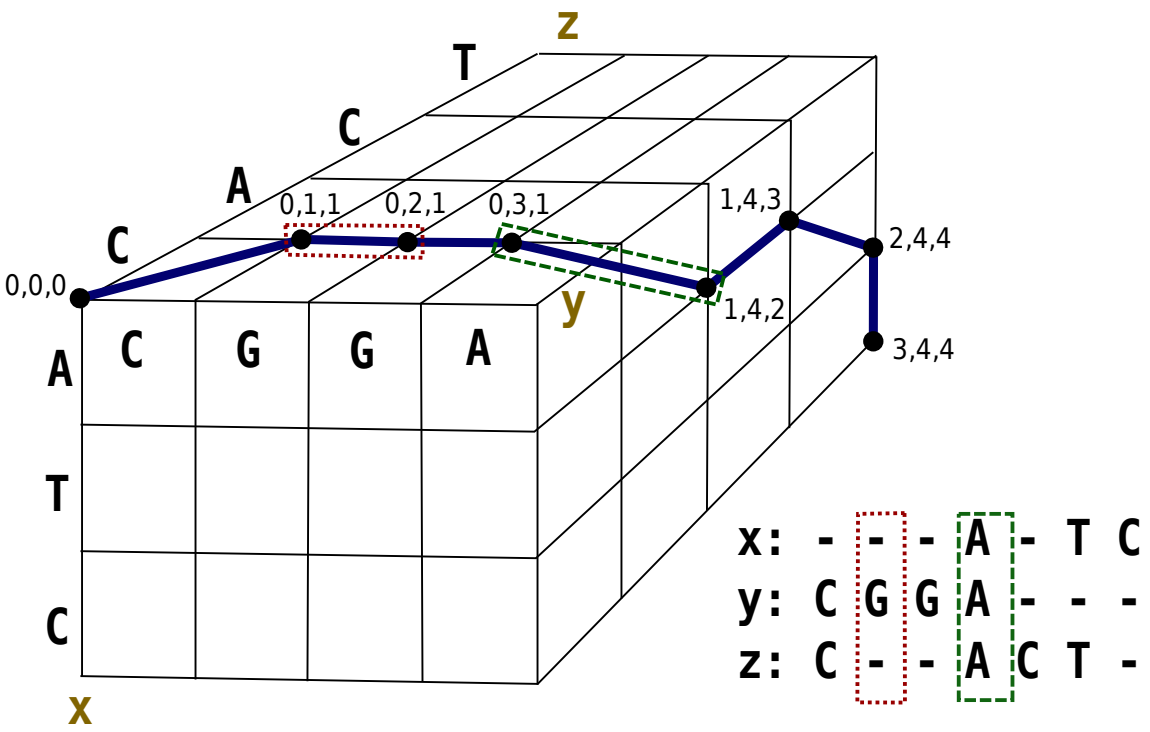

Figura 4: Um alinhamento alternativo para as sequências `ATC`, `CGGA` e `CACT`.

3 A formulação do problema

Desta forma, o problema de alinhamento múltiplo de sequências é especificado assim:

> São dados um inteiro $k > 0$ e k sequências S_1, S_2, ..., S_k. São dados parâmetros r, q e g que são pontuações para *identidade*, *difença* e *espaço*, como visto antes. Queremos obter *um* alinhamento múltiplo entre as k sequências com pontuação máxima, que também é chamado de *alinhamento ótimo*.

Para resolver o problema faremos uso de uma função recorrente cujos valores serão armazenados numa matriz. Para todo i tal que $1 \leq i \leq k$, seja n_i o comprimento da sequência S_i. Seja U o conjunto de pontos de coordenadas inteiras

$$U = \{0, 1, \ldots, n_1\} \times \{0, 1, \ldots, n_2\} \times \cdots \{0, 1, \ldots, n_k\}.$$

Seja $Q = (Q_1, Q_2, \ldots, Q_k) \in U$ um ponto neste espaço k-dimensional. Denotamos por

$$S_i[1..Q_i]$$

o prefixo de comprimento Q_i da sequência S_i. Quando $Q_i = 0$, o prefixo $S_i[1..Q_i]$ é o prefixo vazio. Denotamos por $S(Q)$ o conjuntos dos prefixos

$$S(Q) = \{S_1[1..Q_1], S_2[1..Q_2], \ldots, S_k[1..Q_k]\}$$

e definimos $p(Q)$ como sendo a pontuação de um alinhamento ótimo para as sequências de $S(Q)$.

Para nossa felicidade, ao se remover a última coluna de um alinhamento ótimo $A(Q)$, o alinhamento que permanece é também ótimo. Observe que quando removemos a última coluna de um alinhamento para $S(Q)$, obtemos um alinhamento para as sequências de $S(Q-\delta)$, onde $\delta = (\delta_1, \delta_2, \ldots, \delta_k)$ é tal que cada δ_i é 1 se a linha correspondente à sequência S_i na coluna era diferente de *espaço* e 0 caso contrário. Definimos $p(Q) = 0$ quando $Q = (0, 0, \ldots, 0)$ e não definimos a pontuação de nenhum ponto Q fora de U. De forma geral, para os demais pontos de U, a pontuação do alinhamento ótimo entre os prefixos em $S(Q)$ pode ser definida[1] de forma recorrente por

$$p(Q) = \max(\{p(P) + W(P, Q) \mid P = Q - \delta \in U \text{ para algum } \delta \in \Delta\}). \tag{1}$$

Naturalmente que no caso geral são $2^k - 1$ pontos $Q - \delta$, mas quando alguma coordenada de Q for nula, nem todos os pontos $Q - \delta$ têm valor $p(Q - \delta)$ definido pois alguns têm coordenadas negativas.

Observe que a pontuação máxima de um alinhamento múltiplo (ótimo) para as três sequências dadas é o valor de $p(Q)$, para $Q = (n_1, n_2, \ldots, n_k)$.

Voltemos ao exemplo da Figura 2. Temos $k = 3$ e as três sequências S_1, S_2 e S_3 definidas como respectivamente `ATC`, `CGGA` e `CACT`. Os valores de n_1, n_2 e n_3 por sua vez são 3, 4 e 4, respectivamente. A Tabela 1 lista o conjunto dos pontos de U, neste caso.

Como exemplo, tomemos o ponto $Q = (2, 4, 2)$, em negrito na Tabela 1. O conjunto de $S(Q)$ associado a este ponto é formado pelos prefixos `AT`, `CGGA`, `CA`. Observe que na própria Figura 2 temos um alinhamento para $S(Q)$ ao tomarmos o alinhamento formado pelas cinco primeiras colunas do alinhamento ali exibido. A última coluna deste alinhamento, a quinta coluna, é associada ao deslocamento $\delta = (1, 0, 0)$, de forma que o ponto $Q - \delta = (1, 4, 2)$, e um alinhamento para $S(Q - \delta)$ é o alinhamento obtido ao remover a quinta coluna e ficar apenas com as quatro primeiras colunas. (Caso tenhamos que o alinhamento da figura seja ótimo, aos alinhamentos de $S(Q)$ e $S(Q - \delta)$ descritos acima também deverão ser ótimos, como também deverá ser ótimo o alinhamento de qualquer ponto na trajetória de $(0, 0, 0)$ a $(3, 4, 4)$ que está na linha azul em destaque.) O conjunto de deslocamentos Δ para o caso em questão é aquele da Figura 3, caso $k = 3$. Ao se querer computar $p(Q)$ usando as

[1] Esta definição das recorrências (1) não funcionaria para $Q = (0, 0, \ldots, 0)$ pois o conjunto sob o qual se tomaria o máximo seria vazio. De fato, $(0, 0, \ldots, 0)$ é o único ponto de U em que o conjunto em questão é vazio.

(0,0,0)	(0,0,1)	(0,0,2)	(0,0,3)	(0,0,4)
(0,1,0)	(0,1,1)	(0,1,2)	(0,1,3)	(0,1,4)
(0,2,0)	(0,2,1)	(0,2,2)	(0,2,3)	(0,2,4)
(0,3,0)	(0,3,1)	(0,3,2)	(0,3,3)	(0,3,4)
(0,4,0)	(0,4,1)	(0,4,2)	(0,4,3)	(0,4,4)
(1,0,0)	(1,0,1)	(1,0,2)	(1,0,3)	(1,0,4)
(1,1,0)	(1,1,1)	(1,1,2)	(1,1,3)	(1,1,4)
(1,2,0)	(1,2,1)	(1,2,2)	(1,2,3)	(1,2,4)
(1,3,0)	(1,3,1)	(1,3,2)	(1,3,3)	(1,3,4)
(1,4,0)	(1,4,1)	(1,4,2)	(1,4,3)	(1,4,4)
(2,0,0)	(2,0,1)	(2,0,2)	(2,0,3)	(2,0,4)
(2,1,0)	(2,1,1)	(2,1,2)	(2,1,3)	(2,1,4)
(2,2,0)	(2,2,1)	(2,2,2)	(2,2,3)	(2,2,4)
(2,3,0)	(2,3,1)	(2,3,2)	(2,3,3)	(2,3,4)
(2,4,0)	(2,4,1)	**(2,4,2)**	(2,4,3)	(2,4,4)
(3,0,0)	(3,0,1)	(3,0,2)	(3,0,3)	(3,0,4)
(3,1,0)	(3,1,1)	(3,1,2)	(3,1,3)	(3,1,4)
(3,2,0)	(3,2,1)	(3,2,2)	(3,2,3)	(3,2,4)
(3,3,0)	(3,3,1)	(3,3,2)	(3,3,3)	(3,3,4)
(3,4,0)	(3,4,1)	(3,4,2)	(3,4,3)	(3,4,4)

Tabela 1: Conjunto U para o exemplo da Figura 2.

recorrências (1), somos levados a tomar o máximo sobre 7 valores possíveis, cada um deles em função de uma possível escolha de δ em Δ. Cada uma destas escolhas, define um ponto $Q-\delta$. As recorrências (1) recorrem à aplicação da função p em cada um dos pontos $Q-\delta$, que em nosso caso são: $(2,4,1)$, $(2,3,2)$, $(2,3,1)$, $(1,4,2)$, $(1,4,1)$, $(1,3,2)$ e $(1,3,1)$. (Na Tabela 1, estes 7 pontos $Q-\delta$ estão dentro de um retângulo, enquanto que o ponto Q está em negrito.) Para $\delta=(1,0,0)$, o ponto $Q-\delta$ é $(1,4,2)$ e contribui no cálculo do máximo com o valor

$$p((1,4,2))+W((1,4,2),(2,4,2)).$$

A coluna associada à aresta de $(1,4,2)$ a $(2,4,2)$ é

T
–
–

e que possui pontuação $g+g+0=2g$. Assim, $W((1,4,2),(2,4,2))=2g$.

4 Recorrendo a uma matriz

Para calcular a função recorrente $p: U \longrightarrow \mathbb{R}$ definida acima, a maneira mais apropriada ao conteúdo visto em aula é através do uso de uma matriz k-dimensional M de tamanho $(n_1+1)(n_2+1)\cdots(n_k+1)$. Para cada ponto $Q \in U$, armazenamos em $M[Q]$ o valor de $p(Q)$. Assim, usando as recorrências (1), podemos calcular os valores da Matriz M pela fórmula abaixo:

$$M[Q]=\max(\{M[P]+W(P,Q) \mid P=Q-\delta \in U \text{ para algum } \delta \in \Delta\}), \tag{2}$$

que toma o máximo num conjunto de no máximo 2^k-1 valores (um para cada possível $\delta \in \Delta$).

Através de k laços (loops) encaixados seria possível varrer todos os pontos de U, realizando os cálculos necessários e buscando os valores $M[P]$ vistos na própria matriz M, mas não é possível escrever um número variável k de laços ... Você deve planejar como contornar este problema!

Na Tabela 1, por exemplo, vemos os pontos de U listados numa ordem tal que, ao computarmos p em cada um destes pontos Q e armazenar na posição correspondente da matriz M, todos os valores $p(Q-\delta)$ necessários ao cálculo de $p(Q)$ já terão sido computados e armazenados na posição correspondente da matriz M.

O que deve ser entregue

O aluno deve entregar um programa escrito em C,Python, Java ou C++ que lê da linha de comando três inteiros r, q e g, um inteiro k e k strings S_1, S_2, ..., S_k correspondentes às k sequências em questão. O programa deve imprimir um alinhamento ótimo e sua pontuação.

O programa também deve imprimir um tal alinhamento num formato semelhante ao da Figura 2. Para o exemplo do alinhamento da Figura 2, os argumentos na linha de comando seriam:

```
1 0 0 ATC CGGA CACT
```

Teste com estes dados. Altere g para -1

```
1 0 -1 ATC CGGA CACT
```

e veja o que acontece. O mesmo alinhamento da Figura 2 tem pontuação -5, mas o alinhamento ótimo tem pontuação 0. Também teste com os seguintes prefixos de tRNA do aminoácido Leucina, anticódon CAA:

```
1 0 -1 GTCAGGATGGCCGAGTGGTCTAAGGCGCCAGACTCAAGT \
GTCAGGATGGCCGAGTGGTCTAAGGCGCCAGACTCAAGG \
GCCTCCTTAGTGCAGTAGGTAGCGCATCAGTCTCAAAA \
GTCAGGATGGCCGAGCAGTCTTAAGGCGCTGCGTTCAAAT
```

Depois de rodar por uns cinco minutos, o algoritmo deve exibir um alinhamento ótimo de pontuação 150.

5 Consistência de Dados e Diretrizes

O seu programa não precisa fazer consistência de dados. O programa pode supor que todos os dados lidos (opções, número, ...) estão corretos.

Todo exercício programa deve seguir as observações gerais normalmente adotadas nas disciplinas de ciência da computação, como por exemplo as contidas na página `http://www.ime.usp.br/~mac2166/infoeps`, onde estão descritas as diretrizes para forma de entrega do exercício, aspectos importantes na avaliação etc.