Bacharelado em Ciência da Computação

Programação Orientada a Objetos

Professores: Bernardo Copstein e Isabel H. Manssour

Trabalho Final 2017/2

Título:

Análise de Genes em Genomas

Objetivos:

- Construir um sistema para auxiliar a análise de Genes em Genomas armazenados em arquivos do tipo texto disponíveis em bancos de dados de genoma.
- Explorar os conceitos de programação orientada a objetos desenvolvidos ao longo do semestre, em especial a modelagem das abstrações envolvidas no problema e o uso das APIs de coleções disponíveis na linguagem Java, visando a obtenção de soluções otimizadas.

Contextualização:

Em biologia, o **genoma** é toda a informação hereditária de um organismo que está codificada em seu DNA. Isto inclui tanto os genes como as sequências não-codificadoras que são muito importantes para a regulação gênica, dentre outras funções.

Um gene é a unidade funcional da hereditariedade. Ele é composto de ácidos nucleicos. Existem dois tipos de ácidos nucleicos: ácido desoxirribonucleico ou DNA (fita dupla) e ácido ribonucleico ou RNA (fita simples). Os genes são os portadores de informações genéticas que proporcionam a diversidade dos seres vivos.

O DNA é formado por quatro nucleotídeos: Adenina (A), Citosina (C), Guanina (G) e Timina (T). O RNA é composto das bases Adenina (A), Citosina (C), Guanina (G) e Uracila (U). Portanto, em vez de T, o RNA possui a base U.

As sequencias de DNA/RNA são complementares de maneira que se conhecemos uma podemos deduzir a outra. Para tanto basta saber que A pareia com T e C pareia com G.

Um gene é uma sequência de bases nitrogenadas ou nucleotídeos distintos (DNA) que codificam uma determinada sequência de proteína. Proteínas são as máquinas moleculares que fazem todo ser vivo funcionar.

Por exemplo, a sequência que segue, no formato FASTA, codifica a sequência das quatro bases nitrogenadas de um gene (DNA) do bacilo de Koch (bactéria causadora da tuberculose):

>NC_000962.3:1674202-1675011 *Mycobacterium tuberculosis* H37Rv, complete genome

ATGACAGGACTGCTGGACGGCAAACGGATTCTGGTTAGCGGAATCATCACCGACTCGTCGATCGCGTTTCACATCG
CACGGGTAGCCCAGGAGCAGGGCGCCCAGCTGGTGCTCACCGGGTTCGACCGGCTGCGGCTGATTCAGCGCATCA
CCGACCGGCTGCCGGCAAAGGCCCCGCTGCTCGAACTCGACGTGCAAAACGAGGAGCACCTGGCCAGCTTGGCCG
GCCGGGTGACCGAGGCGATCGGGGCGGGCAACAAGCTCGACGGGGTGGTGCATTCGATTGGGTTCATGCCGCAG
ACCGGGATGGGCATCAACCCGTTCTTCGACGCGCCCTACGCGGATGTGTCCAAGGGCATCCACATCTCGGCGTATT
CGTATGCTTCGATGGCCAAGGCGCTGCTGCCGATCATGAACCCCGGAGGTTCCATCGTCGGCATGGACTTCGACCC
GAGCCGGGCGATGCCGGCCTACAACTGGATGACGGTCGCCAAGAGCGCGTTGGAGTCGGTCAACAGGTTCGTGG
CGCGCGAGGCCGGCAAGTACGGTGTGCGTTCGAATCTCGTTGCCGCAGGCCCTATCCGGACGCTGGCGATGAGTG
CGATCGTCGGCGGTGCGCTCGGCGAGGAGGCCGGCGCCCAGATCCAGCTGCTCGAGGAGGGCTGGGATCAGCGC
GCTCCGATCGGCTGGAACATGAAGGATGCGACGCCGGTCGCCAAGACGGTGTGCGCGCTGCTGTCTGACTGGCTG
CCGGCGACCACGGGTGACATCATCTACGCCGACGGCGGCGCGCACACCCAATTGCTCTAG

O formato FASTA é o principal da área de bioinformática. Ele é simples: uma linha identificadora que começa com o símbolo '**>**' seguido da identificação da sequência. A sequência corresponde basicamente a conjuntos das letras ATCG (os quatro nucleotídeos ou quatro bases que formam o DNA). A combinação dessas quatro bases em conjuntos de 3 (os códons) possibilitam 64 combinações diferentes. Dessas 64 combinações, porém, apenas 20 resultam em aminoácidos estáveis.

A análise da sequência de DNA em códons (grupos de 3 letras) permite identificar os aminoácidos que a sequência codifica. A figura 1 a seguir apresenta a tabela de codificação. Nesta tabela, a primeira coluna contém o nome do aminoácido, a segunda coluna contém a letra pela qual o mesmo é universalmente identificado e a terceira coluna contém os códons que o codificam (mais de um códon pode codificar o mesmo aminoácido).

A decodificação de uma sequência de DNA, porém, não é tão simples. O agrupamento dos códons pode ser feito a partir da primeira, da segunda ou da terceira letra. Também pode ser feito do início para o fim (sentido 5'-3') ou do final para o início (sentido 3'-5'). Desta forma, são geradas 6 sequências de aminoácidos diferentes. Para saber qual é a sequência correta é preciso identificar qual das sequencias apresenta a maior distância entre um aminoácido indicador de início do processo de codificação de uma proteína (Met) e um aminoácido indicador do final deste processo (Stop).

Amino Acid	SLC	DNA codons
Isoleucine	I	ATT, ATC, ATA
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	GTT, GTC, GTA, GTG
Phenylalanine	F	TTT, TTC
Methionine	M	ATG
Cysteine	C	TGT, TGC
Alanine	A	GCT, GCC, GCA, GCG
Glycine	G	GGT, GGC, GGA, GGG
Proline	P	CCT, CCC, CCA, CCG
Threonine	T	ACT, ACC, ACA, ACG
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TAT, TAC
Tryptophan	W	TGG
Glutamine	Q	CAA, CAG
Asparagine	N	AAT, AAC
Histidine	H	CAT, CAC
Glutamic acid	E	GAA, GAG
Aspartic acid	D	GAT, GAC
Lysine	K	AAA, AAG
Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons	Stop	TAA, TAG, TGA

Figura 1 – Relação entre os aminoácidos e os códons que os codificam.

Vamos analisar como exemplo a seguinte sequência hipotética:

GATGACAGGACTGCTGGACTAGAA

Existem 6 possibilidades de tradução para esta sequência:

	Agrupamento dos códons	Sequência de aminoácidos
5'-3' : 1	GAT GAC AGG ACT GCT GGA CTA GAA	D D K T A G L E
5'-3' : 2	**ATG ACA GGA CTG CTG GGA TGA AGA**	**Met T G R R G Stop R**
5'-3' : 3	TGA CAG GAC TGC TGG ACT AGA	Stop Q D C W T R
3'-5' : 1	AAG ATC AGG TCG TCA GGA CAG TAG	K I R C S G Q Stop
3'-5' : 2	AGA AGT AGG GTC GTC AGG ACA GTA	R S R V V R T V
3'-5' : 3	AGA TCA GGT CGT CAG GAC AGT	R S G R Q D S

A sequência que apresenta maior distância entre um "Met" e um "Stop" é a sequência da linha 2. Portanto, é esta sequência que deve ser considerada como a sequência correta de aminoácidos.

Obtendo as Sequências de Genes

Genomas de bacilos da tuberculose podem ser baixados da *National Library of Medicine* https://www.ncbi.nlm.nih.gov/pubmed/29028987 (veja a figura 2).

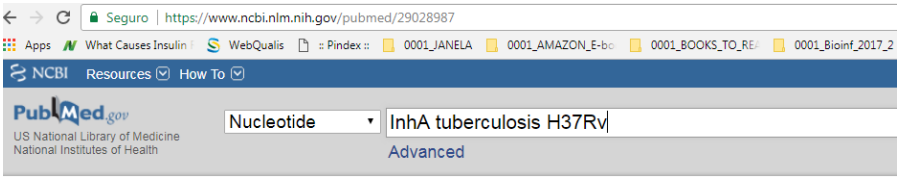

Figura 2 – Como obter genomas de bacilos da tuberculose.

O H37Rv é um bom ponto de partida. Para baixar as informações primeiro selecione um genoma. Se escolher o formato FASTA, o genoma inteiro será obtido, incluindo os genes e as sequências intermediárias de aminoácidos. Para obter apenas os genes selecione "*send to*" e depois "*gene sequences*", como ilustra a figura 3.

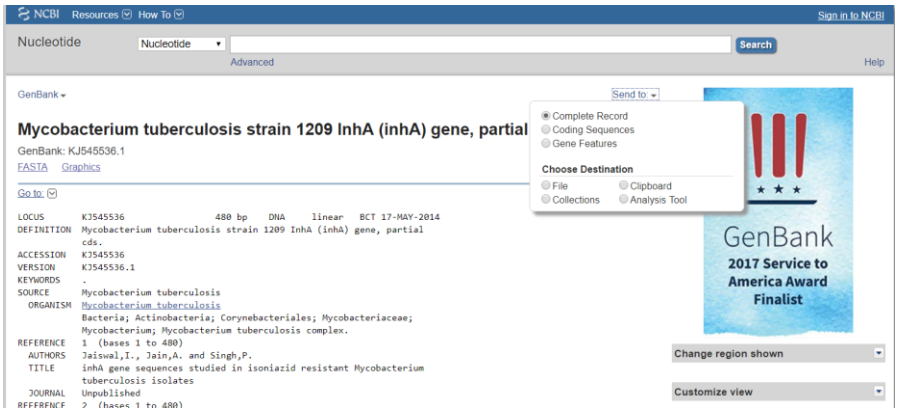

Figura 3 – Selecionando o arquivo correto.

Se tudo estiver correto, um arquivo como o extrato apresentado na figura 4 deve ser obtido.

As linhas que se iniciam por ">" identificam um gene. Várias informações sobre este gene estão anotadas entre "[" e "]". Neste trabalho irão nos interessar o "*locus*" (identificador do gene) e a "*location*" (identifica a posição de início e fim do gene dentro do genoma).

A linhas que seguem contêm a sequência de bases do gene. Esta sequência se encerra quando começa a identificação do próximo gene.

```
>lcl|AL123456.3_cds_CCP42723.1_1 [gene=dnaA] [locus_tag=Rv0001] [db_xref=EnsemblGenomes-Gn:Rv0001,EnsemblGenomes-
Tr:CCP42723,GOA:P9WNW3,InterPro:IPR001957,InterPro:IPR003593,InterPro:IPR010921,InterPro:IPR013159,InterPro:IPR013317,InterPro:IPR
018312,InterPro:IPR020591,InterPro:IPR027417] [protein=Chromosomal replication initiator protein DnaA] [protein_id=CCP42723.1]
[location=1..1524] [gbkey=CDS]
TTGACCGATGACCCCGGTTCAGGCTTCACCACAGTGTGGAACGCGGTCGTCTCCGAACTTAACGGCGACC
CTAAGGTTGACGACCGACCCAGCAGTGATGCTAATCTCAGCGCTCCGCTGACCCCTCAGCAAAGGGCTTG
GCTCAATCTCGTCCAGCCATTGACCATCGTCGAGGGGTTTGCTCTGTTATCCGTGCCGAGCAGCTTTCTC
CAAAACGAAATCGAGCGCCATCTGCGGGCCCCGATTACCGACGCTCTCAGCCGCCGACTCGGACATCAGA
TCCAACTCGGGGTCCGCATCGCTCCGCCGGCGACCGACGAAGCCGACGACACTACCGTGCCGCCTTCCGA
AAATCCTGCTACCACATCGCCAGACACCACAACCGACAACGACGAGATTGATGACAGCGCTGCGGCACGG
GGCGATAACCAGCACAGTTGGCCAAGTTACTTCACCGAGCGCCCGCACAATACCGATTCCGCTACCGCTG
GCGTAACCAGCCTTAACCGTCGCTACACCTTTGATACGTTCGTTATCGGCGCCTCCAACCGGTTCGCGCA
CGCCGCCGCCTTGGCGATCGCAGAAGCACCCGCCCGCGCTTACAACCCCCTGTTCATCTGGGGCGAGTCC
GGTCTCGGCAAGACACACCTGCTACACGCGGCAGGCAACTATGCCCAACGGTTGTTCCCGGGAATGCGGG
TCAAATATGTCTCCACCGAGGAATTCACCAACGACTTCATTAACTCGCTCCGCGATGACCGCAAGGTCGC
ATTCAAACGCAGCTACCGCGACGTAGACGTGCTGTTGGTCGACGACATCCAATTCATTGAAGGCAAAGAG
GGTATTCAAGAGGAGTTCTTCCACACCTTCAACACCTTGCACAATGCCAACAAGCAAATCGTCATCTCAT
CTGACCGCCCACCCAAGCAGCTCGCCACCCTCGAGGACCGGCTGAGAACCCGCTTTGAGTGGGGGCTGAT
CACTGACGTACAACCACCCGAGCTGGAGACCCGCATCGCCATCTTGCGCAAGAAAGCACAGATGGAACGG
CTCGCGGTCCCCGACGATGTCCTCGAACTCATCGCCAGCAGTATCGAACGCAATATCCGTGAACTCGAGG
GCGCGCTGATCCGGGTCACCGCGTTCGCCTCATTGAACAAAACACCAATCGACAAAGCGCTGGCCGAGAT
TGTGCTTCGCGATCTGATCGCCGACGCCAACACCATGCAAATCAGCGCGGCGACGATCATGGCTGCCACC
GCCGAATACTTCGACACTACCGTCGAAGAGCTTCGCGGGCCCGGCAAGACCCGAGCACTGGCCCAGTCAC
GACAGATTGCGATGTACCTGTGTCGTGAGCTCACCGATCTTTCGTTGCCCAAAATCGGCCAAGCGTTCGG
CCGTGATCACACAACCGTCATGTACGCCCAACGCAAGATCCTGTCCGAGATGGCCGAGCGCCGTGAGGTC
TTTGATCACGTCAAAGAACTCACCACTCGCATCCGTCAGCGCTCCAAGCGCTAG
```

Comentado [IH1]: Isto significa que o gene começa com TAG? Seria a primeira sequência "TAG" encontrada?

Comentado [IH2]: No texto aparece: "*location*" (identifica a posição de início e fim do gene dentro do genoma). Neste caso, quer dizer que o gene inicia no primeiro caractere abaixo do identificador e segue até o caractere que está na posição 1524?

Figura 4 – Extrato de um arquivo de sequência de genes.

Sistema a ser Desenvolvido

Deve-se projetar e desenvolver um sistema capaz de ler um arquivo de sequência de genes obtido junto a *National Library of Medicine* e exibir o seguinte conjunto de informações:

- *Locus* do gene;
- Posição de início e fim do gene dentro do genoma;
- Sequência de bases do gene (tal como lida do arquivo);
- Sequência de bases do gene agrupadas em códons (direção e deslocamento indicados);
- Sequência de aminoácidos correspondente à sequência de códons indicada;
- Sequência "correta" de aminoácidos;
- Representação gráfica da sequência de bases (atribuindo cores para as bases).

Entregáveis

- Diagrama de classes;
- Código fonte do sistema;
- Executável do sistema (capaz de executar fora do ambiente de um IDE).

Outras Informações

- O trabalho deve ser feito em dupla ou individualmente até a data especificada.
- Cada aluno ou dupla deverá entregar no Moodle um arquivo zip contendo a implementação feita (todas as pastas e os arquivos .java, arquivo com a modelagem e o executável do sistema). Este arquivo deve ter o nome e sobrenome do(s) aluno(s), da seguinte forma: nome_ultimosobrenome-nome_ultimosobrenome.zip. Deve ser feito o upload deste arquivo na tarefa indicada para isto no Moodle até a data e horário especificados.

- Os trabalhos não podem apresentar erros de compilação e as soluções de cada aluno ou dupla devem ser originais.