ANÁLISE DE *TANDEM REPEATS* CODIFICANTES EM GENOMAS BACTERIANOS

Vinícius A. SILVA[1]; Mozart de A. MARINS[2]

RESUMO

Pesquisas de bioinformática buscam implementar computacionalmente técnicas de biologia molecular envolvidas na compreensão do genoma dos organismos. O presente trabalho apresenta recursos para identificar padrões de repetições adjacentes (*tandem repeats*) em genomas bacterianos e também para verificar se os *tandem repeats* encontrados no genoma estão presentes em regiões codificantes, ou seja, regiões do genoma com requisitos para serem convertidas em proteínas.

INTRODUÇÃO

Após a descoberta de Watson e Crick em 1953 sobre o código genético e o fluxo da informação biológica, dos ácidos nucléicos (DNA) para as proteínas, tais códigos (A – Adenina; T – Timina; C – Citosina; G – Guanina) passaram a constituir os principais objetos de estudo de uma nova ciência, a Biologia Molecular. Logo surgiram métodos de sequenciamento do DNA, que permitiam a definição da sequência das bases do genoma dos organismos, inclusive de bactérias (OKURA, 2002).

Na segunda metade da década de 90, com a criação dos sequenciadores automáticos de DNA, houve uma explosão na quantidade de sequências identificadas de diversas espécies.

De mãos de tantas informações (arquivos contendo a sequência dos nucleotídeos – A, T, C e G do genoma) teve-se que criar bases para gerenciar e armazenar tais dados.

O NCBI (*National Center for Biotechnology Information*) fundado em 1988 tem informações genômicas cadastradas e possui várias aplicações que auxiliam na manipulação das mesmas. A base de dados criada para tal fim foi a GenBank,

[1] Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Passos. Passos/MG, email: vinicius.silva@ifsuldeminas.edu.br;
[2] Universidade de Ribeirão Preto – Câmpus Ribeirão Preto. Ribeirão Preto/SP, email: marins@gmb.bio.br;

banco com os sequenciamentos genéticos de acesso público (disponível no endereço eletrônico www.ncbi.nlm.nih.gov/genbank).

Além do armazenamento ocorria, paralelamente, a necessidade de análise desses dados, o que tornava indispensável a utilização de ferramentas computacionais eficientes para a interpretação dos resultados obtidos. Assim nascia a bioinformática. Essa nova ciência envolveria a união de diversas linhas de conhecimento - a engenharia de software, a matemática, a estatística, a ciência da computação e a biologia molecular. Os trabalhos de bioinformática buscam implementar computacionalmente técnicas e conceitos da biologia molecular para evidenciar alvos no DNA. Por exemplo, a busca de genes que possam estar ligados à patogenia de uma bactéria.

Neste sentido, o objetivo do presente trabalho é utilizar/desenvolver recursos computacionais para correlacionar as técnicas de *tandem repeats* e ORFs (*Open Reading Frames)* em busca de alvos para desenvolvimento de drogas. Em um segundo momento, os dados destacados pelo sistema serão analisados (*in vitro*) no laboratório de Biotecnologia da UNAERP - Universidade de Ribeirão Preto.

Tandem Repeats (TRs) são sequências de DNA que se repetem como múltiplas cópias adjacentes com a mesma orientação na cadeia de DNA, com elementos intervenientes ou não (WATSON, 2009). Estas repetições são classificadas em microssatélites ou minissatélites, sendo os microssatélites compostos por repeats de tamanho 2-10 nucleotídeos e os minissatélites compostos por regiões de 11-200 nucleotídeos. Por exemplo, a sequência AACGTT aparece cinco vezes no TR AACGTT AAC**C**GTT AACGTT**T** A_CGTT AACGTT. Vale ressaltar que as repetições podem ser idênticas ou parcialmente degeneradas. Em bactérias, vários trabalhos (WATSON, 2009) comprovam que os TRs estão presentes em genes relacionados com a patogenicidade da bactéria.

Já os ORFs estão associados aos genes do DNA. Simplificadamente, pode-se admitir que um gene seja um segmento de DNA que codifica uma certa proteína (WATSON, 2009). Em bioinformática, esses segmentos codificantes são chamados de “Quadros Abertos de Leitura” ou ORFs (do inglês, *Open Reading Frames*).

Desta forma, é possível determinar as sequências de nucleotídeos em uma molécula de DNA que tem potencial para codificar uma proteína, ou seja, as partes codificantes do DNA. O conjunto de sequencias ORFs de um genoma é comumente chamado de ORFioma.

Para identificar se os TRs encontrados no genoma das bactérias estão em regiões codificantes, ou seja, se os TR estão dentro do ORFioma da bactéria, será utilizada a técnica de alinhamento de sequências. Em Bioinformática, um alinhamento de sequências é uma forma de organizar sequências DNA para identificar regiões similares entre elas. Existem dois tipos de alinhamento: global e local.

No alinhamento global, as sequências são alinhadas em sua totalidade. O alinhamento local consiste no alinhamento de apenas parte das sequências envolvidas. É feita uma procura por regiões com semelhança local e não é considerada a sequência em todo o seu comprimento. O algoritmo de Smith-Waterman é usado no âmbito do alinhamento local de pares de sequências e será utilizado no trabalho.

MATERIAL E MÉTODOS

Os dados para a análise genômica proposta consistem em dois arquivos: um contendo o genoma completo da bactéria e outro com os ORFs do genoma da bactéria. Tais arquivos podem ser adquiridos em bancos de dados de acesso público, como o já citado GenBank.

O genoma completo é armazenado em um arquivo formato fasta. Nesse tipo de arquivo a linha iniciada com o caractere '>' é utilizada para descrever os dados do genoma em questão, como o nome da bactéria e os códigos de acesso do genoma no banco de dados. As demais linhas possuem a sequência de nucleotídeos que compõem o genoma completo, como pode ser visto na figura a seguir.

```
>gi|72393774|gb|CP000107.1| Ehrlichia canis str. Jake, complete genome
TATGTTTTACTGTATTGTATTTTGCTTTTATAAATAACAACTTATTGGAATTTCTTATTGGAATTT
TTAATAATTAACTTTTCATTTATAAAATATTAAAAAATTTACTATACGTAATTACTTATTGGGGGAA
....
```

Figura 1. Exemplo arquivo fasta

Já os ORFs estão distribuídos em um arquivo formato ffn (*Fasta nucleotide coding regions file*). Esse tipo de arquivo possui somente as regiões do genoma que podem ser traduzidas em proteínas, ou seja, possui as regiões codificantes. Cada ORF é indicada no arquivo por uma linha iniciado pelo caractere '>' seguido por dados da base de dados e nas demais linhas aparecem as sequências de

nucleotídeos. A figura a seguir mostra parte do arquivo com os ORFs da bactéria Ehrlichia Canis.

```
>gi|73666633|ref|NC_007354.1|:475-1647 Ehrlichia canis str. Jake chromosome
ATGAACAGTAGCAAATATTATGATGTAGTTATTTCAGGAGCAGGAACGAATGGATGCACAGTGTGT
TGCCAAAAGATAGAGTATTTGCTTTATCAAGGAGGTCTCAAGAAATACTAGACACACACACCTGTG
>gi|73666633|ref|NC_007354.1|:c2204-1644 Ehrlichia canis str. Jake chromosome
ATGCTTTTTTACAGGTACATGTTATTCTTATTTTTATTAAAAAATATAAAGAACCTCCGCGGCCGTTT
GTTGTTAGGGTGTTTTTTCTTGCGTTATTTCTTTATTTTTCTTCCTTTATTTTTATTCACACACAACG
...
```

Figura 2. Exemplo arquivo ffn com ORFs

Para análise do presente trabalho foram utilizados os dados genômicos de 3 bactérias da família *Anaplasmataceae* (*Ehrlichia Canis, Ehrlichia Chaffeensis, Anaplasma Phagovctophilum*). Os arquivos com o genoma e os ORFs foram adquiridos por meio de acesso à base de dados GenBank, disponível pelo endereço eletrônico http://www.ncbi.nlm.nih.gov.

Para identificação dos TR, foram submetidos ao software SERV (LEGENDRE, 2007) o genoma das três bactérias pesquisadas.

Foram desenvolvidos métodos utilizando a linguagem Java para gravar em um banco de dados (Figura 3) MySql o genoma, os ORFs e TRs encontrados nos genomas pelo software SERV.

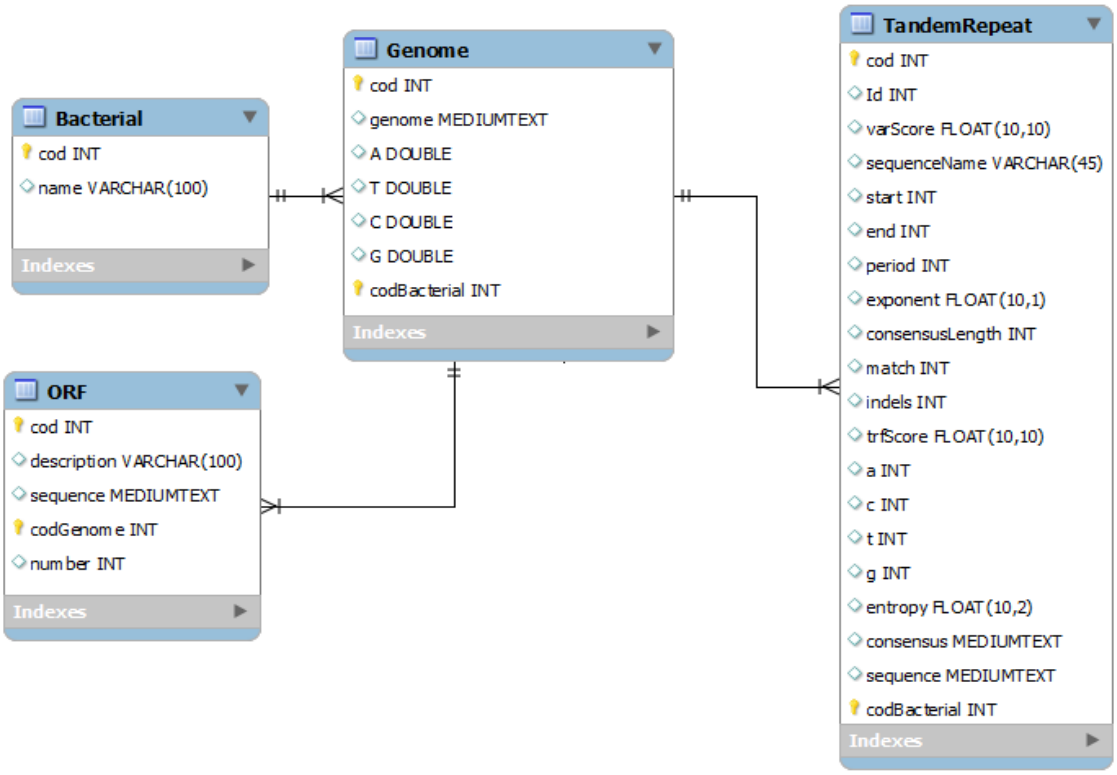

Figura 3. Diagrama E-R para armazenamento dos dados

Para identificar os TR codificantes foram utilizados métodos da biblioteca BioJava. O projeto BioJava (http://biojava.org) é um projeto de código aberto dedicado a fornecer ferramentas Java para o processamento de dados biológicos, como o alinhamento de sequências.

RESULTADOS E DISCUSSÃO

Foram realizadas análises no genoma e no ORFioma das três bactérias citadas. A tabela a seguir mostra o número de nucleotídeos que compõe o DNA.

Genoma (tamanho em nucleotídeos)		
E. canis	***E. chaffeensis***	***A. phagoyctophilum***
1.315.030	1.176.248	1.471.282

Tabela 1 - Número de nucleotídeos no genoma

A tabela 2 mostra o número de ORFs (regiões codificantes) que foram inseridos no banco de dados.

Open Reading Frames		
E. canis	***E. chaffeensis***	***A. phagoyctophilum***
925	1105	1264

Tabela 2 - Número de ORFs presentes no genoma

A tabela a seguir exibe o número de TR encontrados em cada um dos genomas analisados. Utilizando funcionalidades da biblioteca BioJava, cada um dos TR foram alinhandos com todos os ORFs do genoma da bactéira para verificar se o TR estava presente em uma região codificante do genoma. Se o TR tiver um alinhamento de 90% em uma ORF, esse TR foi considerado como um TR presente em uma região codificante.

Tandem Repeats		
Bactérias	**TRs identificados**	**TRs codificantes**
E. canis	901	454
E. chaffeensis	740	322
A. phagoyctophilum	245	96

Tabela 3 - Número de Tandem Repeats encontrados utilizando o software SERV (LEGENDRE, 2007)

CONCLUSÕES

A Bioinformática não substitui a pesquisa de bancada dos biólogos. Os bioinformatas utilizam o conhecimento prévio e fazem comparações e análises

específicas utilizando ferramentas computacionais, com velocidade e capacidade muito superiores às análises feitas manualmente. Sendo assim, vale ressaltar que o presente projeto está sendo realizado em parceria com a UNAERP – Universidade de Ribeirão Preto. Os dados evidenciados pelo software serão testados no laboratório de Biotecnologia da Universidade para verificar se os TR codificantes localizados estão vinculados à patogenia das bactérias, gerando dados para o desenvolvimento de novas drogas e vacinas.

REFERÊNCIAS BIBLIOGRÁFICAS

LEGENDRE M, et. al. Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Research, 2007.

OKURA, Vagner Katsumi. Bioinformática de projetos Genoma de Bactérias. Instituto de Computação – UNICAMP. Dissertação de mestrado, fevereiro de 2002.

SEIBEL, Luiz F. B; LEMOS, Melissa; LIFSCHITZ, Sérgio. Banco de dados de Genoma. Departamento de Informática – PUC-RIO. Artigo técnico, maio de 2008.

WATSON James. D., et. al. DNA Recombinante: Genes e Genomas. 3 ed. Porto Alegre: Artmed, 2009.