Pipeline para identificação e armazenamento de repetições adjacentes

Matheus Eloy Franco[1], Mozart de Azevedo Marins[2], Ana Lucia Fachin[2]

[1]Área de Computação - Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais – Câmpus Machado - MG

[2]Departamento de Biotecnologia – Universidade de Ribeirão Preto - SP (UNAERP)

matheus.franco@ifsuldeminas.edu.br, {mmarins,afachin}@unaerp.br

Abstract. *Historically tandem repeats has been considered as nonfunctional DNA ("junk"), largely due to the fact there is no correlation between the content and the complexity of an organism. Different works have shown that these repeats occur in coding and promoter regions is not random: genes containing tandem repeats are enriched for specific functional classes. This study aimed to develop a pipeline for identification and storage tandem repeats.*

Resumo. *Historicamente repetições adjacentes tem sido consideradas como DNA não funcional ("lixo"), em grande parte devido ao fato de não haver correlação entre seu conteúdo a complexidade e de um organismo. Diferentes trabalhos tem demonstrado que a ocorrência destas repetições em regiões codificantes e promotoras não é aleatória: genes que contêm repetições em série são enriquecidas em classes funcionais específicas. Este trabalho teve como objetivo desenvolver um pipeline para identificação e armazenamento de repetições adjacentes.*

1. Introdução

Repetições adjacentes ou *tandem repeats* são sequências de DNA repetitivas que não ocorrem apenas sequencialmente, mas também diretamente adjacente. Estas sequências de DNA possuem algumas características básicas que são o tamanho da unidade de repetição, a quantidade de vezes que esta unidade se repete em um determinada região do gene e a pureza (perfeição) da repetição. Estas repetição são classificadas em microssatélites caso o tamanho da unidade de repetição seja menor que 10-pb ou minissatélites caso a unidade seja maior ou igual a 10-pb (RICHARD; KERREST; DUJON, 2008). A figura 1 apresenta as características gerais sobre repetições adjacentes de acordo com Boeynaems (2012).

Estas repetições são comuns em organismos eucariotos, o genoma humano por exemplo, contém aproximadamente 260.000 repetições que consiste em cerca de 7% do genoma (RICHARD; KERREST; DUJON, 2008). Análises já realizadas demonstram que mais de 30% dos genes no genoma humano contêm repetições em regiões codificadoras (exons) (LEGENDRE et al., 2007).

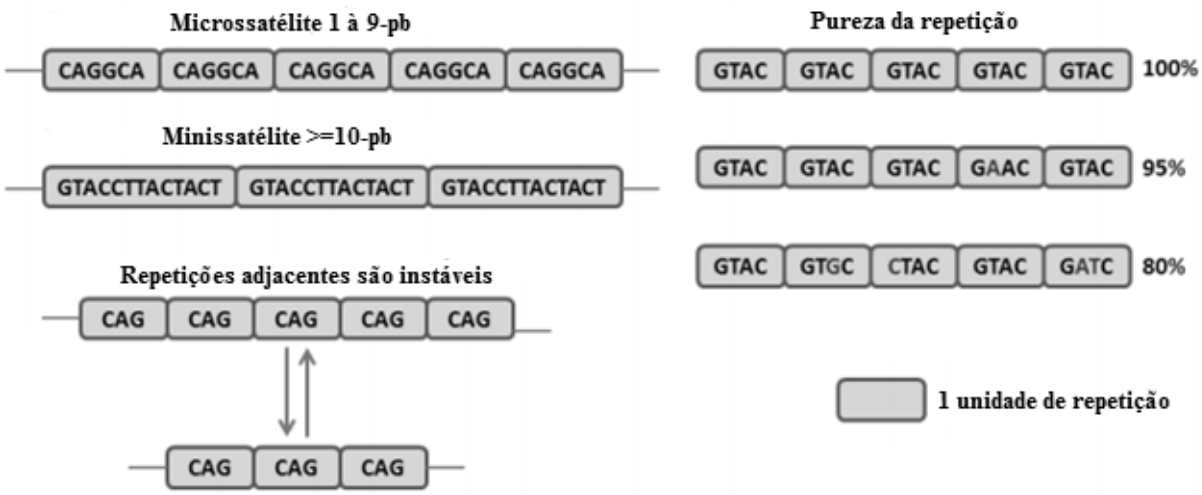

Figura 1: Características gerais das repetições adjacentes
Fonte: Adaptado de Boeynaems (2012)

A identificação destas repetições são importantes pois, no trabalho de Verstrepen et al. (2005) os autores observaram que entre 50% e 60% dos genes de fungos que contém repetições de minissatélites codificam proteínas de parede e adesão celular.

Repetições adjacentes não são encontrados em todos os genes, existindo uma tendência para estas serem encontradas em genes que respondem às mudanças das condições ambientais, sendo que algumas dessas repetições em série podem atuar como mecanismos de ajuste ao ambiente através de alterações fenotípicas (GEMAYEL et al., 2012).

2. Materiais e Métodos

Neste trabalho realizou-se a implementação de um *pipeline* baseado na internet para identificação e armazenamento de repetições adjacentes em sequencias de DNA. Como algoritmo de identificação de repetições adjacentes utilizou-se *Tandem Repeat Finder* (BENSON, 1999) com os seguintes parâmetros: *matching weight 2, mismatching penalty 5, indel penalty 5, match probability 0.8, indel probability 0.1, score ≥40 e maximum period 500*. Através dos parâmetros definidos, a ferramenta permite a identificação de repetições perfeitas e imperfeitas.

Como tecnologia de desenvolvimento para internet utilizou-se ASP.NET *Web Forms* por meio da linguagem de programação C#. Para o armazenamento das repetições identificadas utilizou-se o sistema de gerenciamento de banco de dados relacional MySQL 5.2. Para manipulação das sequências de DNA utilizou-se a biblioteca .NET Bio (.NET BIO, 2014).

A execução do algoritmo de identificação de repetições ocorre em um servidor Linux. Para comunicação entre o sistema web (*pipeline*) e o algoritmo para busca de repetições foi utilizado o serviço de SSH . O diagrama apresentado na figura 2 ilustra o fluxo de execução do *pipeline* desenvolvido.

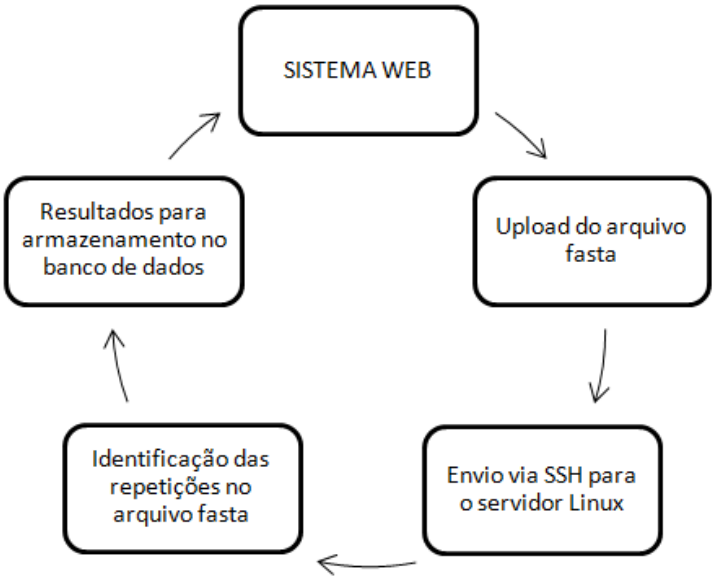

Figura 2: Envio e processamento das repetições.

3. Resultados e discussão

O *pipeline* desenvolvido permite identificar, armazenar e consultar repetições adjacentes a partir de um arquivo multi-fasta com sequências de um determinado organismo. O sistema desenvolvido está hospedado na intranet de nossa instituição estando acessível para consultas através da url http://goo.gl/MWr0yu.

A figura 3A apresenta a interface de envio e armazenamento das repetições. Na interface o usuário deve definir o organismo que receberá as sequências e opcionalmente definir uma descrição para este arquivo, o que possibilita a realização consultas com filtros de posteriormente. Após o processamento das sequências, em um segundo passo é possível realizar o download do arquivo resultante, armazenar os repeats no banco de dados e visualizar o registro de gravação. Após o armazenamento das repetições é possível consultar por um determinado padrão de repetição ou traçar o perfil de repetições existentes em um dados organismo (Figura 3B).

Figura 3: Telas do sistema

A identificação e caracterização destas sequências repetitivas tornou-se uma parte importante de projetos de sequênciamento de genomas, pois este fenômeno contribui significativamente para a variação genética e patogênica entre os organismos.

Porém, apesar da sua prevalência em regiões funcionais do genoma e sua associação com várias doenças neurológicas humanas, e apesar da sua utilidade como marcadores genéticos, estudos sobre repetições em tandem permanecem escassos (DUITAMA et al., 2014). Desta maneira, torna-se importante a possibilidade de identificação e armazenamento destas repetições de forma automatizada.

4. Conclusões

A identificação de repetições adjacentes é uma importante etapa em estudos genômicos. Neste trabalho foi apresentado um sistema web (*pipeline*) que realiza a identificação e armazenamento de repetições adjacentes de forma automatizada, dispensando conhecimentos sobre a execução do algoritmo em um ambiente Linux.

A partir do *pipeline*, é possível consultar padrões de repetições e traçar o seu perfil em determinados genes com objetivo de melhor conhecer as bases moleculares relacionadas a patogenicidade em diferentes organismos, pois genes que contém estas repetições estão associadas a variação fenotípica, incluindo fatores de virulência em organismos patogênicos.

Agradecimentos

Ao Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas - IFSULDEMINAS e a Universidade de Ribeirão Preto - UNAERP pelo apoio no desenvolvimento deste trabalho.

Referências

.NET BIO Framework. Disponível em:<http://bio.codeplex.com/.> Acesso em 06 de Abr. 2014.

BENSON, G. Tandem repeats Finder: a program to analyze DNA sequences., v. 27, p. 573–580, 1999.

BOEYNAEMS, S. Functional Consequences of Variable Tandem Repeats within the Yeast Cyc8 Transcriptional Regulator. Dissertação (Mestrado em Bioengenharia), Katholieke Universiteit Leuven. Leuven, p. 112. 2012.

DUITAMA, J. et al. Large-scale analysis of tandem repeat variability in the human genome. Nucleic Acids Research, v. 1, 2014.

GEMAYEL, R. et al. Beyond Junk-Variable Tandem Repeats as Facilitators of Rapid Evolution of Regulatory and Coding Sequences. Genes, n. 3, p. 461-480, 2012.

LEGENDRE, M. et al. Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Research, v. 17, p. 1787–1796, 2007.

RICHARD, G.-F.; KERREST, A.; DUJON, B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev, v. 72, p. 686–727, 2008.

VERSTREPEN, K. et al. Intragenic tandem repeats generate functional variability. Nat Genet., v. 37, p. 986–990., 2005.