Uma Proposta de Arquitetura Baseada em Algoritmos Inteligentes para a Descoberta de Motifs em Expressões Genéticas

Augusto Garcia Schmidt[1]; Marilton Sanchotene de Aguiar[2]

[1]*Universidade Federal de Pelotas – augustgs@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – marilton@inf.ufpel.edu.br*

1. INTRODUÇÃO

É amplamente reconhecido que a biologia está entre as áreas de pesquisa de maior acúmulo de informações nas últimas décadas. Esse acúmulo foi devido a uma série de avanços tecnológicos nos últimos anos que ocasionaram maior qualidade e quantidade de informação coletada de organismos no nível genômico, transcriptômico e proteômico (PROSDOSCIMI, 2007).

No entanto, à medida que essas informações são coletadas em grande quantidade, as técnicas usadas para análise destas informações devem se tornar mais sofisticadas atendendo a esta demanda em grande escala sujeito a ruídos. Neste contexto, faz-se necessário um estudo sobre algoritmos inteligentes em conjunto com técnicas de bioinformática a serem utilizadas na área de busca e aprendizado de padrões em expressões genéticas de bactérias (K. DAVIES, 2001).

A informação contida no DNA é formada com um alfabeto de quatro letras que correspondem aos quatro nucleotídeos: A, C, T, G. Com essas quatro letras é possível formar aminoácidos. Ao serem combinadas para formar alguma das vinte abreviações utilizadas pelos vinte aminoácidos diferentes que se encontram nas expressões genéticas dos seres vivos (K. DAVIES, 2001). Na genética, um códon é uma sequência de três nucleotídeos de RNA mensageiro que formam um determinado aminoácido ou que indicam o ponto de início ou fim de tradução de RNA mensageiro. Isto mostra que cada conjunto de três nucleotídeos é responsável pela formação de um aminoácido, portanto as expressões genéticas se expressam por trincas de bases, que foram denominadas códons. Motifs são entidades não randômicas encontradas em cadeias de DNA. Um padrão também pode ser definido como um fenômeno não único.

Já motifs, além de possuírem padrões recorrentes na sequência analisada, também possuem uma função biológica. Pode-se definir motifs como um curto segmento compartilhado por múltiplas sequências de DNA que pode conter informações sobre evolução, estrutura ou função (YI-PING PHOEBE CHEN, 2005).

Portanto, é necessária a utilização de algoritmos inteligentes para melhorar o desempenho de ferramentas consolidadas, mas com limitações de desempenho, de uma forma que possa obter melhores resultados e um melhor tempo de execução para um conjunto de ferramentas.

Neste contexto, a implementação desta arquitetura tem o objetivo de melhorar a tecnologia atualmente disponível para os pesquisadores da bioinformática, tornando cada vez mais satisfatórios os resultados obtidos.

2. METODOLOGIA

Este trabalho tem a premissa de fazer um estudo aprofundado dos conceitos e ideias básicas de bioinformática mostrando como são analisados os dados da perspectiva de um profissional da área de biologia.

Serão estudadas as principais técnicas e conceitos de sistemas inteligentes para busca de padrões em expressões genéticas e seus respectivos algoritmos. Como modelos ocultos de markov que possuem a capacidade de calcular a probabilidade de eventos futuros e sequencias com base em acontecimentos passados e algoritmos genéticos que possuem maneiras eficientes de buscar problemas de larga escala como é uma expressão genética (BARRETT, STEVEN J, 2006).

Também deve ser feito um estudo aprofundado de algoritmos de bioinformática amplamente utilizados hoje em dia na busca de padrões em expressões gênicas, investigando suas utilizações, Também deve ser feito um estudo de técnicas de refinamento dos dados encontrados previamente pelos algoritmos citados, como as técnicas de alinhamento local, matrizes de comparação, tratamento de gaps e sobreposição de motifs. (PROSDOCIMI et al, 2002).

A primeira etapa é a utilização do Algoritmo Genético, que utiliza um arquivo contendo as sequencias genéticas necessárias para a criação dos possíveis Motifs aleatórios.

Após a utilização do Algoritmo Genético, é utilizada a ferramenta BLASTP (ALTSCHUL, STEPHEN F., et al. 1990). para realizar o procedimento de alinhamento local utilizando como base de dados o resultado obtido na etapa anterior. O formato escolhido para armazenar as informações foi o .XML, devido a facilidade de manipulação e organização das informações, tornando possível a utilização de *tags* para retirar as informações necessárias para a próxima etapa de execução.

Depois desta etapa é necessário converter o arquivo gerado (.XML) para um formato utilizado pelas ferramentas de bioinformática(.FASTA), nesta etapa também é realizado um filtro entre os possíveis Motifs para realizar uma eliminação de Motifs incorretos.

A próxima etapa é a utilização do CD-Hit uma ferramenta específica de refinamento, para garantir que somente informações válidas sejam utilizadas nas próximas etapas (LI, WEIZHONG, ADAM GODZIK. 2006). O filtro utilizado nesta ferramenta foi de 70% em relação ao valor de acertos encontrados nos possíveis Motifs.

Com informações menos redundantes e errôneas, é utilizada a ferramenta MUSCLE para realizar o procedimento de alinhamento múltiplo, realizando análises sobre cada sequência encontrada (EDGAR, RC. 2004).

A última etapa de execução é a utilização da ferramenta HMMER, que cria um modelo com as informações anteriores. Esta ferramenta é necessária, pois torna possível analisar os resultados de forma eficiente e padronizada, onde cada campo de informação do arquivo anterior contém uma coluna com seus respectivos resultados (FINN RD, et al. 2011).

Para fins de validação dos resultados alcançados com este trabalho são utilizadas sequências com motifs já anotados e serão realizados workshops com pesquisadores do grupo de pesquisa de bioinformática do curso de biotecnologia da Universidade Federal de Pelotas.

3. RESULTADOS E DISCUSSÃO

O tempo de execução de todas as etapas contidas na arquitetura foi de 2 minutos e 51.956 segundos, um tempo considerável para a quantidade de etapas realizadas e para a quantidade de informações em cada procedimento.

A tabela a seguir mostra a quantidade de sequencias encontrada em cada etapa da arquitetura.

Algoritmo Genético	BlastP	Conversão fasta	CD-Hit	Muscle	% Motifs Válidos
1.757 sequencias.	174.914 sequencias.	17.283 sequencias.	28 sequencias.	28 sequencias.	1,59%

Como podemos analisar na tabela acima, as 3 primeiras ferramentas geram uma quantidade significativa de sequencias, deixando claro a necessidade da utilização da ferramenta de refinamento, que reduz drasticamente a quantidade de sequencias encontradas.

4. CONCLUSÕES

Foram obtidas no final da execução foi de 28 sequencias que são possíveis Motifs válidos. Como podemos analisar, foi obtido um ganho de 98,41% sobre a quantidade de sequencias encontrada anteriormente, um ótimo ganho em relação ao maior número de sequencias encontrado (1.757 sequências).

Portanto, é possível analisar que a arquitetura teve um bom resultado, um ótimo ganho de informações relevantes contendo possíveis Motifs e um tempo de execução satisfatório, tornando possível a utilização da arquitetura de forma eficiente e rápida.

Entre os aspectos levantados para continuação do trabalho, devemos melhorar os seguintes pontos:

- Desenvolvimento de uma ferramenta computacional para a descoberta de motifs.
- Otimização do algoritmo genético.
- Criar uma base de dados de motifs já anotados para testes futuros.
- Criar um modelo de dataset para busca de regiões promotoras em bactérias.
- Tornar possível o download da arquitetura para a comunidade acadêmica.

5. REFERÊNCIAS BIBLIOGRÁFICAS

PROSDOCIMI, F. **Introdução à bioinformática.** Belo Horizonte: Biotecnologia ciência e desenvolvimento, (2007).

DAVIES, K. "**Decifrando o genoma: a corrida para desvendar o DNA humano.**" São Paulo: Companhia das Letras. (2001).

YI-PING PHOEBE CHEN. "**Bioinformatics Technologies."** Springer Science & Business Media, (2005).

BARRET STEVEN J. "**Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems.**" Genetic Programming and Evolvable Machines 7.3 (2006): 283-284.

PROSDOCIMI, F, et al. "**Bioinformática: manual do usuário."** Biotecnologia Ciência & Desenvolvimento 29 (2002): 12-25.

PARIDA, LAXMI. "**Pattern discovery in bioinformatics: theory & algorithms.**" CRC Press, 2007.

ALTSCHUL, STEPHEN F., et al. "**Basic local alignment search tool.**" Journal of molecular biology 215.3 (1990): 403-410.

LI, WEIZHONG, and ADAM GODZIK. "**Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.**" Bioinformatics 22.13 (2006): 1658-1659.

EDGAR, RC. "**MUSCLE: multiple sequence alignment with high accuracy and high throughput.**" Nucleic Acids Research 32 (5): 1792–97. doi:10.1093/nar/gkh340. PMC 390337. PMID15034147. (2004).

FINN RD, CLEMENTS J, EDDY SR. "**HMMER web server: Interactive sequence similarity searching. Nucleic Acids Research.**" Web Server Issue 39:W29-W37. (2011).