DESENVOLVIMENTO DE UM SISTEMA PARA ANÁLISE DE SEQUÊNCIAS GENÉTICAS VISANDO IDENTIFICAR PADRÕES DE COMPOSIÇÃO EM PROTEÍNAS EFETORAS

Josenilson G. LINO[1]; Gustavo J. SILVA[1]

[1]IFSULDEMINAS – Câmpus Muzambinho, Muzambinho/MG, e-mail: josenilsonlino@gmail.com; [2]IFSUDESTE – Câmpus Muzambinho, Muzambinho/MG, e-mail: gustavo.jose@muz.ifsuldeminas.edu.br.

RESUMO

As proteínas são essenciais para a vida de qualquer organismo, causando ações benéficas ou maléficas. Este trabalho busca padrões entre proteínas efetoras, aquelas prejudiciais aos organismos, tendo como base os elementos que compõe as proteínas, ou seja, nos aminoácidos. Na busca destes padrões, foi elaborado um banco de dados de proteínas efetoras obtidas do banco de sequências NCBI - National Center for Biotechnology Information. Após o desenvolvimento do banco de dados, as análises para busca de padrão entre os aminoácidos foram realizadas por meio do desenvolvimento e aplicação de algoritmos computacionais. Onde estes algoritmos busca o percentual de cada aminoácido presente em uma proteína e compara com o percentual de outras proteínas do banco de dados a fim de encontrar alguma semelhança. Este procedimento de duas formas: aplicando sobre a sequência total de cada proteína e dividindo uma sequência estrutural de uma proteína em intervalos e em seguida analisando e comparando o percentual de aminoácido existente. Os resultados obtidos foram apresentados em métodos estatísticos, como tabelas e gráficos, para melhor compreensão. Analisando as sequências totais das proteínas, em média, 11,80% apresentaram semelhanças, sendo que, alguns aminoácidos com a Leucina e Lisina, apresentaram maiores influência. Na analise das proteínas em intervalos, das 161 proteínas 86,34% obteve intervalos similares.

INTRODUÇÃO

A biotecnologia busca o desenvolvimento de produtos e a soluções de problemas. Como por exemplo, o desenvolvimento de drogas antibacterianas. Essas drogas são desenvolvidas com base me estudos de proteínas secretadas por bactérias. Durante a interação da bactéria e o hospedeiro, esta secreta proteínas no interior de suas células, ocasionado alterações nas funções celulares, e assim se multiplicam e causam doenças o hospedeiro, podendo acarretar até mesmo o óbito.

A Bioinformática é o estudo e aplicação de técnicas computacionais em diversas áreas da biologia, como a biologia molecular (BITTENCURT, 2005, apud SOUTO et al. 2003), na identificação e análise de sequências genéticas e determinação da estrutura de proteínas (BALDI E BRUNAK, 2001). Essas informações estão presentes em bancos de dados gênicos. Neste trabalho analisou-se proteínas secretadas por diversas bactérias que foram identificadas como efetoras, ou seja, proteínas que interagem com o hospedeiro alterando suas funções celulares. Através de algoritmos computacionais, é possível procurar e identificar padrões em relação à distribuição percentual dos aminoácidos constituintes das proteínas classificadas como efetoras. Embora não tendo utilizado neste trabalho, também são apresentadas algumas das principais técnicas computacionais empregadas na bioinformática.

Este trabalho justifica-se pela importância da análise de proteínas para criação de drogas antibacterianas. A bioinformática, aliada às técnicas de análises estatísticas, possibilitam identificar a existência ou não de padrões em relação à constituição das proteínas bacterianas e a partir deste estudo pesquisadores poderão realizar comparações para identificar proteínas potencialmente efetoras. Outro fator importante para o desenvolvimento do algoritmo para comparação das proteínas é a ausência de softwares correlatos.

Para investigar a existência de padrões em relação ao percentual de aminoácidos constituintes de proteínas efetoras, é necessário construir um banco de dados de proteínas efetoras. E com base neste banco de dados, realizar as analise e comparações entre as proteínas.

MATERIAL E MÉTODOS

O processo metodológico é constituído em duas etapas principais. Primeiramente, foi realizado um processo de seleção e cadastro das proteínas potencialmente efetoras, baseado nos estudos de Lockwood et al. (2011), cujos estudos apresentam a identificação de proteínas efetoras de sistema de secreção Tipo IV da *Anaplasma marginale*. Assim, foram selecionadas 161 proteínas obtidas das bactérias *Anaplasma marginale str. St. Maries, Agrobacterium tumefaciens, Bartonella henselae e Legionella pneumophila.* As sequências genéticas forma obtidas no site do *National Center for Biotechnology Information* (NCBI), disponivel no endereço http://www.ncbi.nlm.nih.gov/. A ferramentas constituída para análise foi desenvolvida com linguagem de programação PHP e o SGBD (Sistema Gerenciador de Banco de Dados) Mysql 5.5.27.

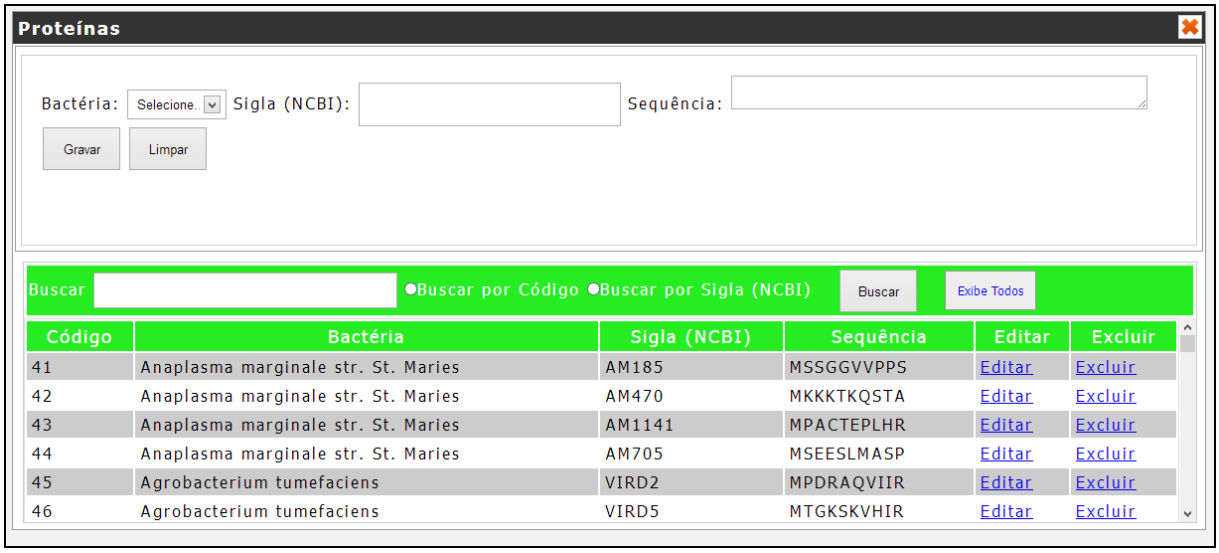

Figura 1: Cadastro de Proteínas

A segunda etapa envolveu a análise das proteínas, que foi realizada de duas formas. Primeiro, baseando-se na quantidade total de cada um dos vinte aminoácidos existentes que compões as sequências das proteínas. Depois, as sequências das proteínas foram divididas em intervalos, e aplicando o mesmo da primeira analise.

O em ambas as análises, o objetivo foi encontrar o percentual de aminoácido de cada que uma proteína possui e com o percentual das outras. Na primeira análise, que envolve a sequência completa, todas as proteínas foram comparadas entre si, procurando quantidades semelhantes. Já na segunda forma de análise, por

intervalo, as proteínas foram divididas em intervalos e a quantidade de aminoácidos encontradas nos intervalos, foi comparada com todas as proteínas da base de dados também dividias no mesmo intervalo. Este intervalo pode ser determinado em cada analise, assim fica a critério do pesquisador.

Como cada proteína pode assumir um tamanho distinto, em ambas as análises, o processo de comparação foi realizado tendo com base dois critérios importantes: Taxa de variação e Taxa de Aceitação. A Taxa de variação define um intervalo para que o percentual da quantidade de um aminoácido na comparação de duas proteínas sejam aceitos como semelhante. Ou seja, não é necessário que um aminoácido possua quantidades idênticas, podendo ser próximas para mais ou para menos. Por exemplo, na analise com taxa de variação de 15%, o algoritmo busca similaridade entre 15% superior e inferior ao valor do aminoácido comparado. A Taxa de Aceitação esta relacionada à quantidade de aminoácidos que satisfazem a Taxa de Variação. Ou seja, considerando uma taxa de aceitação de 70%, o algoritmo verifica se dos vinte aminoácidos, ao menos 70% deles estão entre o valor máximo e mínimo do intervalo analisado.

RESULTADOS E DISCUSSÃO

Os resultados obtidos mostraram que alguns aminoácidos apresentaram a maior concentração e outros atingiram os menores valores da variação como pode ser visto no gráfico 1. Analisando a composição total das proteínas, tendo como base uma taxa de variação de 15% e uma taxa de aceitação de 70%, proteínas com semelhanças entre si equivalem a 11,80%. Em 63,16% das proteínas com composição semelhantes, o Aminoácido Leucina, apresentou maior frequência, seguido pela Lisina e Glutamato com 31,53% e 5,26%. Em contrapartida, Tiptofano foi o aminoácido com menor frequência, com 63,16%, em seguida a Cisteína com 31,53% e a Asparagina com 5,26%.

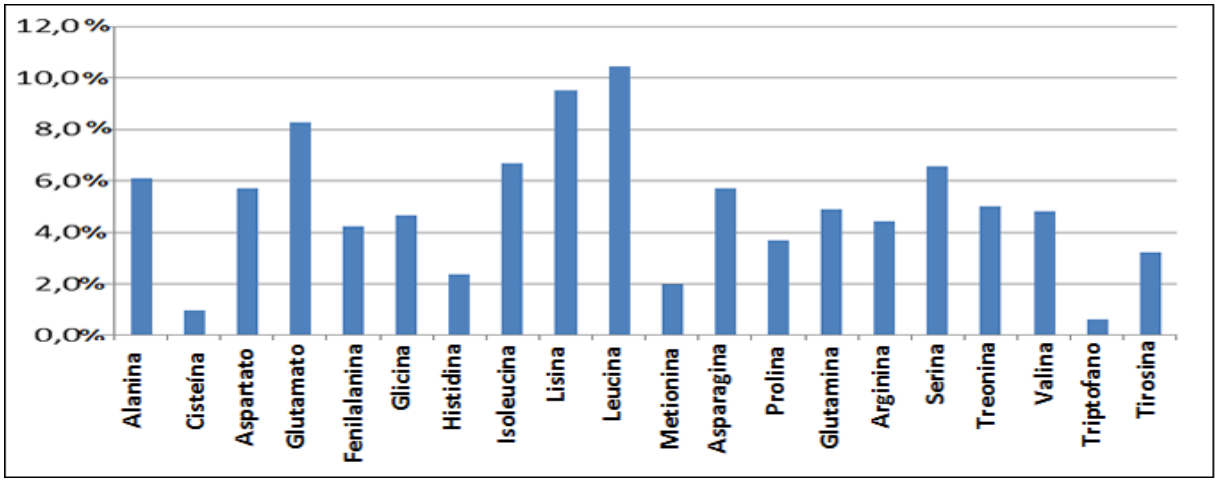

Gráfico 1: Percentual médio de Aminoácidos nas proteínas

A análise realizada tendo como base um intervalo de 200 aminoácidos, uma taxa de variação de 10% e uma taxa de aceitação de 70%, cerca de 86,34% das 161 proteínas, deveram semelhança em ao menos 70% dos 20 aminoácidos no intervalo de 200 aminoácidos por proteína, as proteínas menores que este intervalo não foram analisadas. É importante destacar que nesta forma de análise, o intervalo, a taxa de variação e a taxa de aceitação, são definidos a cada pesquisa. O motivo da escolha dos valores acima, esta ligada ao intuito de apresentar as variações encontradas.

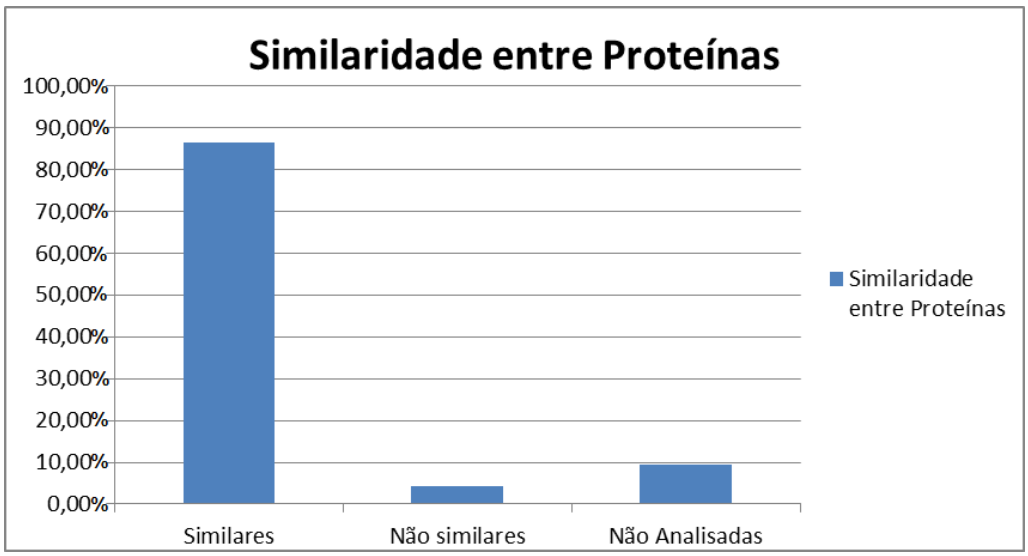

Gráfico 2: Similaridade entre Proteínas

A análise ainda aponta que a proteína com maior quantidade de aminoácidos em comum, obteve 37,89% de semelhança com o total de 161 proteínas. E,

desconsideram aquelas que não apresentaram o mínimo de 70% de similaridade, as que apresentaram a menor índice de igualdade, atingiu o valor de 0,62%. Este percentual de 0,62% foi o que apareceu na amostragem, ou seja, grande parte das proteínas atingiu 0,62% de semelhança. A média geral é de 7,00%, de semelhança entre as 161 proteínas analisadas.

CONCLUSÕES

Com base nos resultados obtidos, conclui-se que existe uma similaridade entre a composição destas proteínas efetoras, sendo que, na análise realizada por intervalos, esta semelhança se mostrou mais intensa. Assim, pode-se afirmar que em algumas partes das proteínas, existe um padrão que pode ser apontado como um ponto crítico relacionado ao potencial efetor destas proteínas.

Com base nestes estudos, uma necessidade que surge é saber se este padrão se é similar em proteínas não efetoras. E também se o percentual de aminoácidos é semelhante nestes padrões, pois assim, será possível apontar um padrão para investigar em outras proteínas que estão sendo analisadas. Este parâmetro associado a outras caraterística pode ser crucial para avaliar uma dada proteína, agilizando o trabalho de pesquisadores.

REFERÊNCIAS BIBLIOGRÁFICAS

Baldi, P. e Brunak, S. (2001). Bioinformatics: the Machine Learning approach. MIT Press, segunda edição.

BITTENCOURT, Valneide Gomes. Aplicações de Técnicas de Aprendizagem de Máquinas no Reconhecimento de Classes Estruturais de Proteínas. 2005. 116 f. Dissertação (Mestrado) - Depto de Centro de Tecnologia, Univ. Fed. do Rio Grande do Norte, Natal, 2005.

LOCKWOOD, Svetlana et al. Identification of Anaplasma marginale Type IV Secretion System Effector Proteins. Minnesota: Plos One, 2011.