Trabalho Prático 1:
Alinhamento de Sequências Proteicas

Hugo Richard Amaral, Luís E. O. Lizardo
Universidade Federal de Minas Gerais
{hucharal, lizardo}@dcc.ufmg.br

Resumo: O Alinhamento de sequências possibilita comparar duas sequências e verificar a similaridade entre elas. Neste trabalho foi implementado um algoritmo de alinhamento de sequências para tentar identificar possíveis mutações em uma proteína. As analises foram feitas para a enzima triose-fosfato isomerase (TIM). O objetivo do trabalho é identificar as mutações que podem alterar a função proteica desta enzima e sugerir possíveis modificações que possam restaurar a função proteica desta enzima.

1. INTRODUÇÃO

O alinhamento de sequências é uma técnica que permite arranjar e/ou organizar sequências de DNA, RNA ou proteína para identificar regiões de similaridade oriundas de relação funcional, estrutural e evolutivas entre as sequências. Dessa forma, o alinhamento possibilita localizar trechos conservados entre genomas, comparar uma sequência desconhecida com um banco de dados e reconstruir sequências a partir da sobreposição de fragmentos.

Proteínas são compostos orgânicos bioquímicos que desempenham funções específicas nos organismos. Elas são formadas por sequências de resíduos de aminoácidos. Cada aminoácido possui características distintas e possuem um papel na função proteica. A troca de um aminoácido por outro, decorrente de uma mutação, pode provocar alterações no funcionamento da proteína.

Neste trabalho utilizamos um algoritmo de alinhamento de sequências para identificar mutações que possam ter alterado a função de uma proteína. Com base nos resultados encontrados, sugerimos modificações que permitam resgatar a funcionalidade desta proteína. Para isso o desenvolvimento do trabalho é utilizado a enzima triose-fosfato isomerase (TIM), que teve sua

função proteica alterada devido a mutações em sua sequência de resíduos.

Com base nos experimentos realizados, identificamos 11 possíveis mutações que podem alterar a funcionalidade da enzima triose-fosfato isomerase. Os experimentos foram realizados alinhando a sequência mutada com a mesma enzima sem mutações, e também alinhando outras sequências da mesma família com essa enzima original e sem mutações. A comparação desses alinhamentos possibilitou concluirmos quais delas poderiam alterar a função da proteína.

2. CONCEITOS BÁSICOS

Nesta seção são apresentados os principais conceitos envolvidos na solução do problema.

2.1 Alinhamento de sequências

O alinhamento de sequências é uma técnica da bioinformática para comparar sequências de DNA, RNA ou proteína para identificar regiões similares. A similaridade pode indicar a ocorrência de relações funcionais, estruturais ou evolucionarias entre as sequências. O alinhamento pode apresentar a ocorrência de **matches**, quando há casamento exato entre dois nucleotídeos (no caso de DNA e RNA) ou resíduos de aminoácidos (no caso de proteínas), de **mismatches**, interpretados como mutações pontuais, e **gaps**, como inserções ou deleções introduzidas em uma ou ambas as sequências.

O alinhamento é chamado de **simples** quando corresponde à comparação entre duas sequências, e **múltiplo**, quando é entre três ou mais sequências. Ele pode ser do tipo **global,** quando a comparação é feita de uma extremidade a outra das sequências, ou **local,** quando o objetivo é encontrar e extrair um ou mais segmentos das sequências comparadas que exibam alta similaridade.

Para o problema apresentado neste trabalho será utilizado o algoritmo de Needleman-Wunsch [1], que realiza alinhamento simples e global. Este algoritmo criado em 1970 é um método de programação dinâmica e utiliza uma tabela de pontuação (explicada na subseção 2.2) que define previamente a similaridade entre dois nucleotídeos ou resíduos de aminoácidos.

2.2 Matriz de pontuação

Uma matriz de pontuação define a similaridade entre dois caracteres (nucleotídeos ou resíduos de aminoácidos) em uma sequência. Ela atribui um valor ou *score* para cada comparação de pares de caracteres. Normalmente os valores são inteiros, sendo os positivos representando os pares similares de caracteres e negativo para pares dissimilares.

As matrizes de pontuação para proteínas são normalmente construídas por meio da analise de frequências de substituições de resíduos em alinhamentos de famílias conhecidas de proteínas. O *score* também leva em conta a frequência com que cada aminoácido ocorre na natureza.

Neste trabalho será utilizado a matriz PAM$_n$ (*Point Accepted Mutation*) [2]. Esta matriz é baseada em alinhamentos globais de proteínas muito semelhantes e evidencia a origem evolutiva de proteínas. Duas sequências são ditas a uma distância evolutiva de 1 PAM (PAM1) se uma pode ser convertida na outra com uma média de 1 mutação a cada 100 aminoácidos, ou seja, a matriz foi calculada a partir da comparação entre sequências apresentando menos de 1% de divergências. As demais matrizes PAMs foram extrapoladas a partir da PAM1. PAMs com taxa de divergências menores, são mais indicadas para alinhamento de sequências que são conhecidamente mais similares. No trabalho serão utilizadas as matrizes[1] derivadas PAM60, PAM120 e PAM250.

3. METODOLOGIA

3.1 Grupos de aminoácidos

A fim de identificar quais alterações de proteínas são mais drásticas, utilizamos uma classificação dos aminoácidos para agrupá-los de forma que em um mesmo grupo tivéssemos aminoácidos com características semelhantes.

A classificação considerada foi a mesma utilizada pela ferramenta VERMONT[2] e está descrita a seguir:

- Grupo 1 (Polar Positivo): Histidina (H), Lisina (K), Arginina (R);

[1] http://www.bioinformatics.nl/tools/pam.html
[2] http://homepages.dcc.ufmg.br/sabrinas/vermont/

- Grupo 2 (Polar Negativo): Aspartato (D), Glutamato (E);
- Grupo 3 (Polar Neutro): Asparagina (N), Glutamina (Q), Serina (S), Treonina (T);
- Grupo 4 (Alifático Apolar): Alanina (A), Glicina (G), Isoleucina (I), Leucina (L), Metionina (M), Valina (V);
- Grupo 5 (Anel Apolar): Fenilalanina (F), Prolina (P), Triptofano (W), Tirosina (Y);
- Grupo 6 (Cristeína): Cristeína ©

Dessa forma, mutações realizadas dentro de um mesmo grupo, ou seja, entre aminoácidos com a mesma classificação, foram consideradas conservativas e, portanto, não alteram a função da proteína.

3.2 Alinhamento entre dTIM e 2YPIA

De posse dos grupos de aminoácidos, realizamos o primeiro alinhamento entre as sequências selvagem (2YPIA) e a mutada (dTIM). Nessa etapa, pretendemos identificar na proteína selvagem quais os pontos que ocorreram algum tipo de mutação, adição ou deleção de aminoácido.

Para isso, guardamos todas as posições em que as duas proteínas diferiram e qual foi a diferença identificada (mutação, adição ou deleção), pois consideramos que uma mutação só é perfeitamente caracterizada se soubermos o que o ocorreu e em qual posição da proteína ocorreu.

3.3 Alinhamento entre as proteínas da família

No passo seguinte, realizamos o alinhamento entre a proteína selvagem e as 133 proteínas da mesma família. O objetivo aqui é identificar quais são as diferenças nas sequências de aminoácidos que ocorrem na família e que, portanto, não alteram a função das proteínas.

3.4 Comparação dos alinhamentos

Com dados obtidos nas etapas anteriores, comparamos estes para tentar identificar aquelas mutações que possivelmente alteram a função da proteína. Nosso objetivo foi tentar resolver o problema utilizando o melhor critério possível, ou seja, apontar aquelas mutações que, com um grau alto de confiança, modificam a função da proteína.

Portanto, para todas as mutações encontradas na dTIM, verificamos se ela também ocorre em

alguma proteína da família. Vale ressaltar que, uma mutação só é considerada como a mesma, se a troca ocorre entre os mesmos aminoácidos e na mesma posição em relação à proteína selvagem.

Considerando que, se há uma diferença entre aminoácidos da proteína selvagem e alguma de sua família e mesmo assim elas possuem a mesma função, então a mesma alteração entre a dTIM e a selvagem possivelmente não causará mudança de função. Essa foi a decisão tomada para tentarmos identificar com o critério mais confiável quais as alterações na proteína mutada que alteram sua função original.

4. RESULTADOS

Além de tentar resolver o problema proposto, também realizamos algumas análises sobre os dados, a saber:

1. Como a qualidade dos resultados era interferida a partir da utilização de diferentes matrizes de pontuação;
2. Qual a diferença entre resultados obtidos ao desconsiderar as mudanças conservativas ou não.

Assim, realizamos alguns testes, variando essas informações, e os resultados são apresentados a seguir.

4.1 Mutações

4.1.1 Mutações entre as proteínas dTIM e 2YPIA

Nessa análise inicial, verificamos qual era o número total de mutações existente na proteína dTIM que a diferenciava da proteína selvagem, considerando as matrizes de pontuação PAM60, PAM120 e PAM250. Podemos ver uma ilustração desse alinhamento na Figura 1 e os dados podem ser verificados na Figura 2.

```
2ypia:   MARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATYLDYSVSLVKKPQVTVGAQNAYLKASGAFTGENSV 80
         MART FVGGN K NG K   KE VE L  A  P  VEVV  PPA YLD      K     V AQN Y  A GAFTGE S
dTIM:    MARTPFVGGNWKMNGTKAEAKELVEALK-AKLPDDVEVVVAPPAVYLDTAREALKGSKIKVAAQNCYKEAKGAFTGEISP 80

2ypia:   DQIKDVGAKWVILGHSERRSYFHEDDKFIADKTKFALGQGVGVILCIGETLEEKKAGKTLDVVERQLNAVLEEVKD-WTN 160
            KD GA  VILGHSERR YF E D   A K   AL  G  VI CIGETLEE  AGKT  VV RQ  A L    D W N
dTIM:    EMLKDLGADYVILGHSERRHYFGETDELVAKKVAHALEHGLKVIACIGETLEEREAGKTEEVVFRQTKALLAGLGDEWKN 160

2ypia:   VVVAYEPVWAIGTGLAATPEDAQDIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGFLVGGASLKPE 240
         VV AYEPVWAIGTG  ATPE AQ  HA IRK LA       A   RILYGGS    NA       D DGFLVGGASLKPE
dTIM:    VVIAYEPVWAIGTGKTATPEQAQEVHAFIRKWLAENVSAEVAESVRILYGGSVKPANAKELAAQPDIDGFLVGGASLKPE 240

2ypia:   FVDIINSRN 249
         F DIINSRN
dTIM:    FLDIINSRN 249
```

Figura 1. Alinhamento entre a proteína selvagem e a mutada

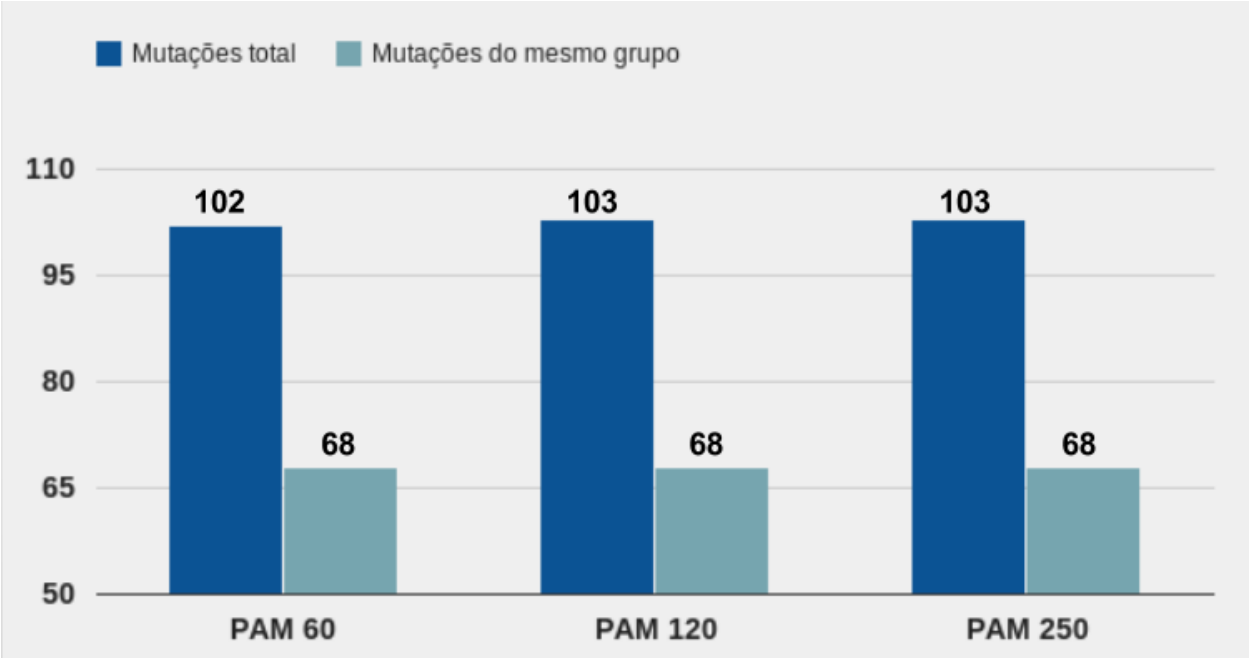

Figura 2. Número de mutações entre dTIM e 2YPIA

Além da análise feita para diferentes matrizes, avaliamos também os resultados quando desconsiderávamos as mutações conservativas (dentro do mesmo grupo) ou não, Conforme podemos observar pela figura, a utilização de qualquer uma das três tabelas gerou o mesmo resultado. Por outro lado, ao desconsiderarmos as mudanças conservativas, ou seja, ao avaliarmos que tais mudanças não impactam na função da proteína, o número total de mutações se reduzia a quase metade.

4.1.1 Mutações entre as proteínas dTIM e 2YPIA que alteram a função

Nesse segundo momento, apresentamos os resultados que tentam resolver o problema proposto. Conforme dito anteriormente, pretendíamos chegar a resultados o mais confiáveis possíveis no intuito de dizer que certa mutação alterou a funcionalidade da proteína.

Na Figura 3, mostramos os resultados obtidos. Podemos perceber que, assim como na análise anterior, não encontramos diferenças significativas ao utilizarmos diferentes versões das matrizes de pontuação PAM. Já ao considerarmos os grupos de proteínas, vemos que houve uma redução no número total de mutações.

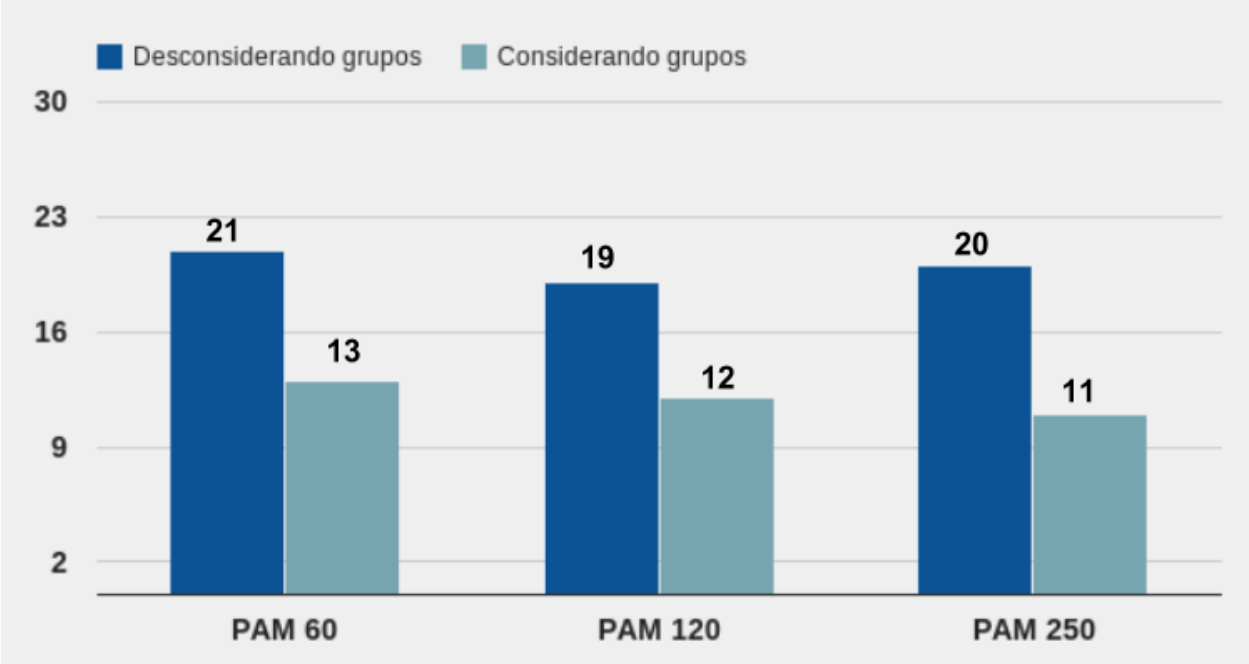

Figura 3. Número de mutações que alteram a função da proteína

Para as três tabelas, aquela que retornou um número menor de mutações foi a PAM250. Para ilustrarmos quais as modificações determinadas por ela, mostramos na Tabela 1 qual o tipo de modificação, qual(is) o(s) aminoácido(s) envolvido(s) e em que posição, com relação a proteína selvagem (2YPIA), a mudança ocorreu.

TIPO	ALTERAÇÃO	POSIÇÃO
Deleção	T	28
Mutação	Y -> T	48
Mutação	K -> G	55
Mutação	D -> K	110
Mutação	Q -> H	118
Adição	G	154
Mutação	S -> F	187
Mutação	D -> A	198
Mutação	S -> E	202
Mutação	N -> K	213
Mutação	G -> P	214

Tabela 1. Relação de mutações utilizando matriz PAM250

4.2 Restauração da funcionalidade

Dado que a proteína selvagem e todas as proteínas da família possuem a mesma função, consideramos que, se a proteína mutada pudesse ser transformada em uma delas, teríamos então a funcionalidade restaurada.

Para isso, devemos considerar, também, o menor esforço necessário para realizar tal transformação. Assim, decidimos analisar a pontuação de cada alinhamento entre a proteína dTIM entre a selvagem e as demais proteínas da família. Aquela que maximizasse a pontuação seria a proteína com maior similaridade com a mutada e que, assim, necessitaria de menos mutações.

De todas as proteínas, aquela que apresentou maior pontuação foi a proteína 1SW3A, cujo valor é 941. O alinhamento, ilustrado na Figura 4, mostrou que existem 102 mutações existentes entre as duas sequências, mas, considerando apenas as não-conservativas, teríamos um total de 61 mutações necessárias para transformar a mutada na proteína 1SW3A.

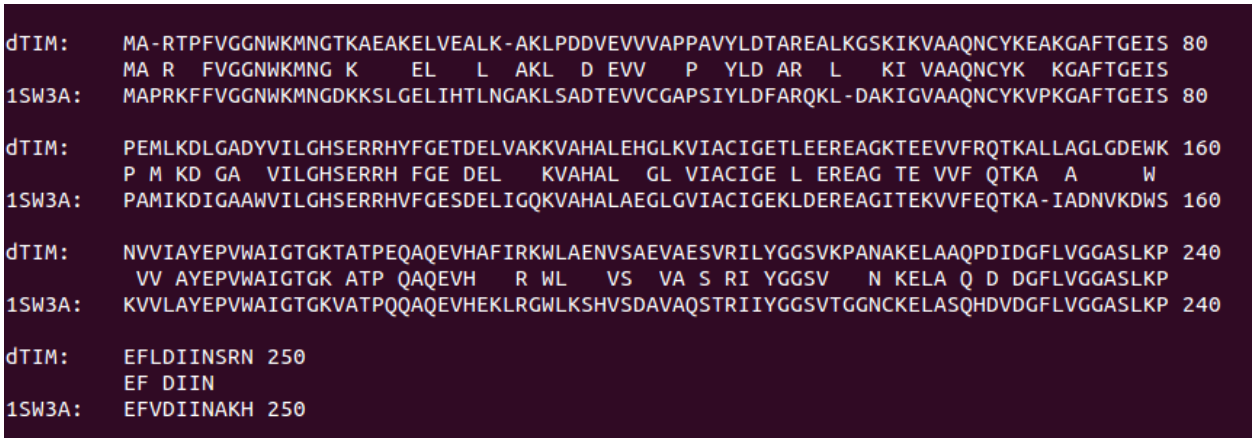

Figura 4. Alinhamento entre dTIM e proteína mais similar 1SW3A

5. CONCLUSÃO

Através da pesquisa e desenvolvimento do trabalho, podemos concluir que o algoritmo de alinhamento implementado (Needleman-Wunsch) é simples e fácil de implementar. Entretanto, apenas a análise do resultado obtido através da execução do algoritmo não é suficiente para darmos um resultado confiável de que determinada mutação alterou a funcionalidade da proteína.

Todas as decisões tomadas durante o desenvolvimento levaram em consideração apenas a posição que mutação ocorreu em relação a sequência original. Entretanto, outras características importantes foram desconsideradas, como a forma adquirida pela proteína após o enovelamento. A estrutura formada influencia diretamente em como a proteína reage com seus ligantes e como é o formato do sítio ativo, por exemplo. Mudanças que ocorrem nesses pontos críticos podem alterar drasticamente em como a proteína se comporta e, consequentemente, sua funcionalidade.

Outros fatores também são determinantes para tentar analisar as mutações, como a escolha do gap fixo ou variável, quais critérios utilizar para classificar dois aminoácidos como semelhantes, a utilização de uma matriz de pontuação mais específica, entre outros.

Entretanto, podemos concluir que o alinhamento de sequências pode servir como ponto de partida, ou um primeiro passo, para tentar descobrir quais pontos, ou regiões, de uma proteína

as mutações ocorridas alteraram a sua função. Além disso, tivemos uma ideia inicial do quão é complicado a resolução desse problema, principalmente por causa da grande quantidade de variáveis que podem influenciar, de certa forma, a função de uma proteína.

REFERÊNCIAS

[1] NEEDLEMAN, S. B. and WUNSCH, C. D., A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology 48 (3): 443–53, 1970.

[2] DAYHOFF, Margaret O.; SCHWARTZ, Robert M. A model of evolutionary change in proteins. In: In Atlas of protein sequence and structure. 1978.