Algoritmos de alinhamento múltiplo: conhecer para entender possíveis vieses na inferência filogenética

Caio César de Melo Freire*

*Laboratório de Evolução Molecular e Bioinformática, Instituto de Ciências Biomédicas, Universidade de São Paulo. Email: freire@ime.usp.br

Introdução

Alinhamento múltiplo de sequências (AMS) é uma ferramenta fundamental na pesquisa biológica e pode ser usada para uma variedade de propósitos abrangendo desde identificação de estruturas secundárias, detecção de ARN não funcionais e inferência filogenética. Uma boa definição para AMS é de um procedimento para converter sequências de diferentes tamanhos em sequências de tamanhos iguais pelo posicionamento de lacunas quando necessárias, com a meta de inferir similaridade entre os caracteres da sequência. (OGDEN, TH; ROSENBERG, M; 2006). No entanto, sequências de mesmo tamanho também podem necessitar de alinhamento, logo AMS também poderia ser definido como a melhor visualização da máxima similaridade entre as sequências estudadas.

Apesar de AMS ser crítico para a reconstrução filogenética pouca atenção tem sido empregado a este importante fator que pode enviesar de sobremaneira a topologia das árvores filogenéticas. Inclusive Hall (2005) afirma que é um truísmo acreditar que a qualidade de uma árvore não é melhor que a qualidade do alinhamento utilizado na estimativa da árvore. Ogden e Rosenberg (2006) demonstram que dependendo da topologia geral de uma árvore esta pode está mais sujeita a erros relacionados a alinhamentos, variando principalmente em função dos comprimentos de ramos das árvores analisadas. Pode-se justificar a relação apresentada pela constatação de que comprimentos de ramos em uma topologia representam a variação genética nas sequências analisadas, logo um alinhamento múltiplo que origina este tipo de topologia é mais diverso que outro AMS relacionado a uma topologia com comprimentos de ramos menores.

O alinhamento ótimo entre duas sequências peptídicas pode ser eficientemente computado em tempo quadrático utilizando um algoritmo de programação dinâmica (NEEDLEMAN, S; WUNCH, C; 1970). Em geral, AMS se refere ao alinhamento de três ou mais sequências. O algoritmo de programação dinâmica para alinhar duas sequências pode ser extrapolado para alinhar L sequências no tempo O((2L-1)nL). No entanto, este algoritmo não pode ser utilizado diretamente par alinhar sequências

múltiplas em virtude da alta complexidade computacional (GUPTA, S ET AL; 1995; SONG, J ET AL; 2007).

Diante da complexidade computacional do AMS uma variedade de métodos heurísticos e programas tem sido propostos para evitar a otimização direta da função do *Score* enquanto eficientemente gera alinhamentos precisos. Estes métodos podem ser divididos em quatro categorias: iterativos, progressivos, baseados em âncoras e probabilísticos (SONG, J ET AL; 2007).

Os métodos iterativos são os que iniciam com um alinhamento inicial e em seguida iterativamente refinam o alinhamento até não ser mais possível melhorá-lo com uma abordagem estocástica ou determinística (SONG, J ET AL; 2007). Os métodos progressivos alinham as sequências em subconjuntos por repetitivamente selecionar duas sequências do conjunto e as substituir pelos respectivos alinhamentos. O processo termina quando o conjunto contém somente uma "sequência" que consiste em um alinhamento múltiplo de todas as sequências do conjunto (THOMPSON, J ET AL; 1994). Ferramentas de AMS baseadas em âncoras iniciam o alinhamento pela busca de motivos locais conservados nas sequências e os usar como âncoras para os alinhamentos. As regiões entre as âncoras são alinhadas para formar um alinhamento médio entre as sequências (KATOH, K ET AL; 2002). Métodos probabilísticos são baseados na distribuição das probabilidades de substituição obtidas dos alinhamentos múltiplos disponíveis, o AMS pode ser construído pela maximização da probabilidade média de substituições (SONG, J ET AL; 2007).

Este trabalho tem o objetivo de analisar os métodos mais utilizados entre os heurísticos e também uma abordagem probabilística baseada em Perfil de Modelo Oculto de Markov (*Profile HMM*) para Alinhamento múltiplo de sequências, considerando as possíveis consequências para a inferência filogenética.

Métodos Heurísticos

ClustalW

O algoritmo básico deste tipo de AMS consiste em três estágios principais: (*i*) todos os pares de sequências são alinhados separadamente para calcular a matriz de distância fornecendo assim a divergência entre cada par de sequências. (*ii*) Uma árvore guia é calculada a partir desta matriz de distância. (*iii*) As sequências são progressivamente alinhadas de acordo com a topologia na árvore guia.

As árvores guias usadas para direcionar o processo final de alinhamento múltiplo são calculadas a partir da matriz de distância utilizando-se o método *Neighbor-joining*. Produzindo-se assim árvores não enraizadas com comprimentos de ramos proporcionais a divergência entre as sequências, em seguida a árvore é

enraizada no ponto médio (*midpoint*), onde a distância média dos comprimentos de ramos de cada lado da árvore é equivalente (THOMPSON, JD ET AL; 1994).

Algumas das melhorias implementadas no algoritmo progressivo básico em *ClustalW* permitem o aumento da sensibilidade do método progressivo para alinhar sequências muito divergentes. As melhorias são ponderação da sequência (Fator de multiplicação do *Score* baseado na conservação do caractere na posição); penalidades para lacunas (duas penalidades para lacunas, uma para abertura e outra para extensão); dependência da matriz de composição (determina fator de escala para penalidade de abertura de lacuna); dependência da similaridade de sequências (o percentual de divergência entre os grupos alinhados é utilizado para aumentar a penalidade de abertura de uma lacuna para um grupo mais relacionado e diminuir a penalidade para um grupo mais distante); dependência do tamanho das sequências (*Scores* crescem com tamanho da sequência); dependência da diferença nos tamanhos das sequências (a penalidade de extensão de lacuna é modificada de acordo com este parâmetro); penalidade de lacunas para posições específicas; diminuição das penalidades de lacuna onde já existem lacunas (a penalidade de abertura é reduzida em relação ao número de sequências que tem uma lacuna nesta posição e a penalidade de extensão é reduzida pela metade); aumento das penalidades de lacunas em regiões próximas a lacunas; penalidades de lacunas reduzidas em trechos hidrofílicos (a penalidade de abertura é reduzida em um terço nestas regiões que são *loops* em potencial); penalidade resíduo-específica (se a posição não está em um trecho hidrofílico nem contém outras lacunas, a penalidade de abertura de lacuna é multiplicada pelos fatores de Passarela e Argos); e matrizes de substituição (THOMPSON, JD ET AL; 1994).

Muscle

O algoritmo deste tipo de AMS inclui estimativa rápida de distância utilizando elementos *kmer*, alinhamento progressivo usando uma função chamada *log-expectation* (LE) e refinamento (método iterativo) baseado em particionamento restrito que é dependente de topologia de árvore filogenética. Em suma, após a construção da árvore guia, a etapa fundamental é o alinhamento de perfil *pairwise*, que é usado na etapa de alinhamento progressivo e depois na etapa de refinamento (EDGAR, R; 2004a).

Muscle utiliza duas medidas de distância para um par de sequências, distância *kmer* em pares não alinhados e distância de Kimura em pares alinhados. Uma *kmer* é uma subsequência contígua de tamanho *k*. Sequências relacionadas tendem a ter mais *kmers* em comum que o esperado pelo acaso. A vantagem desta medida é que ela não requer alinhamento o que fornece uma significativa vantagem de velocidade. A correção de Kimura para múltiplos sítios é aplicada sobre a matriz de identidade para

convertê-la em uma estimativa aditiva de distância. As matrizes de distância são agrupadas utilizando UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*), vale ressaltar essa diferença com *ClustalW* que utiliza *Neighbor-joining*, Edgar (2004a) afirma que a utilização de UPGMA no seu método supera superficialmente os resultados com *Neighbor-joining*, a despeito que a última metodologia é capaz de estimar melhor uma árvore filogenética. Justifica-se o uso de UPGMA na árvore guia pela proposição que no alinhamento progressivo, a melhor precisão é obtida em cada nó da árvore pelo alinhamento de dois perfis que tem menor número de diferenças, mesmo que eles não sejam grupos irmãos (*evolutionary neighbors*). De acordo com Edgar (2004b), para gerar os perfis de alinhamento *Muscle* utiliza uma função de probabilidade chamada *log-expectation* (equação 1).

$$LExy = (1 - f\,xG)\,(1 - fyG)\,\log \sum i \sum j fxi\,fyi\,pij / pi\,pj$$

Equação 1. Função para estimar log-expectation (LE) onde *i* e *j* são os tipos de aminoácidos, *pi* a probabilidade de fundo de *i*, *pij* a probabilidade de *i* e *j* serem alinhados juntos, *fxi* a freqüência observada de *i* na coluna *x* no primeiro perfil e *fxG* a freqüência observada de lacunas naquela coluna na posição *x* na família (similarmente para a posição *y* no segundo perfil). As probabilidades pi e pj são derivadas da 240 VTML PAM que é uma matriz de substituição baseada em máxima verossimilhança (JONES, DT ET AL; 1992).

Durante o processamento deste AMS, tem-se uma primeira etapa caracterizada por construção de uma matriz de distância D1 baseada em *kmer* de cada par de sequências, inferência de uma árvore binária (T1) por UPGMA utilizando D1, contrução de um alinhamento progressivo utilizando T1 como guia e um alinhamento grosseiro construído por método progressivo. Na etapa 2, objetiva-se corrigir a principal fonte de erro em T1 que é a imprecisão de *kmer* para isso aplica-se a distância de Kimura (mais precisa que *kmer* e necessita de um alinhamento) no primeiro alinhamento e se obtém a matriz de distância D2, inferência de uma segunda árvore binária T2 por UPGMA, em seguida alinhamento por método progressivo. A etapa 3 classificada como refinamento é baseada em particionamento de T2 e computação de novos perfis que são realinhados entre si e se o *Score* do novo alinhamento for superior ao anterior este novo alinhamento é aceito. Este processo de particionamento é repetido até um estado estacionário ou até atingir um limite pré-definido pelo usuário (EDGAR, R; 2004a).

Mafft

Mafft é uma ferramenta de AMS que faz uso de duas técnicas: (*i*) Identificação de regiões homólogas por transformação rápida de Fourier (FFT) (uma sequência de aminoácidos é convertida em uma sequência que descreve volume e polaridade de

cada resíduo) e (*ii*) Sistema de *Score* simplificado (utilização de poucos parâmetros). Este método pressupõe que a frequência das substituições de aminoácidos depende fortemente das diferenças entre suas propriedades físico-químicas. Dessa forma, é possível agrupar motivos peptídicos de acordo com suas propriedades e assim ancorar estes motivos ao longo de um alinhamento. Esta ferramenta utiliza janelas padrão de 30 resíduos, os motivos são identificados por picos de correlação que correspondem a dois blocos similares, uma matriz de homologia é gerada (KATOH, K ET AL; 2002).

De acordo com Katoh e colaboradores (2005), o algoritmo do Mafft tem três etapas (duas progressivas e uma de refinamento). Durante a Etapa 1, uma distância grosseira entre pares de sequências é calculada baseada no número de categorias de Aa compartilhadas. Uma árvore-guia é construída por UPGMA e as sequências são alinhadas em relação a esta árvore. Na etapa 2, o alinhamento da etapa anterior é utilizado para construir uma nova matriz de distância que é utilizada em uma nova árvore para um novo alinhamento. A etapa 3 é de refinamento e utiliza a técnica de particionamento restrito semelhante ao já descrito na seção sobre *Muscle*.

Métodos probabilísticos

Apesar de a maioria das ferramentas para AMS utilizarem métodos que otimizam *Score* associado com um alinhamento em particular, nem sempre o alinhamento com o melhor *Score* é o que tem mais significado biológico. Portanto, um método capaz de gerar uma lista de alinhamentos com melhores *Scores* pode ser mais útil em determinadas situações. Se duas sequências de um alinhamento com máxima similaridade forem utilizadas como ponto inicial, um conjunto de alinhamentos com os melhores *Scores* pode ser computado com um algoritmo de computação dinâmica. Para cada alinhamento no conjunto, um perfil de Modelo Oculto de Markov (HMM) pode ser construído para descrever a distribuição estatística de cada aminoácido para cada coluna no alinhamento (SONG, J ET AL; 2007).

Um HMM pode ser definido como um conjunto de estados conectados, cada estado potencialmente capaz de emitir uma série de observações. O processo evolui em uma dimensão, geralmente no tempo, mas não obrigatoriamente. O modelo é parametrizado com probabilidades que governam cada estado no tempo *t+1*, dado que o estado prévio é conhecido. Os pressupostos de Markov são utilizados para excluir a dependência de se conhecer toda a história entre os estados para se acessar o próximo estado. Dessa forma, somente um passo é requerido para se acessar o próximo estado. Como o processo evolui no tempo, cada estado pode potencialmente emitir

observações que são consideradas como fluxo de informação ao longo do tempo. Quando aplicado a sequências biológicas a dimensão tempo é substituída pelo posicionamento na sequência (BIRNEY, E; 2001).

Um perfil de HMM é uma implementação desenvolvida especialmente para Alinhamento Múltiplo de Sequências que combina a ideia de perfil com HMM. Uma das principais vantagens desta abordagem é que os *Scores* não são determinados de maneira heurística e sim em fórmulas estatisticamente consistentes (FONZO, V ET AL; 2007), resultando em última instância em uma sequência consenso com base probabilística. A arquitetura de um perfil (Figura 1) de HMM pode ser simplificada para uma estrutura com três estados M, D e I *(match*, *delete* e *insert*). O estado M corresponde a um aminoácido consenso para esta posição na família proteica. O estado D é não emissor e representa o salto desta posição consenso no alinhamento múltiplo. O estado I representa a inserção de resíduos após a posição consenso (BIRNEY, E; 2001).

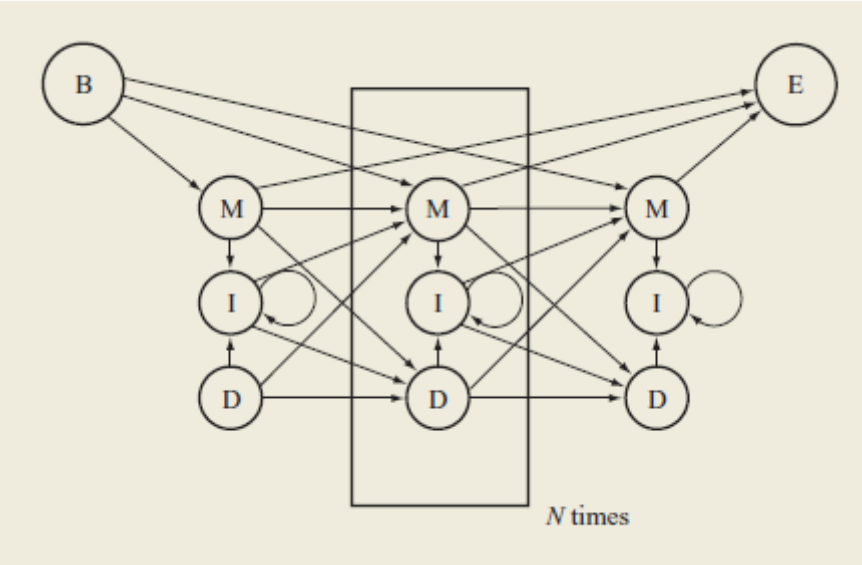

Figura 1: Arquitetura de Perfil de HMM – Estrutura repetitiva de três estados (M, D e I). Cada conjunto de três estados representa uma única coluna no alinhamento. Retirada de Birney (2001).

Dessa forma, um conjunto de perfis pode ser construído a partir do conjunto de alinhamentos com os melhores *Scores*. As sequências remanescentes no conjunto são alinhadas aos perfis de HMM uma por uma. Após uma dada sequência ser alinhada a todos os perfis disponíveis os parâmetros estatísticos nestes perfis são atualizados baseados nos alinhamentos com os melhores *Scores.* O processo termina quando o conjunto de sequências não contém mais sequências desalinhadas. Os alinhamentos com os melhores *Scores* são reportados como possíveis alinhamentos das sequências (SONG, J ET AL; 2007).

Palign é uma ferramenta que utiliza perfis de HMM para AMS. Pode-se dividir seu algoritmo em três etapas. Na primeira, a ordem para alinhar sequências para os perfis de HMM é determinada pelo melhor alinhamento *pairwise* entre cada par de sequências no conjunto. As duas sequências com o máximo *Score* de alinhamento serão utilizadas para construir os perfis de HMM iniciais, pode-se ter mais de um alinhamento ótimo inicial. Este processo determina a ordem em que as sequências serão alinhadas aos perfis de HMM gerados. Na etapa 2, assume-se que o número de perfis de HMM será *k*, um algoritmo de programação dinâmica computa os alinhamentos com os máximos *Scores* no tempo $O(klog_2kn^2)$, onde *n* é o tamanho da sequência. Cada um dos alinhamentos pode ser descrito como um perfil de HMM que contém a distribuição dos aminoácidos de cada coluna do alinhamento. Na etapa 3, as sequências remanescentes são alinhadas a cada um dos perfis por um algoritmo de programação dinâmica semelhante ao da etapa anterior para computar os alinhamentos com *k Scores* máximos. Portanto, tem-se k^2 possíveis alinhamentos no total selecionam-se os alinhamentos com *k Scores* máximos dos k^2 disponíveis e os perfis de HMM são atualizados com base nesses alinhamentos. Este processo é repetidamente aplicado até não existirem mais sequências desalinhadas no banco de dados. Em seguida, uma lista com *k* alinhamentos é reportada e os *Scores* máximos como resultados dos alinhamentos (SONG, J ET AL; 2007).

Considerações Finais

Diante do problema exposto na contextualização da importância deste estudo, a interferência da acuidade do alinhamento múltiplo na inferência filogenética. Uma vez que de acordo com Ogden e Rosenberg (2006), a topologia da árvore filogenética pode ser em algumas situações mais influenciada pela acuidade do alinhamento que pelo próprio método empregado para se reconstruir as genealogias. Conclui-se que de acordo com a natureza do banco de sequências a ser investigado, pode-se escolher entre diversos tipos de abordagem para se realizar o alinhamento múltiplo de sequências de maneira acurada.

Gong e colaboradores (2010) avaliaram as performances de alguns algoritmos de AMS amplamente utilizados (***Dialign***, ***Tcoffee***, ***ClustalW*** e ***Muscle***) por meio da medida de permutação de similaridade, *Score* que foca somente na ordem relativa de distâncias evolucionárias entre proteínas e rejeita pequenas diferenças destas distâncias, o que parece favorecer a robustez desta medida superando os ruídos nos dados analisados. Neste estudo, foi demonstrado que a precisão dos algoritmos varia de acordo com a natureza dos dados, foram testados em duas bases de dados diferentes ROSE e BALIBASE com diferentes performances, ***Tcoffee*** e ***Dialign*** são superiores em BALIBASE e ***Dialign*** é bem superior em ROSE. Também é

interessante que todas as abordagens heurísticas parecem perder acuidade em bancos de sequências muito divergentes, quando a similaridade média entre as sequências é inferior a 35%.

De acordo com Do e colaboradores (2010), o problema da construção de um alinhamento consiste na definição explícita ou implícita de uma função objetiva para avaliar a qualidade do alinhamento e empregar um algoritmo eficiente para encontrar o alinhamento ótimo. Um dos problemas na definição desta função é a estimativa das penalidades atribuídas às lacunas nos alinhamentos, que são baseadas em um esquema de tentativa e erro. Uma das maneiras de contornar a necessidade de estimar penalidades é utilizar um Modelo Oculto de Markov em que os parâmetros são obtidos de maneira probabilística de acordo com a natureza dos dados. Outro fator importante que deve ser considerado nesta metodologia é que abordagens heurísticas estão sujeitas a erros iniciais durante a construção de suas árvores guias que podem enviesar todo o alinhamento. Negativamente, abordagens probabilísticas baseadas em HMM apresentam complexidade computacional elevada o que dependendo do tamanho do banco de dados pode tornar o tempo de análise longo demais.

Portanto, diante da gama de algoritmos disponíveis o filogeneticista deve ter consciência sobre a natureza de seus dados e que a escolha do algoritmo a ser empregado no AMS deve ser consciente e não um ato falho e corriqueiro de “apertar botões” e seguir modismos. Sempre que possível é recomendado que os alinhamentos sejam observados levando em consideração o sentido biológico embargado nesta abordagem.

Referências

Birney, E. Hidden Markov models in biological sequence analysis. **IBM J. RES. & DEV**. 2001; 45: 449-454.

Do, C.B.; Mahabhashyam, M.S.; Brudno, M.; Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. **Genome Res.** 2005; 15(2):330-40.

Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Res**. 2004a; 32(5):1792-7.

Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. **BMC Bioinformatics**. 2004; 5:113.

Fonzo, V.D.; Aluffi-Pentini, F.A.; Parisi, V. Hidden Markov Models in Bioinformatics. **Current Bioinformatics**. 2007; 2:49-61.

Gong, Z.; Li, F.; Dong, L. Performance assessment of protein multiple sequence alignment algorithms based on permutation similarity measurement. **Biochem Biophys Res Commun.** 2010;399(4):470-4.

Gupta, S.K.; Kececioglu, J.D.; Schäffer, A.A. Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. **J Comput Biol.** 1995; 2(3):459-72.

Hall, B.G. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. **Mol Biol Evol.** 2005; 22(4):792-802.

Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic Acids Res.** 2002; 30(14):3059-66.

Needleman, S.B.; Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **J Mol Biol.** 1970; 48(3):443-53.

Ogden, T.H.; Rosenberg, M.S. Multiple sequence alignment accuracy and phylogenetic inference. **Syst Biol**. 2006; 55(2):314-28.

Song J, Liu C, Song Y, Qu J, Hura GS. Alignment of multiple proteins with an ensemble of hidden Markov models. **Proceedings of the Sixth International Conference on Machine Learning and Applications.** 2007.

Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Res.** 1994; 22(22):4673-80.