Alinhamentos múltiplos: alguns problemas, suas implicações e soluções

Vitor Ferreira Onuchic

Resumo

Existem diversas aplicações para alinhamentos múltiplos de sequências, sendo estes amplamente utilizados. Assim, as consequências oriundas de erros nestes alinhamentos nas análises subsequentes podem ser graves. Nesta resenha discutiremos sobre as possíveis implicações desses erros, e apresentamos algumas abordagens que têm como objetivo a melhoria na acurácia desses alinhamentos. As abordagens apresentadas aqui são a utilização de modelos evolutivos em pair-HMMs e a utilização de modelos pair-HMM mais complexos, que levam em conta a diferente taxa de evolução de partes distintas das sequências. Além disso, relataremos ainda como inserções acabam sendo mais penalizadas quando técnicas de alinhamento múltiplo progressivo são utilizadas, e apresentaremos uma solução para este problema, apresentada por Löytynoja e Goldman.

Introdução

Desde que a teoria da evolução de Darwin começou a ser amplamente aceita, tem sido de grande interesse científico a idéia de reconstruir as relações evolutivas entre os organismos. Com o desenvolvimento da biologia molecular e com a postulação do dogma básico, foi desenvolvido um modelo de evolução baseado na alteração física da sequência de ácidos nucléicos dos cromossomos. Assim, a partir do começo da década de 1960, sequências moleculares têm sido uma importante base para estudos de evolução e de filogenia.

Uma vez que podemos abordar a evolução a partir do mecanismo de modificação da sequência nucleotídica de nossos genomas, podemos modelar o processo evolutivo como um processo de edição, com inserções, alterações e remoções de sequências de ácidos nucléicos. Para estimar a distância entre duas sequências, podemos assim partir de uma alinhamento de suas bases. Desta maneira, algoritmos de alinhamento de sequências têm sido parte fundamental destes estudos. Esse tipo de algoritmo visa, em geral, parear caracteres homólogos, sendo eles nucleotídeos ou aminoácidos. Esse problema se torna mais complicado devido à mutações, inserções e deleções que acontecem nas sequências no curso da evolução de cada organismo.

Além de informações filogenéticas, alinhamentos podem ainda prover informações a respeito da função e da estrutura de determinada sequência. Isso é feito através da comparação com sequências de função ou estrutura conhecidas e pareamento de regiões reconhecidamente importantes para a determinação dessas características com regiões da nova sequência.

Existem diversas técnicas para realizar esse tipo de tarefa. O algoritmo clássico [5] utiliza uma pontuação para cada par de nucleotídeos ou aminoácidos alinhados, de forma que sequências mais similares e alinhamentos melhores obtêm uma pontuação

maior. Uma técnica bastante similar para solucionar esse tipo de problema, mas que apresenta um maior formalismo matemático, e possibilita a utilização de métodos estatísticos tanto para calibrar as pontuações mais adequadamente, quanto para analisar os resultados, é a utilização de modelos pair-HMM [1].

Muitas vezes, entretanto, estamos interessados em comparar diversas sequências ao mesmo tempo, e identificar nestas qual foi o processo pelo qual essas sequências divergiram a partir de um ancestral comum. Para isso, é necessário descobrir quais dos caracteres presentes nas sequências são homólogos, e quais foram os eventos de inserção e deleção que ocorreram. Esse tipo de comparação entre sequências, chamado alinhamento múltiplo, é muito utilizado quando deseja-se estimar árvores filogenéticas de organismos, ou em estudos de genômica comparativa, em que queremos identificar as modificações que ocorreram nos genomas de vários organismos que levaram ao aparecimento de determinadas características. Muitas outras aplicações são também conhecidas.

Em todas essas utilizações, entretanto, é de suma importância que os alinhamentos sejam feitos de maneira correta. Wong *et al.* mostrou recentemente algumas possíveis consequências da utilização de alinhamentos múltiplos incorretos. Neste mesmo artigo, mostra ainda que as diferentes abordagens existentes para este problema acabam gerando resultados bastante discrepantes [6]. Este tema será discutido mais a fundo nesta resenha.

Não se conhece, até o momento, nenhuma maneira de resolver o problema de alinhamento de diversas sequências de maneira ótima com complexidade diferente de exponencial no número de sequências [1]. Este fato torna, portanto, esse tipo de solução inviável. Assim, algumas técnicas que utilizam heurísticas tiveram que ser desenvolvidas para gerar alinhamentos múltiplos da maneira mais eficiente e correta possível. A solução mais amplamente utilizada para este problema é chamada alinhamento múltiplo progressivo. Nesta técnica alinhamentos múltiplos são gerados a partir de alinhamentos sucessivos de pares de sequências. Inicialmente, um par de sequências é alinhado, em seguida, uma terceira sequência é alinhada ao alinhamento gerado pelo par de sequências iniciais, e esse processo é iterado até que todas as sequências tenham sido alinhadas [1].

Recentemente Löytynoja e Goldman apontaram o fato de que esse tipo de abordagem para realizar alinhamentos múltiplos acaba penalizando eventos de inserção mais de uma vez, gerando alinhamentos finais com uma proporção de substituições e deleções diferente da correta [4]. Neste mesmo artigo, propõe ainda uma maneira de corrigir este tipo de erro, e mostram as consequências geradas por esta modificação nos alinhamentos. Este tema será também discutido mais a fundo nesta resenha.

Estes mesmos autores sugerem ainda que a utilização de informações sobre a estrutura das sequências analisadas, podem ser de grande ajuda no melhoramento da acurácia de alinhamentos, já que diferentes estruturas em um gene, por exemplo, têm taxas de evolução bastante distintas, devido à pressão seletiva não homogênea. Neste artigo, mencionam ainda a utilização de modelos pair-HMM que levam em conta a distância evolutiva entre as sequências em questão [3].

2. Problemas relacionados à alinhamentos múltiplos

Nesta sessão descreveremos inicialmente o problema de penalização excessiva de inserções quando é utilizada a técnica de alinhamento múltiplo progressivo. Em seguida discorreremos sobre as implicações da não correção, ou correção inadequada deste problema para os alinhamentos múltiplos. Finalmente serão discutidas algumas possíveis consequências da utilização de alinhamentos múltiplos incorretos. Essas implicações serão analisadas a partir do experimento feito por Wong *et al.*.

2.1. Problema de penalização excessiva de inserções [4]

Um problema mais específico apresentado por diversas técnicas de alinhamento múltiplo ressaltado por Löytynoja e Goldman, é o problema da penalização excessiva de eventos de inserção. Esse é um problema que é inerente da metodologia que faz alinhamentos múltiplos progresivamente, que é a metodologia básica da maioria das técnicas de alinhamento múltiplo utilizadas na atualidade.

Nesta técnica o alinhamento múltiplo é feito através do alinhamento de pares de sequências ou de alinhamentos de alinhamentos prévios em ordem decrescente de proximidade evolutiva, de acordo com uma dada árvore filogenética guia. Apesar de em alinhamentos de pares de sequencias gaps abertos devido a inserções não possam ser diferenciados daqueles abertos devido a eventos de deleções, quando fazemos diversos alinhamentos consecutivos, gaps associados à deleções são penalizados apenas uma vez, enquanto gaps que aparecem por causa de inserções têm que ser abertos em todas as iterações seguintes. Cada vez que um gap é aberto, há uma penalidade associada a isso, assim, um único evento de inserção é penalizado várias vezes. Esse fenômeno é evidenciado na Figura 1 (A e B).

Algumas heurísticas previamente desenvolvidas para tentar solucionar este problema dão uma penalidade menor para a abertura de gaps em posições que já têm algum gap aberto. Apesar de esse tipo de abordagem reduzir a penalidade das inserções, ela gera um problema quando acontecem múltiplos eventos de inserção ou deleção próximos, já que as aberturas de gaps vão ficar sistematicamente enviesadas no sentido de serem colocadas sempre no mesmo lugar. Além disso, esse alinhamento forçado de gaps, levará ao alinhamento de resíduos não homólogos. A consequência disso será uma superestimação da taxa de substituição de resíduos, e também leva a uma subestimação do tamanho da sequência ancestral. Esses eventos estão representados na Figura 1 (C e D). Uma solução diferente para esse problema foi sugerida por Löytynoja e Goldman e será descrita mais à frente.

2.2. Implicações de erros em alinhamentos em análises posteriores [6]

Alinhamentos múltiplos de sequências são, como mencionado anteriormente, a base para diversos tipos de estudos. Esses envolvendo genômica comparativa, construção de árvores filogenéticas, entre outros. Até recentemente, entretanto, essas análises eram, em geral realizadas sobre alinhamentos curados, e manualmente editados. O procedimento de geração desse tipo de alinhamento, entretanto, é bastante demorado, e depende de profissionais bastante treinados. Assim, esse tipo de alinhamento tem se

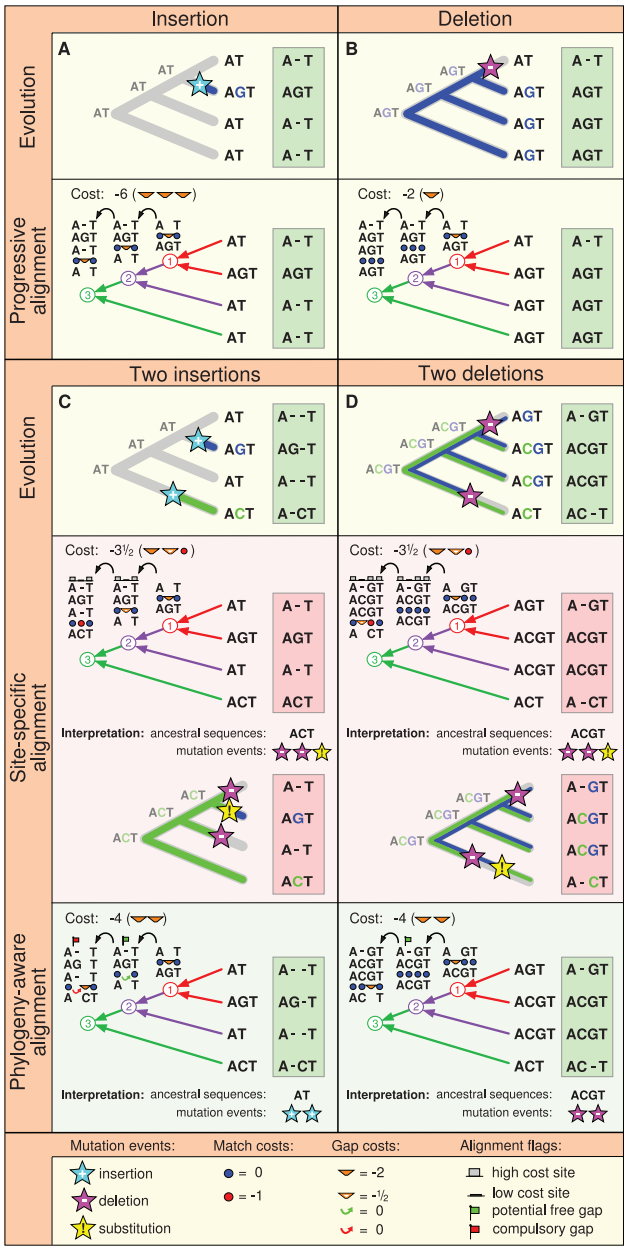

Figura 1. Inserções e deleções não são equivalentes em alinhamentos múltiplos progressivos. **(A)** Inserções necessitam que um gap seja aberto em cada uma das iterações de um alinhamento múltiplo progressivo. Assim, a implementação ingênua desse algoritmo penaliza esse evento diversas vezes. **(B)** Uma deleção é penalizada apenas uma vez. **(C e D)** Implementações que diminuem a penalidade de abertura de um gap quando um gap já foi aberto naquela mesma posição anteriormente. Isso leva ao pareamento incorreto de evento independentes, no caso de mais de um evento de inserção ou deleção próximos. Isso leva ainda ao pareamento de resíduos que na realidade não deveriam estar pareados. Esses dois fatos levam a uma estimativa aumentada da taxa de substituição e do tamanho da sequência ancestral (quadros rosas). A solução sugerida por Löytynoja e Goldman, representada nos quadros verdes, consegue diferenciar entre inserções e deleções através da comparação com sequências próximas, e permite que permite que gaps sejam abertos sem custo em posições em que um gap já foi aberto anteriormente devido a uma inserção, mas não quando esse gap é proveniente de uma deleção. **(referencia)**

tornado cada vez mais difícil de ser obtido, devido à grande quantidade de dados gerados e, consequentemente, de sequências a serem alinhadas. Métodos automatizados de alinhamento, apesar de já terem evoluído bastante, podem não produzir alinhamentos corretos, principalmente quando trata-se de alinhamentos entre sequências bastante divergentes.

Wong em 2008 fez um estudo que evidencia as implicações da utilização de alinhamentos gerados automaticamente, sem levar em conta a incerteza associada a eles. Neste estudo Wong selecionou um conjunto de genes ortólogos de sete espécies diferentes de leveduras. Foram selecionados apenas genes que apareciam em todas as sete espécies. Esse conjunto consistia em 1502 grupos de 7 genes ortólogos. Em seguida foi realizado o alinhamento múltiplo entre cada um desses grupos utilizando 7 programas

diferentes com essa mesma finalidade. Esses programas eram: Clustal W, Muscle, T-Coffee, Dialign 2, Mafft, Dca, e ProbCons.

Dois tipos de análise comumente feitas utilizando alinhamentos múltiplos foram então realizadas para os alinhamentos gerados. A primeira análise foi uma estimação da árvore filogenética dos organismos envolvidos no estudo. Neste experimento, uma árvore foi estimada, através do método da máxima verossimilhança, para cada alinhamento múltiplo gerado. Foi observado que em poucos casos as árvores geradas a partir do alinhamento de um mesmo conjunto de genes ortólogos eram as mesmas para cada método de alinhamento. Havia casos em que os sete métodos de alinhamentos diferentes geravam seis árvores filogenéticas diferentes a partir de um mesmo conjunto de genes ortólogos. Foi observado entretanto, que as árvores geradas a partir de conjuntos de genes mais similares eram mais consistentes, devido, provavelmente, ao fato de sequências mais similares serem mais fáceis de serem alinhadas.

A outra análise realizada sobre esses dados foi a busca por sítios de seleção positiva. Esses sítios são aqueles que são bastante conservados devido a uma pressão seletiva positiva agindo sobre eles. Para a maioria dos ORFs utilizados, nenhum sítio foi inferido como estando sobre seleção positiva fazendo a inferência a partir do alinhamento múltiplo gerado por qualquer método. Entretanto, foi possível observar que em 28,4% dos casos, as inferências feitas utilizando os alinhamentos gerados a partir de diferentes métodos, não davam resultados consistentes, já que o número de sítios sobre seleção positiva inferidos não era o mesmo. Esta proporção dependia do valor do limiar adotado para identificar um sítio como sobre pressão seletiva, mas mesmo utilizando um limiar de 95%, ainda havia inconsistência entre as inferências feitas a partir de cada alinhamento em 14,8% dos casos.

Podemos perceber, portanto, que diferenças nos alinhamentos obtidos, podem ter consequências sérias nas análises feitas posteriormente. Pôde-se perceber ainda que a variabilidade encontrada entre os alinhamentos feitos por diferentes métodos aumenta muito quando a divergência entre as sequências alinhadas aumenta. Assim, métodos que gerem alinhamentos melhores quando as sequências são bastante divergentes são necessários, assim como é necessário levar em conta essa incerteza dos alinhamentos quando fazemos qualquer tipo de análise utilizando esses dados.

3. Abordagens para melhorar a qualidade de alinhamentos múltiplos

Nesta sessão mostraremos algumas abordagens utilizadas para melhorar a qualidade de alinhamentos gerados através da técnica de alinhamentos múltiplos progressivos. Descreveremos inicialmente a abordagem de Löytynoja e Goldman para tentar solucionar o problema da penalização excessiva de inserções. Discorreremos ainda sobre modelos de pair-HMM para alinhamentos de pares de sequências que levam em conta a diferença na taxa de evolução entre diferentes fragmentos de uma sequência, assim como o fato de, no caso de alinhamento de genes, haver regiões codificadoras e não codificadoras. A maneira como as probabilidades de emissão e transição entre os estados do modelo são calculadas neste trabalho é também descrita, já que calcula as pontuações utilizadas para fazer alinhamentos de pares de sequências a partir da distância evolutiva entre as sequências envolvidas, abordagem que pode melhorar os alinhamentos gerados.

3.1. Solucionando o problema de penalização de inserções [4]

No trabalho de Löytynoja e Goldman, é sugerida uma possível solução para o problema de penalização excessiva de inserções. Essa abordagem é chamada alinhamento atento à filogenia e é implementada em uma alinhador múltiplo chamado PRANK$_{+F}$. Nesta metodologia, quando gaps são adicionados em uma etapa do alinhamento múltiplo, eles são marcados como possíveis inserções. Este alinhamento é então alinhado à sequência mais próxima à ele. Essa comparação serve então como base para decidir se aquele gap foi gerado por uma inserção ou por uma deleção. No caso de ele ser, de fato, uma inserção, esta posição é então marcada permanentemente como inserção, e novos gaps podem ser adicionados ali sem penalidade nas próximas iterações. Além disso, como um evento de inserção gera resíduos que não são homólogos à nenhum outro, essas posições marcadas como inserções, são proibidas pelo algoritmo de serem pareadas com outros resíduos. Dessa forma, eventos diferentes de inserção são contados separadamente, mesmo quando eles ocorrem na mesma posição. Por outro lado, quando a comparação com sequências próximas mostra que o gap parece ter sido gerado por uma deleção, a marcação é removida, e a penalidade para abertura de gaps naquela posição é mantida, fazendo com que o efeito seja direcionado apenas para inserções.

Para validar esta abordagem, um conjunto de teste foi gerado sinteticamente. Neste conjunto foram geradas sequências de acordo com árvores filogenéticas contendo 16, 32 e 64 sequências diferentes. Essas sequências foram geradas utilizando parâmetros evolutivos realistas, que mimetizavam a evolução do DNA em regiões sem restrições funcionais ou estruturais. Além disso, elas foram feitas de maneira que no alinhamento real a quantidade de inserções e deleções fosse a mesma. São simuladas nesses

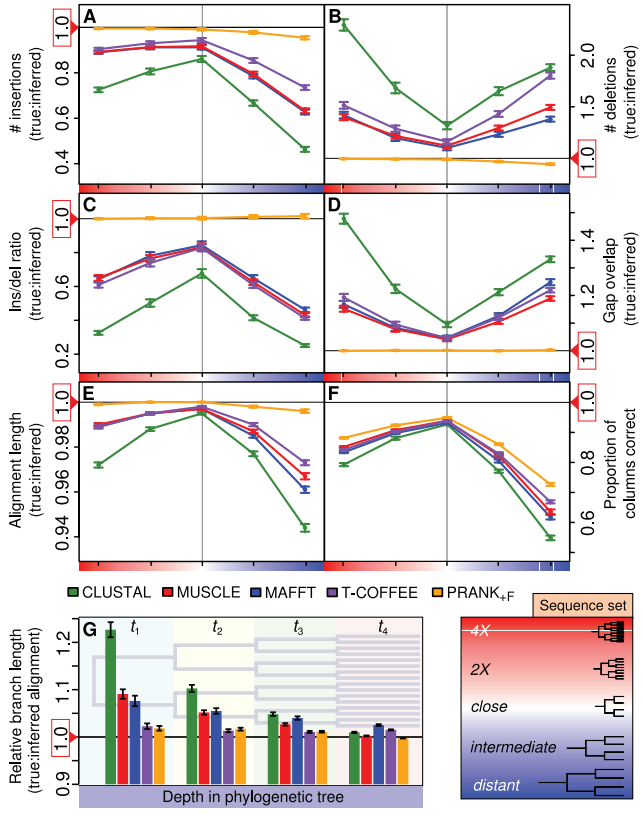

Figura 2. (A,B e C) Pôde-se observar que todos os programas existentes sub-estimavam a quantidade de inserções, e superestimavam o número de deleções, assim, apresentando um valor incorreto para a razão inserção/deleção. **(D)** É representado uma medida da quantidade de gaps que foram alinhados, quando não deveriam ser. **(E)** Razão entre o tamanho do alinhamento encontrado por cada método e o tamanho real. **(F)** Proporção de colunas completamente corretas no alinhamento múltiplo. **(G)** As barras representam a razão entre o tamanho dos ramos das árvores inferidos a partir de cada metodologia e o tamanho real desses ramos, para diferentes profundidades nas árvores (t1 sendo o mais profundo e t4 o menos). Nas figuras (A) até (F), o gradiente branco-azul significa um aumento na distância evolutiva entre as sequências utilizadas, enquanto o gradiente branco-vermelho significa um aumento no número de sequências utilizadas para uma mesma distância evolutiva.

experimentos árvores com as sequências mais próximas e mais distantes evolutivamente. As sequências de cada árvore foram então alinhadas utilizando diversos programas que utilizam a técnica de alinhamento múltiplo progressivo (CLUSTAL W, MAFFT, MUSCLE and T-COFFEE), além daquele implementado neste trabalho (PRANK$_{+F}$). Diversas medidas de acurácia dos alinhamentos foram feitas, e os resultados estão apresentados na Figura 2. Pôde-se observar que a nova metodologia tem resultados melhores do que todas as outras em todos os quesitos. Além disso, é possível observar que todas as metodologias, exceto PRANK$_{+F}$, falham tanto quando a distância evolutiva aumenta, quanto quando a quantidade de sequências analisadas aumenta e mantendo a distância evolutiva constante. Isso é especialmente interessante, pois uma técnica usualmente utilizada em análises filogenéticas para melhorar supostamente a qualidade de alinhamentos múltiplos com sequências com grande distância evolutiva, é a utilização de sequências de distância evolutiva intermediária para auxiliar os programas. Entretanto, podemos ver neste resultado, que esse aumento no número de sequências não parece auxiliar o alinhamento.

3.2. Modelos de pair-HMM sensíveis à diferenças entre as taxas de evolução de partes das sequências [3]

Os pair-HMMs são uma variação do modelo HMM especialmente útil para encontrar alinhamentos entre pares de sequências e para avaliar a relevância dos alinhamentos encontrados. Diferente dos HMMs tradicionais, que geram apenas uma única sequência, os pair-HMMs geram um par de sequências alinhadas.

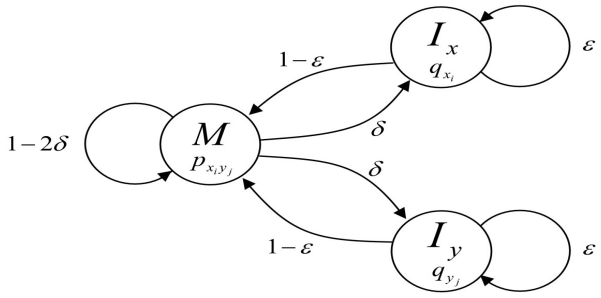

Figura 3. Modelo simples de pair-HMM

Consideremos por exemplo o pair-HMM representado na Figura 3. Esse é um modelo muito simples de pair-HMM, mas é bastante ilustrativo. Nele, são geradas simultaneamente duas sequências alinhadas $x = x_1 x_2 ... x_{L_1}$ e $y = y_1 y_2 ... y_{L_2}$. O estado I_x emite um único símbolo não alinhado x_i na sequência x. Da mesma maneira, o estado I_y também emite um símbolo y_j não alinhado, mas dessa vez na sequência y. Por fim o estado M emite um par de símbolos x_i e y_j alinhados, tais que x_i é emitido na sequência x e y_j é emitido na sequência y [1].

Podemos notar que existe nesse modelo uma relação um pra um entre a sequência de estados e o par de sequências alinhadas emitido. Com isso, o problema de encontrar o alinhamento ótimo entre as duas sequências x e y fica reduzido a encontrar o caminho

ótimo nesse modelo. Essa sequência ótima pode ser encontrada utilizando uma extensão do algoritmo de Viterbi [1].

Uma das vantagens de se usar pair-HMM ao invés de algoritmos tradicionais de alinhamento, é que com esse modelo calculamos a probabilidade de duas sequências serem relacionadas sem que seja preciso nos comprometermos com um alinhamento específico, o que é muito útil quando a similaridade entre elas é baixa e é consequentemente difícil de se identificar o alinhamento correto. Isso pode ser feito eficientemente através do algoritmo forward. Além disso, podemos ainda calcular através de pair-HMMs a probabilidade *a posteriori* de duas bases estarem alinhadas utilizando o algoritmo forward-backward [1].

Modelos de pair-HMM mais complexos que utilizem parâmetros mais adequados para cada região alinhada podem ser utilizados para melhorar a qualidade dos alinhamentos obtidos. Um exemplo disso seriam os pair-HMMs utilizados no trabalho de Löytynoja e Goldman. Esse tipo de modelo está representado na Figura 4. Nesta figura pode-se observar que há basicamente dois pair-HMMs combinados. Cada um desses modelos reconheceria, por exemplo, regiões que evoluem mais rapidamente e mais lentamente, dependendo das probabilidades de emissões e transições utilizadas em cada um deles.

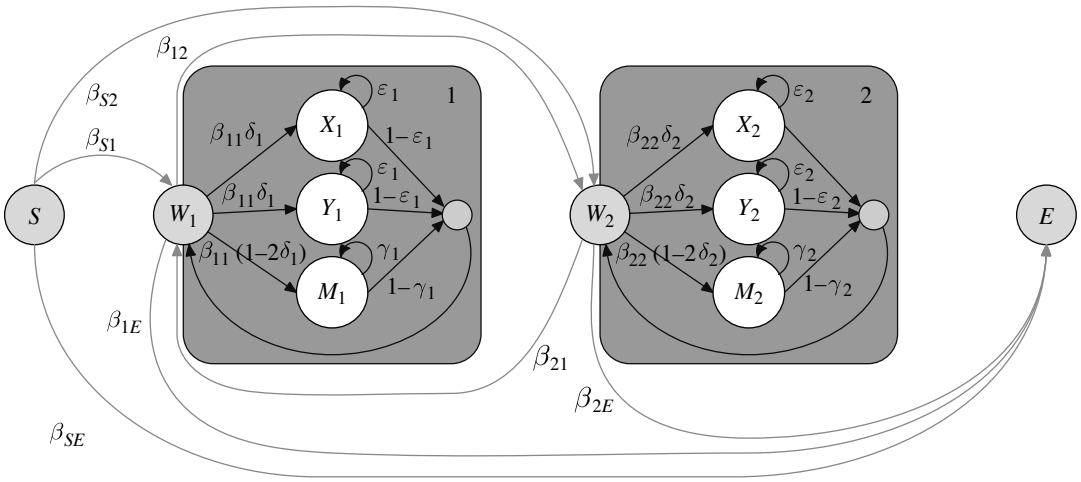

Figura 4. Modelo com cada um das regiões em cinza escuro descrevendo um processo evolutivo distinto.

Neste trabalho, além de serem utilizados modelos diferentes para diferentes regiões das sequências, as probabilidades de transição para um estado de gap e de emissão para cada um dos estados são ainda calculadas baseando-se na distância evolutiva entre as sequências, e no caso das probabilidades de emissão de matrizes 4x4, com seus elementos representando a taxa de mutação instantânea entre cada um dos nucleotídeos definidas inicialmente. Isso permite, em um alinhamento múltiplo progressivo, que as probabilidades de transição e emissão do modelo sejam ajustadas em cada iteração, de forma que a distância evolutiva entre as sequências sendo alinhadas naquele momento seja levada em conta. Como deseja-se modelar dois processos evolutivos distintos, as matrizes de taxa de mutação instantânea entre nucleotídeos para cada um dos processos em questão pode não ser a mesma.

Para validar essa abordagem, foram alinhados os genes CAPZA2 de 15 organismos diferentes, junto com 500 bases que de fato flanqueiam esses genes de cada lado dessas sequências. Foram utilizados dois modelos diferentes. O primeiro é aquele apresentado na figura 4. Neste caso, as probabilidades foram ajustadas de forma que um dos sub-modelos descrevesse regiões de rápida evolução e o outro descrevesse regiões de evolução mais lenta. Essas probabilidades eram estimadas a partir do conhecimento prévio que havia sobre a taxa de mutação, frequência de gaps, e duração média de cada tipo de região, além da distância evolutiva entre as sequências. O segundo modelo utilizado representava, além de regiões de evolução mais rápida e mais lenta, regiões codificadoras, em que cada nucleotídeo de um códon era modelado independentemente. Uma representação gráfica simplificada desse modelo está apresentada na Figura 5 (B e C). O treinamento desse último modelo foi também realizado através de conhecimento prévio sobre a estrutura das sequências. Os alinhamentos foram feitos utilizando o programa PRANK descrito anteriormente.

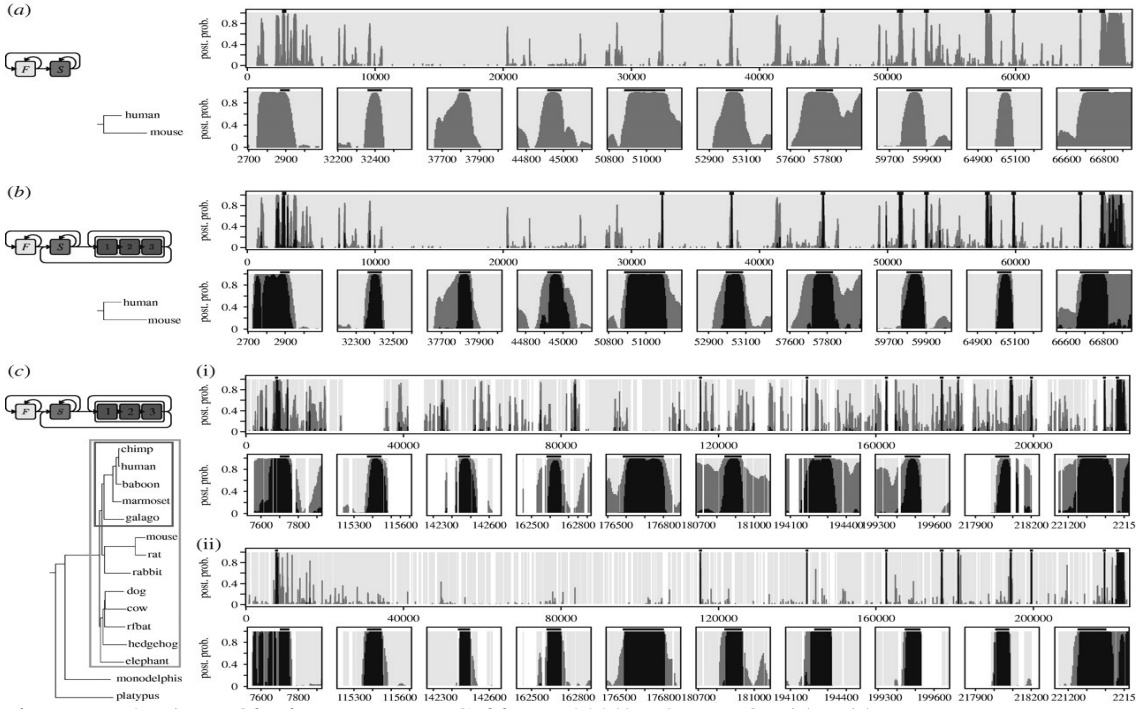

Figura 5: Os painéis (a) a (c) mostram a probabilidade a posteriori de diferentes classes de estruturas no alinhamento completo (gráfico do topo) e nas regiões em torno dos éxons codificantes de proteínas. (gráfico inferior) . Em (a) e (b), respectivamente, os modelos FAST/SLOW e CODON são usados para alinhar sequências de Humano e Camundongo; em (c) o modelo CODON é utilizado para linhar 15 sequências de mamíferos. As cores cinza-claro, cinza-escuro e preto representam, respectivamente, os estados que modelam as regiões de elução rápida, lenta região codificante. Em (c) a adição de sequências mais distantes aumenta a informação evolutiva e tem como consequência que as regiões com alta probabilidade dos estados de região codificante correspondem melhor às regiões dos éxons codificantes.(o painel (i) representa as 5 sequências do quadro menor e o painel (ii) as 15 do maior quadro na árvore guia). As regiões que correspondem a éxons conhecidos são indicadas por barras horizontais no topo dos gráficos.

Neste teste assume-se que alinhamentos melhores serão gerados quando o modelo correto de evolução for utilizado em cada região da sequência. Os resultados deste teste

estão apresentados na Figura 5. Podemos ver que para alinhamento de pares de sequência as regiões codificantes foram identificadas, entretanto, podemos notar que há diversas regiões em que o modelo de evolução lenta é utilizado, mas não há evidências de nenhuma estrutura que justifique isso. Podemos notar ainda no alinhamento de pares de sequências que o modelo de região codificante apresenta melhor resultado. O resultado é bastante melhorado quando tratamos de alinhamentos múltiplos. Neste caso, parece que os alinhamentos de sequências mais próximas acabam ajudando aquele entre sequências mais distantes, já que fornecem informação sobre a variação espacial do processo evolutivo.

Discussão

Pudemos notar nos diferentes trabalhos mencionados nesta resenha que softwares que realizam alinhamentos múltiplos de sequências têm importância fundamental na análise das sequências biológicas. Considerado o fato de que novas técnicas de sequenciamento permitem a geração de cada vez mais sequências, a importância desse tipo de programa só tende a aumentar.

Entretanto, podemos ver pelo trabalho de Wong *et al*. que a utilização desse tipo de técnica sem análise cuidadosa dos resultados gerados, pode levar a conclusões muitas vezes incorretas. Os resultados desse trabalho mostram ainda que as técnicas de alinhamento múltiplo automatizado precisam ainda ser melhoradas, já que num futuro próximo, se não já no presente, será impossível fazer correções manuais nos alinhamentos gerados devido à enorme quantidade de sequências utilizadas.

Podemos ainda questionar se os requisitos computacionais das abordagens para alinhar diversas sequências são adequados para tratar essa nova ordem de grandeza do número de sequências. Uma análise sobre isto é feita em [2]. A partir dela pode-se notar que a maioria das técnicas não têm essa capacidade.

Apesar dessa dificuldade em lidar com a nova quantidade de dados, pudemos ver que as técnicas de alinhamento múltiplo têm também evoluído bastante. A acurácia dessas novas metodologias parece ser muito maior do que a das técnicas inicialmente utilizadas.

É interessante notar ainda que o software de alinhamento múltiplo mais utilizado continua sendo o CLUSTALW [4], mesmo este sendo superado por todas as novas metodologias para realizar este tipo de tarefa. Esta é uma situação que precisa ser mudada, haja vistas as consequências da utilização de alinhamentos múltiplos incorretos.

Referências

[1] R . Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, Biological sequence analysis: Probabilistic models of proteins and nucleic acids, The press syndicate of the University of Cambridge, 1998.

[2] K. Liu, C. R. Linder, T. Warnow, Multiple sequence alignment: a major challenge to large-scale phylogenetics, PLoS Curr. 2010 November 19; 2: RRN1198.

[3] A. Löytynoja, N.Goldman, A model of evolution and structure for multiple sequence

alignment, Phil. Trans. R. Soc. B 363 (2008), 3913-3919

[4] A. Löytynoja, N.Goldman, Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis, Science, Vol. 320, 2008.

[5] S. B. Needleman; and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48 (1970), 443–53

[6] K.M. Wong, M.A. Suchard, J.P. Huelsenbeck Alignment Uncertainty and Genomic Analysis; Science, Vol. 319, 2008.