Trabalho Recuperação

- Victor Azadinho Miranda
- RA: 171042191

Alinhamento Múltiplo de Sequências

Um alinhamento de sequência múltipla (ASM, em português: **a**linhamento de **s**equência **m**últipla; ou MSA, em inglês: ***m**ultiple **s**equence **a**lignment*) é um alinhamento de sequência de três ou mais sequências biológicas, geralmente proteína, DNA ou RNA. Em muitos casos, supõe-se que o conjunto de entrada de sequências de consulta tenha um relacionamento evolutivo pelo qual compartilham uma ligação e são descendentes de um ancestral comum. A partir do ASM resultante, a homologia de sequência pode ser inferida e a análise filogenética pode ser conduzida para avaliar as origens evolutivas compartilhadas das sequências. As representações visuais do alinhamento como na imagem à direita ilustram eventos de mutação, como mutações pontuais (alterações de um único aminoácido ou nucleotídeo) que aparecem como caracteres diferentes em uma única coluna de alinhamento e mutações de inserção ou exclusão (indels ou lacunas) que aparecem como hífens em uma ou mais das sequências no alinhamento. O alinhamento de múltiplas sequências é freqüentemente usado para avaliar a conservação de sequências de domínios de proteínas, estruturas terciárias e secundárias e até mesmo aminoácidos ou nucleotídeos individuais.

Algorítmos

Existem vários métodos de alinhamento usados em várias sequências para maximizar as pontuações e a exatidão dos alinhamentos. Cada um é geralmente baseado em uma determinada heurística com uma visão do processo evolutivo. A maioria tenta replicar a evolução para obter o alinhamento mais realista possível para melhor prever as relações entre as sequências.

Construção de alinhamento progressivo É descrito um método de alinhamento progressivo que utiliza o algoritmo de alinhamento em pares de Needleman e Wunsch iterativamente para atingir os alinhamentos múltiplos de um conjunto de sequências de proteínas e para construir uma árvore evolutiva que descreve seu relacionamento. As sequências são assumidas a priori para compartilhar um ancestral comum, e as árvores são construídas a partir de diferentes matrizes derivadas diretamente dos múltiplos alinhamentos. O impulso do método envolve colocar mais confiança na comparação de sequências recentemente divergentes do que naquelas desenvolvidas no passado distante.

Em particular, esta regra é seguida: “uma vez que uma lacuna, sempre uma lacuna”. O método foi aplicado a três conjuntos de sequências de proteínas: 7 superóxido dismutases, 11 globinas e 9 sequências semelhantes à tirosina quinase. Alinhamentos múltiplos e árvores filogenéticas para esses conjuntos de sequências foram determinados e comparados com árvores derivadas por

tratamentos convencionais de pares. Em vários casos, o método progressivo levou a árvores que pareciam estar mais de acordo com as expectativas biológicas do que as árvores obtidas por métodos mais comumente usados.

Métodos iterativos Os desempenhos relativos de quatro estratégias para alinhar um grande número de sequências de proteínas foram avaliados por referência aos alinhamentos estruturais correspondentes de 54 famílias independentes. O alinhamento de múltiplas sequências de uma família foi construído por um determinado método a partir das sequências de estruturas conhecidas e seus homólogos, e o subconjunto consistindo nas sequências de estruturas conhecidas foi extraído de todo o alinhamento e comparado com a contraparte estrutural em um resíduo forma de resíduo. As penalidades de abertura e extensão foram otimizadas para cada família e método.

Cada um dos quatro métodos de alinhamento múltiplos deu alinhamentos significativamente mais precisos do que o método de pares convencional. Além disso, uma diferença clara no desempenho foi detectada entre três dos quatro métodos de alinhamento múltiplos examinados. O método progressivo mais popular atualmente classificado em pior lugar entre os quatro, e a estratégia iterativa aleatória que otimiza a pontuação de soma de pares classificada em segundo lugar. As duas estratégias de melhor desempenho, uma das quais foi desenvolvida recentemente, buscam uma pontuação ótima de soma de pares ponderada, em que os pesos dos pares foram introduzidos para corrigir representações desiguais de subgrupos em uma família.

O novo método usa iterações duplamente aninhadas para tornar o alinhamento, a árvore filogenética e os pesos dos pares mutuamente consistentes. Mais importante ainda, a melhoria na precisão dos alinhamentos obtidos por esses métodos iterativos em relação ao método par a par ou progressivo tende a aumentar com a diminuição da identidade de sequência média, implicando que o refinamento iterativo é mais eficaz para o alinhamento geralmente difícil de sequências remotamente relacionadas. Quatro matrizes de substituição de aminoácidos bem conhecidas também foram testadas em combinação com os vários métodos. No entanto, os efeitos das matrizes de substituição foram considerados menores no quadro de alinhamentos múltiplos, e a mesma ordem de desempenho relativo dos métodos de alinhamento foi observada com qualquer uma das matrizes.

Métodos de consenso A geração de alinhamentos de sequência múltipla (MSAs) é uma etapa crucial para muitas análises de bioinformática. Assim, melhorar a precisão do MSA e identificar erros potenciais em MSAs é importante para uma ampla gama de pesquisas pós-genômicas. É apresentado um novo método chamado MergeAlign, que constrói MSAs de consenso a partir de vários MSAs independentes e atribui uma pontuação de precisão de alinhamento a cada coluna.

Usando testes de benchmark convencionais, é demonstrado que, em média, os MSAs MergeAlign são mais precisos do que os MSAs gerados usando qualquer

matriz única de substituição de sequência. É mostrado que as pontuações da coluna MergeAlign estão relacionadas à precisão do alinhamento e, portanto, fornecem um método ab initio de estimar a precisão do alinhamento na ausência de MSAs de referência com curadoria. Usando dois testes de desempenho de alinhamento novos e independentes que utilizam um grande conjunto de famílias de genes ortólogos, é demonstrado que o aumento do desempenho do MSA leva a um aumento no desempenho das análises filogenéticas posteriores.

Usando vários testes de desempenho de alinhamento, é demonstrado que este novo método tem ampla aplicação geral na pesquisa biológica.

Modelos ocultos de Markov Modelos ocultos de Markov (HMMs) são um meio altamente eficaz de modelar uma família de sequências não alinhadas ou um motivo comum dentro de um conjunto de sequências não alinhadas. O HMM treinado pode então ser usado para discriminação ou alinhamentos múltiplos. A descrição matemática básica de um HMM e seu procedimento de treinamento de maximização de expectativa é relativamente simples.

São revisadas as extensões matemáticas e heurísticas que movem o método do teórico para o prático. Em seguida, analisa experimentalmente a eficácia da regularização do modelo, modificação do modelo dinâmico e estratégias de otimização. Finalmente, é demonstrado no domínio SH2 como um domínio pode ser encontrado a partir de sequências não alinhadas usando um tipo de modelo especial. O trabalho experimental foi concluído com o auxílio do software Sequence Alignment and Modeling.

Métodos que reconhecem a filogenia Algoritmos de programação dinâmica garantem encontrar o alinhamento ideal entre duas sequências. Para mais do que algumas sequências, algoritmos exatos tornam-se computacionalmente impraticáveis, e algoritmos progressivos que iteram alinhamentos de pares são amplamente usados. Esses métodos heurísticos têm uma séria desvantagem porque algoritmos de pares não diferenciam inserções de exclusões e acabam penalizando eventos de inserção única várias vezes.

Essa penalidade irrealisticamente alta para inserções normalmente resulta em superestimação de sequências e uma subestimação do número de eventos de inserção. É descrito como uma modificação do algoritmo de alinhamento tradicional que pode distinguir a inserção da exclusão e evitar a penalização repetida das inserções e ilustrar esse método com um modelo de Markov oculto de par que usa uma função de pontuação evolutiva. Em comparação com um método de alinhamento progressivo tradicional, nosso algoritmo infere um maior número de eventos de inserção e cria lacunas que são filogeneticamente consistentes, mas espacialmente menos concentradas. Nossos resultados sugerem que alguns "pontos quentes" de inserção / exclusão podem na verdade ser artefatos de algoritmos de alinhamento tradicionais.

Descoberta de *motif* Um sistema é descrito para encontrar e montar as regiões mais altamente conservadas de proteínas relacionadas para pesquisa de banco de dados. Primeiro, uma versão automatizada do algoritmo de Smith para encontrar motivos é usada para detecção sensível de alinhamentos locais múltiplos. Em seguida, os alinhamentos locais são convertidos em blocos e o melhor conjunto de blocos não sobrepostos é determinado.

Quando o sistema automatizado foi aplicado sucessivamente a todos os 437 grupos de proteínas relacionadas no catálogo PROSITE, resultaram 1764 blocos; eles podem ser usados para pesquisas muito confidenciais de bancos de dados de sequência. Cada bloco foi calibrado por meio de busca no banco de dados SWISS-PROT para obter uma medida da distribuição aleatória de correspondências, e os blocos calibrados foram concatenados em um banco de dados que poderia ser pesquisado. São fornecidos exemplos nos quais relacionamentos distantes são detectados usando um conjunto de blocos para pesquisar um banco de dados de sequências ou usando sequências para pesquisar o banco de dados de blocos. O uso prático do banco de dados de blocos é demonstrado pela detecção de relações até então desconhecidas entre oxidorredutases e pela avaliação de uma relação proposta entre a proteína Vif do HIV e as proteases tiol.

Alinhamento de sequência múltipla não codificante Os alinhamentos de sequência múltipla (MSAs) são geralmente pontuados sob a suposição de que as sequências sendo alinhadas evoluíram por descendência comum. Consequentemente, as diferenças entre as sequências refletem o impacto das inserções, deleções e mutações. No entanto, sequências de ligação de DNA não codificantes, tais como sítios de ligação de fator de transcrição (TFBSs), frequentemente não estão relacionadas por descendência comum e, portanto, os métodos de pontuação de alinhamento existentes não são bem adequados para alinhar tais sequências.

É apresentada uma nova metodologia de MSA múltipla que pontua sequências de DNA de TFBS incluindo a interdependência de bases vizinhas. São apresentadas duas variantes suportadas por diferentes hipóteses nulas subjacentes, uma estatisticamente e a outra gerada termodinamicamente. São avaliados os alinhamentos por meio de seu desempenho na previsão do TFBS; ambos os métodos mostram melhorias consideráveis quando comparados com algoritmos MSA padrão. Além disso, a hipótese nula gerada termodinamicamente supera a estatística devido à maior estabilidade na energia livre de empilhamento de base do alinhamento. O método de hipótese nula gerado termodinamicamente pode ser baixado de http://sourceforge.net/projects/msa-edna/

Bibliografia

"Multiple sequence alignment". Wikipedia, the free encyclopedia. Acessado em 28 de Agosto de 2020.

Feng, D., Doolittle, R.F. Progressive sequence alignment as a prereq-

uisitetto correct phylogenetic trees. J Mol Evol 25, 351–360 (1987). https://doi.org/10.1007/BF02603120

Osamu Gotoh, Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinement as Assessed by Reference to Structural Alignments, Journal of Molecular Biology, Volume 264, Issue 4, 1996, Pages 823-838, ISSN 0022-2836, https://doi.org/10.1006/jmbi.1996.0679

Collingridge, P.W., Kelly, S. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. BMC Bioinformatics 13, 117 (2012). https://doi.org/10.1186/1471-2105-13-117

Richard Hughey, Anders Krogh, Hidden Markov models for sequence analysis: extension and analysis of the basic method, Bioinformatics, Volume 12, Issue 2, April 1996, Pages 95–107, https://doi.org/10.1093/bioinformatics/12.2.95

An algorithm for progressive multiple alignment of sequences with insertions Ari Löytynoja, Nick Goldman Proceedings of the National Academy of Sciences Jul 2005, 102 (30) 10557-10562; DOI: 10.1073/pnas.0409137102

Steven Henikoff, Jorja G. Henikoff, Automated assembly of protein blocks for database searching, Nucleic Acids Research, Volume 19, Issue 23, 11 December 1991, Pages 6565–6572, https://doi.org/10.1093/nar/19.23.6565

Rafik A. Salama, Dov J. Stekel, A non-independent energy-based multiple sequence alignment improves prediction of transcription factor binding sites, Bioinformatics, Volume 29, Issue 21, 1 November 2013, Pages 2699–2704, https://doi.org/10.1093/bioinformatics/btt463