Projeto de pesquisa

Augusto Fernandes Vellozo

8 de março de 2010

Projeto de pesquisa apresentado como parte das exigências do Concurso Público de Provas e Títulos para o Magistério de Ensino Superior da Universidade Federal de São Paulo no campus de São José dos Campos.

Algoritmos eficientes são essenciais em muitas atividades na área de bioinformática, principalmente devido à grande quantidade de dados envolvida. Sendo assim, muitos aspectos teóricos da Ciência da Computação são muito importantes e, em alguns casos, até imprescindíveis para resolverem vários problemas nesta área com tempo e memória viáveis para o trabalho de um biólogo ou bioinformata.

1 Tópicos de pesquisa

1.1 Alinhamento de sequências com rearranjos

Na história da evolução alguns eventos introduzem mudanças nas sequências de DNA. Alguns eventos biológicos típicos são as substituições, remoções e inserções de nucleotídeos. Se é esperado identificar uma alta similaridade entre duas sequências então qualquer comparação destas sequências precisa levar em consideração a possibilidade da ocorrência desses

eventos. Um modo comum de se comparar sequências biológicas é através do alinhamento destas sequências. Procedimentos de alinhamento típicos tentam identificar que partes das sequências mudam e os eventos biológicos que originaram estas mudanças. Após, apresentam um alinhamento ótimo de acordo com algum critério de otimização e sistema de pontuação associado aos eventos.

Alinhamentos podem ser associados a um conjunto de operações de edição que transformam uma sequência em outra. Normalmente as únicas operações de edição consideradas são a substituição, a inserção e a remoção de um símbolo. Se os custos são associados a cada operação, existe um procedimento de programação dinâmica clássico, que em tempo e espaço[1] $O(n^2)$ computa o conjunto mínimo de operações de edição com o custo total mínimo e apresenta o alinhamento associado.

Um outro evento biológico que ocorre muitas vezes nas sequências de DNA é a inversão de trechos desta sequência. Neste evento um segmento da sequência é substituído pelo correspondente segmento reverso complementar. Nas sequências de DNA, consideramos que o símbolo A é o símbolo complementar de T e vice-versa. Da mesma forma, consideramos que C e G são complementares.

Dadas duas sequências e custos fixos para as operações de edição de substituição, inserção, remoção e inversão, o problema de alinhamento com inversões é um problema de otimização que indaga o custo total mínimo de um conjunto destas operações de edição que transforma uma sequência em outra. Além disso, pode também haver interesse na apresentação do alinhamento correspondente, ou seja das operações de edição que transformam uma sequência na outra.

Em [3] podemos ver que o problema de decisão, associado ao problema de obter um

[1] Consideraremos neste texto que n é o comprimento da maior sequência analisada

alinhamento com inversões para um alfabeto de tamanho ilimitado, é NP-difícil.

Em 1992, Schöniger e Waterman [12] introduziram uma hipótese simplificadora: todas as regiões envolvendo inversões não se sobrepõem. Isso levou ao problema do alinhamento com inversões sem sobreposições. Eles apresentaram uma solução exata em tempo $O(n^6)$ para esse problema.

Em 2003, Lago [7] propôs um algoritmo exato para o problema do alinhamento com inversões sem sobreposições que executa em tempo $O(n^4)$ e que usa espaço $O(n^2)$. Em 2005, Alves, Lago, e Vellozo [1] propuseram um algoritmo exato para este mesmo problema que executa em tempo $O(n^3 \log n)$ e que usa espaço $O(n^2)$. Em 2006, Vellozo, Alves, e Lago [14] propuseram um algoritmo exato para o mesmo problema que executa em tempo $O(n^3)$ e que usa espaço $O(n^2)$ para um sistema de pontuação das operações de edição que utiliza valores inteiros e fixos.

Outro evento muito comum em sequências de DNA é a duplicação, que introduz dentro da sequência uma cópia (ou repetição) de um segmento da própria sequência. Se a cópia de um segmento é introduzida exatamente depois ou antes deste segmento, ou seja numa posição adjacente, dizemos que a duplicação é em *tandem* (*tandem repeat*).

Ser capaz de identificar de maneira sistemática, e portanto exata os eventos de duplicação é um problema importante em bioinformática que ainda não foi resolvido de maneira satisfatória ou realmente eficaz, especialmente no caso de genomas de organismos eucariotas. Algumas repetições, como as compridas (onde cada cópia pode atingir centenas de nucleotídeos), são particularmente difíceis de serem identificadas, por causa do comprimento e devido a outros eventos ocorridos após a duplicação que podem ter alterado os trechos duplicados.

Em 1997, Benson [2] propôs um modelo para o alinhamento de sequências que considera

somente *tandem repeats* de mesmo tamanho. Ele propôs dois algoritmos exatos para obter um tal alinhamento ótimo. O primeiro algoritmo proposto executa em tempo $O(n^5)$ e espaço $O(n^2)$. O segundo algoritmo proposto executa em tempo $O(n^4)$ e espaço $O(n^3)$.

Durante o meu doutorado elaborei um algoritmo [13] que considera duplicações em qualquer parte da sequência onde não tenha havido outra duplicação (sem sobreposição) que executa em tempo $O(n^3)$ e espaço $O(n^2)$.

1.2 Anotação funcional

Atualmente existem muitos novos projetos de sequenciamento do genoma de organismos. Estes projetos geram uma grande quantidade de sequências de DNA. Um grande desafio é tentar compreender quais são as funções desempenhadas por estas sequências na vida do organismo.

Alguns trechos dessas sequências são classificados como genes. Um gene pode gerar uma ou mais proteínas que podem ter uma ou mais funções. Desta forma podemos atribuir a um gene uma ou mais funções. Estas funções podem ser classificadas, por exemplo, para transporte, regulação ou metabolismo.

Para anotar quais são as funções de um gene, muitas vezes utilizam-se técnicas e algoritmos que tentam encontrar sequências de genes de outras espécies com funções já conhecidas e com alto grau de similaridade com a sequência do gene da qual se quer descobrir a(s) função(ões). Desta forma, é transferida a função (ou as funções) de um gene similar já estudado ao gene que está sendo estudado. Existem vários métodos e algoritmos, como em [4] e [9], que tentam relacionar os genes que são similares a outros genes já estudados e assim possibilitar a transferência de suas funções.

A anotação computacional, ou em *in silico*, atribui uma função a um gene ou proteína

utilizando apenas métodos computacionais. Após a obtenção destas funções anotadas computacionalmente (*in silico*) alguns experimentos biológicos (*in vitro*) podem ser realizados, porém os experimentos *in vitro* são muito custosos e demorados. Desta forma, muitas vezes somente as anotações *in silico* são consideradas. Porém uma pequena diferença de poucos nucleotídeos pode gerar uma função diferente a um gene. Contudo, ainda não são conhecidas totalmente quais alterações nas sequências dos genes podem levar a funções diferentes destes genes. Portanto, a não realização de experimentos *in vitro* pode ocasionar uma propagação de erros nas anotações funcionais dos genes.

Desta forma, dentro do projeto do sequenciamento do genoma do inseto *Acyrthosiphon pisum* [5], desenvolvemos o aplicativo CycADS [15] que gera uma pontuação (ou o número de evidências) para cada anotação funcional do organismo, ou seja, para cada atribuição de uma função a um gene é dada uma pontuação. Esta pontuação da anotação tenta refletir o grau de confiabilidade desta anotação. Desta forma, esta pontuação tenta mostrar a necessidade ou não de experimentos *in vitro* para assegurar a existência da função anotada. Neste projeto somente as anotações funcionais metabólicas estão sendo analisadas através do aplicativo CycADS e a pontuação de cada anotação varia somente de acordo com a quantidade de métodos que concordam com a anotação.

Existem ainda muitas outras características que podem ser consideradas em projetos futuros para gerar uma pontuação mais precisa da confiabilidade de cada anotação funcional. Por exemplo, o sistema CycADS já está preparado para considerar no valor da pontuação de cada anotação, o grau de similaridade e os métodos utilizados na anotação funcional. Métodos com experimentos *in vitro* podem gerar pontuações maiores que métodos *in silico*, assim como funções em genes mais similares podem gerar pontuações maiores e com isto, dar maior confiabilidade a estas anotações.

1.3 *Motifs* e Redes metabólicas

Nas células de um organismo ocorrem várias reações químicas que transformam alguns compostos químicos em outros. A rede metabólica de um organismo contém todas as reações que podem ocorrer dentro de um organismo.

Após o sequenciamento e a anotação funcional do genoma de um organismo é possível conseguir a rede metabólica deste organismo. Uma das maneiras de se conseguir isto é através do aplicativo CycADS [15] que desenvolvemos.

Na tentativa de compreender melhor a rede metabólica de um organismo, muitos estudos e análises estão sendo realizados atualmente. Uma das maneiras de se realizar estes estudos e análises é através da procura e identificação de padrões (ou *motifs*) [8] na rede metabólica.

Podemos representar uma rede metabólica R como um grafo $G_R = (V,E)$ tal que $V = \{r | \forall r$ tal que r é uma reação metabólica em $R\}$ e $E = \{(r_1,r_2) | \forall r_1 \in V$ e $\forall r_2 \in V$ tal que existe pelo menos um composto (substrato ou produto) em r_1 que também é composto (substrato ou produto) em $r_2\}$. Podemos com isto procurar padrões (ou *motifs*) em G_R, como por exemplo buscar os subgrafos isomorfos a um dado subgrafo (padrão).

Dado um grafo $G = (V,E)$, um conjunto de cores C e uma função $c : V \rightarrow C$ que rotula (ou colore) os vértices de V, definimos *motifs* coloridos como um conjunto M de cores não necessariamente distintas, tal que toda cor em M é uma cor em C e existe um subgrafo conexo S de G tal que para toda cor $c \in C$ o número de ocorrências de c em M é igual ao número de vértices em S com a cor c.

Desenvolvemos um algoritmo inédito para encontrar todos os *motifs* coloridos de um tamanho fixo em uma rede metabólica e assim poder descobrir *motifs* coloridos que ocorrem numa quantidade de vezes surpreendentes (muitas ou poucas ocorrências) na rede metabólica de um certo organismo. De forma similar, o algoritmo pode ser usado também em redes de

interações entre proteínas (PPI).

Além dos *motifs* muitos outros aspectos são estudados nas redes metabólicas, como os modos elementares [11] e os conjuntos minimais de precursores de um conjunto de produtos alvo [6].

Outra análise que pode ser feita é a comparação da rede metabólica de dois organismos. Neste sentido, com a recente disponibilização da rede metabólica do inseto *Acyrthosiphon pisum* [5] e da já conhecida rede metabólica da bactéria *Buchnera aphidicola Aps* [10] podemos fazer de forma inédita a comparação das redes metabólicas de dois organismos que mantêm uma relação de simbiose .

2 Objetivos e atividades

Meu objetivo na área de pesquisa é investigar e obter resultados relevantes para a comunidade científica.

Atualmente, tenho interesse e condições de alcançar este objetivo trabalhando nos tópicos citados na seção 1 da seguinte forma:

- na elaboração de algoritmos e estruturas de dados eficientes que permitam tratar e extrair informações de sequências biológicas;
- no desenvolvimento de aplicativos e bases de dados das anotações funcionais dos novos genomas sequenciados;
- no desenvolvimento de algoritmos para a análise e comparação de redes metabólicas.

Devido ao meu passado profissional em desenvolvimento de softwares, outro tópico que também tenho interesse é o de Engenharia de Software. Dada a proximidade dos tópicos que já pesquisei, pretendo também realizar pesquisas em Recuperação de Informação. Além

disto, estou sempre aberto e disposto a trabalhar com outros tópicos que me tragam novos desafios.

Na área de pesquisa como docente na universidade, pretendo realizar as seguintes atividades:

- redigir e submeter artigos em publicações com reconhecimento internacional;
- realizar uma grande troca de conhecimentos científicos e acadêmicos com os pesquisadores da universidade;
- estabelecer contato com pesquisadores brasileiros e estrangeiros da área, assim como estreitar e fortalecer os relacionamentos já existentes e
- orientar alunos de graduação e pós-graduação interessados em realizar pesquisas.

3 Metodologia

As implementações dos algoritmos desenvolvidos serão desenvolvidos em microcomputadores. No caso de algoritmos com foco em dados biológicos, os dados processados serão sequências biológicas reais obtidas de bancos de dados públicos ou através de algum biólogo com o qual haja contato.

Quaisquer informações e artigos a serem utilizados na pesquisa serão obtidos preferencialmente através de meios eletrônicos e em formato digital. Caso isto não seja possível então serão utilizados os periódicos e publicações arquivados na universidade.

Os resultados obtidos com os algoritmos aqui propostos serão comparados com aqueles obtidos de eventuais programas já disponíveis.

Também pretendo obter retorno da relevância e coerência dos resultados obtidos a partir de diálogos com biólogos e pesquisadores com os quais eu tenha contato.

A orientação de alunos interessados em pesquisas será também utilizada para auxiliar na execução das atividades a serem desenvolvidas no campo da pesquisa.

Referências

[1] C. E. R. Alves, A. P. do Lago, and A. F. Vellozo. Alignment with non-overlapping inversions in $O(n^3 \log n)$-time. In *Proceedings of GRACO2005*, volume 19 of *Electron. Notes Discrete Math.*, pages 365–371 (electronic), Amsterdam, 2005. Elsevier.

[2] G. Benson. Sequence alignment with tandem duplication. In *RECOMB '97: Proceedings of the first annual international conference on Computational molecular biology*, pages 27–36, New York, NY, USA, 1997. ACM Press.

[3] X. Chen, J. Zheng, Z. Fu, P. Nan, Y. Zhong, S. Lonardi, and T. Jiang. Assignment of orthologous genes via genome rearrangement. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2(4):302–315, 2005.

[4] C. Claudel-Renard, C. Chevalet, T. Faraut, and D. Kahn. Enzyme-specific profiles for genome annotation: PRIAM. *Nucl. Acids Res.*, 31(22):6633–6639, 2003.

[5] A. T. I. A. G. Consortium. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol*, 8(2):e1000313, 02 2010.

[6] L. Cottret, P. Vieira Milreu, V. Acu na, A. Marchetti-Spaccamela, F. Viduani Martinez, M.-F. Sagot, and L. Stougie. Enumerating precursor sets of target metabolites in a metabolic network. In *WABI '08: Proceedings of the 8th international workshop on Algorithms in Bioinformatics*, pages 233–244, Berlin, Heidelberg, 2008. Springer-Verlag.

[7] A. P. do Lago, I. Muchnik, and C. Kulikowski. A sparse dynamic programming algorithm for alignment with non-overlapping inversions. *Theor. Inform. Appl.*, 39(1):175–189, 2005.

[8] V. Lacroix, C. G. Fernandes, and M.-F. Sagot. Motif search in graphs: Application to metabolic networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 3(4):360–368, 2006.

[9] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*, 35(Web Server issue):W182–5, 2007.

[10] M. D. Prickett, M. Page, A. E. Douglas, and G. H. Thomas. BuchneraBASE: a post-genomic resource for Buchnera sp. APS. *Bioinformatics*, 22(5):641–642, 2006.

[11] S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems (JBS)*, 2(2):165–182, 1994.

[12] M. Schöniger and M. S. Waterman. A local algorithm for DNA sequence alignment with inversions. *Bulletin of Mathematical Biology*, 54(4):521–536, Jul 1992.

[13] A. F. Vellozo. *Alinhamento de sequências com rearranjos*. PhD thesis, USP- Universidade de São Paulo, Abril 2007. `http://www.teses.usp.br/teses/disponiveis/45/45134/tde-04052007-185842`.

[14] A. F. Vellozo, C. E. R. Alves, and A. P. do Lago. Alignment with non-overlapping inversions in $o(n^3)$-time. In *6th Workshop on Algorithms in Bioinformatics*. Springer, 2006. Lecture Notes in Bioinformatics 4175.

[15] A. F. Vellozo and S. Colella. Cycads - cyc annotation database system. 2009. `http://code.cycadsys.org`.