BlastPhen
Agrupamento por Similaridade com Genomas Completos

Aluno: Ricardo Nishikido Pereira

Orientadores:
Marco Dimas Gubitoso (IME) e Paolo Zanotto (ICB)

Universidade de São Paulo

São Paulo, 2004

Resumo

Esta é a proposta de monografia sobre uma iniciação científica realizada na área de bioinformática, orientada pelos professores Paolo Zanotto[1] e Marco Dimas Gubitoso[2].

1 Introdução

A crescente disponibilização de seqüências de genomas completos permitiu o advento da genômica comparativa e conseqüentemente a sistemática genômica. Uma questão chave na transição da sistemática baseada em genes para a baseada em genomas é o desenvolvimento de métodos que, a partir de informações acerca do conteúdo de um genoma, irão considerar relações ancestrais de genes homólogos e parálogos bem como a arquitetura genômica dentro de um mesmo quadro integrado. Técnicas estatísticas sofisticadas estão disponíveis para inferência filogenética, como os implementados nos métodos Bayeseanos e máxima verossimilhança. Esses métodos fazem uso de modelos evolucionários explícitos e testáveis, permitindo testes de significância e odenação de hipóteses. Contudo, a ausência de genes ou de sua

[1]Instituto de Ciências Biomédicas
[2]Instituto de Matemática e Estatística

ordem no genoma podem impor um problema sério quando são feitas tentativas de integrar esses dados com os obtidos através de inferências baseadas em alinhamento de genes.

A sistemática baseada em genes pode ser estendida para genomas completos uma vez que os genomas em questão são alinhados e tratados. Esse procedimento é árduo e provoca a perda de informação sobre características parcialmente compartilhadas.

Alternativamente, podemos comparar genomas e construir distribuições a partir de scores para características genômicas compartilhadas de maneira par-a-par. Essas distribuições são comparadas e diversas de suas características como momentos, ou comparações mais complexas envolvendo distâncias de Kullback-Leibler, Chernoff, Bhattacharyya, são estudadas com relação à sua utilidade para clusterização de genomas durante a reconstrução filogenética.

Neste projeto propomos o *BlastPhen*, um programa que implementa uma técnica de clusterização por similaridade de forma rápida e eficiente, utilizando-se o método de comparação de distribuições citado acima.

O *BlastPhen* utiliza como base de cálculo os resultados obtidos pelo programa *Blast*, que compara seqüências de genes e proteínas, encontrando e atribuindo valores a subseqüências semelhantes.

A partir desses dados, o *BlastPhen* avalia, com o auxílio de técnicas estatísticas, o grau de similaridade dos genomas, fornecendo uma medida de "distância" entre eles. Esses dados são posteriormente utilizados para gerar a clusterização dos seres em questão.

2 Objetivos

- Desenvolvimento de um programa (*BlastPhen*) que faz a clusterização por similaridade de diferentes organismos, através de comparações de distribuições estatísticas, permitindo seu uso para genomas complexos. Tais distâncias serão então utilizadas para formar a árvore filogenética apropriada.

- Encontrar a métrica adequada para ser utilizada pelo programa. Alternativamente, diferentes métricas poderão ser disponibilizadas, se cada uma comportar-se melhor em um determinado caso.

- Utilizar o programa desenvolvido para formar árvores filogenéticas de conjuntos de vários organismos com genomas complexos.

3 Descrição

3.1 Sobre análise genômica

Um *ORF* é uma seqüência de DNA que contém um conjunto contíguo de codons, cada um descrevendo um amino-ácido [2]. O programa *getorf* seleciona determinados ORFs de um genoma baseado no comprimento mínimo especificado pelo usuário.

Os programas *Blast* (*Basic Local Alignment Search Tools*) são um conjunto de algoritmos de comparação de seqüências utilizados para comparar uma seqüência com outras em um determinado banco de dados. Essas comparações são feitas par a par. A cada comparação é atribuída uma pontuação (*score*) que reflete o grau de similaridade entre as seqüências. Quanto mais alta a pontuação, maior é o grau de similaridade.

Para calcular o *raw score* são levados em consideração identidades, substituições e *gaps*. *Bit score* é o valor normalizado do *raw score*. Por isso, ele pode ser utilizado para comparar pontuações.

Identidade é um segmento no qual duas seqüências são invariantes. Uma substituição é a presença de bases diferentes em uma posição de um alinhamento.

Um *gap* é um espaço introduzido em um alinhamento para compensar inserções e remoções em uma seqüência com relação a outra. Os gaps inserem valores negativos no cálculo dos scores.

Em alinhamentos de amino-ácidos, as pontuações para identidades e substituições são dadas por matrizes de substituições. Essas matrizes contém as probabilidades de um amino-ácido mutar para outro. Tais valores foram calculados empiricamente, através da análise de extensas bases de dados.

3.2 Situação atual

Como base de testes foi selecionado um conjunto de 26 Báculo vírus cuja árvore filogenética é conhecida.

Os genomas são submetidos primeiramente ao programa *getorf*, que extrai os segmentos mais significativos das seqüências genéticas. Esses dados são então submetidos ao *Blast*, que compara cada seqüência S com as demais, listando as que têm mais semelhanças com S.

Em uma outra etapa, os arquivos gerados pelo *Blast* são processados e deles são extraídas somente as informações relevantes para o *BlastPhen*: os valores dos *bit scores*, *raw scores*, *identities* e *positives*. Também nesta etapa são calculadas as médias, medianas e modas dos atributos citados para cada comparação entre seqüências.

Tendo reunido os dados necessários, o *BlastPhen* calcula as distâncias entre as espécies de acordo com as métricas abaixo:

- Kullback-Leibler

$$\mathcal{D}(p_1 \parallel p_0) = \int p_1(x) log \frac{p_1(x)}{p_0(x)} dx$$

  Como a distância de Kullback-Leibler não é simétrica, foi utilizada a média harmônica para simetrizá-la [1]:

$$\frac{1}{\mathcal{R}(p_0, p_1)} \equiv \frac{1}{\mathcal{D}(p_1 \parallel p_0)} + \frac{1}{\mathcal{D}(p_0 \parallel p_1)}$$

- Chernoff

$$\mathcal{C}(p_0, p_1) = \max_{0 \leq t \leq 1} -log\mu(t), \qquad \mu(t) = \int [p_0(x)]^{1-t} [p_1(x)]^t dx$$

- Bhattacharyya

$$\mathcal{B}(p_0, p_1) = -\log \mu(\tfrac{1}{2})$$

- Integral da diferença

$$\mathcal{I}(p_0, p_1) = \int |p_0(x) - p_1(x)| dx$$

Como podemos verificar, tais medidas de distâncias referem-se a dados contínuos enquanto que neste projeto estamos lidando com dados discretos. Portanto, com a finalidade de adaptar as informações, estão sendo utilizados histogramas para agrupá-las.

Dois problemas surgiram com a utilização dos histogramas: quantas classes devem ser utilizadas para agrupar os dados e o que fazer quando não há dados em uma classe com relação a uma função e com relação a outra há dados (i.e. $p_0(x) = 0$ e $p_1(x) \neq 0$ e vice-versa).

Os resultados são então organizados em tabelas, facilitando a consulta e a construção da árvore filogenética.

Duas grandes dificuldades estão sendo as escolhas do tamanho do *orf* utilizado pelo *getorf* e da matriz de substituição de aminoácidos aplicada pelo *Blast*.

Após o cálculo das distâncias entre as espécies, o BlastPhen agrupa as mesmas em tabelas de tal forma que cada tabela consiste em um *bonsai*. Se pensarmos em cada espécie como sendo um vértice de um grafo, cada bonsai representa um clique desse grafo.

3.3 Desenvolvimento

Na próxima fase serão feitos testes de sensibilidade para os seguintes parâmetros:

- Intervalo dos histogramas
- Normalizador (utilizado para eliminar os zeros dos histogramas)
- Tamanho mínimo dos *orf*s selecionados pelo *getorf*
- Matriz de substituição de amino-ácidos utilizada pelo *Blast*

Os testes com a base de dados escolhida serão refeitos após a obtenção dos resultados dos testes de sensibilidade.

Depois de encontrados resultados condizentes com os métodos tradicionais para os primeiros testes, serão examinados outros grupos de organismos, também de filogenia conhecida.

Estando encerrada a fase de validação do *BlastPhen* terá início uma etapa de otimização do código, tendo em vista principalmente a parelização e a melhoria da eficiência dos cálculos matemáticos.

4 Cronograma

- 03/2004: Procura das técnicas estatísticas a serem utilizadas.
- 05/2004: Estudo, implementação e escolha das técnicas estatísticas.
- 07/2004: Testes com diversos grupos de vírus e implementação do algoritmo de separação de "bonsais".
- 09/2004: Refinamento das técnicas utilizadas.

5 Estrutura da monografia

- Primeira parte
  - Itrodução
  - Objetivos
  - Metodologia
  - Descrição

  - Conclusão
  - Bibliografia

- Segunda Parte
  - Desafios e frustrações
  - Disciplinas do BCC mais relevantes
  - Interação com os orientadores
  - Outras observações

Referências

[1] Don H. Johnson and Sinan Sinanović. Symmetrizing the kullback-leibler distance. Technical report, Rice University, Houston, TX, march 2001.

[2] David W. Mount. *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor Laboratory Press, 2001.