

UNIVERSIDADE FEDERAL DE
PERNAMBUCO

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA

UMA FERRAMENTA WEB PARA
INFERÊNCIA DE HAPLÓTIPOS

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluno
Orientadora

Ranieri Valença de Carvalho
Katia Silva Guimarães

(rvc4@cin.ufpe.br)
(katiag@cin.ufpe.br)

Março de 2009

Conteúdo

Introdução	3
Contexto biológico.....	3
Polimorfismo de Único Nucleotídeo.....	3
O Problema Biológico	3
O Problema computacional	4
Objetivo	5
Cronograma	6
Referências	7
Assinaturas.....	8

Introdução

Contexto biológico

O código genético humano é um conjunto de informações codificadas em cadeias de caracteres, onde cada um destes caracteres é definido por um nucleotídeo. A cadeia de uma grande quantidade destes nucleotídeos conectados por ligações fosfodiéster forma uma molécula de DNA (ácido desoxirribonucléico) ou de RNA (ácido ribonucléico).

Dentro das células eucarióticas (organismos mais complexos, incluindo seres humanos), as moléculas de DNA situam-se dentro do núcleo celular, e são compactadas e organizadas em estruturas chamadas de cromossomos. Ao conjunto de cromossomos, cujo número e morfologia são característicos de uma espécie ou de seus gametas, dá-se o nome de cariótipo.

Nos seres humanos a maioria das células possui um cariótipo composto de 46 cromossomos combinados em 23 pares. Os dois cromossomos de cada par, contudo, não são totalmente idênticos, sendo um oriundo da mãe e, outro, do pai. O conjunto de informação genética que ocorre na união dos dois conjuntos de 23 cromossomos compõe o genótipo do indivíduo; a sequência de um único conjunto de 23, doado por um dos pais, é denominada haplótipo.

É nos cromossomos que estão contidas as informações codificadas que vão determinar diversas características de um indivíduo e expressão de proteínas. A região da sequência de DNA que carrega informações codificantes é denominada gene.

Variações em determinadas posições (especialmente em posições não freqüentemente variadas) das seqüências de DNA podem ocasionar mudanças fenotípicas ou até anormalidades no metabolismo. Quando as alterações genéticas provocam mudanças dentro de uma mesma espécie, este tipo de variação é conhecido como polimorfismo. Os polimorfismos acarretam expressão protéica diferenciada nos diferentes indivíduos, não raramente gerando problemas.

Polimorfismo de Único Nucleotídeo

A forma predominante de polimorfismo é o SNP (do inglês: *Single Nucleotide Polymorphism*), uma vez que a chance destas flutuações acontecerem em posições separadas é naturalmente maior do que a chance de ocorrerem em sítios adjacentes. Os SNPs são alterações da seqüência de DNA que ocorrem quando um único nucleotídeo (A, T, C, ou G) na seqüência do genoma é modificada. Um exemplo de SNP poderia ser a mudança em seqüência de DNA: ACGGCTAA a ATGGCTAA, ocorrendo uma mudança da base nitrogenada C para T.

Os SNPs são o tipo de polimorfismos mais importante e mais estudados, pois perfazem cerca de 90% de toda a variação genética humana. Muitos destes não têm qualquer efeito sobre a função celular; entretanto há alguns SNPs que modificam o aminoácido resultante, podendo acarretar mudanças na estrutura e/ou função da proteína final. Esse tipo de SNP é o mais comumente estudado.

O Problema Biológico

Acredita-se que o estudo aprofundado de mapas de SNPs poderá ajudar na identificação de vários genes associados a doenças complexas, tais como câncer, diabetes, doenças vasculares e algumas formas de doenças mentais. Estas associações são difíceis de determinar com métodos convencionais, porque um único gene modificado pode fazer apenas uma pequena contribuição para a doença.

Para se estudar essas doenças mais complexas, o haplótipo é o conjunto de dados que é mais informativo e conveniente para se estudar os SNPs, em vez de usar os dados do genótipo. Entretanto, existe uma grande dificuldade em se obter apenas os dados de haplótipo com as técnicas convencionais da biologia molecular.

Diante disso, a Inferência de Haplótipo (*Haplotype Inference* – HI) tem como objetivo extrair as informações do haplótipo a partir da observação de dados do genótipo. A partir deste feito, torna-se possível a realização de diversas aplicações computacionais. Para a inferência, cada genótipo é representado como um vetor de caracteres 0, 1 ou 2, onde cada caracter representa um SNP.

Uma posição no genótipo apresenta os valores 0 (ou 1), quando ambos os haplótipos (o par de alelos dos cromossomos) que formam o genótipo tem valores 0 (ou 1) nesta posição (são homocigotos); caso sejam diferentes, ou seja, tenha o valor 0 em dos alelos e 1 no outro, esta posição do genótipo terá valor 2 (heterocigoto).

O Problema computacional

O Problema da Inferência de Haplótipos é o seguinte. Dados um conjunto de vetores (cadeias de caracteres) de genótipos, cujos sítios são representados por 0, 1 ou 2, encontrar um conjunto de vetores binários, sendo um par destes para cada vetor de genótipo. Para cada vetor de genótipo g , os vetores binários associados (v_1 e v_2) devem ter valor 0 (ou 1) em qualquer posição em que g possui 0 (ou 1). No caso em que o valor de g seja 2, então exatamente um dos vetores binários (v_1 ou v_2) deve ter o valor 0, e o outro 1.

Um vetor é dito “resolvido” se ele possui nenhum ou apenas um caracter 2. Isso porque é possível saber imediatamente quais os vetores binários associados a ele. Caso ele possua dois ou mais caracteres 2, o vetor é chamado de “ambíguo”, e nesse caso é preciso algum algoritmo para inferir quais devem ser os vetores binários associados. Um exemplo de vetor ambíguo é 10221. Nesse caso, 10101 e 10011 podem ser os vetores binários associados, mas também podem ser 10111 e 10001.

Nesses casos, é preciso usar algum algoritmo para inferir quais são os vetores binários, que representam os haplótipos.

Alguns algoritmos amplamente conhecidos são o Algoritmo de Clark e o método da Parcimônia Pura. O Algoritmo de Clark escolhe primeiramente os vetores resolvidos e depois os utiliza para inferir os ambíguos através de uma regra de inferência. Como seu resultado depende diretamente da ordem dos vetores de entrada, ele deve ser executado um grande número de vezes, randomizando a cada vez a ordem da entrada, e depois escolhendo o melhor conjunto de saída de todas as execuções.

O método da Parcimônia Pura sugere que os haplótipos sejam inferidos baseando-se no modelo da parcimônia pura, sugerido por Gusfield, 2003. Este problema é NP-difícil, e utiliza uma abordagem de programação linear inteira para resolvê-lo.

Objetivo

O objetivo deste Trabalho de Graduação é construir uma ferramenta que inclui diversas opções de métodos para resolver o Problema de Inferência de Haplótipos e que poderá ser disponibilizada na Web, além de uma pesquisa aprofundada sobre os algoritmos utilizados.

Um dos métodos será o Algoritmo de Clark, já citado nesta proposta. Outros métodos serão estudados e implementados, a fim de obtermos uma ferramenta mais completa e útil.

A ferramenta receberá uma entrada (genótipos) e retornará uma saída formatada; o usuário poderá escolher entre os algoritmos qual deles quer usar, ajustando seus parâmetros, além de poder comparar seus resultados.

Cronograma

No cronograma abaixo estão listados alguns pontos importantes da evolução deste trabalho.

Atividade	Março			Abril			Maio			Junho		
Pesquisa e análise de algoritmos	■	■	■	■	■	■						
Implementação dos algoritmos			■	■	■	■	■	■				
Desenvolvimento da ferramenta					■	■	■	■	■	■	■	
Elaboração do relatório					■	■	■	■	■	■	■	
Preparação da apresentação											■	■

Referências

D. Gusfield, S. H. Orzack. Haplotype Inference. In: S. Aluru. Handbook of Computational Molecular Biology, cap18. CRC Press. 1104p, 2006

Assinaturas

Katia Silva Guimarães
Orientadora

Ranieri Valença de Carvalho
Aluno