Projeto de Iniciação Científica

Modelagem de dados genotípicos através de processos com interação de alcance variável.

André Jucovsky Bianchi
Aluno

Florencia Leonardi
Orientadora

Instituto de Matemática e Estatística
Universidade de São Paulo

Período: 01/08/2009 - 31/07/2010

1 Introdução

Um dos problemas mais importantes da genética na atualidade é a localização de regiões no genoma direta ou indiretamente associadas com uma certa doença complexa. O estudo dos mapas de "polimorfismos de um único", conhecidos como SNP (do inglês *single nucleotide polimorphsm*) constitui atualmente uma das abordagens mais promissoras para esse fim (Altshuler; 2008). Um SNP é uma variante comum (que ocorre em mais do que 1% da população) em uma única base do DNA. Atualmente, os mapas existentes contêm aproximadamente um milhão de SNPs (plataforma Affymetrics 6.0), o que requer, para sua análise, ferramentas eficientes tanto do ponto de vista estatístico quanto do ponto de vista computacional.

As abordagens tradicionais para este problema são as que tentam estimar a "dependência" entre a variável resposta (doença) e os diferentes sítios do mapa genético, utilizando ferramentas de regressão e análises uniloco (Ziegler et al.; 2008). Mas sabe-se que existe uma grande dependência entre os diferentes SNPs, o que tem motivado o interesse por caracterizar e descrever esse tipo de dependência. Essa caracterização poderia também servir para tentar inferir relações de ancestralidade entre os diferentes grupos estudados, o que pode prevenir ocorrências de resultados de falsos positivos nos estudos de mapeamento de doenças.

Neste projeto propomos modelar os dados de mapas de SNPs utilizando processos estocásticos com interação de alcance variável. Os processos de memória variável estacionários foram introduzidos em Rissanen (1983) e estudados posteriormente por vários autores, entre os quais podemos citar (Bühlmann and Wyner; 1999), Csiszár and Talata (2006b), Duarte et al. (2006) e Galves and Leonardi (2008). Dada a não-estacionariedade intrínseca dos dados genômicos, neste trabalho propomos a utilização de modelos não estacionários para a modelagem desse tipo de dados.

2 Conjunto de dados genotípicos

Os dados que pretendemos analisar foram obtidos pelo Consórcio Norte-Americano para a Artrite Reumatóide (NARAC). Eles foram primeiramente analisados em (Plenge; 2007a) e disponibilizados para a décima sexta edição do *Genetic Analysis Workshop* (GAW16).

Este conjunto de dados compreende a descrição de 545.080 SNPs do genoma de 2.062 indivíduos. Deste total, 1.194 são portadores de artrite reumatóide, enquanto que os outros 868 estão livres da doença. Além dos dados de SNPs, outras variáveis foram coletadas, como por exemplo duas variáveis quantitativas que são utilizadas no diagnóstico de indivíduos com artrite reumatóide.

Diferentes estudos na literatura indicam que a existência da artrite reumatóide em um indivíduo está bastante associada à região mais densa do genoma humano, o complexo chamado HLA (Human Leukocyte Antigen), situado no cromossomo 6 (Plenge; 2007b). A partir da nossa modelagem pretendemos verificar este fato e também achar outras possíveis regiões associadas com esta doença. Além disso, há evidências de que os indivíduos da amostra pertençam a diferentes grupos ancestrais, o que também

poderá ser verificado através da modelagem com processos estocásticos.

3 Processos com interação de alcance variável

Na modelagem dos dados genotípicos, assumiremos que cada SNP é uma variável aleatória assumindo valores no conjunto $A = \{0, 1, 2\}$, onde será atribuído o valor 0 se o SNP não contém o alelo de interesse, 1 se o SNP contém um alelo e 2 se o SNP contém dois alelos de interesse. Dentro de cada cromossomo, existe uma ordem para os SNPs dada pela localização física (em pares de bases) específica de cada um deles no DNA. Portanto assumiremos que os SNPs representam uma sequência de variáveis aleatórias $X_1, \ldots, X_n$ sobre A, onde n representa o número de SNPs.

Um primeiro passo na nossa modelagem dos dados genotípicos será identificar "janelas", isto é subconjuntos de SNPs consecutivos, que apresentem uma forte dependência entre eles. Este passo tem como objetivo identificar regiões genômicas que possam levar a uma caracterização da estrutura do DNA a partir desses dados.

Dito de outra forma, para cada SNP X_i, $i \in \{1, \ldots, n\}$, nosso objetivo será estimar o valor dos inteiros k e l tais que

$$P(X_i = x_i | X_1^{i-1} = x_1^{i-1}, X_{i+1}^n = x_{i+1}^n) = P(X_i = x_i | X_{i-k}^{i-1} = x_{i-k}^{i-1}, X_{i+1}^{i+l} = x_{i+1}^{i+l}),$$

onde dados os inteiros $r < s$, x_r^s representa a sequência $x_r, x_{r+1}, \ldots, x_s$. Num primeiro momento assumiremos que os inteiros k e l dependem da posição i do SNP mas não dependem da sequência específica x_{i-k}^{i+l}. Assim, para cada SNP teremos associada uma janela de "dependência" dada pelo intervalo inteiro $[i - k, i + l]$. Este modelo constitui uma generalização dos campos Markovianos 1-dimensionais, dado que neste caso a interação em cada sítio pode ser diferente para os diferentes sítios. Por outro lado, sabe-se que no caso 1-dimensional, os campos Markovianos são equivalentes aos processos de Markov. Portanto, espera-se que muitos dos resultados obtidos para processos estocásticos de memória variável, como por exemplo os obtidos em Csiszár and Talata (2006b) ou Galves and Leonardi (2008), e os obtidos para campos Markovianos estacionários, como por exemplo Csiszár and Talata (2006a), possam ser adaptados a nossa abordagem.

Num segundo passo do desenvolvimento deste projeto tentaremos identificar janelas de SNPs associadas com a variável resposta Y, definida de tal forma que assuma o valor 1 se a pessoa é doente ou 0 no caso contrário. A estimação das janelas de dependência será feita através da definição de critérios de máxima verossimilhança penalizada, como é feito para os processos de memória variável (Csiszár and Talata; 2006b). As principais vantagens desta abordagem é a sua consistência e o fato de poder ser implementada em tempo linear.

4 Objetivo

O objetivo principal deste projeto é a modelagem dos dados genotípicos utilizando processos com interação de alcance variável e o desenvolvimento de algoritmos eficientes para a sua estimação. Os algoritmos estarão baseados no critério de estimação por

máxima verossimilhança penalizada, introduzido em Csiszár and Talata (2006b) para o caso de processos de memória variável. No nosso caso, os processos estudados serão não estacionários.

A partir da implementação computacional dos algoritmos procederemos à análise do conjunto de dados GAW16 descrito anteriormente. Nesse caso, o objetivo será a identificação de regiões do genoma contendo SNPs com forte dependência entre eles ou com uma forte influência na variável resposta. Um outro objetivo será a caracterização da população analisada e a definição do conceito de ancestralidade para os diferentes grupos.

Referências

Altshuler, D. e. a. (2008). Genetic mapping in human disease, *Science* **322**: 881–888.

Bühlmann, P. and Wyner, A. J. (1999). Variable length Markov chains, *Ann. Statist.* **27**: 480–513.

Csiszár, I. and Talata, Z. (2006a). Consistent estimation of the basic neighborhood of Markov random fields, *Ann. Statist.* **34**(1): 123–145.

Csiszár, I. and Talata, Z. (2006b). Context tree estimation for not necessarily finite memory processes, via BIC and MDL, *IEEE Trans. Inform. Theory* **52**(3): 1007–1016.

Duarte, D., Galves, A. and Garcia, N. (2006). Markov approximation and consistent estimation of unbounded probabilistic suffix trees, *Bull. Braz. Math. Soc.* **37**(4): 581–592.

Galves, A. and Leonardi, F. (2008). *Exponential inequalities for empirical unbounded context trees*, Vol. 60 of *Progress in Probability*, Birkhauser, pp. 257–270.

Plenge, R. e. a. (2007a). TRAF1-C5 as a risk locus for rheumatoid arthritis - a genomewide study, *N Engl J Med.* **357**: 1199–209.

Plenge, R. e. a. (2007b). Two independent alleles at 6q23 associated with risk of rheumatoid arthritis, *Nat Genet* **39**(12): 1477–1482.

Rissanen, J. (1983). A universal data compression system, *IEEE Trans. Inform. Theory* **29**(5): 656–664.

Ziegler, A., König, I. and Thompson, J. (2008). Biostatistical aspects of genome-wide association studies, *Biometrical Journal* **50**(1): 8–28.