Modelagem de Dados Genotípicos através de Processos com Interação de Alcance Variável

André J. Bianchi, Andressa Cerqueira, Júlia P. Soler, Florencia Leonardi[1]
Instituto de Matemática e Estatística, Universidade de São Paulo
[1]florencia.leonardi@yahoo.com.br

Resumo
Um dos maiores problemas no mapeamento de genes associados com doenças complexas é o entendimento de como o espaço de marcadores genômicos está estruturado. Em espaços genômicos esparsos e de alta dimensão, como é o caso das plataformas de SNPs (Single Nucleotide Polimorphsms), este conhecimento pode ajudar na construção de análises multilocos que são mais informativas para a detecção de genes. Neste trabalho propomos modelar longas sequências de marcadores do tipo SNP por meio de processos estocásticos com interação de alcance variável. O objetivo é identificar blocos de SNPs com forte influência uns sobre os outros. Como aplicação da metodologia, consideramos dados do Genetic Analysis Workshop 16 que correspondem a uma sequência de 545.080 SNPs avaliados em 1.194 indivíduos portadores de artrite reumatóide e em 868 livres da doença.

1. Introdução

Um dos problemas mais importantes da genética na atualidade é o mapeamento de genes associados com doenças complexas em humanos. Por mapeamento entende-se o conjunto de procedimentos para a identificação de regiões no genoma direta ou indiretamente associadas com uma certa característica de interesse. O estudo de mapas de "polimorfismos de um único nucleotídeo", conhecidos como SNP (do inglês Single Nucleotide Polimorphsm), constitui uma das abordagens mais promissoras para esse fim (Altshuler et al.,2008). Um SNP é uma variação comum (com prevalência maior que 1%) que ocorre em uma única base do DNA em diferentes pontos do genoma. Atualmente, os mapas existentes contêm aproximadamente um milhão de SNPs (por exemplo, as plataformas 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 diferentes grupos estudados, o que pode prevenir ocorrências de resultados de falsos positivos nos estudos de mapeamento de doenças.

Neste trabalho propomos modelar 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 Buhlmann and Wyner (1999), Csiszár and Salata (2006a), Duarte et al. (2006) e Galves and Leonardi (2008). Dada a não-estacionaridade intrínseca dos dados genômicos, neste

trabalho propomos a utilização de modelos não estacionários para a modelagem desse tipo de dados. Além disso, a estrutura de dependência entre SNPs também é avaliada por meio de testes simultâneos de associação entre pares de SNPs (conhecidos como testes de desequilíbrio de ligação). As análises propostas são realizadas usando os recursos do aplicativo estatístico R e do PLINK (Purcell, 2007).

2. Material e Métodos

2.1. Conjunto de dados de SNP

Os dados analisados neste trabalho foram obtidos pelo Consórcio Norte-Americano para a Artrite Reumatóide (NARAC). Eles foram primeiramente analisados por Plenge et al. (2007b) 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. De maneira geral, para cada indivíduo, o conjunto de dados de SNPs disponível está no formato de genótipos, o que corresponde a uma sequência de valores 0, 1, 2 (codificando os genótipos aa, Aa e AA, respectivamente) de tamanho 545.080.

2.2. Processos com interação de alcance variável aplicado aos dados genotípicos

Na modelagem dos dados genotípicos é assumido que cada SNP é uma variável aleatória com valores no conjunto $A=\{0,1,2\}$, onde é 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), a qual é específica de cada um deles no DNA. Portanto, é natural assumir que os SNPs representam uma sequência de variáveis aleatórias $X_1,\ldots ,X_n$ sobre A, onde n representa o número total de SNPs.

Um primeiro passo na modelagem proposta é 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 $(i=1, \ldots , n)$, há interesse em estimar o valor dos inteiros k e l tais que

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

em que, dados os inteiros $r < s$, x_r^s representa a sequência $x_r, x_{r+1}, \ldots ,x_s$. Num primeiro momento assume-se 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 tem-se 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, muitos dos resultados obtidos para processos estocásticos de memória variável, como, por exemplo, os obtidos em Csiszár and Talata (2006a) ou Galves et al. (2008), e os obtidos para campos

Markovianos estacionários, como, por exemplo, Csiszár and Talata (2006b), podem ser adaptados à análise dos dados de SNP.

A estimação das janelas de dependência é feita através da definição de critérios de máxima verossimilhança penalizada, como é feito para os processos de memória variável em Csiszár and Talata (2006a). As principais vantagens desta abordagem é a sua consistência e o fato de poder ser implementada em tempo linear.

Um segundo passo do desenvolvimento deste trabalho é identificar quais janelas de SNPs estão associadas com a variável resposta Y, definida de tal forma que assume o valor 1 se a pessoa é doente ou 0 no caso contrário.

Por fim, os resultados dessa modelagem são comparados com os obtidos da aplicação de testes de associação simultâneos entre pares de SNPs, um procedimento comumente usado em Genética e implementado no aplicativo PLINK (Purcell, 2007).

3. Resultados

Aplicando o modelo proposto aos dados de SNP de artrite reumatóide, foram obtidas vizinhanças de influência de tamanho médio 2,22418 SNPs e desvio padrão 0,714481 SNP. O tamanho médio das vizinhanças para a esquerda foi de 1,11432 e para a direita de 1,10985 SNP. Isto significa que, no modelo proposto, cada SNP é fortemente influenciado por um pouco mais de 2 SNPs adjacentes, em média.

O número de vizinhanças sobrepostas em cada SNP pode ser visto na Figura 1 (apenas para os 200 primeiros SNPs). Para cada SNP*j*, este é o número de vizinhanças de influência dos SNPs adjacentes a *j* que o contém.

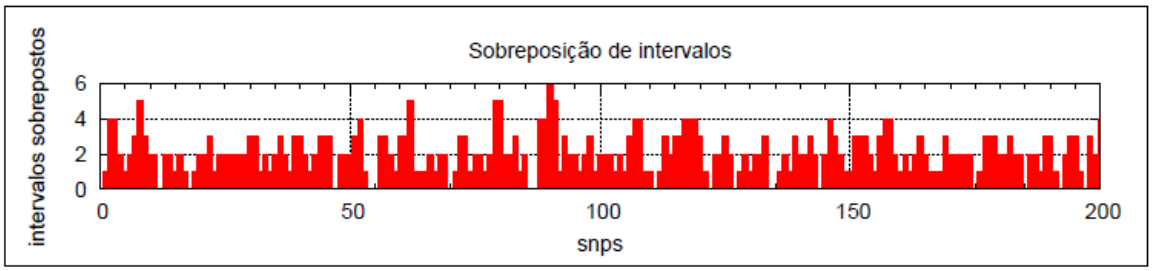

Figura 1. Sobreposição de vizinhanças de influência para os 200 primeiros SNPs do mapa. O valor para cada SNP indica em quantas vizinhanças de influência este SNP está contido.

Para se ter uma idéia de como estas vizinhanças estão relacionadas entre si, na Figura 2 estão representados os valores para cada SNP, com o tamanho das vizinhanças à esquerda indicados como valores negativos, e o tamanho das vizinhanças à direita indicados como valores positivos. Apesar de não ser a melhor representação, é possível ter a noção de que muitas vezes as vizinhanças formam blocos de influência. É frequente observar blocos de SNPs adjacentes em que todos os SNPs do bloco estão contidos nas vizinhanças de influência de todos os outros SNPs do bloco.

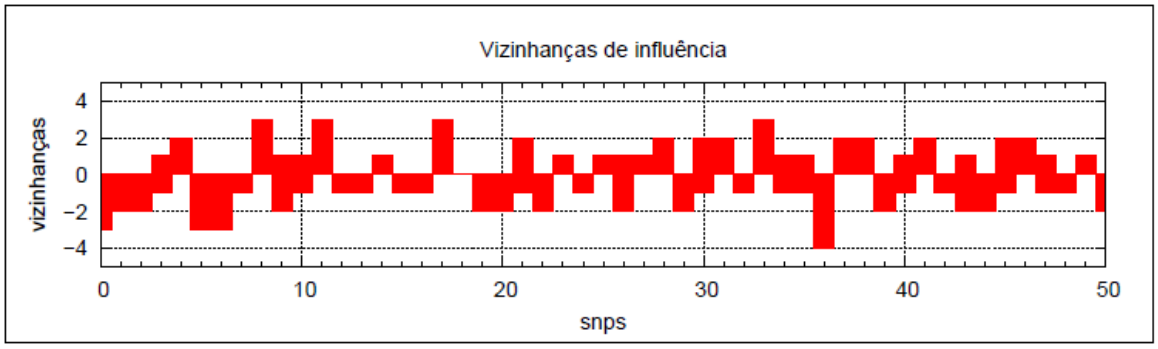

Figura 2. Vizinhanças de influência para os 50 primeiros SNPs. O valor positivo é o tamanho da vizinhança à esquerda, e o valor negativo é o tamanho da vizinhança à direita.

A partir da determinação das vizinhanças de cada SNP, há interesse em verificar se os SNPs se agrupam de alguma forma em "janelas" de influência, ou seja, blocos de informação formados por sequências de SNPs com a propriedade de que todas as vizinhanças de influência de cada SNP da janela estão contidos na própria janela. Analisando as vizinhanças determinadas ao longo da sequência analisada obtém-se que os SNPs estão divididos em 48.697 janelas de influência. O tamanho médio das janelas é 10,27 SNPs, a menor janela tem tamanho 1 e a maior é composta por 83 SNPs. O desvio padrão do tamanho das janelas é de 5,94 SNPs.

Estes resultados indicam que a teoria de processos estocásticos com interação de alcance variável pode ser adaptada e aplicada a sequências de dados de SNPs com o objetivo de identificar blocos de SNPs com forte influência uns sobre os outros. Tal informação pode ser usada para estudar o efeito que tais blocos ou regiões genômicas têm sobre fenótipos de interesse, por exemplo, por meio de testes de associação multilocos (Ziegler et al., 2008).

4. Bibliografia

Altshuler, D. et al. (2008). Genetic mapping in human disease. *Science* **322**: 881-888.

Bühlmann, P. and Wyner, A.J. (1999). Variable lingh Markov chains. *Ann. Stat.* **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.

Duarte, D.; Galves, A. and Garcia, N (2006). Markov approximation and consisten estimarion 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: 257-270.

Plenge, R. et al. (2007a). TRAFI-C5 as a risk locus for rheumatoid arthritis – a genome wide study. *N Engl. J. Med* **357**:1199-209.

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

Purcell, S. et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**: 559-574.

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

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