Janelas de Influência em Polimorfismos de Um Único Nucleotídeo

André Jucovsky Bianchi (drebs@linux.ime.usp.br)

Orientadora: Florencia Graciela Leonardi (florencia@usp.br)

Instituto de Matemática e Estatística, Universidade de São Paulo

1. Introdução

Nosso objetivo neste estudo é aplicar alguns modelos estocásticos simples, de forma a tentar determinar regiões do DNA com forte influência umas sobre as outras, e em seguida estabelecer critérios para agrupar essas regiões em conjuntos com alta dependência entre seus elementos. Estes conjuntos formam blocos de pequenas sequências de código que podem ser avaliadas individualmente com relação a alguma característica fenotípica. Em nosso caso, o conjunto de dados que estamos avaliando contém indivíduos *caso* e *controle* para a doença Artrite Reumatóide.

Até o final da década de 90, os modelos existentes consideravam marcadores com algumas dezenas de nucleotídeos. Com o avanço do sequenciamento do genoma humano, foram identificadas áreas onde longas sequências diferem entre os indivíduos em apenas um nucleotídeo. Foi possível determinar que, nessas áreas, geralmente duas bases alternativas ocorrem com alta frequência (mais do que 1%!) e que, apesar de serem menos informativas do que os outros marcadores até então utilizados, tais áreas são mais abundantes e têm maior potencial para automação de sequenciamento. O nome dado a cada uma dessas áreas é SNP, para Single-Nucleotide Polimorphism – ou Polimorfismo de Um Único Nucleotídeo.

2. Vizinhanças de influência

Os SNPs podem ser modelados como $s \in \mathbb{N}^+$ variáveis aleatórias diferentes, dispostas sequencialmente no espaço, às quais chamamos X_j, com $1 \leq j \leq s$.

Para verificar se o valor de um SNP j fixado é influenciado pelos valores de um certo conjunto $V_j \subseteq \mathbf{S}$ de SNPs, estudamos a probabilidade $P(X_j = x_j \mid X_i = x_i,\ i \neq j)$ da ocorrência de um valor $x_j \in A$ para a variável X_j, condicionada a diferentes tamanhos de vizinhança adjacente. Se os SNPs não forem independentes uns dos outros, então a probabilidade de que x_j ocorra deve variar dependendo dos valores dos outros SNPs:

$$P(X_j = x_j^i \mid X_l = x_l^i,\ l \neq j) = P(X_j = x_j^i \mid X_l = x_l^i,\ l \in V_j)$$

A partir da expressão acima, definimos a pseudoverossimilhança penalizada[1] para uma l, r-vizinhança do SNP j como:

$$\overline{L}_j^{\mathfrak{D}}(l,r) = \sum_{\omega \in A^l} \sum_{\tau \in A^r} \sum_{a \in A} \left(N_j^{\mathfrak{D}}(\omega, a, \tau) \log_{|A|} \left(\frac{N_j^{\mathfrak{D}}(\omega, a, \tau)}{N_j^{\mathfrak{D}}(\omega, \cdot, \tau)} \right) \right) - \frac{(|A|-1)}{2} |A|^{|\omega\tau|} . \log_{|A|}(n)$$

Para encontrar $\hat{l}_j$ e $\hat{r}_j$, as vizinhanças estimadas associadas à posição de SNP j, basta encontrar os argumentos que maximizem a expressão acima. Os resultados da aplicação do modelo podem ser vistos na Figura 1.

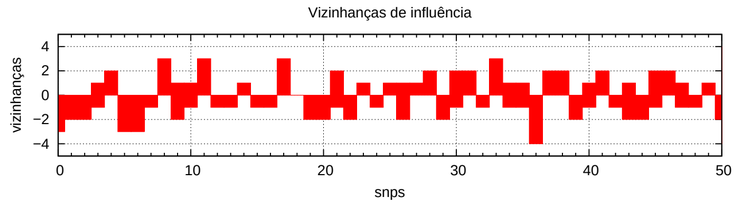

Figura 1: *Vizinhanças de influência para os 50 primeiros SNPs. O valor positivo é o tamanho da vizinhança à esquerda, e o valor negativo é o tamanho da vizinhança à direita. É possível observar a formação de blocos de influência.*

3. Janelas de influência e métricas

A partir dos dados obtidos, é possível determinar *janelas de influência*, ou seja, sequências de SNPs com a propriedade de que todos os intervalos de influência de cada SNP da janela estão contidos na própria janela.

Uma vez determinadas as janelas, podemos usar diversas medidas de distância para verificar se alguma janela distingue indivíduos *caso* de indivíduos *controle* mais do que as outras. Um exemplo de medida utilizada (os resultados podem ser vistos na Figura 2) é a *Divergência de Kullback-Leibler*, uma medida assimétrica com a seguinte fórmula:

$$D(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|w_i|}} \hat{P}_i(\omega) . \log \frac{\hat{P}_i(\omega)}{\hat{Q}_i(\omega)}$$

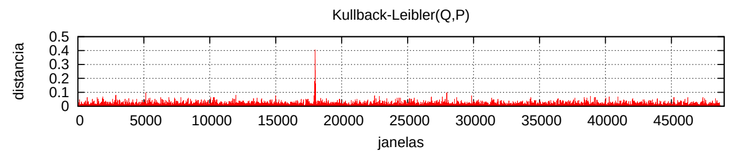

Figura 2: *Distância entre indivíduos caso e controle em cada janela, dada pela Divergência de Kullback-Leibler.*

4. Resultados e Conclusão

A *pseudoverossimilhança penalizada* é um estimador consistente para encontrar *vizinhanças de influência* em SNPs, o que nos permite determinar *janelas de influência* entre SNPs.

As técnicas de medição de distância entre indivíduos caso e controle tiveram sucesso em identificar regiões de SNPs que os diferenciam consideravelmente. Os resultados são consistentes com a literatura existente [2], que indica forte influência de regiões do cromossomo 6 na presença da Artrite Reumatóide.

Referências

[1] Julian Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

[2] Robert M. et al. Plenge. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nature Genetics*, 8(39):1477–1482, 2007.