Trabalho de Formatura Supervisionado

Vizinhanças e Janelas de Influência em Polimorfismos de Um Único Nucleotídeo

André Jucovsky Bianchi
Orientadora: Florencia Graciela Leonardi

1 de dezembro de 2009

Sumário

Parte I
Janelas de Influência em SNPs

1 Introdução

O estudo das estruturas de funcionamento do genoma repetidamente nos mostra que a interação entre os genes no DNA é mais complexa do que qualquer modelo já formulado. De fato, a própria concepção de "gene"está em constante revisão e, o que antes pensávamos ser explicável por modelos simples de hereditariedade de características, hoje o avanço técnico-científico nos mostra ser composto de cada vez mais mecanismos de codificação e funcionamento. Os modelos mais atuais descrevem genes sobrepostos, genes contidos em genes, genes segmentados que são recortados e colados pela célula em diversas combinações, criando mensagens que codificam proteínas diferentes, e até a atuação do RNA como mecanismo corretor de falhas entre gerações de seres vivos [6].

Nosso objetivo neste estudo é aplicar alguns modelos estatísticos a um certo conjunto de dados de sequenciamento genético, de forma a tentar determinar regiões do DNA com forte influência umas sobre as outras. Uma vez realizada esta identificação, é possível estabelecer critérios para agrupar essas regiões em conjuntos com alta dependência entre seus elementos, de modo que formem blocos de código que podem ser avaliados conjuntamente com relação a alguma característica fenotípica.

1.1 Genes e marcadores

Um modelo matemático simples que captura a idéia da herança de uma certa característica na informação passada de dois progenitores para uma prole pode ser formulado por meio de um par de variáveis aleatórias (Y_1, Y_2), cada uma associada a um progenitor e podendo assumir valores em $\{0, 1\}$, de acordo com distribuições de probabilidade que dependem do genótipo de cada indivíduo. Uma realização destas variáveis é a concepção de um novo indivíduo, a prole, e o valor resultante dessa realização é o código genético herdado.

Desprezando interações entre os genes e com o ambiente, a dificuldade tem sido descobrir quais são as características fenotípicas que podem ser associadas ao modelo descrito acima e, para essas características, quais são as regiões do genoma envolvidas na determinação do fenótipo. Estas regiões compõem o *gene* associado à característica. Um *marcador genético* é uma posição já identificada do DNA, nem sempre associada a um gene ou uma característica específica, mas passível de genotipagem.

Podemos generalizar um pouco o modelo acima permitindo que as variáveis Y_1 e Y_2 assumam mais do que apenas dois valores, aumentando assim o

número de características fenotípicas diferentes que podem ser codificadas a partir das variáveis. Cada um dos possíveis valores que uma sequência genômica pode assumir é chamado um *alelo*.

1.2 Single-Nucleotide Polimorphism - SNPs

Até o final da década de 90, os modelos existentes consideravam marcadores genéticos compostos de algumas dezenas de nucleotídeos chamados *micro-satélites*. Este tipo de marcador consiste na identificação de regiões onde ocorre a repetição de pequenos conjuntos de bases em longas sequências, como por exemplo $(CA)^n$, com n variando entre 10 e 100 (ou seja, uma longa sequência da forma $CACA \ldots CACA$). Em alguns casos, os blocos de repetição chegam a ser compostos de até 6 nucleotídeos. O tamanho físico dos micro-satélites permite que a amostragem do genoma seja feita por grandes regiões.

Com o avanço das técnicas de sequenciamento, 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 conterem menos informação do que outros marcadores até então utilizados, tais áreas são mais abundantes e têm maior potencial para automação de sequenciamento [10]. O nome dado a essas áreas é SNP, para Single-Nucleotide Polimorphism – ou polimorfismo de um único nucleotídeo. Um dos problemas mais importantes da genética na atualidade tem sido a determinação da associação direta ou indireta de SNPs com certas doenças complexas [1].

SNPs admitem apenas dois alelos (frequente e raro) e por esta razão são menos informativos do que outras alternativas multi-alélicas. Esta deficiência pode ser contornada utilizando uma densidade maior de pontos: mapas com cerca de 1000 SNPs extraem a mesma quantidade de informação que mapas com algo em torno de 300 regiões de micro-satélites. Hoje a estimativa é de que existam cerca de 13 milhões de SNPs em todo o genoma humano, mas os mapas atuais contém uma média de um milhão de SNPs.

Além de informações de associação genética (*genetic linkage*, em inglês), os SNPs também evidenciam características fundamentais de ancestralidade, como pode ser visto em [9]. Nesta referência são utilizados resultados clássicos de Álgebra Linear para a realização de uma decomposição espectral da matriz de amostras multiplicada por sua transposta. Os resultados mostram que a projeção de cada indivíduo nos subespaços gerados pelos autovetores da decomposição agrupa os indivíduos em diferentes regiões, de acordo com suas características raciais e de ancestralidade.

1.3 O conjunto de dados

Os dados que estamos analisando foram obtidos pelo Consórcio Norte-americano para a Artrite Reumatóide (NARAC)[1], e foram primeiramente analisados em [7] e disponibilizados para a décima sexta edição do Genetic Analysis Workshop (GAW16)[2]. Este conjunto de dados compreende os valores de 500.180 `SNP`s distribuídos ao longo dos 22 autossomos humanos, para 2.062 indivíduos *sem relação de parentesco*. Destes, 868 indivíduos manifestam a artrite reumatóide (indivíduos *caso*) e 1.194 estão livres da doença (indivíduos *controle*).

Originalmente os dados foram fornecidos em uma grande tabela onde cada linha representa um indivíduo diferente. Nas colunas desta tabela, além dos valores dos `SNP`s, também foram fornecidas uma variável indicadora da presença da artrite reumatóide e duas variáveis quantitativas relacionadas à produção de certas substâncias relacionadas à doença.

Para a análise que faremos, uma nova codificação é introduzida com o objetivo de criar uma nova tabela contendo apenas variáveis que indicam a "natureza"do par de alelos de cada `SNP` de cada indivíduo: raros, frequentes ou um de cada. As variáveis quantitativas mencionadas não são utilizadas na nova codificação (e nem neste trabalho como um todo) e a variável indicadora de presença da doença será utilizada de forma implícita sempre que for necessário restringir o conjunto de dados a um dos grupos de indivíduo.

As entradas da tabela original referentes a valores de `SNP`s consistem em pares ordenados sobre os possíveis valores de nucleotídeos: $(m, n) \in \{A, T, C, G\}^2$. Cada par representa os valores sequenciados em uma posição de `SNP` para o par de cromossomos do indivíduo correspondente. Como para cada `SNP` não nos interessa estudar exatamente o "valor"das bases nitrogenadas encontradas, mas sim a "natureza"dos alelos representados pelas bases em termos de frequência, utilizamos uma codificação diferente que captura esta idéia.

Através da realização de uma contagem simples é possível estabelecer para cada `SNP` a frequência de ocorrência de cada valor de nucleotídeo, e determinar se corresponde a um alelo frequente ou raro. Atribuindo valor 0 para alelo frequente e 1 para alelo raro, cada par de valores $(m, n) \in \{A, T, C, G\}^2$ transforma-se em um par $(a, b) \in \{0, 1\}^2$. Note que alelos raros para um certo `SNP` podem ser frequentes para outro `SNP`.

Indivíduos com valores coincidentes para um par – $(0, 0)$ ou $(1, 1)$ – são chamados *homozigotos* e indivíduos com valores diferentes – $(0, 1)$ e $(1, 0)$ – são chamados *heterozigotos*. Uma vez realizadas estas contagens de frequência, podemos definir a nova codificação que utilizaremos. Para cada par representando as frequências dos alelos encontrados nos `SNP`s de cada indivíduo, teremos:

[1] http://www.naracdata.org/

[2] http://www.gaworkshop.org/gaw2008.htm

- valores $(0,0)$ correspondendo a indivíduos homozigoto com alelos de maior frequência dentro do `SNP` são codificados como **0**;
- valores $(0,1)$ ou $(1,0)$ correspondendo a indivíduos heterozigoto são codificados como **1**;
- valores $(1,1)$ correspondendo a indivíduos homozigoto com alelos de menor frequência dentro do `SNP` são codificados como **2**.

Para deixar mais claro, o primeiro passo da codificação é passar cada valor original por uma função $f : \{A,T,C,G\}^2 \to \{0,1\}^2$ que captura a frequência dos alelos em cada `SNP`. O segundo passo é passar estes novos valores por uma função $g : \{0,1\}^2 \to \{0,1,2\}$ que captura a relação entre os tipos dos alelos de cada indivíduo.

Deste modo, em nossa nova codificação, os dados de cada indivíduo (linhas da tabela) correspondem a 500.180 valores do conjunto $A = \{0,1,2\}$. Daqui pra frente, sempre que nos referirmos ao *conjunto de dados*, estaremos nos referindo aos dados representados nesta nova codificação.

1.4 Notação

Em nosso conjunto de dados, um `SNP` pode assumir aleatoriamente um dos valores do conjunto $A = \{0,1,2\}$ e portanto poderíamos representá-lo por uma variável aleatória X, de forma que uma realização desta variável aleatória corresponderia a um valor de `SNP` obtido através do sequenciamento genético do DNA de um indivíduo. Porém, seria ingênuo demais imaginar que todos os `SNP`s obedecem a uma mesma lei probabilística, e então vamos considerar uma variável aleatória diferente para cada posição de `SNP`.

Serão estudadas portanto $s \in \mathbb{N}^+$ variáveis aleatórias diferentes (no caso, $s = 500.180$), dispostas sequencialmente no espaço, às quais chamaremos X_j, com $1 \leq j \leq s$. Para evitar verbosidade excessiva, poderemos nos referir ao "`SNP` j", e no caso deve-se ler "`SNP` na posição j"ou então apenas "X_j". Da mesma forma, conjuntos de índices poderão significar conjuntos de `SNP`s, e nestes casos o significado ficará claro pelo contexto.

Uma realização de X_j, ou seja, o valor do sequenciamento do `SNP` j do DNA, será denotada por $x_j \in A$. Quando nos referirmos às diversas amostras (indivíduos), o valor de X_j na amostra i será indicado por x_j^i . Os valores obtidos de uma sequência adjacente de $q - p$ `SNP`s do genoma de um indivíduo i, com $1 \leq p < q$, serão representados por uma cadeia de caracteres $x_{p,q}^i = x_p^i x_{p+1}^i \ldots x_{q-1}^i x_q^i$, $x_{p,q}^i \in A^{(q-p+1)}$. O símbolo $x_{1,s}^i$ denota, portanto, o sequenciamento completo do DNA do indivíduo i.

Sejam $\omega \in A^n$ e $\omega' \in A^m$, com $m \leq n$. Dizemos que ω' é *prefixo* de ω se existe $\tau \in A^{n-m}$ tal que $\omega = \omega'\tau$. De forma equivalente, ω' é *sufixo* de ω se existe $\tau \in A^{n-m}$ tal que $\omega = \tau\omega'$. Note que, em especial, ω é prefixo e sufixo de ω qualquer que seja ω, pois podemos tomar $\tau \in A^0$, que chamamos

de "sequência vazia"(neste caso τ seria o elemento neutro da operação dada pela concatenação de símbolos).

Uma amostra de n indivíduos, cada um com s valores de SNPs, pode ser representada por uma matriz $\mathfrak{D}_{n\times s}$ com n linhas e s colunas. A linha i de $\mathfrak{D}$, representada por $\mathfrak{D}^i$, contém os valores do sequenciamento genético de todos os SNPs do indivíduo i. A coluna j de $\mathfrak{D}$ contém os valores das realizações da variável aleatória X_j, nos diferentes indivíduos. Será necessário em alguns momentos nos referirmos ao conjunto de dados restrito a cada conjunto de indivíduos, *caso* e *controle*. Para tal, denotaremos por $\mathfrak{D}_{\restriction K}$ o conjunto de dados restrito aos indivíduos *caso* e por $\mathfrak{D}_{\restriction L}$ o conjunto de dados restrito aos indivíduos *controle*.

2 Vizinhanças de influência

Nesta seção, temos como objetivo tentar identificar vizinhanças de influência entre SNPs adjacentes, ou seja, descrever e quantificar de que forma o valor de um SNP pode ser dependente dos valores dos SNPs de sua vizinhança. Para isto, utilizaremos os dados amostrais coletados para estudar as probabilidades de ocorrência de cada sequência em cada região do genoma. Investigaremos vizinhanças de diversos tamanhos e utilizaremos um critério de máxima verossimilhança para determinar vizinhanças ótimas, diminuindo a relevância de uma certa vizinhança quanto maior for o seu tamanho.

Uma técnica parecida é utilizada em [5] para estimar árvores de contexto, úteis para modelar processos nos quais o número de variáveis aleatórias das quais depende o valor de uma certa variável fixada é diferente para cada símbolo considerado, ou seja, depende do contexto.

2.1 Buscando uma vizinhança

Para verificar se o valor de um SNP j é influenciado pelos valores de um certo subconjunto de SNPs, estudaremos a probabilidade $P(X_j = x_j \mid X_l,\ l \neq j)$ da ocorrência de um valor $x_j \in A$ para a variável X_j, condicionada aos valores de diferentes vizinhanças. 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.

É razoável supor que diversas realizações de um mesmo SNP em indivíduos que não são parentes entre si sejam independentes e estejam submetidas a uma mesma lei probabilística. Para verificar se o valor de um certo SNP j é influenciado pelos valores de SNPs vizinhos, temos então que encontrar uma *vizinhança* $V_j \subseteq \{1, \ldots, s\}$ tal que:

$$P(X_j = x_j \mid X_k = x_k,\ k \neq j) = P(X_j = x_j \mid X_k = x_k,\ k \in V_j) \quad (2.1)$$

Para simplificar o problema não olharemos para todas as vizinhanças possíveis, mas somente para vizinhanças $V_j^{l,r}$ centradas em X_j e com alcance adjacente de l SNPs para a esquerda e r SNPs para a direita. Formalmente, $V_j^{l,r} = \{i \in \mathbb{N} \mid j - l \leq i \leq j + r,\ i \neq j,\ j - l \geq 1,\ j + r \leq s\}$. Queremos portanto estimar para cada SNP j valores l e r tais que $V_j = V_j^{l,r}$. O tamanho de uma vizinhança de alcance l para a esquerda e r para a direita é $|V_j^{l,r}| = l + r$.

Também introduziremos uma notação para tratar das sequências numa vizinhança de um SNP. Sejam $\omega \in A^l$, $\tau \in A^r$ e $a \in A$. Se $x_{p,q}$ é uma sequência tal que $q - p = l + r + 1$ e $x_{p,q} = \omega a \tau \in A^{l+r+1}$, diremos que ω é um l-prefixo da sequência, τ é um r-sufixo da sequência, e $(\omega \cdot \tau)$ é uma $l + r$-vizinhança de a em $x_{p,q}$.

2.2 Pseudo-verossimilhança penalizada

A verossimilhança de uma amostra, dado um modelo estatístico, é uma função dos parâmetros deste modelo que nos permite estimar os valores destes parâmetros de acordo com resultados amostrais. Para isso, o que se faz usualmente é variar os parâmetros até encontrar um vetor de valores que maximize a verossimilhança calculada.

Para nossa análise, definiremos uma função $\check{\ell}_j(l, r \mid \mathfrak{D})$ que, dada uma amostra $\mathfrak{D}$ e um SNP j, associa dois números naturais l e r (representando respectivamente o alcance à esquerda e à direita da vizinhança considerada) a um número real negativo. A função será construída de forma a tender a zero quanto maior for a relevância probabilística da vizinhança considerada, mas também de forma a penalizar vizinhanças de tamanhos maiores.

Se considerarmos a probabilidade da ocorrência de cada linha de $\mathfrak{D}$, como cada realização de X_j é independente das demais, podemos escrever:

$$P(\mathfrak{D}) = \prod_{i=1}^{n} P(\mathfrak{D}^i) = \prod_{i=1}^{n} P(X_j = x_j^i,\ \forall\ j) \tag{2.2}$$

Agora, temos que condicionar a probabilidade da ocorrência de cada amostra nos valores das possíveis vizinhanças. A *pseudo-verossimilhança* nos fornece um método consistente e intuitivamente plausível para estimação dos parâmetros:

$$L(\mathfrak{D}) = \prod_{i=1}^{n} \prod_{j=1}^{s} P(X_j = x_j^i \mid X_k = x_k^i,\ \forall\ k \neq j) \tag{2.3}$$

Apesar do fato de que em geral a verossimilhança não é dada pela mesma expressão que a pseudo-verossimilhança, como descrita acima, é possível provar que maximizar a pseudo-verossimilhança é o mesmo que maximizar a verossimilhança, ou seja, que ambas convergem para os mesmos valores, e

que a pseudo-verossimilhança é um estimador consistente para os parâmetros considerados [2].

Supondo então que existe uma $l_j + r_j$-vizinhança que determina a influência dos `SNP`s adjacentes ao `SNP` j no valor assumido pela variável X_j, podemos utilizar a equação (2.3) de forma e restringir os valores de k à vizinhança de interesse:

$$L(\mathfrak{D}) = \prod_{i=1}^{n} \prod_{j=1}^{s} P(X_j = x_j^i \mid X_k = x_k^i, \ \forall \ k \in V_j^{l_j, r_j}) \tag{2.4}$$

Como todos os valores de probabilidade estão no intervalo $[0, 1]$, a multiplicação de qualquer número destes valores também pertence ao mesmo intervalo, e seu logaritmo pertence ao intervalo $]-\infty, 0]$. Consequentemente, maximizar a verossimilhança de uma certa amostra como dada por (2.4) é o mesmo que maximizar o logaritmo da mesma expressão. Assim, podemos começar a definir a função de pseudo-verossimilhança que estamos procurando, passando l e r como parâmetros e definindo:

$$\begin{aligned}
\ell(l, r \mid \mathfrak{D}) &= \log_{|A|} \left(\prod_{i=1}^{n} \prod_{j=1}^{s} P(X_j = x_j^i \mid X_k = x_k^i, \ \forall \ k \in V_j^{l,r}) \right) = \\
&= \sum_{i=1}^{n} \sum_{j=1}^{s} \log_{|A|} \left(P(X_j = x_j^i \mid X_k = x_k^i, \ \forall \ k \in V_j^{l,r}) \right) = \\
&= \sum_{j=1}^{s} \sum_{i=1}^{n} \log_{|A|} \left(P(X_j = x_j^i \mid X_k = x_k^i, \ \forall \ k \in V_j^{l,r}) \right)
\end{aligned} \tag{2.5}$$

Torna-se conveniente introduzir uma notação de contagem para representar o mesmo cálculo com uma parametrização diferente. Chamaremos de $N_j^{\mathfrak{D}}(\omega, a, \tau)$, com $\omega \in A^l$ e $\tau \in A^r$, o número de vezes que o símbolo $a \in A$ aparece rodeado pela $l + r$-vizinhança $(\omega \cdot \tau)$, com a na posição j de cada amostra $\mathfrak{D}^i$. Para compactar a notação, também definiremos $P_j(a \mid \omega, \tau) = P(X_j = a \mid X_{j-l} = \omega_1, \ldots, X_{j-1} = \omega_l, X_{j+1} = \tau_1, \ldots, X_{j+r} = \tau_r)$. Utilizando estas novas notações, (2.5) pode ser escrita como:

$$\ell(l, r \mid \mathfrak{D}) = \sum_{j=1}^{s} \sum_{\omega \in A^l} \sum_{\tau \in A^r} \sum_{a \in A} \left(N_j^{\mathfrak{D}}(\omega, a, \tau) \log_{|A|} \left(P_j(a \mid \omega, \tau) \right) \right) \tag{2.6}$$

Como podemos maximizar cada parcela de (2.6) individualmente, é possível considerar cada `SNP` j de forma separada, e podemos escrever a j-ésima verossimilhança em função de uma certa $l + r$-vizinhança $(\omega \cdot \tau)$ em torno da posição j, da forma:

$$\ell_j(\omega, \tau \mid \mathfrak{D}) = \sum_{a \in A} N_j^{\mathfrak{D}}(\omega, a, \tau) \log_{|A|} \left(P_j(a \mid \omega, \tau) \right) \tag{2.7}$$

Para poder proceder com a maximização de (2.7), devemos levar em consideração a restrição $\sum_{a \in A} P_j(a \mid \omega, \tau) = 1$. Para isto, introduzimos um multiplicador de Lagrange λ multiplicando a restrição, como descrito no apêndice de [4], para obter a seguinte equação:

$$\ell_j^*(\omega, \tau \mid \mathfrak{D}) = \sum_{a \in A} N_j^{\mathfrak{D}}(\omega, a, \tau) \log_{|A|} \left(P_j(a \mid \omega, \tau) \right) - \lambda \left(\sum_{a \in A} P_j(a \mid \omega, \tau) - 1 \right) \tag{2.8}$$

O ponto de máximo de ℓ_j^* tem derivadas parciais iguais a zero para cada um dos parâmetros. Podemos então descobrir o valor de λ derivando (2.8):

$$\frac{\partial \ell_j^*(\omega, \tau \mid \mathfrak{D})}{\partial P_j(a \mid \omega, \tau)} = \frac{N_j^{\mathfrak{D}}(\omega, a, \tau)}{P_j(a \mid \omega, \tau)} - \lambda, \ \ \forall a \in A \tag{2.9}$$

E calculando o valor da probabilidade quando (2.9) se anula:

$$\frac{N_j^{\mathfrak{D}}(\omega, a, \tau)}{P_j(a \mid \omega, \tau)} - \lambda = 0 \ \ \Rightarrow \ \ P_j(a \mid \omega, \tau) = \frac{N_j^{\mathfrak{D}}(\omega, a, \tau)}{\lambda} \tag{2.10}$$

Somando para todo $a \in A$, obtemos o valor de λ:

$$\begin{aligned} \sum_{a \in A} P_j(a \mid \omega, \tau) = \sum_{a \in A} \frac{N_j^{\mathfrak{D}}(\omega, a, \tau)}{\lambda} \ \ \Rightarrow \ \ 1 = \sum_{a \in A} \frac{N_j^{\mathfrak{D}}(\omega, a, \tau)}{\lambda} \ \ \Rightarrow \\ \Rightarrow \lambda = \sum_{a \in A} N_j^{\mathfrak{D}}(\omega, a, \tau) = N_j^{\mathfrak{D}}(\omega, \cdot, \tau) \end{aligned} \tag{2.11}$$

Então, aplicando (2.10) e (2.11) a (2.6), temos:

$$\ell(l, r \mid \mathfrak{D}) = \sum_{j=1}^{s} \sum_{\omega \in A^l} \sum_{\tau \in A^r} \sum_{a \in A} \left(N_j^{\mathfrak{D}}(\omega, a, \tau) \log_{|A|} \left(\frac{N_j^{\mathfrak{D}}(\omega, a, \tau)}{N_j^{\mathfrak{D}}(\omega, \cdot, \tau)} \right) \right) \tag{2.12}$$

Como a fração dentro do logaritmo em (2.12) está sempre no intervalo $[0, 1]$, o valor dos logaritmos é sempre negativo. A equação então nos fornece sempre um valor menor que zero. Para chegar à expressão que utilizaremos, falta notar que quanto menor for a vizinhança, menor será o valor de (2.12)

pois maior será o valor do fator linear que multiplica cada somando. Assim, a vizinhança que maximiza (2.12) é sempre a maior vizinhança possível.

Temos então que adicionar um termo de penalização à (2.12), que dependa explicitamente do alcance da vizinhança considerada. Aqui utilizaremos a penalização como em [3]: uma constante $c = \frac{(|A|-1)}{2}$, multiplicada por um termo que seja função do tamanho da vizinhança $t(\omega, \tau) = |A|^{|\omega\tau|}$. Separando (2.12) para cada j, finalmente temos uma expressão para a pseudo-verossimilhança penalizada para um `SNP` j em função de l e r:

$$\check{\ell}_j(l, r \mid \mathfrak{D}) = \sum_{\omega \in A^l} \sum_{\tau \in A^r} \sum_{a \in A} \left(N_j^{\mathfrak{D}}(\omega, a, \tau) \log_{|A|} \left(\frac{N_j^{\mathfrak{D}}(\omega, a, \tau)}{N_j^{\mathfrak{D}}(\omega, \cdot, \tau)} \right) \right) - \\ - \frac{(|A|-1)}{2} |A|^{|\omega\tau|} . \log_{|A|}(n) \tag{2.13}$$

Para encontrar $\hat{l}_j$ e $\hat{r}_j$, as vizinhanças estimadas associadas à posição de `SNP` j, basta encontrar o argumento que maximize a expressão acima:

$$(\hat{l}_j, \hat{r}_j) = \arg\max_{l,r \in \mathbb{N}} \left(\check{\ell}_j(l, r \mid \mathfrak{D}) \right) \tag{2.14}$$

2.3 Algoritmo

Um algoritmo para calcular (2.14) para todo `SNP` j deve realizar, para cada posição de `SNP`, uma contagem simples de palavras num alfabeto $A = \{0, 1, 2\}$, levando em consideração todas as vizinhanças de diferentes tamanhos à esquerda e a direita. É importante, no entanto, fazer algumas considerações acerca do custo em tempo e espaço de tal algoritmo.

Em nosso caso, estamos analisando 500.180 `SNP`s de 2.062 indivíduos. Se utilizarmos 1 byte para representar cada entrada da matriz $\mathfrak{D}$ de amostras, precisaremos de algo em torno de 1Gb para manter a matriz inteira em memória (em oposição a cerca de 4Gb caso representássemos cada entrada como um inteiro de 32 bits). Hoje em dia os computadores pessoais mais simples possuem 3Gb ou 4Gb de memória principal, então este número, para nosso conjunto de dados, é factível. No entanto, para a análise de um número 10 vezes maior de indivíduos ou `SNP`s, extrapolaríamos os limites razoáveis e precisaríamos otimizar a representação dos dados ou utilizar técnicas numéricas ou de paralelismo para auxiliar nas contas.

Para dimensionar o tempo gasto pelo algoritmo em função do tamanho da entrada, observemos que em (2.13) o termo de penalização cresce exponencialmente com o tamanho da vizinhança, enquanto que a somatória cresce com velocidade $O(n.\log(n))$, e assim, é possível estabelecer limites para o tamanho das vizinhanças que serão analisadas. Na implementação consideramos vizinhanças de tamanho $l \leq 20$ para a esquerda e $r \leq 20$ para

a direita, mas na prática nenhuma das vizinhanças teve tamanho maior do que 5 SNPs.

Então, para cada SNP, são consideradas $20 * 20 = 400$ vizinhanças diferentes, utilizando uma *árvore de vizinhança*, conforme descrito na próxima seção, para cada uma das vizinhanças. Em cada uma destas árvores são feitas n inserções onde cada inserção custa tempo menor ou igual ao tamanho da maior palavra considerada (40 símbolos), constante. A análise subsequente das vizinhanças e o cálculo das vizinhanças ótimas não são influenciados pelo número de indivíduos na amostra. Assim, nosso algoritmo leva tempo linear no tamanho da matriz de amostras, ou seja, $O(n.s)$.

2.3.1 Estrutura de dados: árvore de vizinhança

Um algoritmo que calcule a pseudo-verossimilhança penalizada para um SNP da forma definida na seção anterior precisa considerar diversas vizinhanças diferentes e realizar uma contagem de palavras que ocorrem na amostra, para cada vizinhança considerada. Utilizando estas contagens, o algoritmo deve calcular a pseudo-verossimilhança penalizada para cada vizinhança e decidir qual é a vizinhança ótima estimada para aquele SNP.

Uma forma simples de contabilizar palavras é inseri-las em uma árvore onde cada nó tem no máximo $|A|$ filhos, de forma que a sequência de arestas de cada caminho da raiz até uma folha represente uma palavra diferente. A inserção de uma palavra nova é feita sempre a partir da raiz e a aresta a ser seguida para a inserção do próximo símbolo é decidida com base no valor do símbolo a ser inserido. Se mantivermos um contador de visitas em cada nó, ao final de todas as inserções teremos, em cada nó, o número de palavras inseridas relacionadas ao caminho que leva da raiz até o nó em questão.

A idéia é utilizar uma árvore nova para cada vizinhança de cada SNP e, após inserir todas as palavras de uma certa vizinhança em uma árvore vazia, prosseguir com o cálculo da pseudo-verossimilhança penalizada desta vizinhança. Neste caso, a árvore não contém informação nenhuma sobre a vizinhança considerada, mas somente sobre as palavras que nela ocorrem. As árvores são utilizadas em subrotinas do algoritmo, e a verificação da máxima pseudo-verossimilhança penalizada para cada SNP e criação de novas árvores para cada vizinhança ficam a cargo de uma instância mais geral do algoritmo.

Para cada amostra de cada vizinhança considerada, é necessário identificar o símbolo em torno de qual esta vizinhança ocorre, e para isto utilizamos não apenas um, mas $|A| = 3$ contadores por nó, e associamos cada contador a um símbolo diferente. Assim, basta "lembrar", ao longo da inserção de cada palavra da vizinhança, a informação do símbolo em torno do qual esta vizinhança ocorreu. Ao final teremos em cada contador a quantidade de palavras em torno de cada símbolo para uma vizinhança.

A uma árvore com $|A|$ contadores e no máximo $|A|$ filhos em cada nó,

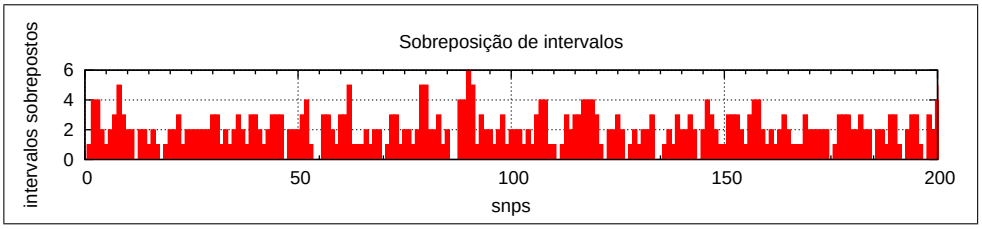

Figura 1: Sobreposição de vizinhanças de influência para os 200 primeiros SNPs do mapa. O valor para cada SNP indica em quantas vizinhanças de influência este SNP está contido.

onde cada contador e cada aresta está associada a um símbolo de A, damos o nome de *árvore de vizinhança*. Esta é a estrutura de dados principal que utilizamos nos algoritmos implementados, tanto para estimar as vizinhanças de cada SNP, quanto para calcular as distâncias e os testes Chi-Quadrado na próxima seção.

2.4 Resultados e estatísticas

Aplicando o modelo descrito na seção 2.2 aos dados descritos na seção 1.3, obtemos vizinhanças de influência de tamanho médio 2,22418 SNPs e desvio padrão 0,714481 SNP. O tamanho médio das vizinhanças para a esquerda é de 1,11432 e para a direita é de 1,10985 SNP. Isto significa que, em nosso modelo, 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 número de vizinhanças de influência dos SNPs adjacentes a j que o contém.

Para ter uma idéia de como estas vizinhanças estão relacionadas entre si, podemos ver na Figura 2 os valores para cada SNP, com o tamanho das vizinhanças à esquerda representados como valores negativos, e o tamanho das vizinhanças à direita 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 onde todos os SNPs do bloco estão contidos nas vizinhanças de influência de todos os outros SNPs do bloco. Na próxima seção formalizaremos esta idéia definindo *janelas de influência*.

3 Janelas de influência

A partir das vizinhanças determinadas na seção anterior, uma pergunta interessante de formular é 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

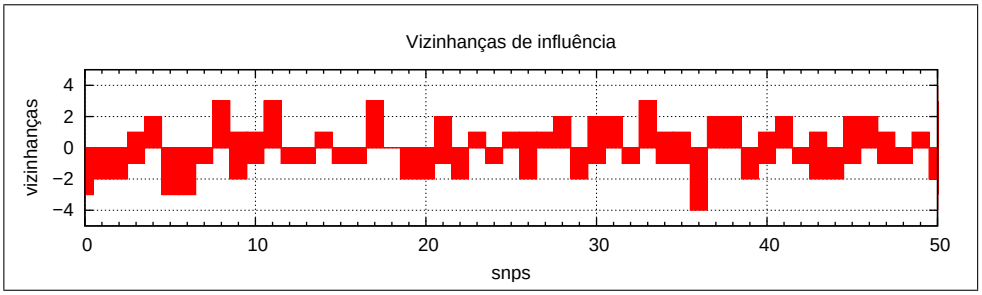

Figura 2: Vizinhanças de influência para os 50 primeiros SNPs. O valor positivo é o tamanho da vizinhança à esquerda, e o valor negativo é o tamanho da vizinhança à direita.

SNP da janela estão contidos na própria janela. Com esta idéia podemos definir formalmente o tipo de estrutura que estamos procurando.

Definição: Uma *janela de influência* é uma sequência $J \subseteq \{1, \ldots, S\}$ *adjacente* de SNPs tal que para todo $j \in J$, $V_j \subseteq J$.

Analisando as vizinhanças determinadas anteriormente obtemos que, segundo nosso modelo, 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.

3.1 Relevância para o desenvolvimento da Artrite Reumatóide

Uma vez determinadas as janelas de influência da forma descrita na seção anterior, precisamos de meios para analisar a relevância de cada janela na diferenciação entre indivíduos *caso* e *controle*.

A seguir, desenvolveremos dois métodos de análise com o objetivo de verificar se as distribuição de probabilidade de indivíduos *caso* e *controle* relacionadas a cada janela são independentes. Primeiro estabeleceremos algumas medidas de distância diferentes entre distribuições e visualizaremos graficamente os resultados das medidas para cada janela. Em seguida, realizaremos um teste de hipótese para cada janela encontrada, de forma a verificar se há evidências para alguma janela de que as distribuições realmente sejam independentes.

3.2 Métricas

Dada uma janela de influência $J_i = (X_{i_1}, X_{i_2}, \ldots, X_{i_n})$, podemos estimar duas distribuições de probabilidade sobre J_i, uma para o conjunto de *casos*, que chamaremos P_i, e outra para o conjunto *controle*, que denotaremos por Q_i. Dada uma sequência $\omega \in A^{|J_i|}$, onde $|J_i|$ é o tamanho de J_i, estimaremos

cada valor $P_i(\omega)$ e $Q_i(\omega)$ através das frequências relativas a cada grupo de indivíduos.

Lembrando a notação introduzida anteriormente, se K é o conjunto de índices dos indivíduos *caso* e L é o conjunto de índices dos indivíduos *controle*, então chamamos de $\mathfrak{D}_{\restriction K}$ a amostra $\mathfrak{D}$ restrita aos indivíduos *caso* e de $\mathfrak{D}_{\restriction L}$ a amostra $\mathfrak{D}$ restrita aos indivíduos *controle*. Aproveitando a notação anterior com uma pequena modificação, $N_{J_i}^{\mathfrak{D}_{\restriction X}}(\omega)$ representará aqui o número de vezes que a sequência ω aparece na amostra $\mathfrak{D}$ restrita ao conjunto de indivíduos de índices em X (no caso, K ou L), na posição da janela J_i. De posse destas notações, temos os seguintes estimadores:

$$\hat{P}_i(\omega) = \frac{N_{J_i}^{\mathfrak{D}_{\restriction K}}(\omega)}{|\mathfrak{D}_{\restriction K}|} \quad \text{e} \quad \hat{Q}_i(\omega) = \frac{N_{J_i}^{\mathfrak{D}_{\restriction L}}(\omega)}{|\mathfrak{D}_{\restriction L}|}. \tag{3.1}$$

Com estes estimadores, podemos definir algumas medidas diferentes para distância entre as distribuições em uma janela.

- **Somatória de diferenças:**

$$D_1(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|J_i|}} |\hat{P}_i(\omega) - \hat{Q}_i(\omega)| \tag{3.2}$$

  Esta medida é obtida somando-se as diferenças entre frequências de cada palavra, para todas as palavras de uma janela.

- **Sequência com maior diferença:**

$$D_2(\hat{P}_i, \hat{Q}_i) = \max_{\omega \in A^{|J_i|}} |\hat{P}_i(\omega) - \hat{Q}_i(\omega)| \tag{3.3}$$

  Em uma variação da medida anterior, ao invés de somarmos todas as diferenças simplesmente escolhemos entre todas as palavras, aquela que dá a maior diferença entre indivíduos *caso* e *controle* naquela janela.

- **Divergência de Kullback-Leibler:**

$$D_3(\hat{P}_i, \hat{Q}_i) = \sum_{\omega \in A^{|J_i|}} \hat{P}_i(\omega). \log \frac{\hat{P}_i(\omega)}{\hat{Q}_i(\omega)} \tag{3.4}$$

  A divergência de Kullback-Leibler é uma medida bastante utilizada para estimar distância entre distribuições e analisar a entropia existente em conjuntos de dados. Note que esta medida não é simétrica, ou seja, $D_3(\hat{P}_i, \hat{Q}_i)$ não é necessariamente igual a $D_3(\hat{Q}_i, \hat{P}_i)$. Inclusive, observando os resultados do cálculo de $D_3(\hat{P}_i, \hat{Q}_i)$, notamos que

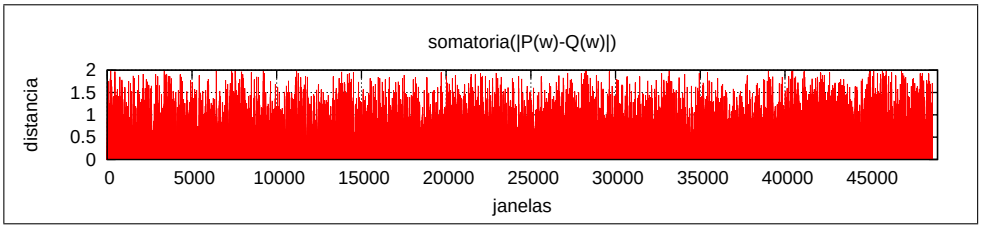

Figura 3: As distâncias entre janelas *caso* e *controle* utilizando a medida D_1 não evidenciam nenhuma área em especial.

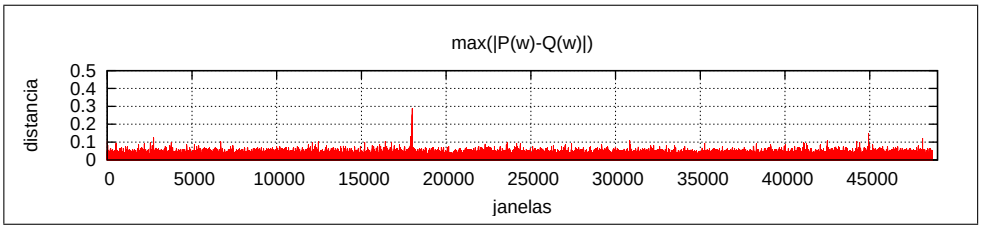

Figura 4: Um pico de distância é detectado no cromossomo 6 utilizando D_2.

nenhuma região é evidenciada por esta medida, ao menos numa primeira vista. Podemos então definir uma nova medida simétrica, apenas calculando a média aritmética entre os dois possíveis arranjos dos argumentos de (3.4):

- **Divergência de Kullback-Leibler ponderada:**

$$D_4(\hat{P}_i, \hat{Q}_i) = \frac{D_3(\hat{P}_i, \hat{Q}_i) + D_3(\hat{Q}_i, \hat{P}_i)}{2} \tag{3.5}$$

3.2.1 Resultados

Ao analisar o resultado do cálculo de D_1 nas janelas de influência, que pode ser visto na Figura 3, verificamos que não há nada que seja especialmente evidenciado por esta medida a uma primeira vista. Já nas Figuras 4, 7 e 9

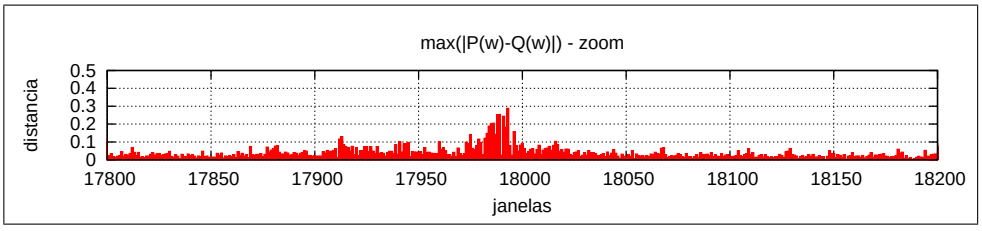

Figura 5: Um pico de distância é detectado no cromossomo 6 utilizando D_2 - visão mais próxima.

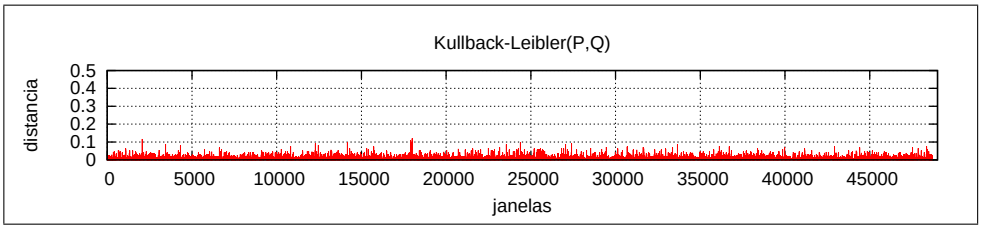

Figura 6: Divergência de Kullback-Leibler, medida assimétrica, não evidencia nenhuma região em especial quando calculamos $D_3(\hat{P}_i, \hat{Q}_i)$.

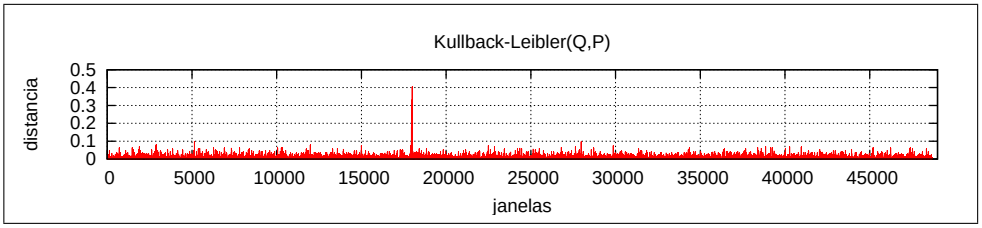

Figura 7: Divergência de Kullback-Leibler, medida assimétrica, evidencia uma região no cromossomo 6 ao calcular $D_3(\hat{Q}_i, \hat{P}_i)$.

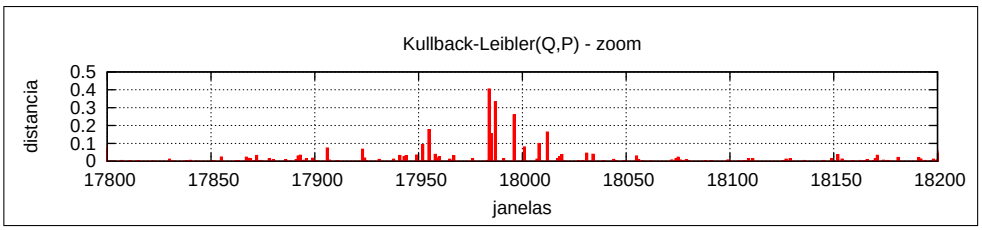

Figura 8: Divergência de Kullback-Leibler detecta um pico de distância no cromossomo 6 - visão mais próxima.

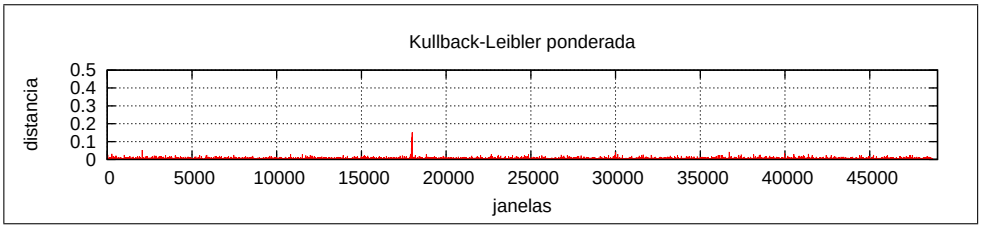

Figura 9: Divergência de Kullback-Leibler simetrizada detecta um pico de distância no cromossomo 6.

podemos ver que o resultado das outras medidas mostra um pico de distância nas janelas em uma região próxima à janela de índice 20.000. Olhando mais de perto, nas Figuras 5 e 8, identificamos a região de cerca de 80 janelas, entre as janelas de índices 19.920 e 19.980, como a mais intensa de toda a extensão.

É interessante notar a diferença dos resultados utilizando a medida assimétrica D_3. Invertendo os argumentos passados para a função de cálculo da medida, passamos de uma situação onde não é possível notar evidenciação de nenhuma região em especial (Figura 6) para uma situação onde é nitidamente evidente uma distância maior na mesma região indicada pelas medidas simétricas (Figura 7). O resultado de D_4, a versão simétrica de D_3, pode ser visto na Figura 9.

Estes resultados estão de certa forma em concordância com a literatura existente no assunto: as janelas compreendidas no intervalo de 17.230 até 20.455 pertencem ao cromossomo 6 do genoma humano e existem na literatura muitos resultados indicando forte influência de `SNP`s desta região na presença da Artrite Reumatóide nos seres humanos[3] [7, 8].

Para cada medida simétrica realizada, podemos calcular a média e desvio padrão dos valores de distância e utilizar estes dados estatísticos para tentar selecionar as janelas mais influentes. Uma forma de fazer isto é utilizar como critério um valor de distância maior do que, digamos, dois desvios padrões acima da média. Este critério, juntamente com os resultados desta seção, serão utilizados ao final do trabalho para determinar quais são os `SNP`s mais influentes na presença da Artrite Reumatóide, de acordo com nosso modelo.

3.3 Teste de Chi-Quadrado

Uma outra forma de avaliar se as sequências que ocorrem em uma certa janela têm relação com o fato de cada indivíduo possuir a doença Artrite Reumatóide é realizar um teste de Chi-Quadrado relacionando as frequências observadas de cada sequência na janela para cada tipo de indivíduo. Tratam-se portanto de 48.697 testes diferentes, um para cada janela. Para isto, é possível utilizar a mesma infra-estrutura desenvolvida anteriormente para contar frequências, de forma a gerar as tabelas para cada teste.

Temos que montar, para cada janela J_i, uma tabela onde cada coluna corresponde a uma sequência $\omega_l \in A^{|J_i|}$ e cada linha representa uma classe de indivíduos. Cada entrada da tabela é o número de vezes que ω_l aparece na janela J_i para cada tipo de indivíduo. A última coluna (linha) contém a soma das entradas de cada linha (coluna) da tabela. Considerando que o número de sequências que pode aparecer em uma janela de tamanho $|J_i|$ é $m = |A^{|J_i|}|$, uma tabela montada para as frequências das sequências na janela J_i tem o seguinte formato:

[3] http://www.naracdata.org/naracpub.asp

	ω_1	...	ω_m	
caso	$N_{J_i}^{\mathfrak{D}_{\restriction K}}(\omega_1)$	...	$N_{J_i}^{\mathfrak{D}_{\restriction K}}(\omega_m)$	$\lvert\mathfrak{D}_{\restriction K}\rvert$
controle	$N_{J_i}^{\mathfrak{D}_{\restriction L}}(\omega_1)$	...	$N_{J_i}^{\mathfrak{D}_{\restriction L}}(\omega_m)$	$\lvert\mathfrak{D}_{\restriction L}\rvert$
	$N_{J_i}^{\mathfrak{D}}(\omega_1)$	...	$N_{J_i}^{\mathfrak{D}}(\omega_m)$	$\lvert\mathfrak{D}\rvert$

O teste de Chi-Quadrado requer que sejam calculadas frequências esperadas para cada evento possível, a partir das frequências totais observadas para cada classe de evento. O valor esperado para uma casela da tabela é calculado multiplicando os totais da linha e coluna da casela e dividindo pelo valor total de amostras. No contexto que estamos estudando, o cálculo do valor esperado para a sequência ω na janela J_i é feito através das seguintes expressões para os indivíduos *caso* e *controle* respectivamente:

$$E_i^K(\omega) = \frac{N_{J_i}^{\mathfrak{D}}(\omega).|\mathfrak{D}_{\restriction K}|}{|\mathfrak{D}|} \quad \text{e} \quad E_i^L(\omega) = \frac{N_{J_i}^{\mathfrak{D}}(\omega).|\mathfrak{D}_{\restriction L}|}{|\mathfrak{D}|} \tag{3.6}$$

A estatística Chi-Quadrado para J_i é então calculada pela somatória das diferenças entre valores observados e esperados elevadas ao quadrado e normalizadas, para todas as entradas da tabela.

$$\chi_i^2 = \sum_{\omega \in A^{|J_i|}} \left(\frac{\left(O_i^K(\omega) - E_i^K(\omega)\right)^2}{E_i^K(\omega)} + \frac{\left(O_i^L(\omega) - E_i^L(\omega)\right)^2}{E_i^L(\omega)} \right) \tag{3.7}$$

3.3.1 Resultados

O teste de Chi-Quadrado nas janelas foi realizado utilizando alguns níveis de significância diferentes. Como nossa amostra é relativamente grande (2.062 indivíduos), o teste com nível de significância de 1%, cujo resultado pode ser visto na Figura 10, não revela muita coisa. Apesar disso, diminuindo o nível de significância para $0,01\%$ (Figura 11) e em seguida para $0,0001\%$ (Figura 12), podemos ver que a região evidenciada está compreendida entre as janelas 17.906 e 18.018. Esta região contém as janelas observadas como influentes na seção anterior através das métricas utilizadas.

Algumas outras regiões também são evidenciadas, o que é perfeitamente normal dado o tamanho de nossa amostra. Testes de Chi-Quadrado em geral são feitos para amostras pequenas, e em amostras muito grandes é comum obter-se eventos infrequentes, o que influencia nos resultados do teste.

4 Conclusão

Pudemos observar que através do cálculo das pseudo-verossimilhanças penalizadas para cada `SNP` foi possível encontrar regiões de alta influência entre

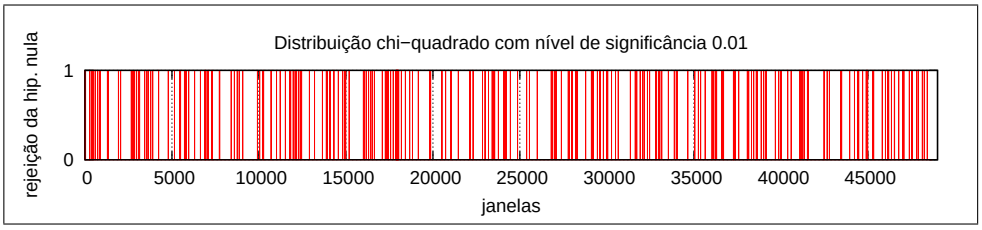

Figura 10: Teste de Chi-Quadrado para as janelas de influência com nível de significância de 1%.

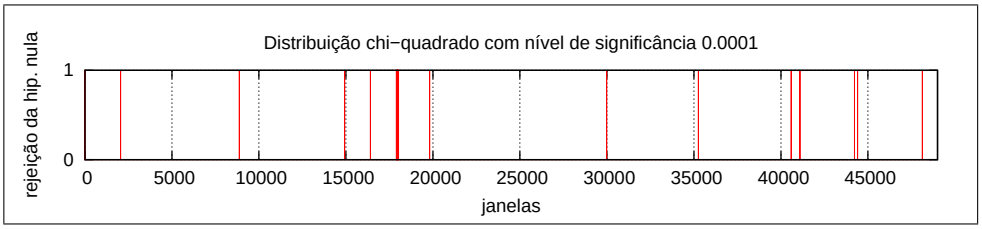

Figura 11: Teste de Chi-Quadrado para as janelas de influência com nível de significância de $0,01\%$.

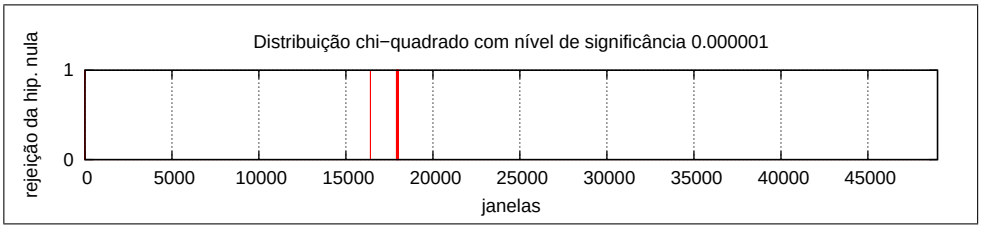

Figura 12: Teste de Chi-Quadrado para as janelas de influência com nível de significância de $0,0001\%$.

SNPs adjacentes. Observando com atenção a Figura 2 identificamos diversas sequências de SNP nas quais as vizinhanças se sobrepõe e, para algumas delas, a vizinhança é a mesma para todos os SNPs da sequência. Isto sugere que existe realmente uma influência dos valores dos SNPs adjacentes no valor assumido por um certo SNP.

Estas observações nos levaram ao conceito de *janela de influência*: uma sequência adjacente de SNPs tal que todas as vizinhanças dos SNPs da sequência estão contidas na própria sequência.

Uma vez definidas as janelas de influência, passamos à observação da relevância de cada janela para o desenvolvimento da Artrite Reumatóide em um indivíduo. Para isto, fizemos dois tipos de análise: a primeira utilizando diversas métricas diferentes para determinar distâncias entre as distribuições dos indivíduos *caso* e *controle* para cada janela, e a segunda realizando um teste de Chi-Quadrado para cada janela, de forma a tentar determinar em que janelas o fato dos indivíduos estarem doentes ou não é independente das sequências específicas de SNPs que ocorrem nas janelas.

Uma forma de tentar obter resultados mais próximos da realidade é levar em conta a intersecção dos resultados dos dois métodos de estimação utilizados. Apenas duas janelas foram identificadas como influentes em todos os testes, as de número 17.942 e 18.002, nas quais estão compreendidos 6 SNPs: rs2844654, rs1264333, rs2051549, rs2071550, rs1573649, rs6903130 e rs6901084.

As manipulações de dados realizadas por softwares estatísticos específicos para análise de dados genotípicos são muito mais complexas do que as feitas neste trabalho, e os pesquisadores da área geralmente realizam diversas filtragens dos dados para eliminar influências de ancestralidade, de erros de genotipagem, e de outros ruídos que podem interferir nos resultados. As análises feitas neste trabalho e os resultados obtidos talvez não estejam livres de defeitos, mas de fato indicam regiões do DNA e, mais especificamente, alguns SNPs que têm relação estatística com o desenvolvimento da Artrite Reumatóide. A vantagem da análise que desenvolvemos é sua simplicidade em termos estatísticos e o evidenciamento tão explícito de certas áreas. Este estudo compõe um ensaio sobre uma tentativa de compreensão das estruturas do DNA utilizando ferramentes estatísticas e matemáticas relativamente simples.

Parte II
O Trabalho de Conclusão e o curso de Computação

5 O Trabalho de Conclusão de Curso

No início de 2008 cursei a disciplina *MAE0228 - Noções de Probabilidade e Processos Estocásticos*, obrigatória para o meu curso, Bacharelado em Ciência da Computação (BCC). O assunto me interessou muito e no semestre seguinte (o segundo de 2008) cursei a disciplina optativa *MAE0326 - Aplicações de Processos Estocásticos*. Na época, conversei com meu professor que me disse que havia muitos trabalhos interessantes que poderiam ser desenvolvidos na área, relacionados a modelagem e análise estocástica de dados de DNA. Este professor me indicou minha atual orientadora, a professora Florencia Graciela Leonardi, do Departamento de Estatística do IME-USP.

A partir do início de 2009, começamos a conversar sobre o trabalho que seria desenvolvido. A professora Florencia estava em contato com outras pesquisadoras da área de bioinformática, pesquisadoras estas que viriam a nos fornecer o conjunto de dados do GAW16 que utilizamos neste trabalho. Fizemos diversas reuniões apenas entre nós, com uma frequência de algo entre duas e três vezes ao mês, e também algumas reuniões com outros pesquisadores do IME-USP. A primeira dessas reuniões teve como objetivos a apresentação do conjunto de dados e de uma visão geral dos modelos já existentes de análise, e uma primeira conversa sobre as possibilidades de novos estudos nesse cenário. Nas reuniões seguintes, conseguimos traçar um caminho para o que deveria ser feito, e de que forma deveria ser conduzido.

A princípio, a idéia era aplicar modelos de Cadeias de Markov com alcance variável e verificar se seria possível extrair dos dados alguma informação relevante. Porém, ao pensar sobre um modelo razoável, verificamos que a idéia de vizinhanças adjacentes para a esquerda e para a direita não encaixava nestes tipos de modelo pois essas vizinhanças não expressam o mesmo tipo de causalidade que as cadeias de Markov. Enquanto nas cadeias de Markov uma mesma sequência sempre gera um símbolo com uma mesma probabilidade, em nosso modelo uma mesma sequência poderia gerar símbolos com probabilidades diferentes, se ocorresse em vizinhanças distintas. Por exemplo, suponhamos que estamos analisando as vizinhanças de 0**1**00 e 00**1**0, onde a posição do símbolo 1 é o `SNP` em questão. Ora, as sequências dadas pelas vizinhanças são a mesma, 000, mas não é razoável imaginar que tenham a mesma relação causal com o `SNP`, pois o tamanho à esquerda e à direita em cada caso é diferente.

No fim, resolvemos utilizar algumas técnicas de cadeias de alcance variável

em um contexto um pouco diferente, e foi quando escolhemos a abordagem da pseudo-verossimilhança penalizada. As idéias de agrupar os SNPs em janelas e de aplicar diferentes medidas e testes de Chi-Quadrado em cada janela vieram ao longo do trabalho. Foi uma experiência bastante agradável em termos de liberdade de investigação dos dados, mas por vezes também causou um pouco de angústia quando não havia imaginação suficiente para levar o trabalho adiante.

O resultado foi extremamente satisfatório como experiência pessoal, e espero poder contribuir com o estudo da influência de SNPs em características fenotípicas, mesmo que o trabalho tenha uma abordagem relativamente simples em termos de análise estatística.

5.1 Relevância do curso de BCC no desenvolvimento do Trabalho

É claro que o curso como um todo teve grande influência no desenvolvimento deste trabalho, por exemplo em matérias cujo assunto não tem nada a ver com probabilidade ou estatística, mas que ajudaram a desenvolver as noções de formalidade na escrita científica e a maturidade necessária para a compreensão de estruturas matemáticas complexas. Foi o caso, por exemplo, de *MAC0430 - Algoritmos e Complexidade Computacional*, na qual tive de fazer trabalhos extensos sobre assuntos de difícil digestão, ou de *MAT0213 - Álgebra II* que me introduziu pela primeira vez a modelos matemáticos abstratos. Apesar disso, é possível identificar algumas disciplinas que tiveram especial relevância para este trabalho.

Listo abaixo as disciplinas cursadas no BCC que foram mais relevantes em termos de uso direto dos conhecimentos adquiridos nessas matérias no presente trabalho:

- **MAE0121 - Introdução à Probabilidade e Estatística 1 e MAE0122 - Introdução à Probabilidade e Estatística 2:** Noções básicas necessárias para compreensão dos problemas e métodos de estimação, noções de probabilidade condicional e teste de hipótese.
- **MAE0228 - Noções de Probabilidade e Processos Estocásticos e MAE0326 - Aplicações de Processos Estocásticos (optativa):** familiaridade com aplicações de probabilidade e modelagem computacional.
- **MAC0323 - Estruturas de Dados:** programação em C e utilização de árvores para representação de dados.
- **MAT0121 - Cálculo Diferencial e Integral II:** derivadas parciais.

6 Trabalhos futuros

Me inscrevi para o Programa de Pós Graduação em Ciência da Computação no IME-USP e no momento da entrega deste trabalho aguardo o resultado da seleção. Ainda estou incerto quanto à área que pretendo seguir, mas sem dúvida este trabalho despertou bastante interesse no campo da Bioinformática. A literatura no assunto é vasta e para a continuidade da aprendizagem seria interessante frequentar congressos e tomar contato com mais pesquisadores da área.

O IME-USP é, com certeza, um dos melhores lugares do mundo para desenvolver trabalhos matemáticos e computacionais, e se for aprovado na Pós Graduação tenho certeza de que conseguirei desenvolver trabalhos interessantes e relevantes, independente da área que escolha.

6.1 Agradecimentos

Agradeço à professora Florencia Leonardi, do Departamento de Estatística do IME-USP, por me orientar neste trabalho e suportar minhas dúvidas e dedicar parte de seu tempo. À professor Júlia Pavan, do mesmo departamento, também dedico um agradecimento especial por ler meu trabalho e me indicar diversas possibilidade de melhoria no texto e na análise dos dados.

Agradeço ao CNPQ que, desde Setembro de 2009, me concedeu uma Bolsa de Iniciação Científica para que realizasse o trabalho.

Obrigado também à sociedade Paulista, que me manteve na universidade todos estes anos através do financiamento público e pagamento de impostos. O conhecimento adquirido, retornará à sociedade de diversas formas, seja em trabalhos científicos desenvolvidos na academia, através de produção de software livre ou desenvolvimento de projetos não acadêmicos informais nos quais terei a oportunidade de passar o que aprendi adiante.

Por fim, agradeço à minha família pelo suporte dado todos esses anos, tanto emocionalmente quanto financeiramente.

Referências

[1] Altshuler, D., Daly, M. J., and Lander, E. S. Genetic mapping in human disease. *Science (New York, N.Y.) 322*, 5903 (November 2008), 881–888.

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

[3] Csiszar, I., and Talata, Z. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *Information Theory, IEEE Transactions on 52*, 3 (March 2006), 1007–1016.

[4] Guttorp, P. *Stochastic Modeling of Scientific Data*. Stochastic Modeling. Chapman & Hall, 1995.

[5] Leonardi, F. Some upper bounds for the rate of convergence of penalized likelihood context tree estimators, 2007.

[6] Pearson, H. Genetics: what is a gene? *Nature 441*, 7092 (May 2006), 398–401.

[7] Plenge, R. M. e. a. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nature Genetics 8*, 39 (2007), 1477–1482.

[8] Plenge, R. M. M., Seielstad, M., Padyukov, L., Lee, A. T. T., Remmers, E. F. F., Ding, B., Liew, A., Khalili, H., Chandrasekaran, A., Davies, L. R. L. R., Li, W., Tan, A. K. S. K., Bonnard, C., Ong, R. T. H. T., Thalamuthu, A., Pettersson, S., Liu, C., Tian, C., Chen, W. V. V., Carulli, J. P. P., Beckman, E. M. M., Altshuler, D., Alfredsson, L., Criswell, L. A. A., Amos, C. I. I., Seldin, M. F. F., Kastner, D. L. L., Klareskog, L., and Gregersen, P. K. K. Traf1-c5 as a risk locus for rheumatoid arthritis – a genomewide study. *N Engl J Med* (September 2007).

[9] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics 38*, 8 (August 2006), 904–909.

[10] Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., and Lander, E. S. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science 280*, 5366 (May 1998), 1077–1082.