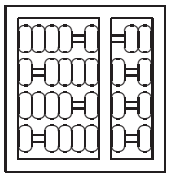
Comparação de métodos para determinação de SNPs com medidas de confiabilidade

Christian Baudet, Miguel Galves, Zanoni Dias



INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Comparação de métodos para determinação de SNPs com medidas de confiabilidade

Christian Baudet *Miguel Galves*
Zanoni Dias

Technical Report - IC-06-15 - Relatório Técnico

September - 2006 - Setembro

Comparação de métodos para determinação de SNPs com medidas de confiabilidade

Christian Baudet [*] Miguel Galves [†] Zanoni Dias [‡]

Resumo

Neste trabalho realizamos estudos sobre métodos para a identificação de polimorfismos de base única, comumente conhecidos pela sigla `SNP` (em inglês, *Single Nucleotide Polymorphism*). Identificar este tipo de polimorfismo é um processo importante pois, devido ao fato de eles poderem influenciar a estabilidade ou, até mesmo, a estrutura das proteínas codificadas pelos genes, os SNPs podem estar relacionados a diversas doenças.

Um bom método de identificação deve ser capaz de apontar as posições polimórfica e fornecer uma medida de confiabilidade para a detecção realizada. Neste sentido, estudamos dois métodos: `polybayes` [8] e `MSASNP`.

O primeiro método trata-se de um programa que realiza análise bayesiana para análise das seqüências e identificação dos SNPs. O segundo método adota uma estratégia simples, utilizado conceitos básicos de probabilidade.

Os testes mostraram que o `polybayes` é o mais indicado por ser capaz de indentificar os SNPs e fornecer valores mais confiáveis sobre a probabilidade de detecção correta de um SNP.

1 Introdução

Dizemos que há um polimorfismo em uma seqüência genética quando existe uma ou mais formas genéticas (alelos) distintas em indivíduos da mesma espécie. Para que um alelo seja considerado um polimorfismo, ele deve aparecer em, pelo menos, 1% da população analisada. Caso contrário, considera-se que o alelo é uma mutação pontual.

Polimorfismo de Base Única (SNP – *Single Nucleotide Polymorphism*, em inglês) é um polimorfismo que ocorre em apenas uma base [1]. Não são considerados SNPs as mutações (inserções ou remoções simples de bases) em uma seqüência genômica.

Os SNPs podem ser polimorfismos bi, tri ou tetra alélicos, ou seja, possuirem duas, três ou quatro formas distintas. Porém os dois últimos tipos são raros. As variações mais freqüentes são substituições entre bases nitrogenadas de mesma caraterística estrutural (um A por um G, um G por um A, um C por um T ou ainda um T por um C). Tais substituições chamamos de transições. As outras substituições chamamos de transversões.

[*]Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.
[†]Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.
[‡]Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.

Podemos classificar um SNP como sinônimo ou não. Dizemos que o SNP é um sinônimo se o aminoácido codificado pelo *codon* contendo o SNP é o mesmo que aquele codificado pelo *codon* sem SNP e é não sinônimo, caso contrário. Apesar de não alterar a seqüência de aminoácidos, um SNP sinônimo pode afetar a estabilidade da proteína codificada. Já um SNP não sinônimo pode modificar a estrutura e a função da proteína codificada. Um dos maiores interesses da pesquisa sobre o genoma humano é determinar se um SNP não sinônimo, chamado de nsSNP, afeta a função da proteína e conseqüentemente tem impacto sobre a saúde do indivíduo. Aproximadamente metade das causas genéticas de doenças originam-se da substituição de aminoácidos [3].

Neste trabalho apresentaremos métodos com medidas de confiabilidade para identificação de SNPs em *clusters*. Na Seção 2 decreveremos a fonte de dados utilizada para testes com o método por nós descrito e pelo sistema `polybayes`, assim como a bases de dados de referência com anotações sobre SNPs na fonte de dados. Os métodos estudados serão descrito na Seção 3 e na Seção 4 apresentaremos os resultados obtidos por eles. Finalmente, na Seção 5 concluíremos o estudo realizado apresentado as avaliações feitas em cima dos resultados apresentados.

2 Fonte de dados e base de referência

Utilizamos como fonte de dados para entrada em nossos testes uma banco de ESTs de cana-de-açúcar com SNPs anotados. Tais informações foram extraídos da base de dados do projeto SUCEST [9]. Inicialmente um conjunto de 291689 ESTs foi produzido. Tal conjunto é composto por seqüências com um tamanho médio de $829,44 \pm 182,60$ bp com qualidade média de $23,15 \pm 15,71$. Posteriormente as sequências genéticas foram agrupadas em *clusters* utilizando o pacote `cap3` [7]. Foram gerados 43141 *clusters* onde 16338 são *singlets* (*clusters* formados por um único EST).

2.1 Descrição do conjunto inicial

O método de obtenção dos polimorfismos está descrito em Grivet *et al.* 2001 [5] e Grivet *et al.* 2003 [6]. A detecção de polimorfismos em cada *cluster* foi feita em dois passos: inicialmente define-se como SNP uma posição onde o alelo menos frequente aparece no mínimo duas vezes na seção transversal do alinhamento, com qualidade superior ou igual a 20. O segundo passo consiste em filtrar os SNPs, mantendo apenas as posições cuja vizinhança de 10 bases (5 para cada lado) esteja perfeitamente alinhada com todos os outros ESTs do *cluster*.

Para cada *cluster*, o projeto anotou as posições de SNPs observados, as bases observadas nos ESTs alinhados e suas respectivas frequências.

2.2 Validação dos SNPs

Ao todo foram obtidos 8198 arquivos representando *clusters* (um *cluster* por arquivo), com 43029 posições de SNPs anotados. Para validar os dados obtidos, todos os *clusters* foram processados pelo `polybayes` [8], utilizando sua configuração padrão. Os arquivos phd

Análise dos SNPs da cana-de-açúcar

		Média/Cluster			
	Clusters	$PB \cap SC$	$SC \setminus PB$	$PB \setminus SC$	Total
$PB = SC$	181	1.8	0.0	0.0	1.8
$PB \supset SC$	6310	4.6	0.0	12.1	16.7
$PB \subset SC$	35	1.9	1.8	0.0	3.7
$PB \cap SC \neq \emptyset$	1261	7.8	2.2	17.4	27.4
$PB \cap SC = \emptyset$	254	0.0	2.0	5.2	7.2
$PB = \emptyset$	157	0.0	2.6	0.0	2.6
Total	8198	4.8	0.5	12.1	17.4

Tabela 1: Comparação entre resultados obtidos pelo projeto SUCEST e `polybayes`. As colunas $PB \cap SC$, $SC \setminus PB$ e $PB \setminus SC$ representam respectivamente SNPs que pertencem tanto ao conjunto SC quanto ao PB, apenas a SC e apenas a PB.

necessários para execução do programa foram gerados a partir das sequências em formato `fasta` e `qual` dos ESTs. Para cada *cluster*, comparamos o conjunto de SNPs obtidos pelo `polybayes`, que chamaremos de PB, e o conjunto de SNPs mapeados pelo projeto SUCEST, que chamaremos de SC. Os resultados foram agrupados da seguinte forma:

- *clusters* onde $PB = SC$
- *clusters* onde $PB \supset SC$ (onde $PB \neq \emptyset$)
- *clusters* onde $PB \subset SC$
- *clusters* onde $PB \cap SC \neq \emptyset$ (onde $PB \not\supset SC$ e $PB \not\subset SC$)
- *clusters* onde $PB \cap SC = \emptyset$ (onde $PB \neq \emptyset$)
- *clusters* onde $PB = \emptyset$

Os resultados estão sumarizados na Tabela 1.

A grande maioria de *clusters* com $PB \supset SC$ deve-se ao fato do projeto SUCEST ter removido SNPs cuja vizinhança de tamanho 10 não tivesse qualidade mínima de 20 e cujo alinhamento no *cluster* não fosse perfeito. Observando as seções transversais de posições onde `polybayes` não detectou SNPs marcados pelo projeto SUCEST, observamos que a grande maioria tem baixa cobertura, e possui apenas duas bases polimórficas.

Para efetuar as análises posteriores, foi montado um conjunto de *clusters* contendo apenas SNPs marcados tanto por `polybayes` quanto pelo projeto SUCEST. O conjunto contém 7787 *clusters*, contendo 39049 posições de SNPs (90.75% dos SNPs marcados inicialmente).

3 Métodos estatísticos para determinação de SNPs com medidas de qualidade

Na busca por métodos estatísticos que forneçam parâmetros de confiabilidade quanto a definição de um SNP, avaliamos dois métodos. O primeiro é a ferramenta `polybayes`. O segundo é um sistema simples que leva em conta as qualidades de cada base, determinadas pelo pacote `phred`.

No texto que segue chamaremos este sistema simples, por nós definido, de Método Simples de Avaliação de SNP ou, simplesmente, MSASNP.

3.1 `Polybayes`: detecção de SNPs por análise bayesiana

O programa `Polybayes` [8] utiliza um algoritmo de inferência Bayesiana para calcular a probabilidade de um dado alelo ser polimórfico. O algoritmo considera uma seção transversal de um alinhamento com N seqüências $R_1, \ldots, R_N$ como sendo uma permutação de N elementos que podem assumir os valores A, C, G ou T, num total de 4^N permutações de nucleotídeos.

A detecção de SNPs em um alinhamento múltiplo é efetuada avaliando-se a probabilidade de heterogeneidade de nucleotídeos em uma seção transversal do alinhamento múltiplo, ou seja, avaliando a probabilidade de que uma dada posição de uma seqüencia possa ter várias formas alélicas. A probabilidade $P(S_i|R_i)$ que um nucleotídeo S_i seja A, C, G, ou T é estimada à partir da probabilidade de erro $P_{ERROR,i}$ obtida da qualidade da base. Atribui-se $(1 - P_{ERROR,i})$ para a base determinada pelo `phred` e $P_{ERROR,i}/3$ para cada uma das três outras bases.

Cada seção do alinhamento é classificada de acordo com sua multiplicidade nucleotídica (número de alelos diferentes na seção transversal), a variação específica e a distribuição de alelos. Utiliza-se o valor $P_{POLY} = 0.003$ (um locus polimórfico a cada 333 bp) como a probabilidade total a priori de que um locus é polimórfico [2, 4]. Este valor é distribuído entre as bases para criar uma probabilidade a priori $P_{Prior}(S_1, \ldots, S_N)$ para cada permutação. Um valor a priori de $(1 - P_{POLY})/4$ é atribuído a cada uma das quatro permutações não polimórficas, correspondendo a uma composição de base uniforme $P_{Prior}(S_i)$.

A probabilidade Bayesiana a posteriori de uma permutação em um nucleotídeo em particular é calculada considerando 4^N permutações diferentes como conjunto de modelos conflitantes:

$$P(S_1, \ldots, S_N|R_1, \ldots, R_N) = \frac{F(S_1, \ldots, S_N)}{G(S_{i1}, \ldots, S_{iN})}$$

onde

$$F(S_1, \ldots, S_N) = \frac{P(S_{i1}|R_1)}{P_{Prior}(S_{i1})} \times \ldots \times \frac{P(S_{iN}|R_{iN})}{P_{Prior}(S_{iN})} \times P_{Prior}(S_1, \ldots, S_N)$$

e

Exemplo de análise de seção transversal

	Qualidade	P_{erro}	A	C	G	T
A	10	0,100	0,900	0,033	0,033	0,033
A	20	0,010	0,990	0,003	0,003	0,003
A	15	0,032	0,968	0,011	0,011	0,011
A	8	0,158	0,842	0,053	0,053	0,053
C	11	0,079	0,026	0,921	0,026	0,026
C	11	0,079	0,026	0,921	0,026	0,026
T	3	0,501	0,167	0,167	0,167	0,499
C	9	0,126	0,042	0,874	0,042	0,042
		P_B	$3,57 \times 10^{-6}$	$7,66 \times 10^{-9}$	$3,04 \times 10^{-13}$	$9,08 \times 10^{-13}$
		$P_{B'}$	1,000	0,999	0,317	0,589

Tabela 2: Exemplo de fase inicial do algoritmo para cálculo de probabilidade de SNP. Cada linha mostra a base determinada pelo `phred` na seção transversal, sua qualidade associada, a probabilidade de erro associada à qualidade e as probabilidades de acerto atribuídas a cada possível base para aquela posição da sequência. As duas últimas linhas indicam a probabilidade de existir somente a base B na seção transversal (P_B) e a probabilidade de a base B aparecer pelo menos uma vez ($P_{B'}$).

$$G(S_{i1},\ldots,S_{iN}) = \sum_{\forall s \in (S_{i1},\ldots,S_{iN})} \left(\frac{P(S_{i1}|R_1)}{P_{Prior}(S_{i1})} \times \ldots \times \frac{P(S_{iN}|R_{iN})}{P_{Prior}(S_{iN})} \times P_{Prior}(S_{i1},\ldots,S_{iN}) \right)$$

A probabilidade a posteriori Bayesiana de um SNP, P_{SNP}, é a soma das probabilidades a posteriori de todas as permutações heterogêneas. O cálculo é efetuado por um algoritmo recursivo. Um locus em um alinhamento múltiplo é considerado como SNP candidato se a probabilidade a posteriori correspondente for maior que o um valor de limiar $P_{SNP,MIN}$.

3.2 Descrição do método MSASNP

Dada uma seção transversal, calcula-se para cada base a probabilidade de erro P_{erro} (ou seja, a probabilidade que a base não seja aquela determinada pelo `phred`), que pode ser obtida a partir da qualidade Q da base, através da fórmula:

$$P_{erro} = 10^{\frac{-Q}{10}}.$$

A probabilidade de acerto P_{acerto} (probabilidade da base ser de fato aquela determinada por `phred`) é portanto definida por $P_{acerto} = 1 - P_{erro}$.

Para cada posição da seção transversal, assume-se que a base definida por `phred` tem probabilidade P_{acerto} de ser a correta, e que as três outras bases têm a mesma probabilidade $P_{erro}/3$ de ser a real base existente na sequência na posição do alinhamento.

Ao final do cálculo de probabilidade para cada base em cada posição da seção transversal, temos uma tabela $4 \times N$, onde N representa tamanho da seção transversal e cada coluna

Exemplo de cálculo de probabilidades de variações de SNP

	A	C	G	T
A	-	0,999	0,317	0,589
C	0,999	-	0,317	0,589
G	0,317	0,317	-	0,187
T	0,589	0,589	0,187	-

Tabela 3: Exemplo de cálculo de valores de probabilidade de variações de SNP ($P_{XY} = P_{X'} \times P_{Y'}$) utilizando os dados de $P_{B'}$ da Tabela 2.

representa as probabidades de uma dada base (A, C, G ou T). Para cada coluna representando uma base calculamos as probababilidades P_A, P_C, P_T, P_G da seção do alinhamento conter apenas aquela base. Assim temos que:

$$P_A = \prod_{i=1}^{N} P_{Ai},\ P_C = \prod_{i=1}^{N} P_{Ci},\ P_G = \prod_{i=1}^{N} P_{Gi},\ P_T = \prod_{i=1}^{N} P_{Ti}$$

onde P_{Ai}, P_{Ci}, P_{Ti} e P_{Gi} representam a probabilidade da seqüência i de uma seção transversal conter respectivamente as bases A, C, T, ou G. Um exemplo desse produtório está mostrado na Tabela 2.

A probabilidade da seção de um alinhamento conter ao menos uma base B é dada por:

$$P_{B'} = 1 - \prod_{i=1}^{N} (1 - P_{Bi})$$

onde B pode ser A, C, G ou T. A Tabela 2 também ilutra o cálculo destes valores.

A probabilidade de uma coluna do alinhamento conter um SNP é dada por:

$$P_{SNP} = 1 - (P_A + P_C + P_T + P_G)$$

e a probabilidade do SNP ser formado por um dado par de bases é calculado pela multiplicação das probabilidades de cada uma das bases aparecer ao menos uma vez, ou seja:

$$P_{XY} = P_{X'} \times P_{Y'}$$

(por exemplo $P_{AC} = P_{A'} \times P_{C'}$). Aqui discutiremos apenas SNPs bialélicos, contudo este conceito pode ser estendido para SNPs tri ou tetra alélicos.

Considerando a seção transversal do exemplo exibido na Tabela 2, a probabilidade de ocorrência de um SNP na posição seria de $P_{SNP} = 1 - 3,57 \times 10^{-6} - 7,66 \times 10^{-9} - 3,04 \times 10^{-13} - 9,08 \times 10^{-13} = 0.999996$. Utilizando o mesmo exemplo, a Tabela 3 exibe as probabilidades de ocorrência de cada uma das possíveis variações bialélicas de SNPs.

4 Comparação dos métodos `Polybayes` e MSASNP

Para avaliar os métodos utilizamos um conjunto de dados formado por 8198 *clusters* de seqüências de cana-de-açúcar que possuíam SNPs anotados pelos pesquisadores do projeto SUCEST. No total, a anotação realizada nestes *clusters* aponta 42853 posições de SNPs ($5,23$ SNP/*cluster*).

Como estamos trabalhando apenas com posições bialélicas, nós filtramos todas as posições que apresentaram variações tri ou tetra alélicas ou que apresentaram eventos de INDEL. Após a filtragem, obtivemos uma lista de 41558 posições, contendo SNPs bialélicos, distribuídas em 8115 *clusters* ($5,07$ SNP/*cluster*).

Executamos o software `polybayes` com e sem filtro de seqüências parálogas. Este filtro do `polybayes` analisa as seqüências e, a partir das discrepâncias que elas apresentarem em relação à âncora (consenso do *cluster*, neste caso), separa-as em dois grupos: nativas e parálogas. Para o cálculo de SNPs, o programa considera apenas as seqüências nativas. Quando o filtro é aplicado o número de seqüências a ser analisada em busca de SNPs é menor e, portanto, a execução é mais rápida. Se o filtro é desligado, além da execução ser mais lenta, geralmente o programa retorna um número maior de posições de SNPs.

A execução do `polybayes` sem o filtro de parálogos produziu um total de 172842 posições de polimorfismo ($21,08$ SNP/*cluster*). Deste total, 131622 posições eram SNPs bialélicos distribuídos em 8195 *clusters* ($16,06$ SNP/*cluster*). O tempo de execução foi de 661 minutos e 45 segundos em uma máquina com 2 processadores INTEL Xeon 3.2 GHz, 4 GB DDR ECC e 4 discos 320 ULTRA SCSI 133 GB rodando Fedora Core 4.

Utilizando o filtro, o número de posições polimórficas obtido foi de 138695, distribuídas em 8042 *clusters* ($16,05$ SNP/cluster). Destas, 103325 eram posições bialélicas distribuídas em 8029 *clusters* ($12,60$ SNP/*cluster*). Para produzir estes dados, o programa gastou 578 minutos e 59 segundos na mesma máquina.

Aplicamos o método MSASNP no mesmo conjunto de *clusters*. O método gastou 302 minutos e 27 segundos na mesma máquina utilizada para executar o `polybayes`.

Os dados brutos, produzido pelo método, indicavam os valores para todas as posições que tinham pelo menos duas bases diferentes em uma seção transversal do alinhamento. Um total de 4144426 posições apresentaram esta característica mínima, resultando em $505,54$ SNP/*cluster*.

Este número de posições é muito maior do que o número de posições indicadas pelo SUCEST, por exemplo. Obviamente, a maior parte não se trata de SNPs e, portanto, um critério deve ser criado para separar as posições que realmente são polimorfismos.

Decidimos utilizar como critério o valor de probabilidade P_{SNP} calculado. Contudo, este número, pela própria natureza do cálculo, tende a ser, na maioria dos casos, muito próximo de 1. Por exemplo, se utilizamos o valor mínimo de $0,9$ para considerar a posição como um SNP, temos um total de 4115998 posições ($502,07$ SNP/*cluster*), ou seja, apenas $0,68\%$ do conjunto total é descartado.

Para avaliar o efeito da escolha de diferentes valores para a probabilidade mínima requerida, utilizamos a fórmula $f(x) = 1 - 10^{-x}$, com x variando no intervalo $[1, 20]$, para definir o conjunto de probabilidades mínimas a ser testado. Os números de posições definidas como sendo SNPs, segundo cada valor utilizado, é exibido no gráfico da Figura 1

(curva vermelha). Neste gráfico podemos ver também o número de posições em que ocorreu correspondência com o conjunto de SNPs definido pelo SUCEST (curva verde). A curva azul representa o número de polimorfismos bialélicos encontrados pelo `polybayes` sem o filtro de parálogos. Já a curva magenta indica a intersecção entre `polybayes` e SUCEST.

Podemos observar no gráfico que o método MSASNP aponta muitas posições, apresentando número maior de SNPs que o `polybayes` em grande parte dos casos. Podemos notar também que apresenta sempre um número maior de SNPs que o apresentado pela intersecção entre SUCEST e `polybayes`.

O gráfico da Figura 2 exibe a porcentagem de posições de SNPs apontadas pelo sistema `polybayes` que conferem com os dados do SUCEST (curva verde). E na curva vermelha apresentamos os valores obtidos com o método MSASNP que conferem com os apresentados pelo SUCEST. Podemos observar que o `polybayes` acerta bastante e que o método MSASNP acerta cada vez menos quando impomos mais restrições.

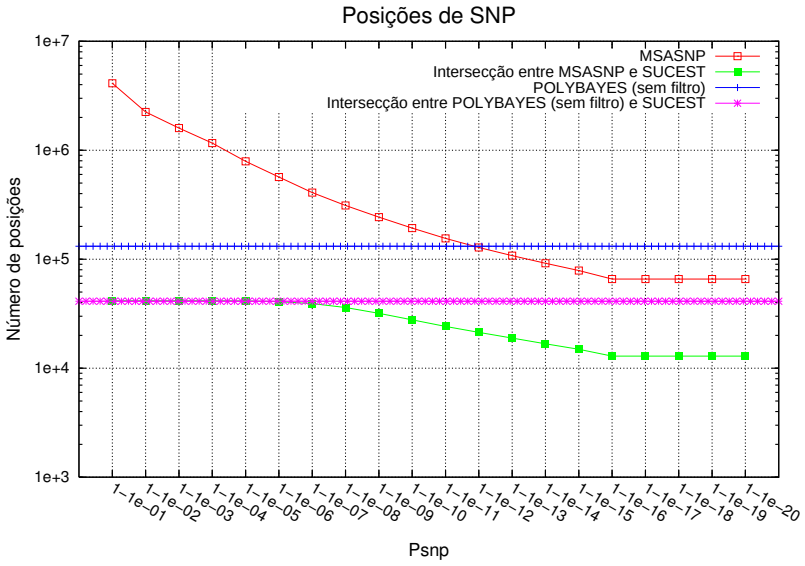


Figura 1: Gráfico comparativo no número de posições marcadas como SNP. No eixo X temos um limite inferior, para a probabilidade de ser SNP, onde consideramos que seja um SNP. No eixo Y temos o número de posições marcadas como sendo SNP. A curva vermelha refere-se ao método MSASNP. A curva azul apresenta o número de SNP apontados pelo `polybayes` (sem filtro). A curva magenta apresenta o número de SNPs que aparecem no SUCEST (dados de referência) e `polybayes` ao mesmo tempo. A curva verde aponta o número de SNPs que aparecem no SUCEST e método MSASNP ao mesmo tempo.

O gráfico nos mostra que mesmo com a utilização de um valor para probabilidade mínima, o método MSASNP continua a apontar muitas posições. Isso ocorre porque os alinhamentos dos *clusters* possuem muitas regiões com baixa qualidade, produzindo uma grande quantidade de SNPs em posições consecutivas.

Assim, decidimos aplicar um filtro de janela deslizante que percorre as posições do alinhamento e elimina SNPs consecutivos. A janela inicia a procura pelo primeiro candidato

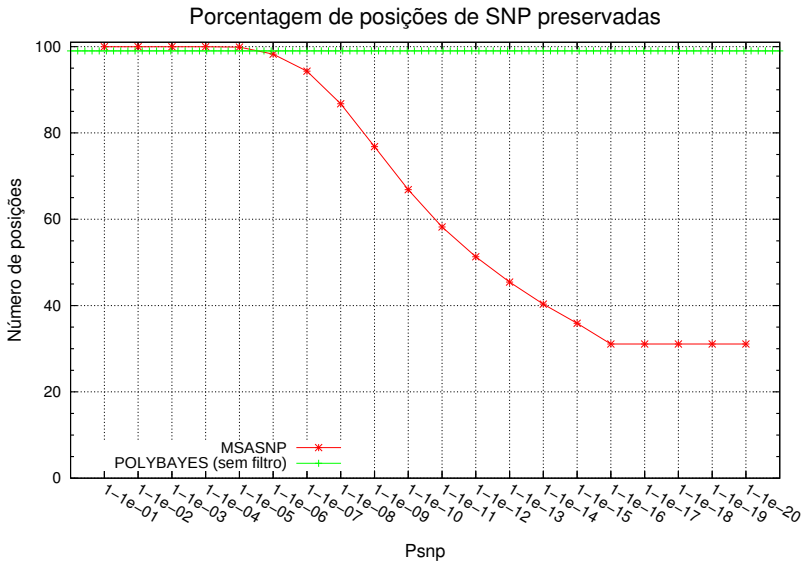


Figura 2: Gráfico comparativo no número de SNPs preservados, tomando como referência os dados do SUCEST. No eixo X temos um limite inferior, para a probabilidade de ser SNP, onde consideramos que seja um SNP. No eixo Y temos o número de posições preservadas. A curva verde refere-se ao `polybayes` (sem filtro) e a curva vermelha ao método MSASNP.

a SNP existente no alinhamento. Ao encontrar esta posição, a janela a indica como SNP e pula 5 posições, ignorando qualquer candidato a SNP existente neste intervalo. Este procedimento, portanto, não permite que exista um SNP distante do outro a menos de 5 posições.

Os gráficos das Figuras 3 e 4 são equivalentes aos das Figuras 1 e 2 só que agora utilizando a janela deslizante. Como podemos ver, o número de posições indicadas como SNP pelo método MSASNP caiu bastante. Contudo, a porcentagem de posições apontadas pelo SUCEST e pelo `polybayes` também caíram. Isso indica que este filtro não é capaz de eliminar falsos positivos sem afetar os verdadeiros positivos.

Além disso, analisando o `polybayes`, verificamos que a intersecção de suas posições de SNP com as do SUCEST é de 41138, ou seja, $98,99\%$ das posições bialélicas.

Para isso, o `polybayes` produziu 131622 posições, ou seja $3,17$ vezes mais do que o apontado pelo SUCEST.

Por outro lado, o método MSASNP usando probabilidade mínima de $1 - 10^{-6}$ produziu uma intersecção de posições de SNP com as do SUCEST é de 40828, ou seja, $98,27\%$ das posições bialélicas. Porém produziu 567618 posições, ou seja $13,66$ vezes mais do que o apontado pelo SUCEST.

Se utilizarmos o filtro de janelas nos dados do `polybayes`, o número de posições de SNP cai para 112247 ($2,70$ vezes mais que o conjunto SUCEST). A intersecção entre estes dois conjuntos foi de 39034 posições, ou seja, $93,96\%$ do total.

Usando o filtro de janelas no método MSASNP com probabilidade mínima de $1 - 10^{-6}$ temos 120089 posições de SNP ($2,89$ vezes mais que o conjunto SUCEST) e uma intersecção

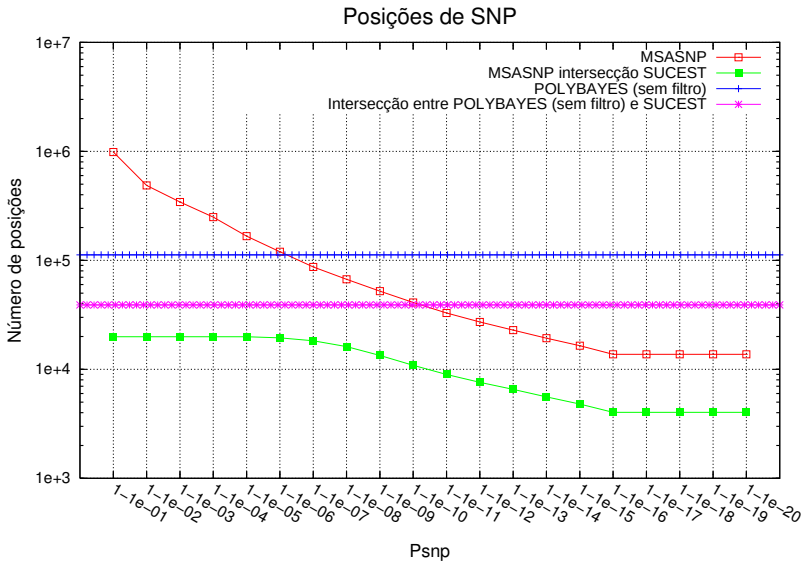


Figura 3: Gráfico comparativo no número de posições marcadas como SNP quando utilizamos uma janela deslizante de cinco posições entre dois SNPs. No eixo X temos um limite inferior, para a probabilidade de ser SNP, onde consideramos que seja um SNP. No eixo Y temos o número de posições marcadas como sendo SNP. A curva vermelha refere-se ao método MSASNP. A curva azul apresenta o número de SNP apontados pelo `polybayes` (sem filtro). A curva magenta apresenta o número de SNPs que aparecem no SUCEST (dados de referência) e `polybayes` ao mesmo tempo. A curva verde aponta o número de SNPs que aparecem no SUCEST e método MSASNP ao mesmo tempo.

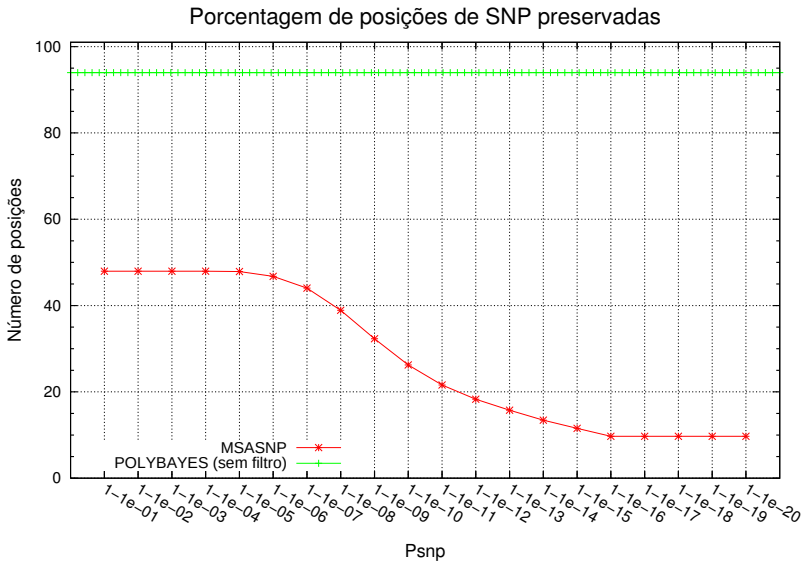


Figura 4: Gráfico comparativo no número de SNPs preservados, tomando como referência os dados do SUCEST, quando utilizamos uma janela deslizante de cinco posições entre dois SNPs. No eixo X temos um limite inferior, para a probabilidade de ser SNP, onde consideramos que seja um SNP. No eixo Y temos o número de posições preservadas. A curva verde refere-se ao `polybayes` (sem filtro) e a curva vermelha ao método MSASNP.

de 19437 posições. Apesar de atingirmos um número de posições próximo do obtido pelo `polybayes`, o número de verdadeiros positivos caiu para $46,79\%$ dos SNPs apontados pelo SUCEST.

5 Conclusão

A análise dos gráficos produzidos nos indica que o método MSASNP não foi capaz de obter resultados satisfatórios por diversos motivos.

O primeiro motivo é a imensa quantidade de posições de SNP apontadas por ele. Por exemplo, enquanto a taxa de posições bialélicas anotadas no projeto SUCEST era de $5,07$ SNP/*cluster*, a taxa apresentada pelo método MSASNP era de $505,54$ SNP/*cluster*. Com esta taxa, o método era capaz de encontrar todos os polimorfismos anotados. Contudo, o número de falsos positivos apresentados torna a utilização do método inviável. Mesmo com a determinação de um valor mínimo de probabilidade para considerar a posição como um SNP, o número de falsos postivos era muito alto.

O segundo motivo está relacionado ao próprio valor de confiabilidade produzido pelo método. Devido à natureza do cálculo e dos valores de qualidades, os valores de probabilidade de uma posição ser um SNP eram sempre muito altos.

Finalmente, a utilização de um filtro de janela adicional, que impedia que dois SNPs estivessem a menos de 5 bases de distância um do outro, mostrou-se pouco eficiente pois, apesar

de ele reduzir o número de falsos positivos, ele também reduzia o número de verdadeiros positivos a menos de 50% do número efetivo de posições de SNP.

O método `polybayes` mostrou resultados melhores. Apesar de inicialmente ele apresentar uma taxa de posições bialélicas de $16,06$ SNP/*cluster* (execução sem filtro de parálogos), a taxa de verdadeiros positivos encontrados ficou em $98,99\%$.

A utilização do filtro de janela nos resultados do `polybayes` faz o número de posições de SNPs bialélicos cair de 131622 para 112247 posições ($13,69$ SNP/*cluster*). Isto diminui a taxa de acerto para $93,96\%$.

Este último resultado nos leva a concluir que o filtro de janela deve ser melhorado. Como ele não considera as probabilidades que as posições possuem de serem SNPs, ele acaba sacrificando SNPs verdadeiros para atender o critério de distanciamento entre polimorfismos consecutivos.

Concluímos, por fim, que entre os métodos estudados, o `polybayes` é o mais indicado para a detecção de SNPs com informação de confiabilidade. Contudo, vale observar que um filtro adicional deve ser desenvolvido para que o número de falsos positivos possa ser diminuído sem o aumento do número de falsos negativos nos resultados produzidos pelo programa.

Referências

[1] A. J. Brookes. The essence of SNPs. *Gene*, 234:177–186, 1999.

[2] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C. R. Lane, E. P. Lim, N. Kalyanaraman, and J. Nemesh. Characterization of single-nucleotide polymorphisms in coding regions od human genes. *Nature Genetics*, 22:231–238, 1999.

[3] D. N. Cooper, E. V. Ball, and M. Krawczak. The human gene mutation database. *Nucleic Acid Research*, 26:285–287, 1998.

[4] M. K. Halushka et al. Patterns of single-nucleotide polymorphisms in candidate genes regulating blood-pressure homeostasis. *Nature Genetics*, 22:239–247, 1999.

[5] L. Grivet, J.C. Glaszmann, and P. Arruda. Sequence polymorphism from EST data in sugarcane: a fine analysis of 6-phosphogluconate dehydrogenase genes. *Genetics and Molecular Biology*, 24(1–4):161–167, 2001.

[6] L. Grivet, J.C. Glaszmann, M. Vincentz, F. da Silva, and P.Arruda. ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. *Theoretical Applied Genetics*, 106:190–197, 2003.

[7] X. Huang and A. Madan. CAP3: a DNA sequence assembly program. *Genome Research*, 9:868–877, 1999.

[8] G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O Stitziel, L. Hillier, P-Y. Kwok, and W. R. Gish. A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, 23:452–456, December 1999.

[9] The Sugar Cane EST Genome Project, September 2002. http://sucest.lad.ic.unicamp.br.