

PROSPECÇÃO AUTOMÁTICA DE SNPs

WAGNER ARBEX*, MARCOS VINÍCIUS G. BARBOSA DA SILVA*, VÍTOR SANTOS COSTA†, LUÍS ALFREDO VIDAL DE CARVALHO‡

**Empresa Brasileira de Pesquisa Agropecuária
Rua Eugênio do Nascimento, 610
36038-330, Juiz de Fora, MG, Brasil*

†*Universidade do Porto
Rua Campo Alegre, 1021/1055 - Gab. 1.45
4169-007, Porto, Portugal*

‡*Universidade Federal do Rio de Janeiro
Centro de Tecnologia - Bloco H-319
21945-970, Rio de Janeiro, RJ, Brasil*

Emails: arbex@cnpgl.embrapa.br, marcos@cnpgl.embrapa.br, vsc@dcc.fc.up.pt, alfredo@cos.ufrj.br

Abstract— The great amount of data generated from genome sequencing projects has led to the development of new computing tools with high level of processing capacity. Many of these tools are software developed as pipelines, what enables to treat the genomic information in an automated fashion, from the obtainment to the storage of such information, passing through its identification, analysis and prospecting. In order to achieve good results in the investigation of single nucleotide polymorphisms (SNPs) inside a genetic sequence, some routines must be established and aggregated or integrated to these tools as a new section inside the pipeline of automatic information analysis. Also, such routines must be able to generate easily apprehensible outputs.

Keywords— Single nucleotide polymorphisms, Automated search of single nucleotide polymorphisms, Pipeline to single nucleotide polymorphisms identification

Resumo— O grande volume de dados gerados a partir dos projetos de seqüenciamento de genomas obrigou a criação de novas ferramentas de computação com alto poder de processamento. Muitas dessas ferramentas são *softwares* desenvolvidos na forma de *pipelines*, o que permite automatizar o processo de tratamento da informação genômica, partindo da obtenção de tal informação, passando por sua identificação, análise e prospecção e indo até seu armazenamento. Para a obtenção de bons resultados na investigação de *single nucleotide polymorphisms* (SNPs), ou seja, polimorfismos de base única, em uma seqüência genética, algumas rotinas devem, necessariamente, ser estabelecidas e agregadas ou integradas a essas ferramentas, como uma nova seção dentro do *pipeline* de análise automática das informações, sendo capazes, ainda, de gerar saídas de fácil compreensão.

Palavras-chave— Polimorfismo de base única, Busca automatizada de polimorfismos de base única, *Pipeline* para identificação de polimorfismos de base única.

1 Introdução

Quando o Department of Energy e o National Institutes of Health tomaram a iniciativa, em meados da década de 80, de investigar o genoma humano buscavam, respectivamente, avaliar os riscos da energia nuclear à saúde e compreender melhor os processos biológicos subjacentes a saúde e a doença (Watson, 2004). Hoje, passadas duas décadas, existem diversos “projetos genoma” que, entre outros efeitos, resultaram na criação e estabelecimento da bioinformática, que veio possibilitar o desenvolvimento de tecnologias para processamento e interpretação das grandes massas de dados genéticos.

O surgimento desse novo campo de pesquisa, cujo nome remete à idéia da interdisciplinaridade em amplo espectro, como a própria junção dos institutos de pesquisa em energia e saúde, há vinte anos, não trouxe consigo uma definição precisa de seu núcleo de investigação, mas pode ser visto como o uso de técnicas aplicadas de matemática,

ciência da computação e estatística para resolver problemas biológicos.

O propósito desse texto é apresentar o *script* polymorp.pl, uma ferramenta de bioinformática que automatiza a prospecção de polimorfismos de base única (*single nucleotide polymorphisms* - SNPs), chamados de “snips”, a partir de sua integração com o *pipeline* phredPhrap, um *script* de leitura e montagem de seqüências genômicas, utilizado no Projeto Genoma Humano (Roberts et al., 2001) e em vários projetos genoma de diversos organismos, que otimiza o funcionamento dos programas Phrep, phd2fasta, Cross_match e Phrap (Ewing et al., 1998) (Ewing and Green, 1998) (Gordon et al., 1998), permitindo a utilização desses em grandes volumes de dados.

Para o cumprimento dessa proposta, esse texto encontra-se organizado em cinco seções, incluindo a *Introdução*, que apresenta o objetivo proposto e as ferramentas de computação automatizadas e discutidas ao longo do artigo. Em seguida, a segunda seção, *Polimorfismos de base*

única, define o que são e mostra a importância da investigação de tais polimorfismo. A terceira seção, *Computação de dados genéticos*, introduz como é a computação automática das grandes massas de dados genéticos e genômicos, sendo esse tópico também abordado na quarta seção, *O polymorp.pl*, onde são mostradas a integração do programa em questão com o *script* phredPhrap, assim como seu funcionamento, a partir dos resultados que gera. Por fim, a quinta seção, *Conclusão*, traz a análise do *polymorp.pl* e a sugestão de trabalhos futuros.

2 Polimorfismos de base única

Os projetos de seqüenciamento de genoma trouxeram muitas revelações para ciência, uma delas foi a descoberta de que o código genético humano mostrou-se mais variado e complexo do que propriamente maior, quando comparado ao de outras espécies¹.

Uma das muitas complexidades, variações e particularidades do genoma, humano ou de qualquer espécie, são os SNPs: modificações de um único nucleotídeo², em uma dada seqüência, quando comparada a outra, que podem alterar a formação das proteínas e, possivelmente, o conjunto dessas alterações pode provocar variações de características individuais.

O que difere um indivíduo dos demais da sua espécie é o código genético, ou seja, em sua essência, as seqüências de nucleotídeos que formam as moléculas e seqüências de DNA e RNA, que são traduzidas em proteínas, que, por sua vez, interagem e formam as células, que também, por sua vez, interagem e formam os tecidos, os órgãos, até que, finalmente, formam os indivíduos.

Dessa forma, se cada ser vivo fosse um livro, as diferenças entre os indivíduos de uma espécie começariam nas letras, mais especificamente, na ordem em que as letras formam as palavras. Ou seja, no código genético, as diferenças se iniciam na ordem em que os nucleotídeos se apresentam para, posteriormente, originarem as proteínas. Essa implicação foi ilustrada por Dani (2005) quando afirma que “como em uma espécie de baralho genético, a combinação das cartas, mais do que simplesmente seu valor individual, é que define o resultado do jogo”. Em outras palavras, essa é a importância dos SNPs, pois, em síntese, a alteração de um único nucleotídeo, uma única base, em uma dada seqüência, pode alterar a produção de uma certa proteína e, se for o caso, o conjunto

¹A investigação do genoma humano mostrou que esse possui em torno de 20 a 25 mil genes (DOE, 2004) (NHGRI, 2008), pouco mais do que os cerca de 19 mil contabilizados no genoma do verme nematóide *C. elegans* (Watson, 2004).

²Nucleotídeos são comumente chamados de “bases” ou, ainda, “pares de bases”, se estiver sendo considerada a “fita dupla” da estrutura do DNA.

dessas alterações pode provocar variações nas características dos indivíduos da espécie.

Os SNPs podem ser bi, tri ou tetra-alélicas, ou seja, possuem duas, três ou quatro formas distintas (Figura 1), sendo que os SNPs bi-alélicos, são os mais comuns (Baudet et al., 2006).

```

...GGCAAACTCCAG... ..GGCATATCTCCAG... ..GGCATACTCCA...
...GGCAAACTCCAG... ..GGCAAACTCCAG... ..GGCAAACTCCA...
...GGCAAACTCCAG... ..GGCATATCTCCAG... ..GGCATACTCCA...
...GGCAGACTCCAG... ..GGCAGACTCCAG... ..GGCAGACTCCA...
...GGCAGACTCCAG... ..GGCAGACTCCAG... ..GGCAGACTCCA...

```

Figura 1: Exemplos hipotéticos de polimorfismos bi, tri e tetra-alélicos, respectivamente.

A investigação de SNPs busca meios para esclarecer as seguintes questões (Arbex, 2007):

- Como identificar um SNP em uma seqüência?
- Como comprovar se o nucleotídeo “trocado”, que caracteriza a seqüência como polimórfica, é realmente um caso de polimorfismo, já que “base diferente” pode ser um simples erro de identificação da própria base?
- O polimorfismo provocará alteração na seqüência de bases a ponto de influenciar ou de participar de algum processo capaz de alterar a conformação de uma proteína, formando uma “nova” proteína?
- Essa “nova” proteína, caso seja formada, quando combinada com as demais, provocará ou suprirá a manifestação de alguma característica específica no indivíduo?

O *polymorp.pl*, que será discutido na Seção 4, atua sobre primeira questão, a identificação de possíveis SNPs em uma dada seqüência, e, ainda, sobre o segundo questionamento, fornecendo medidas para auxiliar na verificação desses.

3 Computação da informação genética

3.1 Tratamento das seqüências genéticas

Prodocimi and Santos (2004) observam que o surgimento da bioinformática clássica ocorreu com o seqüenciamento de biomoléculas e “destas deve permanecer inseparável” e, sem dúvida, uma referência para a origem da bioinformática é o início dos projetos genoma, pelo grande volume de dados gerados por tais projetos e, portanto, pela necessidade de sistemas de computação com grande aporte computacional e de profissionais especializados para a manipulação dos mesmos.

Decorrente dessa necessidade, um dos exemplos mais notórios de utilização das ferramentas de bioinformática é a sua utilização no armazenamento de dados, exigido no tratamento de diversos tipos de dados da “era genômica”. Após o trabalho nos laboratórios e “bancadas” para obtenção dos

mesmos, fica a cargo da bioinformática os procedimentos para identificação, organização, armazenamento, recuperação, classificação e apresentação dos dados gerados, assim como a pesquisa, o estudo e a análise para obtenção das informações em geral.

Nesse contexto, a bioinformática e suas soluções de automação constituem o conjunto de ferramentas que permite o tratamento dos dados oriundos de seqüências genéticas, no que tange ao volume que apresentam.

Como exemplo, em agosto de 2005, o International Nucleotide Sequence Database Collaboration (INSDC) anunciou ter ultrapassado a marca de 100 gigabases armazenados (Mehnert and Cravedi, 2005), ou seja, 100.000.000.000 de pares de bases³. Após um ano, em agosto de 2006, de acordo com NCBI (2006), esse número era superior a 145,7 gigabases, um crescimento de aproximadamente 45% no volume de dados armazenados, em apenas doze meses.

Esses números mostram, em termos amplos, a magnitude do volume de “matéria prima” a ser tratada.

3.2 Processamento e geração automática das seqüências genéticas

A obtenção das seqüências genéticas, ou melhor, dos dados em que essas se transformam, inicia-se por meio dos sequenciadores automáticos de DNA (Watson, 2004), onde o DNA extraído da amostra é submetido a um processo eletroforese, gerando um eletroferograma ou cromatograma (Figura 2).

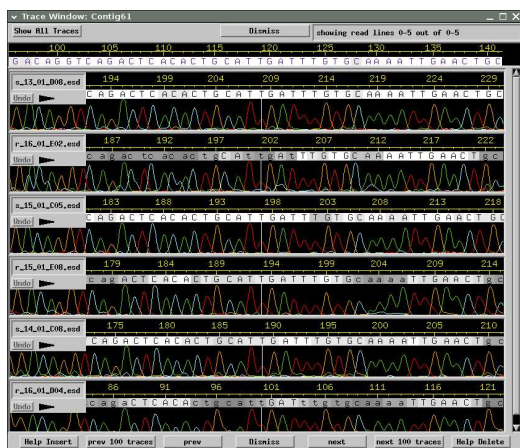


Figura 2: Cromatogramas visualizados por meio do programa Consed.

A partir desse ponto, diversos sistemas computacionais podem ser utilizados para que seja possível estabelecer a seqüência genética propriamente dita, isto é, a longa cadeia de nucleotídeos

³Considerando o número de pares de bases obtidas em projetos de genomas completos e em projetos que utilizaram o método *shotgun* (*whole genome shotgun* - WGS).

e, entre as possibilidades de processamento das seqüências, o presente texto apresenta o *pipeline* phredPhrap, que organiza os programas Phred, phd2fasta, Phrap, Cross_match, e discute a prospecção de SNPs por meio do polymorp.pl. A importância desse *pipeline* está em seus programas, visto a larga utilização, a robustez e a confiabilidade desses (Prosdocimi et al., 2002).

Dessa forma, as seqüências, originalmente cromatogramas, são tratadas nesse *pipeline* de rotinas (Figura 3) e, ainda, se necessário, as seqüências geradas (Figura 4) e os próprios cromatogramas podem ser visualizados e editados por meio do programa Consed.

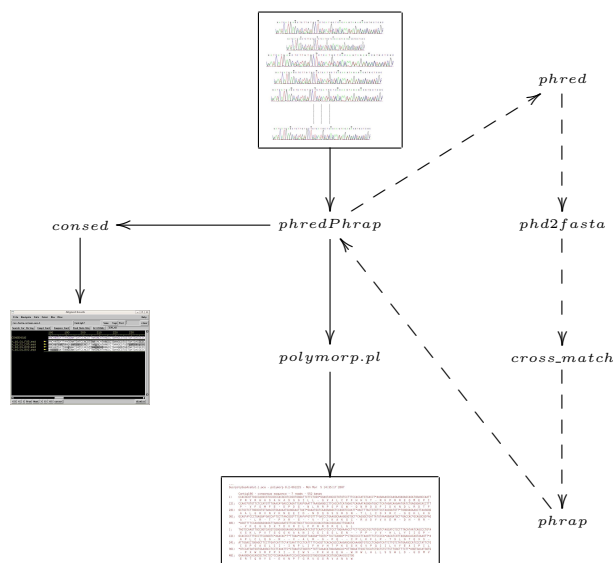


Figura 3: Pipeline para obtenção e montagem das seqüências, feito pelo *script* phredPhrap a partir dos cromatogramas, e a prospecção de SNPs, por meio do polymorp.pl, integrada ao *script*.

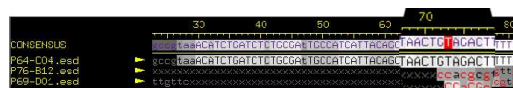


Figura 4: Exemplo de um consenso entre seqüências, originado a partir de 3 *reads* (seqüências), visualizado por meio do programa Consed.

Esses programas, e ainda outros voltados ao processamento de dados genéticos, foram desenvolvidos por pesquisadores da Universidade de Washington e possuem versões⁴ para diversas plataformas e licenças gratuitas para uso geral ou específicas para uso acadêmico. Suas descrições e outras informações podem ser encontradas em

⁴A documentação, bem como, instruções em geral, estão disponíveis em <http://www.phrap.org/>, <http://www.phrap.org/phredphrap/phred.html>, <http://www.phrap.org/phredphrap/phrap.html>, <http://www.phrap.com/>, <http://www.phrap.com/phred> e <http://www.phrap.com/consed>.

Ewing et al. (1998), Ewing and Green (1998) e Gordon et al. (1998).

Com o fim da execução do *pipeline*, são obtidos os consensos, gerados por alinhamentos relativos entre as seqüências, que também podem ser visualizados e editados por meio de *softwares* próprios para tal, como o Consed (Figura 4).

4 O polymorp.pl

4.1 Objetivos do polymorp.pl

Em sua versão atual, o polymorp.pl integra-se ao *pipeline* phredPhrap e implementa uma busca não-supervisionada, gerando saídas de fácil compreensão e um modelo simplificado de descrição de dados que apresentam:

- A similaridade entre os códons da seqüência quando comparada ao seu consenso, de forma visual;
- Uma estatística descritiva, com valores absoluto e percentual dos possíveis aminoácidos identificados nas posições correspondentes das seqüências agrupadas para a geração do consenso.

O polymorp.pl não pode afirmar se cada códon identificado determina um aminoácido, mas dado que as seqüências processadas são oriundas de seqüências já transcritas, e, portanto, seus códons já foram expressos como aminoácidos, então, o polymorp.pl assume tal possibilidade.

4.2 Leitura dos resultados

O relatório de saída do polymorp.pl é dividido em seis seções, que serão apresentadas a seguir:

1. Um cabeçalho (Figura 5) com a identificação dos dados analisados e do processamento dos mesmos.

```
Sus/polySusAceOut.1.ace - polymorp 0.2-061215 - Mon Mar 5 14:35:17 2007
```

Figura 5: Cabeçalho de identificação da saída.

O cabeçalho apresenta a identificação do arquivo analisado, que deve estar no formato “ace”, uma das saídas do phredPhrap, a data e a hora do processamento;

2. A identificação do consenso gerado pelo phredPhrap (Figura 6), o número de *reads* agrupados na formação do consenso e o número de bases consideradas.

```
Contig186 - consensus sequence - 7 reads - 552 bases
```

Figura 6: Dados do consenso.

O número de bases consideradas pelo polymorp.pl pode não coincidir com o número de bases do consenso gerado pelo phredPhrap, como será visto em breve;

3. As seqüências-consenso (Figura 7), propriamente ditas, com a tradução dos códons em possíveis aminoácidos, considerando apenas um *frame* na tradução, e a reversa complementar do consenso, com a mesma tradução.

```
1: CCAACGGTTGGCCAGCGGTCCGCCAAGCGGTCCGGTGGGATTCTCTCGG*GGAGTAGGG
  P R V R H A S A H A S G G I L L - G V A
121: CCAAC*GTGTTCTCCATGTTGAACA*GACCCAGATTCAG*GAATTAAGGAGCCTTCA
  P - V F S M F E - D P D S - N L R R P S
:
:
481: CAACTTGGACTA
  Q L G L
1: TAGTCCAAGTTGCCGGTCAAGTCCGGGGGGAAGCGAGGAAACATCTGTTCAATCTCTCC
  X S K L P V T S G G K A A N I C S I S S
121: GCACCCCTTTGCCCTCAGGGTC*AAACAC*T*TTGAA*CGCGTTAAGGA*TGTC*TCCT
  A P L C L Q G - N - - X - A L R - G - P
:
:
481: GCGAAGCGGTGG
  A N A W
```

Figura 7: O consenso gerado.

A interpretação de uma seqüência para a tradução dos códons em possíveis aminoácidos ocorre partir de “molduras”, que podem ter início na posição i , $i + 1$ ou $i + 2$ de uma dada seqüência. Contudo, a informação sobre a exata posição na qual a tradução foi iniciada pode ter sido perdida no processo de obtenção dos dados. Além disso, diversos fatores influenciam a ocorrência, ou não, da tradução.

Decorrente de tais condições, o polymorp.pl permite que o usuário informe qual a posição do início da “moldura” que será utilizada, possibilitando a análise dos dados, e a busca por SNPs, pelas três possíveis posições de início da tradução nos sentidos original e reverso da seqüência.

4. A identificação dos *reads* (Figura 8) que originaram a seqüência e o tamanho desses, em número de bases, considerado pelo polymorp.pl.

```
seq 1 - 537 bases
```

Figura 8: Dados do *read*.

O número de bases consideradas pode ser inferior ao número de bases informada pelo phredPhrap, pois o polymorp.pl considera como início da “janela” do consenso a posição em que todas seqüências, consenso e seus respectivos *reads*, possuem uma base válida.

A Figura 4 ilustra essa situação, onde o consenso gerado pelo phredPhrap, se inicia na

posição 27, porém o `polymorp.pl` considera a janela a partir da posição 73, pois é a primeira posição onde existe uma base lida no consenso e em todas as seqüências.

5. Os *reads* propriamente ditos (Figura 9), com a tradução dos códons em possíveis aminoácidos, considerando apenas um *frame* na tradução, e a reversa e complementar dos *reads*, com a mesma tradução. Além disso, visualmente, o `polymorp.pl` mostra a similaridade dos códons, codificada como:

- “...” - Códon coincidente com o códon do consenso na mesma posição;
- “1..” - Alteração de uma base, isto é, um candidato a SNP, na primeira posição do códon, quando comparado ao códon do consenso na mesma posição. Raciocínio análogo para as representações “.2.” e “.3.”, indicando alterações de uma base nas segunda e terceira posições do códon, respectivamente;
- “espaço” - O espaço em branco significa que existem diferenças entre a seqüência e o consenso em, no mínimo, duas bases contíguas, quando comparado ao códon do consenso na mesma posição.

```

.....3.....
1: CCACGCGTTCGCCACGCGTCCGCCACGCGTCCGGTGGGAT*CTTCTCGG*GGAAGAGCG
  P R V R H A S A H A S G G - L L - G V A
121: CC AAC*GTGTTCTCCATGTTTGAACA*GACCCAGATTTCAG*GAATTTAAGGAGGCCCTCA
  P - V F S M F E - D P D S - N L R R P S
:
:
...1.....
481: CCGGACGTGNN
  P G R -
:
:
. . . 1. . . 3 .1. . . . .2.
1: NNCACGTCGGGGGGAAGGCAGCGACGGCCGCGCAAGGCCGCGCTGGGAAACGG*GG*
  - T S G G K A A T G G P R P R L G N - -
:
:
. . . 1. . . . .3
121: CAGGGTC*AAGCAG*T*TTGAA*CGCGTTAAGGA*TGTC*TCCTCAGGG*T*CTGCCCC
  Q G - S - - X - A L R - G - P Q - - C P
:
:
.2. . .
481: GCGAACGCGTGG
  A N A W

```

Figura 9: Os reads com as indicações de códons coincidentes, ou não, em relação ao consenso.

O asterisco (*), que pode ser observado em alguns pontos da seqüência, indica *gaps* inseridos pelo `phredPhrap`, para permitir um melhor alinhamento da seqüência com as demais, quando buscava o consenso dessas, ou melhor, quando gerava a seqüência consenso desses *reads*

6. A última seção traz a tabela com os quantitativos e percentuais dos códons, considerando somente os *reads*, ou seja, o consenso não é considerado para essa estatística, visto que o mesmo é um resultado da análise dos *reads*. Essa estatística é feita para a seqüência no

sentido original e no sentido reverso com a seqüência complementar.

Na Figura 10, algumas informações foram suprimidas devido o espaço necessário para reproduzi-las, entretanto, é possível visualizar por completo as informações do último códon (número 184) e utilizá-las na interpretação do exemplo.

Esse códon apresentou uma variação de dois diferentes aminoácidos, que possivelmente poderiam ser originados, lisina (*K*) e leucina (*L*), sendo que o aminoácido de maior ocorrência, *L*, foi detectado três vezes, o que representa 75% das ocorrências.

1 -	GGA/G: 1 14%	GTC/V: 1 14%	AGT/S: 1 14%	TCC/S: 1 14%	...
2 -	GTC/V: 1 14%	ATT/I: 1 14%	GGT/G: 1 14%	CGG/R: 1 14%	...
3 -	GGA/G: 1 14%	TCA/S: 1 14%	GTT/V: 1 14%	CCC/P: 1 14%	...
:					
:					
---	184 -	AAA/K: 1 25%	CTA/L: 3 75%		

1 -	TGT/C: 1 14%	TTT/F: 1 14%	TAG/X: 3 43%	NODET: 2 29%	
2 -	TGC/C: 1 14%	AGT/S: 1 14%	CCT/P: 1 14%	ACG/T: 1 14%	...
3 -	CTT/L: 1 14%	TCC/S: 1 14%	CAC/H: 1 14%	CCC/P: 1 14%	...
:					
:					
---	184 -	GGA/G: 1 25%	ACT/T: 1 25%	GAC/D: 1 25%	TCC/S: 1 25%

Figura 10: Análise e quantificadores dos códons.

São relacionados os resultados do primeiro ao último códon, no caso, o códon 184, para o sentido original das seqüências e, em seguida, do primeiro ao último códon para as seqüências reversas complementares.

Caso algum códon não identifique um possível aminoácido, por exemplo, devido a um erro de leitura, o `polymorp.pl` considera o códon como não determinado (“NODET”), como pode ser visto na primeira seqüência reversa complementar na Figura 10.

Como as seqüências analisadas são originárias de cDNA, então, tais seqüências, sofreram transcrição reversa e, portanto, deve ser considerada a possibilidade de que nucleotídeos podem ter sido transcritos incorretamente, gerando falsos polimorfismos. Uma das formas de se identificar e eliminar tais erros é verificar a freqüência alélica mínima e as medidas descritivas dos códons, em números absoluto e percentual, especificadas nessa seção, auxiliam no cálculo da freqüência alélica mínima.

5 Conclusões

Qualquer número que faça referência à bioinformática parte de uma ordem de grandeza destacada. Por exemplo, os genomas possuem milhões ou bilhões de pares de bases, os genes são contados aos milhares e até mesmo as perspectivas de

retorno aos investimento na indústria movida pela bioinformática estão entre as mais significativas do mercado (RNCOS, 2006).

Além desses, a referência ao crescimento das bases de dados genômicos é um exemplo inequívoco da grandeza desses números. Em agosto de 2005, como já foi dito, as bases do INSDC registraram a marca histórica de 100 gigabases e, com um crescimento próximo de 90%, em dezembro de 2007, de acordo com NCBI (2007), essa mesma massa de dados registrou um número superior a 190 gigabases.

Assim, diante desses números, o tratamento de tais dados somente torna-se factível por meio das ferramentas de bioinformática, automatizando as rotinas e muitas dessas organizadas em *pipelines*. Além disso, segundo Kanehisa and Bork (2003), especificamente para a identificação de SNPs, é necessário o tratamento de dados em larga escala e com grande vazão.

O *polymorp.pl* tem as vantagens de integrar-se ao *phredPhrap*, de tratar os dados com a vazão necessária, ter sua execução em ordem polinomial e de fornecer uma saída de fácil entendimento, podendo ser acompanhada em tela⁵ ou lida e analisada posteriormente em um arquivo texto. Além disso, por ser um *script* desenvolvido em Perl, pode ser executado em qualquer ambiente de computação que possua o interpretador adequado, sem estar preso a nenhuma restrição de arquitetura de *hardware* ou de *software*.

Entre outros possíveis trabalhos futuros, e para as novas versões do *polymorp.pl*, sugere-se a implementação de modelos estatísticos mais sofisticados, permitindo uma visualização analítica mais detalhada das bases de dados analisadas.

Referências

- Arbex, W. (2007). *Modelos computacionais para identificação de informação genômica associada à resistência ao carrapato bovino*, Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia da Universidade Federal do Rio de Janeiro, Rio de Janeiro. Exame de qualificação.
- Baudet, C., Galves, M. and Dias, Z. (2006). Comparação de métodos para determinação de SNPs com medida de confiabilidade, *Relatório técnico*, Instituto de Computação da Universidade Estadual de Campinas, Campinas.
- Dani, S. U. (2005). Marcadores moleculares poderão aumentar a produtividade do rebanho nacional.
URL: www.excegen.com.br/artigos/artigo_4.php
- DOE (2004). How many genes are in the human genome? October 27, 2004.
URL: www.ornl.gov/sci/techresources/Human_Genome/faq/genenumber.shtml
- Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred (II): Error probabilities, *Genome Research* **8**: 186–194.
- Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998). Base-calling of automated sequencer traces using phred (I): Accuracy assessment, *Genome Research* **8**: 175–185.
- Gordon, D., Abajian, C. and Green, P. (1998). Consed: A graphical tool for sequence finishing, *Genome Research* **8**: 195–202.
- Kanehisa, M. and Bork, P. (2003). Bioinformatics in the post-sequence era, *Nature Genetics* **33**: 305–310.
- Mehnert, R. and Cravedi, K. (2005). Public collections of dna and rna sequence reach 100 gigabases. August 22, 2005.
- NCBI (2006). Genbank overview. September 26, 2006.
URL: www.ncbi.nlm.nih.gov/Genbank/index.html
- NCBI (2007). GenBank flat file: Release 163.0. December 15, 2007.
URL: <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>
- NHGRI (2008). An overview of the human genome project. May 9, 2008.
URL: www.genome.gov/12011238
- Prodocimi, F. and Santos, F. R. (2004). Sobre informática, genômica e ciência, *Ciência Hoje* **35**(209): 54–57.
- Prodocimi, F., Coutinho, G., Binneck, E., Silva, A., dos Reis, A., Junqueira, A. C., dos Santos, A. C., Júnior, A. N., Wust, C., Filho, F. C., Kessedjian, J., Petretski, J., Camargo, L. P., Mattos, R., Lima, R., Pereira, R., Jardim, S., Sampaio, V. and Folgueras-Flatschart, A. (2002). Bioinformática: manual do usuário, *Biociência, Ciência & Desenvolvimento* **5**(29): 12–25.
- RNCOS (2006). Bioinformatics market update, *Market research report*, Research & Consultancy Outsourcing Services. Product code: R459-779.
- Roberts, L., Davenport, R. J., Pennisi, E. and Marshall, E. (2001). A history of the Human Genome Project, *Science* **291**(5507): 1194.
- Watson, J. D. (2004). *DNA: the secret of life*, Alfred A. Knopf, New York.

⁵O acompanhamento em tela não é recomendado, pois, em geral, os arquivos de entrada são grandes massas de dados.