DEPARTAMENTO DE BIOQUÍMICA E IMUNOLOGIA
PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

TESE DE DOUTORADO

Racionalizando a utilização do algoritmo PHRED para a análise de seqüências de DNA

FRANCISCO PROSDOCIMI

Orientador: Prof. José Miguel Ortega
Co-orientador: Prof. Fabrício Rodrigues dos Santos

FRANCISCO PROSDOCIMI

"Racionalizando a utilização do algoritmo PHRED para a análise de seqüências de DNA"

Questionando dogmas genômicos

Tese apresentada ao Programa de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de Doutor em Bioinformática.

ÁREA DE CONCENTRAÇÃO: BIOINFORMÁTICA GENÔMICA

Orientador: Dr. José Miguel Ortega
Co-orientador: Dr. Fabrício Rodrigues dos Santos

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-graduação em Bioinformática
Departamento de Bioquímica e Imunologia
Belo Horizonte – MG
Novembro de 2006

"Não podem haver barreiras para a liberdade de questionamento. Não há lugar para dogma na ciência."
J Robert Oppenheimer

"A Maioria das pessoas preferiria morrer a pensar; de fato, muitas o fazem."
Bertrand Russell

"Amar e mudar as coisas me interessa mais"
Belchior, Alucinação

"Carpe diem"
Horácio, Odes 1.11

AGRADECIMENTOS

Gostaria de agradecer, primeiramente, ao Professor Fabrício Santos, orientador oficial da presente tese por, pelo menos, três anos. Apesar de ter percebido meu interesse por uma área ligeiramente diferente daquela que me dispus a realizar quando da minha entrada no doutorado, nunca deixou de me incentivar, de confiar em meu trabalho e jamais me forçou ou obrigou a tomar um rumo diferente daquele que escolhi para mim mesmo. Obrigado, Fabrício.

Agradeço muito aos meus colegas do doutorado e companheiros da vida e da academia, com quem troquei muitas idéias diariamente, seja a respeito de ciência ou amenidades em geral. Dos meus grandes amigos da universidade agradeço, principalmente, ao Chico Lobo, Ana Carolina Simões, Fabiano Comin, Cecile Fleury, Sávio Farias, Débora Garcia, Juliana Alves, Maurício Sant'Anna, Maurício Mudado, Mariana Bertelli, Adriano Barbosa, Fernanda Kehdy e Cristina Ribeiro.

Gostaria também de agradecer a todos os membros do LBEM, do LGB e do Laboratório de Biodados, os três laboratórios onde desenvolvi estudos durante esses últimos quatro anos e onde, em todos, sempre me senti em casa, como se cada um deles fosse o meu próprio local de trabalho. Meu muito obrigado vai também à professora Glória Franco, amiga e co-orientadora não oficial da presente tese.

Agradeço ainda ao Fabiano Peixoto, que deu o ponta-pé inicial em muitas das análises de PHRED apresentadas aqui e que me ensinou a ser um bom aprendiz de computeiro e a saber usar, com eficiência, os comandos *shell* do linux, o tosco editor de texto VI (cujo enorme pôster mostrando os comandos mais simples ficava em cima da minha mesa) e o awk. Vale notar que, caso o Fabiano tivesse resolvido entrar no doutorado em bioinformática, talvez esta tivesse sido a tese dele, não a minha.

Agradeço muito ao meu inteligente, tranqüilo e filosófico amigo, Jerônimo Conceição Ruiz, que me mostrou todas as manhas e artimanhas do processo científico inglês quando cheguei, perdido, para trabalhar no projeto genoma de *Schistosoma mansoni* no Sanger Centre. Sou muito grato, ainda, aos meus grandes amigos brasileiros doutorandos da Universidade de Cambridge e mais conhecidos como "Cambródis" que, em meio à paradeza nerd-cultural de Cambridge, sempre combinavam programas para afastar a melancolia dos dias ingleses, frios e cinzas. Dentre os cambródis, agradeço principalmente à Caroline Gasperin, Leda Sampson, Daniel Nelson, Juliano Yioda, Ronaldo Batista e Pedro Anselmo. Agradeço também aos meus *labmates* da unidade de sequenciamento de patógenos do Sanger e ao meu

orientador inglês Matthew Berriman, por ter tentado resolver com empenho todos os problemas burocráticos que tive durante minha estadia na Inglaterra.

Gostaria de agradecer também ao *Home Office*, a imigração inglesa, por ter me banido do país em meio à copa do mundo e ao meu orientador alemão, Klaus Brehm, por ter me permitido trabalhar nos fins de semana e viajar durante a semana para assistir os jogos da copa. Agradeço muito aos grandes amigos alemães que fiz na cidade de Würzburg, onde estive analisando ESTs de *Echinococcus* durante dois meses. Em especial agradeço ao Peter, Markus, Ali e Dirk pela amizade e por terem me recebido como quem recebe um Ronaldinho.

Agradeço muito, é claro, à minha família, por ter me dado todo o suporte, apoio e incentivo, em todos os momentos, sempre, e sem exceção. Aos meus pais, minhas irmãs, minha madrinha, meus primos, meus tios e à minha avó, que sempre tem alguma pergunta sobre as células-tronco e que não morreu, como pensou que fosse acontecer, quando se despediu de mim, chorando, antes de minha partida para o doutorado sanduíche. Já voltei, vó!

Vale aqui fazer um agradecimento mais filosófico à poesia e beleza que há no mundo, em todas as coisas; poesia esta que me encanta a todo instante mas que é por muitos ignorada, infelizmente. Vale também agradecer à magia e ao mistério que se esconde por trás da alma feminina e que está sempre a nos encantar, a nos ludibriar e também a nos fazer sofrer; mas quem gostaria de viver num mundo sem elas? E vale agradecer também à razão, deusa do cientista, à propulsora curiosidade da alma humana e ao desconhecido, nosso infinito (?) objeto de trabalho.

Todo bioinformata deve também, creio eu, agradecer aos técnicos e cientistas que trabalharam de forma a produzir e publicar as seqüências de DNA utilizadas para a realização de seus trabalhos: obrigado, portanto, a todas essas pessoas!

Por fim, gostaria de agradecer imensamente ao meu orientador, Professor J Miguel Ortega. Espero guardar sempre comigo sua visão da ciência como um processo altamente criativo e prazeroso, a despeito das pressões externas para se produzir conhecimento num ritmo desenfreado, tradicionalista e repetitivo. Agradeço ao Miguelito não só pelos momentos dentro, mas também fora do laboratório, no dia-a-dia e por transformar uma relação que poderia ser de estresse em verdadeira amizade, camaradagem e respeito. Agradeço ainda pela paciência que ele teve comigo quando eu passava um momento difícil na Inglaterra, onde continuou me incentivando a produzir conhecimento da forma correta, sendo que mesmo em meio à uma certa depressão ainda fui capaz de escrever um trabalho e publicá-lo. Valeu demais, Miguelito!

ÍNDICE

LISTA DE ARTIGOS

Número	Título	Autores	Status/ *Revista*	Pg
1	DNA Sequences Base Calling by PHRED: Error Pattern Analysis	**Prosdocimi, F** Peixoto, FC Ortega, JM	Publicado *RTInfo*	25
2	Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values	**Prosdocimi, F** Peixoto, FC Ortega, JM	Publicado *Gen Mol Res*	31
3	Accessing optimal primer distance from insert	**Prosdocimi, F** Ortega, JM	Publicado *In silico Biol*	42
4	Setting PHRED scores to obtain maximum biological information	**Prosdocimi, F** Peixoto, FC Ortega, JM	Submetido *Nucleic Acids Res*	55
5	Effects of sample re-sequencing and trimming on the quality and size of assembled consensus	**Prosdocimi, F** Lopes, DAO Peixoto, FC Ortega, JM	No prelo *Gen Mol Res*	72

* Outros artigos publicados e não relacionados diretamente ao tema da tese podem ser observados nas seções finais: *Produção científica durante o doutorado* e *Anexos*

LISTA DE TABELAS

LISTA DE FIGURAS

SIGLAS E ABREVIATURAS

Sigla/Abreviatura	Significado
ASP	Alignment Starting Position
BcSP	Base-calling Starting Position
BLAST	Basic Local Alignment Search Tool
cDNA	Complementar DNA
CENAPAD	Centro Nacional de Processamento de Alto Desempenho
EST	Expressed Sequence Tag
FAPEMIG	Fundação de Amparo à Pesquisa do Estado de Minas Gerais
GSS	Genome Survey Sequence
INDEL	Inserção e deleção
NCBI	National Center for Biotechnology Information
NHGRI	National Human Genome Research Institute
NMW	Non-masked windows
PCR	Polymerase chain reaction
PERL	Practical Extraction and Report Language
PHRED	Phil's Read Editor
PHRAP	PHRagment Assembly Program
PPW	Predicted Perfect Windows
PQV	PHRED Quality Value
PSP	Polymerization Starting Position
PW	Perfect Windows
RPW	Real Perfect Windows
SAGE	Serial Analysis of Gene Expression
SW	Spoiled Windows
SWAT	Smith-Waterman algoritmo
TIGR	The Institute for Genomic Research
TP	Trimming Position
UFMG	Universidade Federal de Minas Gerais
WC	Weighted correctness

RESUMO

A ciência é, por vezes, dogmática. Mesmo o cientista questionador às vezes é obrigado a tomar como verdade algo que se acredita na comunidade de forma a realizar suas pesquisas em busca do conhecimento. Na área da genômica, alguns dogmas estão ainda arraigados à cultura científica e o objetivo principal da presente tese foi tentar, na medida do possível, questionar e testar alguns desses dogmas com a intenção de trazer à luz da razão um conhecimento mais sólido sobre alguns limitados aspectos relacionados, principalmente, ao processo de nomeação das bases (*base-calling*). Para avaliar, portanto, a utilização do algoritmo PHRED, o principal nomeador de bases utilizado em projetos genoma, desenvolvemos primeiro uma metodologia sólida de análise. Tal metodologia tentou diminuir o número de variáveis a se analisar em uma corrida de seqüenciamento para que nossas análises não levassem em consideração peculiaridades específicas de uma ou outra reação produzida. Dessa forma, realizamos o seqüenciamento de um vetor de clonagem bastante conhecido (pUC18) em um único conjunto, homogeneizando as amostras de forma que a única variável possível fosse a separação eletroforética e o padrão de nomeação de bases. Produzimos, portanto, 846 seqüências do vetor pUC18 que foram comparadas, através de alinhamentos locais, com um controle positivo: a seqüência já publicada para esta molécula. Dessa forma, pudemos identificar os erros do seqüenciamento / nomeação de bases e avaliar diferentes parâmetros de utilização do algoritmo PHRED. Verificamos que o padrão de erros observado era relativamente igual ao esperado, que as bases incorretas não podiam ser previstas através da observação dos valores de qualidade de sua vizinhança e que as trocas (*mismatches*) são mais comuns quando associadas a valores baixos de qualidade, enquanto se nota a presença de erros relacionados a *indels* de alta qualidade. Percebemos também uma aplicação desta abordagem para o processo de desenho de iniciadores de seqüenciamento e realizamos um estudo avaliando este tópico, o qual mostra que a leitura de boa qualidade é iniciada a uma distância mensurável à jusante do iniciador de seqüenciamento. Com o objetivo de tentar mascarar as bases incorretas em letras minúsculas, observamos que o valor de qualidade 7 parece ser o mais adequado para utilizar nesses casos, em boa parte das situações. Além disso, calibramos o programa PHRED para funcionar de forma a apresentar apenas a informação não-ruidosa, biologicamente relevante. Por último, analisamos ainda a formação de consensos a partir dessas seqüências e mostramos a surpreendente ineficiência do re-seqüenciamento de forma a produzir seqüências fiéis à molécula molde.

ABSTRACT

Science is sometimes dogmatic. Even the very thinker scientists are sometimes forced to accept as true something believed by the community in order to advance their research. In the genomic research field, some dogmas are still attached to scientific culture and the main goal of this thesis is the tentative to question some of these dogmas and bring to the light of reason a consistent knowledge about some restrict aspects related to the base-calling process. Therefore, in order to evaluate the execution of PHRED, the main base-caller used in genome projects, we first develop a consistent methodology of analysis. Using this methodology we have tried to reduce the number of variables to be analyzed in sequencing reads, making our analysis free of particularities happening in some specific sequencing reaction. With this in mind, we have performed the sequencing of a well-known cloning vector (pUC18) in a single-pool, homogenizing the samples before and after the sequencing reaction. So, 846 sequences from the pUC18 cloning vector were produced by single-pool and compared, through local alignments, with a positive control: the sequence published for this molecule. This comparison allowed us both to identify precisely the errors happening in the sequencing and/or base-calling and to evaluate different parameters used for PHRED running. We have verified (1) an error pattern very similar to the expected one, (2) the impossibility to predict errors evaluating the base quality values surrounding the neighborhood of miscalled bases, (3) the high presence of mismatches in low quality values and (4) the presence of some indels in high quality regions. We have realized also an application of these base-calling data to the process of designing primers for sequencing and one study was published on this subject. Trying to softmask low quality bases, we have made another study to find the best PHRED quality value to be used to mask most of the errors without masking correct bases. Moreover, we have studied and adjusted PHRED trimming parameters in order to retrieve from the sequence just the biologically relevant information. At last, we have analyzed the consensus production through different number of sequencing reads in order to find the appropriate number of sample re-sequencing to generate a high-fidelity molecule.

1. INTRODUÇÃO

1.1. Sequenciamento de moléculas de DNA e o processo de nomeação de bases

É bem aceito na área da biologia molecular que as seqüências de DNA são produzidas pelas máquinas seqüenciadoras. Mas será esta uma verdade?

Nesta tese trataremos sempre da forma mais comum de sequenciamento de moléculas de DNA utilizada hoje, realizada através de desenvolvimentos do método de Sanger (Sanger and Coulson, 1975). Para a realização desta técnica, utilizam-se didesoxinucleotídeos marcados com moléculas fluorescentes de forma a interromper a síntese durante a reação de polimerização que caracteriza o sequenciamento e permitir a posterior identificação do último nucleotídeo adicionado, unidade interruptora da polimerização. Lembramos aqui que esta reação de "polimerização interrompida" é realizada em placas submetidas a termocicladores através da catálise enzimática por DNA polimerases termoresistentes. Ou seja, ela independe dos sequenciadores automáticos.

Estaremos ainda tratando, salvo quando explicitado, sobre procedimentos de eletroforese capilar, onde as amostras de DNA de diferentes tamanhos, resultantes da reação de sequenciamento, são submetidas a um campo elétrico, dentro de uma matriz capilar. E é justamente essa eletroforese capilar das moléculas, associada à leitura da fluorescência que é emitida a cada instante, a única função das máquinas de sequenciamento de DNA. Portanto, a função do seqüenciador de DNA é apenas realizar a eletroforese capilar e identificar quais as fluorescências foram captadas a cada instante, ao longo do procedimento da eletroforese.

Os dados brutos obtidos a partir dos sinais identificados pelo laser do seqüenciador são então utilizados como entrada para programas de bioinformática conhecidos como nomeadores de bases (*base-callers*). Esses algoritmos serão então responsáveis por transformar esse dado bruto numa seqüência de nucleotídeos que represente, o mais fielmente possível, a molécula de entrada. Além disso, realiza-se também a associação de um valor de qualidade à cada base predita e este valor representa a chance estatística da base (A, C, G ou T) ter sido nomeada corretamente. Muitas máquinas sequenciadoras de DNA vêm com programas próprios para a nomeação de bases. Além disso, vários outros algoritmos de nomeação de bases são conhecidos baseados em diferentes métodos, como análises de Fourier (Ewing and Green, 1998), máxima verossimilhança (Brady et al., 2000), detecção prioritária de

picos (Walther et al., 2001) ou apenas através da detecção de marcação multi-cor (He and McGown, 2001; Song and Yeung, 2000; Giddings et al., 1993).

Vimos, portanto, que a produção de moléculas de DNA está atrelada a três fatores principais: (1) a realização da reação de sequenciamento; (2) a eletroforese capilar e (3) a nomeação das bases, donde se conclui que as máquinas sequenciadoras de DNA atuam apenas em parte do processo de obtenção da seqüência de uma molécula desejada.

FIGURA 1. **Etapas para o sequenciamento de moléculas de DNA**. As moléculas são polimerizadas durante a reação de sequenciamento numa máquina termocicladora, a eletroforese capilar e leitura dos padrões de fluorescência é realizada pelo aparelho seqüenciador e o dado bruto é finalmente lido pelo nomeador de bases (PHRED) de forma a gerar a seqüência de bases e qualidades associadas às bases.

Como já comentado, outra das mais importantes funções de um algoritmo de nomeação de bases, consiste em ser capaz de identificar, com precisão, a probabilidade de determinada base nomeada estar correta. Alguns trabalhos clássicos sobre estes programas foram capazes de definir medidas de confidência para as bases nomeadas, entretanto não apresentaram relatórios atestando sua validade ou seu poder discriminatório (Berno, 1996; Giddings et al., 1993). O primeiro trabalho que realmente mostrava a validade de um sistema de qualidade, baseou-se em análises de discriminantes para diferenciar as bases nomeadas correta e incorretamente, definindo uma probabilidade de erro associada a cada uma das bases nomeadas (Lawrence and Solovyev, 1994). Entretanto, desde que Ewing and Green (1998) definiram seu método que utilizava a probabilidade de erro logarítmica, como realizado pelo algoritmo PHRED, esta acabou se tornando a métrica padrão para se analisar a qualidade das bases nomeadas.

A seguir discutiremos com mais detalhes o algoritmo PHRED.

1.2. O algoritmo PHRED

Ainda que existam diferentes algoritmos e técnicas para a realização da nomeação das bases, a comunidade científica adotou como padrão quase unânime, a utilização do algoritmo PHRED (Ewing and Green, 1998; Ewing et al., 1998) para a realização deste procedimento. Escolhemos, portanto, analisar o comportamento deste algoritmo em relação a diversos aspectos.

Segundo seus autores (Ewing and Green, 1998), o PHRED funciona basicamente em quatro etapas distintas, a saber: (1) *Lane tracking*, onde as extremidades do dado bruto são identificadas; (2) *Lane profiling*, onde o padrão de cada um dos quatro sinais de fluorescência são somados através da extensão do dado eletroforético, com o objetivo de se definir as intensidades dos sinais através de milhares de pontos uniformemente espaçados ao longo da corrida – nesta etapa é produzido o eletroferograma (também vulgarmente chamado de cromatograma); (3) *Trace processing*, onde métodos de processamento de sinais são utilizados para suavizar as estimativas do dado informacional, diminuir o ruído e corrigir possíveis efeitos causados pela diferente mobilidade eletroforética dos marcadores fluorescentes; (4) *Base-calling*, onde o eletroferograma processado é traduzido em uma seqüência de bases com qualidades associadas.

O PHRED foi testado e analisado pelos autores (segundo seu manual em http://www.phrap.org/phredphrap/phred.html) para as seguintes máquinas de sequenciamento de DNA: ABI, modelos 373, 377 e 3700; Molecular Dynamics MegaBACE e LI-COR 4000.

Alguns erros conhecidos durante o sequenciamento estão relacionados aos seguintes fatores: (1) mobilidade de fragmentos pequenos: parece que o PHRED não é capaz de separar muito bem os sinais gerados pelas moléculas menores do que 50 nucleotídeos nem de retirar o ruído dessas moléculas; (2) já em moléculas maiores, por vezes acontece a formação de uma estrutura em grampo (*hairpin*) no fim da molécula, fazendo com que ela migre mais rapidamente do que seria esperado pelo seu tamanho e, portanto, fazendo com que uma base seja lida, incorretamente, antes do momento apropriado e esteja desviada para a esquerda na leitura. Esse problema já havia sido verificado por Sanger (Sanger and Coulson, 1975; Sanger et al., 1977) e não é uma falha na nomeação de bases, mas um erro intrínseco deste método de sequenciamento; (3) sinal fraco ou alto ruído: freqüentemente produzido devido a

problemas na reação de sequenciamento, efeitos relacionados ao contexto de seqüência ou incorporação ineficiente do didesoxinucleotídeo; (4) péssima qualidade de seqüência depois de regiões repetitivas de mono ou dinucleotídeos, onde pode ter havido um escorregamento (*slippage*) da polimerase e onde o fator do peso diferencial de cada fluorocromo pode não ser efetivamente bem calculado e atrapalhar a separação de cada pico.

Chamaremos de PHRED Quality Value (PQV) o valor de qualidade associado a cada uma das bases nomeadas num arquivo de qualidade (.qual) e que representa a chance da mesma ter sido incorretamente nomeada. A figura 1 mostra um arquivo com a seqüência no formato FASTA de um *read* utilizado neste trabalho com sua respectiva seqüência de qualidade .QUAL como exemplo.

```
>G01.esd CHROMAT_FILE: G01.esd TIME: Thu Oct 30 17:13:15 2003
TACGAGCTCGAATTCGTAATCATGTCATAGCTGTTTCCTGTGTGAAATTG
TTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTA
AAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGC

>G01.esd CHROMAT_FILE: G01.esd
29 32 34 32 33 33 40 34 32 32 32 32 28 25 29 27 27
32 32 32 39 31 31 35 35 35 40 40 40 39 29 29 29 27
32 48 34 25 26 29 29 29 34 29 32 27 17 9 12 15 19 21
40 40 48 48 46 40 40 39 39 35 40 40 40 56 56 56 56
56 56 56 56 42 42 56 56 46 46 46 46 46 46 46 46 40
40 40 46 46 46 46 46 46 46 40 40 40 46 40 40 40 46
46 46 46 46 56 56 56 46 46 46 46 40 40 47 56 47 56
47 56 47 40 37 39 37 37 37 37 46 46 46 56 51 51 51
46 46 42 42 42 46 46 42 56 56 56 46 51 51
```

FIGURA 2. **Arquivos FASTA e QUAL nomeados pelo PHRED.** Arquivos representando parte de uma seqüência de pUC18 trimada e nomeada pelo PHRED e parte do arquivo de qualidade da região correspondente, mostrando o PQV associado a cada uma das bases.

O valor de qualidade de PHRED é medido através da seguinte fórmula:

$$PQV = -10 \log_{10}(p),$$

sendo *p* a probabilidade de ocorrência do erro.

Para exemplificar, um PQV de 10 para uma base significa que ela terá 10% de chance de estar incorreta; um valor de 20 dará a ela uma chance em 100 de estar incorreta e um valor de PQV igual a 30 estará associado a uma chance em 1.000 daquela base ter sido incorretamente nomeada.

1.3. A produção de seqüências de DNA em projetos genoma e transcriptoma

Depois de geradas, as bibliotecas de DNA ou cDNA (DNA complementar), a produção de seqüências para a realização de projetos genoma ou transcriptoma baseia-se nos mesmos procedimentos: (1) um iniciador é utilizado para amplificar as seqüências, (2) didesoxinucleotídeos marcados são usados para interromper a síntese em cada uma das posições nucleotídicas, (3) a eletroforese dos fragmentos é realizada, (4) o laser do seqüenciador lê as fluorescências das moléculas migrando no capilar e (5) o processo de nomeação de bases retorna ao usuário a cadeia de As, Cs, Ts e Gs que representam a molécula de interesse. Duas diferenças básicas, entretanto, existem entre esses projetos: a forma de geração da biblioteca e a interpretação dos dados gerados.

A biblioteca de DNA construída para a realização de projetos genoma é montada através da fragmentação do DNA genômico inteiro do organismo de interesse. A fragmentação do DNA pode ser realizada através de métodos físicos – como a sonicação, que normalmente produz segmentos de DNA com extremidades cegas – ou através de cortes por enzimas de restrição, que deixam extremidades coesivas prontas para a clonagem em vetores específicos, mas que podem gerar uma distribuição não aleatória dos tamanhos dos fragmentos cortados devido à ausência do sítio da enzima em determinados pontos da seqüência. Depois do corte, seleciona-se o tamanho adequado do DNA fragmentado para a clonagem e, então, realiza-se a ligação das moléculas de um tamanho apropriado com o vetor de clonagem escolhido. A figura 3 mostra um exemplo genérico para a montagem de uma biblioteca de DNA genômico.

Já a geração de bibliotecas de cDNA para a produção de etiquetas gênicas (ESTs, *Expressed Sequence Tags*, figura 4) é realizada através da produção do DNA complementar a partir do mRNA extraído de uma célula de um organismo sujeita a alguma condição espaço-temporal específica (Adams et al., 1991). A seqüência dos mRNAs extremamente instáveis deve ser transformada em cDNA através da utilização da enzima transcriptase reversa, atuando a partir da seqüência de um iniciador de oligo-dT que se liga à cauda de poli-A presente na grande maioria dos RNAs mensageiros eucarióticos. Então é gerada uma primeira fita híbrida de DNA e RNA e então o RNA original é degradado através da utilização de uma enzima ribonuclease. Alguns pedaços de RNA que ainda permanecem no híbrido são então utilizados como iniciadores para a síntese da segunda fita, feita pela enzima DNA polimerase. Essas moléculas são então inseridas em vetores de clonagem normalmente através da ligação de seqüências adaptadoras contendo sítios de enzimas de restrição em suas

extremidades. Assim, tanto o vetor de clonagem quanto a molécula de cDNA são digeridos pela mesma enzima de restrição, deixando extremidades coesivas que serão posteriormente ligadas através da utilização de uma enzima DNA ligase.

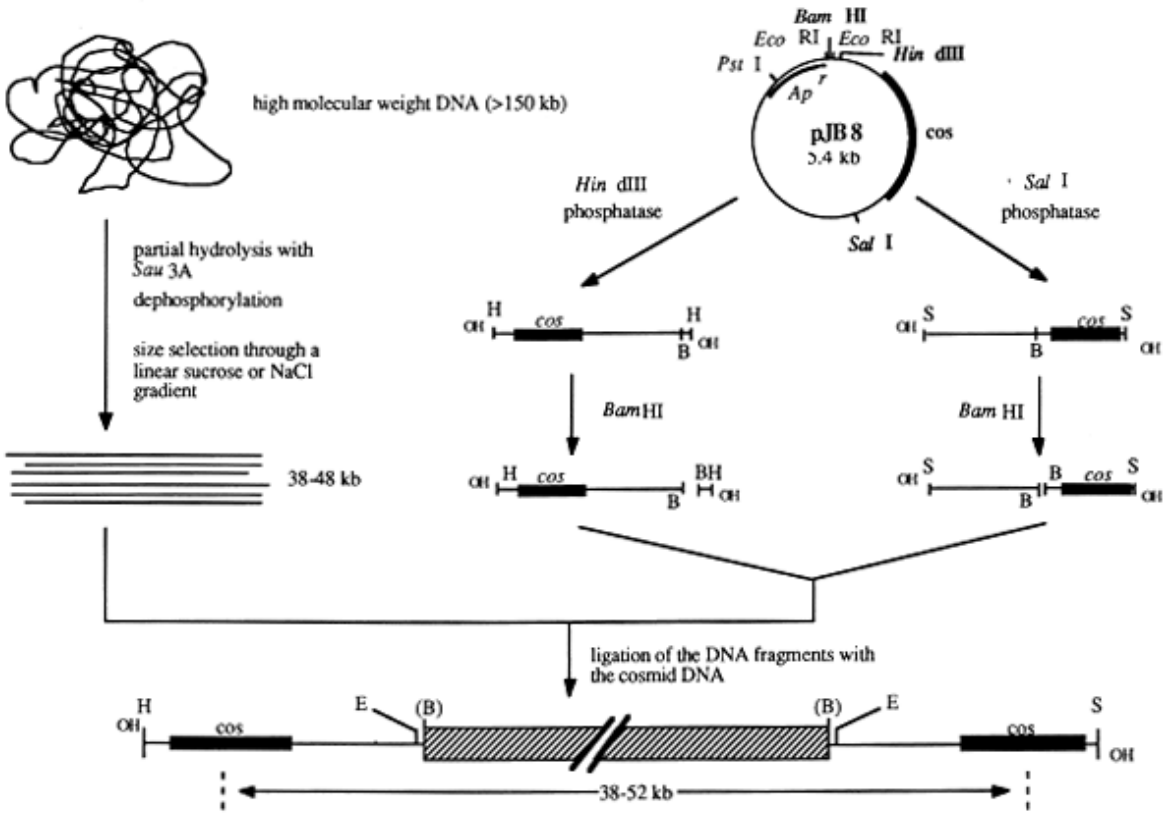


FIGURA 3: **Construção de uma biblioteca de DNA.** O DNA é partido em pedaços por enzimas de restrição e o tamanho do inserto é selecionado. As moléculas recombinantes que constituirão os clones da biblioteca são formadas, na parte inferior da figura, pela ligação do DNA genômico com o vetor de clonagem – no caso, um cosmídeo. (Figura obtida em http://www.fao.org/DOCREP/004/T0094E/T0094E03.htm)

Como já comentado, o processo de sequenciamento das moléculas oriundas de uma biblioteca de DNA genômico ou de cDNA é idêntico. O técnico que sequencia uma ou outra molécula, por exemplo, não precisa sequer ficar sabendo o que está sequenciando, pois os procedimentos são exatamente iguais. Já a interpretação e o processamento dos dados é completamente diferente.

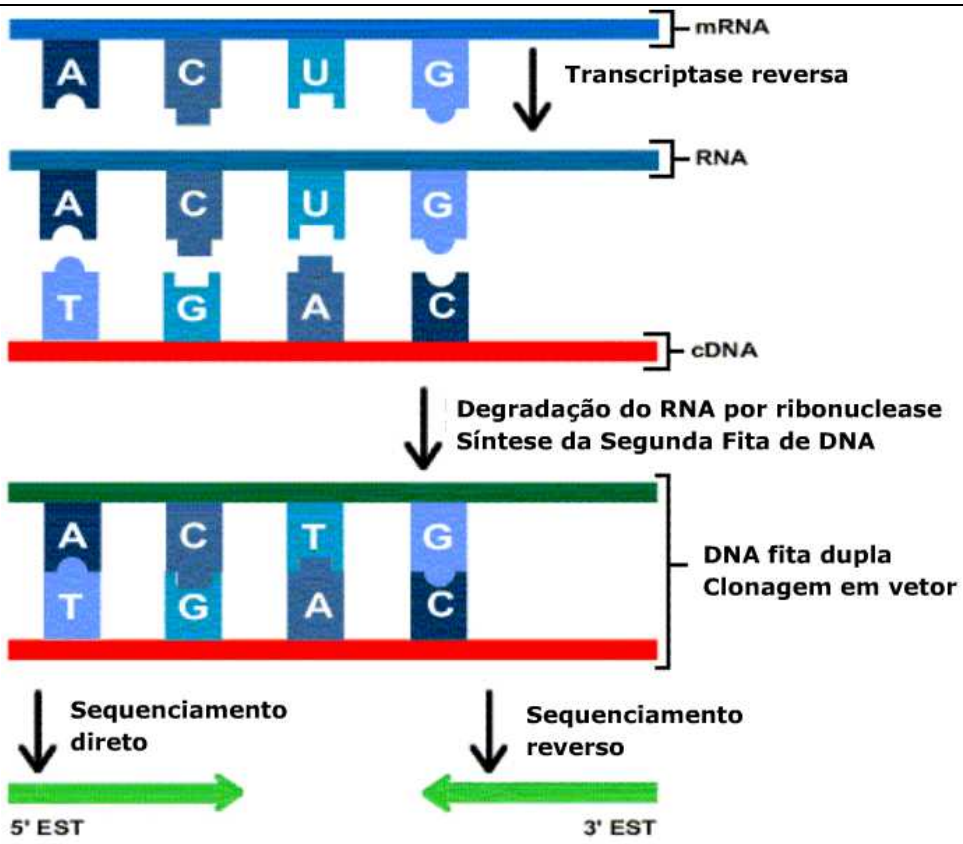


FIGURA 4: **Produção de ESTs em projetos transcriptoma.** As etiquetas de seqüências expressas são obtidas, primeiramente, a partir da transcrição reversa do mRNA, produzindo uma fita de DNA complementar (cDNA). Após esse procedimento, utiliza-se uma RNAse H para digerir a seqüência de RNA inicial e é produzida a segunda fita de DNA, gerando a molécula de DNA de fita dupla. Essa molécula é normalmente clonada direcionalmente em vetores de clonagem e são utilizados iniciadores para o sequenciamento das extremidades 5' ou 3' do cDNA em apenas uma "rodada" de sequenciamento. As seqüências obtidas são as chamadas ESTs.

Quando realizamos projetos genoma, temos o interesse de produzir o maior número de seqüências possível, com o maior tamanho possível, de preferência gerando uma seqüência que contenha apenas dados do inserto, sem gerarmos uma só base relacionada ao vetor de clonagem. Isso vem do fato de que, quanto mais seqüências tivermos do nosso organismo de interesse, mais chance teremos de completar aquele genoma, obtendo todas as informações presentes do DNA de tal organismo. Por isso, podemos selecionar seqüências de tamanho maior para serem clonadas nos vetores, já que não necessariamente desejamos que o sequenciamento

dessas, a partir das duas extremidades do vetor, se sobreponha de forma a produzir um consenso. Na verdade esperamos que, com o sequenciamento de um número muito grande de seqüências, os chamados "gaps virtuais", ou seja, as regiões observadas entre os *reads* produzidos a partir de cada uma das extremidades do mesmo clone, sejam fechados por seqüências oriundas de outros clones.

Já em projetos de sequenciamento de bibliotecas de cDNA, muitas vezes temos o interesse em obter a seqüência completa dos RNAs mensageiros que codificam os genes daquele organismo. Portanto, em projetos de produção de etiquetas gênicas normalmente seleciona-se um tamanho de inserto aproximado de 2Kb, de forma que, quando o seqüenciarmos a partir de uma e outra extremidade, possamos obter ao menos uma pequena região de sobreposição dessas seqüências que permitirão, a um programa de agrupamento, reuni-las em uma só molécula virtual representando uma grande parcela, ou a totalidade, de cada um dos genes. Além disso, quando da realização desses projetos de produção de seqüências gênicas, normalmente existe o interesse em obter, no início da seqüência gerada, uma parte de DNA do vetor de clonagem ou do adaptador. Esse sequenciamento extra de seqüências não informativas serve para mostrar ao pesquisador que o mRNA original provavelmente começava naquela posição e, assim, muitas vezes torna-se possível encontrar o início da seqüência codificadora relacionada ao gene em questão e selecionar aquele clone em especial para um sequenciamento completo (*full-lenght sequencing*) (Ota et al., 2004; Strausberg et al., 2002; Strausberg et al., 1999). Com relação ao outro lado da seqüência, a observação da cauda de poli-A também é importante para caracterizar com fidelidade aquele RNA e atestar o término daquela seqüência transcrita. Alguns serviços de agrupamentos de seqüências gênicas geradas por projetos de sequenciamento de cDNAs consideram a presença da cauda de poli-A como uma evidência importante da expressão daquele gene e as entradas Unigene, por exemplo, exigem que pelo menos um membro de cada agrupamento contenha a cauda poli-A (Pontius et al., 2003; Schuler et al., 1996).

1.4. Agrupamento (*clustering*) de seqüências de DNA

Outra das técnicas básicas da bioinformática também avaliada neste estudo, consiste no agrupamento de seqüências. O agrupamento (também conhecido como *clustering* ou *assembly*) é importante, pois a seqüência das leituras obtidas através do método de Sanger em seqüenciadores capilares dificilmente ultrapassa mil pares de bases. Há de se notar, entretanto, que as moléculas biológicas são, sem dúvida,

maiores do que este tamanho, principalmente se considerarmos dados de DNA genômico que constituem as moléculas cromossomais de organismos eucarióticos ou procarióticos. O agrupamento de seqüências, portanto, é utilizado para reunir em uma só molécula virtual, as seqüências obtidas das moléculas reais, construindo consensos cada vez maiores, que podem chegar a milhões de pares de bases, como é o caso da montagem de cromossomos eucarióticos. Apesar da existência de diversos programas de agrupamento de seqüências, como o SEQAID (Peltola et al., 1984), AMASS (Kim et al., 1999), Celera Assembler (Myers et al., 2000), Euler (Pevzner et al., 2001), GigAssembler (Kent and Haussler, 2001), ARACHNE (Batzoglou et al., 2002) e PCAP (Huang et al., 2003), os algoritmos mais utilizados pelos pesquisadores ainda são o PHRAP (Green, 1998) e o CAP3 (Huang and Madan, 1999), sendo que ambos levam em consideração os valores de qualidade produzidos pelos algoritmos de nomeação de bases de forma a tentar produzir uma versão mais consistente das seqüências consenso.

Além da montagem de genomas, os algoritmos de agrupamento de seqüências são também utilizados para agrupar seqüências parciais de cDNA, as ESTs (Adams et al., 1991), com o objetivo de descobrir novos genes e analisar a expressão gênica de um determinado organismo submetido a uma condição temporal ou espacial específica. No caso do agrupamento de ESTs, a utilização de tais programas de agrupamento é também importante para eliminar a redundância das seqüências (Figura 5), facilitando a anotação (Oliveira e Johnston, 2001) e, acredita-se, aumentando o nível de confiabilidade de cada uma delas (Miller et al., 1999).

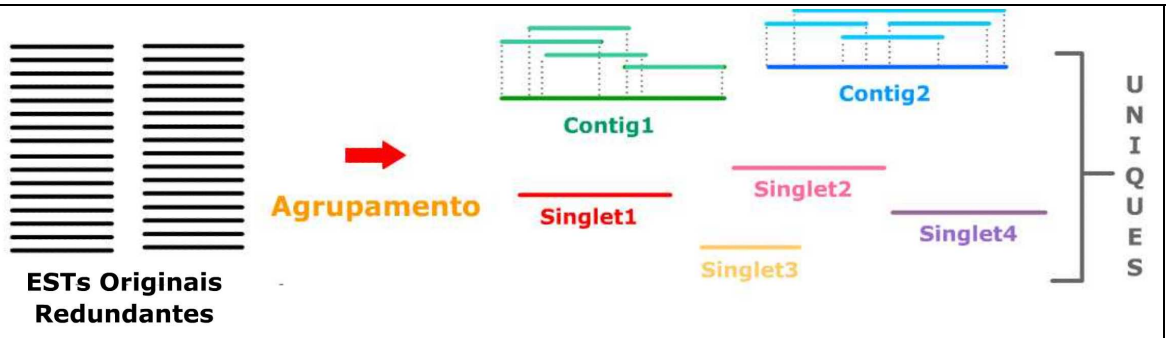


FIGURA 5: **Agrupamento de seqüências de ESTs.** O agrupamento das seqüências aumenta o tamanho médio do conjunto de dados, produz as seqüências não-redundantes chamadas de uniques e, acredita-se, melhora a qualidade das seqüências produzidas. Os uniques são o conjunto das seqüências consenso mais as seqüências *singlets*.

Os algoritmos de agrupamento são freqüentemente executados em duas etapas principais que consistem na (1) separação das seqüências em grupos, baseado na similaridade entre elas ser maior do que um limiar pré-definido e na (2) montagem do consenso, baseado na superposição das seqüências do mesmo grupo e análise dos valores de qualidade para a construção do consenso (Green, 1998; Huang and Madan, 1999; Batzoglou et al., 2002; Huang et al., 2003). Alguns algoritmos, entretanto, realizam apenas uma ou outra função, como é o caso do BLASTclust, do pacote BLAST (Altschul et al., 1997), que apenas mostra quais são as seqüências do mesmo grupo, não realizando nenhum procedimento de montagem dos consensos. No caso do algoritmo PHRAP o escore mínimo para o agrupamento de seqüências é igual a 30 (Green, 1998; Figura 6). Já o CAP3 leva em consideração valores de qualidade também no momento de realizar o agrupamento das seqüências e, portanto, o escore do alinhamento é ponderado por estes valores (Huang and Madan, 1999).

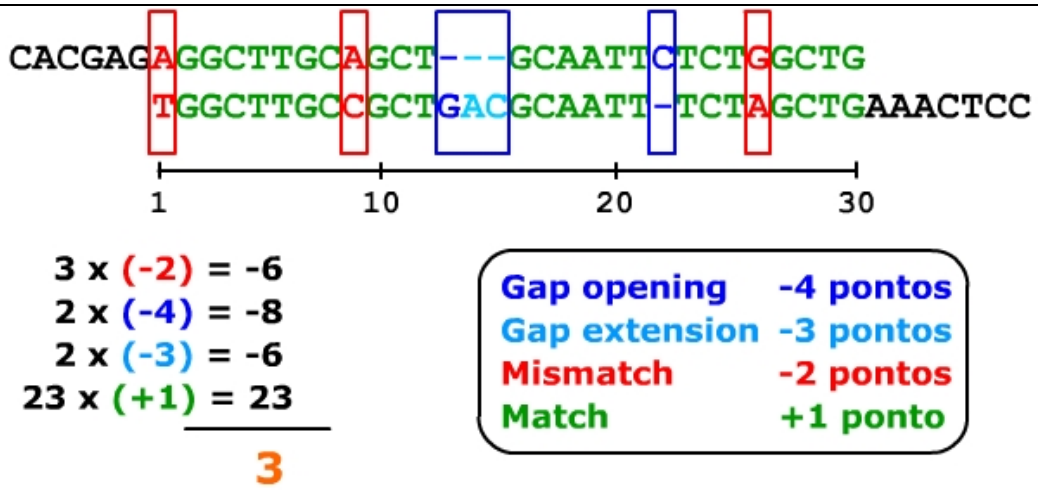


FIGURA 6: **Procedimento básico para o agrupamento de seqüências.** As seqüências são alinhadas e os valores de acertos (*matches*), trocas (*mismatches*), aberturas e extensões de gap (*gap opening* e *gap extension*) são calculados de acordo com os valores padrão, presentes em uma matriz pré-definida. A soma desses valores gera um escore do alinhamento que, neste exemplo apresentado, é igual a 3. As seqüências são agrupadas caso esse valor de escore do alinhamento seja maior do que um limiar pré-definido que, no caso do PHRAP, é igual a 30.

1.5. Agrupamento de seqüências utilizando os algoritmos CAP3 e PHRAP

Os programas de agrupamento mais populares e mais utilizados hoje em dia são o PHRAP e o CAP3. Há algum tempo atrás, notava-se uma maior utilização do PHRAP e do pacote completo PHRED-PHRAP-CONSED na área genômica. Hoje em dia, entretanto, diversos trabalhos parecem ter atestado a melhor adequação do CAP3 em diversas aplicações e parece que este tem sido o programa mais utilizado tanto para o agrupamento de seqüência de DNA quanto de cDNA (Masoudi-Nejad et al., 2006; Lee et al., 2005; Prosdocimi et al., 2002).

Segundo seu manual, o CAP3 funciona através dos seguintes procedimentos:

1. Corte das regiões de baixa qualidade 5' e 3';
2. Realização de um alinhamento global das seqüências entre si;
3. Cálculo do escore de alinhamento entre cada par de seqüências (tamanho da seqüência sobreposta *x* qualidade da região de sobreposição *x* escores de match/mismach/gap) através de alinhamento global;
4. Realização de alinhamentos locais para identificar falsas sobreposições;
5. Observação do arquivo de entrada contendo a identificação das seqüências e o tamanho máximo e mínimo de distâncias entre elas (o CAP3 permite a utilização desse tipo de arquivo, o que proporciona sua utilização em projetos onde há seqüenciamento apenas das extremidades de clones), identificando falsas sobreposições;
6. Comparação do resultado do escore com os valores limites definidos;
7. Se o escore do alinhamento for menor do que o escore mínimo as seqüências não formam um agrupamento, se for maior, as seqüências são agrupadas;
8. Alinhamento global das seqüências de cada consenso;
9. Cálculo dos valores de qualidade dos nucleotídeos de cada seqüência em cada posição do alinhamento global, para definir qual base será adicionada ao consenso e qual sua qualidade final;
10. Análise das deleções e inserções entre as seqüências para definir a montagem do consenso;
11. Montagem final das seqüências consenso.

Já o PHRAP, segundo a documentação do programa, funciona através dos seguintes passos de montagem:

1. Lê a seqüência e o arquivo de qualidade, corta regiões de homo-polímero no fim das seqüências e constrói as seqüências complementares;
2. Encontra pares de seqüências que têm regiões de similaridade. Elimina leituras duplicadas. Realiza comparações SWAT (Smith-Waterman) em pares de seqüências que apresentam regiões de sobreposição e computa o escore SWAT;
3. Procura regiões de sobreposição características de vetores e marca-as de forma que não sejam utilizadas no agrupamento;
4. Encontra regiões duplicadas;
5. Encontra seqüências com regiões de sobreposição em si mesmas;
6. Encontra pares de seqüências que não apresentam regiões boas de sobreposição;
7. Realiza comparações de seqüências aos pares para confirmar sobreposições, utiliza-as para computar valores de qualidade;
8. Computa escores para cada sobreposição (baseado na qualidade de bases iguais e diferentes);
9. Realiza novamente os dois passos anteriores;
10. Encontra o melhor alinhamento para cada par sobreposto que tenha mais de um alinhamento significante numa dada região (utiliza o melhor escore dentre várias sobreposições);
11. Identifica seqüências provavelmente quiméricas e com deleções;
12. Constrói esquema de consensos, utilizando os escores de pares de sobreposições em ordem decrescente. A consistência dos esquemas é checada em nível de comparação entre os pares de seqüências;
13. Constrói a seqüência dos consensos como um mosaico das partes de maior qualidade das leituras;
14. Alinha seqüências aos consensos, observa inconsistências e possíveis locais de alinhamento incorreto. Ajusta os escores das seqüências dos contigs.

Ainda que o CAP3 venha sendo mais utilizado que o PHRAP, é notória a diferença de performance e tempo de execução entre os dois programas. A execução do PHRAP é muito mais rápida que o CAP3 e ele ainda apresenta um algoritmo extra que permite o agrupamento de um número de *reads* maior do que 64.000.

1.6. Erros em seqüências de DNA

Todos sabemos que nenhum tipo de empreendimento humano está isento de erros, sendo que esta máxima vale também para os projetos genoma. Desde 1996, quando da realização de um workshop de validação de seqüências de DNA no NHRGI (*National Human Genome Research Institute*), já se falava que a quantidade de erro aceitável para o genoma humano seria de uma base incorreta a cada 10.000 sequenciadas e que os processos de nomeação das bases e agrupamento de seqüências deveriam passar por estudos de validação, de preferência realizados por outros grupos não relacionados ao NHRGI (Felsenfeld et al., 1999).

Antes disso, porém, o primeiro trabalho que temos notícia tratando de erros em projetos de sequencimento de DNA, foi publicado no início da década de 90 por uma equipe associada ao TIGR (White et al., 1993). Enquanto ainda se parecia questionar a validade da técnica de sequenciamento de etiquetas gênicas (ESTs), White, Adams, Venter e colaboradores estavam preocupados com a contaminação de seqüências de outros organismos em suas bibliotecas. Assim, desenvolveram um algoritmo que verificava a representatividade do conteúdo de seqüências de DNA de seis letras (hexâmeros) em cada uma das espécies. Através, portanto, da verificação da quantidade relativa desses hexâmeros em uma seqüência, seria possível identificá-la como sendo desta ou daquela espécie (White et al., 1993).

Desde então, vários algoritmos vêm sendo descritos para tentar corrigir diferentes tipos de erros mais comuns em projetos genoma, começando pelo processamento das seqüências (Staden, 1996; Smith et al., 1997; Wendl et al., 1998; Scheetz et al., 2003; Adzhubei et al., 2006) e passando para a avaliação de erros de mudança na fase de leitura (*frameshifts*) (Fichant and Quentin, 1995; Guan and Uberbacher, 1996; Medique et al., 1999), erros no sequenciamento de regiões repetitivas (Tammi et al., 2002; Tammi et al., 2003) e, principalmente, novos algoritmos para tentar melhorar o agrupamento de seqüências e a formação dos consensos (Green, 1998; Huang and Madan, 1999; Kim et al., 1999; Myers et al., 2000; Pevzner et al., 2001; Kent and Haussler, 2001; Batzoglou et al., 2002; Huang et al., 2003).

Sem que pudéssemos, entretanto, avaliar todos esses parâmetros sobre qualidade de dados genômicos e acreditando na importância de uma sólida fundamentação das bases do conhecimento, preferimos focar a presente tese em análises de algoritmos nomeadores de bases – no caso, o mais utilizados deles, o

PHRED – tentando avaliá-lo racionalmente e escolher a melhor forma de parametrização e utilização do mesmo, como se verá a seguir.

1.7. Alinhamento de seqüências

O alinhamento de seqüências é outra das técnicas básicas da bioinformática abordada na presente tese. Entretanto, o alinhamento de seqüências é utilizado aqui não como objeto de pesquisa, e sim como uma ferramenta fiel para se analisar tanto o processo de nomeação de bases quanto o processo de agrupamento de seqüências. O alinhamento de seqüências de biomoléculas consiste no processo de comparar duas seqüências (de nucleotídeos ou proteínas) de forma a se observar seu nível de identidade ou similaridade, para que possamos inferir (ou não) a uma delas, alguma propriedade já conhecida da outra. O alinhamento entre duas seqüências pode ser feito de forma global ou local (figura 7).

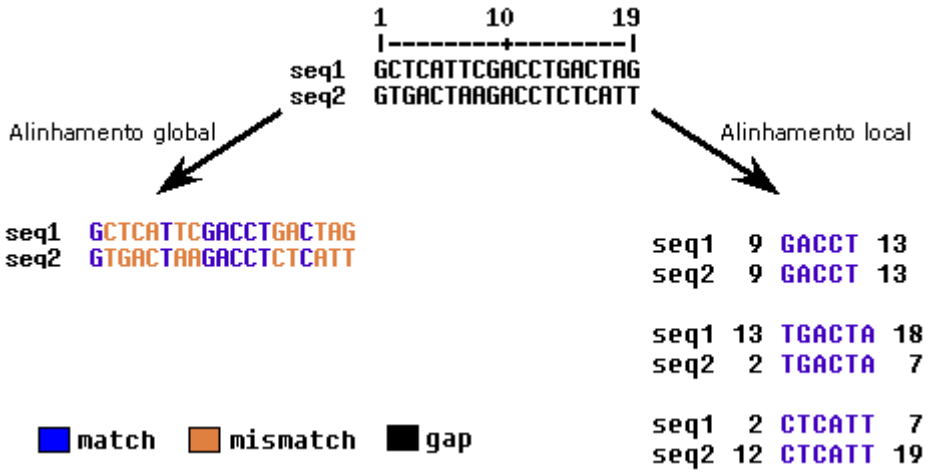


FIGURA 7. **Alinhamento global e local.** À esquerda vemos um exemplo de como é feito um alinhamento global das seqüências e à direita vemos um exemplo da realização de um alinhamento local. Retirado de Prosdocimi et al., 2003.

O alinhamento global é realizado quando comparamos uma seqüência de aminoácidos ou nucleotídeos com outra, ao longo de toda sua extensão (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html). O popular programa MULTALIN, por exemplo, realiza um alinhamento heurístico, múltiplo e global (Corpet, 1988) entre seqüências de biomoléculas. Já o bastante conhecido algoritmo de Needleman-Wunsch realiza o alinhamento global ótimo entre duas

seqüências de biomoléculas quaisquer (Needleman and Wunsch, 1970). Já o alinhamento local acontece quando a comparação entre duas seqüências não é feita ao longo de toda sua extensão, mas sim através de pequenas regiões de similaridade entre elas (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html).

Uma particularidade interessante dos programas de alinhamento local que foi explorada no presente trabalho, está relacionada ao fato de que tais algoritmos iniciam o alinhamento entre duas seqüências nas regiões onde elas mostram similaridades altas e tentam estender esse alinhamento até as bordas dessas seqüências. Acontece que, caso as seqüências comecem a se mostrar excessivamente diferentes, o valor de escore daquele alinhamento cai abaixo de um limiar pré-definido e, assim, o algoritmo desiste do alinhamento e reporta, como resultado, apenas a região onde a similaridade tenha se mostrado acima daquele limiar.

Utilizamos aqui dois dos principais algoritmos de alinhamento local para realizar o agrupamento de seqüências em moléculas de DNA, o *gapped* BLAST (*Basic Local Alignment Search Tool*, Altschul et al., 1997) e o algoritmo Smith-Waterman (SWAT, Smith and Waterman, 1981). Ambos os algoritmos baseiam-se na comparação das seqüências de entrada com seqüências presentes num banco de dados.

As principais diferenças entre o BLAST e o SWAT estão relacionadas à qualidade do alinhamento reportado e ao tempo de execução dos algoritmos. O SWAT apresenta o melhor alinhamento local possível (alinhamento ótimo), pois funciona através da montagem de uma matriz de similaridade entre as seqüências de entrada e aquelas presentes no banco de dados (Smith and Waterman, 1981). Essa matriz de comparação é montada tendo como base valores pré-definidos de similaridade e dissimilaridade entre cada uma das bases, definidos através da matriz de substituição. Neste trabalho, algumas vezes utilizamos o SWAT associado a duas matrizes de substituição diferentes para alinhamentos de seqüências nucleotídicas (*mat50* e *mat70*), de forma que pudéssemos observar melhor o comportamento de nossos dados.

Já o BLAST consiste em uma heurística que acelera a busca do melhor alinhamento (McGinnis and Madden, 2004). Ele funciona partindo a seqüência de entrada em subseqüências sementes e verificando quais as seqüências presentes no banco de dados apresentam tais subseqüências coincidentes (Altschul et al., 1990; Altschul et al., 1997; Zhang et al., 1998). A partir, então, das similaridades de subseqüências semente, o BLAST tenta alongar tais subseqüências para ambos os lados e, caso esse alongamento ultrapasse um escore mínimo pré-definido, considerando também os valores de coincidências e trocas pré-dispostos pela matriz

de substituição, um resultado (*hit*) é apresentado ao usuário (Altschul et al., 1997; Ye et al., 2006). Apesar de não apresentar um resultado ótimo como o algoritmo SWAT, o resultado do BLAST já se mostrou ser muito próximo deste "melhor resultado possível" e, considerando que ele acelera enormemente a busca nas enormes bases de dados em biologia molecular, desde sua criação ele tem sido o algoritmo mais popular para o alinhamento de seqüências de biomoléculas.

1.8. Dogmatismo, paradigmas científicos e questões sócio-econômicas

Segundo Thomas Kuhn (1962), o desenvolvimento da ciência normal – termo definido como aquela ciência que vem sendo praticada pela grande maioria dos laboratórios de pesquisa mundiais –, consiste principalmente na adequação do cientista a um princípio paradigmático em sua linha de atuação e a realização de experimentos do tipo "resolução de quebra-cabeças" dentro dessa linha de pesquisa diretamente atrelada a um determinado paradigma científico vigente. Se observarmos de perto a atuação e a realização comum das práticas científicas, veremos que os cientistas muito poucas vezes questionam os conhecimentos mais básicos em suas áreas de atuação, tendo-os como fixos, imutáveis e corretos. A história das ciências tem demonstrado que durante a atividade padrão dentro do que é considerada a ciência normal, os cientistas não estão preocupados em questionar as bases do conhecimento adquirido e tido como correto em determinada área da pesquisa científica. Assim, desinteressados em validar melhor tais bases e procurar um conhecimento mais sólido, os cientistas passam a tratar freqüentemente apenas da obtenção de mais conhecimento específico – algo que é chamado por Kuhn de "resolução de quebra-cabeças" –, tendo considerado que as bases demonstradas para sua ciência estão corretas e que representam fielmente a natureza do universo, da vida ou da mente.

Assim, a história das ciências mostra que os conhecimentos adquiridos pela última geração de cientistas são freqüentemente considerados pelos cientistas modernos como dogmas indestrutíveis nos quais se deve acreditar e procurar, com sua pesquisa, conhecer novos detalhes sobre os mesmos.

Ao contrário do paradigma vigente, entretanto, a presente tese pretende questionar alguns dogmas arraigados à pesquisa genômica mundial, sendo o principal deles o fato de que "as seqüências de bases que dispomos representa fielmente a composição de bases de um determinado organismo de estudo". Aqui procuramos questionar este dogma e tentar, através de uma metodologia bem delineada,

evidenciar até onde essa afirmativa pode ou não ser verdadeira. Testes extensos foram realizados de forma a racionalizar a questão e verificar até quando e de qual forma podemos e devemos observar e "acreditar" na fidelidade das seqüências de DNA com as quais temos trabalhado correntemente. De forma semelhante, bioinformatas e biólogos tendem a acreditar fielmente nos resultados apresentados pelos algoritmos que executam e esquecem-se que o desenvolvimento de uma metodologia computacional está invariavelmente associada a uma inevitável quantidade de erros metodológicos. A montagem de genomas, por exemplo, exige que diversas seqüências de DNA (já produzidas com erros em três fases distintas, segundo a Figura 1) sejam concatenadas em uma única seqüência que posteriormente irá representar uma molécula biológica inteira, como um cromossomo eucarioto ou um genoma circular de um organismo procarioto. Ao compararmos, neste trabalho, o resultado da execução desses algoritmos com um controle positivo da seqüência que se deseja gerar, fomos capazes de identificar diversos erros comuns que ocorrem durante este procedimento. Esse fato mostra quão sujos tendem a ser os dados biológicos sobre seqüências de biomoléculas que obtemos dos mais afamados bancos de dados existentes para tanto.

Dessa forma, acreditamos que o desenvolvimento da ciência enquanto a busca por novos conhecimentos deve passar por uma análise estreita dos métodos utilizados, de forma que conclusões precipitadas não possam ser obtidas a partir da análise de dados apenas parcialmente corretos. De outra forma, entretanto, compreendemos que a ciência deve avançar em busca de uma interpretação cada vez mais precisa da realidade e que este avanço tem acontecido, a despeito da utilização de técnicas precisas até certo ponto. Nosso objetivo aqui, portanto, foi demonstrar que o questionamento das bases empíricas de uma certa ciência (no caso, a genômica) pode permitir uma melhor e mais eficiente observação dos dados de forma a se produzir conhecimentos mais sólidos.

Com relação a aspectos sociais e econômicos, estivemos também preocupados em realizar o que chamamos de uma ciência "mais limpa" ou a proposição de um tipo de "desenvolvimento sustentável" dentro da ciência acadêmica. Assim, estimulamos os cientistas a atentarem para os gastos extra de dinheiro, tempo e análise que podem ser oriundos de uma não racionalização prévia sobre a realização de projetos científicos. Isso foi aqui demonstrado mais claramente no caso da definição do melhor local onde um iniciador deve ser posicionado durante estudos de genoma e/ou transcriptoma. Com este trabalho, mostramos que somos capazes de evitar o desperdício de dinheiro e recursos com o sequenciamento apenas das porções realmente informativas das moléculas de DNA.

2. OBJETIVOS

Objetivo Geral

Estudar a forma de execução do algoritmo PHRED e estimar formas mais racionais de utilização de seus parâmetros; questionar alguns dogmas arraigados à cultura genômica e buscar uma maior racionalização da prática científica.

Objetivos Específicos

1. Explorar um conjunto eficiente de seqüências que permitirão análises confiáveis dos parâmetros do algoritmo PHRED;
2. Verificar com este conjunto de leituras o funcionamento do algoritmo PHRED: (a) comparando os valores de PHRED com os erros reais, (b) observando quais erros são mais comuns em diferentes valores de qualidade e (c) avaliando se há como prever as bases incorretamente nomeadas baseando-se na qualidade das bases na posição e na vizinhança dos erros;
3. Buscar um valor ótimo de PHRED para utilizar como máscara, de forma a grafar a maior parte das bases incorretas em letras minúsculas (*softmasking*) sem, no entanto, mascarar bases corretamente nomeadas e estudar a co-habitação de bases erradas com bases mascaradas em janelas de diferentes tamanhos;
4. Analisar a posição de início da nomeação das bases pelo PHRED e o início da seqüência com leitura confiável, de forma a determinar a melhor distância para o posicionamento dos iniciadores para seqüenciamento de insertos em bibliotecas, otimizando o número de bases seqüenciadas na região de início do inserto em projetos transcriptoma de organismos eucarióticos;
5. Encontrar o valor mais adequado para poda (*trimming*) das seqüências geradas e nomeadas pelo algoritmo PHRED, de forma a retirar a maior quantidade possível de informação biológica das leituras (*reads*);
6. Definir qual o número ideal de leituras a serem utilizadas quando se deseja produzir um consenso que represente, com fidelidade, a molécula molde, em função de diferentes intensidades de poda com o algoritmo PHRED.

3. JUSTIFICATIVA

O desenvolvimento da ciência cotidiana por vezes nos mostra que algumas técnicas de análises de dados são utilizadas pelos cientistas, como um consenso, sem que para isso tenham sido feitas análises minuciosas que explicitem a forma ideal de utilização de tais técnicas.

Com relação a programas de nomeação de bases, normalmente considera-se como ruim uma seqüência que apresente bases com valor de PHRED menor que 20. Além disso, acredita-se também que o iniciador que realizará o sequenciamento deva ser colocado aproximadamente 100 pares de bases antes do início do inserto – quando isto é lembrado, o que nem sempre é o caso. E, de forma semelhante, não se considera uma seqüência de genoma como validada caso ela não tenha sido seqüenciada ao menos cinco vezes em cada uma das fitas, sendo que para a produção de seqüências de cDNA sem ambigüidade, em larga escala, o tema sequer é discutido.

Mas de onde vieram todos esses axiomas? Por que utilizar PHRED 20? Por que 100 bases deve ser o valor correto? Por que cinco vezes em cada uma das fitas? O senso comum não está apenas na vida cotidiana das pessoas, ele também impera em áreas da ciência, empreendimento que se propõe preciso, técnico e confiável.

O presente trabalho foi desenvolvido de forma a responder algumas destas perguntas que os pesquisadores se fazem e, sem um pilar de apoio, tendem simplesmente a aceitar como verdade aquilo que se acredita no meio acadêmico, sendo que muitos pesquisadores adotam esses padrões como corretos sem jamais questionar a artificialidade dos mesmos. Aqui desenvolvemos análises racionais da utilização dos algoritmos PHRED, principalmente, além de PHRAP e CAP3 com o objetivo de definir padrões através dos quais os pesquisadores possam entender melhor como funciona o mecanismo de nomeação de bases e agrupamento de seqüências utilizando tais programas. Dessa forma, fornecendo esses pontos de apoio teóricos, os pesquisadores serão assim capazes de definir, com um embasamento racional, a forma mais adequada de utilização desses algoritmos em seus trabalhos.

Além disso, o seqüenciamento completo de um genoma gera contribuições inesperadas no conhecimento de outros organismos diversos do que esta sendo estudado, através da genômica comparativa. É notável como a comparação entre um regulador do ciclo celular de leveduras com o humano pode contribuir para a elucidação da origem do câncer, por exemplo. Assim, a busca por melhores formas de aproveitamento da informação biológica revelada por projetos de seqüenciamento em larga escala se justifica, por ser fonte de conhecimento incomensurável.

Acreditamos que os cientistas devem utilizar racionalmente as técnicas em seu trabalho e que ao invés de responderem "fazemos assim porque todos fazem da mesma forma", devem responder "fazemos assim porque lemos este estudo que se baseia nestes dados e que mostrou ser esta a forma mais adequada de utilização desta técnica". Vale salientar que o processo de nomeação de bases, principal tema deste trabalho, é um dos procedimentos mais importantes da bioinformática, pois está diretamente associado à produção das seqüências que são, em si, a base de todo o estudo em nossa área de pesquisa.

4. MATERIAIS E MÉTODOS

4.1 Versões dos softwares utilizados

- PHRED version 0.000925.c
- PHRAP version 0.990329
- BLAST version 2.2.10
- SWAT version 0.990329
- CAP3 version date 08/29/02
- PERL v5.8.0 built for i386-linux-thread-multi

4.2 Sistema operacional

Para todas as análises computacionais foi utilizado o sistema operacional LINUX, nas distribuições mais atuais de Red Hat, Fedora e Suse. Para as análises dos dados, a geração de planilhas, de gráficos e figuras, frequentemente foi utilizado o sistema operacional Windows e o pacote MS Office.

4.3 Bancos de dados

Todos os dados foram armazenados em um banco de dados MySQL (versão 3.23.54), onde foram construídos bancos e tabelas específicas para melhor guardar e obter os dados brutos durante a execução do projeto.

4.4 Computadores

Todas as análises apresentadas aqui foram executadas em estações de trabalho rodando sistema operacional Linux. Ainda que alguns algoritmos desenvolvidos em linguagem PERL tenham demorado dias para completarem sua execução, não houve necessidade da utilização de grandes servidores para a elaboração do presente estudo.

5. RESULTADOS E DISCUSSÕES

5.1. Single-pool sequencing

Todas as análises desenvolvidas para a presente tese foram baseadas em um conjunto de seqüências do plasmídeo pUC18 produzidas segundo o procedimento que chamamos de *single-pool sequencing*. Esse procedimento consistiu na preparação de uma reação de sequenciamento em um único tubo, posteriormente dividida em algumas alíquotas para que fosse realizada a reação de sequenciamento nas máquinas termocicladoras. Depois, o conteúdo dos tubos apresentando as moléculas de DNA já polimerizadas, contendo os terminadores didesoxinucleotídeos, foi novamente reunido em um só tubo, homogeneizado e então, as amostras foram todas novamente separadas em três placas de 96 poços (*wells*) para que a realização do seqüenciamento fosse realizada.

A motivação para realizar esse seqüenciamento em um único conjunto veio da idéia analisar o comportamento médio das moléculas, sem, no entanto, nos atermos a detalhes e pequenos problemas que porventura poderiam ter acontecido em uma ou outra das reações de seqüenciamento. O conteúdo de A, C, G, T em pUC18 é próximo de 25% (24,8 A; 25,2 C; 25,5 G; 24,5 T), sugerindo ausência de viés por esse parâmetro.

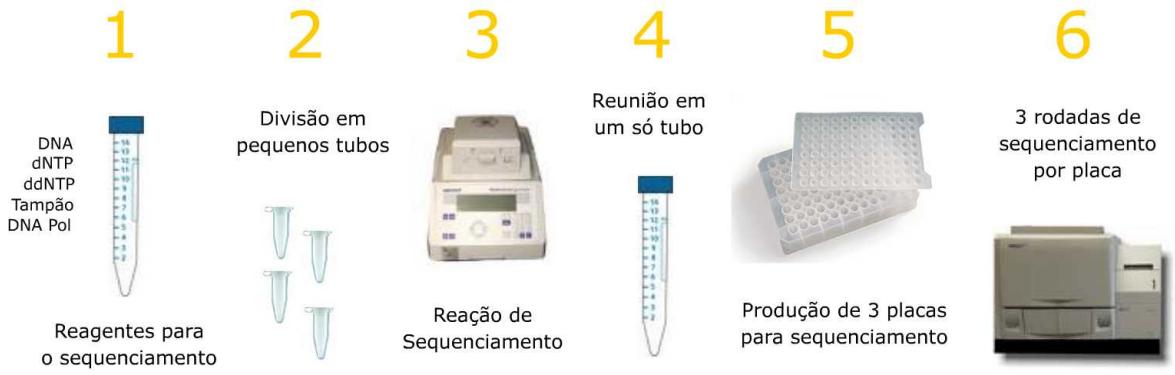


FIGURA 8: **Single-pool sequencing.** Resumo das etapas para a produção das seqüências segundo a técnica de single-pool sequencing.

Cada uma das três placas contendo, teoricamente, um conjunto homogêneo de moléculas amplificadas, foi submetida três vezes à reação de eletroforese capilar em um seqüenciador MegaBACE, produzindo um total de 864 *reads* de sequenciamento,

que possibilitaram a geração de 846 arquivos ESD utilizados nas análises a seguir de diferentes formas.

5.2. Alinhamentos dos *reads* com a sequência do pUC18

Grande parte das análises realizadas na presente tese foi realizada, tendo como base, o alinhamento dos 846 *reads* de pUC18 produzidos com o consenso publicado para seqüência desse vetor de clonagem (GenBank Accession Number L09136), utilizado como controle positivo de toda a análise.

Os 846 *reads* foram alinhados através dos algoritmos BLAST e SWAT -- este último utilizando duas diferentes matrizes de comparação de seqüências de DNA: mat50 e mat70 -- contra a seqüência publicada para este vetor de clonagem. Todos os erros observados nos *reads* foram identificados (troca de base, inserção ou deleção) e localizados com relação à posição na seqüência publicada. Toda essa informação sobre os dados brutos de erros de sequenciamento foi armazenada em um banco de dados MySQL.

Este banco de dados contendo todos os erros de sequenciamento gerados nestas amostras foi utilizado como base para grande parte das análises aqui descritas, como será explicitado em cada um dos artigos a seguir.

5.3. Análise do padrão de bases incorretas nomeadas pelo PHRED em seqüências de DNA

Nossa primeira análise baseou-se em testar a eficiência do algoritmo PHRED e avaliar seu funcionamento.

Considerando que mesmo os autores do trabalho original de descrição do algoritmo PHRED afirmaram que um baixo valor de qualidade de uma base não necessariamente está relacionado ao fato de que a referida base esteja incorreta (Ewing and Green, 1998), resolvemos tentar correlacionar tais valores com a presença de erros em nosso conjunto de dados. Para isso, observamos a correlação entre as bases incorretas e seu valor de qualidade criando índices como "erro observado" e "erro esperado"; verificamos a presença de bases incorretamente nomeadas de acordo com o tipo de erro (inserção, deleção ou troca de bases) para cada valor de qualidade; e verificamos se as bases na vizinhança dos erros apresentam algum padrão que nos pudesse permitir a previsão do erro de sequenciamento.

Nossos resultados mostraram que o PHRED parece adicionar os valores de qualidade corretamente, apesar de que foi mostrado que as regiões com baixos valores de PHRED têm qualidade sub-estimada. Mostramos ainda que, em geral, a maioria dos erros observados representa trocas de bases (*mismatches*) e que, em regiões de alta qualidade, os principais erros encontrados são representados por deleções de bases corretas. A vizinhança das trocas e inserções apresenta PQV médio próximo de 6 em toda a janela vizinha, já as deleções tendem a ser mais problemáticas, pois estas ocorrem em regiões de PQV um pouco maior (em torno de 10, em média). Esse trabalho foi submetido ao congresso WOBII (*Work on Bioinformatics* II) e foi posteriormente publicado na Revista Tecnologia da Informação, da pontifícia universidade de Brasília.

DNA SEQUENCES BASE CALLING BY PHRED: ERROR PATTERN ANALYSIS

Francisco Prosdocimi[1]
Fabiano Cruz Peixoto[2]
José Miguel Ortega[3]

ABSTRACT: PHRED is the most frequently used base caller algorithm in genome projects. An interesting point on PHRED utilization is the fact that a low score on some base may not actually correspond to a miscalling on that base, but it may stand for a putative error on the region around this base. In order to evaluate the efficiency of PHRED on base calling and base quality assigning, we have sequenced pUC 18 and compared sequences called by PHRED with pUC 18 published sequence using Smith-Waterman algorithm. Our results depict a detailed pattern of errors of incorporated by the algorithm, confirm that PHRED provides appropriated base calling but: low-quality regions have their quality usually under-estimated, with most errors being mismatches. On the other side, high-quality regions have super-estimated quality, with errors mainly represened by deletions.

1. Introduction

A mith on genomic science is that DNA sequencing machines are the actual agent that identifies the sequence of bases in DNA molecules. As important as the sequencing equipment is the base caller algorithm. The most known algorithm is PHRED, written by Green an Ewing [1] [2] [3]. Base calling consists in the process of analyzing raw data generated by sequencing equipments and the calling of quality values for each base [4]. Despite the one to one relation between bases called and quality values, it is interesting to notice that sometimes low base quality does not actually correspond to a miscalling of the base, but it may stand for a putative error on the neighborhood region around the base [2]. The lack of a detailed error pattern contributes to a mistaken belief that low quality values stands for miscalled bases while high quality bases stands for extremely accurate calling.

In order to evaluate the pattern of errors introduced by PHRED, a significant number of DNA sequences of the plasmid pUC 18 were generated and raw data was base called by PHRED. The sequences were aligned with the pUC 18 published sequence using the Smith-Waterman (SWAT) algorithm [5]. The errors identified by the alignments were used to pupulate a MySQL database. The analysis of this database showed interesting features related to the frequency ant type of errors occurred in low and high quality regions.

2. Materials and Methods

Sequences used in this work have been provided by laboratories from Universidade Federal de Minas Gerais (UFMG) that integrate the network Rede Genoma de Minas Gerais. The reactions were made in a single pull and divided on tubes for the PCR sequencing reaction. After the PCR sequencing reaction, the sequences were joint again on the same tube, mixed, and then divided on three 96 sequencing well plates. Each plate was run 3 times on a MegaBASE sequencing equipment, yielding a total of 864 reads. From those, 846 processed ESD files (840100 bases) were obtained. All ESD files were processed by PHRED with trimming parameters. Average size and quality were modified from 993 and 19.2 to 747 and 25.6 by trimming, respectively, resulting in a total of 632034 bases. PHRED was executed with trim alt and trim cutoff parameters; the following command line was used. The choice for trim cutoff 0.16 was made based on our previous work presented in I WOB [6].

```
phred traces/trace_i -trim_alt "" -st fasta
-q qual/trace_i.qual
- s fasta/trace_i.fasta -trim_cutoff 0.16
```

[1] Departamento de Biologia Geral, ICB, UFMG
franc@icb.ufmg.br
[2] Laboratório de Computação Científica, UFMG,
fpeixoto@lcc.ufmg.br
[3] Departamento de Bioquímica e Imunologia, ICB, UFMG.
miguel@icb.ufmg.br

R. TECNOL. INF. BRASÍLIA V. 3 N. 2 p. 107-110 DEZEMBRO 2003

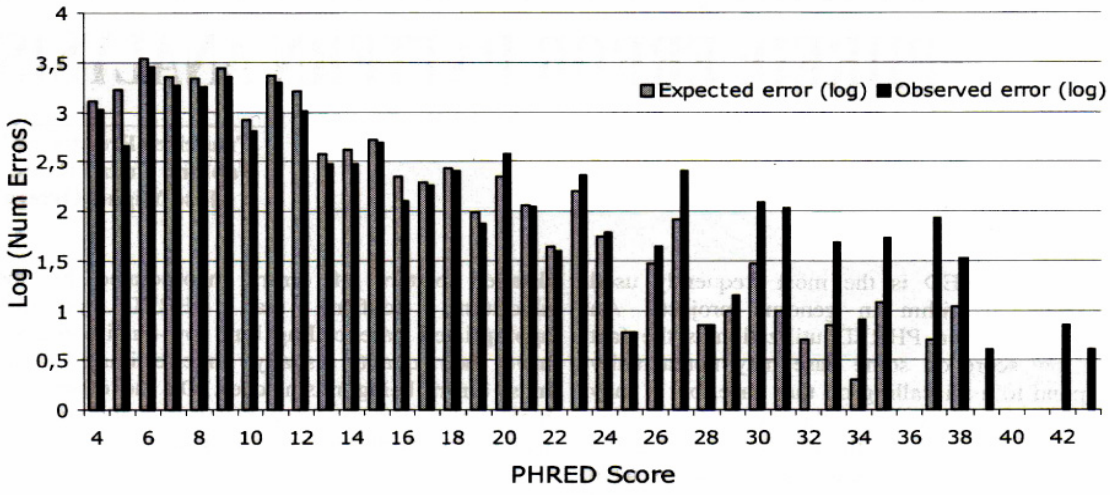


Fig. 1. Predicted x Observed Errors by PHRED Score.

All sequences produced were compared to the pUC 18 published sequence (24.8%A, 35.2%C, 25.5%G, 24.5%T) through SWAT algorithm. A PERL parser was built to populate a MySQL database with the position and the type (mismatch, insertion deletion) of each difference on the alignment result. The command line used for the alignment execution is shown below.

```
swat fasta/trace_i.fasta pUC18 -M mat 70 -
N 1 > trace_i.swat
```

3 Results and Discussion

In order to test if PHRED is defining precisely the chance of error for each position, the bases were classified based on their PHRED score. From now on, this PHRED scores will be called Predicted PHRED Score or pPHRED. We first counted the number of Observed Errors or oErrors based on the SWAT results. We defined Bases[pPHRED] as the total number of bases for each pPHRED. Using this total and the formula below, the number of Predicted Errors or pErrors was calculated.

$$pErrors = Bases[pPHRED] * 10^{\frac{-pPHRED}{10}}$$

We also calculated the Observed PHRED Score or oPHRED for a group of bases predicted with pPHRED based on the oErrors:

$$oErrors = Bases[pPHRED] * 10^{\frac{-oPHRED}{10}}$$

Figure 1 shows the number of putatively miscalled bases predicted for each PHRED score (pErrors) and the number of miscalls actually observed in the analysis of the alignment (oErrors). None out of the 46628 bases with PHRED score higher than 42 were actually miscalled and a significant number of bases (219385, 35%) were between PHRED scores 30 to 42, in which interval observed errrors were higher than expected.

A remarkable goal in the analysis of the database was to reveal the error distribution pattern, considering the type of error introduced by PHRED. Three types of errors can occur - mismatches, insertions and deletions. All of them could be retrieved from the SWAT alignment. The errors were mapped using a SWAT parser and used to populate a MySQL database.

Figure 2 shows the distribution of errors ranked by PHRED score. Mismatches or deletions are the most preponderant errors upon low or high PHRED scores, respectively. Insertions are the less frequent errors for all scores analysed.

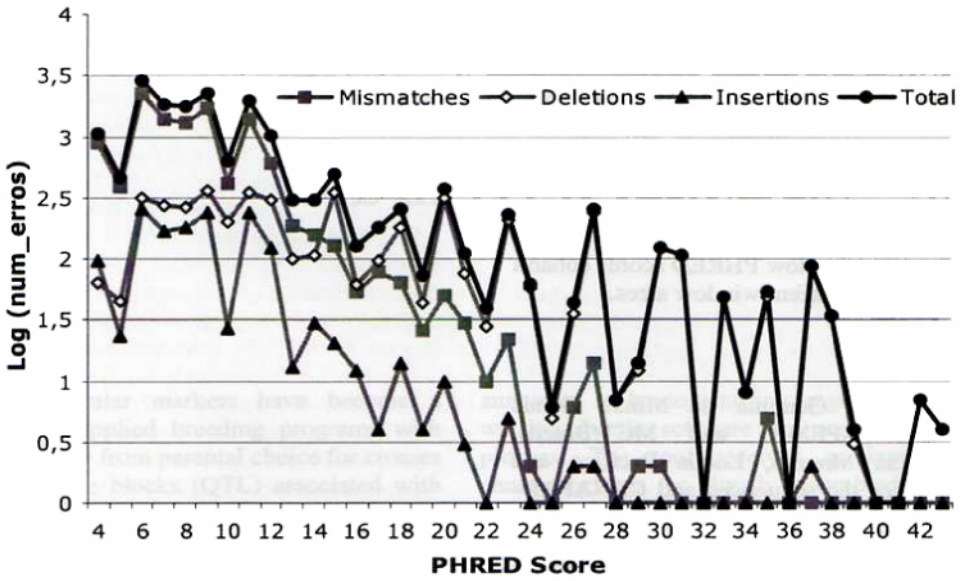


Fig. 2. Error Types by PHRED Score.

PHRED scores on the neighborhood of the errors were also verified in order to test if there were clear patterns of low quality scores surrounding the observed errors. We analysed the quality of the region surrounding the errors, from -5 to +5, being 0 the error position. An average of the quality at each of these positions, for each type of error was measured and plotted on Figure 3. most errors reside in low quality neighborhoods of similar scores. Mismatches are present in the lowest score regions and contrasting with deletions, as seen before. Only mismatches show a mild bias for the position of the error. Thus, in general PHRED scores seem not to provide a hint for what is the miscalled base.

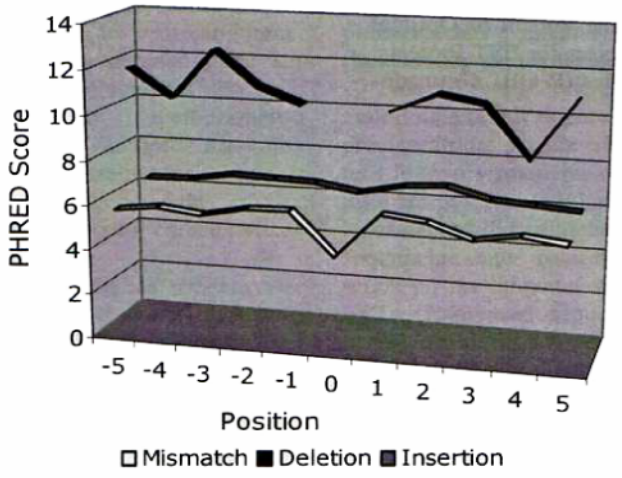


Fig. 3. Average PHRED Score on Error neighborhood.

4 Conclusions

As expected, PHRED base caller algorithm was confirmed to efficiently address the base calling issue and assigning accurate quality values. Errors tended to be underestimated in the higher score bases. A hint for the miscalled base was not observed, but a tendency for mismatches or deletions to be associated with, respectively, lower or higher scores and regions were depicted. We are currently testing whether low PHRED scores cohabit with actual errors in different window sizes.

5 Acknowledgment

We thank Rede Genoma de Minas Gerais (supported by FAPEMIG and MCT/Brazil), especially Marina Mourao, Lucila Pacifico and Renata Ribeiro for the sequences and CENAPAD-MG/CO for machines used in this work.

REFERENCES

1. http://www.phrap.org/phrap.docs/phred.html
2. Ewing B, Green P. Base Calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186-194, 1998.
3. Ewing B, Hillier L, Wendl MC, Green P. Base-Calling of automated sequencer traces using phred. I. Accuracy Assesment. Genome Res 8: 175-185, 1998.
4. Prosdocimi et al. Bioinformatica: Manual do usurio. Biotec Ci Desenvolv. 29: 18-31, 2002.
5. Smith TF, Waterman MS.; Identification of common molecular subsequences. J Mol Biol. 147:195-7, 1981.
6. Peixoto F, Ortega M.; On the Pursuit of Optimal Sequence Trimming Parameters for EST Projects; I Workshop on Bioinformatics (WOB), Gramado-RS, 2002.

5.4. Avaliação da presença de bases incorretas em janelas apresentando baixos valores de PHRED

Considerando os dados mostrados no item anterior, sugerindo que as bases vizinhas parecem não indicar eficientemente qual a base incorreta e embora o valor de PHRED esteja freqüentemente corretamente associado à probabilidade de uma certa base estar incorretamente nomeada, resolvemos tentar predizer a co-habitação de bases incorretas com bases de baixo valor de PHRED.

Portanto, o objetivo desta etapa foi verificar se a utilização de um valor limite de qualidade de PHRED poderia ser utilizado para mascarar a maior parte dos erros em uma determinada seqüência, ou, adicionalmente, em uma janela em torno da base de baixo PQV. Nossa idéia inicial seria tentar mascarar esses erros ao representar, por exemplo, todas as bases com PQV menor do que 10 em letras minúsculas (procedimento conhecido como *softmasking*) nas seqüências a serem depositadas futuramente no GenBank (Benson et al., 2006). O benefício deste procedimento constitui-se no fato de que programas de alinhamento, como BLAST, podem ser programados de forma a evitar o início de alinhamento em seqüências semente que contenham bases grafadas em minúsculas. Portanto, testamos diferentes valores de mascaramento de qualidade para avaliar qual deles mascarava mais bases incorretas (verdadeiro positivo) sem, ao mesmo tempo, mascarar muitas bases corretas (falso positivo). Como se gasta a mesma quantidade de *bytes* para armazenar uma seqüência toda em letras maiúsculas ou uma seqüência contendo maiúsculas e minúsculas, essa informação proveria um nível de informação a mais às seqüências disponibilizadas pelo GenBank sem, no entanto, aumentar o tamanho da informação armazenada neste banco de dados. E de certa forma adicionaria um valor de qualidade limite que, se bem calibrado, seria útil no mapeamento de regiões com maior densidade de erros.

Assim, além de realizarmos o estudo base-a-base, decidimos avaliar o mesmo padrão de mascaramento de erros em um conjunto de bases contínuas que chamamos de "janelas de bases". Desta forma, o estudo do PQV em janelas de diferentes tamanhos poderia auxiliar em vários outros processos, como a identificação inequívoca de sítios para enzimas de restrição, identificação de erros em etiquetas gênicas produzidas pela técnica de SAGE (*Serial Analysis of Gene Expression*, Velculescu et al., 1995) e alinhamentos BLAST, como comentado, que podem utilizar janelas de diversos tamanhos, dependendo do programa. Portanto, realizamos o mapeamento das janelas

incorretas com relação a valores crescentes de PQV e verificamos quantas janelas incorretas e corretas eram mascaradas em diferentes limiares.

Nossos resultados mostraram que o valor de qualidade mais adequado de mascaramento de bases/janelas incorretas para a maior parte das aplicações é 6 ou 7, sendo que valores maiores mascaram muitas bases/janelas corretas, diminuindo o benefício em transformá-las em letras minúsculas, apesar de que o pesquisador pode e deve adotar valores diferentes dependendo do interesse de sua pesquisa. Este trabalho foi publicado na revista Genetics and Molecular Research com o título de "Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values".

Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values

Francisco Prosdocimi[1], Fabiano Cruz Peixoto[2] and José Miguel Ortega[3]
[1]Laboratório de Biodiversidade e Evolução Molecular, Departamento de Biologia Geral, ICB-UFMG, Belo Horizonte, MG, Brasil
[2]Laboratório de Computação Científica, UFMG, Belo Horizonte, MG, Brasil
[3]Laboratório de Biodados, Departamento de Bioquímica e Imunologia, ICB-UFMG, Belo Horizonte, MG, Brasil
Corresponding author: J.M. Ortega
E-mail: miguel@icb.ufmg.br
Genet. Mol. Res. 3 (4): 483-492 (2004)
Received October 4, 2004
Accepted December 6, 2004
Published December 30, 2004

ABSTRACT. When analyzing sequencing reads, it is important to distinguish between putative correct and wrong bases. An open question is how a PHRED quality value is capable of identifying the miscalled bases and if there is a quality cutoff that allows mapping of most errors. Considering the fact that a low quality value does not necessarily indicate a miscalled position, we decided to investigate if window-based analyses of quality values might better predict errors. There are many reasons to look for a perfect window in DNA sequences, such as when using SAGE technique, looking for BLAST seeding and clustering sequences. Thus, we set out to find a quality cutoff value that would distinguish non-perfect windows from perfect ones. We produced and compared 846 reads of pUC18 with the published pUC consensus, by local alignment. We then generated a database containing all mismatches, insertions and gaps in order to map real perfect windows. An investigation was made to find the potential to predict perfect windows when all bases in the window show quality values over a given cutoff. We conclude that, in window-based applications, a PHRED quality value cutoff of 7 masks most of the errors without masking real correct windows. We suggest that the putative wrong bases be indicated in lower case, increasing the information on the sequence databases without increasing the size the files.

Key words: DNA sequence quality, PHRED, Quality window, SAGE, BLAST

INTRODUCTION

Base caller algorithms are as important as sequencing machines for the identification of the sequence of bases in DNA molecules. They are responsible for the analysis of the raw data generated by the sequencing equipment and for the production of the sequence of bases

putatively related to the original molecule, as well as the quality values determined for each of them (Prosdocimi et al., 2002). The best-known and most widely used base caller algorithm is PHRED, written by Green and Ewing (Ewing et al., 1998; Ewing and Green, 1998). An approach frequently used by researchers looking for miscalled bases in DNA sequences is the choosing of a minimum quality value based on intuition, considering the significance of the PHRED quality value (PQV). PQV 20 is the most widely used, and operationally it means that a base has one chance in a hundred to be miscalled. However, a low quality value does not necessarily cohabit with a miscalled position (Ewing and Green, 1998; Prosdocimi et al., 2003).

Beyond the use of a quality cutoff for single bases, many applications can make use of the quality value for a number of bases in tandem, or a window of bases. There are many reasons for researchers to look for a perfect window (PW) in a DNA sequence, defined as a sequence of called bases that putatively do not contain any mismatch or gap (insertion/deletion). This PW is particularly important in the SAGE technique, which consists of single pass sequencing of concatenated fragments of the cDNA tail subsequent to a given restriction site (Velculescu et al., 1995). The bases juxtaposed to the restriction site constitute a tag that has been assigned to genes. One single error on a SAGE tag (containing 14 nucleotides) can generate incorrect associations and false positives (and negatives) in the gene expression inference. Thus, it is quite important to be able to establish an appropriate quality cutoff, under which a window lacks, probabilistically, the potential to be entirely correct, reducing the number of false inferences.

BLAST is another application that could take advantage of PW; it is possible to choose only the perfect windows to be used as a BLAST seeding window (Altschul et al., 1997). In BLAST execution, if one of the letters in the sequence is represented by lower case, it is possible to avoid seeding on them, using, in the stand-alone version, the flag - UT (see README in documentation for stand-alone BLAST). Thus, the alignments will only seed on uppercase PWs, since putatively incorrectly called bases are represented in lower case.

In order to evaluate if the lowest PQV could correctly mask non-perfect windows, we analyzed 846 single-pool reads of pUC18. Aligning the reads to the published sequence for this cloning vector, a database of all mismatches, insertions and gaps generated by the entire sequencing procedure was built. Different window sizes were tested in order to find the best fit between real perfect windows (RPWs) and predicted perfect windows (PPWs), the ones not containing a PQV equal to or below the chosen cutoff. We also evaluated which PQV cutoff showed the best potential to identify the position of sequencing errors without masking, or spoiling, correct windows, so that it could be used in various applications.

MATERIAL AND METHODS

Sequencing reactions

Three laboratories from the Universidade Federal de Minas Gerais (UFMG), which together make up the Rede Genoma de Minas Gerais network, provided the sequences. The reactions were made in a single pool and divided into tubes for the PCR sequencing

reaction. After the PCR sequencing reaction, the sequences were joined again in the same tube, mixed, and then divided on three 96-well sequencing plates. Each plate was run three times on a MegaBACE sequencing equipment, yielding a total of 864 reads. Eight hundred and forty-six processed ESD files were obtained.

Base calling

All ESD files were processed by PHRED, without trimming, and a total of 840,134 bases were called.

Local alignment against the pUC18 published sequence

All the sequences generated were compared to the published pUC18 sequence (24.8% A, 25.2% C, 25.5% G, 24.5% T) using the local alignment algorithm SWAT (Smith and Waterman, 1981). Parser scripts written in PERL were built to populate MySQL tables with the position of errors in the reads, identified through the differences in the alignment results. The SWAT algorithm was run with the DNA matrix mat70, and 156,301 bases were removed from the analysis, since they did not show valid alignment to the pUC18 published sequence. The number of bases removed was similar to what was obtained with a PHRED trimming procedure using a trim cutoff parameter of 0.16 (data not shown).

Window-based analysis

RPW and PPW were defined for different window lengths, in order that they could be used in various applications. Table 1 lists the applications and their respective default window length. The PPW were compared to the RPW ones to identify which PQV cutoff (from 5 up to 15) should be used to mask the majority of the errors without masking (and then spoiling) the correct windows.

Table 1. Window sizes, which were analyzed, and related applications.

Window (number of bases)	Acronym	Application
1	Win1	Base calling
6	Win6	Restriction site at vector clipping
11	Win11	BLASTn seeding step
14	Win14	SAGE
28	Win28	MEGABLAST seeding step
40	Win40	Unigene clustering

Error-main weighted analysis

Some researchers might choose to preferentially mask the real errors, even if this is coupled with undesirable masking of correct windows (spoiled windows). Taking this point into consideration, an index called weighted correctness (WC) was created. There are two types of incorrectly classified windows: the ones containing errors that were not masked (not

masked windows, NMW) and the ones with no errors but which were masked because all their bases were under a certain PQV cutoff (spoiled windows, SW). WC will relate and weight NMW and SW according to the researcher's choice. Considering PSW as the percentage of SW divided by the total percentage of windows classified as wrong and PNW as the percentage of NMW divided by the total percentage of windows classified as correct, we can calculate WC as indicated below. The WC value is therefore a measure of the number of errors (NMW and SW) with weights associated with each type of error (Weight1 for NMW and Weight2 for SW).

$$WC = 100 - \left(\frac{(Weight1 \times PNW) + (Weight2 \times PSW)}{PNW + PSW} \right)$$

RESULTS

Data characterization

Analyzing Figure 1, it is possible to see that some PQV are preferred over others during base calling by PHRED, and the number of called bases for each PQV does not show a clear decay as the quality value increases.

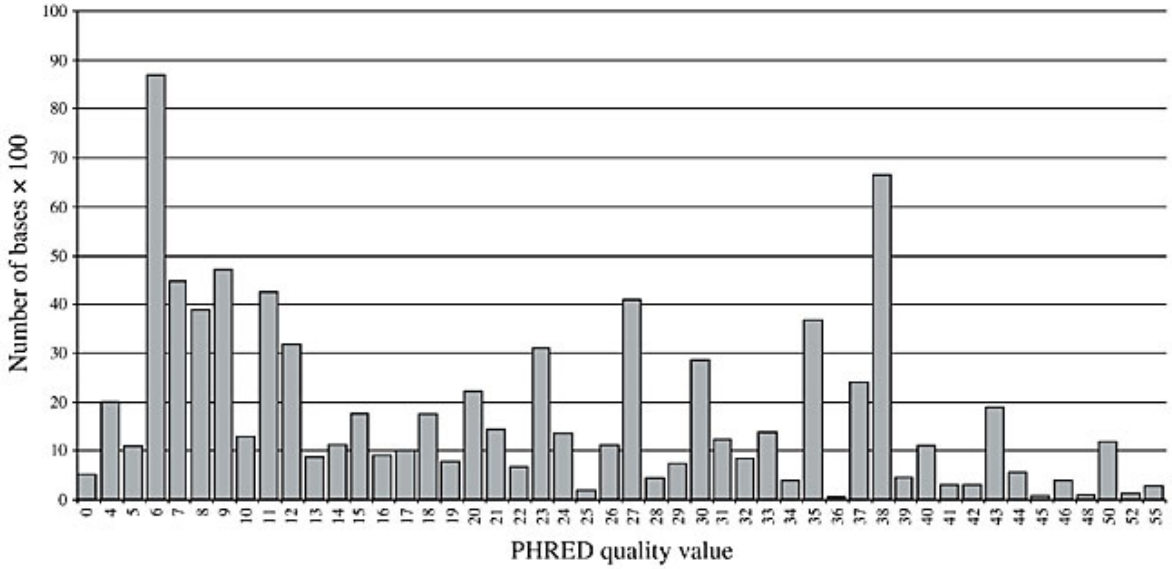


Figure 1. Number of bases called under each PHRED quality value.

The percentage of RPW, the ones with no sequencing errors when aligning the reads to the pUC18 published sequence, was counted for each window size (Table 2). The inverse correlation between the window size and the number of RPWs was expected, since a larger window is likely to shelter more errors than a smaller one.

Table 2. Proportion of real perfect windows (RPW) for each window size.

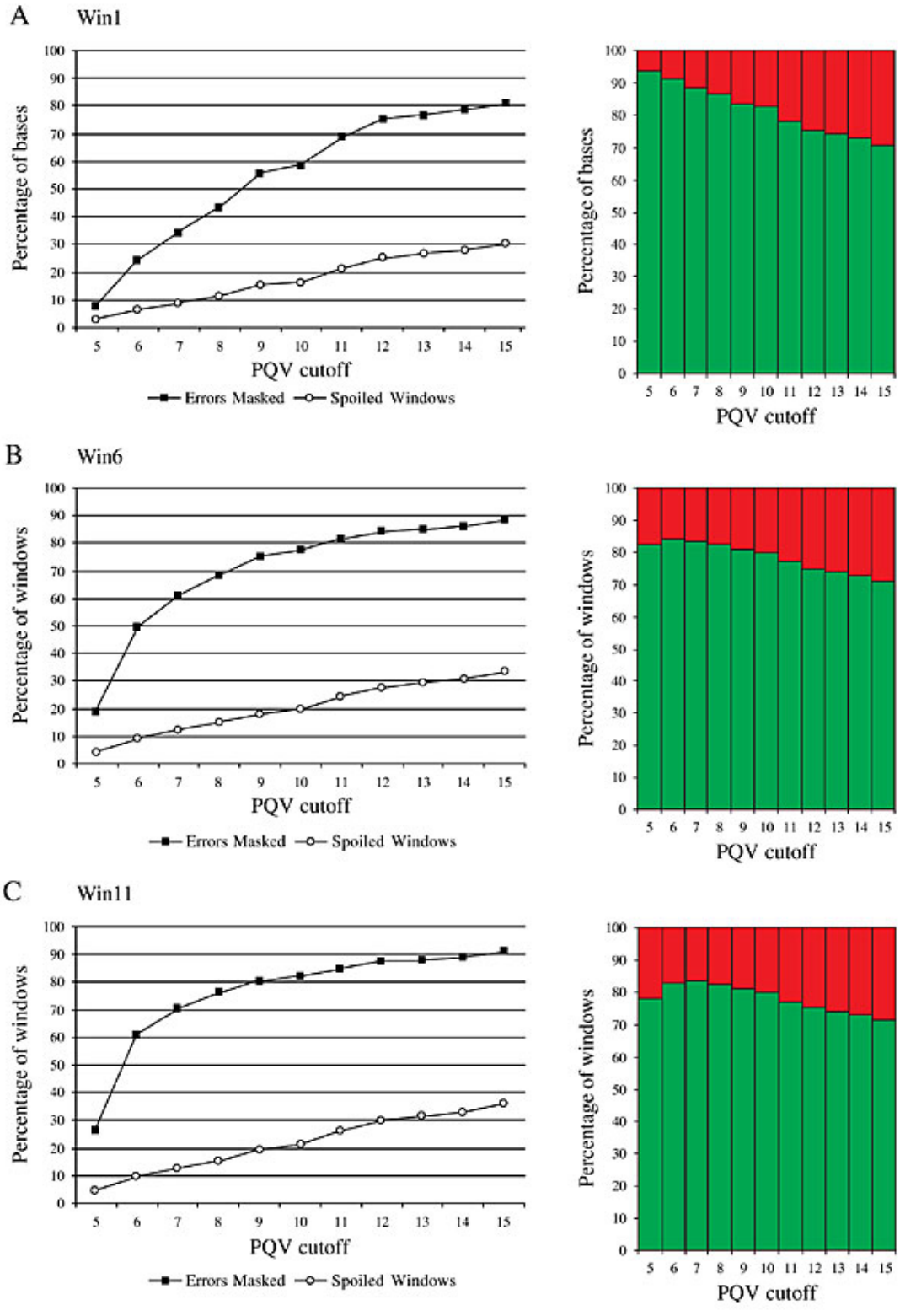
Window	% RPW
Win1	96
Win6	82
Win11	74
Win14	71
Win28	61
Win40	56

PQV cutoff for window sizes

Considering the nature of the PQV, it becomes clear that, when raising the value of the PHRED quality cutoff by which bases are represented with lower case, an increasing number of sequencing windows will be masked, correctly or incorrectly. Here we defined windows containing at least one incorrect base as "errors masked" and windows containing only correct bases that also contain at least one PQV under the cutoff as "spoiled windows". Figure 2 shows the correlation between errors masked and spoiled windows when increasing the PQV cutoff. The graphs on the left side show the percentage of "errors masked" in filled squares and the percentage of "spoiled windows" in empty circles. On the right side of figure we also show the percentage of total window classification that was correct (green) or incorrect (red).

Our main purpose was to find the most adequate PQV cutoff that maximizes the number of windows that do contain errors and are masked by the lowest PQV (errors masked), while minimizing the number of correct windows incorrectly masked (spoiled windows). This PQV cutoff may be used to define unreliable lower case-containing windows. It is clear from Figure 2 that the percentage of errors that are masked tends to saturate, while the percentage of spoiled windows continues to raise. Moreover, it is possible to choose from the data plot a given PQV cutoff for the lower case representation that will mask over 80% of the windows containing errors without spoiling more than 20% of the correct windows. Furthermore, this relationship depends on the size of the working window (e.g., for the Win40, a PQV cutoff of 8 will mask 90% of error-containing windows while avoiding 20% of correct windows to be used by the application).

Another way of looking at the same data is the balance between correct and incorrect classification. Incorrect classification involves a RPW classified as wrong or an error-containing window classified as a PPW. These classifications are represented on the right side of Figure 2.


A
Win1
Percentage of bases
PQV cutoff
Errors Masked
Spoiled Windows
B
Win6
Percentage of windows
PQV cutoff
Errors Masked
Spoiled Windows
C
Win11
Percentage of windows
PQV cutoff
Errors Masked
Spoiled Windows

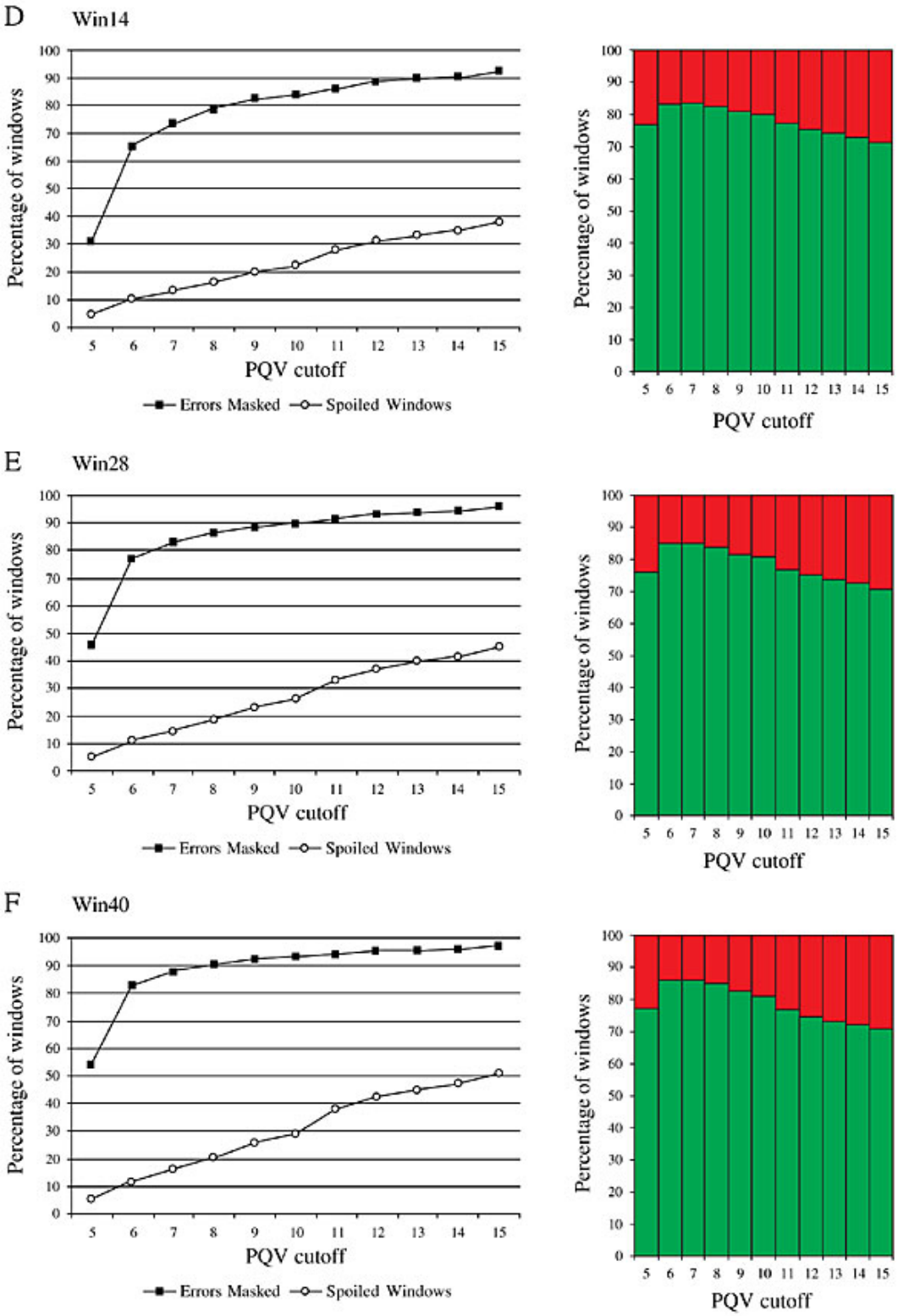


Figure 2. Percentage of errors masked versus spoiled windows (left) and percentage of correct (green) and incorrect (red) window classification (right) for Win1, 6, 11, 14, 28, and 40. PQV = PHRED quality value.

In general, a PQV cutoff of 6 or 7 maximizes the identification of errors, without spoiling a large proportion of correct windows. This best correlation is clear if one looks at the high inclination of the errors masked curve when the PQV cutoff increases from 5 to 6 and from 6 to 7. These inclinations are higher than all other increases of PQV cutoff, and they are much more accentuated than the ones observed for the spoiled window curves at the same cutoffs. However, in general, the PQV cutoff of 6 or 7 still gives a window classification error percentage close to 15% for windows larger than win1.

Error main weighted analysis

All the results for the last section were shown considering that an error that is not masked has the same importance (50-50%) as a correct window that has been incorrectly masked. However, one could argue that it is more relevant to mask the real errors, even if this has been coupled with a high number of spoiled correct windows. Thus, a weighted index WC was developed to obtain the data for many distinct relationships between the weights attributed to NMW and SW (see Methodology).

Data in Figure 3A is similar to the data shown in Figure 2 (green-red graphs). By analyzing the other plots (B, C and D), it is possible to see that the best value of PQV cutoff has been shifted to the right, as expected, when increasing the percentage of priority in error masking. In this way, the best quality cutoff tends to be higher than 7.

DISCUSSION

PHRED software is the most widely used base caller in the genomics field, and its use has been extensively evaluated (Ewing and Green, 1998; Ewing et al., 1998; Richterich, 1998; Walther et al., 2001; Scheetz et al., 2003). Besides being well known that the lowest PQV does not necessarily stand for miscalled bases, the question of whether or not the lowest PQV cohabit in the same sequencing window with the sequencing errors has not been addressed. Our main purpose was to map the miscalled positions and the RPWs on sequencing reads based on a PQV cutoff. With a specific cutoff, e.g., PHRED 15, we decided to represent putatively miscalled bases with lower case letters, so that those with quality values less than or equal to 15, would be considered as unreliable. The advantage of translating bases to lower case letters is that, with the same number of bytes used to store standard sequences, we could add relevant information in an easy-to-see fashion. It would be particularly useful for the publication of single-passed sequences, such as ESTs, GSSs and SAGE tags. Currently, some researchers apply this procedure, but no study has been conducted to evaluate the appropriate cutoff.

Contrary to what was expected by intuition, the PQV cutoff was found to be more effective in masking errors without spoiling PWs at a quality value of 6 to 7. Based on the data that were collected, we would recommend the lower case masking of windows containing at least one base with a quality value of 7 or less, for most of the applications exemplified. However, if the researcher prefers to correctly mask all (or almost all) errors, even with a great chance of increasing the number of spoiled windows, he should use a higher cutoff value, varying from 9 up to 15 (see Figure 3 for a better understanding of this). Although these recommendations are valid for all the data, it is useful to inspect Figure 2, trying to fit

the best PQV cutoff for a particular application, or for the most probable destination of the output.

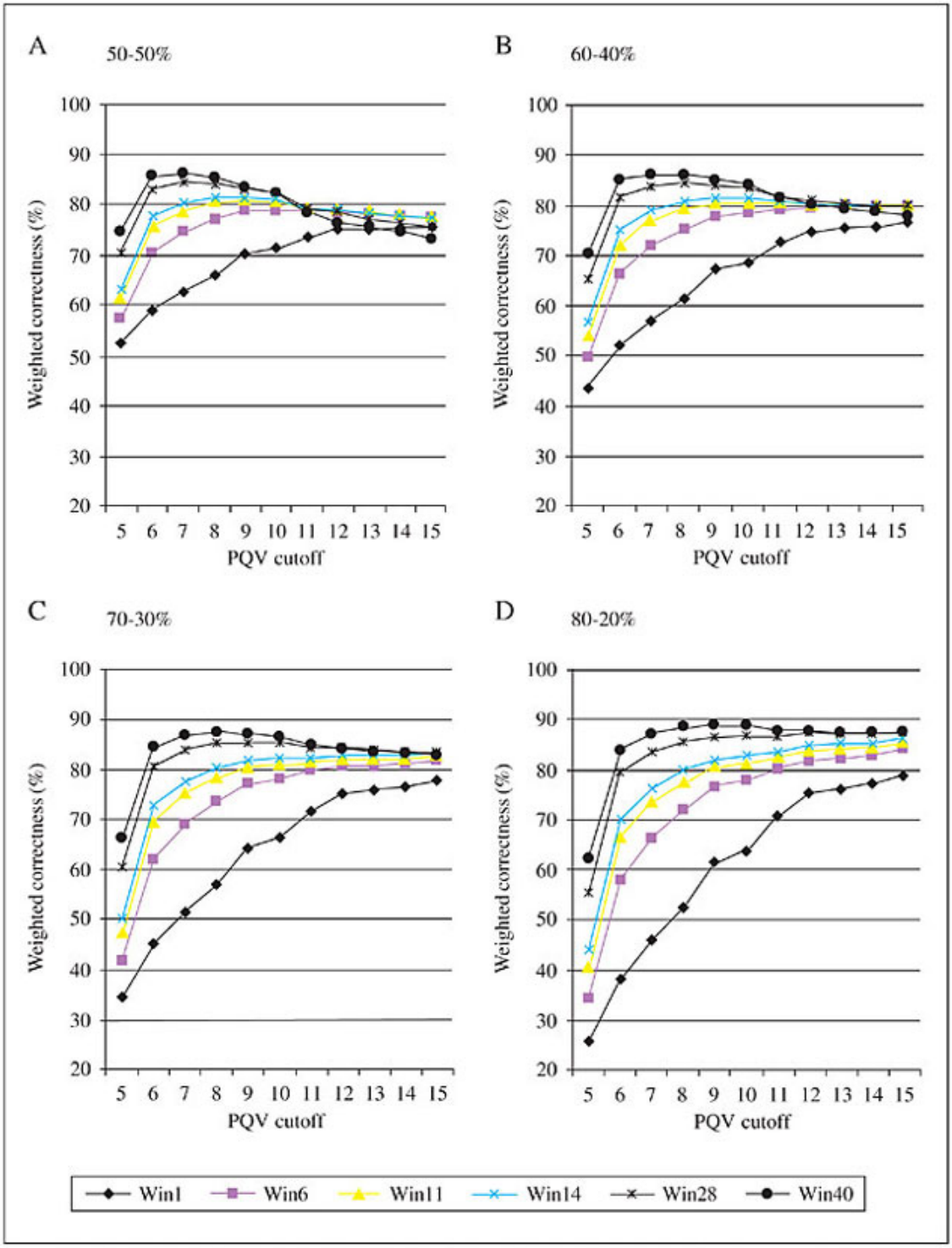


Figure 3. Distinct weights given to not masked windows and spoiled windows. The percentages in the graphs (Weight1-Weight2) indicate the relative importance of error masking as compared to the spoiling of correct windows (see Methodology). PQV = PHRED quality value.

Therefore, we conclude that the use of PQV cutoff masking frequently allows more than 85% of correct identification of windows. We have provided the data necessary for a sequencing research group to balance the number of the PQV cutoff if they preferentially desire either to mask errors or to allow for more correct windows. By examining Figure 3, it is possible to choose the best PQV cutoff for a specific application. We exclusively used sequences from the plasmid pUC18; similar studies using other templates and different sequencing machines, as well as other sequencing substrates (such as PCR or RT-PCR products), are necessary to determine if analogous results will be found.

ACKNOWLEDGMENTS

The authors thank the Rede Genoma de Minas Gerais (supported by FAPEMIG and MCT/Brazil), especially Marina Mourao, Lucila Pacifico and Renata Ribeiro, for the sequences and CENAPAD-MG/CO for the equipment used in the present study.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and **Lipman, D.J.** (1997). Gapped BLAST, PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*: 3389-3402.

Ewing, B. and **Green, P.** (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res. 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 Res. 8*: 175-185.

Green, P. (2004). PHRED Documentation. http://www.phrap.org/phrap.docs/phred.html. Acessed August 29, 2004.

Prosdocimi, F., Cerqueira, G.C., Binneck, E., Silva, A.F., Reis, A.N., Junqueira, A.C.M., Santos, A.C.F., Nhani-Júnior, A., Wust, C.I., Camargo-Filho, F., Kessedjian, J.L., Petretski, J.H., Camargo, L.P., Ferreira, R.G.M., Lima, R.P., Pereira, R.M., Jardim, S., Sampaio, V.S. and **Folgueras-Flatschart, A.V.** (2002). Bioinformática: Manual do usuário (in Portuguese). *Biotec. Cienc. Des. 29*: 18-31.

Prosdocimi, F., Peixoto, F.C. and **Ortega, J.M.** (2003). DNA sequences base calling by PHRED: Error pattern analysis. *R. Tecnol. Inf. 3*: 107-110.

README for stand-alone BLAST (2004). ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blast.txt. Acessed August 29, 2004.

Richterich, P. (1998). Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res. 8*: 251-259.

Scheetz, T.E., Trivedi, N., Roberts, C.A., Kucaba, T., Berger, B., Robinson, N.L., Birkett, C.L., Gavin, A.J., O'Leary, B., Braun, T.A., Bonaldo, M.F., Robinson, J.P., Sheffield, V.C., Soares, M.B. and **Casavant, T.L.** (2003). ESTprep: preprocessing cDNA sequence reads. *Bioinformatics 19*: 1318-1324.

Smith, T.F. and **Waterman, M.S.** (1981). Identification of common molecular subsequences. *J. Mol. Biol. 147*: 195-197.

Velculescu, V.E., Zhang, L., Vogelstein, B. and **Kinzler, K.W.** (1995). Serial analysis of gene expression. *Science 270*: 484-487.

Walther, D., Bartha, G. and **Morris, M.** (2001). Basecalling with life trace. *Genome Res. 11*: 875-888.

Avaliação da posição ótima do *primer* de sequenciamento com relação ao inserto

A seguir, realizamos um estudo com o objetivo de identificar a posição ideal para se posicionar um iniciador (*primer*) de sequenciamento quando se deseja obter seqüências de um determinado inserto. O artigo "Accessing optimal primer distance from insert" foi publicado na revista In Silico Biology.

O artigo trata do fato de que, ao iniciar um projeto genoma, um pesquisador precisa considerar a distância de seu iniciador utilizado para o sequenciamento até a posição onde será clonado o inserto do organismo de interesse. A colocação incorreta deste iniciador pode gerar tanto o sequenciamento excessivo de bases do vetor de clonagem – que não interessam durante o projeto – quanto a perda de elementos importantes da seqüência do inserto, como a seqüência poli(A) ou o códon de iniciação do gene seqüenciado, problemas particularmente relevantes no caso de projetos de sequenciamento de cDNA.

Nossos resultados mostraram que as leituras de sequenciamento contêm aproximadamente 0-20 bases aleatórias e mal nomeadas pelo algoritmo PHRED no início da seqüência e que a boa leitura normalmente se inicia em média 46-54 bases a partir da primeira base a 3' do iniciador. Utilizando abordagens baseadas no algoritmo SWAT, verificamos que 60 bases é uma distância na qual mais de 90% das seqüências exibem informação confiável, apresentando aproximadamente 13 bases de seqüência de vetor, em média. Além disso, os dados mostrados no artigo permitem que um pesquisador possa escolher onde clonar seu inserto ou a qual distância produzir seu iniciador de sequenciamento dependendo de sua aplicação de interesse.

In Silico Biology 5, 0043 (2005); ©2005, Bioinformation Systems e.V.

Accessing optimal primer distance from insert

Francisco Prosdocimi[1] and J. Miguel Ortega[2,*]

[1] Departamento de Biologia Geral, ICB-UFMG, 31270-010, Belo Horizonte/MG, Brazil
Email: franc@icb.ufmg.br
[2] Departamento de Bioquímica e Imunologia, ICB-UFMG, 31270-010, Belo Horizonte/MG, Brazil
Email: miguel@icb.ufmg.br

* Corresponding author

Edited by H. Michael; received June 24, 2005; revised and accepted September 05, 2005; published September 24, 2005

Abstract

When building either DNA or cDNA libraries, a researcher looks at the vector multiple cloning site and chooses which restriction enzyme(s) will be used to clone the inserts. Although this procedure does not seem to be important, the accurate choosing of primer to insert distance can save time and money from genome and transcriptome projects. Here, 846 single-pool pUC18 sequences were produced and compared with the pUC18 consensus using local alignment tools. Data show that reads often contain 0-20 miscalled bases at the beginning of read and noise to signal transition is frequently found at 46-54 bases from the first 3' base downstream the sequencing primer. For SWAT-based approaches, 60 bases was the distance where over 90% of the sequences provided reliable information, presenting 13 vector bases on average. Looking at the data, it is possible to choose the most appropriate primer to insert distance for many applications.

Keywords: pUC18, cloning vector, Smith-Waterman, BLAST, primer positioning, insert, MegaBACE sequencer

Introduction

One of the first steps when a genome or transcriptome project begins is the choice of a cloning vector where the sequences of the organism of interest will be cloned [Preston, 2003]. The map of the vector is analyzed and a position of cloning is chosen taking into account many factors, such as the availability of the restriction enzymes, the cloning directionality and, sometimes, the restriction site distance from primer. But even taking into account this distance, there was not, until now, a clear effort to evaluate what is the best distance to be chosen. This distance is relevant since it affects the actual number of bases sequenced by the project. If the primer is located too far from the beginning of the insert, a great number of non-desired vector nucleotides will be generated on each of the sequences produced, wasting money and time as well as the relevant organisms' data. On the other hand, if the primer is too close to the cloning site, the beginning of the sequence will be frequently lost. For 3' reads of cDNA sequencing projects, the poly(A) tail will often be skipped, since the read will start too far inside the insert. The presence of the poly(A) tail is considered to be an important evidence for the expression of the sequence and, for instance, UniGene entries require that at least one member of a cluster contains the poly(A) tail [Pontius et al., 2003; Schuler et al., 1996]. Otherwise, for the 5' reads, the beginning of a coding sequence (CDS, identified by the ATG start codon) might also be skipped, and this information is relevant to choose the clones that will be used for full-length sequencing, such as done for the Mammalian Genome Collection project [Strausberg et al., 1999]. Therefore, to maximize the insert region that is sequenced, without getting too much vector sequence as well as sequencing the insert from its beginning, it is required to map the position (distance from the primer) where the base calling switches from noise to signal. Moreover, all sequencing applications would benefit with a higher number of relevant bases being included in the process. When dealing with single passed sequences, a greater number of nucleotides would be produced as less vector sequences are obtained without lacking the 5' extremity. When dealing with complete genome assembly, lesser sequences would be necessary to complete a desired genomic coverage since the reads would present more organisms' bases.

One possibility to determine the transition from noise to signal in sequence reads is to align experimental reads to published sequences. Since the complete nucleotide sequence of the cloning vector pUC18 is known (GenBank accession number L09136), we have sequenced many molecules of this vector. So, the reads generated were compared to the literature consensus using different local alignment approaches. A local alignment tool only begins the alignment between two input sequences - in case, the sequencing read produced and the literature consensus - when they present a clear region of similarity; it does not start the alignment when the sequences present different nucleotide patterns. The BLAST software [Altschul et al., 1990] is the most used local alignment algorithm and it has the advantage to be optimized in producing fast and consistent results, mainly when using the great amount of data present in public molecular databases. However, BLAST does not necessarily return the best

result possible, since it uses some heuristic mechanisms to accelerate the search. In order to obtain optimal results in the biomolecules alignment, the Smith-Waterman algorithm (SWAT) is frequently used [Smith and Waterman, 1981]. This algorithm performs the best possible computational alignment between two DNA sequences, despite requiring more time than BLAST. Here, all the 846 reads produced were compared to the vector sequence consensus using either SWAT, BLAST or Cross-match - the most used software to find vector regions on sequencing reads. Using these three programs we stored in a MySQL database the positions where each alignment began. The results show a variation in the sequence beginning but a researcher, analyzing the data, is capable to choose a suitable position where placing the sequencing primer to make the best of the sequencing read.

Methods

Single-pool sequence reaction

The sequencing reactions were made in a single pool and divided into tubes for the PCR amplification, such as described elsewhere [Prosdocimi et al., 2004]. After the PCR, the samples were joined again in the same tube, mixed, and then divided on three 96-well sequencing plates. Each plate was run three times on a MegaBACE sequencing equipment and 846 processed ESD files containing pUC18 reads were obtained.

Base-calling

All the reads were base-called by PHRED algorithm without trimming parameters.

Local alignments

Before doing the alignments, the pUC18 consensus sequence (L09136) was modified by changing the order of the nucleotides. The first nucleotide of the sequence used in the alignment was changed to be first one at 3' of the primer used in the sequencing procedure. Three different alignments were produced, BLASTn was run with default parameters and SWAT was run with two different widely used nucleotide comparison matrices, mat50 and mat70. From now on, the results obtained with SWAT with mat50 matrix will be defined as SWAT50; and SWAT70 when using mat70 matrix. Parser scripts written in PERL were developed to retrieve the information from the alignment tools output files and to populate MySQL tables.

Sequence masking

A secondary approach was developed using Cross-match software. Cross-match is the most used software for vector masking on DNA base-called sequences. It implements a SWAT algorithm inside its code (and, therefore, can be used as a SWAT-based local

alignment tool). The Cross-match software was run with the vector masking parameters given in its manual (-minmatch 10 -minscore 20 -screen). Once more, PERL parsers were developed to retrieve alignment information and populate MySQL database tables.

Data building

The parser script used in the analyses was developed according to the information shown in Fig. 1. First, three conceptual positions were defined: (1) the Polymerization Starting Position (PSP), which consists of the first base polymerized at 3' of the primer; (2) the Base-calling Starting Position (BcSP), consisting in the site where the read begins; and (3) the Alignment Starting Position (ASP), defined as the site on which the read was identified as pUC18 by the sequence alignment algorithms.

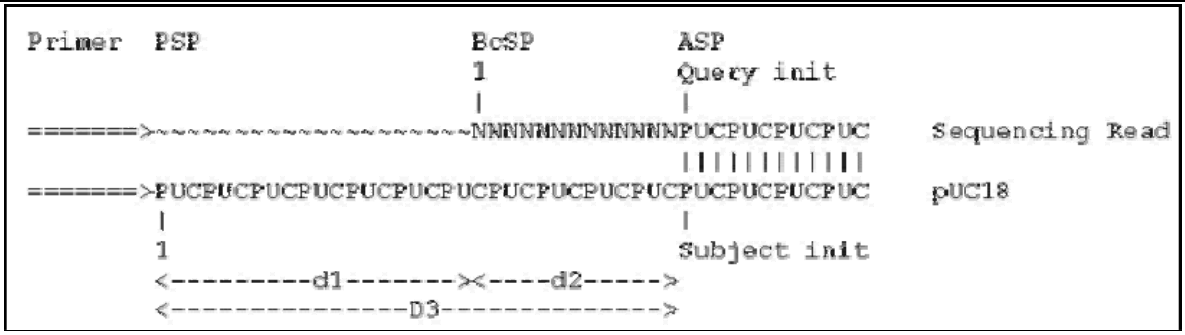


Figure 1: Positions and distances definitions. PSP (Polymerization Starting Position) is defined as the position of the very first base polymerized at 3' of the primer; BcSP (Base-calling Starting Position) consists in the site where the read begins; and ASP (Alignment Starting Position) represents the nucleotide position on which the read was identified as pUC18 by the sequence alignment algorithms.

We defined also some distances such as d1, the distance between PSP and BcSP; and d2, the distance between the BcSP and ASP. Distance d2 was retrieved from the output software data as the Query init position minus one nucleotide (for all alignments done, the pUC18 reads were used as query and the pUC18 consensus sequence was used as subject). Moreover, D3 was defined as the distance between PSP and ASP, retrieved from Subject init position given by the alignment software.

Estimation of the number of cloning vector bases sequenced upon different primer-insert distances

In order to find the number of cloning vector bases sequenced by a hypothetical project, a simulation was conducted supposing the vector to insert transition at increasing distances from the sequencing primer. Therefore, d4 position was defined as the distance between PSP and the nearest base corresponding to the insert. Distance d4 was simulated from 1 to 120 nucleotides, and the average number of vector bases identified after ASP was counted for each position. When D3 resulted to be larger than

d4, the number of cloning vector bases sequenced was considered equal to zero. Otherwise, the number of vector bases identified was given by the difference between d4 and D3, summed and averaged for all 846 sequences analyzed.

Results

We analyzed 846 sequences of 993 base pairs in average, obtained from a single-pool sequencing reaction. We take the advantage of a single-pool reaction in order to make the difference amongst the sequences caused only by the process of capillary electrophoresis and base calling. Frequently the sequence beginning presents a reading extent consisting of randomly called bases. This extent is represented by the distance d2 (Fig. 1) and Table 1 shows, for all the algorithms run, the average size of d2 and the modal class - the most frequent value attained and the number of reads in each class. For cross-match, the values represent the number of bases not masked with X in the beginning of the sequence, which the software does not recognize as pUC18. Moreover, Fig. 2 presents the distribution of d2 size by classes of 10 nucleotides using each of the alignment software. The extent of inefficiently called bases varies upon a pattern that cannot be predicted, often lower than 20 but some with up to 100 miscalled bases before the starting of the reliable base calling.

Table 1: Average and modal size of d2 distance.

Algorithm	Average size	Modal class (number of reads in class)
BLAST	23	16 (33)
SWAT50	16	1 (62)
SWAT70	18	1(56)
Cross-match	16	1 (65)

To evaluate the optimal distance between the primer and the transition insert/sequence of interest, distance D3 (Fig. 1) was determined for each read. Table 2 presents for D3 the same information shown in table one for d2 and also the pUC positions where 10%, 50% or 90% of the reads were identified as efficiently base called by the alignment algorithms.

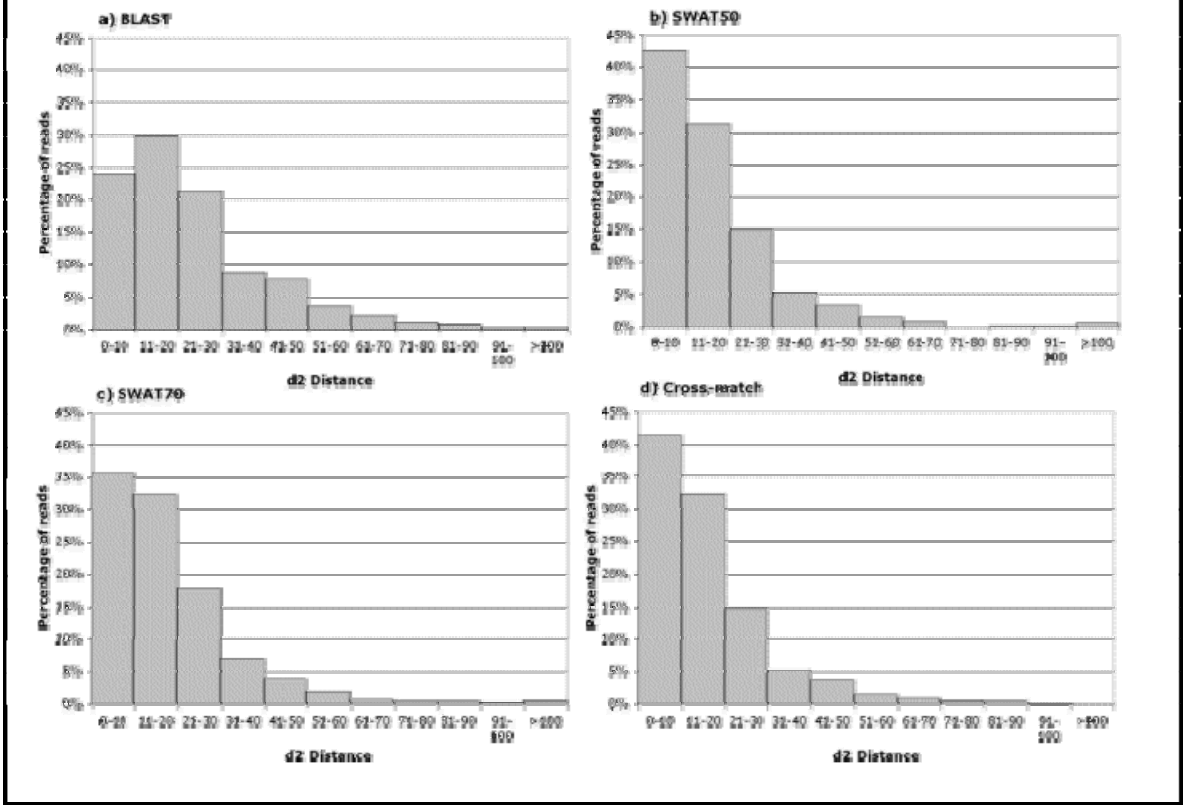


Figure 2: Percentage of reads with distinct values for d2 distance.

Table 2: Determined D3 distance.

Algorithm	Average size	Modal class (Number of reads in class)	pUC Position[a]		
			10% of reads	50% of reads	90% of reads
BLAST	46	48 (180)	41	50	91
SWAT50	54	48 (162)	37	49	60
SWAT70	52	48 (165)	38	49	60
Cross-match	53	48 (144)	39	49	60

[a] Position in pUC, counted from PSP, where the indicated percentage of reads (at least) have reached ASP (see Fig. 1 for definitions).

D3 data can be seen in Fig. 3. The *y*-axis shows the cumulative percentage of sequences that reach ASP, and a dashed line representing the average BcSP for all alignment software was also added. The BcSP positions had shown minimal

discrepancies when using different algorithms (since the base-calling was done only once using PHRED) and they were averaged in a single curve on Fig. 3. Except for BLAST algorithm, there seems to be a tendency of the reliable part of the read (ASP) to start closer to BcSP as the value of this variable rises. Thus, early started reads tend to harbor a larger extend of miscalled bases. Moreover, the distance D3 varies in a way that it can be roughly predicted, between 38 and 60 bases from PSP, around 50 bases (Table 2).

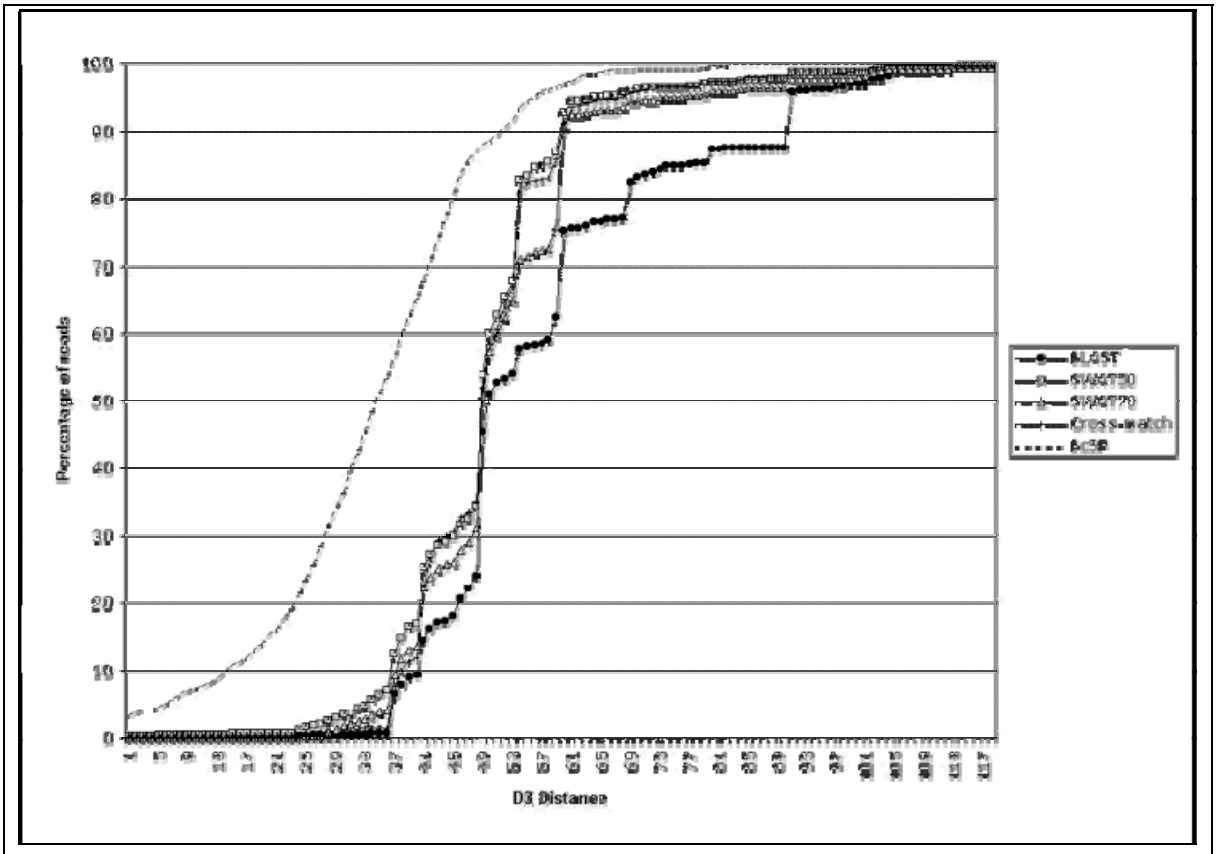


Figure 3: Percentage of sequences reaching ASP using different software. The dashed line represents the average of BcSP.

Additionally, we investigated the relationship between distances d1 and d2 trying to find some correlation between them within the same reads. So, Fig. 4 contains the tuple d1-d2 determined for each read and shows that a high d1 value is often associated with a low d2 value; and vice-versa. This reinforces the indication that the base-called reads that start closer to the primer will present more miscalled bases at their 5' edge. BLAST algorithm, from data in Fig. 4, leaded to similar results than SWAT-based ones, although inspection of Fig. 3 had seemed to suggest that d2 distance could be constant. Data in Fig. 4a shows that BLAST yields an additional series of points consisted of higher D3 distance, although the same inverted relationship between d1 and d2 is verified.

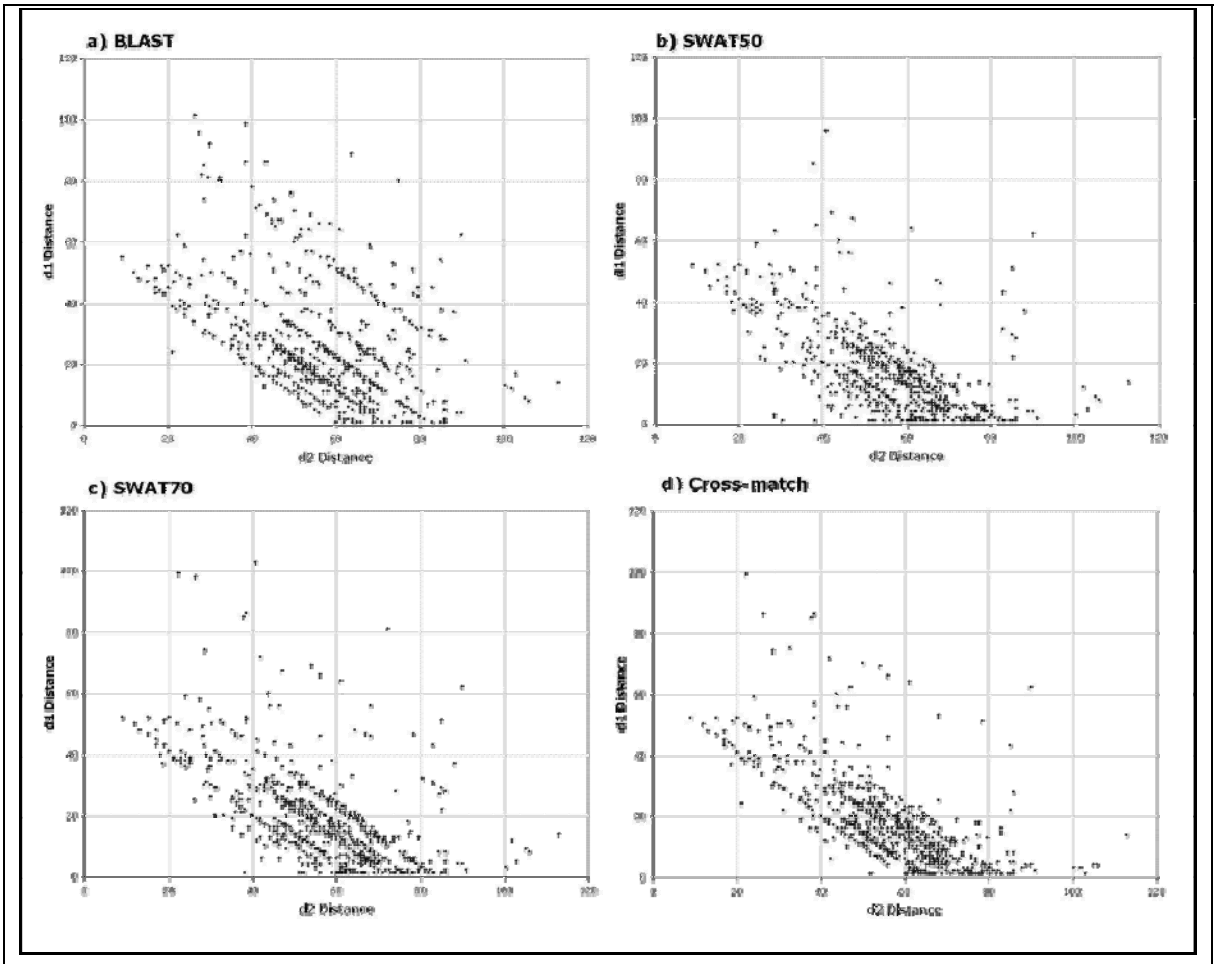


Figure 4: Relationship between distances d1 and d2 inside single reads.

Finally, a simulation has been conducted to estimate the number of cloning vector bases actually sequenced considering the insert cloned at different distances from PSP (d4). Data in Fig. 5 shows that the farthest the distance primer-insert, the higher the number of cloning vector bases attained. Two points were emphasized indicating the positions where over 90% of sequences can be recognized. For SWAT-based approaches, sequences bear 12 vector bases in average at this point, while using BLAST, the estimative is of 35 bases per sequence. Therefore, the result of the simulation as plotted in Fig. 5 may advise researchers to estimate the average number of cloning vector bases that would aggregate in the generated sequences. Data presented in Fig. 5, together with the analysis of the percentage of sequences where all transitions vector-insert will be represented in the reliable portion of read (Fig. 3), it is possible to balance these two parameters and define the optimal primer distance from insert.

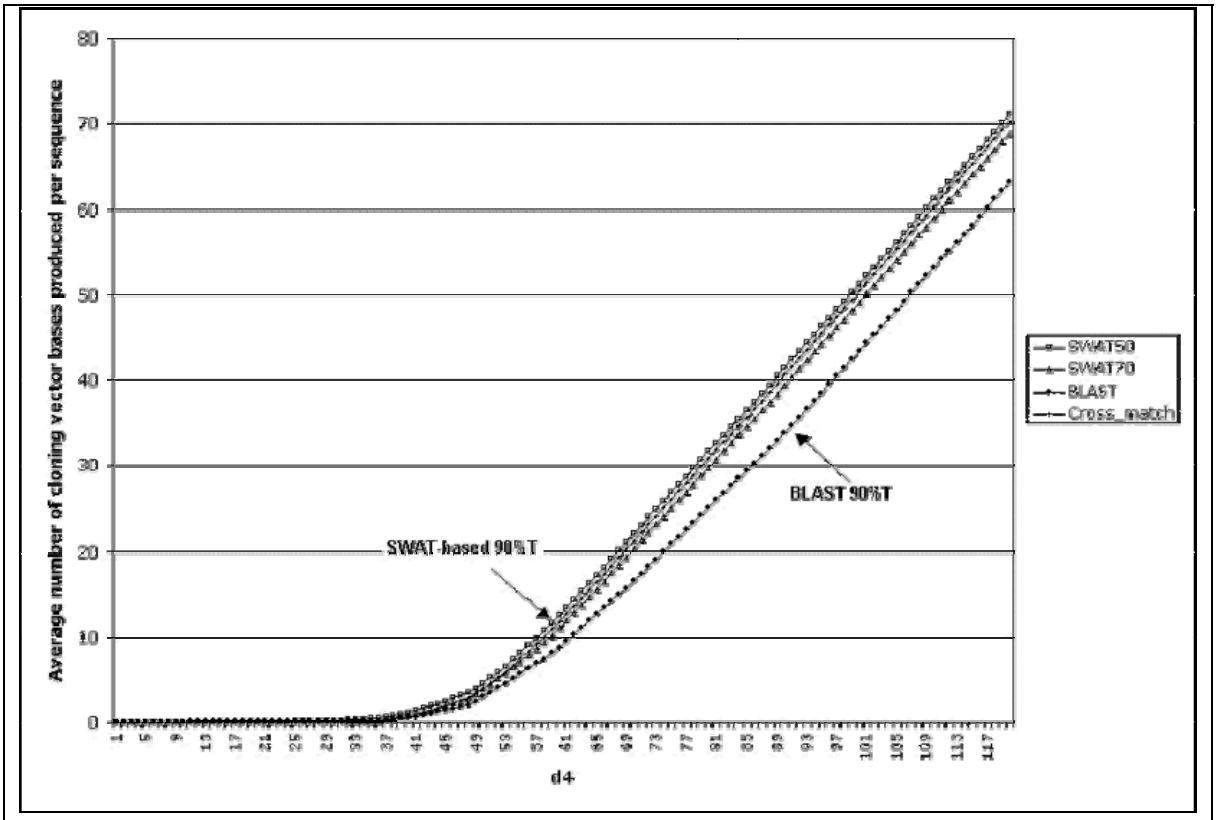


Figure 5: Simulation on the number of cloning vector bases produced per sequence when different insert positions were tested.

Discussion

Here, for the first time, an experiment methodologically well designed has been produced to evaluate the optimal distance where the insert should be cloned from the sequencing primer. Analyzing Table 1 it is possible to conclude that approximately the first 20 bases of a sequence base-called by PHRED (with default parameters) consist of spurious nucleotides. However, often reads present informative region from their very first base, given that the most common value for recognition by SWAT-based software was the first nucleotide. Even considering this data, these first 20 nucleotides should be considered with special care on MegaBACE sequencing reads. Complementing this analysis, it is possible to conclude (Fig. 2) that few sequences present more than 50 spurious nucleotides at the read beginning. Moreover, the broad distribution of this parameter makes difficult to determine the more proper positioning for the sequencing primer.

The best approximation of the desirable primer positioning seems to be reached by determining the distance from the PSP, where polymerization starts, up to the position where the reliable read starts (ASP), as determined by alignment to the known pUC18

sequence. Table 2 suggests that around 50 bases is the average number of bases neither called nor miscalled, as counted from the first position after the primer 3' edge, for all alignment software. The most common D3 distance (modal class), was identical when dealing with any of the algorithms and similarity matrices, and it equals 48 nucleotides. Thus, it is expect from our data that informative reads will start, in average, over 50 bases after the sequencing primer.

However, besides a value around 46-54 nucleotides comprises the modal and average value for the desirable distance (Table 2), from the 846 sequences analyzed, the algorithms BLAST, SWAT50, SWAT70 and Cross-match recognized, respectively, only 24%, 35%, 31% and 35% of the pUC18 sequences as correct, up to position 48 (Fig. 3). Therefore, the option to choose a position where more than 90% of sequences have reached ASP seems to be a good criterion. At this range, BLAST and SWAT produced slightly different results (Fig. 3), being the three SWAT-based approaches producing very similar data. Considering the heuristic nature of the BLAST algorithm and the optimal characteristic of the SWAT alignments, it would be more suitable to take on account preferentially SWAT data. Therefore, inspection of Fig. 3 shows that the number of nucleotides that gives more than 90% of sequences recognized is 60 for all SWAT-based approaches.

It has been also shown that sequences beginning closest to the primer (with low d1 distance) frequently present a larger extent of miscalled bases than reads beginning farther from the primer. Looking at Fig. 4, it is possible to see that d2 distance conversely decreases as d1 distance increases, keeping the D3 nearly constant.

Therefore, a researcher may use the data on Fig. 3 to choose other value but 60 that he/she thinks to be more adequate to place its primer from the cloning site, but this paper shows that this distance value should be somewhat between 20 and 110 nucleotides. If a researcher is interested on the beginning of the molecule, a distance about 60 nucleotides from primer to insert seems to be enough to see the great majority (~90%) of primer-insert transitions, producing an average number of 13 crossmatch-able cloning vector bases per sequence (Fig. 5). Otherwise, if the researcher is sequencing a genome, maybe his/her interest is only to sequence just inserts that will be overlapped later using clustering algorithms and, then, he/she might adjust the distance primer-insert as the minimum possible, avoiding the sequencing of any vector regions.

Further experiments, under the approach suggested, should be done to confirm the results obtained here for other DNA sequences and sequencer machines.

Acknowledgements

The authors thank the "Rede Genoma de Minas Gerais" (supported by FAPEMIG and MCT/Brazil), especially Marina Mourao, Lucila Pacifico and Renata Ribeiro, for providing the sequences used in the analysis. We also thank Maurício Mudado for critical reading of the manuscript.

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403-410.
- Pontius, J. U., Wagner, L. and Schuler, G. D. (2003). UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information.
- Preston, A. (2003). Choosing a cloning vector. Methods Mol. Biol. 235, 19-26.
- Prosdocimi, F., Peixoto, F. C. and Ortega, J. M. (2004). Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values. Genet. Mol. Res. 3, 483-492.
- Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B. B., Butler, A., Castle, A. B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Edwards, C., Fan, J. B., Fang, N., Fizames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hui, L., Hussain, S., Louis-Dit-Sully, C., Ma, J., MacGilvery, A., Mader, C., Maratukulam, A., Matise, T. C., McKusick, K. B., Morissette, J., Mungall, A., Muselet, D., Nusbaum, H. C., Page, D. C., Peck, A., Perkins, S., Piercy, M., Qin, F., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, C., She, X., Silva, J., Slonim, D. K., Soderlund, C., Sun, W.L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wilmer, T., Wu, X., Adams, M. D., Auffray, C., Walter, N. A., Brandon, R., Dehejia, A., Goodfellow, P. N., Houlgatte, R., Hudson Jr, J. R., Ide, S. E., Iorio, K. R., Lee, W. Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M. H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J. C., Sikela, J. M., Beckmann, J. S., Weissenbach, J., Myers, R. M., Cox, D. R., James, M. R., Bentley, D., Deloukas, P., Lander, E. S. and Hudson, T. J. (1996). A gene map of the human genome. Science 274, 540-546.
- Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. J. Mol. Biol. 147, 195-197.
- Strausberg, R. L., Feingold, E. A., Klausner, R. D. and Collins, F. S. (1999). The Mammalian Gene Collection. Science 286, 455-457.

5.5. Definição da melhor posição de poda (*trimming*) de seqüências com o objetivo de obter o máximo de informação biológica

Devido a diversos fatores, como as diferenças na mobilidade eletroforética de segmentos de DNA muito pequenos ou muito grandes (Ewing and Green, 1998), os algoritmos de nomeação de bases são normalmente calibrados para conseguirem identificar com precisão as seqüências de um tamanho intermediário lidas pelo seqüenciador. Durante a leitura dos dados brutos produzidos por moléculas pequenas ou grandes demais, muitas vezes o algoritmo falha durante a definição do tempo entre o aparecimento de um sinal e o aparecimento do próximo (*lane profiling*), nomeando incorretamente, principalmente as bases do início e do fim da seqüência analisada. Isso faz com que a seqüência de nucleotídeos gerada pelo algoritmo de nomeação de bases consista num padrão ruído-sinal-ruído e, portanto, o objetivo do presente trabalho foi tentar definir corretamente as bordas do sinal de forma a separá-lo do ruído e fazer com que o pesquisador venha a obter apenas a parte informativa da seqüência desejada.

Já sabendo disso, os desenvolvedores do algoritmo PHRED produziram um sub-algoritmo interno conhecido como "algoritmo de *trimming*" que permite ao pesquisador ou bioinformata alterar um parâmetro que indica o quanto da região interna da seqüência se deseja obter. Dessa forma, um cientista que deseje uma seqüência da mais alta qualidade possível, deve obter principalmente a região interna do sequenciamento, evitando as extremidades. De forma contrária, ele pode querer obter as pontas, mesmo que estas não representem informações tão fiéis sobre a seqüência de bases da molécula, mas que facilitem sua identificação através da comparação por alinhamento local com outras seqüências nos bancos de dados.

Portanto, durante o procedimento de nomeação das bases realizado aqui, foram utilizados diferentes valores do parâmetro de PHRED conhecido como trim_cutoff. O trim_cutoff está exatamente relacionado ao limite a ser utilizado para considerar o quanto da seqüência interna deve ser considerado como sinal. Assim, variamos esse parâmetro e comparamos as seqüências com a seqüência publicada do pUC18, de forma a identificar qual seria o valor mais adequado do mesmo para a seleção apenas da parte informativa da seqüência, definida como aquela encontrada como hit pelo algoritmo de alinhamento local. Portanto, consideramos "informação biológica" ou "sinal" tudo aquilo que puder ser encontrado por um algoritmo de alinhamento local,

como o BLAST ou o SWAT. Este trabalho, entitulado "Setting PHRED scores to obtain maximum biological information", foi submetido à revista Nucleic Acids Research.

Setting PHRED scores to obtain maximum biological information

Francisco Prosdocimi[1], Fabiano Cruz Peixoto[2], Maurício Mudado[1], J Miguel Ortega[1,*]

[1] Laboratório de Biodados. Depto. Bioquímica e Imunologia, ICB-UFMG, Brazil.

[2] Laboratório de Computação Científica, UFMG

José Miguel Ortega *
miguel@icb.ufmg.br
Laboratório de Biodados. Sala N4-202.
Departamento de Bioquímica e Imunologia, ICB, UFMG
Av. Antônio Carlos, 6627 C.P. 486
31.270-010 Belo Horizonte, MG, Brazil
Tel: +55 31 3499-2654
Fax: +55 31 3499-2570

Running Head
Using SWAT to calibrate ideal PHRED trimming

Keywords
Base-calling, sequence trimming, DNA sequencing, PHRED, pUC18, local alignments, SWAT

ABSTRACT

The presence of noise in DNA sequencing reads is caused by problems in one or more of the following procedures: (1) the sequencing amplification reaction itself, (2) molecule migration during electrophoresis, and (3) base-calling. Accordingly the sequence called contains an aspect of noise-information-noise, where the sequences at the termini of the reads are not trustworthy. To address this problem, terminal sequences are commonly trimmed to ensure retrieval of just the reliable part of the read. Here, we develop an experimental strategy to minimize the effect of errors in the sequencing reaction in order to evaluate precisely PHRED trimming parameters, adjusting them to avoid noise. With this in mind, a large number of pUC18 cloning vector sequences were produced in a single reaction pool. A SWAT-based optimal local alignment approach was used to identify the best trimming parameters when comparing the sequencing reads produced against pUC18 published sequence. Our data suggests that, in the case of EST projects, PHRED should be run with parameters -trim_alt and -trim_cutoff 0.16 in order to obtain maximum biological information from the sequencing read. These results will also allow for the ability to choose the best PHRED parameters for other approaches.

INTRODUCTION

The most common procedure to generate a DNA sequence consists of three main stages: (1) the sequencing amplification reaction, (2) the electrophoresis and signal reading by the sequencing machine and (3) the base-calling. Each stage may have its own errors and new approaches should be developed trying to avoid them. Sanger and collaborators realized early on that a hairpin structure may be generated in some molecules, inducing them to migrate faster during electrophoresis than expected for the size of the molecule, shifting left the reading of some bases and spoiling the sequencing reaction (Sanger and Coulson, 1975; Sanger et al., 1977). Other well-known putative errors for DNA sequencing are related to the fast mobility of small fragments (spoiling the beginning of reads), inefficient incorporation of dideoxynucleotides (producing weak signal), and bad quality after one or two nucleotide repetitive regions (Ewing and Green, 1998).

Although the presence of errors is intrinsic to the DNA sequencing procedure, the correct usage of base-caller algorithms can overcome many of them and help to produce *bona fide* sequences. Base calling algorithms are used to transform raw sequencing machine data to the characters A, C, G, or T, representing each of the DNA nucleotides. Base caller algorithms may use different strategies to perform the base-calling, such as: Fourier analysis (Ewing and Green, 1998), maximum likelihood (Brady et al., 2000) and priority detection of peaks (Walther et al., 2001). The algorithms usually produce the DNA sequence in FASTA format along with a file that contains putative quality values for each of the bases sequenced.

Although many base-calling programs are available (Giddings et al. 1993, Song and Yeung 2000, Brady et al. 2000, Walther et al. 2001, He and McGown 2001), the most widely used in the genomics field is PHRED (Ewing et al. 1998, Ewing and Green 1998). This software is compatible with multiple sequencing machines and chemistries, and it functions in association with other widely used genomics software, such as cross_match, PHRAP (http://www.phrap.org) and consed (Gordon et al. 1998).

Frequently, the quality of a genome sequence is measured through the number of bases presenting a "PHRED quality value" (PQV) above a pre-determined value (Felsenfeld et al. 1999). The PQV is calculated as the negative logarithm of the error percentage for each base; thus, for example a base with PQV=10 has one chance in ten to be miscalled, while a base with PQV=20 has one chance in a hundred (Ewing and Green 1998). The dogma in the genome sequencing community is that the worst

acceptable PQV value is 20. Sequences presenting nucleotides with less than PQV=20 are considered of poor quality.

It is well known that the quality of the sequencing reads is worst at the extremities of the reads. Therefore, in order to separate signal from noise, sequences at the extremities are commonly trimmed (discarded) when the PQV falls under a certain value. However, different genome and transcriptome research consortiums use different PQV cutoffs during their trimming procedure (Felsenfeld et al. 1999, Verjovski-Almeida et al. 2003), while some papers have been described adopting alternative ways of sequence trimming (White et al. 1993, Telles and Silva 2001, Chou and Holmes 2001, Li and Chou 2004). Considering that PHRED is the most widely used base-caller and has its own trimming parameters to be set, we attempted to define the best parameters on which the sequence called should be trimmed to retrieve just the *bona fide* part of the read. This rationalization is relevant to both monetary and scientific factors, since there is a cost for sequencing DNA molecules and it is necessary to distinguish between sequence signal and random noise in all DNA sequence analysis. Moreover, sequences deposited after incorrect trimming might populate the nucleotide databases with miscalled nucleotides, spending extra time to the biologists to analyze this misleading data.

Local alignment tools compare two segments of DNA or protein sequences until the point that they are found to be similar, giving up the alignment when the sequences become too different (Altschul and Gish, 1996). Many local alignment tools are frequently used by molecular biologists, BLAST (Altschul et al. 1997) being the most widely used. The advantage of using BLAST comes from the fact that it is optimized to produce fast and consistent results when using the large amount of data in public molecular databases (McGinnis and Maden 2004). However, BLAST is a heuristic approach and, therefore, it does not produce the best alignment result possible (Ye et al. 2006). Better alignment results can be obtained using the Smith-Waterman (SWAT) algorithm (Smith and Waterman, 1981). This algorithm performs the best possible computational alignment between two sequences, at the cost of using more time than BLAST on the analysis.

Here, a SWAT-based local alignment approach has been used to define the best PQV trimming parameter for genomics projects. The calibration of PHRED shown here should produce the correct separation of electrophoresis/base-calling noise from the sequence information present on the molecule studied, saving money and time from the analysis.

In the present study, hundreds of pUC18 sequences were produced in a single pool, trimmed with PHRED with different trimming parameters and aligned to the consensus sequence of this cloning vector (GenBank Accession Number L09136) using the SWAT algorithm. The single-pool sequencing procedure generates a dilution of the sequencing errors, allowing us to analyze more precisely the base-calling procedure, and the known sequence of pUC18 has allowed a very precise positive control of the analysis. As a result, we were able to define the precise parameters with which PHRED should be run to allow the retrieval of just the informational part of the sequences called. Researchers interested in sequence trimming, looking to the data provided here, should be able to rationally define a precise trim cutoff to use in their projects depending on how informative and trustworthy they want their sequences to be.

METHODS

Single-pool sequence reaction

The sequencing reactions used here were made in a single pool and divided into tubes for the PCR sequencing reaction. After the PCR sequencing reaction, the sequences were joined again in the same tube, mixed, and then divided between three 96-well plates. This single-pool procedure homogenized the samples and thus diluted specific errors that might occur in some sequencing reactions. Each plate was run 3 times on a MegaBACE sequencer, yielding a total of 864 reads. From those, 846 processed ESD files (840100 bases) were obtained.

Base-calling by PHRED

All 846 ESD files were processed by PHRED with different trimming parameters. First PHRED was run on the sequences with no trimming parameters. After, PHRED was executed with *-trim* and *-trim_alt* parameters. When using *-trim_alt*, the parameter *-trim_cutoff*, defining the percentage of trimmed allowed, was varied from 0.01 (1%) to 0.25 (25%).

No trimming:

```
phred traces/trace_i -st fasta -q qual/trace_i.qual -s fasta/trace_i.fasta
```

Trim

```
phred traces/trace_i –trim "" -st fasta -q qual/trace_i.qual -s fasta/trace_i.fasta.trim
```

Trim_alt, trim_cutoff

```
phred traces/trace_i -trim_alt "" -st fasta -q qual/trace_i.qual -s fasta/trace_i.fasta -trim_cutoff j
```

where *j* was varied from 0.01 to 0.25

The trimming position (TP) was defined by a tuple (start, end) where start and end correspond to, respectively, the initial and final trimming positions deduced from the first line of the resulting FASTA formatted file. The TP information was used to populate a table into a MySQL database.

Local alignment by SWAT

All FASTA sequences produced with different trimming parameters were aligned with the pUC18 consensus (L09136) using the SWAT algorithm provided on PHRAP's package. The similarity matrix used was mat70 and only the best scored local alignment produced was considered.

```
swat fasta/trace_i.fasta pUC18 –M mat70 –N 1 > trace_i.swat
```

RESULTS

Figure 1 shows an actual example from our database that justifies the present work. Using average PHRED trim parameters (PQV 15 trimming), 182 informative bases (from base 39 to 220) were lost from the "left" side (5' of the read) and 355 informative bases (from base 469 to 823) were lost on the "right" side of the molecule (3' of the read). Despite the fact that these bases showed valid alignment against the consensus pUC18 sequence, all are lost when a typical PHRED trimming parameter is used.

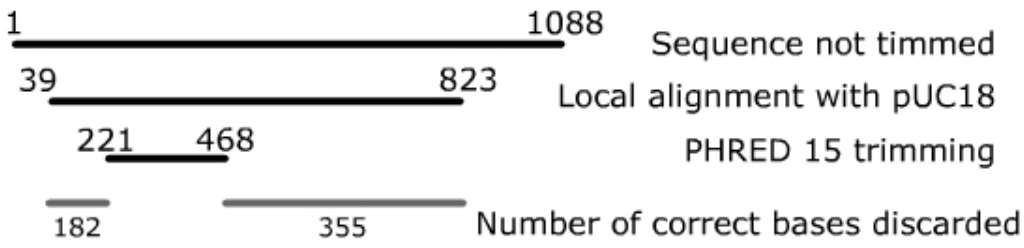


FIGURE 1. Example of informative bases lost when using a typical trimming parameter (PHRED 15).

Although the example shown in Figure 1 shows the potential loss of information, when low trimming parameters are used there is addition of miscalled information, then the base-called sequence contains the entire region produced by SWAT alignment and some extra bases, on both sides, representing noise data. Thus, we found our analysis in one basic, clear and reasonable premise: when you align a base-called sequencing read with its positive control (in case, the published sequence of pUC18), the region observed by optimal local alignment represents the actual biological information present in your nucleotide sequence. Evidently, as the trimming procedure is made less stringent, the greater will be the number of incorrect called bases. Thus, our main goal was to set PHRED trimming parameters to match SWAT alignment between the average read and the pUC18 consensus.

In order to analyze which PHRED parameter threshold should be used to minimize the number of "good" bases discarded, the 846 pUC18 sequences of a single sequencing pool reaction were trimmed with different PHRED parameters. The average number of bases included and discarded by PHRED trim_cutoff parameter is shown in Figure 2 for the "right" side of the molecule (3' end). Figure 3 shows the same data for the "left" side (5' end).

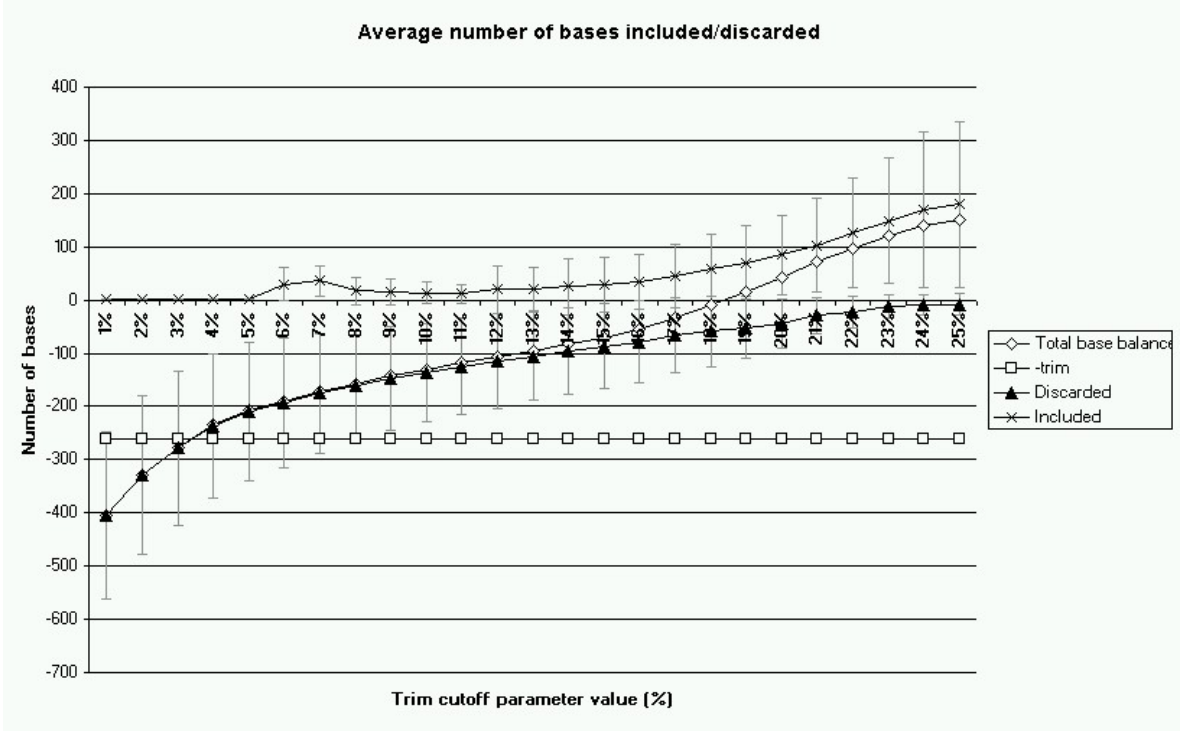


FIGURE 2. Base balance by trim_cutoff for the right side.

The percentage on the x-axis of Figure 2 represents the parameter -trim_cutoff used. PHRED uses three main parameters for trimming. The first and simplest of them is "-trim", which trims the sequence without allowing the specification of any other parameters other than the sequence of the restriction enzyme (user's choice). As one can see in Figure 2, using PHRED *-trim* parameter (squares in the picture), an average of 262 informative bases is lost on the right side of the sequences. The result is in accordance with the information given in the PHRED manual that states: "We recommend that you use '-trim_alt' rather than '-trim' option (…) because we believe that '-trim' trims off too many good bases" (http://www.phrap.org/phredphrap/phred.html). So, we followed the recommendation: the other trimming parameters evaluated here were *-trim_alt* and *-trim_cutoff*. The former one sets PHRED to trim a subsequence of the entire sequence based on trace quality. The trim_alt parameter use an error probability of 0.05 to trim the sequences, except when used together with -trim_cutoff parameter. In this case it trims the sequence based on the error probability given by -trim_cutoff parameter. Using the parameters –trim_alt –trim_cutoff 0.01 (the error probability of 0.01 is related to PQV=20), 403 bases are lost (standard deviation of 158 bases) and using trimming with PQV=15 (trim cutoff 0.03), the number of informative lost bases is 276. The correlation between error

probabilities and PQV is given by the following formula (Ewing et al., 1998) that defines a PQV as a negative logarithm of error probability, times 10:

PQV = -10 log10(p), where *p* is the error probability.

The -trim_cutoff percentage entered in the formula as *p* gives the corresponding PQV.

The value of the -trim_cutoff PHRED parameter that is best to avoid both the incorporation of miscalled bases and the lost of informative ones is derived from the point at which no informative bases are either won or lost; namely, trim cutoff 0.18, which corresponds (rounding to the closest value) to PQV=7, a value very far from the current dogma of PQV 20. Considering the differences between different sequences and sequence contexts, we suggest the adoption of a slightly more conservative position for the right side trimming. We suggest the usage of -trim_alt with –trim-cutoff value of 0.16, which corresponds almost exactly to PQV=8.

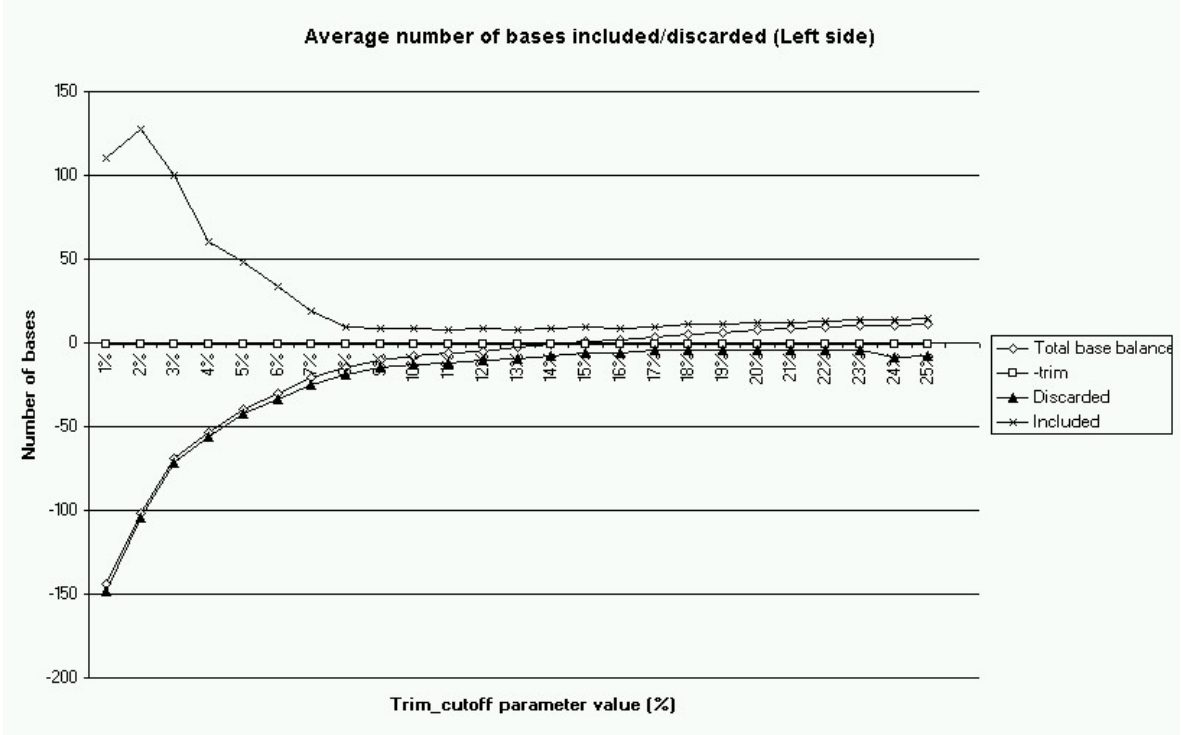


FIGURE 3. Base balance by trim_cutoff for the left side.

Figure 3 shows the number of included and discarded bases on the left side for pUC18 PHRED base-called sequences. On average, one single informative base is lost on the left side of the molecule when using the default -trim parameter. Using PQV=20

trimming parameter (-trim_cutoff 0.01), 144 bases are lost in average (standard deviation of 110 bases) and using PQV=15, the number of informative bases lost is 69. The PQV which minimizes both the loss of informative bases and the addition of miscalled ones corresponds to trim cutoff 0.14, which is closest to PQV=9. Moreover, using PHRED trimming parameter cutoff of 0.16, only 2 non-informative bases are retained, on average (standard deviation of ~14). Despite the high number of informative bases missed under PQV=20 trimming on the right side (Figure 2), we observed a strange behavior regarding the bases included under low trim-cutoff parameters, due to the presence of fifteen base-called sequences with many non-informative bases in the left side: two of them showing more than 400 miscalled bases, leading to the observation of a peak in the left side of Figure 3. As the number of sequences presenting additional bases increases to a significant proportion (around 100 sequences, by 0.08 cutoff, see Figure 5 below), the averaged result decreases approaching more realistic values.

It was also verified, for all base-called sequences analyzed, how many have included or discarded bases by trim_cutoff. Figures 4 and 5 show these data for the right and left side, respectively.

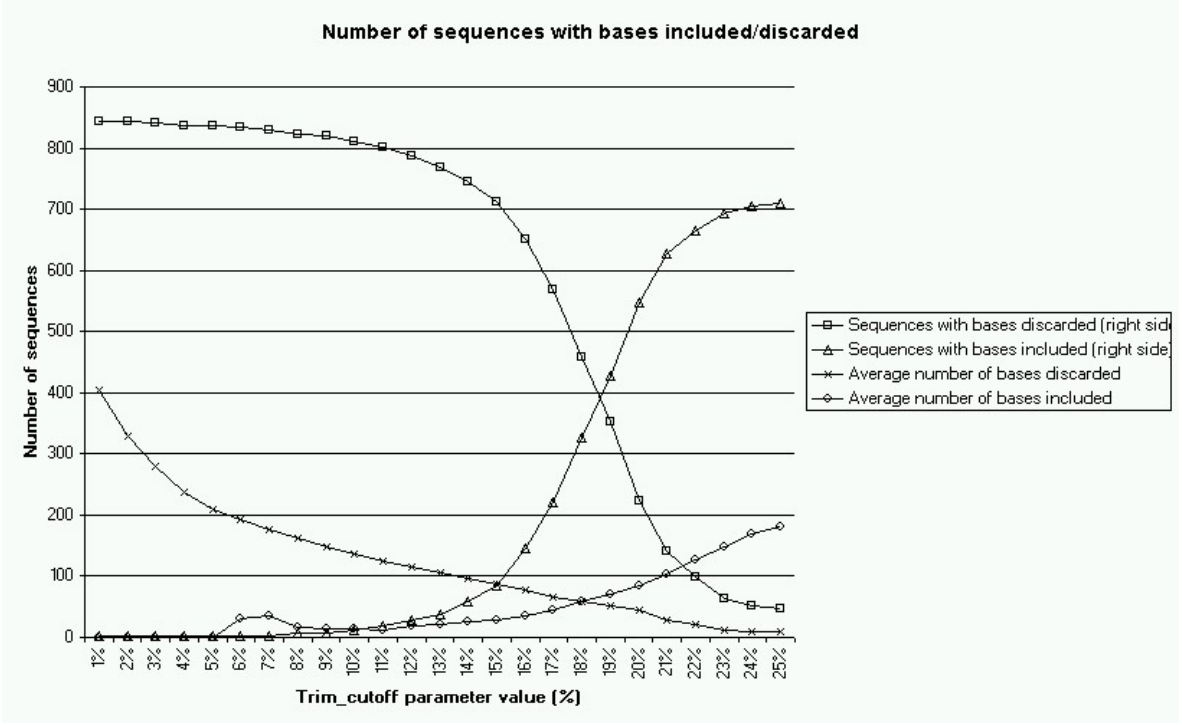


FIGURE 4. Number of sequences with bases included or discarded and average number of these bases for the right side of the sequences.

We also plot on Figures 4 and 5 the number of bases included or discarded, such as seen on Figures 2 and 3, to give the reader a broader view of the results generated and to show the coincidence in the intersection of lines.

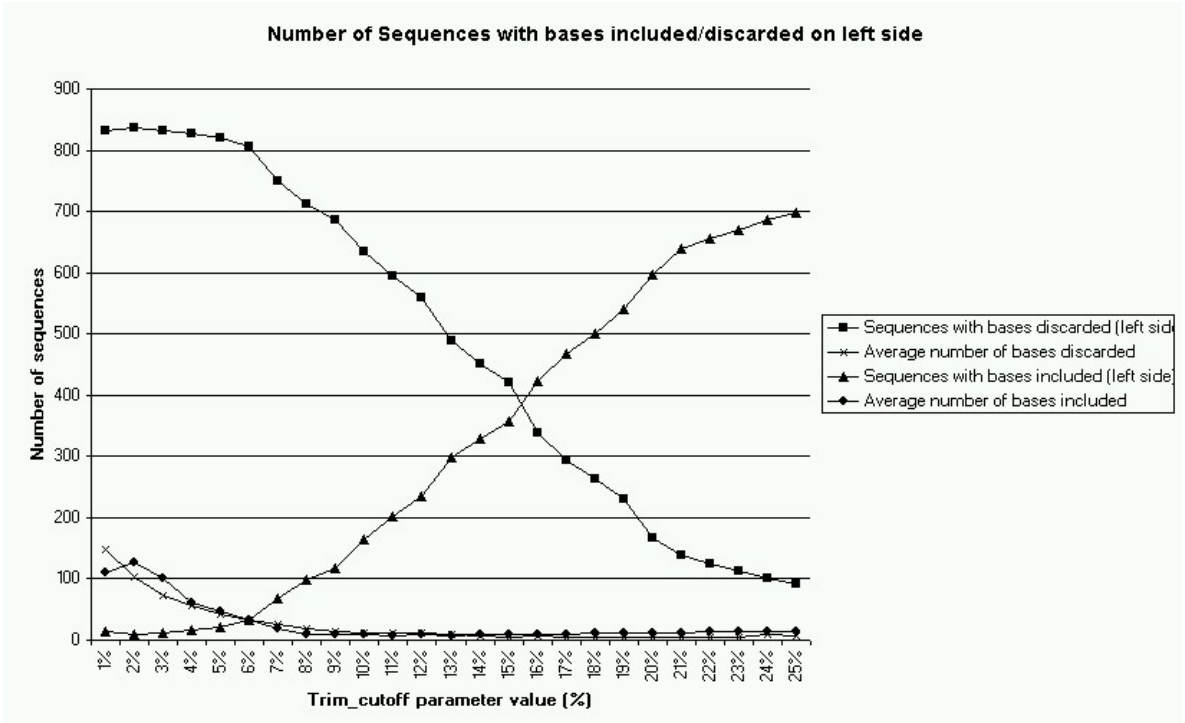


FIGURE 5. Number of sequences with bases included or discarded and average number of these bases for the left side of the sequences.

For the left side of base-called sequences (Figure 5), the average number of bases included/discarded is close to zero for trimming parameters greater than 7%. This observation shows a small effect of varying trimming cutoff parameters for this side of the molecule.

DISCUSSION

We have shown that the correct setting of PHRED trimming parameters results in the obtaining of more information from a sequencing read, avoiding the waste of relevant biological information while preventing the presence of many miscalled bases.

Researchers working on EST data should pay special attention to these results, since the main goal of the EST projects is the identification of genes expressed in a

cell/organism under some conditions (Adams et al., 1991). Moreover, this identification of ESTs (sequence annotation) is frequently done using local alignment (BLAST) against nucleotide and protein databases. The data showed here was also analyzed with BLAST and the results produced about the same of the SWAT approach. However, we prefer to show the SWAT data since it gives us the optimal local alignment. Processing of EST data with a lower PHRED score value may increase up to three times their biological information in some sequences (Figure 1 and data not shown).

Corroborating our observations, PHRED trimmed reads aligned to pUC18 translated consensus sequence using BLASTx support that maximum information is attained under PQV 8. Figure 6 shows that maximum averaged BLAST score is obtained when sequences are trimmed with trim cutoff 0.16 and that sequence trimming using less stringent PQV values do not produce better scores, confirming PQV 8 as the most useful to retrieve maximum biological information from a sequencing read.

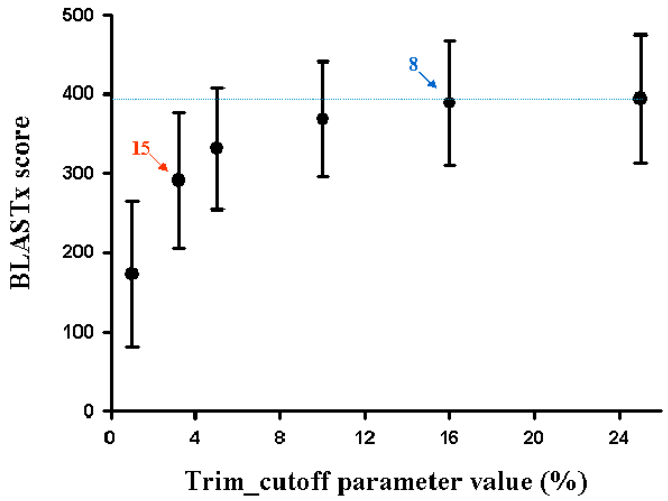


FIGURE 6. BLASTx scores using pUC18 sequence translated to amino acid sequence as subject and reads processed with the indicated trim cutoff (in percentage). PQV 15 and PVQ 8 are indicated by arrows.

Researchers interested in genome projects also could take advantage of the data presented here, since we provide clear results that will give them the foundation to choose if they will accept longer and more error-prone sequences, increasing the coverage with fewer sequences. Alternatively, they will be able to choose to deal with small but high quality sequences by using low trim-cutoff values. In that case, the

ideal trimming cutoff would be somewhere between no trimming and not over trimming.

We recommend here the utilization of PHRED trimming parameter cutoff of 0.16 to trim sequences aiming to obtain maximum biological information (regardless of best sequence quality). Inspecting the data showed here, a researcher can choose which number of bases he/she wants to loose or include. It has been also noticed that, on the left side of the molecule, the trimming parameter has less relevance than for the right side, where many more bases are frequently discarded.

The procedure applied here to a single-pool sequencing reaction might provide the confidence to researchers to admit a greater percentage of errors at the tips of the reads (up to 16% under our recommendation). Moreover, a similar analysis can be promptly tested on sequences of empty cloning vectors, certainly available in the collection of reads in any sequencing laboratory.

ACKNOWLEDGEMENTS

The authors wish to thank the "Rede Genoma de Minas Gerais" (supported by FAPEMIG and MCT/Brazil), especially Marina Mourao, Lucila Pacifico and Renata Ribeiro, for providing the sequences used in the analysis. We specially thank Dr. Darren Natale (PIR) for critically reviewing this manuscript.

REFERENCES

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**(5013), 1651-6.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**(17), 3389-402.
3. Altschul SF, Gish W. (1996) Local alignment statistics. *Methods Enzymol*, **266**, 460-80.
4. Brady D, Kocic M, Miller AW, Karger BL. (2000) A maximum-likelihood base caller for DNA sequencing. *IEEE Trans Biomed Eng.,* **47**(9), 1271-80.

5. Chou HH, Holmes MH. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics,* **17**(12), 1093-104.
6. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. (1998) Accuracy assessment. *Genome Res*., **8**(3), 175-85.
7. Ewing B, Green P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*., **8**(3), 186-94.
8. Felsenfeld A, Peterson J, Schloss J, Guyer M. (1999) Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res*., **9**(1), 1-4.
9. Giddings MC, Brumley RL Jr, Haker M, Smith LM. (1993) An adaptive, object oriented strategy for base calling in DNA sequence analysis. *Nucleic Acids Res*., **21**(19), 4530-40.
10. Gordon D, Abajian C, Green P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res*., **8**(3), 195-202.
11. He H, McGown LB. (2000) DNA sequencing by capillary electrophoresis with four-decay fluorescence detection. *Anal Chem.*, **72**(24), 5865-73.
12. Li S, Chou HH. (2004) LUCY2: an interactive DNA sequence quality trimming and vector removal tool. *Bioinformatics,* **20**(16), 2865-6.
13. McGinnis S, Madden TL. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*., **32**(Web Server issue), W20-5.
14. Sanger F, Coulson AR. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*., **94**(3), 441-8.
15. Sanger F, Nicklen S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*., **74**(12), 5463-7
16. Smith, TF, Waterman MS. (1981) Identification of common molecular subsequences. *J Mol Biol*., **147**(1), 195-7.
17. Song JM, Yeung ES. (2000) Alternative base-calling algorithm for DNA sequencing based on four-label multicolor detection. *Electrophoresis,* **21**(4), 807-15.
18. Telles GP, da Silva FR. (2001) Trimming and clustering sugarcane ESTs. *Gen Mol Biol.,* **24**(4), 17-23.
19. Verjovski-Almeida S, DeMarco R, Martins EA, Guimaraes PE, Ojopi EP, Paquola AC, Piazza JP, Nishiyama MY Jr, Kitajima JP, Adamson RE, Ashton PD, Bonaldo MF, Coulson PS, Dillon GP, Farias LP, Gregorio SP, Ho PL, Leite RA, Malaquias LC, Marques RC, Miyasato PA, Nascimento AL, Ohlweiler FP, Reis EM, Ribeiro MA, Sa RG, Stukart GC, Soares MB, Gargioni C, Kawano T, Rodrigues V, Madeira AM, Wilson RA, Menck CF, Setubal JC, Leite LC, Dias-Neto E. (2003) Transcriptome

analysis of the acoelomate human parasite Schistosoma mansoni. *Nat Genet.,* **35**(2), 148-57.

20. Ye J, McGinnis S, Madden TL. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**(Web Server issue), W6-9.
21. Walther D, Bartha G, Morris M. (2001) Basecalling with LifeTrace. *Genome Res.,* **11**(5), 875-88.
22. White O, Dunning T, Sutton G, Adams M, Venter JC, Fields C. (1993) A quality control algorithm for DNA sequencing projects. *Nucleic Acids Res.*, **21**(16), 3829-38.

5.6. Efeito do número de leituras e de poda (*trimming*) na qualidade e tamanho de consensos

Depois de analisarmos extensivamente o efeito dos escores de PHRED na qualidade das seqüências, decidimos avaliar o efeito do agrupamento de seqüências na qualidade do consenso gerado. Este trabalho partiu da seguinte pergunta inicial, sempre realizada pelos biólogos moleculares aos bioinformatas: "quantas vezes preciso seqüenciar minha molécula de interesse para obter uma seqüência fiel da mesma?". Evidentemente esta questão é ainda mais relevante quando a produção de seqüências é feita em larga escala, pois embora aparentemente complicado, é bastante simples identificar automaticamente clones de interesse e utilizar uma mesma reação de seqüenciamento para obter várias leituras.

Partindo do pressuposto que os erros encontrados nas seqüências são raros, acredita-se que o re-sequenciamento da mesma molécula, seguido por posterior agrupamento das seqüências produzidas, irá aumentar a qualidade da seqüência, gerando uma seqüência de bases altamente confiável. Mas qual será o número ideal de seqüências a ser gerado para minimizar o custo/benefício na produção de moléculas de alta qualidade?

Para responder a esta pergunta, analisamos tanto (1) o número de leituras de seqüenciamento a serem produzidas; (2) o valor de *trim_cutoff* a ser utilizado durante a nomeação das bases e (3) o tamanho final dos consensos. Assim, partimos de 846 seqüências de pUC18 cuja poda com diferentes valores de *trim_cutoff* já havia sido realizada e as separamos em grupos de duas a dez seqüências, gerando 1000 grupos de cada tamanho. Os consensos gerados pelo agrupamento dessas seqüências utilizando PHRAP (não mostrado) e CAP3 foram novamente comparados com o padrão do pUC18 utilizando alinhamentos locais (BLAST).

Os resultados mostraram que a execução do *trimming* reduz eficientemente o número de erros, ainda que isso leve à redução do tamanho do consenso produzido. Notamos ainda que o aumento no número de leituras produzidas não leva ao efeito esperado de melhorar a qualidade do consenso gerado significativamente, mas faz com que o tamanho deste seja relativamente aumentado. Portanto, o pesquisador deve fazer um balanço da quantidade de erros admitida e do tamanho da seqüência desejada ao escolher quantas vezes seqüenciar sua molécula de interesse. Todavia, a expectativa de que o sequenciamento de 10 amostras de uma mesma molécula pudesse assegurar a produção de um consenso idêntico ao molde não se comprova, sendo mais aconselhável, nesses casos, o aumento do rigor de poda. Este trabalho,

entitulado "Effects of the number of reads and trimming on quality and size of assembled consensi" foi submetido à revista Genetics and Molecular Research e estamos esperando as críticas dos revisores.

Effects of sample re-sequencing and trimming on the quality and size of assembled consensus

Francisco Prosdocimi[1], Denize Altiva de Oliveira Lopes[1], Fabiano Cruz Peixoto[2], José Miguel Ortega[1]*

1- Laboratório de Biodados. Depto. Bioquímica e Imunologia, ICB-UFMG
2- Laboratório de Computação Científica, UFMG

José Miguel Ortega *
miguel@icb.ufmg.br
Laboratório de Biodados. Sala N4-202.
Departamento de Bioquímica e Imunologia, ICB, UFMG
Av. Antônio Carlos, 6627 C.P. 486
31.270-010 Belo Horizonte, MG, Brazil
Tel: +55 31 3499-2654
Fax: +55 31 3499-2570

RUNNING HEAD
Number of reads and consensus quality

* To whom correspondence should be addressed

SUMMARY

The production of nucleic acid sequences is always associated with base calling errors. In order to produce a high quality DNA sequence from a molecule of interest, researchers are used to sequence the same sample many times. Therefore, considering base-calling errors as rare events, re-sequencing the same molecule and assembling the reads is frequently thought to be a way to produce reliable sequences. However, a relevant question on this issue is: how many times the sample needs to be re-sequenced to minimize the costs and achieve this high-fidelity sequence? Here, both the effect of re-sequencing numbers and PHRED trimming parameters were observed to verify the accuracy and size of the final consensus sequence. Hundreds of single-pool reaction pUC18 reads were generated and assembled into consensus with CAP3. Using local alignment against the pUC18 cloning vector published sequence, the position and number of errors in the consensus were identified and stored in MySQL databases. We verified that stringent PHRED trimming parameters efficiently reduces the number of errors, although it also reduces the size of the final consensus. Moreover we observed the poor effect of re-sequencing on reducing the number of errors, although this procedure was capable to enlarge the consensus size.

Keywords: Sequencing reads, trimming, assembling, PHRED, CAP3.

INTRODUCTION

Most of the recent developments in the field of genomics and bioinformatics dealt with data generated from genome sequencing projects and it is well-known that all genomes are build *in silico* by the superposition of thousands of overlapping reads joined together by assembly softwares such as PHRAP (Green, 1999) or CAP3 (Huang and Madan, 1999). Some assembly software, including the two cited here, take advantage of base quality values determined by base-caller algorithms such PHRED (Green and Ewing, 1998; Ewing *et al*., 1998) in order to produce more reliable consensus sequences. Although their main application consists in the production of huge genomic sequences, assembly software are also used to cluster EST (Expressed Sequence Tag) data, in this case focusing gene discovery based on single-pass, partial sequencing of cDNA molecules, aiming to analyze the transcriptome (Adams *et al*., 1991; Franco *et al*., 1997). One interesting issue about this consists in the fact that assembled molecules from genome projects are allowed to enter in genome databases whilst assembled ESTs are restricted to specific project websites and they are not allowed to be integrated in any of the very best known public molecular databases. Nevertheless, not so rare is the evolution of an EST project to a full-length cDNA sequencing project, such as the Mammalian Gene Collection (Strausberg *et al*., 1999; MGC Program Team, 2002), where selected clones are introduced into a pipeline of dedicated sequencing to eliminate any ambiguities from the reads generating a consensus sequence. The edited sequences can then be deposited into databases distinct from dbEST (Boguski *et al*., 1993), such as GenBank and GenPept, becoming therefore targets for ordinary BLAST similarity searches. Ideally, a combination of forward and reverse reads should be used in EST sequencing projects, but many of the selected cDNA clones are larger than the distance that could be covered in both orientations with the simple alternative of using vector anchored primers. Thus, the question that rises is whether or not a sufficiently large number of reads could be assembled into an error-free consensus and, if so, what would be the cost/benefit relationship between the number of samples sequenced and the efficiency in the production of this high-quality consensus, which could be promptly deposited as a partial cDNA sequence, either 5’ or 3’.

Another possible alternative is the manual editing of the consensus with software such as Consed (Gordon *et al*., 1998), a procedure that shall be encouraged instead of any automated alternatives, although the operator would certainly benefit from additional information produced by automated tools, such as the expected

number of errors per molecule as a function of (i) the number of available reads and (ii) the amount of errors admitted during trimming procedures. Usually, for genome projects, trimming seems to not be recommended because high quality regions are often overlapping low quality ones. However, this is clearly not the situation in partial sequencing of cDNA molecules, since all reads are expected to start at the same position and, most important, the low quality regions are concentrated in the edges of the sequences.

In this work we report the analysis of sample re-sequencing (from 2 up to 10 times) and PHRED trimming parameters on assembled consensus' errors and size. All procedures were conducted using a set of 846 pUC18 one-direction reads, generated by a single-pool sequencing reaction (Prosdocimi et al., 2004). Assembling was conducted with CAP3 software and errors were analyzed with BLASTn (Altschul *et al*., 1997). Data obtained indicated that trimming efficiently reduces the number of errors but affects the size of the consensus, while the impact of having a large number of reads is not as remarkable as it could intuitively be expected.

METHODS

Sequencing reactions

The sequencing reaction premix was made in a single pool and divided on tubes for the PCR sequencing reaction. After that, products were joined together on the same tube, mixed, and sequenced in 96-well plates with MegaBACE equipment. This single-pool procedure have homogenized all our samples and dilute specific errors occurred in some sequencing reactions. Three laboratories from the Federal University of Minas Gerais (UFMG) that integrate the Minas Gerais Genome Network provided the 846 processed ESD files used in this work.

Base calling and trimming

All ESD files were processed by PHRED using variable trimming parameters. First, PHRED was run on each sequence with no trimming parameters (nT data). Further, PHRED was performed using *-trim_alt* parameter. When using *-trim_alt*, the parameter *-trim_cutoff* was modified from 0.01 (1%) to 0.25 (25%) for each read. This means that each read have been trimmed 26 times, with different PHRED trimming parameters, and the FASTA and QUAL resulting files were stored.

Sequence assembly

From the 846 ESD files, 1,000 groups of two sequences were randomly taken and assembled with CAP3 software. The same procedure was done for groups of 3, 4, 5, 6, 7, 8, 9 and 10 sequences. Therefore, we have done 9,000 sequence group draws and assembly.

Local alignment against pUC18 published sequence

All the generated consensus sequences were compared to the pUC18 published sequence (24.8% A, 25.2% C, 25.5% G, 24.5% T; GenBank accession number L08752) using the local alignment algorithm BLAST. Tabular output data (-m 8 option) was used to populate MySQL tables.

Statistical analyses of data

Since the data did not fit normal distribution, non-parametric ANOVA statistical tests were performed. So, we have run Kruskal-Wallis median tests to analyze the number of errors and size of the consensus generated when using trimming cutoff parameter at 1% or not trimmed sequences.

RESULTS

Here, the efficiency of re-sequencing on the production of error-free consensus was evaluated by sampling thousands of groups containing two up to ten reads from a collection of 846 reads of the pUC18 cloning vector. Reads were base called with PHRED software and assembled with CAP3. Usually, during PHRED processing of data, no trimming of the low quality portion of the reads is performed (denoted by nT - no trimming - in figures). By aligning the 9,000 consensi produced with the published pUC18 sequence using BLASTn program, the errors in these *in silico* sequences (sometimes called contigs) were evaluated. It is noteworthy that BLASTn alignments do not elongate over the low quality portion of the reads, therefore errors per sequence tend to a maximum.

We decided to include additional data applying a trimming cutoff with PHRED internal algorithm -*trim_alt*, varying the trimming cutoff from 1 to 25% of accepted errors at the edge of reads, in order to check the effect of this pre-processing on the amount of errors in consensus sequences.

Figure 1a presents the average number of errors per consensus sequences assembled by CAP3 and accessed with BLASTn, showing that trimming can reduce errors to less than one per sequence. Re-sequencing (increasing from two up to ten reads) was expected to reduce significantly the number of errors from the consensus assembled but, in fact, the reduction was not as significant as one might suppose. For a detailed analysis two regions of the curves where the differences between data were either maximized or minimized (trimming cutoff of 1% or nT). These data were also statistically analyzed and they are shown in Figures 1b and 1c. For the consensus generated from nT reads, the best results were observed with the assembling of ten reads, although the cost/benefit over the use of three reads is clearly higher. When reads were trimmed with cutoff 0.01 (1%), the effect of increasing the number of reads from 2 to 10 has shown a 4.3 fold reduction on errors per molecule (up to 24% of the initial amount, figure 1b). However, with no trimming (figure 1c) the reduction was of 1.5 fold (64% of the initial amount remaining) and not significant from 3 up to 8 reads. Thus, trimming most efficiently decreases the errors while maintaining the highest responsiveness to the increase on the number of reads.

The surprising effect of increasing the number of reads on molecules trimmed under cutoff of 1%, which corresponds to PHRED 20, as opposed to the low effect on not trimmed reads, recommended pattern for genome assembly, lead us to investigate the nature of these errors. Data presented on figure 2a show that at 1% cutoff, mismatches were minimum even when using only two reads. However, from 10% cutoff up to no trimming (nT), the number of mismatches decreased in similar proportion as for total errors as more reads were assembled (compare figures 2a and 1a). In contrast, when we analyzed the gaps, the opposite was observed: under PHRED 20, gaps were efficiently reduced as the number of reads increases, but this was not observed for non trimmed or poorly trimmed reads (figure 2b). This last observation is concordant with data showing that high-quality errors are mainly those generated by the insertion (Prosdocimi *et. al*., 2003), thus producing gaps on the alignment. Therefore, the effect of efficient reduction in the number of reads under PHRED 20 is due to the decrease in the number of gaps and the proportion of decrease of the number of either mismatches or gaps under PHRED 10 up to no trimming is rather similar and low. Curiously, the number of gaps is minimum under 4% cutoff for 2 reads or 2% cutoff for 10 reads.

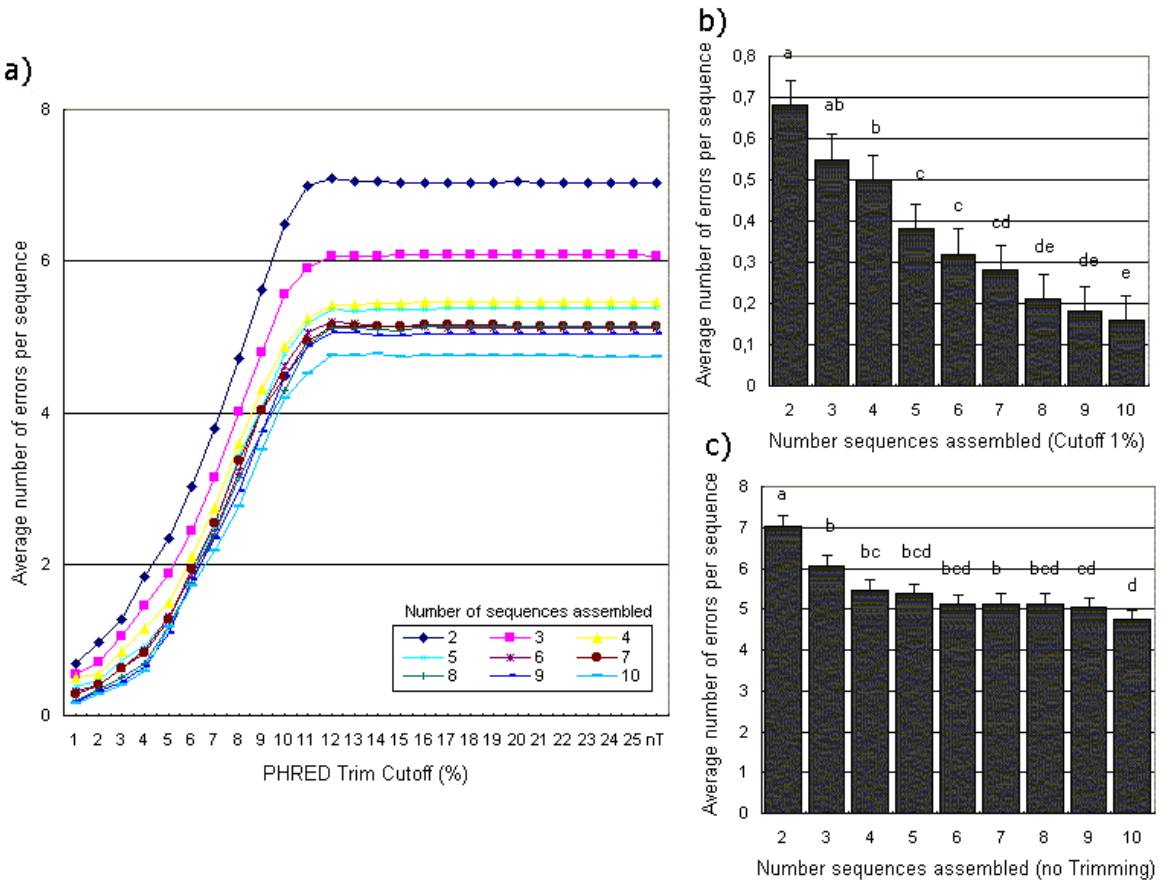


FIGURE 1 – a) Average number of errors per sequence when different number of reads (2 to 10) were assembled with CAP3 and aligned with pUC18 published sequence using blast software, sorted by phred trim cutoff percentage. B) Zoom for most interesting data.

Although under PHRED 10 up to no trimming the use of more than two reads did not significantly improved the quality of the obtained sequences, trimming in this range as opposed to 1% cutoff increased the size of the consensus, thus producing a realistic benefit (Figure 3a). Resultant assembled consensus sequences were greater than 500 base pairs. Moreover, data presented in Figures 3b and 3c depict that consensus size is more responsive to the number of reads under PHRED 20 (1% cutoff, Figure 3b) than when using non trimmed reads (Figure 3c).

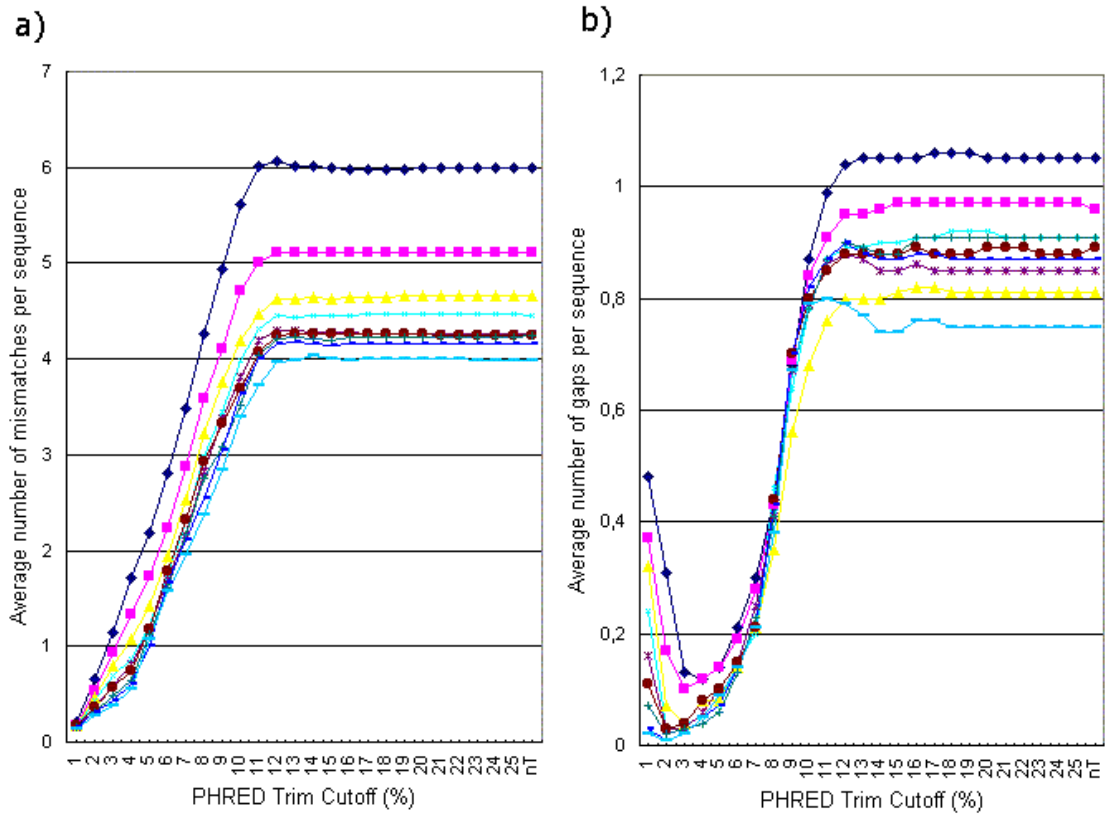


FIGURE 2 – a) Average number of mismatches per sequence when different number of reads (2 to 10) were assembled with CAP3 and aligned with pUC18 published sequence using blast software, sorted by PHRED trim cutoff percentage. b) Average number of gaps per sequence when different number of reads (2 to 10) were assembled with CAP3 and aligned with puc18 published sequence using blast software, sorted by PHRED trim cutoff percentage.

We considered the fact that all reads start at the primer and progressively loose quality as they proceed away from the starting position. This might result in situations where a poor-quality edge of a single read stands for the quality of the consensus, even if ten sequences have been assembled. Thus, we conducted the experiment exemplified in Figure 4. First, three up to ten reads were assembled and the consensus was aligned to the individual reads used in the assembly. After that, any portions of the consensus generated by only one or two reads were eliminated to ensure that each position of the consensus would be covered by at least three reads.

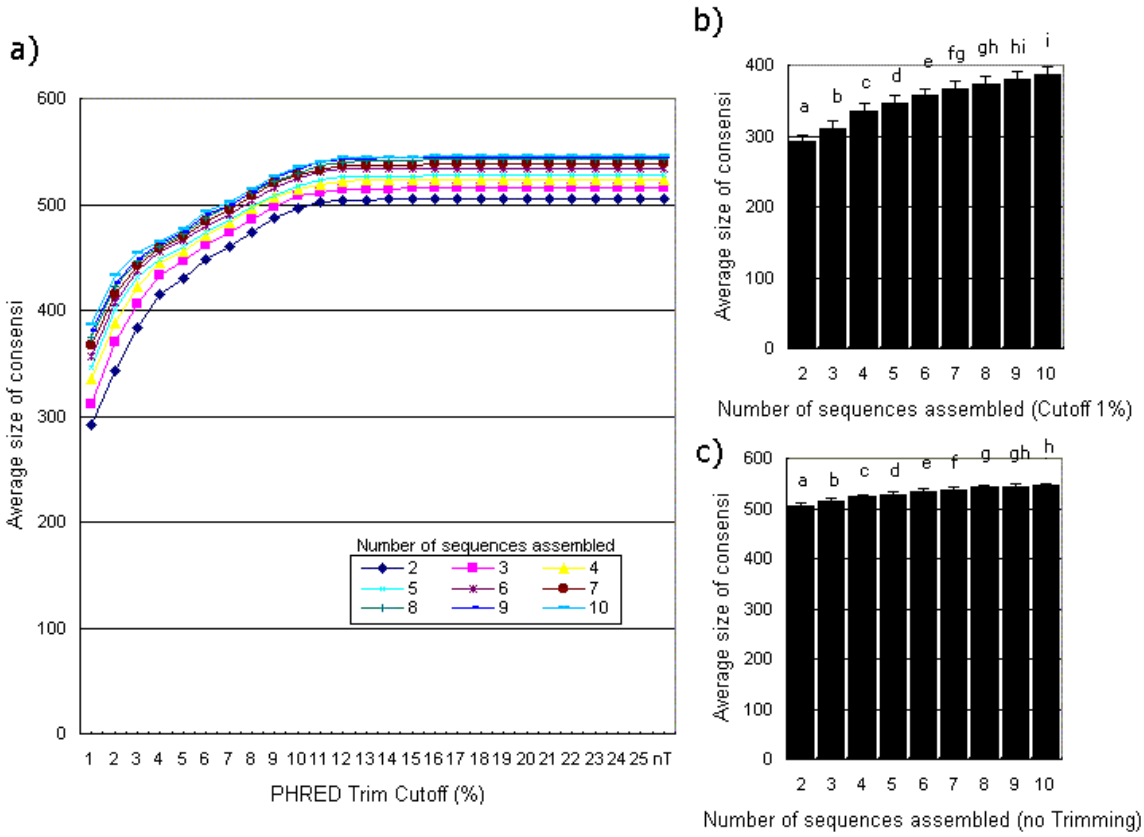


FIGURE 3 – a) Average size of consensi when different number of reads (2 to 10) were assembled with CAP3, sorted by PHRED trim cutoff percentage. B) Zoom for PHRED trim cutoff 1%.

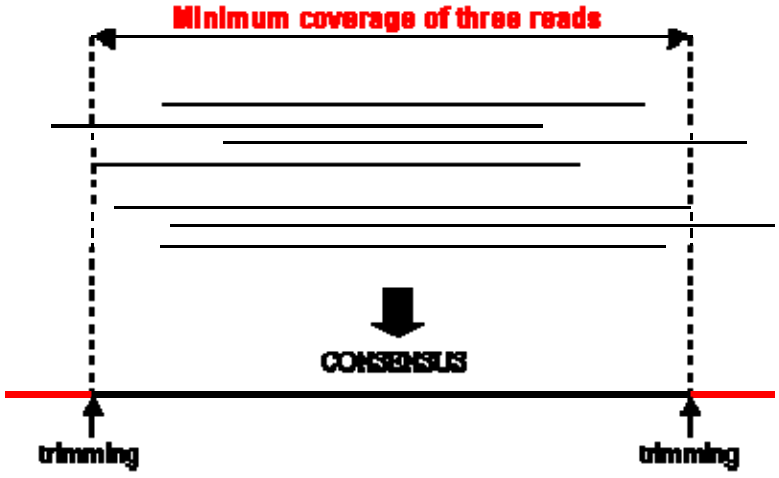


FIGURE 4 – a) Methodology for consensi trimming.

Analyses of the number of errors in these trimmed consensus sequences are presented in Figures 5a to 5c. The maximum number of errors per sequence diminishes from around 6 (Figure 1a) to up to 2.5 (Figure 5a). Again, increasing the number of used reads from three up to ten produced a small effect on the number of errors per consensus when non-trimmed individual reads (nT) were used (figure 5c). Intriguingly, the use of more than four reads raised the number of errors when using PHRED 20 cutoff (Figure 3b) in these 3-reads coverage consensus sequences. The total amount of errors in nT sequences, as compared to the simpler procedure of assembling the reads without taking on account the number of overlapping sequences (Figure 1a), is reduced by around 50% (from 5-7 to up to 2.5 errors per molecule), what it is still lower than the effect of trimming the reads with higher values of such as PHRED 20.

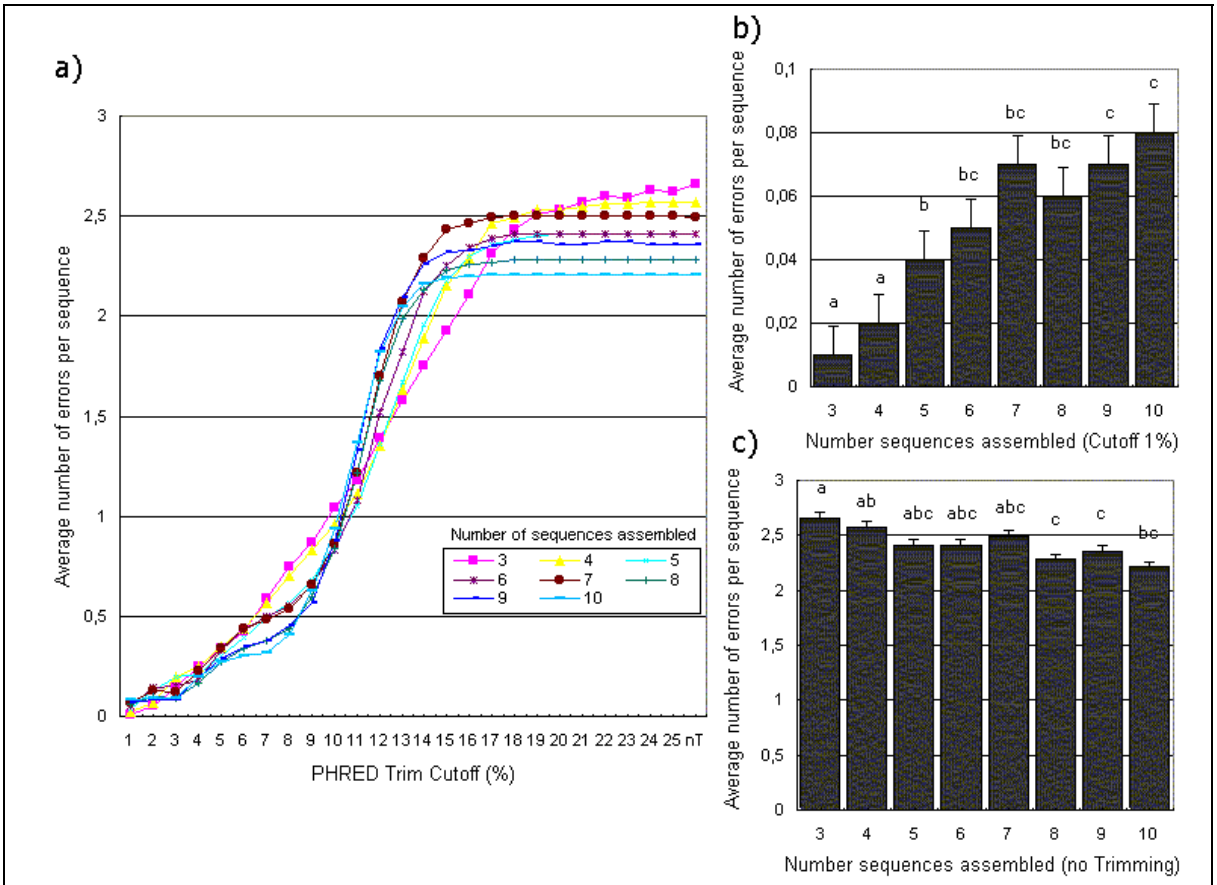


FIGURE 5 – a) Average number of errors per molecule when different number of reads (2 to 10) were assembled with CAP3, trimmed for the regions containing at least 3 sequences and aligned to puc18 published sequence using blast software, sorted by PHRED trim cutoff percentage. B) Zoom for PHRED trim cutoff 1%.

DISCUSSION

Application of PHRED base caller and the assembler software CAP3 is common on both large-scale genome analysis and low scale sequencing as exemplified here. Some works have already addressed issues about their functioning (Ewing and Green, 1998; Ewing *et al*., 1998; Richterich, 1998; Huang and Madan, 1999; Schen and Skiena, 2000; Walther *et al*., 2001), but as long as we know this is the first detailed analysis of sample re-sequencing and trimming parameters on the quality and size of the assembled consensus. Although manual inspection is desirable, here we evaluated the potential of automated procedures on giving the operator qualified information about the expected occurrence of errors. That might be valuable while inspecting 5'UTRs and N-terminal coding region without significant similarity to deposited sequences.

Our results show that the assemblage of large amount of reads (up to ten) does not reduce the average number of errors at the intensity that it might be possibly expected (Figures 1 and 2). It was evidenced that trimming procedures previous to the assemblage are the best choice when the aim is to obtain a high-quality sequence. However, the resultant sizes of these sequences are affected by stringent PHRED trimming cutoff parameters at the ranges shown in Figure 3. In our experiments, size reduction up to 40% was accompanied by a 10 fold error reduction per molecule due to trimming (PHRED 20), as compared to less than 10% gain in size and below 30% reduction of errors for non trimmed (nT) reads by increasing the number of reads up to ten.

This behavior could be restricted to the assembling software used. An alternative to CAP3 is the software PHRAP. We observed that consensus sequences assembled with PHRAP presented higher average number of errors than those produced by CAP3 (data not shown), in accordance with other published data (Huang and Madan, 1999), though the results obtained were very similar in nature.

The poor effect of re-sequencing on the average number of errors per sequence for non-trimmed reads was very equivalent when the type of error was investigated (Figures 2a and 2b), although the contribution of gaps seemed to count for the most of the reduction for PHRED 20 trimmed reads (Figures 2b and 1b). This observation is in agreement with our previous work that evidenced that mismatches are frequently associated with the lowest quality values while inserted bases often show higher quality values than mismatches (Prosdocimi *et al*., 2003). Thus, under stringent trimming cutoff (e.g. PHRED 20), due to the introduction of gaps by the assembler

software, the expected improvement by using more reads shall concentrate on diminishing the occurrence of gaps.

The clipping of consensus regions formed by the assemblage of less than three overlapping reads produced sequences with less number of errors (Figure 5), suggesting a procedure that is possible to be implemented. Even under this treatment of the set of reads, the effect of sample re-sequencing from 3 to 10 times was even less significant.

Thus, we conclude that the production of a large number of reads from the same molecule in a single direction, rather than eliminate consensus errors, is more efficient to enlarge the size of the produced sequence (around 33% and 10%, for PHRED 20 trimmed and non-trimmed reads, respectively, Figures 3b and 3c). The set of evaluation presented here provide the data necessary for research groups to balance between the size of the automated certified sequences and the quality of the generated consensus sequences. With a brief inspection of Figures 1 and 3, it is possible to choose the best PHRED trimming cutoff parameter and the number of reads to be assembled and furthermore to predict the expected average number of errors and size of the resultant consensus sequences.

In general, high-quality sequences are possible to be obtained with two reads (trimmed with PHRED 20) when size is not a constraint and the goal is to give the operator secure information about a specific portion of the read (e.g. when the correct translation start site is being investigated).

ACKNOWLEDGEMENTS

The authors wish to thank the "Rede Genoma de Minas Gerais" (supported by FAPEMIG and MCT/Brazil), especially Marina Mourao, Lucila Pacifico and Renata Ribeiro, for providing the sequences used in the analysis.

REFERENCES

1. Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, B.F., Kerlavage, A.R., McCombie, W.R. and Venter, J.C. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.

2. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST, PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402.

3. Boguski, M.S., Lowe, T.M. and Tolstoshey, C.M. (1993). dbEST- database for "expressed sequence tags". *Nat Genet* 4:332-3

4. Ewing, B. and Green, P. (1998) Base-Calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8: 186-194.

5. Ewing, B., Hillier, L., Wendl, M.C. and Green P. (1998) Base-Calling of automated sequencer traces using phred. I. Accuracy Assessment. *Genome Res* 8: 175-185.

6. Franco, G.R., Rabelo, E.M., Azevedo, V., Pena, H.B., Ortega, J.M., Santos, T.M., Meira, W.S., Rodrigues, N.A., Dias, C.M., Harrop, R., Wilson, A., Saber, M., Abdel-Hamid, H., Faria, M.S., Margutti, M.E., Parra, J.C. and Pena, S.D. (1997). Evaluation of cDNA libraries from different developmental stages of *Schistosoma mansoni* for production of expressed sequence tags (ESTs). *DNA Res* 4: 231-240.

7. Gordon, D., Abajian, C. and Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Res* 8:195-202.

8. Green, P. (1998) Documentation for PHRAP and cross-match. http://www.phrap.org/phrap.docs/phrap.html

9. Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly software. *Genome Biol* 9: 868-877.

10. MGC (Mammalian Gene Collection) Program Team. (2002). Generation and Initial Analysis of more than 15,000 Full-Length Human and Mouse cDNA Sequences. *PNAS* 99: 16899-16903

11. Prosdocimi, F., Peixoto, F.C. and Ortega, J.M. (2003) DNA Sequences Base Calling by PHRED: Error Pattern Analysis. *R Tecnol Inf* 3: 107-110.

12. Prosdocimi, F., Peixoto, F.C. and Ortega, J.M. (2004) Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values. *Gen Mol Res* 3:483-492.

13. README for stand-alone BLAST. ftp://ftp.ncbi.nlm.nih.gov/blast/documents/ blast.txt

14. Richterich, P. (1998) Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res*. 8:251-259.

15. Chen, T. and Skiena, S.S. (2000) A case study in genome-level fragment assembly. Bioinformatics. 16:494-500.

16. Strausberg, R.L., Feingold, E.A., Klausner, R.D. and Collins, F.S. (1999) The Mammalian Gene Collection. *Science* 286: 455-457.

17. Walther, D., Bartha, G. and Morris, M. (2001) Basecalling with LifeTrace. *Genome Res*. 11:875-888.

6. CONSIDERAÇÕES FINAIS

É incrível pensar que ainda hoje, quase 10 anos depois do *boom* de surgimento dos seqüenciadores em larga escala, considerando desde o revolucionário ABI PRISM 3700, desenvolvido em meados de 1998 (Davies, 2001), um trabalho como este ainda seja relevante no âmbito da pesquisa científica e que não se tenha pensado, desde o início, em formas mais racionais de utilização dos processos de sequenciamento e dos algoritmos de nomeação de bases. Um colaborador estrangeiro, ao ler um de nossos trabalhos, não pôde deixar de comentar: "não consigo acreditar que isso não tenha sido estudado antes". Mas não foi.

Parece-me impossível tentar precisar, ainda que através de aproximações grosseiras, quanto dinheiro e quanto tempo de análise especializada já não tenham sido desperdiçados por pesquisadores que, desde os primeiros projetos genoma e transcriptoma, poderiam beneficiar-se em muito do presente trabalho. Quanto será que já não se gastou de dinheiro ao sequenciar exageradamente bases do vetor de clonagem, por não se saber ao certo a distância do iniciador até o inserto? Ao fim de 2005, o GenBank já apresentava mais de 52 milhões de seqüências depositadas em seu banco de dados (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html). Quantas bases, de cada uma dessas seqüências, não terão sido relativas ao vetor de clonagem? Se pensarmos que grande parte das seqüências produzidas nos laboratórios de pesquisa em todo o mundo jamais foi publicada em qualquer banco de dados, podemos concluir que os valores associados ao desperdício é ainda maior, incalculável. Quantos pesquisadores já terão se preocupado com este problema? Quão melhor seria hoje nossa compreensão dos transcriptomas dos organismos caso soubéssemos, desde o início, que valores menos rigorosos de poda permitem uma melhor identificação dos genes através de alinhamentos locais? O dbEST (Boguski et al., 1993) apresenta quase 39 milhões de seqüências (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html) que, muito provavelmente, poderiam ser estendidas nas bordas, propiciando uma mais fácil identificação e anotação dessas seqüências. A ciência capitalista morde o próprio rabo: a produção excessiva de conhecimento, sem tempo para análises minuciosas dos resultados, leva ao desperdício de dinheiro. A academia considera bons àqueles pesquisadores que produzem muito, geram uma infinidade dados e publicam artigos atrás de artigos. De fato, não parece importar se os cientistas questionam ou não seus dogmas, só importa que publiquem cada vez mais e mais. Desenfreada, segue a ciência acadêmica e nossa incessante busca pelo conhecimento.

Por outro lado, é virtualmente impossível, mesmo ao mais questionador dos cientistas, tentar se preocupar e indagar todas as variáveis envolvidas em seu trabalho. Aqui mesmo, fomos obrigados a aceitar alguns dogmas genômicos, sendo o principal deles: "o PHRED é o melhor algoritmo para a nomeação de bases existente". O questionamento dos dogmas apresenta níveis e mesmo nós, pretensos quebradores de dogmas científicos, estamos sujeitos a aceitá-los em diversos níveis.

Outra discussão que cabe no contexto da presente tese é aquela que trata da diferença entre método e ciência. Ao submeter um dos artigos componentes deste estudo a uma revista conceituada da área, o mesmo foi rejeitado rapidamente pelo editor, sem sequer encaminhar aos revisores, por pensar que o presente trabalho era "apenas um avanço técnico". O que caracteriza uma técnica? O que caracteriza um estudo científico? A nós parece claro que a ciência gera a técnica. A técnica é a aplicação da ciência que está muito bem consolidada. O sequenciamento de DNA já foi ciência, hoje é técnica: acredito que Fred Sanger não discordaria disso. Mas o sequenciamento de DNA ainda é ciência quando se busca aumentar a velocidade e eficiência do próprio método, haja vista publicações recentes, em revistas de alto impacto, tratando de novos e promissores métodos de sequenciamento de biomoléculas (Margulies et al., 2005; Ng et al., 2006; Pinard et al., 2006).

E o que é a bioinformática: ciência ou técnica? Parece-nos que boa parte do empreendimento em bioinformática está relacionada a observações que gerem questionamentos científicos e que levem à produção de uma técnica que seja utilizada universalmente pelos pesquisadores biólogos – até que apareçam outros cientistas a questionar esta técnica e apresentar uma solução melhor para o mesmo problema, modificando o paradigma Kuhniano vigente. Foi este tipo de trabalho que tentamos produzir aqui.

Além disso, o trabalho que apresentamos nesta tese não é, definitivamente, um estudo tradicional em biologia, considerando que não produz, diretamente, nenhum avanço no entendimento da biologia ou evolução de uma determinada espécie. Este também não é um trabalho em computação. Uso e abuso do desenvolvimento de algoritmos PERL e da construção e armazenamento da informação em bancos de dados MySQL ao longo desta tese, mas estou sempre a utilizar a computação como uma técnica, não como área de pesquisa. E se este também não é um trabalho em matemática, estatística ou medicina, só nos resta uma única conclusão: este é um trabalho puramente de bioinformática. Não está associado a nenhum organismo em especial, não pretende desenvolver nenhum algoritmo complexo, nenhuma teoria

matemática que explique um evento qualquer. Ele pretende padronizar métodos gerais de estudos genômicos de ampla utilização nas ciências biológicas.

Além, portanto, de ser um trabalho puro em bioinformática, este é também um trabalho importante pois remonta às bases de toda biologia molecular: a produção de uma seqüência de DNA fiel ao molde. Quase todas as outras aplicações – dentro da própria bioinformática (biomolecular) e de muitas das novas ciências "ômicas" que, aos poucos, vão surgindo e desenvolvendo-se como áreas sólidas da pesquisa científica – baseiam-se em análises de seqüências já produzidas, esquecendo-se do erro intrínseco em sua produção e considerando-as, simplesmente, como *perfeitamente corretas*. Portanto, a preocupação com a qualidade das seqüências produzidas é um critério de importância inquestionável, pois está na base tanto da própria bioinformática quanto de todas essas novas ciências moleculares.

Considerando a conhecida evolução nos métodos de nomeação de bases e produção de seqüências de DNA, vale a pena ressaltar e elogiar a iniciativa do Ensembl e do NCBI (*National Center for Biotechnology Information*) na construção de repositórios de dados brutos gerados pelos seqüenciadores, iniciativas estas conhecidas como *Ensembl Trace Server* (http://trace.ensembl.org) e *Trace Archive* (http://www.ncbi.nlm.nih.gov/Traces), dois serviços que compartilham dados diariamente. Esses serviços permitem aos pesquisadores a submissão dos arquivos gerados a partir do sequenciamento para que outros pesquisadores possam acessá-los e reanalisá-los, se desejado, através da execução de seu próprio procedimento de nomeação de bases, por exemplo. Dessa forma, caso alguns pesquisadores leiam os trabalhos publicados por outros e desejem reanalisar dados relativos a organismos de seu interesse, eles podem buscar os dados que estejam presentes nestes bancos de dados para isso. Além do mais, caso seja descoberto no futuro algum método sensivelmente mais eficiente para a nomeação das bases, os dados brutos já estarão automaticamente disponíveis aos pesquisadores para que a reanálise desses dados possa ser realizada. Atualmente tais bancos apresentam mais de um bilhão de *traces* disponíveis para *download*.

O trabalho apresentado aqui teve, como um de seus pontos fortes, o desenvolvimento de um método rigoroso para diminuir os efeitos particulares das reações de sequenciamento e permitir uma análise mais precisa do processo de nomeação de bases. O *single-pool sequencing* nos permitiu uma análise homogênea e eficiente das reações de sequenciamento realizadas, removendo problemas específicos que poderiam acontecer em uma ou outra molécula. Além disso, a utilização de uma seqüência de vetor de clonagem bem conhecida (pUC18) e extensivamente

sequenciada para funcionar como controle positivo mostrou-se extremamente eficaz, já que todos os outros trabalhos de averiguação de qualidade previamente publicados apresentavam, como controle, contigs genômicos (Huang et al., 2003; Batzoglou et al., 2002; Kent and Haussler, 2001; Huang and Madan, 1999) que, apesar de finalizados, ainda poderiam conter erros não-conhecidos que tenham prejudicado a análise precisa dos dados. Por outro lado, entretanto, a utilização da seqüência de um único vetor de clonagem neste trabalho pode ter gerado alguns problemas contexto-específicos e alterado de alguma forma o resultado produzido, sendo que seria muito interessante podermos testar nossas hipóteses e resultados em contextos de seqüências diferentes. Chegamos a realizar um projeto piloto com um outro vetor de clonagem, cujos dados foram gerados durante o sequenciamento de moléculas de cDNA de *Schistosoma mansoni* a partir de vetores que não haviam incorporado o inserto. Os resultados não divergiram muito da média esperado, mas como possuíamos apenas uma pequena quantidade desses dados, estatisticamente insignificante, e como tais moléculas não haviam sido produzidas através do processo de *single-pool sequencing*, preferimos não realizar especulações sobre contexto de seqüência antes de um estudo mais aprofundado, deixando este tema em aberto. No entanto, temos a intenção de podermos produzir novas seqüências de outros vetores de clonagem em *single-pool*, em um futuro próximo, e podermos comparar os resultados com estes apresentados aqui, de forma a corroborá-los (?) em um contexto mais amplo. Assim, esperamos ser capazes de publicar, tão logo quanto possível, um artigo de revisão mais completo tratando dos vários pontos já abordados, mas com um conjunto de evidências ainda mais sólido e diversificado. Todavia, é oportuno ressaltar que, embora o início da leitura de boa qualidade seja praticamente sincrônico, o que nos permitiu sugerir o melhor posicionamento do iniciador de seqüenciamento, a queda de qualidade se dá em diversas regiões do pUC18, o que equivale a contextos de seqüências diferentes. Ainda, verificamos que o mascaramento usando *crossmatch* das seqüências de vetor encontradas no final de leituras onde se usou PQV 8 (como sugerimos) é perfeito em casos onde o inserto é pequeno e a leitura alcança o vetor pelo outro lado, suportando nossos resultados.

Nesta tese, portanto, apresentamos vários trabalhos que tentaram trazer à luz da razão, através de experimentos bem delineados, a melhor forma de utilização de algoritmos de nomeação de bases e agrupamento de seqüências. Alguns dogmas que se acreditava em biologia molecular foram postos abaixo quando mostramos, por exemplo, que uma seqüência podada com PQV 8 permite um aproveitamento melhor da informação biológica e que este valor de poda é o mais adequando quando se

deseja reconhecer uma seqüência através de alinhamentos locais, como é o caso comum em projetos de descoberta gênica através do sequenciamento de ESTs (Artigo 4). Mostramos ainda que o efeito do re-sequenciamento da mesma molécula tem efeitos pequenos sobre a qualidade do consenso produzido e que a utilização de um valor rigoroso de *trimming* é a melhor forma de se obter moléculas fiéis à original, ao custo de produzi-las com um tamanho menor (Artigo 5). Identificamos, ainda, a posição ideal de posicionamento do iniciador para a realização de projetos genoma, algo que é pouco pensado pelos pesquisadores no momento de início dos projetos, mas que pode gerar grandes ganhos quantitativos, qualitativos e monetários ao fim do projeto (Artigo 3). A análise do local ótimo de posicionamento do iniciador pode evitar o desperdício da geração de seqüências do vetor de clonagem e, assim, aumentar a cobertura do genoma estudado com um mesmo número de *reads* seqüenciados, além de permitir encontrar, com precisão, regiões biologicamente importantes e limítrofes que podem estar presentes em seqüências de cDNA como, por exemplo, a seqüência do poli-A e o códon de iniciação da tradução.

A forma de apresentação dos dados realizada aqui permite ainda que os pesquisadores possam exercer seu livre-arbítrio e, analisando o resultado dos experimentos por nós produzidos, escolher diferentes parâmetros de PHRED de acordo com a pesquisa que desejem desenvolver. Observando os resultados produzidos, é possível definir padrões mais ou menos conservadores para cada caso estudado.

Então, como doutores em filosofia científica (PhD) que somos, acreditamos que a boa ciência é feita, primeiro, da criatividade gerada por um questionamento – que vem de se observar algum dado natural e perguntar-se: "*por que isso acontece desta forma*?" – e, baseando-se nesta pergunta proposta, ser capaz de imaginar *modelos experimentais bem delineados* baseados nos melhores controles possíveis, positivos e negativos, afastando dali as informações ruidosas do sistema. E é baseando-se nesta fórmula que acreditamos que os cientistas devam tentar peneirar das minas de dados, com cada vez mais eficiência, apenas o ouro mais precioso que é a *essência da natureza*.

7. REFERÊNCIAS BIBLIOGRÁFICAS *

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC. **Complementary DNA sequencing: expressed sequence tags and human genome project.** Science. 1991 Jun 21;252(5013):1651-6.
2. Adzhubei AA, Laerdahl JK, Vlasova AV. **preAssemble: a tool for automatic sequencer trace data processing.** BMC Bioinformatics. 2006 Jan 17;7:22.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. **Basic local alignment search tool**. J Mol Biol. 1990 Oct 5;215(3):403-10.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** Nucleic Acids Res. 1997 Sep 1;25(17):3389-402. Review.
5. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. **ARACHNE: a whole-genome shotgun assembler.** Genome Res. 2002 Jan;12(1):177-89.
6. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. **GenBank.** Nucleic Acids Res. 2006 Jan 1;34(Database issue):D16-20.
7. Berno AJ. **A graph theoretic approach to the analysis of DNA sequencing data.** Genome Res. 1996 Feb;6(2):80-91.
8. Brady D, Kocic M, Miller AW, Karger BL. **A maximum-likelihood base caller for DNA sequencing.** IEEE Trans Biomed Eng. 2000 Sep;47(9):1271-80.
9. Boguski MS, Lowe TM, Tolstoshev CM. **dbEST--database for "expressed sequence tags"**. Nat Genet. 1993 Aug;4(4):332-3.
10. Corpet F. **Multiple sequence alignment with hierarchical clustering.** Nucleic Acids Res. 1988 Nov 25;16(22):10881-90.
11. Davies K. **Decifrando o genoma:** a corrida para desvendar o DNA humano. Companhia das Letras 2001. 469p.
12. Ewing, B. and Green, P. (1998) **Base-Calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 8: 186-94.
13. Ewing, B., Hillier, L., Wendl, M.C. and Green P. (1998) **Base-Calling of automated sequencer traces using phred. I. Accuracy Assessment**. *Genome Res* 8: 175-85.
14. Fichant GA, Quentin Y. **A frameshift error detection algorithm for DNA sequencing projects.** Nucleic Acids Res. 1995 Aug 11;23(15):2900-8.
15. Felsenfeld A, Peterson J, Schloss J, Guyer M. **Assessing the quality of the DNA sequence from the Human Genome Project.** Genome Res. 1999 Jan;9(1):1-4.
16. Giddings MC, Brumley RL Jr, Haker M, Smith LM. **An adaptive, object oriented strategy for base calling in DNA sequence analysis.** Nucleic Acids Res. 1993 Sep 25;21(19):4530-40.
17. Green P. **Documentation for PHRAP and cross-match.** http://www.phrap.org/phrap.docs/phrap.html. 1998.
18. Guan X, Uberbacher EC. **Alignments of DNA and protein sequences containing frameshift errors.** Comput Appl Biosci. 1996 Feb;12(1):31-40.
19. He H, McGown LB. **DNA sequencing by capillary electrophoresis with four-decay fluorescence detection.** Anal Chem. 2000 Dec 15;72(24):5865-73.
20. Huang X, Madan A. **CAP3: A DNA sequence assembly program.** Genome Res. 1999 Sep;9(9):868-77.

21. Huang X, Wang J, Aluru S, Yang SP, Hillier L. **PCAP: a whole-genome assembly program.** Genome Res. 2003 Sep;13(9):2164-70.
22. Kent WJ, Haussler D. **Assembly of the working draft of the human genome with GigAssembler.** Genome Res. 2001 Sep;11(9):1541-8.
23. Kim S, Segre AM. **AMASS: a structured pattern matching approach to shotgun sequence assembly.** J Comput Biol. 1999 Summer;6(2):163-86.
24. Kuhn TS. **The Structure of Scientific Revolutions**. The University of Chicago, 1962.
25. Lawrence CB, Solovyev VV. **Assignment of position-specific error probability to primary DNA sequence data.** Nucleic Acids Res. 1994 Apr 11;22(7):1272-80.
26. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J. **The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes.** Nucleic Acids Res. 2005 Jan 1;33(Database issue):D71-4.
27. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. **Genome sequencing in microfabricated high-density picolitre reactors.** Nature. 2005 Sep 15;437(7057):376-80. Epub 2005 Jul 31.
28. Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S. **EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments.** Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W459-62.
29. McGinnis S, Madden TL. **BLAST: at the core of a powerful and diverse set of sequence analysis tools.** Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W20-5.
30. Medigue C, Rose M, Viari A, Danchin A. **Detecting and analyzing DNA sequencing errors: toward a higher quality of the Bacillus subtilis genome sequence.** Genome Res. 1999 Nov;9(11):1116-27.
31. Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA. **A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base.** Genome Res. 1999 Nov;9(11):1143-55.
32. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC. **A whole-genome assembly of Drosophila.** Science. 2000 Mar 24;287(5461):2196-204.
33. Needleman SB, Wunsch CD. **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** J Mol Biol. 1970 Mar;48(3):443-53.
34. Ng P, Tan JJ, Ooi HS, Lee YL, Chiu KP, Fullwood MJ, Srinivasan KG, Perbost C, Du L, Sung WK, Wei CL, Ruan Y. **Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes.** Nucleic Acids Res. 2006 Jul 13;34(12):e84.
35. Oliveira G, Johnston DA. **Mining the schistosome DNA sequence database.** Trends Parasitol. 2001 Oct;17(10):501-3.
36. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, Kimura K, Makita H, Sekine M, Obayashi M, Nishi T, Shibahara T, Tanaka T, Ishii S, Yamamoto J, Saito

K, Kawai Y, Isono Y, Nakamura Y, Nagahari K, Murakami K, Yasuda T, Iwayanagi T, Wagatsuma M, Shiratori A, Sudo H, Hosoiri T, Kaku Y, Kodaira H, Kondo H, Sugawara M, Takahashi M, Kanda K, Yokoi T, Furuya T, Kikkawa E, Omura Y, Abe K, Kamihara K, Katsuta N, Sato K, Tanikawa M, Yamazaki M, Ninomiya K, Ishibashi T, Yamashita H, Murakawa K, Fujimori K, Tanai H, Kimata M, Watanabe M, Hiraoka S, Chiba Y, Ishida S, Ono Y, Takiguchi S, Watanabe S, Yosida M, Hotuta T, Kusano J, Kanehori K, Takahashi-Fujii A, Hara H, Tanase TO, Nomura Y, Togiya S, Komai F, Hara R, Takeuchi K, Arita M, Imose N, Musashino K, Yuuki H, Oshima A, Sasaki N, Aotsuka S, Yoshikawa Y, Matsunawa H, Ichihara T, Shiohata N, Sano S, Moriya S, Momiyama H, Satoh N, Takami S, Terashima Y, Suzuki O, Nakagawa S, Senoh A, Mizoguchi H, Goto Y, Shimizu F, Wakebe H, Hishigaki H, Watanabe T, Sugiyama A, Takemoto M, Kawakami B, Yamazaki M, Watanabe K, Kumagai A, Itakura S, Fukuzumi Y, Fujimori Y, Komiyama M, Tashiro H, Tanigami A, Fujiwara T, Ono T, Yamada K, Fujii Y, Ozaki K, Hirao M, Ohmori Y, Kawabata A, Hikiji T, Kobatake N, Inagaki H, Ikema Y, Okamoto S, Okitani R, Kawakami T, Noguchi S, Itoh T, Shigeta K, Senba T, Matsumura K, Nakajima Y, Mizuno T, Morinaga M, Sasaki M, Togashi T, Oyama M, Hata H, Watanabe M, Komatsu T, Mizushima-Sugano J, Satoh T, Shirai Y, Takahashi Y, Nakagawa K, Okumura K, Nagase T, Nomura N, Kikuchi H, Masuho Y, Yamashita R, Nakai K, Yada T, Nakamura Y, Ohara O, Isogai T, Sugano S. **Complete sequencing and characterization of 21,243 full-length human cDNAs.** Nat Genet. 2004 Jan;36(1):40-5.

37. Peltola H, Soderlund H, Ukkonen E. **SEQAID: a DNA sequence assembling program based on a mathematical model.** Nucleic Acids Res. 1984 Jan 11;12(1 Pt 1):307-21.
38. Pevzner PA, Tang H, Waterman MS. **An Eulerian path approach to DNA fragment assembly.** Proc Natl Acad Sci U S A. 2001 Aug 14;98(17):9748-53.
39. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, Egholm M, Rothberg JM, Leamon JH. **Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing.** BMC Genomics. 2006 Aug 23;7:216.
40. Pontius JU, Wagner L, Schuler GD. **Unigene: a unified view of the transcriptome.** *In*: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information.
41. Prosdocimi F, Faria-Campos AC, Peixoto FC, Pena SD, Ortega JM, Franco GR. **Clustering of *Schistosoma mansoni* mRNA sequences and analysis of the most transcribed genes: implications in metabolism and biology of different developmental stages.** Mem Inst Oswaldo Cruz. 2002;97 Suppl 1:61-9.
42. Prosdocimi F, Cerqueira GC, Binneck E, Silva AF, Reis AN, Junqueira ACM, Santos ACF, Nhani-Júnior A, Wust CI, Camargo-Filho F, Kessedjian JL, Petretski JH, Camargo LP, Ferreira RGM, Lima RP, Pereira RM, Jardim S, Sampaio VS and Folgueras-Flatschart AV. **Bioinformática: manual do usuário.** Biotec Ci Des 29: 18-31, 2002.
43. Sanger F, Coulson AR. **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** J Mol Biol. 1975 May 25;94(3):441-8.
44. Sanger F, Nicklen S, Coulson AR. **DNA sequencing with chain-terminating inhibitors.** Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7
45. Scheetz TE, Trivedi N, Roberts CA, Kucaba T, Berger B, Robinson NL, Birkett CL, Gavin AJ, O'Leary B, Braun TA, Bonaldo MF, Robinson JP, Sheffield VC, Soares MB, Casavant TL. **ESTprep: preprocessing cDNA sequence reads.** Bioinformatics. 2003 Jul 22;19(11):1318-24.
46. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Edwards C, Fan JB, Fang N, Fizames

C, Garrett C, Green L, Hadley D, Harris M, Harrison P, Brady S, Hicks A, Holloway E, Hui L, Hussain S, Louis-Dit-Sully C, Ma J, MacGilvery A, Mader C, Maratukulam A, Matise TC, McKusick KB, Morissette J, Mungall A, Muselet D, Nusbaum HC, Page DC, Peck A, Perkins S, Piercy M, Qin F, Quackenbush J, Ranby S, Reif T, Rozen S, Sanders C, She X, Silva J, Slonim DK, Soderlund C, Sun WL, Tabar P, Thangarajah T, Vega-Czarny N, Vollrath D, Voyticky S, Wilmer T, Wu X, Adams MD, Auffray C, Walter NA, Brandon R, Dehejia A, Goodfellow PN, Houlgatte R, Hudson JR Jr, Ide SE, Iorio KR, Lee WY, Seki N, Nagase T, Ishikawa K, Nomura N, Phillips C, Polymeropoulos MH, Sandusky M, Schmitt K, Berry R, Swanson K, Torres R, Venter JC, Sikela JM, Beckmann JS, Weissenbach J, Myers RM, Cox DR, James MR, Bentley D, Deloukas P, Lander ES, Hudson TJ. **A gene map of the human genome.** Science. 1996 Oct 25;274(5287):540-6.

47. Smith TM, Abajian C, Hood L. **Hopper: software for automating data tracking and flow in DNA sequencing.** Comput Appl Biosci. 1997 Apr;13(2):175-82.
48. Smith, T.F. and Waterman, M. S. (1981) **Identification of common molecular subsequences**. *J Mol Biol* 147: 195-7.
49. Song JM, Yeung ES. **Alternative base-calling algorithm for DNA sequencing based on four-label multicolor detection.** Electrophoresis. 2000 Mar;21(4):807-15.
50. Staden R. **The Staden sequence analysis package.** Mol Biotechnol. 1996 Jun;5(3):233-41.
51. Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF, Zeeberg B, Buetow KH, Schaefer CF, Bhat NK, Hopkins RF, Jordan H, Moore T, Max SI, Wang J, Hsieh F, Diatchenko L, Marusina K, Farmer AA, Rubin GM, Hong L, Stapleton M, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Brownstein MJ, Usdin TB, Toshiyuki S, Carninci P, Prange C, Raha SS, Loquellano NA, Peters GJ, Abramson RD, Mullahy SJ, Bosak SA, McEwan PJ, McKernan KJ, Malek JA, Gunaratne PH, Richards S, Worley KC, Hale S, Garcia AM, Gay LJ, Hulyk SW, Villalon DK, Muzny DM, Sodergren EJ, Lu X, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madan A, Young AC, Shevchenko Y, Bouffard GG, Blakesley RW, Touchman JW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Krzywinski MI, Skalska U, Smailus DE, Schnerch A, Schein JE, Jones SJ, Marra MA; Mammalian Gene Collection Program Team. **Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences.** Proc Natl Acad Sci U S A. 2002 Dec 24;99(26):16899-903.
52. Strausberg RL, Feingold EA, Klausner RD, Collins FS. **The mammalian gene collection.** Science. 1999 Oct 15;286(5439):455-7.
53. Tammi MT, Arner E, Britton T, Andersson B. **Separation of nearly identical repeats in shotgun assemblies using defined nucleotide positions, DNPs.** Bioinformatics. 2002 Mar;18(3):379-88.
54. Tammi MT, Arner E, Kindlund E, Andersson B. **Correcting errors in shotgun sequences.** Nucleic Acids Res. 2003 Aug 1;31(15):4663-72.
55. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. **Serial analysis of gene expression.** Science. 1995 Oct 20;270(5235):484-7.
56. Ye J, McGinnis S, Madden TL. **BLAST: improvements for better sequence analysis.** Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W6-9.
57. Walther D, Bartha G, Morris M. **Basecalling with LifeTrace.** Genome Res. 2001 May;11(5):875-88.
58. Wendl MC, Dear S, Hodgson D, Hillier L. **Automated sequence preprocessing in a large-scale sequencing environment.** Genome Res. 1998 Sep;8(9):975-84.
59. White O, Dunning T, Sutton G, Adams M, Venter JC, Fields C. **A quality control algorithm for DNA sequencing projects.** Nucleic Acids Res. 1993 Aug 11;21(16):3829-38.

60. Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF. **Protein sequence similarity searches using patterns as seeds.** Nucleic Acids Res. 1998 Sep 1;26(17):3986-90.

* (Outras referências nos trabalhos da seção Resultados)

PRODUÇÃO CIENTÍFICA DURANTE O DOUTORADO

Artigos Científicos Publicados em Revistas Internacionais

1. *A set of amino acids found to occur more frequently in human and fly than in plant and yeast proteomes consists of non-essential amino acids.* Comp. Biol. Med: **2006** (in press). **Prosdocimi F**, Mudado M, Ortega JM.
2. *Efficient secondary database driven annotation using model organism sequences.* In Silico Biol 6(5):363-72, **2006**. Faria-Campos AC, Campos SVA, **Prosdocimi F**, Franco GC, Franco GR, Ortega JM.
3. *Genetic algorithm for analysis of mutations in Parkinson's disease.* Artif Intell Med, 35:227-41, **2005**. Smigrodski R, Goertzel B, Pennachin C, Coelho L, **Prosdocimi F**, Parker DW Jr.
4. *Accessing optimal primer distance from insert.* In Silico Biol 5(5-6):469-77, **2005**. **Prosdocimi F**, Ortega JM
5. *Diet as a pressure on the amino acid content of proteomes.* Lec Not Comp Sci 3594:153-159, **2005**. **Prosdocimi F**, Ortega JM
6. *Nucleic acid binding properties of SmZF1, a zinc finger protein of* Schistosoma mansoni. Int J Parasitol. 34: 1211-9, **2004**. Calzavara-Silva CE, **Prosdocimi F**, Abath FGC, Pena SD, Franco GR.

Artigos Científicos Publicados em Revistas Nacionais

1. *Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values.* Gen Mol Res 3(4):483-92, **2004**. **Prosdocimi F**, Peixoto FC, Ortega JM.
2. *DNA Sequences Base Calling by PHRED: Error Pattern Analysis.* RTInfo 3: 107-10, **2003**. **Prosdocimi F**, Peixoto FC, Ortega JM.
3. Rede mineira de sequenciamento: estudo do transcriptoma do parasita *Schistosoma mansoni*. Bioscience J Especial-2004: 93-100, **2003**. Franco GR, **Prosdocimi F**, Faria-Campos AC, Ortega JM.
4. Ferramentas bioinformáticas aplicadas à caracterização da expressão gênica. Bioscience J Especial-2004: 109-117, **2003**. Faria-Campos AC, Mudado M, Peixoto FC, Bravo-Neto E, **Prosdocimi F**, Ortega JM.

Artigos no Prelo

1. *Effects of the number of reads and trimming on quality and size of assembled consensi*. (2006). Gen Mol Res, **2006**. **Prosdocimi F**, Lopes DAO, Peixoto FC, Ortega JM.

Artigos Submetidos

1. *Setting PHRED scores to obtain maximum biological information.* (2006) **Prosdocimi F**, Peixoto FC, Ortega JM.

Artigos de Divulgação Científica Publicados

1. *¡Todos somos transgenicos!* **Prosdocimi F**. Revista PENSAR 3 (3): 11-15, **2006**.
2. *Sobre bioinformática, genoma e ciência*. Ciência Hoje 35 (209): 54-57, **2004**. **Prosdocimi F**, Santos FR.
3. *O DNA vai à escola – Seção Bioinformática*. **Prosdocimi F**. Colaboração na escrita da seção de bioinformática do site da pesquisadora Márcia Lachtermacher-Triunfol
   http://www.odnavaiaescola.com/banco.htm
   http://www.odnavaiaescola.com/identificadores.htm
   http://www.odnavaiaescola.com/anotacao.htm
   http://www.eldnavaalaescuela.com/modulo4.pdf

Trabalhos apresentados em congressos

1. Successful clustering of ortholog groups by Bidirectional Best Hit (BBH) using organisms modeled from a single ancestral via stepwise mutation. **Prosdocimi F**, Mudado M, Ortega JM. 2006. XIV ISMB (Intelligent Systems for Molecular Biology), Fortaleza.
2. How protein evolution measured by similarity change and stop codon incidence depends on the genetic code. **Prosdocimi F**, Melo H, Capanema ER, Ortega JM. 2006. XIV ISMB (Intelligent Systems for Molecular Biology), Fortaleza.
3. Bioinformatics analyses of *Schistosoma mansoni* ESTs generated from trans-spliced enriched cDNA libraries. Mourão MM, Lobo FP, **Prosdocimi F**, Franco GR. 2006. XIV ISMB (Intelligent Systems for Molecular Biology), Fortaleza.

4. The simulation of microsatellite haplotype evolution: comparison between genealogy made without and with bottleneck. Freitas L, **Prosdocimi F**, Santos FR. 2005. X-meeting -- 1st international conference of the AB^3C, Caxambu.
5. Diet as a pressure on the amino acid content of proteomes. **Prosdocimi F**, Ortega JM. 2005 - BSB (Brazilian symposium on bioinformatics) - São Leopoldo-RS – Prêmio de **melhor trabalho** apresentado no congresso.
6. Where do I put my sequencing *primer*? **Prosdocimi F**, Ortega JM. 2005 - XXXIV Reunião anual da SBBq – Águas de Lindóia - SP
7. Effects of the number of reads and trimming on quality and size of assembled consensi. **Prosdocimi F**, Lopes DAO, Peixoto FC, Ortega JM. 2004 - International Congress of Bioinformatics and Computational Biology (ICoBiCoBi) - Angra dos Reis-RJ
8. Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values. **Prosdocimi F**, Peixoto FC, Ortega JM. 2004 - International Congress of Bioinformatics and Computational Biology (ICoBiCoBi) - Angra dos Reis-RJ
9. modEST: a simulator of cDNA library construction and EST sequencing. Peixoto F, Oliveira A, **Prosdocimi F**, Carvalho O, Ortega JM. 2004 - International Congress of Bioinformatics and Computational Biology (ICoBiCoBi) - Angra dos Reis-RJ
10. TGFinder automated identification of genes controlled by transcription factors using *Drosophila melanogaster* as a model. **Prosdocimi F**, Calzavara-Silva CE, SANTOS FR, Franco GR. 2004 - International Congress of Bioinformatics and Computational Biology (ICoBiCoBi) - Angra dos Reis-RJ
11. Does diet influence genome? -- a bioinformatics approach on the use of essential amino acids in proteins. **Prosdocimi F**, Ortega JM. 2004 - XXXIII Reunião anual da SBBq - Caxambu-MG
12. MGSim: A simulator of genealogies using Microsatelite Loci. **Prosdocimi F**, Santos FR. 2004 - 50º Congresso Brasileiro de Genética - Florianopolis - SC
13. SAABs e TGFinder: new tools for functional studies of SmZF1, a putative transcription factor of *Schistosoma mansoni*. Drummond MG, Calzavara-Silva CE, **Prosdocimi F**, Franco GR. 2004 – 50º Congresso Brasileiro de Genética - Florianopolis - SC
14. TGFinder: an algorithm to find target genes for transcription factors and its applications. **Prosdocimi F**, Calzavara-Silva CE, Santos FR, Franco GR. 2004 - II Enapebi- Belo Horizonte -MG

15. Tentative to produce high quality DNA molecules using the softwares PHRED, PHRAP and CAP3. Lopes DAO, **Prosdocimi F**, Ortega JM. 2003 - XXXII Reunião anual da SBBq - Caxambu-MG
16. DNA Sequences Base Calling by PHRED: Error Pattern Analysis. **Prosdocimi F**, Peixoto F, Ortega JM. 2003 - II WOB - Macaé
17. Mining GenBank and dbEST *Schistosoma mansoni* sequences. **Prosdocimi F**, Faria-Campos AC, Franco GR. Apresentação oral - International Conference on Bioinformatics and Computational Biology (ICoBiCoBi) - 2003 - Ribeirão Preto
18. Candidate genes for the Late Onset Alzheimer disease in human chromosome 10. **Prosdocimi F** et al. Apresentação oral - International Conference on Bioinformatics and Computational Biology (ICoBiCoBi) - 2003 - Ribeirão Preto
19. Effects of Primer Positioning and Trimming Algorithms in EST Information Retrieval. **Prosdocimi F**, Peixoto F, Ortega JM. 2003 - XXXII Reunião anual da SBBq - Caxambu.

ANEXOS

Infelizmente a forma de apresentação de uma tese de doutorado, conforme padrão acadêmico vigente, não permite que mais de um tema principal seja abordado no corpo principal do referido documento. Por isso, tomamos a liberdade de apresentar aqui ao menos dois outros trabalhos que realizamos e que consideramos de relevância particular, embora não ligados ao tema central da presente tese.

O primeiro deles foi publicado na revista "Ciência Hoje" e vem ressaltar a importância da divulgação científica. A ciência não é e não deve ser um empreendimento que apenas os cientistas entendam: ela deve ser discutida e entendida por tantos indivíduos quanto seja possível. Consideramos dever do cientista levar ao público leigo e à sociedade, de maneira geral, seus trabalhos e seus questionamentos. Dessa forma, apresentamos o artigo "Sobre a bioinformática, o genoma e a ciência" anexo.

Já nosso segundo trabalho anexo mostra um exemplo de utilização da bioinformática como uma ciência de fato. Este trabalho foi produzido através de uma pergunta que só poderia ter sido desenvolvida quando observados dados de seqüências de biomoléculas, um questionamento próprio da *bioinformática lupa*. Sem a observação das seqüências, jamais teríamos pensado ou podido responder à seguinte pergunta: "Será que existe algum viés para que os organismos que usam aminoácidos essenciais estejam trocando os aminoácidos de suas proteínas de forma que usem menos desses aminoácidos e mais daqueles que conseguem produzir independentemente de sua dieta?" Como se não bastasse, a observação e a hipótese terem surgido da análise de seqüências, também a investigação baseou-se apenas em experimentos *in silico* e os modestos resultados obtidos também não se valeram de outras análises senão as computacionais. Uma versão inicial deste trabalho de bioinformática e genômica comparativa ganhou o prêmio de *melhor trabalho* apresentado no congresso *Work on Bioinformatics II*, realizado no Rio Grande do Sul, em 2005. Posteriormente, realizamos análises mais extensas e publicamos, na revista *Computers in Biology and Medicine* (2006), o artigo anexo, entitulado "*A set of amino acids found to occur more frequently in human and fly than in plant and yeast proteomes consists of non-essential amino acids*". Este formato de investigação está ligado às nossas perspectivas de estudos futuros e, portanto, achamos pertinente apresentá-lo aqui.

Francisco Prosdocimi

Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais

Fabrício R. Santos

Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais

Sobre bioinformática, genoma e ciência

As notícias sobre o seqüenciamento do genoma do homem e de outras espécies, e sobre a identificação de genes e de suas funções, tornaram-se freqüentes nos últimos anos. As informações genômicas acumulam-se com tal rapidez que, para sua análise e detalhamento, são necessários processos cada vez mais poderosos e criativos. Essa é a tarefa da bioinformática, ciência que trouxe uma abordagem científica aos dados basicamente descritivos obtidos pelas máquinas seqüenciadoras.

A bioinformática consiste principalmente na análise computacional de seqüências de DNA, RNA e proteínas. Essa nova ciência surgiu na última década em função da necessidade de ferramentas sofisticadas para analisar o crescente volume de dados gerado em biologia molecular. O GenBank, criado no centro norte-americano para informação biotecnológica (NCBI, na sigla em inglês), foi um dos primeiros e ainda é o mais popular banco de dados para o depósito de seqüências de DNA. É lá que pesquisadores de todo o mundo depositam as seqüências de As, Cs, Gs e Ts (iniciais das 'peças' básicas da molécula de DNA, os nucleotídeos adenina, citosina, guanina e timina) que obtêm ao seqüenciar o genoma dos mais diversos organismos.

No final dos anos 90 observou-se um crescimento exponencial do número de seqüências de biomoléculas depositadas no GenBank (figura 1). Esse aumento teve início a partir de 1990, quando surgiram os seqüenciadores de DNA a *laser*, totalmente automatizados. Tais máquinas têm com freqüência 96 capilares (tubos minúsculos por onde passam os fragmentos de DNA) e podem 'ler', em média, 550 letras (A, C, G e T) por capilar em cada análise. Há cerca de 3 bilhões dessas letras no genoma humano. Seqüenciadores ainda mais potentes, com 384 capilares, podem 'ler' mais de um milhão de letras do DNA por dia!

Existem no Brasil dezenas de seqüenciadores, grande parte deles distribuída entre laboratórios de todo o país quando do início do Projeto Genoma da Fundação de Amparo à Pesquisa do Estado de São Paulo, que seqüenciou o DNA da bactéria *Xylella fastidiosa*, praga da laranja (http://aeg.lbi.ic.unicamp.br/xf/), e o Projeto Genoma Brasileiro (http://www.brgene.lncc.br), que já permitiu o seqüenciamento dos DNAs das bactérias *Chromobacterium violaceum* e *Mycoplasma synoviae*.

A grande maioria das seqüências publicadas em bancos de dados internacionais vem de pro-

jetos genoma e transcriptoma (ou de genoma funcional). O primeiro genoma seqüenciado foi o da bactéria *Haemophilus influenzae*, em meados de 1995. Hoje, o NCBI já contém 1.628 genomas de vírus, 174 genomas de procariotos (bactérias e arqueobactérias) e 20 genomas de organismos eucarióticos. Essa imensa quantidade de informação vem se tornando cada vez mais complexa com o estudo das interações entre biomoléculas e das variações existentes entre os indivíduos de uma população (figura 2). Mas, afinal, que informações cientificamente relevantes o genoma trouxe para os pesquisadores, para as pessoas e para a sociedade? Será que projetos genoma são pesquisas meramente descritivas? Então, qual a relevância da genômica e qual o papel da bioinformática na consolidação dessa ciência?

À primeira vista os estudos de genoma não parecem ser pesquisas científicas clássicas, pois não se baseiam em hipóteses elaboradas *a priori* sobre a biologia de um dado organismo. No máximo, a pergunta que se poderia fazer antes de seqüenciar um genoma seria "esse organismo tem algum gene de potencial biotecnológico?" ou "o que há nos genes desse organismo que o faz conseguir viver nessa condição, ou que gera uma doença?". Tais perguntas, porém, dificilmente serão respondidas diretamente pelo seqüenciamento, e certamente exigirão estudos posteriores. E mais: é possível que alguma investigação não-genômica mais minuciosa sobre esse ou aquele aspecto em particular possa esclarecer de modo mais direto essas questões.

Mas isso não tira o mérito dos estudos genômicos. Acreditamos que a ciência vive hoje a era da anatomia molecular. No século 19, quando pouco se conhecia – de forma sistemática – do mundo biológico, os naturalistas, que exploravam o mundo em busca de informação taxonômica, encontrando e classificando animais e plantas então desconhecidos, podiam ser considerados os cientistas da biologia. A descrição e a documentação de novas espécies era especialmente necessária naquela época, já que pouco ou nada se conhecia sobre a grande diversidade da vida na Terra. O mesmo ocorreu quando surgiram os anatomistas, que pela primeira vez documentaram em detalhe o interior do corpo do homem e de outros animais. Eles apenas descreviam, da melhor maneira possível à época, a localização dos órgãos e tecidos.

Portanto, se a genômica não pode ser considerada, classicamente, uma ciência, também não podem ser vistas assim a taxonomia e a anatomia, já que todas elas são empreendimentos principalmente descritivos, e não investigativos. Mas isso, mais uma vez, não lhes tira o mérito. Quanto conhecimento científico foi construído com base nas informações geradas por naturalistas e anatomistas? A ciência biomédica foi montada a partir do trabalho dos anatomistas, e a teoria mais importante e unificadora de toda a biologia – a evolução – surgiu diretamente das observações e estudos descritivos dos naturalistas Charles Darwin (1809-1882) e Alfred Wallace (1823-1913).

E a genômica? O genoma, na verdade, pode ser descrito como a 'anatomia molecular' de uma espécie. E só agora, neste início de século 21, somos capazes de desvendar e descrever como as espécies são constituídas em seu nível mais básico, o da informação molecular. A genômica, portanto, é a 'ciência descritiva' atual. E assim como as ciências biomédicas trouxeram o método científico ao estudo da anatomia, a bioinformática veio trazer cientificidade aos dados genômicos.

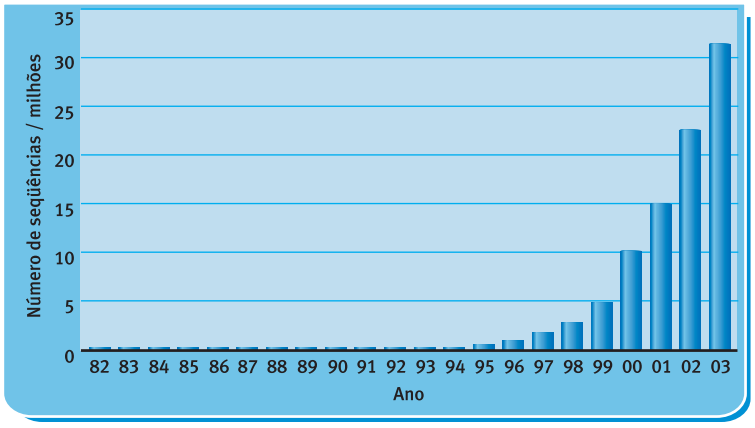


Figura 1. Crescimento do número de seqüências de biomoléculas depositadas no GenBank, o mais popular banco de dados de DNA

Se a genômica não pode ser considerada, classicamente, uma ciência, também não podem ser vistas assim a taxonomia e a anatomia, já que todas elas são empreendimentos principalmente descritivos, e não investigativos

A bioinformática traz uma abordagem científica aos dados gerados em projetos genoma, como já fazem outras ciências bem estabelecidas, como a biologia molecular, a genética e a bioquímica

É importante definirmos bem o que é a bioinformática e em que contexto esse conceito é usado neste ensaio. Muitos crêem que essa ciência consista em qualquer análise computacional de problemas biológicos, mas isso não está de acordo com sua origem. A bioinformática clássica surgiu com o seqüenciamento de biomoléculas e destas permanece inseparável. É possível propor uma definição razoavelmente clara: a bioinformática consiste em 'todo tipo de estudo ou de ferramenta computacional que se pode realizar e/ou produzir de forma a organizar ou obter informação biológica a partir de seqüências de biomoléculas'. Se o estudo envolve seqüências de biomoléculas (DNA, RNA ou proteínas), direta ou indiretamente, trata-se de bioinformática. Se não, trata-se de computação aplicada à biologia, que é extremamente importante em várias áreas e já existia bem antes do início dos seqüenciamentos de biomoléculas.

Definido o conceito de bioinformática que utilizamos, podemos enquadrar muitos estudos nessa área em três princípios paradigmáticos, aos quais daremos os nomes metafóricos de 'tijolo', 'peneira' e 'lupa'.

Estudos de bioinformática 'tijolo' são os relacionados à execução de projetos genoma e normalmente produzem processos para analisar seqüências e interpretar genomas. Algumas dessas ferramentas já são clássicas. Podemos citar o *base calling*, onde as bases do DNA são lidas no seqüenciador a partir dos cromatogramas (perfis de emissão fluorescente que variam entre os nucleotídeos A, C, G e T). Cada cromatograma é transformado em uma seqüência, e um índice de confiabilidade é associado a cada letra do DNA. Em seguida analisam-se as seqüências que têm parte das letras em comum, para eliminar as sobreposições, alinhar os trechos corretos e com isso gerar o 'texto' completo do genoma da espécie estudada (que pode ter milhões ou bilhões de letras). Novas ferramentas para 'conferir' seqüências, alinhá-las (na montagem de um genoma), identificar genes e padronizar processos de *base calling* são alguns exemplos de projetos de bioinformática 'tijolo', sem os quais é impossível a análise sistemática dos 'edifícios genômicos'.

Vale observar que as ferramentas de comparação de seqüências de DNA têm permitido um grande avanço na identificação das funções de genes. Nesse caso, a seqüência de um novo gene é comparada com aquelas armazenadas em um banco de dados de genes de função conhecida, permitindo a rápida dedução da possível função desse gene recém-seqüenciado. Testes experimentais para descobrir a função de cada novo gene descoberto possivelmente exigiriam várias décadas de pesquisa.

A quantidade de informações gerada por um projeto genoma torna virtualmente impossível a análise destas (ou de uma pequena parcela) pelo grupo que gerou essa seqüência completa de DNA. Assim, trabalhos posteriores, envolvendo fragmentos de diferentes genomas, serão necessários para analisar temas específicos (por exemplo, proteínas envolvidas no metabolismo de açúcares). Esses trabalhos de mineração de dados genômicos são característicos da chamada bioinformática 'peneira'.

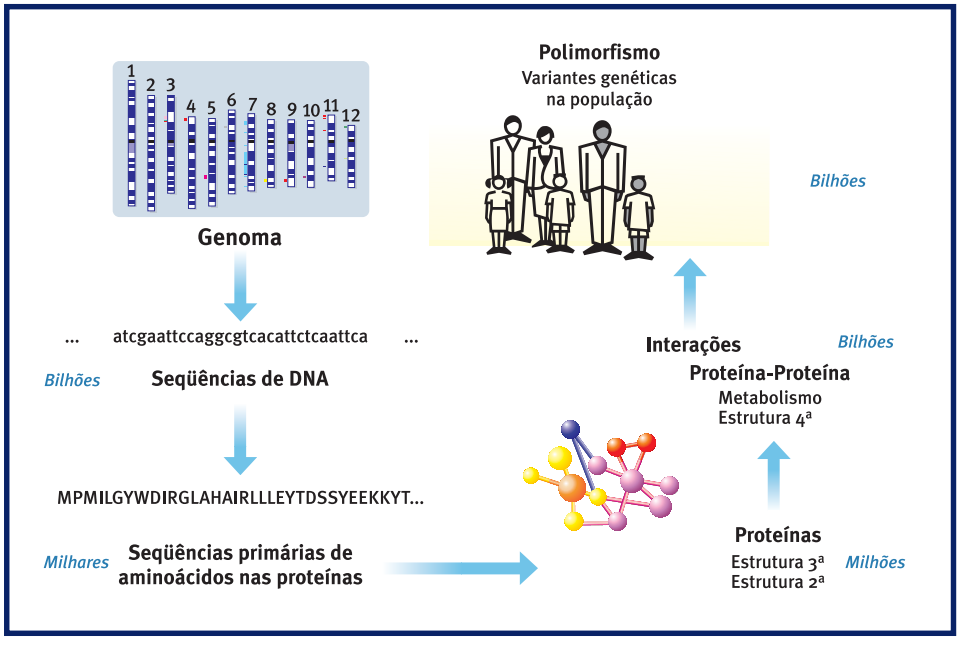


Figura 2. O estudo das interações entre diferentes biomoléculas e das variações genéticas presentes na população, torna mais complexa a imensa quantidade de informação gerada pelos projetos genoma

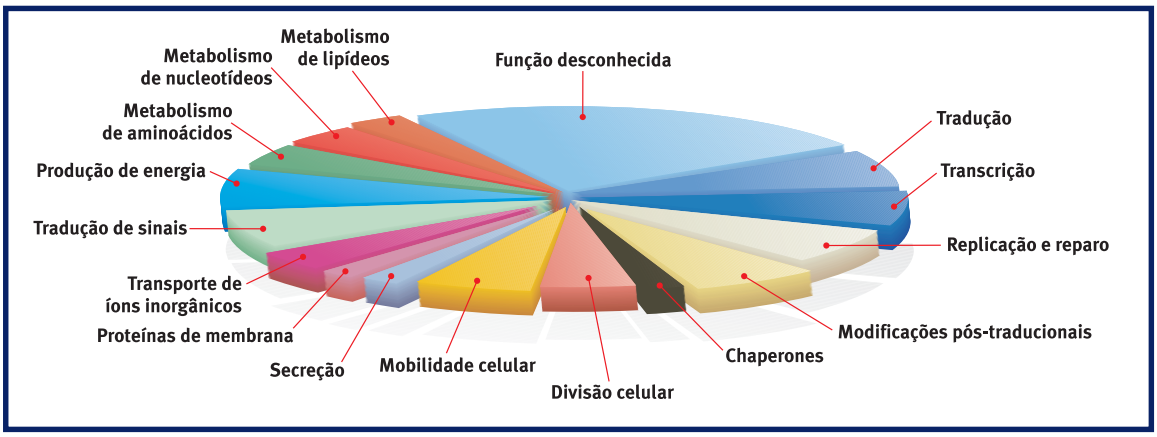


Figura 3. Funções biológicas dos genes já identificadas em estudos genéticos

Como a genômica é em essência uma disciplina descritiva, os trabalhos dessa área exibem muitos dados sem qualquer detalhamento, muitas vezes por limitação do periódico que os publica. A divisão dos genes em grupos, de acordo com sua função biológica (figura 3), é um exemplo da informação descritiva freqüentemente presente em artigos de genoma. Que informação relevante se poderia tirar desse monte de dados? Usando 'peneiras' específicas, os cientistas podem gerar conhecimento mais aprofundado sobre aspectos de seu interesse. A construção de bancos de dados de seqüências de genes que tenham uma ou outra função específica ou de estruturas tridimensionais de proteínas, por exemplo, também está incluída no âmbito da bioinformática 'peneira'. Todo ano, a primeira edição da revista britânica *Nucleic Acids Research* traz um resumo dos bancos de dados mais utilizados na área.

Entretanto, é nos trabalhos de bioinformática 'lupa' que a ciência aparece com maior clareza na genômica. Vale ressaltar que os estudos de genoma e bioinformática citados até agora são indispensáveis para o aumento do conhecimento científico sobre os organismos e sua constituição molecular. Nos estudos do tipo 'lupa', porém, o método científico é rigorosamente aplicado. Aqui, empregando as mais variadas ferramentas computacionais, o processo investigativo científico é retomado: observam-se os dados, criam-se hipóteses e realizam-se experimentos *in silico* (dentro do computador) para comprová-las ou refutá-las através de algoritmos (processos de cálculo que permitem solucionar problemas) bioinformáticos.

É interessante verificar que estudos 'lupa' não são necessariamente publicados em revistas especializadas em bioinformática. Isso acontece porque os algoritmos usados nesses estudos são vistos apenas como a metodologia de um trabalho que tenta buscar um resultado biológico mais específico. A bioinformática não é o centro do trabalho, como ocorre nas abordagens 'tijolo' e 'peneira'. Nos trabalhos 'lupa', a hipótese e os resultados são mais importantes que as ferramentas bioinformáticas usadas como meio investigativo. Assim, tais estudos são freqüentemente publicados nas revistas relacionadas com o organismo em que se está estudando o fenômeno ou em revistas específicas de genética, biologia molecular ou bioquímica.

Exemplos de estudos de bioinformática 'lupa' são aqueles nos quais alguma característica biológica de um organismo é explicada a partir da observação de suas seqüências gênicas ou protéicas e da comparação com seqüências similares em organismos proximamente relacionados. Esses estudos de genômica comparativa permitem associar aspectos da biologia dos organismos comparados à presença ou à ausência de determinado gene, grupo de genes ou processos metabólicos.

Assim, a bioinformática traz uma abordagem científica aos dados gerados em projetos genoma, como já fazem outras ciências bem estabelecidas, como a biologia molecular, a genética e a bioquímica. Vale registrar, no Brasil, a iniciativa pioneira da Coordenação para o Aperfeiçoamento de Pessoal de Nível Superior (Capes) de induzir a criação de cursos de doutorado na área de bioinformática, o que já aconteceu em duas universidades (a de São Paulo e a Federal de Minas Gerais).

Os estudos de genomas, como vimos, são importantes para produzir um grande volume de informações sobre a anatomia molecular de uma espécie. Tais informações podem ser usadas como pontos de partida para a produção de novos conhecimentos científicos através de diferentes modelos experimentais, seja *in vitro*, *in vivo* ou *in silico*. Essa última abordagem é representada por metodologias baseadas na criação de algoritmos dentro dessa nova e importante ciência do século 21, a bioinformática. ■

ELSEVIER

Computers in Biology and Medicine ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

A set of amino acids found to occur more frequently in human and fly than in plant and yeast proteomes consists of non-essential amino acids

Francisco Prosdocimi[a], Maurício A. Mudado[b], J. Miguel Ortega[b,*]

[a]*Departamento de Biologia Geral, ICB-UFMG, Laboratório de Biodados, Av. Antonio Carlos, 6627 C.P. 486, 31.270-010 Belo Horizonte, MG, Brazil*

[b]*Departamento de Bioquímica e Imunologia, ICB-UFMG, Laboratório de Biodados, Av. Antonio Carlos, 6627 C.P. 486, 31.270-010 Belo Horizonte, MG, Brazil*

Abstract

We investigated the hypothesis that essential amino acids are being replaced in proteins by non-essential amino acids. We compared the amino acid composition in human, worm and fly proteomes, organisms that cannot synthesize all amino acids, with the amino acids of the proteomes of plant, bakers yeast and budding yeast, which are capable of synthesizing them. The analysis covered 460,737 proteins (212,197,907 amino acids). The data suggest a bias towards the usage of non-essential amino acids (mostly the set GAPQC) by metazoan organisms, except for the worm, a Pseudocoelomata. Our results support the hypothesis that non-essential amino acids have been substituting essential ones in the Coelomata.

Keywords: Diet; Genome; Proteome; Essential amino acids; Metazoa; Amino acid usage

1. Introduction

Whether dietary habits have been influencing the genomic and proteomic constitution of organisms during evolution is an interesting and unaddressed question. It is well known that many metazoan organisms (MO) are not able to synthesize some amino acids and that these so-called "essential amino acids" (EAA) must be included in the diet. One way to have a glimpse of the influence of dietary habits on these organisms is through the analysis of their proteomes for the use of these special amino acids.

If we think about the strong effect of diet on evolution, two different and, rather opposite, evolutionary scenarios can be conceived. In one of these, diet is not understood as a selective force for modification of the proteome of ancestral organisms. This first scenario could come about in two ways: (1) organisms capable of reproducing themselves have always been well fed and malnourished individuals are unable to reproduce; or (2) the requirement of essential amino acids is so small that even the worst-fed organisms are capable of reproducing competitively. Consequently, diet would not alter selection pressure to promote genome or proteome modification.

In an opposite scenario, diet could act as a mechanism for genomic and proteomic modification. In this situation, if ancestral metazoan organisms presented proteomes enriched by EAA, they would need to ingest a large quantity of food in order to produce a rich and functional set of proteins. But if they could not obtain enough food, they would produce few offspring and many of them probably would not reach reproductive age due to the deficiency in their proteins. In that scenario, DNA mutations in ancestral genes, substituting codons for EAA to codons for non-essential amino acids (NEAA) would be selected for, and the organisms harboring them would be capable of producing the fittest offspring. It is conceivable that these substitutions mostly occurred in the large stretches of the polypeptide chain that do not require strict conservation for their function to be exerted properly. Moreover, even similar substitution into conserved domains might exhibit such a tendency.

* Corresponding author. Tel.: +55 31 3499 2654.
E-mail address: miguel@icb.ufmg.br (J.M. Ortega).

0010-4825/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2006.02.003

One way to study the influence of diet on genome modification is through the evaluation of the EAA content in proteins from different species. Taking into account the second scenario point of view, MO changed their amino acid sequences in proteins from EAA to NEAA along the course of evolution; organisms would then become less and less dependent on the amino acid composition of ingested proteins to attain optimum metabolism and function. We examined if diet has been acting as a selective pressure in proteome modification and, if so, at what scale this has been happening. To accomplish this, we searched for evidence of substitution of EAA by NEAA in metazoan and non-metazoan organisms. Moreover, we examined if this kind of substitution has happened in all proteins or in only a particular set of proteins.

We analyzed the EAA content of all proteins from the completed genomes of six species. Among these six organisms, we choose three MO which require nine EAA, *Homo sapiens* (*hsa*), *Drosophila melanogaster* (*dme*) and *Caenorhabditis elegans* (*cel*), and three non-metazoan organisms (NMO) that have enzymes to synthesize all the amino acids, *Arabidopsis thaliana* (*ath*), *Saccharomyces cerevisiae* (*sce*) and *Schizosaccharomyces pombe* (*spo*). We used four secondary genomic databases to evaluate the differences in EAA usage for these organisms: COG, RefSeq, UniRef and KEGG. COG, the NCBI database that clusters groups of orthologous proteins [1], was used only in the eukaryotic version (KOG); it allows comparisons between evolutionary-related proteins, as do UniRef [2,3] and KEGG [4]. The UniRef100 representatives from Uniprot were used, clustered in UniRef50 entries. Ortholog clusters were also obtained from KEGG. By using those databases we were able to investigate if the amino acids have been changing in proteins with the same origin and function. RefSeq, the NCBI non-redundant Reference Sequence database [5,6], was also used for the evaluation of the amino acid content of proteins throughout the complete proteomes of the selected organisms.

2. Methodology

2.1. Essentiality index ranking

For each KOG, KEGG and UniRef50 cluster, an index called "Essentiality Index" (EI) was calculated, as previously described [7], which represents its proportion of EAA. EAA percentage for RefSeq entries or each KOG, KEGG and UniRef50 clusters were averaged and standard error calculated. The amino acid arginine was removed from this index, since it is often called a semi-essential amino acid and it can be produced in some phases of an organism's life cycle [8,9]. The EAA considered were: H, I, L, K, M, F, T, W, V, and the EI was calculated as

$$\mathrm{EI} = \frac{\mathrm{Number\_of\_EAA}}{(\mathrm{Total\_number\_of\_aa}) - (\mathrm{Number\_of\_R})}, \quad (1)$$

where in clusters presenting more than one protein per organism (putative paralogs), the number of amino acids from all proteins were summed to generate the EI for that cluster. The R, in the formula, represents the number of arginines that were removed from the total, as it is a semi-essential amino acid.

Pair-wise comparisons of the EI were made between all organisms studied for all KOG, KEGG and UniRef50 clusters shared by them. The total number of events and the number of times in which NMO clusters presented a greater index than MO were counted and listed.

2.2. Preference ratio index and clustering

The amino acid usage was calculated for each organism, based on the RefSeq, UniRef and KEGG databases. Pair-wised comparisons were made, mainly between MO and NMO, to determine which amino acids were predominant in one or the other group, through the utilization of the preference ratio (PR) index.

$$\mathrm{PR_{AA}} = \log_2 \frac{\%\mathrm{AA\_Usage_{MO}}}{\%\mathrm{AA\_Usage_{NMO}}}, \quad (2)$$

where for each amino acid, the PR was calculated by comparing its usage in MO and NMO. Considering the theory, a low PR for EAA would be expected (most used in NMO) and a high PR for the NEAA (most used in MO).

Hierarchical clustering of amino acidic PR indexes were produced based on these PR values using cluster software from Eisen [10], with ratios normalized by $\log_2$.

2.3. Selected amino acid index

One more index was created to take into account the proportion of some selected, often-different amino acids between MO and NMO. The selected amino acids (SAA) group was R, H, G, A, P, Q and C and the selected index (SI) was calculated as follows:

$$\mathrm{SI} = \frac{\mathrm{Number\_of\_SAA}}{(\mathrm{Total\_number\_of\_aa})}, \quad (3)$$

where in clusters presenting more than one protein (putative paralogs), the numbers of amino acids from all proteins were summed to generate the SI for that cluster.

SI pair-wise comparisons were made between all organisms studied for all KOG, KEGG and UniRef clusters shared by them.

3. Results

3.1. Raw data analysis

The first analysis performed was simply the calculation of EAA percentage in the proteome (Fig. 1).

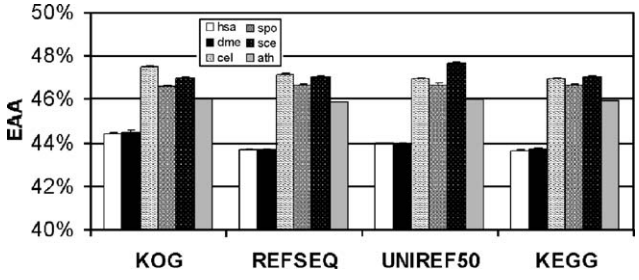


Fig. 1. Average frequency of essential amino acids by organism and database studied. The standard error is also shown for each value.

The EAA percentage was observed to be lower in MO than in NMO, with the exception of *cel*, in all the databases, but it was not statistically relevant (the standard deviation ranged from 5% to 7%). Moreover, the Coelomata organisms, human and fruit fly, showed a significantly lower frequency of EAA than did the non-metazoans, as judged by the differences in the means (Fig. 1) versus the standard deviations.

3.2. Voting of KOGs essentiality index

The EIs for each KOG, KEGG ortholog cluster and UniRef50 clusters shared by a pair of organisms were determined. The number of protein clusters shared by the MO–NMO pairs of organisms was counted and the EI of each shared cluster was compared. Table 1 shows the data for the KOG database, which presents the highest number of shared clusters (42,531 tuples); the same comparison was also made for UniRef and KEGG clusters, which yielded 4989 and 11,623 tuples, respectively (see http://biodados.icb.ufmg.br/EAA for supplementary information).

As expected, *hsa* and *dme* had fewer high-EI KOGs than did *ath*, *sce* and *spo*. Although the results did not show a strong difference between MO and NMO, they point in the direction of the existence of selection pressure. Again, *cel* was an exception when compared to *ath*/*spo*, and taking into account the missing information about amino acid biosynthesis pathways in this worm, further analysis was done separately with *hsa* and *dme*. KEGG data (Table s1, supplementary information) supports the difference between Coelomata (*hsa* and *dme*) and the yeasts (*sce* and *spo*), but produces weak evidence against *ath*, while UniRef based vote analysis (Table s2, supplementary information) did not support the EAA to NEAA switch; bad cluster structures were manually checked in this database.

An evaluation of the most dissimilar KOGs in terms of the EI between *ath* and *hsa,* the ones showing high EI in *ath* and low EI in *hsa* (Table 2), indicated three ribosomal proteins in the four most dissimilar clusters (KOG4752, KOG0002, KOG3445), which might indicate that highly expressed genes have a greater ratio of essential to non-essential substitution between MO and NMO. We found that KOG4293 showed very few expressed sequence tags (ESTs) sampled in *hsa* in comparison with *ath*, while the other proteins seemed to be from genes expressed occasionally (K-EST web-site http://biodados.icb.ufmg.br/K-EST [11]).

3.3. Preference ratio clustering

In order to determine if NEAA occur preferentially in the three MO, the amino acid percentage usage was derived based on RefSeq data, a large non-redundant dataset consisting of complete conserved domains (CDs). Additionally, a confirmation experiment was conducted with protein entries from the KEGG database. A PR was defined as the

Table 1
Percentage of KOGs with a higher essentiality index [7]

	hsa	*dme*	*cel*	*ath*	*sce*	*spo* (%)
hsa higher	–	*53.27%*	*35.31%*	**45.28%**	**34.00%**	**39.54**
dme higher	*2000* (*4280*)	–	*31.88%*	**44.98%**	**32.69%**	**38.44**
cel higher	*2689* (*4157*)	*2724* (*3999*)	–	**60.76%**	**46.18%**	**53.34**
ath higher	**1670** (**3052**)	**1568** (**2850**)	**1087** (**2770**)	–	*36.89%*	*42.83*
sce higher	**1613** (**2444**)	**1538** (**2285**)	**1213** (**2254**)	*1466* (*2323*)	–	*58.51*
spo higher	**1535** (**2539**)	**1456** (**2365**)	**1084** (**2323**)	*1372* (*2400*)	*1033* (*2490*)	–

Note: The percentages in the upper right diagonal represent the percentage of shared KOGs in which the organism in the row presents more essential amino acids (EAA) than the organism in the column. In bold we see the comparisons between MO and NMO and in italics the comparisons inside each group. The numbers in the lower left diagonal represent absolute number of KOGs, in which the organism in the row presents more EAA than the organism in the column. Again, in bold we see the comparisons between MO and NMO and in italics the comparisons inside each group. The total numbers of shared KOGs between the organisms are in parentheses. Meatozoan organisms (MO): *Homo sapiens* (*hsa*), *Drosophila melanogaster* (*dme*), and *Caenorhabditis elegans* (*cel*), and non-metazoan organisms (NMO): *Arabidopsis thaliana* (*ath*), *Saccharomyces cerevisiae* (*sce*) and *Schizosaccharomyces pombe* (*spo*).

Table 2
hsa–ath most different KOGs in the essentiality index (EI)

KOG	DIFF[a] (%)	Description
KOG4752	38	(J) Ribosomal protein L41
KOG0002	24	(J) 60s ribosomal protein L39
KOG3491	19	(S) Predicted membrane protein
KOG3445	17	(J) Mitochondrial/chloroplast ribosomal protein 36a
KOG3500	15	(C) Vacuolar H+-ATPase V0 sector, subunit M9.7 (M9.2)
KOG1793	14	(S) Uncharacterized conserved protein
KOG4293	14	(T) Predicted membrane protein, contains DoH and Cytochrome b-561/ferric reductase transmembrane domains
KOG3423	14	(K) Transcription initiation factor TFIID, subunit TAF10 (also component of histone acetyltransferase SAGA)
KOG2346	13	(S) Uncharacterized conserved protein

[a]Difference of EI values between *hsa* and *ath*.

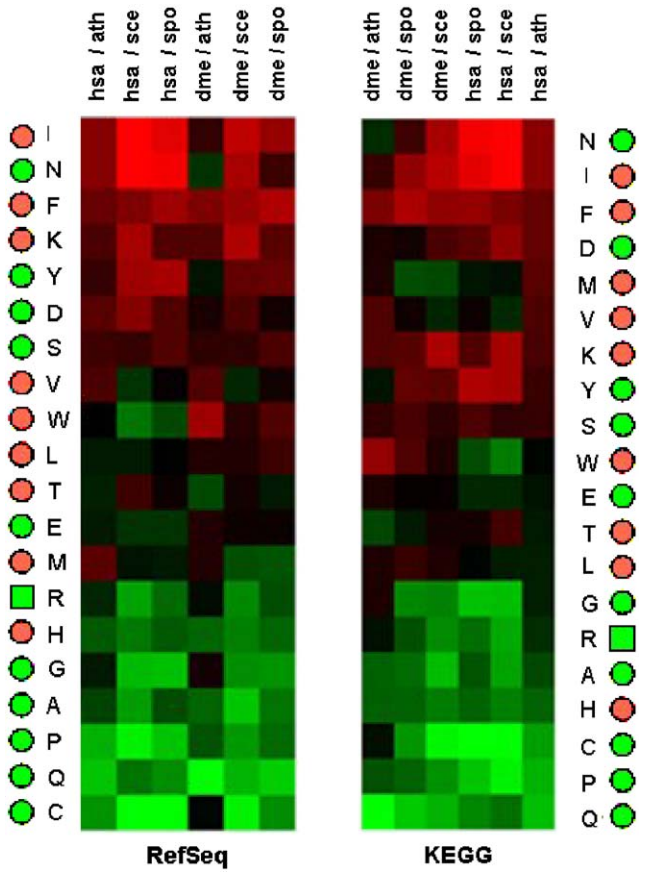


Fig. 2. Clustering analysis of PR indexes between the indicated MO and NMO for RefSeq and KEGG. Green and red filled symbols indicate, respectively, NEAA/semi-essential or EAA. The colors in the plot represent a tendency of the amino acids to be present preferably in MO (green) or NMO (red).

percentage of each amino acid in MO divided by its percentage in NMO, normalized by $\log_2$.

There was a tendency of NEAA (labelled with green symbols close to the letters in Fig. 2) to be more frequent in Coelomata MO than in NMO (green cluster, lower portion of Fig. 2). Remarkably, both Coelomata clades, Protostomia (*dme*) and Deuterostomia (*hsa*) showed a similar tendency of AA usage. The worm, a Pseudocoelomata, was not included for hierarchical clustering (Fig. 2), since its PR indices seemed to follow a random pattern on comparison with the NMO (Figs. 3e and 3f). Results for the KEGG database were similar to those shown in Fig. 3 (data not shown). As a set of amino acids (RHGAPQC, Figs. 2 and 3) tends to be more frequent in the human and the fruit fly than in the plant and yeast proteomes, we evaluated the percentage of the proteomes that consisted of these selected amino acids (SAA). Aside from R, which is semi-essential and is only required under certain conditions, and H, which is essential but showed ratios that were not amongst the highest ones (Fig. 2), they are all NEAA. Thus, we included R and H in SAA.

The SAA percentage was observed to be higher in Coelomata MO than in NMO (once again with exception of *cel*), but the difference was not significant, as judged by the high standard deviation values (around 6%, data not shown), although the standard error was acceptable (Fig. 4). Comparing *hsa* against *ath*, using all four databases, with both KOG and RefSeq data, the difference reached over 5% (Fig. 4), while it was around 2% used all the NEAA in the previous analysis (Fig. 1).

Remarkably, when comparing *hsa* and *sce* with respect to EAA content, about 40% of the entries presented a higher amount in humans (Table 1), while the analysis of the SAA showed that in 91% of the cases, *hsa* is richer in this set than *sce* (Table 3). Determination of this set might be of relevance for further exploration of proteomes in the Coelomata clade, as well as for other Pseudocoelomata data that might become available, although the bias seems to occur similarly in *dme* and *hsa* (Fig. 5).

4. Discussion

As far as we know, this is the first attempt to investigate a hypothesis about the influence of diet on proteome modification along evolution. Considering that many amino acids substitutions can be conservative in terms of protein function, the genome of complex organisms could be modified as a response to selection pressure favoring the preferential usage of NEAA. We found some indications of the occurrence of this kind of modification in the comparison between MO and NMO, although it seems to be happening mainly in some proteins that may be constitutively and highly expressed (Table 2), instead of along the entire proteome; while

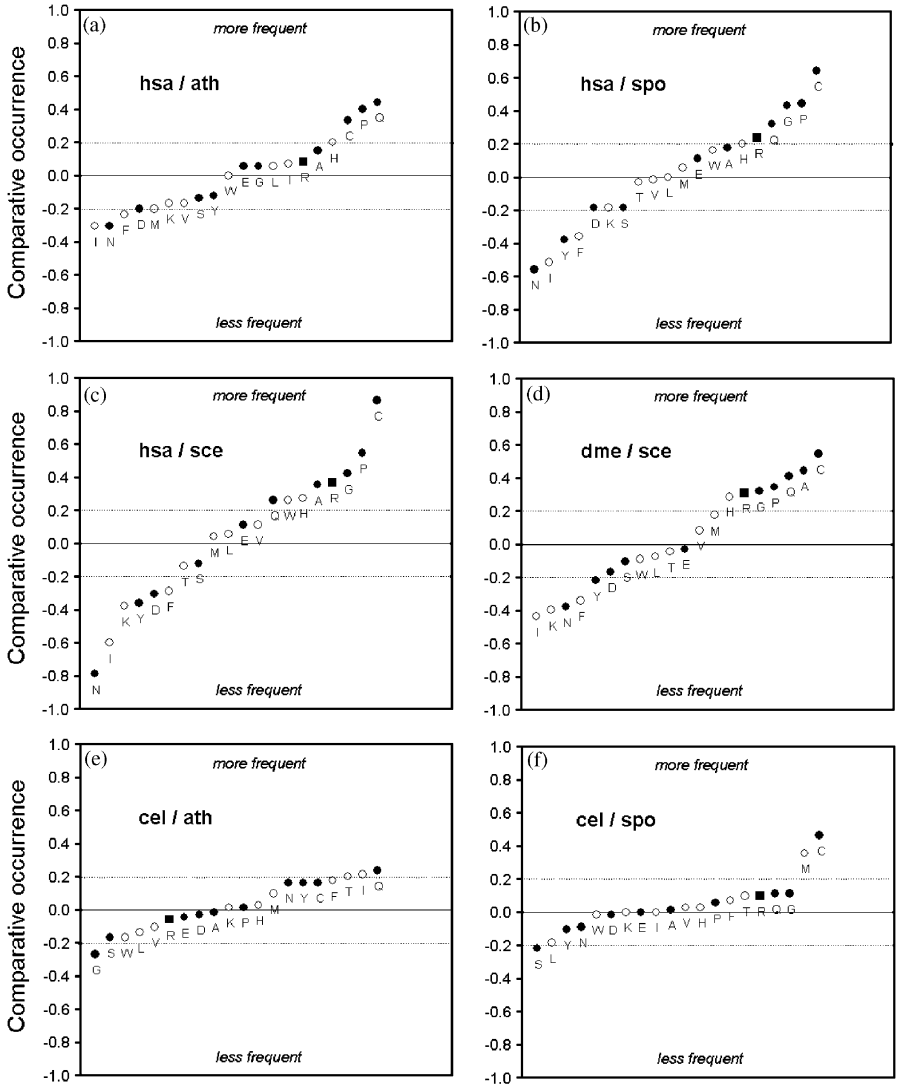


Fig. 3. Preference ratio (PR) plot of RefSeq amino acids for the selected organisms. EAA are shown as empty circles and NEAA are shown as filled circles. Arginine (R), as a semi-essential amino acid, is shown as a filled square. Comparisons involving *Caenorhabditis elegans* (e, f) are shown as controls.

it may be undetected in proteins originated in MO. We are currently extending our investigation to a large number of organisms that might have information about the requirement for EAAs. Our data indicate that both Coelomata clades, *dme* and *hsa* present a bias (Fig. 5), with very similar patterns (Fig. 2), while this bias is absent in the clade Pseudocoelomata (*cel*), which directly diverges from the Coelomata clades (*dme, hsa*). These observations support the evidence for a coelomata clade [12], a complex issue with controversial positions [13].

It would be highly desirable to study position-specific substitutions in orthologous proteins, since we could determine which substitutions are most frequent, and if these happen with minimum nucleotide substitutions, based on the genetic code. However, it is very probable that the substitutions are happening freely in the non-conservative regions of the proteins (the ones not matched by local sequence alignment software). Moreover, the investigation of EAA proportion in specific processes

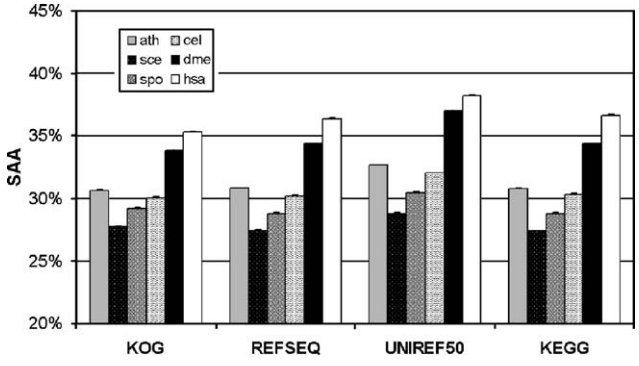


Fig. 4. Average number of selected amino acids by organism and database studied. The standard error is also shown for each value.

and pathways should give us a glimpse into where these substitutions are most relevant. Our analysis pointed out a set of amino acids, RHGAPQC, which are more common in man and

Table 3
Percentage of KOGs with higher selected index (SI)

	hsa	*dme*	*cel*	*ath*	*sce*	*spo* (%)
hsa higher	–	*56.10%*	*75.90%*	**74.30%**	**91.00%**	**84.52**
dme higher	*1879*	–	*72.12%*	**70.78%**	**89.37%**	**80.90**
cel higher	*1001*	*1116*	–	**49.26%**	**81.75%**	**68.55**
ath higher	**785**	**834**	**1406**	–	*82.31%*	*67.97*
sce higher	**220**	**243**	**412**	*411*	–	*26.67*
spo higher	**393**	**452**	**731**	*768*	*1823*	–

Note: The percentages in the upper right hand diagonal represent the percentage of shared KOGs in which the organism in the row presents more essential amino acids (EAA) than the organism in the column. In bold we see the comparisons between metazoans (MO) and non-metazoans (NMO) and in *italics* the comparisons inside each group. The numbers in the lower left hand diagonal represent absolute number of KOGs, in which the organism in the row presents more EAA than the organism in the column. Again, in bold, we see the comparisons between MO and NMO and in *italics* the comparisons inside each group. The total number of shared KOGs between the pairs of organisms can be seen in Table 1 (in parentheses).

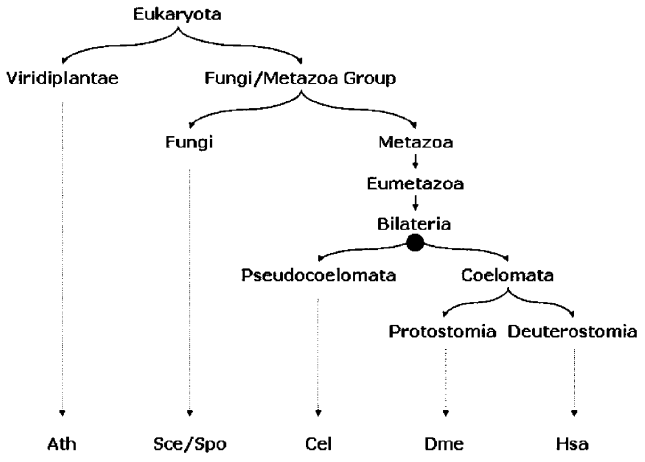


Fig. 5. Clade separation for the organisms investigated. The filled circle marks the putative position of divergence.

fruit fly when their usage in the proteome is compared against plant and yeast. Although chemical properties did not seem to provide any clues on the origin of this set, one of them is synthesized using a key metabolic intermediate, A from pyruvate; R, P and Q are the three amino acids derived from E, which is not included in the set, but is directly derived from α-ketoglutarate. Finally, G and C are both derived from S, which again is not in the set. Moreover, neither E nor S was under represented (Fig. 1). H clustered together in the comparison shown in Fig. 1, and it is an EAA, but the ratios never attained the highest levels (Fig. 2). On the other hand, the often less-frequently-occurring amino acid N (Figs. 1 and 2), besides being a NEAA, is derived from E, as is the EAA cluster M, T, K and I. It is possible that metabolic pathways may be influencing the composition of the set of NEAA that would be most available to substitute for EAA requirements in the proteome composition of MO. We observed that the *cel* proteome did not show the same bias as *dme* and *hsa*, with respect to the usage of NEAA, and it is tempting to suggest that this might be due to the fact that its diet consists of bacterial cells, which provide the worm with all of the necessary amino acids. If this is true, the influence of diet in the genome might be more complex than initially thought, since it would also depend on specific nourishment patterns of organisms and of their ancestors. The only highly sequenced Pseudocoelomata is another *Caenorhabditis*; thus a more detailed inspection of the origin of the bias is not possible. These preliminary results support the hypothesis that EAA were replaced by NEAA, in both clades—(*dme* and *dsa*)—of Coelomata. For reasons that need to be clarified, the set RHGAPQC constitutes a group of NEAA (aside from H and the semi-essential R) found to occur more frequently in humans and fruit flies than in plant and yeast proteomes.

5. Summary

It is still unknown whether diet has been influencing the genomic and proteomic constitution of organisms along evolution. One way to study the influence of diet on the genome modification of organisms is through the evaluation of the essential amino acids (EAA) content in proteins from different species. We investigated the hypothesis that EAA have been replaced by non-essential amino acids (NEAA) in the proteins of some organisms. We compared the amino acid composition of the proteome of humans, a worm and a fly, which cannot synthesize all amino acids, with the proteomes of a plant, bakers yeast and budding yeast, which are capable of synthesizing them. The analysis covered 460,737 proteins (212,197,907) from four well-known molecular databases. Our data suggest a bias in the inclusion of NEAA (mostly the set RHGAPQC) by metazoan organisms, except for in the worm. These preliminary results support the hypothesis that EAA have been replaced by NEAA in both clades, Protostomia and Deuterostomia, of Coelomata metazoans. For reasons yet to be clarified, the set RHGAPQC constitutes a group of NEAA (aside from H and the semi-essential R) found to occur more frequently in human and fly, than in plant and yeast proteomes.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi: 10.1016/j.combiomed.2006.02.003.

References

[1] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, D.A. Natale, The COG database: an updated version includes eukaryotes, BMC Bioinform. 4 (1) (2003) 41.

[2] R. Leinonen, F.G. Diez, D. Binns, W. Fleischmann, R. Lopez, R. Apweiler, UniProt archive, Bioinformatics 20 (2004) 3236–3237.

[3] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S. Yeh, The Univ. Protein Resour. (UniProt) Nucleic Acids Res. 33 (2005) D154–D159.

[4] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 27 (1999) 29–34.

[5] K.D. Pruitt, K.S. Katz, H. Sicotte, D.R. Maglott, Introducing RefSeq and LocusLink: curated human genome resources at the NCBI, Trends Genet. 16 (1) (2000) 44–47.

[6] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res. 33 (1) (2005) D501–D504.

[7] F. Prosdocimi, J.M. Ortega, Diet as a pressure on the amino acid content of proteomes, Lect. Notes Comput. Sci. 3594 (2005) 153–159.

[8] W.C. Rose, M.J. Oesterling, M.J. Womack, Comparative growth on diets containing ten and nineteen amino acids, with further observations upon the role of glutamic and aspartic acid, J. Biol. Chem. 176 (1948) 753–762.

[9] I. Nakagawa, T. Takahashi, T. Suzuki, K. Kobayashi, Amino acid requirements of children: minimal needs of tryptophan, arginine and histidine based on nitrogen balance method, J. Nutr. 80 (1963) 305–310.

[10] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (25) (1998) 14 863–14 868.

[11] M.A. Mudado, A. Barbosa-Silva, J.A. Torres, S. Paula-Pinto, J.M. Ortega, K-EST: KOG expression/sampling tool, Bioinformatics, submitted for publication.

[12] Y.I. Wolf, I.B. Rogozin, E.V. Koonin, Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis, Genome Res. 14 (2004) 29–36.

[13] A.M. Aguinaldo, J.M. Turbeville, L.S. Linford, M.C. Rivera, J.R. Garey, R.A. Raff, J.A. Lake, Evidence for a clade of nematodes, arthropods and other moulting animals, Nature 387 (1997) 489–493.