Maurício de Alvarenga Mudado

Uso da Base de Dados Secundária KOG como Ferramenta para Caracterização de Expressão Gênica e Mineração de Dados em Projetos Transcriptoma

Belo Horizonte – MG
Julho / 2007

Maurício de Alvarenga Mudado

Uso da Base de Dados Secundária KOG como Ferramenta para Caracterização de Expressão Gênica e Mineração de Dados em Projetos Transcriptoma

Tese apresentada ao Programa de Pós-graduação em Bionformática da Universidade Federal de Minas Gerais como pré-requisito para a obtenção do título de Doutor em Bionformática.

Área de Concentração: Bioinformática Genômica.

Orientador:

Dr. J. Miguel Ortega

Co-Orientador:

Dr. Sérgio Vale Aguiar Campos

CURSO DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA
DEPARTAMENTO DE BIOQUÍMICA E IMUNOLOGIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Belo Horizonte – MG
Julho / 2007

“born into this
walking and living through this”
Charles Bukowski

Agradecimentos

- Ao Miguel
- Laboratório de Biodados: Adriano, Alessandra, Rosana, Gabriel, Dudu, Chico, Saulo, Cris Neri, Bellinha, Dani, Elisa
- Outros labs: Céc, Cris Ribeiro, Chico Lobo, Michael, Gustavo Cerca
- CENAPAD
- FUNED
- Secretaria: Kátia, Alberto
- Meus pais, Thais

Resumo

São apresentados neste trabalho um conjunto de novos resultados, técnicas e ferramentas destinadas à mineração de dados e ao auxílio na análise de Etiquetas de Seqüências Expressas (EST) geradas em projetos transcriptoma. A base de dados de proteínas secundárias KOG foi utilizada como ferramenta no alinhamento e anotação automática de seqüências EST de quatro organismos, *A. thaliana, C. elegans, D. melanogaster e H. sapiens*. Estes alinhamentos foram utilizados para diversos fins e em diversos experimentos, entre eles: inferência de valores de corte para similaridade utilizando tBLASTn entre seqüências de EST e proteínas do mesmo organismo; desenvolvimento de um teste de anotação entre EST e proteínas KOG; avaliação da qualidade da anotação usando os valores de corte inferidos; avaliação da qualidade da anotação utilizando *uniques* gerados pelo programa TGICL; caracterização funcional das EST com KOG; caracterização da amostragem de EST ou expressão gênica com KOG; avaliação da cobertura da base KOG por quantidades incrementais de EST e inferência de um número mínimo para cobri-la; criação de uma ferramenta *web* denominada K-EST, que disponibiliza dados de amostragem de EST por KOG e também de conservação entre agrupamentos KOG; inferência de perda de genes ou pelo menos de expressão gênica em organismos pertencentes ou não à base KOG, utilizando dados de amostragem de EST e conservação.

Abstract

A set of new results, techniques and tools are presented in this work for data mining and to help in the analysis of Expressed Sequence Tags (EST) generated by transcriptome projects. The secondary database KOG was utilized as a tool in the alignment and automatic annotation of ESTs from four organisms, *A. thaliana, C. elegans, D. melanogaster e H. sapiens*. These alignments were utilized to many purposes and in many experiments, like: inference of similarity cutoffs utilizing tBLASTn with ESTs and proteins from the same organism; development of an annotation test with EST and KOG proteins; evaluation of the quality of annotation by using the cutoff values discovered; evaluation of the quality of annotation by using uniques generated by the TGICL software; functional characterization of ESTs with KOG; evaluation of KOG coverage with incremental EST number and inference of a minimal number of EST to cover it; creation of a web tool named K-EST that makes available the EST sampling data with KOG and also the conservation data among KOG clusters; inference of gene loss, or at least loss of gene expression in organisms belonging or not to the KOG database, by using EST sampling data and conservation.

Sumário

Lista de Artigos

Nº	Título	Autores	Status	Pgs.
1	Ferramentas Bioinformáticas Aplicadas à Caracterização da Expressão Gênica	Faria-Campos, A. C. Mudado M, A. Peixoto, F. C. Bravo-Neto, E. Prosdocimi, F. Ortega J, M.	Publicado - Bioscience Journal EDUFU - Editora e Livraria da Universidade Federal de Uberlândia Uberlândia, MG Vol 20, Supl. 1, Pgs. 109-117, 2004	**38-46**
2	Tests of automatic annotation using KOG proteins and ESTs from 4 eukaryotic organisms	Mudado M, A. Bravo-Neto, E. Ortega J, M.	Publicado – Lecture Notes Computer Sci. Vol. 3594, Pgs. 141-152 2005	**51-62**
3	Acessing the Quality of Automatic Annotation of ESTs from Model Organisms with the KOG database	Mudado M, A. Fernandes G, R. Ortega J, M.	Manuscrito em preparação	**64-80**
4	On The Improvement of Transcriptome Annotation After Clustering and Assemblage of Incremental Number of ESTs	Mudado M, A. Ortega J, M.	Aceito para publicação - Trabalho completo nos anais do congresso BSB2007 – Angra dos Reis – RJ – Brasil - Agosto – 2007	**81-87**
5	A picture of gene sampling/expression in model organisms using ESTs and KOG proteins	Mudado Mde, A. Ortega J, M.	Publicado - Genet Mol Res Vol. 5, Issue 1 Pgs. 242-253 2006	**89-100**
6	K-EST: KOG Expression / Sampling Tool	Mudado M, A. Barbosa Silva, A. Fernandes G, R. Paula-Pinto S, A. Torres, J. Ortega J, M.	Manuscrito em preparação	**103-111**
7	Data mining and annotation of novel *Schistosoma mansoni* ESTs with the KOG database	Mudado M, A. Oliveira, G. Rede Genoma de Minas Gerais Ortega J, M.	Aceito para publicação - Trabalho completo nos anais do congresso BSB2007 – Angra dos Reis – RJ – Brasil - Agosto – 2007	**114-122**

Lista de Tabelas

Lista de Figuras

Siglas e Abreviaturas

Ath	*Arabidopsis thaliana*
BLAST	*Basic Local Alignment Search Tool*
cDNA	*Complementar DNA*
CDS	*Coding Sequence*
Cel	*Caenorhabditis elegans*
COG	*Clusters of Orthologous Groups*
Dme	*Drosophila melanogaster*
EST	*Expressed Sequence Tag*
Hsa	*Homo sapiens*
KOG	*Eukaryotic Orthologous Groups*
K-EST	*KOG Expression / Sampling Tool*
NCBI	*National Center for Biotechnology Information*
PAS	*Position of Alignment Start*
PCR	*Polymerase chain reaction*
PERL	*Practical Extraction and Report Language*
PHRED	*Phil's Read Editor*
PHRAP	*PHRagment Assembly Program*
SAGE	*Serial Analysis of Gene Expression*
Sma	*Schistosoma mansoni*
TIGR	*The Institute for Genomic Research*
TGICL	*TIGR Gene Indices Clustering Tool*
UFMG	*Universidade Federal de Minas Gerais*
UTR	*Untranslated Region*

1 Introdução

1.1 Bioinformática genômica e seqüenciamento de DNA

Apesar de ser historicamente ligada à análise de genômica e genética , um conceito mais geral de bioinformática pode ser entendido como a "aplicação da informática na área de biologia", cobrindo áreas como *neuroinformática, imunoinformática, filoinformática e informática na genética* (Costa Lda, 2004). Segundo um conceito intermediário de bionformática de João Carlos Setúbal, a bionformática pode ser entendida como um "conjunto de técnicas advindas da matemática, estatística e computação aplicadas a problemas de biologia molecular, em particular aos problemas da genômica" (Moreira-Filho *et al.*, 2004). A área de bionformática genômica, ou *informática na genética* surgiu principalmente para geração de conhecimento e para a análise dos dados provindos do seqüenciamento de DNA em larga escala.

A criação do seqüenciamento de DNA pelo método de terminação didesoxi (Sanger *et al.*, 1977) encorajou novas iniciativas no meio científico na tentativa de se acelerar esta área (Mcbride *et al.*, 1989). Na década de 80 e início da década de 90 foram publicados os trabalhos que trariam as idéias primordiais para o surgimento de um processo de seqüenciamento do DNA rápido e automatizado. Entre elas estão métodos de seqüenciamento rápidos usando iniciadores fluorescentes (Wilson *et al.*, 1990a; Wilson *et al.*, 1990b) uso de capilares ultrafinos para eletroforese (Drossman *et al.*, 1990) e métodos que usavam laser e fluorescência para a detecção de bases marcadas (Chen *et al.*, 1991; Zhang *et al.*, 1991). Máquinas que agrupavam esses processos em sistemas quase totalmente automatizados começaram a surgir (D'cunha *et al.*, 1990) e a serem comercializadas. A figura 1 exemplifica o processo de seqüenciamento automatizado.

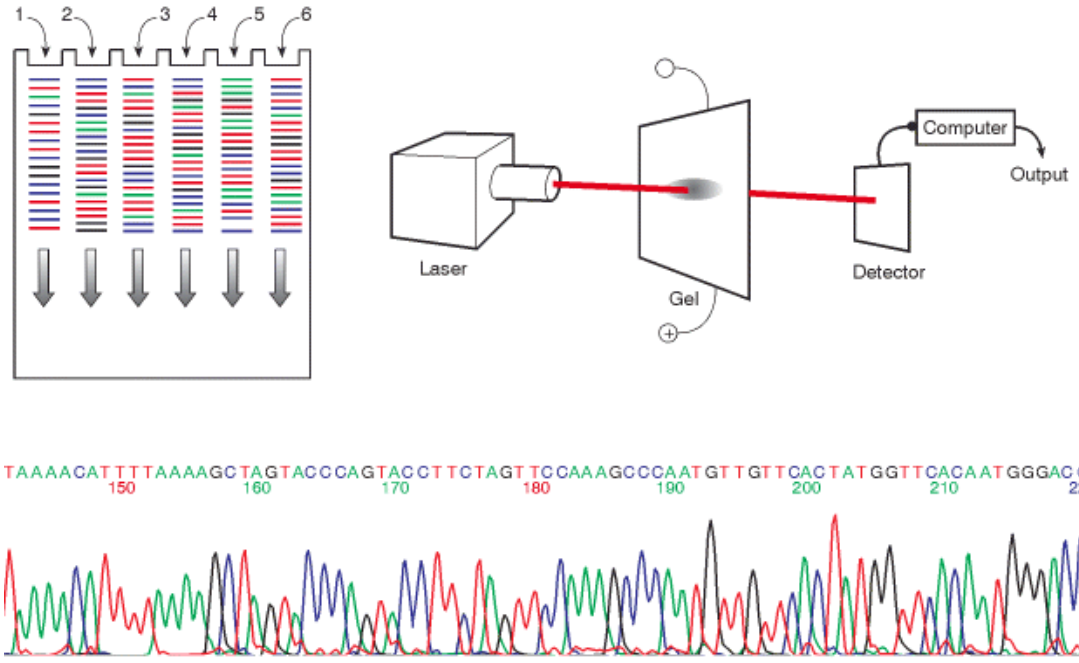


Figura 1. Exemplo do processo de seqüenciamento automatizado com marcadores fluorescentes e detecção a laser. Durante a corrida de um gel de eletroforese com as amostras marcadas, um mecanismo a laser detecta a fluorescência em cada canaleta do gel e gera um fluorograma ou cromatograma. A parte inferior mostra um cromatograma típico, resultado de um seqüenciador ABI PRISM 377. O gel mostrado na figura foi substituído por géis capilares ultrafinos em versões mais modernas de seqüenciadores comerciais. Retirado de (Strachan e Read, 1999).

De fato uma iniciativa de seqüenciar o genoma humano já estava em processo, patrocinada em grande parte pelo Departamento de Energia Americana (DOE), idéia surgida em meados de 1984. Segundo Emmanuel Dias Neto, o projeto genoma humano (PGH) foi uma iniciativa para resolver questões relativas aos ataques nucleares ocorridos no Japão, durante a Segunda Guerra, principalmente sobre os efeitos da radioatividade no DNA humano (Moreira-Filho *et al.*, 2004). O projeto teve forte incentivo em meados de 1990 com o apoio do Instituto Nacional de Saúde americano (NIH).

1.2 Projetos Genoma

Em 1995 foi publicado o primeiro genoma de um organismo celular, a bactéria parasita *Haemophilus influenzae* (Fleischmann *et al.*, 1995) e em seguida outra bactéria parasita o *Mycoplasma genitalium* (Fraser *et al.*, 1995). No ano seguinte, foi seqüenciado o primeiro

eucarioto, a levedura *Saccharomyces cerevisiae* (revisto por Dujon, 1996). Nos anos seguintes uma série de outros organismos tiveram seus genomas seqüenciados, entre eles o humano. Segundo Koonin, a disponibilidade de todas essas seqüências é uma oportunidade para o estudo de relações evolutivas e para o estudo do conteúdo funcional dos genomas (Koonin e Mushegian, 1996).

Sem dúvida um grande marco na ciência foi a publicação do rascunho do genoma humano em 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001). Vários artigos e comentários foram publicados nas revistas científicas *Science* e *Nature*. Uma das grandes dificuldades do projeto genoma humano foi a da montagem do rascunho do genoma. Esse é um processo tipicamente bioinformático, resolvido pelo consórcio publico por Jim Kent, com o uso do programa GigAssembler (Kent e Haussler, 2001), criado por ele. Nesse processo estão passos como descontaminação e mascaramento de seqüências repetitivas, alinhamento de EST, mRNA, BAC, alinhamento e sobreposição de *contigs* e montagem de cromossomos inteiros.

Após a montagem de genomas é necessário tentar decifrar esse conjunto enorme de seqüências, num processo chamado anotação. A anotação pode ser definida como a processo de descoberta de componentes importantes do genoma, principalmente genes e seus produtos (Birney *et al.*, 2001; Stein, 2001). A predição gênica em genomas eucarióticos é até hoje um processo complicado, assim como a predição de fase de leitura aberta (ORF – *Open Reading Frame*) e de estruturas exon-intron. Uma das estratégias mais valiosas até hoje na anotação de genomas é o uso de seqüências de DNA complementar geradas a partir de mRNA, como as EST, que são provindas de regiões transcritas do genoma, para o alinhamento e busca por regiões de estrutura gênica (Brent, 2005). Projetos que produzem esse tipo de seqüência estão comentados abaixo.

1.3 *Transcriptômica*

O transcriptoma pode ser entendido como produto da parte expressa do genoma, ou uma coleção de moléculas de RNA que se dividem em codificadores de proteínas e não codificadores (ncRNA). Apenas uma pequena parte de toda a coleção corresponde a moléculas codificantes, ou mRNA (RNA mensageiros) que são traduzidas em proteínas. A

outra parte é constituída de moléculas de RNA transferência, RNA ribossomal e afins (Brown, 2002).

Porém a figura do transcriptoma está mudando. Recentemente o papel dos ncRNA (*non-coding RNA*) estão tomando uma maior importância na complexidade do transcriptoma, principalmente de mamíferos. Historicamente a presença de ncRNA em bibliotecas de cDNA eram interpretados como artefatos de clonagem ou seqüências truncadas. Porém várias evidências vêm mostrando que os ncRNA são mais comuns em organismos mais complexos, principalmente mamíferos, e especula-se sobre seu papel na complexidade desses organismos (Gustincich *et al.*, 2006). Sugere-se que apesar de que uma pequena parte do genoma (2%) ser composta por regiões codificadoras de proteínas, possivelmente o resto seja também transcrito na forma de moléculas de RNA não codificadores regulatórios (revisto por Mendes Soares e Valcarcel, 2006).

A transcrição do DNA em RNA é um ponto chave na regulação da expressão gênica. A regulação gênica pode atuar via processos pós-transcricionais como edição alternativa do RNA e após a tradução do RNA em proteínas, por meio da metilação, fosforilação, proteólise limitada, glicosilação, etc. Porém o passo mais importante na regulação da expressão gênica em organismos superiores é a transcrição e a presença de um transcrito é um forte indício de que aquele gene está sendo expresso. O transcriptoma fornece características sobre o padrão de expressão daquele organismo, tecido ou célula em questão (Moreira-Filho *et al.*, 2004).

1.3.1 SAGE, *Microarray* e EST

Dentre as metodologias atuais de estudo de transcriptomas estão o SAGE (*Serial Analysis of Gene Expression)*, os microarranjos (*microarrays*) e as EST ou (*Expressed Sequence Tags*).

1.3.1.1 SAGE

O método SAGE (Figura 2) provê uma análise quantitativa da expressão gênica, com a vantagem de conseguir detectar transcritos desconhecidos. Por outro lado, um intenso trabalho de atribuição de cada *tag* (explicado abaixo) a um transcrito é necessário. Em vias gerais, o método é iniciado pelo uso de uma enzima de restrição denominada enzima de ancoragem (*Nla*III por exemplo) e os fragmentos de dez bases subseqüentes a esse sítio de

restrição são colecionados em uma biblioteca usando uma enzima de *tagging* (*Bsm*FI). Esses fragmentos chamados *tags* são seqüenciados na forma de um concatâmero longo, o que aumenta em muito o rendimento de aquisição de informação. A freqüência de cada *tag* de 14 bases (pois o padrão inclui as quatro bases do sítio de restrição e as dez subseqüentes) representa a quantidade de cada transcrito detectado (revisto por Wang, 2007). Outras variações como *longSAGE*, *superSAGE* e MPSS (método comercial protegido por patente) foram desenvolvidas mais recentemente (Brenner *et al.*, 2000; Wei *et al.*, 2004; Matsumura *et al.*, 2005). Em 2003 foi criado um método para a análise de regiões de início de transcrição e promotores, chamado CAGE (*Cap Analysis Gene Expression*) (Shiraki *et al.*, 2003). CAGE é realizado com uma metodologia semelhante à do método SAGE, porém são coletados *tags* de 21 pares de bases a partir das pontas 5' encapadas de moléculas cDNA purificadas. Esses *tags* são concatenados, clonados, seqüenciados e então mapeados no genoma na busca de promotores e regiões de início de transcrição.

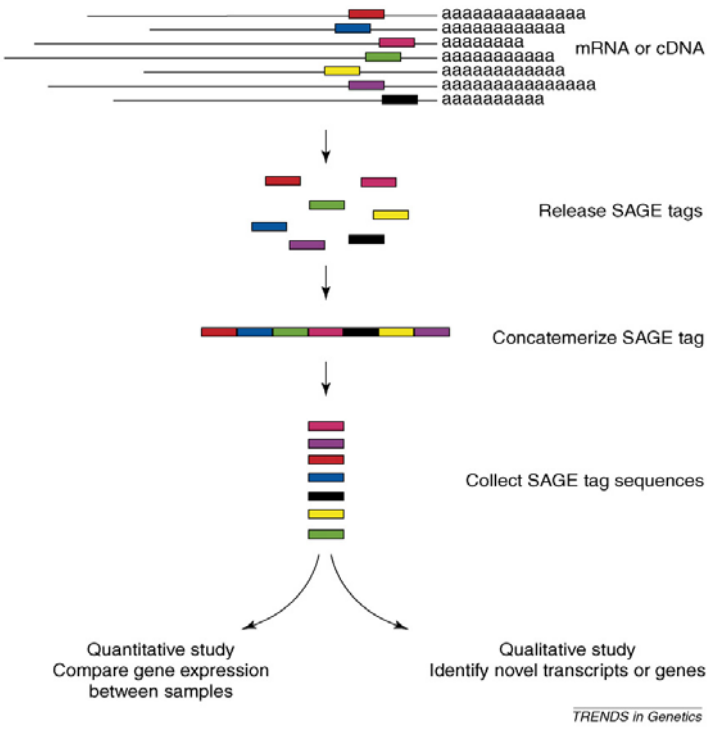


Figura 2. O método SAGE. Ver detalhes no texto. Retirado de (Wang, 2007).

1.3.1.2 Microarray

O método de microarranjos (Figura 3) é um dos melhores métodos de se analisar a expressão gênica atualmente, principalmente por permitir avaliar simultaneamente a expressão de milhares de genes. Esse método se baseia no princípio da hibridação de fragmentos de ácidos nucléicos marcados com fluoróforos (geralmente cDNA provindos da transcrição reversa de transcritos isolados de células ou tecidos), em coleções de genes já conhecidas (fixadas em anteparos de vidro) ou então representadas por conjuntos únicos de sondas que são sintetizadas em *chips* e vendidas comercialmente. Após a hibridação, a fluorescência de cada gene (ou *spot,* na lâmina ou *chip*) é medida por um *scanner* e a intensidade do sinal caracterizará a expressão daquele gene. O uso de fluoróforos que emitem luz em comprimentos de onda diferentes (representados por verde a vermelho) é feito para diferenciar o tecido do qual o transcrito se originou (Moreira-Filho *et al.*, 2004). Diferentemente do método SAGE e EST, o método de microarranjos necessita do conhecimento prévio de coleções de genes, às quais os cDNA referentes aos transcritos em estudo, se hibridarão. Uma segunda diferença é que não é determinada com precisão a expressão relativa à transcrição global. Geralmente duas condições experimentais são comparadas e a expressão diferencial de cada gene é avaliada. Trata-se de um método de triagem, sendo desejável a aplicação de uma análise específica (seja por *Northen blot* ou *Real-Time* PCR) para validação da suposta expressão diferencial.

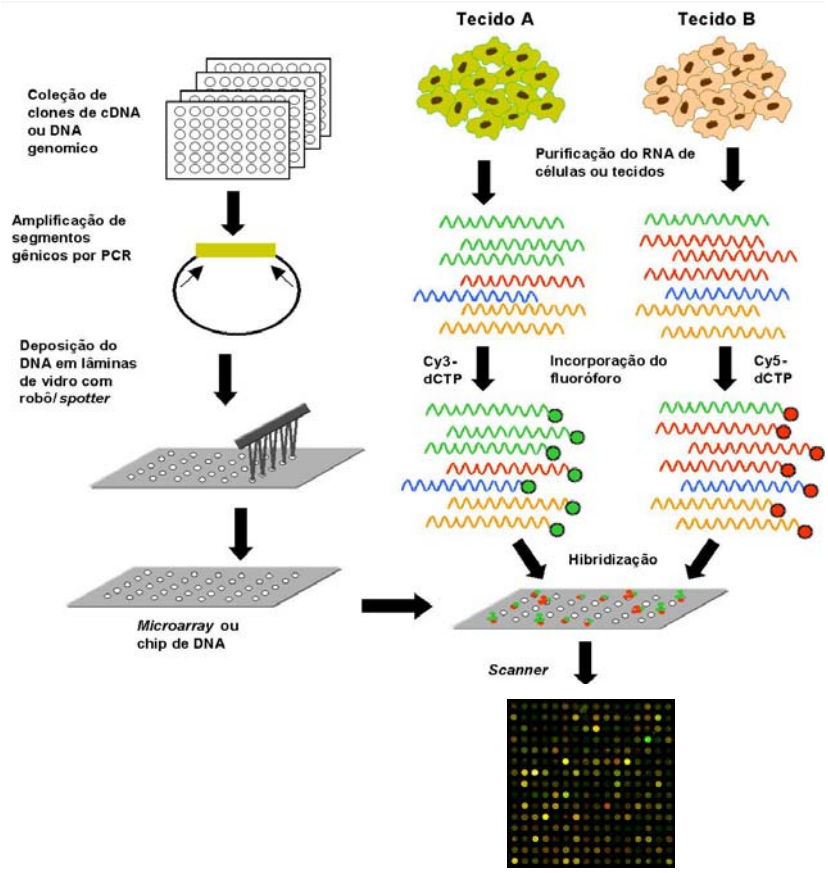


Figura 3 – Esquema de um experimento de microarranjo. Adaptado de (Moreira-Filho *et al.*, 2004).

1.3.1.3 EST

O termo EST surgiu em 1991 em uma publicação de Adams e colaboradores, numa iniciativa de descoberta gênica no genoma humano usando-se amostras do transcriptoma (Adams *et al.*, 1991). Uma seqüência EST (tecnicamente entendida como o seqüenciamento em único passo da extremidade do cDNA) é em teoria um representante de um transcrito expresso, daí o nome *Expressed Sequence Tag*. Essas seqüências são geradas em grande escala, em um processo relativamente barato. Em linhas gerais, seqüências EST são curtas (entre 200 e 800 pares de bases), não editadas, seqüenciadas em passo único e selecionadas randomicamente de bibliotecas de cDNA (revisto em Nagaraj *et al.*, 2007). Um dos processos convencionais de produção e análise de EST está descrito na figura 4. Um processo alternativo privilegia a produção de seqüências com viés para a região central do transcrito. Essas seqüências são chamadas de ORESTES (ORF *expressed sequence tags*) (de Souza *et al*. 2000).

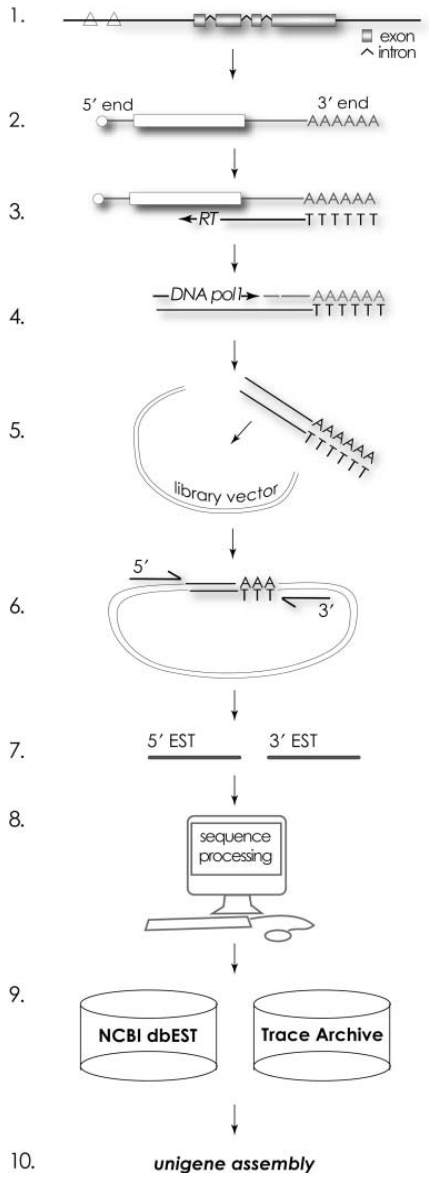


Figura 4. Processo de produção e análise de seqüências EST. ***1.Região de DNA genômico contendo introns e éxons e motivos reguladores (triângulos). 2. Introns são removidos do mRNA maduro no processo denominado edição; os mRNA são "encapados" na região 5' e caudas poli A são adicionadas na região 3'. 3. Transcriptase reversa é usada para a produção de DNA complementar (cDNA) pela molécula de mRNA. 4. Fita dupla de cDNA é produzida utilizando-se RNAse H e DNA polimerase. 5. Os cDNA são inseridos em vetores de clonagem para produzir uma biblioteca de cDNA. 6. Os insertos são seqüenciados por um ou ambos os lados (5' e/ou 3'). 7. As seqüências 5' e 3' resultantes são chamadas EST. 8. As EST são editadas para a remoção de seqüências de vetor, contaminantes e bases de baixa qualidade. 9. Depósito em bancos de dados públicos de EST (dbEST) e de cromatogramas (Trace Archive). 10. Passo alternativo de geração de unigenes.*** **Retirado de (Bouck e Vision, 2007).**

O seqüenciamento de EST provê um potencial enorme para a geração de informação, seja do genoma ou do transcriptoma. Dentre as diversas aplicabilidades dessa técnica estão a descoberta gênica, complemento para a anotação de genomas, ajuda na identificação da estrutura gênica, estabelecimento da viabilidade de transcritos alternativos, guiar a caracterização de polimorfismos de base única ou SNP (*Single Nucleotide Polimorfism*) e facilitar a análise de proteomas (revisto por Nagaraj *et al.*, 2007). Os mRNA de eucariotos apresentam regiões não traduzíveis, chamadas UTR (*Untranslated Region*), em ambas as extremidades 5' e 3', que atuam no controle da expressão, tradução e regulação gênica. Regiões 5' UTR em seqüências EST podem estar ausentes, todavia, caso a transcrição reversa não seja eficiente, o que será mais freqüente quanto maior for o transcrito. As caudas poli A da região 3' UTR são implicadas no metabolismo de mRNA assim como na predição de fronteiras gênicas e diferentes mecanismos pós-transcricionais (revisto por Nagaraj *et al.*, 2007) e em uma EST adicionam uma prova de que trata-se do seqüenciamento de um transcrito. Para se tentar predizer a estrutura ou a função de uma proteína assim como tentar isolá-la é ideal saber sua seqüência completa, ou seja, determinar a região codificante completa do clone de cDNA. Apesar de serem seqüências parciais, as EST podem conter informações que levam à descoberta de clones que contêm seqüências completas de produtos gênicos. Sabendo disso, foi criado em 1999 o MGC (*Mammalian Gene Collection*), um projeto que tem o intuito de produzir bibliotecas de cDNA de mamíferos enriquecidas com clones com seqüências completas. A busca por clones com seqüências completas é feita produzindo-se EST, avaliando-as na busca de informações que possam distinguir clones completos (Strausberg *et al.*, 1999).

Contudo as EST apresentam diversas limitações. Dentre elas as principais são: i) a representação global dos genes do tecido ou organismo em questão em uma dada biblioteca e ii) a qualidade das seqüências. Bibliotecas de cDNA são uma representação aproximada da taxa de mRNA presentes em um dado tecido ou organismo, em uma dada condição e tempo. Genes mantenedores da vida (ou *housekeeping*) têm em teoria expressão constante nas células e tecidos e podem apresentar amostragem bastante redundante. Já outros tipos de genes podem ter uma oscilação em sua expressão de acordo com o tecido ou condições ambientais. Genes pouco expressos terão pouca representação enquanto genes ausentes não serão representados nessas bibliotecas. Apesar da presença de uma EST ser um indício

confiável da expressão de um gene, a ausência não indica necessariamente que um gene não existe, mesmo porque alguns genes são muito pouco conservados entre organismos diferentes. Somente pode-se afirmar que não foi possível encontrar um dado transcrito naquela amostra (Rudd, 2003).

As EST são seqüências curtas e com baixa qualidade. Apresentam em torno de 4% de erro, e a qualidade da seqüência é em média bem melhor no centro da seqüência do que nas pontas. Evidentemente isto depende do tipo de seqüenciador utilizado e do pós-processamento que foi aplicado, mas principalmente pelo fato de ser seqüenciada em um único passo. Ainda, seqüências contaminantes provindas de vetores de clonagem e sítios de *polylinkers* estão presentes frequentemente nas extremidades de uma EST depositada em bancos públicos (Nagaraj *et al.*, 2007).

1.3.1.3.1 Edição e processamento de seqüências EST

O produto primário do seqüenciamento em seqüenciadores automáticos são os eletroferogramas ou cromatogramas (ver Fig. 1), que precisam ser processados por programas *base-callers* (nomeadores de bases) no intuito de se gerar as seqüências em um formato padrão, como o FASTA por exemplo (Fig. 5). Alguns seqüenciadores automáticos já são comercializados com programas nomeadores de bases próprios, porém um dos mais utilizados pela comunidade científica é o PHRED (*Phil's Read Editor*) (Ewing e Green, 1998; Ewing *et al.*, 1998). O PHRED, assim como todo programa nomeador de bases, gera e analisa os cromatogramas determinando as bases de acordo com a intensidade da fluorescência medida pelo laser. De acordo com a acurácia da medida, são atribuídos valores de qualidade para as bases nomeadas. O valor de qualidade das bases nomeadas pelo programa PHRED é calculado pela fórmula: $-10 \times \log_{10}(p)$. Onde p é a probabilidade da base estar errada. Como exemplo, um valor resultante 10 significa que a base tem 10% de chance de estar incorreta, assim como um valor de 20 significa que ela tem 1% de chance de estar incorreta. A nomeação de bases assim como o próprio processo de seqüenciamento em um único passo são processos que geram diversos erros que são incorporados pela seqüência resultante. Por isso a EST é dita uma seqüência de baixa qualidade. Regiões de baixa qualidade podem ser removidas pelo próprio algoritmo PHRED, utilizando-se o comando "trim_alt". O algoritmo PHRED foi extensivamente

discutido no trabalho de tese de doutorado de Francisco Prosdocimi (Prosdocimi, 2006) de nosso laboratório, e em suas publicações (Prosdocimi *et al.*, 2003; Prosdocimi *et al.*, 2004).
Outra parte importante na edição e processamento das EST é a remoção de seqüências indesejadas como sítios de *polylinkers* e seqüências do vetor de clonagem. Para isso existem programas que mascaram essas seqüências. Um dos mais conhecidos é o *cross-match* que é distribuído em conjunto com o programa PHRED. Esse programa alinha as EST com seqüências conhecidas de vetores e outros contaminantes, e mascara essas seqüências, trocando as bases por 'X' (de onde o nome *cross-match*). Diversas metodologias para edição e processamento de EST foram (Chou e Holmes, 2001) e até hoje são criadas (Lee *et al.*, 2007).

```
>001_4_1_A09.esd    166      0    166  ESD trimmed
CGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCT
GCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAG
AATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAG
GCCAGGAACCGTAAAA
```

Figura 5. Amostra de seqüência no formato FASTA, gerada pelo programa PHRED. Note que essa seqüência de 166 bases já foi podada (*trimmed,* tendo zero bases de baixa qualidade no início da seqüência*)* para manter apenas 166 bases com qualidade de PHRED 20 (apenas 1% de chance de erro).

1.3.1.3.2 Cálculo de Expressão diferencial de genes usando EST

Bibliotecas de cDNA podem ser usadas para encontrar genes diferencialmente expressos usando-se seqüências EST provindas de diferentes tecidos usados na construção das bibliotecas. Encontrar esses genes pode ser de importância médica e farmacêutica (Stekel *et al.*, 2000). Alguns trabalhos foram inicialmente publicados tratando do assunto em bibliotecas construídas de células e tecidos de organismos como *H. sapiens* e *S. mansoni* (Lee *et al.*, 1995; Franco *et al.*, 1997). Um dos trabalhos mais importantes nessa área foi o projeto CGAP (*Cancer Genome Anatomy Project)*, destinado ao estudo de genes diferencialmente expressos em tecidos tumorais humanos (O'brien, 1997). Atualmente, apesar da aparente massificação do uso de microarranjos na análise de expressão gênica diferencial, ainda são publicados trabalhos usando seqüências EST nesta área (Aouacheria *et al.*, 2006).
Técnicas estatísticas para a análise e descoberta de genes diferencialmente expressos em bibliotecas de cDNA foram utilizadas e aprimoradas no final da década de 90 e início do

ano 2000. Uma dessas técnicas é o teste exato de Fisher usado no DDD ou *Digital Differential Display*, procedimento utilizado no UniGene (Pontius *et al.*, 2003) pelo projeto CGAP (Scheurle *et al.*, 2000). O DDD utiliza os agrupamentos Unigene, permitindo ao usuário formar dois conjuntos (*pool*) de bibliotecas e reporta os Unigenes diferencialmente presentes nos dois conjuntos O teste exato de Fisher é um teste de significância estatística usado para verificar a probabilidade de erro ao se constatar genes diferencialmente expressos nesses dois *pools*. Essa validação é realizada pela função de distribuição de probabilidade hipergeométrica.

Também foram desenvolvidas outras ferramentas para cálculo de expressão diferencial de genes no CGAP como o *DGED*, também baseado no UniGene. O DGED usa uma relação de probabilidade de ocorrência e estatística bayesiana para realizar e validar os cálculos de diferença de expressão entre duas bibliotecas de cDNA (Lal *et al.*, 1999). No SAGE DGED também podem ser usados SAGE *tags*, onde é usada uma taxa de distribuição de *tags* conhecida *a priori* para calcular a probabilidade a posteriori desta *tag* ocorrer, no intuito de validar a freqüência daquela *tag* em uma dada biblioteca.

Algumas críticas surgiram quanto ao uso do teste exato de Fisher para o cálculo de genes diferencialmente expressos e novas técnicas foram desenvolvidas na tentativa de se criar testes mais acurados estatisticamente (Audic e Claverie, 1997) assim como técnicas que usam comparação entre mais de uma biblioteca simultaneamente (Greller e Tobin, 1999; Stekel *et al.*, 2000).

Uma das vantagens do método de cálculo de diferença de expressão descrito por Stekel e colaboradores em 2000 é a possibilidade da comparação entre múltiplas bibliotecas de cDNA ou *pools* no cálculo de expressão diferencial. Esse método consiste no cálculo de um valor denominado R que representa o valor de heterogeneidade real da amostra em relação à variação amostral, resultando em uma taxa logarítmica da probabilidade de diferença de expressão. Subseqüentemente é feita uma verificação do valor de R por geração de *pools* de amostra com valores randômicos com o uso da função de distribuição de Poisson. Esses *pools* randômicos são usados para calcular a probabilidade esperada, tornando possível comparar com a probabilidade amostral. É calculado então um valor de credibilidade de R, denominado *believability*, que indica a porcentagem de probabilidade daquele valor representar uma variação verdadeira, ou expressão diferencial.

1.3.1.3.3 *Agrupamento de EST e montagem de consensos*

Seqüências EST convencionais representam as extremidades 5’ e 3’ de um transcrito. É possível tentar sobrepor EST na tentativa de eliminar a redundância e formar agrupamentos que possuam mais informação sobre o transcrito. Agrupamentos de EST 5’ que se sobreponham a agrupamentos de EST 3’ podem formar um agrupamento único que represente um transcrito mais longo e possivelmente completo (Bouck e Vision, 2007). A forma mais fácil de produzir agrupamentos de EST é via similaridade por métodos de alinhamento entre seqüências. A primeira iniciativa em grande escala de se produzir agrupamentos de EST com esse propósito foi a do UniGene (Boguski e Schuler, 1995) e a mais tarde a do STACKPACK (Miller *et al.,* 1999). A idéia do UniGene é a de reunir EST, vindas principalmente do banco dbEST do NCBI (Boguski *et al.*, 1993), em agrupamentos que representem um único transcrito ou produto gênico. O método de agrupamento do UniGene utiliza o programa MegaBLAST (Zhang *et al.*, 2000) para realizar o alinhamento de todas as seqüências contra todas. Pela técnica de *single-linkage*, seqüências que alinham com 96% de similaridade em 70% da região potencialmente alinhável vão sendo agrupadas (Fig. 6, Lucas Wagner, comunicação pessoal).

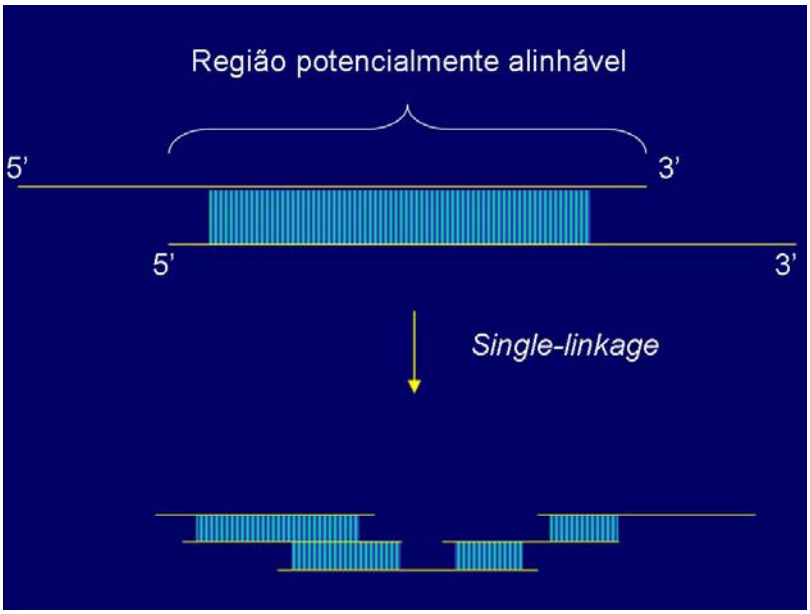


Figura 6. Esquema, mostrando o alinhamento entre duas seqüências (parte superior) mostrando o alinhamento (em azul) e a região potencialmente alinhável (chave). O método de *single-linkage* agrupa seqüências cujos alinhamentos possuem 96% de similaridade e ocupam 70% da parte potencialmente alinhável.

O UniGene porém não produzia seqüências únicas formadas pelo consenso do agrupamento de EST (Fig. 7A, letras b e c, e Fig. 7B) . Programas utilizados na montagem de genomas são usados nesse intuito, que apresentam bons resultados, como o PHRAP (http://www.phrap.org) e o CAP3 (Huang e Madan, 1999). Ainda, algumas iniciativas utilizam ambas as técnicas de clusterização e formação de consenso em uma metodologia híbrida como é o caso do TIGR *gene índices* (Pertea *et al.*, 2003). Alguns trabalhos foram publicados no intuito de avaliar essas metodologias de clusterização e montagem de consensos (Liang *et al.*, 2000). Mais recentemente foram avaliados critérios como divisão de consensos que representavam um único transcrito (erro tipo I), assim como montagem de quimeras (consensos com EST de mais de um transcrito, erro tipo II) (Wang *et al.*, 2004).

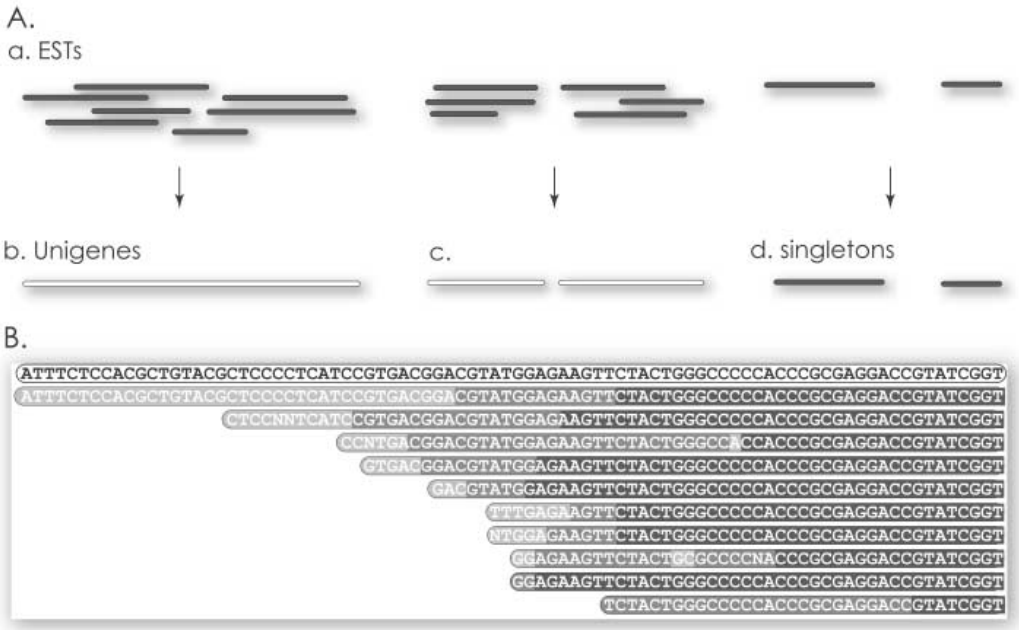


Figura 7. Montagem de *Unigenes ou Uniques.* A. Formação dos consensos, também chamados *Unigenes* ou *Uniques* (b e c) pelas EST (a); seqüências *singleton* (d) são EST órfãs. B. Exemplo de seqüência consenso (em branco, parte superior), e as seqüências que a representam (tons de cinza). Retirado de (Bouck e Vision, 2007).

1.4 Anotação de seqüências via alinhamento

Uma maneira de se identificar genes em seqüências biológicas é realizando buscas por homologia entre as seqüências desconhecidas e seqüências de genes conhecidas. A maneira mais usual de se executar esse processo é usando bancos de dados de seqüências já conhecidas e programas de alinhamento entre seqüências como BLAST (Altschul *et al.*, 1990) ou algoritmos como Smith-Waterman (Smith e Waterman, 1981). Com o resultado de alinhamento é possível verificar a similaridade entre as seqüências e inferir homologia (revisto por Koonin e Galperin, 2003).

1.4.1.1 Algoritmos para alinhamento de seqüências

Alinhamento de seqüências é o procedimento de se comparar duas (alinhamento par a par) ou mais seqüências (alinhamento múltiplo), na busca por caracteres ou padrões de caracteres que seguem a mesma ordem nas seqüências. O alinhamento de seqüências é útil na descoberta de informações funcionais, estruturais e evolutivas entre seqüências biológicas (Mount, 2001). Seqüências muito similares entre dois organismos provavelmente realizam a mesma função e se são de organismos diferentes, provavelmente se originaram de um ancestral comum (Koonin e Galperin, 2003).

Existem dois métodos mais conhecidos de alinhamentos entre seqüências, o global e o local (figura 8). No global é feita uma tentativa de se alinhar toda a seqüência, enquanto no local são alinhadas subseqüências, ou regiões. O alinhamento local é usado de maneira mais geral por algumas razões, como por exemplo, a existência de regiões conservadas (domínios conservados) entre seqüências. Outro exemplo é o fato de que regiões de proteínas ou DNA acabam divergindo demais, apenas sobrando pequenas regiões com homologia detectável (Koonin e Galperin, 2003).

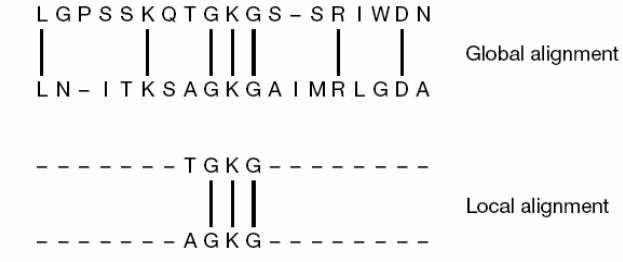


Figura 8. Exemplo de alinhamento global e local. Retirado de (Mount, 2001).

Em ambos os casos de alinhamentos (global e local), são montadas matrizes de pontuação entre as seqüências e aplicadas regras de pontuação positiva (*score*), no caso de bases iguais (*match*) e de pontuação negativa, ou penalidades, no caso de bases diferentes (*mismatch*) ou para a abertura de espaços (*gaps*) e alongamento de espaços. Os valores de pontuação mudam de acordo com o programa de alinhamento utilizado, ou são alteráveis pelo usuário.

Algoritmos de alinhamento tratam as seqüências como cadeias de caracteres e usam do sistema de pontuação para encontrar a melhor solução (no caso de algoritmos de alinhamento ótimos) ou soluções aproximadas (algoritmos heurísticos). É importante ressaltar que essa pontuação não está ligada à significância estatística do alinhamento, que deve ser computada separadamente, usando por exemplo a estatística de Karlin-Altschul (Karlin e Altschul, 1990), utilizada no programa BLAST.

Algoritmos de alinhamento ótimos são aqueles que encontram sempre a melhor solução, ou o melhor alinhamento. Os algoritmos ótimos têm um custo computacional muito alto e em alguns casos podem ser impossíveis de serem executados em tempo viável. Existem algoritmos que contornam esses casos, e usam de heurísticas para acelerar a execução. Seu resultado poderá não ser ótimo e apenas se aproximar do resultado ótimo (Koonin e Galperin, 2003).

Os algoritmos mais conhecidos de alinhamentos global e local utilizam uma técnica computacional chamada programação dinâmica. A idéia básica da programação dinâmica vem do fato de que um caminho que leve a uma solução ótima pode ser dividido em sub-caminhos ótimos. De forma que ao se estender sub-caminhos ótimos, chega-se sempre à solução ótima. No caso de seqüências, o alinhamento ótimo global é feito comparando-se seqüências de ponta à ponta, sempre guardando em uma matriz as pontuações para os alinhamentos em cada base comparada. No final, essa matriz possuirá em cada ponto do alinhamento, o melhor valor de pontuação, de forma que os melhores valores indicarão o melhor alinhamento (Baxevanis e Ouellette, 2001).

No alinhamento entre proteínas, são utilizadas matrizes de pontuação pré-computadas, que levam em conta a probabilidades de trocas e mutações entre aminácidos. Leva-se em conta estudos evolutivos e de propriedades bioquímicas de aminoácidos. As matrizes mais

conhecidas são a PAM (*Point Accepted Mutation*), criadas por Margareth Dayhoff e as BLOSUM utilizadas como padrão pelo programa BLAST.

1.4.1.1.1 Alinhamento Global

O algoritmo mais conhecido para alinhamento global ótimo entre seqüências é o Needleman-Wunsch (Needleman e Wunsch, 1970). Algoritmos de alinhamento global são a melhor forma de se analisar relações evolutivas entre seqüências. Existem diversos algoritmos de alinhamento global não ótimos. Entre eles estão o MultiAlign (Corpet, 1988) e o Clustal (Higgins e Sharp, 1988).

1.4.1.1.2 Alinhamento Local

O primeiro algoritmo para alinhamento local ótimo, utilizando programação dinâmica foi o Smith-Waterman (Smith e Waterman, 1981). Esse algoritmo é uma variação do algoritmo de Needleman-Wunsch adaptado para realizar alinhamentos locais. Por ser um algoritmo ótimo, sua execução pode ser não factível em computadores atuais. Diversos programas que usam algoritmos heurísticos para alinhamento local surgiram, na tentativa de gerar soluções aproximadas. Entre eles os mais conhecidos são o FASTA (Pearson e Lipman, 1988) e o BLAST (Altschul *et al.*, 1990).

1.4.1.1.3 BLAST

Distribuído pelo NCBI (http://www.ncbi.nlm.nih.gov/BLAST) o software BLAST é um dos programas para alinhamento local, heurístico, mais utilizados atualmente. Existem diversas variedades de BLAST (Tabela 1), as quais realizam alinhamentos entre seqüências *query* e seqüências *subject*, que podem ser nucleotídeos e/ou aminoácidos ou variações.

Tabela 1. Variedades de BLAST.

Tipo	*Query*	*Subject*
BLASTn	Nucleotídeo	Nucleotídeo
BLASTp	Proteína	Proteína
BLASTx	Nucleotídeo traduzido em proteína	Proteína
tBLASTn	Proteína	Nucleotídeo traduzido em proteína
tBLASTx	Nucleotídeo traduzido em proteína	Nucleotídeo traduzido em proteína

1.4.1.1.4 HSP e a estatística Karlin-Altschul

Como todo programa de alinhamento local, ao invés de procurar por alinhamentos perfeitos entre duas seqüências inteiras, o BLAST realiza buscas por fragmentos ou regiões de alto *score*. Isso é feito utilizando-se um pedaço (semente) da seqüência *query* de um determinado tamanho e realizando uma busca por um alinhamento perfeito dessa semente contra o *subject.* Encontrado o alinhamento perfeito da semente, o BLAST tenta estendê-lo para as vizinhanças, até que o *score* não possa ser mais incrementado (a similaridade cai demais). Esse fragmento de alinhamento (da semente, somado à extensão para as bases vizinhas) é chamado de HSP (*High Scoring Segment Pair*). Um HSP deve ser estatisticamente validado, o que significa que ele tem de ter baixa probabilidade de ocorrer ao acaso. Por inferência, um HSP deve pertencer provavelmente a homólogos e ser biologicamente relevante (Koonin e Galperin, 2003).

Karlin e Altschul descobriram que os valores máximos de *scores* de HSP seguem a distribuição de valor extremo (Karlin e Altschul, 1990) e chegaram à seguinte fórmula:

- $E = Kmne^{-\lambda S}$

Onde *E* é o valor que mede a probabilidade daquele HSP ter ocorrido ao acaso. *S* é o valor do sistema de *score* utilizado, *K* e λ são valores naturais de escala para o espaço de busca e o valor de *score*. O produto *mn* é o espaço de busca, onde *m* representa o tamanho da seqüência *query* e *n* representa o tamanho do *subject* (quando o *subject* representar múltiplas seqüências, então BLAST multiplica n por (n / N), onde N é o somatório de bases de todas as seqüências do *subject*). O espaço de busca é crítico nesse cálculo já que o tamanho da base de dados assim como o tamanho da sequencia *query* influem diretamente no valor de *E*. Portanto, um dado HSP pode ser estatisticamente significativo ao se usar bases de dados com poucas seqüências mas não em bases de dados maiores, assim como o mesmo HSP pode ser significativo ao se alinhar seqüências *query* de tamanho pequeno e não ser significativo ao se alinhar uma seqüência maior (Koonin e Galperin, 2003).

1.5 Bases de Dados de Seqüências Biológicas

Com a extensa comercialização dos seqüenciadores automáticos de DNA, projetos de seqüenciamento em larga escala, principalmente de genomas, se tornaram bastante comuns. Grandes projetos de seqüenciamento, principalmente de organismos modelo como *C. elegans, D. melanogaster, A. thaliana* e *H. sapiens* terminaram e suas seqüências foram depositadas em bases de dados públicas. A função dessas bases, além de conter as próprias seqüências, também é a de prover acesso direto a elas e a informações relacionadas, como autores, organismo a que pertence, e anotação (Mount, 2001). A indexação de bases para agilizar o acesso a seqüências, também se tornou um fato comum. Além disso, o sistema de busca em bases de dados se tornou simplificado e, principalmente, rápido (Lesk, 2002).

Os sítios e núcleos de pesquisa em bioinformática que comportam as bases de dados de seqüências biológicas mais importantes atualmente são o americano NCBI – *National Center for Biotechnology Information* (http://www.ncbi.nlm.nih.gov), o europeu EMBL-EBI – *European Bioinformatics Institute* (http://www.ebi.ac.uk) e o japonês GenomeNet (http://www.genome.ad.jp). Esses sítios além de prover acesso às bases de dados também provêm serviços como a realização de buscas por homologia nas seqüências de suas bases.

1.5.1 Bases de dados primárias e secundárias

Uma base de dados primária é uma base de dados onde seqüências são depositadas com critérios de avaliação relaxados ou nulos. De fato, muitos fragmentos de seqüências, informações erradas e seqüências de baixa qualidade são encontradas nessas bases (Krawetz, 1989). Um exemplo de base de dados primária é o GenBank (Burks *et al.*, 1985) situado no NCBI.

Bases de dados secundárias são bases que passaram por curadoria manual onde apenas seqüências com qualidade e anotação que respondessem aos critérios desejados são mantidas. Essas bases de dados são menores em tamanho, mas mais confiáveis em termos de informação, o que as torna atrativas para a realização de buscas por homologia mais acuradas. Exemplos de bases de dados secundárias são o RefSeq (Pruitt *et al.*, 2000), UniRef (Suzek *et al.*, 2007a), COG e KOG (Tatusov *et al.*, 2000; Tatusov *et al.*, 2001; Tatusov *et al.*, 2003).

1.5.1.1 Bases de dados de seqüências ortólogas

As bases de dados de seqüências ortólogas são bases secundárias cujas seqüências são separadas em agrupamentos por critérios evolutivos. Exemplos são os COG e KOG, citados acima, KEGG Orthology (Kanehisa *et al.*, 2002), PIRSF (Wu *et al.*, 2004) , OrthoMCL-DB (Li *et al.*, 2003) e InParanoid (O'brien *et al.*, 2005).

Os principais critérios evolutivos supracitados são os de ortologia e paralogia. Esses termos começaram a ser muito utilizados em 1995, exatamente quando os primeiros genomas de organismos celulares foram seqüenciados e suas seqüências puderam ser comparadas (Koonin, 2005). Genes ortólogos podem ser definidos como genes derivados do último ancestral gênico comum entre as espécies comparadas (Figura 9). Parálogos são genes relacionados por duplicação gênica (Figura 10).

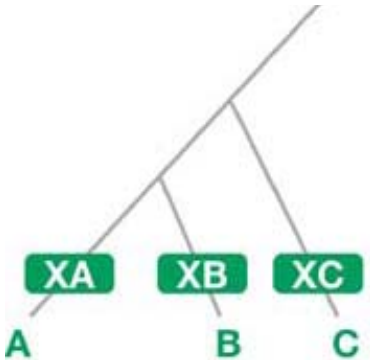


Figura 9. Definição de ortologia. Os organismos A, B e C, após eventos de especiação, possuem os genes XA, XB e XC que são ortólogos entre si numa relação um para um. Retirado de (Koonin, 2005).

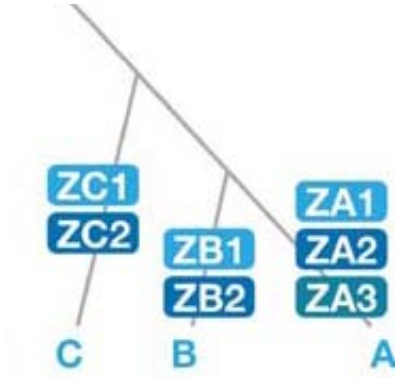


Figura 10. Definição de paralogia. Duplicações gênicas nas espécies A, B e C levam ao surgimento de genes parálogos entre si (ZC1 e ZC2; ZB1 e ZB2; ZA1-3). Retirado de Koonin, 2005.

Porém uma típica característica de genes ortólogos é a de realizarem a mesma função, apesar de ser difícil comprovar essa suposição em contextos biológicos diferentes.

A situação se complica um pouco mais, quando levamos em conta os termos co-ortólogos, *inparalogs*, *outparalogs*. Co-ortólogos são genes que duplicaram em uma espécie e são simultaneamente ortólogos de genes em outra espécie (exemplo na Figura 10, os genes ZA1-3 são co-ortólogos dos genes ZB1-2 e ZC1-2). *Inparalogs* são genes parálogos resultantes de uma duplicação logo após uma especiação. *Outparalogs* são genes parálogos resultantes de duplicação antes de uma especiação (revisto por Koonin, 2005).

1.5.1.1.1 Base de dados COG e KOG

As bases de dados COG e KOG são bases de dados de proteínas ortólogas de organismos procariotos (COG) e eucariotos (KOG). A base COG foi montada usando-se mais de 100 mil proteínas de 69 organismos cujo genoma foi totalmente seqüenciado. A base KOG é uma extensão da COG e usa mais de 88 mil seqüências de proteínas de 7 organismos eucariotos, com genoma também totalmente seqüenciado.

As bases COG e KOG foram criadas de forma a tentar identificar proteínas que tivessem uma relação de ortologia. Dessa forma, proteínas de um mesmo COG/KOG evoluíram de um ancestral gênico comum por uma série de eventos de especiação e duplicação (Tatusov *et al,* 1997*)*.

Essas bases foram criadas sobre a premissa de que proteínas ortólogas são mais próximas ou similares entre si do que qualquer outra na base de dados. Por isso foram usados os melhores alinhamentos (*BeT,* do inglês *Best Hit*) entre todas as proteínas contra todas de modo a aglomerar as seqüências em *clusters* de pelo menos três organismos. Dessa forma um *BeT* é uma proteína de um genoma *query* que é mais semelhante a uma proteína do genoma *subject*. Essa metodologia consegue lidar com dois problemas. A questão de proteínas linhagem específicas, ou seja, que surgiram após a divergência das espécies em questão e também o problema de proteínas com múltiplos domínios, que podem ser originadas de um único gene em uma espécie mas por dois ou mais genes separados em outras espécies, o que poder gerar um agrupamento artefactual.

Dessa forma, os *BeT* de cada proteínas são analisados e usados para formar a unidade mínima dos COG/KOG, que é um triângulo. Ou seja, COG/KOG são formados por proteínas de pelo menos 3 linhagens / espécies diferentes. Em seguida, tenta-se unir as

bases de triângulos, de forma a aumentar os agrupamentos. A base KOG ainda provê grupos provisórios de proteínas com somente duas espécies, ou TWOG, e grupos de linhagem específica, ou LSE. O uso de *BeT* na geração dos triângulos não depende do nível absoluto de similaridade entre as proteínas comparadas, o que permite aglomerar proteínas que vieram de genes que evoluem mais lenta ou rapidamente.

Tanto COG como KOG foram criados por metodologias que não são totalmente automatizadas. O protocolo utilizado inclui um procedimento automatizado para detectar candidatos a ortólogos, separação manual dos domínios componentes da proteína e inspeção manual e anotação. A análise dos *BeT* em KOG ainda passa por um processo de identificação de domínios conservados pelo banco CDD (Marchler-Bauer *et al.*, 2002), tendo em vista a grande gama de arquiteturas multidomínio de proteínas eucarióticas, além do fato de que muitas vezes proteínas ortólogas diferirem na composição domínios (Tatusov *et al.,* 2003).

As bases KOG e COG possuem também um atrativo para a anotação automática de seqüências que é a classificação das proteínas e agrupamentos em grupos funcionais.

2 Objetivos

2.1 Objetivo Geral

Utilizar a base de dados secundária KOG como ferramenta para mineração de dados e caracterização de expressão gênica, usando anotação automática de seqüências de EST provindas de projetos transcriptoma.

2.2 Objetivos Específicos

1. Construir um banco de dados local relacionando seqüências de EST públicas, de organismos presentes na base de dados KOG, assim como as informções provenientes dessa base.
2. Definir um limiar de corte de similaridade para o programa tBLASTn, compatível com o de 96% usado no UNIGENE para o programa MegaBLAST usando proteínas KOG (muito expressas e sem parálogos) e seqüências de EST de um mesmo organismo.
3. Definir uma metodologia para avaliar a anotação automática de seqüências de EST na base de dados KOG, usando organismos presentes na própria base.
4. Verificar essa mesma metodologia usando seqüências de EST clusterizadas com o programa TGICL.
5. Usar a base de dados KOG para caracterizar funcionalmente as seqüências de EST obtidas e tentar revelar informações biológicas pertinente sobre a amostragem de seqüências de EST ou expressão gênica desses organismos.
6. Definir o nível de cobertura da base KOG usando quantidades crescentes de seqüências de EST para verificar impacto do número de seqüências de EST na caracterização do transcriptoma.
7. Desenvolver uma ferramenta *web* "K-EST" que permita o acesso a dados de amostragem de seqüências de EST por agrupamentos KOG, diferença de expressão entre organismos, conservação desta amostragem dentre outras funcionalidades.
8. Verificar a hipótese de perda ou ausência de genes utilizando os dados de amostragem de seqüências de EST e a base de dados KOG, disponibilizados pela

ferramenta *web* construída. Usar ESTs de *Schistosoma mansoni* para testar a metodologia.

3 Justificativa e Relevância

Desde o surgimento da técnica de seqüenciamento de EST, bancos de dados públicos dessas seqüências estão em constante crescimento e a demanda para análise e anotação desse grande conjunto de informação virtual cresce simultaneamente. A construção e seqüenciamento de bibliotecas de cDNA e a geração de EST são uma das técnicas atuais mais usadas nas análises de transcriptomas. Apesar de ainda ter custo elevado em relação a outras (ex.: SAGE), as EST possibilitam estudos mais complexos que vão além da caracterização da expressão gênica. Exemplos são: o estudo de polimorfismos de base única (SNP), auxílio na descoberta gênica e montagem de genomas, possibilidade da identificação e seqüenciamento de clones completos que podem prover a caracterização da seqüência de uma proteína completa. Além disso, o sequenciamento de cDNA é uma técnica cujos custos estão em tendência de queda.

O seqüenciamento completo de genomas de diversos organismos possibilitou surgimento de diversos bancos de dados públicos de seqüências ortólogas curadas, amplamente disponíveis para análises bioinformáticas. Como exemplo estão os bancos de dados do NCBI, COG e KOG e também diversos outros como o Uniref, PIRSF e KEGG Orthology. Esses bancos, que também estão em constante crescimento, possuem coleções de seqüências já anotadas e curadas que constituem uma rica fonte para a anotação de novas seqüências. A anotação automática de novas seqüências, além de ser amplamente utilizada no meio científico, é uma maneira rápida e eficiente de se obter informações sobre os milhares de seqüências de EST novas e sem função definida. Todavia, não existem metodologias para medida de performance da base de dados na anotação automática.

Esse trabalho descreve um conjunto de novas técnicas, dados e ferramentas gerados e obtidos a partir da anotação automática de coleções públicas de EST e o banco de seqüências protéicas KOG. Acreditamos que os resultados obtidos nessa tese serão de grande valia como recurso bioinformático na mineração de dados e análise de seqüências provindas de projetos transcriptoma geradores de EST em andamento ou já concluídos. Além disso, esse trabalho pode prover auxílio na avaliação dos bancos de dados de proteínas ortólogas já existentes, como os próprios COG/KOG e também na criação de bancos de dados similares.

4 Materiais e Métodos

4.1 Hardware

A maior parte dos programas e bancos de dados foram instalados e usados em estações de trabalho comuns. Sistemas operacionais, programas, bancos de dados, servidores de páginas *web*, e demais softwares foram instalados e usados em sistema operacional Linux, em distribuições diversas, com o predomínio de RedHat 8.0 ou superior (http://www.redhat.com), Fedora (http://www.redhat.com/fedora), CentOS (http://www.centos.org) e Suse (http://www.novell.com/linux).

4.2 Bancos de Dados

Foram usadas diversas versões de MySQL sendo que a atual é a 5.01.

4.3 Servidores de Páginas Web

Servidores de páginas *web* Apache em conjunto com a linguagem PHP foram usados preferencialmente.

4.4 Softwares

De um modo geral, foram usados softwares, linguagens de programação e demais serviços nativos ou próprios para as distribuições Linux, como:

- PHP
- PERL e demais pacotes adquiridos do sítio CPAN (http://www.cpan.org).
- MySQL
- Apache
- Servidores de SSH, FTP, impressão.
- LaTeX

Softwares específicos para bioinformática usados:

- BLAST, versões 2.2.8 até 2.2.13
- PHRED 0.000925.c

- PHRAP 0.990329
- TGICL - adiquirido do sítio http://compbio.dfci.harvard.edu/tgi

Softwares Windows utilizados:

- SigmaPlot, versão 8.0 e 10.0 de demonstração; Systat Software, Inc.1735, Technology Drive, Ste 430 San Jose, CA 95110, EUA. (http://www.systat.com/products/SigmaPlot/)

4.5 Livros e páginas da Internet consultados

O aprendizado de linguagens e técnicas de programação, banco de dados, publicação na *web*, manutenção de máquinas, instalação de programas e sistemas operacionais, manutenção de espaço em disco, *becape,* administração de contas de usuários, etc, foi obtido a partir de consultas feitas nos sítios e referências a seguir.:

4.5.1 Linguagens de programação

4.5.1.1 Perl

- Cozens, S. and Safari Tech Books Online. (2005). "Advanced Perl programming." 2nd. from http://proquest.safaribooksonline.com/0596004567 *Acesso online*.
- Tisdall, J. D. (2001). Beginning Perl for bioinformatics. Sebastopol, CA, EUA. O'Reilly.
- Tisdall, J. D. and Safari Tech Books Online. (2003). "Mastering Perl for bioinformatics." *Acesso online* http://proquest.safaribooksonline.com/0596003072.
- Wall, L., T. Christiansen, et al. (2000). "Programming Perl." 3a. edição. *Acesso online* http://proquest.safaribooksonline.com/0596000278.
- Schwartz, R. L., T. Christiansen, et al. (1997). "Learning Perl." 2a. edição *Acesso online* http://proquest.safaribooksonline.com/1565922840.

4.5.1.2 PHP

- Sklar, D., A. Trachtenberg, et al. (2003). "PHP cookbook." *Acesso online* http://proquest.safaribooksonline.com/1565926811.

- http://www.php.net

4.5.2 Banco de dados

4.5.2.1 MySQL

- http://www.mysql.com

4.5.3 Sistema Operacional Linux / Unix

4.5.3.1 Linux e *Shell*

- Dougherty, D., A. Robbins, et al. (1997). "Sed & awk." 2a. edição. *Acesso online* http://proquest.safaribooksonline.com/1565922255.
- Robbins, A. and Safari Tech Books Online. (1999). "UNIX in a nutshell a desktop quick reference for System V Release 4 and Solaris 7." 3ª. Edição. *Acesso online* http://proquest.safaribooksonline.com/1565924274.
- Rosenblatt, B. and Safari Tech Books Online. (1993). "Learning the Korn shell." 1a. edição. *Acesso online* http://proquest.safaribooksonline.com/1565920546.
- GNU Awk User's Guide: http://www.gnu.org/software/gawk/manual/gawk.html

5 Resultados e discussão

5.1 Descarregando e alinhando seqüências EST com a base KOG

Obtivemos coleções públicas de seqüências de EST de quatro organismos modelo, *Arabidopsis. thaliana* (Ath), *Caenorhabditis elegans* (Cel), *Drosophila melanogaster* (Dme) *e Homo sapiens* (Hsa), (ver tabela de número 1, artigo 2 e também materiais e métodos do artigo 3) organismos presentes na base de dados KOG, e realizamos o alinhamento dessas seqüências com suas proteínas da base de dados KOG usando o software tBLASTn do pacote BLAST (Altschul *et al.*, 1990), usado para alinhar seqüências de proteínas (*query*) contra seqüências de nucleotídeos (*subject*). Foi usado um valor de corte de E-value bastante rigoroso (10^{-10}) em uma tentativa de remover alinhamentos espúrios. A descrição dos parâmetros usados no programa está nos artigos 2, 3, 4 e 5. Esse software foi escolhido em detrimento ao BLASTx (que alinha seqüências de nucleotídeos contra de proteínas), para o caso de um possível incremento da base de dados KOG com seqüências de outros organismos, de modo a não se fazer necessário realizar todos os alinhamentos novamente, e também para que os alinhamentos incluídos tenham a mesma significância para diferentes conjuntos de EST (já que o valor de E é dependente do tamanho da base de dados, ou *subject*). Esses alinhamentos foram uilizados para popular um banco de dados MySQL e usados subseqüentemente nos artigos 1, 2, 5, 7 e 8. Em outro momento, novos conjuntos de EST para os mesmos organismos foram adquiridas (materiais e métodos do artigo 3) o mesmo procedimento de alinhamento foi realizado e um novo banco de dados populado. Dessa vez, esses dados foram utilizados nos artigos 3, 4 e 6. A base de dados KOG, por sua vez foi previamente descarregada e também populada em banco de dados, assim como descrito nos artigos 2, 3, 4 e 5. Dessa forma, as informações relativas a esses alinhamentos foram colocadas de maneira a serem facilmente extraídas e manipuladas.

5.2 Ferramentas bioinformáticas aplicadas à caracterização da expressão gênica

O primeiro artigo, publicado em 2004 na revista *Bioscience Journal* de Uberlândia, MG (Faria-Campos *et al.*, 2004), trata de uma revisão sobre os trabalhos em bioinformática em

andamento no Laboratório de Biodados – ICB, UFMG em meados de 2004. Em linhas gerais, é explicado como é feita a nomeação de bases após o seqüenciamento de EST com o software PHRED (Ewing e Green, 1998; Ewing *et al.*, 1998) e a medição de erros dessa nomeação em um experimento controlado com o vetor pUC18, experimento este utilizado nos artigos 2 e 3, comentados adiante no item 4.3. O uso parcimonioso do software PHRED para maximizar a informação contida em EST foi assunto de trabalho de tese de Francisco Prosdocimi no laboratório (Prosdocimi, 2006).

O uso de proteínas de organismos modelo (*C. elegans*, Cel e *D. melanogaster,* Dme) para pesquisa de homologia com o software BLAST de seqüências EST do organismo *S. mansoni* (Sma) se mostrou eficiente e suficiente na anotação das seqüências desse verme. Esse estudo fez parte do trabalho de tese de Alessandra C. Faria-Campos no laboratório, foi publicado em 2006 na revista *In-Silico Biology* (Faria-Campos *et al.*, 2006a), e foi referencia para o artigo 7, onde seqüências de EST de Sma provenientes da Rede Genoma de Minas Gerais e públicas são anotadas com a base KOG. Esses dados foram também usados em estudos de diferença de expressão e predição de genes não existentes em Sma em comparação com Cel e Dme (artigo 7).

Também é delineada a anotação automática de seqüências EST em bases secundárias e cunhada a expressão "anotação reversa" para designar o uso coleções de proteínas já conhecidas para prover função a seqüências EST que podem designar clones completos. Os alinhamentos das seqüências de EST e as proteínas da base KOG, explicados no item 4.1, foram utilizados no cálculo de saldo de códons e caracterização de clones completos. Além disso, o uso da base de dados KOG para anotação das EST e o teste de eficiência dessa anotação, que se mostrou em torno de 90% acurada, são citados nesse artigo e tratados detalhadamente nos artigos 2, 3 e 4. O artigo 2 foi submetido ao congresso BSB 2005 (*Brazilian Symposium on Bioinformatics*) realizado em São Leopoldo, RS, em 2005 e publicado como trabalho completo na revista *Lecture Notes in Computer Sciences* (Mudado M *et al.*, 2005). O artigo 4 foi aceito para publicação também como trabalho completo nos anais do congresso BSB 2007, que será realizado em Angra dos Reis, RJ, em Agosto de 2007.

Assim, o artigo 1 relata uma das contribuições dessa tese, a performance de anotação automática com a base KOG (figura de número 18, artigo 1).

Original Article

FERRAMENTAS BIOINFORMÁTICAS APLICADAS À CARACTERIZAÇÃO DA EXPRESSÃO GÊNICA

USE OF BIOINFORMATIC TOOLS IN GENE EXPRESSION CHARACTERIZATION

Alessandra C. FARIA-CAMPOS; Maurício A. MUDADO; Fabiano C. PEIXOTO; Estevam BRAVO-NETO; Francisco PROSDOCIMI; José Miguel ORTEGA

RESUMO: Projetos transcriptoma são o retrato mais fiel do conjunto de seqüências expressas de um dado genoma, fornecendo as melhores evidências da existência de regiões que constituem a informação para a produção das proteínas do organismo. Uma das abordagens utilizadas neste tipo de projeto é a produção de etiquetas de seqüência transcrita (EST, do inglês *expressed sequence tag*), que vem se firmando como uma forte tendência na ciência nacional. A geração das ESTs em larga escala deve ser acompanhada de sua anotação, a qual converte a informação produzida em conteúdo biológico, explicitando quais genes são expressos em um dado organismo em um dado momento. Neste artigo relatamos a aplicação de ferramentas bioinformáticas à caracterização de ESTs e o uso de uma potente estratégia de escolha de etiquetas que representam genes de interesse, para que estes genes sejam seqüenciados com detalhe a ponto de se obter não somente a etiqueta, mas a dedução da seqüência completa da proteína codificada. São também relatadas a avaliação de diferentes valores de PHRED com o objetivo de maximizar a informação obtida da etiqueta e uma estimativa da qualidade da estratégia de escolha de clones, através da avaliação das porcentagens de erro. O procedimento de escolha foi denominado "anotação reversa" e representa uma importante contribuição para projetos transcriptomas adicionando a estes a dimensão de geradores de informação a ser utilizada por outros grupos de pesquisa.

UNITERMOS: ESTs, BLAST, anotação

INTRODUÇÃO

A produção de informação sobre os genes presentes em diversos organismos, de bactérias ao homem, cresceu exponencialmente nas últimas décadas (disponível em http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html). Apesar de muitos projetos de seqüenciamento terem escolhido o DNA genômico como alvo, o seqüenciamento do transcriptoma indiscutivelmente provê as melhores evidências da existência de regiões que constituem a informação para a produção das proteínas do organismo. Uma das abordagens utilizadas em projetos transcriptoma, é a produção em larga escala de etiquetas de seqüência transcrita (EST, do inglês *expressed sequence tag*). A geração das ESTs em larga escala deve ser acompanhada de sua anotação, a qual converte a informação produzida em conteúdo biológico, explicitando quais genes são expressos em um dado organismo em um dado momento. A produção de ESTs foi iniciada no Brasil pela iniciativa de descoberta gênica em *Schistosoma mansoni* (abordada por G. Franco e col. neste volume), resultando em duas publicações pioneiras no ano de 1995 (FRANCO et al., 1995a e b), antecipando desta forma em cinco anos a publicação do primeiro genoma completo seqüenciado no Brasil (SIMPSON et al., 2000). A obtenção de informação a partir das ESTs mostrou ser uma forte tendência na ciência nacional (BRANDAO et al., 1997, RAMALHO-ORTIGAO et al., 2001; FELIPE et al., 2003) e neste artigo relatamos a aplicação de ferramentas bioinformáticas à caracterização de ESTs e o uso de uma potente estratégia de escolha de etiquetas que representam genes de interesse, para que estes genes sejam seqüenciados com detalhe a ponto de se obter não somente a etiqueta, mas a dedução da seqüência completa da proteína codificada.

CONTEÚDO

Extraindo toda a informação da etiqueta

Uma das maneiras de se identificar a função do gene retratado pela seqüência parcial do mRNA representada pela etiqueta (figura 1) é pesquisar a homologia entre ela e seqüências de aminoácidos presentes em bases de dados públicas, como o GenBank. Para tanto, utilizam-se programas de busca de homologia, como os do pacote BLAST (ALTSCHUL et al., 1997). É comum para a comunidade científica verificar criteriosamente os

alinhamentos comparativos entre a EST e as seqüências depositadas, devido ao fato de os equipamentos seqüenciadores freqüentemente produzirem leituras automatizadas da seqüência de DNA com erros, em ambas extremidades. A quantidade de erros nesses locais pode em alguns casos ser tão significativa que softwares de alinhamento nem mesmo conseguem reconhecer a leitura gerada. É importante notar que o equipamento seqüenciador, na verdade, não gera a seqüência de bases, mas um "cromatograma" como o da figura 2, que é então processado por um software nomeador de bases PHRED (EWING et al., 1998a e b), o qual a partir da análise dos picos, interpreta o resultado como uma seqüência de bases e atribui uma chance de erro a cada base nomeada. Na escala de PHRED, 10% de erro corresponde a PHRED 10, 1% de erro a PHRED 20, 0,1% de erro a PHRED 30 e assim por diante. Uma questão que se impõe é: qual a quantidade de erro que se pode aceitar para que a leitura da etiqueta contenha o máximo de informação aproveitável, ou seja, corresponda a uma leitura que, quando comparada à ideal com o software BLAST, nem adicione bases extras (indesejáveis, pois o software de alinhamento não as reconheceria), nem tampouco perca bases que poderiam ser reconhecidas como parte da etiqueta daquela seqüência. A figura 3 mostra o resultado obtido quando 864 leituras do plasmídio pUC18, produzidas pela Rede Genoma de Minas Gerais, foram comparadas com a seqüência publicada do referido vetor. Percebe-se que se pode admitir uma densidade de erros de cerca de 18%, os quais são concentrados na extremidade da leitura (não mostrado), para que não sejam perdidas bases representando informação, nem tampouco sejam adicionadas bases que não seriam reconhecidas pelo software BLAST de alinhamento local. Nós admitimos, na Rede Genoma de Minas Gerais, 16% de erro nas extremidades, o que corresponde na fórmula de cálculo de PHRED ao valor igual a oito. Com isso, maximizamos a informação contida na etiqueta. Os projetos concorrentes mais arrojados neste aspecto utilizam PHRED 15 (3,2% de erro). Pode-se perceber, pelo nosso experimento, que o uso de PHRED neste valor acarreta significativa perda de informação e que o valor utilizado por nosso grupo apresenta uma relação custo-benefício melhor.

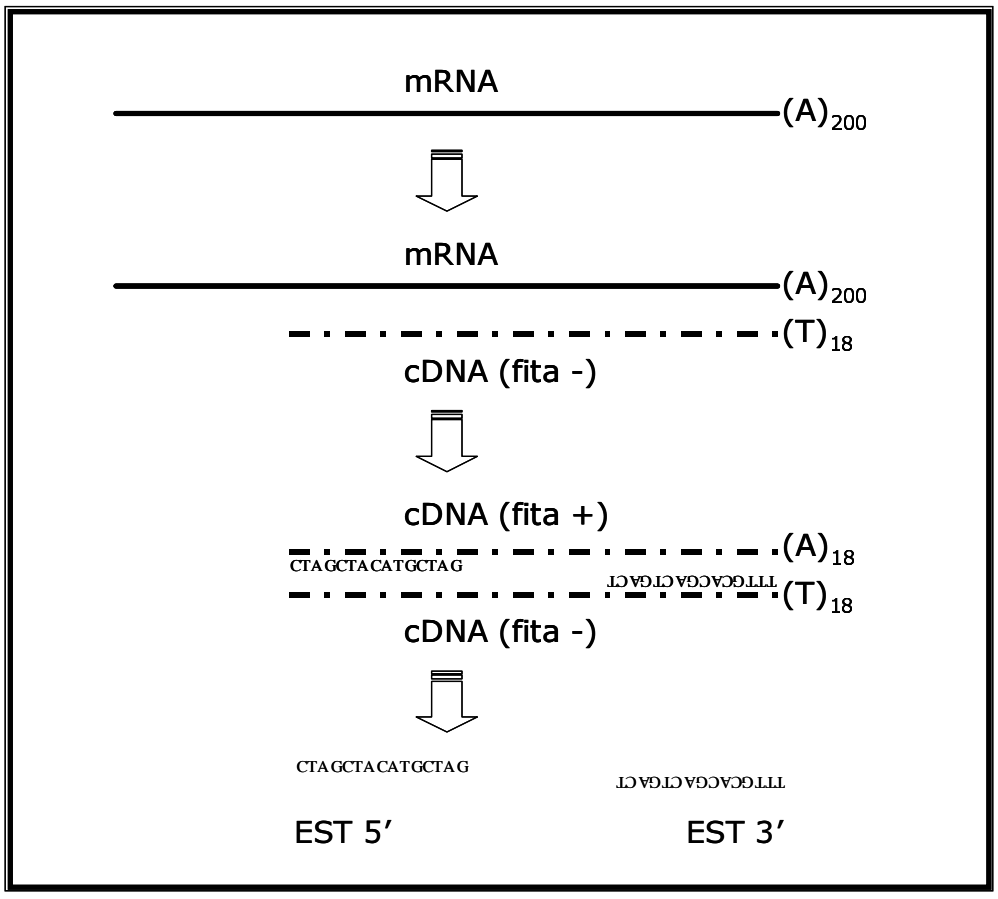
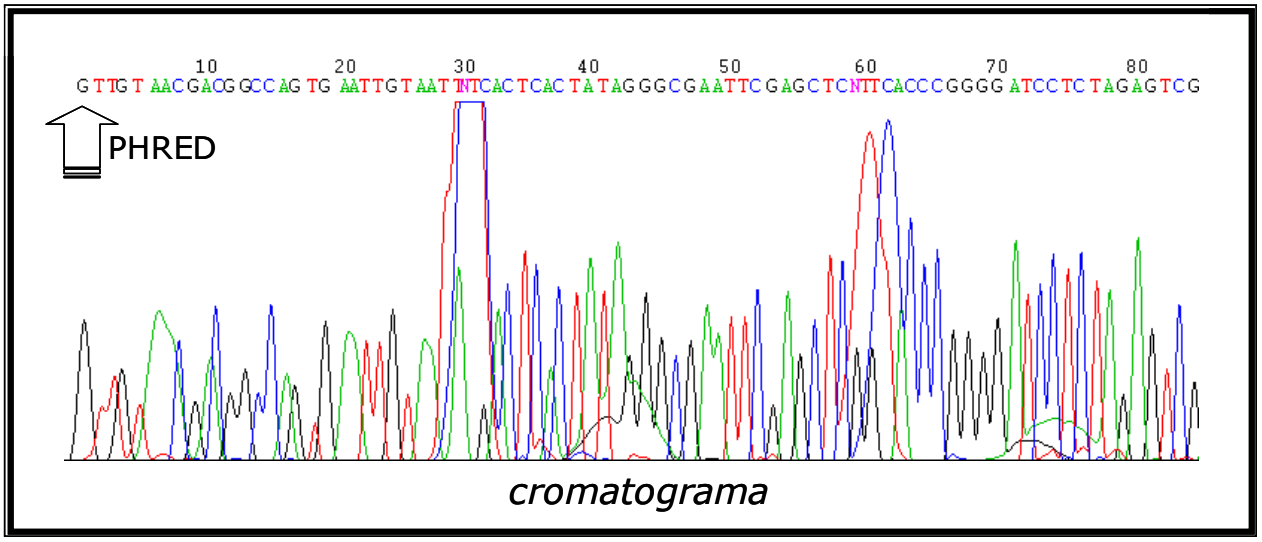


Figura 1: Esquema explicativo do procedimento de geração de ESTs, seqüências parciais em única tentativa de moléculas de cDNA, derivadas do mRNA pela transcrição reversa.



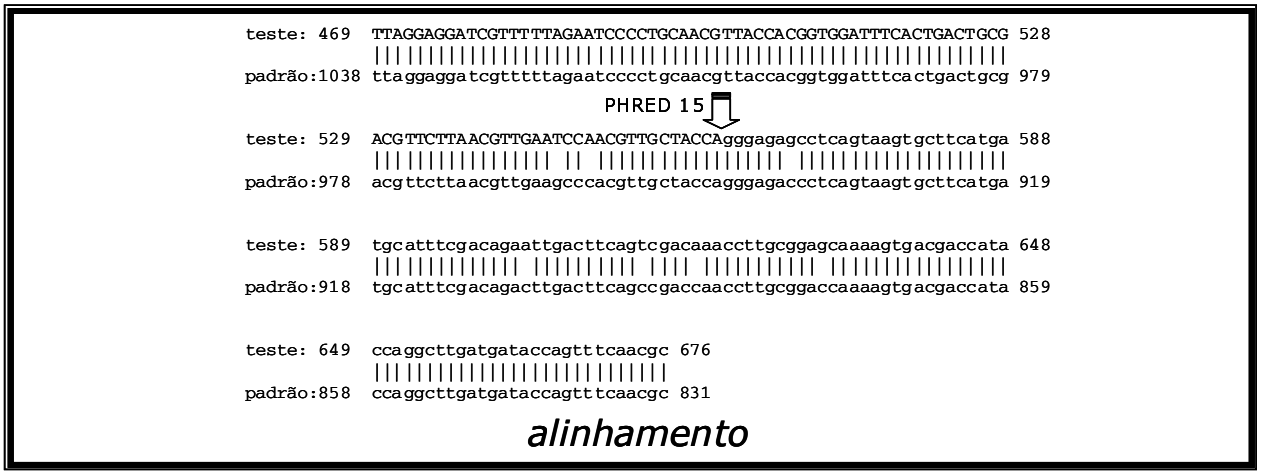


Figura 2: Cromatograma gerado a partir da leitura dos dados brutos do seqüenciador MegaBACE por um programa nomeador de bases (PHRED). Na parte inferior vemos, em letras maiúsculas, as bases que apresentam valor de qualidade acima de 15 e, em letras minúsculas, as que apresentam valor inferior. Tais bases minúsculas, apresentando informação (como visto no alinhamento com o padrão), são freqüentemente descartadas em projetos genoma.

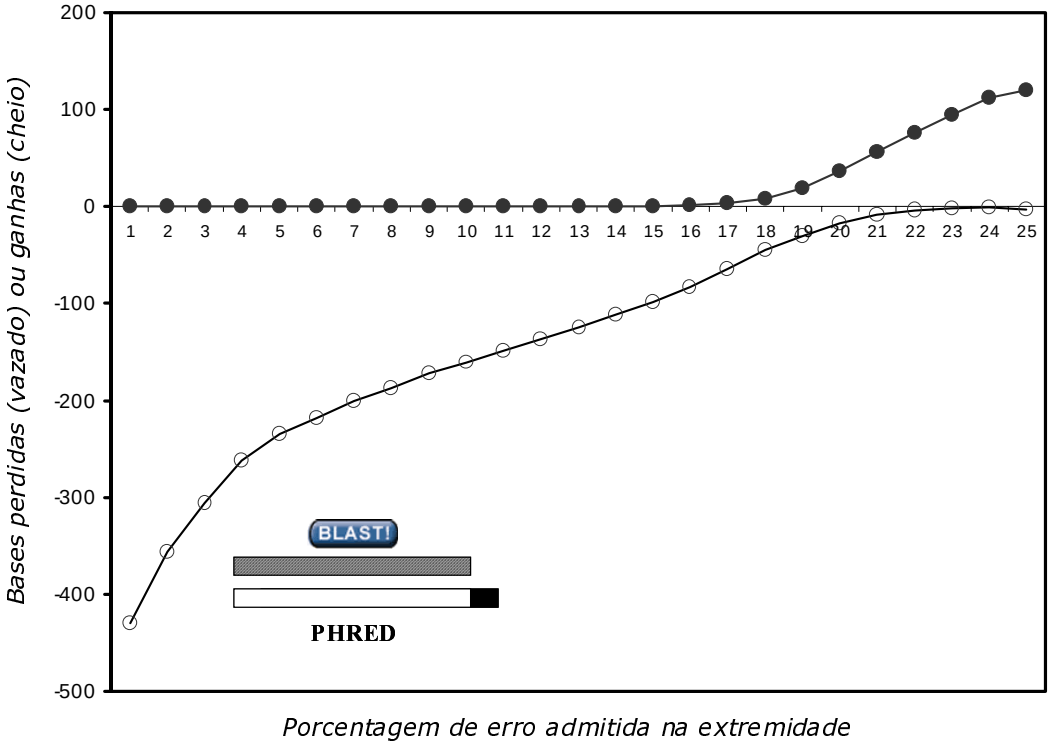


Figura 3: Cada ponto representa o número médio de bases descartadas (símbolos vazios) ou incorporadas (símbolos preenchidos) quando 846 seqüências foram produzidas e alinhadas com o programa BLAST com seu padrão (pUC18). Percebe-se que, utilizando-se valores de qualidade muito elevados (PHRED variando de 1 a 10%), muitas bases informativas são perdidas. Enquanto, utilizando valor muito baixo (maior que 20) bases não informativas são incorporadas. Dados similares foram obtidos com o programa de alinhamento SWAT (não mostrado).

Identificando as proteínas de interesse antes do projeto começar

Recentemente, bases de dados denominadas secundárias começaram a ser publicadas. A relevância destas bases de dados se deve ao fato de que, nestas bases, proteínas são organizadas de acordo com critérios evolutivos, categorias funcionais, ou possuem algum valor informativo agregado a elas. Além disso, o complemento gênico completo de organismos modelo está representado nessas bases, o que provê uma lista completa dos genes do organismo de interesse a serem obtidos, a fim de se fazer comparações evolutivas, moleculares, etc. Em outras palavras é possível, seguindo a listagem de genes nessas bases de dados secundárias, definir os genes a serem estudados antes mesmo do projeto se iniciar. Todavia, quando se concentra a busca utilizando-se apenas genes de organismos modelo, uma questão que surge é se o conjunto de toda a biota (seqüências de todos os organismos disponíveis no GenBank) não reconheceriam muito mais eficientemente as etiquetas de seqüências expressas (EST). A figura 4 mostra uma comparação feita com o software BLAST utilizando ESTs de *Schistosoma mansoni* para pesquisa de homologia contra seqüências protéicas de dois organismos modelo, *C. elegans* e *D. melanogaster*, tendo seus resultados de escore plotados na abscissa, enquanto na ordenada estão os resultados de escore pesquisando contra todo o restante da biota (ou do GenBank), com exceção dos dois organismos modelo escolhidos e também de seqüências de *S. mansoni*. A inclinação de 45 graus da quase totalidade da distribuição dos pontos indica que, quando um organismo modelo adequado possui um gene, dificilmente a limitação de pesquisa de homologia às suas seqüências atribui uma baixa qualidade ao resultado. Esta conclusão é muito importante, porque as bases secundárias são constituídas de seqüências de apenas alguns organismos modelo.

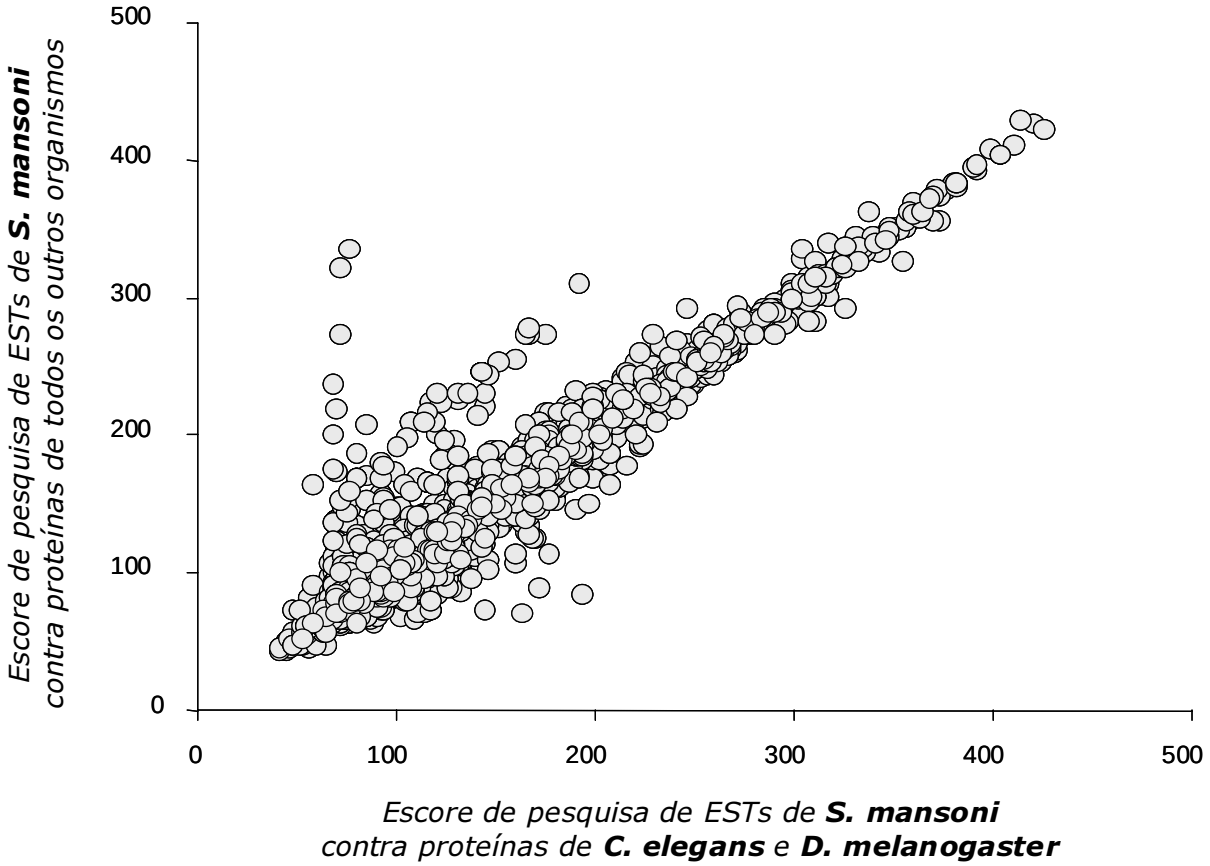


Figura 4: Similaridade de escore na pesquisa de homologia utilizando proteínas de *C. elegans* e *D. melanogaster* em comparação ao uso de proteínas de todos os outros organismos (com exceção de *C. elegans*, *D. melanogaster* ou do próprio *Schistosoma*), para pesquisar a homologia de ESTs de *Schistosoma mansoni*.

Saldo de códons positivo é um importante impulso para a iniciativa de caracterização completa da seqüência das proteínas

Nós cunhamos a expressão saldo de códons para expressar a quantidade de códons que uma EST possui em relação à proteína homóloga presente nas bases de dados secundárias. Como exemplificado na figura 5, quando o saldo é positivo, isto indica que a molécula a partir da qual a etiqueta foi gerada tem material codificador suficiente para conter inclusive o códon inicial da proteína (uma metionina). Nem sempre isso acontece, aliás, trata-se de um evento não muito freqüente. Raramente, na produção do cDNA (a molécula derivada do RNA mensageiro pela ação da transcriptase reversa), consegue-se uma cópia completa de toda a extensão da molécula (figura 1). Etiquetas geradas a partir de clones de cDNA onde a transcriptase reversa não completou a cópia fornecem etiquetas com saldo de códons negativo, ou seja, faltam códons em relação ao necessário para retratar toda a região codificadora. A figura 6 mostra um experimento onde, para a pesquisa, foram utilizadas etiquetas do organismo *C. elegans* contra proteínas do próprio organismo e calculou-se o saldo de códons real presente nas ESTs. Subseqüentemente foram utilizadas as mesmas ESTs de *C. elegans* para pesquisa de homologia com proteínas de outros organismos modelo – *D. melanogaster*, *H. sapiens* e *A. thaliana* – e obteve-se o saldo de códons dito calculado. Pode-se apreciar que, na maior parte das vezes, a diferença entre o saldo de códons real e o calculado não é maior que cinco códons. Portanto, a previsão da completude da informação é bastante confiável.

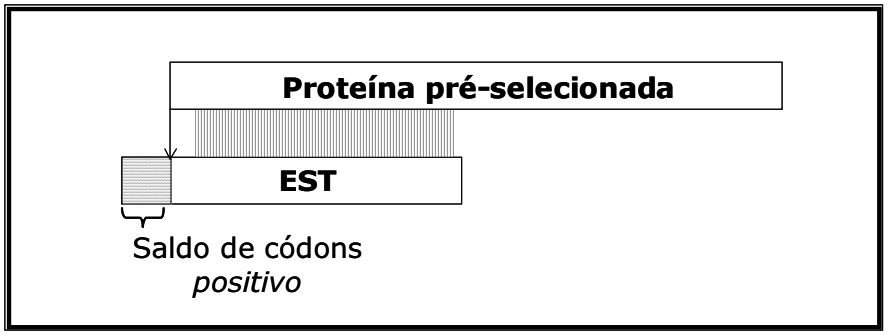


Figura 5: Esquema explicativo do método de cálculo do saldo de códons em ESTs. A ponta da flecha aponta a provável localização do códon da metionina.

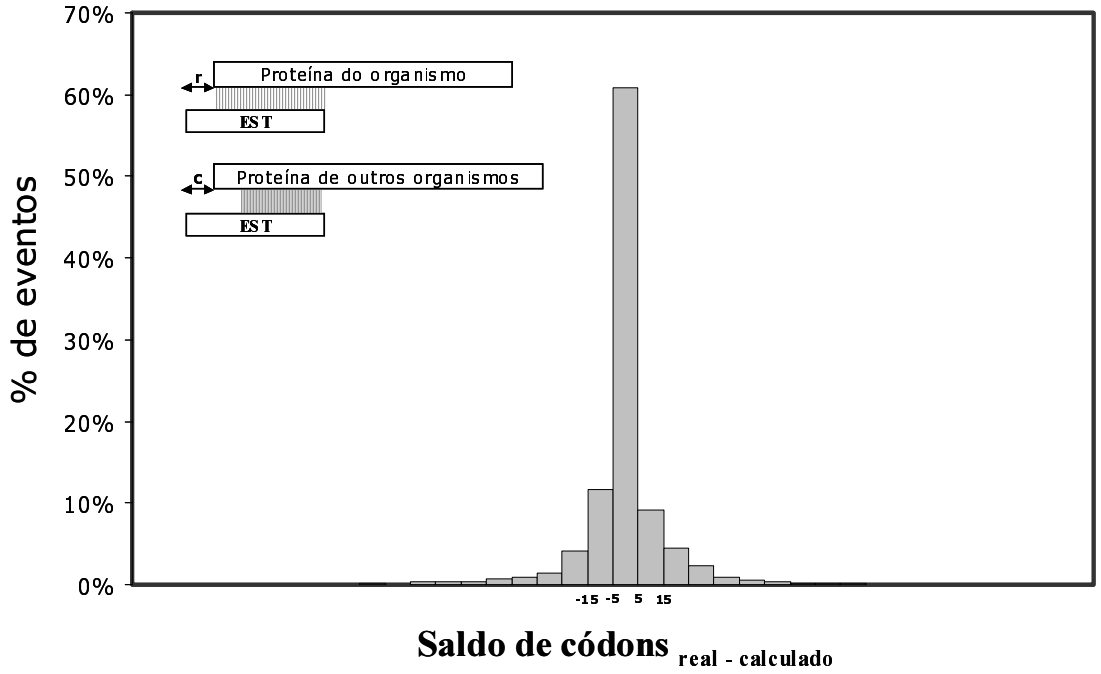


Figura 6: A diferença entre o saldo de códons real e o calculado usando o organismo *C. elegans* e proteínas KOG na maior parte dos eventos é inferior a cinco códons. O inserto mostra um esquema de como são medidos os saldos real e calculado.

Estimativa do erro na anotação de etiquetas com a base de dados secundária KOG

A base de dados KOG (TATUSOV et al., 2003; disponível em: http://www.ncbi.nlm.nih.gov/COG.) foi produzida no Nacional Center for Biotechnological Information (NCBI, ver: http://www.ncbi.nlm.nih.gov/) utilizando proteínas de organismos modelo, dentre os quais os metazoários *A. thaliana*, *C. elegans, D. melanogaster* e *H. sapiens*. Nesta base de dados, cada enzima ou outra proteína proveniente desses organismos é agrupada em entradas numeradas (por exemplo, as enolases desses organismos constituem o KOG2670). As seqüências desta base podem ser usadas com sucesso na anotação automática de ESTs como demonstrado no experimento esquematizado na figura 7. Neste experimento, utilizamos 215.200 ESTs de *C. elegans* para duas buscas de homologia: (i) inicialmente designamos cada EST a uma proteína de *C. elegans* pertencente a um dado KOG e, assim, obtivemos o resultado ideal da pesquisa; (ii) utilizamos as mesmas ESTs de *C. elegans* na busca contra as proteínas dos demais organismos da base de dados, não incluindo obviamente as proteínas do próprio organismo. Três tipos de resultados eram esperados: correto (o mesmo número KOG era obtido nas duas pesquisas), trocado (números KOG diferentes) ou

especulativo (a EST não havia sido designada a nenhuma proteína de *C. elegans*, todavia os demais organismos "sugeriram" uma possível classificação. Apesar de muita crítica já ter sido levantada contra a anotação automatizada de ESTs (KARTER et al., 2001), os resultados obtidos, notoriamente demonstram que o acerto é superior a 90%! Nós cunhamos a expressão "anotação reversa" para designar a utilização de proteínas de uma base de dados secundária na seleção de clones representados por ESTs que devem ser encaminhados ao processo de dedução da região codificadora completa. É importante ressaltar que a análise automatizada não prescinde de avaliação manual. Entretanto, o processo de caracterização de um transcriptoma é significativamente agilizado com a introdução da técnica proposta, a qual torna o processo de produção de seqüências de aminoácidos editadas mais eficiente por passar por uma análise automatizada e portanto mais rápida antes da avaliação manual.

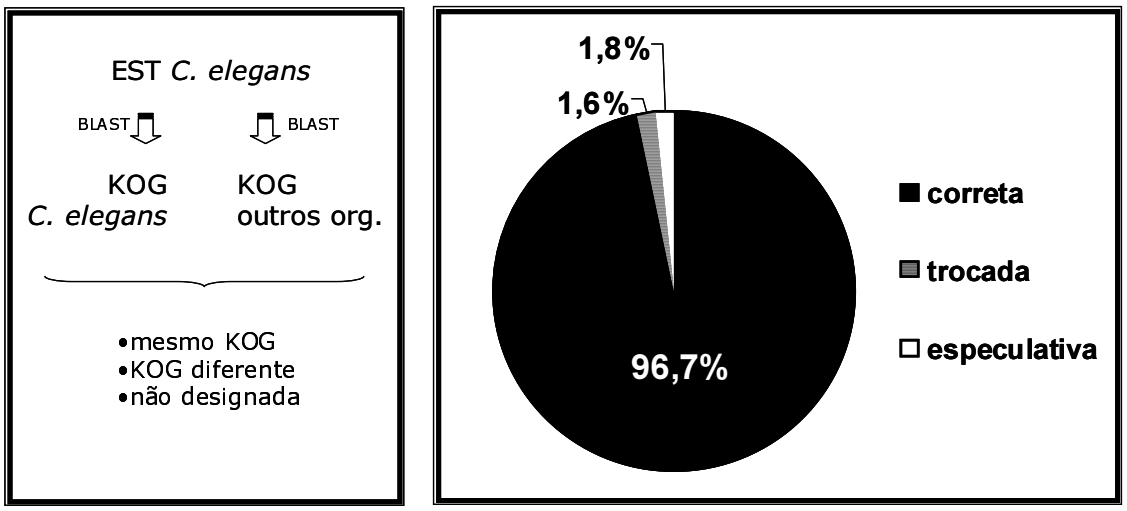


Figura 7: Eficiência de anotação usando proteínas KOG e seqüências EST de *C. elegans*. O quadro à esquerda mostra um diagrama do procedimento realizado e, à direita, o resultado (detalhes no texto).

CONCLUSÃO

Quando se produzem ESTs, vive-se a instigante experiência de gerar seqüências "pesquisa" (*query*), que podem ser comparadas com seqüências de aminoácidos presentes em bases de dados públicas e, assim, chegar-se à identificação dos genes constituintes do transcriptoma de um organismo de interesse. Através de técnicas de anotação reversa, é possível acrescentar aos projetos transcriptoma uma nova dimensão: a de geração de seqüências "alvo" (*subject*) que serão objeto de pesquisa para outros projetos. A geração de seqüências de aminoácidos, sejam estas completas ou, ao menos parciais das extremidades N e C terminais, é o que nos permite participar de bases de dados de domínios, de comparações evolutivas, dentre outras. O Laboratório de Biodados (UFMG) tem sua equipe intensamente dedicada a este tipo de iniciativa visando contribuir como grupo de pesquisa para as iniciativas de genômica funcional não somente desempenhando o papel de pesquisa, mas também de alvo da mesma.

ABSTRACT: The best way to characterize expressed sequences of a given genome is through transcriptome projects, which give clear evidence of regions harboring the information to produce specific proteins in an organism. EST (expressed sequence tags) production is one of the approaches used in this kind of project and is the main tendency of nacional science with respect to sequencing. EST production must be followed by annotation to convert sequence in information with biological meaning. In this work we report the use of bioinformatic tools in EST characterization associated to a powerful new strategy to choose ESTs representing specific genes for full-length sequencing. The chosen clones can be sequenced and characterized in detail allowing the complete aminoacid sequence to be determined. We report also studies on diferent PHRED values and the information content associated with the use of them. In addition to this we describe here tests that give an evaluation of the quality of the strategy used in the

clone selection. This strategy has been named "reverse annotation" and represents an important contribution of our research group to transcriptome projects, generating information that can be used by these projects

Uniterms: EST, BLAST, annotation

Referências Bibliográficas

ALTSCHUL S.F., MADDEN T.L., SCHAFFER A.A., ZHANG J., ZHANG Z., MILLER W., LIPMAN D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res.**, v.25, n.17, p.3389-402, set. 1997.

BRANDAO A., URMENYI T., RONDINELLI E., GONZALEZ A., DE MIRANDA A.B., DEGRAVE W. Identification of transcribed sequences (ESTs) in the Trypanosoma cruzi genome project. **Mem. Inst. Oswaldo Cruz**. v. 92, n. 6, p.863-6, nov.-dez. 1997.

EWING B., GREEN P., Base-calling of automated sequencer traces using phred. II. Error probabilities. **Genome Research**, v.8, p.186-194, 1998a.

EWING B., HILLIER L., WENDL M.C., GREEN P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Research**.v.8, p:175-185, 1998b.

FELIPE M.S., ANDRADE R.V., PETROFEZA S.S., MARANHAO A.Q., TORRES F.A., ALBUQUERQUE P., ARRAES F.B., ARRUDA M., AZEVEDO M.O., BAPTISTA A.J., BATAUS L.A., BORGES C.L., CAMPOS E.G., CRUZ M.R., DAHER B.S., DANTAS A., FERREIRA M.A., GHIL G.V., JESUINO R.S., KYAW C.M., LEITAO L., MARTINS C.R., MORAES L.M., NEVES E.O., NICOLA A.M., ALVES E.S., PARENTE J.A., PEREIRA M., POCAS-FONSECA M.J., RESENDE R., RIBEIRO B.M., SALDANHA R.R., SANTOS S.C., SILVA-PEREIRA I., SILVA M.A., SILVEIRA E., SIMOES I.C., SOARES R.B., SOUZA D.P., DE-SOUZA M.T., ANDRADE E.V., XAVIER M.A., VEIGA H.P., VENANCIO E.J., CARVALHO M.J., OLIVEIRA A.G., INOUE M.K., ALMEIDA N.F., WALTER M.E., SOARES C.M., BRIGIDO M.M. Transcriptome characterization of the dimorphic and pathogenic fungus Paracoccidioides brasiliensis by EST analysis. **Yeast.** v. 20, n.3, p.263-71, fev. 2003.

FRANCO G.R., SIMPSON A.J., PENA S.D. Sequencing and identification of expressed *Schistosoma mansoni* genes by random selection of cDNA clones from a directional library. **Mem Inst Oswaldo Cruz**.; v. 90, n. 2, p.215-6 , mar.-abr. 1995a.

FRANCO G.R., ADAMS M.D., SOARES M.B., SIMPSON A.J., VENTER J.C., PENA S.D. Identification of new *Schistosoma mansoni* genes by the EST strategy using a directional cDNA library. **Gene.** Vol. 152, n.2 , p.141-7 , Jan, 1995b.

http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

http://www.ncbi.nlm.nih.gov/COG

http://www.ncbi.nlm.nih.gov

KARTER K., OKA A., TAMIYA GEN, BELLGARD M.I. Bioinformatics Issues for Automating the Annotation of Genomic Sequences. **Genome Informatics.** v.12, p.204–211, 2001.

RAMALHO-ORTIGAO J.M., TEMPORAL P., DE OLIVEIRA S.M., BARBOSA A.F., VILELA M.L., RANGEL E.F., BRAZIL R.P., TRAUB-CSEKO Y.M. Characterization of constitutive and putative differentially expressed mRNAs by means of expressed sequence tags, differential display reverse transcriptase-PCR and randomly amplified polymorphic DNA-PCR from the sand fly vector *Lutzomyia longipalpis.* **Mem Inst Oswaldo Cruz**. , Vol. 96, n. 1, p.105-11, Jan. 2001.

SIMPSON A.J., REINACH F.C., ARRUDA P., ABREU F.A., ACENCIO M., ALVARENGA R., ALVES L.M., ARAYA J.E., BAIA G.S., BAPTISTA C.S., BARROS M.H., BONACCORSI E.D., BORDIN S., BOVE J.M., BRIONES M.R., BUENO M.R., CAMARGO A.A., CAMARGO L.E., CARRARO D.M., CARRER H., COLAUTO N.B., COLOMBO C., COSTA F.F., COSTA M.C., COSTA-NETO C.M., COUTINHO L.L., CRISTOFANI M., DIAS-NETO E., DOCENA C., EL-DORRY H., FACINCANI A.P., FERREIRA A.J., FERREIRA V.C., FERRO J.A., FRAGA J.S., FRANCA S.C., FRANCO M.C., FROHME M., FURLAN L.R., GARNIER M., GOLDMAN G.H., GOLDMAN M.H., GOMES S.L., GRUBER A., HO P.L., HOHEISEL J.D., JUNQUEIRA M.L., KEMPER E.L., KITAJIMA J.P., KRIEGER J.E., KURAMAE E.E., LAIGRET F., LAMBAIS M.R., LEITE L.C., LEMOS E.G., LEMOS M.V., LOPES S.A., LOPES C.R., MACHADO J.A., MACHADO M.A., MADEIRA A.M., MADEIRA H.M., MARINO C.L., MARQUES M.V., MARTINS E.A., MARTINS E.M., MATSUKUMA A.Y., MENCK C.F., MIRACCA E.C., MIYAKI C.Y., MONTERIRO-VITORELLO C.B., MOON D.H., NAGAI M.A., NASCIMENTO A.L., NETTO L.E., NHANI A JR., NOBREGA F.G., NUNES L.R., OLIVEIRA M.A., DE OLIVEIRA M.C., DE OLIVEIRA R.C., PALMIERI D.A., PARIS A., PEIXOTO B.R., PEREIRA G.A., PEREIRA H.A. JR, PESQUERO J.B., QUAGGIO R.B., ROBERTO P.G., RODRIGUES V., DE M ROSA A.J., DE ROSA V.E. JR., DE SA R.G., SANTELLI R.V., SAWASAKI H.E., DA SILVA A.C., DA SILVA A.M., DA SILVA F.R., DA SILVA W.A. JR., DA SILVEIRA J.F., SILVESTRI M.L., SIQUEIRA W.J., DE SOUZA A.A., DE SOUZA A.P., TERENZI M.F., TRUFFI D., TSAI S.M., TSUHAKO M.H., VALLADA H., VAN SLUYS M.A., VERJOVSKI-ALMEIDA S., VETTORE A.L., ZAGO M.A., ZATZ M., MEIDANIS J., SETUBAL J.C. The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis. **Nature**. v. 406, n. 6792 p.151-7, jul. 2000.

9. TATUSOV R.L., FEDOROVA N.D., JACKSON J.D., JACOBS A.R., KIRYUTIN B., KOONIN E.V., KRYLOV D.M., MAZUMDE R., MEKHEDOV S.L., NIKOLSKAYA A.N., RAO B.S., SMIRNOV S., SVERDLOV A.V., VASUDEVAN S., WOLF Y.I., YIN J.J. AND NATALE D.A. The COG database: an updated version includes eukaryotes. **BMC Bioinformatics**., v.4, n.1, p.41. set. 2003.

5.3 Encontrando um valor de corte de similaridade para alinhamentos aminoácido-nucleotídeo compatível com o valor de 96% para alinhamentos nucleotídeo-nucleotídeo

Esse tema foi tratado no artigo 2 e repetido com uma nova coleção de seqüências EST no artigo 3. O intuito era descobrir inicialmente um limiar de corte de similaridade que pudesse gerar um resultado de anotação mais acurado usando o software tBLASTn. O banco UniGene (Miller *et al.*, 1997) usa como valor de corte para similaridade em alinhamentos entre seqüências de nucleotídeos o valor de 96%, que leva em conta os erros intrínsecos existentes em etiquetas de seqüências expressas. Assim, seqüências com menos de 96% de identidade não são consideradas o mesmo gene. Segundo (Tatusov *et al.*, 1997), valores de corte de similaridade para alinhamentos entre seqüências de nucleotídeos e seqüências protéicas podem gerar perda de resultados positivos (perda em pesquisas de homologia para encontrar seqüências ortólogas). Porém, nada se sabia sobre o valor de corte de similaridade para alinhamentos entre seqüências de nucleotídeo e protéicas provindas de um mesmo gene. Haveria um efeito protetor para erros na terceira base dos códons, dada a redundância do código genético? Foram usadas inicialmente seqüências EST de um vetor de clonagem (pUC18) alinhadas com sua seqüência traduzida e sua seqüência de nucleotídeos, em um experimento controlado. Neste experimento os códons de terminação na seqüência traduzida foram codificado por "*", já que a região alinhada não se tratava de uma única ORF. Depois, os resultados foram estendidos para as seqüências de EST dos quatro organismos modelo, alinhadas com suas próprias seqüências protéicas KOG e suas respectivas seqüências de nucleotídeos (apenas os CDS). Neste caso foram escolhidas as entradas KOG que tinham as maiores taxas de expressão (maior ocorrência de EST alinhadas) e apenas entradas KOG de um único gene foram utilizadas, evitando-se possíveis alinhamentos com parálogos. Verificou-se que a similaridade média entre alinhamentos proteína-nucleotídeo é sempre mais baixa do que entre alinhamentos nucleotídeo-nucleotídeo, de modo que erros nas primeiras duas bases dos códons parecem gerar um efeito mais deletério nos alinhamentos do que o efeito protetor da terceira base.

A estratégia usada para encontrar o valor de corte de identidade para alinhamentos aminoácido-nucleotídeo foi a de inferir esse valor usando o mesmo número de alinhamentos descartados com o corte de 96% de similaridade entre seqüências

nucleotídeo-nucleotídeo. Os valores de corte encontrados nos artigos 2 e 3 se encontram aproximadamente na mesma faixa, entre 70% e 80%, para todos os organismos. Como se pode observar visualmente no artigo 2 e mais criteriosamente no artigo 3, o uso dos valores de corte para similaridade usando tBLASTn encontrados, resulta em um aumento da precisão, ou PPV, sem comprometer a parcela correta, ou de verdadeiros positivos da anotação. O uso de um valor de cutoff na designação de uma EST à sua proteína correspondente evitou que EST contendo domínio conservado fosse designada incorretamente, todavia isto não a impede de ser anotada pelos outros organismos. Esta categoria, definida por anotação de uma EST não designada, foi denominada nos artigos 2 e 3 como "especulação". Em conclusão, a anotação automática com a base KOG não é absolutamente sujeita a erros como se poderia sugerir antes da avaliação de sua performance, como discutido a seguir.

5.4 Teste da eficiência de anotação de seqüências EST e a base KOG

A base de dados KOG provê uma vantagem por ser construída usando relações de ortologia entre proteínas. Dessa forma, agrupamentos KOG podem ser constituídos por proteínas de vários organismos (pelo menos três) e ainda assim representar uma mesma proteína ancestral com provavelmente uma mesma função. A função é apenas provável de ser a mesma devido ao agrupamento de *inparalogs e outparalogs* em agrupamentos C/KOG (revisto por Koonin, 2005). Essa relação de ortologia tornou possível realizar um teste de anotação. Inicialmente foi preciso "designar" as EST dos organismos presentes na própria base a suas próprias proteínas como controle. Em um próximo passo, removem-se as proteínas do próprio organismo da base e a "anotação" é feita com as proteínas dos outros organismos. Dessa forma compara-se a designação de uma EST com as proteínas KOG do organismo de onde foram originadas, com a anotação dessas EST com as proteínas KOG dos outros organismos da base. Essa comparação é feita verificando-se o agrupamento KOG designado e anotado. Esse teste tem as funções de (i) verificar a eficiência de um experimento de anotação de EST com proteínas KOG (testar a própria base KOG) e (ii) verificar se os valores de corte de identidade estavam adequados, pois podem ser utilizados em outras aplicações. Além disso, esse teste é expansível para bases semelhantes à KOG,

formadas com a mesma relação de ortologia como KO/KEGG, PIRSF, Ortho-MCL-DB, InParanoid ou UniRef (Kanehisa *et al.*, 2002; Wu *et al.*, 2004; Suzek *et al.*, 2007b).

Como visto nos artigos 2 e 3, a anotação com a base KOG foi eficiente (acima de 90% correta) para as coleções de EST de todos os quatro organismos (ver material suplementar do artigo 3 em http://biodados.icb.ufmg.br/KOG_paper_2). Além disso, os valores de corte para designação de EST se mostraram eficientes e aumentaram a precisão do teste de anotação (ver artigo 3).

O erro na anotação foi avaliado mais criteriosamente no artigo 3, onde as anotações trocadas foram separadas por classes. Verificou-se que uma parcela da anotação trocada é devida a agrupamentos KOG com características muito semelhantes (classes *String* e *Family*). Foram criadas duas ferramentas *web* (*EST search tool* e *KOG annotation search tool*) que disponibilizam informações sobre toda a anotação das EST com a base KOG. Essas ferramentas permitem busca usando os valores de corte utilizados, visualização de domínios conservados CDD (Marchler-Bauer *et al.*, 2002) pré-anotados em proteínas KOG, informações sobre EST, entre outras facilidades. Com a ferramenta *KOG annotation search tool* tornou possível verificar visualmente um dos motivos das trocas de anotação (do tipo *Family*): o alinhamento de seqüências EST em regiões de domínios conservados presentes em ambas as proteínas KOG usadas na designação e anotação (ver figura 11). Esse tipo de trocas sugere que seria interessante reunir agrupamentos KOG semelhantes, o que iria resultar em anotação coincidente com a designação. Optamos por computar como uma troca a anotação para um KOG de mesma *string* da designação, pois a separação em duas entradas KOG foi um critério da base de dados. Todavia, como a troca seria corrigida, propomos aqui um critério para agrupamento de entradas de grupos de homólogos que foram separados.

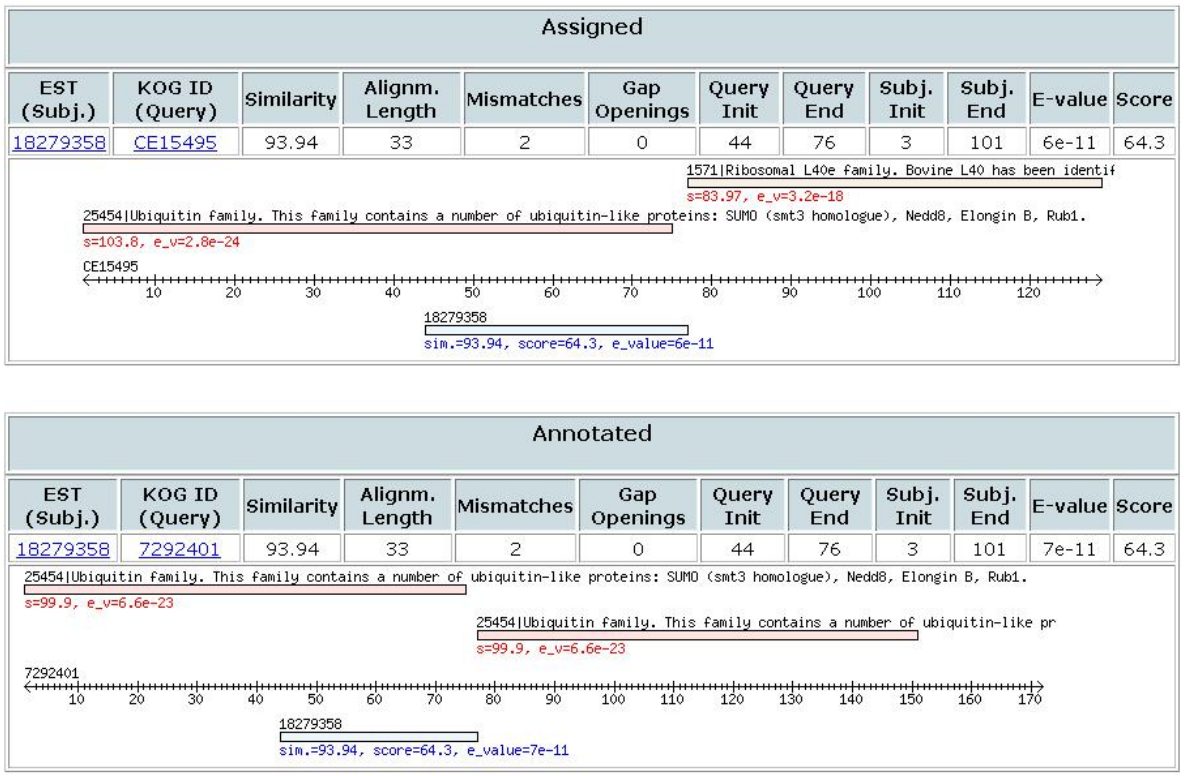


Figura 11 – Exemplo de anotação trocada de uma EST de *C. elegans* (gi 18279358) com proteínas KOG (designada para CE15495 – KOG0003 e anotada para 7292401 – KOG0001). A EST está colorida em azul embaixo da proteína KOG (flecha preta, graduada, ao centro) e domínios CDD previamente anotados à proteína KOG estão coloridos de tons vermelhos acima. Notar a presença de um mesmo domínio conservado (*Ubiquitin familiy* – pfam 25454) em ambas as proteínas, na mesma região onde a EST alinhou.

Tests of Automatic Annotation Using KOG Proteins and ESTs from 4 Eukariotic Organisms

Maurício de Alvarenga Mudado, Estevam Bravo-Neto, and José Miguel Ortega

Universidade Federal de Minas Gerais, Avenida Antonio Carlos, 6627,
Belo Horizonte MG, Postal Code 486, Brazil
miguel@icb.ufmg.br
http://www.biodados.icb.ufmg.br

Abstract. BLAST homology searches have been largely used to annotate function to novel sequences. Secondary databases like KOG can be used in this intention since their sequences have functional classification. We devised an experiment where public ESTs from four eukariotic organisms, which protein sequences are present in the KOG database, are classified to functional KOG categories using tBLASTn. First we assigned the ESTs from one organism to KTL (KOG, TWOG and LSEs) proteins and then we searched the database depleted of the same organism's proteins to simulate a novel transcriptome. Data show that classification was correct (assignment equals annotation) 87.2%, 96.8%, 92.0%, 88.7% for *A. thaliana*(Ath), *C. elegans*(Cel), *D. melanogaster*(Dme) and *H. sapiens*(Hsa) respectively. We have estimated identity cutoffs for all organisms to use with tBLASTn. These cutoffs trim the same amount of events that a BLASTn in order to minimize false positives in consequence of sequence errors. We found values of 80%, 78%, 78% and 84% for amino-acid identity cutoff for Hsa, Dme, Cel and Ath, respectively. We then evaluated our system by comparing the KTL categories of the assigned ESTs with the KTL categories that the ESTs were classified without the organism's KTL proteins. Moreover, we show the potential of annotation of the KOG database and the ESTs used. Suplementary Information can be found at: http://www.biodados.icb.ufmg.br

1 Introduction

Homology searches have been largely used to annotate the putative function of novel described sequences, either nucleotides or aminoacids. Usually software from the BLAST package [8] is used in this type of search [2] and the best hit (higher bit score) associated with a cutoff requirement of low E value is sufficient to establish a relationship of homology between query and subject [12][13]. Quality of annotation remarkably depends on the quality of the database that is being used as subject in homology searches. Secondary databases are currently available where sequences are not only deposited, but classified into functional categories. These databases are being widely used in the categorization of ESTs [4][18][5]. One of these databases is KOG [17], at NCBI, which organizes protein

J.C. Setubal and S. Verjovski-Almeida (Eds.): BSB 2005, LNBI 3594, pp. 141–152, 2005.

entries from seven organisms with complete sequenced genome into three classes of occurrence: KOG, TWOG and LSE, which occur in three, two or only one organism, respectively. Each protein has received a KOG ID - e.g. enolase is KOG0047. All orthologs, and eventually occurring paralogs, are classified under the same ID, and there are IDs for the three classes of KOG. Thus, this database is an attractive subject for testing automated annotation procedures. ESTs and transcriptome projects have showed its importance not only for gene discovery [1] but also for analysis of differential expression of genes [14][6][16]. ESTs are known to bear up to 4% of sequencing errors due to its single-pass nature. Development of automated annotation for ESTs is already being issued [3][15]. Annotation is mostly solved with the use of best hit to the subject database, but the identity of a nucleotide sequence, that contains errors, to an aminoacid sequence of the proper organism have not been addressed yet. It has been largely accepted (e.g. UniGene database - Lukas Wagner, personal communication,[19]) that 96% identity at nucleotide level is sufficient to assign an EST to the correspondent nucleotide cDNA sequence. Errors occurring in the third base of the codons tend to be silent in either tBLASTn or BLASTx searches. However, errors in the first two bases of the codon are expected to be hazardous to the alignment. In this work we set up to define a cutoff in BLASTx / tBLASTn searches that would be equivalent to 96% identity cutoff in nucleotide to nucleotide comparisons (BLASTn).Then we devised an experiment where we initially assigned ESTs to proteins from the KOG database of the proper organism and further annotated the ESTs with the entire KOG database lacking the proteins from the organism whose ESTs were used to query the database. This procedure simulates the annotation of a novel transcriptome. Furthermore, we evaluated our procedure verifying if the annotation was either correct, resulting in the same database ID, changed to a different one or even speculative (ESTs not assigned to any organism's protein but annotated by other organism's sequences).

2 Material and Methods

2.1 Vector Sequences

The pUC18 sequences used in this work have been provided by 3 laboratories from *Universidade Federal de Minas Gerais* (UFMG) that integrate the network *Rede Genoma de Minas Gerais*. The reactions were made in a single pool and divided into tubes for the PCR sequencing reaction. After the reaction, the sequences were joint again in the same tube, mixed, and then divided into 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 were obtained.

2.2 Other Sequences

The EST sequences were downloaded from dbEST database [9] at Mai/2003. All KTL proteins and KOG conserved domains were downloaded from NCBI

homepage [7] from the "kyva" file. We then selected the 88,613 classified KTL proteins found at the "kog, "twog" and "lse" files at the same address, to use in the BLAST searches. The KTL proteins are divided in 60,758 KOG, 4,451 TWOG and 23,404 LSE proteins. To retrieve the 50 CDS relative to the 50 KOG proteins from the four organisms, we selected the 100 more expressed KOG proteins that were hit by the ESTs from the four organisms (data not shown). We chose the 50 ones that have only one representing ortholog protein, to avoid ESTs being aligned to paralogs. We downloaded the respective mRNA sequences from these 50 KOG proteins (NCBI provides a list of proteins from KOG database assigned to their relative mRNAs - called "kyva=gb"). We then selected only the CDS of these mRNAs and removed the stop codons, by parsing the genbank file with a PERL script, to assure the proportion of identity between the alignments of ESTs to its proteins/nucleotides.

2.3 Data Processing

All data were processed using MySQL version 3.23.58 and scripts wrote in PERL language, version 5.8.0. The BLAST software package version 2.2.8 was obtained from NCBI. PHRED software version 0.020425.c was obtained (see [10]), thanks to Phill Green. All processing was made on a Linux Red Hat 9 machine, Pentium IV HT, 2.4 GHz and 1 GB RAM. The BLAST searches were run additionally on four other machines with similar power of processing and same operational system.

2.4 BLASTs

The tBLASTn/BLASTn were run with the following parameters: -m 8 -b 10e6 -e 1e-10 -F f . These parameters activate the tabular output of BLAST, allows up to 10 million hits to one protein (the default is 250) and deactivates the low complexity filter, respectively. The low-complexity filter was deactivated in order to permit tBLASTn to achieve 100% identity in the alignments.

2.5 PHRED

The software PHRED was run with the following parameters: -trim_alt "" -st fasta -trim_cutoff <n> Which activates the trimming algorithm selects the file type and activates the trimming with error cutoff (n) respectively.

2.6 Statistics

When necessary data were reported as means ± SEM (standard error of the mean).

3 Results

3.1 Defining Cutoffs

To define a cutoff for tBLASTn that is equivalent to 96% for BLASTn, we took advantage of 846 sequence reads of pUC18 (see material and methods)

and aligned these sequences with either BLASTn to the published nucleotide sequence (genbank access number L09136) or with tBLASTn to a single frame translation starting at the first nucleotide downstream to the primer. We solved the problem of alignments to stop codons by representing the respective positions with the "*" character, what leads to 100% identity to tBLASTn alignments (not shown). Reads were trimmed with PHRED basecalling software under increasing error acceptance, using trim_alt PHRED internal algorithm. (E.g. 1% of error corresponds to PHRED 20 , 10% to PHRED 10). Data presented in figure 1A show that, for all error densities used, alignments of single-pass pUC18 reads (here simulating controlled ESTs) to the nucleotide sequence result, in average, to more than 96% identity , while alignments to the aminoacid sequences yielded lower levels of identity.

In order to investigate the behavior of the actual cDNA sequences we downloaded large sets of ESTs (Table 1) from the four organisms present in the KOG database (ath: *A. thaliana*; Cel : *C. elegans*; Dme: *D. melanogaster*; *hsa: H. sapiens*) from dbEST. We then selected 50 KOG proteins from each organism requiring that they corresponded to the most occurring ESTs and did not have paralogs, thus hits should probably point to a single protein. Data in figure 1B show that ESTs, aligned with tBLASTn to the amino-acid sequences, consistently show average levels of identity lower than the correspondent complete CDS nucleotide sequences. Moreover, the identities observed for each EST collection seem to represent the error density characteristic of the collection, as judged by comparison with the data presented in figure 1A.

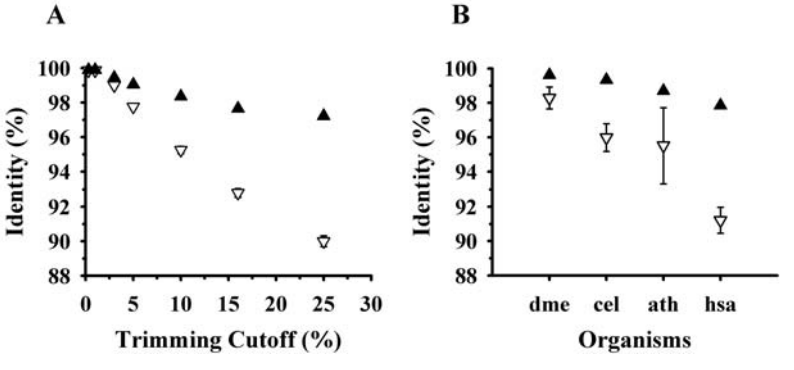


Fig. 1. A. Plot of mean identity ± mean standard error obtained from tBLASTn and BLASTn experiments with PUC18 using variable PHRED trim cutoff parameters. B. Plot of average identity and mean standard error obtained from tBLASTn and BLASTn experiments with a set of ESTs and 50 nucleotide/aminoacid sequences from the four organisms. Inverted open triangles are tBLASTn experiments and straight full triangles are BLASTn experiments. The n for pUC18 ranged from 846 reads with 25% of error to 673 reads with 0.3% of error. The n for Dme, Cel, Ath and Hsa was 23,630, 14,096, 1,891 and 14,484 sequences respectively

Table 1. Organisms and the respective ESTs, KOGs and proteins used in this work

Organisms	ESTs	KOGs	Proteins
Arabidopsis thaliana	178,538	4,872	24,154
Caenorhabditis elegans	215,200	5,306	17,101
Drosophila melanogaster	261,404	5,145	10,517
Homo sapiens	1,941,556	6,572	26,324
pUC18*	846 **	-	1 ***

* pUC18 stands for the commercial vector (see GenBank accession number L09136). ** pUC18 reads obtained by sequencing. *** The nucleotide sequence of pUC18 was translated into 1 protein sequence.

To estimate a cutoff for tBLASTn that would correspond to the 96% cutoff for BLASTn and to find the amount of events that these two cutoffs represent, we performed BLASTn and tBLASTn alignments using the pUC18 reads and its respective nucleotide/aminoacid sequences (Figure 2A). Reads were processed with 16% trim_alt cutoff (this procedure yields a maximum score plateau when aligning pUC18 reads to its nucleotide/amino-acid sequences - data not shown). Figure 2B shows the same tuples as in figure 2A but grouped by number of events, so it is possible to conclude that 96% of identity cutoff, when aligning nucleotides, corresponds to 93% of the totality of tuples. Therefore, the identity cutoff value that retrieves the same amount of tuples when using aminoacid

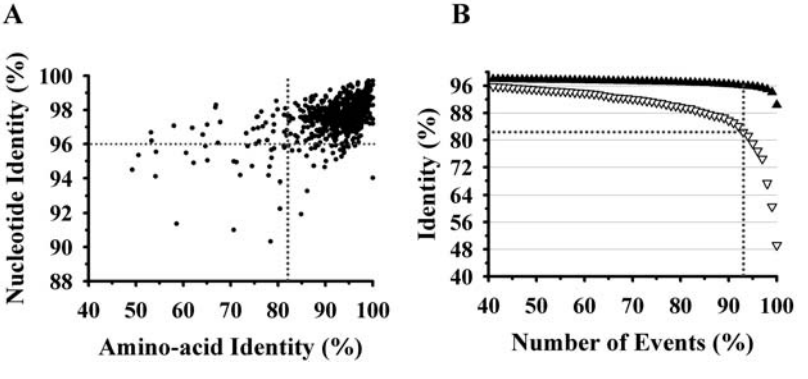


Fig. 2. A. Plot of tBLASTn - BLASTn tuples result of BLASTs performed with the translated and the nucleotide sequence of pUC18 with the reads obtained by automatic sequencing and PHRED 8. The dotted lines limit 96% and 82% of identity cutoff for BLASTn and tBLASTn. B. Same plot organized by number of events. The open inverted triangles represent tBLASTn plots and the full straight triangles BLASTn. The horizontal dotted line shows the cutoff for tBLASTn when 96% of identity for BLASTn is the reference. The vertical dotted line shows the ammount of events that the two cutoffs are representing

sequences as a target is 82.3% (depicted by the dotted lines in figure 2A). This same procedure was performed with the 4 organisms sequences and we found values of 80%, 78%, 78% and 84% for amino-acid identity cutoff for Hsa, Dme, Cel and Ath, respectively (data not shown). The mean cutoff value found for the 4 organisms (80% ± 2.8) is very close to the one found for pUC18 (82.3%). The percentage of events that these values collect ranged from 84% to 99%, suggesting that these cutoffs are discarding a minority of correct events when using tBLASTn.

3.2 Simulating Novel Transcriptomes

To test if these cutoff values are adequate, and to verify the accuracy of an automatic annotation experiment using ESTs and KTL proteins with tBLASTn, we conducted the pipeline as explained in figure 3.A. First, all ESTs from one organism (eg. Dme) are searched against the KTL proteins from the same organism. The best matches from this experiment are therefore assigning Dme ESTs to KTL proteins. Second, all ESTs from Dme are searched against the KOG database, but lacking Dme proteins, simulating in this way an annotation of a novel transcriptome. The best matches from this second experiment can be classified into 3 groups: correct annotation, when an EST from the second experiment is assigned to the same KOG ID as in the first experiment; speculative annotation, when an EST has not been assigned to a KOG ID in the first experiment, but it found a hit to a KOG ID in the second experiment; changed annotation, when an EST points to a KOG ID in the second experiment that is different from the KOG ID it was assigned to in the first experiment. The first experiment was conducted using the respective cutoff value for each of the four organisms and the accuracy of the annotation measured with the second experiments. There is two further possibilities of missing annotations (figure 3B), where ESTs are assigned but miss annotations in the simulation of a novel transcriptome (assigned but no hit), and where ESTs have no hit at all in neither experiments (no hit). When analyzing the totality of annotated ESTs, this methodology is able to correctly process around 90% (87.2%, 96.8%, 92.0%, 88.7% for *A. thaliana*, *C. elegans*, *D. melanogaster* and *H. sapiens* respectively), using the cutoffs determined. The percentage of changed annotation remains very small for all organisms, never overscoring 5%. The speculative annotation is more expressive in Hsa and Ath, but with values below 10% (data not shown).

We tested if annotability is altered by using different identity cutoffs when assigning ESTs to KOG IDs. We performed rounds of annotation, starting from 45% up to 100% of identity cutoff. Figure 4 shows the percentage of ESTs that are found in the 3 categories, when using these cutoffs and the 4 organisms sequences. Together these 3 categories and all assigned ESTs forms the group of ESTs that are potentially annotable. We found that, in most cases, the use of low cutoff values augments the group of correct annotation relative to the other groups. Moreover, the changed annotation stays at low values (below 2% of the ESTs in most cases). Changed annotation slightly diminishes when the cutoff is raised, probably because fewer errors are permitted in the alignment.

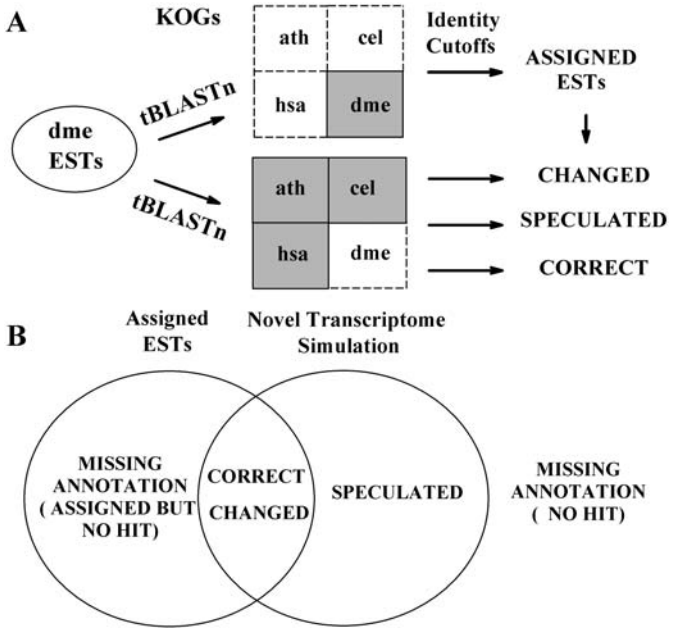


Fig. 3. Schema of the experiment devised to test the annotation of the ESTs from the four organisms with the KOG database. A. The experiment is made using ESTs and KTLs (KOGs, TWOGs and LSEs) from Dme as an example. It is divided in two steps, first a tBLASTn is made against KTL proteins only from Dme (grey square), to assign ESTs to KTL classes. The next step is a tBLASTn of ESTs from Dme against KTL proteins from the other organisms (the tree grey squares) but not from Dme, then simulating a novel transcriptome. The classification of the annotation is obtained by comparing the classes of KTL that the ESTs were assigned in the first and second experiments. Three classes are possible: changed, speculated and correct annotations. B. The products obtained from A. 5 classes are possible: correct, changed and speculated and 2 classes of missing annotations: the "no hit" and "assigned but no hit" ESTs

On the contrary, when raising the cutoff values above 80%, the percentage of speculative annotation raises because less ESTs are being assigned to a KOG ID in the first experiment (less alignments pass this filter). This can be assumed by the diminishment of the correct annotation class in the same proportion of the increase of the speculative class. We found that, for Dme and cutoff values below 90%, annotation of almost 50% of the total ESTs is to correct KOG IDs. This is followed by Cel annotation, with 40% of total ESTs. Distinctly, Ath and Hsa annotation might be classified as poorer, with only 20% and 13% of all ESTs being correctly annotated.

In order to show the potential of annotation that is gained or lost by using different identity cutoffs, we plotted in figure 5 the amount of ESTs that are potentially annotable: correct, changed, speculated and assigned but with miss-

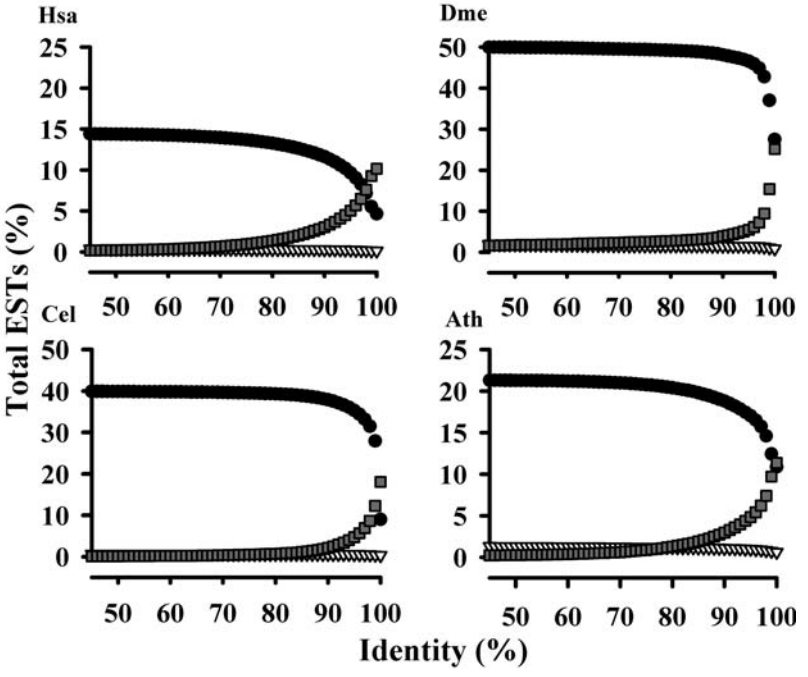


Fig. 4. Testing the annotation with KOG using different identity cutoffs. Full circles represent the correct annotation. Gray squares are representing speculations of a right KOG and inverted open triangles represent changed annotations. Hsa: *Homo sapiens*. Cel: *C. elegans*. Dme: *D. melanogaster*. Ath: *A. thaliana*

ing annotation when simulating new transcriptomes (the black + grey areas). The definition of potentially annotable is used because these ESTs have had a hit to a KOG protein in any of the experiments. The amount of ESTs that this methodology was unable to classify (no hit to any database when assigning and when simulating novel transcriptomes), can be observed by calculating the complementary area of the black + grey areas. The black area represent the correct, speculated and changed annotations. In other words, the ESTs that had been annotated by any KOG protein in the second experiment. The grey areas represent the ESTs that had been assigned to a KOG ID but had no annotation in the second experiment. This can be caused by assignments to unique proteins of the organism (dark grey areas) in the first experiment. Using lower cutoff values, Dme ESTs have the best potential of annotability with around 77% of all ESTs being annotable and less than 23% unable to be classified. It is followed by Cel (75% and 25%) and Ath with a good anotability (around 80%) but with a poor potential of classification by the system (57% loss). This loss can be explained by a large amount of assignments to LSE proteins (discussed below). *H. Sapiens* is a special case, with a very low efficiency of around 35% and with almost 20% of its ESTs unable to be classified by our system.

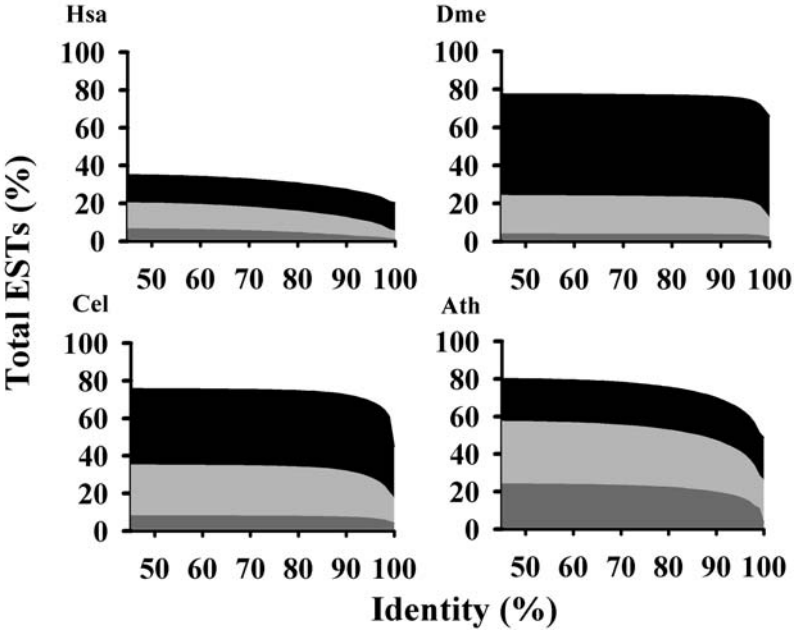


Fig. 5. Test of annotation with KOG using different identity cutoffs showing the porcentage of ESTs that are annotable and the missing assigned ESTs. The black + grey areas show the total percentage of annotable ESTs. The black areas are representing the correct, speculated and changed annotations. The light grey areas are representing the assigned ESTs to KOG proteins but with no hit when simulating a new organism transcriptome. The dark grey areas are representing the amount of ESTs from the light grey areas that were assigned to LSE proteins. Hsa: *Homo sapiens*. Cel : *C. elegans*. Dme: *D. melanogaster*. Ath: *A. thaliana*

3.3 Searching LSEs

When simulating a novel transcriptome, a low potential of annotability and wrong KOG ID classification can be explained by ESTs being assigned to LSE proteins, genes that are present only in the organism being annotated. In this case, annotation can result in two cases: a no hit to any KOG, TWOG or LSE from other organisms or an undesired changed annotation. We did a survey into all assigned EST sequences using the organism's cutoffs, to test if assignment to LSE proteins were biasing these results. We found that Ath, as expected for being the only plant in the database, has the higher number of ESTs assigned to LSE proteins (41% of the total 129,100 EST sequences assigned), followed by Hsa (16% of 574,091), Cel (11% of 160,065) and Dme (6% of 194,838). Furthermore we surveyed all set of changed annotation and missed assigned ESTs to obtain the percentages of these groups that were initially assigned to LSE proteins. We found that, from the group of changed annotation, Hsa and Ath have 30.8% and 35.5% of sequences assigned initially to LSE proteins respectively. Cel and

Dme have a smaller amount with 10.3% and 16.6% respectively. From the set of sequences that were assigned but are missing an annotation, Hsa and Ath have again the greater percentage with 29.0% and 42.4% belonging to LSE proteins respectively. Cel has 23.2% followed by Dme with only 7.0%. Altough in some cases, like Hsa and Ath that have around 30% of all changed annotation caused by EST sequences assigned to LSE proteins, these numbers are representing a very small portion of the total EST sequences used. Thus, the most part of sequences that were initially assigned to LSE proteins are not being classified as a changed annotation. This result indicates that ESTs assigned to LSE proteins are not causing a strong bias on changed annotation. In most cases, less than 3% of all EST sequences assigned to LSE proteins are contributing to this group. Except by Ath, the group of EST sequences that had been assigned but is missing an annotation is not greatly increased by assignments to LSE proteins. The plant shows a significant ammount of LSE proteins being assigned but are missing an annotation.

4 Discussion and Conclusion

Assuming a cutoff value for identity when using tBLASTn is necessary since the big volume of data is diminished, requiring less computational effort and storage space.

Our system was able to correctly classify around 90% of all annotable ESTs which passed the first 10^{-10} BLAST E-value cutoff.

The identity cutoff values found for the 4 organisms are therefore suitable as changed annotation is almost not altered for all cutoffs and always have low values, representing less than 5% of all ESTs. Our results also show that the use of high identity cutoffs can be harmful to an automatic annotation procedure. This is depicted by the raise of speculative annotation and diminishment of correct annotation, when using high identity cutoffs (above 80%), in figure 4.

Hsa lacks a good potential of annotability and we speculate two possible explanations. First, probable low quality EST sequences in the database: sequences with low lengths and high error rates (see Fig.2B). We are currently investigating this fenomena.

The second explanation is that KOG is not yet a complete database to annotate Hsa and it may lack more than 60% of the necessary proteins to annotate a human transcriptome. To explain this, we'll perform a future annotation experiment with larger secondary databases like Uniprot [11] or the NCBI's nr database.

However, all annotable Hsa ESTs showed a high rate of correctness (above 87%). Furthermore, Dme, Cel and Ath showed a better automatic annotation potential with around 80% of annotability.

We conclude that KOG is a reliable database for EST annotation depicted by the results obtained with the four organisms studied. Suplementary information was made available with APACHE/PHP and can be found at http:// www.biodados.icb.ufmg.br .

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, R.F. et al.: Complementary DNA sequencing: expressed sequence tags and human genome project. Science **252** (1991) 1651–1656
2. Altschul, S.F., Madden, T.L., Schaffer, AMINO-ACID, Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25** (1997) 3389–3402
3. Cuff J.A., Birney E., Clamp M.E., Barton G.J.: ProtEST: protein multiple sequence alignments from expressed sequence tags. Bioinformatics. **16(2)** (1999) 111–6
4. Faria-Campos A.C., Cerqueira G.C., Anacleto C., Carvalho C.M.B., Ortega J.M.: Mining microorganism EST databases in the quest for new proteins. Genet. Mol. Res. **2(1)** (2003) 169–177
5. Felipe M.S., Andrade R.V., Petrofeza S.S., Maranhao A.Q., Torres F.A., Albuquerque P., Arraes F.B., Arruda M., Azevedo M.O., Baptista A.J., Bataus L.A., Borges C.L., Campos E.G., Cruz M.R., Daher B.S., Dantas A., Ferreira M.A., Ghil G.V., Jesuino R.S., Kyaw C.M., Leitao L., Martins C.R., Moraes L.M., Neves E.O., Nicola A.M., Alves E.S., Parente J.A., Pereira M., Pocas-Fonseca M.J., Resende R., Ribeiro B.M., Saldanha R.R., Santos S.C., Silva-Pereira I., Silva M.A., Silveira E., Simoes I.C., Soares R.B., Souza D.P., De-Souza M.T., Andrade E.V., Xavier M.A., Veiga H.P., Venancio E.J., Carvalho M.J., Oliveira A.G., Inoue M.K., Almeida N.F., Walter M.E., Soares C.M., Brigido M.M.: Transcriptome characterization of the dimorphic and pathogenic fungus Paracoccidioides brasiliensis by EST analysis. Yeast. **20(3)** (2003) 263–71
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., Pena S.D.: Evaluation of cDNA libraries from different developmental stages of Schistosoma mansoni for production of expressed sequence tags (ESTs). DNA Res. **4(3)** (1997) 231–40
7. ftp://ftp.ncbi.nih.gov/pub/COG/KOG/
8. http://www.ncbi.nlm.nih.gov/BLAST/
9. http://www.ncbi.nlm.nih.gov/dbEST
10. http://www.phrap.org
11. http://www.uniprot.org
12. Koonin E.V., Fedorova N.D., Jackson J.D., Jacobs A.R., Krylov D.M., Makarova K.S., Mazumder R., Mekhedov S.L., Nikolskaya A.N., Rao B.S., Rogozin I.B., Smirnov S., Sorokin A.V., Sverdlov A.V., Vasudevan S., Wolf Y.I., Yin J.J., Natale D.A.: A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. **5(2)** (2004) R7
13. Koonin, Eugene V. and Galperin, Michael Y.: Sequence - Evolution - Function Computational Approaches in Comparative Genomics. Norwell (MA) 2003
14. Lee N.H., Weinstock K.G., Kirkness E.F., Earle-Hughes J.A., Fuldner R.A., Marmaros S., Glodek A., Gocayne J.D., Adams M.D., Kerlavage A.R., et al.: Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 Cel ls before and after nerve growth factor treatment. Proc Natl Acad Sci. **92(18)** (1995) 8303–7
15. McCallum J, Ganesh S.: Text mining of DNA sequence homology searches. Appl Bioinformatics. **2(3 Suppl)** (2003) S59–63

16. Stekel D.J., Git Y., Falciani F.: The Comparison of Gene Expression from Multiple cDNA Libraries. Gen. Res. **10** (2000) 2055–2061
17. Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., Koonin E.V., Krylov D.M., Mazumder R., Mekhedov S.L., Nikolskaya A.N., Rao B.S., Smirnov S., Sverdlov A.V., Vasudevan S., Wolf Y.I., Yin J.J., Natale D.A.: The COG database: an updated version includes eukaryotes. BMC Bioinformatics. **4(1)** (2003) 41.
18. Vettore A.L., da Silva F.R., Kemper E.L., Souza G.M., da Silva A.M., Ferro M.I., Henrique-Silva F., Giglioti E.A., Lemos M.V., Coutinho L.L., Nobrega M.P., Carrer H., Franca S.C., Bacci Junior M., Goldman M.H., Gomes S.L., Nunes L.R., Camargo L.E., Siqueira W.J., Van Sluys M.A., Thiemann O.H., Kuramae E.E., Santelli R.V., Marino C.L., Targon M.L., Ferro J.A., Silveira H.C., Marini D.C., Lemos E.G., Monteiro-Vitorello C.B., Tambor J.H., Carraro D.M., Roberto P.G., Martins V.G., Goldman G.H., de Oliveira R.C., Truffi D., Colombo C.A., Rossi M., de Araujo P.G., Sculaccio S.A., Angella A., Lima M.M., de Rosa Junior V.E., Siviero F., Coscrato V.E., Machado M.A., Grivet L., Di Mauro S.M., Nobrega F.G., Menck C.F., Braga M.D., Telles G.P., Cara F.A., Pedrosa G., Meidanis J., Arruda P. Telles G.P., Braga M.D.V., Dias Z., Lin T. Quitazau J.AMINO-ACID, da Silva F. R., Meidanis J. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res. **13(12)** (2003) 2725–35
19. Wheeler D.L., Church D.M. ,Federhen S., Lash A.E., Madden T.L., Pontius J.U., Schuler G.D.,Schriml L.M., Sequeira E., Tatusova T.A, Wagner L.: Database Resources of the National Center for Biotechnology. Nucl Acids Res **31** (2003) 28–33

5.5 Verificando a anotação com KOG após agrupamento e montagem das EST em contigs

Esse tema foi abordado no final do artigo 3 e em especial no artigo 4, que acabou sendo todo incorporado no artigo 3. Este último, por ser um trabalho mais completo e abrangente, será submetido a publicação em um periódico. Como comentado no item 4.4, *clustering* e montagem de consenso de EST, procedimento amplamente usado em projetos transcriptoma, foram avaliados usando-se o método de teste de anotação com a base KOG. Métodos para avaliação de procedimento de *clustering* e montagem de EST já foram descritos anteriormente, porém usando dados de anotação diretamente no genoma (Burke *et al.*, 1999; Wang *et al.*, 2004) e não ainda com proteínas. Além disso, é possível verificar o efeito do número de EST utilizadas na montagem sobre a performance da anotação, já que projetos EST de portes diferentes são financiados. O programa para agrupamento *e* montagem de consenso de EST usado foi o TGICL, criado no antigo *TIGR Institute of Genomic Research* (Pertea *et al.*, 2003). O programa é na realidade um *script* PERL que controla os programas MegaBLAST (Zhang *et al.*, 2000) e CAP3 (Huang e Madan, 1999). Os artigos 3 e 4 mostram que o resultado do TGICL é aparentemente saturável em torno de 80% ao se usar mais de 50.000 sequências de EST (figuras 29 e 31). Interessantemente, as EST de *Homo sapiens* formaram menos *uniques*, ou montagens de consensos de EST. Uma das hipóteses para o fenômeno é a da presença de um número maior de RNA não codificadores em transcriptomas de mamíferos, que tenderiam a não se agrupar, comparados com outros transcriptomas (Gustincich *et al.*, 2006).

Além disso, o procedimento de agrupamento e montagem de consensos de EST aumentaram a performance do teste de anotação. O resultado foi mais eficiente ao se usar um número de seqüências acima de cinco mil (figuras 30 e 33). Aparentemente a anotação de consensos é menos performática que a de EST, todavia quando o número de EST componente dos consensos é recuperado, verifica-se que a anotação é melhor em cerca de 20%.

Performance of Automatic Annotation of EST data using the secondary database KOG as a model

Maurício A. Mudado, Gabriel R. Fernandes and J. Miguel Ortega

[1] Laboratório de Biodados, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, UFMG, Av. Antônio Carlos 6627, Belo Horizonte - MG, Brasil
{mudado, gfernandes, miguel}@icb.ufmg.br

Abstract. A method to access the performance of automatic annotation of EST data from the Model Organisms (MO) *Arabidopsis thaliana* (Ath), *Caenorhabidtis elegans* (Cel), *Drosophila melanogaster* (Dme) *and Homo sapiens* (Hsa) was reproduced and incremented. Large collections of public ESTs from MO were downloaded and filtered. Similarity cutoffs for alignment of ESTs with amino acid sequences that correspond to the identity of 96% in nucleotide – nucleotide alignments were defined (78%, 81%, 71% and 72% for Ath, Cel, Dme and Has, respectively). ESTs were aligned to the cognate organism proteins ("assignment") and subsequently with the database depleted of them ("annotation"). Annotation was classified as either correct, changed or speculated, the later consisting of non-assignable ESTs that match a KOG entry from anther organism. ESTs attained levels of up to 45% of annotation. All four organisms showed high precision in annotation, with Positive Predicted Value higher than 94%. Changed annotation was evaluated and classified. It revealed that these ESTs were annotated by KOGs having similar description and conserved domain composition similar to the assigned KOG. We suggest that these KOGs should be joined. Random EST subsets (5K, 50K, 100K and 150K) were used as input to the TGICL clustering software package. The clustering percentage saturated at ~80%, by using up to 50K ESTs. Compared to non-clustered ESTs, annotation of assembled ESTs yielded better results which improved as the number of clustered ESTs increased. Compared to non-clustered ESTs, results for *C. elegans* and *D. melanogaster* depicted an increment in the annotation of up to 1.48 and 1.28 fold. Correct annotation was 5.4% and 4.8% higher (up to 1.21 and 1.20 fold). The other organisms showed similar results. The KOG database showed to be a good secondary database for EST annotation. The analyses applied here are promptly applicable to evaluate the performance of other similar databases.

Keywords: KOG, EST, annotation, BLAST, best hit, similarity, clustering

1 Introduction

Expressed sequence tags, or ESTs [1], are a fast and economic way of sampling the expressed genes of an organism [2, 3]. Transcriptome projects [4, 5] and smaller EST sequencing projects [6, 7] are becoming more attractive and frequent as costs in

cDNA sequencing lowers. There is a vast EST data being produced (for actual numbers see http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html) and stored in online available databases [8, 9].

ESTs serve to support genome projects for discovery of novel genes and gene mapping [10]. ESTs are also used for sampling of genes by similarity searches with protein databases and clustering [11-14]. Many pipelines for EST annotation have been developed (reviewed in [15]) as many secondary protein databases are being created [16-18]. One of these databases is KOG [19], which have clusters of proteins that are most likely orthologs by using the best hit approach. KOG have proteins from seven Model Organisms with complete sequenced genomes. KOG clusters have three levels: KOG and TWOG, with proteins from at least three and two different organisms, respectively; and LSE, which are species specific clusters (proteins from one organism only). KOG clusters have a different identification ID (e. g.: KOG0052 for Translation elongation factor EF-1 alpha/Tu), which are supposed to represent different proteins or groups of proteins that evolved distinctively. Also KOG has a higher classification by letters that correspond to functional categories (e. g.: J - Translation, ribosomal structure and biogenesis) which makes it even more interesting for functional annotation.

Before annotation, ESTs are usually filtered for low quality sequences and then clustered to remove redundancy. Transcriptome clustering is a widely used procedure, initiated by the construction of the Human Unigene [14]. Many strategies for EST clustering have been set up and many are still being developed [20, 21]. TIGR have created the Tiger Gene Indices Clustering Tool (TGICL)[22], which contains an initial step that allows for cluster generation in a similar manner to the Unigene procedure, based on MegaBLAST [23] comparisons of the EST sequences, and in a second stage by running Cap3 [24] on each cluster for the production of uniques, which are contigs and singlets (not assembled ESTs).

Functional annotation is generally based on homology searches through alignments with the BLAST software package [25]. The best alignment (best hit) coupled with a low E value threshold usually suffices to suggest a relationship of homology [26, 27]. Because of the intrinsic sequencing errors, alignments of ESTs with nucleotide sequences (other ESTs or cDNA sequences) are usually imposed a similarity cutoff of ~96%, such as in UniGene (Lucas Wagner, personal communication). However, identity cutoffs for EST alignments with protein sequences from the same source are still not a deeply explored issue [28]. Errors in the third base of the codon tend to be silent although errors in the first and second bases would lead to errors in the alignment.

In this work we set up to define identity cutoffs for tBLASTn alignments of large collections of ESTs from four Model Organisms, or its uniques generated by TGICL, with proteins from the KOG database (we speak of identity instead of similarity, since the nucleotide and protein sequences are from the same source organism). Also, we presented a novel method for accessing the performance of this annotation. Results show that KOG database performs acceptably on the annotation of both ESTs and contigs. Moreover, clustering of large amounts of ESTs with TGICL improved annotation by diminishing no hit events and also raising correct annotation of ESTs that build the contigs.

2 Methods

2.1 Sequences

We selected only the 88,613 classified KTL proteins found in the "kog", "twog" and "lse" files at the NCBI KOG FTP site (ftp://ftp.ncbi.nih.gov/pub/COG/KOG/) to set up a MySQL database. This database was used to retrieve the fasta sequences from the "kyva" file available in the same site. Those sequences were used in the BLAST searches as queries. The KTL proteins consist of 60,758 KOG (present in three or more organisms), 4,451 TWOG (present in two organisms) and 23,404 LSE (lineage specific expansion) proteins.

EST sequences from 4 model organisms were downloaded: *Arabdopsis thaliana* (Ath), *Caenorhabditis elegans* (Cel), *Drosophila melanogaster* (Dme) and *Homo sapiens* (Hsa) from NCBI web site (www://ncbi.nlm.nig.gov). EST were filtered to select only health tissues and organs. The total number of EST downloaded was 360,833 for Ath, 302,080 for Cel, 375,360 for Dme and 365,619 for Hsa.

The pUC18 sequences used in this work have been provided by 3 laboratories from Universidade Federal de Minas Gerais (UFMG) that integrate the network Rede Genoma de Minas Gerais. The reactions were made in a single pool and divided into tubes for the PCR sequencing reaction. After the reaction, the sequences were joint again in the same tube, mixed, and then divided into three 96 sequencing well plates. Each plate was run 3 times on a MegaBACE sequencing equipment, yielding a total of 864 reads. From those, 846 processed ESD files were obtained.

2.2 PHRED

The software PHRED used was version was run with the following parameters: -trim alt "" –st fasta -trim cutoff <n> Which activates the trimming algorithm selects the file type and activates the trimming with error cutoff (n) respectively.

2.3 BLAST searches.

BLAST software (version 2.2.13) was used in EST alignment against KTL proteins. tBLASTn program was used with the following parameters:
-m 8 -b 1000000 -e 1e-10 -F f. These parameters activate the tabular output of BLAST, allowing up to 1 million hits to each query protein (default is 250) and deactivates the low-complexity filter, respectively. The low-complexity filter was deactivated in order to allow tBLASTn to achieve 100% identity in the alignments. RPS-BLAST of CDD conserved domains against KOG was also performed, and best hits selected.

2.4 Clustering

The software TGICL (http://compbio.dfci.harvard.edu/tgi) was used to cluster the EST and generate uniques. TGICL was run with the following parameters: –p 95 –l 40 –v 30. These parameters cluster sequences which overlap with at least 95% similarity, at least 40 bp identical and 30 bp distance from overlap to sequence end. Also, TGICL script was modified to include the following parameters to run the tclust software: SCOV=70 PID=95. These parameters force building of high stringency clusters with at least 95% of identity and 70% coverage of the shorter sequence. The PERL package Math::Random (http://www.cpan.org) was used in order to select random subsets of 5K, 50K, 100K and 150K ESTs for clustering.

3 Results and Discussion

3.1 Establishing the default identity cutoffs

Errors in EST sequences may lead to either fewer or greater errors in the alignments with the correspondent protein. In [28] we defined amino acid identity cutoffs that are equivalent to 96% nucleotide identity cutoffs for the EST collections from the four Model Organisms Ath, Cel, Dme and Hsa. In this work, the same approach was reproduced with newer EST collections (~300K EST for each organism), with a more proper database selection, where cDNA libraries were filtered for normal tissues, organs and environment conditions where available, as well as for the size of the

libraries (most of them presenting more than 5K EST). 50 KOG proteins were selected for each organism, requiring that they had high EST sampling and no paralog. The respective 50 CDS nucleotide sequences were retrieved from GenBank [29]. The pUC vector was translated *in silico* and stop codons substituted with the "*" character, known to be recognized by BLAST. EST sequences were then aligned with either BLASTn or tBLASTn with the respective nucleotide and amino acid sequences and best hits selected. Fig. 1 shows that the tBLASTn alignments always show lower mean identity compared to BLASTn ones. Also, by varying the error densities of the pUC reads (Fig. 1A) with PHRED trimming cutoff algorithm, it is feasible to predict the error densities of the organism's EST collections (Fig. 1B). We selected the tBLASTn identity cutoff that would ensure the correct assignment of a given EST to its correlated protein by comparing the number of events (tuples of amino acid and nucleotide alignments) that the 96% cutoff for BLASTn limited. Fig. 1C visually shows the amount of tuples (94% of the total, data not shown) selected by the 96% identity cutoff for BLASTn (horizontal dashed line), which equals and the amount selected by 81% identity cutoff for tBLASTn (vertical dashed line). It can be noticed that nucleotide identity cutoffs of 96% are not perfect and removes 6.71% of good tuples from pUC aligments. It also removed 5.52%, 1.60%, 0.54% and 6.51% of good tuples from Ath, Cel, Dme and Hsa respectively (data not shown). Fig. 1D shows how the amino acid identity cutoff (short vertical dashed line) for Cel was selected by fixating the number of events (long horizontal dashed line) selected by the nucleotide cutoff (dotted line). By doing this to all four organisms, we found the default values for amino acid identity cutoffs of 78%, 81%, 71% and 72% for Ath, Cel, Dme and has, that would grant reliable alignment of a given EST to its corresponding protein (for full results, see supplementary figures S1 and S2). These values are close to the one find from the controlled experiment with the pUC vector (82%) and the cutoffs for the former not-filtered EST collection for these same organisms (84%, 78% 78% and 80%), see [28] for more details.

3.2 Measuring the quality of annotation

These cutoffs were then used in the assignment of the EST to the correspondent KTL protein by selecting the best hit from tBLASTn alignments (Fig. 2, right side). The annotation of the EST was performed by removing the proteins from the proper organism from the database, and aligning the EST with the remaining six organism proteins with tBLASTn (Fig. 2 left side) and selecting the best hit. Tthis step is a simulation of the annotation of a novel transcriptome. Five types of annotation are obtatined: correct, changed, speculated, 'assigned but no hit' and 'no hit'. When the assignment and the annotation of an EST are both to the same KOG ID, the annotation is correct and when they are not, the annotation is changed. When an EST annotates to a KOG ID but do not assign we consider that the database is speculating an annotation. When an EST is assigned to a KOG ID but does not annotate to any KOG ID, we classify it as 'assigned but no hit'. Finally, when there occurs nor assignment neither annotation, it is defined as 'no hit' (see Table 1).

From all annotated EST, over 90% of correct annotation was obtained to all four organisms and changed annotation was lower than 5.2% (see supplementary fig. S3).

Assignment cutoffs from 45% to 100% and no cutoff at all were tested in order to verify its influence in the performance. Fig. 3 shows the resultant annotation to all EST. The default assignment cutoffs (78%, 81%, 71% and 72% for Ath, Cel, Dme and Hsa) seemed to be appropriate, since they did not lead to higher speculative annotation (see dashed lines in Fig. 3) and tended to decrease changed annotation. The defined cutoffs improved changed annotation by 11.2%, 18.4%, 10.16% and 24.7% and prompted small loss of correct annotation of 2.5%, 1.5%, 1.5% and 3.8% for Ath, Cel, Dme and Hsa respectively, in relation to no cutoff usage (See supplementary figure S4 for full results). It is expected that incorrect assignment of a given EST to a related protein (e.g. bearing a conserved domain) might occasionally result on a coincident annotation to the same KOG ID, moreover the performance measured without a cutoff probably will deeply depend on the database used. Although the relative small loss of correct annotation when the cutoff is applyied, it represents a significant number of EST (2,324 for Ath; 1,957 for Cel; 1,750 for Dme and 5,038 for Hsa) compared to the number of EST from changed annotation loss (547 for Ath; 462 for Cel; 869 for Dme and 535 for Hsa). However, performance is not significantivelly different if no cutoff is used.

Total annotation (correct + changed + speculated) added up to around 40% for all organisms except for Ath, that was under 30%. This high fraction of "no hit" and "assigned but no hit" are probably due to the divergence of Ath proteins and because it is the only plant in the database. Moreover, some assignments are to LSE (see below). Changed annotation was always small, lower than 1.5% to all EST collections, Correct and speculative annotations show a dependent pattern because inadequately higher assignment cutoffs avoid EST sequences to assign to the correspondent protein, and therefore they end up entering the speculated classification. "Assignment but no hit" in the annotation step attained values of around 30% to Cel, Dme and Hsa, but the value raised to over 50% to Ath (see supplementary figure S5). This class of annotation is partially caused by assignment to LSE proteins (exclusive of the organism), that therefore would lead to an incapability of annotation. Ath have more LSE proteins than the other organisms (9,563 against 5,837 for Cel, 5,960 for Hsa and 1,288 for Dme), also probably because it is the only plant between the seven KOG organisms. Assignment to LSE proteins leads to no annotation up to 5% in Dme, 10% in Cel and Hsa, and up to 25% in Ath . We decided to keep LSE proteins in the study to show that the KOG database still have important expressed LSE proteins that probably would be clustered into TWOG or KOG clusters if more proteins from other organisms were used.

"No hit" was more pronounced in Hsa EST (around 40% of the total). The other organisms showed values of around 30% (Dme) and 20% (Cel and Ath, see supplementary Fig. S5). The EST sequences were also surveyed for size and percentage of 3'–5' oriented ESTs. No bias was found for any class of annotation to these parameters (see supplementary table T1). It is possible that the Hsa EST might correspond to a large fraction of non KOG entries, due to the diversity of cDNA libraries generated, with some contribution of a larger UTR region.

3.3 Changed annotation Analysis

Changed annotation was analyzed in more detail. Figure 4 shows an example of this analysis for Cel. Correct and changed annotated EST showed different distributions per assignment identity; changed annotated EST presented a significant fraction of hits of lower identity during the assignment step, what does not occur for correct annotated ones (Fig. 4A and 4B). Similar results were obtained for the other 3 organisms (data not shown). This explains why PPV raises when higher assignment cutoffs are used (discussed below, Fig. 4D). Higher cutoffs brought the mean identity to almost the same level of the Correct annotated EST (Fig. 4C).
Another measure of the effect of the cutoffs, the positive predictive value (PPV), also called precision [30], was calculated by the following formula (for a review, see [31]):

PPV = Correct annotation / (Correct annotation + changed annotation)

PPV showed values above 95% to all organisms and above 94% for Ath (Fig. 4D). Higher values of PPV are obtained with higher assignment cutoffs although the loss tend to increase (speculative annotation raises and correct annotation diminishes with excessively higher assignment cutoffs, see Fig. 3). Therefore the default assignment cutoffs tended to augment annotation precision without critical annotation loss.

Increasing subsets of 50K, 100K and 150K randomly selected EST (n=5) from all four organisms were analyzed. The results showed very close annotation values with small mean and S. E. M., meaning that the sets of EST used have stable annotation values (data not shown).

3.3 Changed annotation classification

All changed annotated EST were manually surveyed and classified into five groups (see Fig 5): LSE, when the EST was assigned to a LSE protein (therefore any proposed annotation would result in changed annotation); String, when the KOG IDs from assigned and annotated ESTs had the same description, an option from the KOG database of to split a cluster of related proteins into distinct entries, probably due to divergent evolution, e.g. KOG3525 and KOG3526, Subtilisin-like proprotein convertase; Family, when the KOG IDs from assigned and annotated ESTs had a very similar description and mostly the same conserved domains (e.g. KOG1543-Cysteine proteinase Cathepsin L and KOG1544-Predicted Cysteine proteinase TIN-ag – for information on conserved domains, see supplementary information); FOG when EST where assigned to KOG - Fuzzy Ortholog Groups (FOG), that are huge KOG clusters with many similar proteins with promiscuous conserved domains; and Other, when there was other reasons for changed annotation. Fig. 5 shows the distribution of classes of changed annotation for all organisms. Considering the Cel changed annotation, over 63.8% (75.0%, 37.9% and 61.3% for Ath, Dme and Hsa respectively) is explained by the classifications LSE, FOG, Family and String and around 46.2% are in the Other group (25.0%, 61.1% and 38.7% for Ath, Dme and Hsa). The Family and String groups are an indicative that KOG database have different clusters that might be joined into single KOG entries. Cel showed low values

for these groups (7.4% and 7.9%, respectively). Hsa (6.6% and 4.3% respectively) and Dme (4.5% and 2.0% respectively) presented similar results. Ath, on the other hand, presented higher values for these groups (49.2% and 15.4%). This, added to the higher number of LSE, is an indicative that the KOG database is poorer structured for Ath, which is reasonable for being the only plant among the other organisms. The FOG group was responsible for around the same amount of changed annotation for all four organisms (10.3%, 13.8%, 13.4% and 9.9% for Ath, Cel, Dme and Ath).

3.5 Clustering randomly selected EST

EST sequences were randomly selected in incremental sets of 5K, 10K, 50K, 100K and 150K, with the Math::Random PERL package. TGICL was run with these subsets in order to know if assemblage of EST into contigs saturates and how many EST sequences are needed in order to achieve a clustering plateau. Fig. 6 shows that the percentage of EST in clusters (non-singlets) rises exponentially from ~40% to ~70% when using 5K to 50K EST for Cel, Dme and Ath. Clustering then stalls to 80% when using more than 100K EST. This result proves that assemblage is dependent of the number of EST sequences used in the clustering with TGICL. *H. sapiens* had a much lower clustering percentage compared to the other organisms EST. This phenomenon is still under investigation, but might be related to a more boadly exploitation of the tissues and organs in the composition of the EST collection. As it can be observed in Fig. 6, 5K EST are below the ideal number of EST for clustering as this number of sequences are not enough to achieve the clustering plateau. Also, clustering of over 50K EST lead to an improvement in annotation (see Fig. 7).

As seen by the linear pattern in Fig. 7 C and F, non-clustered EST tend to have the same performance of annotation in all sets of EST, compared to the annotation of uniques (Fig. 7 A and D) and non-clustered EST (Fig. 7 B and E). Non-clustered EST for Cel and Dme show ~44% and 41% of correct annotation, around 32% of "no hit" annotation and almost 1.5% of changed annotation for both organisms in all sets of sorted EST. Thus, performance of KOG database for automated annotation is the same independently of the size of the project. On the other hand, uniques and clustered EST showed different patterns of annotation in all subsets of EST. This result shows that incrementing the number of clustered EST leads to an input of novel information to the annotation process.

The annotation of uniques shows that no hit annotation raises 6% and 6.8% (up to 1.24 and 1.86 fold compared to non-clustered EST) and correct annotation diminishes 7% and 8% (up to 0.7 fold for both organisms, compared to non-clustered EST), for Cel and Dme, respectively, as the collection raises from 5K to 150 K EST (Fig. 7 A and D). Furthermore, the annotation of the EST comprised by these uniques (Fig. 7 B and E) show an increase in correct annotation of 5.4% and 4.8% (up to 1.21 and 1.20 fold compared to non-clustered EST) and diminishing of "no hit" annotation of 3.1% and 5.9% (up to 0.67 and 0.77 fold, leading to an increment in the annotation of up to 1.48 and 1.28 fold, compared to non-clustered EST) for Cel and Dme, respectively. The difference in annotation between uniques and its ESTs are showing that the number of changed annotated contigs and "no hit" contigs are raising but the size of correct annotated contigs are also raising (more correct EST are being assembled)

with incrementing number of input EST to TGICL. Thus, EST sequences that represent these uniques are being annotated with more quality and in large number. Results were similar to Ath and Hsa (data not shown).

4 Conclusion

We presented a novel method for accessing the quality of automatic annotation of public EST sequences from four model organisms with the KOG database, using tBLASTn best hits. The methodology consisted of calculating identity cutoffs, proportional to 96% cutoff for BLASTn, which were found to be 78%, 81%, 71% and 72% for Ath, Cel, Dme and Hsa respectively. These cutoffs were then used in the assignment of the EST to its cognate organism KOG protein. The best hits were compared to the annotation of the same EST with the KOG database, depleted of the cognate organism proteins used in the assignment in first place.

The methodology had a good annotation performance with over 90% correct annotation and less than 10% of changed and speculated annotation for all organisms. Total EST annotation had a high "no hit" number, and covered between 40% and 50% of the EST from Cel, Dme and Hsa and below 30% of the EST from Ath. This result shows that transcriptome data have much yet to be discovered. One may speculate that much of the expressed transcripts are non-coding (for a review see [32]).

The evaluation of the method showed that the cutoffs are adequate, improving annotation with small loss in correct annotation (up to 3.8% for Hsa) compared to the loss in changed annotation (up to 24.7% also for Hsa). Furthermore, the annotation had high precision, with PPV values around 95%. Analysis of the changed annotation showed KOG clusters that might have been joined together. These KOG had different IDs in assignment and annotation but showed the same or very close descriptions with similar composition of conserved domains.

Clustering EST improved annotation as shown by the results obtained with TGICL. Correct annotated EST increased and "no hit" EST fraction diminished. Clustering proved to be a saturating process, showing saturation at around 80% when using over 50K EST. The cutoffs used also showed better results in annotation than without cutoffs.

Overall, the KOG database presented as a good secondary database for the automatic annotation of EST. Since more genomes are being completed and annotated, we expect the KOG database to have more organisms incorporated. As this happens, the annotation of transcriptome data with KOG is most likely to improve. The method described here can also be utilized for evaluation of the performance of other secondary databases organized in the same way. Data is already been prepared for the KEGG/KO orthology database [33] and OrthoMCL-DB [34].

Acknowledgements

Research supported by CAPES, FAPEMIG and CNPq/MCT. Thanks to Rede Genoma de Minas Gerais, for the pUC sequences.

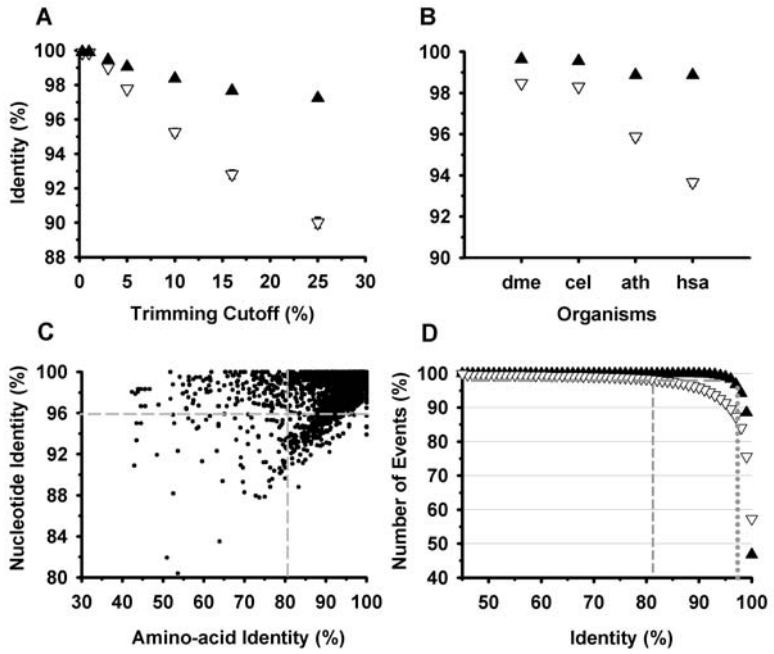


Fig. 1. A and B. Mean identity ± S.E.M. obtained from tBLASTn (*open inverted triangles*) versus BLASTn (*full triangles*) alignments. A. Alignments of pUC18 vector ESTs against its own in silico translated protein and nucleotide sequences. B. Alignments of the four organisms ESTs against 50 KOG proteins and the respective nucleotide sequences. C, BLASTn x tBLASTn tuples obtained from the alignment of Cel ESTs and the 50 KOG protein and nucleotide sequences (*dashed lines represent the identity cutoffs, see text for details*). D, data from C , showing the tBLASTn (*open inverted triangles*) versus BLASTn (*full triangles*) alignments sorted by number of events versus identity; (*short dashed and dotted lines*) represent the amino acid and nucleotide identity cutoffs respectively; (*the long dashed line*) represent the number of events that the cutoffs are limiting.

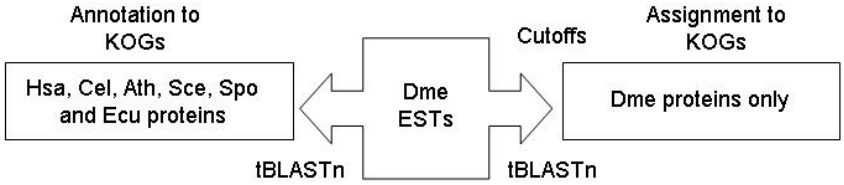


Fig. 2. Schema of Dme ESTs (*center*) assignment/annotation with tBLASTn. The ESTs are assigned to Dme's own KOG proteins with the use of similarity cutoffs (*right side)* and annotated to all KOG proteins but Dme's KOG proteins (*left side).*

Table 1. All possible types of annotation.

Type of Annotation	Assignment	Annotation	KOG ID
Correct	+	+	Same
Changed	+	+	Different
Speculated	-	+	Any
Assign. But no Hit	+	-	Any
No Hit	-	-	-

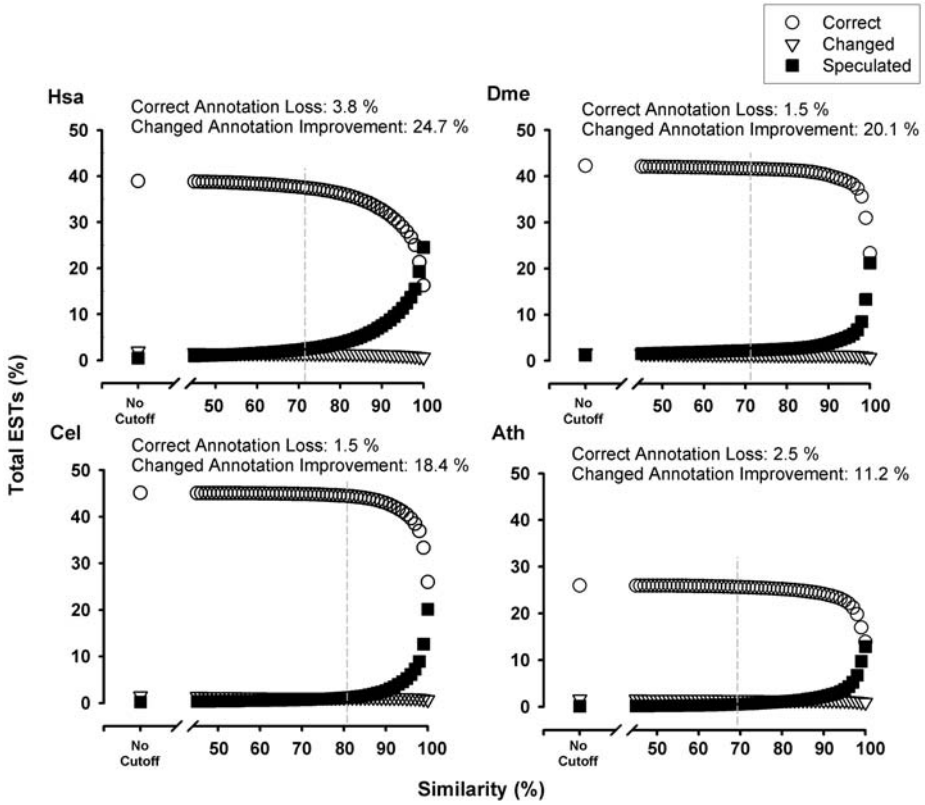


Fig. 3. Result of annotation of ESTs using no assignment identity cutoffs and variable assignment identity cutoffs (45% to 100%). Correct annotation (*open circles*), speculated annotation (*full squares*) and changed annotation (*open inverted triangles*) are shown. The selected assignment identity cutoffs: 78%, 81%, 71% and 72% for Ath, Cel, Dme and Hsa are shown (*vertical dashed lines*).

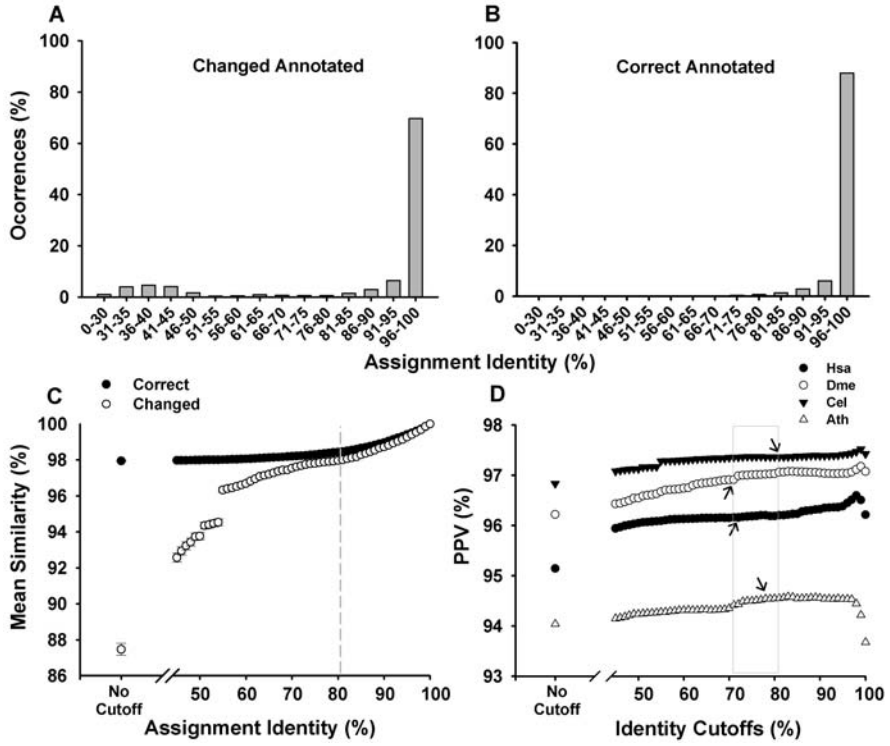


Fig. 4. *C. elegans* Changed annotation analysis. *A and B.* Distribution by identity assignment.*C.* Mean similarity ± S.E.M. distribution per assignment identity for Correct and Changed annotated ESTs (*black circles and white circles respectively*). *D.* PPV for Hsa (*full circles*), Dme (*open circles*), Cel (*inverted full triangles*) and Ath (*open triangles*); arrows show the default identity cutoff for assignment (*region marked by a gray square*).

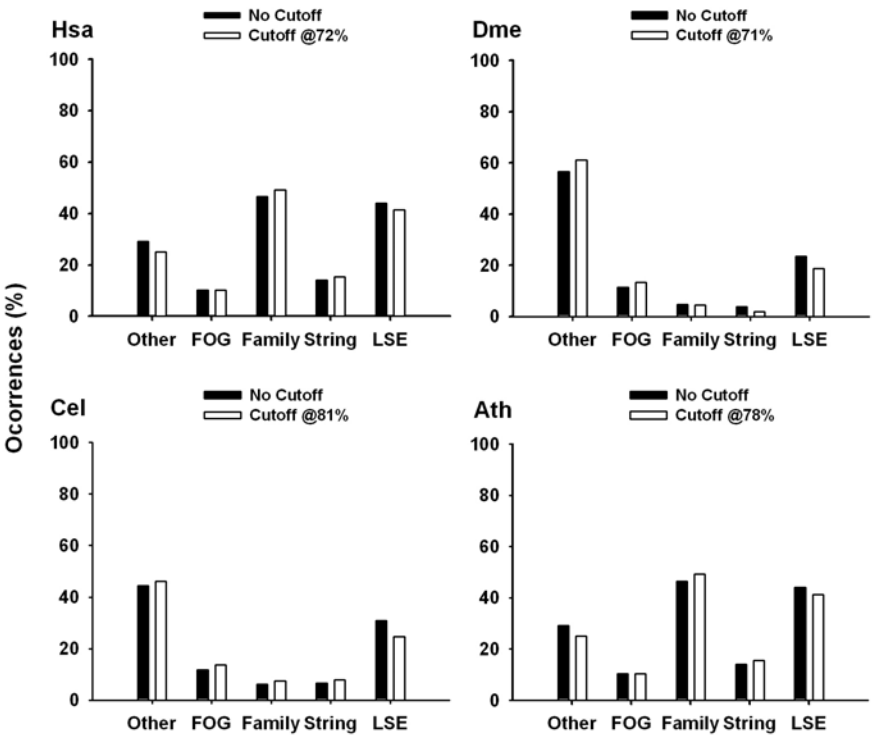


Fig. 5 Histogram of Changed annotated ESTs, for Hsa, Cel, Dme and Ath, distributed by classes (*Other, FOG, Family, String and LSE*; *black bars for no assignment cutoff and white bars when the default cutoff is used*); the mean similarity ± S.E.M for each class, when using no assignment cutoff (*black circles*) and the default cutoff of 81% for Cel (*white circles*) are showed.

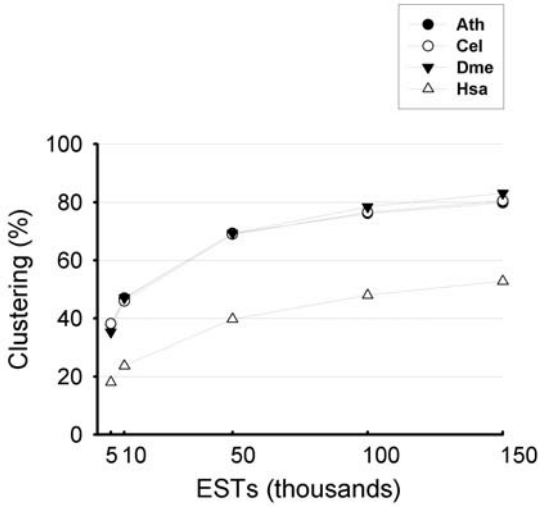


Fig. 6. A.Percentage of ESTs in clusters (clustering) of incrementing number of ESTs selected at random (5K, 10K, 50K, 100K and 150K). *A. thaliana* (*Ath*), *C. elegans* (*Cel*), *D. melanogaster* (*Dme*) and *H. sapiens* (|*Hsa*) are shown (*full circles, open circles, full inverted triangles, open inverted triangles respectively*).

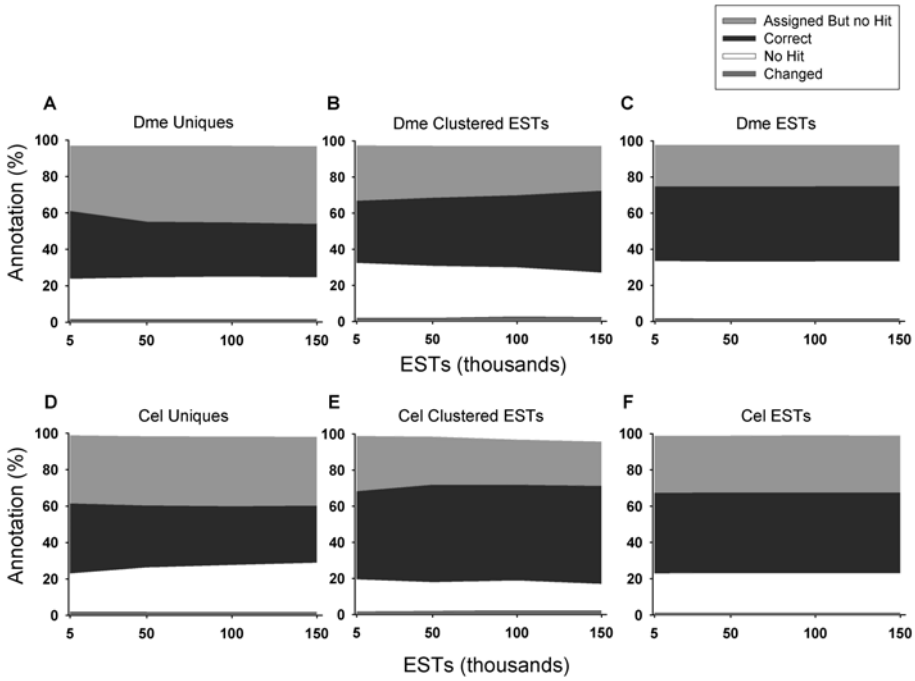


Fig. 7. *D. melanogaster* and *C. elegans* annotations of uniques (*A and D*), clustered ESTs (*B and E*) and ESTs only (*C and F*), using the default assignment cutoffs. Annotations represented, from top to bottom: Assigned but No Hit (*lighter grey*), Correct (*black*), No Hit (*white*), and Changed (*dark gray*).

References

1. Adams, M.D., et al., *Complementary DNA sequencing: expressed sequence tags and human genome project.* Science, 1991. **252**(5013): p. 1651-6.
2. Adams, M.D., et al., *3,400 new expressed sequence tags identify diversity of transcripts in human brain.* Nat Genet, 1993. **4**(3): p. 256-67.
3. Franco, G.R., et al., *Identification of new Schistosoma mansoni genes by the EST strategy using a directional cDNA library.* Gene, 1995. **152**(2): p. 141-7.
4. Verjovski-Almeida, S., et al., *Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni.* Nat Genet, 2003. **35**(2): p. 148-57.
5. Vettore, A.L., et al., *Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane.* Genome Res, 2003. **13**(12): p. 2725-35.
6. Semova, N., et al., *Generation, annotation, and analysis of an extensive Aspergillus niger EST collection.* BMC Microbiol, 2006. **6**: p. 7.
7. Vizcaino, J.A., et al., *Generation, annotation and analysis of ESTs from Trichoderma harzianum CECT 2413.* BMC Genomics, 2006. **7**: p. 193.
8. Boguski, M.S., T.M. Lowe, and C.M. Tolstoshev, *dbEST--database for "expressed sequence tags".* Nat Genet, 1993. **4**(4): p. 332-3.
9. Strausberg, R.L., et al., *The mammalian gene collection.* Science, 1999. **286**(5439): p. 455-7.
10. de Souza, S.J., et al., *Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags.* Proc Natl Acad Sci U S A, 2000. **97**(23): p. 12690-3.
11. Faria-Campos, A.C., et al., *Mining microorganism EST databases in the quest for new proteins.* Genet Mol Res, 2003. **2**(1): p. 169-77.
12. Mudado Mde, A. and J.M. Ortega, *A picture of gene sampling/expression in model organisms using ESTs and KOG proteins.* Genet Mol Res, 2006. **5**(1): p. 242-53.
13. Faria-Campos, A.C., et al., *Efficient secondary database driven annotation using model organism sequences.* In Silico Biol, 2006. **6**(5): p. 363-72.
14. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology.* Nucleic Acids Res, 2003. **31**(1): p. 28-33.
15. Nagaraj, S.H., R.B. Gasser, and S. Ranganathan, *A hitchhiker's guide to expressed sequence tag (EST) analysis.* Brief Bioinform, 2007. **8**(1): p. 6-21.
16. Apweiler, R., A. Bairoch, and C.H. Wu, *Protein sequence databases.* Curr Opin Chem Biol, 2004. **8**(1): p. 76-80.
17. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.
18. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Res, 2000. **28**(1): p. 33-6.
19. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**: p. 41.
20. Masoudi-Nejad, A., et al., *EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W459-62.
21. Ptitsyn, A. and W. Hide, *CLU: a new algorithm for EST clustering.* BMC Bioinformatics, 2005. **6 Suppl 2**: p. S3.
22. Pertea, G., et al., *TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets.* Bioinformatics, 2003. **19**(5): p. 651-2.
23. Zhang, Z., et al., *A greedy algorithm for aligning DNA sequences.* J Comput Biol, 2000. **7**(1-2): p. 203-14.
24. Huang, X. and A. Madan, *CAP3: A DNA sequence assembly program.* Genome Res, 1999. **9**(9): p. 868-77.

25. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
26. Koonin, E.V., *Orthologs, paralogs, and evolutionary genomics.* Annu Rev Genet, 2005. **39**: p. 309-38.
27. Koonin, E.V. and M.Y. Galperin, *Sequence - evolution - function: computational approaches in comparative genomics*. 2003, Boston: Kluwer Academic. xiii, 461 p., [11] p. of plates.
28. Mudado M, A., E. Bravo-Neto, and M. Ortega J, *Tests of automatic annotation using KOG proteins and ESTs from 4 eukaryotic organisms.* Lecture Notes Computer Sci., 2005. **3594**: p. 141-152.
29. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2007. **35**(Database issue): p. D21-5.
30. Wu, J., et al., *KOBAS server: a web-based platform for automated annotation and pathway identification.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W720-4.
31. Gambino, B., *Reflections on accuracy.* J Gambl Stud, 2006. **22**(4): p. 393-404.
32. Mendes Soares, L.M. and J. Valcarcel, *The expanding transcriptome: the genome as the 'Book of Sand'.* Embo J, 2006. **25**(5): p. 923-31.
33. Kanehisa, M., et al., *The KEGG databases at GenomeNet.* Nucleic Acids Res, 2002. **30**(1): p. 42-6.
34. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.

On The Improvement of Transcriptome Annotation After Clustering and Assemblage of Incremental Number of ESTs

Maurício A. Mudado and J. Miguel Ortega

Laboratório de Biodados, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, UFMG, Av. Antônio Carlos 6627, Belo Horizonte - MG, Brasil
{mudado,miguel}@icb.ufmg.br

Abstract. EST clustering is a widely used procedure in transcriptome projects and is a common sense to improve annotation. In this work we demonstrate a method to test the BLAST annotation of four Model Organism's ESTs and its uniques assembled with the TGICL assemblage software, with the KOG database. Increased numbers of ESTs were used and results show that clustering is reduced by using 5K ESTs but approaches saturation around 80%, by using over 50K ESTs. Compared to non-clustered ESTs, annotation of assembled ESTs shows better results and improves as increased number of ESTs is used. Compared to non-clustered ESTs, results for *C. elegans* and *D. melanogaster* show an increment in the annotation, by diminishing no hit annotations (around 0.8 fold) and raising correct annotation by around (1.2 fold) in both organisms. Thus, a 20% improvement of correct annotation is attained by EST clustering and assemblage

Keywords: Transcriptome, EST, Clustering, TGICL, BLAST, annotation, KOG

1 Introduction

The clustering of a transcriptome is a widely used procedure, initiated by the construction of the Human Unigene [1]. In Unigene, single-pass partial cDNA sequences also known as Expressed Sequence Tag or EST [2] are compared to each other and to available cDNA sequences with the program MegaBLAST, a greedy version [3] of BLAST software [4] developed to increase the speed up the clustering procedure. Other initiatives relied on the use of the assemblage software Cap3 [5] to simultaneously cluster and assemble clustered sequences into consensus sequences also known as contigs. Besides not being designed for clustering, the main difficulty inherent from the use of Cap3 for processing large EST collections is the intense use of memory. Some researchers used to break transcriptomes in reasonable samples to generate contigs and later submit their contigs to subsequent rounds of assemblage with Cap3 [6]. TIGR has produced Transcript Index (TI) for several organisms using such approach, but its bioinformatics team has later developed a software package known as TGICL Tool, which contains a initial step that allows for cluster generation

in a similar manner to the Unigene procedure, based on MegaBLAST comparisons of the EST sequences, and in a second stage by running Cap3 on each cluster, producing contigs and singlets (non clustered ESTs), which constitutes the uniques or transcript index (TI) sequences [7].

A common sense in the literature is that the assemblage of individual EST sequences into contigs shall improve annotation, due to the fact that the larger the sequence, higher the score in BLAST comparisons to public available databases. However, no exhaustive experimental investigation has been conducted on this issue. Our group has developed an approach that suits to this demand. The procedure makes use of KOG database, in which proteins from diverse Model Organisms are clustered into groups of orthologs and paralogs [8], In a first round, ESTs are assigned to the cognate organism proteins, providing with a positive control for the annotation and, in a second round, already assigned ESTs are annotated by KOG entries from other organisms, and the annotation is compared to the initial assignment step. Thus, resultant annotation can lay in three categories: correct, changed and speculated. The last occurs when a EST is not assigned to the cognate organism KOG entry, but the database speculate an annotation for it, by aligning it to a KOG entry from other organism. Together with these three categories for annotated ESTs there are also ESTs in the "no hit" category, which can be either too short to provide a hit of alternatively representing genes not present in KOG dataset, and "assigned but no hit" during annotation procedure, which might concentrate genes that are specific to the analyzed organism (e.g. *A. thaliana*, the only plant in KOG database).

Here we present a test of TIGCL clustering and assemblage of incremental number of ESTs from *C. elegans* and *D. melanogaster* and tests of annotation with KOG database. Assemblage was sensible to the input number of ESTs (reduced with 5K but saturating by 150K ESTs). We confirm that the generation of contigs improved annotation without adding potential errors that could result from chimerical assemblage of distinct genes. This effect, as calculated in terms of total ESTs analyzed, led to a very small increase in changed annotation (up to 1.08%) for 150K. Moreover, an important bias on clustering and assemblage of genes that are prompted to correct annotation drives the apparent result that correct annotation is poorer (0.71 fold) if uniques are annotated as opposite to individual ESTs (around 1.2 fold). Furthermore, similar results have been obtained with the *A. thaliana* and *H. sapiens* ESTs.

2 Methods

2.1 Sequences

Large sets of ESTs from four model organisms were downloaded from GenBank at the NCBI web site (http://www.ncbi.nlm.nig.gov): 360,833 for *Arabdopsis thaliana* (Ath); 302,080 for *Caenorhabditis elegans* (Cel); 375,360 for *Drosophila melanogaster* (Dme) and 365,619 for *Homo sapiens* (Hsa). ESTs were filtered for health tissues and organs.

The KOG database was filtered for the 88,613 classified KTL proteins from seven Model Organisms, found in the "kog", "twog" and "lse" files at (ftp://ftp.ncbi.nih.gov/pub/COG/KOG/). A MySQL database was populated with this data and used to select the respective fasta sequences from the "kyva" file ate the same site. The KTL proteins were divided into 60,758 KOG, 4,451 TWOG and 23,404 LSE proteins.

2.2 BLASTs

BLAST software (version 2.2.13) was used in the alignment of uniques against KTL proteins. tBLASTn was used with the following parameters:
-m 8 -b 1000000 -e 1e-10 -F f. These parameters activate the tabular output of BLAST, allowing up to 1 million hits to one protein (default is 250) and deactivates the low-complexity filter, respectively. The low-complexity filter was deactivated in order to allow tBLASTn to achieve 100% identity in the alignments.

2.3 Clustering

The software TGICL (http://compbio.dfci.harvard.edu/tgi) was used to cluster the ESTs and generate uniques. TGICL was run with the following parameters: –p 95 –l 40 –v 30. These parameters put together sequences which overlap with at least 95% similarity, at least 40 bp identical and 30 bp distance from overlap to sequence end. Also, TGICL script was changed to include the following parameters to run the tclust software: SCOV=70 PID=95. These parameters force building of high stringency clusters with at least 95% of identity and 70% coverage of the shorter sequence. The PERL package Math::Random (http://www.cpan.org) was used in order to select random subsets of ESTs for clustering.

3 Results and Discussion

3.1 Clustering randomly selected ESTs

ESTs were randomly selected in incremental sets of 5K, 10K, 50K, 100K and 150K, with the Math::Random PERL package. TGICL was run with these subsets in order to know if assemblage of ESTs saturates and how many ESTs are needed in order to achieve a clustering plateau. As seen in Fig.1 the percentage of ESTs in clusters (non-singlets) raises exponentially from ~40% to ~70% when using 5K to 50K ESTs for Cel, Dme and Ath. Clustering then stalls to 80% when using more than 100K ESTs. This result proves that assemblage is dependent of the number of ESTs used for clustering with TGICL. *H. sapiens* had a much lower clustering percentage compared to the other organisms ESTs, probably because of greater number of 3'-5' ESTs.

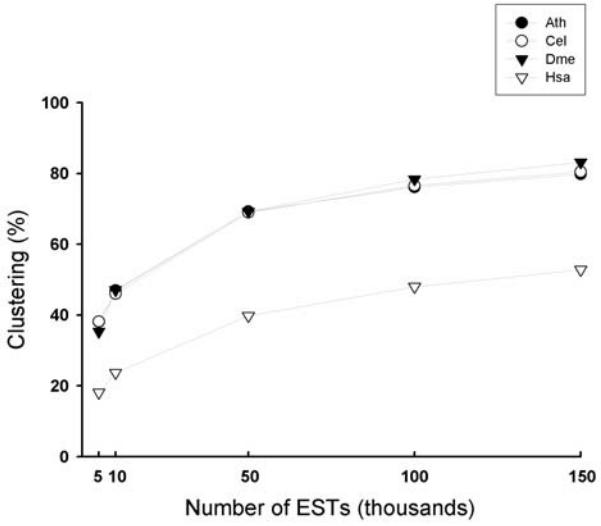


Fig. 1. Percentage of ESTs in clusters (*clustering*) of incrementing number of ESTs selected at random. *A. thaliana* (*Ath*), *C. elegans* (*Cel*), *D. melanogaster* (*Dme*) and *H. sapiens* (|*Hsa*) are shown (*full circles, open circles, full inverted triangles, open inverted triangles*).

3.2 Measuring the annotation quality of uniques

Uniques were then aligned with tBLASTn with the KTL proteins from KOG database (see methods). The annotation experiment for the uniques with the KOG database has two steps. First, uniques are assigned to its proper organism's KTL proteins by selecting the best hits from tBLASTn alignments (Fig.2, right side). Second, uniques are annotated by removing the proper organism's proteins from the database, aligning the uniques with the remaining six organism's proteins with tBLASTn (Fig.2 left side) and always selecting the best hits.

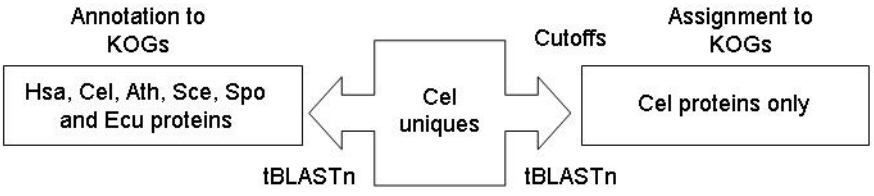


Fig. 2. Schema of the assignment and annotation of *C. elegans* uniques with KOG proteins. The uniques are assigned to Cel's own KOG proteins with the use of similarity cutoffs (*right side)* and annotated to all KOG proteins but Cel's KOG proteins (*left side)*. All organisms' uniques passed by the same pipeline.

Five types of annotation are allowed: correct, changed, speculated, 'assigned but no hit' and 'no hit'. When the assignment and the annotation of a unique are both to the same KOG ID, the annotation is correct and when they are not, the annotation is changed. When a unique annotates to a KOG ID but did not assign we say that the database is speculating an annotation. When a unique is assigned to a KOG ID but didn't annotate to any KOG ID, we say that it is 'assigned but no hit'. Finally, when there is nor assignment neither annotation, it is defined as 'no hit' (see Table 2).

Table 1. Types of annotation.

Type of Annotation	Assignment	Annotation	KOG ID
Correct	+	+	Same
Changed	+	+	Different
Speculated	-	+	Any
Assign. But no Hit	+	-	Any
No Hit	-	-	-

Fig. 3A and C shows a comparison of the quality of annotation of the uniques (white symbols), formed from subsets of 5K, 50K, 100K and 150K ESTs from Cel and Dme, to the direct annotation of the same set of non-clustered ESTs (no TGICL used) (black symbols). Fig. 3B and D shows results from a similar experiment, where the quality of annotation is shown by directly computation of the number of ESTs that are comprised by the uniques formed previously (white symbols). The same comparison against non-clustered ESTs is shown (black symbols).

As seen in Fig. 3 A through D, non-clustered ESTs (black symbols) tend to have the same pattern of annotation (almost linear) in all sets of ESTs compared to the annotation of uniques (white symbols). Non-clustered ESTs from Cel and Dme show ~44% and 41% of correct annotation and around 32% of no hit annotation and almost 1.5 % of changed annotation for both organisms in all sets of ESTs annotated. On the other hand, the result for uniques and the clustered ESTs comprised by these uniques, shows that incrementing the number of clustered ESTs leads to an input of novel information to the annotation process.

The annotation of uniques shows that no hit annotation raises 6% and 6.8% (up to 1.24 and 1.34 fold compared to non-clustered ESTs) and correct annotation diminishes 7% and 8% (up to 0.7 fold for both organisms, compared to non-clustered ESTs), for Cel and Dme respectively, by using up to 150 K ESTs (Fig. 3 A and C). The assigned but no hit annotation is also higher in uniques by 6.4% in Cel and 0.1% in Dme (1.2 and 1.0 fold respectively), compared to non-clustered ESTs in the same range of ESTs. However, the annotation of the ESTs comprised by these uniques (Fig. 3 B and D) depicts a different picture. Clustering is already effective by using 5K ESTs in both Cel and Dme (see small increment in correct annotation and diminishing of assigned but no hit and no hit annotations, compared to non-clustered ESTs). By using up to 150K ESTs there is an augment in correct annotation of 5.4% and 4.8% (up to 1.21 and 1.20 fold compared to non-clustered ESTs) and a diminishing of no hit annotation of 3.1% and 5.9% (up to 0.67 and 0.77 fold compared to non-clustered ESTs) for Cel and Dme respectively. The assigned but no hit annotation is also

diminished by 6.9% and 3% (0.78 and 0.87 fold) for Cel and Dme respectively, in the same range (150K ESTs). Changed annotation is only augmented by 1.08% and 0.89% for Cel and Dme respectively. Speculative annotation was unchanged, with values around 2% and was suppressed from the graphic. The difference in annotation between uniques and its ESTs (Fig. 3 A and C versus Fig. 3 B and D) is due to a numerical artifact. Although the number of correct annotated uniques is diminishing and no hit uniques are rising, the number of correct annotated ESTs comprised by these uniques is also rising at the same time. There is a lower number of contigs represented by a great number of ESTs annotated correctly, against a higher number of contigs represented by a fewer number of ESTs with no hit annotations. Results were similar to Ath and Hsa (data not shown, see supplementary in www.biodados.icb.ufmg.br).

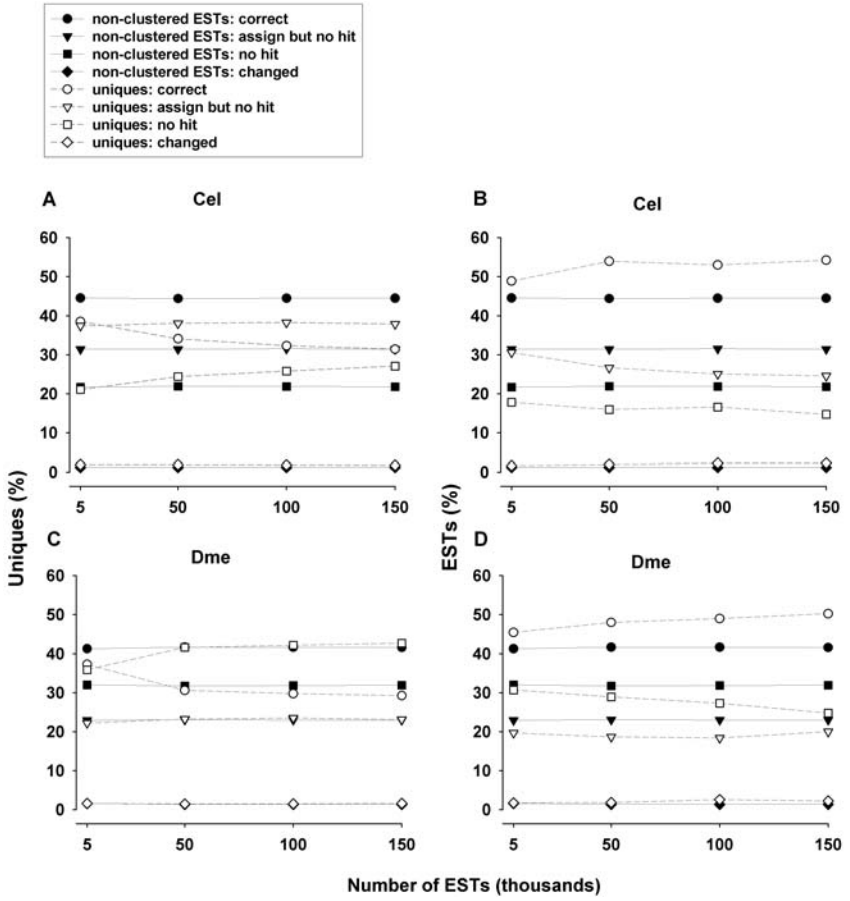


Fig. 3. Comparison of the annotation of uniques (*A and C*) and the ESTs (*B and D*) comprised by these uniques (*white symbols)* with non-clustered ESTs (*black symbols*). Increasing numbers of ESTs were used (*5K, 50K, 100K and 150K*). Annotation results are shown in percentage by symbols (*circle:correct; inverted triangle:assigned but no hit; square:no hit and lozenge:changed annotations*).

In conclusion, this work showed that EST clustering with the software TGICL approaches saturation at around 80% by using over 50K ESTs. Furthermore, we demonstrated a method to evaluate the annotation of uniques generated by TGICL and its ESTs with the KOG database. Results showed, by comparison with non-clustered ESTs, that clustering ESTs with TGICL improved annotation, by diminishing assigned but no hit and no hit annotations (from 0.67 to 0.87 fold) and rising correct annotated ESTs - around 1.2 fold - for both organisms. Thus, a 20% improvement of correct annotation is attained by EST clustering and assemblage. Changed annotation is only slightly augmented (up to 1.08% of all ESTs). Annotation had better improvement as the number of input ESTs to TGICL was increased up to 150K.

Although the clustering of ESTs into contigs is expected to yield a gain in accuracy, this issue has not yet been investigated since the proper positive control was not available. That was possible using KOG clusters. The evaluations presented here indicate that the clustering of ESTs improve the accuracy of annotation for the user as the project generates large amounts of ESTs, thus justifying the analysis of individual ESTs as they are generated if the goal is to produce short amount of data (e.g. under 10K ESTs). TGICL as well as Unigene approach, by clustering ESTs with a BLAST search prior to Cap3 assemblage, improves the scalability of the process, since only increased clusters are subject to novel rounds of Cap3 assemblage.

Acknowledgements
Research supported by CAPES, FAPEMIG and CNPq/MCT.

References

1. Wheeler, D. L., et al.: Database Resources of the National Center for Biotechnology. Nucl Acids Res 31 (2003) 28-33
2. Adams, M.D. *et al*. Complementary DNA sequencing: Expressed sequence tags and human genome project. Science. 252 (1991) 1651−1656
3. Zhang, Z., Schwartz, S., Wagner, L., Miller, W. J.: A greedy algorithm for aligning DNA sequences. Comput Biol. 7(2000) 203-14
4. Altschul, S. F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. 215 (1990) 403-410
5. Huang, X., Madan, A.: CAP3: A DNA sequence assembly program. Genome Res. 9 (1999) 868-77.
6. Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L., Quackenbush, J.: An optimized protocol for analysis of EST sequences. Nucleic Acids Res. 28 (2000) 3657–3665
7. 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. 33 (2005) D71-4
8. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., et al.: The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4 (2003) 41.

5.6 Caracterizando a expressão / amostragem gênica com a base KOG

A base KOG provê uma classificação funcional própria para seus aglomerados de seqüências de proteínas (*clusters*). Essa classificação foi usada na caracterização funcional das EST dos quatro organismos usados, como mostrado no artigo de número 5. Além da caracterização funcional, foram feitos alguns ensaios comparando o número de genes mais e menos expressos entre os quatro organismos e combinações. Um tema semelhante será abordado no artigo 6, onde foi feita uma comparação entre genes diferencialmente expressos entre os quatro organismos e combinações. Porém o tema mais importante do artigo 5 sem dúvida é a abordagem da cobertura da base KOG por seqüências EST. A base KOG possui um viés para aglomerados representando proteínas *housekeeping*, conservadas em vários organismos modelo, alguns bem distantes evolutivamente. Pode-se dessa forma usar a cobertura da base KOG por EST de organismos modelo como parâmetro para se estimar um número mínimo de seqüências EST a serem seqüenciadas em novos projetos transcriptoma, que abrangessem essas proteínas. Todavia, uma amostragem irreal é obtida quando as proteínas do organismo cognato da EST estão presentes na pesquisa (artigo 5, figura 5A). Esta situação somente seria obtida para um organismo proximamente relacionado a algum presente na base de dados, *M. musculus*, por exemplo. Uma estimativa do processo de descoberta gênica mais realista é mostrado neste trabalho utilizando apenas as proteínas de organismos não cognatos para a descoberta (artigo 5, figura 5B). Embora cada categoria funcional apresente uma cobertura diferente (artigo 5, figura 6), percebe-se que a descoberta de toda a base KOG não é completa nem com um número bastante expressivo de EST (artigo 5, figuras 4, 6 e 7), embora uma cobertura bastante relevante seja alcançada com cerca de 150 mil EST.

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

A picture of gene sampling/expression in model organisms using ESTs and KOG proteins

Maurício de Alvarenga Mudado and José Miguel Ortega

Laboratório de Biodados, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Pampulha, Caixa Postal 486, 31270-010 Belo Horizonte, MG, Brasil
Corresponding author: J.M. Ortega
E-mail: miguel@ufmg.br

Genet. Mol. Res. 5 (1): 242-253 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

ABSTRACT. The expressed sequence tag (EST) is an instrument of gene discovery. When available in large numbers, ESTs may be used to estimate gene expression. We analyzed gene expression by EST sampling, using the KOG database, which includes 24,154 proteins from *Arabidopsis thaliana* (Ath), 17,101 from *Caenorhabditis elegans* (Cel), 10,517 from *Drosophila melanogaster* (Dme), and 26,324 from *Homo sapiens* (Hsa), and 178,538 ESTs for Ath, 215,200 for Cel, 261,404 for Dme, and 1,941,556 for Hsa. BLAST similarity searches were performed to assign KOG annotation to all ESTs. We determined the amount of gene sampling or expression dedicated to each KOG functional category by each model organism. We found that the 25% most-expressed genes are frequently shared among these organisms. The KOG protein classification allowed the EST sampling calculation throughout the glycolysis pathway. We calculated the KOG cluster coverage and inferred that 50 to 80 K ESTs would efficiently cover 80-85% of the KOG database clusters in a transcriptome project. Since KOG is a database bi-

ased towards housekeeping genes, this is probably the number of ESTs needed to include the more commonly expressed genes in these organisms. We also examined a still unaddressed question: what is the minimum number of ESTs that should be produced in a transcriptome project?

Key words: EST, Transcriptome projects, KOG, COG, Annotation

INTRODUCTION

The expressed sequence tag (EST) is an instrument of gene discovery. Although it bears around 3-4% sequencing errors (Hillier et al., 1996), this tag suffices for identification of orthologous genes in other organisms through homology searches, thus providing functional annotation of the EST and also demonstrating the presence of the gene in the organism and/or developmental stage of interest (Adams et al., 1991; Faria-Campos et al., 2003). Large numbers of ESTs, coming from independent cDNA libraries, can be used to estimate gene expression (Lee et al., 1995; Franco et al., 1997; Ewing et al., 1999). EST occurrence also allows a gene sampling estimate in novel transcriptome projects.

The total number of EST sequences deposited in public databases has grown considerably (see http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Transcriptome projects are still a good alternative to genome projects as they are less expensive and generate information about gene expression (reviewed by Lindlof, 2003). Large numbers of EST sequences are being produced, although there are still some questions to be answered regarding this approach, such as: i) Is transcript redundancy really directly connected to gene expression? ii) How many cDNA libraries from different tissues and developmental stages are needed to yield a complete picture of an organism's transcriptome? iii) What is the minimum number of ESTs that need to be produced in a transcriptome project in order to give a good representation of the most ubiquitous genes (e.g., housekeeping genes)? We evaluated this last point.

An initial approach to answer this question is to perform automatic transcriptome annotation using secondary databases, where ESTs are annotated by similarity to characterized proteins. The KOG database (eukaryotic representatives of the COG database - Tatusov et al., 2001 and 2003) is one of the many secondary databases, such as GOA/UniProt and KEGG (Kanehisa and Goto, 2000; Camon et al., 2003) that have sequences classified into functional categories and groups, and can be used for this kind of study. The KOG database contains 24,154 proteins from *Arabidopsis thaliana* (Ath), 17,101 from *Caenorhabditis elegans* (Cel), 10,517 from *Drosophila melanogaster* (Dme), and 26,324 from *Homo sapiens* (Hsa). KOG proteins are clustered by function so it is plausible to assume that the KOG database is biased towards ubiquitous clusters - genes that are simultaneously present in at least three model organisms among the seven composing the database.

In order to estimate the efficiency of gene sampling in transcriptome projects, BLAST similarity searches were conducted using ESTs and KOG proteins from these four organisms, requiring different similarities for different organisms. We previously determined that similarity cutoffs should be 78% for Dme and Cel, 80% for Hsa and 84% for Ath (Mudado et al., 2005).

Briefly, experiments were conducted with either pUC18 sequence reads or ESTs; the alignment of these sequences to their respective edited nucleotide sequences was selected based on an identity cutoff of 96% (since single-pass reads may bear up to 4% errors); the similarity cutoff determined for their alignments to the respective edited amino acid sequences was over 80%.

The KOG database allows comparison of all the gene samples of the organisms by functional categories, or with respect to a specific pathway, such as glycolysis, since enzymes that compose the pathways have already been classified in the database. In addition it is possible to compare genes sampled simultaneously from one up to all organisms present in the database.

Public EST databases and secondary databases are depositories of novel and growing information that can be extracted with appropriate bioinformatics approaches. We made use of dbEST (Boguski et al., 1993) and KOG databases, from the NCBI, for sampling of KOG genes within the transcriptome collection available for four model organisms. This approach allows one to predict the number of EST sequences that need to be produced in order to represent the genes present in KOG. We also estimated the minimum number of reads necessary in a novel transcriptome project.

MATERIAL AND METHODS

BLAST

The EST sequences were downloaded from the dbEST database by May 2003. All KTL (KOG, TWOG and LSE) proteins and KOG-conserved domains were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/pub/COG/KOG/) and were in the “kyva” file. All ESTs and proteins that were used are shown in Table 1.

BLAST software (version 2.2.8) was used to search EST sequences against a KOG database depleted of conserved domains. We selected only the 88,613 classified KTL proteins found in the “kog”, “twog” and “lse” files at the KOG site, for use in the BLAST searches as queries. The KTL proteins were divided into 60,758 KOG, 4,451 TWOG and 23,404 LSE proteins. The subjects of the BLAST searches were the organisms’ ESTs. tBLASTn was used with the following parameters: -m 8 -b 10e6 -e 1e-10 -F f. These parameters activate the tabular output of BLAST, allowing up to 10 million hits to one protein (default is 250) and deactivates the low-complexity filter, respectively. The low-complexity filter was deactivated in order to allow tBLASTn to achieve 100% identity in the alignments.

Data processing

All BLAST results, obtained in tabular output (m8 option) with an e-value cutoff of 10^{-10}, were added to an MySQL (version 3.23.58) database, which was populated with all the information related to the KOG database, including the functional classification assigned to each protein. When necessary, PERL (version 5.8.0) scripts were generated to solve computational problems.

The sampling/expression was defined by the number of ESTs that were a best hit in the BLAST searches and the number of hits per KOG functional category was counted. The best scores were always selected to avoid assigning any given EST to more than one protein. Simi-

larity cutoffs of 78% for Cel and Dme, 80% for Hsa and 84% for Ath were used, as previously described (Mudado et al., 2005).

To create random EST datasets, PERL scripts were created with the Math::Random package (thanks to John Venier and Barry W. Brown) downloaded from CPAN (http://www.cpan.org). Ten thousand up to 150,000 ESTs were selected in increasing rounds of 10,000 EST selections. Every round was repeated 10 times in order to obtain the sampling error. All EST selections were made in an independent manner (in every round the ESTs were reselected from the database).

The KOG coverage was calculated in two different manners: cluster coverage (Figures 5 and 6) and protein coverage (Figure 7). A KOG cluster was assumed "covered" when at least one protein from that cluster had a hit to an EST sequence. On the other hand, protein coverage was stricter, as it demanded that all KOG proteins from one cluster had hits in order to yield 100% coverage. The KOG coverage was calculated using only KOG clusters that represent genes of at least three model organisms. TWOGs and LSEs were not included in incremental experiments (Figures 5 to 7), since they contain too many non-categorized proteins (functional category X), which are not well annotated (unnamed proteins) and are more organism-specific than KOGs (see supplementary Tables S2 and S3 at http://biodados.icb.ufmg.br). However, experiments including TWOGs and LSEs are shown in supplementary Figures S4 to S6. Supplementary material is available at Laboratório de Biodados, UFMG (http://biodados.icb.ufmg.br). To perform the cluster coverage and protein coverage experiments, 4,597, 3,285, 4,235, and 4,351 KOG clusters were selected, which contained 19,039, 13,744, 10,581, and 8,445 proteins from Hsa, Ath, Cel, and Dme, respectively.

Statistical analysis

Data were reported as means ± SEM (standard error of the mean). All *t*-tests performed were unpaired, with 50 degrees of freedom.

RESULTS AND DISCUSSION

Gene sampling

A set of ESTs corresponding to all ESTs available for Ath, Cel and Dme (by May 2003; Table 1) was downloaded. About 10 times more ESTs from Hsa were downloaded (almost 2 million), which was assumed to be a sufficiently large sample to obtain a precise analysis of Hsa gene sampling. Large sets of ESTs tend to dilute biases in cDNA libraries (e.g., more libraries from a specific organ or specific time of development) and redundant sequence deposits in dbEST. Conversely, EST production driven by sequencing centers is balanced to cover the transcriptome. However, the term 'gene expression' was avoided and substituted by the term 'gene sampling', as the main goal was not the EST per gene index, but the probability of gene discovery in a transcriptome project. All sampling data are available as part of K-EST: the KOG expression/sampling tool (http://biodados.icb.ufmg.br/K-EST/, Mudado et al., submitted).

Individual sampling lists were generated for each of the four organisms, resulting in sampling profiles by functional category (Figure 1). This methodology tends to depict, by the amount of gene sampling, how each organism differentially produces transcripts related to the

Table 1. Numbers of sequences used for comparing gene expression.

Organisms	ESTs	KTLs	KTL proteins	KOGs	KOG proteins
Arabidopsis thaliana	178,538	4,872	24,154	3,285	13,744
Caenorhabditis elegans	215,200	5,306	17,101	4,235	10,581
Drosophila melanogaster	261,404	5,145	10,517	4,351	8,445
Homo sapiens	1,941,556	6,572	26,324	4,597	19,039

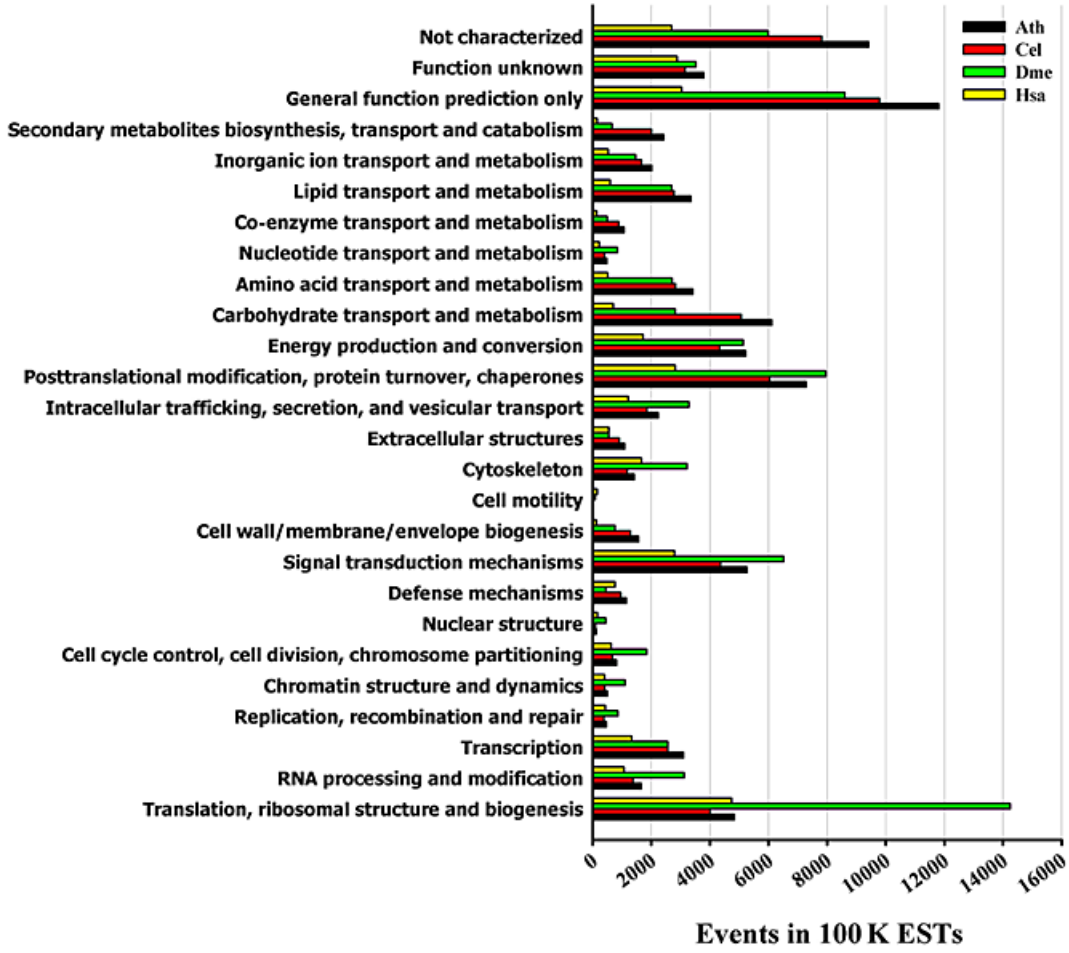


Figure 1. Gene sampling using KOG functional categories. The black, red, green, and yellow bars represent *Arabidopsis thaliana* (Ath), *Caenorhabditis elegans* (Cel), *Drosophila melanogaster* (Dme), and *Homo sapiens* (Hsa), respectively.

various types of biological processes. When doing the BLAST searches for gene sampling of the organisms, the other proteins were preserved in the query, as the best hits were almost totally composed of the cognate organism proteins. Most of the proteins of Cel, Ath and Hsa annotated their own ESTs (less than 0.5% of the annotation was from the proteins of other

organisms). Only Dme appears to have had a more considerable cross-annotation, since around 3.8% of other organisms' KOG proteins were used to annotate its ESTs.

The differences in sampling of genes amongst organisms were analyzed. In Figure 2A, the 25% most and least expressed genes from the set of 2,523 KOG genes common to the four organisms were examined to determine the proportion of sharing involving most/least categories. As expected, all four eukaryotes shared 50-62% of the genes in the 25most category, while sharing 36-40% of the genes in the 25% least sampled set of genes, indicating that the more frequently sampled genes are shared more often amongst these organisms ($P < 0.05$). Evolutionary distance seems to count, since Dme and Cel share more genes per category than when they are compared with Ath and Hsa. The next step was to compare the set of KOG genes common to the four organisms, among all organisms for the two categories (25most and 25least). Figure 2B shows that sharing genes is more common in the 25most category. The 25least category produced the inverse situation, as expected.

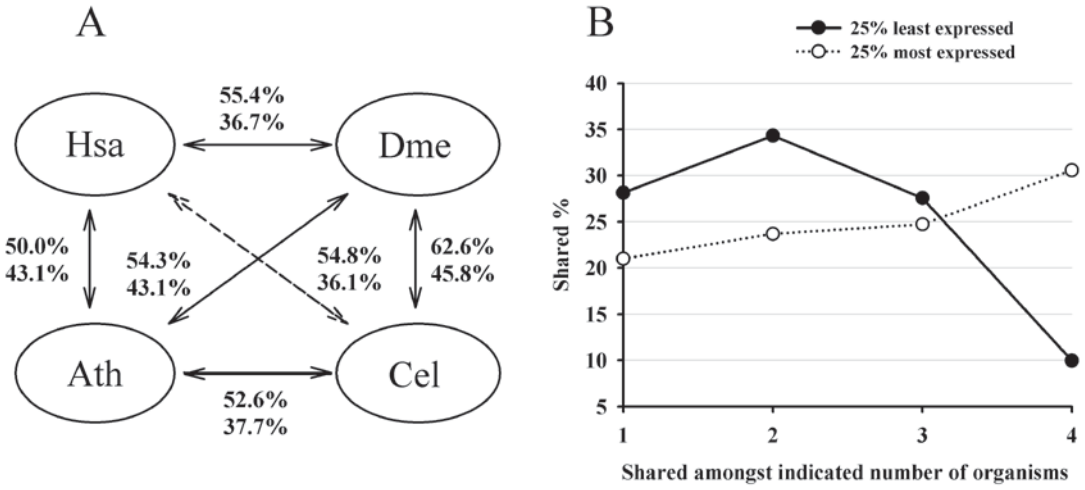


Figure 2. A. Comparison of the 25% most and least expressed genes for all four organisms. The upper and lower numbers of all tuples represent the 25% most and least expressed genes, respectively. **B.** Global comparison of the 25% most and least expressed genes among all four organisms. Hsa = *Homo sapiens*; Dme = *Drosophila melanogaster*; Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*.

The KOG database allows a more detailed analysis of gene sampling, since all genes are described and classified individually. Enzyme sampling within the glycolysis pathway was examined as an example. Figure 3 illustrates the similarities in the pathway of the four organisms. GAPDH was highly sampled in all four eukaryotes, followed by fructose biphosphate aldolase.

KOG coverage

The global cluster coverage was calculated for all functional categories in order to determine the intensity at which the KOG database clusters were covered by all organisms' ESTs. KOGs, TWOGs and LSEs specific for every organism were selected for this objective.

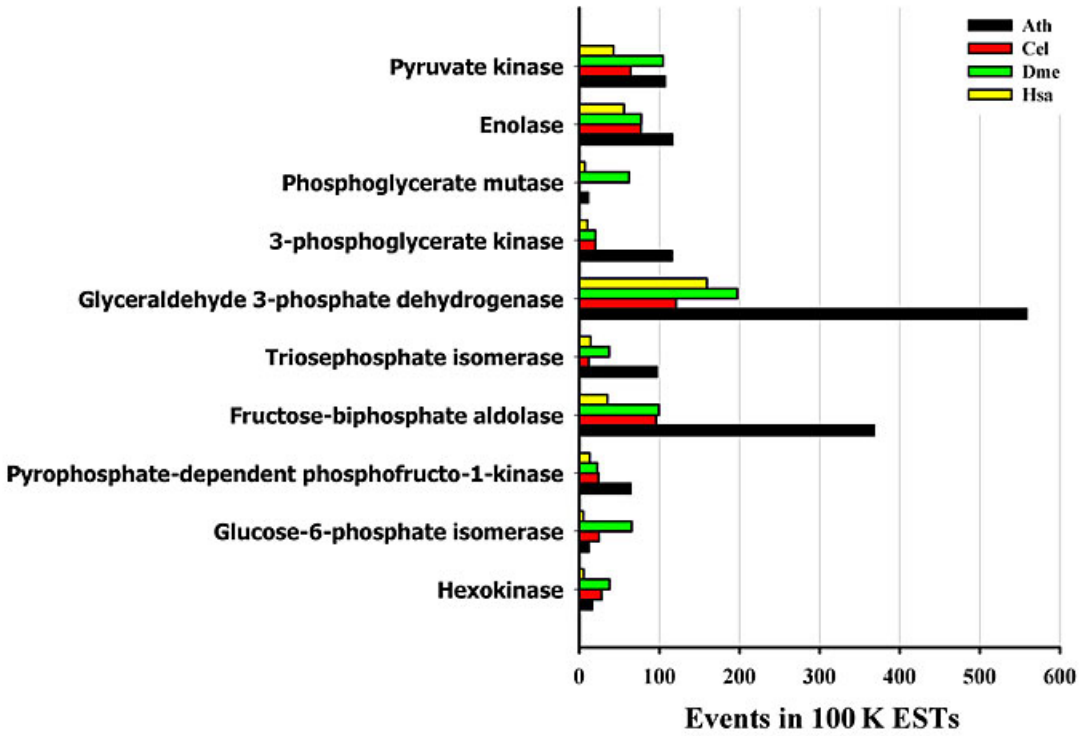


Figure 3. Sampling of glycolysis pathway enzymes. Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*; Dme = *Drosophila melanogaster*; Hsa = *Homo sapiens*.

The number of ESTs used appears to cover more than 95% of the KTL clusters of the four organisms (Figure 4).

In order to estimate an optimal number of ESTs that are sufficient to fully cover the KOG clusters, transcriptome projects with different sizes were simulated by creating EST pools, selected at random, from the EST databases of each organism. We selected only KOG entries (genes present in at least three model organisms) specific for each organism. TWOGs and LSEs were discarded since they represent genes that are more organism-specific. Two distinct experiments were executed. First, by maintaining the KOG proteins from all organisms in the database, curves of coverage tended to saturate (Figure 5A). Sets from 10 to 150 K ESTs were generated, with 10 repetitions in order to obtain the sampling errors. Saturation of the coverage curve was expected to occur since all ESTs probably have their correlated proteins in the database. A similar coverage result was expected when annotating novel ESTs, using databases that contain proteins with high similarity to the query organism (closely related organisms). The cluster coverage rose exponentially when using 10 to 80 K ESTs and then increased in a linear pattern. We suggest that 50 to 80 K is a reasonable minimum number of ESTs to be produced in order to obtain around 80-85% of the genes that are common among organisms, such as the genes represented by KOG. Also, all organisms but Hsa required at least 60 K ESTs to cover 80% of KOG clusters. *Homo sapiens* ESTs had a different behavior and showed less coverage

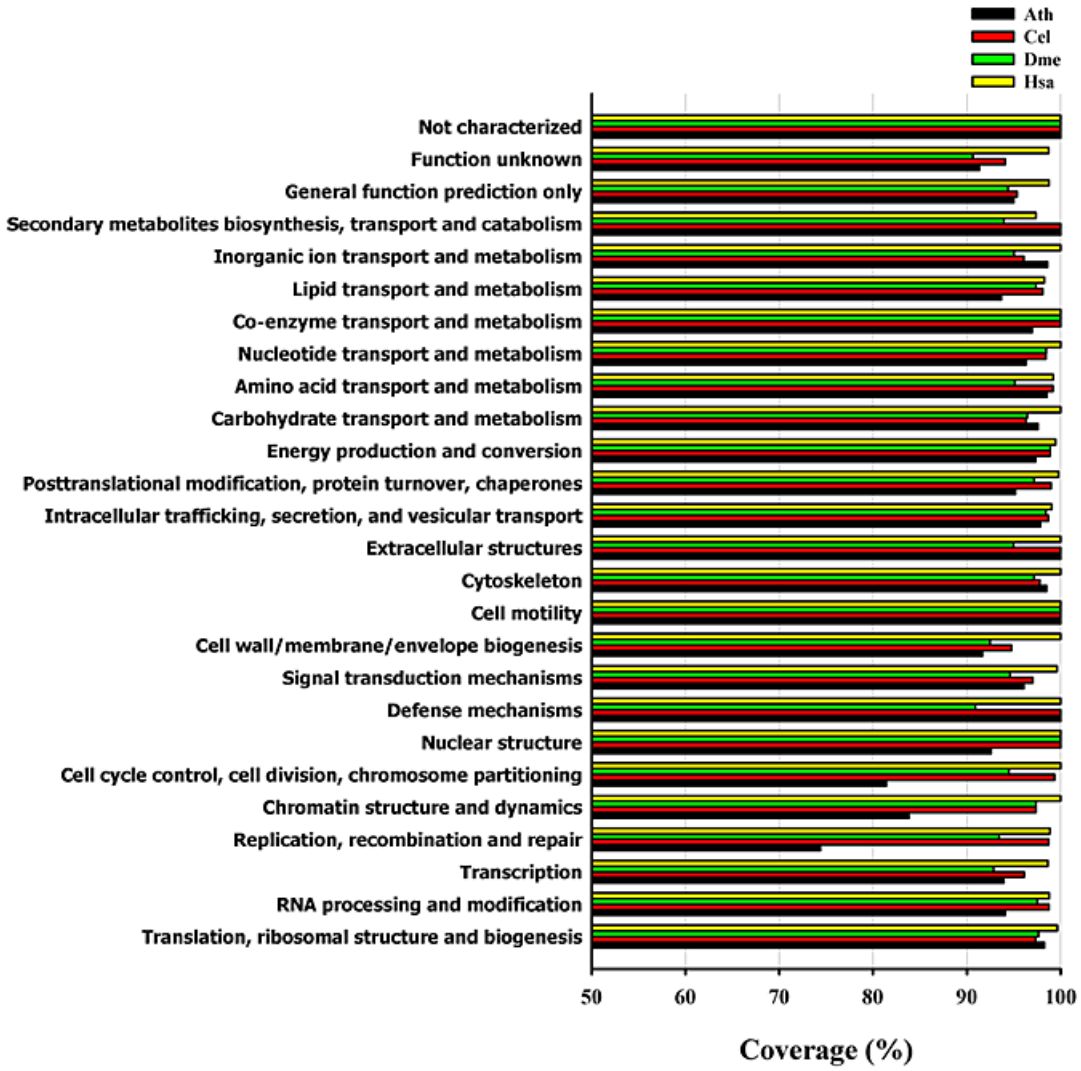


Figure 4. Coverage of KOG database (KOG, TWOG, LSE) by ESTs of the four eukaryotes (178,538, 215,200, 261,404, and 1,941,556 for Ath, Cel, Dme and Hsa, respectively). Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*; Dme = *Drosophila melanogaster*; Hsa = *Homo sapiens.*

potential, probably because the database contained shorter sequences (see supplementary Table S1 at http://biodados.icb.ufmg.br). In the second experiment (Figure 5B), the organisms' proteins were removed from the database when the cognate organisms' ESTs were annotated (e.g., ESTs from Dme would be annotated only by the other organisms' proteins, except Dme proteins). As expected, less effective coverage (around 10-20% loss compared to Figure 5A) was obtained. Figure 5B shows that Ath proteins were less efficiently covered by the incremental EST collections than Ath clusters. This behavior was expected, as it is the only plant in the database. Dme and Cel seemed not to be as affected in this second experiment, probably because their proteomes are relatively more related.

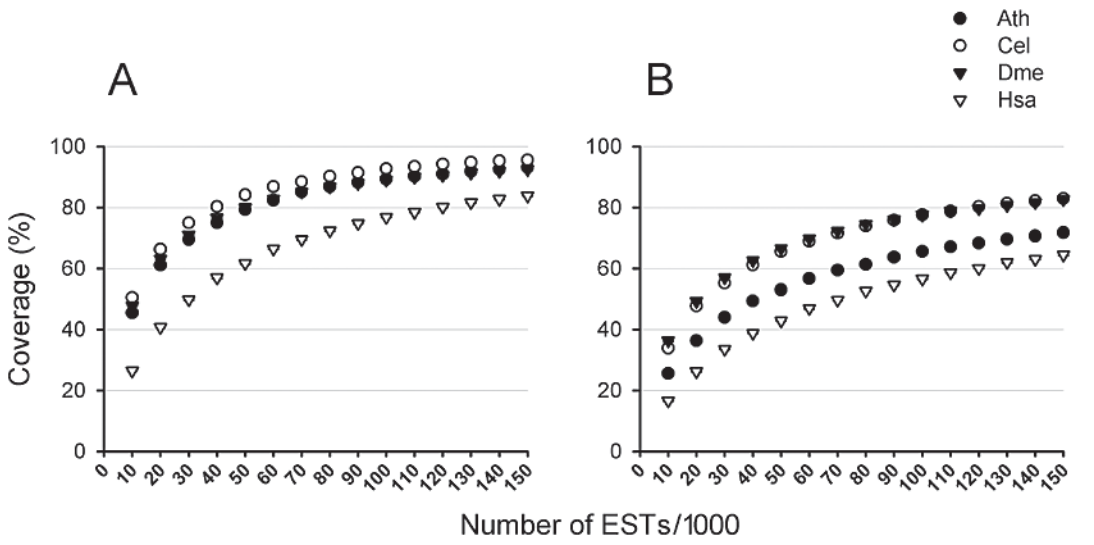


Figure 5. KOG coverage calculated by using 10 to 150 K ESTs (N = 10) from the four eukaryotes. **A.** Annotation with all proteins. **B.** Annotation with the cognate proteins depleted from the database. The standard error of the mean was under 1% in all events. Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*; Dme = *Drosophila melanogaster*; Hsa = *Homo sapiens*.

Figure 6 shows the same data from the last experiment, but with coverage distributed by KOG functional category. As seen in Figure 6A and B, Cel, Dme and Ath had similar patterns of KOG functional category coverage when using 10, 50, 100, and 150 K ESTs. By using unpaired *t*-tests with the 50 and 100 K coverage data, all organisms gave P values lower than 0.01. When the same test was run with the 100 and 150 K categories, Cel, Dme and Ath gave P values above 0.05. It is possible that after the exponential phase of the coverage curve (Figure 5A and B) producing more ESTs would be less effective in covering the KOG clusters, as these two coverage sets (100 and 150 K) were not statistically different. Hsa did not show this characteristic, as expected by the latter experiment. 'Cell Motility' and 'Not categorized' categories gave larger error bars since they were composed of small numbers of clusters (see supplementary Table S3).

The coverage of proteins from an organism, considering only KOG proteins (not TWOG or LSE), was analyzed (Figure 7). Figure 7A shows the protein coverage by the cognate organisms' ESTs. Less than 50% of all proteins were covered by using up to 150 K ESTs. This happens because only a few paralog representatives of the clusters were being preferentially sampled (data not shown). Moreover, the coverage seemed to saturate when more ESTs were used, with the same exponential phase, followed by a linear plateau (Figure 5A). As little as 20% of the Ath KOG proteins were being covered, probably because of the large number of duplicated genes in this plant; few of them were sampled by the ESTs (data not shown). Hsa again covered fewer proteins than for the other organisms, possibly due to the smaller size of the ESTs added to putatively larger 3' and 5' UTRs. Figure 7B shows the quantity of KOG proteins covered when the species from which the ESTs originated was excluded. Comparing Figures 5B and 7B, it appears that no more than 6% of the non-cognate organisms' proteins

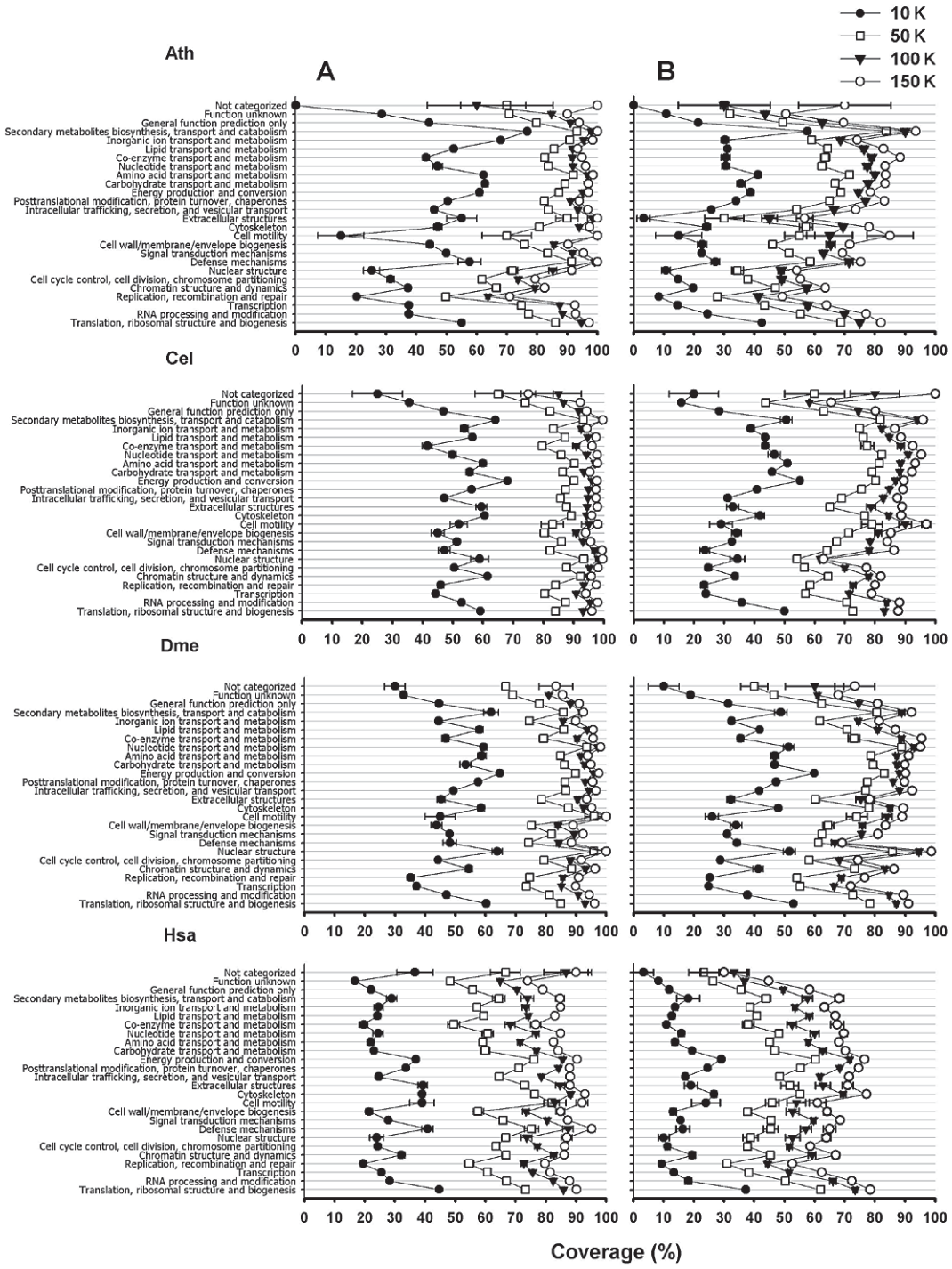


Figure 6. Coverage of KOG functional categories by using 10, 50, 100, and 150 K ESTs (N = 10) from the four eukaryotes. **A.** Annotation with all proteins. **B.** Annotation with the cognate proteins depleted from the database. Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*; Dme = *Drosophila melanogaster*; Hsa = *Homo sapiens.*

cover up to 60-80% of the KOG clusters. This shows that the KOG database does not entirely depend on the presence of proteins of the same organism, from which the ESTs originate, to annotate its ESTs, if KOG clusters are used instead of individual protein entries.

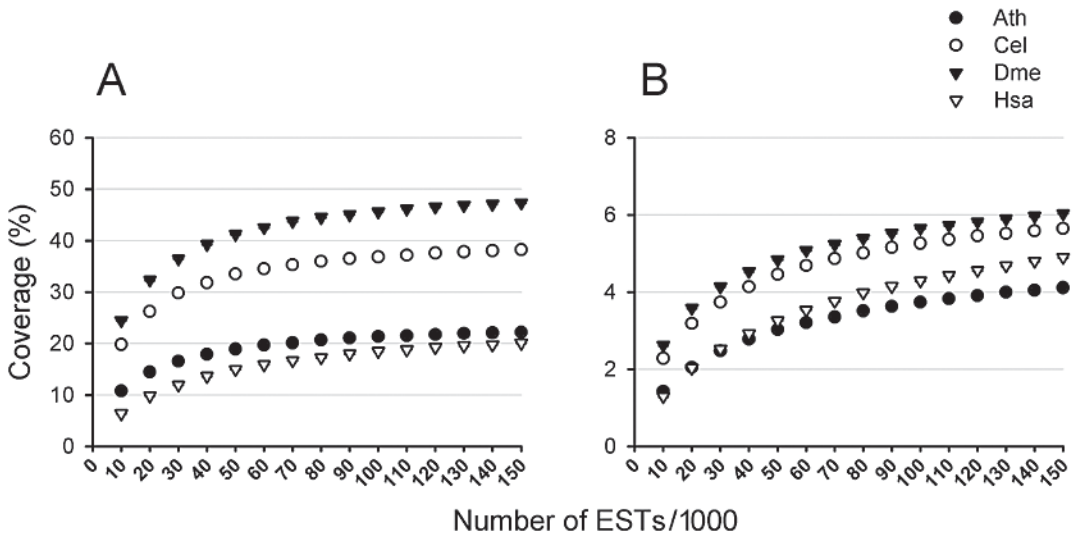


Figure 7. KOG protein coverage calculated by using 10-150 K ESTs (N = 10) from the four eukaryotes. **A.** Database including all proteins. **B.** Database without the cognate proteins. The standard error of the mean was under 1% in all cases. Ath = *Arabidopsis thaliana*; Cel = *Caenorhabditis elegans*; Dme = *Drosophila melanogaster*; Hsa = *Homo sapiens.*

CONCLUSIONS

The KOG protein coverage and cluster coverage results give a good indication on how to conduct a transcriptome project efficiently. Researchers can make predictions on how many ESTs should be produced in order to determine the genes that are most and least commonly expressed in other species. Results presented by each KOG entry can be accessed online in K-EST.

The study of sampling/expression of genes by ESTs with secondary databases generates answers to questions such as how many reads a transcriptome project should generate to cover a reasonable number of genes, or how frequently specific genes are expected to be sampled within a transcriptome project, given their sampling in model organism transcriptomes.

As the KOG database grows and incorporates more organisms, broader answers may be generated to these questions. In addition, new secondary databases are being developed, with more sequences and different organisms. The UniProt database (Bairoch et al., 2005) is such an example that we are currently investigating.

ACKNOWLEDGMENTS

Research supported by CAPES, FAPEMIG and CNPq/MCT.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.

Bairoch A, Apweiler R, Wu CH, Barker WC, et al. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res*. 33: D154-D159.

Boguski MS, Lowe TM and Tolstoshev CM (1993). dbEST-database for "expressed sequence tags". *Nat. Genet*. 4: 332-333.

Camon E, Barrell D, Brooksbank C, Magrane M, et al. (2003). The Gene Ontology Annotation (GOA) project: Application of GO in SWISS-PROT, TrEMBL and InterPro. *Comp. Funct. Genom*. 4: 71-74.

Ewing RM, Ben KA, Poirot O, Lopez F, et al. (1999). Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res*. 9: 950-959.

Faria-Campos AC, Cerqueira GC, Anacleto C, de Carvalho CM, et al. (2003). Mining microorganism EST databases in the quest for new proteins. *Genet. Mol. Res*. 2: 169-177.

Franco GR, Rabelo EM, Azevedo V, Pena HB, et al. (1997). Evaluation of cDNA libraries from different developmental stages of *Schistosoma mansoni* for production of expressed sequence tags (ESTs). *DNA Res*. 4: 231-240.

Hillier LD, Lennon G, Becker M, Bonaldo MF, et al. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res*. 9: 807-828.

Kanehisa M and Goto S (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28: 27-30.

Lee NH, Weinstock KG, Kirkness EF, Earle-Hughes JA, et al. (1995). Comparative expressed-sequence tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. *Proc. Natl. Acad. Sci. USA* 92: 8303-8307.

Lindlof A (2003). Gene identification through large-scale EST sequence processing. *Appl. Bioinformatics* 2: 123-129.

Mudado MA, Bravo-Neto E and Ortega JM (2005). Tests of automatic annotation using KOG proteins and ESTs from 4 eukaryotic organisms. *Lecture Notes Computer Sci*. 3594: 141-152.

Mudado MA, Barbosa-Silva A, Torres JA, Paula-Pinto S, et al. K-EST: KOG Expression Sampling Tool. *Bioinformatics* (submitted).

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, et al. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*. 29: 22-28.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.

5.7 K-EST: uma ferramenta para comparação de amostragem de EST de organismos modelo em KOG

O K-EST, assunto do artigo de número 6 desta tese, é uma ferramenta desenvolvida para disponibilizar parte dos dados obtidos nos artigos anteriormente discutidos. Se as coleções de EST dos quatro organismos (Ath, Cel, Dme e Hsa) forem encaradas como resultado de grandes projetos transcriptoma únicos, o alinhamento dessas EST com as proteínas KOG poderá ser usado para estimar a expressão gênica ou pelo menos de amostragem de EST nesses projetos e, mais importante, em projetos novos. Utilizamos uma normalização padrão para cem mil EST de modo a permitir uma extrapolação para outros projetos transcriptoma de outros tamanhos. Com isso permitimos que projetos transcriptoma em andamento possam ser analisados com base em coleções conhecidas de EST e proteínas KOG. O K-EST é dividido basicamente em duas páginas diferentes, uma chamada *sampling,* que é a disponibilização dos dados de expressão gênica explicados acima, e a página de *conservation*. Esta última, provê uma informação derivada do teste de anotação discutido nos artigos 1, 2, 3, 4 e 5. Compara-se o resultado de designação de EST de um organismo (por exemplo, Dme) para suas proteínas, com a anotação dessas mesmas EST para as proteínas dos outros organismos (no caso Ath, Cel, Hsa, Sce, Spo e Ecu) . Se as proteínas que representam um dado KOG, presentes nos outros organismos, conseguirem anotar o mesmo número de EST que foram designadas para proteínas de Dme, do mesmo KOG, dizemos que essa proteína é muito conservada. Se o número de EST recuperado é baixo, dizemos que a proteína é pouco conservada. É evidente que entradas KOG para as quais seja designada uma quantidade alta de EST e simultaneamente estas sejam anotadas com alta eficiência por outros organismos, devem ter também um grande número de EST em organismos novos, exceto se este KOG representar um gene não expresso no novo organismo. Os dados das páginas de *conservation* e *sampling* foram usados na tentativa de prever genes que não existem em um dado organismo. Assim. agrupamentos KOG com amostragem de EST e conservação de anotação elevadas em diversos organismos modelo simultaneamente, representam teoricamente proteínas muito expressas e conservadas. Se um dado agrupamento KOG com essas características não for amostrado em algum outro organismo, então há uma possibilidade desse gene ter sido perdido durante o processo evolutivo, e por isso nenhuma EST é amostrada. Essa hipótese é testada no final do artigo 6

e no artigo 7. Esse método se mostrou aparentemente eficaz na predição de genes que não existem em um organismo. Outras funcionalidades interessantes do K-EST é o uso de estatísticas para inferência da probabilidade de um dado KOG ser diferencialmente expresso em um dado organismo ou em combinações dos quatro organismos. Para isso usamos a estatística descrita por (Stekel *et al.*, 2000). Usou-se para esse fim a coleção inteira de EST de cada organismo como uma única biblioteca de cDNA. Apesar de ser um método inovador, a comparação da diferença de expressão entre os quatro organismos, e combinações, gerou resultados com possíveis significados biológicos. Um exemplo é a indicação de uma menor divergência entre Cel e Dme do que entre outras combinações de organismos.

Outras funcionalidades, como buscas por indentificadores KOG, BLAST para seqüências do usuário, perguntas ao banco de dados e uma página de ajuda também foram construídos.

O K-EST, ferramenta desenvolvida para a *web,* usando ferramentas livres como PHP, Apache, MySQL, Linux e BLAST, além de seqüências públicas de EST e proteínas KOG pode ser utilizada no endereço http://www.biodados.icb.ufmg.br/K-EST. Está em andamento um projeto para disponibilizar uma versão para instalação local, assim como para inserção de dados de amostragem de EST de organismos exteriores ao KOG.

K-EST: KOG Expression / Sampling Tool

Maurício A. Mudado, Adriano Barbosa-Silva, Gabriel R. Fernandes, Saulo A. Paula-Pinto, João Torres and J. Miguel Ortega
Universidade Federal de Minas Gerais, Avenida Antonio Carlos 6627, Belo Horizonte, MG, Postal Code 486, Brazil

ABSTRACT

Summary: K-EST is a web-based application that shows the sampling of large sets of ESTs from four model organisms with the KOG database. K-EST may be used as a tool to predict the EST sampling or, roughly, gene expression in novel transcriptome projects by comparison to the four model organism expression / sampling of KOG entries. K-EST uses statistical methods to analyze differential expression between organisms and internal cDNA libraries. The expression / sampling may be normalized by constitutively expressed transcripts: actin, GAPDH and KOG0052. Other important feature of K-EST is to show the conservation between genes of the four organisms used, represented by the fraction of EST from a given KOG that are sampled by the KOG entries from other organisms.
Availability: http://biodados.icb.ufmg.br/K-EST
Contact: miguel@icb.ufmg.br

1. INTRODUCTION

Transcriptome projects result in a collection that is supposed to sample the sequences from cell transcripts and are frequently used as a complement to genome projects as a support for gene mapping. Generally single-pass sequences are produced. The digital representation of these sequences, known as EST or Expressed Sequence Tag, besides bearing up to 4% of sequencing errors, is intended to allow the identification of the codified protein throughout similarity searches of orthologous genes of other organisms [1, 2]. However, when a large number of ESTs originated from several independent cDNA libraries are available, they may be used to estimate gene expression [3, 4]. For the skeptic observer, EST occurrence provides, at least, an estimative of a given chance of gene sampling in an EST-based gene discovery program. At the present, a wealth of information can be mined from large EST collections available in public primary databases. Secondary databases, a repository of curate biological sequences, are a good source of information for annotation of novel sequences such as ESTs. The KOG (Eukaryote Cluster of Orthologs Group) database [5] is an interesting example for this purpose since its protein entries are classified into functional categories and groups. Moreover, KOG congregate into clusters, the proteins that exert analogous function in several model organisms with complete genome sequenced. These clusters are named KOGs, TWOGs or LSEs (when they join together proteins from at least 3, 2 or 1 organism, respectively), and represent one gene or protein that are therefore conserved during evolution time (e.g. enolase, represented by the ID “KOG0047”). In this work we present a tool that enables one to see the sampling of ESTs in four model organisms (Ath - *Arabidopsis thaliana*; Cel – *Caenorhabditis elegans*; Dme - *Drosophila melanogaster* and Hsa – *Homo sapiens*) that were sampled with protein entries from the

KOG database, using BLAST similarity searches [6]. It does not only show differences of expression / sampling between these organisms, but also allows one to evaluate whether some genes would appear or not in a novel transcriptome project by simply comparing EST sampling throughout organisms. In the K-EST conservation pages, the fraction of EST from a given KOG that can be sampled by the KOG entries from other organisms is shown, thus allowing the user to distinguish conserved and not conserved KOG entries and to estimate the actual sampling potential.

2. PURPOSE OF K-EST

K-EST provides two major features that are most informative: sampling and conservation. These features can be used to compare and predict EST sampling of novel transcriptome projects. This is an important strategy, especially for ongoing projects, where one may want to predict the percentage of coverage of certain functional categorized genes (e. g. DNA repair genes) as a function of the EST producing effort. This information can support the inference of the number of EST sequences to be produced. One way of accessing this information is measuring the percentage of KOG coverage (for more information, see [7]). More specifically, researchers seeking for a specific gene can compare and predict its sampling by simply navegating the K-EST sampling page. This page provides the sampling of large sets of EST from four model organisms with KOG clusters and divided by KOG functional categories. Differential expression is shown within each organism cDNA libraries, but also by comparing the whole organism EST sampling for a given KOG or functional category, supporting a comparative analysis between the model organisms.
The lack of sampling of a specific KOG in a transcriptome can be a signal of gene loss in the organism being studied. K-EST provides some parameters that can be used to augment the reliability of sampling and gene loss prediction. The conservation page allows a comparison of the sampling of the EST sequences by any given KOG from a model organism and the fraction of these EST that are annotated with KOG proteins from the other model organisms present in KOG database. High conservation means that ortholog proteins from other organisms show as high similarity to the EST set that had been sampled by the cognate KOG protein. This indicates that they are evolutively conserved. Genes with high conservation are expected to have therefore a more reliable sampling. Taken together, conservation, internal differential expression and differential expression amongst organisms can be used as parameters to verify the reliability of sampling and used to predict sampling and to infer gene loss by lack of expression/sampling in novel transcriptome data.

3. K-EST

The K-EST (KOG Expression / Sampling Tool) is a web-based application that helps researchers to predict EST sampling in novel transcriptome projects. It was developed using PHP (hypertext preprocessor) and a relational database (MySQL). The database was populated with BLAST results (best hits only, 10^{-10} E-value cutoff) from large sets of EST sequences (Table 1), from the four model organisms cited above, queried against proteins from the KOG database. BLAST results were processed with PERL scripts to depict the EST sampling by each KOG entry and or its associated functional category.
K-EST homepage was assembled to allow the user to select the combinations of 1, 2, 3 or 4 organisms studied and its respective expression / sampling within combinations of KOG

functional categories. All expression / sampling data were normalized by 100K ESTs in order to allow the researchers to predict the level of transcript appearance and whether or not a gene would be sampled in novel transcriptome projects of different sizes. The user can also select EST expression / sampling normalized by generally more expressed transcripts like GAPDH, Actin and Translation elongation factor EF-1 (KOG0052), in order to minimize sampling bias.
User can perform a real-time annotation of a query sequence by means of a BLAST search (version 2.3.13, installed locally) against the KOG database that directly reports to the sampling or conservation of the homologous genes.
Conservation pages allow user to compare the sampling of EST against the cognate organism proteins and against the other organism proteins. This procedure allows one to verify how conserved is the chance of sampling of a gene by the complementary organisms in database.

4. EST SAMPLING

To analyze the difference in EST sampling between organisms, two strategies were used. First, the more general and simple method was to calculate the fold of expression, by dividing the highest EST sampling by the lowest. The fold gives an idea of how different is the expression between the selected organisms. The other strategy makes use of a method described by [8]. This method has been developed to investigate differential expression between multiple cDNA libraries. Its output is a single real value, called 'R', in which values above a threshold suggests differential expression between libraries of a given gene. We have adapted this method in order to calculate difference in expression between the collections of EST from the four organisms. Considering the whole set of EST of an organism as a single library is an innovative procedure that appears to produce biological information about expression/sampling of the KOG clusters amongst organisms.
We performed the R calculation with all combinations of four organisms allowing the user to access these values. We also calculated the difference in expression within the cDNA libraries from organisms alone, using the original method, in order to show if a KOG cluster was also differentially expressed in an organism. User can access the expression /sampling in individual libraries. Randomized data was used in order to find the R value threshold (believability) for false positives. R values are shown with different cell colors: green, gold and red representing low (<20%), even (between 20% and 99%) and high (>=99%) probability of differential expression in the webpage tables.

Table 1. Information about the EST (number of sequences), origin of EST (cDNA libraries, development stages, organs, tissues and authors), KOG clusters (KOG/TWOG/LSE) clusters and proteins used in this work.

ORGANISMS	EST	cDNA libraries	Dev. Stages	Organs	Tissues	Authors	KOG	Proteins
Arabidopsis thaliana	360833	15	4	3	10	8	4,872	24,154
Caenorhabditis elegans	293530	10	7	1	2	9	5,306	17,101
Drosophila melanogaster	370672	15	9	8	1	10	5,145	10,517
Homo sapiens	360398	31	6	9	28	8	6,572	26,324

5. CONSERVATION

Since all the organisms used in K-EST are present in the KOG database, we initially performed BLAST searches with proteins and EST sequences originated from the cognate organism (e.g. EST and proteins from Dme), to reveal the actual EST sampling for each organism. Thereafter, we removed the cognate organism proteins from the database and used only the proteins from the other organisms in the BLAST searches (e.g. EST from Dme against proteins from Ath, Cel and Hsa) to show whether the ratio of EST sampling was maintained (thus, KOG hits would be conserved between organisms) or the sampling was diminished (indicating that KOG proteins would be less conserved between organisms, so sampling will suffer from this lack of conservation). This procedure was performed with all organisms and all KOG clusters, so the user can switch the analysis from EST sampling to KOG cluster conservation. The level of conservation is also depicted by colors (green, gold and red, representing above 80%, between 20-80% or lower than 20% of conservation). It is supposed that a direct comparison between each organism proteins might conduct to equivalent results, although we believe that the results shown in this chosen format are more operational.

6. EXAMPLES OF DATA MINED FROM K-EST

Table 2 shows interesting examples of data mined from K-EST that can be used for exploration of novel and ongoing transcriptome projects. The four organisms share less KOG entries evenly expressed (believability lower than 50%) compared to the combinations of 3 or 2 organisms. On the same way, the four organisms share more KOG clusters differentially expressed (believability >=99%). The two organisms that share more KOG evenly expressed and less KOGs differentially expressed are Cel and Dme (42.9% and 17.6% respectively). Interestingly, different expressed KOG are more distinguishable (believability>=99%) whenever the plant is in the combination, compared to combinations of Cel, Dme and Hsa. This information can be an indication that Cel and Dme have a

shorter evolutionary distance between them and that Cel, Dme and Hsa are also closer compared to the plant. Under our knowledgement this is the first initiative of comparison of transcription rates and differential expressed clusters of genes (homologous genes) between organisms. Furthermore, it seems that using the whole set of EST as a single pool , to compare different organisms may constitute a novel procedure that allows the differentiation of organisms by its gene expression/sampling profile.
The categories that appear to have more KOG clusters evenly expressed within the combinations of all organisms are V (Defense Mechanisms), L (Replication, recombination and repair), W (Extracellular structures) and S (Function Unknown). The categories that share more differentially expressed KOG entries are Q (Secondary metabolites biosynthesis, transport and catabolism), Y (Nuclear Structure), B (Chromatin structure and dynamics), C (Energy production and conversion) and W (Extracellular structures). The values of fold agreed with those obtained for R and proved to be a good clue on showing differential expression information.
Because of the great variability of cDNA libraries (see Table 1), as expected, the majority of KOGs showed internal believability >50%. Interestingly, Ath and Hsa seemed to share more KOGs with internal believability <50% (5.6%) on the contrary to the four organisms (0%). Dme and Hsa showed the greatest number of shared conserved (above the 80% index) KOG proteins (28.9%). On the other hand, Ath, Cel and Dme show the lowest index of conserved KOG proteins (8.7%). When organisms are taken individually, the frequency of conserved KOG clusters is 17.7%, 32.2%, 39.3% and 40.9% for, respectively, Ath, Cel, Dme and Hsa (see Table 3). The categories that show more KOG entries highly conserved are J (Translation, ribosomal structure and biogenesis) and C (Energy production and conversion). Conversely, the categories that show less conserved KOGs are W (Extracellular structures), Y (Nuclear Structure) and S (Function Unknown).

PREDICTING LACK OF GENES WITH K-EST

One can use the sampling and annotation information from K-EST conservation page to try to predict if a given organism lacks a gene, by comparing its EST sampling to model organisms sampling. It is important here to take the K-EST conservation index into account. We used Hsa, Cel and Dme sampling and conservation information to predict if the lack of sampling to a given KOG indicates that the organism do not have the proteins present in KOGs. In other words, we aimed to distinguish statistical zero sampling from NO KOG status, based on the sampling ratio by other organisms and conservation. The hypothesis is that if a given organism lacks sampling in a given KOG but other organisms show high sampling to the same KOG, then it is probable that this organism does not have this protein or cluster of proteins represented by a KOG entry. Figure 1, A-C, show KOG clusters that show sampling different from zero for Cel and Dme but not to Hsa (A), and the other two combinations of these three organisms (B and C). The Y axis shows the number of KOG clusters that had no EST hit in the analyzed organism, while the X axis is the exigency of minimum number of EST per 100K that hit the KOG in the other two organisms. As expected, the number of KOG lacking sampling/annotation in one organism is less abundant as the minimum sampling/annotation exigency to the other two organisms is increased (Fig.1 A-C). By determining True Positives (TP) as the KOG entries with zero hits in the analyzed organism and simultaneously not present in it, and False Positives (FP) as KOG with zero hits in the analyzed organism although present in it, the PPV or positive

predictive value (TP/(TP+FP)) can be used to measure the predictability of this hypothesis. Fig.1 D-F (full circles), shows the PPV using the sampling/annotation of the combinations of organisms. It can be seen that raising the minimum hit to the other organisms compared to Cel (Fig. 1D) and Hsa (Fig.1E) did not produce the expected result as the predictability of non-existing KOGs oscillated at 0.4 (Hsa) and 0.8 (Cel), although culminated in PPV of 1 to one gene only, at 27 and 13 minimum hits to other organisms. On the other hand, the predictability of non-existing KOG from Dme (Fig. 1F) rose accordingly to the minimum hits to Hsa and Cel. Dme compared to Cel and Hsa showed PPV of around 0.8 and 0.9 rising to 1 before 13 and 6 minimum hits in the other organisms, respectively. This procedure was able to predict 7 and 23 lacking KOGs with 100% positive rates for Cel and Dme respectively. The use of conservation information from K-EST (Fig.1, open circles) made it possible to augment predictability for Hsa but had little effect in Dme and Cel. Although a lack of a given KOG in an organism could represent the real lack of that gene, problems with the database or low expressed transcripts could lead to errors. It is desirable to further verify the reliability of this information directly in the genome if available, other databases or by ‘wet lab’ experiments.

CONCLUSION

K-EST is a useful tool for researchers participating in novel transcriptome projects. It provides information that can be used for comparison of EST sampling and prediction of gene loss. K-EST is available at http://www.biodados.icb.ufmg.br/K-EST

Table 2. Examples of differential sampling/expression and conservation that exists simultaneously in different combination of organisms: Ath, Cel, Dme and Hsa, respectively represented by A, C, D and H.

Combination of organisms	% KOGs with believab. <50%	% KOGs with believability >=99 %	Cat. with higher % of KOGs with believability <50%* [fold of cat.]	Cat. with higher % of KOGs with Believability >=99%** [fold of cat.]	% KOGs with internal believab. < 50%	% KOGs with conserv. >=80%
1) A C D H	35 / 2523 (1.4%)	2189 / 2523 (86.7%)	V (10.0%) [3.5]	Y (100.0%) [8.2]	0 / 2273 (0.0%)	304 / 2284 (13.3%)
2) A C D -	81 / 2573 (3.1%)	1988 / 2573 (77.3%)	V (10.0%) [3.5]	Q (95.4%) [3.1]	1 / 2350 (4e-4%)	323 / 3713 (8.7%)
3) A C - H	78 / 2702 (2.9%)	2072 / 2702 (76.7%)	V (13.3 %) [1.7]	Y (95.6%) [8.2]	9 / 2432 (0.3%)	343 / 2444 (14.0%)
4) A – D H	84 / 2794 (3.0%)	2108 / 2794 (75.4%)	V (15.4%) [3.5]	Q (100.0%) [7.4]	2 / 2548 (7.8e-4%)	375 / 2548 (14.7%)
5) - C D H	188 / 3952 (4.7%)	2685 / 3952 (67.9%)	L (8.82%) [1.5]	Y (96.0%) [2.1]	7 / 3697 (0.2%)	785 / 3713 (21.1%)
6) A C - -	227 / 2779 (8.2%)	1639 / 2779 (59.0%)	V (11.76%) [1.5]	B (79.0%) [1.2]	33 / 2535 (1.3%)	370 / 2560 (14.4%)
7) A - - H	368 / 3106 (11.85%)	1557 / 3106 (50.1%)	W (33.3%) [1.6]	Q (76.6%) [7.4]	156 / 2780 (5.6%)	444 / 2832 (15.7%)
8) A – D -	252 / 2856 (8.8%)	1676 / 2856 (58.7%)	V (15.4%) [3.1]	Y (84.0%) [4.1]	7 / 2647 (0.2%)	424 / 2651 (16.0%)
9) - C - H	609 / 4277 (14.2%)	2028 / 4277 (47.4%)	S (17.1%) [1.1]	C (64.74%) [2.7]	112 / 3938 (2.8%)	686 / 4056 (16.9%)
10) - - D H	651 / 4601 (14.1%)	2023 / 4601 (44.0%)	L (24.5%) [1.0]	Q (75.0%) [2.41]	30 / 4095 (0.7%)	1282 / 4433 (28.9%)
11) - C D -	717 / 4075 (17.6%)	1748 / 4075 (42.9%)	V (31.0%) [2.05]	W (68.3%) [3.1]	14 / 3851 (0.3%)	945 / 3941 (24.0%)

*;** The category X (Not categorized) was excluded. Only categories with more than 5 KOGs were used

Table 3. Single organism conservation*

	Ath	Cel	Dme	Hsa
Conserved KOGs (>=80%)	17.7%	32.2%	39.3%	40.9%
Poor-conserved KOGs (<20%)	41.3 %	21.2 %	19.8 %	15.6 %
Most conserved KOG category	J (37.4%)	C (65.7%)	J (68.2%)	C (74.4%)
Less conserved KOG category *	W (83.3%)	Y (42.8%)	S (29.3%)	S (26.0%)

*LSEs not included

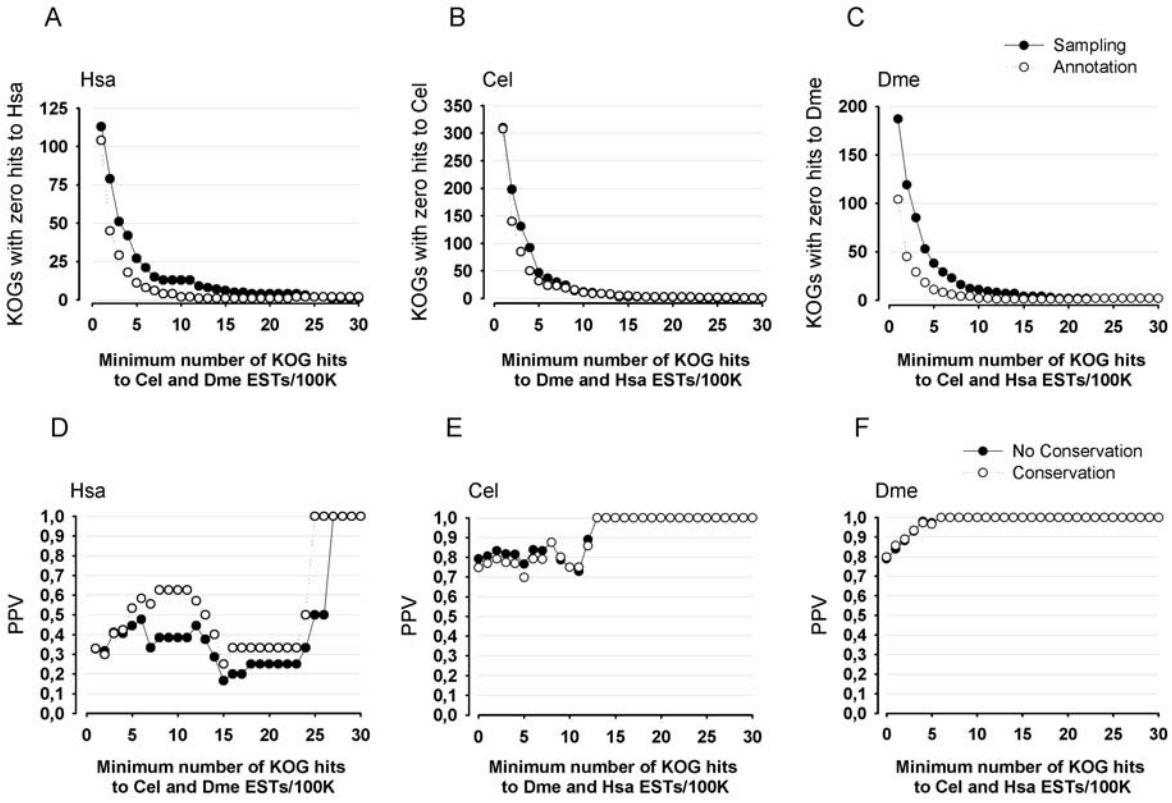


Figure 1. Gene loss prediction with the use of sampling and annotation information from Hsa, Cel and Dme. **A-C**: minimum number of KOG hits to ESTs that had zero hits with to the compared organism, using sampling (assigning ESTs to the cognate proteins, *full circles*) and annotation (annotating ESTs to the other organisms' proteins, *open circles*). **D-F**: PPV, or positive predictive value (TP/TP+FP), of non-existing KOGs by using the information from A-C with the usage of conservation information (exigency of at least 20% conservation or more, in both organisms, *open circles*) or without conservation (*full circles*).

ACKNOWLEDGEMENTS

This work was supported by FAPEMIG and CNPq. MAM received a fellowship from CAPES (Coordenação de Aperfeiçoamento de Nível Superior).

REFERENCES

1. Adams, M.D., et al., *Complementary DNA sequencing: expressed sequence tags and human genome project.* Science, 1991. **252**(5013): p. 1651-6.
2. Faria-Campos, A.C., et al., *Mining microorganism EST databases in the quest for new proteins.* Genet Mol Res, 2003. **2**(1): p. 169-77.
3. Franco, G.R., et al., *Identification of new Schistosoma mansoni genes by the EST strategy using a directional cDNA library.* Gene, 1995. **152**(2): p. 141-7.
4. Lee, N.H., et al., *Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment.* Proc Natl Acad Sci U S A, 1995. **92**(18): p. 8303-7.
5. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**: p. 41.
6. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
7. Mudado Mde, A. and J.M. Ortega, *A picture of gene sampling/expression in model organisms using ESTs and KOG proteins.* Genet Mol Res, 2006. **5**(1): p. 242-53.
8. Stekel, D.J., Y. Git, and F. Falciani, *The comparison of gene expression from multiple cDNA libraries.* Genome Res, 2000. **10**(12): p. 2055-61.
9. Kanehisa, M., et al., *The KEGG databases at GenomeNet.* Nucleic Acids Res, 2002. **30**(1): p. 42-6.
10. Kanehisa, M., *Linking databases and organisms: GenomeNet resources in Japan.* Trends Biochem Sci, 1997. **22**(11): p. 442-4.

5.8 Anotação e mineração de dados de EST de *Schistosoma mansoni* com KOG

Muito das informações e técnicas desenvolvidas e obtidas até agora para anotação e verificação de anotação de EST com a base KOG vieram por uma demanda do consórcio Rede Genoma de Minas Gerais (RGMG), durante o projeto em andamento do seqüenciamento do transcriptoma do platelminto *Schistosoma mansoni* (Sma). O Laboratório de Biodados, ICB – UFMG, participou na análise das seqüências desse verme geradas por esse consórcio. Com a disponibilização dessas seqüências, foi possível empregar algumas das informações e técnicas de anotação de EST com a base KOG. Os resultados foram organizados em um manuscrito (artigo de número 7 dessa tese), aceito para publicação como trabalho completo nos anais do congresso BSB 2007, Angra dos Reis, RJ, Brasil. Além disso, uma página *web* foi desenvolvida para disponibilizar os dados de anotação automática de Sma com a base de dados KOG, no intuito de ajudar na anotação manual e análise das seqüências, além de outras funcionalidades. Essa página pode ser acessada em: http://bioinfo.cpqrr.fiocruz.br/pct.

No artigo de número 7, foi comparada a caracterização da anotação das EST produzidas pela RGMG com as EST públicas, disponíveis no dbEST. Compararam-se os dados de cobertura da base KOG por EST de Cel e Dme com Sma, e foi inferido um número de proteínas KOG que potencialmente deveriam ter sido encontradas com as EST de Sma (parte desta falta pode se dever a viés de bibliotecas, mas parte pode representar genes ausentes no parasito). Os resultados sugerem que as EST provindas da RGMG ajudaram a complementar a caracterização do transcriptoma de Sma (achava-se que um máximo de descoberta gênica havia sido atingido (Verjovski-Almeida *et al.*, 2003). Sugeriu-se ainda que existem proteínas potencialmente anotáveis não descobertas, comparando-se com dados de anotação de EST de Cel e Dme com KOG (excluindo-se os organismos cognatos da EST). Similarmente, aglomerados KOG com expressão não detectável em Sma em comparação com Cel e Dme sugerem a perda de expressão de alguns genes em Sma. Assim como no K-EST, usou-se toda a coleção de EST de Sma para inferir a amostragem de EST por KOG e comparou-se com a dos outros organismos – a conservação foi levada em conta. Os resultados sugerem que Cel e Dme divergem menos quanto ao padrão de expressão por KOG, em comparação com Dme, já que uma parcela menor do transcriptoma apresenta

expressão diferencial com credibilidade alta. Esta é, em nosso conhecimento, a primeira tentativa de comparar as intensidades de transcrição entre genes de organismos diferentes. Embora a anotação com KOG tivesse sido gerada automaticamente pelo sistema de processamento de EST de Sma, o restante das comparações corresponde a este trabalho de tese.

Data mining and annotation of novel *Schistosoma mansoni* ESTs with the KOG database

Maurício A. Mudado[1], Guilherme Oliveira[2, 3], Rede Genoma de Minas Gerais[4], J. Miguel Ortega[1]

[1] Laboratório de Biodados, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, UFMG, Av. Antônio Carlos 6627, Belo Horizonte - MG, Brasil
{mudado,miguel}@icb.ufmg.br
[2] Centro de Pesquisas Rene Rachou, Av. Augusto de Lima, Belo Horizonte - MG, Brasil
{oliveira} @cpqrr.fiocruz.br
[3] Programa de Pós-Graduação e Pesquisa, Santa Casa de Belo Horizonte, Av. Francisco Sales 1111, Belo Horizonte - MG, Brasil
[4] Glória Franco, Vasco Azevedo, Fabrício Santos, Santuza Teixeira, Sérgio Costa, Instituto de Ciências Biológicas, UFMG; Ieso Castro, UFOP; Newton Portilho Carneiro, Cláudia Guimarães, EMBRAPA Milho e Sorgo; Luiz Goulart, UFU; Luciano Paiva, UFLA; Sergio Brommonschenkel, UFV.

Abstract. The KOG database was used to annotate the novel ESTs of the parasite *S. Mansoni*, sequenced by the RGMG (Rede Genoma de Minas Gerais) consortium. As the KOG database provides functional classification to its group of proteins it provided a good tool for the automatic annotation of the ESTs. The coverage of KOG clusters were used to estimate the number of proteins still to be discovered. This prediction made use of public ESTs of Model Organisms to compare the gene expression of *C. elegans* and *D. melanogaster* to *S. mansoni*. By using differential expression statistics we demonstrated that the *S. mansoni* transcriptome is more similar to *C. elegans* than the *D. melanogaster* transcriptome.

Keywords: *Schistosoma mansoni*, KOG, ESTs, BLAST, gene expression

1 Introduction

Intestinal schistosomiasis is caused mainly by the blood fluke *Schistosoma mansoni* in 54 countries, including Brazil. In Brazil, over 8 million people are infected and it is estimated that over 30 million are under the risk of contracting the disease [1, 2]. Schistosomiasis mansoni is treatable with a single dose of Praziquantel [3]. Mass chemotherapy has been effective in decreasing morbidity in endemic areas [4]. However this is the only effective drug for mass chemotherapy. In addition, there have been isolated reports of resistance or decreased susceptibility to Praziquantel in Egypt and Senegal and it is a consensus among the specialists that the development of new drugs and vaccines will be a major advance towards the control of this disease [5]. The approaches used to date towards the development of a vaccine have not been

successful and alternative approaches, especially with the use of the genomic and transcriptomic data are being sought [6].

The diploid genome of *S. mansoni* is approximately 270 megabase pairs contained within seven pairs of autosomal chromosomes and one pair of sex chromosomes, ZZ for male and ZW for female worms [7, 8, 9, 10]. Currently the genome has been sequenced and is in the process of annotation [11]. Advances have also been obtained with the transcriptome. Three major efforts towards the study of the transcriptome have been undertaken. The first, with support of WHO/TDR [12], a second at a much larger scale funded by FAPESP [13] and a third, yet unpublished supported by FAPEMIG [14]. Today, over 155,000 ESTs are available at dbEST at GenBank and over 60,000 are to be made available soon.

Many secondary databases have been serving as automatic annotation source for novel sequenced ESTs [15, 16, 17], like KEGG [18], COG [19] and KOG [20]. These databases are interesting for this task because are curated and have a proper protein categorization and classification. This work relates to how the transcriptomic data is categorized by the KOG database. Since KOG database provides a great source for further comparative genomics, it was chosen as a model analysis. Automated annotation with KOG database has been extensively evaluated in a previous work [21]. Other secondary databases such as OrthoMCL-DB, Kegg Orthology and PIR-SF are being currently processed. We also attempt to predict the gene coverage and compare the transcriptome of *Schistoma mansoni* with other Model Organisms and use that information to predict significant differences between them.

2 Methods

2.1 Library construction and sequencing

Libraries were constructed from total RNA obtained from the life cycle (egg and adult worm) maintained in mice and *B. glabrata* snails at the Centro de Pesquisas René Rachou – Fundação Oswaldo Cruz. Libraries were constructed using the Lambda ZAPII kit according to the manufacturer's instructions (Stratagene). Clones were obtained by selection on ampicilin accorind to the manufacturer's instructions. Clones were cultured in Circle growth medium (Bio 101) and plasmid DNA was obtained using a 96 well format Qiaprep (Qiagen). DNA sequencing was performed on MegaBace500 or MegaBace1000 sequencers (GE Healthcare Life Sciences) using the DyeNamic sequencing kit (GE Healthcare Life Sciences). Most of the data used here correspond to adult worm ESTs (roughly 50%), followed by egg ESTs (around 20 %) as judged by the retrieved information from the public data.

2.2 Website

A website was hosted under a Linux machine with Apache Webserver and constructed with PHP and MySQL. Please visit http://bioinfo.cpqrr.fiocruz.br/pct

2.3 BLASTs

BLASTx was conducted to annotate *S. mansoni*'s contigs with the KOG protein database (http://www.ncbi.nlm.nih.gov/COG/new/) and proteins downloaded from the 'kyva' file (ftp://ftp.ncbi.nih.gov/pub/COG/KOG/kyva). BLAST was used with the following parameters: -F f -m8 -e 1e-10, in order to turn off the low complexity filter, activate tabular output and establish an e-value cutoff of 1e-10. To avoid selecting more than one KOG protein to the same EST, the best alignments were always selected. A RPS-BLAST of the Conserved Domain Database (CDD - http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) against the KOG proteins was also performed, and best hits selected.

2.4 R value and believability statistics

R statistics [22] were used to show differences in expression between RGMG, dbEST, *C.elegans* and *D. melanogaster* ESTs. Data were randomized in order to find the R value threshold (believability) for false and true positives.

3 Results and Discussion

We are interested in revealing which fraction of the KOG database is covered by the ESTs generated by the RGMG *S. mansoni* transcriptome project. In Fig. 1 we show the result of BLASTX searches against the KOG amino acid sequences (60,758 *S. mansoni* sequences organized in 5,384 clusters) as a query and either public *S. mansoni* ESTs downloaded from dbEST or those generated in this project as a subject (under a stringent E-value cutoff of 10^{-10}). Each KOG entry with at least one EST hit was considered covered. The results show that, contrary to a previous prediction [13], gene discovery as understood by KOG coverage was not complete and the RGMG project contributed with a total of 117 novel hits to KOG database (black part of the bars in Fig. 1). The total of ESTs representing cDNA clones sequenced by the RGMG project also includes 1,332 hits to KOG database that are also common with dbEST sequences (INTERSECTION, grey part of the bars in Fig. 1). Because parasites may lack some commons biological functions in comparison to free living organisms, we estimated the potential of KOG coverage of EST collections from *C. elegans* and *D. melanogaster* (average KOG coverage represented by dashed part of the bars in Fig. 1). A total of 875 KOGs are missing in *S. mansoni* transcriptome by this comparison. These KOGs may either remain to be discovered or conversely, may represent genes that are absent in the parasite. The total of KOG entries are represented in Fig. 1 by dashed lines comprising the total KOG database including additional 920 KOG entries not detected in *S. mansoni*. Therefore, it seems more appropriated to compare gene discovery with the maximum represented by Model Organisms, dashed part of bars in Fig. 1.

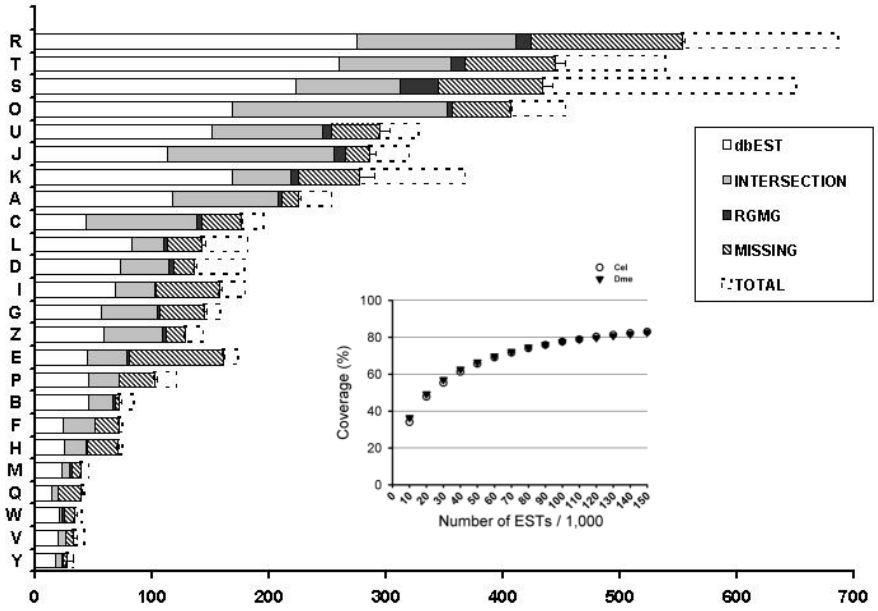


Fig. 1. Discovery of genes similar to KOG clusters in *S. mansoni* transcriptome. A total of 5384 KOG entries (LSE and TWOG entries, present in only one or two organisms in KOG database were not considered) separated by functional categories were used for BLAST similarity searches using *S. mansoni* ESTs. Bars are divided in KOG cluster representatives discovered using only ESTs present in dbEST (*white*), those common to dbEST and this work (*grey*) and the genes discovered exclusively by this project (*black*). Bars were classified by decreasing discovery in *S. mansoni* transcriptome. Conversely, the averaged KOG coverage attained by using up to 150,000 ESTs from either *C. elegans* or *D. melanogaster* was used to estimate the KOG representatives that are missing (*inset describes KOG coverage by increasing EST collections from these two Model Organisms*, as in [22]). Bars (*dashed*) show the complementary KOG entries in each category. KOG functional categories are represented by letters (*CELLULAR PROCESSES AND SIGNALING: D - Cell cycle control, cell division, chromosome partitioning; M - Cell wall/membrane/envelope biogenesis; N - Cell motility; O - Posttranslational modification, protein turnover, chaperones; T - Signal transduction mechanisms; U - Intracellular trafficking, secretion, and vesicular transport; V - Defense mechanisms; W - Extracellular structures; Y - Nuclear structure; Z – Cytoskeleton. METABOLISM: C - Energy production and conversion; E - Amino acid transport and metabolism; F - Nucleotide transport and metabolism; G - Carbohydrate transport and metabolism; H - Coenzyme transport and metabolism; I - Lipid transport and metabolism; P - Inorganic ion transport and metabolism; Q - Secondary metabolites biosynthesis, transport and catabolism. INFORMATION STORAGE AND PROCESSING: A RNA processing and modification; B - Chromatin structure and dynamics; J - Translation, ribosomal structure and biogenesis; K – Transcription; L - Replication, recombination and repair. POORLY CHARACTERIZED: R - General function prediction only; S - Function unknown*).

This type of comparison can be useful in several ways. One possibility is to investigate genes that are not being transcribed in *S. mansoni* by comparing its relative expression, as measured by the number of ESTs (best hits) that point to a given KOG, to the expression of the same KOG cluster determined by Model

Organisms ESTs best hits. Thus, it is possible to qualify the occurrence of zero hits to KOG in *S. mansoni* transcriptome by considering the rate of sampling of these same KOG in Model Organisms. Pre-computed data involving KOG database and up to 300 thousand ESTs from four Model Organisms was been used to build a bioinformatics tool that allows one to predict the sampling of KOG clusters in a novel transcriptome, named KOG Expression/Sampling Tool or K-EST [23]. We have selected for analysis those KOGs with zero hits to either dbEST ESTs, RGMG ESTs or contigs assembled using both sources of EST sequences. As an example, a total of 231 KOGs have more than 5 hits out of 100 thousand ESTs in *C. elegans* and *D. melanogaster* transcriptomes, while showing zero hits in the *S. mansoni* transcriptome (Fig. 2, open circles). However, since the parasite represents a novel transcriptome, a better analysis can be attained by considering the rate of sampling of Model Organism ESTs without the use of KOG amino acid sequences from that same species. This analysis is shown in Fig. 2 by the filled circles. The result indicates, for example, that a total of 64 KOGs have no hit to the *S. mansoni* transcriptome while producing over 5 hits per 100,000 ESTs in both *C. elegans* and *D. melanogaster* transcriptomes. Three KOGs - NADP-dependent isocitrate dehydrogenase, Myo-inositol-1-phosphate synthase and Carboxylesterase and related proteins show more than 30 ESTs (42-95 ESTs) per 100,000 in these Model Organisms transcriptomes, and are absenct in *S. mansoni*.

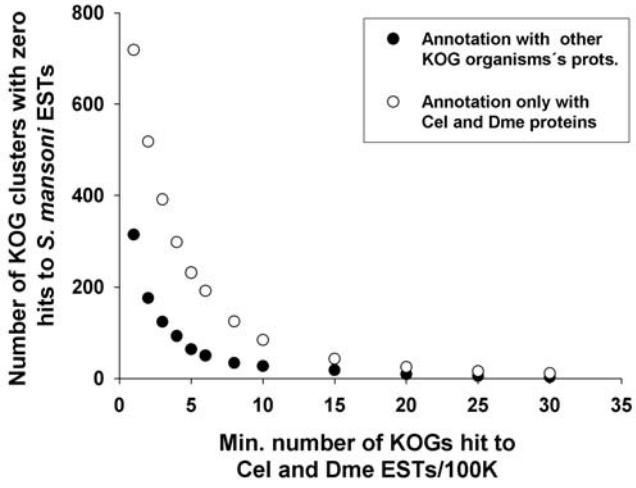


Fig. 2. Number of KOGs with undetectable expression in *S. mansoni* and sampled in Model Organisms. KOG proteins were used to run a tBLASTn analysis of S. mansoni transcriptome (dbEST ESTs, RGMG ESTs and contigs assembling both sources of ESTs) and *C. elegans* and *D. melanogaster* transcriptomes (over 300,000 ESTs). The number of KOG clusters with zero hits in *S. mansoni* transcriptome are shown (*open symbols*) in function of the minimum range of sampling of ESTs from Model Organisms. Also, similar analysis where KOG proteins from the cognate Model Organisms were not used is shown (*closed symbols*, see text for details).

Interestingly, Table 1 lists two biochemical pathways, Fatty Acid metabolism and Androgen and Estrogen metabolism, comprehending respectively eight and two KOGs with apparently undetectable expression in *S. mansoni*. The fact that these KOGs compile single pathways reinforce the suspicion that they are absent from the parasite transcriptome.

Table 1. KEGG / KO pathways with zero hits to *S. mansoni* ESTs/contigs and with multiple hits to *C. elegans* and *D. melanogaster* ESTs. *, ** Up numbers are hits to KOG proteins exclusive from the column's organism proteins. Bottom numbers inside parenthesis are hits to KOG proteins from other organisms but the column's organism proteins (*C. elegans or D. melanogaster;* see http://www.biodados.icb.ufmg.br/K-EST for more information). Numbers normalized by 100K ESTs.

KOG ID	Definition	*C. elegans (Other orgs.)	**D. melanogaster (Other orgs.)
ko00071 - Fatty acid metabolism			
KOG0135	Pristanoyl-CoA/acyl-CoA oxidase	9.76 (7.90)	12.62 (8.80)
KOG0136	Acyl-CoA oxidase	28.35 (22.30)	24.10 (14.92)
KOG0139	Short-chain acyl-CoA dehydrogenase	72.96 (61.80)	22.19 (21.42)
KOG1391	Acetyl-CoA acetyltransferase	33.46 (19.98)	24.10 (22.95)
KOG1392	Acetyl-CoA acetyltransferase	19.98 (19.05)	18.74 (18.74)
KOG1681	Enoyl-CoA isomerase	10.22 (9.76)	6.50 (6.50)
KOG1683	Hydroxyacyl-CoA dehydrogenase /enoyl-CoA hydratase	32.99 (24.63)	35.58 (29.84)
KOG3072	Long chain fatty acid elongase	38.57 (27.88)	14.54 (12.62)
ko00150 - Androgen and estrogen metabolism			
KOG2987	Fatty acid desaturase	6.51 (5.11)	12.24 (11.86)
KOG1600	Fatty acid desaturase	19.98 (18.12)	105.97 (73.83)

Sampling of gene expression in Model Organisms is an interesting source of information to investigate genes putatively differentially expressed in *S. mansoni*. Moreover, it would be possible to compare the parasite to other Model Organisms as to the level of expression. Table 2 shows the result of comparison of KOG sampling in *S. mansoni*, *C. elegans* and *D. melanogaster*. A statistical test developed by Stekel et al 2000 [22] was applied to the KOG sampling results aiming to determine the fraction of KOG clusters that are differentially expressed. The parameter believability indicates the level of confidence for the differential expression. The results show that the source of ESTs used in the comparison seem not to significantly influence the

analysis, since both RGMG or dbEST ESTs, or the sum of those data result in around 52 to up to 60 KOGs showing differential expression amongst the parasite, *C. elegans* and *D. melanogaster*, under the more stringent cutoff of believability (> 90%). Using the total of the ESTs available, about 59% of KOGs appear to be differentially expressed between *S. mansoni* and *D. melanogaster*, while only 27.52% of KOGs showed differential expression between the nematode and trematode worms. Conversely, 44.16% of KOGs were differentially represented between both Model Organisms. Thus, worm transcriptomes seem to be less divergent. Here, sampling was again measured in Model Organism transcriptomes without the use of the cognate organism KOG proteins, to make it fairer to the *S. mansoni* ESTs.

Table 2. Differential expression believability of *S. mansoni* (RGMG, dbEST, RGMG + dbEST), *C. elegans* (*Cel*) and *D. melanogaster* (*Dme*) ESTs annotated against the KOG database. The percentage of EST hits against the 2,902 common KOGs is used (raw numbers inside parenthesis).

	Believability < 50%	**Believability between 50% and 99%**	**Believability > 99%**
RGMG, Cel and Dme	9.43% (259)	34.98% (961)	55.59% (1,527)
dbEST, Cel and Dme	15.54% (427)	32.22% (885)	52.24% (1,435)
RGMG + dbEST, Cel and Dme	11.50% (316)	28.14% (773)	60.36% (1,658)
RGMG + dbEST and Dme	17.91% (492)	22.57% (620)	59.52% (1,635)
RGMG + dbEST and Cel	36.11% (992)	36.37% (999)	27.52% (756)
Cel and Dme	28.54% (784)	27.30% (750)	44.16% (1,213)

A website (http://bioinfo.cpqrr.fiocruz.br/pct) was created in order to show *S. mansoni*'s automatic annotation and help manual curators. The website shows the annotation of *S.mansoni* ESTs with the KOG database. Contig annotation is also shown allowing the comparison of EST and contig annotations with KOG. Also, a pre-computed annotation of KOG proteins with the NCBI's Conserved Domains Database (CDD) was used in order to annotation. We believe that, in addition to provide sequencing projects with a method to investigate expected outcomes of their projects, the data presented will be useful for the understanding of the biological functions of schistosomes by comparing the rate of gene expression with Model Organisms.

Acknowledgements
This work was funded a grant from FAPEMIG (EDT 17001/0 and REDE-281/05). GO is a CNPq fellow.

References

1. Bergquist, N. R.: Schistosomiasis: from risk assessment to control. Trends Parasitol 18 (2002) 309-314
2. Ferrari, M. L., Coelho, P. M., Antunes, C. M., Tavares, C. A., da Cunha, A. S.: Efficacy of oxamniquine and praziquantel in the treatment of Schistosoma mansoni infection: a controlled trial. Bull World Health Organ. 81(2003) 190-196
3. Cioli, D., Pica-Mattoccia, L.: Praziquantel. Parasitol Res. 90 Supp1 (2003) S3-S9
4. Kheir, M. M., Baraka, O. Z., el Tom, I. A., Mukhtar, M. M. , Homieda, M. M.: Effects of single-dose praziquantel on morbidity and mortality resulting from intestinal schistosomiasis. East Mediterr Health J. 6 (2000) 926-931
5. Doenhoff, M. J., Kusel, J. R., Coles, G. C., Cioli, D.: Resistance of Schistosoma mansoni to praziquantel: is there a problem? Trans R Soc Trop Med Hyg. 96 (2002) 465-469
6. Gryseels, B.: Schistosomiasis vaccines: a devil's advocate view. Parasitol Today. 16 (2000) 46-48
7. Marx, K. A., Bizzaro, J. W., Blake, R. D., Tsai, M-H., Tão, L-F.: Experimental DNA melting behavior of the three major Schistosoma species. Mol Biochem Parasitol. 107 (2000) 303-307.
8. Simpson, A. J. G., Sher, A., McCutchan, T. F.: The genome of Schistosoma mansoni: isolation of DNA, its size, bases and repetitive sequences. Mol Biochem Parasitol. 6 (1982) 125-137
9. Grossman, A. I., McKenzie, R., Cain, G. D.: Sex heterochromatin in Schistosoma mansoni. J Parasitol. 66 (1980) 368-370.
10. Short, R. B., Menzel, M. Y.: Somatic chromosomes of Schistosoma mansoni. J Parasitol. 65 (1979) 471-473
11. El-Sayed, N. M., Bartholomeu, D., Ivens, A., Johnston, D. A., LoVerde, P. T.: Advances In schistosome genomics. Trends Parasitol. 20 (2004)154-7
12. Oliveira, G., Johnston, D. A.: Mining the schistosome DNA sequence database. Trends Parasitol.17 (2001) 501-3
13. Verjovski-Almeida, S., DeMarco, R., Martins E. A., Guimaraes P. E., et al.: Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni. Nat Genet. 35 (2003) 148-57
14. Oliveira, G., Rodrigues, N. B., Romanha, A. J., Bahia, D.: Genome and Genomics of schistosomes. Canadian Journal of Zoology. 82 (2004) 375-390.
15. Vettore, A. L., da Silva, F. R., Kemper, E. L., Souza, G. M., et al.: Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res. 13 (2003) 2725-35
16. Faria-Campos, A. C., Cerqueira, G. C., Anacleto, C., de Carvalho, C. M., Ortega, J. M.: Mining microorganism EST databases in the quest for new proteins. Genet Mol Res. 2 (2003) 169-77
17. Deng, Y., Dong, Y., Thodima, V., Clem, R. J., Passarelli, A. L.: Analysis and functional annotation of expressed sequence tags from the fall armyworm Spodoptera frugiperda. BMC Genomics. 19 (2006) 7:264
18. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27 (1999) 29-34.
19. Tatusov, R. L., Galperin, M. Y., Natale, D. A., Koonin, E. V.: The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28 (2000) 33-6
20. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., et al.: The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4 (2003) 41.
21. Mudado, M. A., Ortega J. M.: A picture of gene sampling/expression in model organisms using ESTs and KOG proteins. Genet. Mol. Res. 5 (2006): 242-253

22. Stekel, D. J., et al.: The Comparison of Gene Expression from Multiple cDNA Libraries. Gen. Res. 10 (2000) 2055-2061
23. Mudado, M. A., Barbosa-Silva, A., Torres, J. A., Paula-Pinto, S., et al.: K-EST: KOG Expression Sampling Tool. (submitted for publication).

6 Considerações Finais

Neste trabalho foram apresentadas medidas de similaridade em alinhamentos envolvendo (i) seqüências EST de um organismo e seqüências de aminoácidos de proteínas KOG do organismo cognato e (ii) as mesmas EST alinhadas com a seqüência de nucleotídeos codificadora das proteínas KOG. Foi determinado que um cutoff entre 70 – 80% designava corretamente as EST às proteínas. Isto foi considerado em nossos ensaios como equivalente a um valor de corte de similaridade de 96% em alinhamentos de EST com seqüências nucleotídicas. Este valor de corte foi variável, conforme a coleção de EST, e parece estar refletindo a porcentagem de erro nas bibliotecas de EST, como mostrado na figura 19. Trata-se de um resultado não obtido anteriormente por nenhum grupo e foi peça fundamental para a realização dos testes de performance de anotação automática.

Uma segunda contribuição deste trabalho foi a descoberta de que a anotação automática de EST com a base de dados KOG não é ineficiente como poderia ter sido sugerido. A ausência de um teste transparente por vezes leva a conclusões errôneas. Cerca de 90% das anotações realizadas pela base KOG concorda com a designação feita a proteínas do organismo cognato. É plausível supor que agrupamentos incorretos de genes homólogos levassem a uma performance menor, e não tão alta. A quantidade de anotação especulativa, ou seja, anotação de EST que não havia alcançado o valor limite de corte para designação, foi pequeno, o que indica que a base de dados KOG possui sempre os ortólogos esperados em cada organismo, pois de outra forma a especulação seria mais alta. Isso pode ser observado em experimentos iniciais que realizamos com a base de dados *KEGG Orthology*.

O experimento que realizamos pode ser prontamente aplicável a outras bases de dados. Atualmente, um outro trabalho de tese foi iniciado no nosso laboratório para avaliar as bases *KEGG Orthology*, *OrthoMCL-DB*, Inparanoid, *PIR families*, *HomoloGene*, dentre outras, pelo aluno Gabriel Fernandes.

Projetos EST são realizados em tamanhos bastante variáveis, gerando desde 5 mil até acima de 300 mil seqüências de EST. Verificamos a performance de anotação da base KOG em anotar corretamente diferentes quantidades sorteadas de EST, concluindo que a performance é sempre a mesma. Todavia, como era de se esperar, o agrupamento e geração de consensos de EST depende do número de EST usadas, pois se torna mais fácil agrupar

EST quando o número é maior e a coleção fica mais redundante. Não seria possível, dado ao custo computacional, simular agrupamentos a partir de quantidades incrementais de EST. Por isso determinamos a curva de saturação da produção de *uniques* e avaliamos a performance da anotação em alguns pontos. Verifica-se que para projetos singelos de cerca de 5 mil EST a anotação de consensos tem praticamente a mesma performance que EST isoladas. Quando este número é maior (cerca de 150 mil EST), a análise de *uniques* traz uma conclusão surpreendente: a anotação correta diminui. Supostamente isso se deve ao fato de que as EST que teriam anotação correta aglomeram com maior facilidade. Todavia, o resultado mais apropriado é obtido quando são analisadas as EST componentes dos consensos. Verifica-se um aumento de por volta de 20% na anotação correta e, além disso, um aumento da quantidade de EST que são anotadas. Isso atesta um bom funcionamento do programa TGICL, que nada mais é do que a aglomeração de EST de maneira similar ao UniGene e a geração de consensos com Cap3, programa este que utiliza grande quantidade de recursos computacionais, principalmente memória, caso grandes conjuntos de EST sejam utilizados. Desta forma o TGICL permite que o Cap3 lide com cada agrupamento por vez e consiga chegar ao resultado final sem comprometer o uso de memória.

A designação de EST às entradas KOG do organismo cognato nos permitiu ter em mãos uma coleção de amostrabilidade de cada KOG. Evidentemente essa amostrabilidade é função do número de EST disponíveis. Assim, cunhamos a expressão “anotação reversa” para exemplificar a pesquisa de EST homólogas à uma coleção curada de proteínas, trazendo à genômica comparativa novas seqüências homólogas àquelas que já tínhamos disponíveis, objetivando conhecer a expressão relativa dessas proteínas antes mesmo que um novo transcriptoma seja seqüenciado.Assim, verificamos a expressão relativa de cada entrada KOG nos quatro organismos estudados e os resultados sugerem que cerca de 300 mil EST conseguem gerar uma amostragem bastante completa da coleção de genes contemplados na base KOG. Dependendo da categoria KOG e da coleção de EST, uma boa cobertura já acontece com cerca de 150 mil EST. É interessante notar que a inspeção dos resultados pode contribuir para uma previsão do resultado que se vai conseguir. De acordo com o conceito de “anotação reversa”, esta previsão indica a facilidade ou dificuldade de se descobrir em um novo transcriptoma os genes que se procura para a análise comparativa.

Utilizando os dados obtidos da amostragem de EST para cada entrada KOG foi publicado um *website,* K-EST. O usuário pode navegar pelas categorias funcionais para identificar genes de seu interesse, mas pode também iniciar a pesquisa por intermédio de um alinhamento com sua seqüência de interesse através de BLAST. A par da estimativa da expressão de cada KOG, o site K-EST prevê a eficiência de descoberta do gene em um projeto EST sendo o organismo de interesse tão diverso dos organismos modelo quanto eles são entre si. Assim, o usuário pode nas páginas de "*conservation*" detectar não somente o quantidade relativa de transcritos para o gene de interesse, mas também obter uma idéia de se ele pode ser detectado por organismos diversos do objeto de sua pesquisa. Várias iniciativas de anotação de EST a vias bioquímicas, utilizando KOG ou *KEGG Pathways*, tem sido reportadas. Todavia, em nosso conhecimento o *website* K-EST é a primeira tentativa de reportar o efeito do grau de conservação na anotação funcional. É bem possível que alguns genes raros e pouco conservados sejam de detecção muito difícil, sendo necessário um número muito grande de EST para que sejam detectados. Isto pode ser facilmente verificado pelas páginas de "*conservation*" do K-EST.

Baseado na estimativa prévia da amostrabilidade e da conservação dessa medida. É possível caminhar na direção de identificar genes que fogem a esta expectativa. Isto foi verificado avaliando-se a ocorrência de zero EST para cada KOG. Quando este KOG era razoavelmente expresso em outros organismos modelo e não em um deles, era proposta a sua não existência nesse organismo. Como a base de dados KOG indica quais genes estão ausentes em quais organismos, era possível verificar quantas EST seriam necessárias serem detectadas no outro organismo modelo para que a ocorrência de EST no organismo em questão fosse acurada. Verificamos que é possível indicar genes supostamente não existentes, ou não expressos, baseado na expressão e conservação dos mesmos em organismos modelo. Em nosso conhecimento esta é a primeria vez que o grau de expressão de genes em transcriptomas conhecidos é utilizado para prever a ausência de expressão de um gene em outros organismos. A abordagem foi aplicada a *S. mansoni*, com resultados plausíveis.

A expressão diferencial pode ser medida pela estatística proposta por Stekel, obtendo-se inicialmente o valor do parâmetro R e, posteriormente, determinado-se a parcela de genes que têm alta credibilidade de expressão diferencial. Verificamos que *S. mansoni* tem menos

genes com expressão diferencial de credibilidade alta em comparação com *C. elegans* de que com *D. melanogaster*. Em nosso conhecimento é a primeira vez que são usados níveis de expressão gênica em genômica comparativa.

Os artigos aqui descritos acomodam idéias que ainda estão vivas e podem ser aprimoradas com a chegada de novas seqüências. É interessante reforçar que o número de seqüências de proteínas curadas e de EST de diversos organismos estão em contínuo crescimento nos bancos de dados públicos. O surgimento da base KOG em 2003 foi um evento decisivo na elaboração e execução dos trabalhos aqui apresentados. Bases secundárias com seqüências curadas, que usam informações de homologia, e, portanto correlações evolutivas entre seqüências, só puderam aparecer com o sequenciamento em larga escala de genomas completos de organismos. Este é um fenômeno típico do momento da bioinformática em que estamos, e que vem se expandindo ainda mais. Uma versão atualizada da base KOG, com mais oito organismos, já está anunciada (http://www.ncbi.nlm.nih.gov/COG/new). Um número maior de organismos e seqüências na base KOG poderá levar a novas hipóteses bem como a ajudar a melhorar os resultados obtidos aqui, trazendo uma melhor compreensão das questões levantadas nessa tese. Outros métodos de construção de bases de dados secundárias de proteínas ortólogas estão sendo aperfeiçoados e novas bases estão surgindo. Exemplos são o *OrthoMCL* e o *InParanoid* (Li *et al.*, 2003; O'brien *et al.*, 2005; Chen *et al.*, 2006). Como seriam os resultados mostrados nessa tese usando essas outras bases de dados de proteínas ortólogas no lugar do KOG? Dessa forma, acredito que esta tese é composta por trabalhos que ainda podem gerar novos frutos e têm grande relevância no âmbito científico atual. Os resultados descritos aqui podem fornecer novas informações e auxiliar pesquisadores e cientistas que atuam em bioinformática de cunho genômico e que desenvolvem projetos relacionados na atualidade. Imaginemos que por exemplo, *C. elegans* não tivesse seu genoma seqüenciado ainda, e portanto não estivesse presente na base KOG. Porém com suas EST poderíamos avaliar seu grau de cobertura com KOG , alinhando essas EST com a base KOG. Seria possível descobrir que 150 mil EST de *C. elegans* são suficientes para cobrir 80% do KOG. Além disso poderíamos usar esses alinhamentos e utilizá-los no K-EST, descobrindo dessa forma que Cel possui um número menor de genes diferencialmente expressos (*believability>99%*) quando comparado com Dme (42,9%) em relação aos outros organismos (47,4% com Hsa, 59,0% com Ath). Além disso, utilizando

os dados de *sampling* e *conservation* de Hsa e Dme, seria possível tentar prever genes que não existem, ou têm expressão nula em Cel. Dessa forma, obteríamos 7 possíveis genes com maior expressão em Dme e Hsa (de pelo menos 13 EST em 100 mil), cuja expressão é nula em Cel.

Como visto nos manuscritos, a grande maioria dos resultados foi gerada e obtida a partir de seqüências, dados e ferramentas públicas e livres. Dessa forma, esse pode ser entendido como um trabalho genuíno em bioinformática; usou-se o computador e técnicas provindas da informática e computação, na tentativa de resolver os problemas, gerar ferramentas, e análises sobre dados públicos. Dessa forma, essa tese reforça uma idéia do aluno de doutorado da primeira turma do Curso de Pós-graduação em Bioinformática da UFMG, Carlos Henrique da Silveira, que dizia: "Um bioinformata só precisa de idéias, um computador ligado à Internet e uma garrafa térmica com café para realizar seus trabalhos".

7 Referências Bibliográficas

1. Adams, M. D., *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. Science, v.252, n.5013, Jun 21, p.1651-6. 1991.

2. Altschul, S. F., *et al.* Basic local alignment search tool. J Mol Biol, v.215, n.3, Oct 5, p.403-10. 1990.

3. Aouacheria, A., *et al.* Bioinformatic screening of human ESTs for differentially expressed genes in normal and tumor tissues. BMC Genomics, v.7, p.94. 2006.

4. Audic, S. e J. M. Claverie. The significance of digital gene expression profiles. Genome Res, v.7, n.10, Oct, p.986-95. 1997.

5. Baxevanis, A. D. e B. F. F. Ouellette. Bioinformatics: a practical guide to the analysis of genes and proteins. New York: Wiley-Interscience. 2001. xviii, 470 p., [13] p. of plates p. (Methods of biochemical analysis; v. 43)

6. Birney, E., *et al.* Mining the draft human genome. Nature, v.409, n.6822, Feb 15, p.827-8. 2001.

7. Boguski, M. S., *et al.* dbEST--database for "expressed sequence tags". Nat Genet, v.4, n.4, Aug, p.332-3. 1993.

8. Boguski, M. S. e G. D. Schuler. ESTablishing a human transcript map. Nat Genet, v.10, n.4, Aug, p.369-71. 1995.

9. Bouck, A. e T. Vision. The molecular ecologist's guide to expressed sequence tags. Mol Ecol, v.16, n.5, Mar, p.907-24. 2007.

10. Brenner, S., *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol, v.18, n.6, Jun, p.630-4. 2000.

11. Brent, M. R. Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res, v.15, n.12, Dec, p.1777-86. 2005.

12. Brown, T. A. Genomes. New York: Wiley-Liss. 2002. xxvii, 572 p. p.

13. Burke, J., *et al.* d2_cluster: a validated method for clustering EST and full-length cDNAsequences. Genome Res, v.9, n.11, Nov, p.1135-42. 1999.

14. Burks, C., *et al.* The GenBank nucleic acid sequence database. Comput Appl Biosci, v.1, n.4, Dec, p.225-33. 1985.

15. Chen, D. Y., *et al.* Low-cost, high-sensitivity laser-induced fluorescence detection for DNA sequencing by capillary gel electrophoresis. J Chromatogr, v.559, n.1-2, Oct 18, p.237-46. 1991.

16. Chen, F., *et al.* OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res, v.34, n.Database issue, Jan 1, p.D363-8. 2006.

17. Chou, H. H. e M. H. Holmes. DNA sequence quality trimming and vector removal. Bioinformatics, v.17, n.12, Dec, p.1093-104. 2001.

18. Corpet, F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res, v.16, n.22, Nov 25, p.10881-90. 1988.

19. Costa Lda, F. Bioinformatics: perspectives for the future. Genet Mol Res, v.3, n.4, p.564-74. 2004.

20. de Souza SJ, Camargo AA, Briones MR, Costa FF, *et al.* Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. Proc Natl Acad Sci U S A, v. 97, n.23, Nov 7, p.12690-3. 2000.

21. D'cunha, J., *et al.* An automated instrument for the performance of enzymatic DNA sequencing reactions. Biotechniques, v.9, n.1, Jul, p.80-5, 88-90. 1990.

22. Drossman, H., *et al.* High-speed separations of DNA sequencing reactions by capillary electrophoresis. Anal Chem, v.62, n.9, May 1, p.900-3. 1990.

23. Dujon, B. The yeast genome project: what did we learn? Trends Genet, v.12, n.7, Jul, p.263-70. 1996.

24. Ewing, B. e P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res, v.8, n.3, Mar, p.186-94. 1998.

25. Ewing, B., *et al.* Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res, v.8, n.3, Mar, p.175-85. 1998.

26. Faria-Campos, A. C., *et al.* Efficient secondary database driven annotation using model organism sequences. In Silico Biol, v.6, n.5, p.363-72. 2006a.

27. Faria-Campos, A. C., *et al.* Production of full-length cDNA sequences by sequencing and analysis of expressed sequence tags from Schistosoma mansoni. Mem Inst Oswaldo Cruz, v.101 Suppl 1, Sep, p.161-5. 2006b.

28. Faria-Campos, A. C., *et al.* Ferramentas Bioinformáticas Aplicadas à Caracterização da Expressão Gênica. Bioscience Journal, v.20, n.Suplem. 1, 2004, p.109-117. 2004.

29. Fleischmann, R. D., *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, v.269, n.5223, Jul 28, p.496-512. 1995.

30. Franco, G. R., *et al.* Evaluation of cDNA libraries from different developmental stages of Schistosoma mansoni for production of expressed sequence tags (ESTs). DNA Res, v.4, n.3, Jun 30, p.231-40. 1997.

31. Fraser, C. M., *et al.* The minimal gene complement of Mycoplasma genitalium. Science, v.270, n.5235, Oct 20, p.397-403. 1995.

32. Greller, L. D. e F. L. Tobin. Detecting selective expression of genes and proteins. Genome Res, v.9, n.3, Mar, p.282-96. 1999.

33. Gustincich, S., *et al.* The complexity of the mammalian transcriptome. J Physiol, v.575, n.Pt 2, Sep 1, p.321-32. 2006.

34. Higgins, D. G. e P. M. Sharp. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, v.73, n.1, Dec 15, p.237-44. 1988.

35. Huang, X. e A. Madan. CAP3: A DNA sequence assembly program. Genome Res, v.9, n.9, Sep, p.868-77. 1999.

36. Kanehisa, M., *et al.* The KEGG databases at GenomeNet. Nucleic Acids Res, v.30, n.1, Jan 1, p.42-6. 2002.

37. Karlin, S. e S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A, v.87, n.6, Mar, p.2264-8. 1990.

38. Kent, W. J. e D. Haussler. Assembly of the working draft of the human genome with GigAssembler. Genome Res, v.11, n.9, Sep, p.1541-8. 2001.

39. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet, v.39, p.309-38. 2005.

40. Koonin, E. V. e M. Y. Galperin. Sequence - evolution - function: computational approaches in comparative genomics. Boston: Kluwer Academic. 2003. xiii, 461 p., [11] p. of plates p.

41. Koonin, E. V. e A. R. Mushegian. Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. Curr Opin Genet Dev, v.6, n.6, Dec, p.757-62. 1996.

42. Krawetz, S. A. Sequence errors described in GenBank: a means to determine the accuracy of DNA sequence interpretation. Nucleic Acids Res, v.17, n.10, May 25, p.3951-7. 1989.

43. Lal, A., *et al.* A public database for gene expression in human cancers. Cancer Res, v.59, n.21, Nov 1, p.5403-7. 1999.

44. Lander, E. S., *et al.* Initial sequencing and analysis of the human genome. Nature, v.409, n.6822, Feb 15, p.860-921. 2001.

45. Lee, B., *et al.* ESTpass: a web-based server for processing and annotating expressed sequence tag (EST) sequences. Nucleic Acids Res, May 25. 2007.

46. Lee, N. H., *et al.* Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. Proc Natl Acad Sci U S A, v.92, n.18, Aug 29, p.8303-7. 1995.

47. Lesk, A. M. Introduction to bioinformatics. New York: Oxford University Press 2002.

48. Li, L., *et al.* OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res, v.13, n.9, Sep, p.2178-89. 2003.

49. Liang, F., *et al.* An optimized protocol for analysis of EST sequences. Nucleic Acids Res, v.28, n.18, Sep 15, p.3657-65. 2000.

50. Marchler-Bauer, A., *et al.* CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res, v.30, n.1, Jan 1, p.281-3. 2002.

51. Matsumura, H., *et al.* SuperSAGE. Cell Microbiol, v.7, n.1, Jan, p.11-8. 2005.

52. Mcbride, L. J., *et al.* Automated DNA sequencing methods involving polymerase chain reaction. Clin Chem, v.35, n.11, Nov, p.2196-201. 1989.

53. Mendes Soares, L. M. e J. Valcarcel. The expanding transcriptome: the genome as the 'Book of Sand'. Embo J, v.25, n.5, Mar 8, p.923-31. 2006.

54. Miller, G., *et al.* IMAGE cDNA clones, UniGene clustering, and ACeDB: an integrated resource for expressed sequence information. Genome Res, v.7, n.10, Oct, p.1027-32. 1997.

55. 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. v.9, n.11, Nov, p.1143-55. 1999.

56. Moreira-Filho, C. A., *et al.* Genômica. São Paulo: Editora Atheneu. 2004

57. Mount, D. W. Bioinformatics: sequence and genome analysis. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press. 2001. xii, 564 p. p.

58. Mudado M, A., *et al.* Tests of automatic annotation using KOG proteins and ESTs from 4 eukaryotic organisms. Lecture Notes Computer Sci., v.3594, p.141-152. 2005.

59. Nagaraj, S. H., *et al.* A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform, v.8, n.1, Jan, p.6-21. 2007.

60. Needleman, S. B. e C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol, v.48, n.3, Mar, p.443-53. 1970.

61. O'brien, C. Cancer genome anatomy project launched. Mol Med Today, v.3, n.3, Mar, p.94. 1997.

62. O'brien, K. P., *et al.* Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res, v.33, n.Database issue, Jan 1, p.D476-80. 2005.

63. Pearson, W. R. e D. J. Lipman. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A, v.85, n.8, Apr, p.2444-8. 1988.

64. Pertea, G., *et al.* TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics, v.19, n.5, Mar 22, p.651-2. 2003.

65. Pontius, J. U., *et al.* UniGene: A Unified View of the Transcriptome. The NCBI Handbook. 2003.

66. Prosdocimi, F. Racionalizando a utilização do algoritmo PHRED para a análise de seqüências de DNA. Tese (doutorado em Bioinformática), Instituto de Ciências Biológicas - UFMG, Belo Horizonte, 2006. 110 f. p.

67. Prosdocimi, F., *et al.* DNA Sequences Base Calling by PHRED: Error Pattern Analysis. RTInfo, v.3, p.107-110. 2003.

68. Prosdocimi, F., *et al.* Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values. Genet Mol Res, v.3, n.4, p.483-92. 2004.

69. Pruitt, K. D., *et al.* Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet, v.16, n.1, Jan, p.44-7. 2000.

70. Rudd, S. Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci, v.8, n.7, Jul, p.321-9. 2003.

71. Sanger, F., *et al.* DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A, v.74, n.12, Dec, p.5463-7. 1977.

72. Scheurle, D., *et al.* Cancer gene discovery using digital differential display. Cancer Res, v.60, n.15, Aug 1, p.4037-43. 2000.

73. Shiraki, T., *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A, v.100, n.26, Dec 23, p.15776-81. 2003.

74. Smith, T. F. e M. S. Waterman. Identification of common molecular subsequences. J Mol Biol, v.147, n.1, Mar 25, p.195-7. 1981.

75. Stein, L. Genome annotation: from sequence to biology. Nat Rev Genet, v.2, n.7, Jul, p.493-503. 2001.

76. Stekel, D. J., *et al.* The comparison of gene expression from multiple cDNA libraries. Genome Res, v.10, n.12, Dec, p.2055-61. 2000.

77. Strachan, T. e A. P. Read. Human molecular genetics 2. New York: Wiley-Liss. 1999. xxiii, 576 p. p.

78. Strausberg, R. L., *et al.* The mammalian gene collection. Science, v.286, n.5439, Oct 15, p.455-7. 1999.

79. Suzek, B. E., *et al.* UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, v.23, n.10, May 15, p.1282-8. 2007a.

80. Suzek, B. E., *et al.* UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters. Bioinformatics, Mar 22. 2007b.

81. Tatusov, R. L., *et al.* The COG database: an updated version includes eukaryotes. BMC Bioinformatics, v.4, Sep 11, p.41. 2003.

82. Tatusov, R. L., *et al.* The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res, v.28, n.1, Jan 1, p.33-6. 2000.

83. Tatusov, R. L., *et al.* A genomic perspective on protein families. Science, v.278, n.5338, Oct 24, p.631-7. 1997.

84. Tatusov, R. L., *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res, v.29, n.1, Jan 1, p.22-8. 2001.

85. Venter, J. C., *et al.* The sequence of the human genome. Science, v.291, n.5507, Feb 16, p.1304-51. 2001.

86. Verjovski-Almeida, S., *et al.* Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni. Nat Genet, v.35, n.2, Oct, p.148-57. 2003.

87. Wang, J. P., *et al.* EST clustering error evaluation and correction. Bioinformatics, v.20, n.17, Nov 22, p.2973-84. 2004.

88. Wang, S. M. Understanding SAGE data. Trends Genet, v.23, n.1, Jan, p.42-50. 2007.

89. Wei, C. L., *et al.* 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. Proc Natl Acad Sci U S A, v.101, n.32, Aug 10, p.11701-6. 2004.

90. Wilson, R. K., *et al.* Development of an automated procedure for fluorescent DNA sequencing. Genomics, v.6, n.4, Apr, p.626-34. 1990a.

91. Wilson, R. K., *et al.* Optimization of asymmetric polymerase chain reaction for rapid fluorescent DNA sequencing. Biotechniques, v.8, n.2, Feb, p.184-9. 1990b.

92. Wu, C. H., *et al.* PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res, v.32, n.Database issue, Jan 1, p.D112-4. 2004.

93. Zhang, J. Z., *et al.* High-sensitivity laser-induced fluorescence detection for capillary electrophoresis. Clin Chem, v.37, n.9, Sep, p.1492-6. 1991.

94. Zhang, Z., *et al.* A greedy algorithm for aligning DNA sequences. J Comput Biol, v.7, n.1-2, Feb-Apr, p.203-14. 2000.