Efficient secondary database driven annotation using model organism sequences

Article *in* In Silico Biology · February 2006

Source: PubMed

CITATION
1

READS
34

6 authors, including:

Alessandra C Faria-Campos
Federal University of Minas Gerais
31 PUBLICATIONS **164** CITATIONS
SEE PROFILE

Sergio Campos
Federal University of Minas Gerais
123 PUBLICATIONS **2,913** CITATIONS
SEE PROFILE

Francisco Prosdocimi
Federal University of Rio de Janeiro
253 PUBLICATIONS **3,922** CITATIONS
SEE PROFILE

J Miguel Ortega
Federal University of Minas Gerais
87 PUBLICATIONS **922** CITATIONS
SEE PROFILE

In Silico Biology 6 (2006) 363–372
IOS Press

Efficient Secondary Database Driven Annotation Using Model Organism Sequences

Alessandra C. Faria-Campos[a], Sérgio V.A. Campos[b], Francisco Prosdocimi[c], Glaura C. Franco[d], Glória R. Franco[a] and J. Miguel Ortega[a,*]

[a]*Departamento de Bioquímica e Imunologia, ICB-UFMG, 31270-010, Brazil*
[b]*Departamento de Ciência da Computação, ICEX-UFMG, 31270-010, Brazil*
[c]*Departamento de Biologia Geral, ICB-UFMG, 31270-010, Brazil*
[d]*Departamento de Estatística, ICEX-UFMG, 31270-010, Brazil*
E-mail: alessa@icb.ufmg.br; scampos@dcc.ufmg.br; franc@icb.ufmg.br; glaura@est.ufmg.br; gfranco@icb.ufmg.br

Edited by H. Michael; received 21 November 2005; revised 26 June 2006; accepted 28 June 2006; published 30 July 2006

ABSTRACT: The use of sequences from specific organisms for annotation requires that it does not represent great loss of information and that the sequences available suffice for annotation. In order to investigate whether or not sequences from model organisms may suffice for annotation of sequences from the trematode *Schistosoma mansoni*, we performed local BLAST searches of *S. mansoni* sequences against other organisms sequences present in the NCBI database *nr*. Results have been inserted into a relational database and hits to sequences from three model organisms, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens* have been computed and compared to hits to sequences from other organisms present in *nr*; score values of each alignment have also been registered. Our observations have shown that a large fraction of orthologous proteins exists in the set of sequences from the three model organisms selected, and therefore a similar fraction of transcripts can be annotated when using either *nr* or model organism datasets. Moreover, hits to model organisms' sequences are largely as informative as *nr*. Results suggest that model organisms provide a reliable set of sequences to use as a reference database for *S. mansoni* sequence annotation, showing the clear possibility of using a restricted dataset of expected better quality for functional annotation and therefore supporting secondary database driven annotation approaches.

KEYWORDS: Annotation, EST, *schistosoma mansoni*

INTRODUCTION

The genomes of several organisms have been partially or completely sequenced and annotated in the past decade and the resultant cumulative information has been successfully used to annotate novel sequences [*C. elegans* consortium, 1998; Celniker, 2000; Collins *et al.*, 2001; Venter *et al.*, 2001]. Amongst the investigated organisms are those considered Model Organisms (MO) such as *Saccharomyces cerevisiae*, *C. elegans* and *D. melanogaster*, which have concentrated the most intense efforts aiming the identification and classification of gene products. These genomes constitute a rich source of information

*Corresponding author. E-mail: miguel@icb.ufmg.br.

Electronic publication can be found in *In Silico Biol.* **6**, 0034 <http://www.bioinfo.de/isb/2006/06/0034/>, 30 July 2006.

with biological relevance and provide an unprecedented opportunity for automated mining of novel sequences [Pandey and Lewitter, 1999; Hsu, 2004]. Moreover, there is a considerable tendency of MO sequences to constitute dedicated databases (e.g. Flybase, ACedb, RefSeq) and, most importantly, these sequences often compose secondary databases such as KOG, CGAP-BioCarta, KEGG and others. Up to this moment, there is a lack of evidence of whether or not gene annotation and mining could be efficiently processed exclusively with MO sequences. Gene annotation consists on the analysis of sequences from a target organism and their interpretation aiming to extract from them biological information [Stein, 2001]. It is generally expected that divergence of the target of annotation from the available sequences used for its annotation could impair the procedure. This could become even more dramatic if the comparison is restricted to sequences from a specific set of organisms. On the other hand, annotation depends critically on the reliability and completeness of the sequences used as reference for the annotation [Bork *et al.*, 1996; Boffelli *et al.*, 2003; Boffelli *et al.*, 2004]. The exponential grow of information has been suggested as one of the causes for the lower quality of annotation found in sequences present in primary databases [Boeckmann *et al.*, 2003; Benson *et al.*, 2004]. Therefore, the use of sequences from primary databases as reference for annotation, such as the entire *nr* sequence collection (the complete non-redundant protein database available at NCBI – Wheeler *et al.*, 2003), might not be always the appropriate choice, since the use of inadequately annotated sequences as the basis for annotation would produce error-ridden and incomplete annotation of the novel sequences, which in turn would generate additional source for low quality annotation [Natale *et al.*, 2002; Ouzounis and Karp, 2002; Misra *et al.* 2002].

Additionally, the use of MO sequences for annotation and mining may support a new approach for gene discovery. A possible alternative to speed up sequence mining is to reverse the process, by initially choosing a collection of proteins derived from MO proteomes to annotate sequences that are being produced over the time by transcriptome sequencing projects. A good reason for using this approach is to propose a dedicated full-length sequencing schedule based on clones that are similar to a MO proteome "checking list". Although a large amount of sequence information from public databases is not used in this process, three remarkable advantages of the procedure stand out: (a) the redundancy and completeness of the protein collection is easily controlled; (b) comparisons to few species sequences increase the chance that 5' and 3' EST hit the same ortholog; (c) it would be possible to promptly direct gene discovery towards a large selection of genes, which can be used in evolutionary and functional comparisons. Although it is conceivable that interesting genes would be missed if annotation was done exclusively against MO datasets, annotation-based on a structured search can quickly reveal the core of common pathways. Thus, it remains to be tested how MO datasets would perform in comparison to the entire set of GenBank *nr* sequences.

In this work, we investigated the use of sequences from metazoa MO to annotate sequences from the trematode *S. mansoni*, an important human parasite, which has been the target of several initiatives to discover and characterize its genes but whose genome has not yet been completely sequenced [Franco, *et al.*, 1995; Santos *et al.*, 1999; Franco *et al.*, 2000; Prosdocimi *et al.*, 2002; Verjovski-Almeida *et al.*, 2003]. Through similarity searches we investigated the efficiency and accuracy of the annotation provided by MO sequences as opposed to that provided by *nr* sequences. Sequences from the lower eukaryote *S. cereviseae* have been used in a similar comparison to provide an unbiased analysis, since this organism has been traditionally placed at a farther evolutionary distance from *S. mansoni*. Results have shown that MO provide a reliable set of sequences to use as database for *S. mansoni* sequence annotation, thus supporting the possibility of using a restricted dataset of sequences of high quality for annotation and mining.

MATERIAL AND METHODS

Sequences

The following sets of sequences have been used in this study: *S. mansoni* (*sma*) ESTs from NCBI – dbEST (152,749 ESTs – August/2004); *sma* Uniques and SmAEs sequences, which have been generated through clustering of *sma* ESTs using the software CAP3 (6329 Uniques – January/2002; 30988 SmAEs – September/2003). Sequences have been downloaded from the National Center for Biotechnology Information – NCBI (http://www.ncbi.nlm.nih.gov) and The *Schistosoma mansoni* Genome Project from São Paulo – Brazil (http://cancer.lbi.ic.unicamp.br/schisto6/) except *sma* Uniques, which have been provided by the author [Prosdocimi *et al.*, 2002]. A database holding the sequences from all organisms was created locally using sequences from *nr*, which have been retrieved from NCBI's ftp site (3,296,422 sequences – August/2004; ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz).

Similarity Searches and SQL Queries

Similarity searches have been performed using BLAST [Altschul *et al.*, 1990] with the default parameters. The EXPECT threshold was set to 10^{-10}, what has been previously observed to suffice for filtering fortuitous hits although not preventing significant alignments from happening [Prosdocimi *et al.*, 2002]. BLAST results have been parsed and inserted into a MySQL database. The remaining processing was performed through SQL queries. Once we determined the total number of hits to *nr*, we divided them in two categories: (i) hits to *Schistosoma* spp. sequences – with identity over 80% and (ii) hits to other sequences. The last set of hits was then divided into three categories: hits to *C. elegans – D. melanogaster* (*cel-dme*), hits to *H. sapiens* (*hsa*) and hits to *nr**, defined as a set of sequences that have not shown hits neither to *Schistosoma* spp, *cel-dme* nor *hsa* sequences. The best hits against MO and *nr** for each *sma* sequence have been selected for analysis of quality of alignments. Score values of these hits have been computed and the ratios of scores obtained have been transformed by the log in base 2, so that events with value zero represent equal scores for both sets of sequences, events with negative values represent highest scores to *nr** sequences while positive values, highest scores to model organisms proteome.

Statistical Tests

Pearson correlation and Spearman rank-correlation [Lehmann, 1975] coefficients were calculated for comparisons of scores of *sma* sequences (EST, Uniques or SmAE) to either MO or *nr** sequences and significance statistics test applied. Coefficients presented large positive values (greater than 0.80) for all cases ($p < 0.001$), as shown in Table 1.

RESULTS AND DISCUSSION

In order to evaluate the use of sequences from MO proteomes as reference for annotation we have investigated the efficiency of these sequences in annotation of different sequence collections from *S. mansoni* (ESTs and assembled ESTs). For that, results of similarity searches using *S. mansoni* sequences against MO sequences have been compared to results obtained in searches against the complete set of amino acid sequences present in the *nr** database.

Table 1
Comparison of *nr** and model organism scores – percentage of total number of hits

Sequence type	MO	Sector[a]	Range of scores to MO sequences					Coefficient (p-value)	
			<100	100–199	200–299	>300	all scores	Spearman	Pearson
EST	*cel* + *dme* + *hsa*	*nr** > 1.25 MO	3.6	1.7	0.1	0.0	5.4	0.94	0.96
		intermediary	**32.4**	**47.1**	**12.9**	**2.0**	**94.5**	(<0.001)	(<0.001)
		MO > 1.25 *nr**	0.0	0.1	0.0	0.0	0.1		
		all sectors	36.1	48.9	13.0	2.0	100.0		
EST	*cel* + *dme*	*nr** > 1.25 MO	7.3	3.0	0.1	0.0	10.4	0.93	0.94
		intermediary	**29.5**	**46.0**	**12.2**	**1.9**	**89.6**	(<0.001)	(<0.001)
		MO > 1.25 *nr**	0.0	0.0	0.0	0.0	0.0		
		all sectors	36.9	49.0	1.5	1.9	100.0		
EST	*sce*	*nr** > 1.25 MO	27.1	26.3	1.5	0.0	54.8	0.84	0.87
		intermediary	**11.3**	**25.8**	**7.4**	**0.6**	**45.1**	(<0.001)	(<0.001)
		MO > 1.25 *nr**	0.0	0.0	0.0	0.0	0.0		
		all sectors	38.4	52.1	8.9	0.6	100.0		
Uniques	*cel* + *dme*	*nr** > 1.25 MO	5.5	3.7	0.5	0.4	10.0	0.85	0.89
		intermediary	**28.5**	**40.6**	**12.4**	**8.5**	**90.0**	(<0.001)	(<0.001)
		MO > 1.25 *nr**	0.0	0.0	0.0	0.0	0.0		
		all sectors	34.0	44.3	12.9	8.9	100.0		
SmAE	*cel* + *dme*	*nr** > 1.25 MO	9.3	4.7	0.5	0.1	14.6	0.87	0.93
		intermediary	**28.2**	**34.4**	**12.9**	**9.8**	**85.3**	(< 0.001)	(<0.001)
		MO > 1.25 *nr**	0.0	0.0	0.0	0.0	0.1		
		all sectors	37.5	39.1	13.5	10.0	100.0		

[a]Sectors of ratio between scores to *nr** and MO; intermediary sector represents ratio $= 1 +/- 25\%$.

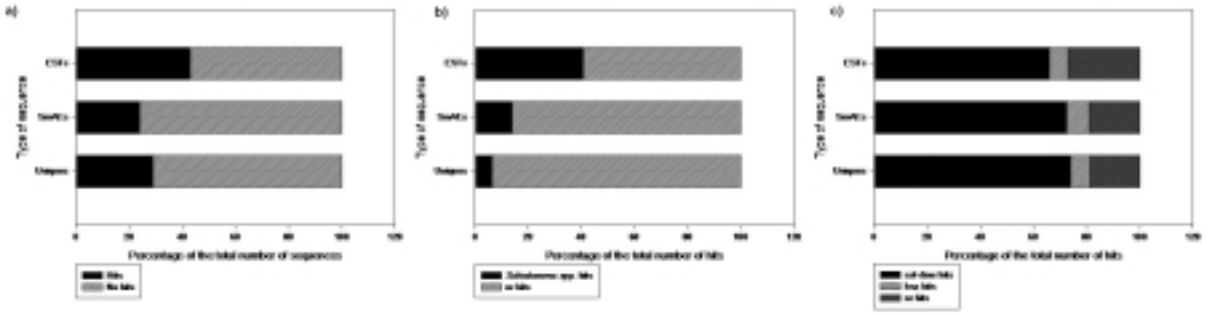

Fig. 1. Percentage of total number of hits of *S. mansoni* sequences on BLAST searches against *nr* grouped by category. A: Hits and no hits. B: Hits to *Schistosoma* spp. and *nr*; C: Hits to *C. elegans-D. melanogaster*, *H. sapiens* and *nr**. Abbreviations: EST (expressed sequence tag); Uniques and SmEA (assembled ESTs); *cel* (*C. elegans*); *dme* (*D. melanogaster*); *hsa* (*H. sapiens*); *nr* (non-redundant NCBI dataset).

MO proteomes show similarity to a large fraction of the S. mansoni sequences that are annotated by nr

S. mansoni (*sma*) is a human parasite that has been the subject of gene discovery projects based on the EST approach for several years [Franco *et al.*, 1995; 2000; Verjovski-Almeida *et al.*, 2003]. However, even with the large number of sequences available in public databases, only a fraction of its transcriptome can be annotated by similarity searches against all amino acid sequences available in the *nr* database. This behavior can be observed in Fig. 1a, where it can be seen that 43% of all *sma* ESTs show hits to *nr* sequences. From this total, that is the sequences that can be annotated, a significant number (41%) are directly assigned to *Schistosoma* proteins already present in public databases (either in the form of partial or complete CDS), as can be seen in Fig. 1b. These sequences represent 17.5% of the total *sma*

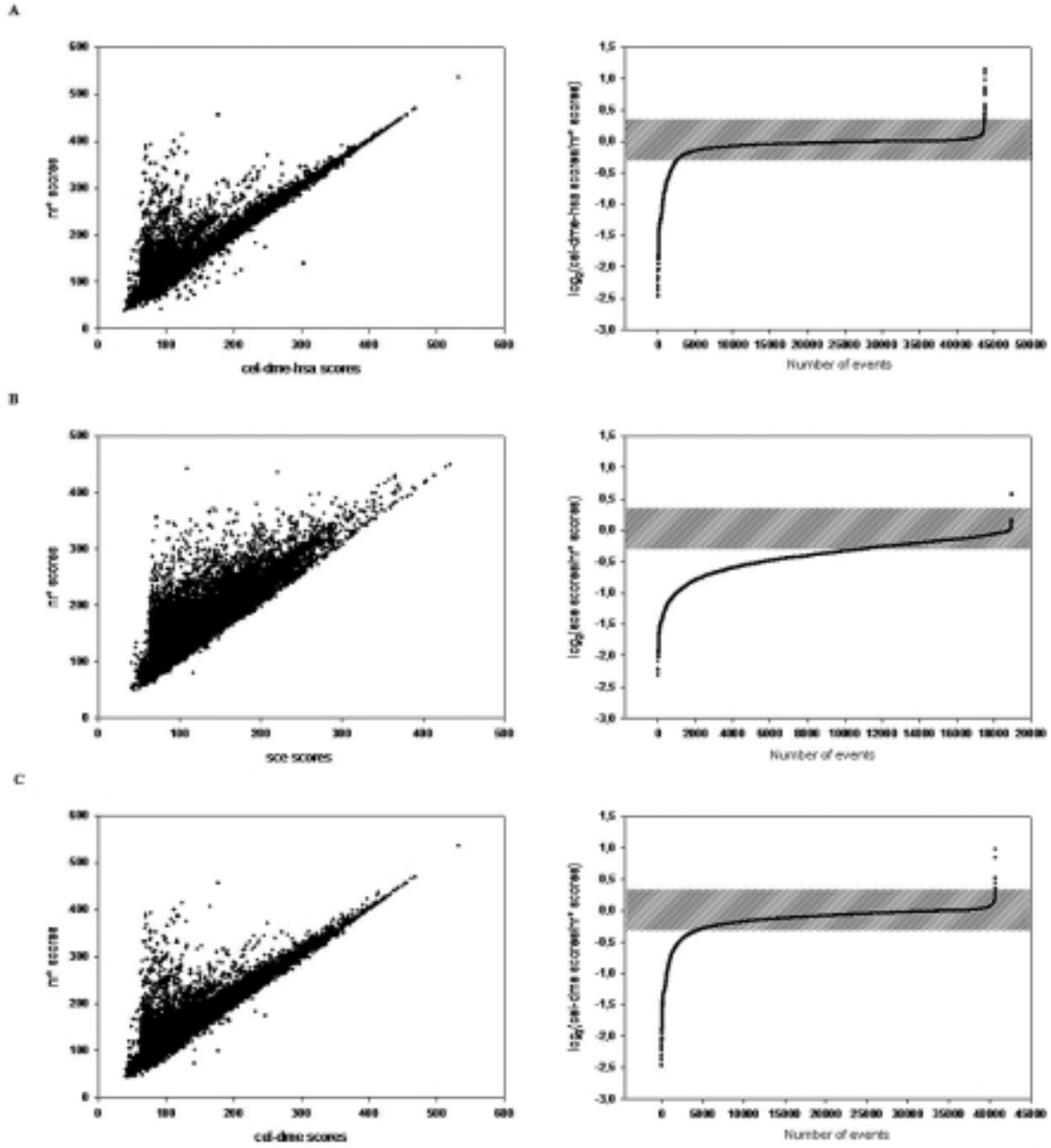

Fig. 2. Comparison of score values of BLAST searches of *S. mansoni* ESTs against *nr** and model organisms sequences and ratio of scores obtained against model organisms sequences to scores against *nr** transformed by the log in base 2. A: hits to *nr** versus hits to *cel-dme-hsa*; B: hits to *nr** versus hits to *sce*; C: hits to *nr** versus hits to *cel-dme*. Abbreviations: same as in Fig. 1 and *sce* (*S. cerevisiae*); *nr** (*nr* dataset excluding *Schistosoma* spp. and organisms in x-axis).

ESTs available in dbEST. In order to consider an EST as a hit to a protein entry, a similarity of over 80% of identity was required [Mudado *et al.*, 2005]. Assignment might not grant correct annotation, since many of the *Schistosoma* proteins have been annotated automatically by similarity; however these

sequences are not a substrate for gene discovery.

The behavior of contigs resulting from the assembling of ESTs follows a similar pattern. Two previous studies had assembled ESTs into contigs, generating a list of contigs and singlets named by the respective authors as either Uniques [Prosdocimi *et al.*, 2002] or SmAEs [Verjovski-Almeida *et al.*, 2003]. Uniques represent the assembling of 16,813 ESTs available in 2002 while SmAEs consist of the assembling of a recent EST collection of 124,681 ESTs. As shown in Fig. 1b, the percentage of sequences of both types that can be annotated directly by *Schistosoma* proteins is significantly smaller than that of ESTs (6.4% and 13.9% respectively). This behavior suggests that *Schistosoma* known proteins consist mainly of highly expressed genes and are not capable of providing annotation to a significant portion of the new genes represented by the contigs.

Figure 1b shows additionally that a significant portion of the worm transcriptome can be annotated by similarity searches against proteins from organisms other than *Schistosoma* (represented by hits to *nr* sequences). However, hits to *nr* yield strings that are poorly informative in a significant number of cases [Andrade *et al.*, 1999]. Moreover, using annotation-based on hits to *nr* does not guarantee the production of a reliable list of proteins to be investigated in the subject organism, because often orthologs from prokaryotes and eukaryotes receive conspicuously different definitions, even upon the advance of Gene Ontology. Therefore, an attractive alternative is to rely on restricted sequence sets of known proteomes from MO. In this case, a MO proteome driven annotation approach can be more effective, since such approach can contribute significantly to characterize new sequences at the amino acid level and to increase the number of full-length genes available. This will increase the information on genes that are expressed less frequently, which can result in new perspectives for control strategies [Verjovski-Almeida *et al.*, 2004].

The use of a subset of sequences may, however, raise questions regarding the quality of the resulting annotation. Thus, tests investigating the annotation both quantitatively and qualitatively must be conducted.

Regarding *S. mansoni* sequences, an intuitive choice of model organisms to use as reference for annotation would certainly point to *C. elegans* (cel) and *D. melanogaster* (*dme*) as the best candidates, due to the relative evolutionary proximity [Hausdorf, 2000]. Sequences from these organisms often participate in secondary databases, such as KOG (NCBI) and KEGG databases, in which frequently a classification into functional categories is added to the curated and non-redundant annotation [Tatusov *et al.*, 2000; Tatusov *et al.*, 2001; Tatusov *et al.*, 2003; Kanehisa *et al.*, 2004]. Fig. 1c shows that, from the *sma* sequences that have not been annotated directly by *Schistosoma* spp. proteins, most entries show significant similarity (under 10^{-10} E-value cutoff) to *cel* + *dme* amino acid sequences. A small fraction is annotated if *H. sapiens* (*hsa*) proteome is added to the reference set, even though *hsa* is an evolutionarily distant organism. Figure 1c also shows that just a small fraction of the sequences (less than 27% for all types of sequences investigated) depend on the use of other *nr* sequences to be identified. Thus, from a quantitative point of view the annotation achieved using *cel* + *dme* sequences can be nearly as good as that achieved using all other organisms sequences. ESTs are the type of sequence that is less annotated by *cel* + *dme* sequences, suggesting that some proteins highly expressed in *S. mansoni* are absent in these model organisms. These genes seem also not to be present in *hsa*, since the increase in the number of hits when *hsa* sequences are added is about the same for all types of sequences.

Model organism (*MO*) *sequences often provide hits of quality similar to that of nr sequences*

It can be argued that similarity to *nr* entries could be significantly higher than to members of a limited collection of proteins from MO. To evaluate this possibility, we have compared the scores in alignments

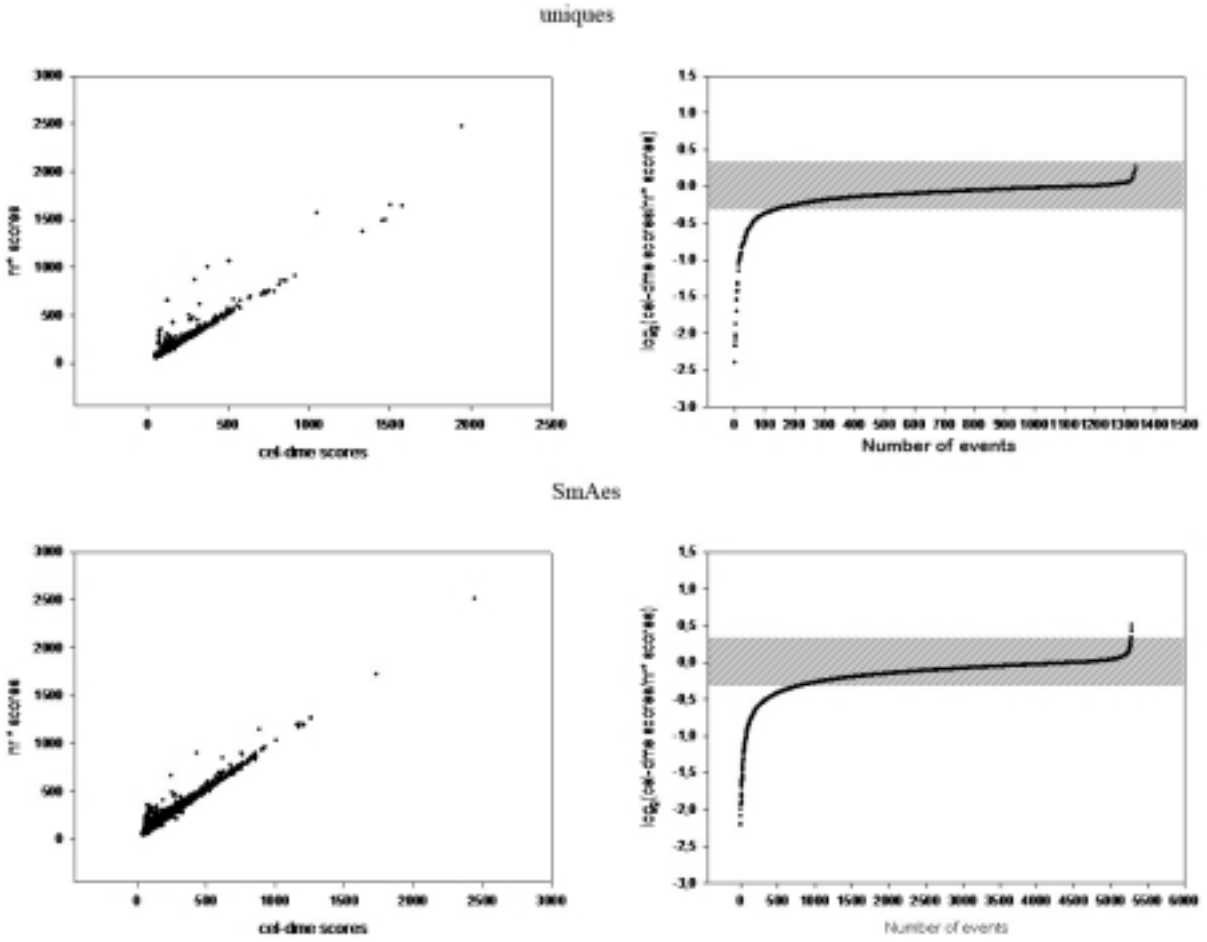

Fig. 3. Comparison of score values of BLAST searches of *S. mansoni* Uniques and SmAEs against *nr** and model organisms sequences and ratio of scores obtained against model organisms sequences to scores against *nr** transformed by the log in base 2. A: Uniques – hits to *nr** versus hits to *cel-dme*; B: SmAEs – hits to *nr** versus hits to *cel-dme*. Abbreviations: same as in Figs 1 and 2.

of all types of *sma* sequences to either MO proteomes or the complement of those in the *nr* collection (*nr**). Data presented in Fig. 2 argues against this conservative view. When sequences from three model organisms were used (*cel* + *dme* + *hsa*) most of the data has presented similar scores to either model organisms proteins or to *nr** (*nr*, not including *cel*, *dme*, *hsa* or *Schistosoma* spp), what is depicted by the fact that most of the points lie on the diagonal. A subset of events shows slightly higher scores to *nr** than to MO sequences and those are concentrated in the range of the lowest scores in alignments to the MO proteomes. A quantitative analysis of this distribution is presented in Table 1. Similarity searches that result in scores on hits to *nr** sequences 25% higher than to MO sequences (*nr** $>$ 1.25 MO) or the opposite (MO $>$ 1.25 *nr**), have been counted and processed as percentage of the total of points in the graphics in Fig. 2. It becomes clear that less than 5.5% of the points fall apart from the diagonal (intermediary sector) in more than 25% percent when EST sequences are used and the model organisms are *cel* + *dme* + *hsa*. As the scores in alignments to model organism sequences rise (e.g. over 100), the number of points that are far from the diagonal decreases, suggesting that model organisms are able to

ensure the best annotation possible at better score ranges. Correlation coefficients showed large values (Spearman: 0.94; Pearson 0.96; p-value 0.000) for this comparison. A detailed overview of events discussed above for each type of *sma* sequence and collection of model organism sequences is depicted in right panels in Fig. 2, where the ratio of scores obtained for MO to *nr** have been transformed by the log in base 2, so that events with value zero represent equal scores for both sets of sequences, events with negative values represent highest scores to *nr** while positive values, highest scores to model organisms proteome. It is clear that the majority of the events present values closer to zero (between -0.332 and $+0.332$, what represents 25% deviation from even scores) showing that, in terms of score, annotation with model organisms is equivalent to the use of a non-organized set of proteins from all other organisms. In order to provide a different type of comparison, the analysis was repeated using *S. cerevisiae* (*sce*) as the model organism. In this case a significant number of points are far from the diagonal showing higher scores to *nr** (*nr* here subtracted of *Schistosoma* spp and *sce* sequences). Representation of the ratio of scores in the right panel of Fig. 2b clearly shows the worse quality of annotation with *sce* in comparison to *nr**. Moreover, Table 1 shows that only 45.1% of the points are in the intermediary sector. Another type of comparison can be performed leaving out *hsa* and using only *cel* + *dma* as model organisms. In this case, 10.4% of the points are far from the diagonal showing score differences over 25% (4.1% of the points show score differences over 50%, not shown). This analysis shows equally good results, as shown in Fig. 2c and Table 1, which means that in this case there is not a strong need for using *hsa* sequences and a simplified approach can also be used.

Similar results have been obtained for Uniques and SmAEs (Table 1). A detailed view of the behavior for these types of sequences is shown in Fig. 3. Thus, the suggested model organisms are sufficient for providing support for gene discovery in *S. mansoni*. However, it may be reasonable to use sequences from other model organisms that constitute a given secondary, curated and classified database, since some gain in efficiency can be obtained by including a more evolutionary distant organism as *H. sapiens*, as shown here.

S. mansoni is a good example of a novel organism to be annotated by the proposed approach because its evolutionary proximity to the invertebrates chosen here is significant [Hausdorf, 2000]. Besides that, our data has shown that if an orthologous protein is present in the model organisms (yielding a score over $\sim$100 to the MO sequence), these organisms are able to provide a hit with quality similar or very close to the best *nr* hit, decreasing the necessity of using less reliable or redundant sequence databases.

Clearly, as the generation of amino acid and nucleotide sequences further advances in the direction of gene discovery, all resources available to propagate annotation should be used. As characterization of new sequences and classification in secondary databases progresses, the use of restricted sets of sequences of expected better quality constitutes one such resource along with the analysis of domains conserved in gene families. Future developments using such resources will depend critically on the full length sequencing of the coding region, justifying and endorsing the use of the reverse annotation procedure as described here. Furthermore, the use of specific databases is particularly promising because the restricted sets are smaller and require less computational resources without sacrificing the quality of the analysis. As the sequence databases grow, this advantage will become more important since processing large sets of sequences tend to become too timing consuming, slowing down the annotation process.

ACKNOWLEDGEMENTS

The authors would like to thank J.A. Tôrres, E. Bravo-Neto and M.A. Mudado for assistance with SQL queries and Conselho Nacional de Pesquisa e Desenvolvimento (CNPq) and Fundaç ão de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG) for financial support.

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. **215**, 403-410.
- Andrade, M. A., Brown, N. P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. and Sander, C. (1999). Automated genome sequence analysis and annotation. Bioinformatics **15**, 391-412.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2004). GenBank: update. Nucleic Acids Res. **32**, D23-D26.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31**, 365-370.
- Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovchrarenko, L., Patchter, L. and Rubin, E. M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human of the human genome. Science **299**, 1391-1394.
- Boffelli, D., Weer, C. V., Weng, L., Lewis, K. D., Shoukry, M. I., Pachter, L., Keys, D. N. and Rubin, E. M. (2004). Intraspecies sequence comparison for annotating genomes. Genome Res. **14**, 2406-2411.
- Bork, P. and Bairoch, A. (1996). Go hunting in sequence databases but watch out for the traps. Trends Genet. **12**, 425-427.
- The *C. elegans* Sequencing Consortium. (1998). Genome Sequence of the nematode *C. elegans*: A platform for investigating biology. Science **282**, 2012-2018.
- Celniker, S. E. (2000). The *Drosophila* genome. Curr. Opin. Gen. Dev. **10**, 612-616.
- Landers, E. S., *et al.*; International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature **409**, 860-921.
- Franco, G. R., Adams, M. D., Soares, M. B., Simpson, A. J., Venter, J. C. and Pena, S. D. (1995). Identification of new *Schistosoma mansoni* genes by the EST strategy using a directional cDNA library. Gene **152**, 141-147.
- Franco, G. R., Valadªo, A. F., Azevedo, V. and Rabelo, E. M. L. (2000). The *Schistosoma* gene discovery program: State of the art. Int. J. Parasit. **30**, 453-463.
- Hausdorf, B. (2000). Early evolution of the bilateria. Syst. Biol. **49**, 130-142.
- Hsu, S. Y. T. (2004). Bioinformatics in reproductive biology - functional annotation based in comparative sequence analysis. J. Reprod. Immunol. **63**, 75-83.
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res. **32**, D277-D280.
- Lehmann, E. L. (1975). Nonparametrics, Statistical Methods Based on Ranks. Holden-Day, San Francisco.
- Misra, S., *et al.* (2002). Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol. **3**, RESEARCH0083.
- Mudado, M. A., Bravo-Neto, E. and Ortega, J. M. (2005). Tests of automatic annotation using KOG proteins and ESTs from 4 eukariotic organisms. Lecture Notes in Computer Sciences **3594**, 141-152.
- Natale, D. A., Shankavaram, U. T., Galperin, M. Y., Wolf, Y. I., Aravind, L. and Koonin, E. V. (2000). Towards understanding the first genome sequence of a crenarcheon by genome annotation using clusters of orthologous groups of proteins (COGs). Genome Biol. **1**, RESEARCH0009.
- Ouzounis, C. A. and Karp, P. D. (2002). The past, present and future of genome-wide re-annotation. Genome Biol. **3**, COMMENT2001.
- Pandey, A. and Lewitter, F. (1999). Nucleotide sequence databases: a gold mine for biologists. Trends Biochem Sci. **24**, 276-280.
- Prosdocimi, F., Faria-Campos, A. C., Peixoto, F., Ortega, J. M. and Franco, G. R. (2002). Clustering of *Schistosoma mansoni* mRNA sequences and analysis of the most transcribed genes: implications in metabolism and biology of different developmental stages. Mem. Inst. Oswaldo Cruz **97**, 61-69.
- Santos, T. M., Johnston, D. A., Azevedo, V., Ridgers, I. L., Martinez, M. F., Marotta, G. B., Santos, R. L., Fonseca, S. J., Ortega, J. M., Rabelo, E. M., Saber, M., Ahmed, H. M., Romeih, M. H., Franco, G. R., Rollinson, D. and Pena, S. D. (1999). Analysis of the gene expression profile of *Schistosoma mansoni* cercariae using the expressed sequence tag approach. Mol. Biochem. Parasitol. **103**, 79-97.
- Stein, L. (2001). Genome annotation: from sequence to biology. Nat. Gen. **2**, 493-503.
- Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. **28**, 33-36.
- Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. and Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29**, 22-28.
- Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, EV., 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. and

Natale, D. A (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4**, 41.
- Venter, J. C., *et al.* (2001). The sequence of the human genome. Science **291**, 1304-1352.
- Verjovski-Almeida, S., *et al.* (2003). Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. Nat. Genet. **35**, 148-157.
- Verjovski-Almeida, S., Leite, L. C. C., Dias-Neto, E., Menck, C. F. M. and Wilson, R. A. (2004). Schistosome transcriptome: insights and perspectives for functional genomics. Trends Parasitol. **20**, 304-308.
- 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. and Wagner, L. (2003). Database resources of the National Center for Biotechnology. Nucleic Acids Res. **31**, 28-33.