Production of full-length cDNA sequences by sequencing and analysis of expressed sequence tags from *Schistosoma mansoni*

Alessandra C Faria-Campos, Fernanda S Moratelli, Isabella K Mendes, Paula L Ortolani, Guilherme C Oliveira*, Sérgio V A Campos, J Miguel Ortega, Glória R Franco/+**

Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas **Laboratório de Universalização de Acesso a Internet, Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901 Belo Horizonte, MG, Brazil *Centro de Pesquisas René Rachou-Fiocruz, Belo Horizonte, MG, Brasil

The number of sequences generated by genome projects has increased exponentially, but gene characterization has not followed at the same rate. Sequencing and analysis of full-length cDNAs is an important step in gene characterization that has been used nowadays by several research groups. In this work, we have selected Schistosoma mansoni *clones for full-length sequencing, using an algorithm that investigates the presence of the initial methionine in the parasite sequence based on the positions of alignment start between two sequences. BLAST searches to produce such alignments have been performed using parasite expressed sequence tags produced by Minas Gerais Genome Network against sequences from the database* Eukaryotic Cluster of Orthologous Groups *(KOG). This procedure has allowed the selection of clones representing 398 proteins which have not been deposited as* S. mansoni *complete CDS in any public database. Dedicated sequencing of 96 of such clones with reads from both 5' and 3' ends has been performed. These reads have been assembled using PHRAP, resulting in the production of 33 full-length sequences that represent novel* S. mansoni *proteins. These results shall contribute to construct a more complete view of the biology of this important parasite.*

Key words: expressed sequence tags - sequencing - *Schistosoma* - full-length cDNA

While the number of sequences generated by genome projects has increased exponentially, this phenomenon has not been followed by gene characterization at the same rate (Saghatelian & Cravat 2005). Aiming to diminish that gap, several approaches have been used for gene discovery through searches on the genomic sequence or analysis of the transcriptome of the organisms (Okazaki et al. 2002). One such approach is gene discovery using expressed sequence tags (ESTs), short sequences produced by sequencing the ends of cDNA molecules, which represent a snapshot of the gene expression for a given organism at a certain time (Adams et al. 1991). However, a complete picture of the gene products of the organism can only be obtained when the full-length sequence of specific proteins is determined. For that, it is essential to select clones that potentially present the complete coding sequence, up to the initial methionine and proceed to sequencing and characterization of such clones (Das et al. 2001). The initial step in the characterization is to determine the completeness of the cDNAs from which the ESTs were generated, followed by the annotation using the biological description of sequences present in function-oriented databases. Several authors developed systems to attain this goal which use among other tools similarity searches, statistical information and genome mapping (Salamov et al. 1998, Nishikawa et al. 2000, Del Val et al. 2003, Furuno et al. 2003, Hotz-Wagenblatt et al. 2003). *Schistosoma mansoni* cDNAs have been selected in this work for full-length sequencing using an algorithm based on BLAST searches of parasite ESTs against sequences from the database *Eukaryotic Cluster of Orthologous Groups* (KOG) (Tatusov et al. 2003). The implementation of the algorithm uses SQL searches to predict the presence of the initial methionine in the parasite sequence, resulting in the selection of ESTs representing clones putatively containing the complete sequence. By this procedure we have been able to select 3988 ESTs representing 398 proteins which have not yet been deposited as *S. mansoni* complete CDS in any public database. Dedicated 5' and 3' end sequencing of 96 clones has been performed and reads have been assembled using PHRAP. As a final result, 33 full-length sequences have been produced which represent novel *S. mansoni* proteins.

MATERIALS AND METHODS

Selection of ESTs representing putative full-length clones - S. mansoni ESTs have been selected for full-length sequencing using the algorithm described by Nishikawa et al. (2000) with modifications. The completness of the clones represented by the ESTs was determined by comparison of these to sequence of proteins from the secondary database KOG using tblastn. The positions of start and end of alignment assigned by BLAST in both sequences were determined through SQL queries. When the length of the not-aligned 5'-terminal of the EST was longer than that of the beggining of the aligned protein multiplied by three, the EST has been considered to rep-

Financial support: CNPq, Fapemig
+Corresponding author: gfranco@icb.ufmg.br
Received 26 May 2006
Accepted 26 June 2006

resent a clone that had the entire coding region and therefore has been selected for full-length sequencing (Fig. 1).

A total of 63,960 ESTs produced by Minas Gerais Genome Network (RGMG) has been used in BLAST searches against 88,644 sequences from the database KOG. Start and end alignment position from BLAST results have been inserted into a MySQL database for processing, allowing clone selection through SQL queries.

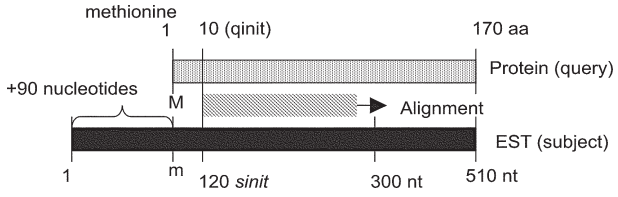

Fig. 1: model for prediction of clone completeness. BLAST searches (tBLASTn) of *Schistosoma mansoni* ESTs (*subject*) are performed against KOG protein sequences (*query*) and results are inserted into MySQL database. SQL searches retrieve positions of alignment start for both sequences (*qinit*: alignment start position for ortholog protein; *sinit*: alignment start position for EST) and use them to calculate the presence of the methionine using the following equation: [(*sinit*) / 3] – *qinit= m];* M: initial methionine in the ortholog protein; m: initial methionine in the *S. mansoni* translated sequence; aa: amino acid; nt: nucleotide.

Sequencing and assembling - DNA has been prepared for sequencing as described by Brazilian National Genome Project Consortium (2003), using 400 ηg of DNA. Sequencing reactions have been performed using the kit DYEnamic™ ET Dye Terminators according to instuctions provided by the manufacturer with the primers M13 reverse (5'-GGAAACAGCTATGACCATG-3') and foward (5'-GTTTTCCCAGTCACGAC-3') and run on MegaBACE™ 1000 (GE Healthcare). Six reads (three for each primer direction) have been generated for each clone ranging from 0.4 to 0.6 Kb. Base calling and sequence assembling have been performed using phred/phrap/consed (http://www.phrap.org) and a quality value cutoff phred = 20.

Annotation and search of ORFs - Contig annotation has been performed using BLAST searches against KOG sequences and the tool Blast2GO (Conesa et al. 2005). Manual inspection of sequences has been performed after automatic annotation and ORFs have been located by Consed analysis. Complete sequences have been translated using RevTrans (Wernersson & Pedersen 2003) and ORFs aligned to their ortologs using Multialin (Corpet 1988) to confirm completeness.

RESULTS

By running an implementation of the clone selection algorithm on EST sequences generated by RGMG consortium, a total of 3988 clones putatively having the full-length sequence has been selected. These clones represented 398 proteins which have not been deposited as *S. mansoni* complete CDS in any public database. From these, 96 clones have been re-sequenced and assembled. Assembly resulted in 33 contigs representing novel *S. mansoni* proteins with sizes ranging from 79 to 375 amino acids with identities in the alignments varying from 30.6% to 78.65% (Table I). These proteins belonged to 11 different KOG biochemical pathways/functional categories as seen in Table II: Energy production and conversion; Intracellular trafficking, secretion, and vesicular transport; Cell cycle control, cell division, chromosome partitioning and cell motility; RNA processing and modification; Defense mechanisms; Amino acid transport and metabolism; Transcription; Translation, ribosomal structure, and biogenesis; Posttranslational modification, protein turnover, chaperones, and Cytoskeleton. Eight proteins have not been classified regarding a biochemical pathway, belonging to the category of General Function or Function Unknown. Selected sequences have been translated and aligned to their orthologs, to verify completeness. Fig. 2 shows an alignment for one of the proteins as an example. Alignments for all proteins can be seen on supplementary material (www.icb.ufmg.br/~alessa/pesquisa/pesquisa.html). Complete sequences have been submitted to NCBI.

DISCUSSION

Full-length cDNAs are extremely useful for determining the genomic structure of genes, especially when analyzed within the context of genomic sequence (Strausberg et al. 2002). Knowing the full-length sequence of a gene allows the prediction of the entire sequence of a protein which can be used for functional and evolutionary studies and for improving the knowledge of the biology of this species. However, the selection of specific cDNAs for full-length sequencing made without the aid of bioinformatics tools is very laborious and time-consuming since it involves the screening of a number of cDNA libraries of variable quality and/or direct strategies to process individual clones (Das et al. 2001). In this study we proposed a method that can be used to increase the availability of full-length sequences, and applied the algorithm to *S. mansoni* sequences. The selection of such clones by computer speeds up the investigation process and sequencing of such clones provides an approach to investigate S. *mansoni* sequences that have not been used by other groups yet. Many initiatives for the investigation of complete CDS in several organisms, including *S. japonicum*, are currently under way and this work now integrates these efforts (Strausberg et al. 1999, Collins 2002, Hu et al. 2003, Baross et al. 2004).

The number of *S. mansoni* complete CDS sequences publicly available before this work (Feb/2006) was of 437 nucleotide sequences described as complete CDS in GenBank flat files and 1108 protein sequences having a suggested CDS pointed in the GenPept files. However, a great number of the CDS sequences available at the moment represent redundant entries in the database. The sequences obtained in this study are unique sequences representing proteins not previously described as complete and therefore representing an important contribution in functional gene characterization of this parasite. The proteins analyzed in our study indicate a high degree of conservation with the orthologs used in the selection with small variations in size and sequence (see alignments

TABLE I

Clones selected for full-length sequencing and orthologs used in selection

Accession No.	Protein	Ortholog ID	Ortholog size (aa)	ORF size (aa)	Identity (%)	Align. size (aa)	Score
DQ480533	Cytochrome c	7298326	108	94	72.62	84	131.0
DQ480534	Vacuolar assembly/sorting protein DID2	7293876	198	79	60.00	75	94.74
DQ480535	40S ribosomal protein S20	Hs4506697	119	117	66.33	98	135.6
DQ480536	60s ribosomal protein L15	Hs15431293	204	227	78.65	192	321.6
DQ480537	Dynein-associated protein Roadblock	Hs7661822	96	97	63.04	92	125.6
DQ480538	Alternative splicing factor SRp20/9G8	Hs4506901	130	211	41.14	158[a]	92.43
DQ489539	40S ribosomal protein S4	Hs4506725	263	264	71.32	258	396.4
DQ480540	Vacuolar H+-ATPase V1 sector, subunit E	Hs4502317	226	161	66.43	140	184.1
DQ480541	60S ribosomal protein L26	Hs4506621	145	142	65.96	141	187.2
DQ480542	B-cell receptor-associated protein	7292004	135	240	36.40	239[a]	153.3
DQ480543	60S ribosomal protein L36	Hs16117794	105	102	57.89	95	114.8
DQ480544	D-Tyr-tRNA (Tyr) deacylase	CE20080	150	170	55.41	148	175.3
DQ480545	Vacuolar sorting protein VPS24	7296875	223	189	36.77	223[a]	151.8
DQ480546	Uncharacterized conserved protein	7300169	119	117	56.14	114	127.1
DQ480547	60s ribosomal protein L34	7301438_1	160	120	61.29	124[a]	155.2
DQ480548	Ribosomal protein RPL1/RPL2/RL4L4	Hs16579885	427	375	61.30	354	454.9
DQ480549	Nuclear DNA-binding protein	7292004	228	140	30.60	134	77.41
DQ480550	CDGSH-type Zn-finger containing protein	Hs8923930	108	118	52.63	95	106.7
DQ480551	Protein tyrosine phosphatase-like	CE13650	271	215	38.26	230[a]	172.2
DQ480552	RNA polymerase III subunit C11	CE25620	108	115	42.48	113	84.73
DQ480553	U2 snSRNP Auxiliary Factor - small subunit	Hs5032083	298	221	76.47	221	352.8
DQ480554	Predicted membrane protein	Hs18559233	149	148	51.80	139	166.4
DQ480555	Clathrin adaptor complex, small subunit	CE09797	157	157	69.87	156	240.0
DQ480556	Nuclear transport factor 2	Hs5031985	127	129	40.83	120	96.29
DQ480557	Uncharacterized conserved protein DREV/CGI-8	Hs19923449	283	294	45.32	278	226.1
DQ480558	U2 snRNP splicing factor, small subunit, and related proteins	7296221	264	165	62.07	116	141.4
DQ480559	60S ribosomal protein L14	YHL001w	138	146	46.72	122	124.4
DQ480560	U-snRNP-associated cyclophilin type peptidyl-prolyl cis-trans isomerase	Hs5454154	177	174	75.14	173	288.9
DQ480561	Ubiquitin/40S ribosomal protein S27a fusion	Hs4506713	156	153	53.38	148	165.2
DQ480562	Predicted actin-bundling protein	7293670	262	288	40.83	289[a]	205.3
DQ480563	Predicted membrane protein	Hs7657595	157	152	50.75	134	140.2
DQ480564	ER lumen protein retaining receptor	Hs5803050	212	195	69.52	210	303.5
DQ480565	Transcription elongation factor SPT4	Hs4507311	117	117	55.86	111	149.1

a: alignment sizes larger than ORF sizes are due to insertions of gaps.

TABLE II

Distribution of full-length cDNA clones according to KOG biochemical pathway/functional category

Biochemical pathway/Functional category	No. of *Schistosoma mansoni* complete proteins
Energy production and conversion	2
Intracellular trafficking, secretion, and vesicular transport	5
Cell cycle control, cell division, chromosome partitioning and cell motility	1
RNA processing and modification	3
Defense mechanisms	1
Amino acid transport and metabolism	1
Transcription,	2
Translation, ribosomal structure and biogenesis	9
Posttranslational modification, protein turnover, chaperones	1
Cytoskeleton	1
General function	2
Function unknown	5
Total	33

```
1                                                  50
        Hs15431293  MGAYKYIQEL WRKKQSDVMR FLLRVRCWQY RQLSALHRAP RPTRPDKARR
     GF_AW2_013_B04 MGAYKFMQEL YRRKQSDVMR YIQRMRCWEY RQSHAIVRVT RPSRPDRARR
          Consensus  MGAYK%iQEL wRrKQSDVMR %iqRmRCW#Y RQlhAihRap RPsRPDrARR

                     51                                                 100
        Hs15431293  LGYKAKQGYV IYRIRVRRGG RKRPVPKGAT YGKPVHHGVN QLKFARSLQS
     GF_AW2_013_B04 LGYKAKQGYV IYRIRIRRGG RKRQVPKGAT YGKPAGEGVN QLKNQRSIQA
          Consensus  LGYKAKQGYV IYRIR!RRGG RKRqVPKGAT YGKPageGVN QLKnaRSiQa

                     101                                                150
        Hs15431293  VAEERAGRHC GALRVLNSYW VGEDSTYKFF EVILIDPFHK AIRRNPDTQW
     GF_AW2_013_B04 VAEERVGRKC GALRILNSYW VGEDSTFKFF EVICIDPFHK AIRRDPQIRW
          Consensus  VAEERaGRhC GALR!LNSYW VGEDST%KFF EVIcIDPFHK AIRR#P#irW

                     151                                                200
        Hs15431293  ITKPVHKHRE MRGLTSAGRK SRGLGKGHKF HHTIGGSRRA AWRRRNTLQL
     GF_AW2_013_B04 ICKAHQKHRE MRGLTSAGSK SRGLGKGHKF HKTIGGSRRA AWDAPQYCRD
          Consensus  IcKahqKHRE MRGLTSAGrK SRGLGKGHKF HhTIGGSRRA AWrar#tcrd

                     201                        227
        Hs15431293  HRYR...... .......... .......
     GF_AW2_013_B04 ATKALKIHLV TCLRYAKKIK KNNKQKK
          Consensus  arka...... .......... .......
```

Fig. 2: alignment of the translated ORF of a *Schistosoma mansoni* full-length cDNA to its ortholog using Multalin. Hs15431293 – ortholog protein; GF_AW2_013_B04 – *S. mansoni* cDNA – RGMG clone (accession number DQ480536). Identical amino acids are shown in bold, capital letters. For non identical amino acids the following symbols are used: % and !: conservative substitutions; #: non conservative substitutions.

in supplementary material). This is also suggested by the values of identities in the alignments presented in Table I. We have not observed a high variation in any specific portion of the proteins studied, although previous studies suggested a higher degree variation in the amino terminal portion of eukaryotes proteins (Kisselev & Frolova 1995). We observed that the size of the proteins for which the full-length sequences have been obtained was relatively small, as should be expected given the characteristics of the procedure used for library construction for production of ESTs.

In this study we have proposed a method that can be used to increase the availability of full-length sequences and used this method to study *S. mansoni* sequences. This represents an important contribution in the research on this parasite, for which the complete functional characterization of the genome is still under way.

ACKNOWLEDGMENTS

To Minas Gerais Genome Network for providing ESTs for analysis.

REFERENCES

Adams MD, Kelley JM, Goccaayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science 252*: 1651-1656.

Baross A, Butterfield YS, Coughlin SM, Zeng T, Griffith M, Griffith OL, Petrescu AS, Smailus DE, Khattra J, McDonald HL, McKay SJ, Moska M, Holt RA, Marra MA 2004. Systematic recovery and analysis of full-ORF human cDNA clones. *Genome Research 14*: 2083-2092.

Brazilian National Genome Project Consortium 2003. The complete geneome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability. *Proc Natl Acad Sci USA 100*: 11660-11665.

Collins FS 2002. Generation and analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci USA 99*: 16899-16903.

Conesa A, Götz S, Gracía-Gómez JM, Terol J, Talon M, Robles M 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomic research. *Bioinformatics 21*: 3674-3676.

Corpet F 1988. Multiple sequences alignment with hierarchical clustering. *Nucleic Acids Res 16*: 10881-10890.

Das M, Harvey I, Chu LL, Sinha M, Pelletier J 2001. Full-length cDNAs: more then just reaching ends. *Physiol Genomics 6*: 57-80.

Del Val C, Glatting K, Suhai S 2003. cDNA2 Genome: a tool for mapping and annotating cDNAs. *BMC Bioinformatics 4*: 39.

Furuno M, Kasukawa T, Saito R, Adachi J, Suzuki H, Baldarelli R, Hayashizaki Y, Okazaki Y 2003. CDS annotation in full-length cDNA sequence. *Genome Research 13*: 1478-1487.

Hötz-Wagenblatt Hankeln T, Ernst P, Glatting KH, Schmidt ER, Suhai S 2003. ESTAnnotator: a tool for high throughput EST annotation. *Nucleic Acids Res 31*: 3716-3719.

Hu W, Yan Q, Shen DK, Liu F, Zhu ZD, Song HD, Xu XR, Wang ZJ, Rong YP, Zeng LC, Wu J, Zhang X, Wang JJ, Xu XN, Wang SY, Fu G, Zhang XL, Wang ZQ, Brindley PJ, McManus DP, Xue CL, Feng Z, Chen Z, Han ZG 2003. Evolutionary and biomedical implications of a *Schistosoma*

japonicum complementary DNA resource. *Nat Genet 35*: 139-147.

Kisselev LL, Frolova LY 1995. Termination of translation in eukaryotes. *Biochem Cell Biol 73*: 1079-1086.

Nishikawa T, Ota T, Isogai T 2000. Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. *Bioinformatics 16*: 960-967.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schonbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, Brusic V, Chothia C, Corbani LE, Cousins S, Dalla E, Dragani TA, Fletcher CF, Forrest A, Frazer KS, Gaasterland T, Gariboldi M, Gissi C, Godzik A, Gough J, Grimmond S, Gustincich S, Hirokawa N, Jackson IJ, Jarvis ED, Kanai A, Kawaji H, Kawasawa Y, Kedzierski RM, King BL, Konagaya A, Kurochkin IV, Lee Y, Lenhard B, Lyons PA, Maglott DR, Maltais L, Marchionni L, McKenzie L, Miki H, Nagashima T, Numata K, Okido T, Pavan WJ, Pertea G, Pesole G, Petrovsky N, Pillai R, Pontius JU, Qi D, Ramachandran S, Ravasi T, Reed JC, Reed DJ, Reid J, Ring BZ, Ringwald M, Sandelin A, Schneider C, Semple CA, Setou M, Shimada K, Sultana R, Takenaka Y, Taylor MS, Teasdale RD, Tomita M, Verardo R, Wagner L, Wahlestedt C, Wang Y, Watanabe Y, Wells C, Wilming LG, Wynshaw-Boris A, Yanagisawa M, Yang I, Yang L, Yuan Z, Zavolan M, Zhu Y, Zimmer A, Carninci P, Hayatsu N, Hirozane-Kishikawa T, Konno H, Nakamura M, Sakazume N, Sato K, Shiraki T, Waki K, Kawai J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Imotani K, Ishii Y, Itoh M, Kagawa I, Miyazaki A, Sakai K, Sasaki D, Shibata K, Shinagawa A, Yasunishi A, Yoshino M, Waterston R, Lander ES, Rogers J, Birney E, Hayashizaki Y, FANTOM Consortium, RIKEN Genome Exploration Research Group Phase I & II Team 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature 420*: 563-573.

Saghatelian A, Cravatt BF 2005. Assignment of protein function in the postgenomic era. *Nat Chem Biol 1*: 130-142.

Salamov AA, Nishikawa T, Swindells MB 1998. Assessing protein coding region integrity in cDNA sequencing projects. *Bioinformatics 14*: 384-390.

Strausberg RL, Feingold EA, Klausner RD, Collins FS 1999. The mammalian gene collection. *Science 286*: 455-457.

Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF, Zeeberg B, Buetow KH, Schaefer CF, Bhat NK, Hopkins RF, Jordan H, Moore T, Max SI, Wang J, Hsieh F, Diatchenko L, Marusina K, Farmer AA, Rubin GM, Hong L, Stapleton M, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Brownstein MJ, Usdin TB, Toshiyuki S, Carninci P, Prange C, Raha SS, Loquellano NA, Peters GJ, Abramson RD, Mullahy SJ, Bosak SA, McEwan PJ, McKernan KJ, Malek JA, Gunaratne PH, Richards S, Worley KC, Hale S, Garcia AM, Gay LJ, Hulyk SW, Villalon DK, Muzny DM, Sodergren EJ, Lu X, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madan A, Young AC, Shevchenko Y, Bouffard GG, Blakesley RW, Touchman JW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Krzywinski MI, Skalska U, Smailus DE, Schnerch A, Schein JE, Jones SJ, Marra MA, Mammalian Gene Collection Program Team 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci USA 99*: 16899-16903.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics 4*: 41.

Wernersson R, Pedersen AG 2003. RevTrans - Constructing alignments of coding DNA from aligned amino acid sequences. *Nucl Acids Res 31*: 3537-3539.