Clustering of *Schistosoma mansoni* mRNA Sequences and Analysis of the Most Transcribed Genes: Implications in Metabolism and Biology of Different Developmental Stages

Francisco Prosdocimi, Alessandra C Faria-Campos*, Fabiano C Peixoto, Sérgio DJ Pena*, José M Ortega*, Glória R Franco*/+**

Departamento de Biologia Geral *Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, **Laboratório de Computação Científica, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901 Belo Horizonte, MG, Brasil.

The study of the Schistosoma mansoni *genome, one of the etiologic agents of human schistosomiasis, is essential for a better understanding of the biology and development of this parasite. In order to get an overview of all* S. mansoni *catalogued gene sequences, we performed a clustering analysis of the parasite mRNA sequences available in public databases. This was made using softwares PHRAP and CAP3. The consensus sequences, generated after the alignment of cluster constituent sequences, allowed the identification by database homology searches of the most expressed genes in the worm. We analyzed these genes and looked for a correlation between their high expression and parasite metabolism and biology. We observed that the majority of these genes is related to the maintenance of basic cell functions, encoding genes whose products are related to the cytoskeleton, intracellular transport and energy metabolism. Evidences are presented here that genes for aerobic energy metabolism are expressed in all the developmental stages analyzed. Some of the most expressed genes could not be identified by homology searches and may have some specific functions in the parasite.*

Key words: *Schistosoma mansoni* - bioinformatics - expressed sequences tag - clustering analysis - metabolism

Schistosoma mansoni is a dioiceous trematode and one of the etiologic agents of schistosomiasis, the second more significant tropical disease concerning public health. Despite recent efforts undertaken to contain its progress, the disease is still endemic in several countries, with around 200 million people infected by the parasite (http://www.who.int/ctd/schisto/epidemio.htm). The study of *S. mansoni* is, therefore, very important in human parasitology. Gaining knowledge on the genome of this parasite is essential for a better understanding of its metabolism and biology and will help to elucidate important aspects of the mechanisms of drug resistance and antigenic variation that allow it to escape from the host immune system (Franco et al. 2000).

The size of *S. mansoni* genome is estimated in 270Mb with the number of expressed genes ranging from 15000 to 20000 (Simpson et al. 1982, Franco & Simpson 2001). Although some genomic sequences of *S. mansoni* have been produced, the *Schistosoma* Genome Network (SGN) has chosen as first priority the sequencing of cDNA using the expressed sequence tags (ESTs) strategy, from which is possible to obtain fast and relevant information about the genes of the parasite (Franco et al. 2000). Since the creation of the SGN, sponsored by the World Health Organization, 16813 *S. mansoni* ESTs have been deposited in the EST database (dbEST release 031502, http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

Although resulting in fast and very important information, ESTs available from public databases, such as dbEST, show some degree of redundancy and present a great number of errors, because they are single pass sequences (Miller et al. 1999). To overcome these problems and to increase the length of the sequences, facilitating identification by homology searches, clustering procedures are performed (Oliveira & Johnston 2001). In this kind of procedure, sequences that have some region of similarity are joined into a cluster. Therefore, sequences possessing overlapping regions and representing a single gene are joined into the same cluster, decreasing redundancy. Sequences of each cluster are then aligned to generate a consensus sequence. In this approach, the base (and, if available, the quality value designated by the base caller program) present in each sequence position is considered in the construction of a high quality consensus (Huang & Madan 1999). The clustering procedure can, therefore, have two outcomes: consensus are generated by the alignment of the sequences of a cluster and singlets result from sequences that have not been grouped to any others. Theoretically, each sequence (either a consensus or a singlet) should represent an individual gene, and so, these sequences are called uniques. As is expected that each sequence represents a single gene, the comparison of the number of uniques with the total number of predicted genes make it possible to know, approximately, how many genes have not been discovered yet.

This work was supported by a grant from Brazilian Research Council (CNPq).
+Corresponding author. Fax: +55-31-3499.2984. E-mail: gfranco@icb.ufmg.br.
Received 18 June 2002
Accepted 15 August 2002

In order to identify which gene is represented by each unique, a process of sequence annotation should be performed. Among other approaches, this can be carried out by homology searches, using BLAST programs, against nucleotide and protein databases (Stein 2001).

In this work we performed a clustering analysis of *S. mansoni* sequences available in public databases in order to investigate roughly how many of the parasite genes have been sequenced so far. We also investigated which of these genes are highly expressed and looked for a correlation between these genes and the biology of the worm. Attempts to establish an association between genes frequently transcribed in the parasite and the developmental stages they are expressed may result in the identification of stage-specific genes. This information could then be used in a strategy to specifically repress the gene expression in such stage of the worm life cycle, leading to the blockage of the parasite development.

MATERIALS AND METHODS

S. mansoni sequences - All *S. mansoni* sequences were obtained from National Center for Biotechnology Information - NCBI - US (http://www.ncbi.nlm.nih.gov) in FASTA format in May/2001. The number of sequences present in the databases at that moment was 14275, which includes ESTs (dbEST) as well as complete coding sequences (GenBank).

Treatment of sequences - Although GenBank sequences have a quality assurance review we performed a filtering procedure to avoid contaminants that could generate errors in sequence clustering. Vector regions were masked using the software cross_match, used according to documentation instructions (http://www.phrap.org/phrap.docs/phrap.html). For this purpose, a vector file was generated containing three of the most used cDNA library cloning vectors (pGEM, pBluescript SK- and pUC 18). Besides that, two PERL scripts were developed to filter the sequences. The first (filter1.pl), was used to remove small (< 150 nucleotides) and poor quality (> 4% ambiguities) sequences (Franco et al. 1997b). The second script (filter2.pl) was used to remove regions such as cloning enzymes restriction sites, linkers, poly Ts and poly As.

Clustering procedure - To perform sequence clustering, we used the softwares PHRAP and CAP3 (Huang & Madan 1999). We performed preliminary tests to adjust the parameters of the softwares to more stringent values, which produced better results than the default parameters (data not shown). In this way, only sequences that overlap in at least 40 bases with 100% sequence identity were joined in the same cluster. Two PERL scripts were generated to count the number of sequences present in each CAP3 or PHRAP cluster. Clusters containing the same number of sequences were grouped in classes, and half of the total number of classes was analyzed. These classes were considered to contain clusters of the most frequently expressed genes.

Homology searches - The BLAST software was run locally with all default flags and an e-value cut-off of 10^{-10} to identify the uniques representing the highly expressed genes produced by the clustering procedure. Databases used in the BLAST procedure were the non-redundant nucleotide (nt) and protein (nr) databases of NCBI (Weller et al. 2002).

RESULTS

Treatment of sequences - After the treatment of the sequences with cross_match, the files containing the masked sequences were submitted to filter1 script and results can be seen in Table I. Of the 14275 sequences used in this step, 4.7% have more than 4% ambiguities and 4.3% have less than 150 nucleotides. Thus, of the total sequences submitted to round one of filtering, 12982 had an acceptable number of ambiguous bases and were not too small. They were submitted to filter2 and the results of this second round of filtering can be seen in Table II. Of the 12982 sequences used in this step, 22.5% had regions of linkers, 1.2% had the *Xho*I cloning site, 10.5% had poly Ts in the beginning of the minus strand sequence and 0.96% had poly As in the end of the plus strand sequence.

Clustering procedure - The results of the clustering procedure can be seen in Table III. Using the software CAP3, 70.1% of the 12982 sequences were clustered, producing 1832 consensus. On the other hand, using PHRAP, 66.6% of the 12982 sequences were clustered producing 1763 consensus. Therefore, 5708 uniques were produced

TABLE I

Number of sequences rejected by filter1.pl

	Input File	More than 4% ambiguities [a]	Less than 150 nucleotides	Output File
Number of sequences	14275	677 (4.7%)	616 (4.3%)	12982 (90.9%)

a: if the sequence has both more than 4% ambiguities and small size it is included only in this class.

TABLE II

Number of sequences modified by filter2.pl

	Linker	*Xho*I Sites 5'or 3'	Poly T 5'	Poly A 3'
Number of sequences	2923 (22.5%)	160 (1.2%)	1,366 (10.5%)	125 (1%)

TABLE III
Clustering analysis by CAP3 and PHRAP

Software [a]	Number of singlets	Number of contigs	Number of uniques [b]	Number (%) of clustered sequences
CAP3	3876	1832	5708	9106 (70.1%)
PHRAP	4237	1763	6000	8652 (66.6%) [c]

a: both softwares were used at more stringent parameters (see methodology); *b*: number of singlets + number of contigs; *c*: if the number of singlets and sequences grouped by the PHRAP clustering procedure is added, the total number of sequences is 12889. This total is 93 sequences fewer than the expected number (12982 sequences from the input file). That can be explained because these 93 sequences are duplicated in the input sequences and were discarded during the clustering procedure performed by this software.

using CAP3 and 6000 uniques were produced using PHRAP.

Identification of the most expressed genes - Both softwares produced 51 classes of clusters regarding to the number of sequences joined, as can be seen in Table IV. The clusters representing highly expressed genes were the ones with more sequences than the clusters at the class 25 (presenting 25 sequences). Therefore, all contigs produced by the alignment of more than 25 sequences were identified and annotated for CAP3 (Table V) and PHRAP (Table VI).

DISCUSSION

Regarding the data shown in Tables I and II, it can be seen that a great number of *S. mansoni* sequences available in public databases were rejected by the quality criteria used in this analysis (close to 10%) and also contain regions of linkers, *Xho*I cloning sites, poly Ts 5' and poly As 3' (more than 35%). These "contaminant" regions are necessary in the process of cDNA library construction. Some of them are not removed by softwares based on sequences comparison, such as cross_match (used in the filter procedure), because of their small size. Considering the great number of sequences presenting these "contaminants", the use of softwares developed specifically to remove or mask them, like the PERL script filter2 performed here, is essential.

The softwares used here for clustering of sequences produced a very similar number of uniques (around 6000) with a slight difference of approximately 0.05% (Table III). This difference is probably due to distinct specifications in the algorithms of the softwares and to different values of matches, mismatches and gaps scores being used during sequence alignment procedures. This non-redundant number of uniques is, therefore, the most recent estimate that can be made about the number of genes already sequenced in the parasite. Thus, considering the putative number of genes present in the worm (15000-20000), it can be inferred that approximately 30%-40% of *S. mansoni* genes is already available in public databases.

Other authors obtained a different number of clusters for *S. mansoni* sequences. The researchers from TIGR suggested 1711 clusters and 4985 singlets as a result of their clustering procedure using CAP3. However, their initial number of sequences was 13540 and their clusters contain sequences that overlap for at least 40 bases, with more than 95% sequence identity (http://www.tigr.org/tdb/tgi_info.html; *Schistosoma mansoni* Gene Index *release* 3.0 - September 06, 2001- http://www.tigr.org/tdb/smgi). Considering the level of stringency used in our analysis and the small size and quality of the sequences removed by the filtering procedure, it is probable that great part of these filtered sequences are present as singlets in TIGR gene index. The WHO/UNDP/World Bank *Schistosoma* Genome Network, release 4 (March/2000) suggested 1830 clusters and 5439 singlets, generated from 13154 input sequences (http://www.nhm.ac.uk/hosted_sites/schisto). Their clustering procedure was carried out using the software Sequencer 3.1.1, where sequences with more than 90% of similarity in regions greater than 60 bases were grouped. In this case, it is possible that the software used for clustering is more stringent than the ones used by us, which explain the discrepancies observed between the results.

Regarding the annotation of the most expressed genes, we observed that, as expected, the majority of them belong to the category of housekeeping genes. For carbohydrate metabolism, three contigs with great number of sequences representing genes from the glycolytic pathway were found, when using either CAP3 or PHRAP: glyceraldehyde-3-phosphate dehydrogenase (105 sequences), fructose 1,6 bisphosphate aldolase (68 sequences) and enolase (26 sequences) (Tables V, VI). It is interesting to notice that PHRAP and CAP3 joined exactly the same *S. mansoni* mRNA and EST sequences to construct these clusters, which increases the confidence of the result. Another interesting point regarding the glycolytic pathway enzymes is the high proportion of these genes expressed in cercariae. This agrees well with previous work from our group suggesting that in this developmental stage the parasite needs a high amount of energy for rapid swimming and for body contraction in order to find and penetrate the host skin (Santos et al. 1999).

A great number of sequences representing three subunits of the enzyme cytochrome oxidase (CO), the fourth complex of the respiratory chain and also responsible for proton pumping across the inner mitochondrial membrane, were found in clusters constructed by both softwares. These are genes from the mitochondrial genome and because of their large sizes our analysis produced more than one cluster, formed by different number of sequences, sharing the same gene annotation. This suggests that the three clusters of cytochrome oxidase I (COI) produced by CAP3 represent different parts of the same gene. This reasoning does not apply to the two clusters of COI, COII and COIII produced by PHRAP as was observed by the

alignment of these contigs to the complete mitochondrial DNA sequence of *S. mansoni* (data not shown). In this case, it seems that the two clusters of each gene represent the same region. The high expression of these genes supports the idea that *S. mansoni*, in all developmental stages analyzed here, is an organism presenting, mainly, aerobic metabolism. Previous studies have described *S. mansoni* adult worms as presenting anaerobic metabolism (Bueding & Fisher 1982, Thompson et al. 1984, Rumjanek 1987, Van Oordt et al. 1989). This idea was corroborated by the observation that high amounts of lactate are produced by the worms (Rumjanek 1987). It is important to notice that the enzyme lactate dehydrogenase was identified as one of the most expressed genes,

TABLE IV

Distribution of PHRAP and CAP3 uniques by number of sequences

No. sequences	No. PHRAP uniques	*Class* [a]	No. sequences	No. CAP3 uniques
1	4237	*1*	1	3876
2	830	*2*	2	846
3	339	*3*	3	363
4	183	*4*	4	193
5	105	*5*	5	110
6	73	*6*	6	85
7	33	*7*	7	38
8	32	*8*	8	30
9	25	*9*	9	27
10	15	*10*	10	14
11	20	*11*	11	20
12	13	*12*	12	10
13	8	*13*	13	12
14	5	*14*	14	3
15	14	*15*	15	12
16	5	*16*	16	10
17	6	*17*	17	6
18	5	*18*	18	3
19	3	*19*	19	3
20	2	*20*	20	5
21	2	*21*	21	2
22	1	*22*	22	3
23	5	*23*	23	4
24	1	*24*	24	1
25	1	***25***	25	2
26	1	***26***	26	1
29	3	***27***	27	1
30	2	***28***	29	1
31	1	***29***	30	3
32	2	***30***	31	1
35	1	***31***	33	1
36	1	***32***	37	1
37	2	***33***	38	1
39	1	***34***	41	1
40	2	***35***	44	1
41	1	***36***	48	1
42	2	***37***	49	1
44	2	***38***	53	1
47	1	***39***	54	1
51	2	***40***	61	1
57	1	***41***	68	2
61	1	***42***	69	2
68	3	***43***	70	1
70	1	***44***	84	1
81	1	***45***	91	1
101	1	***46***	105	1
105	1	***47***	112	1
111	1	***48***	132	1
114	1	***49***	158	1
118	1	***50***	166	1
145	1	***51***	286	1

a: number of classes used to define the analyzed clusters. In bold we see the most expressed uniques. Their identification is shown in Tables V and VI.

when clusters were constructed using PHRAP, and this could explain the above assumption, reinforcing the idea that *S. mansoni* may produce part of its energy through anaerobic oxidation of glucose. However, the number of sequences representing subunits of cytochrome oxidase, in all developmental stages analyzed here, is greater than the number of lactate dehydrogenase sequences. Therefore, our results agree with other authors that describe an interruption of the oxygen consumption in an adult worm extract when it is submitted to mitochondrial inhibitors, demonstrating the existence of an active respiratory chain (Coles 1972).

As for specific developmental stages, our results showing a high number of cytochrome oxidase sequences in cercariae are in accordance with previous studies that verified a high expression of enzymes from the tricarboxylic acid cycle and oxidative phosphorylation in this stage of the parasite life cycle (Skelly et al. 1993). Our results agree as well with other studies about the consumption of oxygen and production of carbon dioxide from metabolized glucose in this stage (Van Oordt et al. 1989, Bruce et al. 1971). Concerning the lung stage, not even a single sequence representing lactate dehydrogenase was found. Meanwhile, sequences representing the subunits 1 and 2 of cytochrome oxidase were present. Thus, it seems that the worm is capable of carrying aerobic metabolism also in this stage. To confirm these results, however, it is necessary to obtain more sequences or construct other libraries from the same stage, once the number of ESTs generated from it is still small.

TABLE V

Annotation of CAP3 contigs with higher number of sequences

Annotation	No. sequence [a]	E [b]	C [c]	L [d]	M [e]	F [f]	A W [g]	? [h]
Clustering error	286	11	4	5	45	126	93	2
Similar to *Schistosoma mansoni* cytochrome c oxidase subunit 1 (COI)	166	4	35	2	17	36	71	1
Similar to *S. mansoni* cytochrome c oxidase subunit 2 158 (COII)	7	19	2	24	44	62	0	
Similar to *S. mansoni* cytochrome c oxidase subunit 1 (COI)	132	6	13	3	29	49	32	0
Similar to *S. mansoni* eggshell (chorion) protein	112	2	3	0	6	67	33	1
S. mansoni glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	105	2	20	4	12	4	63	0
Similar to *S. mansoni* cytochrome c oxidase subunit 3 (COIII)	91	7	8	0	3	6	67	0
Similar to *S. mansoni* calcium binding protein mRNA	84	0	81	0	0	0	2	1
Putative gene for *S. mansoni* heat shock protein 90	70	1	0	0	2	18	48	1
Similar to *S. mansoni* cathepsin B (Sm31)	69	0	4	1	19	23	21	1
Similar to *S. mansoni* fibrillin 2 gene	69	11	3	2	12	15	24	2
Similar to *S. mansoni* elongation factor 1- alpha gene	68	0	1	2	12	25	28	0
Similar to *S. mansoni* fructose 1,6 bisphosphate aldolase gene	68	2	24	1	11	7	21	2
Similar to *S. mansoni* hemoglobinase precursor (antigen SM32)	61	0	0	0	8	23	27	3
Similar to *S. mansoni* mRNA gene for eggshell protein	54	0	0	0	0	20	34	0
Similar to *S. mansoni* actin 2	53	2	5	0	6	6	34	0
Similar to *S. mansoni* actin	49	2	1	0	5	16	24	1
Similar to *S. mansoni* Pro-His-rich protein	48	0	0	0	0	17	30	1
Similar to *S. mansoni* myosin heavy chain (MYH)	44	0	0	1	24	7	10	2
Similar to *S. mansoni* glutathione S- transferase 28 Kd (GST 28) (SM28 antigen)	41	2	1	0	3	3	30	2
Putative gene for myosin regulatory light chain A	38	0	0	0	8	2	28	0
Similar to *S. mansoni* cytochrome c oxidase subunit 1 (COI)	37	2	10	1	6	6	12	0
Similar to *Homo sapiens* neuroendocrine- specific protein B	33	0	0	0	0	6	27	0
Similar to *S. mansoni* unknown protein	31	0	30	0	0	0	0	1
Putative gene for dynein light chain 2 (8 Kda dynein light chain)	30	13	0	0	3	3	11	0
Similar to *S. mansoni* tubulin alpha	30	2	2	1	7	6	11	1
Unknown	30	0	30	0	0	0	0	0
Similar to *S. mansoni* Y-box binding protein gene	29	0	3	0	1	1	23	1
Similar to *S. mansoni* actin 2 gene	27	2	0	0	1	1	23	0
Putative gene for enolase (2-phosphoglycerate dehydratase)	26	0	3	0	8	4	10	1
Putative gene for eggshell protein precursor	25	0	0	0	0	19	6	0
Unknown	25	1	0	0	2	11	11	0
Total	2189	79	300	25	274	571	916	24

a: total number of sequences in the cluster; *b*: number of sequences found in egg libraries; *c*: number of sequences found in cercariae libraries; *d*: number of sequences found in lung stage libraries; *e*: number of sequences found in male adult worm libraries; *f*: number of sequences found in female adult worm libraries; *g*: number of sequences found in mixed adult worm libraries; *h*: number of sequences where library information could not be retrieved.

Regarding cell motility and the cytoskeleton, we observed the presence of three clusters representing actin genes (two similar to actin 2 and one similar to actin 1), in the clustering results obtained by both softwares. Previous studies have shown that the abundance of both actins is greater in males than in females, being less represented in eggs and cercariae (Davies et al. 1985). Considering that the total number of female sequences in our analysis is twice that of male sequences, it seems that actin 2 is more expressed in males, while actin 1, in females. However, the number of sequences used here is still low and more sequences should be produced to support this statement.

At this point, it is interesting to notice that the third cluster with the highest number of sequences produced by PHRAP, represents an actin related protein (ARP1),

TABLE VI

Annotation of PHRAP contigs with higher number of sequences

Annotation	No. sequence [a]	E [b]	C [c]	L [d]	M [e]	F [f]	AW [g]	? [h]
Similar to *Schistosoma mansoni* cytochrome c oxidase subunit 1 (COI)	145	9	2	4	34	47	49	0
Clustering error	118	6	0	0	21	61	29	1
Similar to Homo sapiens ARP1 actin- related protein 1 homolog B, centractin beta	114	2	0	4	23	58	27	0
Similar to *S. mansoni* cytochrome c oxidase subunit 2 (COII)	111	6	0	2	24	44	35	0
Similar to *S. mansoni* glyceraldehyde 3-phosphate dehydrogenase (gapdh)	105	2	20	4	12	4	63	0
Similar to *S. mansoni* eggshell protein	101	0	0	0	6	67	26	2
Similar to *S. mansoni* calcium-binding protein	81	0	78	0	0	0	2	1
Putative gene for *S. mansoni* heat shock protein 90	70	1	0	0	2	18	48	1
High similar to *S. mansoni* elongation factor 1-alpha	68	0	1	2	12	24	29	0
Similar to fructose 1,6-bisphosphate aldolase	68	2	24	1	11	7	21	2
Similar to *S. mansoni* fibrillin 2	68	11	3	2	12	15	24	1
Similar to *S. mansoni* cytochrome c oxidase subunit 1 (COI)	61	1	2	1	12	37	8	0
Unknown	57	0	30	0	0	0	26	1
S. mansoni myosin heavy chain	51	1	0	1	25	7	15	2
Similar to *S. mansoni* eggshell protein	51	0	0	0	0	20	31	0
Similar to *S. mansoni* Pro-His-rich protein	47	0	0	0	0	17	29	1
S. mansoni actin 1	44	2	1	0	5	15	20	1
Similar to *S. mansoni* cytochrome c oxidase subunit 3 (COIII)	44	0	8	0	0	0	36	0
S. mansoni actin 2	42	3	5	0	4	2	28	0
Similar to *S. mansoni* cytochrome c oxidase subunit 2 (COII)	42	0	17	0	0	0	25	0
Similar to *S. mansoni* cytochrome c oxidase subunit 3 (COIII)	41	6	0	0	3	6	26	0
S. mansoni glutathione S-transferase 28 kd (GST 28) (SM28 antigen)	40	2	1	0	3	3	29	2
High similar to *S. mansoni* cathepsin B like cysteine proteinase precursor (antigen SM31)	40	0	2	1	12	14	11	0
S. mansoni actin 2	39	1	0	0	3	5	30	0
Similar to *S. mansoni* hemoglobinase precursor (Antigen SM32)	37	0	0	0	4	11	21	1
Similar to *S. mansoni* AUT1	37	0	0	0	1	3	33	0
Putative gene for myosin regulatory light chain	36	0	0	0	8	2	26	0
Unknown	35	2	10	1	6	6	10	0
Unknown	32	0	4	0	0	0	28	0
Putative gene for L lactate dehydrogenase	32	1	4	0	5	10	11	1
Similar to *S. mansoni* neuroendocrine-specific protein B	31	0	0	0	0	4	27	0
S. mansoni cathepsin B-like cysteine proteinase precursor (antigen SM31)	30	0	2	0	7	9	11	1
Similar to *S. mansoni* unknown protein	30	0	29	0	0	0	0	1
S. mansoni alpha tubulin	29	2	2	1	7	6	10	1
Similar to *S. mansoni* Y-box binding protein	29	0	3	0	1	1	23	1
Unknown	29	0	29	0	0	0	0	0
S. mansoni enolase (2-phosphoglycerate dehydratase)	26	0	3	0	8	4	10	1
Putative gene for dynein light chain	25	9	0	0	3	3	10	0
Total	2086	69	280	24	274	530	887	22

a: total number of sequences in the cluster; *b*: number of sequences found in egg libraries; *c*: number of sequences found in cercariae libraries; *d*: number of sequences found in lung stage libraries; *e*: number of sequences found in male adult worm libraries; *f* : number of sequences found in female adult worm libraries; *g*: number of sequences found in mixed adult worm libraries; *h*: number of sequences where library information could not be retrieved.

and this cluster have not been found among the most expressed sequences clustered by CAP3. This may be explained by the fact that ARP1 sequences were grouped by CAP3 with other ones containing a minisatellite region on them, forming then a cluster with an excessive number of sequences (the first cluster of Table V). Sequences harboring minisatellites also produced a clustering error for the group with the second highest number of sequences, when PHRAP was used.

Despite of this problem, ARP1 was identified as one of most expressed genes by PHRAP. This protein is the major subunit of dynactin, a multiprotein complex that acts, in vitro, as a cofactor for dynein-mediated movement of vesicles along the microtubules (Frankel & Mooseker 1996). The presence of both alpha tubulin and dynein light chain between the most expressed genes could suggest the great importance of vesicular transport for *S. mansoni*. Other motor proteins were identified among the most expressed genes: a myosin heavy chain and a myosin regulatory protein. These motor proteins are responsible for a great number of intracellular movements essential for reproduction and survival of an eukaryotic cell (Valle & Gee 1998).

More than 40 years ago, a classic experiment showed that the main source of amino acids for schistosome adult worms is the digestion of hemoglobin (Timms & Bueding 1959). Since then, a significant number of proteases discovered in the worm were classified as hemoglobinases, although the precise proteolytic pathway for the degradation of hemoglobin is still unsolved (Brindley et al. 1997). Among the worm cysteine proteinases, the proteins Sm31 (similar to mammalian cathepsin B) and Sm32 (similar to asparaginyl endopeptidases, a family of cysteine proteinases) (Klinker et al. 1989) are the best characterized and represent a class of highly immunogenic proteins, both used in schistosomiasis diagnosis (Li et al. 1996). Two contigs representing the enzyme Sm31 and one representing Sm32 were found among the most expressed genes produced by PHRAP. As for those produce by CAP3, we found just one cluster for each of these enzymes. Looking at the developmental stages where these proteins were expressed, we found Sm32 only in adult worms, whereas Sm31 is found also in cercariae and lung stage.

The production of eggs by the worm pairs and their deposit in the liver and intestine is a major cause of pathogenesis in schistosomiasis (Chen et al. 1992). Two eggshell proteins were found among the most expressed sequences grouped by PHRAP and CAP3. These are proteins expressed, usually, by mature females during the egg production. However, we found these proteins in male, egg and cercariae libraries. Analysis of the alignments used by CAP3 to construct this cluster, showed that the egg and cercariae sequences were clustered in this group because a CT microsatellite present in some sequences. The male sequences, on the other hand, were clustered correctly. An interesting fact is that all male sequences come from the same library. It is possible that this library contained some female contaminant sequences, as it is very difficult to separate the worm pairs for mRNA isolation.

One chaperone molecule was found among the *S. mansoni* most expressed genes. The consensus was similar to the heat shock protein 90 (Hsp90) of different organisms. Hsp90 is a molecular chaperone, which seems to operate in a cytoplasmic multichaperone machinery, including Hps70, peptidyl-prolyl isomerases and other co-chaperone like protein kinases. The clustering of the same Hsp90 sequences by both softwares increases our confidence in this contig (Richter & Buchner 2001, Young et al. 2001).

Two clusters representing calcium-binding proteins were found among the most expressed genes generated by both softwares. One of these clusters is related to a family of proteins called fibrillins. The human fibrillin is a calcium-binding protein, which is the major structural component of microfibrils located in the extracellular matrix of connective tissue (Hanford 2000). This protein seems to be highly expressed in *S. mansoni* eggs but its function in this organism is unknown. The other calcium-binding protein is found almost exclusively in the cercariae stage. This gene seems to be turned on within a short period of time after the release of the cercariae into the water and shut off when the cercariae transforms into schistosomula (Ram et al. 1989). The calcium-binding proteins are interesting because several metabolic and physiological events are triggered by calcium ions with the help of these molecules (Goodman et al. 1979). Moreover, these proteins are necessary in the process of host skin invasion by cercariae, once that Ca^{2+} ions play crucial roles in many matrix-matrix, cell-matrix and cell-cell contacts (Maurer & Hohenester 1997).

Several of the most expressed sequences could not be identified by homology searches. Seven of these uniques resulted from the PHRAP clustering and four from the CAP3 clustering. Three of these unknown highly expressed genes were clustered by both softwares: two of them were present exclusively in cercariae and a pro-his rich protein was found only in females (and adult-worms). Of the cluster resulting exclusively from the PHRAP clustering, two have been found only in adults and cercariae, one just in adults (AUT1) and the other in all developmental stages. The other unknown unique produced in the CAP3 procedure was found in adults and eggs.

Another poorly characterized gene was the one encoding the protein presenting similarity to human neuroendocrine-specific protein B. The neuroendocrine-specific proteins are thought to be involved in neuroendocrine secretion or in membrane trafficking in neuroendocrine cells (Hens et al. 1998). These proteins are called reticulons because they associate with the endoplasmic reticulum (van de Velde et al. 1994). The cluster representing this protein is composed of sequences only from adult and female libraries and was produced using both softwares. The function of this protein, in *S. mansoni*, is still unknown.

The elongation factor 1 alpha is known as the second most abundant transcript in eukaryotic cells, constituting 1-2% of total protein in normal cells (Condeelis 1995). Its main function is to catalyze the GTP-dependent binding of aminoacyl-transfer RNA to the ribosomes. One cluster representing this gene was found by each program and,

curiously, there are no sequences of this gene in egg libraries, perhaps due to the small amount of sequences analyzed from this stage.

Glutathione S-transferases are detoxification enzymes that catalyze the conjugation of electrophilic substrates to glutathione (Sheehan et al. 2001). This protein is one of the best characterized antigens and the most tested vaccine candidate against schistosomiasis (Riveau et al. 1998). This enzyme was found in all *S. mansoni* developmental stages, except the lung stage, using both PHRAP and CAP3.

The Y-box binding proteins are multifunctional regulators of gene expression (Matsumoto & Wolffe 1998) and can bind to both double-stranded and single-stranded DNA and also to mRNA (Franco et al. 1997a). Twenty-nine sequences of this gene were found in libraries of males, females, adult worms and cercariae.

In conclusion, the putative most expressed *S. mansoni* genes were analyzed here, considering the presence of each one in different stages of the worm life cycle. We have observed that the majority of these genes is related to the maintenance of the cell homeostasis, encoding genes related to carbohydrate and energy metabolism, cytoskeleton, chaperones, transcription factors and detoxification enzymes. Genes related to pathogenesis, encoding eggshell proteins and proteases have also been found. Several genes could not be identified and may have some specific functions in the parasite. Its important to notice that the identification of some genes as the most expressed could be the result of redundancy in a given library. This observation should be applied, mainly, to the genes found only in one library.

We have identified some problems during our analysis due to clustering errors that occurred mainly when repetitive regions (micro or minisatellites) were present in the sequences. However, some clusters contained exactly the same sequences, when comparing the results generated by the two clustering programs and a great number of the most expressed genes were found by both softwares. This reinforces the reliability of this analysis.

Unfortunately, sequences from some developmental stages of *S. mansoni* (such as miracidium and sporocysts) are not yet available in public databases. It seems necessary, therefore, a community effort in order to fill in these blanks, producing more sequences from each stage in order to complete the characterization of the parasite transcriptome. Fortunately, new expressed gene sequencing projects were recently approved in Brazil and these, in cooperation with the efforts from all over the world, will generate more sequences to fill the gaps existing today.

ACKNOWLEDGEMENTS

To Dr Osvaldo Carvalho from Cenapad, MG/CO for providing computers and technical support.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res 25*: 3389-3402.

Brindley PJ, Kalinna BH, Dalton JP, Day SR, Wong JYM, Smythe ML, Mcmanus DP 1997. Proteolytic degradation of host hemoglobin by schistosomes. *Mol Biochem Parasitol 89*: 1-9.

Bruce JI, Ruff MD, Hasegawa H 1971. *Schistosoma mansoni*: endogenous and exogenous glucose and respiration of cercariae. *Exp Parasitol 29*: 86-93.

Bueding E, Fisher J 1982. Metabolic requirements of schistosomes. *J Parasitol 68*: 208-212.

Chen LL, Rekosh DM, LoVerde PT 1992. *Schistosoma mansoni* p48 eggshell protein gene: characterization, developmentally regulated expression and comparison to the p14 eggshell protein gene. *Mol Biochem Parasitol 52*: 39-52.

Coles GC 1972. Oxidative phosphorylation in adult *Schistosoma mansoni*. *Nature 240*: 488-489.

Condeelis J 1995. Elongation factor 1 alpha, translation and the cytoskeleton. *Trends Biochem Sci 5*: 169-170.

Davis A, Blanton R, Klich P 1985. Stage and sex specific differences in actin gene expression in *Schistosoma mansoni*. *Mol Biochem Parasitol 16*: 289-298.

Franco GR, Simpson AJG 2001. The structure and expression of the *Schistosoma* genome. In AAF Mahmoud, *Schistosomiasis*, Imperial College Press, London, p. 85-113.

Franco GR, Garratt RC, Tanaka M, Simpson AJG, Pena SDJ 1997a. Characterization of a *Schistosoma mansoni* gene encoding a homologue of the Y-box binding protein. *Gene 198*: 5-16.

Franco GR, Rabelo EML, Azevedo V, Pena HB, Ortega JM, Santos TM, Meira WSF, Rodrigues NA, Dias CMM, Harrop R, Wilson A, Saber M, Abel-Hamid H, Faria MSC, Margutti MEB, Parra JC, Pena SDJ 1997b. Evaluation of cDNA Libraries from different developmental stages of *Schistosoma mansoni* for production of expressed sequence tags (ESTs). *DNA Res 4*: 231-240.

Franco GR, Valadão AF, Azevedo V, Rabelo EML 2000. The *Schistosoma* gene discovery program: state of the art. *Int J Parasitol 30*: 453-463.

Frankel S, Mooseker MS 1996. The actin-related proteins. *Curr Opin Cell Biol 8*: 30-37.

Goodman M, Pechere JF, Haiech J, Demaille G 1979. Evolutionary diversification of structure and function in the family of intracellular calcium-binding proteins. *J Mol Evol 13*: 331-352.

Hanford PA 2000. Fibrillin-1, a calcium-binding protein of extracellular matrix. *Biochim Biophys Acta 1498*: 84-90.

Hens J, Nuydens R, Geerts H, Senden NH, Van de Ven JM, Roebroek AJ, van de Velde HJ, Ramaekers FC, Broers JL 1998. Neuronal differentiation is accompanied by NSP-C expression. *Cell Tissue Res 292*: 229-237.

Huang X, Madan A 1999. CAP3: A DNA Sequence Assembly Program. *Genome Biol 9*: 868-877.

Klinkert MQ, Felleisen R, Link G, Ruppel A, Beck E 1989. Primary structures of Sm31/32 diagnostic proteins of *Schistosoma mansoni* and their identification as proteases. *Mol Biochem Parasitol 33*: 113-122.

Li YL, Idris MA, Corachan M, Han JJ, Kirschfink M, Ruppel A 1996. Circulating antigens in schistosomiasis: detection of 31/32 kDa proteins in sera from patients infected with *Schistosoma japonicum, S. mansoni, S. haematobium,* or *S. intercalatum*. *Parasitol Res 82*: 14-18.

Maurer P, Hohenester E 1997. Structural and functional aspects of calcium binding in extracellular matrix proteins. *Matrix Biol 15*: 569-580.

Matsumoto K, Wolffe AP 1998. Gene regulation by Y-box proteins: coupling control of transcription and translation. *Trends in Cell Biol 8*: 318-323.

Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA 1999. A compreensive approach to clustering of expressed gene sequence: the se-

quence tag alignment and consensus knowledge base. *Genome Res 9*: 1143-1155.

Oliveira G, Johnston DA 2001. Mining the schistosome DNA sequence database. *Trends Parasitol 17*: 501-503.

Ram D, Grossman Z, Markovics A, Avivi A, Ziv E, Lantner F, Schechter I 1989. Rapid changes in the expression of a gene encoding a calcium-binding protein in *Schistosoma mansoni*. *Mol Biochem Parasitol 34*: 167-176.

Richter K, Buchner J 2001. Hsp90: chaperoning signal transduction. *J Cell Physiol 188*: 281-290.

Riveau G, Poulain-Godefroy O, Dupré L, Remoué F, Mielcarek N, Locht C, Capron A 1998. Glutathione S-transferases of 28kDa as major vaccine candidates against schitosomiasis. *Mem Inst Oswaldo Cruz 93*: 87-94.

Rumjanek FD 1987. Biochemistry and physiology. In *The Biology of Schistosomes*, Academic Press, p. 163-183.

Santos TM, Jonhston DA, Azevedo V, Ridgers IL, Martinez MF, Marotta GB, Santos RL, Fonseca SJ, Ortega JM, Rabelo EML, Saber M, Ahmed HM, Romeih MH, Franco GR, Rollinson D, Pena SDJ 1999. Analyses of the gene expression profile of *Schistosoma mansoni* cercariae using the expressed sequence tag approach. *Mol Biochem Parasitol 103*: 79-97.

Sheehan D, Meade G, Foley VM, Dowd CA 2001. Structure, function and evolution of glutathione transferases: implications for classfication of non-mammalian members of an ancient enzyme superfamily. *Biochem J 360*: 1-16.

Simpson AJG, Sher A, McCutchan TF 1982. The genome of *Schistosoma mansoni*: isolation of DNA, its size, bases and repetitive sequences. *Mol Biochem Parasitol 6*: 125-137.

Skelly PJ, Lincoln LD, Shoemaker CB 1993. Expression of *Schistosoma mansoni* genes involved in anaerobic and oxidative glucose metabolism during the cercaria to adult transformation. *Mol Biochem Parasitol 60*: 93-104.

Skelly PJ, Shoemaker CB 1995. A molecular genetic study of the variations in metabolic function during schistosome development. *Mem Inst Oswaldo Cruz 90*: 281-284.

Stein L 2001. Genome annotation: from sequence to biology. *Nat Reviews 2*: 493-505.

Timms AR, Bueding E 1959. Studies of a proteolytic enzyme from *Schistosoma mansoni*. *Br J Pharmacol 14*: 68-73.

Thompson DP, Morrison DD, Pax RA, Bennet JL 1984. Changes in glucose metabolism and cyanide sensitivity in *Schistosoma mansoni* during development. *Mol Biochem Parasitol 13*: 39-51.

Vallee RB, Gee MA 1998. Make room for dynein. *Trends in Cell Biol 8*: 490-493.

van de Velde HJ, Senden NH, Roskams TA, Broers JL, Ramaekers FC, Roebroek AJ, Van de Ven WJ 1994. NSP-encoded reticulons are neuroendocrine markers of a novel category in human lung cancer diagnosis. *Cancer Res 54*: 4769-4776.

Van Oordt BEP, Tielens AGM, Van den Bergh SG 1989. Aerobic to anaerobic transition in the carbohydrate metabolism of *Schistosoma mansoni* cercariae during transformation in vitro. *Parasitology 98*: 409-415.

Van Oordt BEP, Van del Reuvel JM, Tielens AGM, Van den Bergh SG 1985. The energy production of the adult *Schistosoma mansoni* is for large part aerobic. *Mol Biochem Parasitol 16*: 117-126.

Weller DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA 2002. Database resources of the National Center for Biotechnology information: 2002 update. *Nucleic Acids Res 30*: 13-16.

Young JC, Moarefi I, Hartl FU 2001. Hsp90: a specialized but essential protein-folding tool. *J Cell Biol 154*: 267-273.