DEPARTAMENTO DE BIOLOGIA GERAL
PÓS-GRADUAÇÃO EM GENÉTICA

DISSERTAÇÃO DE MESTRADO

Análise dos genes mais expressos e do *status* atual do transcriptoma de *Schistosoma mansoni* utilizando ferramentas de bioinformática

FRANCISCO PROSDOCIMI

FRANCISCO PROSDOCIMI DE CASTRO SANTOS

"Análise dos genes mais expressos e do *status* atual do transcriptoma de *Schistosoma mansoni* utilizando ferramentas de bioinformática"

Dissertação apresentada ao Programa de Pós-graduação em Genética do Departamento de Biologia Geral do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de Mestre em Genética.

ÁREA DE CONCENTRAÇÃO: BIOINFORMÁTICA

Orientadora: Profa. Glória Regina Franco
Co-Orientador: Prof. Fabrício Rodrigues dos Santos

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Departamento de Biologia Geral
Belo Horizonte – MG
Março de 2003

"A dupla hélice é realmente uma molécula extraordinária. O homem moderno tem talvez 50 mil anos, a civilização existe há apenas 10 mil anos (...). Mas o DNA e o RNA existem há pelo menos vários bilhões de anos. Durante todo esse tempo, a dupla hélice esteve por aí, ativa; no entanto, somos as primeiras criaturas sobre a Terra a nos tornarmos conscientes da sua existência."

Francis Crick

"Eles estão em mim e em você. Eles nos criaram, corpo e mente. E sua preservação é a razão última de nossa existência. Transformaram-se muito, esses replicadores. Agora eles recebem o nome de genes e nós somos suas máquinas de sobrevivência."

Richard Dawkins, O gene egoísta

A G R A D E C I M E N T O S

Agradeço a meus pais simplesmente por tudo, não há como citar ou descrever o que fizeram por mim ao longo desses 23 anos.

Agradeço a toda minha família, aos amigos da biologia e colegas do CEFET pelo convívio e a distração que me proporcionaram. Agradeço aos meus colegas do mestrado em genética – Fred, Simone, Lu e Camilli – pelo companheirismo, pelas discussões, pelas mágicas, pelos butecos da biologia e por terem tornado mais divertidas as disciplinas da pós-graduação. Dentre os amigos pessoais agradeço ao Rodrigo Loyola, Flávio Pimenta, Hamilton Cruz, Rafael Prosdocimi, Luís Flávio Prosdocimi e Flávio Garcia por me agüentarem nas discussões sobre os mais diversos assuntos e pelo companheirismo.

Devo muito e gostaria de agradecer em especial à secretária do curso de pós-graduação em genética Marina Miranda. Educada e atenciosa, resolveu todos os meus problemas de aproveitamento de créditos, cartas para o colegiado e confusões com trancamento e matrícula em disciplinas, mesmo depois do prazo.

Agradeço aos amigos que participaram do 2º curso de especialização em bioinformática que cursei no LNCC, em Petrópolis, entre Agosto e Novembro de 2002. O apoio, a amizade, a união, o truco, as brigas, o violão, o playstation, os golos e os jogos de dardo foram indispensáveis para o êxito dos trabalhos produzidos e para aliviar a saudade que sentia da minha família, dos amigos e, principalmente, da Cacá. Agradeço também à Telemar, por não consertar (durante algum tempo) um orelhão, próximo ao hotel Quitandinha em Petrópolis, permitindo que ele aceitasse chamadas interurbanas a cobrar. Talvez eles tenham percebido o defeito depois que o orelhão-do-amor (ou disque-patroa, como era chamado por alguns) passou a receber chamadas a cobrar de mais de trinta minutos de quase todos os estados do Brasil.

Antes de sair de Petrópolis devo agradecer muito a um dos meus mais novos melhores-amigos e companheiro intelectual, Gustavo Cerqueira. Apesar da bagunça que aprontava em nosso mísero AP foi muito bom tê-lo conhecido e convivido com ele por três meses consecutivos. Nunca havia encontrado ninguém que levasse tão a sério e tivesse a coragem de mudar completamente sua área de atuação com o objetivo de seguir à risca uma filosofia de vida completamente racional. Super-profissional, inteligente e *workaholic* tem tudo para se destacar brevemente como um dos principais bioinformatas do país. Ou como filósofo.

Agradeço aos amigos do Laboratório de Genética-Bioquímica (LGB) pelo convívio, pelas ajudas, pelas dicas, pelas brincadeiras e por tudo mais. Obrigado ao Chico Lobo, Carlos Eduardo, Marina Mourão, Patrícia Souza, Débora Naves, Alessandra Campos e Carol Loque; ao Charles Anacleto, Cláudia Benedetto e Flávia Parra; ao Luís Augusto, Alice Machado, Carlos Gustavo e Débora Aline; à Juliana Pimenta, Jorge, Simone e Paula; ao professor Carlos Renato Machado. Agradeço ao Professor Miguel Ortega pela amizade, pelas discussões sempre inteligentes sobre a bioinformática e pelas sugestões dadas sobre este trabalho.

Devo um agradecimento especial ao Fabiano Peixoto e ao professor Osvaldo Carvalho por disponibilizarem as máquinas e um espaço físico adequado para o desenvolvimento da minha dissertação dentro do LCC/CENAPAD.

Gostaria de agradecer também a alguns dos grandes cientistas da atualidade que me fizeram compreender, mais do que ninguém, através da leitura de diversos de seus livros, a maravilha e o poder do empreendimento científico e a importância da divulgação de ciência. Meu muito obrigado a Richard Dawkins, Carl Sagan, Stephen Jay Gould, Antonio Damasio e a todos os que divulgam a boa ciência.

Devo muito à minha querida Chefa, Dra. Glória Regina Franco, exemplo a ser seguido tanto no campo pessoal como profissional. Sempre alegre, educada, estudiosa, vibrante e empolgada, minha ídola. Desejo-lhe tudo de bom em sua nova vida de mamãe.

Gostaria de agradecer ainda ao meu co-orientador Fabrício Santos pelas sugestões precisas em alguns pontos do trabalho e pelo apoio. Admiro-o por sua inteligência, eficiência e dinâmica e espero que possamos nos dar bem no doutorado em bioinformática que irei começar em breve sob sua orientação.

Finalmente agradeço à minha querida Carolina Furtado por todos os momentos que passamos juntos (física ou virtualmente) desde o início do nosso namoro. Agradeço por ter segurado as pontas enquanto estive em Petrópolis e por ter me recebido de braços abertos na volta. Agradeço ainda por ter me adotado em sua família e por tudo o que passamos até aqui e pelo que ainda iremos passar. Talvez seja um paradoxo agradecê-la aqui, já que ela é responsável pela ausência de alguns resultados que poderiam estar presentes nessa dissertação, sempre me incentivando a estudar e trabalhar menos. Mas ao mesmo tempo, sem ela eu não teria a mesma tranqüilidade e a mesma inspiração para trabalhar, para fazer ciência e para viver. Obrigado, Cacá.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

ABREVIATURAS

ACP	*Acyl-carrier protein* ou proteína carregadora de acila
AdoMET	S-adenosilmetionina
ADP	Adenosina difosfato
AN	*Accession number* ou número de acesso do GenBank
AMP	Adenosina monofosfato
AP	Apurínico ou Apirimidínico
APRT	Adenina fosforibosil transferase
ARP	*Actin-related proteins* ou proteínas relacionadas a actina
ATP	Adenosina trifosfato
BLAST	*Basic Local Alignment Search Tool*
CAP3D	CAP3 executado nas condições DEFAULT
CAP3OS	CAP3 executado em condições mais rigorosas (-o 40 –s 800)
cAMP	AMP cíclico
CDC	*Cell Division Cycle protein* ou proteína de divisão celular
cDNA	DNA complementar
cGMP	GMP cíclico
CoA	Coenzima A
COG	*Clusters of Ortologous Groups*
CTP	Citosina trifosfato
dbEST	Banco de dados de ESTs do NCBI
DDBJ	DNA *Data Bank of Japan*
dTTT	*Desoxitimidina trifosfato*
dUTP	*Desoxiuridina trifosfato*
dUMP	*Desoxiuridina monofosfato*
EF	*Elongation factor* ou fator de alongamento
EGF	*Epidermal Growth Factor ou* Fator de crescimento epidérmico
EGFR	*Epidermal Growth Factor Receptor* ou receptor para o fator de crescimento epidérmico
EMBL	*European Molecular Biology Laboratory*
ER	Retículo Endoplásmico
eRF	*Eukaryotic Releasing factor* ou Fator de liberação eucariótico
EST	*Expressed Sequence Tags* ou Etiquetas de Seqüências Expressas
FAD	Flavina adenina dinucleotídeo
GTP	Guanosina trifosfato
GMP	Guanosina monofosfato
GDP	Guanosina difosfato
FTP	*File Transfer Protocol*
GI	*GenInfo Identifier*
GO	*Gene Ontology*
GSH	Glutationa
HDL	*High-density lipoprotein* ou lipoproteína de alta densidade
HGM-CoA	Beta-hidroxi-beta-metil-glutaril CoA
HIP	*HSP-70 Interacting Proteins* ou proteínas que interagem com HSP-70
HSC	*Heat-Shock cognate proteins* ou proteínas cognates a choque térmico
HSP	*Heat Shock Protein* ou Proteínas de Choque Térmico
HTML	*HyperText Markup Language*
HTTP	*HyperText Transfer Protocol*

IMP	Inosina monofosfato
KEGG	*Kyoto Encyclopedia for Genes and Genomes*
LDL	*Low-density lipoprotein* ou lipoproteína de baixa densidade
LTR	*Long Terminal Repeat* ou repetição terminal longa
mtDNA	DNA mitocondrial
NAD	Nicotinamida adenina dinucleotídeo
NCBI	*National Center for Biotechnology Information*
NMP/NDP/NTP	Nucleosídeo mono/di/tri fosfato
NIH	*National Institutes of Health*
NLM	*National Library of Medicine*
nr	Banco de dados não redundante de proteínas
nt	Banco de dados não redundante de nucleotídeos
ORESTES	*Open Reading Frame ESTs*
PCNA	*Proliferation Cellular Nuclear Antigen* ou Antígeno nuclear de proliferação celular
PCR	Reação em cadeia da polimerase
PDB	*Protein Data Bank* (Banco de dados de estruturas de proteína)
PDI	*Protein Dissulfite Isomerases* ou proteínas dissulfito isomerases
PHRAP	*PHil Revised Assembly Program* ou *PHRagment Assembly Program*
PKA	*Protein kinase A* ou proteína quinase A
PKC	*Protein kinase C* ou proteína quinase C
PPIases	*Peptidil-prolil cis-trans isomerases*
PRPP	Fosforibosil pirofosfato
RAM	Memória de acesso aleatório
RAP	*RAS-related protein* ou proteína relacionada a RAS
rRNA	RNA ribossômico
SGP	*Schistosoma mansoni* Genome Project
SMOX	*Schistosoma mansoni* homeobox
SMS	*The Sequence Manipulation Suite*
snRNP	Pequena ribonucleoproteína nuclear
SP	Banco de dados Swiss-Prot
SRP	*Signal Recognition Particle* ou partícula reconhecedora do sinal
SWAT	Algoritmo Smith-Waterman para comparação de seqüências
TF	*Transcription Factor* ou fator de transcrição
TIGR	*The Institute for Genomic Research*
TrEMBL	Banco de dados *Translated* EMBL
tRNA	RNA transportador
TCP	*T-complex protein*
TIM	*Transport across the inner membrane* ou transportador através membrana interna
TOM	*Transport across the outer membrane* ou transportador através membrana externa
TRAP	*Translocon-Associate proteins*
TTP	Timidina trifosfato
UDP	*Uradina difosfofato*
WHO	*World Health Organization*
YAC	Cromossomo artificial de levedura

RESUMO

O desenvolvimento da tecnologia de geração de Etiquetas de Seqüências Expressas (ESTs) proporcionou aos cientistas uma forma rápida e barata de produzir dados de seqüências de DNA em larga escala. Entretanto, seqüências de ESTs são pequenas em extensão e podem apresentar uma alta taxa de erros, o que dificulta o processo de identificação do gene através de ferramentas de alinhamento local (BLAST). Para contornar este problema, utilizam-se ferramentas de agrupamento de ESTs, que agrupam seqüências similares e formam um consenso, de tamanho maior. Dessa forma, pode-se realizar a montagem dos genes, aumentando o tamanho das seqüências disponíveis e permitindo uma melhor identificação baseada em genes e proteínas ortólogos. Muitas ferramentas têm sido utilizadas para realizar o agrupamento de seqüências, inclusive programas para montagem de *contigs* genômicos, como o PHRAP e o CAP3. Utilizamos esses dois programas para agrupar as seqüências de *Schistosoma mansoni* do GenBank. Observamos os grupos de seqüências mais expressas em cada fase de desenvolvimento do organismo e tentamos relacionar sua alta expressão com a biologia do verme. O resultado do agrupamento realizado pelo CAP3 produziu 2017 consensos e 4311 *singlets* e esse número apresenta uma estimativa, *grosso modo*, de que a comunidade científica já seqüenciou cerca de um terço dos 15.000 a 20.000 possíveis genes do verme. Entretanto, mais importante do que saber quantos genes já foram seqüenciados é saber quais são eles. Com esse propósito, realizamos o processo de anotação gênica para todas as 6328 seqüências *unique*. O principal método de anotação que tem sido utilizado pelos bioinformatas é baseado em pesquisas de similaridade (BLASTs) contra diferentes bancos de dados, incluindo bancos “não redundantes” de nucleotídeos e proteínas e bancos curados de proteínas. As anotações foram realizadas manualmente, considerando os resultados das buscas de similaridade para a identificação gênica. As seqüências anotadas foram classificadas através das categorias funcionais dos COGs (*Clusters of Ortologous Groups*) e foi criado um *website* para disponibilizar toda essa informação.

ABSTRACT

The technology of generation of Expressed Sequence Tags (EST), developed in last decade, provides to the scientists an easy and unexpensive way to produce DNA sequence data in large scale. However, EST sequences are small and have a high error rate, what makes harder the process of gene identification through local alignment tools. To circumvent this problem, EST clustering tools are used to join overlapping sequences, creating a consensus larger than the original sequences. This makes it possible to increase the sequence size and allows a better identification based on ortologous genes or proteins. Several different tools have been used to carry out the sequence clustering, such as PHRAP and CAP3, softwares originally developed to produce genomic contigs. We have used these two softwares in order to cluster all the GenBank and dbEST *Schistosoma mansoni* cDNA sequences. Furthermore, we have identified the most expressed sequences in each developmental stage of the worm and we tried to relate their high expression with the biology of the organism. CAP3 has produced, after clustering analysis, 2017 contigs and 4311 singlets, or 6328 non-redundant sequences. This number represents an estimate of the number of genes that have already been sequenced by the scientific community, corresponding approximately to a third part of the 15.000-20.000 putative genes of the worm. However, even more relevant than knowing how many genes have already been sequenced is to know what they are. With this in mind, we performed a process of gene annotation for all *Schistosoma* 6328 unique sequences (the non-redundant ones). The main annotation method nowadays used by the researchers in bioinformatics is based on similarity search (BLASTs) against different databases, including non-redundant nucleotide and protein databases and protein curated databases. Our annotations were manually performed, considering the results of the similarities searches for gene identification. The identified sequences were classified in the COG (Clusters of Ortologous Groups) functional categories and a website was created to make all these information available.

1. INTRODUÇÃO

1.1. ORGANISMO DE ESTUDO – *Schistosoma mansoni*

O *Schistosoma mansoni* é um verme trematódeo dióico (Neves 1983) e seu estudo é de grande interesse na parasitologia humana. A esquistossomose, doença causada por espécies do gênero *Schistosoma*, é a segunda doença tropical mais importante com relação à saúde pública (só perdendo para a malária) (http://www.who.int/ctd/schisto/epidemio.htm). Apesar dos recentes esforços feitos no sentido de conter seu avanço, a doença permanece endêmica em 74 países (FIGURA 1), deixando mais de 600 milhões de pessoas em risco de infecção, com cerca de 200 milhões de pessoas infectadas (http://www.who.int/ctd/schisto/epidemio.htm).

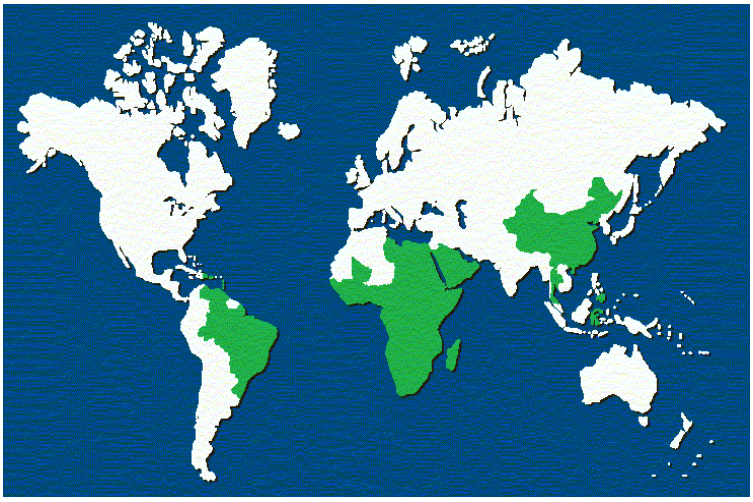

Figura 1. Países afetados pela esquistossomose*. Em verde estão mostrados os países nos quais a esquistossomose é, ainda hoje, uma doença endêmica. Obtido em http://www.who.int/tdr/diseases/schisto/diseaseinfo.htm.*

O *S. mansoni* é um organismo que apresenta diferentes fases de desenvolvimento e passa seu ciclo de vida alternando entre formas parasitárias em diferentes hospedeiros e formas intermediárias de vida livre (FIGURA 2).

Os vermes adultos vivem no sistema porta-hepático humano. Através da corrente sanguínea, os casais de *S. mansoni* migram para o plexo hemorroidário, onde as fêmeas fazem a postura de ovos ao nível da submucosa. Oito a dez dias após a postura, os ovos passam por uma fase de maturação até que o miracídio esteja formado (Neves 1983). Nesse período, os ovos migram da camada submucosa até a luz intestinal, aqueles que conseguem alcançar a luz intestinal vão ser eliminados até o exterior juntamente com o bolo fecal. Ao alcançar a água, os ovos liberarão os

miracídios, que nadarão ativamente até encontrarem algum caramujo do gênero *Biomphalaria* (Neves 1983). A expectativa de vida do miracídio é de cerca de 8 horas e, se não encontrarem algum caramujo durante esse período, começarão a morrer. Ao entrarem através das partes moles de algum caramujo susceptível do gênero *Biomphalaria*, os miracídios transformam-se nos chamados esporocistos I, que dão origem a cerca de 150-200 dos chamados esporocistos II. Esses últimos migram para as glândulas digestivas ou para a ovoteste do caramujo, onde dão origem ao estágio de cercária (Neves 1983). A cercária passa ao meio exterior aquoso atravessando o epitélio do manto e da pseudobrânquia, normalmente nas horas mais quentes e luminosas do dia. Depois de liberadas, as cercarias podem viver até 3 dias, mas têm sua principal atividade nas primeiras 8 horas. Elas nadam ativamente na água e, ao alcançarem a pele do homem, fixam-se, preferencialmente, entre os folículos pilosos, auxiliadas por suas ventosas (Neves 1983). Através de ação lítica e mecânica inicia-se a penetração ativa na pele do hospedeiro, que demora cerca de 15 minutos. Após a penetração da cabeça, as cercárias perdem a cauda e passam a ser conhecida como esquistossômulos, que vão do tecido subcutâneo para os vasos linfáticos ou venosos, onde são espalhados por todo o corpo. Entretanto, apenas os esquistossômulos que chegam ao sistema porta é que são capazes de se desenvolver (Neves 1983). Ali eles se alimentam e se transformam em machos ou fêmeas cerca de 30 dias após a penetração. E, completando o ciclo, um casal de vermes adultos migra para o plexo hemorroidário para realizar a oviposição cerca de 40 dias após a infecção (Neves 1983).

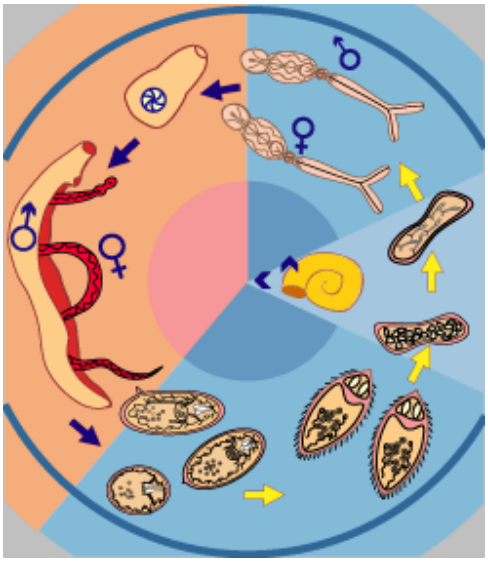

***Figura 2. Fases de desenvolvimento do parasita S. mansoni**. Em bege está a fase encontrada no hospedeiro humano (esquistossômulo e verme adulto) e em azul claro a fase encontrada no hospedeiro molusco do gênero Biomphalaria (esporocistos I e II). Em azul escuro são mostradas as fases de vida livre do organismo (abaixo miracídio, acima cercária). Obtido em http://www.who.int/tdr/diseases/schisto/lifecycle.htm.*

Os schistossomatídeos são organismos diplóides e seu conteúdo de DNA é distribuído em sete pares de autossomos e um par de cromossomos sexuais. Ao contrário dos humanos, o macho é o sexo homogamético (ZZ) e a fêmea, o heterogamético (ZW). Cerca de 60% de seu DNA é composto por DNA alta a moderadamente repetitivo, com apenas 29,4% de conteúdo de GC (Marx *et al*. 2000).

Estima-se que o genoma de *S. mansoni* contenha cerca de 270 megabases, o que corresponde a cerca de 10% do tamanho do genoma humano (Simpson *et al*. 1982). Esse enorme tamanho torna inviável, ao menos por enquanto, a realização do seqüenciamento do genoma completo do parasita.

A obtenção de conhecimentos sobre o genoma de parasitas é extremamente importante para o entendimento de sua biologia e metabolismo (Degrave *et al.* 2001), permitindo-nos encontrar diferentes frentes de ataque para o tratamento dos doentes. O projeto genoma do *S. mansoni* (SGP) foi criado em 1992 através de uma iniciativa brasileira e foi apoiado pelo TIGR (*The Institute for Genomic Research*) (Franco *et al.* 2000).

1.2. A PRODUÇÃO DE ESTS DE *S. mansoni*

O diretor do TIGR em 1991, J. Craig Venter, em um trabalho chave para a genômica moderna, popularizou a estratégia de seqüenciamento de etiquetas de seqüências expressas, as ESTs (Adams *et al.* 1991) (FIGURA 3). Nessa época, o GenBank possuía cerca de 3000 genes e ele, juntamente com seus colaboradores, conseguiram descobrir outros 337 a partir de seqüências de ESTs, mais de 10% do que toda a comunidade científica havia feito até a época (Davies 2001). Rapidamente, o método de seqüenciamento de ESTs foi adotado como uma forma eficiente de produzir etiquetas de identificação gênica em larga escala, e um banco de dados próprio para essas seqüências, o dbEST, foi criado no *National Center for Biotechnology Information*, o NCBI (Boguski *et al.* 1993).

Desde o início do SGP, vários grupos espalhados pelo mundo já seqüenciaram 16.815 ESTs de *S. mansoni*, disponíveis no dbEST (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html) (dbEST release 010303 – 03/01/2003). Entretanto, esse número de ESTs não nos diz quantos e nem quais dos 15.000 a 20.000 genes, que se estima existir no parasita (Franco *et al.* 2000), já foram seqüenciados e é nesse ponto que entra a bioinformática. Como as ESTs possuem normalmente ente 150bp e 600bp, na maioria das vezes, elas não representam um gene completo. Por isso, sempre que possível, seus clones são seqüenciados a partir das duas extremidades,

gerando duas seqüências que podem (ou não, se o cDNA for muito grande) ser agrupadas gerando uma seqüência maior, através da análise de *clustering*.

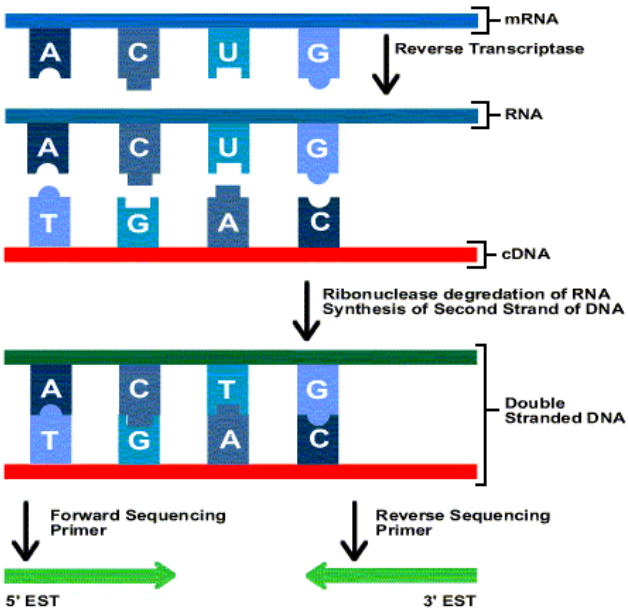

***Figura 3. Produção de ESTs**. As etiquetas de seqüências expressas são obtidas através, primeiramente, da transcrição reversa do mRNA, produzindo uma fita de DNA complementar (cDNA). Após esse procedimento, utiliza-se uma RNAse H para digerir a seqüência de RNA inicial e é produzida a segunda fita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

1.3. A Bioinformática na Descoberta de Novas Drogas

Podemos considerar a bioinformática como uma linha de pesquisa que envolve aspectos multidisciplinares e que surgiu a partir do momento em que se iniciou a utilização de ferramentas computacionais para a análise de dados genéticos, bioquímicos e de biologia molecular. A bioinformática envolve a união de diversas linhas de conhecimento – a ciência da computação, a engenharia de *softwares*, a matemática, a estatística e a biologia molecular – e tem como finalidade principal desvendar a grande quantidade de dados que vem sendo obtida através de seqüências de DNA e proteínas. Para o desenvolvimento de genomas completos, a informática é imprescindível e a biologia molecular moderna não estaria tão avançada hoje, não fossem os recursos computacionais existentes.

A bioinformática e a genômica comparativa estão se tornando as principais metodologias para a descoberta de novos alvos para drogas, vacinas e diagnósticos de doenças humanas, parasitárias ou não (Terstappen & Reggiani 2001). No caso de doenças parasitárias, através da comparação dos dados de genoma dos parasitas e hospedeiros pode-se tentar descobrir alguma susceptibilidade do parasita a uma droga em uma determinada via metabólica ou pode-se encontrar genes presentes apenas no parasita que lhe sejam essenciais e possam ser utilizados como alvos específicos para drogas ou diagnóstico (FIGURA 4). É importante notar que a bioinformática pode apenas sugerir genes candidatos e compostos de partida para a produção de novas drogas (Terstappen & Reggiani 2001). Os genes e os compostos encontrados devem ser testados e validados através de ensaios biológicos e bioquímicos, para testar se poderão ser utilizados clinicamente.

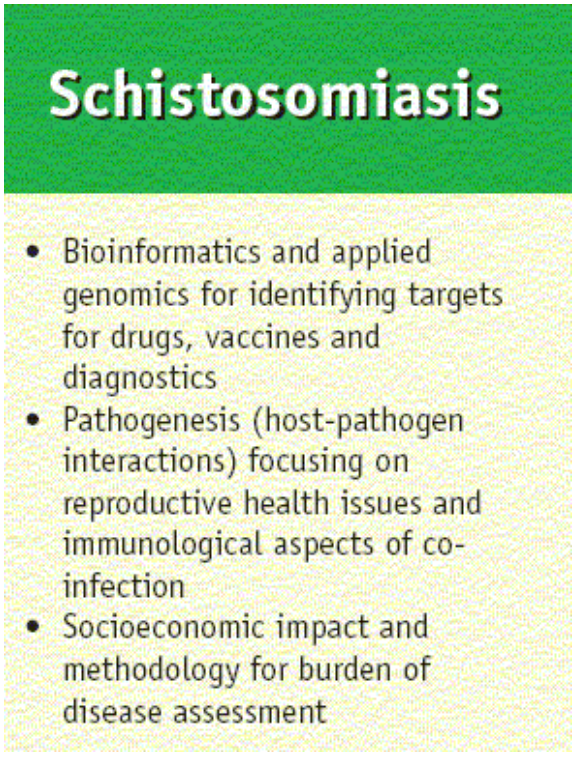

***Figura 4. Estratégias propostas pela OMS para pesquisas básicas em esquistossomose**. Como prova da importância da bioinformática na descoberta de alvos para novas drogas, vacinas e diagnósticos, essa figura foi retirada da página da OMS que trata de estratégias em pesquisas básicas sobre esquistossomose. Obtido em http://www.who.int/tdr/diseases/schisto/strategy.htm.*

1.4. Bases de Dados em Biologia Molecular

As bases de dados em biologia molecular são importantes principalmente para proporcionar à comunidade científica uma forma de tornar os dados (produzidos em todo o mundo) acessíveis de forma fácil, rápida e inteligente (http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html).

A primeira base de dados de biologia molecular parece ter surgido por volta de 1960, quando Dayhoff e colaboradores construíram um catálogo contendo todas as seqüências de proteínas conhecidas até a data. Essas seqüências foram publicadas num livro chamado "*Atlas of Protein Sequences and Structure*", de 1965. O conteúdo dessa base de dados não deveria conter mais de 1Mb de informação, se transferida para computadores modernos (Baxevanis & Ouellette 2001).

Com o advento do seqüenciamento do DNA e, principalmente, a partir da década de 1990, do seqüenciamento em larga escala, foi necessária a construção de bancos de dados mais robustos para abrigar a explosão no número de seqüências obtidas pelos pesquisadores (Figura 5). O NCBI, por exemplo, foi criado pelo NIH (*National Institutes of Health*) em 1988 para abrigar esse tipo de informação (Wheller *et al.* 2002).

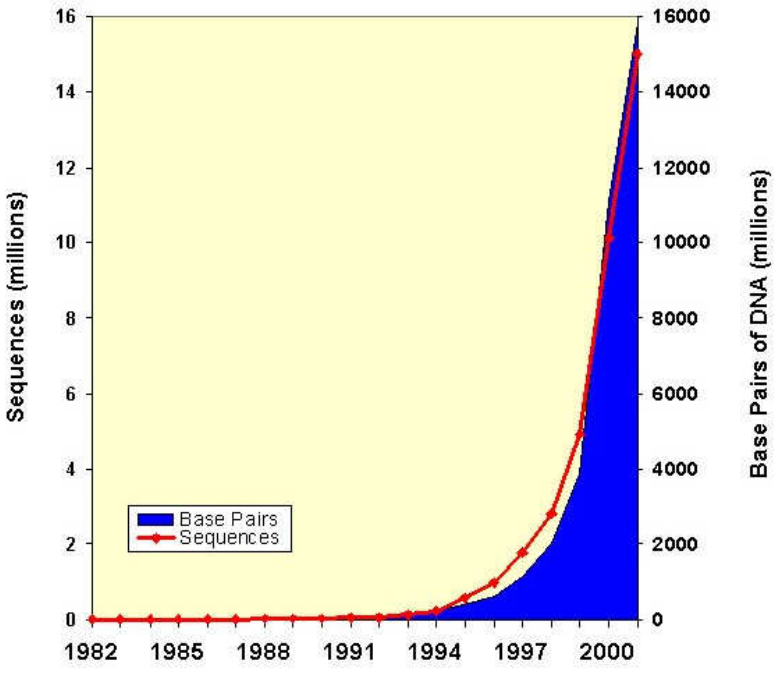

***Figura 5. Crescimento do Genbank**. Crescimento exponencial do número de seqüências contidas no GenBank ao longo das duas últimas décadas. Obtido em http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html.*

Dessa forma foi criada uma colaboração internacional para montar um banco de dados de seqüências de nucleotídeos, a INSDC (*International Nucleotide Sequence Database Colaboration*). Essa instituição contém o NCBI, o EMBL (*European Molecular Biology Laboratory*) e o DDBJ (*DNA Data Bank of Japan*) (Tateno *et al.* 2002). Cada um desses centros possibilita a submissão individual de seqüências de DNA e trocam informações entre si diariamente, sendo que todos os três possuem informações atualizadas de todas as seqüências disponíveis para os pesquisadores (Stoesser *et al.* 2002). Apesar disso, cada centro apresenta os dados de forma particular, apesar de bastante semelhante.

Existem basicamente dois tipos de bancos de dados disponíveis para utilização e pesquisa de genes e proteínas (Baxevanis & Ouellette 2001). Os bancos de dados primários apresentam resultados de dados experimentais que são publicados com alguma interpretação, mas não há uma análise cuidadosa desses dados com relação aos outros publicados anteriormente. Esse é o caso, por exemplo, do GenBank, EMBL e PDB (*Protein Data Bank*). Já os secundários são aqueles onde há uma compilação e interpretação dos dados de entrada de forma que podem ser obtidos dados mais representativos e interessantes. Esses são os bancos de dados curados, como o SWISS-PROT e o TrEMBL.

1.4.1. IDENTIFICADORES DE SEQÜÊNCIAS NOS BANCOS DE DADOS NCBI

O primeiro identificador de seqüência criado no NCBI foi o LOCUS, que era o único identificador de um registro no GenBank. O nome do loco é definido como uma seqüência de 10 ou menos letras em caixa alta que apresentam um mnemônico para a função e o organismo de origem da seqüência. Assim o nome HUMHBB era utilizado para representar a região da β-globina humana (Baxevanis & Ouellette 2001). Entretanto, com a descoberta de cada vez mais locos e alelos diferentes, e com o aumento exponencial do número de seqüências no GenBank, ficou impossível a invenção e a atualização dos nomes de forma controlada. Assim os nomes de LOCUS, apesar de ainda aparecerem nos arquivos de formato GenBank, não têm nenhuma utilidade prática.

Devido a essas dificuldades de utilização da informação armazenada em LOCUS, o conselho internacional de colaboradores para seqüências de nucleotídeos (NCBI, EMBL e DDBJ) introduziu o conceito de **accession number (AN)** ou número de acesso. Esse número não carrega, intencionalmente, nenhuma informação biológica, de forma a permanecer estável. Originalmente consistia de uma letra seguida por cinco números, sendo que cada letra corresponderia ao centro (NCBI, EMBL ou DDBJ) ao qual a seqüência foi submetida (Baxevanis & Ouellette 2001). Entretanto, logo esse número também começou a apresentar problemas já que as seqüências eram atualizadas contendo o mesmo AN. No arquivo GenBank há um campo nomeado de

accession, onde há a informação sobre o histórico de uma determinada seqüência; se ela se juntou a outra, se foi atualizada, etc. Apesar desses problemas, o AN é o índice mais controlado e confiável dos registros do NCBI/EMBL/DDBJ. Para melhorar a identificação de seqüências antigas, os membros do INSDC resolveram, em 1999, acrescentar, ao AN, o número de sua versão (Benson *et al.* 2002). Dessa forma pode-se ver o número de acesso, um ponto, e o número de atualizações feitas em uma determinada seqüência. Por exemplo, o número de acesso A21645.3 é a terceira atualização da seqüência A21645 e as versões mais velhas permanecem armazenas e acessíveis através dos números de submissão A21645.1 e A21645.2. Um código similar de **AN.versão** é dado também para seqüências de proteínas.

Para criar um índice mais robusto para suas entradas, o NCBI, em 1992, criou um novo identificador, o **GenInfo Identifier** (GI), um número inteiro simples. Esse é um identificador único para cada seqüência, independente de atualizações ou de qualquer outra coisa. Toda entrada no NCBI possui um GI único da seqüência que não é alterado de forma alguma, permanecendo na base de dados para o acesso (Benson *et al.* 2002). Se uma seqüência difere-se da outra por apenas um par de bases, as duas possuirão diferentes GIS, apesar de possuírem, por exemplo, o mesmo AN (com diferentes números de versão). Todos os processos internos do NCBI utilizam o número de GI para sua a execução.

1.5. TRATAMENTO DAS SEQÜÊNCIAS

As seqüências depositadas no GenBank podem conter uma grande quantidade de erros e contaminantes, já que não há uma curadoria desse banco. Assim, os pesquisadores que depositam as seqüências é que são os principais responsáveis por sua qualidade. Entretanto, muitas vezes os pesquisadores utilizam alguma forma de submissão automática das seqüências, devido ao grande número de seqüências produzidas em seu laboratório. Devido a esse fato muitas seqüências do GenBank apresentam contaminantes e devem ser tratadas antes de sua utilização.

O principal tratamento ao qual as seqüências devem ser submetidas é a retirada das regiões correspondentes aos vetores de clonagem. Para isso, o NCBI apresenta um banco de dados contendo apenas seqüências de vetores, disponível para FTP anônimo em ftp://ftp.ncbi.nlm.nih.gov/repository/vector-ig/. Esse banco de dados contém informações sobre 2.534 vetores, dos quais 1.102 apresentam seqüência completa. Plasmídeos, fagemídeos, cosmídeos, fagos e YACs (cromossomos artificiais de levedura) que se replicam em cepas bacterianas, leveduras e linhagens celulares de *Drosophila*, macacos e humanos estão disponíveis (*Vector-IG cloning vector database*, 1995 – ftp://ftp.ncbi.nlm.nih.gov/repository/vector-ig/vectors.txt).

Entretanto, para realizar a procura de seqüências de vetores dentro das seqüências de DNA, o NCBI, dispõe de um banco de dados especial, **o UniVec**, utilizado para identificar rapidamente seqüências originadas de vetores. A procura dessas seqüências contra esse banco é eficiente e rápida, pois várias pequenas seqüências encontradas em diversos vetores estão presentes nele apenas uma vez, eliminando a redundância. O UniVec contém também seqüências de adaptadores, *linkers* e *primers* freqüentemente utilizados no processo de clonagem de cDNA e de DNA genômico, permitindo que essas seqüências sejam encontradas durante a procura (VecScreen – *The univec database*, 2001 – http://www.ncbi.nlm.nih.gov/ VecScreen/UniVec.html). Seu *download* pode ser feito através do *site* ftp://ncbi.nlm.nih.gov/pub/UniVec/. Existem, na verdade, dois arquivos de UniVec que podem ser obtidos no *site* de FTP acima mencionado: o **UniVec** e o **UniVec_Core**. Ambos são arquivos contendo seqüências de vetores no formato FASTA.

O arquivo UniVec inclui várias seqüências e foi feito para maximizar a detecção de contaminantes e, exatamente por isso, ele pode produzir falsos positivos. Ou seja, quando utilizado como entrada em programas de mascaramento de vetores, o resultado poderia mostrar regiões mascaradas que representam realmente dados de cDNA dos organismos e não seqüências de vetores. Assim, quando o UniVec é utilizado com esses propósitos, é necessária uma revisão humana para verificar a presença de falsos positivos (Kitts *et al.* 2002).

Já o arquivo UniVec_Core, que é uma parte do arquivo UniVec, foi feito para minimizar o número de falsos positivos, de forma que não seja necessária uma revisão humana das seqüências. Esse arquivo não contém vetores com origem de replicação de mamíferos, dentre outros, e, portanto, algumas regiões que seriam mascarados utilizando-se o UniVec não serão ao utilizar o UniVec_Core (Kitts *et al.* 2002). Nesse trabalho utilizamos, primeiramente, o arquivo Univec_Core para a retirada de regiões de vetores das seqüências de *S. mansoni* (veja Metodologia para mais informações).

1.6. Programas de Alinhamento de Seqüências

O alinhamento de seqüências consiste no processo de comparar duas seqüências (de nucleotídeos ou proteínas) de forma a se observar seu nível de identidade. Isso é feito para descobrirmos o grau de similaridade entre as seqüências de forma que possamos inferir (ou não) a uma delas, alguma propriedade já conhecida da outra (Prosdocimi *et al.* 2003). O alinhamento entre duas seqüências pode ser feito de forma global ou local (Figura 6).

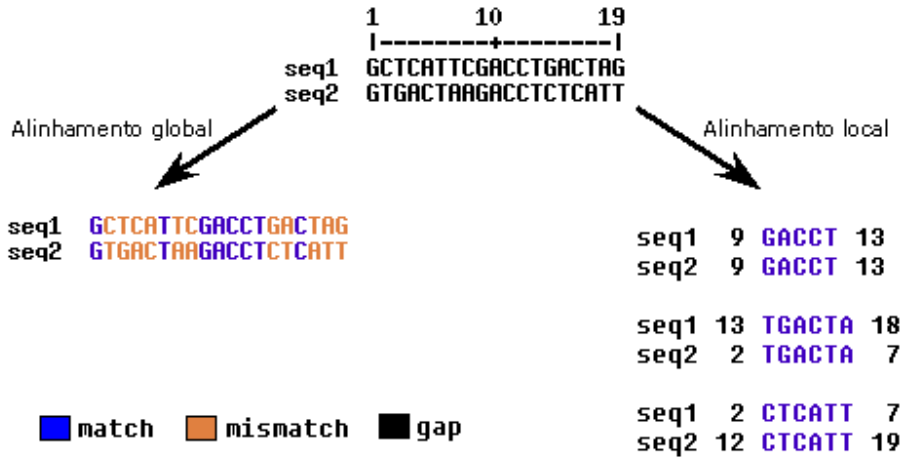

***Figura 6. Alinhamento global e local**. À esquerda vemos um exemplo de como é feito um alinhamento global das seqüências e à direita vemos um exemplo da realização de um alinhamento local. Retirado de Prosdocimi et al. 2003.*

1.6.1. ALINHAMENTO GLOBAL

O alinhamento global é feito quando comparamos uma seqüência de aminoácidos ou nucleotídeos com outra, ao longo de toda sua extensão (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html). O programa MULTALIN realiza esse tipo de alinhamento (Corpet 1988). Nesse caso são dados valores em uma matriz de comparação para as similaridades (*matches*), diferenças (*mismatches*) e falhas (*gaps*) encontrados durante o alinhamento das seqüências. As somas dos valores do alinhamento, de acordo com essa matriz de comparação, resulta num valor, que é um escore de similaridade entre as seqüências (FIGURA 7). No MULTALIN não é dado escore de similaridade (já que ele permite o alinhamento de várias seqüências ao mesmo tempo), e a semelhança entre as seqüências deve ser medida através de inspeção visual.

1.6.2. ALINHAMENTO LOCAL

O alinhamento local acontece quando a comparação entre duas seqüências não é feita ao longo de toda sua extensão, mas sim através de pequenas regiões destas (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html) (FIGURA 7).

O principal programa utilizado para o alinhamento local de seqüências é o BLAST (*Basic Local Alignment Search Tool*), encontrado em http://www.ncbi.nlm.nih.gov/BLAST/. Esse *software* compreende um conjunto de algoritmos de comparação de seqüências montado de forma a explorar toda a informação contida em bases de dados de DNA e proteínas (*http://www.ncbi.nlm.*

nih.gov/BLAST/blast_overview.html). Os programas BLAST foram desenvolvidos de modo a aumentar ao máximo a velocidade da busca por similaridade, já que as bases de dados são grandes e vêm crescendo exponencialmente, mesmo correndo o risco de perder um pouco na sensibilidade do resultado (Altschul *et al.* 1997). A rapidez da busca deve-se ao fato de que o programa quebra as seqüências de entrada e das bases de dados em fragmentos – as palavras (*words*) – e procura, inicialmente, similaridades entre elas. A busca é então feita com palavras de tamanho **W** que devem apresentar pelo menos um escore **T** de alinhamento entre si, dado de acordo com uma matriz de valores. Assim, as palavras que apresentam esse escore **T** (maior responsável pela velocidade e sensibilidade da busca) (Altschul *et al.* 1997) são estendidas em ambas as direções para ver se geram um alinhamento com um escore maior do que **S**. Uma outra vantagem de se utilizar o alinhamento local feito pelo BLAST é que, dessa forma, é possível identificar relações entre seqüências que apresentam apenas regiões isoladas de similaridade (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/similarity.html). Nesse projeto os valores das variáveis **W**, **S** e **T** não foram alterados, de forma que utilizamos seus valores DEFAULT para obtermos os resultados desejados.

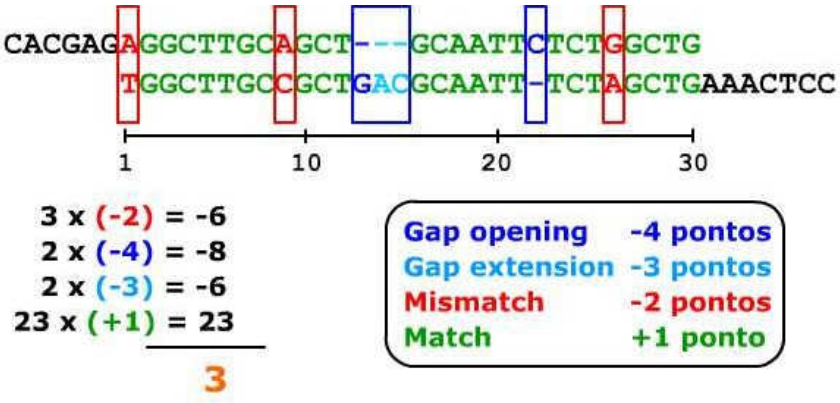

***Figura 7. Alinhamento de seqüências**. O alinhamento de seqüências de DNA é feito através da procura de uma região de similaridade entre duas seqüências. Quando essa região é encontrada são dados pontos para similaridades (match), diferenças (mismatches), abertura de falhas (gap opening) e extensão de falhas (gap extension) que possam ser encontradas no seu alinhamento. A somatória dos pontos desse alinhamento é chamado de escore do alinhamento e, no exemplo mostrado, o escore do alinhamento é 3. Tais escores são contabilizados tanto nos alinhamentos globais quanto locais.*

Os resultados do BLAST são então apresentados de acordo com dois parâmetros: o valor do escore (*Score bits*) e o valor E (*e-value*). O valor de escore depende do tamanho do alinhamento, do número de *matches*/*mismatches*/*gaps* e da matriz de comparação de seqüências utilizada e é normalizado através de variáveis estatísticas (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/

Blast_output.html). Já o valor E representa o número de alinhamentos com escores iguais ou melhores que "S" que seria de se esperar que ocorressem ao acaso numa base de dados do tamanho da utilizada. Assim, quanto menor o valor E, melhor o alinhamento, de forma que (num banco de dados de grandes proporções) um valor de E igual a 0 significa que não há chance de que um alinhamento entre as duas seqüências tenha ocorrido por mero acaso (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html).

O BLAST apresenta diferentes subprogramas que devem ser utilizados de acordo com o tipo de seqüência de entrada e os bancos de dados que se deseja pesquisar. A TABELA 1 apresenta as possibilidades de entrada, bancos de dados e programa a ser utilizado.

TAB1. UTILIZAÇÃO DOS PROGRAMAS BLAST

FORMATO DA SEQÜÊNCIA DE ENTRADA	BANCO DE DADOS	FORMATO DA SEQÜÊNCIA QUE É COMPARADO	PROGRAMA BLAST ADEQUADO
Nucleotídeos	Nucleotídeos	Nucleotídeos	BLAST**n**
Proteínas	Proteínas	Proteínas	BLAST**p**
Nucleotídeos	Proteínas	Proteínas	BLAST**x**
Proteínas	Nucleotídeos	Proteínas	**t**BLAST**n**
Nucleotídeos	Nucleotídeos	Proteínas	**t**BLAST**tx**

***Tabela 1**: Programas BLAST utilizados de acordo com o formato de entrada de seqüência e banco de dados desejados. Adaptada de http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/query_tutorial.html.*

Nesse trabalho, onde utilizamos seqüências de nucleotídeos como entrada, executamos os programas BLASTn (para busca no banco de dados de nucleotídeos) e BLASTx (para a busca no banco de dados de proteínas). Nesse último a seqüência de entrada é traduzida nas seis fases de leitura, produzindo seis seqüências de proteínas, que são, então, utilizadas como entrada para a execução de um BLASTp (http://www.ncbi.nlm.nih.gov/blast/html/BLASThomehelp.html).

1.7. AGRUPAMENTO DE SEQÜÊNCIAS DE DNA (*CLUSTERING ANALYSIS*)

Na análise de *clustering*, as seqüências de ESTs tratadas do organismo em questão são utilizadas como entrada em um programa. Este deve comparar essas seqüências entre si, de forma a encontrar quais delas são idênticas ou contêm regiões parecidas o suficiente para que sejam reunidas em uma só. Assim, o programa apresenta uma saída contendo as seqüências que foram agrupadas – chamadas de **consensos** ou **contigs** – e as seqüências que não foram reunidas (por não apresentarem similaridade suficiente com nenhuma outra) – chamadas de **singlets**. Cada uma

das seqüências resultantes do agrupamento (seja ela uma *singlet* ou um *contig*) é chamada de ***unique*** (FIGURA 8). Considerando uma análise ideal (e utópica), cada uma das seqüências *unique* deve representar um gene distinto. Entretanto, na prática, a presença de famílias gênicas e de genes duplicados dificulta a obtenção desse resultado ideal e, muitas vezes, a seqüência *unique* pode representar mais de um gene. Em outras ocasiões, um mesmo gene pode estar representado por mais de um *unique*, sendo que um dos *uniques* pode corresponder, por exemplo, à extremidade 5' de um determinado gene e outro à extremidade 3' do mesmo.

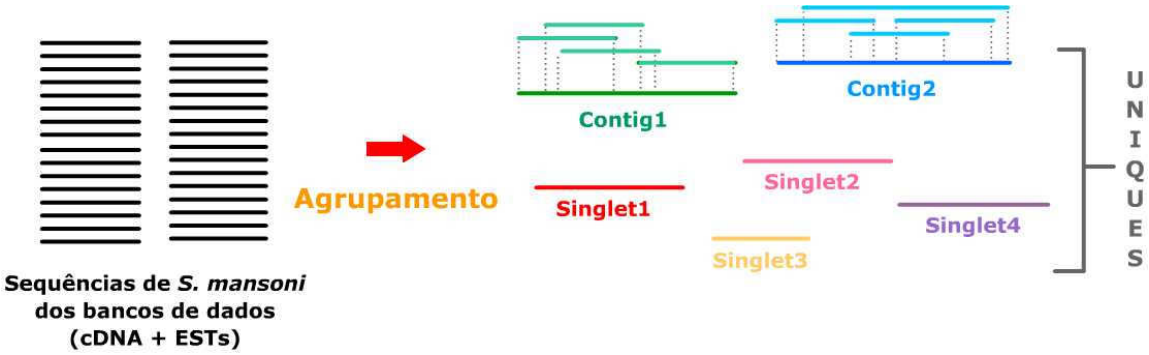

Figura 8. Agrupamento das seqüências*. O agrupamento das seqüências produz as seqüências não-redundantes, chamadas de uniques. As uniques são o conjunto das seqüências consenso mais as seqüências singlets.*

O agrupamento das seqüências é importante devido, principalmente, aos seguintes fatores: (1) elimina a redundância das seqüências, (2) aumenta o tamanho das seqüências facilitando a anotação por homologia (Oliveira & Johnston 2001), (3) aumenta o nível de confiabilidade de cada seqüência (Miller *et al.* 1999). Diferentes abordagens têm sido utilizadas para o agrupamento de seqüências de ESTs. O Unigene do NCBI utiliza comparações de seqüências em vários níveis de rigor para agrupar as seqüências em consensos (http://www.ncbi.nlm.nih.gov/ UniGene/build.html) (Schuler 1997). No TIGR, os índices gênicos são formados utilizando um *software* desenvolvido por eles mesmos, o TIGR *Assembler*, ou o CAP3 (Liang *et al.* 2000). Já no projeto genoma humano (HPG) as seqüências são agrupadas utilizando-se o *software* PHRAP (*International Human Genome Sequencing Consortium* 2001).

Já foram publicados trabalhos mostrando o agrupamento de seqüências de *S. mansoni* utilizando outros *softwares* e outras abordagens (Franco *et al.* 1997). Atualmente alguns serviços disponibilizam seqüências *uniques* de *S. mansoni* agrupadas de diferentes formas e utilizando diferentes *softwares*. Esses serviços podem ser encontrados nos seguintes *websites* (Oliveira & Jonhston 2001):

- ✓ *S. mansoni* and *S. japonicum clusters* at the Schistosome Genome Network *website*: http://www.nhm.ac.uk/hosted_sites/schisto/clusters/intro.html;
- ✓ *S. mansoni* gene index at The Institute for Genomic Research: http://www. tigr.org/tdb/smgi (Quackenbush *et al.* 2001);
- ✓ *S. mansoni clusters* at University of Pennsylvania: http://www.cbil.upenn.edu/ParaDBs/Schistosoma_2/index.html.

Apesar disso, preferimos realizar nossa própria análise, utilizando dois dos principais *softwares* para o agrupamento de seqüências: PHRAP e CAP3. Analisando suas capacidades podemos definir melhor a qualidade desejada em nossos *clusters* e mantê-los atualizados de acordo com nossa necessidade.

1.8. Descrição do Algoritmo Utilizado pelo CAP3 para o Agrupamento de Seqüências

Segundo a documentação do programa (Huang 1999) e o artigo publicado (Huang & Madan 1999), o CAP3 realiza os seguintes procedimentos para definir as seqüências presentes em cada consenso e montar sua seqüência:

1. Corte das regiões de baixa qualidade 5' e 3';
2. Realização do alinhamento global das seqüências entre si;
3. Cálculo do escore de alinhamento de cada par de seqüências (tamanho da seqüência sobreposta x qualidade da região de sobreposição x escores de *match/mismach/gap*) através de alinhamento global;
4. Realização de alinhamentos locais para identificar falsas sobreposições;
5. Observação do arquivo de entrada contendo a identificação das seqüências e o tamanho máximo e mínimo de distâncias entre elas (o CAP3 permite a utilização desse tipo de arquivo, o que proporciona sua utilização em projetos onde há seqüenciamento apenas das extremidades de clones), identificando falsas sobreposições;
6. Comparação do resultado do escore com os valores limites definidos

   Se o escore do alinhamento for menor do que o escore mínimo as seqüências não formam um agrupamento, se for maior, as seqüências o formam;
7. Alinhamento global das seqüências de cada consenso;
8. Cálculo dos valores de qualidade dos nucleotídeos de cada seqüência em cada posição do alinhamento global, para definir qual base será adicionada ao consenso;

9. Análise das deleções e inserções entre as seqüências para definir a montagem do consenso;
10. Montagem das seqüências consenso finais.

1.9. Descrição do Algoritmo Utilizado pelo PHRAP para o Agrupamento de Seqüências

Segundo a documentação do programa (Green 1999), o PHRAP funciona da seguinte forma:

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

1.10. ANOTAÇÃO DE GENOMAS

As seqüências genômicas são fontes ricas de informações sobre a biologia dos organismos, mas devem ser traduzidas através de análises computacionais e de interpretação biológica para que possamos extrair delas a maior quantidade possível de dados úteis (Lewis *et al.* 2000). A anotação genômica consiste num processo de vários passos e Stein (2001) divide-a, em três categorias básicas: a anotação de nucleotídeos, de proteínas e de processos (FIGURA 9).

***Figura 9. Anotação de genomas completos**. Esquema representando as fases e as perguntas que se deseja responder em cada uma das fases da anotação de genomas. Retirado de Prosdocimi et al. 2003.*

A anotação de nucleotídeos é feita quando existem informações sobre o genoma completo (ou segmentos de DNA) de algum organismo. Assim, procura-se encontrar a localização física (posição cromossômica) de cada parte da seqüência e descobrir onde estão os genes (Rouzé 1999), RNAs, elementos repetitivos, etc. Na anotação de proteínas, que é feita quando existem informações sobre os genes (obtidos por seqüenciamento genômico ou de cDNA) de algum organismo, procura-se identificar os genes já descobertos e descobrir sua função. Assim é possível saber quais são aqueles que determinado organismo possui e quais ele não possui. A anotação de processos procura identificar as vias e processos nos quais diferentes genes interagem, montando uma anotação funcional eficiente.

Nesse trabalho realizamos, principalmente, a anotação em nível protéico e abordamos vários aspectos da anotação de processos.

1.10.1. ANOTAÇÃO DE PROTEÍNAS

Nessa etapa da anotação genômica procura-se montar um catálogo das proteínas e genes presentes nos organismos, nomeá-los e associá-los a prováveis funções através, principalmente, de buscas por homologias (Aubourg & Rouzé 2001).

Várias técnicas recentes têm sido desenvolvidas para identificar automaticamente as proteínas pertencentes a diferentes grupos isofuncionais (chamados erroneamente de grupos de ortologia –

Jensen 2001), entretanto muitas dessas técnicas podem gerar classificações ambíguas. Na prática, o que é normalmente feito é a classificação das proteínas preditas com base em domínios funcionais, configurações espaciais e presença de padrões conservados, além de pesquisa ampla de similaridade contra proteínas bem caracterizadas.

Uma forma comum de se realizar a anotação de proteínas é procurar similaridades das seqüências com proteínas presentes em diferentes bancos de dados, utilizando ferramentas de alinhamento local como o BLASTp ou PSI-BLAST (Altschul *et al.* 1997). As coleções mais valiosas de seqüências de proteínas são os bancos de dados SWISS-PROT e TrEMBL. O primeiro apresenta uma coleção de seqüências de proteínas confirmadas e extensivamente anotadas. Ele contém ainda referências para outros bancos de dados de seqüência e estrutura, referências bibliográficas, identificação da família protéica e descrições sobre a provável função e papel biológico da proteína (Bairoch & Apweiler 2000). Entretanto, a velocidade do seqüenciamento genômico é maior que a dos curadores e, por isso, foi criado o banco de dados TrEMBL, que contém uma tradução automática das seqüências codificadoras (cds) submetidas aos bancos de dados de nucleotídeos (Lang 1997, Apweiler 2000).

Uma análise complementar seria a procura de domínios funcionais, sendo que as bases de dados mais utilizadas nesse processo são: PFAM, PRINTS, PROSITE, ProDom, SMART e BLOCKS. Esses vários bancos de dados de padrões são altamente sobreponíveis, mas cada um possui seu próprio sistema de nomenclaturas e método de procura, o que torna difícil a interpretação dos resultados (Stein 2001). Por isso foi desenvolvido, recentemente, um banco integrado de assinaturas de proteínas, conhecido como InterPro, que procura integrar as informações dos bancos anteriormente citados. Cada entrada do InterPro contém uma breve descrição da família ou domínio, uma lista de proteínas do SWISS-PROT ou TrEMBL que o contém, referências bibliográficas e *links* para cada um dos bancos membros (Apweiler *et al.* 2001). Infelizmente, o InterPro ainda não apresenta um sistema de busca em larga escala de fácil utilização e, portanto, não pudemos utilizá-lo no presente trabalho.

O banco InterPro tem sido utilizado para a anotação de diversos genomas, como o de leveduras, vermes, moscas, mostardas e homens. Desses, cerca de 40% a 50% das proteínas preditas possuem pelo menos uma entrada no InterPro, de onde se conclui que metade das proteínas eucarióticas pertence a novas famílias protéicas e que muito ainda precisa ser aprendido (Apweiler *et al.* 2001).

1.10.2. Anotação de Processos

A parte mais interessante e desafiadora do processo de anotação gênica é relacionar, finalmente, a genômica com os processos biológicos. Para isso foi criado um consórcio chamado **Gene Ontology** (GO), que busca criar um vocabulário padrão para descrever a função dos genes eucarióticos. Ele consiste em três divisões: função molecular (atividade específica do gene em questão, por exemplo: atividade enzimática), processos biológicos (processo no qual o gene está inserido, como a meiose) e componentes celulares (descreve a estrutura celular na qual o gene está localizado, como organelas ou ribossomos) (*The Gene Ontology Consortium* 2000). Apesar de que o GO realmente parece ser o processo mais adequado para a anotação gênica preferimos, nesse trabalho, não o utilizar. Essa escolha deveu-se ao fato de que faltam informações pertinentes nos artigos e no *site* do projeto (http://www.geneontology.org) sobre como anotar e identificar corretamente os genes e proteínas através desse serviço. Acreditamos que, num futuro próximo, todos os organismos utilizarão o padrão GO de anotação.

Para a anotação de processos é necessário mais do que trabalho computacional. Técnicas biológicas em larga escala, como mutagênese mediada por transposons, análise de expressão em *microarrays*, *RNA interference*, identificação de proteínas por espectroscopia de massa, ensaios baseados em *green-fluorescent-protein* para determinar a localização subcelular e padrões temporais de expressão de proteínas e estudos de duplo-híbrido em leveduras têm sido de fundamental importância para identificar o papel de genes e proteínas nos processos biológicos (Stein 2001). Cada novo experimento adiciona mais informação e permite um melhor entendimento do genoma.

No presente trabalho a anotação de processos foi realizada de modo a tentar correlacionar os genes mais expressos em diferentes fases de desenvolvimento de *S. mansoni* com a biologia de cada uma delas. Além disso, foi realizada a divisão de todos os *uniques* anotados em categorias funcionais de acordo com as categorias dos COGs (*Clusters of Ortologous Groups* – http://www.ncbi.nlm.nih.gov/COG) e os genes encontrados em cada categoria foram analisados com relação às vias bioquímicas das quais participam. Realizamos ainda a comparação do número de genes observados em cada uma das categorias com o número de genes observados na categorização de outros organismos eucarióticos de genoma completo.

1.10.3. A REALIZAÇÃO DA ANOTAÇÃO GENÔMICA

Stein (2001) propõe alguns modelos bastante pertinentes para explicar como é realizada, passo a passo, a anotação genômica. Segundo ele, esses processos de identificação gênica normalmente seguem algum dos seguintes modelos organizacionais: a fábrica, o museu e a festa. Cada modelo é adequado para alguma das fases do trabalho de anotação (Stein 2001).

Durante a primeira fase, quando o principal trabalho é encontrar genes e mapear variações e marcadores, o modelo da fábrica é o melhor. Nesse modelo uma rede de computadores trabalha seguindo uma série de programas de anotação. A seqüência de entrada é jogada numa série de programas para predição de genes, procura de similaridades entre seqüências de nucleotídeos e proteínas e procura de domínios funcionais. Isso permite a geração de grandes quantidades de dados sobre o genoma.

Então se inicia a fase de museu, quando a ênfase passa da localização dos dados para a sua interpretação. Nesse modelo um conjunto de curadores deve classificar e catalogar o genoma de forma sistemática, encontrando e corrigindo erros feitos pelos programas na primeira etapa. A maior parte dessa etapa é feita a mão e deve basear-se também na literatura obtida sobre o organismo em questão para uma melhor integração com os dados genômicos.

Após o tédio da curadoria é hora da festa. Nesse modelo, vários biólogos e bioinformatas são colocados juntos em um mesmo ambiente para discutir, anotar e realizar o fechamento do genoma. Os biólogos procuram associar os dados de genoma à biologia do organismo, montando várias hipóteses de trabalho e os bioinformatas montam ferramentas e dão o suporte técnico para ajudar a produzir os resultados desejados. Esse modelo já foi utilizado para a anotação do genoma de *Drosophila* (Adams *et al.* 2000) e do camundongo (*The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium* 2001).

É interessante notar que, enquanto o seqüenciamento genômico é uma tarefa bastante especializada, a anotação genômica é algo bastante multidisciplinar, no qual toda a comunidade científica (biológica) pode e deve contribuir.

Em nosso trabalho realizamos a etapa de fábrica (agrupamento das seqüências e realização de BLASTs) e de museu (anotação e classificação dos *uniques*). Entretanto, como o genoma ainda não está completo e esse é um trabalho preliminar de estudo do *status* atual do transcriptoma de *S. mansoni*, preferimos adiar a festa.

2. OBJETIVOS

2.1. OBJETIVO GERAL

Compilar as informações produzidas pela comunidade científica sobre o *status* atual de desenvolvimento do projeto genoma (transcriptoma) do verme *Schistosoma mansoni* e publicar um *website* contendo informações relevantes.

2.2. OBJETIVOS ESPECÍFICOS

1. Produzir uma versão mais confiável de seqüências de *S. mansoni*, livres de regiões contaminantes, como vetores, adaptadores e sítios de restrição;
2. Identificar os genes mais expressos em cada fase de desenvolvimento do parasita e correlacionar esses dados com a biologia do organismo;
3. Produzir uma coleção de seqüências não-redundantes através do agrupamento (utilizando os programas PHRAP e CAP3) e descobrir qual a porcentagem aproximada de genes de *S. mansoni* já foi seqüenciada e está disponível nos bancos de dados de seqüências;
4. Identificar quais genes do verme já foram seqüenciados, realizando a anotação dos *uniques*;
5. Classificar os *uniques* identificados em categorias funcionais;
6. Construir um *website* simples, agradável e de fácil navegação para abrigar todas essas informações.

3. MATERIAIS E MÉTODOS

3.1. MATERIAIS UTILIZADOS

3.1.1. COMPUTADORES

- ✓ **Máquina Rigel**
  Partição da estação Sun de alto-desempenho E-10.000 (com 32 processadores e 17Gb de memória RAM), contendo 4 processadores e 4Gb de RAM.
- ✓ **Máquina Genoma**
  Pentium II com processador de 500MHz e 256Mb de memória RAM.
- ✓ **Máquina Proteu**
  Pentium II com processador de 500MHz e 128Mb de memória RAM.

3.1.2. *SOFTWARES*

- ✓ **PHRAP e CROSS_MATCH**
  Ambos os *softwares* são fornecidos gratuitamente, num mesmo pacote, para o uso acadêmico. Para adquirí-los foi necessário entrar no *site* que contém o contrato acadêmico (http://www.phrap.org/consed/academic_agreement. txt) e redigir um e-mail contendo uma cópia do contrato, informações sobre nome, e-mail, instituição e os programas desejados. É necessário constar nesse *mail* que o usuário concorda com o que está escrito no contrato. A mensagem eletrônica foi enviada para o criador do *software*, Dr. Phil Green, da universidade de Washington, phg@u.washington.edu, e, dessa forma, os programas foram enviados, compactados, para o e-mail que constava no documento.
- ✓ **CAP3**
  O programa CAP3 está disponível gratuitamente através de pedido a Xiaoqiu Huang (xqhuang@cs.iastate.edu). Mandamos, então, uma mensagem eletrônica para o autor, dizendo em qual plataforma gostaríamos de executar o *software*. Recebemos, conforme pedido, o programa compilado para Solaris e Linux.
- ✓ **UNIX**
  Sistema operacional já instalado nos computadores do CENAPAD. Vários comandos do pacote UNIX foram utilizados, a saber: grep, awk, sort, vi, vim, time, etc.
- ✓ **Pacote BLAST**
  Através de FTP anônimo fizemos também o *download* dos arquivos executáveis do pacote BLAST do NCBI, através do endereço ftp://ncbi.nlm. nih.gov/blast/executables. O BLAST foi instalado em plataforma Sun (máquina Rigel). Fizemos também o *download* das bases de

dados não-redundantes de nucleotídeos, proteínas, da base de dados do PDB e do SWISS-PROT através de FTP em ftp://ncbi.nlm.nih.gov/blast/db.

- ✓ **Interpretador PERL**
  Instalado junto com o sistema operacional linux (máquina genoma) ou solaris (máquina rigel). Utilizado para executar os programas gerados na linguagem PERL.
- ✓ **Microsoft Frontpage 2000 e Microsoft Word 2000**
  Utilizados para a montagem do *website*. Previamente comprados e instalados.
- ✓ **Macromedia Flash MMX**
  Utilizado para a montagem das figuras do *website*. Previamente comprado e instalado.

3.2. OBTENÇÃO DAS SEQÜÊNCIAS DE *S. mansoni*

As seqüências de nucleotídeos de *S. mansoni* foram obtidas no GenBank (http://www.ncbi.nlm.nih.gov). Para selecionarmos apenas as seqüências de mRNA do parasita, fizemos a seguinte busca no GenBank.

```
("Schistosoma mansoni"[Organism] AND biomol_mrna[PROP])
```

A primeira etapa do projeto (até seção 3.6, inclusive) foi realizada com um banco de dados de seqüências de *S. mansoni* já presente nas máquinas do CENAPAD e o conjunto de dados foi atualizado para a segunda etapa do projeto. Assim a primeira etapa do trabalho foi realizada com as seqüências obtidas através de busca no GenBank no dia 25/05/2001. A partir daqui iremos nos referir ao arquivo contendo essas seqüências como Sm0501. Para a segunda etapa do projeto o conjunto de dados foi atualizado através da obtenção das seqüências de mRNA de *S. mansoni* do GenBank, no dia 19/04/2002. O arquivo contendo as seqüências atualizadas será chamado de Sm0402.

3.3. TRATAMENTO DAS SEQÜÊNCIAS OBTIDAS DO GENBANK

As seqüências de entrada (Sm0501 e Sm0402) foram tratadas de forma a retirar as regiões que não representam o mRNA de *S. mansoni*. Assim, foram retiradas as seqüências de vetores, adaptadores e sítios de restrição mais comumente utilizadas na montagem de bibliotecas de cDNA. Além disso, para evitar erros no agrupamento das seqüências, foram retiradas as seqüências menores do que 150 nucleotídeos e com mais de 4% de ambigüidade (N ou X). Das seqüências

restantes, foram retirados os sítios de poli T 5' e poli A 3' para evitar o agrupamento incorreto pelos programas. Um esquema dessa metodologia pode ser conferido na FIGURA 10, a seguir:

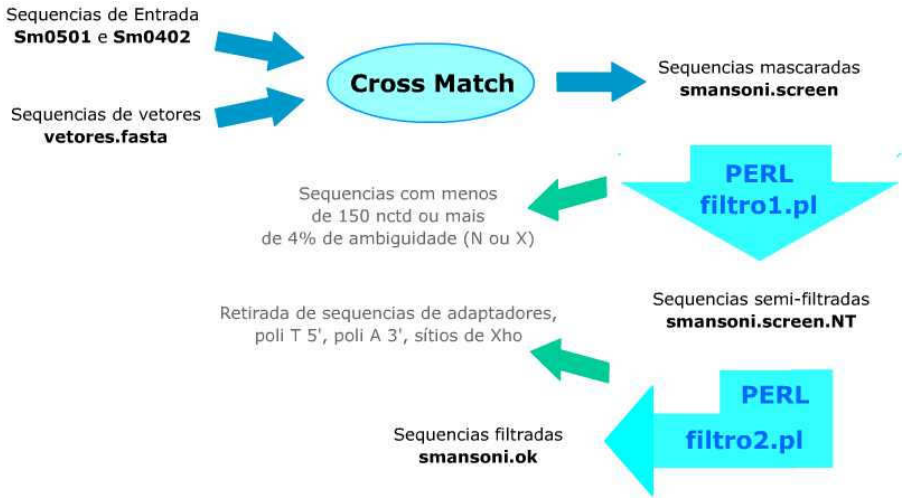

Figura 10. Esquema da metodologia utilizada para a filtragem das seqüências. *Seqüência de passos utilizada para a filtragem das regiões que não representam mRNA, das seqüências pequenas e de baixa qualidade e de regiões que poderiam atrapalhar o posterior agrupamento das seqüências de S. mansoni.*

3.3.1. RETIRADA DE SEQÜÊNCIAS DE VETORES

As seqüências de entrada do GenBank foram, primeiramente, filtradas para retirar as seqüências de vetores que poderiam existir em suas extremidades. Para isso utilizamos o *software* CROSS_MATCH.

Começamos o trabalho utilizando o arquivo **UniVec_Core** para mascarar as seqüências de vetor do nosso arquivo de seqüências. Entretanto, ao observar o arquivo contendo as seqüências mascaradas, notamos que havia falsos positivos e que, mesmo o UniVec_Core, estava mascarando algumas regiões que não eram vetores. Além disso, o programa deixava de mascarar um número muito grande de seqüências adaptadoras.

Devido a esse fato resolvemos montar nosso próprio arquivo de vetores contendo apenas os principais vetores utilizados na montagem de bibliotecas de cDNA. O nosso **Univec_Schisto** foi montado, contendo apenas as seqüências dos seguintes vetores de clonagem:

- ✓ **pGEM** Promega
  http://www.promega.com/vectors/pgem5zfm.txt;

- ✓ **pBlueScript SK-** Stratagene http://www.stratagene.com/vectors/sequences/ bl2ksm_s. txt;
- ✓ **pUC 18/19** (*Accession Number* - U03991.1).

Utilizando o *software* CROSS_MATCH realizamos a busca das regiões de vetor em nossas seqüências. Os parâmetros foram utilizados de acordo com a documentação do programa (Green 1999) e a linha de comando utilizada é mostrada a seguir:

```
cross_match smansoni univec_schisto –minmatch 10 –minscore 20 –screen
```

onde,
smansoni é o nome do arquivo contendo seqüências de mRNA de *S. mansoni* obtidas do banco de dados (Sm0501 ou Sm0402);
univec_schisto é o arquivo montado com a seqüência dos vetores;
-minmatch 10 é o parâmetro que mostra o tamanho mínimo da seqüência a ser comparada entre os dois arquivos FASTA (o arquivo com as seqüências de *S. mansoni* e o arquivo de seqüências de vetores);
-minscore 20 é o parâmetro que define o escore mínimo do alinhamento para o mascaramento das seqüências (nucleotídeos idênticos nas duas seqüências geram um valor +1). Se o escore de um alinhamento contra as seqüências de vetores é menor do que o MINSCORE, a seqüência não é considerada como contaminação;
-screen é o parâmetro que define que o arquivo de saída deve conter o arquivo de entrada contendo letras Xs nas regiões relativas a seqüências de vetores.

O programa deve gerar, portanto, um arquivo de saída – **smansoni.screen** – onde as seqüências que representavam vetores foram mascaradas com letras Xs.

3.3.2. RETIRADA DE SEQÜÊNCIAS PEQUENAS E DE BAIXA QUALIDADE

Segundo Adams e colaboradores (1991), uma seqüência de EST deve ser utilizada se apresentar menos de 4% de ambigüidade e mais de 150 nucleotídeos. Para obtê-las foi desenvolvido um programa em linguagem PERL (filtro1.pl) que utiliza o seguinte procedimento:

1. Conta o número de nucleotídeos de cada seqüência;
2. Observa a quantidade de Ns e Xs na seqüência;

3. Se o número de Ns + Xs for maior do que 4% do tamanho da seqüência, ela é jogada num arquivo de saída **smansoni.screen.nx** (tais seqüências não serão aproveitadas nas etapas posteriores do trabalho);
4. Se o número de nucleotídeos da seqüência for menor do que 150, ela é jogada num arquivo de saída **smansoni.screen.tam** (tais seqüências não serão aproveitadas nas etapas posteriores do trabalho);
5. Se o tamanho da seqüência for maior ou igual a 150 nucleotídeos, e o número de Ns + Xs for menor do que 4% do tamanho da seqüência, ela é jogada no arquivo **smansoni.screenNT** (essas serão as seqüências aproveitadas para a etapa posterior do trabalho).

Os arquivos **smansoni.screen.nx** e **smansoni.screen.tam** foram observados para que fosse confirmada a correta execução do programa.

Para que fossem retiradas seqüências adaptadoras mais freqüentemente utilizadas (CACGAG), sítios da enzima de restrição *Xho*I (CTCGAG) (utilizada para a clonagem da maioria das seqüências de *S. mansoni* dos bancos de dados), seqüências de poliT 5' (TTTTTTTTTT) e seqüências de poliA 3' (AAAAAAAAAA) do arquivo **smansoni.screenNT** foi feito um novo programa PERL (filtro2.pl) que realiza os seguintes passos:

- ✓ Observa na primeira linha de cada seqüência FASTA:
  Existência de seqüência adaptadora, poliT, e sítio de *Xho*I;
  Se encontrar, retirar tal seqüência e tudo que estiver 5' a ela (provavelmente seqüências de vetor);
- ✓ Observa na última e penúltima linha de cada seqüência FASTA:
  Existência de seqüência poliA, e sítio de *Xho*I;
  Se encontrar, retirar tal seqüência e tudo que estiver 3' a ela (provavelmente seqüências de vetor);

Cada linha de um arquivo no formato FASTA contém, normalmente, 50 nucleotídeos. Como a última linha de cada seqüência pode conter um único nucleotídeo (ou cinqüenta), resolvemos observar também a penúltima linha no segundo caso.

Para fins de confirmação do correto funcionamento, após sua execução, o programa apresenta na tela um resumo de tudo o que fez. Nesse resumo ele mostra o GI de cada seqüência alterada, mostrando o que foi alterado em cada uma delas (retirada de adaptador, sitio de *Xho*I, poliT ou poliA). Esse resumo foi utilizado para calcular o número e a porcentagem de seqüências que

apresentavam cada um dos contaminantes citados e também para a conferência da correta execução do programa.

Os arquivos resultantes dessas etapas foram chamados de **sm0501.ok** ou **sm0402.ok** e foram utilizado como base (entrada) para a execução dos programas de agrupamento de seqüências.

3.4. AGRUPAMENTO DAS SEQÜÊNCIAS DE *S. mansoni* UTILIZANDO OS SOFTWARES PHRAP E CAP3

O arquivo filtrado obtido na primeira etapa, sm0501.ok, foi então utilizado como entrada para a execução dos softwares de agrupamento de seqüências: PHRAP e CAP3. Devido a estudos anteriores com os dois softwares (Prosdocimi 2002, monografia de bacharelado), ambos foram utilizados segundo critérios mais rigorosos de agrupamento. Em ambos os casos as seqüências agrupadas foram aquelas que apresentavam pelo menos uma sobreposição de 40 nucleotídeos com um escore de alinhamento igual a 40.

As linhas de comando utilizadas para os dois softwares foram as seguintes:

```
$> cap3  sm0501.ok –o 40 –s 800
$> phrap sm0501.ok –minmatch 40 –minscore 40
```

Dessa forma, ambos os programas produziram arquivos de *contigs* e *singlets*, além de arquivos LOG, mostrando o que o programa realizou para sua execução.

3.5. IDENTIFICAÇÃO DOS *UNIQUES* MAIS EXPRESSOS POR CADA PROGRAMA

Foram produzidos, nessa etapa, dois programas PERL para calcular o número de seqüências agrupadas em cada cluster. Cada um dos programas desenvolvidos acessava algum dos arquivos de LOG produzidos pelo PHRAP ou CAP3 para obter essa informação. Dessa forma foram geradas duas tabelas por cada programa, mostrando o número de seqüências presente em cada cluster ou o número de *clusters* contendo determinado número de seqüências. Por exemplo, através da execução desse programa, observou-se, para o CAP3, que o Cluster 1 continha 5 seqüências e que havia 12 *clusters* contendo 15 seqüências.

3.6. IDENTIFICAÇÃO DAS FASES DE DESENVOLVIMENTO E ANOTAÇÃO DOS *UNIQUES* MAIS EXPRESSOS POR CADA PROGRAMA

Nesse momento tivemos que voltar ao GenBank para obter novamente as seqüências de mRNA de *S. mansoni* no formato Genbank, já que no formato FASTA, como havíamos obtido, não há informação sobre a fase de desenvolvimento da qual a seqüência foi produzida. Fizemos então o *download* das seqüências que compunham os *uniques* mais expressos no formato Genbank.

A identificação das fases de desenvolvimento foi realizada através da utilização de um programa PERL que procurava a informação sobre biblioteca no arquivo GenBank. Tal programa procurava, no local relativo à biblioteca, alguma das seguintes palavras: *egg*, *cercaria*, *lung*, *male*, *female*, *adult*, *adult worm* ou *AW*. Assim o programa produzia um arquivo de saída contendo os identificadores das seqüências de cada uma das fases para posterior conferência. Além disso, o programa recebia como entrada o número do identificador (GI) das seqüências de cada cluster e produzia uma tabela contendo, para cada cluster mais expresso, o número de seqüências presentes em cada uma das bibliotecas. Dessa forma descobrimos em quais fases de desenvolvimento estes genes estavam sendo transcritos.

A anotação dos *uniques* mais expressos foi realizada através de inspeção visual do resultado do BLASTx contra o banco não-redundante de proteínas do NCBI (nr) e do resultado do BLASTn contra o banco de dados não-redundante de nucleotídeos do NCBI (nt).

3.7. ANOTAÇÃO DOS *UNIQUES* DE *S. mansoni*

3.7.1. EXECUÇÃO DE PROGRAMAS DE ALINHAMENTO LOCAL (BLAST) CONTRA BANCOS DE DADOS DE SEQÜÊNCIAS PARA A PESQUISA DE HOMOLOGIA

Depois da análise dos dados obtidos na primeira etapa, os dados de seqüências de mRNA de *S. mansoni* foram atualizados e filtrados. Assim, o arquivo Sm0402.ok foi utilizado como entrada para o programa CAP3. Os *uniques* (consensos + *singlets*) gerados foram anotados manualmente através da observação visual da homologia contra diferentes bancos de dados. O arquivo de *uniques* foi gerado através da concatenação do arquivo de *singlets* e de *contigs*, realizada através do comando CAT do Unix. Utilizando o programa BLAST, rodando localmente, realizamos a busca por homologia contra os seguintes bancos de dados:

- ✓ Banco de dados **não-redundante de proteínas (NR)**: contém todas as proteínas depositadas no GenBank;
- ✓ Banco de dados **não-redundante de nucleotídeos (NT)**: contém todas as seqüências de nucleotídeos depositadas no GenBank (menos ESTs e seqüências genômicas finalizadas);
- ✓ Banco de dados do **SWISS-PROT (SP)**: é um dos bancos de dados mais bem organizado de proteínas, muito bem anotado, apresentando várias informações funcionais sobre as proteínas catalogadas;
- ✓ Banco de dados de **genoma de *S. mansoni***: produzido através do *download* (do GenBank) e formatação de todas seqüências de nucleotídeos gênicas e genômicas de *S. mansoni*;

De forma a limitar o tempo de processamento dos BLASTs – que chegava a durar uma semana inteira ocupando uma das melhores máquinas do CENAPAD (Rigel) – utilizamos um valor de *cut-off* para o e-value do programa BLAST de 1×10^{-10}. Esse valor foi escolhido devido à observação de BLASTs realizados previamente com valor de *cut-off* DEFAULT (10) e diminuiu sensivelmente o tempo de execução do programa. Observando resultados do BLAST com valor de *cut-off* DEFAULT, pudemos perceber que as seqüências homólogas com e-value maior do que 1×10^{-10}, na grande maioria das vezes, não apresentavam uma boa similaridade com a seqüência de entrada.

As linhas de comando executadas para as comparações com diferentes bancos de dados podem ser vistas abaixo:

NR

```
$> blastall -p blastx -d nr -i sm0402.capos –e 0.0000000001 -o sm.nr.x10
```

NT

```
$> blastall -p blastn -d nt -i sm0402.capos –e 0.0000000001 -o sm.nt.n10
```

SP

```
$> blastall -p blastx -d swissprot –i sm0402.capos -e 0.00000000001 -o sm.sp.x10
```

S. mansoni genome

```
$> blastall -p blastn -d schisto_genome –i sm0402.capos -e 0.00000000001 -o sm.ge.x10
```

3.7.2. Anotação Manual dos *Uniques* através de Observação dos Resultados dos BLASTs

A anotação manual dos *uniques* de *S. mansoni* foi realizada através da observação dos resultados obtidos, de cada um dos *uniques*, nas pesquisas de homologia (BLASTs) contra os bancos de dados do SWISS-PROT, NR e NT (Figura 11).

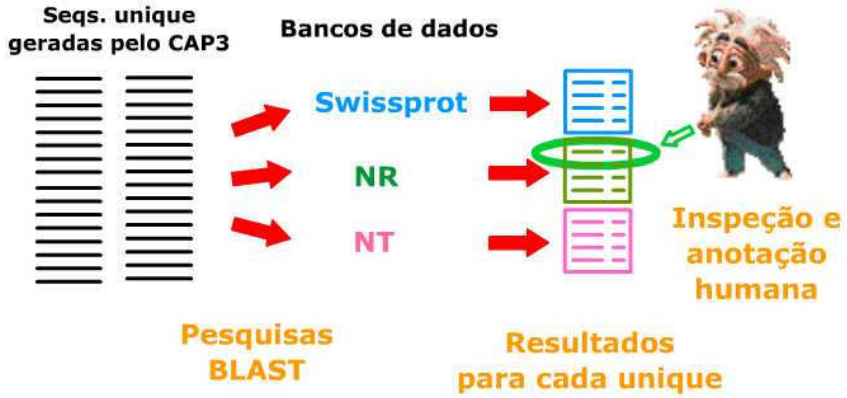

Figura 11. Esquema da metodologia utilizada para a anotação das seqüências. *As seqüências uniques eram alinhadas localmente contra os bancos de dados NR, NT e SWISS-PROT e a anotação era realizada observando o resultado obtido nesses alinhamentos.*

Alguns critérios básicos foram utilizados para realizar a anotação, dentre eles:

- ✓ **Prioridade para os resultados do BLAST-SP (BLAST SWISS-PROT)**
  Como o banco de dados SP é mais bem organizado e curado do que todos os outros utilizados, preferimos adotar sua anotação a utilizar anotações de outros bancos, como, por exemplo, o NR. Dessa forma, sempre que possível, utilizamos as anotações do SP, mais claras e completas, para identificar os genes de *S. mansoni*.
- ✓ **Apresentação dos números de classificação das enzimas (ECs)**
  Durante a anotação das enzimas nesse trabalho, tentamos, sempre que possível, identificar seu número de acordo com a atual nomenclatura enzimática (Bairoch 2000) da União Internacional de Bioquímica e Biologia Molecular (IUBMB - *International Union of Biochemistry and Molecular Biology*), de forma a tornar nossa anotação universal e de fácil comparação com outros sistemas de classificação e bancos de dados. O código de 4 dígitos EC define a classe das enzimas: 1, oxidoredutases; 2, transferases; 3, hidrolases; 4, liases; 5, isomerases; e 6, sintetases. O significado dos dígitos posteriores depende da classe da enzima e dá informações sobre substratos e cofatores. O último dígito representa

a especificidade ao substrato, mecanismo molecular ou o tipo de ligação química (Devos & Valencia 2001).

- ✓ **Consensos no BLAST-SP com diversos organismos**
  Quando eram observadas, no BLAST-SP, similaridades com a mesma proteína de diversos organismos diferentes, a seqüência foi identificada como:
  Putative gene for (nome da proteína em questão).
- ✓ **Altas similaridades com genes/proteínas de *S. mansoni***
  Quando eram observadas altas similaridades com genes/proteínas de *S. mansoni*, estes eram classificados como sendo do próprio organismo. Nessas ocasiões, o que acontecia era que, na formação dos consensos, os genes/proteínas originais sofreram pequenas alterações em suas seqüências e mostraram-se um pouco diferentes das seqüências depositadas nos bancos de dados. Assim, tais genes/proteínas foram identificadas como:
  Schistosoma mansoni (nome do gene/proteína em questão)
- ✓ **Baixas similaridades com genes/proteínas de *S. mansoni***
  Quando eram observados genes ou proteínas com baixas similaridades com proteínas de *S. mansoni*, estes eram classificados como similares a genes/proteínas do verme. Esta classe provavelmente representa genes duplicados, membros de famílias gênicas ou ESTs com erros de sequenciamento e foram identificadas como:
  Similar to *Schistosoma mansoni* (nome do gene/proteína em questão)
- ✓ **Similaridades com genes/proteínas de outros organismos**
  Quando não encontramos um consenso no resultado do BLAST-SP, identificamos as seqüências através de seus homólogos em outros organismos. Assim, eram observados os primeiros 10 *hits* e, se havia algum consenso (a mesma proteína era identificada entre vários deles), mesmo que o *best hit* não fosse identificado assim, preferimos identificar a seqüência como similar ao organismo que apresentava mais alto hit e estava dentro do consenso. Nos casos onde não havia consenso ou quando o valor de similaridade (e-value) do *best hit* era significativamente maior do que o das seqüências a seguir, preferimos identificar a seqüência como similar ao *best hit*. Assim, esses *uniques* ficaram identificados da seguinte maneira:
  Similar to (nome do organismo com o melhor hit no BLAST-NR ou BLAST-NT) (nome do gene/proteína em questão).

O BLAST contra o genoma de *S. mansoni* não foi utilizado para a anotação em si, pois seus resultados estavam já incluídos no resultado do BLASTn contra o banco de dados NT. Sua principal utilização foi para confirmar que determinados genes já foram encontrados no genoma do parasita.

3.8. CLASSIFICAÇÃO DOS *UNIQUES* DE *S. mansoni* EM CATEGORIAS FUNCIONAIS

Durante a anotação, os *uniques* foram também classificados de acordo com as categorias funcionais presentes nos COGs (*Clusters of Ortologous Groups* – http://www.ncbi.nlm.nih.gov/COG) do NCBI (Tatusov *et al.* 2001). A classificação foi realizada manualmente utilizando, para isso, informações funcionais presentes em arquivos do NCBI (Weller *et al.* 2002), em informações do SWISS-PROT, em programas de busca na Internet (Google – http://www.google.com), *sites* com informações compiladas sobre genes, como o KEGG (*Kyoto Encyclopedia of Genes and Genomes* – http://www.genome.ad.jp/kegg/) (Kanehisa *et al.* 2002) e em artigos e livros especializados das mais diversas áreas (Watson *et al.* 1997; Griffiths *et al.* 1998; Nelson e Cox 2000). As categorias presentes nos COGs podem ser observadas na FIGURA 12.

PROCESSAMENTO E ARMAZENAMENTO DE INFORMAÇÕES	
J	Tradução, estrutura ribossomal e biogênese
K	Transcrição
L	Replicação, recombinação e reparo do DNA
PROCESSOS CELULARES	
D	Divisão celular e particionamento cromossômico
O	Modificação pós-traducional, *turnover* de proteínas e chaperones
M	Biogênese do envelope celular, membrana externa
N	Mobilidade celular e secreção
P	Metabolismo e transporte de íons inorgânicos
T	Mecanismos de transdução de sinais
METABOLISMO	
C	Produção e conversão de energia
G	Metabolismo e transporte de carboidratos
E	Metabolismo e transporte de aminoácidos
F	Metabolismo e transporte de nucleotídeos
H	Metabolismo de coenzimas
I	Metabolismo de lipídeos
Q	Biossíntese, catabolismo e transporte de metabólitos secundários
POBREMENTE CARACTERIZADOS	
R	Somente predição de função geral
S	Função desconhecida

Figura 12. Categorias funcionais dos COGs. *O projeto COG classifica os genes em 4 grupos funcionais principais, sendo cada um deles dividido em subgrupos.*

Entendemos que um padrão de classificação eficiente, simples e de qualidade é difícil de ser criado e que a natureza não exibe os padrões fixos que gostaríamos de criar para que pudéssemos entender melhor o funcionamento dos genomas. Dessa forma, preferimos utilizar um padrão que já é amplamente conhecido na comunidade científica e que foi criado e é utilizado num dos maiores centros de bioinformática do mundo, o NCBI. É importante notar que, devido à artificialidade dos grupos de classificação, alguns genes foram incluídos em mais de uma categoria

funcional. Além disso, devido ao fato de que o COGs foi desenvolvido para lidar com organismos procariotos (apesar de conter também classificações de eucariotos), tivemos a necessidade de criar novas subcategorias e de adaptar certas famílias de proteínas em categorias já existentes.

3.9. MONTAGEM DE WEBSITE PARA A ANOTAÇÃO DO TRANSCRIPTOMA DE *S. mansoni*

De modo a facilitar e agilizar o processo de anotação dos milhares de *uniques* de *S. mansoni* foram criadas páginas *web* apresentando um formulário de preenchimento dos dados necessários para a anotação e todas as informações necessárias para preenchê-lo (FIGURA 13).

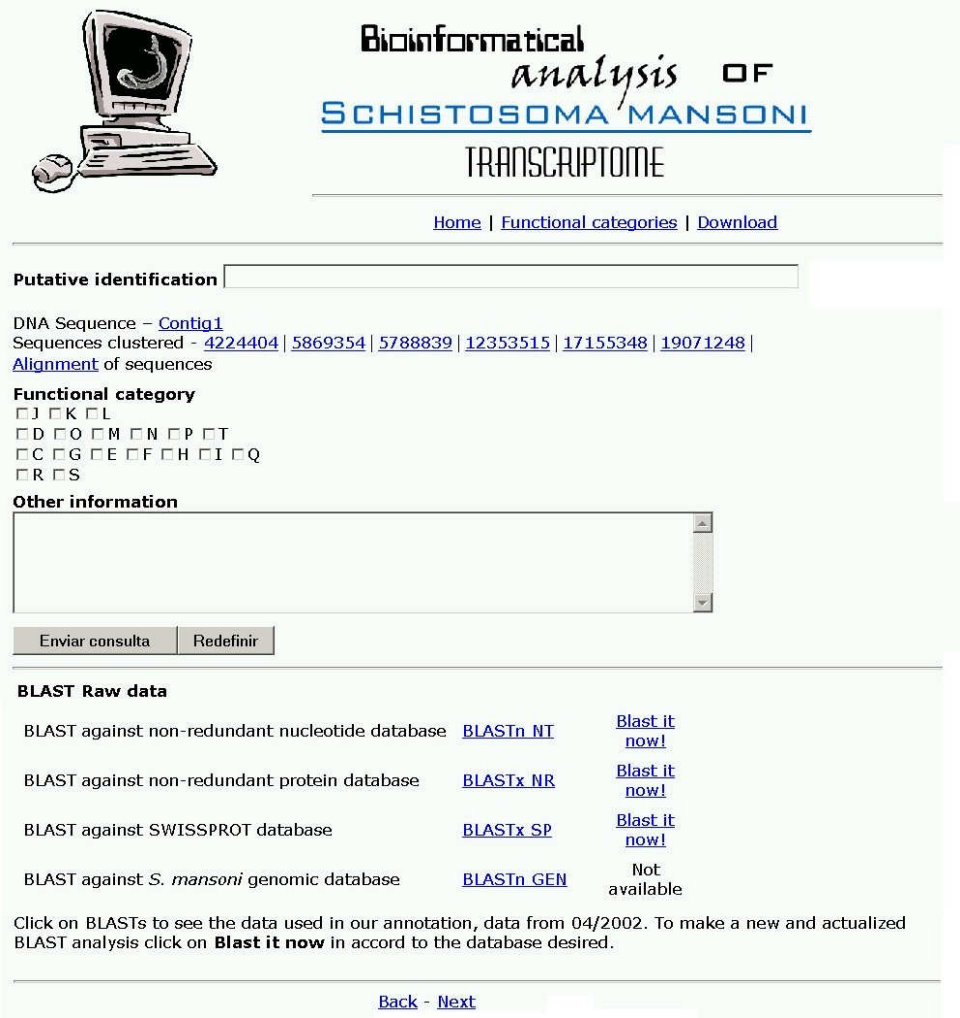

Figura 13. Página com formulário para a anotação. *Para facilitar a anotação foram criadas páginas para cada unique com seus respectivos BLASTs e com campos para a anotação dos nomes, categorias e comentários relativos a cada um dos uniques.*

3.10. MONTAGEM DE WEBSITE PARA PUBLICAÇÃO DAS INFORMAÇÕES OBTIDAS

O *site* foi desenvolvido utilizando-se o programa **Microsoft Frontpage**, versão 2000. As figuras foram feitas utilizando o *software* **Macromedia Flash MMX** . Através de programação PERL um *site* padrão montado no Frontpage foi modificado para gerar páginas com informações sobre cada *unique*. Assim, foi montada uma página para cada um dos *uniques*, contendo as seguintes informações:

- ✓ **Identificação do *unique*** (anotação);
- ✓ **Seqüências de nucleotídeos**

  No caso de consensos, foi apresentado o número dos GIs das seqüências que o formam, com link para o NCBI; a seqüência de nucleotídeos do *contig*; e o alinhamento das seqüências realizado pelo CAP3 para sua montagem;

  No caso de *singlets*, seu número de GI com link para o NCBI, sua seqüência deve ser obtida ali;
- ✓ **Categoria Funcional** do *Unique*, baseada nos COGs;
- ✓ **Principais homologias de nucleotídeos**

  As páginas de cada *unique* apresentam os quatro melhores *hits* obtidos no BLASTn, quando possível.
- ✓ **Principais homologias de proteínas**

  As páginas de cada *unique* apresentam os quatro melhores *hits* obtidos no BLASTx, quando possível.
- ✓ **Principais homologias contra o genoma de *S. mansoni***

  As páginas de cada *unique* apresentam os quatro melhores *hits* obtidos no BLASTn contra o genoma do verme, quando possível.
- ✓ **Resultados brutos**

  Foram apresentados os resultados de todos os BLASTs que utilizamos para realizar a anotação nas páginas. Assim os usuários podem ver os resultados de BLAST de cada *unique*.
- ✓ **Blast it now**

  Além do BLAST que utilizamos foi colocado na página um link para realizar um BLAST instantâneo no NCBI. Para ver o resultado do BLAST do *unique* em questão, atualizado, basta clicar no link **blast it now** de cada banco de dados. O botão aciona o serviço no NCBI e o resultado é visto na hora.

Além disso, o *site* contém informações sobre toda a metodologia utilizada no trabalho, assim como outras informações pertinentes. Um mecanismo de busca permite que o usuário procure seu gene

desejado de *S. mansoni* através do número de GI, AN, ou palavra-chave. O endereço é http://www.icb.ufmg.br/~lgb/schisto.

3.11. Análises dos Resultados da Anotação Gênica

Para a análise dos dados obtidos no processo de anotação gênica, identificamos o número de *uniques* em cada uma das seguintes classes:

1. *Uniques* sem anotação (*no match*);
2. *Uniques* anotadas através de homologia com genes de outros organismos;
3. *Uniques* anotadas através de homologia com genes de *S. mansoni*;
4. *Uniques* que representam genes de *S. mansoni*;

Foi contabilizado o número de *uniques* presentes em cada uma das categorias funcionais, de forma a possibilitar uma visão macro sobre o *status* atual do transcriptoma de *S. mansoni*.

4. RESULTADOS E DISCUSSÕES

4.1. OBTENÇÃO E INSTALAÇÃO DOS *SOFTWARES* UTILIZADOS

Os *softwares* PHRAP e CAP3 foram recebidos através de e-mail e instalados com sucesso, de acordo com as informações contidas na documentação.

O pacote BLAST, obtido através de *download* anônimo do *site* de FTP do NCBI (ftp://ftp.ncbi.nlm.nih.gov), foi instalado de acordo com instruções, assim como seus bancos de dados e o programa de formatação desses bancos (formatdb).

4.2. OBTENÇÃO DAS SEQÜÊNCIAS DE *S. mansoni*

A primeira busca das seqüências de *S. mansoni* no GenBank, que gerou o arquivo de entrada utilizado na primeira etapa do trabalho, foi realizada no dia 25/05/2001 e retornou 14.275 seqüências, dentre ESTs, seqüências completas (*complete cds*) e incompletas (*parcial cds*) de cDNA. A atualização dos dados foi realizada no dia 19/04/2002 e retornou as 17.071 seqüências utilizadas na segunda etapa do trabalho. Uma busca similar por seqüências realizada no dia 22/01/2003 retornou 17.104 genes, de modo que a segunda parte desse trabalho representa, quase completamente, o *status* atual do transcriptoma de *S. mansoni*.

4.3. TRATAMENTO DAS SEQÜÊNCIAS OBTIDAS DO GENBANK

4.3.1. RETIRADA DE SEQÜÊNCIAS DE VETORES

Ao executarmos o programa CROSS_MATCH contra nosso banco de dados de vetores especialmente construído, obtivemos arquivos de saída contendo as regiões que representavam vetores mascaradas com letras Xs.

Um exemplo de seqüência (GI: 9429129) mascarada pode ser observado a seguir:

```
> gi|9429129 (antes do CROSS_MATCH)
CCGCTGAGTGTGTTGGTTATTTAGGACAGAGATATTGTAATTATTATGTGATAGGGTCTGAGATTAAGAT
TGTCTTAGGTTAGTAAAGTTATTTACATAGCTTAGAAGATTGATTAATAAAATCTTGTTGGTTAAGAGAG
GTCATTTAGTAAAATTAGATATATAATGATGGTGGAATAAGTAGGTTAGATTATTGAGTTGTTCCTTGTT
TTAGGAGTTGCCACCAAAGTGTTTATTAAAAACATTGCTATTAGTAATGTGTTAATAGTAGGCTCTGCGG
AGAATTCGATATCACGCTTATCGATACCGTCGACCTCGGGGGGGGCGCCGATACCA


> gi|9429129 (depois do CROSS_MATCH)
CCGCTGAGTGTGTTGGTTATTTAGGACAGAGATATTGTAATTATTATGTGATAGGGTCTGAGATTAAGAT
TGTCTTAGGTTAGTAAAGTTATTTACATAGCTTAGAAGATTGATTAATAAAATCTTGTTGGTTAAGAGAG
GTCATTTAGTAAAATTAGATATATAATGATGGTGGAATAAGTAGGTTAGATTATTGAGTTGTTCCTTGTT
TTAGGAGTTGCCACCAAAGTGTTTATTAAAAACATTGCTATTAGTAATGTGTTAATAGTAGGCTCTGCGG
AXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXA
```

4.3.2. RETIRADA DE SEQÜÊNCIAS PEQUENAS E DE BAIXA QUALIDADE

Os arquivos contendo as seqüências cuja porcentagem de letras Ns + Xs foi maior do que 4% e aqueles contendo as seqüências cujo tamanho foi menor do que 150 nucleotídeos foram observados para que fosse confirmada a correta execução do programa. Na tabela abaixo podemos ver o número de seqüências rejeitadas com relação a cada um dos casos.

TAB2. NÚMERO DE SEQÜÊNCIAS REJEITADAS PELO PROGRAMA FILTRO1.PL

	ARQUIVO INICIAL	ARQUIVO NX	ARQUIVO TAM	ARQUIVO FINAL
SM0501	14275	677 (4,74%)	616 (4,31%)	12982 (90,94%)
SM0402	17071	683 (4,00%)	823 (4,82%)	15565 (91,18%)

***Tabela 2**: Número de seqüências rejeitadas pelo programa **filtro1.pl** devido ao pequeno tamanho (arquivo TAM) ou excesso de Ns e Xs (arquivo NX). Vale notar que, se a seqüência possui pequeno tamanho e também excesso de Ns e Xs, ela é contada no arquivo NX.*

Observando a TABELA 2 podemos perceber que cerca de 9% das seqüências foram rejeitadas devido à baixa qualidade (alta porcentagem de Ns e Xs) ou ao pequeno tamanho. Notamos também que as seqüências depositadas entre as duas atualizações contiveram boa qualidade já que apenas 6 delas tinham mais de 4% de N ou X. Ainda com relação às duas atualizações, observamos que mais de 200 seqüências de pequeno tamanho foram depositadas e que, de modo geral, a porcentagem de seqüências aproveitadas ficou mesmo em torno de 91%.

4.3.3. RETIRADA DE OUTROS CONTAMINANTES QUE PODERIAM ATRAPALHAR NO AGRUPAMENTO DAS SEQÜÊNCIAS

O programa **filtro2.pl** foi então executado sobre o arquivo de saída do programa filtro1.pl, contendo as seqüências com as regiões de vetores mascaradas e sem as seqüências pequenas e de baixa qualidade. O programa filtro2.pl é responsável por retirar as seqüências de adaptadores (CACGAG), sítios da enzima de restrição *Xho*I (CTCGAG), seqüências de poliT 5' (TTTTTTTTTT) e seqüências de poliA 3' (AAAAAAAAAA).

Após sua execução, o programa filtro2.pl, apresenta, para fins de confirmação de seu correto funcionamento, um resumo de tudo o que fez. Nesse resumo ele mostra o GI de cada seqüência alterada, mostrando o que foi alterado em cada uma delas (retirada de adaptador, sitio de *Xho*I, poliT ou poliA). Na TABELA 3 é possível observar o número e porcentagem de seqüências que apresentavam cada um dos contaminantes observados (o número total de seqüências aqui é 12.982):

TAB3. NÚMERO DE SEQÜÊNCIAS ALTERADAS PELO PROGRAMA FILTRO2.PL				
	ADAPTADOR	SÍTIO DE *XHO*I 5'OU 3'	POLI T 5'	POLI A 3'
SM0501	2.923 (22,51%)	160 (1,23%)	1.366 (10,52%)	125 (0,96%)
SM0402	2.945 (18,92%)	218 (1,40%)	1.366 (8,78%)	134 (0,86%)

***Tabela 3**: Número de seqüências alteradas pelo programa **filtro2.pl** devido à presença de regiões adaptadoras, sítios da enzima de restrição Xho*I*, seqüências poliT e poliA.*

Analisando esta tabela podemos perceber que parece estar sendo uma preocupação recente dos autores a retirada de seqüências de adaptadores das seqüências. Mas ainda observa-se a presença de sítios de restrição nas seqüências recentemente depositadas.

A grande quantidade de seqüências pequenas e de baixa qualidade (cerca de 9% do total), assim como a grande quantidade de seqüências contendo regiões de adaptadores (cerca de 20% do total) nos mostra, claramente, quão ruins podem ser as seqüências depositadas no GenBank.

Os arquivos resultantes dessa etapa (sem vetor, com seqüências maiores do que 150 nucleotídeos, com menos de 4% de ambigüidade, sem adaptadores, sítios de restrição, poliTs e poliAs) foram utilizado como entrada para a execução dos programas de agrupamento de seqüências.

4.4. AGRUPAMENTO DAS SEQÜÊNCIAS DE *S. mansoni* UTILIZANDO OS SOFTWARES PHRAP E CAP3

Apesar de termos utilizado parâmetros de entrada similares para os dois programas, houve diferença entre o número de seqüências agrupados por eles (TABELA 4). Isso se deve, provavelmente, a diferenças em seus códigos de execução (ver seção 1.7 e 1.8).

TAB4. AGRUPAMENTO DE SEQÜÊNCIAS PELOS PROGRAMAS PHRAP E CAP3

ARQUIVO DE ENTRADA	PROGRAMAS	NÚMERO DE *SINGLETS*	NÚMERO DE *CONTIGS*	NÚMERO DE *UNIQUES*	NÚMERO DE SEQÜÊNCIAS AGRUPADAS
Sm0501	PHRAP	4237	1763	6000	8652 (66,6%)
Sm0501	CAP3	3876	1832	5708	9106 (70,1%)

***Tabela 4**: Resultados do agrupamento de seqüências pelos programas PHRAP e CAP3 para o arquivo Sm0501. Ambos os softwares foram utilizados em parâmetros mais rigorosos de agrupamento (veja seção 3.4).*

Comparando os resultados dos programas poderíamos concluir que o PHRAP é mais rigoroso do que o CAP3 para o agrupamento de seqüências, pois apresentou um menor número de consensos e um maior número de seqüências não agrupadas. É claro que, para um comparação eficiente da execução dos dois programas, outros fatores deveriam ser levados em conta.

Comparamos também nossos resultados com outros dados publicados na internet sobre o agrupamento de seqüências *S. mansoni*. No TIGR, o *release* 3.0 (06/09/2001) do SmGI (*Schistosoma mansoni* Gene Index – http://www.tigr.org/tdb/smgi), apresenta 1.711 *contigs* e 4.985 *singlets*, montados a partir de 13.540 seqüências de entrada. O TIGR utiliza o programa CAP3, com alteração dos parâmetros de entrada (de forma que sejam agrupadas seqüências com similaridade de 95% em pelo menos 40 bases).

Para comparamos esse resultado com o obtido em nossa análise, devemos lembrar que fizemos a filtragem de 1.293 seqüências pequenas ou de baixa qualidade antes de executarmos o programa de agrupamento (ver TABELA 2). É bem provável que as seqüências que foram filtradas por nós estejam representadas como *singlets* nessa análise do TIGR já que, para o agrupamento, a seqüência deve ter ao menos 40 bases sem ambigüidade e bem alinhadas. Assim, se somarmos o número de seqüências filtradas com o número de *singlets* obtidas pelos programas, ficamos com 4.866 seqüências fora de *contigs* no caso do PHRAP e 5.149 no caso do CAP3. Esse é, portanto, o número que deve ser comparado entre as duas análises. Ainda sim parece que o número de singlets obtido por eles é maior do que aquele que obtivemos e, como os parâmetros de

agrupamento utilizados aqui parecem mais rigorosos do que os do TIGR, acreditamos que a diferença esteja mesmo no arquivo de entrada para a análise.

Já na análise de agrupamento apresentada no *site* "*The WHO/UNDP/World Bank Schistosoma Genome Network*" (http://www.nhm.ac.uk/hosted_sites/schisto), o *release* 4, publicado em 03/2000, apresenta 1.830 *clusters* e 5.439 *singlets*, produzidos através da análise de 13.154 seqüências. Esta análise foi realizada com o programa **Sequencer 3.1.1** e foram agrupadas as seqüências com mais de 90% de similaridade em regiões maiores do que 60 bases. O número de *clusters* é bastante similar ao que obtivemos pelo CAP3, entretanto seu número de *singlets* é aproximadamente 10% maior que o nosso (mesmo se considerarmos que as seqüências filtradas devem se somar às *singlets*, como no caso anterior). Isso pode significar que esse programa é mais rigoroso em seu agrupamento do que o PHRAP e o CAP3.

4.5. IDENTIFICAÇÃO DOS UNIQUES MAIS EXPRESSOS

Após o agrupamento, realizamos a montagem de uma tabela apresentando o número de seqüências presentes em cada *unique*, de modo a identificar aqueles que apresentavam um maior número de seqüências e seriam considerados os mais expressos. Tais dados foram obtidos através da execução de programas que obtinham dados a partir dos arquivos de LOG produzidos pelo PHRAP ou CAP3.

Os genes considerados mais expressos foram escolhidos através da análise da TABELA 5. Os *clusters* foram categorizados de acordo com o número de seqüências que continham. A tabela apresentou 51 classes de *clusters* para os dois programas e foram considerados mais expressos aqueles *clusters* que estavam após a classe 25, a metade da distribuição. Por uma coincidência numérica, esses foram também os *clusters* que apresentavam mais de 25 seqüências, para ambos os softwares. Tais *clusters* foram então identificados, comparados entre os dois programas e tiveram sua função biológica discutida.

TAB5. NÚMERO DE SEQÜÊNCIAS AGRUPADAS EM *CLUSTERS*				
NÚMERO DE SEQÜÊNCIAS	NÚMERO DE *UNIQUES* DO PHRAP	CLASSE*	NÚMERO DE SEQÜÊNCIAS	NÚMERO DE *UNIQUES* DO CAP3
1	4237 (singlets)	1	1	3876 (singlets)
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

Tabela 5*: Nesta tabela vemos o todos os clusters produzidos pelo PHRAP e CAP3 divididos de acordo com o número de seqüências que cada um deles contém. * A coluna classe é determinada por um número seqüencial e foi utilizada apenas para definir a metade da tabela. Em negrito estão os genes que foram considerados mais expressos, ou seja, aqueles da metade posterior da tabela.*

4.6. ANOTAÇÃO DOS *CLUSTERS* MAIS EXPRESSOS

Os *clusters* mais expressos foram então anotados de acordo com similaridade no BLAST contra o banco de dados NR e NT. Além disso, observamos de qual biblioteca era proveniente cada uma das seqüências destes *clusters* e montamos as tabelas abaixo:

TAB6. ANOTAÇÃO DOS GENES MAIS EXPRESSOS PELO CAP3

Annotation	No. sequence [a]	E [b]	C [c]	L [d]	M [e]	F [f]	A Wg	? [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 (COII)	158	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

***Tabela 6**: Tabela apresentando a anotação dos clusters mais expressos resultantes do agrupamento pelo CAP3. As seqüências de cada cluster foram divididas de acordo com as bibliotecas de origem. [a] Número total de seqüências do cluster. [b] Número de seqüências encontradas em bibliotecas de ovo. [c] Número de seqüências encontradas em bibliotecas de cercária. [d] Número de seqüências encontradas em bibliotecas de esquistossômulo (ou fase pulmonar, lung stage). [e] Número de seqüências encontradas em bibliotecas de macho adulto. [f] Número de seqüências encontradas em bibliotecas de fêmea adulta. [g] Número de seqüências encontradas em bibliotecas mistas de adultos. [h] Número de seqüências nas quais não foi possível obter informações sobre a biblioteca de origem. Tabela adaptada de Prosdocimi et al. 2002.*

TAB7. ANOTAÇÃO DOS GENES MAIS EXPRESSOS PELO PHRAP

Annotation	No. sequence [a]	E [b]	C [c]	L [d]	M [e]	F [f]	A W [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

***Tabela 7**: Tabela apresentando a anotação dos clusters mais expressos resultantes do agrupamento pelo PHRAP. As seqüências de cada cluster foram divididas de acordo com as bibliotecas de origem. [a] Número total de seqüências do cluster. [b] Número de seqüências encontradas em bibliotecas de ovo. [c] Número de seqüências encontradas em bibliotecas de cercária. [d] Número de seqüências encontradas em bibliotecas de esquistossômulo (ou fase pulmonar, lung stage). [e] Número de seqüências encontradas em bibliotecas de macho adulto. [f] Número de seqüências encontradas em bibliotecas de fêmea adulta. [g] Número de seqüências encontradas em bibliotecas mistas de adultos. [h] Número de seqüências nas quais não foi possível obter informações sobre a biblioteca de origem. Tabela obtida de Prosdocimi et al. 2002.*

Com relação à anotação dos genes mais expressos, observamos que, segundo seria esperado, a maioria deles pertence à categoria dos genes de manutenção da função celular. Para o metabolismo de carboidratos, por exemplo, três *clusters* com grande quantidade de seqüências representando genes da via glicolítica foram encontrados tanto pelo CAP3 quanto pelo PHRAP, são eles: gliceraldeído-3-fosfato-desidrogenase (105 seqüências), frutose 1,6 bisfosfato aldolase (68 seqüências) e enolase (26 seqüências). É interessante observar que o PHRAP e o CAP3 reuniram exatamente as mesmas seqüências de mRNA e ESTs de *S. mansoni* para construir esses *clusters*, o que aumenta a confiabilidade do resultado. Outro ponto interessante relacionado a enzimas da via glicolítica foi a alta quantidade desses genes encontradas em bibliotecas de cercaria. Isso concorda com trabalhos anteriores de nosso grupo sugerindo que nesse estágio de desenvolvimento o parasita necessita de uma grande quantidade de energia para nadar com velocidade e para a contração de seu corpo de modo a penetrar ativamente na pele do hospedeiro (Santos *et al.* 1999).

Um grande número de seqüências representando as três unidades da enzima citocromo oxidase foram encontradas por ambos os softwares. Tal enzima compõe o quarto complexo da cadeia respiratória e é responsável pelo bombeamento de prótons através da membrana mitocondrial interna. O gene que codifica esta enzima é mitocondrial e é muito extenso, por esse último motivo foram encontrados vários *clusters* diferentes representando esse gene, tais *clusters* apresentaram a mesma anotação. Isso sugere que os três *clusters* de citocromo oxidase I (COI) produzidos pelo CAP3 representam simplesmente três partes diferentes do mesmo gene. A alta expressão dos genes de citocromo oxidase suporta a idéia de que o *Schistosoma mansoni*, em todos os estágios analisados aqui, é um organismo que apresenta principalmente um metabolismo aeróbico. Alguns estudos anteriores descrevem o *S. mansoni* como um organismo apresentando metabolismo anaeróbico em estágio adulto (Bueding & Fisher 1982, Thompson *et al.* 1984, Rumjanek 1987, VanOordt *et al.* 1989). Essa idéia foi corroborada pela observação das altas quantidades de lactato produzidas pelos vermes (Rumjanek 1987). É importante notar que a enzima lactato desidrogenase foi identificada como um dos genes mais expressos quando os *clusters* foram construídos utilizando o PHRAP e isso poderia explicar as suposições acima, reforçando a idéia de que o *S. mansoni* pode produzir parte de sua energia através da oxidação anaeróbica de glicose. Entretanto, o número de seqüências representando subunidades de citocromo oxidase, em todas as fases de desenvolvimento analisadas aqui, foi maior do que o número de seqüências de lactato desidrogenase.

Com relação às diferentes fases de desenvolvimento, foi observado um grande número de seqüências de citocromo oxidase principalmente na fase de cercaria, o que está de acordo com estudos anteriores que verificaram uma alta expressão de enzimas do ciclo do ácido carboxílico e da fosforilação oxidativa neste estágio de vida do parasita (Skelly *et al.* 1993). Nosso resultado

também concorda com outros estudos sobre o consumo de oxigênio e a produção de gás carbônico a partir de glicose metabolizada em cercárias (Van Oordt *et al.* 1989, Bruce *et al.* 1971). Com relação à fase de esquistossômulo, nenhuma seqüência de lactato desidrogenase foi encontrada. Entretanto foram encontradas seqüências das subunidades 1 e 2 da citocromo oxidase. Portanto parece que o verme apresenta metabolismo aeróbico também neste estágio de desenvolvimento. Para confirmar esses resultados, entretanto, é necessário obter mais seqüências ou construir mais bibliotecas deste estágio, um vez que o número de ESTs é bastante pequeno.

Com relação à mobilidade celular e ao citoesqueleto, observamos a presença de três *clusters* representando genes para actina (dois similares à actina 2 e um similar à actina 1), no agrupamento obtido por ambos os programas. Estudos anteriores mostraram que a abundância de ambas as actinas é maior em fêmeas do que em machos, estando menos representada nos estágios de ovo e cercária (Davis *et al.* 1985). Considerando que o número total de seqüências de fêmeas em nossa análise é o dobro do número de seqüências de machos, parece que a actina 2 é mais expressa em machos, enquanto a actina 1, em fêmeas. Entretanto o número de seqüências utilizado aqui ainda é pequeno e mais seqüências devem ser produzidas para corroborar esta hipótese.

Nesse ponto é interessante observar que o terceiro cluster com maior número de seqüências produzido pelo PHRAP, representando um proteína relacionada à actina (*actin related protein* – ARP1), não foi encontrada entre as seqüências mais expressas pelo CAP3. Isso pode ser explicado pelo fato de que as seqüências representando ARP1 foram agrupadas pelo CAP3 junto com outras seqüências apresentando uma região de minissatélite, o que levou à formação de um cluster com um número excessivo de seqüências (o primeiro cluster da TABELA 6). Seqüências apresentando regiões de minissatélite também produziram um erro de agrupamento representando o segundo cluster mais expresso pelo PHRAP.

Apesar desse problema, o gene ARP1 foi identificado como um dos genes mais expressos segundo o PHRAP. Essa proteína consiste da maior subunidade da dinactina, um complexo multiprotéico que atua, *in vitro*, como cofator para o movimento de vesículas mediado por dineína ao longo dos microtúbulos (Frankel & Mooseker 1996). A presença tanto da alfa como da beta tubulina e de cadeias leves de dineína entre os genes mais expressos sugere uma grande importância do transporte vesicular em *S. mansoni*. Outras proteínas motoras foram identificadas entre os genes mais expressos: uma cadeia pesada da miosina e proteínas reguladoras da miosina. Essas proteínas motoras são responsáveis por um grande número de movimentos intracelulares essenciais para a reprodução e a sobrevivência de células eucarióticas (Valle & Gee 1998).

Há mais de 40 anos atrás, um experimento clássico mostrou que a principal fonte de aminoácidos para o verme adulto de *S. mansoni* é a digestão de hemoglobina (Timms & Bueding 1959). Apesar da via bioquímica para a degradação da hemoglobina pelo verme ainda não ter sido descoberta, um grande número de proteases de *S. mansoni* foram caracterizadas como hemoglobinases (Brindley *et al.* 1997). Entre as cisteína proteinases, as proteínas Sm31 (similar a catepsina B de mamíferos) e a Sm32 (similar a asparaginil endopeptidases) (Klinker *et al.* 1989) são as mais bem caracterizadas e representam uma classe de proteínas altamente imunogênicas, ambas utilizadas no diagnóstico da esquistossomose (Li *et al.* 1996). Dois *clusters* representando a enzima Sm31 e um representando a enzima Sm32 foram encontradas entre os genes mais expressos produzidos pelo PHRAP. Com relação aos mais expresso pelo CAP3, foi encontrado apenas um cluster para cada uma dessas proteínas. Observando os estágios de desenvolvimento onde essas proteínas são expressas encontramos a Sm31 expressa apenas em vermes adultos. Já a Sm32 parece estar sendo expressa também nas fases de cercária e esquistossômulo.

A produção de ovos pelos pares de vermes adultos e seu depósito no fígado e intestino são as principais causas da patogênese na esquistossomose (Chen *et al.* 1992). Duas proteínas de casca de ovo (*eggshell*) foram encontradas entre os genes mais expressos pelo PHRAP e CAP3. Essas proteínas são expressas freqüentemente por fêmeas maduras durante a produção de ovos. Entretanto encontramos também essas proteínas em bibliotecas de machos e cercária, além da biblioteca de ovos. Análises dos alinhamentos utilizados pelo CAP3 para construir esse cluster mostrou que as seqüências de ovo e cercária foram agrupadas neste cluster devido à presença de um microssatélite CT presente em algumas das seqüências. As seqüências de machos, entretanto, parecem ter sido agrupadas corretamente neste cluster. Um fato interessante que observamos foi que todas as seqüências de macho desse cluster eram provenientes da mesma biblioteca. É possível que essa biblioteca tenha sido contaminada com algumas fêmeas, já que a separação dos casais de vermes para a montagem das bibliotecas é um trabalho bastante árduo.

Uma molécula de chaperone foi encontrada entre os genes mais expressos do verme. A seqüência do consenso era similar a moléculas de *Heat Shock Protein* 90 (HSP 90) de diversos organismos. A HSP90 é um chaperone molecular que parece operar na maquinaria citoplasmática multichaperônica, que inclui a HSP70, peptidil-prolil isomerases e outras moléculas co-chaperones como proteína quinases (Richter & Buchner 2001, Young *et al.* 2001). O agrupamento das mesmas seqüências para a formação do cluster de HSP90 pelos dois programas aumenta a credibilidade deste agrupamento.

Dois *clusters* representando proteínas de ligação a cálcio foram encontrados entre os genes mais expressos gerados pelos dois programas. Um desses *clusters* está relacionado a uma família de proteínas conhecidas como fibrilinas. A fibrilina humana é uma proteína de ligação a cálcio que é o

maior componente estrutural das microfibrilas localizadas na matriz extracelular dos tecidos conectivos (Hanford 2000). Essa proteína parece ser altamente expressa nos ovos de *S. mansoni, mas* sua função nesse organismo ainda é desconhecida. A outra proteína de ligação a cálcio é encontrada praticamente apenas no estágio de cercária. Esse gene parece ser expresso durante um período curto de tempo após a liberação da cercária na água e parece ser desligado logo após a transformação da cercária em esquistossômulo (Ram *et al.* 1989). As proteínas de ligação a cálcio são interessantes porque vários eventos metabólicos e fisiológicos são ativados por íons cálcio com a ajuda dessas moléculas (Goodman *et al.* 1979). Devemos lembrar também que essas proteínas são necessárias para os processos de invasão da pele pelas cercarias, uma vez que o íon cálcio tem um papel fundamental em diversos contatos célula-célula, célula-matriz e matriz-matriz (Maurer & Hohenester 1997).

Vários dos consensos representando os genes mais expressos não puderam ser identificados através de busca por homologia. Sete dos mais expressos, segundo o PHRAP, e quatro, segundo o CAP3. Três desses genes desconhecidos e altamente expressos foram agrupados pelos dois programas: dois deles estavam presentes apenas na fase de cercária e uma proteína rica em prolina e histidina foi encontrada apenas em fêmeas e vermes adultos. Dos *clusters* desconhecidos observados apenas pelo PHRAP, dois foram encontrados apenas em machos e cercária, um apenas em adultos (AUT1) e outro em todas as fases de desenvolvimento. O outro *unique* desconhecido produzido pelo agrupamento do CAP3 foi encontrado apenas em ovos.

Outro gene pobremente caracterizado foi o que apresentava similaridade a uma proteína humana, a proteína neuroendócrino específica B. Acredita-se que essas proteínas estejam envolvidas na secreção neuroendócrina ou no tráfego de membranas nas células neuroendócrinas (Hens *et al.* 1998). Essas proteínas são chamadas de reticulons porque elas se associam com o retículo endoplásmico (van de Velde *et al.* 1994). O cluster representando esta proteína é composto de seqüências oriundas apenas de bibliotecas de vermes adultos e fêmeas e foi observado em ambos os programas. A função desta proteína em *S. mansoni* é desconhecida.

O fator de alongamento 1 alfa é conhecido como o segundo transcrito mais abundante nas células eucariotas, constituindo cerca de 1-2% do total de proteínas em uma célula normal (Condeelis 1995). Sua principal função é catalisar a ligação dos aminoacil-tRNAs aos ribossomos, mediada por GTP. Um cluster representando este gene foi encontrado por ambos os programas.

As glutationas-S-transferases são enzimas detoxificantes que catalisam a conjugação dos substratos eletrofílicos com a glutationa (Sheehan *et al.* 2001). Essa proteína é um dos antígenos mais bem caracterizados e o candidato a vacina mais testado contra a esquistossomose (Riveau *et*

al. 1998). Esta enzima foi encontrada em todos os estágios de desenvolvimento de *S. mansoni*, exceto no estágio pulmonar, tanto pelo PHRAP como pelo CAP3.

As proteínas de ligação a Y-box são reguladores multifuncionais da expressão gênica (Matsumoto & Wolffe 1998) e podem ligar-se tanto a DNAs de fita dupla quanto a DNAs de fita-simples e mRNAs. Vinte nove seqüências desse gene foram encontradas em bibliotecas de machos, fêmeas, vermes adultos e cercárias.

4.7. AGRUPAMENTO DAS SEQÜÊNCIAS DE *S. mansoni* (SM0402) UTILIZANDO O CAP3

Após a atualização dos dados de seqüências de mRNA de *S. mansoni*, estas foram novamente agrupadas. Como pareceria redundante realizar a anotação de todos os *clusters* utilizando ambos os programas de agrupamento, realizamos nesta etapa apenas o agrupamento com o software CAP3, que já havia mostrado resultado melhor em análises anteriores em nosso laboratório (Prosdocimi 2002, monografia de bacharelado).

A FIGURA 14 mostra a porcentagem de seqüências agrupadas pelo programa e a porcentagem de seqüências não redundantes. Foram utilizadas neste estudo 15.565 sequencias de entrada (TABELA 2).

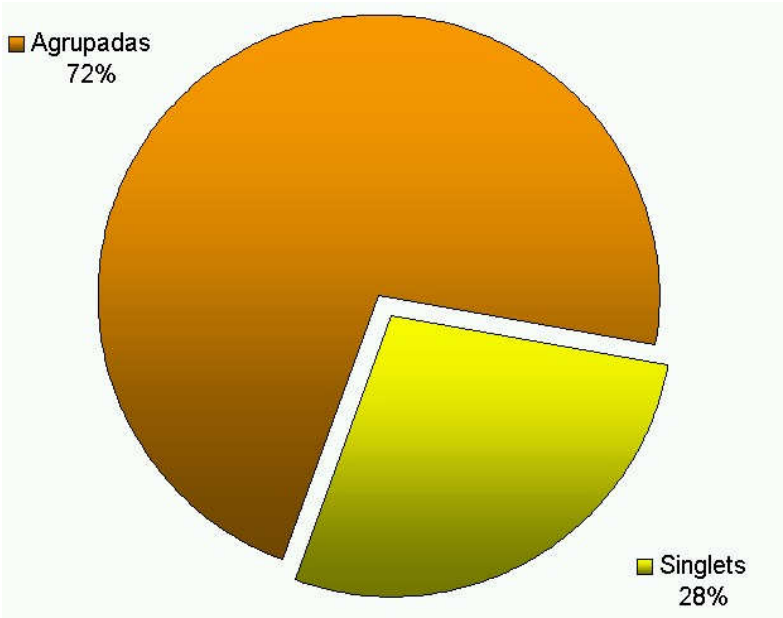

Figura 14. Redundância das seqüências de entrada (Sm0402) após agrupamento pelo CAP3*. Em laranja vemos as seqüências redundantes que foram agrupadas em clusters pelo CAP3. Em amarelo estão as seqüências que não foram agrupadas.*

As 11254 seqüências que foram agrupadas pelo programa CAP3 deram origem a 2017 contigs (FIGURA 15). Nesta figura podemos observar a proporção de contigs com relação a todos os *uniques*. O número de *singlets* nos dois gráficos é o mesmo: 4311.

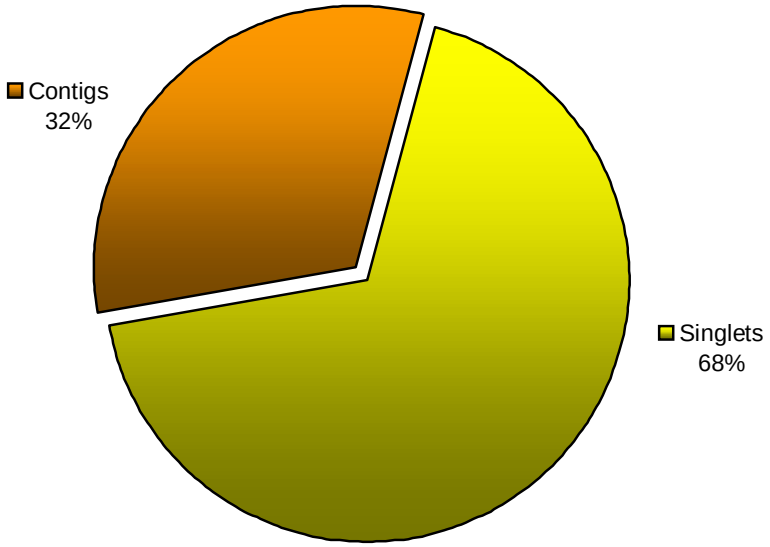

***Figura 15. Classificação dos uniques obtidos após o agrupamento pelo CAP3 (Sm0402)**. Em laranja vemos a porcentagem de contigs entre as seqüências não-redundantes (uniques) e em amarelo estão as seqüências singlets.*

4.8. BUSCAS DE HOMOLOGIAS DOS *UNIQUES* DE *S. mansoni*

É interessante notar na FIGURA 16 que 71% (4470) dos 6328 *uniques* obtidos não apresentaram similaridade significativa com nenhuma seqüência presente nos bancos de dados (de acordo com o *cut-off* utilizado). Dados obtidos em estudos anteriores de bibliotecas isoladas de *S. mansoni* apresentaram resultados semelhantes. Em um estudo de 1997, onde foram analisadas 1401 ESTs de sete bibliotecas diferentes (Franco *et al.* 1997), 71,5% dos 427 consensos obtidos não apresentaram similaridades com proteínas ou genes em bancos de dados. Em outro estudo de 1997, onde foi utilizada uma estratégia de seqüenciamento gênico diferente da estratégia tradicional de produção de ESTs, Dias-Neto e colaboradores realizaram a amplificação de seqüências de mRNA utilizando iniciadores de PCR com seqüências aleatórias a partir de amostras de verme adulto. Eles observaram que 65% das seqüências de *S. mansoni* produzidas não apresentavam homologias com bancos de dados (Dias-Neto *et al.* 1997). Uma menor proporção de genes *no match* foi observada na produção de ESTs em bibliotecas de cercaria, (57,6% das seqüências) (Santos *et al.* 1999). Já na biblioteca de ovo de *S. mansoni*, não foi

possível encontrar similaridade em busca contra bancos de dados de cerca de 78% dos 1104 consensos (Faria 2000).

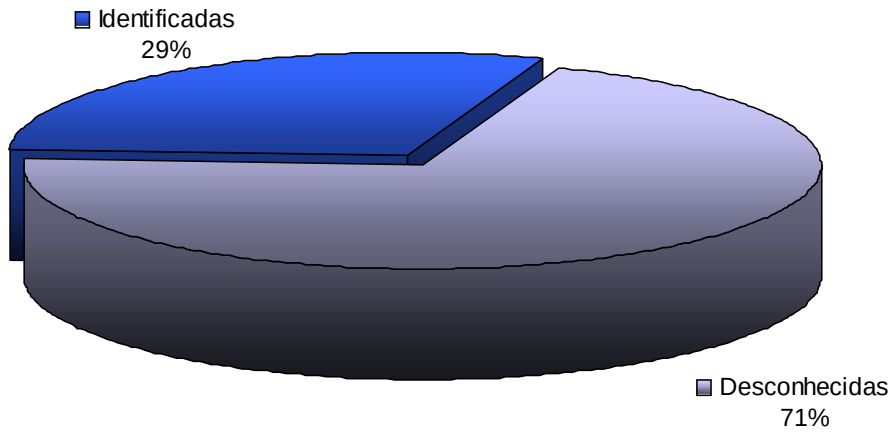

***Figura 16. Resultado da identificação das seqüências unique submetidas ao BLAST**. Em azul vemos a porcentagem dos uniques que apresentaram similaridades em buscas de homologia e em cinza vemos a porcentagem dos uniques que não apresentaram similaridades em buscas BLAST contra o banco de dados NR e nem contra o banco de dados NT. Foi utilizado um cut-off de 10^{-10} para o e-value na realização dessas buscas.*

Deve-se notar que as diferenças e similaridades entre essas análises de genes *no match* podem estar relacionadas também ao valor de *cut-off* do BLAST utilizado em cada uma delas. Entretanto é difícil manter uma comparação precisa entre esses valores de *cut-off*, já que eles dependem do número de seqüências presentes nos bancos de dados na data de execução dos programas. Assim, uma análise BLAST apresentando *cut-off* de *e-value* igual a 10^{-10} hoje é diferente da mesma análise feita há alguns anos ou meses atrás, já que as bases de dados vêm crescendo vertiginosamente.

Comparando ainda a quantidade de genes *no match* com outros genomas eucarióticos já terminados, parece que cerca de 30% dos genes dos organismos entra nessa categoria (Rubin *et al.* 2000), não apresentando homologia com nenhum gene ou proteína dos bancos de dados. Podemos tentar explicar esse número extremamente alto observado em *S. mansoni* através de várias maneiras. Grande parte dos *uniques* (4311 seqüências, as singlets) não representa um gene e sim de pedaços de genes – as ESTs –, assim pode não ter sido possível encontrar similaridades devido ao pequeno tamanho e baixa qualidade dessas seqüências. Além disso, vários contigs podem também ser pequenos suficientes para que não tenha sido possível encontrar similaridades

contra seqüências dos bancos de dados. Deve-se notar também que existem poucos organismos evolutivamente próximos à *S. mansoni* que apresentam grande quantidade de seqüências disponíveis nos bancos de dados. Isso pode ser um fator importante para explicar esse alto número de seqüências *no match*.

Considerando os *uniques* identificados, 1520 apresentaram similaridade com um mesmo gene/proteína em vários organismos, sendo considerados com putativos representantes desse gene em *S. mansoni*. Cento e quarenta e seis apresentaram baixa ou média similaridade com genes/proteínas de *S. mansoni* (sendo identificados como similares aos genes/proteínas deste) e 192 apresentaram alta similaridade com seqüências do verme, sendo identificadas como pertencentes ao próprio. Os resultados obtidos podem ser visualizados na FIGURA 17.

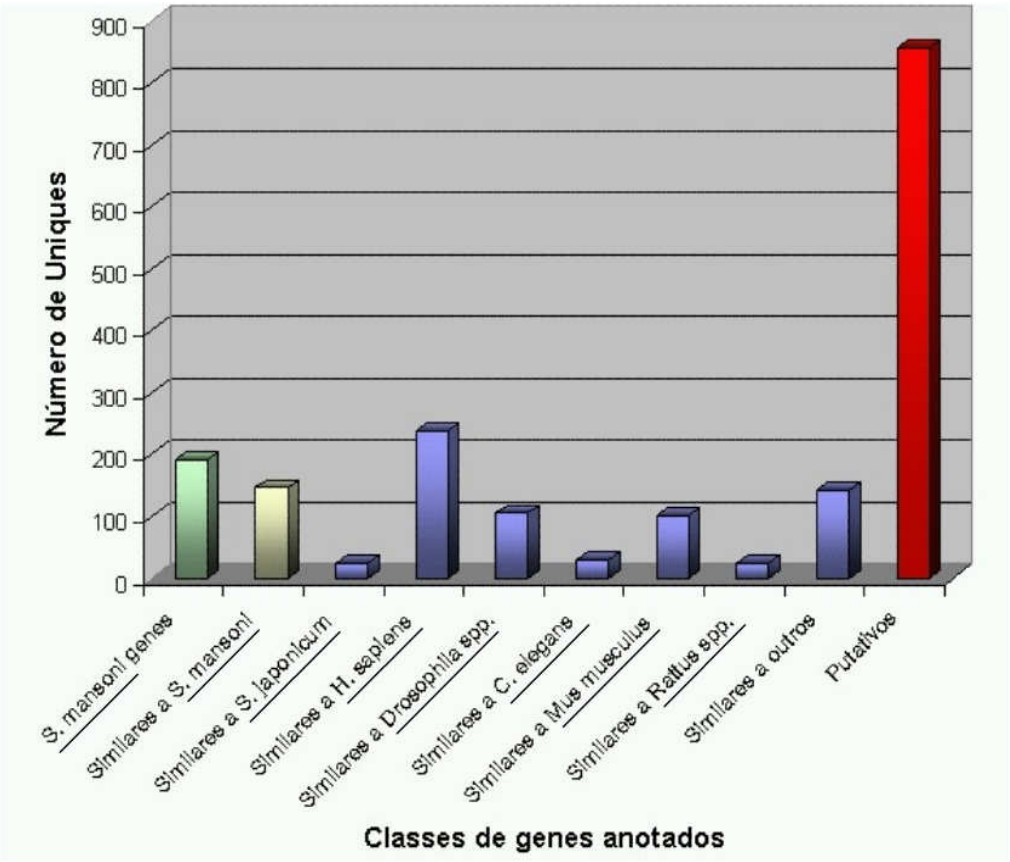

***Figura 17. Resultados das anotações dos uniques identificados**. Em verde vemos a porcentagem dos uniques que representaram genes de S. mansoni, em laranja vemos os uniques que apresentaram similaridades baixas e intermediárias (< 90% ao longo da extensão do unique) a genes do verme e podem ser caracterizados como possíveis parálogos. Em azul estão os genes que apresentaram similaridades a organismos específicos e em vermelho estão os genes que apresentaram similaridades a vários organismos e formaram genes putativos.*

Com os dados mostrados na FIGURA 17, fizemos uma estimativa do número de possíveis genes duplicados e expressos pelo genoma de *S. mansoni*. Segundo nossa análise, contamos 192 genes do verme e mais 146 genes que foram classificados como "Similar to *Schistosoma mansoni*". Isso pode significar que cerca de 76% dos 192 genes observados apresentam cópias (possíveis parálogos) ou isoformas expressas. É claro também que alguns genes podem ter mais de uma isoforma (principalmente se considerarmos que foram seqüenciadas várias cepas diferentes do verme) e, além disso, os erros gerados no seqüenciamento e no agrupamento de ESTs podem ter feito com que tenhamos classificados vários genes de *S. mansoni* como similares. Esse valor de 76% é bastante excessivo se comparado ao número de duplicações gênicas que têm sido observados nos genomas completos de eucariotos, totalizando 49% em *Caenorhabditis elegans* e 41% em *Drosophila melanogaster* (Rubin *et al.* 2000). Acreditamos que uma abordagem gênomica seja mais adequada para descobrir a quantidade de duplicação gênica já que a abordagem transcriptômica está desviada, representando apenas os genes mais expressos.

Os genes identificados passaram então por um processo de classificação funcional, onde foram divididos nas categorias funcionais propostas pelo COG. Assim, a divisão destes *uniques* nas principais categorias dos COGs produziu o resultado mostrado na FIGURA 18.

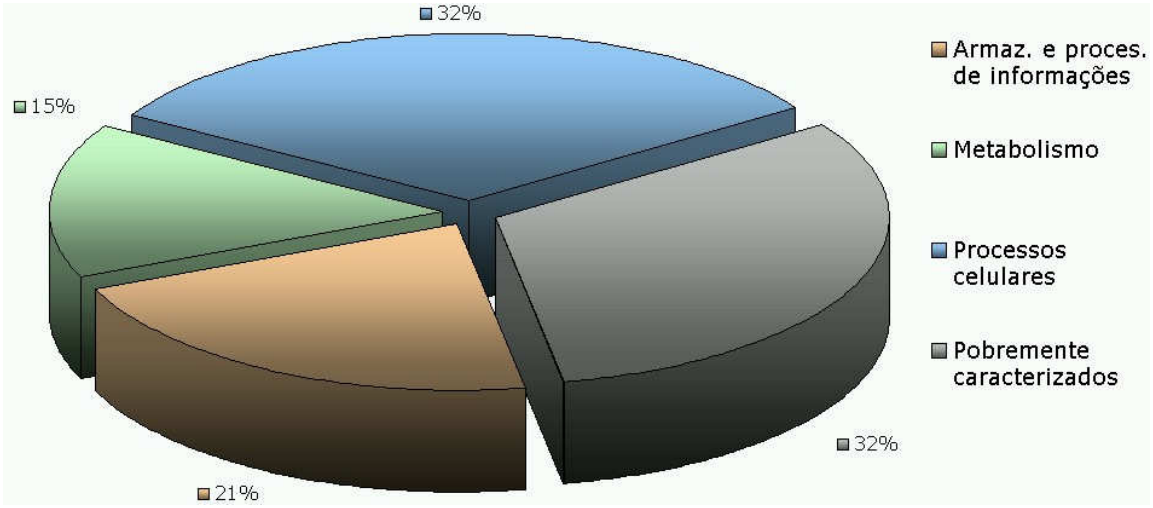

***Figura 18**: **Divisão dos uniques anotados nas principais categorias dos COGs.** As principais categorias dos COGs são: armazenamento e processamento de informações, metabolismo, processos celulares e genes pobremente caracterizados.*

Analisando essa figura, podemos perceber que, apesar de apresentarem similaridades com genes ou proteínas em bancos de dados, boa parte dos *uniques* anotados, 32%, não têm sua função bem caracterizada. Nesse ponto vemos a importância da genômica funcional e das abordagens

bioquímicas tradicionais para identificar a função dos genes nos processos celulares. É bom lembrar que as classificações mostradas aqui são em nível de funções celulares (processamento e armazenamento de informações, metabolismo e processos celulares) e que muitos desses genes pobremente caracterizados com relação a esse tipo de função, podem ter sua função bioquímica e molecular plenamente caracterizada. Nesse ponto lembramos que outras formas de classificação são mais completas e eficientes do que os COGs, como por exemplo, o **GeneOntology**, que é dividido em três níveis de informação (função molecular, processos biológicos e localização celular) (*The Gene Ontology Consortium* 2000).

Para obtermos uma visão mais detalhada e ampla da anotação realizada, dividimos os genes de *S. mansoni* em cada uma das sub-categorias funcionais dos COGs. O resultado desse processo é mostrado na TABELA 8.

TAB8. NÚMERO E PORCENTAGEM DE GENES EM CADA CATEGORIA DOS COGS

CATEGORIAS FUNCIONAIS DOS COGS	Nº Genes	% Genes
Tradução, estrutura ribossomal e biogênese	160	8,42
Transcrição	139	7,32
Replicação, recombinação e reparo do DNA	108	5,68
Divisão celular e particionamento cromossômico	50	2,63
Modificações pós-traducionais, *turnover* de proteínas, chaperones	199	10,47
Biogênese do envelope celular, membrana externa	0	0
Mobilidade celular, secreção, transporte intracelular e citoesqueleto	201	10,58
Metabolismo e transporte de íons inorgânicos	43	2,26
Mecanismos de transdução de sinais	113	5,95
Produção e conversão de energia	67	3,53
Metabolismo e transporte de carboidratos	79	4,16
Metabolismo e transporte de aminoácidos	35	1,84
Metabolismo e transporte de nucleotídeos	32	1,68
Metabolismo de coenzimas	8	0,42
Metabolismo de lipídeos	39	2,05
Biossíntese, catabolismo e transporte de metabólitos secundários	26	1,37
Somente predição da função geral	241	12,68
Função desconhecida	360	18,95
	1900	**100**

***Tabela 8**: Número e porcentagem dos genes anotados em cada uma das sub-categorias funcionais dos COGs. As cores das categorias são correspondentes às mostradas na FIGURA 17.*

Trabalhos anteriores já haviam mostrado classificações de genes de *S. mansoni* de acordo com sua provável função celular. Entretanto muitas das classificações não são sobreponíveis à classificação dos COGs. Franco e colaboradores, por exemplo, classificam suas ESTs em

enzimas, maquinaria traducional/transcricional, membrana/citoplasma, proteínas de transporte e armazenamento, antígenos e outros (Franco *et al.* 1997). Outras classificações já permitem compararmos determinadas categorias. Santos e colaboradores, trabalhando com uma biblioteca de cercaria, identificaram que 13,1% das ESTs informativas desse estágio estão relacionadas ao metabolismo de energia (Santos *et al* 1999). Nossa análise mostra apenas 3,53% dos genes nessa categoria. Isso pode significar que esses genes são mais importantes na fase de cercária, onde o organismo deve gastar uma grande quantidade energia para nadar e para realizar a contração muscular necessária para penetrar a pele do hospedeiro (Santos *et al.* 1999), o que está de acordo com o que mostramos na análise dos resultados das TABELAS 6 e 7.

Podemos ainda comparar os dados obtidos na TABELA 8 com as informações já obtidas sobre genomas completos de outros eucariotos.

TAB9. COMPARAÇÃO ENTRE PORCENTAGEM DE GENES EM CADA CATEGORIA DOS COGS DE DIVERSOS ORGANISMOS EUCARIOTOS

CATEGORIAS	*Schistosoma mansoni*	*Saccharomyces cerevisiae*	*Caenorhabditis elegans*	*Drosophila melanogaster*
Tradução...	8,42	14,12	7,76	9,50
Transcrição	7,32	5,48	4,71	5,88
Replicação...	5,68	8,44	6,53	5,59
Divisão celular...	2,63	1,18	0,77	0,67
Chaperones...	10,47	8,55	9,13	9,61
Envelope celular...	0	1,48	2,18	0,49
... secreção...	10,58	0,51	0,60	1,65
... íons inorgânicos	2,26	4,35	6,99	4,64
Transdução...	5,95	Não disponível	Não disponível	Não disponível
... energia	3,53	5,94	4,85	6,44
... carboidratos	4,16	9,01	10,36	10,56
... aminoácidos	1,84	9,21	6,78	7,81
... nucleotídeos	1,68	4,35	3,09	3,48
... coenzimas	0,42	4,40	2,21	2,43
... lipídeos	2,05	2,66	8,32	5,95
... metabólitos 2º	1,37	Não disponível	Não disponível	Não disponível
... função geral	12,68	17,60	23,63	22,34
... desconhecida	18,95	2,71	2,11	2,96
	100	**100**	**100**	**100**

***Tabela 9**: Porcentagem dos genes anotados em cada uma das categorias funcionais dos COGs de S. mansoni e de outros eucariotos com o genoma já completo (adaptado de Tatusov et al. 2001).*

É interessante notar, na TABELA 9, a semelhança de proporções entre diversas categorias funcionais observadas em *S. mansoni* e em outros eucariotos de genoma completo. Entretanto, devemos lembrar que, enquanto para este parasita foi adotado um estudo de genes expressos baseados no seqüenciamento de ESTs e ORESTEs, para os outros organismos essa tabela foi feita com base em dados do genoma completo. Isso faz com que a abordagem em *S. mansoni*

esteja desviada em relação aos genes de categorias mais expressas. Por isso não achamos interessante discutir cada uma das categorias separadamente. Vale notar, entretanto, a altíssima porcentagem de genes na categoria de "genes desconhecidos" observada em *S. mansoni*, cerca de dez vezes maior do que a porcentagem para os outros organismos. É possível que, pelo fato de estarmos lidando com seqüências parciais de genes de *S. mansoni*, não tenhamos conseguido identificar grande parte dos genes, que acabaram caindo nessa categoria. O fato de existirem poucos organismos taxonomicamente próximos a *S. mansoni* que tenham tido seu genoma completamente seqüenciado deve ser considerada também aqui. Devido a esse fato podemos ter deixado de identificar uma boa parte dos genes, por exemplo, específicos do filo platelminto. Além disso, é de se esperar que a inclusão de novas seqüências de *S. mansoni* nos bancos de dados possa aumentar o tamanho de cada consenso (após a análise de *clustering*) de forma a permitir uma melhor identificação dos genes com base em buscas por homologia, o que diminuiria, talvez, a quantidade de *uniques* na categoria "genes desconhecidos". Uma última hipótese é a de que esses genes desconhecidos apresentem uma alta taxa de expressão e que sua porcentagem não seja tão grande em nível genômico quanto em nível transcriptômico.

4.9. Publicação do WebSite

O website foi publicado em http://www.icb.ufmg.br/~lgb/schisto. Os arquivos intermediários e os programas PERL utilizados na análise estão disponíveis na seção ***download***. Clicando em ***Functional categories*** é possível entrar na página de cada categoria funcional, observando os genes encontrados na categoria e a divisão destes em subcategorias. Em ***Other information*** podem ser encontradas informações sobre a metodologia utilizada no tratamento das seqüências, no agrupamento, na anotação e na classificação funcional. Preenchendo o campo ***Search*** é possível encontrar o gene desejado através de seu número de GI, AN ou através de palavra-chave (presente no nome do gene). A Figura 19 apresenta a página principal de acesso ao site.

4.10. Análises Biológicas dos Genes Presentes nas Categorias Funcionais

Cada *unique* foi identificado através de sua similaridade com genes e proteínas presentes em bancos de dados. Uma vez identificados, pudemos dividí-los em classes funcionais. Agora é hora de tentarmos analisar, dentro do âmbito de cada classe funcional, quais proteínas e famílias protéicas estão presentes no genoma do parasita, assim como sua inter-relação. Com esse objetivo, apresentamos a identificação dos principais *uniques* das principais classes funcionais

analisadas e tentamos correlacionar as funções biológicas e moleculares dos genes e proteínas representados por eles.

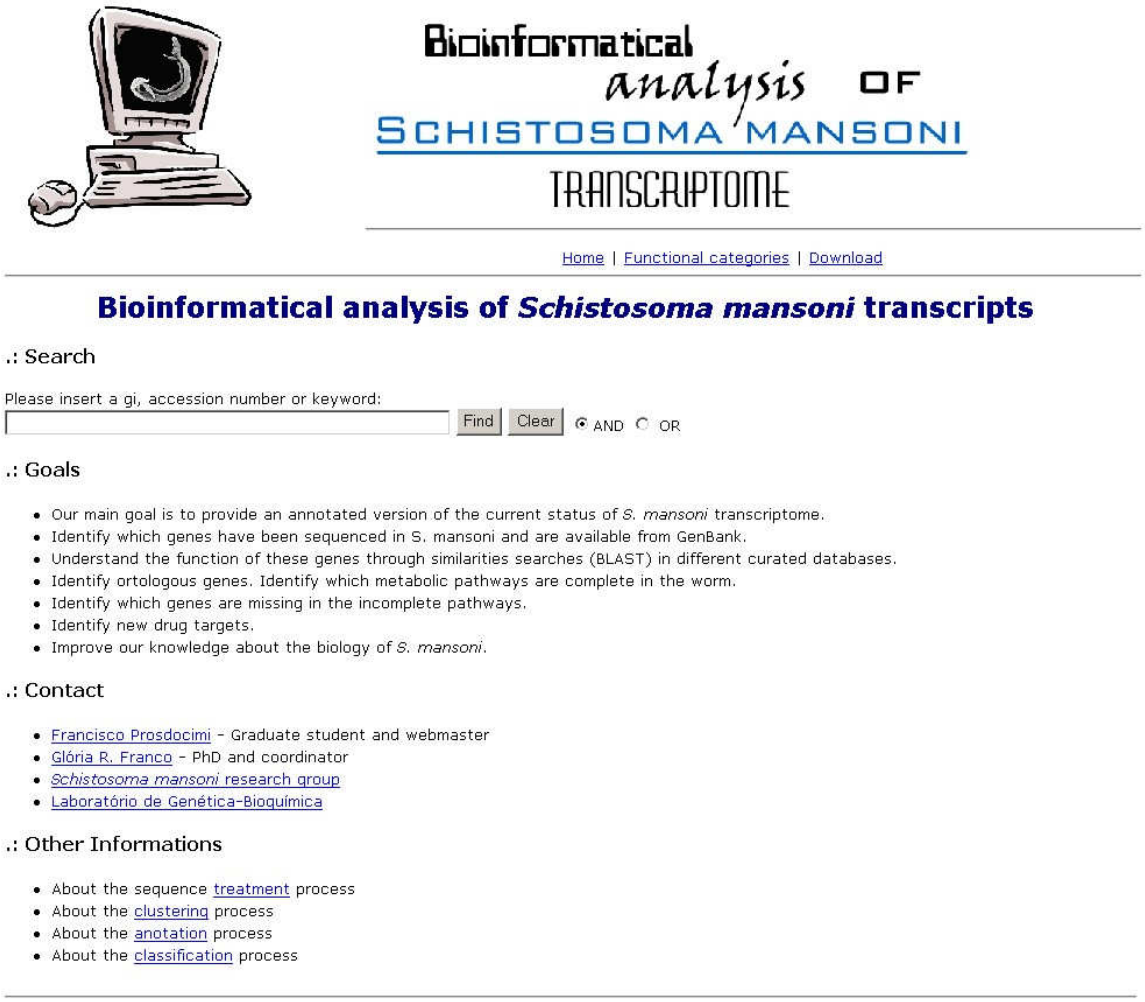

***Figura 19**: **Página inicial do site de publicação das informações**. Mais informações em http://www.icb.ufmg.br/~lgb/schisto.*

4.10.1. TRADUÇÃO, ESTRUTURA RIBOSSOMAL E BIOGÊNESE

Dos vinte genes que codificam aminoacil-tRNA sintetases (necessárias à ligação de cada aminoácido a seu tRNA específico), identificamos quinze em *uniques* de *S. mansoni*. Os outros cinco ainda não encontrados são responsáveis pela ligação dos seguintes aminoácidos ao tRNA: cisteína, fenilalanina, isoleucina, treonina e valina. Desses cinco que faltam ser descobertos, três deles codificam tRNA-aminoacil-sintetases de classe 1 (cisteína, isoleucina e valina) e os dois outros codificam tRNA-aminoacil-sintetases de classe 2 (Francklyn *et al.* 2002). Encontramos quatro *uniques* representando aspartil-tRNA sintetases e dois representando as tRNA aminoacil-

sintetases de histidina, lisina e tirosina. Isso pode representar a existência de diferentes isoformas dessas proteínas ou simplesmente cada um dos *uniques* pode representar regiões 3' e 5' do mesmo gene. É interessante notar que foi encontrado também uma aminoacil-sintetase bifuncional, que pode funcionar tanto como prolil quanto glutamil-tRNA sintetase. Foi encontrada também uma outra tRNA sintetase específica para a ligação do glutamato ao tRNA e talvez essa enzima bifuncional funcione principalmente na síntese do prolil-tRNA, já que não foi encontrada uma proteína específica para a ligação da prolina ao tRNA.

Foram encontrados também cinco *uniques* representando proteínas tRNA metiltransferases, responsáveis por modificações pós-transcricionais em tRNAs.

Com relação às proteínas ribossomais, foram identificados vinte e nove *uniques* representando componentes da subunidade menor do ribossomo e quarenta e quatro da subunidade maior. Estima-se que os ribossomos eucarióticos contenham, em média, oitenta diferentes proteínas (Nelson & Cox 2000, Doudna & Rath 2002), dessa forma podemos dizer que boa parte das proteínas ribossomais de *S. mansoni* já foram descobertas. Foram encontrados também *uniques* representando subunidades de proteínas ribossomais mitocondriais.

Ainda nessa classe observamos que foram encontrados dezesseis *uniques* representando fatores eucarióticos de iniciação da tradução e um representando um fator procariótico. Um *unique* foi identificado como inibidor da iniciação da tradução e outro como uma proteína ligadora de fator de tradução. Foram também encontrados treze *uniques* que representam fatores eucarióticos de alongamento da tradução e dois representando fatores procarióticos, EF-Tu (*Elongation Factor* Tu) e EF-Ts. É interessante notar que um dos fatores de alongamento eucarióticos é específico para a selenocisteína e sugere que *S. mansoni* deve utilizar esse aminoácido em algumas de suas proteínas.

Com relação à terminação da tradução, foi encontrado um *unique* representando o fator eucariótico de liberação da cadeia peptídica 1 (eRF1) e outro representando o fator de liberação da cadeia peptídica 3. Ambas proteínas ficam ligadas diretamente aos ribossomos e acredita-se que interajam entre si para realizar a terminação da tradução em eucariotos (Stansfield *et al.* 1995).

4.10.2. TRANSCRIÇÃO E MODIFICAÇÕES PÓS-TRANSCRICIONAIS

A maior parte dos genes desta categoria representa fatores eucarióticos de transcrição: trinta e seis *uniques*. Além deles foram encontrados também *uniques* representando RNAs polimerases, fatores de splicing, ribonucleoproteínas nucleares, histona desacetilases e DNA e RNA helicases.

Dos fatores de transcrição encontramos cinco *uniques* representando proteínas contendo o motivo de dedos de zinco. Esse motivo típico de proteínas que interagem com o DNA consiste em uma região de cerca de 30 aminoácidos, sendo que quatro deles, normalmente duas cisteínas e duas histidinas, coordenam um único íon de zinco (FIGURA 20, Nelson & Cox 2000, Klug 1999).

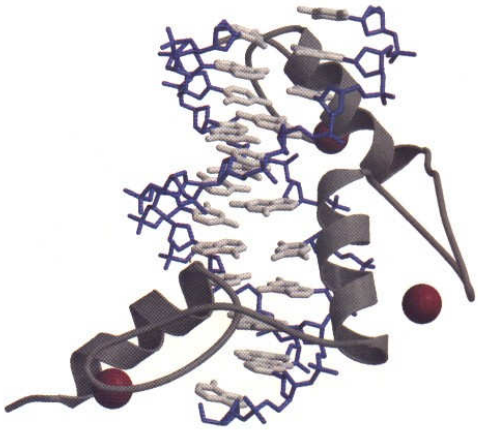

***Figura 20**: **Exemplo de proteína contendo o motivo dedo de zinco**. Na figura vemos os três dedos de zinco da proteína Zif 268 (cinza) complexada com o DNA (branco e azul). Cada dedo de zinco (vermelho) está coordenado a partir de dois resíduos de histidina e dois de cisteína. (Obtido de Nelson & Cox 2000).*

Entre os fatores responsáveis pela iniciação da transcrição em eucariotos pudemos identificar a presença de TFIIB, TFIID e TFIIH. O TFIIB, identificado em um *unique*, é uma proteína monomérica que tem a função de reconhecer a proteína de ligação ao TATA box e recrutar o complexo TFIIH-RNA polimerase para o início da transcrição (Nelson & Cox 2000). O TFIID consiste em um complexo protéico contendo 12 subunidades, das quais a mais importante é exatamente a proteína de ligação ao TATA Box (TBP, do inglês *TATA binding protein*) (Nelson & Cox 2000), que foi identificada aqui em um *unique*. O TFIIH, um complexo que apresenta também 12 subunidades, realiza uma grande diversidade de funções, estando presente inclusive durante o reparo do DNA por excisão de nucleotídeos. No que concerne à transcrição, o TFIIH tem a função de DNA helicase e promove a helicoidização negativa do DNA na região do promotor, além de possuir uma função de quinase, fosforilando a RNA polimerase em sua região C-terminal, de forma

a torná-la ativa e proporcionar o início da transcrição (Nelson & Cox 2000). Cinco *uniques* representam subunidades desse complexo.

Foram encontrados ainda cinco *uniques* representando fatores de alongamento da transcrição, três representando proteínas repressoras de transcrição, três reguladores da transcrição e onze fatores gerais de transcrição, incluindo o fator de ligação ao elemento Y-box, que foi encontrado como um dos genes mais expressos na análise anterior.

É interessante notar também que foram encontrados quatro *uniques* representando diferentes genes homeobox de *S. mansoni*: SMOX-1, SMOX-2, SMOX-3 e SMOX-4. Os genes homeobox são fatores de transcrição envolvidos na diferenciação celular e no desenvolvimento, descobertos originalmente através de estudos de mutações em *Drosophila* que causavam modificação no padrão de segmentação do corpo desses organismos (Ford 1998, Griffiths *et al.* 1998).

A RNA polimerase II é a principal enzima responsável pela transcrição gênica em eucariotos e consiste num complexo protéico formado por 12 subunidades (Nelson & Cox 2000). Seis dessas subunidades estão representas por um *unique* em nossa análise. A RNA polimerase III é responsável pela transcrição de determinados tipos de RNA, principalmente dos tRNAs e 5S rRNAs, dentre outros pequenos RNAs, como aqueles envolvidos no processo de splicing (Paule & White 2000). É interessante notar que alguns promotores para a RNA polimerase III encontram-se dentro dos genes que são transcritos (Nelson & Cox 2000). Identificamos aqui um unique representando uma subunidade desta enzima.

Com relação à acessibilidade dos fatores de transcrição à cromatina, acredita-se que a acetilação de histonas esteja relacionada a um afrouxamento da cromatina, permitindo o acesso de fatores de transcrição e tornando uma região transcricionalmente ativa (Nelson & Cox 2000, Tong 2002). Encontramos dois *uniques* representando a enzima histona desacetilase 3 e dois outros representando a enzima histona desacetilase 8, além de duas outras histonas acetiltransferases.

Encontramos também cinco *uniques* representando RNA helicases com domínio DEAD-box, responsáveis por abrir e desenrolar a dupla fita, permitindo o acesso do complexo de transcrição ao DNA (Nelson & Cox 2000). Alguns trabalhos recentes sugerem também que tais proteínas podem funcionar como chaperones de RNA, agindo de forma a corrigir a estrutura de moléculas de RNA erroneamente enoveladas (Lorsch 2002).

A maioria dos genes eucarióticos apresenta seqüências espaçadoras, os íntrons, que devem ser removidos pela maquinaria de splicing de forma a produzir o RNA maduro, que vai para o citoplasma (Hamm & Lamond 1998, Nelson & Cox 2000, Griffiths *et al.* 1998). O splicing dos RNAs

nas células eucarióticas é realizado principalmente por complexos ribonucleoprotéicos que fazem parte do spliceossomo (FIGURA 21).

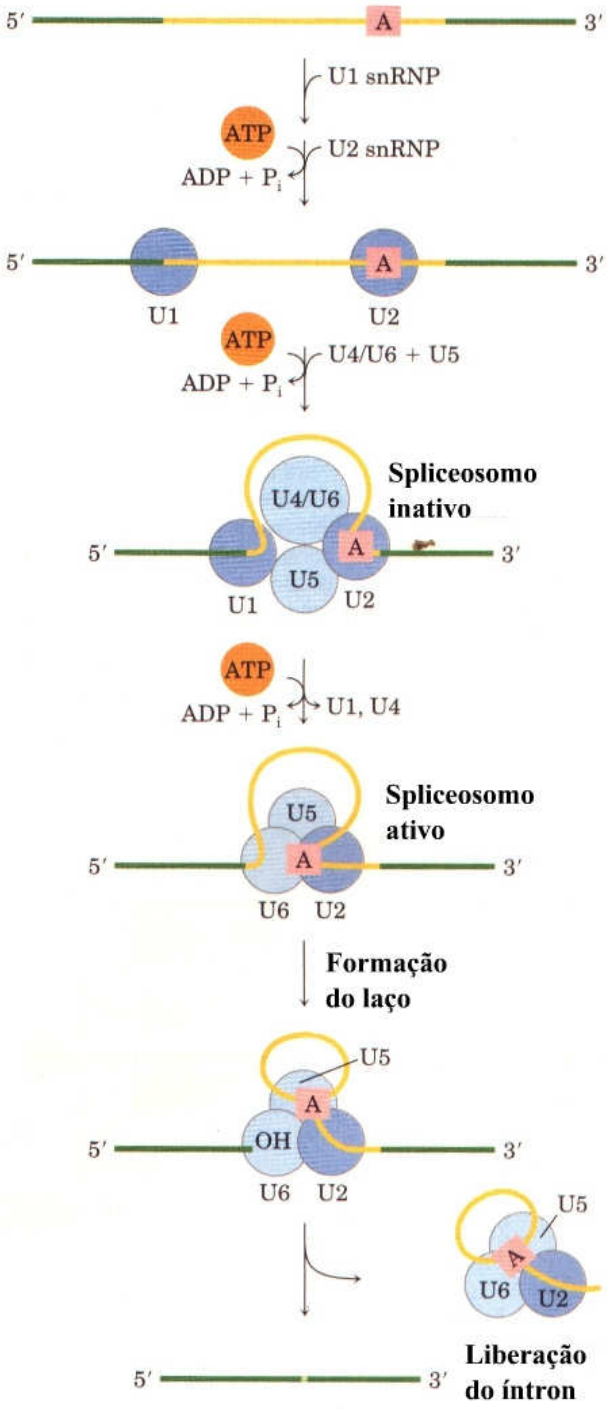

Figura 21: ***Mecanismo de splicing do mRNA***. *Ribonucleoproteínas presentes durante o splicing do mRNA, explicação no texto. (Adaptado de Nelson & Cox 2000).*

Encontramos *uniques* para cada uma das cinco ribonucleoproteínas do spliceossomo: um *unique* representando um homólogo da ribonucleoproteína U1, que se liga à extremidade 5' do íntron; quatro *uniques* representando subunidades da ribonucleoproteína U2, que se liga a uma adenosina da extremidade 3' do íntron, na posição 2' OH; dois *uniques* representando o complexo ribonucleoprotéico U4/U6 e três *uniques* representando subunidades da ribonucleoproteína U5. O complexo U4/U6, juntamente com U5 promove a aproximação das proteínas U1 e U2, cada uma em uma extremidade do íntron. Um rearranjo promovido por ATP acarreta na liberação das subunidades U1 e U4, sendo que U6 passa a se parear com a extremidade 3' do éxon e a extremidade 5' do íntron liga-se à adenosina pareada com U2 formando uma estrutura em forma de laço que é removida para a posterior ligação dos éxons (Nelson & Cox 2000). Foram encontrados ainda seis *uniques* representando fatores de splicing, quatro *uniques* representando proteínas de associação ao spliceossomo e um representando uma proteína reguladora do splicing.

Ainda com relação ao processamento do RNA foi encontrado um *unique* representando uma proteína poli-A polimerase, que catalisa a formação da cauda de poli-A dos mRNAs (Nelson & Cox 2000).

Foram encontradas também várias proteínas de ligação à seqüência de poli-A ou de poli-pirimidinas, além de várias ribonucleoproteínas heterogêneas nucleares. Existe evidência que tais proteínas sejam responsáveis por promover a estabilidade de mRNAs e de realizar um controle pós-transcricional da expressão gênica (Makeyev & Liebhaber 2002).

4.10.3. REPLICAÇÃO, RECOMBINAÇÃO E REPARO DO DNA

Mais da metade dos *uniques* dessa categoria talvez devessem estar em uma outra categoria, a de seqüências repetitivas encontradas em genomas. Sessenta e dois, dentre os cento e oito *uniques* dessa categoria representam transposons, retrotransposons ou transcriptases reversas encontradas no genoma de *S. mansoni*. Apesar dessa grande quantidade de genes relacionados a parasitas do DNA, pudemos encontrar também diversos genes relacionados à replicação do DNA, ao reparo por excisão de base e de nucleotídeos e genes representando proteínas constituintes da cromatina, como as histonas.

Onze *uniques* encontrados nessa categoria apresentaram similaridade a uma seqüência retroviral gag-pol já identificada em um outro trematódeo da subclasse Digenea, o *Clonorchis sinensis* (Bae *et al.* 2001). É interessante notar que este organismo também é um parasita humano, presente principalmente no leste da Ásia, que vive nos dutos biliares do fígado e tem peixes e lesmas como

hospedeiros intermediários (http://www.dpd.cdc.gov/dpdx/HTML/Clonorchiasis.htm, http://www.stanford.edu/class/humbio103/parasitepages/ParaSites/clonorch/ClonorchiasisWebsite.html). Este organismo é taxonomicamente próximo a *S. mansoni* (*NCBI TaxBrowser Information*) e é bem provável que a inserção desse retrovírus tenha acontecido no ancestral comum a esses dois organismos, antes da separação entre as ordens Strigeidida (*S. mansoni*) e Opisthorchiida (*C. sinensis*).

Três *uniques* apresentaram similaridade a um retrotransposon de *S. mansoni*, o SR1, similar a elementos encontrados em galinhas e em outros vertebrados. Drew e Brindley (1997) sugerem que este retrotransposon tenha aparecido no genoma do verme como resultado de transferência horizontal a partir dos hospedeiros vertebrados.

Quinze dos *uniques* dessa categoria mostraram similaridade a uma transcriptase reversa construída sinteticamente a partir de seqüências genômicas de *S. mansoni*. Essa seqüência representa uma parte de outro retrotransposon não-LTR da família SR2 e a baixa similaridade deste com relação a outros elementos repetitivos de vertebrados leva à sugestão de que este não teria surgido através de transferência lateral (Drew *et al.* 1999). Seis outros *uniques* apresentaram similaridade a retrotransposons da família SR2.

Além destes, mais cerca de 25 *uniques* foram identificados com similaridades a diferentes seqüências de transcriptases reversas ou a seqüências de retrovírus (gag-pol poliproteínas) e uma investigação mais minuciosa dessas seqüências poderia levar à descoberta de novos elementos repetitivos em *S. mansoni*.

Para que seja realizada a replicação do DNA é necessário o acesso de enzimas e fatores de replicação às fitas de DNA e exige-se que essas fitas sejam separadas. Essa função é normalmente exercida por enzimas helicases, que se movem ao longo do DNA e separam as fitas num processo que consome ATP (Nelson & Cox 2000). Encontramos três *uniques* representando helicases de DNA. As helicases, ao abrirem a fita de DNA acabam gerando uma superhelicoidização na dupla fita ao redor das regiões “abertas” e essa superhelicoidização deve ser resolvida pela ação de topoisomerases (Nelson & Cox 2000), identificadas em dois *uniques*. A enzima DNAA, que tem a função de reconhecer as regiões de origem de replicação (Nelson & Cox 2000) foi identificada em um *unique*.

Encontramos ainda dois *uniques* representando a proteína MCM5, uma das responsáveis pela ligação de proteínas com domínios MCM ao local de origem de replicação, formando um complexo pré-replicação (Lei & Tye 2001). Um *unique* foi identificado como o fator de replicação O e três *uniques* representando subunidades do fator de replicação C, que se liga à extremidade primer-

molde e atua em conjunto com o antígeno nuclear de proliferação celular (PCNA, identificado em um *unique*) de forma a recrutar as DNA polimerases para a posição inicial de síntese do DNA (Mossi & Hübscher 1998, Takisawa *et al.* 2000). Outro *unique* representava a enzima DNA primase que monta pequenos primers de RNA contendo extremidades 3' OH livres, a partir das quais a DNA polimerase poderá acrescentar desoxiribonucleotídeos para sintetizar uma nova cadeia (Nelson & Cox 2000). Um *unique* foi identificado como uma subunidade da enzima DNA polimerase, a principal responsável pela síntese do DNA.

É interessante notar a presença de histonas entre os *uniques* identificados. As histonas, apesar de extremamente abundantes nas células, não são normalmente clonadas em bibliotecas de cDNA por não possuírem uma cauda de poli-A e escaparem da coluna de oligo-dT utilizada na montagem das bibliotecas. Entretanto, cerca de 1500 seqüências de cDNA presentes nos bancos de dados são dados de ORESTES, onde acontece a amplificação dos cDNAs a partir de primers arbitrários sem a passagem das seqüências de cDNA por colunas de oligo-dT (Dias-Neto *et al.* 2000). Assim é possível que boa parte das seqüências de histonas tenham vindo de dados oriundos desta técnica e não da clonagem tradicional de seqüências de cDNA. Pudemos identificar cinco *uniques* representando seqüências de histonas (H1, H2, H2 variante, H3 e H4).

Com relação ao reparo do DNA, pudemos identificar dois *uniques* similares a genes da via de reparo por erro de pareamento (*mismatch repair*), PMS1 e MLH1. Tais proteínas interagem entre si e ligam-se ao complexo homólogo ao MutS de *Escherichia coli* que reconhece a lesão por erro de pareamento. O complexo MLH1-PMS1 parece funcionar aumentando a eficiência do reconhecimento de erros pelo complexo homólogo ao MutS (Kolodner & Marsischky 1999).

Três *uniques* da via de reparo por excisão de base foram identificados: um representando uma uracil-glicosilase e dois outros representando endonucleases de sítios apurínicos. A uracil-glicosilase cliva a ligação glicosídica entre a base e o açúcar-fosfato do esqueleto do DNA, eliminando a base uracila que resulta da desaminação de citosina (Memisoglu & Samson 2000, Seeberg *et al.* 1995). É interessante notar que se acredita que o DNA tenha evoluído em conter o nucleotídeo timina ao invés de uracila justamente para que o resultado da desaminação de uma citosina pudesse ser identificado (Nelson & Cox 2000). Tal alteração teria produzido um menor número de mutações em moléculas de DNA contendo timina, que teriam sido selecionadas em detrimento àquelas primordiais contendo bases nitrogenadas de uracila.

No reparo por excisão de base, as DNA-glicosilases, portanto, cortam a base incorretamente pareada e deixam, no lugar, um sítio apurínico ou apirimidínico (FIGURA 22). Nesse momento é necessária a atuação de enzimas que retirem esses nucleotídeos sem base e esse é o papel das enzimas AP endonucleases, que cortam a fita do DNA contendo a lesão (Memisoglu & Samson

2000, Seeberg *et al.* 1995). A DNA polimerase I entra em ação nesse momento e polimeriza a região cortada pela AP endonuclease que continha a lesão. A região recém-sintetizada é unida ao restante da fita por uma DNA ligase, enquanto fosfodiesterases cortam a ligação fosfodiéster do nucleotídeo apurínico ou apirimidínico com a cadeia de DNA (Nelson & Cox 2000).

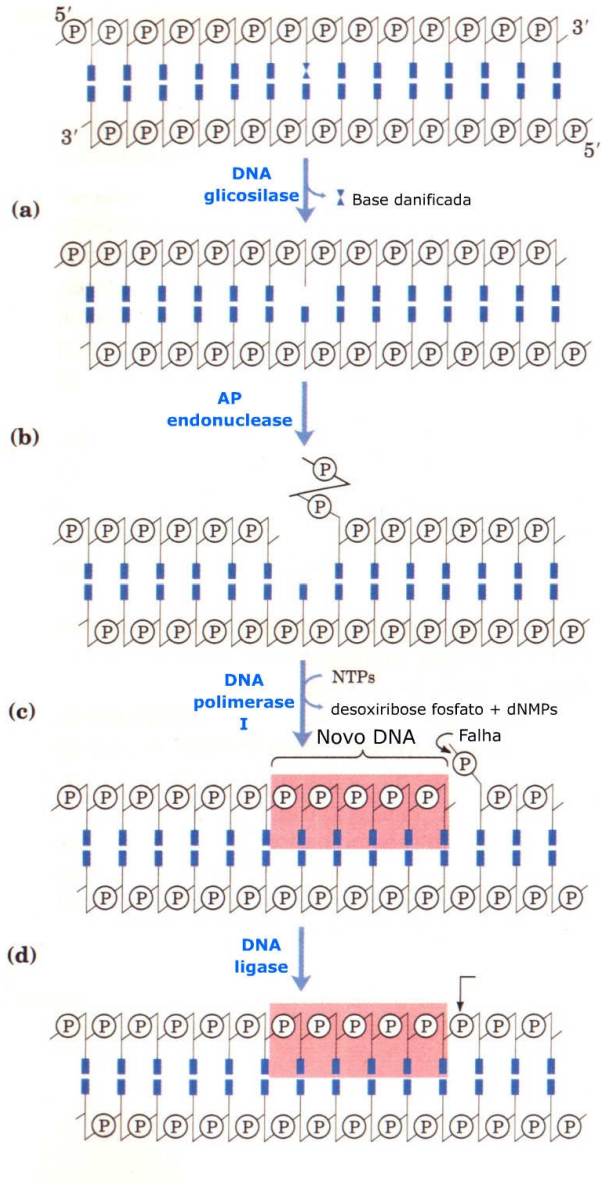

Figura 22: ***Reparo de DNA através da via de reparo por excisão de bases***. *(a) Uma DNA glicosilase reconhece a base incorreta e cliva a ligação entre a base e a desoxirribose. (b) Uma AP endonuclease cliva a ligação fosfodiéster na região do sítio AP. (c) A DNA polimerase I inicia a síntese a partir da extremidade 3'OH livre, removendo uma região da fita danificada e produzindo uma nova cadeia de DNA. A falha após o reparo pela DNA polimerase I é consertada pela DNA ligase, que completa o processo. (Adaptado de Nelson & Cox 2000).*

Vários *uniques* representando genes da via de reparo por excisão de nucleotídeos foram identificados: duas proteínas homólogas a RAD23 (que participa do reconhecimento da lesão segundo Prakash & Prakash 2000), cinco *uniques* representando subunidades de TFIIH (que atua como helicase, necessária para abrir a dupla fita no local da lesão e permitir o acesso de outras proteínas necessárias para o reparo (Lehmann 1995)), um *unique* representando uma proteína de ligação a XPA (que também é necessária para o reconhecimento da lesão (Lehmann 1995)) e outro representando uma proteína do grupo de complementação de XPE.

Encontramos também dois *uniques* responsáveis pelo reparo de quebra de dupla-fita e recombinação, um deles representando o gene MRE11A e outro representando um gene similar à RuvB.

4.10.4. Divisão Celular e Particionamento Cromossômico

Através de nossa análise pudemos identificar vários *uniques* representando genes responsáveis pelo controle do ciclo celular em *S. mansoni*. Um *unique* foi identificado como similar à CDC91 de *S. cerevisiae*, outro foi identificado como à CDC42 que parece ser responsável para a definição da polaridade celular (http://biology.unm.edu/biology/maggieww/Public_Html/spellman-cdc.htm) e mais um como homólogo à CDC48, gene associado ao processo de degradação de proteínas via ubiquitinilação (http://biology.unm.edu/biology/maggieww/Public_Html/spellman-cdc.htm). Encontramos ainda algumas proteínas quinases importantes para a progressão do ciclo celular: um *unique* foi identificado como uma isoforma da CDK10, outro como uma proteína quinase 5 da divisão celular e dois *uniques* representando serina/treonina quinases (IPL1 e PLK). Três *uniques* representavam, ainda, proteínas fosfatases importantes para o ciclo celular. Encontramos um *unique* representando a ciclina C, que parece funcionar como um componente da maquinaria de transcrição associada à RNA polimerase II (Polly *et al.* 2000). Uma proteína associada à ciclina A/CDK2 e que auxilia o alongamento da transcrição pela RNA polimerase II foi identificada em um *unique*.

Um *unique* foi identificado como o regulador da divisão celular p21. A proteína p21 parece ter a função de inibir a atividade de quinases dependentes de ciclinas e pode também se ligar diretamente ao PCNA e inibir a replicação do DNA (http://www.histol.chuvashia.com/tab-en/ccyc-en.htm, Jaime *et al.* 2002)

Mais três *uniques* foram identificados como proteínas da família das septinas, proteínas conservadas que formam filamentos utilizados como sustentação para a formação de diversos

complexos funcionais (Lew 2000). Um papel importante das septinas é direcionar elementos de citoesqueleto para os sítios de clivagem durante a citocinese (Lew 2000, Longtine *et al.* 1996).

Com relação à proteínas de controle da mitose, encontramos dois *uniques* similares à uma proteína promotora de anáfase humana e um homólogo da proteína de controle da mitose dis3. Encontramos ainda dois *uniques* representando proteínas responsáveis pela transição entre a fase G1 e S do ciclo celular.

Quatro *uniques* foram identificados como culinas, proteínas hidrofóbicas (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001373) que medeiam a degradação de proteínas específicas durante transições importantes no ciclo celular (Rubin *et al.* 2000).

Três *uniques* representam receptores do fator de crescimento epidérmico (EGFR), uma proteína envolvida em transdução de sinais que funciona também como um proto-oncogene (Arteaga 2001). Dois *uniques* representam inibidores do gene pró-apoptótico Bax-1 e um *unique* representa o gene pró-apoptótico *Defender against cell death* 1. Um *unique* foi identificado como uma proibitina, proteína citoplasmática que atua em pontos de checagem do ciclo celular, inibindo a síntese de DNA (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR000163).

Um *unique* foi identificado como a proteína *Diaphanous*, responsável pela reorganização dos filamentos de actina necessária para a formação das fibras do fuso durante a mitose e meiose. Algumas evidências mostram que talvez essa proteína tenha um papel ainda mais importante, participando de todos os processo de invaginação de membranas mediados pela actina (Afshar *et al.* 2000).

Além disso, alguns outros *uniques* apresentaram similaridades com proteínas envolvidas em transformações celulares.

4.10.5. Modificações Pós-traducionais, Turnover de proteínas e Chaperones

A degradação de proteínas evita os danos que poderiam ser causados pela produção de proteínas incorretas ou não desejadas e permite a reciclagem dos aminoácidos (Nelson & Cox 2000), além de funcionar como uma das etapas de regulação da expressão gênica.

A degradação de proteínas em organismos eucariotos normalmente acontece através da ligação covalente da proteína ubiqüitina à proteína que deve ser degradada (Figura 23). A ligação da ubiqüitina à proteína que deve ser degradada acontece através de uma via dependente de ATP,

constituída por três enzimas: a proteína de conjugação à ubiqüitina E1, identificada em cinco *uniques*; a proteína de conjugação à ubiqüitina E2, que teve suas subunidades identificadas, neste trabalho, em oito *uniques*; e a proteína de conjugação à ubiqüitina E3, que não foi identificada em nenhum *unique* e que não é sempre necessária para a ubiquitinilação de proteínas (Hochstrasser 2000, Bonifacino & Weissman 1998).

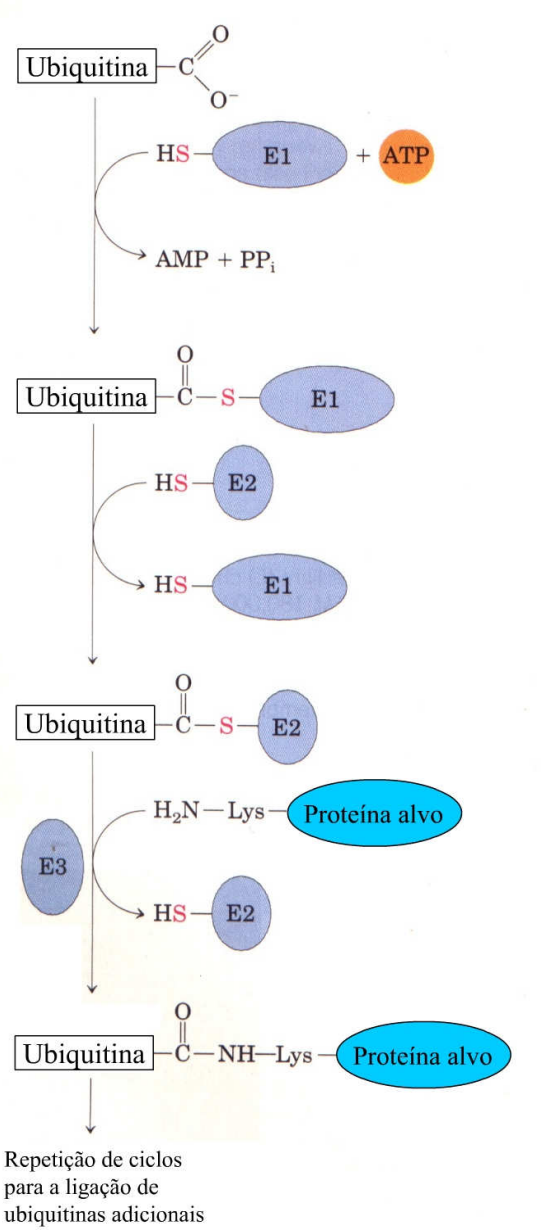

***Figura 23**: **Via de ligação da ubiqüitina a proteínas**. Dois intermediários ubiqüitina estão envolvidos. O grupo carboxil de um resíduo de glicina é ligado através de uma ligação amido a um resíduo de lisina da proteína alvo. Ciclos adicionais levam a poliubiquitinilação das proteínas, que se tornam marcadas para a degradação. (Adaptado de Nelson & Cox 2000).*

Identificamos um *unique* representando a proteína SUMO (*Small Ubiquitin Related Modifier*) da família das proteínas UBL (*ubiquitin-like modifiers*), que se assemelham à ubiqüitina em seu mecanismo de conjugação ao substrato e que possuem uma grande variedade de funções (Hochstrasser 2000). Encontramos ainda sete *uniques* representando proteínas relacionadas ao processo ubiquitinilação de proteínas, inclusive uma proteína de-ubiquitinilante que pode funcionar na reciclagem da ubiqüitina (Hochstrasser 2000).

É normalmente aceito que a ligação de uma simples ubiqüitina não é suficiente para fazer com que uma determinada proteína seja endereçada para o proteasoma, entretanto o tamanho mínimo de ubiqüitinas necessárias para isso ainda não foi corretamente definido (Bonifacino & Weissman 1998). O que se sabe é que a ligação de cadeias de poliubiqüitina a proteínas constitui o principal sinal para a proteólise a partir do proteasoma. O proteasoma 26S é um complexo multimérico contendo cerca de 65 subunidades divididas entre três complexos principais, que apresenta atividades de tripsina, quimiotripsina e glutamil hidrolase (Bonifacino & Weissman 1998). O proteasoma 26S é formado por um núcleo chamado de proteasoma 20S, que consiste num complexo multimérico formado por sete cadeias alfas e sete cadeias betas (Baumeister *et al.* 1998). As cadeias alfas 3, 4, 5, 6 e 7 estão representadas neste trabalho por um *unique* cada uma e as cadeias betas 1, 2 e 3 também estão representadas por um *unique* cada.

O proteasoma possui, além do complexo 20S, dois outros complexos 19S, sendo que cada um deles localiza-se em uma das extremidades do complexo 20S (Baumeister *et al.* 1998). Esse complexo tem a função de reconhecer as proteínas ubiqüitiniladas e convertê-las a uma forma passível de degradação pelo complexo 20S (Baumeister *et al.* 1998) e é conhecido como a subunidade regulatória do proteasoma, contendo cerca de 15 enzimas. Com relação ao proteasoma 19S, as seguintes subunidades foram identificadas em *uniques*: S1 (um *unique*), S2 (um *unique*), S3 (um *unique*), S4 (um *unique*), S5 (um *unique*), S6 (quatro *uniques*), S7 (dois *uniques*), S8 (dois *uniques*), S9 (dois *uniques*) e S10 (cinco *uniques*). As subunidades S11, S12, S13, S14 e S15 não foram identificadas em *uniques*.

As proteínas de choque térmico (HSPs, do inglês *Heat Shock Proteins*) representam componentes ubíquos das células eucarióticas que são responsáveis pela manutenção da correta estrutura das proteínas, funcionando como chaperones moleculares, seja em situações de estresse ou não (Feder & Hoffman 1999). Encontramos cinco *uniques* representando a proteína HSP70 e quatro *uniques* representando proteínas relacionadas a ela. Essa proteína está envolvida em uma grande variedade de funções dentro da célula (Chamberlain & Burgoyne 1997). Quando localizada no citoplasma, parece ter a função de manter outras proteínas em um estado não enovelado, de modo que elas sejam capazes de passar por poros da membrana mitocondrial. Um complexo constituído de HSP60, identificada neste trabalho em dois *uniques*, e HSP10, identificada em um *unique*,

participam do enovelamento da proteína importada dentro da mitocôndria (Ryan *et al.* 1997). As proteínas conhecidas como HSC70 (*Heat-Shock cognate proteins*) são chaperones moleculares que complexam a proteínas não enoveladas quando ligados a ADP e liberam-nas na forma enovelada quando ligados a ATP (Wu *et al.* 2001). É conhecido que a HSC70 deve interagir com outras proteínas para poder exercer suas funções de chaperone e a família de proteínas mais conhecida que interage com HSC70 é a família das DnaJ (Wu *et al.* 2001). Nesse trabalho, cinco *uniques* foram identificados como membros dessa família. Outra família de proteínas na qual já foi evidenciada a ligação com HSC70 é a família das *HSP70-interacting proteins*, ou HIP. Três *uniques* dessa família foram identificados neste trabalho. Um *unique* ainda foi identificado como a proteína HSP82.

Sete *uniques* foram identificados como ciclofilinas (ou peptidil-prolil *cis-trans* isomerases, ou PPIases), que catalisam a isomerização *cis-trans* da ligação entre um aminoácido X e a prolina situada logo a seguir (Ou *et al.* 2001). Essa é uma das etapas mais demoradas do enovelamento de proteínas e, portanto, as PPIases possuem um papel fundamental em acelerar esse processo (Nelson & Cox 2000).

Quatro *uniques* representavam proteínas isomerases de dissulfito (PDI, do inglês *Protein Disulfide Isomerase*). Essas proteínas são tiol dissulfito oxidoredutases extremamente importantes para o correto enovelamento das proteínas, já que quebram ou constroem ligações dissulfeto entre as cisteínas de proteínas incorretamente enoveladas (Noiva 1999, Nelson & Cox 2000).

A proteína do complexo T (TCP, do inglês *T-complex protein*) é formada por um complexo multimérico composto por oito unidades que auxilia no enovelamento de actinas e tubulinas (Liou *et al.* 1998). Encontramos *uniques* representando cada uma das subunidades desse complexo: alfa (dois *uniques*), beta (um *unique*), gama (um *unique*), delta (dois *uniques*), epsilon (três *uniques*), zeta (dois *uniques*), eta (um *unique*), teta (um *unique*).

Quatorze *uniques* foram identificados como proteínas da família das catepsinas, sendo que um *unique* representava a catepsina A, quatro a B, um a C, dois a D e seis a L. As catepsinas são a classe de proteinases mais abundante nas células eucarióticas, contendo proteases de cisteína, serina e ácido aspártico, e estão localizadas normalmente dentro dos lisossomos (Almeida *et al.* 1999).

Quando alcançam o lúmen do reticulo endoplásmico, oligossacarídeos são muitas vezes ligados a resíduos de asparagina de muitas proteínas, no processo conhecido como glicosilação (Nelson & Cox 2000). Vinte e nove *uniques* foram identificados como genes relacionados a este processo, o que reflete sua importância para o metabolismo de *S. mansoni*. Uma enorme diversidade de

oligossacarídeos pode ser ligada a proteínas, mas as vias de glicosilação apresentam um passo inicial comum, onde um esqueleto de oligossacarídeo contendo 14 resíduos é formado e transferido de um doador dolicol-fosfato para o resíduo de asparagina da proteína (Nelson & Cox 2000). Tal reação é catalisada pela enzima oligossacaril transferase, cujas subunidades foram identificadas em seis *uniques*. Depois da transferência, o esqueleto de oligossacarídeos pode ser cortado e modificado a partir de diferentes vias em diferentes proteínas. Várias outras enzimas responsáveis pela modificação e o acréscimo de diferentes oligossacarídeos em proteínas foram encontradas.

Vale ressaltar também o grande número de peptidases que encontramos nessa categoria, representadas por vinte e quatro *uniques*. Tais proteínas realizam a clivagem proteolítica de ligações peptídicas em proteínas e polipeptídeos, modificando e controlando suas atividades (Isaac *et al.* 2000).

Além disso, encontramos também outras proteínas responsáveis por modificações pós-traducionais, como metiltransferases, acetiltransferases, desulfurases, fosfodiesterases e etc.

4.10.6. Mobilidade Celular, Secreção, Transporte Intracelular e Citoesqueleto

Com relação ao transporte intracelular, encontramos oito *uniques* apresentando similaridade a cinco das seis subunidades da partícula reconhecedora do sinal (SRP, do inglês *Signal Recognition Particle*), dois apresentando similaridade com a subunidade alfa de seu receptor e um *unique* representando uma peptidase de sinal. As seqüências sinais estão presentes nas proteínas que devem ser internalizadas no retículo endoplásmico (ER) e são constituídas por 13 a 36 aminoácidos, sendo um ou mais resíduos positivamente carregados na região amino-terminal seguidos por cerca de 13 resíduos hidrofóbicos e contendo mais um resíduo polar perto do sítio onde acontecerá a clivagem (Nelson & Cox 2000). Se uma proteína começa a ser sintetizada e apresenta essa seqüência, a partícula reconhecedora do sinal liga-se ao ribossomo onde ela está sendo sintetizada e interrompe seu alongamento até que o complexo ribossomo-SRP ligue-se a um receptor de SRP na membrana do ER (Nelson & Cox 2000). A síntese da proteína então continua e a proteína será produzida dentro do lúmen do ER, enquanto sua seqüência sinal será clivada por uma peptidase de sinal (Nelson & Cox 2000).

Cinco *uniques* foram identificados como membros da família das importinas alfa e beta, que reconhecem o sinal de localização nuclear e são responsáveis pelo transporte de proteínas para o núcleo da célula (Moroianu *et al.* 1996). No transporte nuclear (Figura 24), um heterodímero de proteínas alfa e beta liga-se à proteína apresentando o sinal de localização nuclear. O complexo

proteína-importina passa pelo poro nuclear através de um processo que requer a presença de um Ran-GTPase. Uma vez dentro do núcleo, as importinas separam-se entre si e da proteína importada e são levadas novamente para o citoplasma de forma a serem reaproveitadas em uma próxima importação (Nelson & Cox 2000). Dois *uniques* foram identificados como D-importinas, também responsáveis por direcionar proteínas para o núcleo.

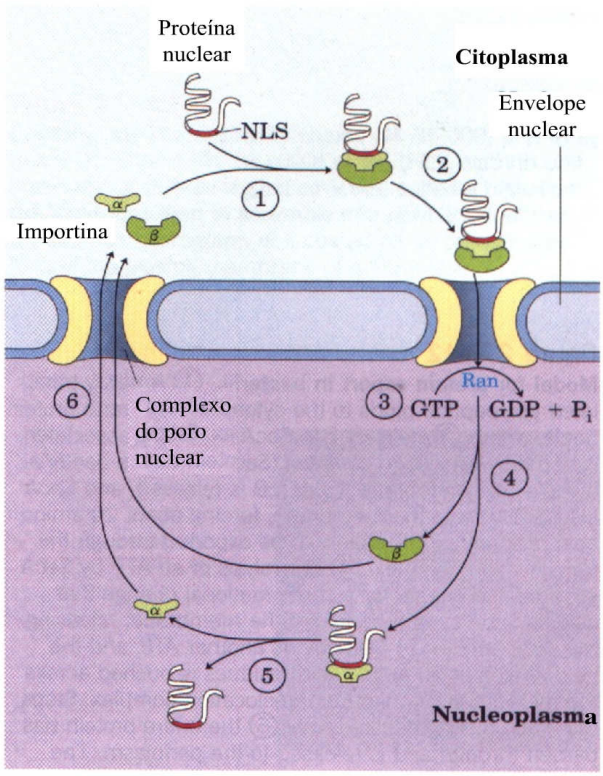

***Figura 24**: **Transporte nuclear mediado por importinas alfa e beta**. (1) Uma proteína apresentando o sinal de localização nuclear é ligada ao complexo importina alfa-beta. (2) O complexo resultante se liga ao poro nuclear e a (3) translocação é mediada por uma Ran GTPase. (4) Dentro do núcleo a importina beta se dissocia da importina alfa e a (5) importina alfa se dissocia da proteína nuclear. (6) As importinas alfa e beta são transportadas para fora do núcleo e recicladas. (Adaptado de Nelson & Cox 2000).*

Seis *uniques* foram identificados como subunidades da proteína coatâmero, responsáveis pelo transporte de vesículas entre o retículo endoplásmico e o complexo de Golgi (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR000804). Dezessete outros *uniques* representavam proteínas com essa mesma função: auxiliar no transporte vesicular ER-Golgi (inclusive quatro representantes da família de proteínas RAB, relacionadas a RAS), sete representavam proteínas importantes para o tráfego intracelular, como nexinas e sintaxinas, dois representam proteínas ER-específicas e três representavam proteína Golgi-específicas. Outros quatro foram identificados como subunidades do

complexo protéico TRAP (do inglês, *Translocon-Associate proteins*) que possuem a função de ligar íons cálcio à membrana do ER e regular a retenção das proteínas residentes no retículo. Dois *uniques* apresentavam similaridade à subunidade alfa e um apresentava similaridade à subunidade beta da proteína SEC61, que parece ter a função especial de importar proteínas de membrana e secretórias para o interior do ER. Seis *uniques* foram identificados representando proteínas de importação mitocondrial, sendo cinco deles subunidades do TOM (transportador através da membrana externa) que transporta as proteínas citoplasmáticas para o espaço entre as duas membranas; e o outro uma subunidade do TIM (transportador através da membrana interna) que importa as proteínas para a matriz mitocondrial (Nelson & Cox 2000).

Algumas proteínas são importadas para dentro das células a partir do meio externo e fazem isso se ligando a receptores presentes em invaginações da membrana chamadas de fossas recobertas (*coated pits*) que concentram receptores para a endocitose (Nelson & Cox 2000). Tais fossas são recobertas, do lado citosólico, por uma rede de proteínas conhecidas como clatrinas. Foram encontrados cinco *uniques* representando subunidades desse complexo protéico.

Três classes principais de proteínas motoras são conhecidas: as miosinas, as dineínas e as quinesinas (Klopfenstein & Vale 2000, Vallee & Gee 1998, Nelson & Cox 2000). A dineína consiste num grande complexo protéico que pode ser dividido em quatro cadeias: a pesada (identificada em dois *uniques*), a intermediária (não identificada), a leve intermediária (identificada em um *unique*) e a leve (identificada em dez *uniques*) (Vallee & Gee 1998). Essas proteínas, compostas de subunidades para a ligação de ATP, atuam como motores para a mobilidade celular de vesículas e organelas ao longo dos microtúbulos (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001372). Foram encontrados também três *uniques* representando subunidades do complexo da dinactina, essencial para o movimento mediado por dineína de organelas e microtúbulos. Este complexo modula a ligação da dineína a organelas celulares e tem um papel importante na formação do fuso durante a mitose. Cinco *uniques* foram identificados como sendo miosinas de cadeia pesada e quatro como sendo miosinas de cadeias leves, uma unidade regulatória e uma proteína de interação à miosina. A molécula de miosina é um complexo multimérico formado por duas moléculas de cadeia pesada e quatro de cadeias leves e é uma proteína contrátil fundamental, encontrada em todos os tipos celulares eucarióticos (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR002928, Nelson & Cox 2000). Foi encontrado também um *unique* representando o gene de paramiosina, principal componente estrutural de muitos filamentos isolados de músculos de invertebrados (http://www.expasy.ch/cgi-bin/niceprot.pl?P06198). Com relação à última proteína motora, a quinesina, encontramos apenas um *unique* representando uma de suas cadeias pesadas. Essa proteína é composta de um complexo oligomérico contendo duas cadeias leves e duas pesadas, sendo que o domínio com a função motora encontra-se nas cadeias pesadas (http://www.ebi.ac.uk/interpro/IEntry?ac=

IPR002151). É interessante notar que o movimento gerado pela quinesina é bastante similar ao produzido pela miosina e até mesmo suas estruturas são parecidas (Cross 2000).

A ligação entre a actina e a miosina deve ser regulada de forma que a contração muscular ocorra apenas quando estiverem presentes os sinais apropriados vindos do sistema nervoso central (Nelson & Cox 2000). Essa regulação é executada por um complexo de duas proteínas: a troponina (identificada em um *unique* – troponina I) e a tropomiosina (identificada em dois *uniques*). A tropomiosina liga-se aos filamentos de actina impedindo sua interação com a miosina (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR000533). Quando chega o impulso nervoso, ocorre a liberação de cálcio. Os íons Ca^{2+} ligam-se à troponina e causam uma alteração conformacional nos complexos troponina-tropomiosina, expondo os sítios de ligação à miosina e promovendo a contração muscular (Nelson & Cox 2000). A troponina é formada por um complexo com três subunidades: uma ligante de cálcio (TnC), uma inibitória (TnI) e uma subunidade de ligação a tropomiosina (TnT). A família das troponinas incluem a troponina T (que se liga a tropomiosina) e a troponina I (que se liga a actina) (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001978). Encontramos ainda três *uniques* representando a proteína calponina, que é capaz de se ligar à actina, troponina, tropomiosina e calmodulina, sendo importante para modulação da contração muscular.

O citoesqueleto é constituído por uma série de filamentos que se estendem ao longo do citoplasma da célula, sendo formados principalmente por filamentos de actina, microtúbulos e filamentos intermediários (Nelson & Cox 2000, Pollard 2001). A tubulina é o principal constituinte dos microtúbulos, que é formado por um dímero de cadeias alfas e betas dessa proteína. Encontramos cinco *uniques* representando subunidades alfas das tubulinas e dois representando subunidades betas. Além disso, encontramos também um *unique* representando uma tubulina gama e dois representando centrinas, ambas proteínas encontrada em centros de organização dos microtúbulos, como os pólos do fuso ou o centrossomo. Quatro outros *uniques* foram identificados como proteínas associadas a microtúbulos.

Nove *uniques* foram identificados como actina, sendo seis actinas 1, duas actinas 2 e uma proteína similar a actina. As actinas são proteínas ubíquas, envolvidas na formação de filamentos que, ao interagirem com as miosinas, produzem um efeito de deslocamento, formando a base para a contração muscular e para muitos aspectos da mobilidade celular (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR004001, Nelson & Cox 2000). Quatro *uniques* representam subunidades do complexo protéico ARP2/3, um representa uma ARP1 (alfa-centractina), outro uma ARP2 e outro uma ARP3. As ARPs são proteínas relacionadas à actina (do inglês, *Actin-related proteins*) cuja função ainda não está bem caracterizada, mas acredita-se que atuem na proteção das extremidades do filamento de actina e que ajudem no processo de polimerização e remodelagem

de tais filamentos dentro das células (Frankel & Mooseker 1996). Um *unique* foi identificado como a proteína N-WASP, de despolimerização da actina, que se liga ao complexo ARP2/3 e atua modificando sua atividade. Outra proteína relacionada à despolimerização da actina, a severina, foi identificada em um *unique*. A alfa-actinina, representada por dois *uniques*, parece também estar envolvida no processo de polimerização dos filamentos de actina, assim como a proteína que se liga a ela, a zixina, encontrada em um *unique*. A proteína F-actina, identificada em dois *uniques* (um para uma cadeia alfa e outro para beta), funciona ligando-se à extremidade dos filamentos de actina de maneira independente de cálcio e promovendo seu crescimento. Sete outros *uniques* representam proteínas relacionadas ou de ligação à actina, um dos quais foi identificado como uma gama filamina, que promove a ligação dos filamentos de actina a glicoproteínas da membrana celular. A proteína anquirina 2 (*ankyrin 2*), encontrada em dois *uniques*, também tem a função de ligar proteínas de membrana ao citoesqueleto.

Quatro *uniques* foram classificados como espectrinas, as principais proteínas constituintes da rede de citoesqueleto da membrana plasmática do eritrócito. Essas proteínas associam-se com a actina para formar uma super-estrutura de citoesqueleto (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001605). Três *uniques* foram identificados como inexinas, componentes estruturais das junções gap (*gap junctions*). Além disso, quatro *uniques* representavam proteínas de ADP ribosilação, envolvidas no tráfego intracelular de proteínas.

4.10.7. Transporte e Metabolismo de Íons Inorgânicos

Nessa categoria identificamos oito *uniques* representando proteínas de canais iônicos sensíveis a voltagem. Destes, cinco representam subunidades de canais de cálcio, dois de canais de potássio e um de canal de cloreto, que medeiam a entrada de cada um desses íons nas células. Tais processos são importantíssimos para uma grande diversidade de funções celulares (Nelson & Cox 2000). Foi identificado também um *unique* representando uma proteína associada a um receptor de GABA(A), que se liga a canais de cloreto mediando a neurotransmissão inibitória.

Dois *uniques* foram identificados como proteínas mitocondriais carreadoras de fosfato. Três *uniques* da família de carreadores de solutos foram também identificados, tais proteínas realizam a troca de íons específicos entre o meio extracelular e o meio intracelular. É interessante notar que foi encontrado também um *unique* representando um transportador de ânions multiespecífico que medeia a excreção de diversos íons orgânicos e é responsável pela resistência a diversas drogas.

Uma das características que define um determinado compartimento celular é a diferença entre o pH de seu lúmen e o pH do citoplasma. Esse pH diferencial é produzido por proteínas conhecidas

como ATP sintases vacuolares ou V-ATPases que são bastante similares às F-ATP sintases que produzem o ATP (Grabe *et al.* 2000). As V-ATPases realizam a acidificação do interior de determinadas organelas ou compartimentos celulares, sendo importantes para suas funções, como é o caso dos lisossomos (http://www.ebi.ac.uk/interpro/Ientry?ac=IPR000245). As V-ATPases são enzimas multiméricas que apresentam um complexo catalítico V1 ligado a um complexo que forma o poro para prótons na membrana. Neste trabalho encontramos sete *uniques* representando as seguintes subunidades do complexo V1: A, B (dois *uniques*), D, E, F e H; e um *unique* representando a subunidade proteolipídica do complexo V0.

Dois *uniques* foram identificados como componentes (cadeias alfa e beta) de uma proteína que realiza a troca (dependente de ATP) de sódio e potássio entre o meio intra e extracelular. Três outros *uniques* apresentaram similaridade a proteínas de *S. mansoni* transportadoras de cálcio para o ER.

Dois *uniques* foram identificados como cadeias pesadas da proteína ferritina, que é a principal proteína sem grupo heme utilizada para o armazenamento de ferro em animais e plantas (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001519) e dois outros apresentaram similaridades com a proteína apoferritina de *S. japonicum*. Ainda com relação ao metabolismo de ferro, um outro *unique* foi identificado como a proteína siderofilina (*siderophilin*), uma proteína transportadora de ferro que pode carregar até duas moléculas do íon quando ligada a uma molécula aniônica, como o bicarbonato.

Cinco *uniques* apresentaram similaridades com proteínas ligantes de cálcio, um *unique* foi identificado como uma proteína que forma um canal específico para água e outro foi identificado como transportador de uma proteína necessária para a utilização de ferro pela mitocôndria. Um *unique* foi identificado como uma proteína do metabolismo de zinco, outro como um transportador de ferro específico de macrófagos e um último como uma proteína similar a outra de levedura envolvida no metabolismo de manganês.

4.10.8. Mecanismos de Transdução de Sinais

A capacidade das células de perceberem e responderem a sinais específicos que alcançam a membrana celular é uma característica fundamental para a vida (Nelson & Cox 2000). Em células animais, os sinais que alcançam uma determinada célula são normalmente divididos em três categorias: os sinais autócrinos, que atuam na mesma célula que o produziu, os sinais parácrinos, que atuam na vizinhança da célula e os sinais endócrinos, exportados até a corrente sanguínea pela célula que o produziu e que alcançam distantes células-alvo (Nelson & Cox 2000). Em todos

esses casos o sinal é detectado por um receptor específico e convertido em uma determinada resposta celular.

O cálcio constitui um mensageiro intracelular ubíquo em células animais e muito do seu efeito é produzido através da ligação com a proteína calmodulina, identificada aqui em dois *uniques*. O complexo cálcio-calmodulina ativa algumas proteínas ligantes de calmodulina e algumas proteínas quinases e fosfatases dependentes de calmodulina, como é o caso das calcineurinas (representadas por dois *uniques*) e das proteínas 14-3-3 (Hardie 1999).

Seis *uniques* foram identificados como proteínas 14-3-3. Essa classe de proteínas homodiméricas é altamente conservada e é encontrada em todas as células eucarióticas (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR000308). Acredita-se que elas tenham um papel importante em uma grande variedade de vias de transdução de sinais, incluindo aquelas que envolvem o ciclo celular e a sobrevivência da célula (Thorson *et al.* 1998). Essas proteínas podem se ligar a seqüências contendo fosfoserinas e, portanto, parece ter papel fundamental nas vias de sinalização mediadas por serina/treonina quinases (Thorson *et al.* 1998). Parecem ainda ativar fortemente a proteína quinase C (PKC), em meio a uma cascata de fosforilações resultantes da entrada de cálcio na célula.

As isozimas da PKC são constituídas de um único polipeptídeo, apresentando um domínio regulatório na região amino-terminal e um domínio de quinase na região carboxi-terminal. Elas são normalmente divididas em três subclasses: as PKCs convencionais, reguladas por diacilglicerol, fosfatidilserina e cálcio; as novas PKCs, reguladas apenas por diacilglicerol e fosfatidilserina; e as PKCs atípicas, cuja forma de regulação ainda não está bem estabelecida (Newton 1997). Três *uniques* foram identificados como proteínas quinases C, sendo dois deles do tipo alfa (convencional) e outro do tipo epsilon (nova proteína quinase). Além disso, foi identificado um *unique* como um substrato para PKC em neurônios e um outro representando uma proteína de ligação à PKC.

Dois *uniques* foram identificados como caseína quinases 2, que fosforilam resíduos de serina ou treonina quando estes estão próximos de cadeias laterais ácidas. Essas proteínas fosforilam uma grande quantidade de proteínas importantes, como a p53 e a DNA topoisomerase II, e uma mutação neste gene é letal para leveduras (Hardie 1999).

Um dos mecanismos mais conhecidos e bem estudados de transdução de sinais é aquele onde está presente a proteína G. Tal mecanismo é definido basicamente por três componentes principais: um receptor de membrana contendo sete segmentos transmembrana, uma enzima na membrana plasmática que gera um segundo mensageiro intracelular e uma proteína de ligação a

GTP que se dissocia do receptor e se liga à enzima, ativando-a (Nelson & Cox 2000). A chegada de um ligante ao receptor de membrana promove uma mudança conformacional no domínio intracelular do receptor, o que afeta sua ligação com uma proteína muito importante na transdução de sinais, a proteína G, formada por um complexo heterotrimérico (Nelson & Cox 2000, FIGURA 25). Assim, um GDP que estava ligado à proteína G é trocado por um GTP, o que ativa esta proteína. Quando ativa, as subunidades beta e gama da proteína G dissociam-se da subunidade alfa, onde está ligado o GTP (Nelson & Cox 2000, Hall *et al.* 1999). Esta subunidade move-se então através da membrana até encontrar a molécula de adenilil ciclase mais próxima. A associação da subunidade alfa da proteína G com a adenilil ciclase estimula esta enzima a sintetizar moléculas de cAMP (AMP cíclico), o que aumenta a concentração dessa molécula no citosol, levando, dentre outras coisas, à ativação da proteína quinase A (Nelson & Cox 2000). O estímulo gerado pela proteína G ligada ao GTP é alto-limitante e logo uma atividade GTPásica da própria subunidade alfa desta proteína provoca sua inativação, convertendo seu GTP novamente a GDP (Nelson & Cox 2000). Inativa, essa proteína dissocia-se da adenilil ciclase e reassocia-se com as subunidades beta e gama ligadas ao receptor, estando a proteína G novamente disponível para interagir com outro receptor de ligação de hormônios (Nelson & Cox 2000). Em nosso trabalho, seis *uniques* foram identificados como subunidades alfas da proteína G, três como subunidades betas e outro *unique* como um receptor de histamina associado a proteína G. Além disso, um *unique* foi identificado como uma subunidade da proteína quinase A.

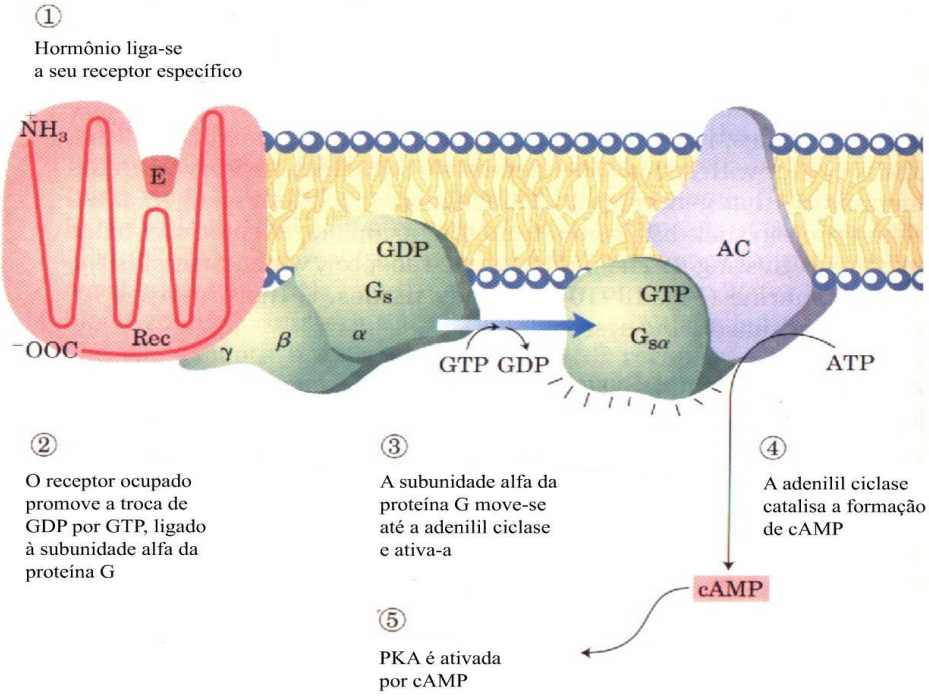

***Figura 25**: **Transdução de sinais através da proteína G**. As cinco etapas que levam à ativação da proteína PKA e à produção do segundo mensageiro cAMP. (Adaptado de Nelson & Cox 2000).*

As proteínas da família RAS compartilham várias características estruturais com a subunidade alfa das proteínas G (Nelson & Cox 2000). Tais proteínas são importantes transdutores de sinais presentes em todas as células e são particularmente mais expressas em células em proliferação (Denhardt 1996). Encontramos um *unique* representando uma proteína homóloga a RAS, outro como uma proteína supressora de RAS e três como proteínas ativadoras da atividade GTPásica de RAS. Dois outros *uniques* foram identificados como a proteína relacionada a RAS, RAP. Essas proteínas, encontradas em grânulos no Golgi e ER funcionam como antagonistas de RAS (Denhardt 1996). A família de proteínas RHO, que compreende pequenas proteínas G que representam papéis importantes na regulação dos filamentos de actina, foram identificadas em dois *uniques* (Denhardt 1996). Identificamos ainda um *unique* representando um receptor acoplado a RHO, outro representando uma proteína ativadora da capacidade GTPásica de RHO e mais um identificado como inibidor da dissociação de GDP a RHO. Um *unique* foi identificado como ativador da atividade GTPásica de RAN, proteínas envolvidas no transporte de RNA e proteínas através da membrana nuclear (Denhardt 1996). Cinco *uniques* foram identificados como a proteína relacionada a RAS, RAB. Membros dessa grande família de proteínas estão envolvidos na regulação do tráfego intracelular entre doadores e aceptores ligados a membranas de compartimentos celulares e no controle da exocitose e endocitose de diferentes tipos de vesículas (Denhardt 1996). Além disso, foram encontrados também dois *uniques* representando as subnidades alfa e beta de um inibidor da dissociação de RAB a GDP, outro representando uma proteína de ligação à RAB5 e mais um representando uma proteína ligante de GTP similar a RAB.

Uma grande quantidade de *uniques* dessa categoria foram identificados como serina-treonina quinases e proteína fosfatases, que representam importantes papéis nas vias transdução de sinais.

4.10.9. TRANSPORTE E METABOLISMO DE CARBOIDRATOS

Os carboidratos são os principais combustíveis químicos utilizados por grande parte dos organismos biológicos e ocupam função central no metabolismo, sendo que sua quebra é de suma importância para a obtenção e o armazenamento de energia (Nelson & Cox 2000).

Na fase preparatória da glicólise, duas moléculas de ATP são investidas e a molécula de hexose é quebrada em duas trioses (Nelson & Cox 2000, FIGURA 26). A primeira enzima a atuar nessa etapa é a hexoquinase, que transforma a glicose em glicose 6-fosfato. Encontramos dois *uniques* representando essa enzima. A próxima etapa é a isomerização da glicose 6-fosfato em frutose 6-fosfato e a enzima que catalisa essa reação é a glicose 6-fosfato isomerase, representada por um *unique*. O passo seguinte, catalisado pela enzima fosfofrutoquinase 1 é a fosforilação da frutose 6-fosfato em frutose 1,6-bisfosfato, reação onde é gasta uma molécula de ATP (Nelson & Cox 2000).

Um *unique* foi identificado como essa enzima. A clivagem da hexose bisfosfato (frutose 1,6-bisfosfato) em duas trioses fosfato (dihidroxiacetona fosfato e gliceraldeído 3-fosfato) é realizada pela enzima frutose 1,6 bisfosfato aldolase, representada por um *unique*. Apenas uma das trioses fosfato pode ser degradada nos passos subseqüentes da glicólise e, portanto, a enzima triose fosfato isomerase catalisa conversão reversiva de dihidroxiacetona fosfato em gliceraldeído 3-fosfato, que será utilizado nas próximas etapas (Nelson & Cox 2000). Essa enzima, representada por um *unique*, já foi identificada em todas as etapas da vida de *S. mansoni* e, sendo uma molécula altamente imunogênica, é uma forte candidata na produção de vacinas contra o parasita (Dunne & Mountford 2001).

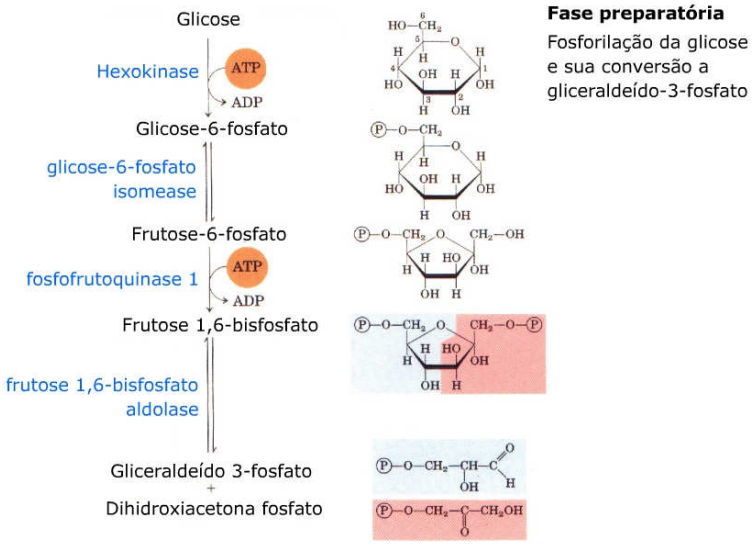

***Figura 26**: **A fase preparatória da glicólise**. Intermediários e enzimas utilizados na etapa preparatória da glicólise. (Adaptado de Nelson & Cox 2000).*

A fase final da glicólise é onde se recupera a molécula de ATP gasta na fase preparatória e produzem-se outras (FIGURA 27). A primeira reação dessa fase é a oxidação do gliceraldeído 3-fosfato em 1,3-bisfosfoglicerato, catalisada pela enzima gliceraldeído 3-fosfato desidrogenase (Nelson & Cox 2000), identificada em três *uniques*. A próxima reação é a fosforilação de uma molécula de ADP em ATP, acoplada à transformação de 1,3-bisfosfoglicerato em 3-fosfoglicerato. Essa reação é catalisada pela enzima fosfoglicerato quinase, representada por um *unique*. Em seguida, a enzima fosfoglicerato mutase, identificada em dois *uniques*, promove a isomerização de 3-fosfoglicerato em 2-fosfoglicerato (Nelson & Cox 2000). Esse produto é então desidratado pela enzima enolase, representada por um *unique*, de forma a produzir o fosfoenolpiruvato. Então, no último passo da glicólise, o fosfoenolpiruvato é transformado em piruvato pela enzima piruvato quinase, que produz também uma molécula de ATP a partir de ADP (Nelson & Cox 2000). A piruvato quinase foi identificada em quatro *uniques*.

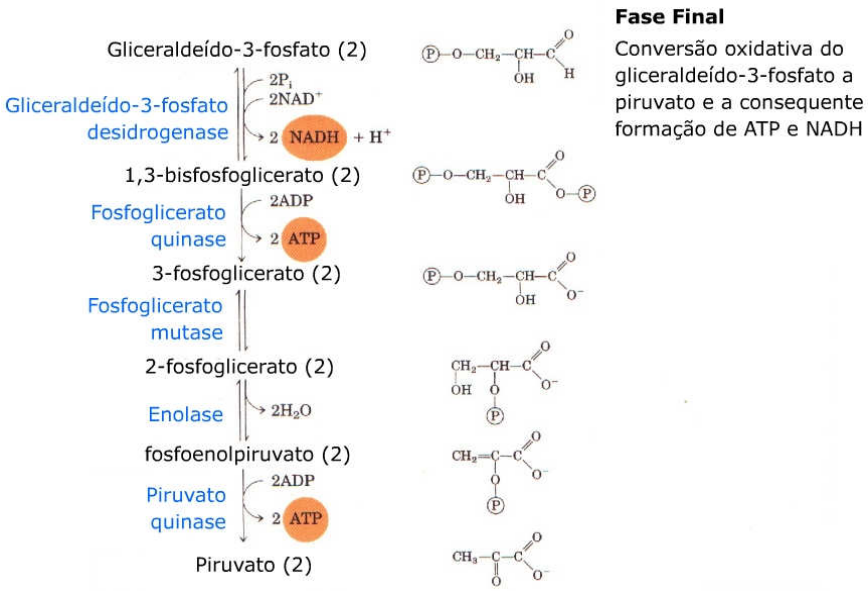

***Figura 27**: **A fase final da glicólise**. Intermediários e enzimas da fase compensatória da glicólise. (Adaptado de Nelson & Cox 2000).*

As enzimas necessárias para a conversão de galactose em glicose para posterior entrada na via glicolítica foram todas identificadas em *uniques*, mostrando que o verme também consegue metabolizar esse açúcar. Assim, identificamos as enzimas: galactoquinase (um *unique*), galactose 1-fosfato uridilil transferase (um *unique*) e a UDP-glicose 4-epimerase (um *unique*). Além disso, encontramos também, em um *unique*, um transportador de UDP-glicose.

Em condições anaeróbias, o piruvato é reduzido a lactato de forma a regenerar o NAD^+ utilizado na oxidação do gliceraldeído 3-fosfato em 1,3-bisfosfoglicerato. Essa reação permite um balanço entre o número de NADH e NAD^+ gastos e utilizados e é catalisada pela enzima lactato desidrogenase (Nelson & Cox 2000), representada aqui por três *uniques*, sendo dois da subunidade A e um da subunidade B.

Na via de síntese de glicogênio encontramos dois *uniques* representando a enzima fosfoglucomutase, que transforma a glicose 6-fosfato em glicose 1-fosfato. A enzima UDP-glucose pirofosforilase gasta uma molécula de ATP para ligar uma molécula de UDP ao grupamento fosfato da glicose 1-fosfato, liberando um pirofosfato (Nelson & Cox 2000). Essa enzima foi identificada em um *unique*. A UDP-glicose então doa seu resíduo de glicose a um terminal não reduzido de molécula de glicogênio, reação catalisada pela enzima glicogênio sintase, identificada em um *unique*. Apesar de que não foi encontrada uma enzima para produzir ramificações na cadeia do glicogênio, encontramos um *unique* representando a enzima que retira essas ramificações, utilizada na via de degradação do glicogênio. Encontramos também 3 *uniques* representando a

enzima glicogênio fosforilase a, que quebra a molécula de glicogênio para produzir glicose 1-fosfato e um *unique* representando a enzima glicogênio fosforilase b, uma forma menos ativa da mesma enzima (Nelson & Cox 2000). A molécula de glicose-1-fosfato deverá ser convertida em glicose 6-fosfato, a ser utilizada em outras vias, pela enzima já comentada, fosfoglucomutase.

O principal destino da glicose é sua quebra em piruvato na via glicolítica e posterior oxidação deste na via do ácido cítrico (Nelson & Cox 2000). Entretanto a glicose pode ter outros destinos na célula, sendo que um deles é a produção de NADPH e ribose 5-fosfato, através da via das pentoses (FIGURA 28).

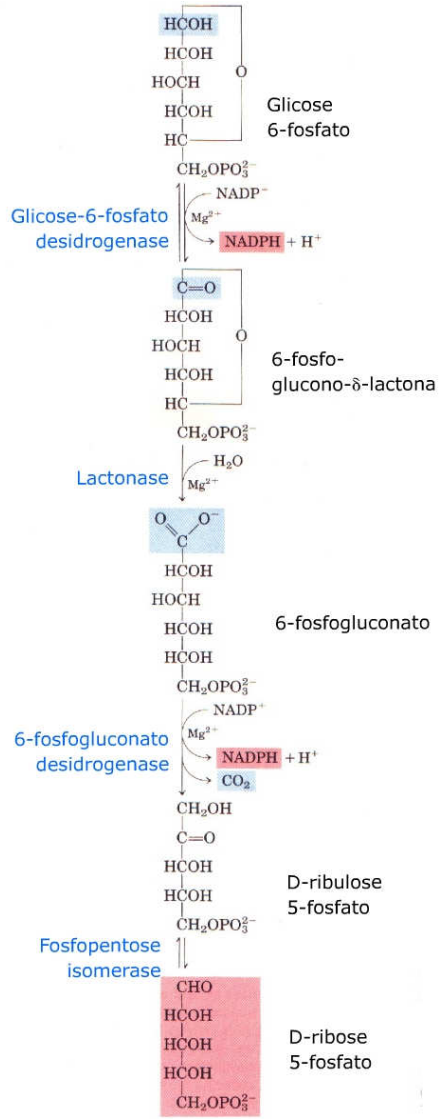

***Figura 28**: **Reações oxidativas da via das pentoses**. Intermediários e enzimas que levam à produção de NADPH e ribose-5-fosfato através da via das pentoses. (Adaptado de Nelson & Cox 2000).*

A primeira reação dessa via é a transformação da glicose 6-fosfato em 6-fosfo-glucono-δ-lactona, acoplada à produção de uma molécula de NADPH a partir de um $NADP^+$, catalisada pela enzima glicose 6-fosfato desidrogenase (Nelson & Cox 2000), representada por dois *uniques*. A próxima etapa é transformação do produto anterior em 6-fosfogluconato, realizada pela enzima identificada em um *unique*, 6-fosfoglucolactonase. A enzima que promove a descarboxilação e oxidação do 6-fosfogluconato para formar a D-ribulose-5-fosfato e NADPH, a 6-fosfogluconato desidrogenase, está representada por dois *uniques*. O composto D-ribulose-5-fosfato deve passar por uma isomerização para gerar o produto final da via, a D-ribose 5-fosfato (Nelson & Cox 2000). A enzima que catalisa esse último passo é a ribose 5-fosfato isomerase, representada por um *unique*. O NADPH produzido aqui será utilizado em vias biossintéticas redutoras e a ribose 5-fosfato será utilizada como precursor para a síntese de nucleotídeos (Nelson & Cox 2000).

É interessante notar que a via das pentoses pode também produzir a glicose 6-fosfato a partir da Ribose 5-fosfato, o que é realizado através de uma via não-oxidativa de reações (Nelson & Cox 2000, FIGURA 29). Analisando essa via, encontramos a enzima ribose 5-fosfato epimerase que transforma a D-ribose 5-fosfato em xilulose 5-fosfato, representada por um *unique*. Dois *uniques* representavam a enzima transcetolase que, utilizando a ribose 5-fosfato e a xilulose 5-fosfato, produz a sedoheptulose 7-fosfato e a gliceraldeído 3-fosfato. A enzima transaldolase, que transforma esses dois últimos produtos em frutose 6-fosfato e eritrose 4-fosfato foi identificada em um *unique*. As outras enzimas que completam essa via são comuns à via glicolítica e já foram comentadas.

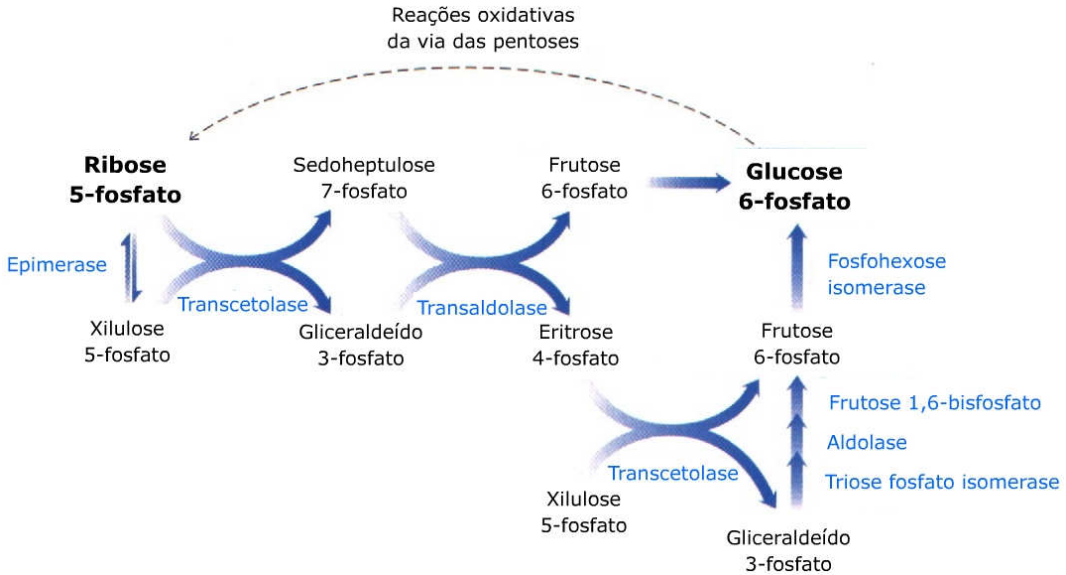

***Figura 29**: **Reações não-oxidativas da via das pentoses**. Intermediários e enzimas que levam à produção de glicose-6-fosfato a partir de ribose-5-fosfato através da via não-oxidativa das pentoses. (Adaptado de Nelson & Cox 2000).*

O piruvato produzido na via glicolítica deve ser transformado em grupos acetil, que entram na via do ácido cítrico para que sejam oxidados a CO_2 (Nelson & Cox 2000). Portanto, a etapa preparatória para a entrada no ciclo do ácido cítrico é a descarboxilação oxidativa do piruvato, que é realizada pelo complexo enzimático piruvato desidrogenase, de forma a produzir uma molécula de acetil-CoA e um NADH (Nelson & Cox 2000). Sete *uniques* representando enzimas desse complexo foram encontrados, sendo dois representando o complexo E1 (piruvato desidrogenase), três representando o complexo E2 (dihidrolipoil acetiltransferase) e os últimos dois representando o complexo E3 (dihidrolipoil desidrogenase).

O ciclo do ácido cítrico constitui uma série de reações importantes para a respiração celular, onde moléculas de acetil-CoA produzidas em vias anteriores são oxidadas (FIGURA 30). A energia liberada por essa oxidação é armazenada em moléculas de NADH e $FADH_2$, coenzimas que serão reduzidas posteriormente para a produção de ATP (Nelson & Cox 2000).

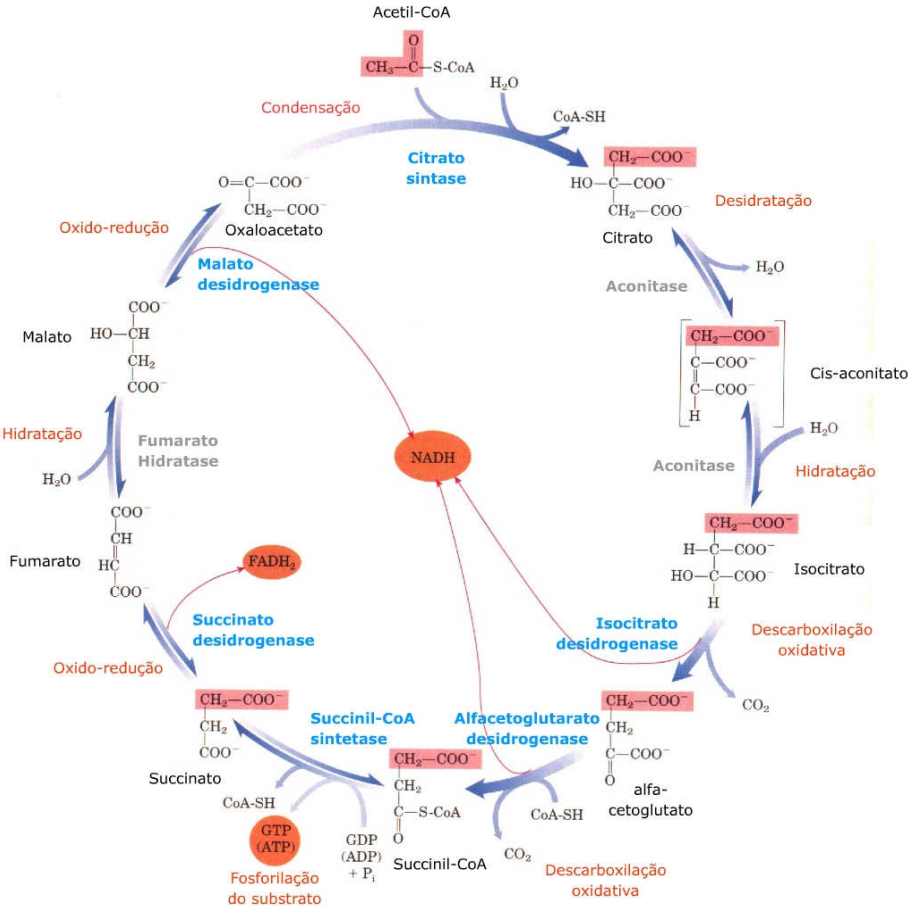

Figura 30: ***Reações do ciclo do ácido cítrico***. *Intermediários, enzimas e reações do ciclo do ácido cítrico. Em vermelho está identificado o que acontece em cada uma das reações do ciclo, em azul temos as enzimas identificadas em uniques e em cinza as enzimas não identificadas em uniques. (Adaptado de Nelson & Cox 2000).*

Na primeira reação deste ciclo, uma molécula de acetil-CoA deve ser condensada a uma molécula de oxaloacetato para produzir o citrato (Nelson & Cox 2000). Essa reação é catalisada pela enzima citrato sintase, identificada em dois *uniques*. Não foi encontrado nenhum *unique* representando a enzima que catalisa a próxima reação do ciclo, a aconitase, que transforma o citrato em isocitrato (produzindo também o intermediário cis-aconitato). Já a enzima isocitrato desidrogenase, que catalisa a descarboxilação oxidativa do isocitrato em alfa-cetoglutarato (ou oxoglutarato), gerando também NADH e CO_2 (Nelson & Cox 2000), está representada por um *unique*. Um *unique* foi encontrado representando o componente E1 do complexo enzimático alfa-cetoglutarato desidrogenase, que catalisa a descarboxilação oxidativa do alfa-cetoglutarato, produzindo succinil-CoA, NADH e CO_2. Identificamos um *unique* representando uma subunidade da enzima succinil-CoA sintetase, que catalisa a formação de succinato e ATP a partir de uma molécula de succinil-CoA (Nelson & Cox 2000). A enzima succinato desidrogenase, responsável pela oxidação do succinato a fumarato, produzindo $FADH_2$, foi identificada em quatro *uniques*. O fumarato deve então ser hidratado para formar o L-malato, reação catalisada pela enzima fumarato hidratase, que não foi identificada em nenhum de nossos *uniques*. A última etapa do ciclo, onde o L-malato deve ser oxidado a oxaloacetato, produzindo NADH, é realizada pela enzima malato desidrogenase (Nelson & Cox 2000), identificada em três *uniques*. Considerando a importância do ciclo do ácido cítrico no metabolismo energético dos eucariotos acreditamos que as enzimas deste ciclo não identificadas em *uniques* estejam presentes no genoma de *S. mansoni* e não tenham sido ainda encontradas devido à quantidade de ESTs seqüenciadas. Com o seqüenciamento de mais ESTs tais enzimas devem ser identificadas.

Quatro *uniques* dessa categoria foram identificados como proteínas transportadoras de glicose e dois *uniques* foram identificados como aldeído desidrogenases, proteínas importantes no segundo passo da utilização de etanol durante a respiração anaeróbica. Encontramos também uma enzima álcool desidrogenase III entre nossos *uniques*, mas como ela é pouco eficiente na produção de etanol para a respiração anaeróbica, tal enzima foi classificada em outra categoria (síntese de metabólitos secundários).

Identificamos ainda dois *uniques* representando a enzima piruvato carboxilase. Essa enzima realiza a carboxilação reversível do piruvato (com CO_2) em oxaloacetato. Tal reação consiste em uma reação anaplerótica, um tipo de reação importante para repor a falta de algum intermediário do ciclo do ácido cítrico (Nelson & Cox 2000).

4.10.10. PRODUÇÃO E CONVERSÃO DE ENERGIA

A energia resultante dos processos de oxidação produzidos durante a quebra de carboidratos foi conservada na forma de moléculas carregadoras reduzidas, como o NADH e $FADH_2$. Durante a fosforilação oxidativa, essas moléculas devem ser oxidadas produzindo (prótons e) elétrons, que serão transferidos para o oxigênio molecular através de uma cadeia de carregadores (Nelson & Cox 2000). No curso dessa transferência de elétrons, uma grande quantidade de energia será liberada e armazenada na forma de moléculas de ATP.

O complexo I da fosforilação oxidativa é composto pela enzima NADH-ubiquinona oxidoredutase que promove a redução de uma molécula de ubiquinona (produzindo ubiquinol) e a expulsão de quatro prótons da matriz mitocondrial para o espaço intermembrana, utilizando uma molécula de NADH (Nelson & Cox 2000). Vinte e três *uniques* representando subunidades dessa enzima (que parece conter quarenta e duas subunidades) foram encontrados em nosso estudo.

O complexo II é formado pela enzima succinato desidrogenase, a única enzima ligada à membrana do ciclo do ácido cítrico e que também é responsável pela redução de moléculas de ubiquinona. As moléculas de ubiquinol reduzidas gerarão um fluxo de prótons em etapas posteriores da fosforilação oxidativa (Nelson & Cox 2000). Quatro *uniques* foram identificados como subunidades dessa enzima.

O complexo III é responsável pela transferência de elétrons entre o ubiquinol, produzido nos complexos I e II, e o citocromo C. Para cada molécula de ubiquinol reduzida, dois citocromos C são oxidados e quatro prótons são bombeados para o espaço intermembrana da mitocôndria (Nelson & Cox 2000). Cinco *uniques* foram identificados como componentes do complexo ubiquinol-citocromo C redutase, que parece apresentar 11 subunidades. Dois *uniques* foi identificado como citocromos C.

A citocromo C oxidase participa da etapa final (complexo IV) da cadeia respiratória, levando os elétrons do citocromo C para o oxigênio molecular, que é reduzido a água (Nelson & Cox 2000). Essa redução promove a produção de um gradiente de prótons através da membrana mitocondrial que, posteriormente, será convertido em ATP. A citocromo oxidase é um complexo formado por treze subunidades. Encontramos aqui sete *uniques* representando a subunidade I dessa enzima e um *unique* representando cada uma das seguintes subunidades: II, III, IV, V e VI.

Depois de todo esse bombeamento de prótons para o espaço intermembrana da mitocôndria, sua matriz passa a apresentar uma menor concentração química de íons H^+, o que gera uma força

próton-motriz (Nelson & Cox 2000). Isso faz com que os prótons do lado de fora tendam a entrar novamente na organela, o que é feito através da enzima ATP sintase. Dessa forma, a enzima consegue produzir ATP às custas da reentrada dos prótons na matriz mitocondrial (Nelson & Cox 2000). A ATP sintase é uma F-ATPase que contém um componente F0 que forma o poro para entrada de prótons, composto por três subunidades (A, B e C); e um componente F1, com cinco subunidades (alfa, beta, gama, delta e epsilon), que é responsável pela produção do ATP. Neste trabalho identificamos um *unique* representando uma subunidade de F0, três *uniques* representando a subunidade alfa de F1, um representando a subunidade beta, dois representando a unidade gama e dois outros representando a subunidade delta.

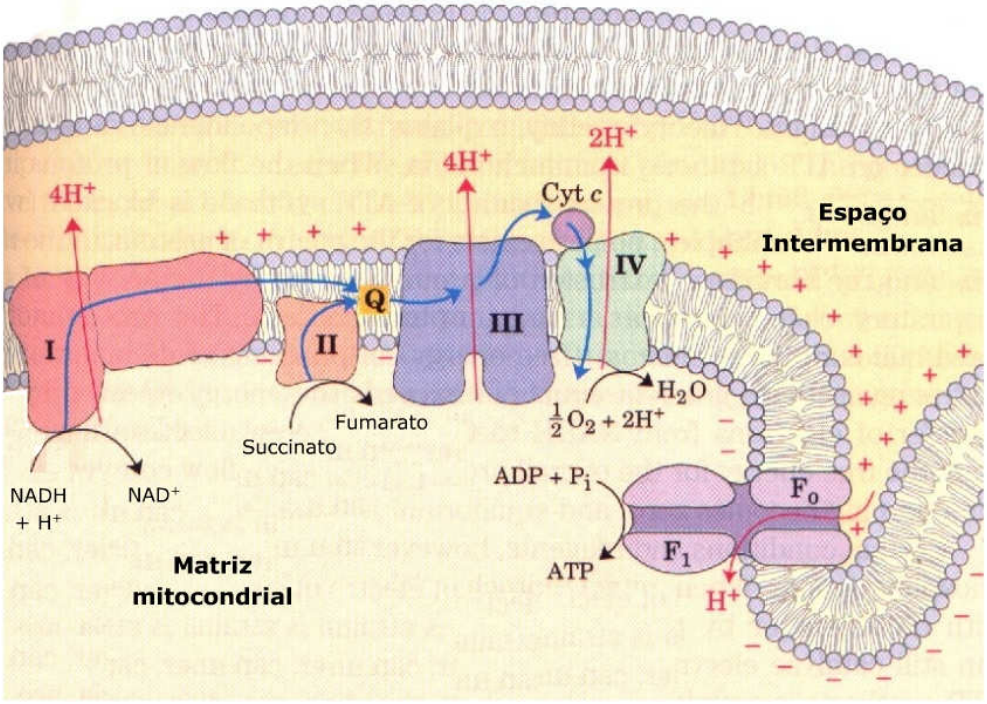

***Figura 31**: **Complexos da fosforilação oxidativa e ATP sintase**. Nessa figura podemos ver os quatro complexos da fosforilação oxidativa, sendo que Q representa a molécula de ubiquinona. A reentrada dos prótons na matriz mitocondrial produz ATP. (Adaptado de Nelson & Cox 2000).*

Ainda nessa categoria identificamos um *unique* como uma proteína que catalisa a troca de ADP por ATP a partir da membrana interna mitocondrial e dois como proteínas *surfeit locus protein*, envolvidas na formação do complexo citocromo oxidase.

4.10.11. TRANSPORTE E METABOLISMO DE AMINOÁCIDOS

O nitrogênio reduzido na forma de íons amônio é assimilado por organismos eucarióticos principalmente através da ingestão de aminoácidos como o glutamato e a glutamina. Tais aminoácidos têm papéis centrais no metabolismo da amônia e de grupos amino, sendo que o glutamato funciona como doador de grupo amino para quase todos os outros aminoácidos (Nelson & Cox 2000). Cinco *uniques* foram identificados como aminotransferases, que realizam a transferência reversível do grupo amino do glutamato para um alfa-cetoácido utilizando o piridoxal fosfato como grupo prostético, sendo dois como aspartato aminotransferases, dois como ornitina aminotransferases e um como alanina aminotransferase.

A enzima glutamato sintase, identificada em dois *uniques*, é importante para a produção de moléculas de glutamato a partir do alfa-cetoglutarato. Essa enzima não está presente na maioria dos animais (inclusive no homem), onde o glutamato é produzido pela enzima L-glutamato desidrogenase (Nelson & Cox 2000). Dada a importância do aminoácido glutamato para a síntese de proteínas e para a produção de outros aminoácidos e, considerando que essa proteína não está presente no homem, talvez ela seja um bom candidato para a produção de drogas antiesquistossomose. Resta saber se o parasita possui alguma via alternativa de produção de glutamato não identificada aqui e se o gene é essencial para sua sobrevivência.

Ainda com relação à síntese e degradação de aminoácidos envolvidos no metabolismo de nitrogênio, um *unique* foi identificado como glutamina sintetase, enzima que catalisa a reação de formação de glutamina a partir de glutamato e ATP (Nelson & Cox 2000). Essa enzima tem um papel fundamental no metabolismo de aminoácidos, pois converte o íon amônio livre, que é tóxico para as células, em glutamina, permitindo seu transporte através do sangue (Nelson & Cox 2000).

Um *unique* foi identificado como a enzima glutamato desidrogenase. Essa enzima é uma das três que levam à produção de prolina a partir do glutamato (Nelson & Cox 2000). A arginina é outro aminoácido produzido através de uma via que parte do glutamato, da qual encontramos apenas uma enzima em nossos *uniques*, a já comentada ornitina aminotransferase. É interessante buscar outros genes dessa via em *S. mansoni*, uma vez que a produção de arginina pelos mamíferos acontece por uma via diferente, passando pelo ciclo da uréia.

A serina, a glicina e a cisteína são aminoácidos derivados do 3-fosfoglicerato, intermediário da via glicolítica (Nelson & Cox 2000). Um *unique* representando uma enzima da via de produção de serina a partir de 3-fosfoglicerato foi identificado, a fosfoserina fosfatase, última enzima da via. Com relação à glicina, foram identificados três *uniques* representando enzimas do sistema

mitocondrial de clivagem deste aminoácido, que o transforma em gás carbônico e metileno-tetrahidrofolato. Com relação à cisteína, foram identificados três *uniques* representando a enzima cisteína desulfurase, utilizada no primeiro passo da degradação deste aminoácido de forma a produzir piruvato (Nelson & Cox 2000). Identificamos também dois *uniques* similares à enzima cisteína dioxigenase de *S. japonicum*.

Quatro *uniques* foram identificados como a proteína adenosilhomocisteinase que participa do ciclo do metil e da produção do aminoácido metionina (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR000043). Essa proteína transfere um grupo metil de uma molécula de S-adenosilmetionina para um aceptor, produzindo uma molécula de S-adenosilcisteína (Nelson & Cox 2000). Essa última molécula deve ser quebrada produzindo uma homocisteína que, através da ação de mais uma enzima, produz metionina.

Um *unique* foi identificado como a proteína fenilalanina hidroxilase, uma oxidase de função mista importante na degradação deste aminoácido. É interessante notar que o defeito nesta enzima é responsável, em humanos, pela fenilcetonúria (Nelson & Cox 2000). Um *unique* foi identificado como a enzima difitina sintase, responsável pela formação de difitina, um aminoácido similar à histidina.

Dois *uniques* foram ainda identificados como a enzima sulfito oxidase, responsável pela última reação na degradação oxidativa dos aminoácidos contendo enxofre. Dois outros *uniques* foram identificados como transportadores de aminoácidos grandes e neutros, enquanto outros dois foram identificados como aminoácido permeases.

4.10.12. Transporte, Biossíntese e Catabolismo de Metabólitos Secundários

A glutationa (GSH) é uma molécula derivada da glicina, glutamato e cisteína que tem a função de manter os grupos sulfidril das proteínas em um estado reduzido e o ferro dos grupos heme no estado ferroso, atuando como um tampão de redox (Nelson & Cox 2000). Essa função redutora da glutationa é utilizada para remover as moléculas tóxicas de peróxido de hidrogênio formadas normalmente durante o metabolismo oxidativo. Dois *uniques* foram identificados como a proteína hidroxiacil glutationa hidrolase (ou glioxalase II) e um como a proteína glutamato-cisteína ligase, ambas utilizadas na via de produção da GSH. Três *uniques* foram identificados como glutationa S-transferases, cuja função é reduzir a GSH de forma a neutralizar hiperóxidos, sendo importantes no sistema de detoxificação de parasitas (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR003080). Essa proteína é também um antígeno bastante imunogênico, sendo considerada a proteína mais bem estudada na busca de vacinas contra a esquistossomose (Dunne & Mountford 2001).

Outra enzima com uma função similar, de proteger a célula contra danos oxidativos, a glutationa peroxidase (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR000889), foi identificada em dois *uniques*. É interessante notar que essa enzima contém um átomo de selênio na forma do raro aminoácido selenocisteína, que é essencial para sua função (Nelson & Cox 2000). A selenocisteína não é adicionada à proteína através de modificações pós-traducionais e sim como um vigésimo primeiro aminoácido codificado pelo códon opal: UGA (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR004536). A presença da proteína "*selenide, water dikinase*", encontrada em dois *uniques*, e necessária para a via de produção da selenocisteína, parece mostrar que o verme é capaz de produzir e utilizar esse raro aminoácido.

Ainda com relação à GSH, foi encontrado um *unique* representando a proteína glutationa redutase, responsável por converter a glutationa oxidada de volta à forma reduzida, capaz de evitar os dados oxidativos gerados pelo peróxido de hidrogênio.

Um *unique* foi identificado como a proteína glicina amidotransferase, que participa da via de produção mitocondrial de creatina, um tampão energético do músculo esquelético (Nelson & Cox 2000). Dois *uniques* foram identificados representando as proteínas responsáveis pelo segundo e terceiro passo da via de síntese de porfirinas, moléculas que apresentam grande importância para proteínas contendo grupo heme, como a hemoglobina e os citocromos (Nelson & Cox 2000). Ainda nessa categoria foram identificados *uniques* de vias responsáveis pela produção de catecolaminas, melanina, serotonina, lipoato, pantotenato e heparan sulfato.

4.10.13. Metabolismo de Coenzimas

A S-adenosilmetionina (adoMet) é o principal cofator enzimático utilizado para transferências de grupos metil entre diferentes moléculas (Nelson & Cox 2000). Esse cofator é sintetizado a partir de uma molécula de ATP e uma metionina através da enzima metionina adenosil transferase (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR002133), identificada em um dois *uniques*.

Um *unique* representando uma proteína participante da via de síntese de tetrahidrobiopterinas foi identificado. As tetrahidrobiopterinas são cofatores importantes no catabolismo de aminoácidos e participam de reações de oxido-redução (Nelson & Cox 2000).

Identificamos um *unique* representando uma proteína que faz parte da via de síntese *de novo* de NAD e NADP, a nicotinamida-nucleotídeo pirofosforilase. O NAD e NADP são cofatores muito importantes nas reações metabólicas de oxido-redução, estando presentes em inúmeras vias bioquímicas.

Um *unique* foi identificado como a enzima trans-preniltransferase, importante na produção da coenzima Q (ubiquinona), um dos elementos-chave da fosforilação oxidativa. Os outros três *uniques* dessa categoria foram identificados como enzimas da via de produção de poli beta-hidroxibutirato, ADP-ribose cíclica e receptores para carboxiamidas.

4.10.14. TRANSPORTE E METABOLISMO DE NUCLEOTÍDEOS

Os nucleotídeos apresentam diversos papéis importantes nas células, sendo os blocos de construção para o DNA e RNA, constituindo as principais moléculas de armazenamento de energia (ATP e GTP), segundos mensageiros em vias de transdução de sinais (cAMP e cGMP) e sendo componentes de diversos cofatores enzimáticos e intermediários biossintéticos, como o NAD, FAD, adoMet, CoA, UDP-glicose o CTP-diacilglicerol (Nelson & Cox 2000).

As vias de síntese de nucleotídeos podem ser divididas em: vias de síntese *de novo* e vias de salvação. Nas vias de síntese *de novo*, os nucleotídeos são construídos a partir de aminoácidos, ribose 5-fosfato, gás carbônico e amônia. Já na via de salvação acontece a reciclagem das bases e dos nucleosídeos liberados durante quebra de ácidos nucléicos (Nelson & Cox 2000).

Vários trabalhos já identificaram a ausência da via de síntese *de novo* de nucleotídeos de purina em *S. mansoni* (Dovey *et al.* 1984, Senft *et al.* 1972). Concordando com estes artigos, não encontramos, entre nossos *uniques,* nenhum que representasse alguma das enzimas da via de síntese *de novo* de purinas, que leva à produção de inosinato (IMP) a partir de uma molécula de fosforibosil-pirofosfato (PRPP).

Um *unique* foi identificado como a enzima ribose fosfato pirofosfoquinase (PRPP sintetase), que produz o fosforibosil-pirofosfato a partir de uma molécula de ribose-5-fosfato. O PRPP é importante tanto para as vias de síntese *de novo* quanto para a via de salvação das purinas (Nelson & Cox 2000). Na via de salvação, a enzima adenina fosforibosil transferase (APRT) realiza catalisa uma reação de conjugação de uma adenina livre a uma molécula de PRPP, produzindo AMP mais pirofosfato. A APRT foi identificada em dois *uniques*. As guaninas e hipoxantinas (molécula derivada da desaminação da adenina) livres são reaproveitadas da mesma forma, produzindo GMP através da ação da enzima hipoxantina-guanina fosforibosiltransferase, encontrada em dois outros *uniques*. Esses nucleosídeos devem ser fosforilados de forma a produzir os nucleotídeos trifosfato, o que é feito pela enzima purina-nucleosídeo fosforilase, encontrada em três *uniques*.

Considerando a ausência da via de síntese *de novo* de nucleotídeos em *S. mansoni*, a procura por drogas inibitórias da via de salvação de nucleotídeos no verme tem sido adotada como uma

estratégia contra a esquistossomose (Foulk *et al.* 2002, Craig *et al.* 1995, Yuang L *et al.* 1992). A descoberta da estrutura tridimensional dessas proteínas tem sido procurada, assim como inibidores específicos para as proteínas do parasita, uma vez que a inibição da via de salvação de purinas em humanos pode acarretar em problemas para o paciente (Nelson & Cox 2000).

Foram identificados também *uniques* responsáveis pela conversão do IMP em GMP e AMP. As enzimas adenilosuccinato sintetase e adenilosuccinato liase, identificadas, cada uma, em dois *uniques*, são responsáveis pela transformação do IMP em AMP (Nelson & Cox 2000). A conversão do IMP em GTP é catalisada por duas enzimas, a IMP desidrogenase, identificada em dois *uniques* e a XMP-glutamina amidotransferase, que não foi identificada no presente trabalho.

Seis *uniques* da família das nucleosídeo monofosfato quinases foram encontrados, sendo dois *uniques* representando a adenilato quinase, dois representando a guanilato quinase, um representando a timidilato quinase e outro representando uma enzima de função mais geral, a nucleosídeo monofosfato quinase. Essas enzimas produzem duas moléculas de NDP (nucleotídeo difosfato) a partir de um NTP e um NMP. Os NDPs formados são fosforilados por outras enzimas para formar os nucleotídeos trifosfatos (Nelson & Cox 2000).

Com relação à síntese *de novo* de pirimidinas, encontramos um *unique* representando a enzima carbamil-fosfato sintetase e a dihidro-orotato desidrogenase. Ainda com relação ao metabolismo de pirimidinas, encontramos um *unique* representando a enzima dUTPase, responsável pela transformação de dUTP em dUMP para a síntese de dTTP durante a via de síntese de desoxiribonucleotídeos contendo timina. Encontramos também um *unique* representando a enzima dihidropirimidinase, utilizada na via de degradação de pirimidinas (Nelson & Cox 2000).

Uma enzima importante para a formação dos desoxiribonucleotídeos a partir dos ribonucleotídeos é a tioredoxina redutase, que foi identificada em um *unique*. Ela é utilizada pela enzima ribonucleotídeo redutase para reduzir nucleosídeos difosfatos em desoxirribonucleosídeos difosfato (Nelson & Cox 2000).

Ainda nesta categoria, encontramos dois *uniques* representando fosfodiesterases, enzimas que geram nucleosídeos monofosfato a partir de diversos nucleotídeos e derivados; um carreador mitocondrial de desoxiribonucleotídeos; uma purina nucleotidase; e uma 3'-5' bisfosfato nucleotidase.

4.10.15. METABOLISMO DE LIPÍDEOS

Os lipídeos apresentam diversas funções celulares. Tais moléculas são as principais formas de armazenamento de energia em grande parte dos organismos e estão presentes tanto em membranas celulares quanto em proteínas de membrana, sendo indispensáveis nessa função (Nelson & Cox 2000). A maioria dos ácidos graxos sintetizados ou ingeridos por um organismo acaba sendo transformado em triacilgliceróis, para o armazenamento de energia, ou sendo incorporado a fosfolípides de membrana (Nelson & Cox 2000).

A formação de triacilgliceróis pode ser realizada por via da dihidroxiacetona fosfato gerada durante a glicólise. Isso é feito através da ação da enzima citosólica glicerol 3-fosfato desidrogenase, representada aqui por três *uniques*, que transforma a dihidroxiacetona em uma molécula de L-glicerol 3-fosfato (Nelson & Cox 2000). Entretanto esse último intermediário pode também ser formado através da fosforilação de uma molécula de glicerol, reação que é catalisada pela enzima glicerol quinase (Nelson & Cox 2000), identificada em um *unique*. Um *unique* representando um transportador da molécula de L-glicerol 3-fosfato foi encontrado. Essas duas enzimas citadas, a glicerol quinase e a glicerol 3-fosfato desidrogenase também catalisam a reação reversa a essas comentadas e, portanto, são também importantes para a posterior quebra do glicerol para produzir energia. Voltando à síntese de triacilgliceróis, a molécula de L-glicerol 3-fosfato deve ser então acilada na posição de seus dois grupos hidroxila, produzindo uma molécula de ácido fosfatídico (Nelson & Cox 2000). Um *unique* foi identificado representando uma das enzimas que realizam essa acilação. O ácido fosfatídico é uma molécula central no metabolismo de lipídeos e pode ser convertido em triacilglicerol ou em um glicerofosfolípide através da ação de diversas enzimas. Uma dessas enzimas, a diacilglicerol acil-transferase, está presente na via de produção de triacilgliceróis a partir do ácido fosfatídico (Nelson & Cox 2000) e foi identificada em um *unique*. Três *uniques* relacionados com a formação de fosfolípides de membrana foram encontrados. Um representa uma quinase de fosfatidil-inositol (que também apresenta um papel importante na transdução de sinais), outro representa uma fosforiletolamina transferase e o último, uma fosfatidil-colina aciltransferase.

Três *uniques* relacionados à síntese de colesterol foram encontrados. Dois deles representam as enzimas acetoacetil-CoA tiolase e beta-hidroxi-beta-metil-glutaril-CoA (HGM-CoA) redutase, importantes na síntese de mevalonato a partir de acetil-CoA. O outro representava a enzima farnesil pirofosfato sintetase, que produz moléculas de farnesil pirofosfato, que se unem para formar o esqualeno, importante intermediário na formação do colesterol. Foi encontrado também um *unique* correspondente à enzima acil-CoA-colesterol aciltransferase, utilizada para produzir ésteres de colesteril (Nelson & Cox 2000).

Um *unique* foi identificado como a proteína soro albumina, que é capaz de transportar até dez moléculas de triacilgliceróis ligadas a ela de forma não covalente, através do sangue. O transporte de lipídeos através do sangue é realizado normalmente por proteínas conhecidas como apolipoproteínas. Diferentes combinações de lipídeos e proteínas podem alterar a densidade dessas moléculas transportadoras do sangue, sendo que dependendo da densidade essas moléculas apresentam receptores de membrana diferentes. Em nosso trabalho um *unique* foi identificado como receptor para HDL (lipoproteínas de alta densidade) e dois outros foram identificados como proteínas relacionadas a receptores de LDL (lipoproteínas de baixa densidade).

As enzimas utilizadas para a degradação de ácidos graxos estão presentes na mitocôndria das células e é necessário um transportador para levá-los até lá. Assim, dois *uniques* foram identificados como transportadores de ácidos graxos para a mitocôndria (transportador de carnitina/acil-carnitina). Uma vez na mitocôndria, os ácidos graxos são oxidados através da remoção de sucessivas unidades de dois carbonos na forma de acetil-CoA, na via conhecida como beta-oxidação dos ácidos graxos (Nelson & Cox 2000). Pudemos identificar um *unique* dessa via, representando a enzima hidroxiacil-CoA desidrogenase. Além disso, *S. mansoni* parece apresentar os três genes necessários para a omega oxidação de ácidos graxos: NADPH-citochrome p450 redutase (representada por três *uniques*); álcool desidrogenase (um *unique*); e aldeído desidrogenase (dois *uniques*) (Nelson & Cox 2000). Considerando que a beta oxidação de ácidos graxos é a via mais importante para a degradação de lípides é estranho que tenhamos encontrado mais enzimas para a via da omega oxidação em *S. mansoni*. Esse fato pode representar que essa via de oxidação, que acontece no ER, tenha uma maior importância neste verme, mas essa hipótese deve ser testada experimentalmente, já que a falta de enzimas identificadas da via de beta oxidação pode refletir apenas um fator aleatório ocorrido durante a clonagem das bibliotecas.

Um *unique* foi identificado como uma proteína carregadora de acila (ACP), que se complexa com a molécula de malonil-CoA e transfere um grupo acila para a molécula de malonil, liberando CoA. Essa enzima é importante na formação de ácidos graxos (Nelson & Cox 2000).

Muitas células degradam continuamente e substituem seus lipídeos de membrana (Nelson & Cox 2000). Uma das enzimas envolvidas nesse processo é a lisofosfolipase, identificada em dois *uniques*.

Ainda nessa categoria, foram identificados dois *uniques* representando enzimas para a produção de ácidos graxos de cadeia longa, um representando uma tioesterase utilizada na degradação desses ácidos graxos de cadeia longa e dois outros representando proteínas de ligação a ácidos graxos. Além disso, encontramos também um *unique* representando uma redutase envolvida na

síntese de colesterol, outro representando uma proteína envolvida no tráfego intracelular de colesterol e um último representando uma ATPase transportadora de fosfolípides.

4.10.16. PREDIÇÃO DA FUNÇÃO GERAL

Considerando que o método de classificação adotado aqui, o mesmo utilizado no COGs do NCBI, classifica os genes apenas a partir de sua função biológica, alguns *uniques* apresentando uma função molecular bem caracterizada acabaram ocupando essa categoria por não terem uma função biológica conhecida, como foi o caso de uma carbonil redutase. Além disso, os genes que apresentavam uma função identificada como provável foram classificados na categoria relacionada a tal função e também nesta categoria, já que não havia bons indícios de sua função precisa, como foi o caso de algumas anexinas e proteínas com o motivo de ligação ao DNA de dedos de zinco.

Mas, realmente, a maioria dos *uniques* caracterizados nesta categoria representa genes onde apenas a função geral pode ser predita. Por exemplo, uma porção importante desses genes compreende os genes de antígenos. Assim, treze *uniques* foram identificados como antígenos de *S. mansoni*, quatorze *uniques* foram identificados como similares a antígenos de *S. mansoni* e podem representar novas classes de antígenos do parasita, um *unique* foi identificado como um provável autoantígeno do ciclo celular, um *unique* apresentou similaridade com um antígeno humano expresso no soro de pacientes com neoplasias, dois outros apresentaram similaridade com os antígenos humanos CD53 e CD63 e um *unique* apresentou similaridade com um antígeno CD36 de rato. Um estudo mais detalhado desses *uniques* seria interessante na tentativa de produção de vacinas contra o parasita.

Como a divisão em categorias funcionais feita pelos COGs (e talvez qualquer outra) não representa com fidelidade todos os genes que podem ser encontrados em um organismo, alguns *uniques* representando proteínas importantes foram identificados e não couberam em nenhuma das categorias funcionais utilizadas, sendo então classificados aqui. Esse foi o caso, por exemplo, de proteínas responsáveis pela detoxicação celular por radicais peróxidos.

Dessa forma, cinco *uniques* identificados nesta categoria representavam genes para a superóxido dismutase, enzima que inibe danos causados por radicais livres de oxigênio produzidos durante a respiração aeróbica. Essas enzimas catalisam a transformação do radical superóxido em oxigênio molecular e peróxido de hidrogênio (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001424). Da mesma forma quatro *uniques* dessa categoria representavam a enzima tioredoxina peroxidase, que também atua como um antioxidante.

4.10.17. FUNÇÃO DESCONHECIDA

Nessa categoria foram classificados os *uniques* que apresentavam similaridade com genes sem função conhecida ou com genes hipotéticos de *S. mansoni* ou de outros organismos.

Deixamos ainda nesta função (mas não contabilizamos o valor nas TABELAS 8 e 9) os 87 *uniques* que apresentavam similaridades ao DNA genômico de *S. mansoni, Homo sapiens, Mus musculus* e *Escherichia coli*, já que muito provavelmente essas seqüências representam contaminações e não o cDNA de *S. mansoni*.

5. CONCLUSÕES E CONSIDERAÇÕES FINAIS

"Seqüenciar o DNA é agora uma das tarefas mais fáceis
de se realizar, além de servir hambúrgueres."
Kary Mullis, prêmio Nobel.

Acreditamos que o seqüenciamento de genomas e de transcriptomas devem continuar sendo tarefas incentivadas e patrocinadas pelos governos e pelas agências de financiamento de pesquisas em todo o mundo. Entretanto é importante ressaltar que a análise desses dados genômicos integrada com informações bioquímicas e biológicas do organismo que está sendo estudado é uma etapa essencial do processo. Parece-nos que tem sido gasto pouco tempo e esforço na análise dos genomas e transcriptomas já "concluídos", sendo que muito mais informação relevante poderia estar sendo obtida através desses dados. A corrida pelo seqüenciamento do próximo genoma, do qual será possível obter informações mais fáceis, rápidas e superficiais, impede uma análise detalhada do genoma estudado em um certo momento. Em vista dessa corrida incessante em busca do "próximo genoma", uma grande quantidade de dados permanece por ser analisada através de trabalhos como esse, onde nada além de computadores, horas na internet e pessoal capacitado foram gastos. Talvez aproveitar o lapso daqueles que seqüenciam os genomas e deixam de explorar todas as suas potencialidades possa ser a melhor alternativa para a ciência em países subdesenvolvidos, já que o custo desse tipo de pesquisa é bastante baixo. Nosso objetivo foi tentar extrair informações relevantes das seqüências de mRNA de *S. mansoni* disponíveis nos bancos de dados públicos e acreditamos termos podido contribuir para o aumento do conhecimento científico sobre esse parasita com esse trabalho, uma vez que um artigo já foi publicado e outro, onde será apresentado o *web-site* de divulgação das informações, está sendo preparado.

Assim analisamos primeiramente, neste trabalho, os prováveis genes mais expressos em *S. mansoni*, considerando a presença de cada um deles em diferentes fases de desenvolvimento do ciclo de vida do verme. Observamos que a maioria deles está envolvida na manutenção da homeostasia celular, codificando genes relacionados ao metabolismo de açúcares e de energia, citoesqueleto, chaperones, fatores de transcrição e tradução e enzimas de detoxicação. Genes relacionados à patogênese da esquistossomose também foram encontrados, codificando proteínas da casca de ovo e proteases. Vários genes não puderam ser identificados e podem ter funções específicas neste parasita, podendo ser utilizados posteriormente para fins de diagnóstico ou como alvos para drogas. É importante notar que a identificação de alguns genes como mais expressos pode ter sido resultado da redundância de alguma biblioteca. Essa observação deve ser aplicada principalmente aos genes encontrados apenas em uma biblioteca.

Foram identificados alguns problemas durante nossa análise devido a erros de agrupamento que ocorreram principalmente devido a regiões repetitivas dentro das seqüências (micro ou minissatélites) e pretendemos montar um script específico para mascarar tais regiões. Entretanto alguns *clusters* continham exatamente as mesmas seqüências quando os resultados obtidos por cada um dos programas utilizados foram comparado, o que reforça a confiabilidade desta análise.

Infelizmente, seqüências de outros estágios de desenvolvimento de *S. mansoni*, como o estágio de miracídio, ainda não estão disponíveis nos bancos de dados públicos. Assim, parece necessário um esforço da comunidade científica para preencher esses buracos de informação, produzindo mais seqüências de cada estágio de forma a completar a caracterização do transcriptoma desse parasita.

Na segunda parte de nosso trabalho realizamos a identificação de todas as seqüências de *S. mansoni* presentes nos bancos de dados públicos e pudemos encontrar vários fatos interessantes. Identificamos uma grande quantidade de seqüências de retrotransposons, inclusive algumas similares a outras já identificadas em outro trematódeo, e imaginamos que seja possível realizar a busca de novos elementos repetitivos no genoma de *S. mansoni* através da observação dos dados já obtidos para *C. sinensis*. Identificamos a presença de vias metabólicas completas no verme, como a glicólise e a via das pentoses (fase oxidativa e não-oxidativa). Mostramos a presença de todos os genes necessários para a utilização de galactose e vários necessários para o metabolismo anaeróbico. Sugerimos algumas vias, como a via de salvação das purinas e as vias de produção da arginina, do glutamato e da selenocisteína, e proteínas específicas, como a purina-nucleosídeo fosforilase, a glutamato sintase e a "*selenide water-dikinase*", nas quais seria interessante realizar uma investigação mais minuciosa no sentido de obter possíveis alvos para a fabricação de drogas antiesquistossomose. Identificamos também algumas proteínas similares a antígenos de *S. mansoni, H. sapiens* e *R. norvegicus* ainda não caracterizadas que poderiam, talvez, servir de ponto de partida para a produção de novas vacinas antiesquistossomose.

Novamente devemos dizer que acreditamos que o número de seqüências presentes nos bancos de dados ainda é baixo para a realização de um trabalho mais apurado para a descoberta de alvos para drogas e vacinas contra a esquistossomose. Felizmente, novos projetos de seqüenciamento de milhares de ESTs deste verme estão sendo produzidos em projetos nos estados de São Paulo e Minas Gerais e, em breve, novas informações sobre o transcriptoma desse verme devem ser produzidas e podem preencher as lacunas presentes atualmente. Acreditamos também que deveria ser dada mais ênfase para projetos de seqüenciamento do genoma completo desse organismo tão importante na parasitologia humana.

6. REFERÊNCIAS BIBLIOGRÁFICAS

1. **Adams, M. D.; Kelley, J. M.; Gocayne, J. D.; Dubnick, M.; Polymeropoulos, M. H.; Xiao, H.; Merril, C. R.; Wu, A.; Olde, B.; Moreno, R. F.; Kerlavage, A. R.; McCombie, W. R. and Venter, J. C.** (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-6.

2. **Adams, M. D.; Celniker, S. E.; Holt, R. A.; Evans, C. A.; Gocayne, J. D.; Amanatides, P. G.; Scherer, S. E.; Li, P. W.; Hoskins, R. A.; Galle, R. F.; George, R. A.; Lewis, S. E.; Richards, S.; Ashburner, M.; Henderson, S. N.; Sutton, G. G.; Wortman, J. R.; Yandell, M. D.; Zhang, Q.; Chen, L. X.; Brandon, R. C.; Rogers, Y. H.; Blazej, R. G.; Champe, M.; Pfeiffer, B. D.; Wan, K. H.; Doyle, C.; Baxter, E. G.; Helt, G.; Nelson, C. R.; Gabor, G. L.; Abril, J. F.; Agbayani, A.; An, H. J.; Andrews-Pfannkoch, C.; Baldwin, D.; Ballew, R. M.; Basu, A.; Baxendale, J.; Bayraktaroglu, L.; Beasley, E. M.; Beeson, K. Y.; Benos, P. V.; Berman, B. P.; Bhandari, D.; Bolshakov, S.; Borkova, D.; Botchan, M. R.; Bouck, J.; Brokstein, P.; Brottier, P.; Burtis, K. C.; Busam, D. A.; Butler, H.; Cadieu, E.; Center, A.; Chandra, I.; Cherry, J. M.; Cawley, S.; Dahlke, C.; Davenport, L. B.; Davies, P.; de, Pablos, B.; Delcher, A.; Deng, Z.; Mays, A. D.; Dew, I.; Dietz, S. M.; Dodson, K.; Doup, L. E.; Downes, M.; Dugan-Rocha, S.; Dunkov, B. C.; Dunn, P.; Durbin, K. J.; Evangelista, C. C.; Ferraz, C.; Ferriera, S.; Fleischmann, W.; Fosler, C.; Gabrielian, A. E.; Garg, N. S.; Gelbart, W. M.; Glasser, K.; Glodek, A.; Gong, F.; Gorrell, J. H.; Gu, Z.; Guan, P.; Harris, M.; Harris, N. L.; Harvey, D.; Heiman, T. J.; Hernandez, J. R.; Houck, J.; Hostin, D.; Houston, K. A.; Howland T. J.; Wei, M. H.; Ibegwam, C.; Jalali, M.; Kalush, F.; Karpen, G. H.; Ke, Z.; Kennison, J. A.; Ketchum, K. A.; Kimmel, B. E.; Kodira, C. D.; Kraft, C.; Kravitz, S.; Kulp, D.; Lai, Z.; Lasko, P.; Lei, Y.; Levitsky, A. A.; Li, J.; Li, Z.; Liang, Y.; Lin, X.; Liu, X.; Mattei, B.; McIntosh, T. C.; McLeod, M. P.; McPherson, D.; Merkulov, G.; Milshina, N. V.; Mobarry, C.; Morris, J.; Moshrefi, A.; Mount, S. M.; Moy, M.; Murphy, B.; Murphy, L.; Muzny, D. M.; Nelson, D. L.; Nelson, D. R.; Nelson, K. A.; Nixon, K.; Nusskern, D. R.; Pacleb, J. M.; Palazzolo, M.; Pittman, G. S.; Pan, S.; Pollard, J.; Puri, V.; Reese, M. G.; Reinert, K.; Remington, K.; Saunders, R. D.; Scheeler, F.; Shen, H.; Shue, B. C.; Siden-Kiamos, I.; Simpson, M.; Skupski, M. P.; Smith, T.; Spier, E.; Spradling, A. C.; Stapleton, M.; Strong, R.; Sun, E.; Svirskas, R.; Tector, C.; Turner, R.; Venter, E.; Wang, A. H.; Wang, X.; Wang, Z. Y.; Wassarman, D. A.; Weinstock, G. M.; Weissenbach, J.; Williams, S. M.; WoodageT.; Worley, K. C.; Wu, D.; Yang, S.; Yao, Q. A.; Ye, J.; Yeh, R. F.; Zaveri, J. S.; Zhan, M.; Zhang, G.; Zhao, Q.; Zheng, L.; Zheng, X. H.; Zhong, F. N.; Zhong, W.; Zhou, X.; Zhu, S.; Zhu, X.; Smith, H. O.; Gibbs, R. A.; Myers, E. W.; Rubin, G. M. and Venter, J. C.** (2000). The genome sequence of *Drosophila melanogater*. *Science* 287: 2185-2195.

3. **Afshar, K.; Stuart, B. and Wasserman, S. A.** (2000). Functional analysis of the *Drosophila* diaphanous FH protein in early embryonic development. *Development* 127: 1887-1897.

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

5. **Almeida, P. C.; Nantes, I. L.; Rizzi, C. C. A.; Júdice, W. A. S.; Chagas, J. R.; Juliano, L.; Nader, H. B.;. and Tersariol, I. L. S.** (1999). Cysteine proteinase activity regulation. *J Biol Chem* 274: 30433-30438.

6. **Apweiler, R.** (2001). Functional information in SWISS-PROT: The basis for large-scale characterisation of protein sequences. *Brief Bioinform* 2: 9-18.

7. **Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Birney, E.; Biswas, M.; Bucher, P.; Cerutti, L.; Corpet, F.; Croning, M. D.; Durbin, R.; Falquet, L.; Fleischmann, W.; Gouzy, J.; Hermjakob, H.; Hulo, N.; Jonassen, I.; Kahn, D.; Kanapin, A.; Karavidopoulou, Y.; Lopez, R.; Marx, B.; Mulder, N. J.; Oinn, T. M.; Pagni, M.; Servant, F.; Sigrist, C. J. and Zdobnov, E. M.** (2001). The InterPro Database; an integrated documentation resource for protein families; domains and functional *sites*. *Nucleics Acid Res* 29: 37-40.

8. **Arteaga, C. L.** (2001). The Epidermal Growth Factor Receptor: From Mutant Oncogene in Nonhuman Cancers to Therapeutic Target in Human Neoplasia. *J Clin Onc* 19: 32-40.

9. **Aubourg, S. and Rouzé P.** (2001)**.** Genome annotation. *Plant Physiol Biochem* 39: 181-193.

10. **Bae, Y-A.; Moon, S-Y.; Kong, Y.; Cho, S-Y. and Rhyu, M-G.** (2001). CsRn1; a novel active retrotransposon in a parasitic trematode; *Clonorchis sinensis*; discloses a new phylogenetic clade of Ty3/gypsy-like LTR retrotransposon. *Mol Biol Evol 18*: 1474-1483.

11. **Bairoch, A.** (2000). The ENZYME database in 2000. *Nucleic Acids Res* 28: 304-305.

12. **Bairoch, A. and Apweiler, R.** (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28: 45-48.

13. **Baumeister, W.; Walz, J.; Zühl, F. and Seemüller, E.** (1998). The proteasome: paradigm of a self-compartmentalizing protease. *Cell* 92: 367-380.

14. **Baxevanis, A. D. and Ouellette, B. F. F.** (2001). Bioinformatics: A practical guide to the analysis of genes and proteins. *Ed. Wiley-interscience. 2nd* ed.

15. **Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Rapp, B. A. and Wheeler, D. L.** (2002). GenBank. *Nucleic Acids Res* 30: 17-20.

16. **Bonifacino, J. S. and Weissman, A. M.** (1998). Ubiquitin and the control of protein fate in the secretory and endocytic pathways. *Annu Rev Cell Dev Biol* 14: 19-57.

17. **Boguski, M. S.; Lowe, T. M. and Tolstoshey, C. M.** (1993). dbEST--database for "expressed sequence tags". *Nat Genet* 4: 332-333

18. **Brindley, P. J.; Kalinna B. H.; Dalton, J. P.; Day, S. R.; Wong, J. Y. M.; Smythe, M. L. and Mcmanus, D. P.** (1997). Proteolytic degradation of host hemoglobin by schistosomes. *Mol Biochem Parasitol* 89: 1-9.

19. **Bruce, J. I.; Ruff, M. D. and Hasegawa, H.** (1971). *Schistosoma mansoni*: endogenous and exogenous glucose and respiration of cercariae. *Exp Parasitol* 29: 86-93.

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

21. **Chamberlain, L. H. and Burgoyne, R. D.** (1997). Activation of the ATPase activity of heat-shock proteins Hsc70/Hsp70 by cysteine-string protein. *Biochem J* 322: 853-858.

22. **Chen, L. L.; Rekosh, D. M.; LoVerde, P. T.** (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.

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

24. **Corpet, F.** (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* 16: 10881-10890.

25. **Craig, S. P. 3rd; McKerrow, J. H.; Newport, G. R. and Wang, C. C.** (1988). Analysis of cDNA encoding the hypoxanthine-guanine phosphoribosyltransferase (HGPRTase) of Schistosoma mansoni; a putative target for chemotherapy. *Nucleic Acids Res* 16: 7087-7101.

26. **Cross, R. A.** (2000). Molecular motors: kinesin's dynamically dockable neck. *Curr Biol* 10: R124-R126.

27. **Davies, K.** (2001). Decifrando o genoma. *Companhia das letras*.

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

29. **Denhardt, D. T.** (1996). Signal-transducting protein phosphorylation cascades mediated by Ras/Rho proteins in the mammalian cell: the potential for multiplex signalling. *Biochem J* 316: 729-747.

30. **Degrave, W. M.; Melville, S.; Ivens, A. and Aslett, M.** (2001). Parasite genome iniciatives. *Int J Parasitol* 31: 532-536.

31. **Dias-Neto, E.; Harrop, R.; Correa-Oliveira, R.; Wilson, R. A.; Pena, S. D. J. and Simpson, A. J. G.** (1997). Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normalized libraries for the generation of ESTs from nanogram quantities of mRNA. *Gene* 186: 135-142.

32. **Dias-Neto, E.; Garcia, Correa, R.; Verjovski-Almeida, S.; Briones, M. R.; Nagai, M. A.; da, Silva, W. Jr.; Zago, M. A.; Bordin, S.; Costa, F. F.; Goldman, G. H.; Carvalho, A. F.; Matsukuma, A.; Baia, G. S.; Simpson, D. H.; Brunstein, A.; de, Oliveira, P. S.; Bucher, P.; Jongeneel, C. V.; O'Hare, MJ.; Soares, F.; Brentani, R. R.; Reis, L. F.; de, Souza, S. J. and Simpson, A. J.** (2000). Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc Natl Acad Sci USA* 97: 3491-3496.

33. **Devos, D. and Valencia, A.** (2001). Intrinsic errors in genome annotation. *Trends Genet* 17: 429-431.

34. **Doudna, J. A. and Rath, V. L.** (2002). Structure and function of the eukaryotic ribosome: The next frontier. *Cell* 109: 153-156.

35. **Dovey, H. F.; McKerrow, J. H. and Wang, C. C.** (1984). Purine salvage in *Schistosoma mansoni* schistosomules. *Mol Biochem Parasitol* 11: 157-167.

36. **Drew, A. C. and Brindley, P. J.** (1997). A retrotransposon of the non-long terminal repeat class from the human blood fluke *Schistosoma mansoni*. Similarities to the chichen-repeat-1-like retroelements of vertebrates. *Mol Biol Evol* 14: 602-610.

37. **Drew, A. C.; Minchella, D. J.; King, L. T.; Rollinson, D. and Brindley, P. J.** (1999). SR2 elements; non-long terminal repeat retrotransposon of the RTE-1 lineage from the human blood fluke *Schistosoma mansoni*. *Mol Biol Evol* 16: 1256-1269.

38. **Dunne, D. and Mountford, A.** (2001). Resistance to infection in humans and animal models. In: Schistosomiasis. *Imperial College Press*. 133-212.

39. **Faria, M. S. C.** (2000). Programa de descoberta gênica em *Schistosoma mansoni*: Geração e análise de 1335 etiquetas de seqüências transcritas (ESTs) de duas bibliotecas de cDNA de ovo. *Dissertação apresentada para obtenção do título de Mestre ao Programa de Pós-graduação em Genética; UFMG; Belo Horizonte*.

40. **Feder, M. E. and Hoffman, G. E.** (1999). Heat shock proteins; molecular chaperones; and the stress response: Evolutionary and Ecological physiology. *Annu Rev Physiol* 61: 243-282.

41. **Ford, H. L.** (1998). Homeobox genes: a link between development; cell cycle; and cancer? *Cell Biol Int* 22: 397-400.

42. **Foulk, B. W.; Pappas, G.; Hirai, Y.; Hirai, H. and Williams, D. L.** (2002). Adenylosuccinate lyase of *Schistosoma mansoni*: gene; structure; mRNA expression; and analysis of the predicted peptide structure of a potential chemotherapeutic target. *Int J Parasitol* 32: 1487-1495.

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

44. **Franco, G. R.; Valadao, A. F.; Azevedo, V. and Rabelo, E. M.** (2000). The *Schistosoma* gene discovery program: state of the art. *Int J Parasitol* 30: 453-463.

45. **Francklyn, C.; Perona, J. J.; Puetz, J. and Hou, Y. M.** (2002). Aminoacyl-tRNA synthetases: Versatile players in the changing theater of translation. *RNA* 8: 1363-1372.

46. **Frankel, S. and Mooseker, M. S.** (1996). The actin-related proteins. *Curr Biol* 8: 30-37.

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

48. **Grabe, M.; Wang, H. and Oster, G** (2000). The mechanochemistry of V-ATPase proton pumps. *Biophys J* 78: 2798-2813.

49. **Green, P** (1999)**.** Documentation for PHRAP and CROSS_MATCH (version 0.990)319) (http://www.phrap.org/phrap.docs/phrap.html).

50. **Griffiths, A. J. F.; Miller, J. H.; Suzuki, D. T.; Lewontin, R. C. and Gelbart, W. M.** (1998). Introdução à genética. *Ed. Guanabara Koogan*: Rio de Janeiro. 6ª ed.

51. **Hall, R. A.; Premont, R. T. and Lefkowitz, R. J.** (1999). Heptahelical receptor signaling: beyond the G protein paradigm. *J Cell Biol* 145: 927-932.

52. **Hamm, J. and Lamond, A. I.** (1998). The unwinding role of DEAD-box proteins. *Curr Biol* 8: R532-R534.

53. **Hanford, P. A.** (2000). Fibrillin-1, a calcium-binding protein of extracellular matrix. *Biochim Biophys Acta* 1498: 84-90.

54. **Hardie, D. G.** (1999). Plant protein serine/threonin kinases: classification and functions. *Annu Rev Plant Physiol Plant Mol Biol* 50: 97-131.

55. **Hens, J.; Nuydens, R.; Geerts, H.; Senden, N. H.; Van de Ven, J. M.; Roebroek, A. J.; van de Velde, H. J.; Ramaekers, F. C. and Broers, J. L.** (1998). Neuronal differentiation is accompanied by NSP-C expression. *Cell Tissue Res* 292: 229-237.

56. **Hochstrasser, M.** (2000). Evolution and function of ubiquitin-like protein-conjugation systems. *Nat Cell Biol* 2: E153-E157.

57. **Huang, X.** (1999). CAP3 Sequence Assembly Program (http://genome.cs.mtu.edu/cap/cap3.html).

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

59. **International Human Genome Sequencing Consortium** (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 859-

60. **Isaac, R. E.; Siviter, R. J.; Stancombe, P.; Coates, D. and Shirrasã A. D.** (2001). Conserved roles for peptidases in the processing of invertebrate neuropeptides. *Biochem Soc Trans* 28: 460-464.

61. **Jaime, M.; Pujol, M. J.; Serratosa, J.; Pantoja, C.; Canela, N. and Casanovas, O.** (2002). The $p21^{Cip1}$ Protein; a Cyclin Inhibitor; Regulates the Levels and the Intracellular Localization of CDC25A in Mice Regenerating Livers. *Hepatol* 35: 1063-1071.

62. **Jensen, R. A.** (2001). Orthologs and paralogs – we need to get it right. *Genome Biol* 2: 1002.1-1002.3.

63. **Junker, V.; Contrino, S.; Fleischmann, W.; Hermjakob, H.; Lang, F.; Magrane, M.; Martin, M. J.; Mitaritonna, N.; O'Donovan, C. and Apweiler, R** (2000). The role SWISS-PROT and TrEMBL play in the genome research environment. *J Biotechnol* 78: 221-234.

64. **Kanehisa, M.; Goto, S.; Kawashima, S. and Nakaya, A** (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res* 30: 42-46.

65. **Kitts, P. A.; Madden, T. L.; Sicotte, H. and Ostell, J. A.** (2002). Univec and Univec_core Databases. *Manuscript in preparation*. (READ ME – Univec – ftp://ncbi.nlm.nih.gov/pub/UniVec/README.uv)

66. **Klinker, M. Q.; Felleisen, R.; Link, G.; Ruppel, A. and Beck, E.** (1989). Primary structure of Sm31/32 kDa proteins diagnostic proteins of *Schistosoma mansoni* and their identification as proteases. *Mol Biochem Parasitol* 33: 113-122.

67. **Klopfenstein, D. R. and Vale, R. D.** (2000). Motor protein receptors: moonlighting on other jobs. *Cell* 103: 537-540.

68. **Klug, A.** (1999). Zinc finger peptides for the regulation of gene expression. *J Mol Biol* 293: 215-218.

69. **Kolodner, R. D. and Marsischky, G. T.** (1999). Eukaryotic DNA mismatch repair. *Curr Opin Gen & Dev* 9: 89-96.

70. **Lang, F.** (1997). TREMBL. *Trends Genet* 13: 417.

71. **Lehmann, A. R.** (1995). Nucleotide excision repair and the link with transcription. *Trends Biochem Sci* 20: 402-405.

72. **Lei, M. and Tye, B. K.** (2001). Initiating DNA synthesis: from recruiting to activating the MCM complex. *J Cell Sci* 114: 1447-1454.

73. **Lew, D.** (2000). Cell-cycle checkpoints that ensure coordination between nuclear and cytoplasmic events in *Saccharomyces cerevisiae. Curr Opin Gen & Dev* 10: 47-53.

74. **Lewis, S.; Ashburner, M. and Reese, M. G.** (2000). Annotating eukaryote genomes. *Curr Opin Struct Biol* 10: 349–354.

75. **Li, Y. L.; Idris, M. A.; Corachan, M.; Han, J. J.; Kirschfink, M.; Ruppel, A.** (1996) Circulating antigens in schistosomiasis: detection of 31/32. in sera from patients infected with *Schistosoma japonicum, S. mansoni, S. haematobium*, or *S. intercalatum*. *Parasitol Res* 82: 14-18.

76. **Liang, F.; Holt, I.; Pertea, G.; Karamycheva, S.; Salzberg, S. L. and Quackenbush, J.** (2000). An optimized protocol for analysis of EST sequences. *Nucleic Acids Res* 28: 3657-3665.

77. **Liou, A. K. F.; McCormack, E. A. and Willison, K. R.** (1998). The chaperonin containing TCP-1 (CCT) displays a single-ring mediated disassembly and reassembly cycle. *Biol Chem* 379: 311-319.

78. **Longtine, M. S.; DeMarini, D. J.; Valencik, M. L.; Al-Awar, O. S.; Fares, H.; De, Virgilio, C. and Pringle, J. R.** (1996). The septins: roles in cytokinesis and other processes. *Curr Opin Cell Biol* 8: 106-119.

79. **Makeyev, A. V. and Liebhaber, S. A.** (2002). The poly(C)-binding proteins: a multiplicity of functions and a search for mechanisms. *RNA* 8: 265-278.

80. **Marx, K. A.; Bizzaro, J. W.; Blake, R. D.; Hsien Tsai, M. and Feng Tao, L.** (2000). Experimental DNA melting behavior of the three major *Schistosoma* species. *Mol Biochem Parasitol* 107: 303-307.

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

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

83. **Memisoglu, A. and Samson, L.** (2000). Base excision repair in yeast and mammals. *Mutat Res* 451: 36-51.

84. **Miller R. T.; Christoffels A. G.; Gopalakrishnan C.; Burke J.; Ptitsyn A. A.; Broveak T. R. and Hide W. A.** (1999). A Compreensive approach to clustering of expressed gene sequence: The sequence tag alignment and consensus knowledge base. *Genome Res* 9: 1143-1155.

85. **Moroianu, J.; Blobel, G. and Radu, A.** (1996). The binding site of karyopherin α for karyopherin β overlaps with a nuclear localization sequence. *Proc Natl Acad Sci USA* 93: 6572–6576.

86. **Mossi, R. and Hübscher, U.** (1998). Clamping down on clamps and clamp loaders: The eukaryotic replication factor C. *Eur J Biochem* 254: 209-216.

87. **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 crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biol* 1: 9.1–9.19

88. **Nelson, D. L. and Cox, M. M.** (2000). Lehninger Principles of Biochemistry*. Worth publishers*. 3rd ed.

89. **Neves, D. P.** (1983). Parasitologia Humana. 5 ed*. Livraria Atheneu* ltda.

90. **Newton, A. C.** (1997). Regulation of protein kinase C. *Curr Opin Cell Biol* 9: 161-167.

91. **Noiva, R.** (1999). Protein disulfide isomerase: the multifunctional redox chaperone of the endoplasmic reticulum. *Cell Dev Biol* 10: 481-493.

92. **Oliveira, G. and Johnston, D. A.** (2001). Mining the schistosome DNA sequence database. *Trends Parasitol* 17: 501-503.

93. **Ou, W-B.; Luo, W.; Park, Y-D. and Zhou, H-M.** (2001). Chaperone-like activity of peptidyl-prolyl cis-trans isomerase during creatine kinase refolding. *Protein Sci* 10: 2346-2353.

94. **Paule, M. R. and White, R. J.** (2000). Transcription by RNA polymerases I and III. *Nucleic Acids Res* 28: 1283-1298.

95. **Pollard, TD** (2001). Genomics; the cytoskeleton and motility. *Nature* 409: 842-843.

96. **Polly, P.; Danielsson, C.; Schräder, M. and Carlberg, C** (2000). Cyclin C is a primary 1α-25-dihydroxyvitamin D3 responding gene. *J Cell Biochem* 77: 75-81.

97. **Prakash, S. and Prakash, L** (2000). Nucleotide excision repair in yeast. *Mutat Res* 451: 13-24.

98. **Prosdocimi, F** (2002). Análises bioinformáticas de transcritos de *Schistosoma mansoni*. *Monografia apresentada para a obtenção do título de Bacharel em Ciências Biológicas/Genética ao Departamento de Biologia Geral; UFMG; Belo Horizonte*.

99. **Prosdocimi, F.; Faria-Campos, A. F.; Peixoto, F.; Pena, S. D. J.; Ortega, 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.

100. **Prosdocimi, F.; Cerqueira, G. C.; Binneck, E.; Silva, A. F.; Reis, A. N.; Junqueira, A. C. M.; Santos, A. C. F.; Nhani-Júnior, A.; Wust, C. I.; Camargo-Filho, F.; Kessedjian, J. L.; Petretski, J. H.; Camargo, L. P.; Ferreira, R. G. M.; Lima, R. P.; Pereira, R. M.; Jardim, S.; Sampaio, V. S. and Folgueras-Flatschar, A. V.** (2003). Bioinformática: Manual do usuário. *Revista Biotecnologia* 29: 18-31.

101. **Pruitt, K. D. and Maglott, D. R.** (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 29: 137-140.

102. **Quackenbush J.; Cho J.; Lee D.; Liang F.; Holt I.; Karamycheva S.; Parvizi B.; Pertea G.; Sultana R. and White J.** (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res* 29: 159-164.

103. **Ram, D.; Grossman, Z.; Markovics, A.; Avivi, A.; Ziv, E; Lantner, F. and 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.

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

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

106. **Rouzé P.; Pavy, N. and Rombauts, S.** (1999). Genome annotation: which tools do we have for it? *Curr Opin Struct Biol* 2: 90-95.

107. **Rubin, G. M.; Yandell, M. D.; Wortman, J. R.; Gabor Miklos, G. L.; Nelson, C. R.; Hariharan, I. K.; Fortini, M. E.; Li, P. W.; Apweiler, R.; Fleischmann, W.; Cherry, J. M.; Henikoff, S.; Skupski, M. P.; Misra, S.; Ashburner, M.; Birney, E.; Boguski, M. S.; Brody, T.; Brokstein, P.; Celniker, S. E.; Chervitz, S. A.; Coates, D.; Cravchik, A.; Gabrielian, A.; Galle, R. F.; Gelbart, W. M.; George, R. A.; Goldstein, L. S.; Gong, F.; Guan, P.; Harris, N. L.; Hay, B. A.; Hoskins, R. A.; Li, J.; Li, Z.; Hynes, R. O.; Jones, S. J.; Kuehl, P. M.; Lemaitre, B.; Littleton, J. T.; Morrison, D. K.; Mungall, C.; O'Farrell, P. H.; Pickeral, O. K.; Shue, C.; Vosshall, L. B.; Zhang, J.; Zhao, Q.; Zheng, X. H. and Lewis, S.** (2000). Comparative Genomics of the Eukaryotes. *Science* 287: 2204-2215.

108. **Rumjanek, F. D.** (1987). Biochemistry and physiology. In The Biology of Schistosomes, Academic Press, p. 163–183.

109. **Ryan, M. T.; Naylor, D. J.; Hoj, P. B.; Clark, M. S. and Hoogenraad, N. J.** (1997). The role of molecular chaperones in mitochondrial protein import and folding. *Int Rev Cytol* 174:127–193.

110. **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). Analyses of the gene expression profile of *Schistosoma mansoni* cercariae using the expressed sequence tag approach. *Mol Bioch Parasit* 103: 79-97.

111. **Schuler, G. D.** (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* 75: 694-698.

112. **Seeberg, E.; Eide, L. and Bjoras, M.** (1995). The base excision repair pathway. *Trends Biochem Sci* 20: 391-396.

113. **Senft, A. W.; Miech, R. P.; Brown, P. R. and Senft, D. G.** (1972). Purine metabolism in *Schistosoma mansoni*. *Int J Parasitol* 2: 249-260.

114. **Sheehan, D.; Meade, G.; Foley, V. M. and Dowd, C. A.** (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.

115. **Simpson, A. J. G.; Sher, A. and McCutchan, T. F.** (1982). The genome of *Schistosoma mansoni*: isolation of DNA; its size; bases and repetitive sequences. *Mol Biochem Parasitol* 6: 125-137.

116. **Skelly P. J.; Lincoln L. D. and Shoemaker C. B.** (1993). Expression of Schistosoma mansoni genes involved in anaerobic and oxidative glucose metabolism during the cercária to adult transformation. *Mol Biochem Parasitol* 60: 93-104.

117. **Stansfield, I.; Jones, M. J. and Tuite, M. F.** (1995). The end in sight: terminating translation in eukaryotes. *Trends Biochem Sci* 20; 489-491.

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

119. **Stoesser, G.; Baker, W.; van, den, Broek, A.; Camon, E.; Garcia-Pastor, M.; Kanz, C.; Kulikova, T.; Leinonen, R.; Lin, Q.; Lombard, V.; Lopez, R.; Redaschi, N.; Stoehr, P.; Tuli, M. A.; Tzouvara, K. and Vaughan, R** (2002). The EMBL nucleotide sequence database. *Nucleic Acids Res* 30: 21-26.

120. **Stothard, P.** (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques* 28: 1102-1104.

121. **Takisawa, H.; Mimura, S. and Kubota, Y.** (2000). Eukaryotic DNA replication: from pre-replication complex to initiation complex. *Curr Opin Cell Biol* 12: 690-696.

122. **Tateno, Y.; Imanishi, T.; Miyazaki, S.; Fukami-Kobayashi, K.; Saitou, N.; Sugawara, H. and Gojobori, T.** (2002). The DNA Data Bank of Japan (DDBJ) for genome scale research in life sciences. *Nucleic Acids Res* 30: 27-30.

123. **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.

124. **Terstappen, G. C. and Reggiani, A.** (2001). In silico research in drug discovery. *Trends Pharmacol Sci* 22: 23-26.

125. **The Gene Ontology Consortium** (2000). Gene Ontology: tool for the unification of biology. *Nat Genet*. 25: 25-29.

126. **The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium** (2001)**.** Functional annotation of a full-length mouse cDNA collection. *Nature* 409: 685-690.

127. **Thompson, D. P.; Morrison, D. D.; Pax, R. A. and Bennet, J. L.** (1984). Changes in glucose metabolism and cyanide sensitivity in *Schistosoma mansoni* during development. *Mol Biochem Parasitol* 13: 39-51.

128. **Thorson, J. A.; Yu, L. W. K.; Hsu, A. L.; Shih, N. Y.; Graves, P. R.; Tanner, W.; Allen, P. M.; Piwnica-Worms, H. and Shaw, A. S.** (1998). 14-3-3 proteins are required for maintenance of raf-1 phophorylation and kinase activity. *Mol Cell Biol* 18: 5229-5238.

129. **Timms, A. R. and Bueding, E.** (1959). Studies of a proteolytic enzyme from *Schistosoma mansoni*. *Br J Pharmacol* 14: 68-73.

130. **Tong, J. K.** (2002). Dissecting histone deacetylase function. *Chem & Biol* 9: 688-670.

131. **Vallee, R. B. and Gee, M. A.** (1998). Make room for dynein. *Trends Cell Biol* 8: 490-493.

132. **Van Oordt, B. E. P.; Tielens A. G. M. and Van den Bergh S. G.** (1989). The energy production of the adult *Schistosoma mansoni* is for large parte aerobic. *Mol Biochem Parasitol* 16: 117-126.

133. **van de Velde, H. J.; Senden, N. H.; Roskams, T.A.; Broers, J. L.; Ramaekers F. C.; Roebroek, A. J.; Van de Ven, W. J.** (1994). NSPencoded reticulons are neuroendocrine markers of a novel category in human lung cancer diagnosis. *Cancer Res* 54: 4769-4776.

134. **Watson, J. D.; Gilman, M; Witkowski, J. and Zoller, M.** (1997). O DNA recombinante. *Ed. UFOP*: Ouro Preto. 2ª ed.

135. **Wheeler, D. L.; Church, D. M.; Lash, A. E.; Leipe, D. D.; Madden, T. L.; Pontius, J. U.; Schuler, G. D.; Schriml, L. M.; Tatusova, T. A.; Wagner, L. and Rapp, B. A.** (2002). Database resources of the National Center for Biotechnology information: 2002 update. *Nucleic Acids Res* 30: 13-16.

136. **Wu, S-J.; Liu, F-H.; Hu, S-M. and Wang, C.** (2001). Different combinations of the heat-shock cognate protein 70 (hsc70) C-terminal functional groups are utilized to interact with distinct tetratricopeptide repeat-containing proteins. *Biochem J* 359: 419-426.

137. **Young, J. C.; Moarefi, I. And Hartl, F. U.** (2001). Hsp90: a specialized but essential protein-folding tool. *J Cell Biol* 154: 267-273.

138. **Yuang, L.; Craig, S. P. 3rd; McKerrow, J. H. and Wang, C. C.** (1992). Steady-state kinetics of the schistosomal hypoxanthine-guanine phosphoribosyltransferase. *Biochemistry* 31: 806-810.