

FUNDAÇÃO OSWALDO CRUZ
CENTRO DE PESQUISAS RENÉ RACHOU

Programa de Pós-graduação em Ciências da Saúde

**Detecção de polimorfismo de base única em
etiquetas de seqüências expressas de
*Schistosoma mansoni***

por

Mariana Crivellari Machado Simões

Belo Horizonte, MG
2005

Ministério da Saúde
Centro de Pesquisas René Rachou
Programa de Pós-graduação em Ciências da Saúde

**Detecção de polimorfismo de base única em
etiquetas de seqüências expressas de
*Schistosoma mansoni***

por

Mariana Crivellari Machado Simões

Dissertação apresentada com vistas à obtenção do Título de Mestre em Ciências.
Área de Concentração: Biologia Celular e Molecular.

Orientador: Dr. Guilherme Oliveira
Co-orientadora: Dra. Diana Bahia

Maio, 2005

Ficha catalográfica

Simões, Mariana Crivellari Machado

“Detecção de polimorfismo de base única em etiquetas de seqüências expressas de *Schistosoma mansoni*”.

“Detection of single nucleotide polymorphisms in expressed sequence tags of *Schistosoma mansoni*”

Dissertação: Mestrado em Biologia Celular e Molecular.

Palavras-chave: Esquistossomose, *Schistosoma mansoni*, Polimorfismos de base única, SNPs, Etiquetas de seqüências expressas, ESTs, Catepsina B, Seqüenciamento, Bioinformática, Modelagem estrutural.

Ministério da Saúde
Centro de Pesquisas René Rachou
Programa de Pós-graduação em Ciências da Saúde

**Detecção de polimorfismo de base única em
etiquetas de seqüências expressas de
*Schistosoma mansoni***

por

Mariana Crivellari Machado Simões

Este trabalho foi realizado no Laboratório de Parasitologia Celular e Molecular do Centro de Pesquisas René Rachou/FIOCRUZ sob a orientação do Dr. Guilherme Oliveira¹ e co-orientação da Dra. Diana Bahia¹. O projeto contou com o suporte financeiro da CAPES, FIOCRUZ e Fogarty-NIH.

¹ – Laboratório de Parasitologia Celular e Molecular, Centro de Pesquisas René Rachou – FIOCRUZ, Belo Horizonte, MG, Brasil.

Ministério da Saúde
Centro de Pesquisas René Rachou
Programa de Pós-graduação em Ciências da Saúde

**Detecção de polimorfismo de base única em
etiquetas de seqüências expressas de
*Schistosoma mansoni***

apresentada por

Mariana Crivellari Machado Simões

foi avaliada pela banca examinadora composta pelos seguintes membros:

Dra. Glória Franco
Dr. Luciano Andrade Moreira
Dra. Silvane Fonseca Murta

Dissertação defendida em 31 de Maio de 2005.

*Dedico esse trabalho
à minha mãe Silvia e ao meu
pai Henrique pelo exemplo
de vida, apoio, determinação
e amor.*

AGRADECIMENTOS

Agradeço primeiramente ao Dr. Guilherme Oliveira, meu orientador desde a Iniciação Científica, pela confiança, respeito, carinho, apoio, amizade, paciência e pelo oferecimento de inúmeras oportunidades que tanto contribuíram para a minha formação pessoal e profissional.

À Dra. Diana Bahia, minha co-orientadora, por toda amizade, incentivo e cobrança, que foram imprescindíveis durante a minha caminhada.

Ao Dr. Álvaro Romanha por toda disponibilidade e sabedoria.

Ao Dr. Goran Neshich e a Dra. Paula Kuser Falcão, pesquisadores do Núcleo de Bioinformática da EMBRAPA/Unicamp, pela colaboração, boa vontade e disponibilidade durante os experimentos de modelagem molecular e análise estrutural. Aos empregados Fábio, Eduardo e ao pesquisador Adalto pela constante ajuda nas soluções dos problemas.

À Ariane (Universidade de São Paulo-IME) por toda disponibilidade de uso das máquinas.

Ao Centro de Pesquisas René Rachou e ao Dr. Roberto Sena Rocha, atual diretor, pelas facilidades concedidas para o desenvolvimento deste trabalho.

Aos meus pais, exemplo de vida, por todo apoio nos momentos difíceis e pela vibração nos momentos de alegria. Ao Felipe, meu irmão, a minha Luluzinha, Mega e a Aurita por todo carinho!!! Amo muito vocês!!!

Ao Rô, por todo amor, toda amizade e por fazer parte da minha vida a tanto tempo, de uma forma muito especial.

Aos queridos colegas e amigos do Laboratório de Parasitologia Celular e Molecular, Nilton, Regina, Flávio, Elisângela, Bernardo, Fernanda Freire, Fernanda Barbosa, Andréa Carla, Luiza, Lívia, Talita, Silvane, Marcela, Maureen, Hélida, Núbia, Thaís, Rosana e Solange (Imuno), pelos agradáveis momentos que passamos no laboratório. A Silvia e a Kênia que tanto contribuíram na limpeza do laboratório e organização dos materiais, facilitando muito o nosso trabalho.

Ao pessoal da bioinformática, François, Kleider, Anderson, Rômulo e principalmente ao Adhemar, excelente colaboração, por ter passado horas do final de semana rodando programas e solucionando problemas.

Ao Dr. Rodrigo Corrêa de Oliveira e ao pessoal do Laboratório de Imunologia, pela amizade e pela constante disponibilidade. Em especial a Clari.

Ao pessoal do Laboratório de Malacologia, Dr. Omar Carvalho, Roberta, Paula, Liana, Larissa e ao meu grande amigo Ronaldo pelo apoio e colaboração.

À Paola, “my english teacher”, pelas conversas e pelo apoio durante essa última etapa.

Ao Segemar, bibliotecário, pelos inúmeros artigos disponibilizados.

Ao pessoal do setor administrativo e da Pós-graduação do Centro de Pesquisas que sempre procurou viabilizar com eficiência e boa vontade as questões burocráticas relacionadas a este trabalho, em especial ao Paulo.

As minhas queridas amigas que estiveram comigo durante todo esse tempo proporcionando muita alegria, Júlia, Lorena, Inês e aos meus colegas de pós-graduação, em especial às minhas duas grandes amigas Fernanda Ludolf e Ana Carolina Campi, que acompanharam e apoiaram cada detalhe dessa etapa.

Às agências que financiaram o projeto, CNPq e Forgy-NIH e a CAPES, pelo fornecimento da bolsa de estudo.

ÍNDICE

	PÁGINA
LISTA DE TABELAS.....	iii
LISTA DE FIGURAS.....	iv
LISTA DE ABREVIATURAS.....	vii
RESUMO.....	x
ABSTRACT.....	xi
I – INTRODUÇÃO.....	01
1.1 – Aspectos gerais da esquistossomose.....	02
1.2 – Genoma de <i>S. mansoni</i>	07
1.3 – Seqüenciamento Genômico de <i>S. mansoni</i>	08
1.4 – Estudo da variabilidade genética de <i>S. mansoni</i>	10
1.5 – Polimorfismos de base única ou SNPs.....	12
II – JUSTIFICATIVA & OBJETIVOS.....	18
2.1 – Justificativa.....	19
2.2 – Objetivo geral.....	20
2.3 – Objetivos específicos.....	20
III – MATERIAIS & MÉTODOS.....	21
3.1 – Detecção de SNPs <i>in silico</i>	22
3.2 – Validação Experimental.....	30
3.2.1 – Amostra biológica e extração de RNA.....	31
3.2.2 – Amplificação do gene da Catepsina B de <i>S. mansoni</i>	32
3.2.3 – Eletroforese.....	33
3.2.4 – Clonagem e Seqüenciamento.....	33
3.3 – Modelagem por Homologia.....	36
3.3.1 – Identificação e seleção de proteínas modelo.....	36

3.3.2 – Alinhamento das seqüências de resíduos de aminoácidos.....	36
3.3.3 – Construção do modelo (geração das coordenadas cartesianas).....	37
3.3.4 – Análise dos modelos PDB.....	38
IV – RESULTADOS.....	39
4.1 – Detecção de SNPs <i>in silico</i>	40
4.2 – Validação de SNPs no gene da Catepsina B de <i>S. mansoni</i>	50
4.3 – SNPs em cepas de campo de <i>S. mansoni</i>	57
4.4 – Modelagem da Catepsina B de <i>S. mansoni</i>	58
V – DISCUSSÃO.....	80
VI – CONCLUSÕES.....	89
VII – ENDEREÇOS ELETRÔNICOS.....	91
VIII - REFERÊNCIAS BIBLIOGRÁFICAS.....	93
IX – APÊNDICE.....	114
Experimento Modelo: cSNPer	115

LISTA DE TABELAS

PÁGINA

Tabela I	Iniciadores utilizados na reação de PCR no gene da catepsina B de <i>S. mansoni</i>	32
Tabela II	Genes que apresentaram o maior número de polimorfismos.....	48
Tabela III	Polimorfismos nos genes candidatos à vacina, na seqüência consenso.....	49
Tabela IV	Frequência de SNPs em cepa de laboratório e isolados de campo.....	57

LISTA DE FIGURAS

	PÁGINA
Figura 1	Ovos das diferentes espécies de <i>Schistosoma</i> 04
Figura 2	Ciclo de vida de <i>Schistosoma mansoni</i> 05
Figura 3	Representação esquemática da presença de SNPs em 2 alelos diferentes..... 14
Figura 4	Parâmetros de qualidade utilizados pelo novo programa, cSNPer, na identificação dos SNPs..... 24
Figura 5	Exemplo do arquivo de saída “resultados.txt” do programa cSNPer..... 26
Figura 6	Exemplo do arquivo de saída “r_de_qualidade1.csv.” do programa cSNPer..... 27
Figura 7	Exemplo do arquivo de saída “saída.csv” contendo a análise geral dos SNPs identificados..... 28
Figura 8	Exemplo do arquivo de saída “frame.csv.” do programa cSNPer..... 29
Figura 9	Estratégia utilizada na detecção de SNPs experimentalmente..... 35
Figura 10	Janela de visualização dos agrupamentos pelo Consed..... 42
Figura 11	Gráfico da variação do número de SNPs nos agrupamentos..... 43
Figura 12	Número de SNPs por agrupamento e a profundidade do agrupamento..... 44
Figura 13	Frequência dos SNPs identificados <i>in silico</i> , de acordo com a mudança da base nucleotídica, em transição ou transversão..... 45
Figura 14	Classificação dos SNPs de acordo com a posição no códon..... 46
Figura 15	Classificação dos SNPs em sinônimos e não-sinônimos..... 47

Figura 16	Desenho esquemático representando o gene da catepsina B de <i>S. mansoni</i>	51
Figura 17	Detecção de SNPs no gene da catepsina B de <i>S. mansoni</i> por sequenciamento.....	52
Figura 18	Frequência dos SNPs no gene da catepsina B de <i>S. mansoni</i> de acordo com a mudança da base nucleotídica, em transição ou transversão.....	53
Figura 19	Classificação dos SNPs de acordo com a posição no códon.....	54
Figura 20	Classificação dos SNPs em sinônimos ou não-sinônimos, dependendo da conservação do aminoácido codificado.....	55
Figura 21	Identificação do sítio de clivagem do peptídeo sinal no gene da catepsina B de <i>S. mansoni</i>	56
Figura 22	Alinhamento (Formato PIR) gerado pelo programa Modeller da procatepsina B modelo e da catepsina B de <i>S. mansoni</i>	59
Figura 23	Estrutura tridimensional da catepsina B de <i>S. mansoni</i> , na conformação inativa.....	60
Figura 24	Gráfico do módulo Ramachandram do programa Sting. Este gráfico é referente à estrutura da catepsina B de <i>S. mansoni</i> gerado por modelagem comparativa.....	61
Figura 25	Superposição da estrutura tridimensional da proteína modelo e da proteína contendo o SNP 2 (mudança nucleotídica G-A), que resulta na substituição do aminoácido Glu(E)27Lys(K).....	63
Figura 26	Aproximação da imagem dos aminoácidos presentes na estrutura tridimensional da proteína original e da proteína contendo a substituição (SNP 2).....	64
Figura 27	Gráficos gerados pelo módulo <i>Graphical Contacts</i> do programa Sting para o SNP 2.....	65

Figura 28	Interações e a distância do aminoácido Glu 27 (em branco) com os aminoácidos da sua vizinhança de acordo com o <i>Graphical Contacts</i> do programa Sting.....	66
Figura 29	Superposição da estrutura tridimensional da proteína modelo e da proteína contendo o SNP 4 (mudança nucleotídica T-A), que resulta na substituição do aminoácido Asp(D)84Glu(E).....	67
Figura 30	Aproximação da imagem dos aminoácidos na estrutura tridimensional da proteína original e proteína contendo a substituição (SNP 4).....	68
Figura 31	Gráficos gerados pelo módulo <i>Graphical Contacts</i> do programa Sting para o SNP 4.....	69
Figura 32	Módulo ConSSeq: conservação do SNP 4.....	70
Figura 33	Alinhamento do modelo da catepsina B mostrando, também, a estrutura secundária do modelo.....	72
Figura 34	Aproximação da imagem contendo a substituição Asn(N)92Ser(S).....	73
Figura 35	Gráficos gerados pelo módulo <i>Graphical Contacts</i> do programa Sting para o SNP 5.....	74
Figura 36	Módulo ConSSeq: conservação do SNP 5.....	75
Figura 37	Superposição da estrutura tridimensional da proteína modelo e da proteína contendo o SNP 6 (mudança nucleotídica G-C) que resulta na substituição do aminoácido Gly (G)101Arg (R).....	76
Figura 38	Aproximação da imagem contendo os aminoácidos da estrutura da proteína original e da proteína contendo a substituição.....	77
Figura 39	Ligação da glicina 101 com o triptofano 74.....	78
Figura 40	Figuras geradas pelo módulo <i>Graphical Contacts</i> do programa Sting para o SNP 6.....	79

LISTA DE ABREVIATURAS

Å	Angstrom
A.L.F	Seqüenciador Automático Fluorescente – “Automated Laser Fluorescence”
AP-PCR	Reação em cadeia da polimerase com iniciadores arbitrários – “Arbitrarily Primed Polymerase Chain Reaction”
BACs	Cromossomos Artificiais de Bactérias
BLAST	Ferramenta de Alinhamento Local – “Basic Local Alignment Search Tool”
cDNA	DNA Codificante
cm	Centímetro
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CsCl	Cloreto de Sésio
CPqRR	Centro de Pesquisas René Rachou
CGAP	Projeto de Anatomia do Genoma do Câncer
COI	Citocromo Oxidase I
DALYS	Disability Adjusted Life Years
dbSNP	Banco de Dados de Polimorfismos de Base Única
ddNTPs	Dideoxynucleotídeos Tri Fosfato
DNA	Ácido Desoxirribonucléico
dNTP	Desoxirribonucleotídeos Tri Fosfato
DHPLC	Cromatografia Líquida de Alta Pressão Desnaturante
EST	Etiqueta de Seqüência Expressa – “Expressed Sequence Tag”
EDTA	Ácido Etilenodiaminotetracético
FAPEMIG	Fundação de Amparo à Pesquisa do Estado de Minas Gerais
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FIOCRUZ	Fundação Oswaldo Cruz
GST	Glutathione-S-Transferase
HCl	Ácido Clorídrico
INDELS	Inserções/Deleções
Kb	Kilobases, 10^3 pares de bases
KCl	Cloreto de Potássio
M	Molar
mA	Mili-Amperagem
Mb	Megabases, 10^6 pares de bases

MCT	Ministério da Ciência e da Tecnologia
MG	Minas Gerais
MgCl ₂	Cloreto de Magnésio
MGEs	Elementos Genéticos Movéis
min	minutos
ml	Mililitros, 10 ⁻³ litros
mM	Milimolar, 10 ⁻³ Molar
mRNA	RNA mensageiro
mtDNA	DNA mitocondrial
NQS	Padrão de Qualidade da Vizinhaça – “Neighbourhood Quality Standard”
ONSA	Organization for Nucleotide Sequencing and Analysis
µg	Microgramas, 10 ⁻⁶ gramas
µl	Microlitros, 10 ⁻⁶ litros
µM	Micromolar, 10 ⁻⁶ Molar
NCBI	Centro Nacional para Informações em Biotecnologia National – “Center for Biotechnology Information”
ng	Nanogramas, 10 ⁻⁹ gramas
NIH	Institutos Nacionais de Saúde – “National Institutes of Health”
ORF	Janela de Leitura Aberta – “Open Read Frame”
pb	Pares de base
PCR	Reação em cadeia da polimerase – “Polymerase Chain Reaction”.
PDB	Banco de Dados de Proteínas – “Protein Database”
pH	Potencial Hidrogeniônico
pmol	Picomoles, 10 ⁻¹² moles
PRE	Elemento repetitivo polimórfico – “Polymorphic Repetitive Element”
QTLs	Locus de Trato Quantitativo – “Quantitative Trait Locus”
rRNA	RNA ribossomal
RFLP	Polimorfismo de tamanho de fragmento de restrição - “Restriction Fragment Length Polymorphism”.
RNA	Ácido Ribonucléico
SAGE	Análise Seriada de Expressão Gênica – “Serial Analysis of Gene Expression”
SNP	Polimorfismos de Base Única – “Single Nucleotide Polymorphisms”
Taq	<i>Thermus aquaticus</i>

TBE	Tris Borato EDTA
TDR	Programa Especial para Pesquisa e Treinamento em Doenças Infecciosas – “Special Programme for Research and Training in Tropical Diseases”
TE	Tris EDTA
TIGR	Instituto de Pesquisa Genômica – “The Institute for Genomic Research”
TPI	Triose Fosfato Isomerase
Tris	Tri-Hidroximetil Amino Metano
U	Unidade
WTSI	Instituto Sanger da Well-Trust – “Wellcome Trust Sanger Institute”
WHO	Organização Mundial da Saúde – “World Health Organization”
YACS	Cromossomos Artificiais de Leveduras

RESUMO

A aplicação de marcadores moleculares tem sido útil na investigação genética de parasitos de importância médica e seus hospedeiros. A correlação de tipos genéticos com dados clínicos tem sido realizada com o uso de marcadores moleculares polimórficos. Marcadores tipo SNPs (polimorfismos de base única) são o tipo mais comum de variação de seqüência encontrados em eucariotas. Sabe-se que os SNPs são marcadores dialélicos altamente estáveis, cuja estimativa de freqüência pode ser realizada com relativa facilidade. Não obstante, até o presente trabalho, SNPs não foram descritos em larga escala no parasito *Schistosoma mansoni*. Nesse contexto, nosso principal objetivo foi identificar polimorfismos de base única em etiquetas de seqüências expressas (ESTs) de *S. mansoni*. As 61.002 ESTs utilizadas na detecção de SNPs foram geradas pela Rede Genoma de Minas Gerais, através da técnica de seqüenciamento parcial de cDNAs. As seqüências foram processadas por um sistema automatizado que foi desenvolvido a partir do uso de programas públicos (PHRED/CROSS_MATCH/PHRAP/CONSED) e um novo programa de busca por SNPs, cSNPer. O cSNPer foi escrito na linguagem C pelo nosso grupo e identifica polimorfismos em arquivos do tipo .ace gerados por programas de agrupamento, como o Phrap, e analisa a consequência da mutação para a proteína codificada. Foram identificados 2.303 possíveis SNPs em 863 agrupamentos, sendo utilizado o valor de qualidade nas ESTs de Phred ≥ 20 , na seqüência consenso de Phred ≥ 40 e nas bases vizinhas de Phred ≥ 15 . Foi analisada a presença de SNPs em genes candidatos à vacina, nas quais Sm14, Sm23, catepsina B e GST apresentaram-se polimórficos. Em seguida, foi realizada a validação da presença de SNPs no gene da catepsina B. A catepsina B é responsável por degradar hemoglobina presente no sangue do hospedeiro humano, que é a principal fonte de nutrição do *Schistosoma*. Foi identificada a presença de 16 SNPs na seqüência da catepsina B, sendo que 6 resultaram na mudança do aminoácido codificado. Com o propósito de verificar se a presença dessas mutações não-sinônimas poderiam estar modificando a estrutura da proteína e, possivelmente, alterando a sua função, foi realizada a modelagem por homologia da catepsina B do parasito, utilizando o programa Modeller. As análises dos aminoácidos na estrutura modelo e na estrutura mutante foram realizadas utilizando o programa STING. Foi observado que as mudanças de aminoácido encontradas, provavelmente, não alteram a estrutura da proteína e, conseqüentemente, a sua função.

ABSTRACT

Molecular markers have shown to be useful in genetic investigations on parasites of medical importance and their hosts. A correlation between genetic types and clinical data has been established by using polymorphic molecular markers. Single nucleotide polymorphisms (SNPs) molecular markers are the most common kind of sequence change found in eucaryotes. It is well known that SNPs are highly stable diallelic markers, whose frequency estimate may be easily made. Nevertheless, up to the present investigation, such molecular markers have not been studied in *Schistosoma mansoni*. Within this context, we aimed at identifying single nucleotide polymorphism in expressed sequence tags (ESTs) of *S. mansoni*. A total of 61.002 ESTs were generated by the *Schistosoma mansoni* Genome Project of the State of Minas Gerais by partially sequencing cDNAs using the ESTs classic strategy used for detecting SNPs. The sequences were processed by an automated system developed using PHRED/CROSS_MATCH/PHRAP/CONSED and a new program of search for SNPs, cSNPer. cSNPer was written in C language by our group to identify polymorphisms in ace files generated by clusters tools such as Phrap, and also the consequence of the mutation on the codificate protein . A total of 2.303 possible SNPs were identified in 863 clusters using a quality value of Phred ≥ 20 for ESTs, Phred ≥ 40 for the consensus sequence and Phred ≥ 15 for neighbouring bases (NQS). The presence of SNPs was analyzed in vaccine candidate genes, which are polymorphic: Sm14, Sm23; catepsin B and GST. We validated the SNPs in the catepsin B gene. This enzyme is responsible for degrading hemoglobin in human host's blood, which is the main nutrition source for *Schistosoma*. A total of 16 SNPs were identified in catepsin B, 6 which resulted in non-synonym aminoacid change. Aimed at verifying whether the presence of such non-synonym mutations could be modifying the protein structure, and, consequently, its function, a model based on homology of the parasite catepsin B was produced using the software Modeller. Structure analysis was carried out with the software STING. The results obtained indicate that the aminoacid sequence changes probably did not alter the protein structure, and consenquently its function.

I - INTRODUÇÃO

1.1 - Aspectos gerais da esquistossomose

A esquistossomose é uma doença parasitária, crônica, debilitante e em alguns casos fatal, que afeta principalmente indivíduos em áreas rurais, sendo endêmica em países tropicais e subtropicais (WHO 2001). A esquistossomose é hoje endêmica em, aproximadamente, 76 países subdesenvolvidos (Engels *et al.*, 2002). A doença foi introduzida no Brasil, no período colonial, com a vinda de escravos africanos (Files 1951). Apesar dos esforços para controlar essa endemia, a esquistossomose permanece ainda como uma grande causa de morbidade, afetando aproximadamente 200 milhões de pessoas em todo o mundo, sendo 85% dos casos na África (Chitsulo *et al.*, 2000). De acordo com a Organização Mundial de Saúde, o crescimento da doença calculado pelo DALYS é 1.760.000 vezes maior que doenças como Tripanossomíase Africana (1.598.000), Dengue (653.000), Chagas (649.000) e Lepra (177.000) (<http://www.who.int/tdrdisease/default.htm>). No Brasil, estima-se que existam 8 milhões de pessoas infectadas atingindo quase todos os estados brasileiros, principalmente nas regiões Nordeste, Sudeste e Centro-oeste (Oliveira *et al.*, 2004).

Várias espécies do gênero *Schistosoma* são importantes parasitos humanos: *S. mansoni*, *S. japonicum*, *S. haematobium*, *S. intercalatum*, *S. mekongi*, *S. matthei* e *S. malayensis*, sendo as três primeiras as mais relevantes. Em geral, as espécies que infectam o homem são facilmente identificadas através do tamanho e morfologia do ovo, a origem geográfica do isolado e a especificidade pelo hospedeiro intermediário (McManus & Hope 1993). Os ovos de *S. mansoni* possuem uma espinha lateral, os ovos das espécies *S. haematobium* e *S. intercalatum* possuem espinha terminal, enquanto a espécie *S. japonicum* possui ovos sem espinha (Figura 1). Das várias espécies conhecidas de *Schistosoma*, sabe-se que *S. mansoni* é a espécie com maior distribuição global e a única espécie causadora da esquistossomose no Brasil (Bergquist 2002). Os principais sintomas da doença são febre, dor de cabeça, apatia por hepatoesplenomegalia, podendo até causar a morte (Van der Werf *et al.*, 2003).

Os parasitos do gênero *Schistosoma*, pertencem ao Filo Platyhelminthes e Classe Trematódeas. São digenéticos, apresentam dimorfismo sexual na fase adulta e possuem o corpo achatado dorso-ventralmente. A fêmea mede cerca de 1,5 cm e possui o tegumento liso. O macho mede cerca de 1 cm, tem o tegumento coberto por tubérculos e espinhas, e um canal ginecóforo para albergar a fêmea e fecundá-la. A fêmea não é capaz de completar a sua maturação sem o acasalamento com o parasito macho. O macho possui um mecanismo

desconhecido que é capaz de regular a expressão de genes na fêmea (Kunz 2001). Foi visto que a fêmea necessita do macho não só no processo de fertilização, mas também na estimulação de fatores imprescindíveis para o seu crescimento e desenvolvimento (Grevelding *et al.*, 1997; Kunz 2001).

A transmissão da doença ocorre pelo contato do homem com águas onde existam moluscos infectados. O ciclo biológico de *Schistosoma mansoni* apresenta uma alternância de gerações entre o hospedeiro intermediário, moluscos do gênero *Biomphalaria spp.*, e os hospedeiros definitivos vertebrados, dentre eles o homem. O ciclo de vida de *S. mansoni* inicia-se quando as fezes de indivíduos contaminados, contendo ovos do parasito, entram em contato com a água doce. Os ovos, em contato com a água, eclodem, liberando miracídios, que são a forma infectante do hospedeiro invertebrado. Os miracídios infectam novos caramujos e cada miracídio se transforma em esporocisto I. Cada esporocisto I, por poliembrionia, origina 150 a 200 esporocistos II, que migram para as glândulas digestivas e ovoteste do caramujo, originando as cercárias que serão liberadas na água. A cercária, forma infectante do hospedeiro vertebrado, infecta o homem por penetração ativa na pele. Ao penetrar na pele, as cercárias perdem a cauda, transformando-se em esquistossômulos. Os esquistossômulos migram para os pulmões cerca de 7 dias após a penetração e, posteriormente, para o sistema porta hepático. Após a maturação, aproximadamente 45 dias após a infecção, os vermes adultos se alojam no plexo mesentérico e vivem por vários anos, podendo viver até 20 anos, no hospedeiro definitivo (Coelho 1970). O ciclo se completa com a postura de ovos pela fêmea (Figura 2), aproximadamente 300 ovos por dia (Pellegrino & Coelho 1978; Valadares *et al.*, 1981). A grande parte dos ovos é eliminada junto às fezes. Contudo, alguns ainda ficam retidos na mucosa intestinal e nos capilares do sistema porta do hospedeiro, onde desencadeiam uma reação inflamatória granulomatosa. A reação granulomatosa que se forma ao redor dos ovos é a principal causa da patogenia da esquistossomose. A reação granulomatosa resulta em fibrose do tecido. A fibrose dos órgãos e a obstrução do plexo venoso podem levar à hipertensão portal, hepatomegalia, esplenomegalia, ascite (aumento do volume abdominal) e formação de varizes esofágicas.

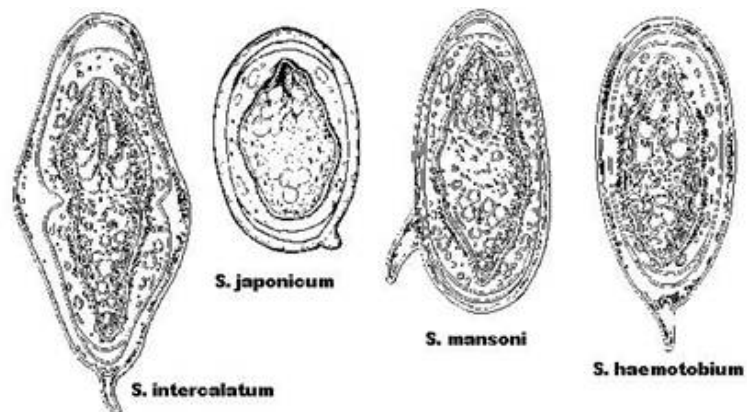


Figura 1 – Ovos das diferentes espécies de *Schistosoma*. O tamanho e a morfologia do ovo variam de acordo com cada espécie.

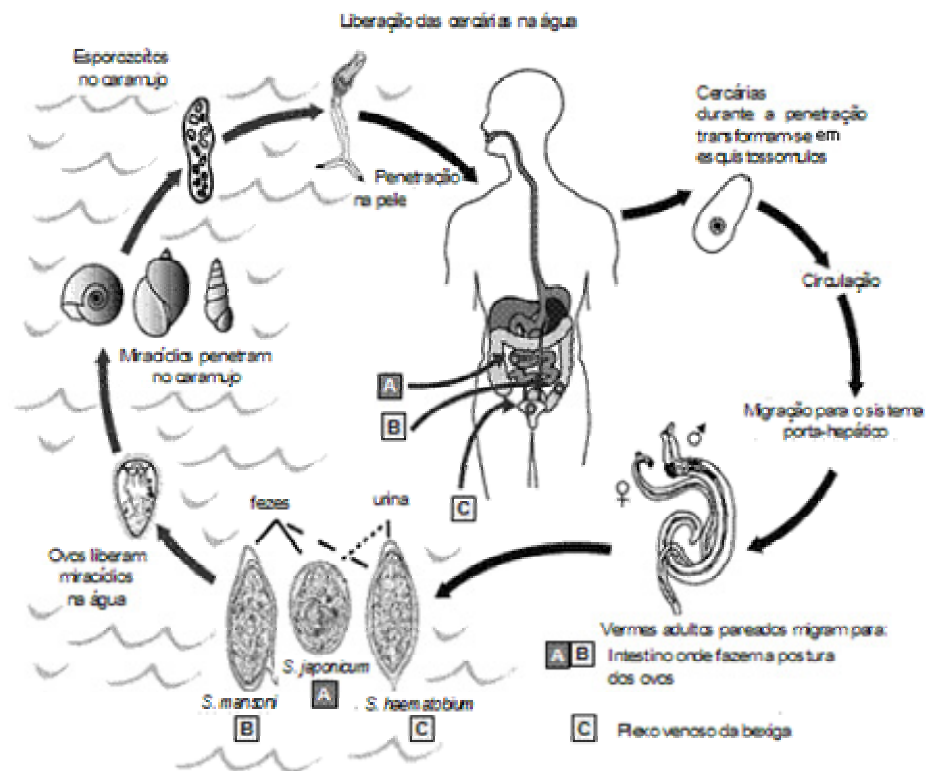


Figura 2 – Ciclo de vida de *Schistosoma*. O ciclo biológico de *Schistosoma mansoni* apresenta uma alternância de gerações entre o hospedeiro intermediário, moluscos do gênero *Biomphalaria spp.*, e os hospedeiros definitivos vertebrados, dentre eles o homem (WHO 2001).

Oxaminiquine e Praziquantel são as drogas mais utilizadas no tratamento da esquistossomose mansoni. O Oxaminiquine aumenta a mobilidade do parasito (Hillman & Senft 1975) e inibe a síntese de ácido nucléico (Pica-Mattoccia *et al.*, 1989). A droga é mais efetiva em parasitos machos do que em parasitos fêmeas. Contudo, apesar de ter apresentado 96% de cura parasitológica, quando realizada biópsia retal o tratamento apresentou apenas 38,3% de cura (Cunha 1982). Atualmente, o medicamento mais efetivo contra o verme é o Praziquantel, que tem contribuído para a diminuição da morbidade em áreas endêmicas (Chitsulo *et al.*, 2000; Kheir *et al.*, 2000; Ferrari *et al.*, 2003). O Praziquantel é menos tóxico e afeta principalmente parasitos fêmeas, causando uma alteração no seu tegumento (Redman *et al.*, 1996) e uma redução no nível de concentração de glutathione (Ribeiro *et al.*, 1998). Um dos possíveis alvos do Praziquantel é um canal de cálcio existente nas células da superfície do parasito, gerando um descontrole no fluxo de íons para dentro e para fora delas (Cioli & Pica-Mattoccia 2003), levando à morte do parasito. Apesar da disponibilidade de drogas de ação rápida, administradas em dose única por via oral, a situação da esquistossomose ainda se apresenta bastante grave (Katz *et al.*, 1989). A prevalência da doença permanece inalterada em muitas regiões endêmicas, devido, na maioria dos casos, aos altos níveis de reinfecção de indivíduos já tratados, seguido do possível aparecimento de populações de parasitos naturalmente resistentes ao tratamento (Bennett *et al.*, 1997). O desenvolvimento de uma vacina contra o parasito seria de grande importância no controle da endemia.

Na tentativa de encontrar possíveis candidatos à vacina, genes foram selecionados e estudados como a glutathione-S-transferase (GST) (Rao *et al.*, 2003), paramiosina (Al-Sherbiny *et al.*, 2003), IrV-5 (miosina) (Nascimento *et al.*, 2002), a triose fosfato isomerase (TPI) (Reynolds *et al.*, 1994), o antígeno de membrana 23 kDa (Sm23/MAP3) (Da'dara *et al.*, 2002), o antígeno de membrana 14 kDa (Sm14) (Fonseca *et al.*, 2005) e a enzima catepsina B (Noya *et al.*, 2001). Não obstante, apesar de grandes avanços, nenhum resultado efetivo foi ainda alcançado. Assim, desde 1993, a Organização Mundial de Saúde (WHO) induziu o estudo da genômica como uma nova abordagem para o desenvolvimento de novas ferramentas de controle do parasito. O estudo da genômica do parasito gera expectativa para um melhor entendimento da biologia, do metabolismo destes organismos e também na busca de novos candidatos para diagnóstico, produção de vacinas e, principalmente, novos alvos de drogas (Degraeve *et al.*, 2001).

1.2 - Genoma de *Schistosoma mansoni*

S. mansoni é um organismo diplóide que possui oito pares de cromossomos (Short & Menzel 1960, Short *et al.*, 1979), sendo 7 pares autossômicos e 1 par sexual. Telômeros típicos foram identificados pela técnica de FISH nestes cromossomos (Hirai & LoVerde 1996). O sexo heterogamético é a fêmea (ZW), enquanto o macho é homogamético (ZZ) (Short & Grossman 1981). O tamanho do genoma haplóide compreende 270 Mb, constituído de um conteúdo A+T bastante elevado (~66%) (Hiller 1974; Simpson *et al.*, 1982; Marx *et al.*, 2000). A presença da metilação não foi observada no DNA do parasito (Fantappie *et al.*, 2001). Com base no tamanho do seu genoma e posição evolutiva, estima-se que *S. mansoni* possua cerca de 15 a 20 mil genes (Simpson *et al.*, 1982). O genoma é constituído 4% a 8% de seqüências de DNA altamente repetitivas (>1000 cópias), 35% a 40% de seqüências de DNA de média repetitividade (~100 cópias) e aproximadamente 60% representam famílias de genes ou regiões de cópia única (Simpson *et al.*, 1982).

A primeira região de DNA repetitivo descrita em *S. mansoni* consistiu do complexo gênico que codifica o RNA ribossomal (rDNA) (Simpson *et al.*, 1984; Van Keulen *et al.*, 1985). Este complexo é constituído de unidades repetitivas em tandem, presentes em torno de 100 unidades por genoma haplóide. Cada unidade contém aproximadamente 10 Kb que codificam três espécies altamente conservadas de rRNA em eucariotas: 5,8S, 18S, e 28S, cujas regiões codificadoras são separadas por uma região menos conservada, não transcrita denominada DNA espaçador (Simpson *et al.*, 1984; Van Keulen *et al.*, 1985). Várias seqüências de DNA repetitivo descritas em *S. mansoni* (Hamburger *et al.*, 1991), distribuídas de maneira dispersa ou em arranjos pelo genoma, foram encontradas em transcritos de mRNA do parasito, similares a regiões hipervariáveis no genoma humano (Spotila *et al.*, 1991). Um exemplo é o clone de cDNA SM750, contendo elemento repetitivo polimórfico (PRE) de 62 pb, que foi encontrado em diferentes tamanhos de transcritos e seqüências. Seqüências repetitivas também apresentaram utilidade em ensaios de hibridização na determinação do sexo do parasito (Spotila *et al.*, 1987). Algumas, ultrapassando 500 cópias, foram encontradas somente no cromossomo W das fêmeas (Spotila *et al.*, 1989; Oliveira & Bahia 2004).

Os avanços genômicos foram seguidos da descoberta e caracterização dos MGEs (elementos genéticos móveis) em helmintos, incluindo a identificação de retrotransposons em *S. mansoni*. Elementos SINE-like, como as famílias *sma*-like, representam, aproximadamente, 100 cópias por genoma haplóide. O elemento SR1 foi o primeiro retrotransposon identificado

em Platelmintos, apresentando mais de 200 cópias por genoma (Drew & Brindley 1997). Os MGEs têm sido utilizados em outros sistemas para estudos epidemiológicos e clínicos, associados a fatores de virulência (Hide & Tilley 2001), análises filogenéticas e em estudos de evolução gênica (Brindley *et al.*, 2003).

1.3 - Seqüenciamento Genômico de *S. mansoni*

Para realizar o seqüenciamento da porção expressa do genoma de *S. mansoni*, foi adotada a técnica de geração de etiquetas de seqüências expressas ou ESTs. As ESTs são seqüências parciais de genes expressos geradas a partir de bibliotecas de cDNA, através do seqüenciamento de uma das suas extremidade 5' ou 3'. Esta é uma metodologia simples e de custo moderado. A técnica permite trabalhar apenas com seqüências expressas e identificar rapidamente diferentes genes por comparações com seqüências já depositadas em banco de dados. Além disso, as ESTs também são utilizadas na construção de mapas físicos, na caracterização de grandes seqüências genômicas e transcritas, na identificação de novos genes e análises de expressão gênica diferencial (Zweiger & Scott 1997). Não obstante, as ESTs são resultantes de seqüenciamento único, sendo muito comum observar erros de seqüenciamento. As seqüências são, geralmente, curtas (aproximadamente 300 pb) e há alta redundância de dados (Liang *et al.*, 2000; Wang *et al.*, 2004). Na tentativa de superar esses problemas faz-se necessário o agrupamento dessas seqüências, com o objetivo de gerar uma seqüência consenso (Oliveira & Johnston 2001). Uma abordagem alternativa as ESTs convencionais é a geração de etiquetas do tipo Orestes. Esta técnica tem como característica predominante o uso de iniciadores aleatórios. Utilizando a técnica de RT-PCR estes iniciadores, em condições de baixa adstringência, se ligam preferencialmente nas porções centrais do mRNA, aumentando preferencialmente o número de ESTs correspondentes ao centro do gene.

O seqüenciamento do genoma humano criou novas oportunidades e iniciativas de seqüenciamento do genoma de inúmeros organismos, inclusive de vários parasitos, entre eles *S. mansoni*. Os primeiros esforços para o seqüenciamento do transcriptoma de *S. mansoni* foi financiado pela OMS com o objetivo de descobrir novos genes utilizando a estratégia de geração de ESTs (Oliveira 2001). Após seqüenciamento e análise por pesquisas de homologia, as primeiras 607 ESTs do parasito foram geradas (Franco *et al.*, 1995). Contudo, a primeira grande iniciativa de seqüenciamento do transcriptoma de *S. mansoni* ocorreu no ano de 2001 pelo grupo ONSA de São Paulo (Verjovski-Almeida *et al.*, 2003). A iniciativa

recebeu financiamento do programa FAPESP com o objetivo de gerar 120.000 seqüências. Estas foram geradas de bibliotecas das diferentes fases do parasito, através da técnica de Orestes (Dias-Neto *et al.*, 2000). Foram geradas 124.640 ORESTES, formando 30.988 agrupamentos. O total de 23% dos nucleotídeos correspondentes a genes conhecidos e 77% a genes novos (Verjovski-Almeida *et al.*, 2004).

Outra grande iniciativa de seqüenciamento do transcriptoma de *S. mansoni* foi criada no ano de 2002 com a união de vários laboratórios do Estado de Minas Gerais, financiados pela FAPEMIG e MCT/CNPq (rgmg.cpqrr.fiocruz.br). O principal objetivo foi aumentar o número de seqüências disponíveis, possibilitando um conhecimento mais aprofundado dos genes expressos e a chance de descoberta de novos genes. Além disso, como foram geradas ESTs convencionais, estas seqüências serão indispensáveis em estudos de proteoma e SAGE (Oliveira *et al.*, 2004). Como consequência, no ano de 2004 foram geradas novas 61.002 ESTs de bibliotecas enriquecidas das diferentes fases do parasito, ainda não disponibilizadas em banco de dados públicos. Em conjunto, espera-se que os projetos transcriptomas gerem uma cobertura quase completa dos genes deste organismo.

Apesar do tamanho e da complexidade do genoma de *S. mansoni*, uma iniciativa internacional para o seqüenciamento genômico do parasito está sendo conduzida pelo TIGR em associação com o Wellcome Trust Sanger Institute (WTSI), com financiamento do NIH e do Wellcome Trust (El-Sayed *et al.*, 2004). O projeto iniciou-se com o seqüenciamento de extremidades 5' e 3' de BACs (cromossomos artificiais de bactérias) e YACs (cromossomos artificiais de leveduras), contendo insertos de 100-140 Kb e 358 Kb, respectivamente, de DNA de *S. mansoni*, que foram obtidos a partir da seleção aleatória de clones. A partir do seqüenciamento das extremidades de BACs, 21 Mb de seqüências genômicas descontínuas foram geradas. Estas seqüências foram úteis não somente para os projetos de descoberta gênica, mas também para prover marcadores para a produção de mapa físico em alta resolução (Venter *et al.*, 1996). Em Outubro de 2002, o TIGR e o Sanger iniciaram o seqüenciamento pela técnica de *Shotgun* (digestão do DNA utilizando endonucleases ou por um processo físico-químico) e, atualmente, possuem juntos 9 vezes a cobertura do genoma de *S. mansoni*, com a estimativa de 0,5% ainda não seqüenciado. O acesso à seqüência genômica será de grande utilidade, pois permitirá a identificação de seqüências reguladoras e, em conjunto com os projetos transcriptoma, uma descrição genômica completa do organismo. Os genes transcritos descobertos serão essenciais para uma anotação mais precisa do genoma,

pois podem ser prontamente localizados na seqüência genômica através do uso de ferramentas de bioinformática (Hu *et al.*, 2004).

1.4 - Variabilidade genética de *S. mansoni*

Fatores tais como a variação na intensidade de infecção de caramujos e de pacientes, a resistência a drogas (Katz *et al.*, 1973; Campos *et al.*, 1976; Araújo *et al.*, 1980; Dias *et al.*, 1982; Coles *et al.*, 1987; Drescher *et al.*, 1993; Bennett *et al.*, 1997), o desenvolvimento do parasito no hospedeiro definitivo, produção de ovos e as características da patogenicidade e da imunogenicidade induzida pelo *Schistosoma* são provavelmente, em parte, resultantes da expressão da variabilidade genética do parasito (Rollinson *et al.*, 1986a; McManus & Hope 1993). Desta forma, o conhecimento da biologia molecular e da variabilidade genética intra e inter-específica do parasito são de fundamental importância para a compreensão dos mecanismos biológicos envolvidos nas interações parasito-hospedeiro, na patogenia e epidemiologia da doença.

A técnica eletroforética de isoenzimas foi a primeira metodologia utilizada em estudos de variabilidade genética do *S. mansoni*. O princípio básico da técnica consiste no uso da eletroforese de isoenzimas em gel de amido (Smithies 1955), seguido da visualização do produto por métodos histoquímicos (Hunter & Markert 1957). Fletcher e colaboradores (1981) foram pioneiros no uso da técnica em *Schistosoma*. Esses pesquisadores realizaram estudos medindo diferenças no nível de infectividade de 22 cepas de *S. mansoni*. Utilizaram 14 enzimas diferentes e detectaram apenas 3 *loci* polimórficos dos 18 testados. Em estudo subsequente, LoVerde e colaboradores (1985) utilizaram 34 enzimas na mesma população, observaram outros 3 novos *loci* polimórficos, contudo não estabeleceram uma clara relação entre o nível de infectividade e o polimorfismo isoenzimático encontrado. Posteriormente, Navarro e colaboradores (1992), utilizando a mesma técnica, diferenciaram cepas de diferentes regiões. Foram encontrados 14 *loci* em cepas da Venezuela e 13 *loci* em cepas do Brasil. Contudo, quando analisaram vermes individuais de uma mesma cepa, verificaram que todos *loci* eram monomórficos.

O avanço de métodos de biologia molecular possibilitou o uso de novas ferramentas analíticas. Uma variedade de genes e fragmentos provenientes de DNA genômico foram clonados e caracterizados, notadamente o RNA ribossomal (rRNA). A utilização da técnica de *Southern blotting* possibilitou as primeiras análises em nível molecular, com sondas de rRNA

hibridizadas ao DNA genômico, permitindo estudos filogenéticos e discriminação inter e intraespecífica entre as espécies (Ali *et al.*, 1991; Després *et al.*, 1992; McManus & Hope 1993).

O DNA mitocondrial (mtDNA) tem sido abordado em estudos de variabilidade genética, e mostrou-se polimórfico em diferentes clones em *S. mansoni* (Pena *et al.*, 1995). O tamanho do genoma mitocondrial completo varia de 16.5 a 24.9 kb. Comparando produto de PCR de mtDNA de 6 cepas de *S. mansoni* de diferentes regiões geográficas, foi demonstrada uma variedade significativa de tamanho dos genes de mtDNA devido à presença de elementos repetitivos que variavam entre 2.000 a 10.000 nucleotídeos (Després *et al.*, 1991). Em seguida, através de análises por RFLP de mtDNA foi corroborada a hipótese da esquistossomose ter sido introduzida no Brasil por escravos africanos, através da comparação de populações da América e da África (Després *et al.*, 1993). Dentre 40 sítios específicos, 5 mostraram-se polimórfico. Contudo, apesar de existir trabalhos demonstrando que a taxa de divergência do mtDNA é maior que a taxa de divergência do DNA genômico (Harrison *et al.*, 1989), este genoma não expressa diploidia e nem herança do tipo mendeliana, não sendo, portanto, aplicável em estudos populacionais ou em estudos de evolução gênica (Jannotti-Passos *et al.*, 1997).

Um outro marcador importante utilizado é o microssatélite. Os microssatélites consistem de pequenas seqüências de DNA repetidas em tandem, amplamente distribuídos nos cromossomos de eucariotos (Chalersworth *et al.*, 1994). Devido ao caráter polimórfico, os microssatélites são amplamente usados em análises de variabilidade genética, na diferenciação de indivíduos (Hagelberg *et al.*, 1992; Oliveira *et al.*, 2004), estimativa de fluxos gênicos (Curtis & Minchella 2000) e em estudos populacionais (Shrivastava *et al.*, 2005). No primeiro trabalho publicado utilizando marcadores do tipo microssatélite em *S. mansoni* (Durand *et al.*, 2000), foram detectados 33 *loci* em banco de dados e biblioteca genômica, sendo 11 polimórficos. Em seguida, Blair e colaboradores (2001) descrevem mais 20 marcadores e verificam uma maior diversidade em populações Africanas quando comparados com os resultados descritos por Durand *et al.*, (2000) em Guadalupe. Usando outros marcadores, Rodrigues e colaboradores (2002) observaram uma maior diversidade em cepas brasileiras de campo quando comparadas com cepas mantidas em laboratório, porém essa diversidade maior não foi significativa quando comparada entre diferentes isolados do campo, sugerindo ocorrência de troca genética (fluxo gênico) no campo.

Com o advento do seqüenciamento em larga escala, principalmente no estudo dos genomas funcionais, grande número de seqüências de DNA para genes expressos foram geradas (ESTs) (Picoult-Newberg *et al.*, 1999; Useche *et al.*, 2001; Batley *et al.*, 2003). Esse grande avanço aumentou o interesse no seqüenciamento comparativo, na busca de diferenças genéticas entre indivíduos e permitiu a descoberta de um novo tipo de marcador, os polimorfismos de base única ou SNPs. Os SNPs são hoje descritos como o tipo de variação genética mais freqüente no genoma humano (Sachidanandam *et al.*, 2001) e têm recebido atenção especial em estudos genéticos devido à sua abundância e estabilidade, comparado aos microsatélites (Gray *et al.*, 2000), além de poderem estar diretamente ligados a fenótipos de interesse (Broman *et al.*, 2004; Suh & Vijg 2004).

1.5 - Polimorfismos de base única ou SNPs

Polimorfismos de base única ou SNPs representam uma fonte abundante de variação genética (Figura 3). SNPs são gerados pela substituição de uma única base nucleotídica ou pequenos eventos de inserção ou deleção (INDELS), ocorrendo com uma taxa de mutação muito baixa, de aproximadamente 1×10^{-9} a 5×10^{-9} por nucleotídeo/por ano, a maioria em posições neutras do genoma humano (Martinez-Arias *et al.*, 2001). A grande vantagem dos SNPs em comparação a outros marcadores reside na abundância de polimorfismos entre alelos de um determinado gene e ampla distribuição, podendo estar presentes em, praticamente, todos os *loci* gênicos. No genoma humano, por exemplo, estima-se que existam aproximadamente 10 milhões de SNPs (Lai 2001), ocorrendo numa freqüência de 1 SNP a cada 500-1.000 pb (Collins *et al.*, 1999; Sachidanandam *et al.*, 2001). Já se sabe que os SNPs são marcadores dialélicos altamente estáveis (<1% são trialélicos, <http://snp.cshl.org/>), já que a probabilidade de alteração de duas bases independentes, presentes numa mesma posição, é pequena. Assim, SNPs são capazes de gerar haplótipos com descendência idêntica, o que permite a comparação estatística de haplótipos entre grupos de estudo e controles. Além disso, sua identificação é passível de automação em alta capacidade (Heaton *et al.*, 2001).

SNPs são classificados de acordo com o tipo de variação de nucleotídeo em transições, purina-purina (A/G) ou pirimidina-pirimidina (C/T), e transversões, purina-pirimidina ou pirimidina-purina (A/C, A/T, G/C, G/T) (Brookes 1999). Apesar do número de possibilidades de ocorrer variações do tipo transversões ser o dobro de transições, o contrário tem sido observado por vários pesquisadores (Picoult-Newberg *et al.*, 1999; Smith *et al.*, 2001; Cheng

et al., 2004). Uma provável explicação, geralmente aceita em DNA de eucariotas, é o processo espontâneo de deaminação da 5-metilcitosina (5mC) para timidina (T) (5' CpG 3', onde p é a ligação fosfodiéster 3' para 5' entre nucleotídeos adjacentes), dando origem a um maior número de transições de C > T (G > A, na fita reversa) (Cooper & Krawczak 1989; Fryxell & Moon 2004). Além disso, a metilação é apontada como a possível responsável pelo aumento do processo de deaminação e, conseqüentemente, aumento do número de transições nas seqüências CpG (Tomso & Bell 2003). A deaminação de citosinas não metiladas geram uracilas, que são normalmente removidas pela enzima uracil glicosilase. Contudo, citosinas encontram-se, geralmente, metiladas e a sua deaminação gera timidinas, que não podem ser removidas por esta enzima. Como conseqüência, o nível de mutações correspondentes a 5mC para T varia entre 10 a 50 vezes mais quando comparado às outras variações. Além disso, mais de um terço das mutações que causam doenças genéticas e que, também, somatizam para o desenvolvimento do câncer são causadas pela hipermutabilidade das seqüências CpG (Fryxell & Moon 2004).

SNPs podem ser encontrados tanto em regiões codificadoras como em regiões não codificadoras do genoma, como introns e regiões intergênicas (Wong *et al.*, 2003). Com isso, além de ESTs, seqüências genômicas também podem ser utilizadas para identificação de polimorfismos (Bao *et al.*, 2005). A presença de SNPs no genoma pode ter conseqüência para o modo como o genoma é expresso, podendo causar edição alternativa do mRNA, alterações no padrão de expressão de genes, geração ou supressão de códons de terminação ou poliadenilação na molécula de RNA mensageiro e alteração nos códons de iniciação de tradução (Guimarães & Costa 2002). Contudo, o grande benefício de detectar SNPs em ESTs deve-se ao fato da possibilidade de identificação de mutações que possuam características funcionais (Kim *et al.*, 2003). Mutações presentes em regiões codificadoras são classificadas em sinônimas ou não-sinônimas. Mutações não-sinônimas resultam em uma substituição de aminoácido codificado na seqüência protéica, podendo ser conservativas ou não conservativas em função das características dos aminoácidos envolvidos na troca. Nesses casos, a presença do polimorfismo pode levar a uma mudança estrutural da proteína codificada e, conseqüentemente, a uma possível alteração da sua função (Stitzel *et al.*, 2004). Por outro lado, mutações do tipo sinônimas são aquelas nas quais a presença do polimorfismo não causa alteração do aminoácido codificado. Embora as mutações sinônimas não alterem a seqüência protéica, elas podem modificar a estrutura e a estabilidade do RNA mensageiro e, conseqüentemente, afetar a quantidade de proteína produzida.

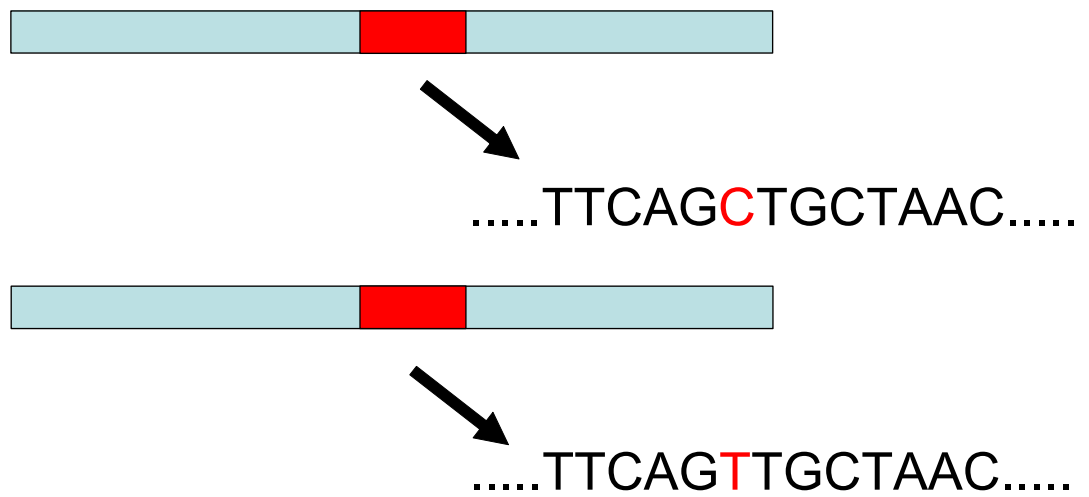


Figura 3 – Representação esquemática da presença de SNPs em 2 alelos diferentes. A cor vermelha representa uma região codificadora e na seqüência nucleotídica, também em vermelho, a presença do SNP (C/T) no respectivo gene.

SNPs vêm sendo utilizados com propósitos forenses (Vallone & Butler 2004), em estudos farmacogenômicos (Riley *et al.*, 2000) e detectado nos mais diferentes tipos de organismos: mamíferos (Heaton *et al.*, 2001, Zhang & Zhao 2004), aves (Kim *et al.*, 2003; Fitzsimmons *et al.*, 2004; Wong *et al.*, 2004), peixes (Hahn *et al.*, 2004), insetos (Berger *et al.*, 2001; Wang *et al.*, 2005), plantas (Ching *et al.*, 2002; Grivet *et al.*, 2003; Somers *et al.*, 2003), fungos (Forche *et al.*, 2005) e outros microrganismos, como vírus (Pavlovic-Lazetic *et al.*, 2004).

Existem dois métodos principais para detectar estes polimorfismos: *in silico* e experimentalmente. O método *in silico* envolve o uso de ferramentas de bioinformática. Com o número crescente de informações, faz-se necessário o desenvolvimento e implementação de novos algoritmos de análise, com o objetivo de facilitar o acesso, uso e gerenciamento de vários tipos de informações. As ferramentas de bioinformática são utilizadas em várias etapas do estudo de polimorfismos, desde a identificação das variações até a predição do efeito delas. Outra aplicação importante é na construção e manutenção de bancos de dados e de ferramentas que podem ser acessadas pela internet. Esses bancos atuam como sedes de referência para a deposição de SNPs, como o *National Center for Biotechnology Information* (NCBI) (dbSNP, <http://www.ncbi.nlm.nih.gov/SNP>), onde já estão depositados mais de 4,1 milhões de SNPs humanos e um número considerável de polimorfismos de outras espécies. Outro exemplo é o ToxoDB, banco de dados contendo informações relativas sobre o genoma do parasito *Toxoplasma gondii* (Kissinger *et al.*, 2003), como ESTs, QTLs, SNPs, microarranjos, proteoma e ferramentas para análises e anotação genômica. Apesar da indiscutível contribuição da bioinformática na identificação de SNPs, a necessidade de validação experimental dos dados é indispensável para se avaliar o potencial informativo de cada polimorfismo. Dessa forma, o surgimento e a evolução de métodos de validação e de genotipagem de amostras para SNPs confirmados também merecem reconhecimento. Hoje, é possível genotipar um grande número de SNPs em várias amostras em um mesmo tubo de reação por *chips* de DNA, ou microarranjos, cromatografia líquida de alta pressão desnaturante (DHPLC), espectrometria de massa, PCR em tempo real e mini-seqüenciamento (Vignal *et al.*, 2002). Entretanto, em análises de menor escala, métodos como seqüenciamento e digestão por enzimas de restrição RFLP são rotineiramente empregados (Mullikin *et al.*, 2000).

Após estudos confirmatórios, SNPs podem ser usados para identificar genes que estão associados à severidade de doenças, resistência a drogas, na identificação de fenótipos de

interesse e por desequilíbrio de ligação (Sachidanandam *et al.*, 2001, Bader 2001). Além disso, também SNPs são úteis na construção de mapas genéticos de alta resolução, diagnóstico genético, análises filogenéticas (Rafalski 2002) ou em estudos de história e genética de populações (Halushka *et al.*, 1999, Weiss 1998; Evans & Relling 1999; Stephens *et al.*, 2001). A possibilidade de se identificar haplótipos adiciona ainda mais valor aos SNPs identificados (Escary *et al.*, 2000), contribuindo no mapeamento de genes susceptíveis a doenças e em estudos de associação (Zhang *et al.*, 2002).

SNPs tornaram-se muito populares em estudos de genética humana sendo usados em pesquisas de doenças complexas, como desordem bipolar (Fridman *et al.*, 2003), câncer (Brentani *et al.*, 2003; Koed *et al.*, 2005), mal de Alzheimer, entre outras (Fridman *et al.*, 2003). Meyer e colaboradores em 2003 detectaram 6 SNPs e 2 haplótipos do gene BRAF significativamente associados ao melanoma (câncer de pele), o que também foi comprovado em outros tipos de câncer, como adenocarcinoma (Naoki *et al.*, 2002) e câncer da tireóide (Kimura *et al.*, 2003). Com o projeto de anatomia do genoma do câncer (CGAP), mais de 10.000 SNPs (Clifford *et al.*, 2000) foram identificados em ESTs humanas e depositados em banco de dados públicos.

Já foram descritos na literatura vários exemplos de identificação de SNPs associados a genes, cSNPs, com o uso de etiquetas de seqüências expressas, ESTs, utilizando ferramentas de bioinformática (Picoult-Newberg *et al.* 1999; Clifford *et al.*, 2000). Essa metodologia utiliza ESTs redundantes existentes em bancos de dados. A comparação de ESTs obtidas de indivíduos diferentes gera resultados similares à busca de polimorfismos em estudo de populações (Gu *et al.* 1998). Com isso, as ESTs tornaram-se, nos últimos tempos, a fonte mais rica para detecção de polimorfismos, devido à redundância de seqüências de genes, a representação de diferentes genótipos depositados em banco de dados e por permitir a associação de SNPs com genes expressos (Barker *et al.*, 2003).

No campo da parasitologia, a presença de SNPs tem sido descrita em alguns parasitos. Podemos citar a identificação de SNPs em diferentes espécies do gênero *Plasmodium* (Feng *et al.*, 2003), com especial referência à espécie *Plasmodium falciparum*, parasito causador da malária em humanos (Purfield *et al.*, 2004; Myrick *et al.*, 2005) e também do seu vetor *Anopheles gambiae* (Morlais *et al.*, 2004), em *Toxoplasma gondii*, parasito intracelular do Filo Apicomplexa, causador da Toxoplasmose em humanos (Su *et al.*, 2004; Peyron *et al.*, 2004) e em *Trypanosoma cruzi*, parasito causador da doença de Chagas (Augusto-Pinto *et al.*, 2003). No campo da parasitologia animal, podemos citar a identificação de SNPs em

Haemonchus contortus, parasito gastrointestinal de ovinos e caprinos (Ruiz *et al.*, 2004). Apesar da identificação de SNPs no genoma de diferentes parasitos, ainda sabe-se muito pouco sobre a real influência desses marcadores e sua consequência fenotípica. Acredita-se que esses marcadores possam ser úteis no estudo de virulência, resistência a drogas e severidade da doença. Até o momento apenas seqüências comumente utilizadas para estudos filogenéticos como COI, DNA ribossomal e mitocondrial foram usados para a verificação de polimorfismos em *Schistosoma*. Entretanto, nos últimos anos, devido às iniciativas de seqüenciamento do genoma de *S. mansoni*, um número relativamente grande de ESTs tornou-se disponível em bancos de dados públicos (Oliveira *et al.*, 2004). Contudo, até o presente trabalho, essa informação não havia sido utilizada para identificação de SNPs em grande escala, apesar do seu potencial uso ser reconhecido. Portanto, existe uma necessidade de produzir este tipo de marcador molecular para *Schistosoma mansoni* e de demonstrar sua utilidade para estudos da esquistossomose.

II - JUSTIFICATIVA & OBJETIVOS

2.1 – Justificativa

No Brasil, a esquistossomose é uma importante endemia parasitária. Estima-se que existam, cerca de 6,3 milhões de portadores da doença (Katz & Peixoto 2000). Focos de esquistossomose já foram descritos em vários estados brasileiros. Contudo, vem se observando o surgimento de novos focos em áreas antes consideradas indenes, como os focos encontrados, recentemente, no Rio Grande do Sul (Graeff-Teixeira *et al.*, 2004). Uma estratégia completa que combine a avaliação simultânea da genética dos organismos e a influência de fatores ambientais será um importante meio para o entendimento dos aspectos biológicos da doença, do parasito e, conseqüentemente, para o desenvolvimento de métodos de controle.

Estudos focalizando a compreensão da biologia do parasito e do tratamento da doença mostram diferenças marcantes entre diferentes populações de parasitos de uma mesma espécie em relação a fatores tais como: tratamento, resistência a drogas, preferência por hospedeiros, patogenicidade e desenvolvimento do parasito no seu hospedeiro definitivo (Drescher *et al.*, 1993, Bennett *et al.*, 1997). Nos parasitos do gênero *Schistosoma*, alguns destes fatores já foram citados como conseqüência da variabilidade genética (Rollinson *et al.*, 1986b; McManus & Hope 1993), e desde então, a aplicação de marcadores moleculares tem sido útil na investigação genética do parasito e da sua relação os com hospedeiros.

As iniciativas de seqüenciamento genômico e o conseqüente aumento do número de seqüências disponíveis em bancos de dados públicos viabilizaram o uso e desenvolvimento de novas ferramentas de análises genéticas. Dentre as possibilidades está o desenvolvimento de novos marcadores baseados em seqüências de genes, em especial os polimorfismos de base única ou SNPs. O uso de SNPs pode ser extremamente útil na identificação de genes responsáveis por fenótipos de interesse (Bader 2001; Fernandez-Mestre *et al.*, 2004). Além disso, SNPs podem ser utilizados para identificar genes associados à severidade da doença ou resistência a drogas, em trabalhos de farmacogenômica, ou em estudos de história e genética de populações (Kleyn & Vesell 1998; Evans & Relling 1999; Lee & Koh 2001).

Não obstante, até o presente momento essa informação não havia sido utilizada para identificação de SNPs em grande escala em *S. mansoni*, apesar do seu potencial uso ser reconhecido. Apenas seqüências comumente utilizadas para estudos filogenéticos foram usadas para a verificação de polimorfismos em *Schistosoma* (Littlewood & Johnston 1995; Barker 1996). Nesse contexto, nos propomos a identificar *in silico* potenciais marcadores moleculares do tipo SNPs para *S. mansoni* e verificar alguns experimentalmente,

disponibilizando informações para futuros estudos de caracterização genética do parasito e o seu uso no desenvolvimento de novos candidatos à vacina e novos alvos de drogas.

2.2 – Objetivo geral

Identificar polimorfismos de base única (SNPs) em etiquetas de seqüências expressas (ESTs) de *Schistosoma mansoni*.

2.3 – Objetivos específicos

- 1 Desenhar uma estratégia de bioinformática para identificar SNPs *in silico* em ESTs de *S. mansoni*.
- 2 Validar o novo programa de busca de polimorfismo, o cSNPer.
- 3 Classificar os possíveis SNPs encontrados em transição/transversão e sinônimos/não-sinônimos.
- 4 Analisar o grau de polimorfismo de genes candidatos à vacina para esquistossomose.
- 5 Determinar e validar a presença de SNPs no gene codificante para a catepsina B.
- 6 Realizar a modelagem por homologia da proteína catepsina B modelo e variantes.
- 7 Analisar o possível efeito das mutações do tipo não-sinônimas na estrutura tridimensional da catepsina B de *S. mansoni*.

III - MATERIAIS & MÉTODOS

3.1 –Detecção de SNPs *in silico*

As 61.002 ESTs utilizadas para detecção de SNPs foram geradas pela Rede Genoma de Minas Gerais pela técnica de seqüenciamento parcial de cDNAs, utilizando a estratégia de ESTs. Estas seqüências foram processadas por um sistema automatizado que foi desenvolvido a partir do uso de programas públicos (PHRED/CROSS_MATCH/PHRAP/CONSED) e um novo programa para identificação de SNPs em ESTs foi desenvolvido pelo nosso grupo, o cSNPer.

Primeiramente, as seqüências foram processadas utilizando o programa de identificação de bases Phred (Ewing & Green 1998; Ewing *et al.*, 1998), que é responsável pela leitura binária dos cromatogramas gerados pelo seqüenciador, convertendo-os em formato texto. Ou seja, Phred atribui uma base nucleotídica referente a cada pico de fluorescência identificado, com uma taxa de erro menor que aquela atribuída pelo programa de identificação de base padrão. O programa usa métodos para examinar os sinais das quatro diferentes fluorescências, na região em volta de cada ponto do conjunto de dados de sinais em função do tempo gerado pelo seqüenciador. Em seguida, ele atribue valores de qualidade as bases (valor de Phred = $q = -10 \times \log_{10}(p)$, onde q é valor de qualidade e p probabilidade estimada de erro de uma base), baseando-se na estimativa da taxa de erro e resolução do pico que foi calculado para cada base individualmente. A linha de comando utilizada para a execução do programa foi: `phred -id . -as RGMG.fasta -qa RGMG.fasta.qual -pd RGMG.phd`. O programa Phred gerou os arquivos de saída contendo as seqüências, os valores de qualidade das seqüências no formato FASTA e um arquivo PHD, utilizado, posteriormente, para visualização dos arquivos.

Em seguida, as seqüências foram mascaradas contra uma biblioteca específica contendo seqüências de vetores. Para isso utilizamos o programa Cross_Match através da linha de comando: `cross_match RGMG.fasta maskfile.lib -minmatch 10 -minscore 20 -screen`, que utiliza a implementação do algoritmo de Smith-Waterman-Gotoh (Smith & Waterman 1981; Gotoh 1982), substituindo a base original por um x. Foi criado um arquivo com a extensão .fasta.screen contendo as seqüências mascaradas e o arquivo de qualidade foi renomeado para .fasta.screen.qual.

O próximo passo foi realizado o agrupamento das ESTs através do programa Phrap (<http://www.phrap.org>), utilizando a linha de comando: `phrap RGMG.fasta.screen -minmatch 20 -qual_show 20 -view`. O arquivo de entrada utilizado foi o arquivo contendo as seqüências

mascaradas pelo Cross_Match e o arquivo de qualidade das seqüências gerado pelo Phred. Foi considerado para o agrupamento das seqüências o valor de qualidade de phred ≥ 20 (-qual_show 20). O programa Phrap comparou as seqüências entre si (*minmatch 20*, ou seja, no mínimo 20 pb alinhadas) de forma a encontrar quais delas eram similares ou continham regiões similares o bastante para serem agrupadas. Assim, o programa gerou um arquivo de saída contendo um arquivo multifasta das seqüências consensos ou contigs. As ESTs que não apresentaram homologia com nenhuma outra foram reunidas em um outro arquivo e designadas como seqüências únicas ou *singlets*. Além disso, o Phrap gerou um arquivo no formato ACE contendo os contigs e suas respectivas qualidades e as ESTs correspondentes a cada contig. Este arquivo serviu como entrada para o programa de detecção de polimorfismos.

A detecção dos polimorfismos foi realizada utilizando um novo programa de busca de SNPs em ESTs, que foi escrito pelo aluno de iniciação científica Kleider Torres do nosso grupo de Bioinformática, o cSNPer. Ele foi projetado para identificação de SNPs baseando no parâmetro de *Neighbourhood Quality Standard/NQS* proposto por Altshuler e colaboradores (2000). O programa foi desenvolvido através da linguagem de programação C apresentando 6 módulos, cada um com uma função específica. O módulo cSNPer.c contém as funções de execuções do programa, o módulo cSNPer.h contém as estruturas de registro e as inclusões de bibliotecas, o módulo archive.c mantém a estrutura da função de leitura, o módulo archive.h apresenta a definição dos parâmetros de leitura e escrita, o módulo main.c contém a função principal do programa e o módulo main.h as definições de *input* (entrada). O programa foi estruturado através de alocação dinâmica de memória, permitindo ao programador alocar memória para variáveis enquanto o programa está sendo executado. O programa utilizou como entrada o arquivo gerado pelo programa de agrupamento Phrap, no formato ACE, e o arquivo contendo os valores de qualidade das bases das ESTs. Para a identificação dos polimorfismos o programa levou em consideração o valor de qualidade da base polimórfica e das bases vizinhas a ela (NQS). O valor de qualidade das bases considerado foi o valor atribuído pelo programa Phred. Foram utilizados como parâmetros, na detecção de SNPs, os seguintes valores de qualidade: Phred ≥ 20 nas ESTs, Phred ≥ 40 na seqüência consenso e um alinhamento das bases vizinhas ao SNP, de pelo menos 10 pb a esquerda e a direita da base em questão, apresentando Phred ≥ 15 (Figura 4).

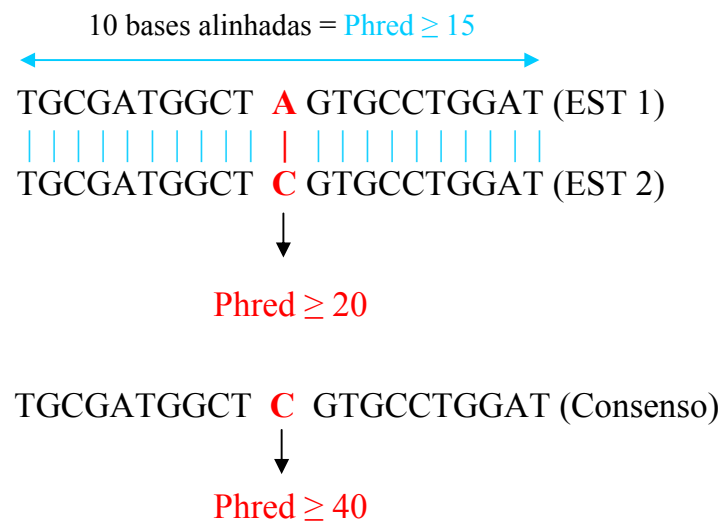


Figura 4 – Parâmetros de qualidade utilizados pelo novo programa, cSNPer, na identificação dos SNPs.

Como arquivos de saída (*output*), o programa gerou: um arquivo “resultados.txt” contendo a análise detalhada de SNPs por contig, como o número de SNPs detectados e a posição do polimorfismo na EST e na sequência consenso (Figura 5), um arquivo “r_de_qualidade1.csv.” contendo a qualidade da base do SNP detectado e a qualidade das bases vizinhas (Figura 6), um arquivo “saida.csv” da análise geral de toda a busca com o total de SNPs detectados, o número de pares de bases analisados, média de SNPs por pb, número de contigs com SNPs, o total de cada mudança de nucleotídeo, número de mutações sinônimas e não-sinônimas e o número de mutações em cada posição do códon (Figura 7). Em seguida, o programa foi implementado de forma a determinar a possível janela de leitura (ORFs) de cada contig e, conseqüentemente, o efeito dos polimorfismos identificados quanto à mudança de aminoácido em mutações sinônimas e não-sinônimas. Durante este processo as seqüências consensos foram sendo traduzidas em aminoácidos de acordo com as seis janelas de tradução (-1, -2, -3, +1, +2, +3) e foi gerado o arquivo de saída “frame.csv.” (Figura 8). A variável *Frame* determinou a janela de tradução de acordo com o tamanho estipulado, que pode ser variado, e a direção da tradução (5'→3' ou 3'→5'). No presente trabalho, a janela de leitura dos contigs foi determinada em ORFs contendo, no mínimo, 60 aminoácidos. No caso de detecção de mais de uma janela com este tamanho, foi considerada aquela na qual havia a presença do SNP ou ambas as janelas. Inserções e deleções não foram incluídas no algoritmo de busca. Para visualização de alguns dos SNPs identificados nos contigs foi utilizado o programa de edição, o Consed (Gordon *et al.*, 1998).

```

1:bioinfo.cpqrr.fiocruz.br - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

ID, CoNTiG621
AAAACCAGCC C TGCTACAGTT, (422), consensus
AAAACCAGCC T TGCTACAGTT, (124), Seq58155,
NQS, 10
Numcontigs:1 \ N_snp:1

ID, CoNTiG630
CAGCACTGAA C TTGACAGTGT, (96), consensus
CAGCACTGAA T TTGACAGTGT, (241), Seq60516,
NQS, 0
Numcontigs:2 \ N_snp:1

ID, CoNTiG631
TGGATTTTTT G GGCAAATATG, (329), consensus
TGGATTTTTT T GGCAAATATG, (330), Seq35558,
ID,
TGAAGTCGTT G GTGTAGAGCT, (517), consensus
TGAAGTCGTT T GTGTAGAGCT, (518),
NQS, 0
Numcontigs:3 \ N_snp:2

ID, CoNTiG636
TTTCAAGGAA G CCCAAATGGC, (146), consensus
TTTCAAGGAA A CCCAAATGGC, (146), Seq60194,
ID,
GTATGGCCAC G GCCACGATGT, (505), consensus
TATTGCCACG T GCCACGATGT, (505),
NQS, 8
Numcontigs:4 \ N_snp:2
    
```

Figura 5 – Exemplo do arquivo de saída “resultados.txt” do programa cSNPer. Em vermelho, o número de identificação do contig (ID). Em azul, os SNPs identificados e as suas respectivas posições, na seqüência consenso e na EST (em rosa). A variável Numcontigs (em amarelo), contabiliza a quantidade de agrupamentos que foram processados e a variável N_snp (em verde), o número de SNPs identificados em cada agrupamento.

```

1:bioinfo.cpqrr.fiocruz.br - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles
Sequencia:58155 ID,CoNTiG621
82 79 82 82 82 74 74 59 59 59 -59- 51 51 66 74 90 90 90 89 82 72(422)
Sequencia:60516 ID,CoNTiG630
76 71 70 68 57 57 48 48 48 54 -54- 57 57 68 68 82 82 82 84 84 59(96)
Sequencia:35558 ID,CoNTiG631
90 58 58 43 46 46 46 50 48 48 -52- 52 52 52 67 57 86 85 89 89 90(329)
Sequencia:35558 ID,
78 78 78 66 53 38 38 23 23 41 -41- 37 37 53 53 53 51 52 67 62 62(517)
Sequencia:60194 ID,CoNTiG636
55 55 55 55 55 55 55 50 50 50 -50- 50 43 50 32 29 29 29 29 41 52(146)
Sequencia:60194 ID,
52 52 52 52 52 52 52 52 52 52 -52- 52 52 52 52 52 52 52 52 52 52(505)
Sequencia:445 ID,CoNTiG637
42 59 59 59 37 53 53 53 53 52 -44- 44 44 44 44 80 78 75 81 85 78(144)
Sequencia:36418 ID,CoNTiG645
90 90 79 65 65 76 68 68 53 53 -50- 50 41 44 59 68 68 68 90 90 69(270)
Sequencia:60879 ID,CoNTiG646
90 90 90 90 83 77 77 62 62 62 -62- 62 62 77 77 90 90 90 90 90 90(274)
Sequencia:39957 ID,CoNTiG651
53 53 56 50 35 37 50 52 52 52 -52- 67 67 87 87 90 90 90 90 90 90(399)
Sequencia:39957 ID,
90 90 67 67 67 52 50 46 52 52 -52- 67 67 67 76 76 76 73 78 81 76(420)

```

Figura 6 – Exemplo do arquivo de saída “r_de_qualidade1.csv.” do programa cSNPer. Este arquivo contém o valor de qualidade da base dos SNPs identificados (em vermelho) e o valor de qualidade das bases vizinhas ao SNP (em azul).

```

Quick Connect  Profiles

|-----|
|C->A | C->T | C->G|
221  | 266  | 134  |
|-----|
|A->C | A->T | A->G|
739  | 263  | 355  |
|-----|
|T->C | T->A | T->G|
389  | 286  | 782  |
|-----|
|G->C | G->A | G->T|
171  | 340  | 258  |
Numero total de bases nos contigs:3702915
Numero total de SNPs: 4204
Numero total SNPs por base:880
Numero de contigs com SNPs:1799
sinonimos: 514      nsinonimos: 1482
posicao no codon: 672      posicao no codon: 641      posicao no codon: 683
./marianafinal112/saidal.csv (END)

```

Figura 7 – Exemplo do arquivo de saída “saida.csv” contendo a análise geral dos SNPs identificados. Este arquivo contém o total de SNPs detectados (em vermelho), o número de pares de bases analisados (em verde), média de SNPs por pb (em amarelo), número de contigs com SNPs (em azul), a soma de cada tipo de mudança nucleotídica (em lilás), número de mutações sinônimas e não-sinônimas e o número de mutações em cada posição do códon (em cinza).

```

1:bioinfo.cpqrr.fiocruz.br - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

CoNTiG621
IKSLLNVTRYQRDHYSSVPLFSPQKDGCVRNRTQRCPANCITCCRFPPTQFIQFLLSVLKHCQ -> VQCCC*
frame: -1
Consensu
SNP: G posicao: 96 codon: GTT aminoacido: V posicao no codon: 1
Sequencia
SNP: A posicao: 241 codon: ATT aminoacido: I
sinonimos: 0 nsinonimos: 1
posicao no codon1: 1 posicao no codon2: 0 posicao no codon3: 0

GTVHRDAQLTASHAVGSRHSSFSFSLSSNTV -> KFSAAAENSKNLSSPAGRNGGDRGGRENREH*
frame: -2
Consensu
SNP: G posicao: 96 codon: AAG aminoacido: K posicao no codon: 3
Sequencia
SNP: A posicao: 241 codon: AAA aminoacido: K
sinonimos: 1 nsinonimos: 1
posicao no codon1: 1 posicao no codon2: 0 posicao no codon3: 1

```

Figura 8 – Exemplo do arquivo de saída “frame.csv.” do programa cSNPer. Este arquivo contém informação relacionada à sequência de aminoácido e o efeito do SNP identificado, como o SNP e a respectiva janela de tradução (ORF) (em amarelo), o códon e o aminoácido codificado na presença do SNP (em azul), classificação dos SNPs em sinônimos e não-sinônimos (em rosa) e a posição do polimorfismo no códon (em verde). A seta vermelha indica a posição do aminoácido variante.

Pelo fato do cSNPer ser um programa novo, foram realizados experimentos controlados com seqüências modelo para verificar a performance do programa de acordo com os parâmetros exigidos. Para isso, foi selecionado um contig modelo e zerada as qualidades das bases. O programa detectou vários SNPs quando o valor de qualidade não foi considerado. Em seguida, manipularam-se os valores de qualidade. Os polimorfismos foram corretamente detectados de acordo com a alteração manual dos parâmetros, quanto às suas características (vide apêndice).

Após a detecção dos SNPs nas ESTs, foi calculado pelo programa SPSS© (versão 11) a correlação entre a profundidade dos contigs (número de pares de base/número de ESTs) e a quantidade de SNPs. O programa calculou todos os desvios pelo comando *compute* e analisou a correlação através do comando *correlate*.

Foi analisada a presença de polimorfismos em genes candidatos à vacina contra *S. mansoni*. Foram selecionados 7 genes: glutationa-S-transferase (GST) (McNair *et al.*, 1993), paramiosina (Laclette *et al.*, 1991), triose fosfato isomerase (TPI) (Shoemaker *et al.*, 1992), o antígeno de membrana 23 kDa (Sm23) (Wright *et al.*, 1990), o antígeno de membrana 14 kDa (Sm14) (Moser *et al.*, 1991), IrV-5 (miosina) (Weston *et al.*, 1993) e catepsina B (Klinkert *et al.*, 1989). As seqüências foram obtidas do GenBank no formato FASTA para cada gene. As seqüências foram, em seguida, formatadas para Blast com o uso do programa formatdb (linha de comando: *formatdb -i RGMG.fasta.screen.contigs -p F -o T*) para a realização do Blast, *Basic Local Alignment Search Tool* (Altschul *et al.*, 1990). Foi rodado o programa blastall através da linha de comando: *blastall -p blastn -i vaccinecandidates.fasta -d RGMG.fasta.screen.contigs -a 3 -b 0 -o blast.out*, utilizando o programa blastn, responsável por comparar seqüências de nucleotídeo contra banco de dados de nucleotídeos. Os contigs correspondentes às seqüências de interesse foram determinados pelo valor máximo do *score* (medido pela similaridade das bases e menor número de *gaps*), apresentando o *e-value* igual a zero.

3.2 - Validação Experimental

A validação experimental foi realizada em possíveis SNPs identificados manualmente, pois o desenvolvimento do programa cSNPer foi finalizado recentemente. Foram utilizadas as seqüências públicas disponíveis no ano de 2002 no GenBank. As seqüências foram agrupadas

e aquelas que tinham 10 ou mais ESTs por agrupamento foram analisadas. Os critérios utilizados para determinar os possíveis SNPs foram:

- Deveriam se localizar, pelo menos, a 30 pb do início ou do fim das seqüências;
- Deveriam estar presentes no mínimo em duas ESTs;
- Não deveriam estar localizados em regiões repetitivas.

O tamanho total de agrupamentos analisados foi de 40.966 pb, nos quais foram identificados 200 possíveis SNPs em 40 agrupamentos diferentes. Dentre os diferentes genes correspondentes aos agrupamentos analisados, o agrupamento 200 correspondente ao gene da catepsina B de *S. mansoni* foi escolhido para verificação experimental dos SNPs detectados.

3.2.1- Amostra biológica e extração de RNA

Foram utilizados dois grupos representantes de *pool* de verme adulto de *S. mansoni* para realização da extração de RNA. O primeiro grupo foi proveniente da cepa Porto Rico, mantida em laboratório. O segundo grupo foi proveniente do campo, de área endêmica. As amostras do campo foram coletadas na comunidade de Caju no Vale do Jequitinhonha e coletadas por Regina Coeli do Laboratório de Parasitologia Celular e Molecular/CPqRR.

Vermes adultos foram homogeneizados em solução de guanidina (4M Tiocinato de Guanidina, 1mM EDTA pH=7,4, 25mM Acetato de Sódio pH=5,5, 5% mercaptoetanol, 2% lauril sarcosinado) e centrifugados a 6.000 rpm por 30 minutos. O sobrenadante foi cuidadosamente transferido para outro tubo contendo solução de CsCl (densidade de 1,77) e posteriormente centrifugado a 28.000 rpm por 20 horas (Ultracentrífuga Beckman L80, USA/SW55ti) a 18°C. Foi desprezado o sobrenadante, as paredes do tubo foram lavadas e o sedimento ressuspensionado em água autoclavada, estéril e livre de RNase. O produto foi tratado com solução de DNase e livre de RNase (Invitrogen, CA, USA). Para síntese de cDNA foi utilizado o Kit Thermoscript (Invitrogen), seguindo as instruções do fabricante. A concentração do material foi determinada por espectrofotometria (aparelho Bio Photometer – AG 22331 Hamburg – Eppendorf) a 260nm, sendo o grau de pureza das amostras determinado pela relação da absorbância: $A_{260/280nm}$. O cDNA foi armazenado a -70 °C

3.2.2 - Amplificação do gene da Catepsina B de *S. mansoni*

Utilizando o programa OLIGO® versão 3.3 (Molecular Biology Insights – Cascade, USA), dois pares de iniciadores foram desenhados flanqueando as regiões polimórficas (Invitrogen, São Paulo). O fragmento de 536 pb, correspondente a porção 5' do gene (1pb – 536pb), foi amplificado através da reação de cadeia em polimerase (PCR). A reação foi realizada utilizando 1ng de cDNA (obtido como descrito acima) de verme adulto em um volume final de 10 µl, contendo 5 pmoles dos iniciadores Sm31-X5 e Sm31-X6 (Tabela I), 200 µM deoxinucleotídeos trifosfato (dNTPs) (Invitrogen, São Paulo), 2,5 U da enzima *Taq* DNA polymerase (Invitrogen, São Paulo) e tampão da enzima fornecido pelo fabricante (MgCl₂ 1,5 mM, Tris-HCl 10 mM pH=8,0, KCl 50 mM). As amostras foram amplificadas em termociclador Thermo Hybaid - PCR Express, aquecendo, primeiramente, a 95^oC por 2 min, seguidos de 40 ciclos a 94^oC por 1 min, temperatura de anelamento de 58^oC por 1 min e 72^oC por 1 min e uma incubação final a 72^oC por mais 10 min. O segundo, foi o fragmento de 210 pb, correspondente a porção 3' do gene (912pb - 1121pb). O fragmento foi amplificado por PCR utilizando 1 ng de cDNA em um volume final de 20 µl, contendo 5 pmoles dos iniciadores Sm31-X7 e Sm31-X8 (Tabela I), 200 µM deoxinucleotídeos trifosfatos (dNTPs) (Invitrogen, São Paulo), 2,5 U da enzima *Taq* DNA polimerase (Invitrogen, São Paulo) e tampão da enzima fornecido pelo fabricante (MgCl₂ 1,5 mM, Tris-HCl 10 mM pH=8,0, KCl 50 mM). As amostras foram amplificadas, como descrito anteriormente, mas com a temperatura de anelamento de 59^oC. Cada reação foi acompanhada de um controle negativo, que consistiu na inclusão de todos os reagentes necessários à amplificação com exceção do DNA modelo.

Tabela I – Iniciadores utilizados na reação de PCR no gene da catepsina B de *S. mansoni* (Número de acesso do GenBank #M21309).

Iniciador	Posição no Gene	Seqüência do Iniciador
Sm31-X5	1 pb – 23 pb	5' ATTCAAGAGTTATTTGGACATGC 3'
Sm31-X6	512 pb – 536 pb	5' CCTGCTTGGGATTACTGGGTGAAGG 3'
Sm31-X7	912 pb – 930 pb	5' ATAAAAGCTTACAAGACTCCTTATTGGTT 3'
Sm31-X8	1103 pb – 1121 pb	5' AATAAAGCTTTTTGAAGTATTCAGTATACA 3'

3.2.3- Eletroforese

Para visualização do DNA amplificado, 5 µl da reação diluída em volume igual de tampão de amostra 2X (azul de bromofenol 0,25%; xilenocianol 0,25% e 30% glicerol) foram aplicados em gel de poliacrilamida 6%, contendo 20 ml de Bis-Acrilamida 30% (29:1) – (Bio-Rad, EUA), 20 ml de TBE 5X [54 g de Tris base (Pharmacia Biotech, Comunidade Européia), 27,5 g de ácido bórico (Gibco, Brasil), 20 ml de EDTA 0,5 M (Pharmacia Biotech) pH=8,0 e H₂O q.s.p. 1 L] e 60 ml de H₂O, em tampão TBE 1X. A eletroforese foi realizada no aparelho Mini-Protean II (Bio-Rad, Hercules, CA, Estados Unidos) a 50 Volts (~15mA/gel) até a separação dos corantes, aumentando-se a tensão para 100 Volts até o final da corrida. O gel foi corado com 0,5 µg/ml de brometo de etídio (Sigma, EUA) e fotografado com aparelho *Eagle Eye II* (Stratagene, La Jolla, CA, Estados Unidos). A presença do amplicon foi determinada pelo peso molecular da banda esperada.

3.2.4 - Clonagem e Seqüenciamento

Os fragmentos amplificados foram clonados em vetor TOPO utilizando o Kit TA-TOPO cloning 2.1 (Invitrogen - Life - Technologies, Europa), seguindo as instruções do fabricante. Em seguida, foi realizada transformação química utilizando células TOPO 10 F' competentes. A 100 µl de células competentes acondicionados em microtubos de 1,5 ml, foram adicionados 6 µl do produto de ligação. Essa mistura foi incubada 30 minutos no gelo, submetida a choque térmico por incubação a 42°C por 50 segundos e, imediatamente, transferida para o gelo. Em ambiente asséptico, foram adicionados 200 µl de meio SOC a cada tudo [1% de triptona (Difco, Brasil); 0,5% de extrato de levedura (Biobrás, Brasil) pH=6.8-7.2); 8,5% de NaCl (Quimex, Brasil – PM=58.44); 2,5 mM de KCl (Synth, São Paulo); 0,01 mM MgCl₂ (Invitrogen, São Paulo); 0,02 mM glicose, água q.s.p 1000 ml]. A mistura foi incubada a 37°C por 45 minutos, sob agitação orbital de 300 rpm. Em seguida, as células foram centrifugadas a 13.000 g (Eppendorf Mini Spin, Alemanha) por 1 minuto e, ressuspensas, com o auxílio de pipeta. Aproximadamente, 150 µl dessa suspensão foram distribuídos em placas de Petri contendo meio LB sólido, adicionado de 100 µg/ml ampicilina (Invitrogen, USA), 40 µg de X-gal (5-bromo-4cloro-3indolil-β-D-galactopiranosida) (Invitrogen, USA) e 0,1 mM IPTG (isopropil-β-D-tiogalactopiranosida) (Ge Healthcare – Amersham, Europa). As placas de Petri foram incubadas a 37°C por 12-14 horas. Para extração do DNA plasmidial, as colônias recombinantes foram transferidas para tubos de 50

ml contendo 5ml de meio LB líquido acrescido de 100 µg/ml de ampicilina e cultivadas sob agitação orbital a 220 rpm por 16 horas a 37°C. A extração foi realizada utilizando o Kit QIAprep Spin Miniprep (Qiagen, São Paulo), seguindo as instruções do fabricante e armazenadas a -20°C. Foi realizada reação de PCR, utilizando os iniciadores específicos descritos, confirmando a presença do inserto. A visualização e a quantificação das amostras foram resolvidas em gel de Agarose 1% (Promega). A agarose foi fundida, aquecendo-a em microondas, em tampão de corrida TBE 1X e aplicada à cama de transferência até sua solidificação. A corrida foi realizada com tampão de transferência TBE 1X a 70Volts em cuba de eletroforese BRL Horizontal Gel Electrophoresis Horizon 11.14 (Gibco - Life Technologies). O gel foi corado em brometo de etídeo (0,5 µg/ml) (Sigma) e a imagem digitalizada pelo aparelho *Eagle Eye II* (Stratagene).

O seqüenciamento dos insertos dos clones foi realizado pelo método Sanger (1997) com o kit DYEnamic ET dye terminator (Amersham Pharmacia Biotech, Madison - USA). Foram utilizados, aproximadamente, 200 ng de DNA e 3,3 pmoles do iniciador universal M13, com volume final de 10 µl. As amostras foram submetidas às seguintes condições de amplificação em termociclador (Mastercycler, Eppendorf): 25 ciclos, 95°C por 20 segundos, 50°C por 15 segundos e 60°C por 1 minuto. Os produtos amplificados foram purificados por precipitação em placas de 96 poços. Na realização do processo de precipitação foram adicionados, primeiramente, 1 µl de acetato de amônio (Amersham Biosciences-UK limited, England) e 30 ml de etanol 96% (Merck, Brasil) em cada poço. A mistura foi, rapidamente, vortexada e incubada por 20 minutos a temperatura ambiente. Em seguida, a placa foi centrifugada (Mastercycler, Eppendorf) a 4.000 rpm por 45 minutos a 7°C, o conteúdo descartado e foram adicionados 100 ml de etanol 70% gelado, lentamente, pela parede do poço. A placa foi centrifugada a 4.000 rpm por 10 minutos a 7°C. Em seguida, o líquido foi descartado e a placa, invertida sobre papel absorvente, foi centrifugada por 1 segundo a 900 rpm. Após a precipitação, as amostras foram suspensas em 10 µl de tampão de amostra (Amersham Biosciences-UK limited, England) adicionados com pipeta multicanal e agitada no vortex por, no mínimo, 2 minutos. O seqüenciamento das amostras foi realizado pelo seqüenciador automático MEGABACE 500 DNA Analysis System (Amersham Pharmacia Biotech).

As seqüências de nucleotídeo dos clones foram alinhadas utilizando o programa GeneTool (<http://www.biotoools.com>). As seqüências de baixa qualidade foram eliminadas. Discrepâncias entre as seqüências foram averiguadas por inspeções visuais dos

cromatogramas e editadas manualmente, quando necessário. A comparação das seqüências permitiu a identificação dos possíveis SNPs (Figura 9). Cada SNP foi categorizado de acordo com a sua característica específica quanto à mudança de nucleotídeo em transição ou transversão, a troca ou não do aminoácido codificado em sinônimo ou não-sinônimo e a respectiva posição no códon.

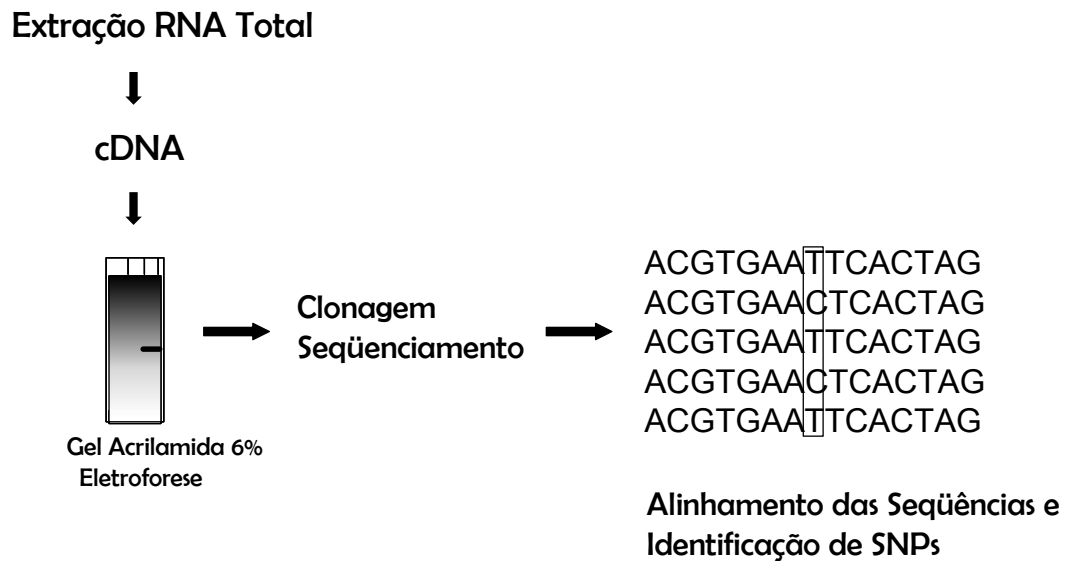


Figura 9 – Estratégia utilizada na detecção de SNPs experimentalmente. Após a síntese de cDNA, a partir da extração de RNA de verme adulto, foi realizada a reação de PCR e, posteriormente, os produtos foram clonados e seqüenciados. As seqüências foram alinhadas utilizando o programa GeneTool e os SNPs identificados.

3.3 - Modelagem por Homologia

Para realização da modelagem por homologia ou modelagem comparativa foi utilizado o programa Modeller versão 7.0 (Marti-Renom *et al.*, 2000). Primeiramente, foi realizada a modelagem da proteína sem a presença de mutações e em seguida, alterando a seqüência de aminoácido, foi realizada a modelagem da proteína para cada mutação não-sinônima.

O método de modelagem de proteínas por homologia implica, basicamente, em seguir quatro passos sucessivos que estão descritos abaixo.

3.3.1 - Identificação e seleção de proteínas modelo

Nesta primeira etapa foi identificada e selecionada uma proteína de estrutura tridimensional conhecida, que serviu como modelo para a determinação da estrutura da catepsina B de *S. mansoni*. Para esta identificação foi realizado blastp (<http://www.ncbi.nlm.nih.gov/BLAST>), utilizando a seqüência de aminoácido da catepsina B do parasito, contra estruturas depositadas no banco de dados PDB. Dentre as várias seqüências que apresentaram similaridade com a proteína alvo, foi escolhida aquela que apresentou o maior valor de *score* com a melhor resolução. A proteína escolhida foi a procatepsina B humana (gi|2982152|pdb|3PBH) com resolução de 2.5 Angstroms.

3.3.2 - Alinhamento das seqüências de resíduos de aminoácidos

Em seguida foi necessário reformatar a seqüência para o formato PIR (arquivo X.ali) e, a partir disso, foi realizado o alinhamento da seqüência de aminoácido da catepsina B de *S. mansoni* com a seqüência da proteína modelo. Utilizando o comando *mod7v7 Align2d.top*, o programa Modeller gerou um alinhamento que foi em seguida otimizado manualmente na tentativa de eliminar os espaços vazios ou *gaps*. O alinhamento levou em consideração características estruturais comuns, tais como elementos de estrutura secundária e resíduos catalíticos.

Exemplo dos comandos utilizados pelo programa Modeller para realizar o alinhamento da proteína modelo e uma das mutações não-sinônima:

```
# Align2d.top file for Modeller
# CAT - cathepsin - Mariana Simões
#####

# Leitura do arquivo PDB
READ_MODEL FILE = '3pbh.pdb'
# Definindo o código do PDB
SEQUENCE_TO_ALI ALIGN_CODES = '3pbh'
# Definindo local de input do arquivo PDB
SET ATOM_FILES_DIRECTORY = 'C:\documentsandSettings\Mariana
Crivellari/MyDocuments\Mestrado\db\PDB'
# Leitura do arquivo
READ_ALIGNMENT FILE = 'CATHBSNP6.txt', ALIGN_CODES = ALIGN_CODES
'CATHBSNP6', ADD_SEQUENCE = ON
# Alinhamento no formato PIR
WRITE_ALIGNMENT FILE='CATHBSNP6.ali', ALIGNMENT_FORMAT = 'PIR'
```

3.3.3 - Construção do modelo (geração das coordenadas cartesianas)

Após a realização do alinhamento, o programa Modeller calculou o modelo tridimensional da seqüência alvo de forma automática através do comando *mod7v7 model.top*. O programa utilizou as coordenadas espaciais da proteína modelo para gerar um conjunto de restrições espaciais, que foram aplicadas à seqüência alvo, como por exemplo à distância entre resíduos similares na estrutura modelo e na estrutura alvo. Além das restrições espaciais, baseadas na similaridade entre as seqüências, o programa, também, utiliza um campo de força controlando a estereoquímica mais apropriada, limitando assim o número de configurações que o modelo pode assumir. A partir disso, dez possíveis modelos foram calculados para a proteína da catepsina B de *S. mansoni* e gerados no formato PDB. Os modelos foram avaliados de acordo com os parâmetros estereoquímicos e foi escolhido aquele que apresentou o valor mais baixo da função objetiva do programa Modeller. Após a modelagem comparativa da catepsina B de *S. mansoni*, foi necessário realizar o mesmo processo de modelagem da proteína para cada mutação não-sinônima.

3.3.4 - Análise dos modelos PDB

Os modelos gerados pelo programa Modeller no formato PDB foram analisados e comparados através do programa STING Millennium Suíte – SMS. O SMS é uma suíte de programas desenvolvida pelo Núcleo de Bioinformática da Embrapa Informática Agropecuária para análise de estruturas de proteínas e a relação com a sua função (Neshich *et al.*, 2003). Utilizando a interface gráfica foi possível visualizar a disposição das cadeias laterais dos aminoácidos e analisar em detalhe as mudanças de aminoácidos decorrentes das mutações não-sinônimas encontradas. Pelo módulo *Java Protein Dossier* (JPD) (Neshich *et al.*, 2004) foi possível analisar e comparar as várias características específicas destes aminoácidos, como a posição na seqüência e na estrutura, parâmetros físico-químicos, identificação de vizinhança, ângulos e distância entre átomos e pelo leque de contatos a relação entre os contatos intra-cadeia, como forças de atração e de repulsão existentes entre átomos distintos. O gráfico de Ramachandran possibilitou a análise da estereoquímica dos modelos e, conseqüentemente, sua validação. O módulo ConSSeq (Higa *et al.*, 2004) analisou a frequência relativa dos aminoácidos em posições específicas da seqüência, baseando na seqüência consenso gerada pelo alinhamento múltiplo de seqüências homólogas.

IV - RESULTADOS

4.1 – Detecção de SNPs *in silico*

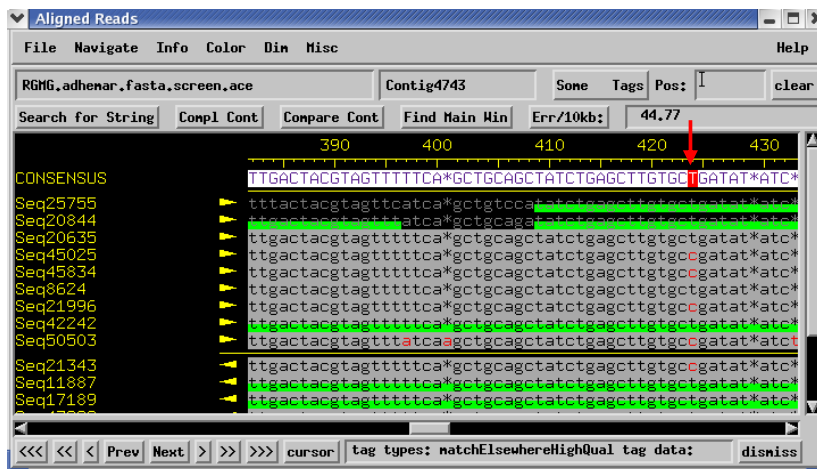
Os SNPs detectados foram provenientes de 61.002 ESTs de *S. mansoni*, geradas pelo Projeto Genoma de Minas Gerais. Essas seqüências foram processadas pelo programa de chamada de bases Phred. O total de 5.082 contigs foi gerado pelo programa Phrap, com uma média de 8,32 ESTs por contig. O programa Phrap também não agrupou 16.196 seqüências denominadas *singlets*, que foram aquelas ESTs que não apresentaram similaridade suficiente com outras seqüências para serem agrupadas. Utilizando o novo programa de detecção de SNPs, o cSNPer, 2.303 possíveis SNPs foram detectados em 863 agrupamentos. Alguns dos SNPs identificados, foram visualizados através do programa de edição de seqüências, Consed (Figura 10). Dentre os agrupamentos com SNPs, o número de polimorfismos variou entre 1 e 49 (Figura 11). O tamanho total de bases nos agrupamentos analisados foi de 3.702.915 pb, com um possível SNP a cada 1.607 pb. A Figura 12 mostra a comparação do número de SNPs e a profundidade do agrupamento (razão entre o número de ESTs e o tamanho total da seqüência consenso). Pode-se observar que o aumento do número de possíveis SNPs não está correlacionado com o aumento da profundidade dos agrupamentos. Esta observação enfatiza que os possíveis SNPs encontrados são provavelmente reais e não se devem a um acúmulo de erros de seqüenciamento, freqüentes em seqüências do tipo ESTs. A forma indicada para medir a relação entre a profundidade e o número de SNPs foi a correlação de Spearman. Observou-se que os dados não seguem a suposição de normalidade indicando a correlação de Spearman, visto que é uma versão não-paramétrica da correlação de Pearson. O coeficiente de correlação foi aproximadamente de 0,015 (valor de $p=0,603$), indicando que não existe associação significativa entre a profundidade e o número de SNPs com nível de significância de 0,05.

Os possíveis SNPs foram analisados de acordo com a variação da base nucleotídica, em transição ou transversão. O tipo de substituição mais freqüente foi por transversão A/C (19,75%), A/T (12,46%), C/G (6,38%) e G/T (25,74%), seguida da substituição por transição A/G (19,19%) e T/C (16,45%) (Figura 13). SNPs também foram classificados considerando o seu efeito sobre o aminoácido codificado, em mutações sinônimas ou mutações não-sinônimas e sua posição no códon. As janelas de leitura (ORFs) dos contigs analisados foram selecionadas por apresentarem no mínimo 60 aminoácidos. Foram detectados 30,54%, 30,32% e 39,13% SNPs na primeira, segunda e terceira base do códon, respectivamente (Figura 14). Sendo que desses, 30,32% representam mutações sinônimas e 69,67% mutações

não-sinônimas (Figura 15). O restante, 1.429 SNPs (62,04%), não estavam em regiões de ORFs.

O próximo objetivo foi analisar os genes que apresentavam o maior número de polimorfismos (Tabela II). É interessante ressaltar que alguns dos genes mais polimórficos observados são antígenos conhecidos como: aldolase (Argiro *et al.*, 2000), GAPDH (Argiro *et al.*, 2000), catepsina B (Noya *et al.*, 2001), proteína 14-3-3 (Schechtman *et al.*, 2001), superoxide dismutase (Cook *et al.*, 2004) e uma proteína do tegumento (Hoffmann & Strand 1996). Foi verificado que retrotransposons e genes mitocondriais, também, são variáveis, sendo que polimorfismos em retrotransposons, provavelmente, signifiquem genes parálogos. Em seguida, foram identificados os contigs referentes aos genes candidatos à vacina de *S. mansoni* e verificado a presença de SNPs nesses genes. Foram selecionados 7 genes: glutationa-S-transferase (GST), paramiosina, triose fosfato isomerase (TPI), o antígeno de membrana 23 kDa (Sm23/MAP3), o antígeno de membrana 14 kDa (Sm14) e catepsina B. Na maioria dos genes foram detectados polimorfismos (Tabela III), exceto nos genes da TPI, miosina e paramiosina.

A



B

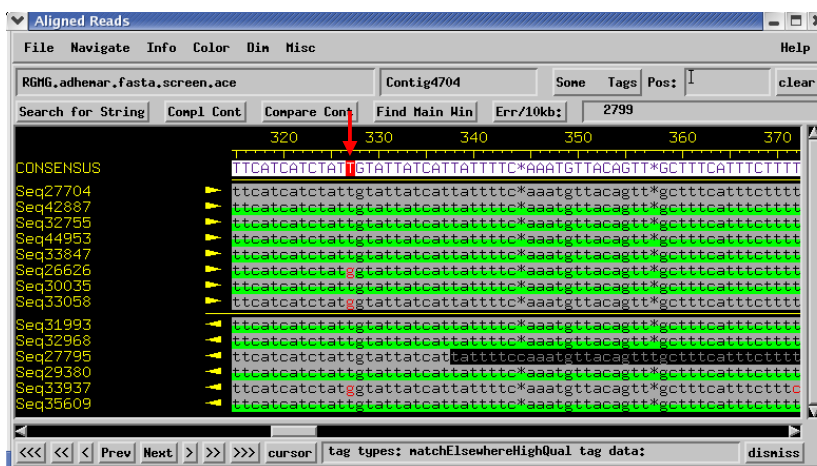


Figura 10 – Janela de visualização dos agrupamentos pelo Consed. A - Figura da sequência consenso do contig 4743 que possui 85% de identidade com o clone SJCHGC09320 de *Schistosoma japonicum*; B - Contig 4704 possui 81% de identidade com o clone SJCHGC06319 de *Schistosoma japonicum*. A seta vermelha indica a posição dos SNPs. A base que varia também está indicada pela cor vermelha.

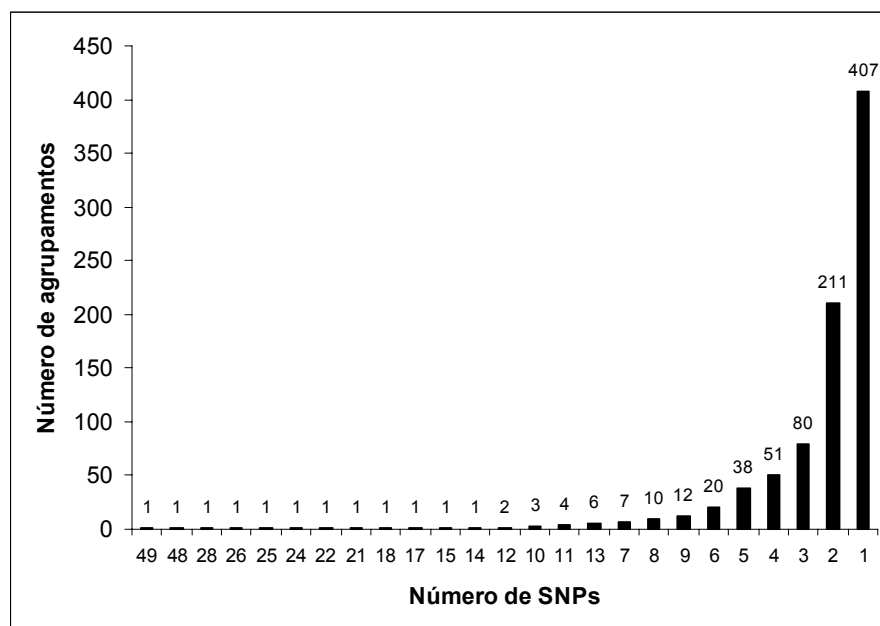


Figura 11 – Gráfico da variação do número de SNPs nos agrupamentos. O número de SNPs variou entre 1 e 49 nos agrupamentos, sendo que, apenas, um agrupamento apresentou 49 polimorfismos, que foi o contig 5054 referente ao gene de *Schistosoma mansoni* for eggshell protein.

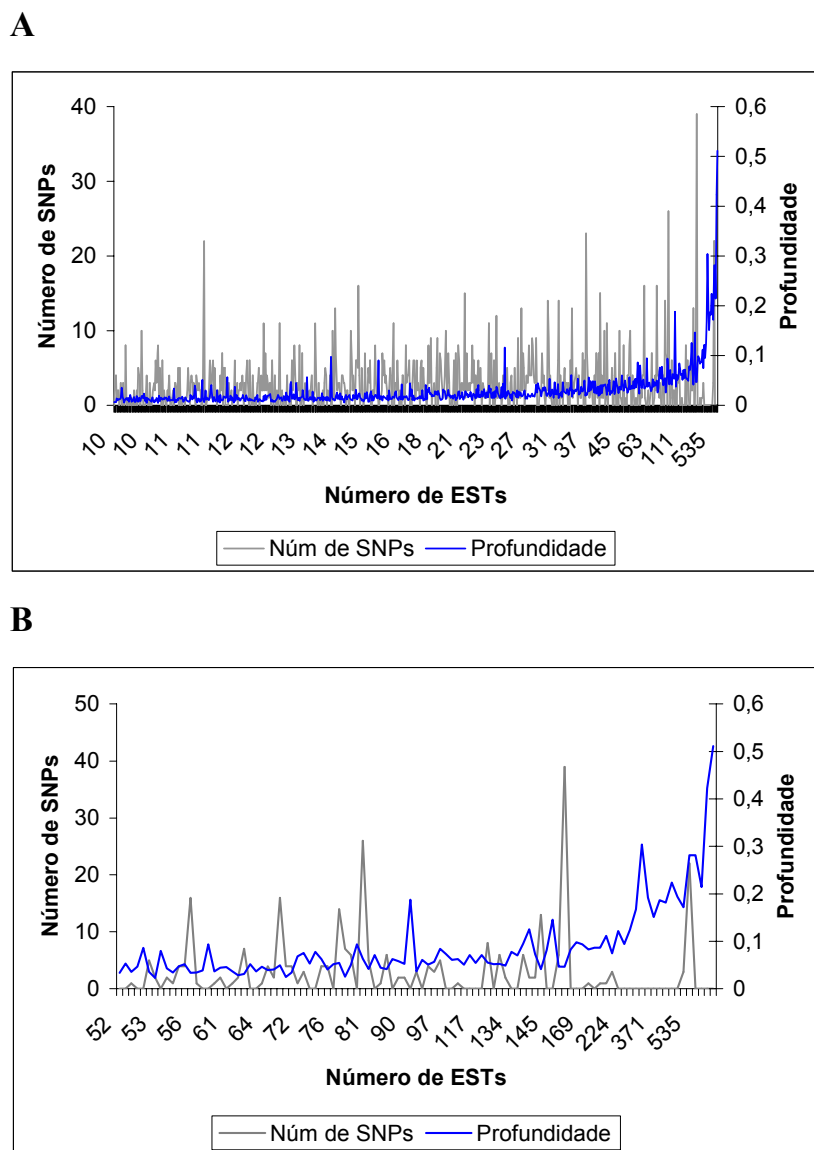


Figura 12 – Número de SNPs por agrupamento e a profundidade do agrupamento. **A** - A profundidade foi definida pelo número de ESTs no agrupamento dividido pelo total de pares de bases da sequência consenso do agrupamento. O valor $R^2 = 0.2532$ calculado demonstrou que não existe associação entre a profundidade dos contigs e o número de SNPs ($p > 0,05$); **B** - Aproximação da imagem da figura anterior contendo agrupamentos com mais de 52 ESTs.

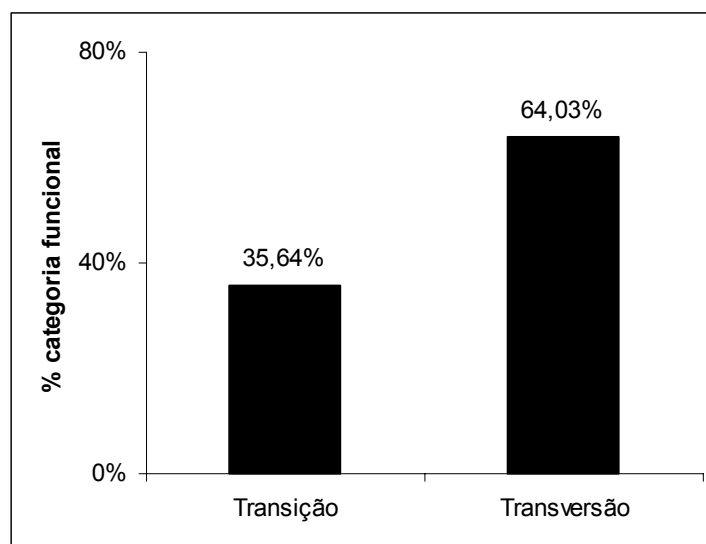


Figura 13 – Frequência dos SNPs identificados *in silico*, de acordo com a mudança da base nucleotídica, em transição ou transversão.

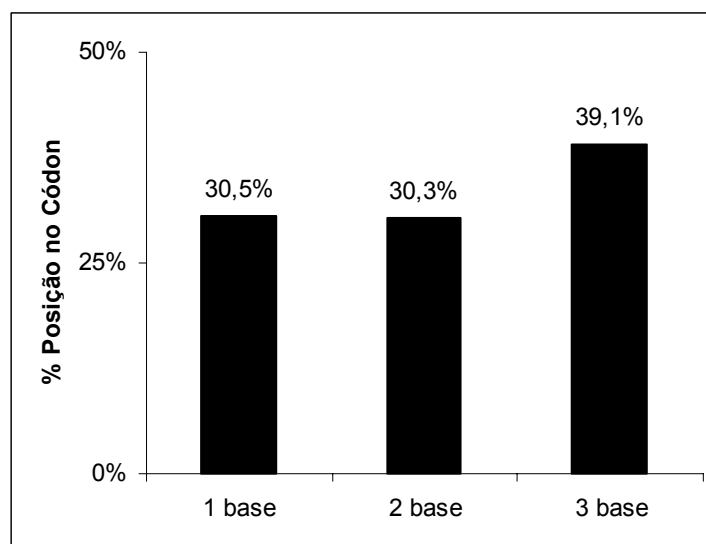


Figura 14 – Classificação dos SNPs de acordo com a posição no códon.

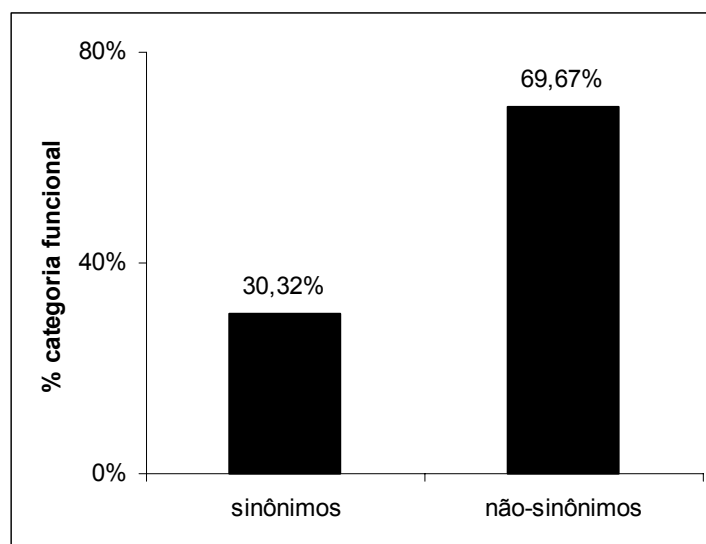


Figura 15 – Classificação dos SNPs em sinônimos e não-sinônimos, dependendo da mudança do aminoácido codificado.

Tabela II – Genes que apresentaram o maior número de polimorfismos. O gene correspondente ao contig de acordo com o resultado de Blastn e o score obtido são mostrados.

Número do Contig	Número deSNPs	Gene	Score
5054	49	<i>S. mansoni</i> mRNA for eggshell protein	1320
5064	48	<i>S. mansoni</i> mRNA for elongation factor 1-alpha	2531
5067	28	<i>S. mansoni</i> mitochondrial coding	2607
4464	26	A retrotransposon of the non-long terminal repeat class	152*
5076	25	<i>S. mansoni</i> fructose 1,6 biphosphate aldolase isoenzima	2464
4816	24	sem identidade significativa	-
5080	22	<i>S. mansoni</i> Saci-3 LTR retrotransposon mRNA	3538
5059	21	<i>S. mansoni</i> lactate dehydrogenase mRNA	896
5055	18	<i>S. mansoni</i> mitochondrial coding region	1223
5078	17	<i>S. mansoni</i> glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene -antígeno de superfície	1233
4582	15	sem identidade significativa	-D5
5029	14	<i>S.mansoni</i> cathepsin B (Sm31) mRNA, complete cds	2050
5049	13	<i>S. mansoni</i> 14-3-3 epsilon mRNA - receptors	712
4670	13	<i>S. mansoni</i> enzyme Cu/Zn superoxide dismutase mRNA	1273
4807	13	<i>S. mansoni</i> tegumental protein with similarity to cytoplasmic dynein light chains	811

Tabela III – Polimorfismos em genes candidatos à vacina, na seqüência consenso.

<i>Contig</i>	<i>Gene</i>	<i>Score do Blastn</i>	<i>SNP</i>	<i>Posição SNP</i>
4883	GST	1427	A/C	511
4883	GST	1427	A/C	570
4883	GST	1427	A/C	572
4883	GST	1427	A/C	144
5075	GST	954	G/C	648
5075	GST	954	A/C	761
4725	Sm23	2097	C-T	243
4725	Sm23	2097	A-T	783
4725	Sm23	2097	C-T	585
4725	Sm23	2097	A-T	858
4683	Sm23	2006	T-A	229
4683	Sm23	2006	T-A	230
4683	Sm23	2006	G-C	334
4683	Sm23	2006	G-C	403
4683	Sm23	2006	G-T	116
4995	Sm23	839	A-C	285
4995	Sm23	839	T-G	348
4995	Sm23	839	T-G	385
5015	Sm14	1231	C-G	383
5015	Sm14	1231	A-T	540
5015	Sm14	1231	C-A	526
5015	Sm14	1231	A-C	562
5015	Sm14	1231	T-C	591
4927	Sm14	700	A-C	110
4927	Sm14	700	A-C	480
4927	Sm14	700	A-C	237
4651	Sm14	664	T-G	468
4651	Sm14	664	T-G	141
5029	Catepsina B	2050	T-C	349
5029	Catepsina B	2050	T-C	523
5029	Catepsina B	2050	C-T	307
5029	Catepsina B	2050	A-C	697
5029	Catepsina B	2050	T-C	460

4.2 – Validação de SNPs no gene da Catepsina B de *S. mansoni*

A catepsina B de *S. mansoni* (GenBank #M21309) é uma cisteína peptidase, similar à de mamíferos, encontrada no estômago do parasito, cuja função é degradar a hemoglobina do hospedeiro, permitindo a nutrição e o desenvolvimento do *Schistosoma* (Sajid *et al.*, 2003). Além disso, a enzima possui carácter altamente antigênico, podendo ser utilizada em métodos de imunodiagnósticos (Noya *et al.*, 2001). Devido a esta função essencial, ela foi descrita como um dos possíveis candidatos à vacina (Loukas *et al.*, 2004). O gene contém 3 exons e 2 introns. A proteína codificada contém um domínio peptidase C1 (Figura 16). Utilizando-se iniciadores específicos que flanqueavam as regiões contendo possíveis SNPs, os fragmentos foram amplificados (Figura 17A) e seqüenciados em ambas as direções (Figura 17B). O total de 83 seqüências foram clonadas.

Foram identificados 16 SNPs, sendo 7 transições (44%) e 9 (56%) transversões (Figura 18). A distribuição da posição no códon foi 17%, 12% e 71% na primeira, segunda e terceira base, respectivamente (Figura 19), sendo 63% mutações sinônimas e 38% não-sinônimas (Figura 20). As mudanças de aminoácidos encontradas para as mutações não-sinônimas foram: Val21Ile, Glu27Lys, Lys75Thr, Asp84Glu, Asn92Ser e Gly101Arg, sendo que Lys75Thr e Asp84Glu encontram-se em regiões de reconhecimento da molécula. Foi identificado o sítio de corte do peptídeo sinal e a maioria dos SNPs foram encontrados após o sítio de clivagem, no domínio C1 (Figura 21).

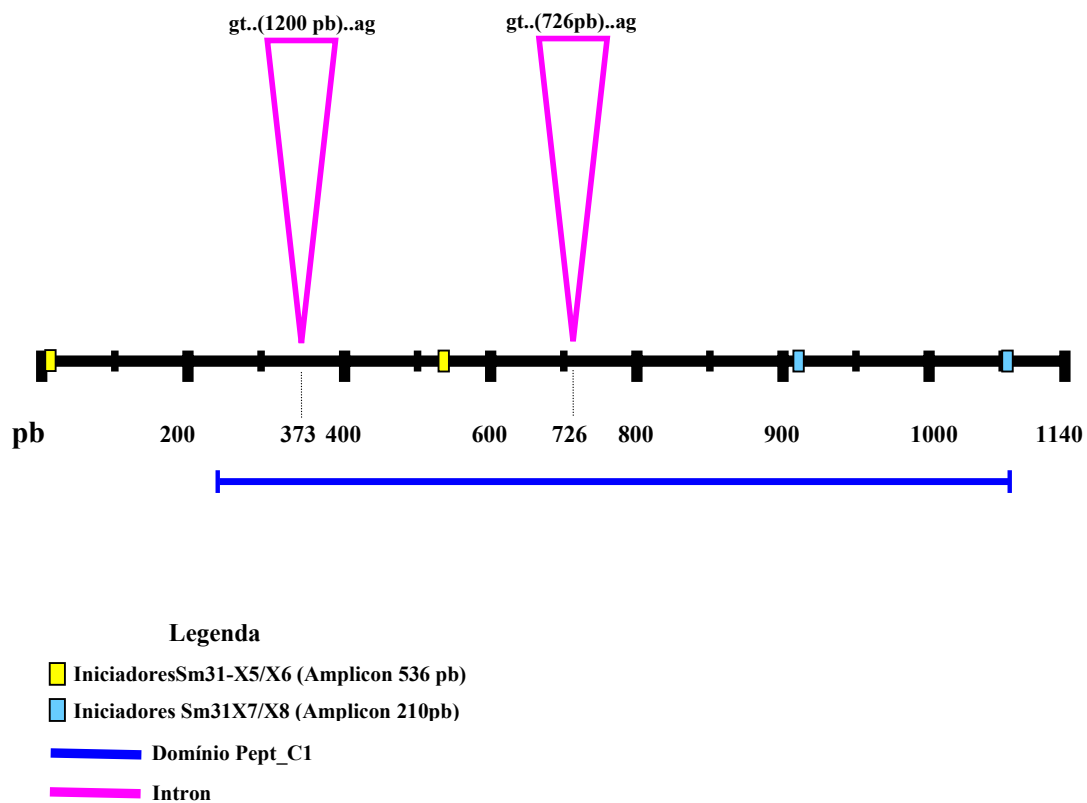


Figura 16 – Desenho esquemático representando o gene da Catepsina B de *S. mansoni*.

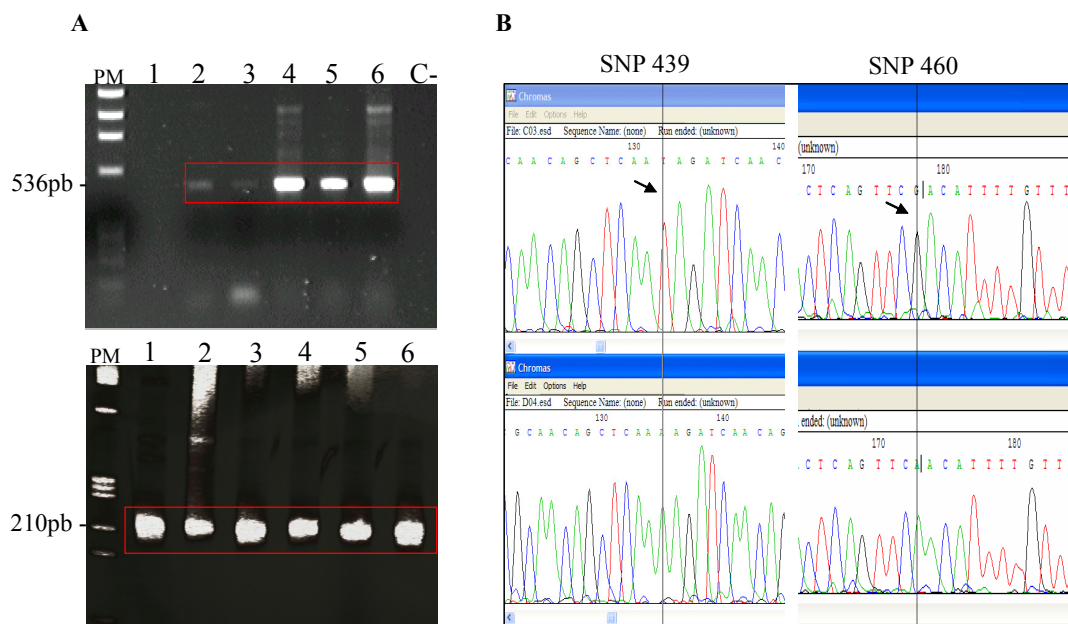


Figura 17 – Detecção de SNPs no gene da catepsina B de *S. mansoni* por seqüenciamento.

A – Gel de poliacrilamida 6% corado pelo brometo de etídeo, em destaque o produto amplificado de verme adulto de *S. mansoni* pela reação PCR utilizando os iniciadores Sm31-X5/X6 (536 pb) e Sm31-X7/X8 (210 pb). Canaleta PM, representa o padrão de peso molecular e a canaleta c- representa o controle negativo da reação de PCR; **B** – Exemplo de cromatogramas resultantes do seqüenciamento de clones de verme adulto de *S. mansoni*, representando o SNP na base 460 e o SNP na base 439, respectivamente (seta). A linha indica a base polimórfica

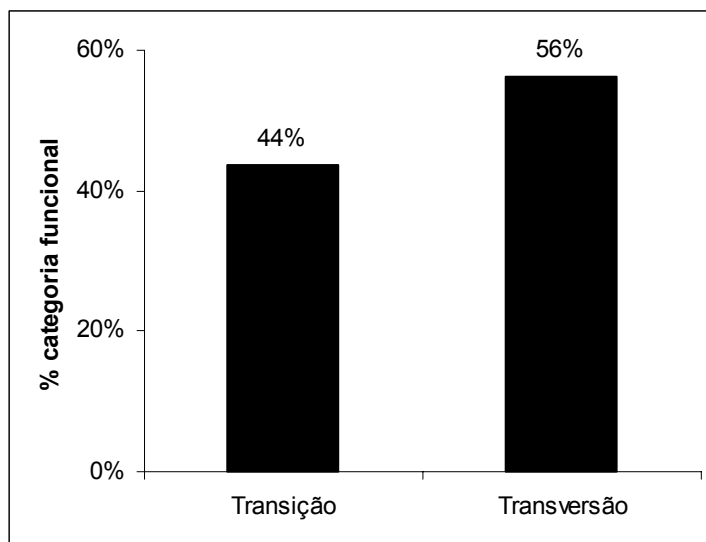


Figura 18 – Frequência dos SNPs no gene da catepsina B de *S. mansoni*, experimentalmente verificados, de acordo com a mudança da base nucleotídica, em transição ou transversão.

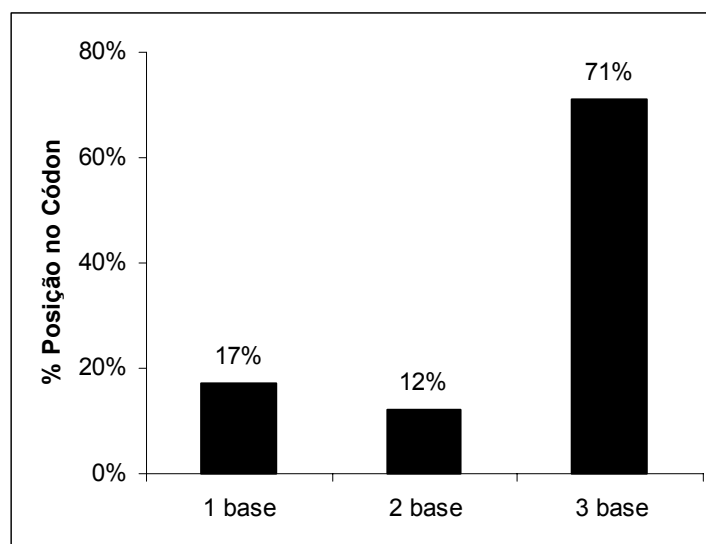


Figura 19 – Classificação dos SNPs, experimentalmente verificados, de acordo com a posição no códon.

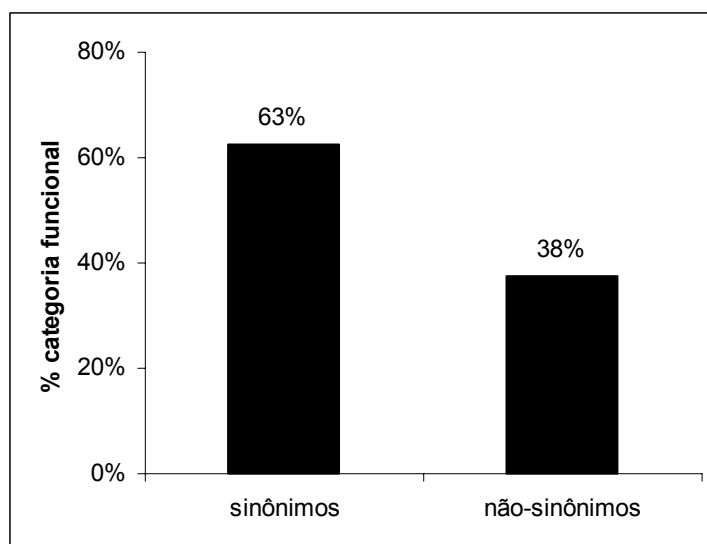


Figura 20 – Classificação dos SNPs, experimentalmente verificados, em sinônimos ou não-sinônimos, dependendo da conservação do aminoácido codificado.

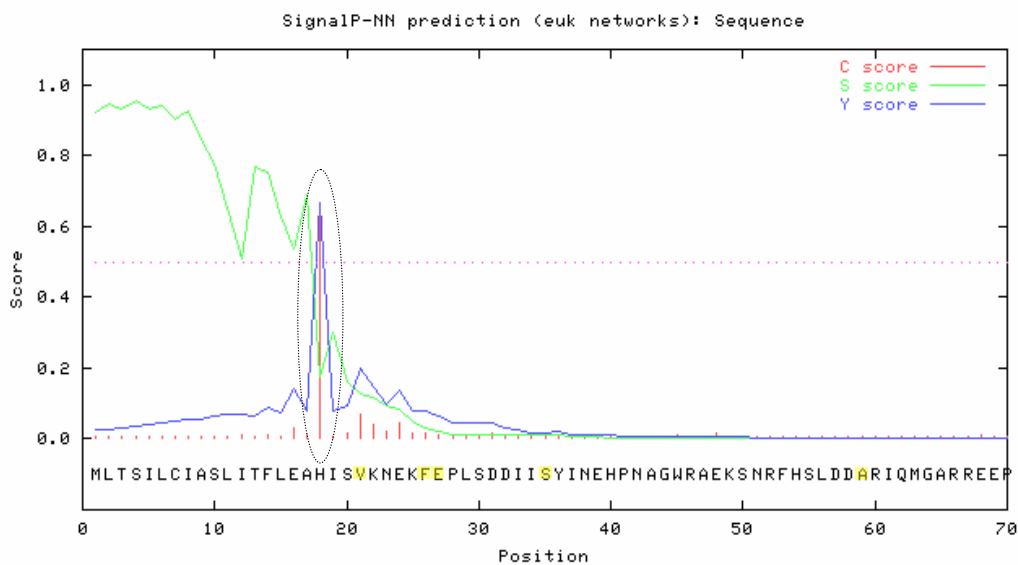


Figura 21 – Identificação do sítio de clivagem do peptídeo sinal no gene da catepsina B de *S. mansoni*. Todos os SNPs, representado alguns em amarelo na seqüência parcial, foram encontrados após o sítio de clivagem (círculo), no domínio peptidase C1.

4.3 – SNPs em cepas de campo de *S. mansoni*

Após identificação dos SNPs em vermes adultos de *S. mansoni* de cepa mantida em laboratório (LE), foi analisada a presença de polimorfismos em isolados de *S. mansoni* provenientes da área endêmica da região da vila de Caju no Vale do Jequitinhonha/M.G. Foram comparados 50 seqüenciamentos de diferentes isolados. O número de SNPs encontrados nos diferentes isolados foi o mesmo. Contudo, em isolados de campo foi observado uma frequência maior dos polimorfismos quando comparado com cepas mantidas em laboratório (Tabela IV).

Tabela IV – Frequência de SNPs em cepa de laboratório e isolados de campo.

SNP	Não-Sinônimos	Posição (pb)	LE	Isolados
A-G	X	80	49%	76%
T-A		82	53%	74%
T-C		97	55%	58%
G-A	X	98	56%	79%
C-A		109	44%	67%
G-A		119	64%	71%
A-C		196	48%	83%
T-C	X	243	69%	76%
A-T		265	72%	86%
C-A	X	271	35%	63%
G-C		292	59%	82%
G-A	X	294	82%	79%
T-C		295	62%	73%
A-T	X	320	41%	79%
A-T		387	55%	89%
G-A		459	63%	74%

4.4 – Modelagem da catepsina B de *S. mansoni*

Para realização da modelagem comparativa da catepsina B de *S. mansoni*, a proteína utilizada como modelo foi a procatepsina humana (gi|2982152|pdb|3PBH). Esta proteína apresentou identidade de 43,7% (similaridade 60%) sobre os 341 aminoácidos da seqüência da catepsina B de *S. mansoni*. As seqüências foram alinhadas pelo programa Modeller. O programa gerou um alinhamento que foi otimizado manualmente eliminando-se os primeiros 25 aminoácidos e um *gap* que não apresentaram alinhamento satisfatório (Figura 22). A eliminação dos aminoácidos na região amino terminal não comprometeu a modelagem da proteína. Não obstante, o primeiro SNP presente nessa região foi excluído. O segundo *gap* presente na seqüência, não foi eliminado do modelo, mas como havia um SNP nesta região este, também, não pôde ser modelado. Com isso, das 6 mutações não-sinônimas encontradas apenas 4 foram modeladas.

Segundo o modelo obtido, a estrutura da enzima catepsina B madura de *S. mansoni* contém dois domínios (domínios L e R), um propeptídeo, um sítio ativo localizado entre estes domínios, formando uma fenda em ‘V’ e uma alça (*loop*). A única diferença observada entre a estrutura da catepsina B ativa e a estrutura dos domínios enzimáticos na proenzima (inativa) é a conformação de uma alça (Ile 105 - Pro 126), responsável pela atividade de exopeptidase da enzima, presente acima da região do sitio ativo (Sajid & McKerrow 2002). A alça é flexível sendo capaz de adotar duas conformações diferentes. Isso só é possível devido à presença de uma ponte de sulfeto interna e dois pares de resíduos de prolina. Quando a enzima está inativa esta alça interage com a cadeia do propeptídeo, formando uma barreira no topo da estrutura, impossibilitando a entrada de substrato no sítio ativo. Esta conformação inativa da enzima faz com que os resíduos Cys 29 e His 199, presentes no sitio ativo, permaneçam próximos (Turk *et al.*, 1996). Contudo, o modelo gerado foi baseado na estrutura inativa da enzima (Figura 23), já que a estrutura da procatepsina B humana, utilizada como modelo, estava nessa conformação. A qualidade do modelo foi avaliada pelo gráfico de Ramachandram gerado pelo programa Sting, que apresentou 83,64% dos resíduos em regiões consideradas no gráfico como favoráveis, 14,5% em regiões aceitáveis e apenas 1,86% dos resíduos em regiões não permitidas (His 15, Asp 33, Glu 127, Thr 174, Asn 285) (Figura 24), sendo assim considerado um bom modelo.

***>P1;3pbh**

MRSR-----OS**F**HPLSDELVNYVNK-----RNTTWQAGHNFYNVDMS
 YLKRLCGTFLGGPKPPQRMFTEDLKLPA SF DAREQWPQCPTIKEIRDQGSCGSCWAFGAVEAISDRIC
 IHTNAHVSVEVSAEDLLTCCGSMCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGC RPYSIPPCEHHV
 NGSRPPCTGE*GDTPKCSKICEPGYSPTYKQDKHYGYNYSVSNSEKDIMAEIYKNGPVEGAFSVYSDF
 LLYKSGVYQHVTGEMMGHAI R I LGWGVENGT PYWLVANSWNTDWGDN GFFKILRGQDHC GIESEVVAG
 IPRTD

***>P1;CATHB**

MLTSILCIASLITFLEAHISVKNEK**F**EPLSDDIISYINEHPNAGWRAEKSNRFHSLDDARIQMGARREE
 PDLRRKRPTVDHND**D**WNVEIPS**N**FDSRKKWP**G**CKSIATIRDQSRCGSCWSFGAVEAMSDRSCIQSGGKQ
 NVELSAVDLLTCCESCGLGCEGGILGPAWDYWVKEGIVTASSKENHTGCEPYFPKCEHHTKGKYPCCG
 SKIYNTPRCKQTCQRKYKTPYTQDKHRGKSSYNVKNDEKAIQKEIMKYGPVEASFTVYEDFLNYKSGIY
 KHITGEALGGHAIRIIGWGVENKTPYWLIANSWNEDWGENGYFRIVRGRDECSIESEVIAGRIN

Figura 22 – Alinhamento (formato PIR) gerado pelo programa Modeller da procatepsina B humana (ref. >P1;3pbh) e da catepsina B de *S. mansoni* (ref. >P1;CATHB). Na seqüência da procatepsina B humana, tracejado em azul, estão os *gaps*. O primeiro deles foi eliminado da modelagem. Na seqüência da catepsina B de *S. mansoni*, sublinhado em roxo, estão os aminoácidos que foram eliminados da modelagem e em amarelo os aminoácidos analisados. O alinhamento final usado para a modelagem por homologia foi a partir do aminoácido F, em vermelho nas seqüências.

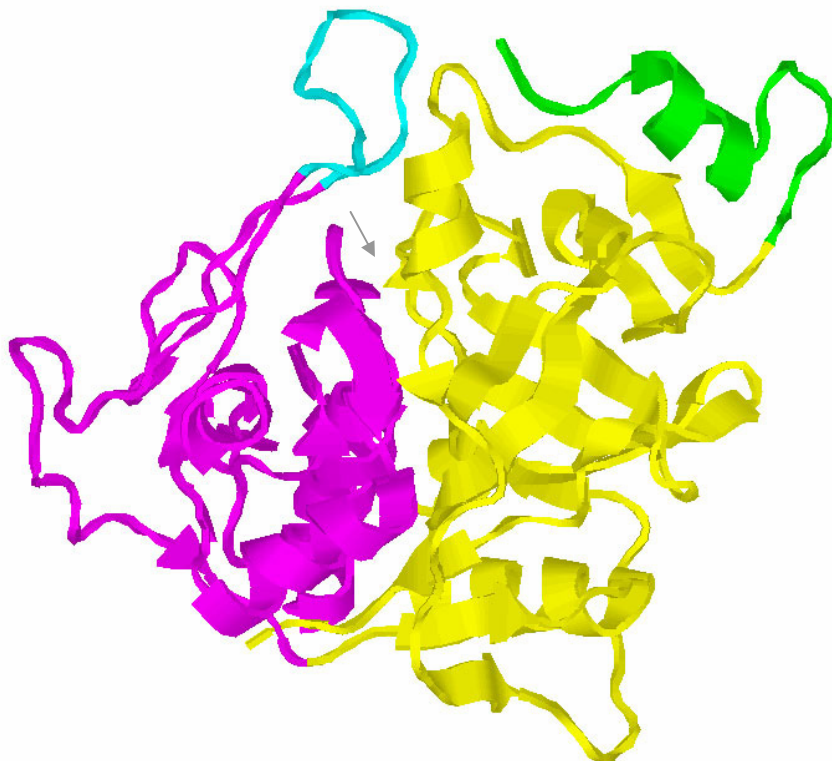


Figura 23 – Estrutura tridimensional da catepsina B de *S. mansoni*, na conformação inativa. As cores rosa (domínio L) e amarelo (domínio R) representam os domínios da proteína, em azul a região da alça e, em verde, o propeptídeo. A seta cinza indica a região do sítio ativo, presente entre os domínios

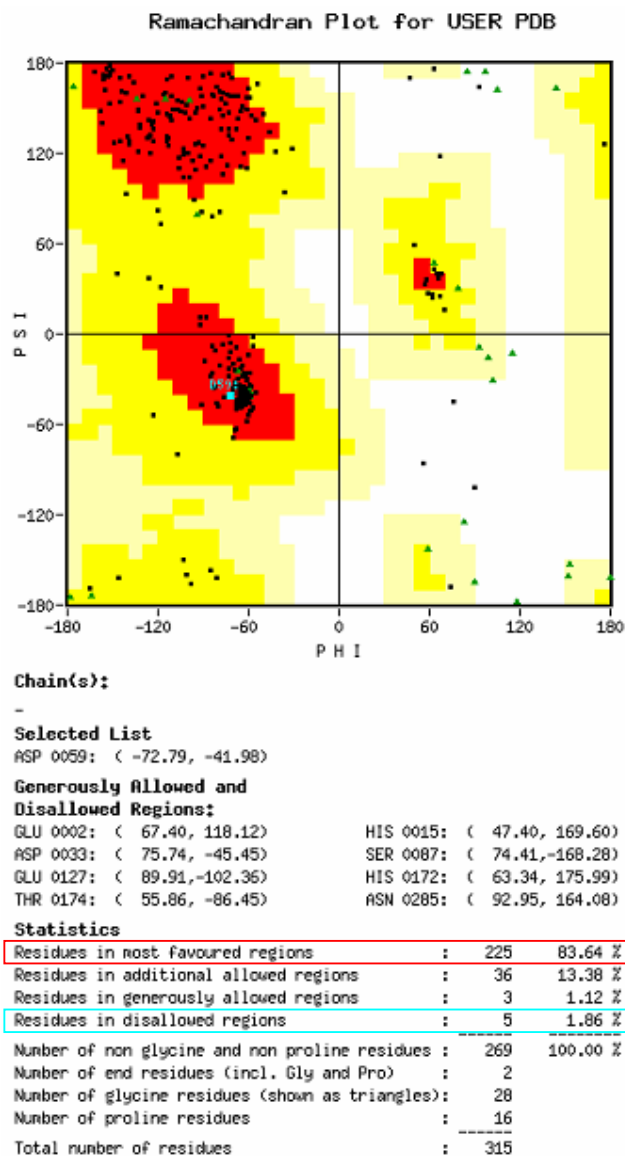


Figura 24 – Gráfico gerado pelo módulo Ramachandram do programa Sting. Este gráfico é referente à estrutura da catepsina B de *S. mansoni*, gerado por modelagem comparativa. O modelo apresentou 83,64% dos seus resíduos em regiões favoráveis (em vermelho) e apenas 1,86% dos resíduos em regiões não permitidas (em verde). Abaixo do gráfico está parte do arquivo gerado pelo programa.

As 4 mutações não-sinônimas detectadas foram analisadas quanto ao seu possível efeito sobre a estrutura tridimensional da catepsina B. As estruturas foram comparadas e analisadas, quanto à localização da mutação na estrutura, distância entre os aminoácidos e indução de novos contatos na presença do aminoácido original e do aminoácido variante.

O primeiro SNP (SNP 2) resultou na alteração de um ácido glutâmico (Glu 27), que é um aminoácido ácido com um grupo carboxílico na sua cadeia lateral, por uma lisina (Lys), que é um aminoácido básico que possui radical com o grupo aceno (Figura 25). Essa variação está localizada na região externa da estrutura (Figura 26). Apesar da diferença de carga entre os aminoácidos envolvidos nesta variação, verificado pelo leque do *Graphical Contacts* do programa Sting, nenhuma indução de novos contatos foi visualizada na presença do aminoácido variante, indicando que provavelmente nenhuma mudança possa ter ocorrido na estrutura dessa proteína (Figura 27A/B). Contudo, o módulo do programa considera interações em um raio de aproximadamente 7Å, que é um número acima do esperado para uma interação. Considerando que este é apenas um modelo, uma análise de dinâmica ou a determinação experimental faz-se necessária para analisar de fato alterações nos contatos intra-moleculares, uma vez que o aminoácido analisado está em contato com outros aminoácidos que possuem diferença de carga (Figura 28). Assim, uma pequena alteração na sua distância em relação a sua vizinhança pode estar causando interações de atração e/ou repulsão, que não puderam ser vistas.

A outra variação analisada (SNP 4) foi consequência da mudança de um ácido aspártico (Asp 84) por um ácido glutâmico (Glu) (Figura 29), ambos carregados negativamente e também localizados na região externa da estrutura (Figura 30). Pelo leque do *Graphical Contacts* foi observado que no modelo original, o aspartato interage atrativamente com outro aminoácido carregado, uma histidina 57. Essa mesma interação foi visualizada com o aminoácido substituído, provavelmente devido à posição da cadeia lateral do ácido glutâmico (Figura 31). Supõe-se assim que, provavelmente, a presença da mutação não causa uma alteração na estrutura da proteína por manter a mesma interação na presença de ambos os aminoácidos. Através da figura do módulo ConSSeq (Figura 32), gerado pelo Sting, foi visto que através do alinhamento de várias proteínas, a segunda opção encontrada na natureza para esse resíduo de aminoácido é um ácido glutâmico, como o que foi encontrado nos resultados. Em 35% das seqüências alinhadas foi encontrado o aminoácido Asp enquanto em 19% dos casos o aminoácido Glu foi encontrado.

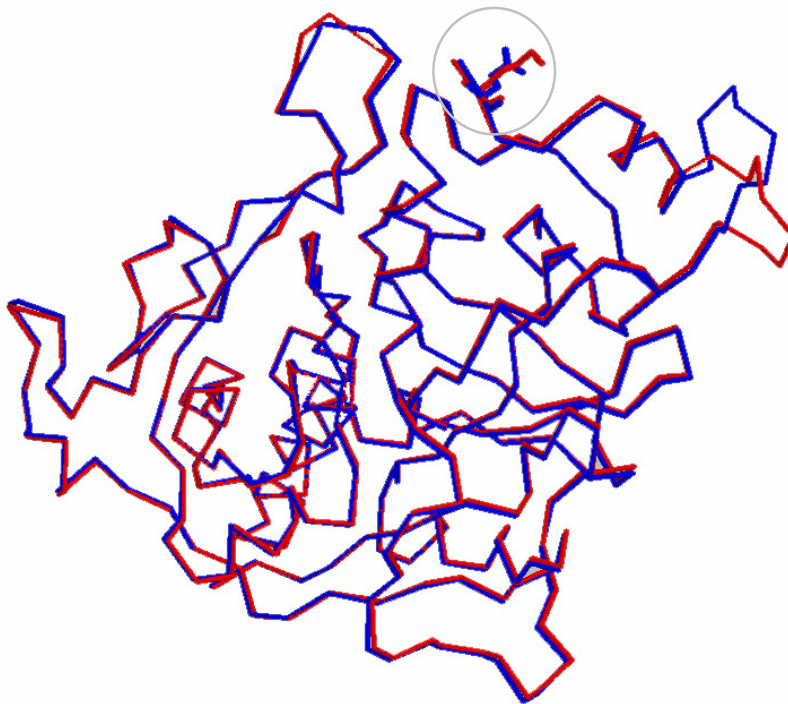


Figura 25 – Superposição da estrutura tridimensional da proteína modelo e da proteína contendo o SNP 2 (mudança nucleotídica G-A) que resulta na substituição do aminoácido Glu(E)27Lys(K). Em azul está a estrutura original e em vermelho a variante. O círculo em cinza indica a posição da variação.

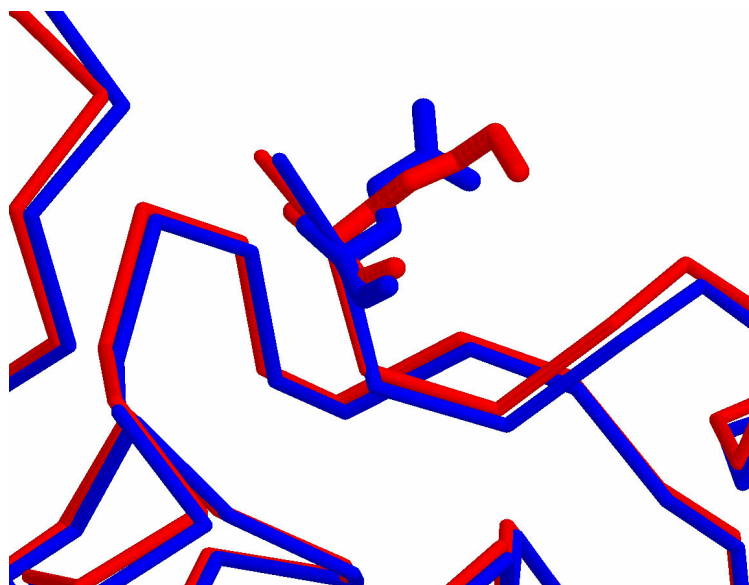


Figura 26 – Aproximação da imagem dos aminoácidos presentes na estrutura tridimensional da proteína original e da proteína contendo a substituição (SNP 2).

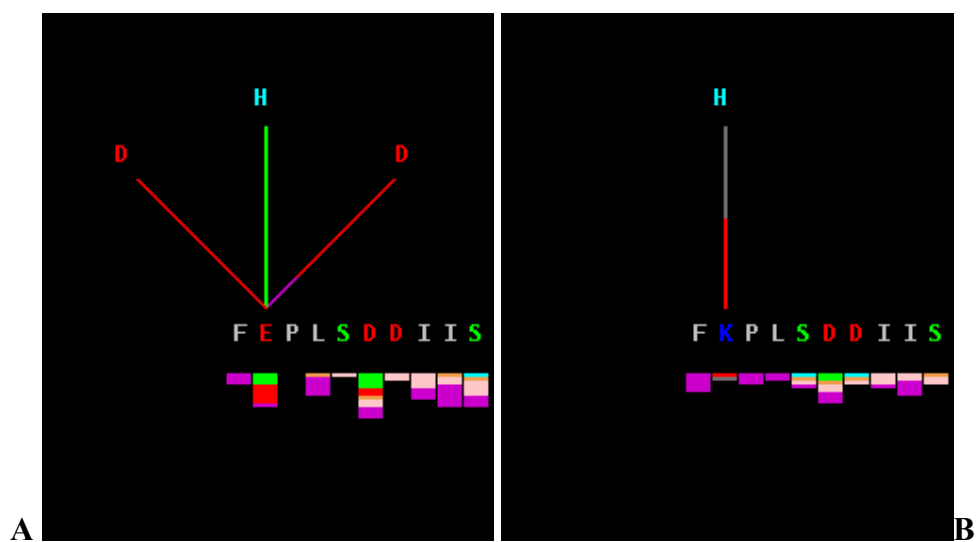


Figura 27 – Gráficos gerados pelo módulo *Graphical Contacts* do programa Sting para o SNP 2. **A** - Este gráfico mostra as interações do aminoácido original Glu 27 (E) com aminoácidos localizados a um raio de até, aproximadamente, 7Å. **B** - Esse gráfico demonstra a interação na presença do aminoácido variante, indicando que não houve indução de novos contatos.

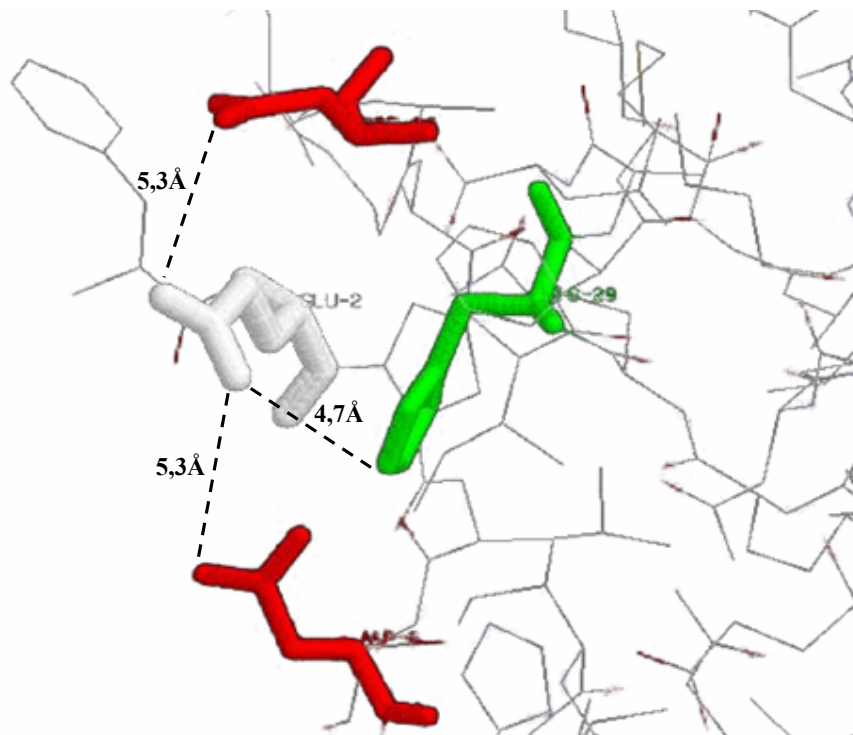


Figura 28 – Interações e a distância do aminoácido Glu 27 (em branco) com os aminoácidos da sua vizinhança de acordo com o *Graphical Contacts* do programa Sting. Pode ser observada, uma interação atrativa com a His 29 (em verde) e uma interação repulsiva com uma Asp 6 e uma Asp 32.

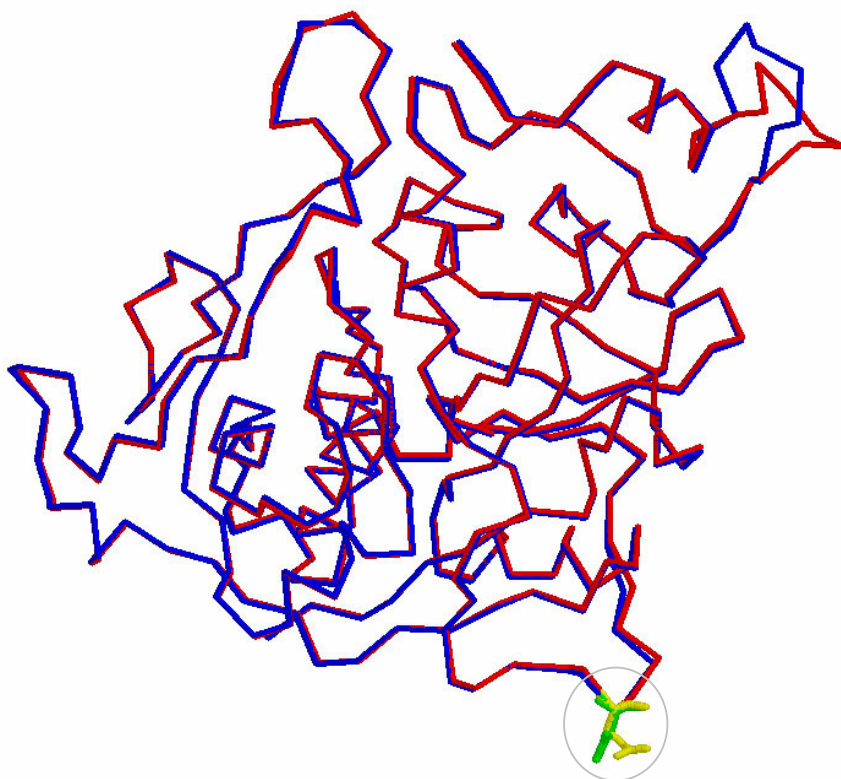


Figura 29 – Superposição da estrutura tridimensional da proteína modelo e da proteína contendo o SNP 4 (mudança nucleotídica T-A), que resulta na substituição do aminoácido Asp(D)84Glu(E). Em verde está o aminoácido original e em amarelo o variante. O círculo em cinza indica a posição da variação.

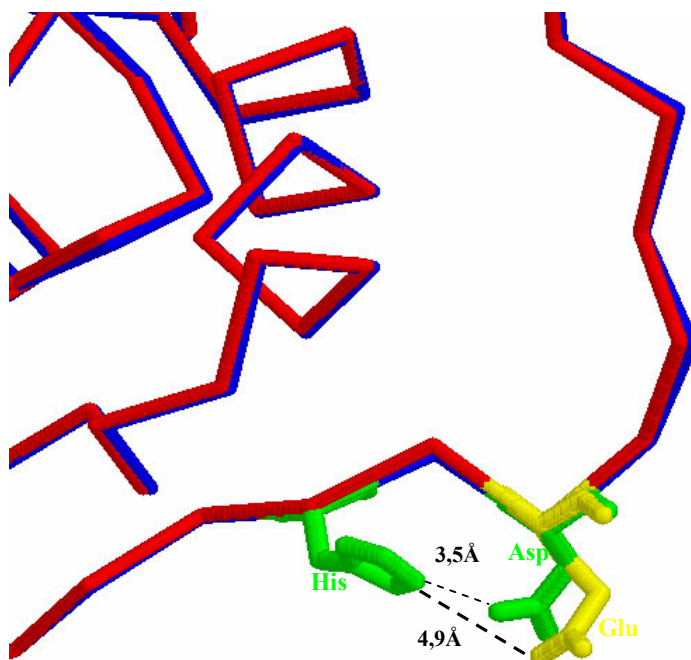


Figura 30 – Aproximação da imagem dos aminoácidos na estrutura tridimensional da proteína original e proteína contendo a substituição (SNP 4). Em verde, o aminoácido original Asp 84 interage com a His 57. E em amarelo, o aminoácido variante Glu 84 mantém a interação com a His 57.

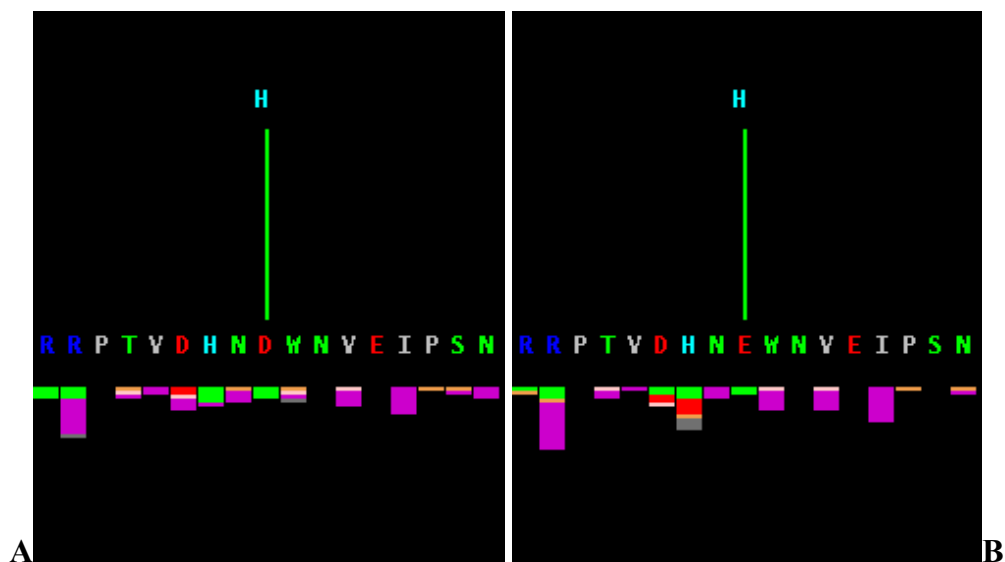


Figura 31 – Gráficos gerados pelo módulo *Graphical Contacts* do programa Sting para o SNP 4. **A** - Asp 59 faz uma ligação com o aminoácido His 57; **B** - Após a mudança do aminoácido presente na estrutura original pelo aminoácido variante, a interação com a His 57 é mantida, não havendo indução de novos contatos.

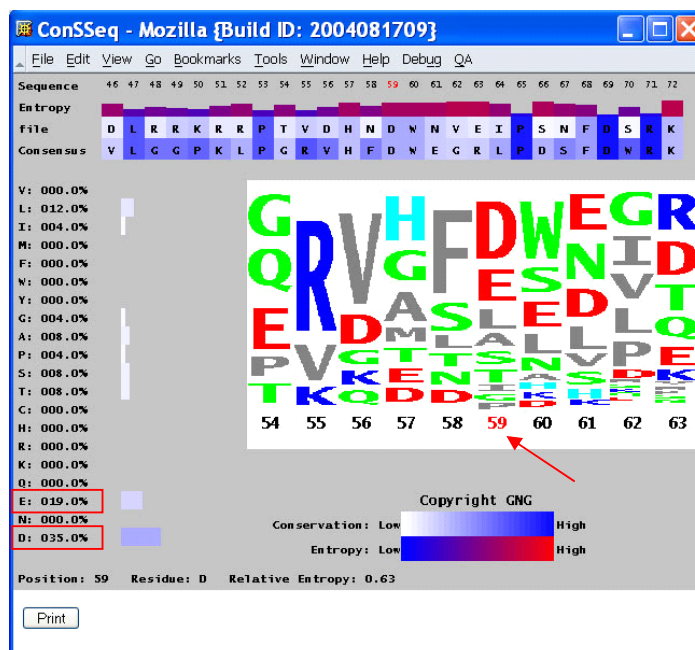


Figura 32 – Módulo ConSSeq: conservação do SNP 4. A coluna do meio é respectiva ao local do aminoácido Asp 84. Na seta, a numeração aparece como 59 devido à eliminação dos 25 aminoácidos iniciais. Como verificado na parte esquerda da figura, em 35% das seqüências alinhadas foi encontrado o aminoácido Asp, enquanto que, em 19% dos casos o aminoácido foi encontrado o aminoácido Glu (retângulo vermelho).

O outro SNP (SNP 5) encontrado resultou na mudança de uma asparagina (Asn 92) por uma serina (Ser), ambos os aminoácidos polares neutros que tendem a formar pontes de hidrogênio (Figura 33). Como a mutação foi encontrada em uma região N-terminal da folha beta da proteína (Figura 34), ela poderia causar uma possível mudança conformacional. Contudo, nenhuma ligação de hidrogênio é realizada entre este aminoácido e a fita beta vizinha. Além disso, pelo leque do *Graphical Contacts* nenhuma alteração das interações originais foi observada na presença do aminoácido variante (Figura 35). Também foi verificado que o aminoácido serina é muito conservado nesta região de acordo com o alinhamento múltiplo de seqüências gerado pelo módulo ConSSeq (Figura 36).

A outra mutação analisada (SNP 6) está localizada em uma alça que liga os dois domínios da estrutura entre duas pequenas hélices (ver Figura 19 e 37), sendo decorrente da variação de um aminoácido apolar, no caso a glicina (Gly 101) por um aminoácido básico, a arginina (Arg) (Figura 38). Para visualizar essa mutação, foi necessário girar o modelo da proteína 180 graus no eixo vertical em relação às figuras anteriores. Embora seja uma mudança de um aminoácido pequeno, sem cadeia lateral como a glicina, por um aminoácido com cadeia lateral grande e carregada, como a arginina, o modelo parece não ter sido alterado, já que a cadeia lateral da arginina ficou posicionada para o lado externo da estrutura (Figura 39). Além disso, segundo mostra o leque obtido com o módulo *Graphical Contacts* do programa Sting, o resíduo glicina 101, que estava no modelo original faz uma ligação de Hidrogênio com a cadeia principal do triptofano 74 (Figura 40A). Essa ponte de hidrogênio provavelmente é o que garante a alça responsável por causar a mudança na direção da cadeia e dá início à formação do segundo domínio da proteína. No entanto, a ligação de hidrogênio entre a cadeia principal do aminoácido triptofano foi mantida mesmo após a Glicina ter sido substituída pela arginina. Sendo assim, a conformação que depende dessa ligação de hidrogênio foi mantida mesmo com a mudança dos resíduos de aminoácido, sugerindo que provavelmente a presença dessa mutação também não altera a estrutura tridimensional da proteína. Contudo, uma nova interação hidrofóbica foi observada entre a Arg 101 e a Gly 111. Embora as ligações hidrofóbicas sejam de fundamental importância para o dobramento das proteínas, esta é a ligação que possui menor energia (0,6Kcal) comparada às outras. Assim, normalmente existe uma rede de interações hidrofóbicas e apenas uma mudança é considerada uma energia muito baixa para de fato mudar a conformação local (Figura 40B).

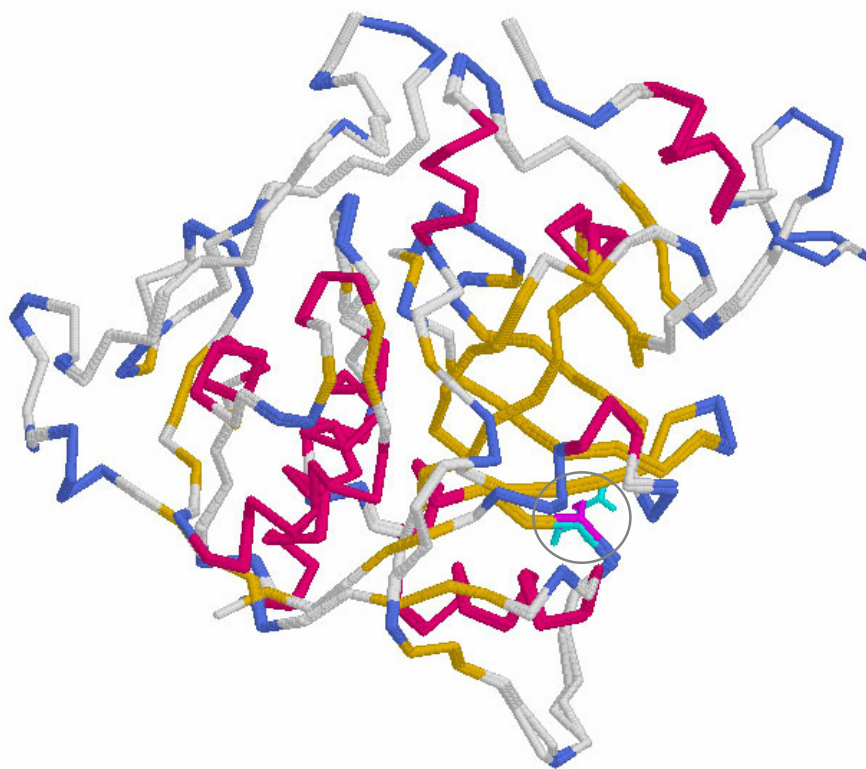


Figura 33 – Alinhamento do modelo da catepsina B mostrando, também, a estrutura secundária do modelo. As partes em vermelho são regiões de alfa-hélice e, em amarelo, estão as regiões de fita beta. Na cor verde claro, está o aminoácido Asn 92 original e, na cor rosa, está o aminoácido variante Ser. O círculo cinza indica a posição da variação do SNP 5 (mudança nucleotídica A-G).

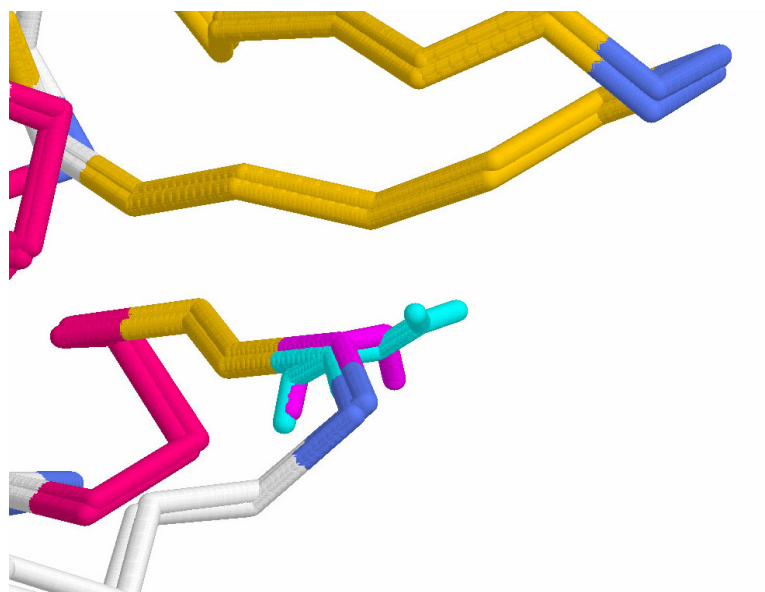


Figura 34 – Aproximação da imagem contendo a substituição Asn(N)92Ser(S). Em verde claro, está a estrutura original e, em rosa, a variante.

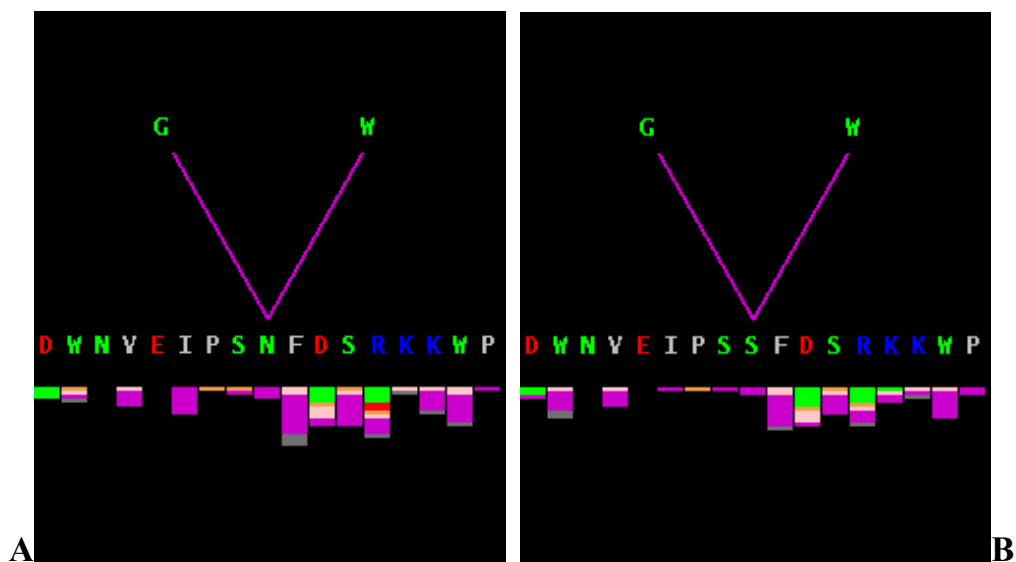


Figura 35 – Gráficos gerados pelo módulo *Graphical Contacts* do programa Sting para o SNP 5. **A** – Esse gráfico mostra que existe uma interação hidrofóbica entre o aminoácido original Asn 92 com dois outros 2 aminoácidos, a Gly 270 e o Trp 269; **B** – As mesmas ligações foram mantidas na presença do aminoácido variante Ser.

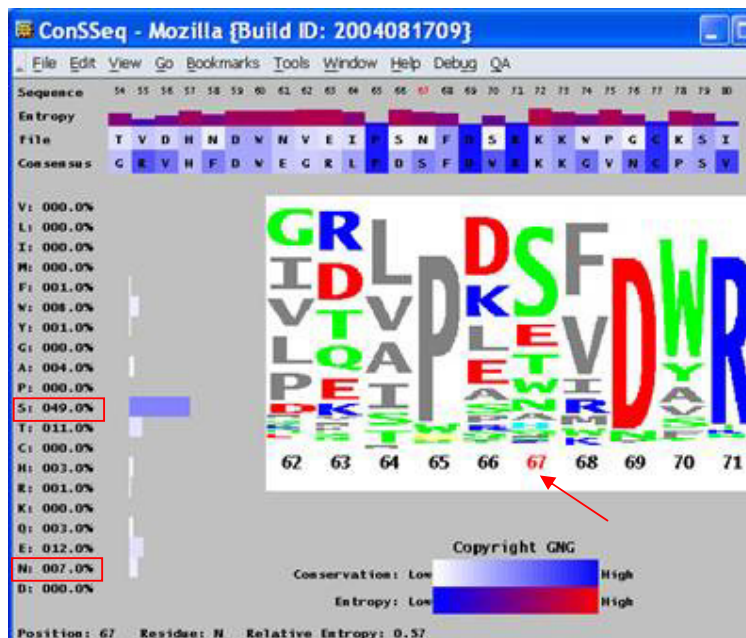


Figura 36 – Módulo ConSSeq: conservação do SNP 5. A coluna do meio é respectiva ao local do aminoácido Asn 92. Na seta, a numeração aparece como 67 devido à eliminação dos 25 aminoácidos iniciais. Como verificado na parte esquerda da figura, o aminoácido serina é muito conservado nessa posição, sendo encontrado em 49% das seqüências alinhadas, enquanto que o aminoácido Asn foi encontrado em somente 7% dos casos (retângulo vermelho).

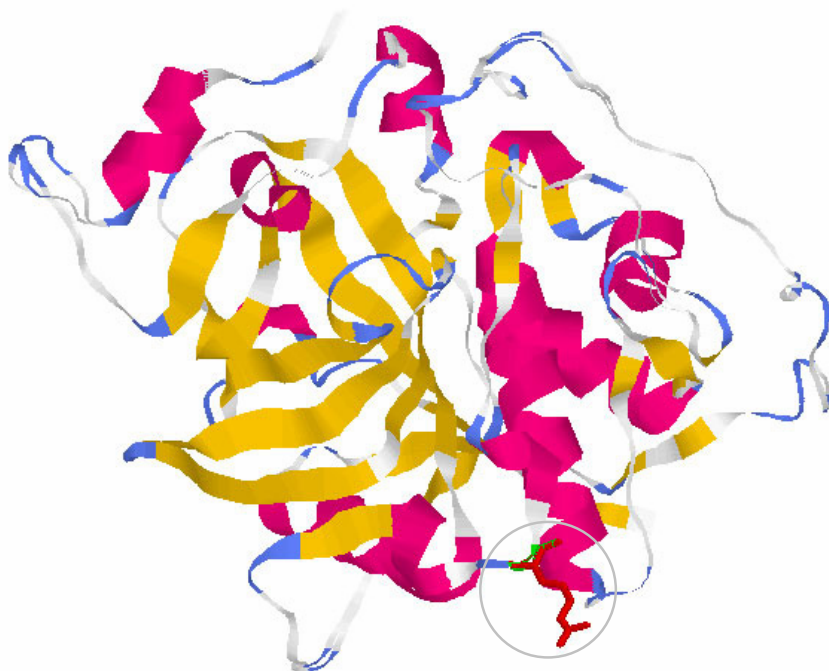


Figura 37 – Superposição da estrutura tridimensional da proteína original e da proteína contendo o SNP 6 (mudança nucleotídica G-C), que resulta na substituição do aminoácido Gly(G)101Arg(R). Em verde está o aminoácido original e, em vermelho, o variante. O círculo cinza indica a posição da variação.



Figura 38 – Aproximação da imagem contendo os aminoácidos na estrutura tridimensional da proteína original e na proteína contendo a substituição. Em verde está a Gly 101 e, em vermelho, a Arg.



Figura 39 – Ligação da glicina 101 com o triptofano 74. Na figura, o resíduo glicina (em verde), presente no modelo original, faz uma ligação de hidrogênio (3,16Å) com a cadeia principal do triptofano 74 (em amarelo).

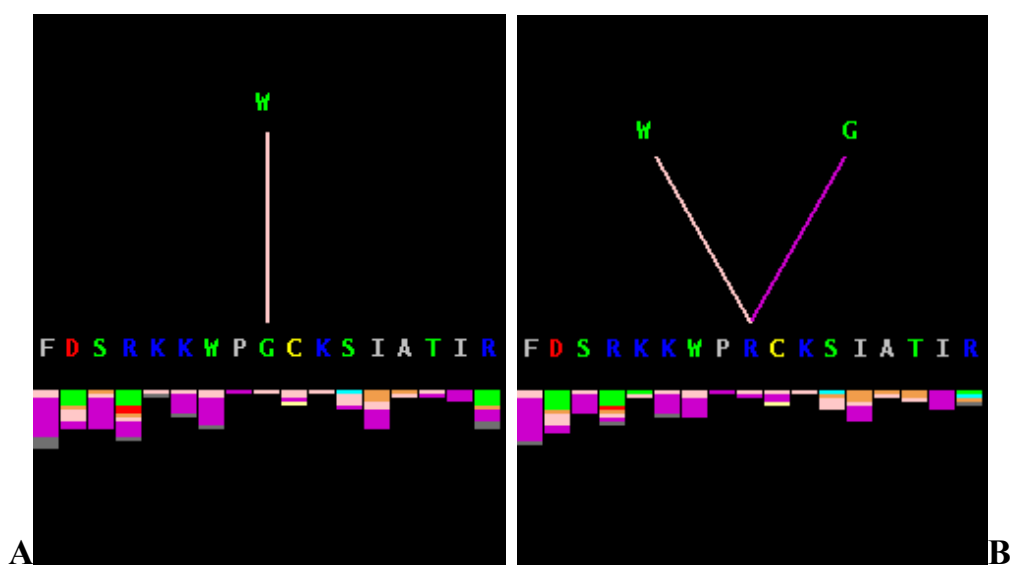


Figura 40 – Figuras geradas pelo módulo *Graphical Contacts* do programa Sting para o SNP 6. **A** - Este gráfico mostra que existe uma ligação de hidrogênio entre a cadeia principal da glicina 101 com a cadeia principal do triptofano 74, que segundo a tabela de contatos essa ligação ocorre entre os átomos N da glicina e O do triptofano; **B** - Mesmo após a mutação, a ligação de hidrogênio é mantida entre o aminoácido variante Arg 101 e cadeia principal do triptofano 74.

V - DISCUSSÃO

A crescente disponibilização de seqüências de genomas completos permitiu o advento da genômica comparativa. Desde o seqüenciamento completo do primeiro organismo procarionte (bactéria *Haemophiluz influenzae*, com aproximadamente $1.830.10^3$ pb), em 1995 (Fleischmann *et al.*, 1995), o número de organismos cujos genomas estão totalmente seqüenciados já alcança algumas centenas e vários outros projetos de seqüenciamento encontram-se atualmente em andamento. Com tamanha fonte de informação e dificuldade na manipulação de grande quantidade de dados, um conjunto adequado de ferramentas foi desenvolvido com o uso de técnicas estatísticas e computacionais sofisticadas que permitiram a análise e entedimento biológico dos dados produzidos. A integração de diferentes áreas, a biologia molecular, a estatística e a computação, permitiu o surgimento de um novo campo de estudo na última década, a bioinformática (Kanehisa & Bork 2003).

A princípio só era possível gerar um pequeno número de seqüências de DNA. Contudo, o desenvolvimento de métodos mais avançados permitiu o seqüenciamento em larga escala, conseqüentemente, a necessidade de processamento computacional. Em *S. mansoni* existe, atualmente, uma iniciativa internacional para o seqüenciamento genômico do parasito através da técnica de *Shotgun* e duas grandes iniciativas de seqüenciamento do transcriptoma. A primeira iniciativa do transcriptoma do parasito, correspondente ao grupo ONSA, gerou 124.640 ESTs, sendo que 23% corresponderam a genes conhecidos e 77% a genes novos. A outra iniciativa, organizada pela rede Minas que gerou aproximadamente 60.000 novas ESTs, ainda não disponíveis em banco de dados. Nesse contexto, nosso principal objetivo foi descrever uma metodologia automatizada de identificação de SNPs em larga escala. SNPs são marcadores que ainda não haviam sido amplamente estudados em *S. mansoni*. No presente trabalho, as ESTs utilizadas para a detecção de SNPs foram geradas pelo Projeto Genoma de Minas Gerais. Vários grupos descreveram com sucesso a utilização de ESTs na identificação de SNPs (Picoult-Newberg *et al.*, 1999; Gu *et al.*, 1998; Halushka *et al.*, 1999; Evans & Relling 1999; Useche *et al.*, 2001). Contudo, o maior problema de se usar ESTs é a baixa qualidade das seqüências, especialmente quando os cromatogramas dessas seqüências não estão disponíveis para avaliação (Picoult-Newberg *et al.*, 1999). Erros em ESTs são, normalmente, conseqüência da baixa qualidade do seqüenciamento ou causados pela enzima transcriptase reversa, durante a construção de bibliotecas de cDNA. Não obstante, a grande vantagem do uso de ESTs na identificação de SNPs é a eliminação da necessidade de ter que gerar seqüências completas de DNA e uma associação imediata com genes expressos (Risch 2000, Jalving *et al.*, 2004). Além disso, a confirmação do SNP agrega maior valor nas

informações contidas nas ESTs (Smith *et al.*, 2001), uma vez que os SNPs são marcadores polimórficos de grande utilidade (Qiu *et al.*, 2004; Halldorsson *et al.*, 2004).

Na primeira etapa do presente trabalho foi necessário selecionar uma ferramenta para agrupar as EST de *S. mansoni*. Todas as ferramentas de montagem recebem como entrada um arquivo contendo seqüências em formato FASTA e, em muitos casos, um arquivo com informações sobre a qualidade das seqüências e ao final da execução é gerado um arquivo de saída contendo os contigs, que são resultantes do agrupamento de fragmentos de seqüências de DNA similares que, idealmente, correspondem a um único gene. Contudo, apesar dessas ferramentas seguirem basicamente os mesmos princípios, elas possuem algoritmos diferentes que as fazem abordar características distintas, gerando montagens com diferenças. Devido a isso, o processo de montagem foi uma etapa complexa e mais de uma ferramenta foi testada, com o objetivo de se obter o melhor resultado para os nossos dados. O programa escolhido para a montagem das 61.002 ESTs de *S. mansoni* foi o Phrap (<http://www.phrap.org>). Este programa é um dos mais utilizados atualmente e foi utilizado na montagem de ESTs em vários trabalhos publicados (Useche *et al.*, 2001; Kim *et al.*, 2003; Dantec *et al.*, 2004). O Phrap utiliza informações relacionadas à qualidade dos dados de entrada, permitindo assim calcular a probabilidade das bases terem sido erroneamente identificadas, possibilitando uma seqüência consenso mais próxima da seqüência original. A utilização de uma ferramenta com esta característica foi uma grande vantagem para o nosso trabalho, já que aumentou a confiabilidade dos polimorfismos encontrados serem reais e não erros de seqüenciamento, uma vez que não foram utilizadas seqüências públicas sem informação sobre a qualidade das bases identificadas. Além disso, o Phrap utiliza o fragmento por inteiro na montagem, ao contrário de outras ferramentas que mascaram os fragmentos de entrada com o intuito de eliminar seqüências de baixa qualidade. Este processo tende a excluir uma quantidade significativa de dados da montagem inicial, exigindo mais tarde um esforço manual para a montagem completa dos grupos. O Phrap também determina os pares de fragmentos cujas sobreposições não são relevantes e detecta a presença de fragmentos quimeras (junções de dois ou mais fragmentos provenientes de regiões completamente distintas daquela molécula), evitando assim erros de montagens no final do processo (Ewing & Green 1998). Apesar do satisfatório desempenho desta ferramenta de montagem, um dos fatores limitantes na utilização da técnica é a necessidade de memória disponível (> 2 Gb) e um número limitado de ESTs na realização dos agrupamentos (até no máximo 64.000 ESTs).

Algumas ferramentas públicas de busca de SNPs foram testadas. Contudo, a maior dificuldade encontrada na execução desses programas foi a otimização dos parâmetros de busca e análise dos arquivos gerados, devido à escassez de documentação sobre a metodologia exata para a determinação do SNP (Useche *et al.*, 2001). Por este motivo, um novo programa de busca de polimorfismos foi escrito pelo nosso grupo, chamado cSNPer, permitindo um conhecimento claro dos critérios de indentificação e otimização dos parâmetros. O programa levou em consideração o valor de qualidade das bases polimórficas na EST (Phred ≥ 20), na seqüência consenso (Phred ≥ 40) e um alinhamento de 10 pb das bases vizinhas, com Phred ≥ 15 . Estes valores são manipuláveis e podem ser adaptados de acordo com cada objetivo. O valor de Phred estima a probabilidade do número de bases incorretas, sendo para Phred 20, a probabilidade de encontrar uma base incorreta a cada 100 pb e para phred 40, uma base incorreta a cada 10.000 pb (Ewing *et al.*, 1998). Com isso, não só a qualidade da base foi analisada como, também, da região em que ela se encontrava. Além disso, através de experimentos controlados (vide apêndice), o programa foi testado para cada parâmetro e validado.

Os SNPs detectados foram classificados como transição e transversão dependendo da variação do nucleotídeo. Ao contrário do que é mencionado na literatura, foi detectado, no presente trabalho, um valor maior de variações do tipo trasversão. Uma explicação provável para esta observação é que transições em grande número deve-se a um elevado conteúdo de G+C no genoma e ao processo de deaminação da 5-metilcitosina em timidina (Fryxell & Moon 2004). Não obstante, o genoma de *S. mansoni* possui um índice baixo de conteúdo G+C (~30%) (Marx *et al.*, 2000) e o seu genoma não é metilado (Fantappie *et al.*, 2001).

SNPs também são classificados de acordo com a mudança do aminoácido codificado em mutações sinônimas ou não-sinônimas. Alguns trabalhos descrevem uma frequência maior de mutações sinônimas comparada às mutações não-sinônimas (Kim *et al.*, 2003; Fitzsimmons *et al.*, 2004). Uma explicação para esta observação é o maior conteúdo G+C no genoma de diferentes organismos e o maior número de transições (Sachidanandam *et al.*, 2001). Por exemplo, o parasito *Plasmodium vivax* possui 55% do seu genoma de conteúdo A+T e apresentou a razão entre mutações não-sinônimas/sinônimas igual a 1,75 (Feng *et al.*, 2003). O parasito *Plasmodium falciparum* possui 80% de seu genoma de conteúdo A+T e a razão encontrada foi 2,34 (Feng *et al.*, 2003). Já no genoma humano, que apresenta o índice de conteúdo G+C elevado, a razão encontrada foi 0,89 (Cargill *et al.*, 1999). *S. mansoni* possui, aproximadamente, 66% do seu genoma de conteúdo A+T e foi verificado, no presente

trabalho, que a razão entre mutações não-sinônimas/sinônimas foi igual a 2. Acreditamos, portanto, que as observações feitas neste trabalho encontram-se dentro do esperado.

Após a detecção geral dos polimorfismos nas ESTs, o próximo objetivo foi analisar o perfil polimórfico dos genes candidatos à vacina de *S. mansoni*. Na maioria dos genes foi detectada a presença de SNPs, exceto nos genes codificantes para paramiosina, miosina e TPI. É extremamente importante o conhecimento sobre o perfil polimórfico de genes candidatos a vacina ou alvos de droga, uma vez que a presença de mutações não-sinônimas podem causar uma alteração na estrutura da proteína codificada por este gene e, conseqüentemente, alterar a sua função. Entretanto, vale a pena ressaltar que as mutações sinônimas, mesmo que raramente, podem de certa forma estar alterando também o processo de transcrição destes genes, como por exemplo, causando alteração do sitio de *splicing* ou redução da solubilidade da proteína codificada (Stitzel *et al.*, 2004).

Apesar da indiscutível contribuição da bioinformática na identificação de SNPs, a necessidade de validação experimental dos dados é indispensável para avaliar o potencial informativo de cada polimorfismo. A partir disso, escolhemos um gene candidato à vacina e comprovamos a presença de SNPs experimentalmente. O gene escolhido para a validação dos polimorfismos foi o gene da catepsina B de *S. mansoni* (GenBank #M21309). Esta enzima foi descrita como sendo a mais abundante cisteína protease secretada no lúmen do intestino de *Schistosoma*, presente em células gastrointestinais e tem papel essencial no desenvolvimento e na reprodução do parasito. O verme utiliza esta enzima para degradar hemoglobinas que estão presentes nas células vermelhas do sangue do hospedeiro, durante sua alimentação (Brindley *et al.*, 1997). A digestão de hemoglobinas é considerada a fonte de nutrientes principal utilizada por *S. mansoni*. Foi demonstrado que a interrupção na atividade dessa enzima pode limitar a habilidade do parasito durante a alimentação e reprodução (Sajid *et al.*, 2003). Outros estudos utilizando inibidores específicos de cisteínas proteases demonstraram uma diminuição significativa no desenvolvimento do verme e uma redução na produção de ovos pelas fêmeas em camundongos infectados por *S. mansoni* (Wasilewski *et al.*, 1996). A técnica utilizada para a identificação dos SNPs foi o seqüenciamento. Apesar das inúmeras técnicas modernas já descritas para identificação de SNPs como PCR em tempo real e microarranjos, entre outras, o seqüenciamento foi escolhido por ser uma técnica já empregada em nosso laboratório, eficiente na detecção de polimorfismos e que não necessitava da compra de novos equipamentos. No presente trabalho, utilizamos inicialmente amostras extraídas de DNA genômico de verme adulto para realização da PCR. Contudo, o produto

amplificado utilizando os iniciadores Sm31-X5/X6, não apresentavam o tamanho de banda esperado. A banda amplificada deveria apresentar 536 pb, porém o tamanho que estava sendo observado apresentava aproximadamente 1.700 pb. As hipóteses levantadas para a obtenção deste perfil foi a amplificação de isoformas ou a presença de introns. A primeira hipótese foi descartada, uma vez que os iniciadores foram desenhados em regiões específicas da enzima Sm31, impossibilitando a amplificação de um outro produto. Para a segunda hipótese foi preciso analisar a seqüência genômica do gene. A seqüência da enzima depositada no banco de dados (NCBI) por Klinkert e colaboradores em 1989 foi definida a partir de clones de cDNA, logo não continha informações sobre regiões intergênicas ou seqüências de introns. Através da montagem e comparação dos fragmentos da seqüência, por blastn, contra seqüências genômicas depositadas em bancos de dados (TIGR- <http://www.tigr.org/>), foi possível detectar a presença de duas regiões de introns na seqüência do gene da catepsina B de *S. mansoni*. Como esperado, um intron estava presente na região amplificada pelo iniciador Sm31-X5/X6, e o outro intron em uma região não polimórfica (Figura 12). Desta forma, foram utilizadas amostras de cDNA para amplificação das amostras e obtenção da banda de peso molecular esperado.

Uma vez comprovada a presença de SNPs em cepas de laboratório foi verificada a presença de polimorfismos em cepas provenientes do campo (Caju-Vale do Jequitinhonha). Trabalhos utilizando marcadores microssatélites já demonstraram uma diversidade genética e heterozigose maior em populações de cepas de campo quando comparadas com cepas de laboratório de *S. mansoni* (Rodrigues *et al.*, 2002). Um claro exemplo de perda alélica foi demonstrado por Norris e colaboradores (2001). O grupo utilizou marcadores do tipo microssatélites e demonstraram que a diversidade dos alelos em amostras de campo do mosquito *Anopheles gambiae*, agente causador da malária em humanos, diminuiu em 50% após 15 gerações mantidas em laboratório. Esse resultado de perda de uma proporção significativa de alelos ocorre principalmente com aqueles alelos raros presentes em população de cepa selvagem (Norris *et al.*, 2001; Stohler *et al.*, 2004). Neste estudo, detectamos o mesmo perfil de SNPs em cepas de laboratório e isolados do campo. Nossas observações estão de acordo com estudos de variabilidade intra-específica usando seqüências de DNA ribossomal, nos quais observaram muito poucas mudanças nucleotídicas entre cepas americanas e africanas (Després *et al.*, 1993). Foi, em seguida, analisada a frequência das variações nucleotídicas. Observamos que na maioria dos casos a frequência das bases polimórficas foi maior em cepas de campo. Uma possível explicação é um maior

endocruzamento entre cepas de laboratório, com uma respectiva queda da diversidade alélica, ou o efeito fundador. Isso evidencia a importância de estar relacionando e analisando o perfil polimórfico de diferentes cepas e o desenvolvimento das vacinas de DNA (Quinlivan *et al.*, 2005), uma vez que a presença destas variações, em especial as mutações não-sinônimas, podem causar uma mudança na estrutura da proteína codificada e, conseqüentemente, uma alteração da sua função.

A explosão da genômica, colocando à disposição dos pesquisadores seqüências genômicas inteiras, fez aumentar o número de seqüências primárias de proteínas depositadas em bancos de dados públicos (Cantor & Little 1998). No entanto, em proporção, o número de proteínas contendo sua estrutura tridimensional definida permaneceu com menores mudanças (<http://www.pdb.org>). Este aumento no número de seqüências permitiu que as estruturas fossem modeladas, baseando-se nas informações de estruturas já resolvidas (Singh & Singh 2005). Métodos de análise de estruturas de proteínas estão tornando-se fundamentais para entender o funcionamento destas macromoléculas, para a modelagem de novas moléculas e para a modelagem de pequenas moléculas complexadas às proteínas. Existe uma total correlação entre a estrutura da proteína e a sua atividade. Desta forma, procura-se conhecer as mudanças que a proteína pode sofrer de acordo com as suas características estruturais (Falcão *et al.*, 2002). Existem dois métodos experimentais para determinação da estrutura tridimensional de uma proteína: cristalografia de Raio X (modelo estático) e espectroscopia de ressonância nuclear magnética (modelo dinâmico) (Liu & Hsu 2005). Entretanto, estes métodos experimentais requerem grande quantidade de material biológico, são demorados e têm alto custo. O método alternativo aos métodos experimentais mais bem sucedido de predição de estruturas tridimensionais de proteínas é a modelagem por homologia, também conhecida como modelagem comparativa (Marti-Renom *et al.*, 2000). Neste trabalho, realizamos a modelagem por homologia da estrutura da catepsina B de *S. mansoni* a partir de uma estrutura modelo. O uso de perfis de seqüências para determinar a função das proteínas tem se tornado um instrumento essencial (Park *et al.*, 1997). A modelagem de uma proteína pelo método da homologia baseia-se no conceito de evolução molecular. Em outras palavras, parte do princípio de que a semelhança entre as estruturas primárias desta proteína e de proteínas homólogas com estruturas tridimensionais conhecidas implique possivelmente em similaridade estrutural entre elas (Teichmann *et al.*, 2001). Se o grau de identidade entre a estrutura primária da proteína alvo e da proteína cuja estrutura foi determinada com uma resolução igual ou superior a 2 Å, for maior que 25% quando o número de resíduos é acima

de 80, existe grande probabilidade de que estas proteínas tenham estruturas tridimensionais semelhantes e pode-se construir um modelo para a proteína desejada (Sander & Schneider 1991). A proteína utilizada como modelo no nosso estudo foi a procathepsina humana (PDB 3pbh), que apresentou um índice de similaridade de sequência de 60% com a com a catepsina B de *S. mansoni*. Devido ao limite de tempo e devido ao fato de que a identidade entre as entre a seqüência modelo e a seqüência alvo foi relativamente alta, uma modelagem minuciosa, das regiões de alça e rotâmeros, não foi realizada.

As proteínas agrupam-se em um número limitado de famílias tridimensionais. Estima-se que existam, aproximadamente, 5.000 famílias protéicas (Godzik 2003). Conseqüentemente, quando se conhece a estrutura de pelo menos um representante de uma família, é geralmente possível modelar, por homologia, os demais membros da família. Contudo, a determinação da estrutura terciária de proteínas por este método é um desafio, exigindo algoritmos capazes de simular, com precisão, a ação das leis que regem o processo de enovelamento ou empacotamento e interações entre resíduos da proteína. Contudo, estas leis ainda não são perfeitamente conhecidas, o que torna muito difícil a obtenção de conformações simultaneamente estáveis e funcionais a custo computacional razoável (Santos-Filho & Alencastro 2003). Apesar das limitações, a ampliação do número de trabalhos utilizando metodologias capazes de gerar modelos *in silico* vem aumentando (Contreras-Moreira *et al.*, 2003; Liko *et al.*, 2004; Droit *et al.*, 2005). O primeiro requisito que uma proteína adequadamente modelada deve atender é ter uma estrutura terciária satisfatória (Marti-Renom *et al.*, 2000). É importante verificar se existem grandes diferenças conformacionais não explicadas (como alças) entre os elementos de estrutura secundária das estruturas modelo e da estrutura alvo modelada que possam estar causando distorções de enovelamento que são difíceis de avaliar (Santos-Filho & Alencastro 2003). No presente trabalho a estrutura tridimensional da proteína modelo e da proteína alvo apresentaram alto índice de similaridade de estrutura primária e funcional, sendo portanto consideradas homólogas, ou seja, derivadas de um ancestral comum. Isso também foi observado quando a procathepsina B humana foi comparada com procathepsina de outros organismos (Cygler *et al.*, 1996). Ambas proteínas, humanas e de *S. mansoni*, apresentam uma inserção de aproximadamente 20 resíduos correspondendo a uma alça responsável pela ativação ou inibição da atividade da proteína. Illy e colaboradores (1997) mostraram que uma proteína catepsina B mutante, sem a presença dessa alça, perde sua atividade essencial como exopeptidase. Além disso, a estrutura da enzima madura possui dois domínios e um sítio ativo

interno aos domínios. A cadeia do propetídeo começa no topo da molécula, contorna a superfície do domínio direito e termina no sitio ativo. Uma vez modelada a catepsina B de *S. mansoni*, foi analisado o possível efeito das mutações não-sinônimas sobre a estrutura da proteína.

Mutações não-sinônimas têm sido foco de interesse pelo fato da possível alteração da função de proteínas e, conseqüentemente, estarem ligadas a genótipos de interesse (Kim *et al.*, 2003). Vários estudos já sugeriram distorções em proteínas como conseqüência da variação de apenas um aminoácido (Alexandrie *et al.*, 2002). No presente trabalho, provavelmente, as mutações não-sinônimas identificadas não causam uma alteração na conformação estrutural da catepsina B de *S. mansoni*. Contudo, as mudanças de aminoácidos se encontram, em sua maioria, na superfície da estrutura tridimensional da proteína sendo possível que estejam alterando regiões de epitopo e, conseqüentemente, afetando o reconhecimento dos anticorpos. Não obstante, a posição dos aminoácidos substituídos requer comprovação experimental. A determinação de possíveis alterações em epitopos também pode ser comprovada experimentalmente. Além disso, foi observado pelo módulo ConSSeq para os SNPs 4 e 5 que o aminoácido codificado na presença do SNP foi encontrado um maior número de vezes que o aminoácido original, sugerindo que esses aminoácidos, provavelmente, possam estar sofrendo algum tipo de pressão seletiva.

Em resumo, o presente trabalho fornece evidências de que os SNPs são marcadores genéticos amplamente distribuídos pelo genoma de *S. mansoni*. Além disso, demonstra a eficácia dos métodos computacionais automatizados nas análises em grande escala e, conseqüentemente, o fornecimento de um grande número de informações sobre os genes expressos no *S. mansoni*. Como perspectiva do presente trabalho, os SNPs identificados e suas respectivas características serão depositados no SchistoDB. O SchistoDB consiste na iniciativa de montar um banco de dado público contendo todas as informações relativas ao genoma de *Schistosoma*. Pesquisadores interessados em um gene poderão obter as informações geradas por este trabalho. As observações geradas por este trabalho geram várias hipóteses sobre a presença e a conseqüência de SNPs que podem ser confirmadas experimentalmente. Isso será extremamente importante, uma vez que as informações sobre o genoma do parasito encontram-se fragmentada e, desta forma, todas as informações poderão ser depositadas em um lugar comum, o que permitirá buscas simples ou complexas, contribuindo para projetos de descoberta de vacinas e de novas formas de tratamento da doença.

VI - CONCLUSÕES

1. O presente estudo demonstrou a grande colaboração da bioinformática nas análises de ESTs e detecção de marcadores moleculares do tipo SNPs a partir do uso de uma metodologia automatizada.
2. Foram detectados 2.303 SNPs em 863 contigs, sendo 64,03% transversões e 35,64% transições, 30,32% mutações sinônimas e 69,67% mutações não-sinônimas, sendo 30,05%, 30,03% e 39,01% na primeira, segunda e terceira base do códon respectivamente.
3. O novo programa de busca de SNPs, o cSNPer, escrito pelo nosso grupo foi validado.
4. Os genes que possuem o maior número de mutações são codificados para conhecidos antígenos.
5. Genes candidatos à vacina apresentaram, em sua maioria, a presença de SNPs. Dentre os genes selecionados: Sm14, catepsina B, Sm23, GST, miosina, TPI e paramiosina, apenas os 3 últimos genes não apresentaram polimorfismos.
6. Foi validada a presença de SNPs no gene da catepsina B de *S. mansoni*.
7. O método de modelagem molecular mostrou-se eficiente para prever a estrutura tridimensional da catepsina B de *S. mansoni*.
8. As mutações não-sinônimas analisadas por modelagem comparativa da catepsina B provavelmente não alteram a estrutura secundária da proteína.

VII – ENDEREÇOS ELETRÔNICOS

<http://www.who.int/tdrdisease/default.htm> - Página da TDR - Doenças Parasitárias Endêmicas.

<http://www.ncbi.nlm.nih.gov/BLAST> - Página da ferramenta Blast, utilizada na comparação de seqüências de nucleotídeos/ou aminoácidos.

<http://www.ncbi.nlm.nih.gov/SNP> - Link da página do NCBI, banco de dados de SNPs.

<http://www.phrap.org> - Página contendo informações sobre as ferramentas Phred/Phrap/Consed.

<http://www.biotools.com> - Página contendo ferramentas de bioinformática para análises de seqüências em geral.

<http://www.tigr.org> – Página do Instituto de Pesquisa Genômica contendo dados sobre o genoma de diferentes organismos.

rgmg.cpqrr.fiocruz.br - Página contendo informação sobre o Projeto Transcriptoma de *S. mansoni* coordenado pela equipe de Minas Gerais.

<http://snp.cshl.org/> - Página do banco de dados de SNPs: *The SNP Consortium Ltda.*

<http://www.pdb.com.br> – Página do banco de dados de estruturas de proteínas.

VIII - REFERÊNCIAS BIBLIOGRÁFICAS

- Alexandrie AK, Rannug A, Juronen E, Tasa G, Warholm M. Detection and characterization of a novel functional polymorphism in the GSTT1 gene. **Pharmacogenetics**, v. 12, p. 613-619, 2002.
- Ali PO, Simpson AJG, Allen R, Waters AP, Humphries CJ, Jonhston DA, Rollinson D. Sequence of small subunit rRNA gene of *Schistosoma mansoni* and its use in phylogenetic analysis. **Molecular and Biochemical Parasitology**, v. 46, p. 201-208, 1991.
- Al-Sherbiny M, Osman A, Barakat R, El Morshedy H, Bergquist R, Olds R. In vitro cellular and humoral responses to *Schistosoma mansoni* vaccine candidate antigens. **Acta Tropica**, v. 88, p. 117-130, 2003.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, p.403-10, 1990.
- Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. An SNP map of the human genome generated by reduced representation shotgun sequencing. **Nature**, v. 407, p. 513-516, 2000.
- Araújo N, Katz N, Dias EP, Souza CP. Susceptibility to chemotherapeutic agents of strains of *Schistosoma mansoni* isolated from treated and untreated patients. **American Journal of Tropical Medicine and Hygiene**, v. 29, p. 890-894, 1980.
- Argiro L, Henri S, Dessein H, Kouriba B, Dessein AJ, Bourgois A. Induction of a protection against *Schistosoma mansoni* with a MAP containing epitopes of Sm37-GAPDH and Sm10-DLC. Effect of coadsorption with GM-CSF on alum. **Vaccine**, v. 18, p. 2033-2038, 2000.
- Augusto-Pinto L, Teixeira SM, Pena SD, Machado CR. Single-nucleotide polymorphisms of the *Trypanosoma cruzi* MSH2 gene support the existence of three phylogenetic lineages presenting differences in mismatch-repair efficiency. **Genetics**, v. 164, p. 117-126, 2003.
- Bader JS. The relative power of SNPs and haplotype as genetic markers for association tests. **Pharmacogenomics**, v. 2, p.11-24, 2001.
- Barker SC, Blair D. Molecular phylogeny of *Schistosoma* species supports traditional groupings within the genus. **Journal of Parasitology**, v. 82, p. 292-298, 1996.
- Barker G, Batley J, Sullivan H, Edwards K, Edwards D. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. **Bioinformatics**, v. 19, p. 421-422, 2003.
- Bao Y.P, Huber M, Wei T.F, Marla SS, Storhoff JJ, Muller UR. SNP identification in unamplified human genomic DNA with gold nanoparticle probes. **Nucleic Acids Research**, v. 19, p. 33(2):e15, 2005.

- Batley J, Barker G, O'sullivan H, Edwards KJ, Edwards D. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. **Plant Physiology**, v. 132, p. 84-91, 2003.
- Bennett JL, Day T, Feng-Tao L, Ismail M, Farghaly A. The development of resistance to anthelmintics: A perspective with an emphasis on the antischistosomal drug praziquantel. **Experimental Parasitology**, v. 87, p. 260-267, 1997.
- Berger J, Suzuki T, Senti KA, Stubbs J, Schaffner G, Dickson BJ. Genetic mapping with SNP markers in *Drosophila*. **Nature Genetics**, v. 29, p. 475-481, 2001.
- Bergquist NR. Schistosomiasis: from risk assessment to control. **Trends in Parasitology**, v. 18, p. 309-314, 2002.
- Blair L, Webster JP, Barke GC. Isolation and characterization of polymorphic microsatellite markers in *Schistosoma mansoni* from Africa. **Molecular Ecology Notes**, v. 01, p. 93-95, 2001.
- Brentani H, Caballero OL, Camargo AA, Da Silva AM, Da Silva WA, Dias Neto E, Grivet M, Gruber A, Guimarães PE, Hide W, Iseli C, Jongeneel CV, Kelso J, Nagai MA, Ojopi EP, Osorio EC, Reis EM, Riggins GJ, Simpson AJ, De Souza S, Stevenson BJ, Strausberg RL, Tajara EH, Verjovski-Almeida S, *et al.*, Zalcberg H. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. **Proceedings of the National Academy of Sciences**. v. 100, p. 13418-13423, 2003.
- Brindley PJ, Kalinna BH, Dalton JP, Day SR, Wong JY, Smythe ML, McManus DP. Proteolytic degradation of host hemoglobin by Schistosomes. **Molecular and Biochemical Parasitology**, v. 89, p. 1-9, 1997.
- Brindley PJ, Laha T, Mcmanus DP, Loukas A. Mobile genetic elements colonizing the genomes of metazoan parasites. **Trends in Parasitology**, v. 19, p. 79-87, 2003.
- Broman KW, Feingold E. SNPs made routine. **Nature**, v. 1, p. 104-105, 2004.
- Brookes AJ. The essence of SNPs. **Gene**, v. 234, p. 177-186, 1999.
- Campos R, Moreira AAB, Sette H Jr, Chamone DAF, Silva LC. Hycanthon resistance in a human strain of *Schistosoma mansoni*. **Transactions of the Royal Society of Tropical Medicine Hygiene**, v. 70, p. 261-262, 1976.
- Cantor CR; Little DP. Massive attack on high-throughput biology. **Nature Genetics**, v. 20, p. 5-6, 1998.

- Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterization of single-nucleotide polymorphisms in coding regions of human genes. **Nature Genetics**, v. 22, p. 231-238, 1999.
- Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC Genetics**, v. 03, p. 1-14, 2002.
- Chalersworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. **Nature**, v. 371, p. 215-220, 1994.
- Chitsulo L, Engels D, Montresor A, Saviolli L. The global status of schistosomiasis and its control. **Acta Tropica**, v. 77, p. 41-51, 2000.
- Cheng TC, Xia QY, Qian JF, Liu C, Lin Y, Zha XF, Xiang ZH. Mining single nucleotide polymorphisms from EST data of silkworm, *Bombyx mori*, inbred strain Dazao. **Insect Biochemistry and Molecular Biology**, v. 34, p. 523-530, 2004.
- Cioli D, Pica-Mattocia L. Praziquantel. **Parasitology Research**, v. 90, p. 3-9, 2003.
- Clifford R, Edmonson M, Hu Y, Nguyen C, Scherpbier T, Buetow KH. Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. **Genome Research**, v. 10, p. 1259-1265, 2000.
- Coelho MV. O parasito *Schistosoma mansoni*. In: CUNHA, A.S. Esquistossomose mansoni. **São Paulo: Editora Savier**, Cap. 1, p. 1-12, 1970.
- Coles GC, Mutahi WT, Kinoti GK, Bruce JI, Katz N. Tolerance of kenyan *Schistosoma mansoni* to oxaminiquine. **Transactions of the Royal Society of Tropical Medicine and Hygiene**, v. 81, p. 782-785, 1987.
- Collins A, Lonjou C, Morton NE. Genetic epidemiology of single-nucleotide polymorphisms. **Proceedings of the National Academy of Sciences**, v. 96, 15173-15177, 1999.
- Contreras-Moreira B, Fitzjohn PW, Bates PA. *In silico* protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. **Journal of Molecular Biology**, v. 328, p. 593-608. 2003.

- Cook RM, Carvalho-Queiroz C, Wilding G, LoVerde PT. Nucleic acid vaccination with *Schistosoma mansoni* antioxidant enzyme cytosolic superoxide dismutase and the structural protein filamin confers protection against the adult worm stage. **Infect Immunology**, v. 72, p. 6112-6124, 2004.
- Cooper DN, Krawczak M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genome. **Human Genetics**, v. 83, p. 181-188, 1989.
- Cunha AS. Avaliação Terapêutica da oxaminiquine na esquistossomose humana pelo método do oograma por biópsia retal. **Revista do Instituto de Medicina Tropical de São Paulo**, v. 24, p. 88-94, 1982.
- Curtis J, Minchella DJ. Schistosome population genetic structure: when clumping worm is not just splitting hairs. **Parasitology Today**, v. 16, p. 68-71, 2000.
- Cygler M, Sivaraman J, Grochulski P, Coulombe R, Storer AC, Mort JS. Structure of rat procathepsin B: model for inhibition of cysteine protease activity by the proregion. **Structure**, v. 4, p. 405-416, 1996.
- Da'dara AA, Skelly PJ, Fatakawala M, Visovatti S, Eriksson E, Harn DA. Comparative efficacy of the *Schistosoma mansoni* nucleic acid vaccine, Sm23, following microseeding or gene gun delivery. **Parasite Immunology**, v. 24, p. 179-187, 2002.
- Dantec LL, Chagne D, Pot D, Cantin O, Garnier-Gere P, Bedon F, Frigerio JM, Chaumeil P, Leger P, Garcia V, Laigret F, De Daruvar A, Plomion C. Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. **Plant Molecular Biology**, v. 54, p. 461-470, 2004.
- Degrave WM, Melville S, Ivens A, Aslett M. Parasite genome initiatives. **International Journal of Parasitology**, v. 31, p. 532-536, 2001.
- Drescher KM, Rogers EJ, Bruce JI, Katz N, Dias LCS, Coles GC. Response of drug resistant isolates of *Schistosoma mansoni*, to antischistosomal agents. **Memórias do Instituto Oswaldo Cruz**, v. 88, p. 89-95, 1993.
- Després L, Imbert-Establet D, Combes C, Bonhomme F, Monnerot M. Isolation and polymorphism in mitochondrial DNA from *Schistosoma mansoni*. **Molecular and Biochemical Parasitology**, v. 47, p. 139-142, 1991.

- Després L, Imbert-Establet D, Combes C, Bonhomme F. Molecular evidence linking hominid evolution to recent radiation of Schistosomes (Platyhelminthes: Trematoda). **Molecular Phylogenetics and Evolution**, v. 1, p. 295-304, 1992.
- Després L, Imbert-Establet D, Monnerot M. Molecular characterization of mitochondrial DNA provides evidence for the recent introduction of *Schistosoma mansoni* into America. **Molecular and Biochemical Parasitology**, v. 60, p. 221-230, 1993.
- Dias LC, De Jesus Pedro R, Deberaldini ER. Use of praziquantel in patients with *Schistosoma mansoni* previously treated with oxaminiquine and/or hycanthone: resistance of *Schistosoma mansoni* to schistosomicidal agents. **Transactions of the Royal Society of Tropical Medicine and Hygiene**, v. 76, p. 652-659, 1982.
- Dias-Neto E, Correa RG, Verjovski-Almeida S, Briones MRS, Nagai MA, da Silva W, Zago MA, Bordin S, Costa FF, Goldman GH, Carvalho AF, Matsukuma A, Baia GS, Simpson DH, Brunstein A, de Oliveira PSL, Bucher P, Jongeneel CV, O'Hare MJ, Soares F, Brentani RR, Reis LFL, de Souza SJ, Simpson AJG. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. **Proceedings of the National Academy of Sciences**, v. 97, p. 3491-3496, 2000.
- Drescher KM, Rogers EJ, Bruce JI, Katz N, Dias LCS, Coles GC. Response of drug resistant isolates of *Schistosoma mansoni*, to antischistosomal agents. **Memórias do Instituto Oswaldo Cruz**, v. 88, p. 89-95, 1993.
- Drew AC, Brindley PJ. A retrotransposon of the non-long terminal repeat class from the human blood fluke *Schistosoma mansoni*. Similarities to the chicken-repeat-1-like elements of vertebrates. **Molecular Biology and Evolution**, v. 14, p. 602-610, 1997.
- Droit A, Poirier GG, Hunter JM. Experimental and bioinformatic approaches for interrogating protein-protein interactions to determine protein function. **Journal of Molecular Endocrinology**, v. 34, p. 263-280, 2005.
- Durand P, Sire C, Théron A. Isolation of microsatellite markers in the digenetic trematode *Schistosoma mansoni* from Guadeloupe island. **Molecular Ecology**, v. 9, p. 997-998, 2000.
- Escary J, Bottius E, Prince N, Reyes C, Fiawoumo Y, Caloustian C, Bruls T, Fujiyama A, Cooper RS, Adeyemo AA, Lathrop GM, Weissenbach J, Gyapay G, Foglio M, Beckmann JS. A first high-density map of 981 biallelic markers on human chromosome 14. **Genomics**, v. 70, p. 153-164, 2000.
- El-Sayed NM, Bartholomeu D, Ivens A, Johnston DA, LoVerde PT. Advances in Schistosome genomics. **Trends in Parasitology**, v. 20, p. 154-157, 2004.

- Engels D, Chitsulo L, Montresor A, Savioli L. The global epidemiological situation of schistosomiasis and new approaches to control and research. **Acta Tropica**, v. 82, p. 139-146, 2002.
- Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. **Science**, v. 15, p. 487-491, 1999.
- Ewing B, Green P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. **Genome Research**, v. 8, p. 186-194, 1998.
- Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. **Genome Research**, v. 8, p. 175-185, 1998.
- Falcão PK, Baudet C, Higa RH, Neshich G. Incorporação das Propriedades Rotâmeros e Ocupância em Métodos de Análise Estrutural de Proteínas. Comunicado Técnico. **International Standard Serial Number (ISSN) 1677-8464**, v. 34, p. 1-6, 2002.
- Fantappie MR, Gimba ER, Rumjanek FD. Lack of DNA methylation in *Schistosoma mansoni*. **Experimental of Parasitology**, v. 98, p. 162-166, 2001.
- Feng X, Carlton JM, Joy DA, Mu J, Furuya T, Suh BB, Wang Y, Barnwell JW, Su XZ. Single-nucleotide polymorphisms and genome diversity in *Plasmodium vivax*. **Proceedings of the National Academy of Sciences**, v. 100, p. 8502-8507, 2003.
- Fernandez-Mestre MT, Montagnani S, Layrissé Z. Is the CCR5-59029-G/G genotype a protective factor for cardiomyopathy in Chagas disease? **Human Immunology**, v. 65, p. 725-728, 2004.
- Ferrari ML, Coelho PM, Antunes CM, Tavares CA, da Cunha AS. Efficacy of oxamniquine and praziquantel in the treatment of *Schistosoma mansoni* infection: a controlled trial. **Bull World Health Organization**, v. 81, p. 190-196, 2003.
- Files VS. The study of the vector-parasite relationships in *Schistosoma mansoni*. **Parasitology**, v. 41, p. 264-269, 1951.
- Fitzsimmons CJ, Savolainen P, Amini B, Hjalms G, Lundeberg J, Andersson L. Detection of sequence polymorphisms in red junglefowl and White Leghorn ESTs. **Animal Genetics**, v. 35, p. 391-396, 2004.
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. **Science**, v. 269, p. 496-512, 1995.
- Fletcher M, LoVerde PT, Woodruff DS. Genetic variation in *Schistosoma mansoni*: enzyme polymorphisms in populations from Africa, Southwest Asia, South America and the West Indies. **American Journal of Tropical Medicine and Hygiene**, v. 30, p. 406-421, 1981.

- Fonseca CT, Cunha-Neto E, Goldberg AC, Kalil J, de Jesus AR, Carvalho EM, Correa-Oliveira R, Oliveira SC. Human T cell epitope mapping of the *Schistosoma mansoni* 14-kDa fatty acid-binding protein using cells from patients living in areas endemic for schistosomiasis. **Microbes and Infection**, v. 7, p. 204-212, 2005.
- Franco GR, Adams MD, Soares MB, Simpson AJ, Venter JC, Pena SD. Identification of new *Schistosoma mansoni* genes by the EST strategy using a directional cDNA library. **Gene**, v. 152, p. 141-147, 1995.
- Forche A, May G, Magee PT. Demonstration of loss of heterozygosity by single-nucleotide polymorphism microarray analysis and alterations in strain morphology in *Candida albicans* strains during infection. **Eukaryoty Cell**, v. 4, p. 156-165, 2005.
- Fridman C, Ojopi EP, Gregorio SP, Ikenaga EH, Moreno DH, Demetrio FN, Guimaraes PE, Vallada HP, Gattaz WF, Dias-Neto E. Association of a new polymorphism in ALOX12 gene with bipolar disorder. **European Archives of Psychiatry and Clinical Neuroscience**, v. 253, p. 40-43, 2003.
- Fryxell KJ, Moon WJ. CpG mutation rates in the human genome are highly dependent on local GC content. **Molecular Biology and Evolution**, v. 22, p. 650-658, 2004.
- Godzik A. Fold recognition methods. **Methods of Biochemical Analysis**, v. 44, p. 525-546, 2003.
- Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. **Genome Research**, v. 8, p. 195-202, 1998.
- Gotoh O. An improved algorithm for matching biological sequences. **Journal of Molecular Biology**, v. 162, p. 705-708, 1982.
- Klinkert MQ, Felleisen R, Link G, Ruppel A, Beck E. Primary structures of Sm31/32 diagnostic proteins of *Schistosoma mansoni* and their identification as proteases. **Molecular and Biochemical Parasitology**, v. 33, p. 113-122, 1989.
- Graeff-Teixeira C, Valar C, Moraes CK, Salvany AM, Brum Cde O, Maurer RL, Bem R, Mardini LB, Jobim MB, Amaral RS. The initial epidemiological studies in the low endemicity schistosomiasis area in Esteio, Rio Grande do Sul, the southernmost Brazilian state, 1997 to 2000. **Memórias do Instituto Oswaldo Cruz**, v. 99, p. 73-78, 2004.
- Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. **Human Molecular Genetics**, v. 16, p. 2403-2408, 2000.
- Grevelding CG, Sommer G, Kunz W. Female-specific gene expression in *Schistosoma mansoni* is regulated by pairing. **Parasitology**, v. 115, p. 635-640, 1997.

- Grivet L, Glaszmann JC, Vincentz M, Da Silva F, Arruda P. ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. **Theoretical and Applied Genetics**, v. 106, p. 190-197, 2003.
- Gu Z, Hillier L, Kwok P. Single nucleotide polymorphism hunting in cyberspace. **Human Mutation**, v. 12, p. 221-225, 1998.
- Guimarães PE, Costa MC. SNPs: sutis diferenças de um código. **Biotecnologia Ciência & Desenvolvimento**, v. 26, p. 24-27, 2002.
- Hagelberg E, Gray IC, Jeffreys AJ. Identification of the skeletal remains of a murder victim by DNA analysis. **Nature**, v. 352, p. 427-429, 1992.
- Hahn ME, Karchner SI, Franks DG, Merson RR. Aryl hydrocarbon receptor polymorphisms and dioxin resistance in Atlantic killifish (*Fundulus heteroclitus*). **Pharmacogenetics**, v. 14, p. 131-143, 2004.
- Halldorsson BV, Istrail S, De La Vega FM. Optimal selection of SNP markers for disease association studies. **Human Heredity**, v. 58, p. 190-202, 2004.
- Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. **Nature Genetics**, v. 22, p. 239-247, 1999.
- Hamburger J, Turetski T, Kapeller I, Deresiewicz R. Highly repeated short DNA sequences in the genome of *Schistosoma mansoni* recognized by a species-specific probe. **Molecular and Biochemical Parasitology**, v. 44, p. 73-80, 1991.
- Harrison RG. Animal mitochondrial DNA as a genetic marker in populational and evolutionary biology. **Trends in Ecology & Evolution**, v. 46, p. 06-11, 1989.
- Heaton MP, Grosse WM, Kappes SM, Keele JW, Chitko-Mckown CG, Cundiff LV, Braun A, Little DP, Laegreid WW. Estimation of DNA sequence diversity in bovine cytokine genes. **Mammalian Genome**, v. 12, p. 32-37, 2001.
- Hide G, Tilley A. Use of mobile genetic elements as tools for molecular epidemiology. **International Journal of Parasitology**, v. 31, p. 599-602, 2001.
- Higa RH, Montagner AJ, Togawa RC, Kuser PR, Yamagishi ME, Mancini AL, Pappas G Jr, Miura RT, Horita LG, Neshich G. ConSSeq: a web-based application for analysis of amino acid conservation based on HSSP database and within context of structure. **Bioinformatics**, v. 20, 1983-1985, 2004.
- Hiller GV. Buoyant density and thermal denaturation profiles of schistosome DNA. **Journal of Parasitology**, v. 60, p. 725-727, 1974.

- Hillman GR, Senft AW. Anticholinergic properties of the antischistosomal drug hytacone. **American Journal of Tropical Medicine and Hygiene**, v. 24, p. 827-834, 1975.
- Hirai H, LoVerde PT. Identification of telomeros on *Schistosoma mansoni* chromosomes by FISH. **Journal of Parasitology**, v. 82, p. 511-512, 1996.
- Hoffmann KF, Strand M. Molecular identification of a *Schistosoma mansoni* tegumental protein with similarity to cytoplasmic dynein light chains. **Journal of Biological Chemistry**, v. 271, p. 26117-26123, 1996.
- Hu W, Brindley PJ, McManus DP, Feng Z, Han ZG. Schistosome transcriptomes: new insights into the parasite and schistosomiasis. **Trends in Molecular Medicine**, v. 10, p. 217-225, 2004.
- Hunter RL, Markert CL. Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. **Science**, v. 125, p. 1294-1295, 1957.
- Illy C, Quraishi O, Wang J, Purisima E, Vernet T, Mort JS. Role of the occluding loop in cathepsin B activity. **Journal of Biological Chemistry**, v. 272, p. 1197-1202. 1997
- Jalving R, Van't Slot R, Van Oost BA. Chicken single nucleotide polymorphism identification and selection for genetic mapping. **Poult Science**, v. 83, p. 1925-1931, 2004.
- Jannotti-Passos LK, Vidigal THDA, Dias-Neto E, Pena SDJ, Simpson AJG, Dutra WO, Souza CP, Carvalho-Parra JF. PCR amplification of the mitochondrial DNA minisatellite region to detect *Schistosoma mansoni* infection in *Biomphalaria Glabrata* snails. **Journal of Parasitology**, v. 83, p. 395-399, 1997.
- Lai E. Application of SNP technologies in medicine: lessons learned and future challenges. **Genome Research**, Vol. 11, p. 927-929, 2001
- Lee JH, Koh I. Drug to SNP: A pharmacogenomics database for linking drug response to SNP. **Genome Informatics**, v. 12, p. 482-483, 2001.
- Liu HL, Hsu JP. Recent developments in structural proteomics for protein structure determination. **Proteomics**, 2005 [Epub ahead of print].
- Kanehisa M, Bork P. Bioinformatics in the post-sequence era. **Nature Genetics**, v. 33, p. 305-310, 2003.

- Katz N, Dias EP, Araújo N, Souza CP. Estudos de uma cepa humana de *Schistosoma mansoni* resistente a agentes esquistossomicidas. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 7, p. 381-387, 1973.
- Katz N, Dias EP, Souza CP, Bruce JI, Coles GC. Rate of action of schistosomicides in mice infected with *Schistosoma mansoni*. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 22, p. 183-186, 1989.
- Katz N, Peixoto SV. Análise crítica da estimativa do número de portadores de esquistossomose no Brasil. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 33, p. 303-308, 2000.
- Kheir MM, Baraka OZ, El-Tom IA, Mukhtar MM, Homieda MM. Effects of single-dose praziquantel on morbidity and mortality resulting from intestinal schistosomiasis. **Eastern Mediterranean Health Journal**, v. 6, p. 926-931, 2000.
- Kim H, Schmidt CJ, Decker KS, Emara MG. A double-screening method to identify reliable candidate non-synonymous SNPs from chicken EST data. **Animal Genetics**, v. 34, p. 249-254, 2003.
- Kimura ET, Nikiforova MN, Zhu Z, Knauf JA, Nikiforov YE, Fagin JA. High prevalence of BRAF mutations in thyroid cancer: genetic evidence for constitutive activation of the RET/PTC-RAS-BRAF signaling pathway in papillary thyroid carcinoma. **Cancer Research**, v. 63, p. 1454-1457, 2003.
- Kissinger JC, Gajria B, Li L, Paulsen IT, Roos DS. ToxoDB: accessing the *Toxoplasma gondii* genome. **Nucleic Acids Research**, v. 31, p. 234-236, 2003.
- Kleyn PW, Vesell ES. Pharmacogenomics: genetic variation as a guide to drug development. **Science**, v. 281, p. 1820-1821, 1998.
- Klinkert MQ, Felleisen R, Link G, Ruppel A, Beck E. Primary structures of Sm31/32 diagnostic proteins of *Schistosoma mansoni* and their identification as proteases. **Molecular and Biochemistry Parasitology**, v. 33, p. 113-122, 1989.
- Koed K, Wiuf C, Christensen LL, Wikman FP, Zieger K, Moller K, Von Der Maase H, Orntoft TF. High-density single nucleotide polymorphism array defines novel stage and location-dependent allelic imbalances in human bladder tumors. **Cancer Research**, v. 65, p. 34-45, 2005.
- Kunz W. Schistosome male-female interaction: induction of germ-cell differentiation. **Trends in Parasitology**, v. 17, p. 227-231, 2001.

- Laclette JP, Landa A, Arcos L, Willms K, Davis AE, Shoemaker CB. Paramyosin is the *Schistosoma mansoni* (Trematoda) homologue of antigen B from *Taenia solium* (Cestoda). **Molecular and Biochemical Parasitology**, v. 44, p. 287-295, 1991.
- Liang F, Ingeborg H, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. An optimized protocol for analysis of EST sequences. **Nucleic Acids Research**, v. 28, p. 3657-3665, 2000.
- Liko I, Igaz P, Patocs A, Toth S, Pazmany T, Toth M, Racz K. Sequence variants of the ligand-binding domain of the glucocorticoid receptor gene and their functional consequences on the three-dimensional protein structure. **Current Medicinal Chemistry**, v. 11, p. 3229-3237, 2004.
- Littlewood DT, Johnston Da. Molecular phylogenetics of the four *Schistosoma* species groups determined with partial 28S ribosomal RNA gene sequences. **Parasitology**, v. 111, p. 167-175, 1995.
- Loukas A, Bethony JM, Williamson AL, Goud GN, Mendez S, Zhan B, Hawdon JM, Elena Bottazzi M, Brindley PJ, Hotez PJ. Vaccination of dogs with a recombinant cysteine protease from the intestine of canine hookworms diminishes the fecundity and growth of worms. **Journal of Infectious Disease**, v. 189, p. 1952-1961, 2004.
- LoVerde PT, Dewald J, Minchella DJ. Further studies on genetic variation in *Schistosoma mansoni*. **Journal of Parasitology**, v. 71, p. 732-734, 1985.
- Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. **Annual Review of Biophysics and Biomolecular Structure**, v. 29, p. 291-325, 2000.
- Martinez-Arias R, Mateu E, Bertranpetit J, Calafell F. Profiles of accepted mutation: from neutrality in a pseudogene to disease-causing mutation on its homologous gene. **Human Genetics**, v. 109, p. 7-10, 2001.
- Marx K A, Bizzaro JW, Blake R D, Tsai MH, Tao LF. Experimental DNA melting behavior of the three major *Schistosoma* species. **Molecular and Biochemical Parasitology**, v. 107, p. 303-307, 2000.
- McManus DP, Hope M. Molecular variation in the human schistosomes. **Acta Tropica**, v. 53, p. 255-276, 1993.

- McNair AT, Dissous C, Duvaux-Miret O, Capron A. Cloning and characterisation of the gene encoding the 28-kDa glutathione S-transferase of *Schistosoma mansoni*. **Gene**, v. 124, p. 245-249, 1993.
- Meyer P, Sergi C, Garbe C. Polymorphisms of the BRAF gene predispose males to malignant melanoma. **Journal of Carcinogenesis**, v. 2, p. 02-05, 2003.
- Morlais I, Poncon N, Simard F, Cohuet A, Fontenille D. Intraspecific nucleotide variation in *Anopheles gambiae*: new insights into the biology of malaria vectors. **American Journal of Tropical Medicine and Hygienic**, v. 71, p. 795-802, 2004.
- Moser D, Tendler M, Griffiths G, Klinkert MQ. A 14-kDa *Schistosoma mansoni* polypeptide is homologous to a gene family of fatty acid binding proteins. **Journal of Biological Chemistry**, v. 266, p. 8447-8454, 1991.
- Myrick A, Sarr O, Dieng T, Ndir O, Mboup S, Wirth DF. Analysis of the genetic diversity of the *Plasmodium falciparum* multidrug resistance gene 5' upstream region. **American Journal of Tropical Medicine and Hygienic**, v. 72, p. 182-188, 2005.
- Mullikin JC, Hunt SE, Cole CG, Mortimore BJ, Rice CM, Burton J, Matthews LH, Pavitt R, Plumb RW, Sims SK, Ainscough RM, Attwood J, Bailey JM, Barlow K, Bruskiwich RM, Butcher PN, Carter NP, Chen Y, Clee M, Coggill PC, Davies J, Davies RM, Dawson E, Francis MD et al., Bentley DR. An SNP map of human chromosome 22. **Nature**, v. 407, p. 516-520, 2000.
- Mullis KB, Faloona F. Specific synthesis of DNA *in vitro* via a polymerase catalyzed chain reaction. **Methods In Enzymology**, v. 155, p. 335-350, 1987.
- Navarro MC, Cesari IM, Incani RM. Isoenzyme studies in one Brazilian and two Venezuelan strains of *Schistosoma mansoni*. **Compendiums of Biochemistry and Physiology**, v. 102, p. 471-474, 1992.
- Naoki K, Chen TH, Richards WG, Sugarbaker DJ, Meyerson M. Mis-sense mutations of the Braf gene in human lung adenocarcinomas. **Cancer Research**, v. 62, p. 7001-7003, 2002.
- Nascimento E, Leao IC, Pereira VR, Gomes YM, Chikhlikar P, August T, Marques E, Lucena-Silva N. Protective immunity of single and multi-antigen DNA vaccines against schistosomiasis. **Memórias do Instituto Oswaldo Cruz**, v. 97, p. 105-109, 2002.

- Neshich G, Togawa RC, Mancini AL, Kuser PR, Yamagishi MEB, Pappas Junior G, Torres WV, Campos TFE, Ferreira LL, Luna FM, Oliveira AG, Miura RT, Inoue MK, Horita LG, Souza DF De, Dominiquini F, Álvaro A, Lima CS, Ogawa FO, Gomes GB, Palandrani JF, Santos GF Dos, Freitas EM De, Mattiuz AR, Costa IC, Almeida CL De, Souza S, Baudet C, Higa RH. STING Millennium: a Web based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. **Nucleic Acids Research**, v. 31, p. 3386-3392, 2003.
- Neshich G, Rocchia W, Mancini AL, Yamagishi ME, Kuser PR, Fileto R, Baudet C, Pinto IP, Montagner AJ, Palandrani JF, Krauchenco JN, Torres RC, Souza S, Togawa RC, Higa RH. JavaProtein Dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. **Nucleic Acids Research**, v. 32, p. 595-601, 2004.
- Norris DE, Shurtleff AC, Toure YT, Lanzaro GC. Microsatellite DNA polymorphism and heterozygosity among field and laboratory populations of *Anopheles gambiae ssp.* (Diptera: Culicidae). **Journal of Medical Entomology**, v. 38, p. 336-340, 2001.
- Noya O, De Noya BA, Ballen DE, Bermudez H, Bout D, Hoebeke J. Immunogenicity of synthetic peptides from the Sm31 antigen (cathepsin B) of the *Schistosoma mansoni* adult worms. **Parasite Immunology**, v. 23, p. 567-573, 2001.
- Oliveira G. The *Schistosoma* gene discovery project, an update. **Trends in Parasitology**, v. 17, p. 108-109, 2001.
- Oliveira GC, Rodrigues NB, Romanha A, Bahia D. Genome and genomics of Schistosomes. **Canadian Journal of Zoology-Revue Canadienne de Zoologie**, v. 82, p. 375-390, 2004.
- Oliveira GC, Bahia D. The genome of *Schistosoma mansoni*. **Proceeding of the third Brazilian Symposium of Mathematical and Computational Biology**, v. 1, p. 101-115, 2004.
- Oliveira GC, Johnston D. Mining the Schistosome DNA sequence database. **Trends in Parasitology**, v. 117, p. 501-503, 2001.
- Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. **Journal of Molecular Biology**, v. 273, p. 349-354, 1997.
- Pavlovic-Lazetic GM, Mitic NS, Beljanski MV. Bioinformatics analysis of SARS coronavirus genome polymorphism. **BMC Bioinformatics**, v. 25, p. 65, 2004.

- Pellegrino J, Coelho PMZ. *Schistosoma mansoni*, wandering capacity of a worm couple. **Journal of Parasitology**, v. 64, p. 181, 1978.
- Pena HB, Souza CP, Simpson AJG, Pena SDJ. Intracellular promiscuity in *Schistosoma mansoni*: Nuclear transcribed DNA sequences are part of a mitochondrial minisatellite region. **Proceeding of National Academy of Sciences USA**, v. 92, p. 915-919, 1995.
- Peyron F, Eudes N, De Monbrison F, Wallon M, Picot S. Fitness of *Toxoplasma gondii* is not related to DHFR single-nucleotide polymorphism during congenital toxoplasmosis. **International Journal of Parasitology**, v. 34, p. 1169-1175, 2004.
- Pica-Mattoccia L, Dias LCS, Archer S. Binding of oxaminiquine to DNA of schistosomes. **Transactions of the Royal Society of Tropical Medicine and Hygiene**, v. 83, p. 89-96, 1989.
- Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M. Mining SNPs from EST databases. **Genome Research**, v. 9, p. 167-174, 1999.
- Purfield A, Nelson A, Laoboonthai A, Congpuong K, Mcdaniel P, Miller R.S, Welch K, Wongsrichanalai C, Meshnick SR. A new method for detection of pfm₁ mutations in *Plasmodium falciparum* DNA using real-time PCR. **Malaria Journal**, v. 3, p. 9, 2004.
- Qiu P, Wang L, Kostich M, Ding W, Simon JS, Greene JR. Genome wide in silico SNP-tumor association analysis. **BMC Cancer**, v. 29, p. 4, 2004.
- Quinlivan M, Gershon AA, Steinberg SP, Breuer J. An evaluation of single nucleotide polymorphisms used to differentiate vaccine and wild type strains of varicella-zoster virus. **Journal of Medical Virology**, v. 75, p. 174-180, 2005.
- Rafalski A. Applications of single nucleotide polymorphisms in crop genetics. **Current Opinion in Plant Biology**, v. 94, p. 94-100, 2002.
- Rao KV, He YX, Kalyanasundaram R. Expression of a 28-kilodalton glutathione S-transferase antigen of *Schistosoma mansoni* on the surface of filamentous phages and evaluation of its vaccine potential. **Clinical and Diagnostic Laboratory Immunology**, v. 10, p. 536-41, 2003.
- Redman C, Robertson A, Fallon PG, Modena J, Kusel GR, Doenhoff MJ. Praziquantel, an urgent and exciting challeng. **Parasitology Today**, v. 12, p. 14-20, 1996.
- Reynolds SR, Dahl CE, Harn DA. T and B epitope determination and analysis of multiple antigenic peptides for the *Schistosoma mansoni* experimental vaccine triose-phosphate isomerase. **The Journal of Immunology**, v. 152, p. 193-200, 1994.

- Ribeiro F, Coelho PMZ, Vieira LQ, Watson DG, Kusel GR. The effect of Praziquantel treatment on glutathione concentration in *Schistosoma mansoni*. **Parasitology**, v. 116, p. 229-236, 1998.
- Riley JH, Allan CJ, Lai E, Roses A. The use of single nucleotide polymorphisms in the isolation of common disease genes. **Pharmacogenomics**, v. 1, p. 39-47, 2000.
- Risch NJ. Searching for genetic determinants in the new millennium. **Nature**, v. 405, p. 847-856, 2000.
- Rodrigues NB, LoVerde PT, Romanha AJ, Oliveira GC. Characterization of new *Schistosoma mansoni* microsatellite loci in sequences obtained from public DNA databases and microsatellite enriched genomic libraries. **Memórias do Instituto Oswaldo Cruz**, v. 97, p. 71-75, 2002.
- Rollinson, D, Walker TK, Simpson AJG. New approaches to schistosome identification. **Parasitology Today**, v. 2, p. 24-25, 1986 a.
- Rollinson D, Walker TK, Simpson AJG. The application of recombinant DNA technology to problems of helminthes identification. **Parasitology**, v. 91, p. 853-871, 1986 b.
- Ruiz A, Molina JM, Njue A, Prichard RK. Genetic variability in cysteine protease genes of *Haemonchus contortus*. **Parasitology**, v. 128, p. 549-559, 2004.
- Sajid M, McKerrow. Cysteine proteases of parasitic organisms. **Molecular & Biochemical Parasitology**, v. 120, p. 1-21, 2002.
- Sajid M, McKerrow HJ, Hansell E, Mathieu M, Lucas K, Hsieh I, Geenbaum D, Bogyo M, Salter J, Lim K, Franklin C, Kim J, Caffrey C. Functional expression and characterization of *Schistosoma mansoni* cathepsin B and its *trans*-activation by an endogenous asparaginyl endopeptidase. **Molecular and Biochemical Parasitology**, v. 131, p. 65-75, 2003.
- Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. **Proteins**, v. 9, p. 56-68, 1991.
- Santos-Filho O, Alencastro R. Modelagem de proteínas por homologia. **Química Nova**, v. 26, p. 253-259, 2003.

- Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, Mcpherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Atshuler D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. **Nature**, v. 409, p. 928-933, 2001.
- Schechtman D, Tarrab-Hazdai R, Arnon R. The 14-3-3 protein as a vaccine candidate against schistosomiasis. **Parasite Immunology**, v. 23, p. 213-217, 2001.
- Shoemaker C, Gross A, Gebremichael A, Harn D. cDNA cloning and functional expression of the *Schistosoma mansoni* protective antigen triose-phosphate isomerase. **Proceedings of the National Academy of Sciences**, v. 89, p. 1842-1846, 1992.
- Short RB, Menzel MY. Chromosomes of nine species of schistosomes. **Journal of Parasitology**, v. 46, p. 273-287, 1960.
- Short RB, Menzel MY, Pathak S. Somatic chromosomes of *Schistosoma mansoni*. **Journal of Parasitology**, v. 65, p. 471-473, 1979.
- Short RB, Grossman AI. Conventional geimsa and C-banded karyotypes of *Schistosoma mansoni* and *S. rodhaini*. **Journal of Parasitology**, v. 67, p. 661-667, 1981.
- Shrivastava J, Qian BZ, Mcvean G, Webster JP. An insight into the genetic variation of *Schistosoma japonicum* in mainland China using DNA microsatellite markers. **Molecular Ecology**, v. 14, p. 839-849, 2005.
- Simpson AJG, Sher A, Mccutchan TF. The genome of *Schistosoma mansoni*, isolation of DNA, its size and repetitive sequences. **Molecular and Biochemical Parasitology**, v. 6, p. 125-137, 1982.
- Simpson AJG, Dame JB, Lewis FA, Mccutchan TF. The arrangement of ribosomal RNA genes in *Schistosoma mansoni* - Identification of polymorphic structural variants. **European Journal of Biochemistry**, v. 139, p. 41-45, 1984.
- Singh GB, Singh H. Databases, models, and algorithms for functional genomics: a bioinformatics perspective. **Molecular Biotechnology**, v. 29, p. 165-183, 2005.
- Smith TF, Waterman MS. Overlapping genes and information theory. **Journal of Theoretical Biology**, v. 91, p. 379-380, 1981.

- Smith EJ, Shi L, Drummond P, Rodriguez L, Hamilton R, Ramlal R, Smith G, Pierce K, Foster J. Expressed sequence tags for the chicken genome from a normalized 10-day-old White Leghorn whole embryo cDNA library: 1. DNA sequence characterization and linkage analysis. **Journal of Heredity**, v. 92, p. 1-8, 2001.
- Smithies O. Zone electrophoresis in starch gels: group variation in the serum proteins of normal human adults. **Biochemical Journal**, v. 61, p. 629-641, 1955.
- Somers DJ, Kirkpatrick R, Moniwa M, Walsh A. Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. **Genome**, v. 46, p. 431-437, 2003.
- Spotila LD, Rekosh DM, Boucher JM, LoVerde PT. A cloned DNA probe identifies the sex of *Schistosoma mansoni* cercariae. **Molecular and Biochemical Parasitology**, v. 26, p. 17-20, 1987.
- Spotila LD, Hirai H, Rekosh DM, LoVerde PT. A retroposon-like short repetitive DNA element in the genome of the human blood fluke *Schistosoma mansoni*. **Cell & Chromosome**, v. 97, p. 421-428, 1989.
- Spotila LD, Rekosh DM, LoVerde PT. Polymorphic repeated DNA element in the genome of *Schistosoma mansoni*. **Molecular and Biochemical Parasitology**, v. 48, p. 117-120, 1991.
- Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han J, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shanner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Rúaño G, Vovis GF. Haplotype variation and linkage disequilibrium in 313 human genes. **Science**, v. 293, p. 489-493, 2001.
- Stitzel NO, Binkowski TA, Tseng YY, Kasif S, Liang J. TopoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. **Nucleic Acids Research**, v. 32, p. 520-522, 2004.
- Stohler RA, Curtis J, Minchella DJ. A comparison of microsatellite polymorphism and heterozygosity among field and laboratory populations of *Schistosoma mansoni*. **International Journal of Parasitology**, v. 34, p. 595-601, 2004.
- Su C, Hott C, Brownstein BH, Sibley LD. Typing single-nucleotide polymorphisms in *Toxoplasma gondii* by allele-specific primer extension and microarray detection. **Methods of Molecular Biology**, v. 270, p. 249-262, 2004.
- Suh Y, Vijg J. SNP discovery in associating genetic variation with human disease phenotypes. **Mutation Research**, v. 573, p. 41-53, 2004.

- Teichmann SA, Murzin AG, Chothia C. Determination of protein function, evolution and interactions by structural genomics. **Current Opinion in Structural Biology**, v. 11, p. 354-363, 2001.
- Tomso DJ, Bell DA. Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands. **Journal of Molecular Biology**, v. 327, p. 303-308, 2003.
- Turk D, Podobnick M, Kuhelj R, Dolinar M, Turk V. Crystal structures procathepsin B at 3.2 and 3.3 Å resolution reveal an interaction motif between a papain-like cysteine protease and its propeptide. **FEBS Letter**, v. 384, p. 211-214, 1996.
- Useche FJ, Gao G, Harafey M, Rafalski A. High-throughput identification, database storage and analysis of SNPs in EST sequences. **Genome Informatics**, v. 12, p. 194-203, 2001.
- Valadares TE, Coelho PMZ, Pellegrino J, Sampaio IBM. *S. mansoni*: Aspectos da oviposição da cepa LE em camundongos infectados com um casal de vermes. **Revista do Instituto de Medicina Tropical de São Paulo**, v. 23, p. 6, 1981.
- Vallone P, Butler J. Y-SNP typing of U.S. African American and caucasian samples using allele – Specific hybridization and primer extension. **Journal of Forensic Sciences**, v. 49, p. 723-731, 2004.
- Van Der Werf MJ, Borsboom GJ, De Vilas SJ. No effect of recall period length on prevalence of self-reported haematuria in *Schistosoma haematobium*-endemic areas. **Transactions of the Royal Society of Tropical Medicine and Hygiene**, v. 97, p. 373-374, 2003.
- Van Keulen H, LoVerde PT, Bobek LA, Rekosh DM. Organization of the ribosomal RNA genes in *Schistosoma mansoni*. **Molecular and Biochemical Parasitology**, v. 15, p. 215-230, 1985.
- Verjovski-Almeida S, Demarco R, Martins EA, Guimarães PE, Ojopi EP, Paquola AC, Piazza JP, Nishiyama MY Jr, Kitajima JP, Adamson RE, Ashton PD, Bonaldo MF, Coulson PS, Dillon GP, Farias LP, Gregorio SP, Ho PL, Leite RA, Malaquias LC, Marques RC, Miyasato PA, Nascimento AL, Ohlweiler FP, Reis EM, Ribeiro MA, Sa RG, Stukart GC, Soares MB, Gargioni C, Kawano T, Rodrigues V, Madeira AM, Wilson RA, Menck CF, Setubal JC, Leite LC, Dias-Neto E. Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. **Nature Genetics**, v. 35, p. 148-157, 2003.
- Verjovski-Almeida S, Leite LC, Dias-Neto E, Menck CF, Wilson RA. Schistosome transcriptome: insights and perspectives for functional genomics. **Trends in Parasitology**, v. 20, p. 304-308, 2004.

- Venter JC, Smith HO, Hood L. A new strategy for genome sequencing. **Nature**, v. 30, p. 364-366, 1996.
- Vignal A, Milan D, Sancristobal M, Eggen A. A review on SNP and other types of molecular markers and their use in animal genetics. **Genetics Selection Evolution**, v. 34, p. 275-305, 2002.
- Zhang K, Calabrese P, Nordborg M, Sun F. Haplotype block structure and its applications to association studies: power and study designs. **The American Journal of Human Genetics**, v. 71, p. 1386-1394, 2002.
- Zhang F, Zhao Z. The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. **Genomics**, v. 84, p. 785-795, 2004.
- Zweiger G, Scott RW. From expressed sequence tags to 'epigenomics': an understanding of disease processes. **Current Opinion of Biotechnology**, v. 8, p. 684-687, 1997.
- Wang JP, Lindsay BG, Leebens-Mack J, Cui L, Wall K, Miller WC, Depamphilis CW. EST clustering error evaluation and correction. **Bioinformatics**, v. 20, p. 2973-2984, 2004.
- Wang J, Xia Q, He X, Dai M, Ruan J, Chen J, Yu G, Yuan H, Hu Y, Li R, Feng T, Ye C, Lu C, Wang J, Li S, Wong GK, Yang H, Wang J, Xiang Z, Zhou Z, Yu J. SilkDB: a knowledgebase for silkworm biology and genomics. **Nucleic Acids Research**, v. 1, p. 399-402, 2005.
- Wasilewski MM, Lim KC, Phillips J, McKerrow JH. Cysteine protease inhibitors block schistosome hemoglobin degradation in vitro and decrease worm burden and egg production in vivo. **Molecular Biochemistry of Parasitology**, v. 81, p. 179-189, 1996.
- Weiss KM. In search of human variation. **Genome Research**, v. 8, p. 691-697, 1998.
- Weston D, Schmitz J, Kemp WM, Kunz W. Cloning and sequencing of a complete myosin heavy chain cDNA from *Schistosoma mansoni*. **Molecular and Biochemical Parasitology**, v. 58, p. 161-164, 1993.
- Wong GK, Yang Z, Passey DA, Kibukawa M, Paddock M, Liu CR, Bolund L, Yu J. A population threshold for functional polymorphisms. **Genome Research**, v. 13, p. 1873-1879, 2003.

Wong GK, Liu B, Wang J, Zhang Y, Yang X, Zhang Z, Meng Q, Zhou J, Li D, Zhang J, Ni P, Li S, Ran L, Li H, Zhang J, Li R, Li S, Zheng H, Lin W, Li G, Wang X, Zhao W, Li J, Ye C, Dai M, Ruan J, Zhou Y, Li Y, He X, Zhang Y, Wang J, Huang X, Tong W, Chen J, Ye J, Chen C, Wei N, Li G, Dong L, Lan F, Sun Y, Zhang Z, Yang Z, Yu Y, Huang Y, He D, Xi Y, Wei D, Qi Q, Li W, Shi J, Wang M, Xie F, Wang J, Zhang X, Wang P, Zhao Y, Li N, Yang N, Dong W, Hu S, Zeng C, Zheng W, Hao B, *et al.*, Yang H; International Chicken Polymorphism Map Consortium. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. **Nature**, v. 432, p. 717-722, 2004.

World Health Organization (WHO). Tropical Diseases Research. Schistosomiasis and soil transmitted helminth infections. Fifty-fourth World Health Assembly, resolution WHA54.19. Geneva: World Health Organization, 2001.

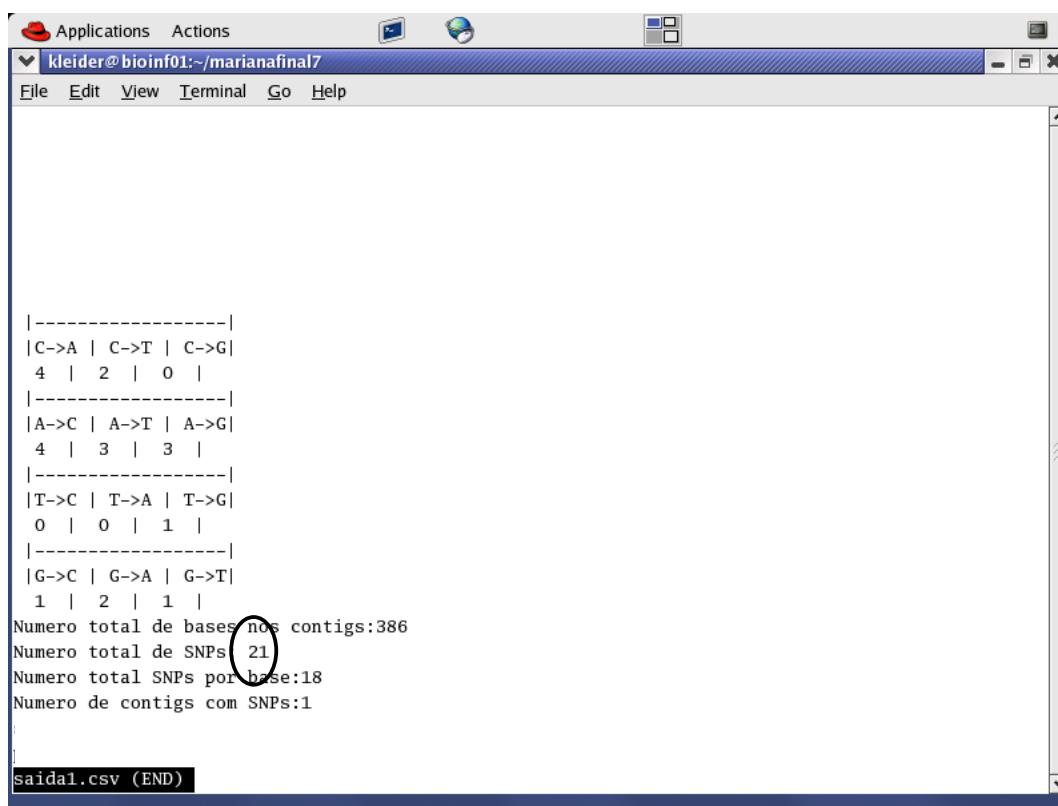
Wright MD, Henkle KJ, Mitchell GF. An immunogenic Mr 23,000 integral membrane protein of *Schistosoma mansoni* worms that closely resembles a human tumor-associate antigen. *The Journal of Immunology*, v. 144, p. 3195-3200, 1990.

IX - APÊNDICE

Novo Programa de detecção de SNPs : cSNPer

Experimento Modelo

Para a realização deste experimento nós selecionamos um contig modelo, zeramos todas os valores de qualidade das bases, tanto na seqüência consenso quanto nas respectivas ESTs. Cada parâmetro foi alterado por vez e o programa executado dentro das metas exigidas. As metas foram variando de acordo com cada objetivo na identificação dos SNPs. O mesmo contig foi utilizado em todas as etapas. Os resultados foram analisados e as Figuras de cada alteração estão representadas a seguir.

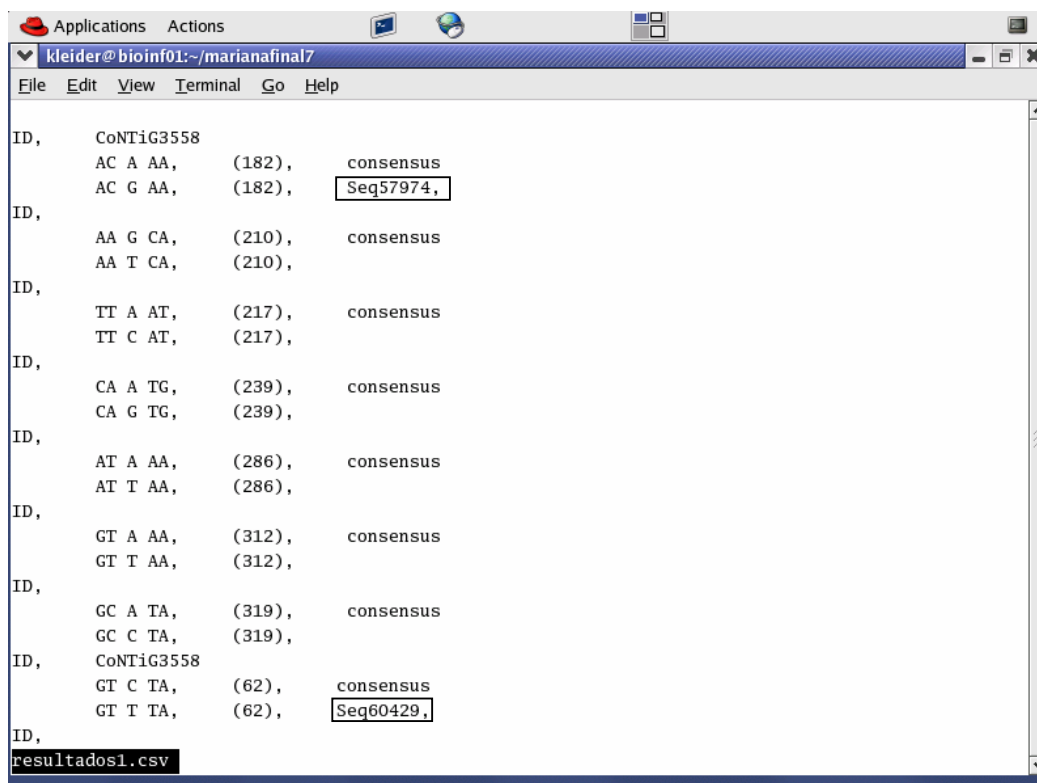
Arquivo de Saída:

```
Applications  Actions
kleider@bioinf01:~/marianafinal7
File Edit View Terminal Go Help

|-----|
|C->A | C->T | C->G|
 4 | 2 | 0 |
|-----|
|A->C | A->T | A->G|
 4 | 3 | 3 |
|-----|
|T->C | T->A | T->G|
 0 | 0 | 1 |
|-----|
|G->C | G->A | G->T|
 1 | 2 | 1 |
Numero total de bases nos contigs:386
Numero total de SNPs: 21
Numero total SNPs por base:18
Numero de contigs com SNPs:1

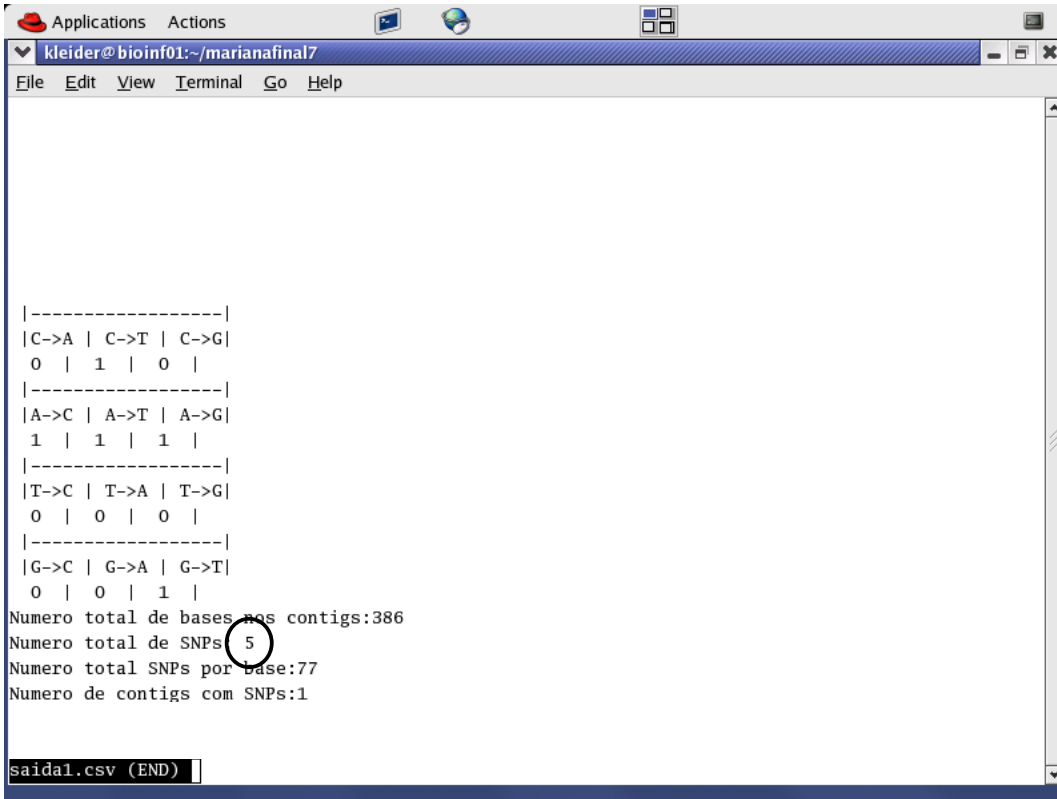
saida1.csv (END)
```

Figura 2 – Arquivo de saída gerado pelo cSNPer, com os valores totais, detectando o total de 21 SNPs, desconsiderando os valores de qualidade das bases.



```
Applications  Actions
kleider@bioinf01:~/marianafinal7
File Edit View Terminal Go Help
ID, CoNTiG3558
AC A AA, (182), consensus
AC G AA, (182), Seq57974,
ID,
AA G CA, (210), consensus
AA T CA, (210),
ID,
TT A AT, (217), consensus
TT C AT, (217),
ID,
CA A TG, (239), consensus
CA G TG, (239),
ID,
AT A AA, (286), consensus
AT T AA, (286),
ID,
GT A AA, (312), consensus
GT T AA, (312),
ID,
GC A TA, (319), consensus
GC C TA, (319),
ID, CoNTiG3558
GT C TA, (62), consensus
GT T TA, (62), Seq60429,
ID,
resultados1.csv
```

Figura 3 – Resultado gerado pelo cSNPer para cada contig, mostrando os SNPs detectados e suas respectivas posições, na seqüência consenso e nas ESTs. Nessa Figura está presente, somente, os SNPs referentes a Seq57974 e um SNP da Seq 60429.

Arquivo de Saída:

```
Applications Actions
kleider@bioinf01:~/marianafinal7
File Edit View Terminal Go Help

|-----|
|C->A | C->T | C->G|
0 | 1 | 0 |
|-----|
|A->C | A->T | A->G|
1 | 1 | 1 |
|-----|
|T->C | T->A | T->G|
0 | 0 | 0 |
|-----|
|G->C | G->A | G->T|
0 | 0 | 1 |
Numero total de bases:386 contigs:386
Numero total de SNPs: 5
Numero total SNPs por base:77
Numero de contigs com SNPs:1

saida1.csv (END)
```

Figura 5 – Arquivo de saída gerado pelo cSNPer, com os valores totais, detectando o total de 05 SNPs, com o valor de qualidade 10 na seqüência consenso. Foram realizadas apenas 4 alterações (Figura 1), porém existem duas variações na mesma posição para diferentes ESTs (182 pb).

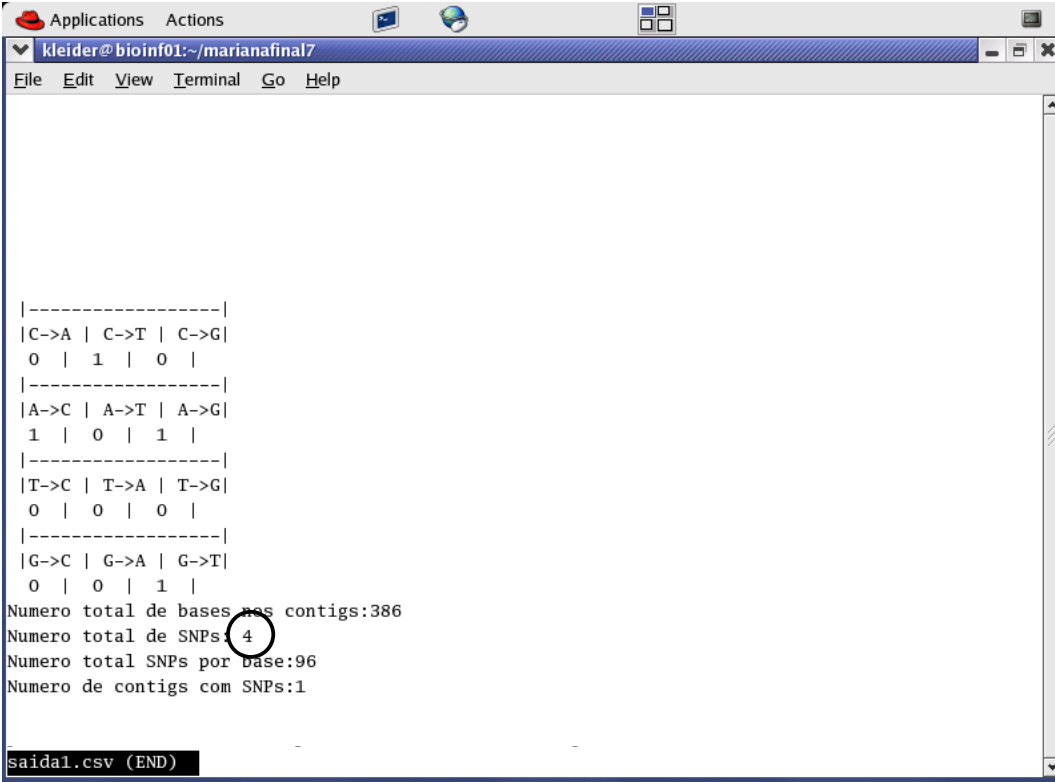
```

Applications Actions
kleider@bioinf01:~/marianafinal7
File Edit View Terminal Go Help

ID, CoNTiG3558
AC A AA, (182), consensus
AC G AA, (182), Seq57974,
ID,
AA G CA, (210), consensus
AA T CA, (210),
ID,
TT A AT, (217), consensus
TT C AT, (217),
ID, CoNTiG3558
GT C TA, (62), consensus
GT T TA, (62), Seq60429,
ID, CoNTiG3558
AC A AA, (182), consensus
AC T AA, (86), Seq58290,
NQS, 0
Numcontigs:1 \ N_snp:5
resultados1.csv (END)

```

Figura 6 – Resultado gerado pelo cSNPer, mostrando os 5 SNPs identificados apresentando qualidade 10 na seqüência consenso.

Arquivo de Saída:

```
Applications Actions
kleider@bioinf01:~/marianafinal7
File Edit View Terminal Go Help

|-----|
|C->A | C->T | C->G|
0 | 1 | 0 |
|-----|
|A->C | A->T | A->G|
1 | 0 | 1 |
|-----|
|T->C | T->A | T->G|
0 | 0 | 0 |
|-----|
|G->C | G->A | G->T|
0 | 0 | 1 |
Numero total de bases:386 contigs:386
Numero total de SNPs: 4
Numero total SNPs por base:96
Numero de contigs com SNPs:1

saida1.csv (END)
```

Figura 9 – Arquivo de saída gerado pelo cSNPer; no círculo o total de 4 SNPs identificados nos parâmetros exigidos.

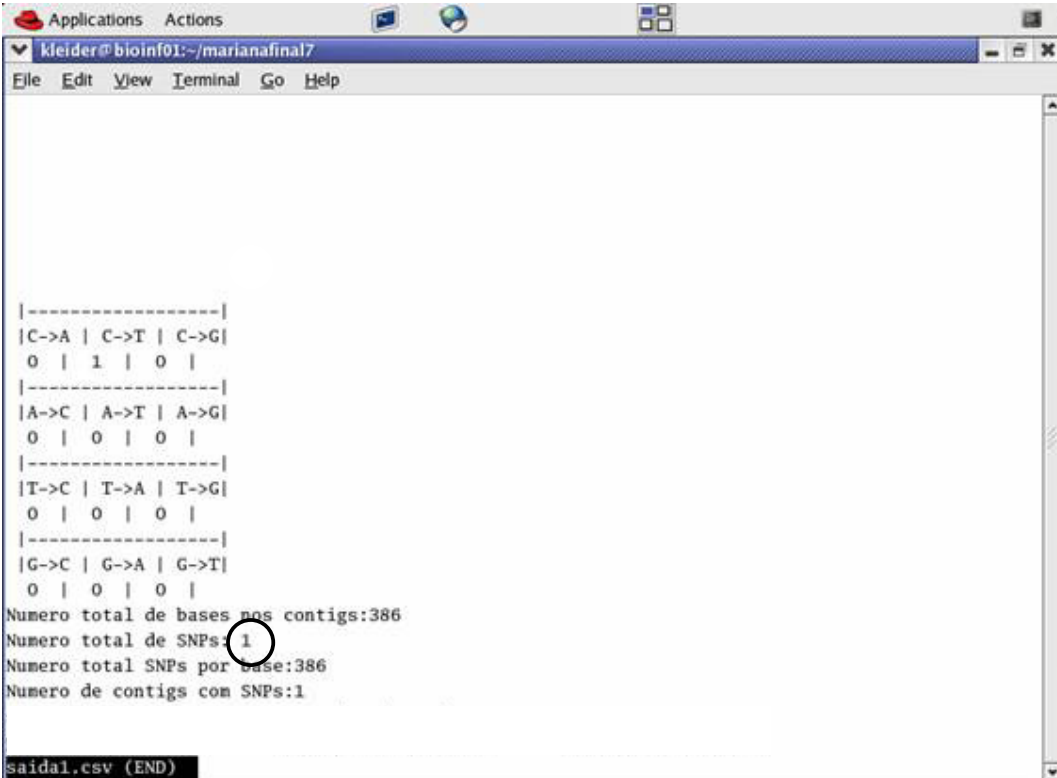
```

Applications  Actions
kleider@bioinf01:~/marianafinal7
File  Edit  View  Terminal  Go  Help

ID,      CoNTiG3558
AC A AA, (182),  consensus
AC G AA, (182),  Seq57974,
ID,
AA G CA, (210),  consensus
AA T CA, (210),
ID,
TT A AT, (217),  consensus
TT C AT, (217),
ID,      CoNTiG3558
GT C TA, (62),   consensus
GT T TA, (62),   Seq60429,
NQS,    1
Numcontigs:1 \ N_snp:4
resultados1.csv (END)

```

Figura 10 – Resultado gerado pelo cSNPer, mostrando detalhadamente os 4 SNPs identificados com qualidade 10 na sequência consenso e qualidade 20 na EST.

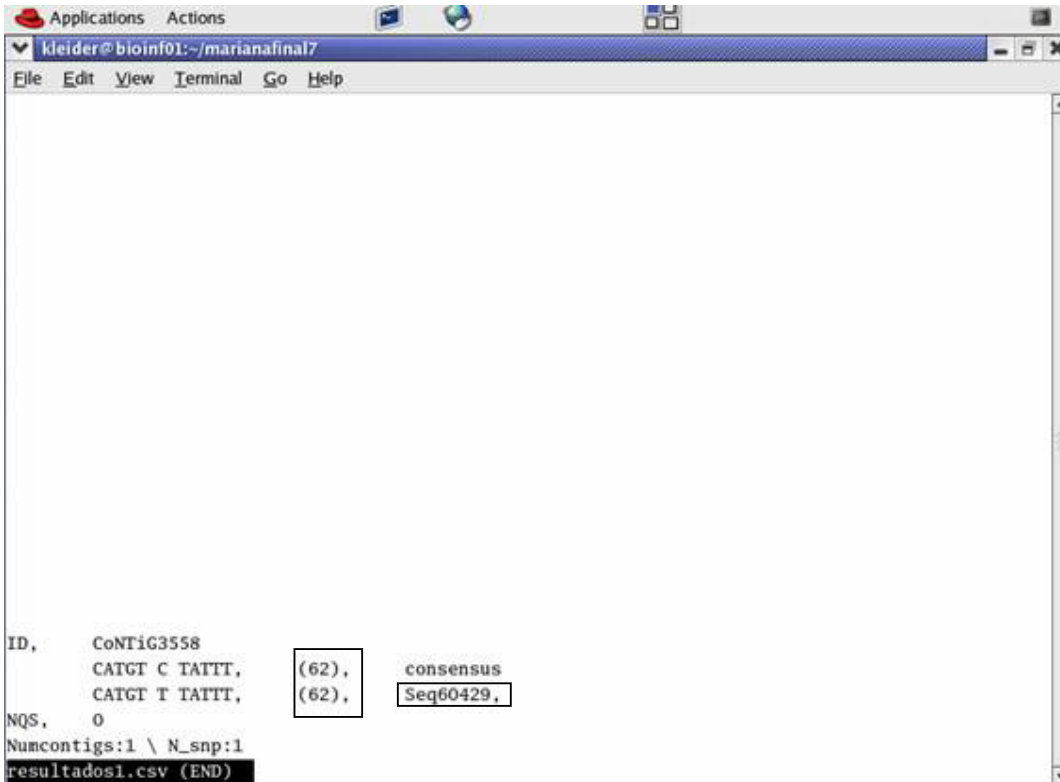
Arquivo de Saída:

```
Applications Actions
kleider@bioinf01:~/marianafinal7
File Edit View Terminal Go Help

|-----|
|C->A | C->T | C->G|
0 | 1 | 0 |
|-----|
|A->C | A->T | A->G|
0 | 0 | 0 |
|-----|
|T->C | T->A | T->G|
0 | 0 | 0 |
|-----|
|G->C | G->A | G->T|
0 | 0 | 0 |
Numero total de bases nos contigs:386
Numero total de SNPs: 1
Numero total SNPs por base:386
Numero de contigs com SNPs:1

saida1.csv (END)
```

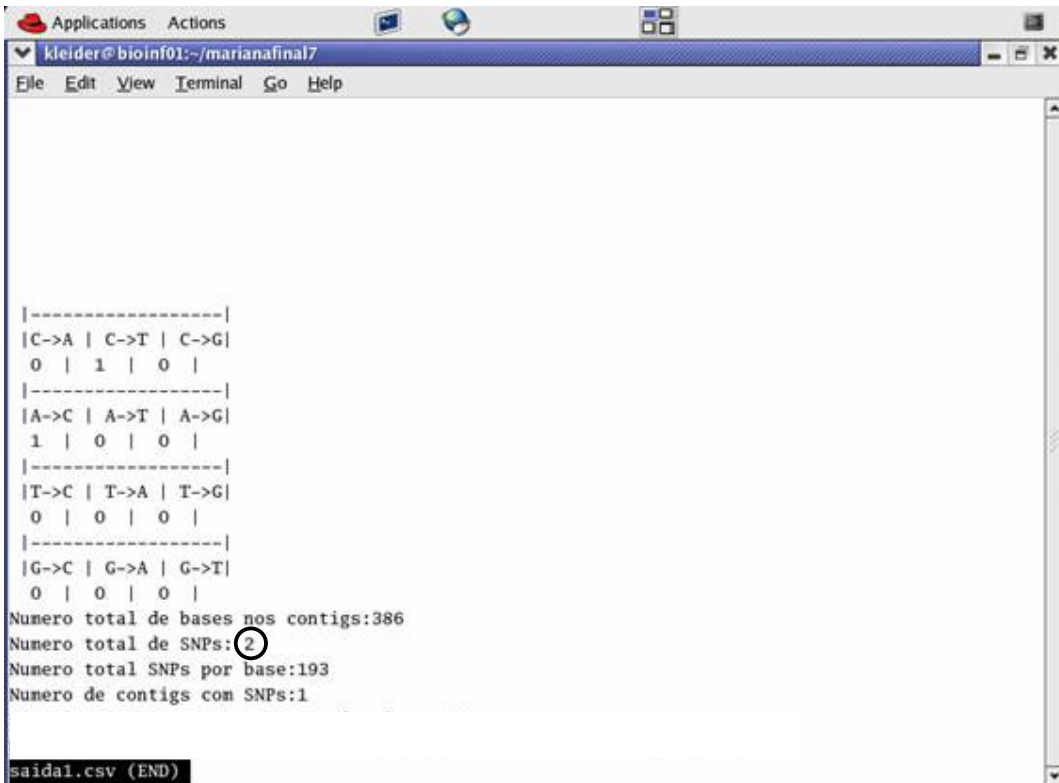
Figura 12 – Resultado do cSNPer, apenas um SNP foi identificado na posição 62 com qualidade na seqüência consenso 10, qualidade na EST 20, qualidade das bases vizinhas na seqüência consenso 15 e alinhamento de pelo menos 5 pb.



```
Applications Actions
kleider@bioinf01:~/marianafinal7
File Edit View Terminal Go Help

ID,      CoNTiG3558
        CATGT C TAITT, (62),  consensus
        CATGT T TAITT, (62),  Seq60429,
NQS,    0
Numcontigs:1 \ N_snp:1
resultados1.csv (END)
```

Figura 13 – Resultado mostrando o SNP na posição 62 identificado dentro dos parâmetros.

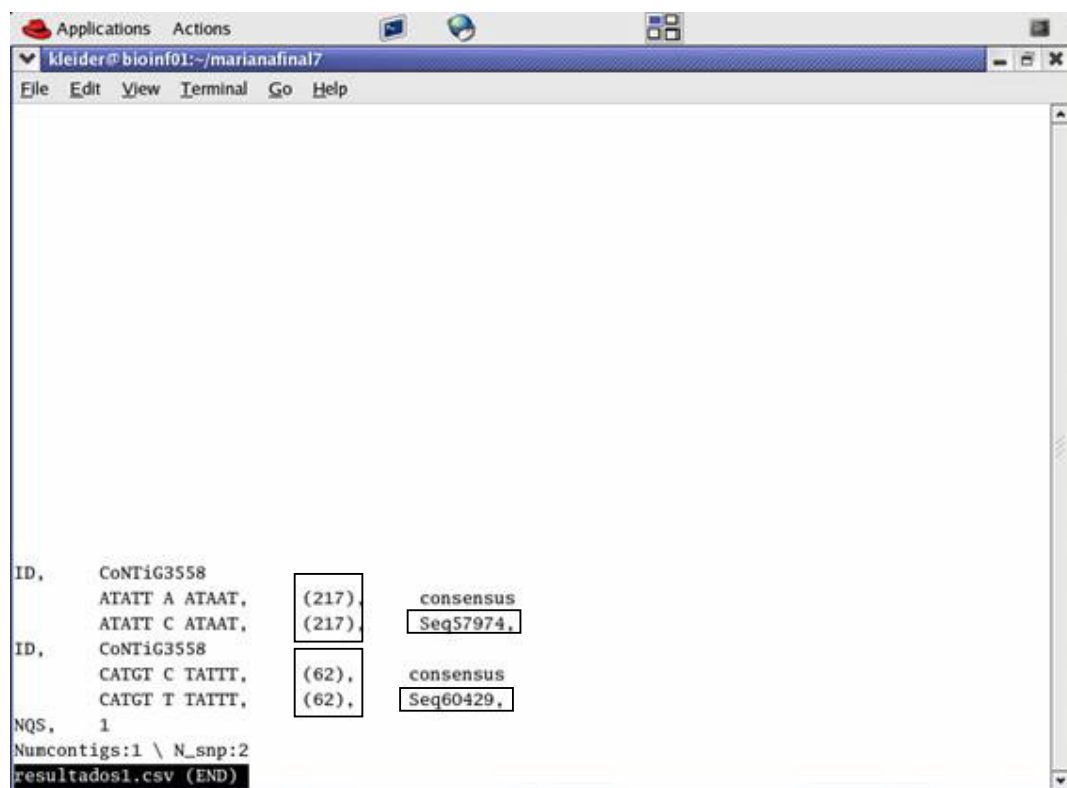
Arquivo de Saída:

```
Applications Actions
kleider@bioinf01:~/marianafinal7
File Edit View Terminal Go Help

|-----|
|C->A | C->T | C->G|
0 | 1 | 0 |
|-----|
|A->C | A->T | A->G|
1 | 0 | 0 |
|-----|
|T->C | T->A | T->G|
0 | 0 | 0 |
|-----|
|G->C | G->A | G->T|
0 | 0 | 0 |
Numero total de bases nos contigs:386
Numero total de SNPs:2
Numero total SNPs por base:193
Numero de contigs com SNPs:1

saida1.csv (END)
```

Figura 15 – Arquivo de saída geral do cSNPer, total de 2 SNPs identificados.



```
Applications  Actions
kleider@bioinf01:~/marianafinal7
File Edit View Terminal Go Help

ID,      CoNTiG3558
         ATATT A ATAAT, (217), consensus
         ATATT C ATAAT, (217), Seq57974,
ID,      CoNTiG3558
         CATGT C TATTT, (62), consensus
         CATGT T TATTT, (62), Seq60429,
NQS,    1
Numcontigs:1 \ N_snp:2
resultados1.csv (END)
```

Figura 16 – Resultado mostrando os dois SNPs identificados, suas posições e sequência referência.

6 - Detecção de SNPs com Phred ≥ 10 na seqüência consenso e Phred ≥ 20 na EST e 5 bases vizinhas alinhadas na seqüência consenso com Phred ≥ 15 e na EST com Phred ≥ 30

Qualidade do SNP na seqüência consenso: **10**;

Qualidade do SNP na EST: **20**;

Qualidade das bases vizinhas ao SNP na consenso: **15**;

Qualidade das bases vizinhas ao SNP na EST: **30**;

Alinhamento das bases vizinhas ao SNP: **5**.

Arquivo de entrada:

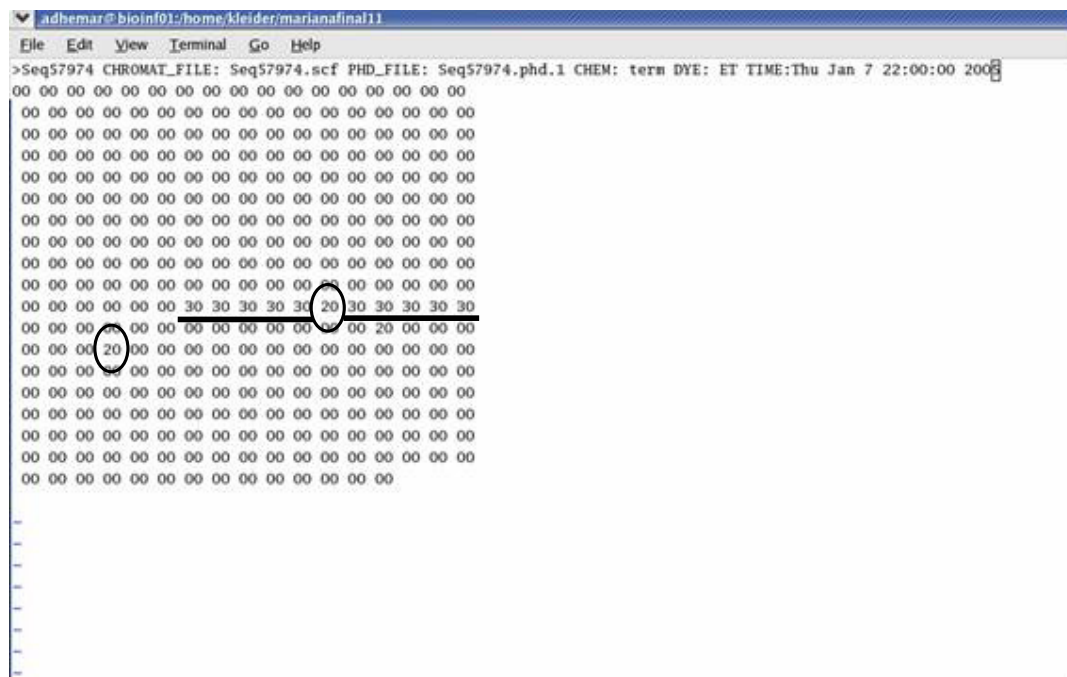


Figura 17 – Arquivo de entrada mostrando a alteração realizada nas qualidades das bases na EST. Circulado a base polimórfica com valor de Phred = 20 e sublinhado a vizinhança com Phred = 30.

Arquivos de saída:

```
|-----|
|C->A | C->T | C->G|
0 | 0 | 0 |
|-----|
|A->C | A->T | A->G|
0 | 0 | 1 |
|-----|
|T->C | T->A | T->G|
0 | 0 | 0 |
|-----|
|G->C | G->A | G->T|
0 | 0 | 0 |
Numero total de bases: 386 contigs:386
Numero total de SNPs: 1
Numero total SNPs por base: 386
Numero de contigs com SNPs:1
saida.csv (EMD)
```

Connected to bioinfo.cpqr.fiocruz.br SSH2 - aes128-cbc - hmac-md5 - none 142x40

Figura 18 – Dentro dos parâmetros exigidos apenas 1 SNP foi detectado, como esperado.

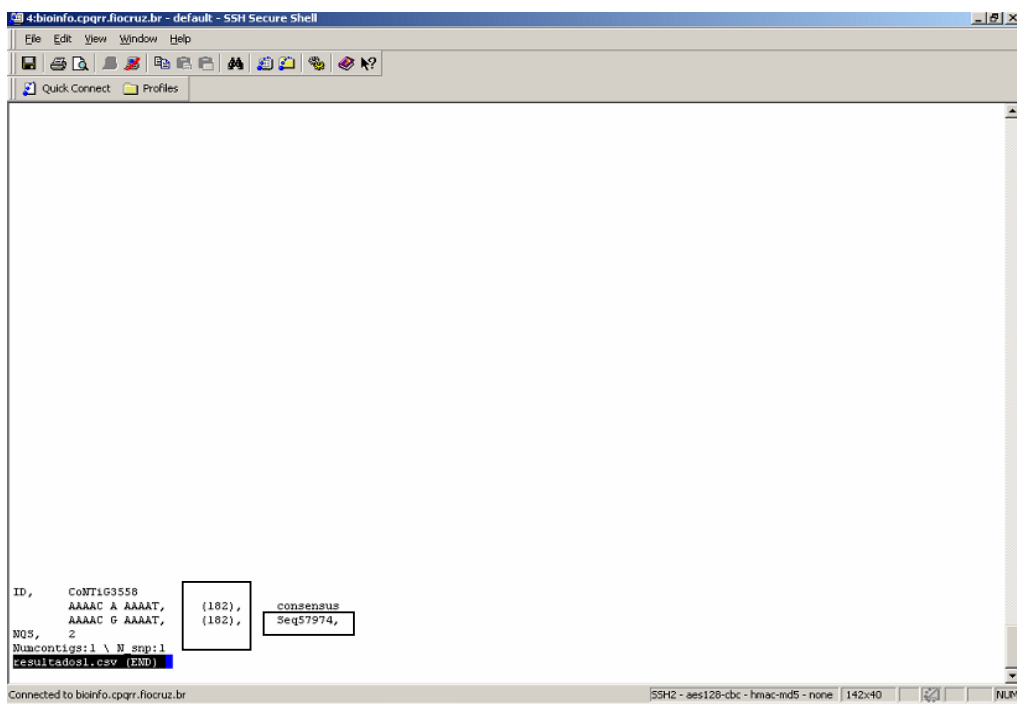


Figura 19 – Arquivo de saída mostrando o SNP detectado e sua respectiva posição.

7) Análise do SNP da posição 182 e sua tradução nas 6 janelas de leitura, aceitando ORF com no mínimo 30 aminoácidos.

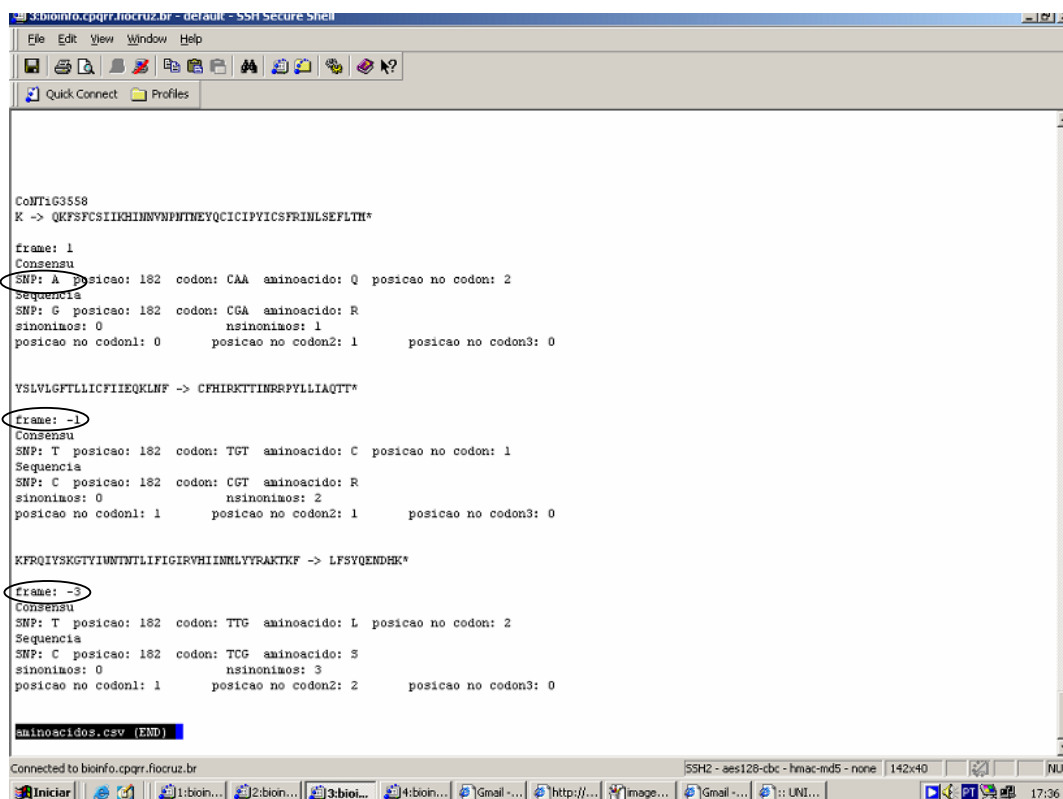


Figura 20 – Detecção de 3 possíveis ORFs para o SNP 182, contendo no mínimo 30 aminoácidos.

Translate Tool - Results of translation

Please select one of the following frames:

[5'3' Frame 1](#)

```

gtgacttgaaaacattaaaagtttcatcaacaaaagtacataaaaactttgatogaacatg
V T - K H - K F H Q Q K Y I K L - S N M
tctatTTTTAAAAAAGTATTTTTGTTTGTATTCATCCAGGTAATCAAATCOA
S I F - K S Y F L F R I H P R V N Q I Q
gttgttgagcaatgagtaaatatggtctctattatggtgtttctgatagaaa
V V - A M S K Y G L L F M V V F L I | - K
cgaaaatttagttttgctctataataaatatcataatggaacctaataccact
Ⓢ K F S F C S I I N H I H N V N P N T T
gaatocagtgtattgtatttctatatatatattagaattaattgtctgaattt
E Y Q C I C I S Y I Y I F R I N C L N F
taacgatgtaaatgcctat
- R C - S P

```

[5'3' Frame 2](#)

```

gtgacttgaaaacattaaaagtttcatcaacaaaagtacataaaaactttgatogaacatgt
- L E N I K S F I N K S T - N F D R T C
ctatTTTTAAAAAAGTATTTTTGTTTGTATTCATCCAGGTAATCAAATCOAAG
L F F K K V I F C F V F I H E L I K F K
ttgttgagcaatgagtaaatatggtctctattatggtgtttctgatagaaa

```

Figura 21 – A seqüência do contig foi traduzida pelo programa Expasy e foi analisado manualmente a detecção do programa. A janela +1 detectada pelo programa está dentro da ORF mínima exigida. O polimorfismo está circulado.

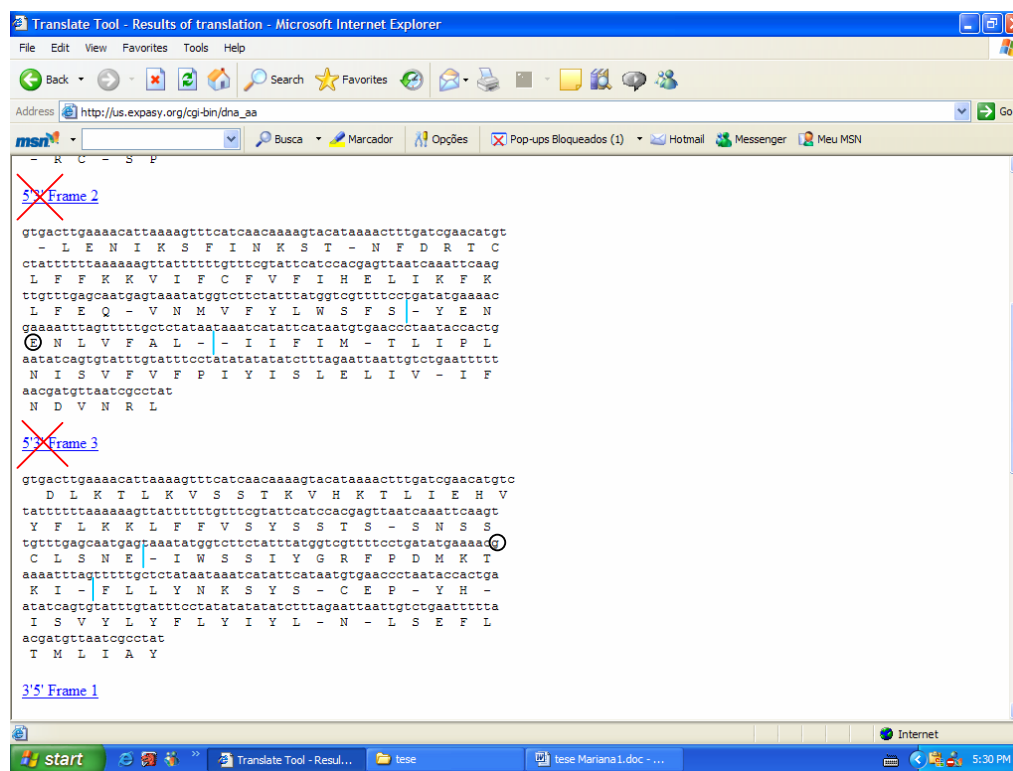


Figura 22 – As janelas com um X não foram selecionados pelo programa e de acordo com o programa Espasy, o polimorfismo está presente em um ORF abaixo de 30 aminoácidos, não sendo assim selecionada. As demais janelas de leitura foram analisadas da mesma forma.