Available online at www.sciencedirect.com
ScienceDirect
Molecular & Biochemical Parasitology 154 (2007) 134–140
MOLECULAR & BIOCHEMICAL PARASITOLOGY

Single nucleotide polymorphisms identification in expressed genes of *Schistosoma mansoni*

Mariana Simões [a], Diana Bahia [a], Adhemar Zerlotini [a], Kleider Torres [a], François Artiguenave [c], Goran Neshich [b], Paula Kuser [b], Guilherme Oliveira [a,d,*]

[a] *Laboratory of Cellular and Molecular Parasitology, Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Av. Augusto de Lima 1715, Belo Horizonte 30190-002, MG, Brazil*
[b] *Department of Structural Bioinformatics, Embrapa/Informática Agropecuária, Rua André Tosello, 209 Cidade Universitária, Campinas 13083-886, SP, Brazil*
[c] *INRA-Unité de Recherche Genomique Info, 2 rue Gaston Crémieux, CP5708, 91057 Evry Cedex, France*
[d] *Programa de Pós-Graduação e Pesquisa, Santa Casa de Belo Horizonte, Av. Francisco Sales 111, 9⁰ andar, Ala C, Belo Horizonte 30150-221, MG, Brazil*

Received 24 January 2007; received in revised form 30 March 2007; accepted 7 April 2007
Available online 13 April 2007

Abstract

Single nucleotide polymorphism (SNP) markers have been shown to be useful in genetic investigations of medically important parasites and their hosts. In this paper, we describe the prediction and validation of SNPs in ESTs of *Schistosoma mansoni*. We used 107,417 public sequences of *S. mansoni* and identified 15,614 high-quality candidate SNPs in 12,184 contigs. The presence of predicted SNPs was observed in well characterized antigens and vaccine candidates such as those coding for myosin; Sm14 and Sm23; cathepsin B and triosephosphate isomerase (TPI). Additionally, SNPs were experimentally validated for the cathepsin B. A comparative model of the *S. mansoni* cathepsin B was built for predicting the possible consequences of amino acid substitutions on the protein structure. An analysis of the substitutions indicated that the amino acids were mostly located on the surface of the molecule, and we found no evidence for a significant conformational change of the enzyme. However, at least one of the substitutions could result in a structural modification of an epitope.
© 2007 Elsevier B.V. All rights reserved.

Keywords: *Schistosoma mansoni*; Single nucleotide polymorphism; Computational biology; Cathepsin B; Comparative modeling

1. Introduction

The availability of genome sequences and a large number of transcriptome sequencing initiatives opens new doors for the discovery of a class of polymorphic molecular markers called single nucleotide polymorphisms (SNPs). SNPs are the most abundant type of genetic variation between individuals and can provide information about phenotypic differences. Owing to their high density, the exploitation of SNPs for marker assays has the potential to provide answers to a large number of important biological, genetic, pharmacological and medical questions [1]. Identifying the polymorphisms in relation to disease predisposition and drug response is a major aim of the post genomic era.

Many of the recent efforts to describe the genomes of organisms focus on the generation of expressed sequence tags (ESTs) by partial sequencing of cDNAs. ESTs have been extensively used for gene discovery, expression analysis and transcript

Abbreviations: bp, base pair; DNA, deoxyribonucleic acid; PCR, polymerase chain reaction; RNA, ribonucleic acid; SNP, single nucleotide polymorphism; TPI, triosephosphate isomerase; Sm31 and SMCB1, cathepsin B-like cysteine proteinase precursor; Sm14, *Schistosoma mansoni* fatty acid-binding protein; Sm23, *Schistosoma mansoni* integral membrane protein; cDNA, complementary DNA; cSNP, coding region single nucleotide polymorphism; RT-PCR, reverse transcriptase-polymerase chain reaction; NQS, neighborhood quality standard; PDB, Protein Data Bank; nsSNP, nonsynonymous SNP; ORF, open reading frame

* Corresponding author at: Laboratory of Cellular and Molecular Parasitology, Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Av. Augusto de Lima 1715, Belo Horizonte 30190-002, MG, Brazil. Tel.: +55 31 3349 7785; fax: +55 31 3295 3115.
E-mail address: oliveira@cpqrr.fiocruz.br (G. Oliveira).

0166-6851/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.molbiopara.2007.04.003

mapping of genes from a wide variety of organisms, including *Schistosoma mansoni* [2]. The transcriptome, however, lacks information on regulatory sequences, intergenic regions and introns. Currently, in depth information on genetic variation in Schistosomes is obtained with polymorphic microsatellite markers, generally located in non-coding regions [3]. In contrast, SNPs have been identified directly in coding regions (cSNPs) with a software-based approach using large numbers of redundant ESTs data sets [4–6]. Nevertheless, up to the present investigation, such molecular markers have not been studied on a large scale in *S. mansoni*. Therefore, in this study we focused on SNPs in gene coding regions of *S. mansoni*.

Schistosomiasis remains a major public health problem in Africa, Asia and parts of South America, despite strenuous efforts to control its impact on human populations. The disease is caused by digenetic blood trematodes, with *S. mansoni* being the only human infecting species in South America and one of the two most relevant species in Africa. Disease control efforts are mainly based on mass chemotherapy, as there is no available vaccine [7]. The study of the genetic variation in *S. mansoni* parasites has practical significance for developing additional strategies to control the disease. This information could be used for the study of transmission dynamics (as genetic markers) or for observing the variability of antigens and drug targets [8,9]. In this study, we developed an automated pipeline to detect SNPs *in silico* in ESTs of *S. mansoni* using high-quality sequences and alignment parameters. Furthermore, we observed the predicted SNPs in vaccine target candidates, validated putative SNPs in the cathepsin B gene and analyzed model variant proteins for possible conformational modifications. Detailed experiment outcomes, including SNP information and EST assemblies are available at http://bioinfo.cpqrr.fiocruz.br/snp.

2. Materials and methods

2.1. Sequence data sets and polymorphism identification

We used public expressed sequence tags (ESTs) generated by Verjovski-Almeida et al. [2], including quality information of the bases obtained with Phred download from the web site mentioned in the manuscript [10,11]. The sequences were assembled into contigs using CAP3 [12].

To automate the process of SNP prediction, we developed cSNPer—a new program to detect SNPs. cSNPer reads the ACE file generated by CAP3 to identify candidate SNPs. To calculate a Neighborhood Quality Standard (NQS), the software considered qualities of the putative SNP in the ESTs (Phred $Q \geq 20$), in the consensus sequence (Phred $Q \geq 40$) and of the 10 bases 5′ and 3′ the putative SNP (Phred $Q \geq 15$). After putative SNP identification, cSNPer also detected ORFs containing a minimum of 150 amino acids, the position of the SNP in the codon and the coded amino acid.

Sequence coding for vaccine candidates were identified by Blast ($E = 0$) against the following: Sm14 (GenBank accession no. M60895), Sm23 (GenBank accession no. M34453), cathepsin B (GenBank accession no. M21309), Sm28 (GST, 28 kDa glutathione *S*-transferase, GenBank accession no. S71584), myosin (GenBank accession no. X65591), paramyosin (GenBank accession no. M35499) and TPI (triose phosphate isomerase, SGTPI, GenBank accession no. AH001087).

2.2. SNP validation

2.2.1. RNA extraction

To validate SNPs, RNA was extracted from pools of *S. mansoni* adult worms of the Puerto Rican strain and field isolates. Worm tissues were homogenized in guanidinium solution as previously described [13]. The total RNA obtained was treated with DNAse I and 5 μg of total RNA were used for cDNA synthesis with Thermoscript RT-PCR System (Invitrogen, Carlsbad, CA, USA).

2.2.2. PCR amplification and sequencing of S. mansoni cathepsin B gene

For PCR, primers flanking the polymorphic regions were automatically designed using Primer3 software [14]. PCR amplification was conducted with primers pairs: PCR-1: Sm31-X5 (5′-ATTCAAGAGTTATTTGGACATGC-3′)/Sm31-X6 (5′-

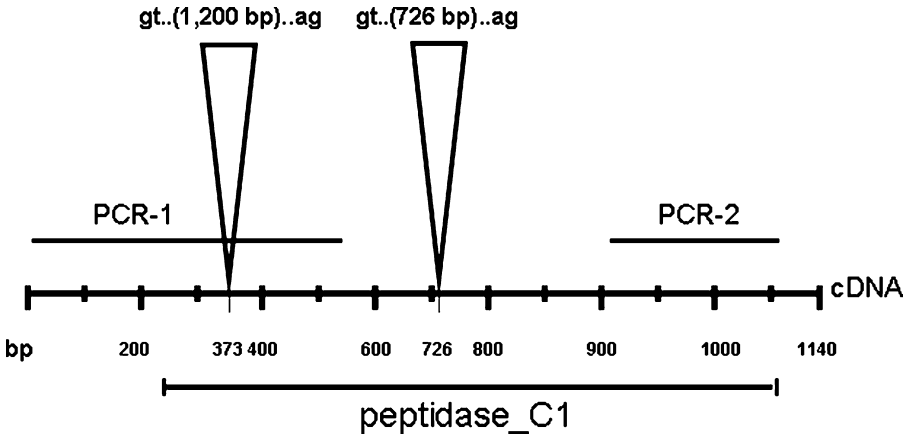

Fig. 1. The cathepsin B gene of *Schistosoma mansoni*. There are two introns (triangles) and three exons (black line). The introns are of 1200 and 726 bp and flanked by gt/ag donor and acceptor sites, respectively. Primers Sm31-X5/X6 were used to amplify a 536 bp cDNA fragment (PCR-1) and primers Sm31X7/X8 were used to obtain the 210 bp (PCR-2), both containing putative SNPs. The peptidase_C1 domain is shown below the cDNA.

CCTGCTTGGGATTACTGGGTGAAGG-3′) and PCR-2: Sm31-X7 (5′-ATAAAAGCTTACAAGACTCCTTATTGG-TT-3′)/Sm31-X8 (5′-AATAAAGCTTTTTGAAGTATTCAGT-ATACA-3′) (Fig. 1). The size-selected fragments of 536 and 210 bp, respectively, were cloned into the TOPO vector using the TOPO TA cloning kit 2.1 (Invitrogen). Plasmid DNA was prepared using the QIAprep Spin Miniprep Kit (Qiagen, USA), and sequenced using the ThermoSequence II dye terminator cycle sequencing kit (GE Healthcare, USA) in a MegaBace 500 sequencer (GE Healthcare, USA). The nucleotide sequences were analyzed and aligned using GeneTool version 2.0 [15]. The quality of the DNA sequences and individual SNPs were assessed by visual inspection of the chromatograms.

2.3. Comparative modeling

Comparative protein structure modeling by satisfaction of spatial restraints, as implemented in the program Modeller (version 7.0) [16] (http://salilab.org/modeller), was used to build three-dimensional models of cathepsin B and to perform structural modeling of amino acids substitutions. The models relied on a template of the human procathepsin B [17], previously determined by X-ray crystallography. This template was selected due to the high identity (∼51%) between the two sequences. For each point mutation, Modeller was used to perform side chain replacement on the cathepsin B model. Modeller was also used to perform energy minimization while employing constraints to the atomic coordinates of aligned residues in the parent structure, which allowed for the generation of a refined homology model. Features of the amino acid substitutes were observed with the Java Protein Dossier module of the software STING [18,19]. The evaluated features for each amino acid substitution were: conservation, change in solvent accessibility, side-chain volume change, effect on amino acid interactions, amino acid electrostatics and other physicochemical properties of amino acids.

3. Results and discussion

3.1. SNP identification

A large number of studies have focused on investigating genetic polymorphisms in individual genes in order to estimate the genetic contribution to the disease outcome. ESTs have been used to mine SNPs in several model organisms, including parasites in a limited manner [20,21]. In this paper, we describe the use of 107,417, representing the near complete transcriptome of *S. mansoni*, ESTs from cDNA libraries of different stages of *S. mansoni* to identify SNPs. A summary of the results is in Table 1, indicating a total of 8,938,265 bp scanned and 15,615 possible SNPs identified in 12,184 clusters. We believe that the stringent parameters used improved the quality of the predicted SNPs. In fact, the number of SNPs did not correlate with the depth of the cluster (number of EST/consensus sequence length, data not shown). This observation indicates that the putative SNPs are not due to the accumulation of sequencing errors in the ESTs.

Table 1
Summary of single nucleotide polymorphisms identified *in silico*

Reads	107,417
Clusters	12,184
Total bp	8,938,265
SNPs detected	15,615
bp/SNP	572
Transitions	11,658
Transversions	3,957
Synonimous	1,832
Nonsynonimous	758

SNPs were classified on the basis of nucleotide substitution as either transversions or transitions. Transitions were observed in 11,658 SNPs (74.66%) and transversions in 3957 (25.34%). The frequency of substitutions was 493 (3.16%) G/C, 784 (5.02%) A/C, 821 (5.26%) G/T, 1859 (11.91%) A/T, 5672 (36.32%) C/T, and 5986 (38.33%) A/G. The ratio of transitions to transversions was 2.95. The higher number of transitions we observed was also detected by other authors in different organisms [22,23]. A higher frequency of transitions is usually related to the deamination process of 5-methilcytosine into uracil [24]. However, Fantappie et al. [25] did not detect the presence of methylation on the genome of *S. mansoni*. In this context, it is unclear why a higher frequency of transitions was observed.

SNPs can also be classified according to the change of coded amino acids into synonymous and nonsynonymous mutations. We observed that a total of 1832 (70.73%) putative SNPs resulted in synonymous amino acid substitutions and 758 (29.26%) in nonsynonymous amino acid substitutions (synonymous/nonsynonymous ratio of 2.41). Other studies have also reported a higher frequency of synonymous mutations when compared with nonsynonymous [26,27]. Thus, the relative proportions of synonymous and nonsynonymous substitutions observed for *S. mansoni* is in agreement with observations made by other groups in different organisms. This observation also confirms the SNP codon position analysis, which indicated 777 (30%) were in the first codon base, 880 (33.98%) in the second, and 933 (36.02%) in the third. A total of 7 stop codons were eliminated and 97 created.

3.2. Identification of SNPs in known antigen coding genes

After a global analysis of EST polymorphisms, we next analyzed the polymorphic profile of *S. mansoni* genes coding for vaccine candidates. The total number of SNPs in coding and non-coding regions for each vaccine candidate were 13 and 4 for Sm14 (two contigs), 26 for Sm23, 17 and 9 for cathepsin B (two contigs), 26 for GST, 6 for myosin, 6 for paramyosin and 3 for TPI (see on line material).

Knowledge of possible polymorphisms in vaccine candidates and drug targets is extremely important, as the presence of nonsynonymous mutations may lead to protein structural alterations that may affect epitope or drug binding sites [28]. However, it is worth noting that nonsynonymous mutations, in rare cases, may also alter the gene transcription process by changing the splice site, for example [29].

The characterized SNPs may be used in the future for the study of possible antigenic variants for vaccine and drug development. In addition, genetic mapping of populations will benefit from polymorphic markers such as SNPs. Finally, the use of these genomic markers, when present in higher numbers, will be useful in efforts to localize genes of interest.

3.3. Validation of SNPs in cathepsin B vaccine gene

Despite the fact that bioinformatics provides an invaluable contribution for SNP identification, experimental validation is necessary. We sought to verify predicted SNPs in the *S. mansoni* cathepsin B vaccine candidate gene (Sm31 or SmCB1, GenBank ID: M21309). Cathepsin B is an essential peptidase for hemoglobin digestion by the parasite [30]. Parasite growth retardation was shown in adult worms with decreased transcript levels of cathepsin B by RNAi [31]. In addition to the vaccine potential, papain-like cysteine endopeptidases have been recognized as potential targets for chemotherapy and serodiagnostic reagents in infections with the human parasitic *Schistosoma* [32].

Polymorphic regions of the cathepsin B transcript were amplified from adult worm cDNA generating products of 536 bp (PCR-1) and 210 bp (PCR-2) (Fig. 1). A total of 150 clones were sequenced in both directions. The sequences were aligned and visual SNP verification was conducted in 83 high-quality sequences (Phred $\geq$ 20). As a result, a total of 16 different SNPs were identified, 44% were transitions and 56% were transversions. The frequency of codon distribution of the SNPs identified was of 17% for the first base, 12% for the second base, and 71% for the third base. Furthermore, synonymous substitutions were observed in 62% of the SNPs and nonsynonymous substitutions (nsSNPs) in 38%. The following nsSNPs amino acids changes were found: Val21Ile, Glu27Lys, Lys75Thr, Asp84Glu, Asn92Ser and Gly101Arg. The signal peptide cleavage site was identified and most of SNPs were found after the cleavage site at the C1 domain.

After identification of SNPs in *S. mansoni* adult worms of the LE strain, we checked for the presence of polymorphisms in individual parasites isolated from an endemic transmission site (Caju, Jequitinhonha river valley, State of Minas Gerais, Brazil). We compared the variations observed with the most frequent allele. The polymorphic positions found in the field isolates were the same as in the laboratory strain, again confirming the predicted SNPs. However, a higher polymorphism frequency was observed in most of field isolate cDNAs (Table 2). The observation that the same polymorphic sites were seen in both the LE strain and field isolates is in agreement with other studies of intra-specific variability using mitochondrial DNA [33]. As expected, a higher allele frequency was observed in field isolates, which is an observation also made with microsatellite markers [34].

3.4. Modeling the cathepsin B gene

To assess whether the amino acids substitutions related to identified SNPs may have a significant impact on the protein structure, comparative three-dimensional models of *S. mansoni* cathepsin B were built based on the human procathepsin B (PDB 3pbh). The human procathepsin molecule consists of a propeptide helix composed of 62 residues (numbered P1 to P62) and 254 enzyme residues (1–254). The homology model included residues 26–315 of *S. mansoni* cathepsin B (corresponding to amino acids 6P to 254 of human procathepsin B). The first 25 residues of *S. mansoni* cathepsin B were not modeled, and amino acids 40–86 (corresponding to amino acids 20P to 60P in the template) had very low similarity to the template. As expected from the conservative nature of comparative modeling, the three-dimensional model of *S. mansoni* cathepsin B shared high overall structure similarity with the template. The modeled structure contained two domains (R and L), with the known active site cleft of the human procathepsin B enzyme between the domains. The modeled structure also had the occluding loop region residues 192–214 (105–126 in human procathepsin B) packed onto the surface of the enzyme, which as similar to the template (Fig. 2).

Each nonsynonymous substitution was analyzed for their potential consequences in the modification of protein function. Two SNPs at bases 80 and 243 (Table 2) were not included in the model because the former was present in an N-terminal region of low similarity and the latter in a gap region of the alignment. All of the amino acid substitutions were in solvent-exposed residues (Fig. 2A and B) and were located away from the catalytic residues, namely Cys117 and His287 (Cys29 and His 199 in human procathepsin B), without causing significant variation on the modeled protein structure. The evolutionary-conservation features were evaluated through the multiple alignments of sequences obtained from the ConSSeq module of Sting [35].

Polymorphisms occurring on solvent-exposed residues may lead to changes in antigenic epitopes. The first observed cathepsin B SNP resulted in the substitution of a negatively charged glutamic acid (Glu27) for a positively charged lysine, involving a significant change in charge. These amino acids were found

Table 2
Frequency of SNPs in the cathepsin B gene of LE strain and of field isolates (FI)

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

Frequency: % of worms containing the SNP; NS: nonsynonymous amino acid substitutions; LE: LE laboratory strain; FI: field isolates; *SNPs analyzed by comparative modeling of the cathepsin B protein of *Schistosoma mansoni*.

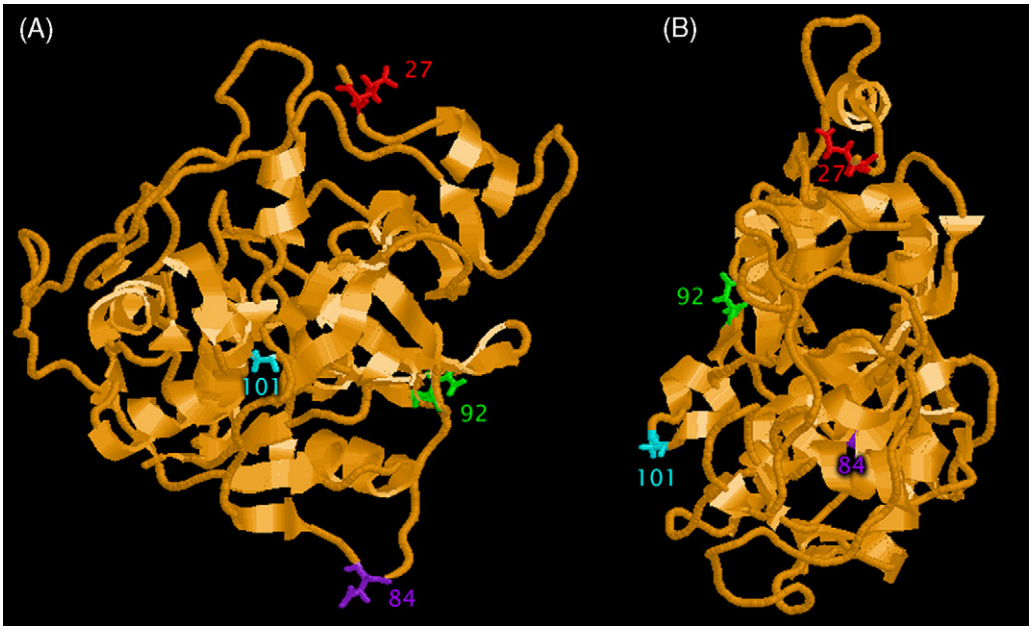

Fig. 2. (A) Ribbon plot of the *S. mansoni* cathepsin B model. The view is along the active site cleft. The N terminus of the model is at the top of the molecule. The side chains of the amino acid substitutions are in different colors and numbered. Noticeably, all substitutions occurred at the surface of the protein structure. (B) Same figure rotated to show the position of amino acid 101 on the surface of the molecule.

to be located in the propeptide stretch of the template human procathepsin B, just before the first helix of the structure, on the surface of the protein. The glutamic acid in the *S. mansoni* cathepsin B possibly interacts with a histidine (His54), and when the replacement for a lysine was introduced, this interaction was lost. The glutamic acid residue was not strictly conserved in the multiple sequence alignment obtained for this protein. However, the amino acids observed in other sequences were a proline in 40% of the sequences aligned and a histidine in 40% of the sequences aligned. Although a lysine residue was not found in any of the sequences, the amino acid lysine has the possibility of making a productive electrostatic interaction with a neighboring aspartic acid (Asp57). The accessibility to the solvent increased slightly from 82 to 104 Å^2, and the volume of the residue remained similar, since glutamic acid and lysine are both large residues. Therefore, the substitution of a glutamic acid for a lysine did not seem to cause a significant conformational change in the structure. In contrast, the change in charge on the surface could alter exposed epitopes and, consequently, antibody recognition. The charge on the surface may also lead to the appearance of previously hidden conformational epitopes [36].

The second SNP observed resulted in the substitution of an aspartic acid (Asp84) for a glutamic acid, both of which are negatively charged amino acids. The glutamic acid was the second most frequent choice of amino acid in this position, as observed from the multiple sequence alignment (19% of the sequences aligned have a glutamic acid in this position). The interaction which occurred between the side-chains of the aspartic acid and a histidine residue in the *S. mansoni* cathepsin B (Asp84 and His82) was broken because of the positioning of the side-chain of the glutamic acid which is also slightly larger than the aspartic acid. However, this interaction could be maintained with a movement of the side-chain of the glutamic acid, since the charges are preserved and the glutamic acid is not making any other interaction. The accessibility to the solvent increased from 119 to 151 Å^2, and the volume of the residue also increased.

The third modeled SNP resulted in the substitution of an asparagine (Asn92) for a serine, a conservative substitution which maintained all of the interactions with the neighboring residues, without causing detectable changes in the accessibility to the solvent or in the electrostatic potential at the surface. Serine was the most frequent choice of amino acid in this position from the multiple sequence alignment, found in 50% of the sequences aligned.

The fourth SNP resulted in the substitution of a glycine (Gly101), a small amino acid, for a positively charged arginine (Fig. 3). A significant modification could be expected from this change. From all the sequences aligned, a glycine residue was found in only 6% of the cases, while an arginine was not found in any sequence. However, the amino acid was positioned at the surface of the molecule, in a turn between two helices (shown at the very bottom of the structure). The only interaction observed was a main-chain interaction with residue Trp99, which was maintained in the presence of both glycine and the arginine residues. There was a large increase in the accessibility to the solvent (62–179 Å^2) due to the large and exposed side-chain of the arginine residue. The electrostatic potential in this region became more negative, showing a possible site for interaction with antibodies. The figures for the other amino acid substitutions analysis can be viewed as supplementary material on line.

In conclusion, we utilized an automated and stringent method for SNP prediction. We were able to clearly demonstrate the occurrence of a large number of putative SNPs in well characterized *S. mansoni* antigens. Several predicted

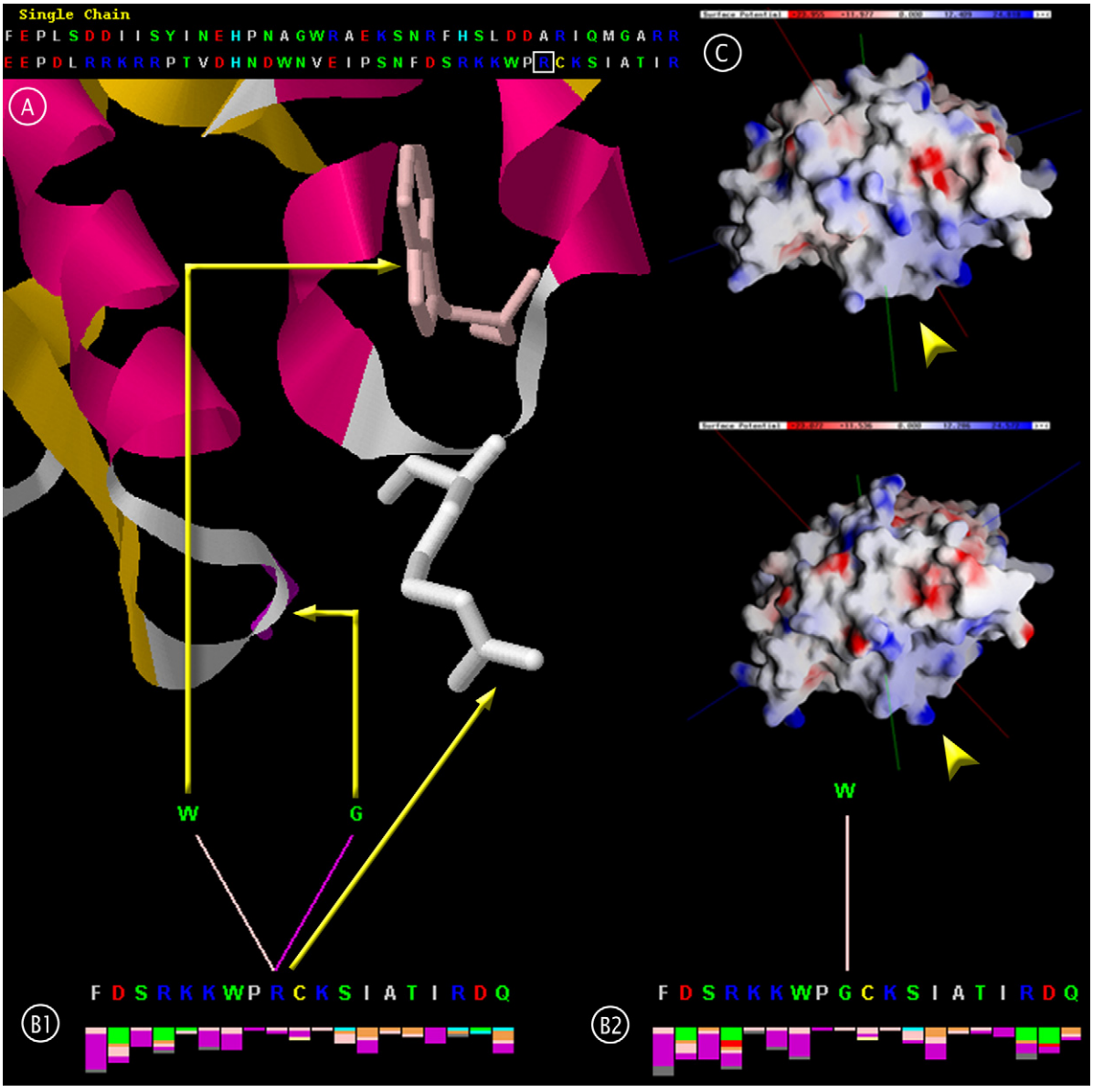

Fig. 3. (A) Schematic representation of the amino acid substitution Gly101Arg located at the bottom of the active site. The arginine residue is shown in white, and the amino acids that interact with arginine are shown in pink (tryptophan W99) and purple (glycine G133). (B1 and B2) Amino acid sequence of the protein with a histogram (below the sequence) containing the number and classes of interactions made by the residues. The lines above the amino acid represent a virtual contact line between residues. For each interaction, a different color is attributed (both in the histogram and in virtual contact lines presentation), according to a color-code. The light-pink line represents a main-chain hydrogen bond, and the dark pink represents a hydrophobic interaction. The color-code of contacts shown for other residues in the chain are: orange: side chain/main chain hydrogen bond; cyan: side chain/side chain hydrogen bond; grey: aromatic stacking; green: charged attractive; red: charged repulsive; yellow: disulphide bridges. Both glycine and arginine in position 101 make a main-chain hydrogen bond with tryptophan W99 (light-pink line). However, when the arginine is in position 101, a new interaction is observed—a hydrophic contact with glycine G133 (dark-pink line). (C) Molecular surface of the protein color-coded by electrostatic potential calculated with GRASP. The molecular surfaces were calculated with the glycine (top) and with the arginine (bottom) in position 101. The surfaces are colored blue for positive, red for negative and white for neutral. The arrowhead indicates the position of the amino acid substitution. When the arginine is in position 101, it is clearly distinguishable as a region of intense positive potential.

SNPs present in the cathepsin B gene were experimentally validated in laboratory and field strains. In the cathepsin B protein model, at least one amino acid substitution may impact antibody binding to the protein. The presence of SNPs on known antigens and other genes of the parasite may be relevant for the development of new vaccines and diagnostic tools.

Acknowledgements

This work was partially funded by NIH-Fogarty training grant 5D43TW007012-03 and FAPEMIG grants 17001/01 and 407/02 to G.O. MS received financial support by funds from a NIH Fogarty Training grants (5D43TW006580-05) and CNPq.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.molbiopara.2007.04.003.

References

[1] Sachidanandam R, Weissman D, Schmidt S, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409:928–33.

[2] Verjovski-Almeida S, Demarco R, Martins E, et al. Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. Nat Genet 2003;35:148–57.

[3] Rodrigues N, Loverde P, Romanha A, Oliveira G. Characterization of new *Schistosoma mansoni* microsatellite loci in sequences obtained from public DNA databases and microsatellite enriched genomic libraries. Mem Inst Oswaldo Cruz 2002;97:71–5.

[4] Somers D, Kirkpatrick R, Moniwa M, Walsh A. Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome 2003;46:431–7.

[5] Nelson M, Marnellos G, Kammerer S, et al. Large-scale validation of single nucleotide polymorphisms in gene regions. Genome Res 2004;14:1664–8.

[6] Guryev V, Berezikov E, Malik R, Plasterk R, Cuppen E. Single nucleotide polymorphisms associated with rat expressed sequences. Genome Res 2004;14:1438–43.

[7] Vennervald B, Dunne D. Morbidity in schistosomiasis: an update. Curr Opin Infect Dis 2004;17(5):439–47.

[8] Criscione DC, Poulin R, Blouin MS. Molecular ecology of parasites: elucidating ecological and microevolutionary processes. Mol Ecol 2005;14(8):2247–57.

[9] Conway DJ, Machado RL, Singh B, et al. Extreme geographical fixation of variation in the *Plasmodium falciparum* gamete surface protein gene Pfs48/45 compared with microsatellite loci. Mol Biochem Parasitol 2001;115(2):145–56.

[10] Ewing B, Hillier L, Wendl M, Green P. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998;8:175–85.

[11] Ewing B, Green P. Base calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998;8:186–94.

[12] Huang X, Madan A. CAP3: a DNA Sequence Assembly Program. Genome Res 1999;9:868–77.

[13] Chirgwin J, Przybyla A, MacDonald R, Rutter W. Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry 1979;18:5294–9.

[14] Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. In: Miseners, Krawetz SA, editors. Bioinformatics methods and protocols: methods in molecular biology. Humana Press; 2000. p. 365–86.

[15] Wishart D, Fortin S. The BioTools Suite. A comprehensive suite of platform-independent bioinformatics tools. Mol Biotechnol 2001;19:59–77.

[16] Sali A, Blundell T. Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.

[17] Podobnik M, Kuhelj R, Turk V, Turk D. Crystal structure of the wild-type human procathepsin B at 2.5. A resolution reveals the native active site of a papain-like cysteine protease zymogen. J Mol Biol 1997;271:774–88.

[18] Neshich G, Rocchia W, Mancini A, et al. Java Protein Dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. Nucleic Acids Res 2004;32:595–601.

[19] Neshich G, Borro L, Higa R, et al. The Diamond STING server. Nucleic Acids Res 2005;33:29–35.

[20] Morlais I, Poncon N, Simard F, Cohuet A, Fontenille D. Intraspecific nucleotide variation in *Anopheles gambiae*: new insights into the biology of malaria vectors. Am J Trop Med Hyg 2004;71:795–802.

[21] Myrick A, Sarr O, Dieng T, Ndir O, Mboup S, Wirth D. Analysis of the genetic diversity of the *Plasmodium falciparum* multidrug resistance gene 5′ upstream region. Am J Trop Med Hyg 2005;72:182–8.

[22] Picoult-Newberg L, Ideker T, Pohl M, et al. Mining SNPs from EST databases. Gen Res 1999;9:167–74.

[23] Cheng T, Xia Q, Qian J, et al. Mining single nucleotide polymorphisms from EST data of silkworm, *Bombyx mori*, inbred strain Dazao. Insect Biochem Mol Biol 2004;34:523–30.

[24] Fryxell K, Moon W. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol 2004;22:650–8.

[25] Fantappie M, Gimba E, Rumjanek F. Lack of DNA methylation in *Schistosoma mansoni*. Exp Parasitol 2001;98:162–6.

[26] Kim H, Schmidt C, Decker K, Emara M. A double-screening method to identify reliable candidate nonsynonymous SNPs from chicken EST data. Animal Genet 2003;34:249–54.

[27] Fitzsimmons C, Savolainen P, Amini B, Hjalm G, Lundeberg J, Andersson L. Detection of sequence polymorphisms in red junglefowl and White Leghorn ESTs. Animal Genet 2004;35:391–6.

[28] Polson H, Conway D, Fandeur T, Mercereau-Puijalon O, Longacre S. Gene polymorphisms of *Plasmodium falciparum* merozoite surface protein 4 and 5. Mol Biochem Parasitol 2005;142:110–5.

[29] Stitziel N, Binkowski T, Tseng Y, Kasif S, Liang J. TopoSNP: a topographic database of nonsynonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 2004;32:520–2.

[30] Sajid M, McKerrow H, Hansell E, et al. Functional expression and characterization of *Schistosoma mansoni* cathepsin B and its trans-activation by an endogenous asparaginyl endopeptidase. Mol Biochem Parasitol 2003;131:65–75.

[31] Correnti J, Brindley P, Pearce E. Long-term suppression of cathepsin B levels by RNA interference retards schistosome growth. Mol Biochem Parasitol 2005;143:209–15.

[32] Loukas A, Bethony J, Williamson A, et al. Vaccination of dogs with a recombinant cysteine protease from the intestine of canine hookworms diminishes the fecundity and growth of worms. J Infect Dis 2004;189:1952–61.

[33] Despres L, Imbert-Establet D, Monnerot M. Molecular characterization of mitochondrial DNA provides evidence for the recent introduction of *Schistosoma mansoni* into America. Mol Biochem Parasitol 1993;60:221–30.

[34] Rodrigues N, Coura-Filho P, de Souza C, Jannoti-Passos L, Dias-Neto E, Romanha A. Populational structure of *Schistosoma mansoni* assessed by DNA microsatellites. Int J Parasitol 2002;32:843–51.

[35] Higa R, Montagner A, Togawa R, et al. ConSSeq: a web-based application for analysis of amino acid conservation based on HSSP database and within context of structure. Bioinformatics 2004;20:1983–5.

[36] Bahia D, Font J, Khaouja A, et al. Antibodies to yeast Sm motif 1 cross-react with human Sm core polypeptides. Eur J Biochem 1999;261:371–7.