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.

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′-

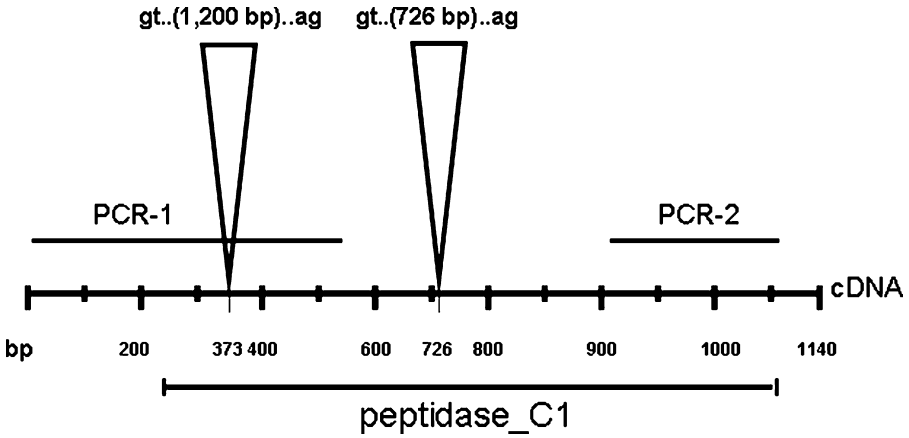

Fig. 1. The cathepsin B gene of *Schistosoma mansoni*. There are two introns (triangles) and three exons (black line). The introns are of 1200 and 726 bp and flanked by gt/ag donor and acceptor sites, respectively. Primers Sm31-X5/X6 were used to amplify a 536 bp cDNA fragment (PCR-1) and primers Sm31X7/X8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