J Membrane Biol (2012) 245:125–129
DOI 10.1007/s00232-012-9420-z

BRIEF COMMUNICATION

Assignment of Putative Functions to Membrane "Hypothetical Proteins" from the *Trypanosoma cruzi* Genome

Ariel Mariano Silber · Claudio Alejandro Pereira

Received: 20 December 2011 / Accepted: 31 January 2012 / Published online: 22 February 2012

Abstract Protozoan parasites cause thousands of deaths each year in developing countries. The genome projects of these parasites opened a new era in the identification of therapeutic targets. However, the putative function could be predicted for fewer than half of the protein-coding genes. In this work, all *Trypanosoma cruzi* proteins containing predicted transmembrane spans were processed through an automated computational routine and further analyzed in order to assign the most probable function. The analysis consisted of dissecting the whole predicted protein in different regions. More than 5,000 sequences were processed, and the predicted biological functions were grouped into 19 categories according to the hits obtained after analysis. One focus of interest, due to the scarce information available on trypanosomatids, is the proteins involved in signal-transduction processes. In the present work, we identified 54 proteins belonging to this group, which were individually analyzed. The results show that by means of a simple pipeline it was possible to attribute probable functions to sequences annotated as coding for "hypothetical proteins." Also, we successfully identified the majority of candidates participating in the signal-transduction pathways in *T. cruzi*.

Keywords *Trypanosoma cruzi* · Hypothetical protein · Membrane protein · Signal transduction · Receptor · Chagas disease

Introduction

Trypanosomes are etiological agents of several veterinary infections, but only two of them cause important human diseases. In sub-Saharan Africa, *Trypanosoma brucei* causes sleeping sickness, and in America *Trypanosoma cruzi* causes Chagas disease. Both trypanosomiases affect mainly poor and marginalized populations. Chagas disease is limited to Central and South America, where about 7.7 million people are infected (Rassi et al. 2010). It is also the first cause of cardiac lesions in young, economically productive adults in endemic countries (Aufderheide et al. 2004).

In 2005, the genomes of the trypanosomatids *T. brucei*, *T. cruzi* and *Leishmania major* were partially completed by the TriTryp sequencing consortium (El-Sayed et al. 2005b). A major problem that occurs in the genome projects in general, and particularly in the TriTryp genomes, is failure to predict gene function; consequently, more than half of the gene products have been annotated as "hypothetical proteins." Genes coding for hypothetical proteins are predominant in trypanosomatid genomes. For example, in the TriTryp database, there are over 105,060 protein-coding genes, with 62,068 codes for hypothetical proteins (59%). In the specific case of *T. cruzi*, there are 11,062 hypothetical proteins (56%) in the genome database, corresponding to 6,526 different genes (Aslett et al. 2010). Considering the present situation, development of novel

Electronic supplementary material The online version of this article (doi:10.1007/s00232-012-9420-z) contains supplementary material, which is available to authorized users.

A. M. Silber
Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, Brazil

C. A. Pereira (✉)
Laboratorio de Biología Molecular de Trypanosoma Cruzi (LBMTC), Instituto de Investigaciones Médicas "Alfredo Lanari", Universidad de Buenos Aires and CONICET, Combatientes de Malvinas 3150, 1427 Buenos Aires, Argentina
e-mail: cpereira@retina.ar

strategies for automated gene-function prediction is especially relevant in protozoan parasites. Putative functions could be assigned to fewer than half of the genes, and these predictions were based on the similarity to previously characterized proteins or known functional domains. For example, many of the predicted metabolic pathways are truncated or incomplete, where the end products would not be used by the parasite (Kanehisa and Goto 2000). Probably, some of these "missing enzymes" exist, but they have been annotated as hypothetical proteins.

Membrane proteins constitute the connecting interface between the intra- and extracellular environments, which mediates the interchange of molecules, medium sensing and cell communication. To date, studies on the membrane components of signal-transduction pathways in protozoan parasites are very scarce. In addition, little is known about the nature of macromolecules with regard to the extracellular environment and their ability to recognize specific environmental signals. The identification of surface proteins able to recognize molecules secreted by the host, which would alter parasite behavior, and the type of response that is expected are critical issues to understand host–parasite interactions (Parsons and Ruben 2000).

In this work, using an automated bioinformatic routine, we determined the distribution of *T. cruzi* membrane proteins according to the transmembrane span number. Based on this initial information, the whole proteins as well as the N- and C-terminal regions were analyzed by BLAST, which allowed prediction of functions for a relevant subpopulation. Novel groups of membrane proteins, including putative receptors, were identified.

Materials and Methods

Databanks Employed

T. cruzi sequence data were obtained from TriTrypDB version 3.3. The Swiss Prot Database, used as a reference for protein alignment, was obtained at the NCBI FTP Databases Repository (ftp://ftp.ncbi.nih.gov/blast/db/). This database, containing 303,518 amino acid sequences, was chosen because it includes only high-quality annotated and nonredundant protein sequences. Other sequences used in this work are from the DNA Database from Japan (DDBJ, http://www.ddbj.nig.ac.jp/) and Interpro Database (http://www.ebi.ac.uk/InterProScan/).

Automated Routines for Sequence Processing

Automated routines were programmed in Perl language (http://www.perl.org/) using Bioperl (http://www.bioperl.org/) code and proceed as indicated in the scheme in Fig. 1a. The routine input was a text report from TriTrypDB for all 5,174 predicted membrane proteins, containing the following data: gene ID, protein length, transmembrane spans (TMS) count, sequence type (gene or pseudogene), amino acid sequence and a table containing the TMS positions. Protein sequences were divided into 37 groups according to TMS number, and then each group was subdivided into three individual fasta files containing the full-length sequence, the N- and C-terminal domains defined as the sequences before the first TMS and after the last TMS. Sequences were compared using the standalone BLAST and the Swissprot protein database. BLASTP version 2.2.24 (Altschul et al. 1997) was run under default parameters using a cut-off score for sequence hits with E values $<10^{-10}$. The final report of the functional groups of hits associated to each query was manually created. Sequences were grouped into 19 categories according to the predicted biological process. Hits related to proteins of particular interest were individually analyzed.

Other Bioinformatic Tools

Further sequence analysis were performed using different resources: TMpred (http://www.ch.embnet.org/), TMHMM version 2.0 (http://www.cbs.dtu.dk/) for topology prediction (Krogh et al. 2001), PredGPI (http://gpcr.biocomp.unibo.it/predgpi/pred.htm) for glycosylphosphatidylinositol PredGPI (http://gpcr.biocomp.unibo.it/predgpi/pred.htm)-anchor sites and Vector NTI 10 package (Invitrogen Corporation, Carlsbad, California) for general sequence analysis.

Results and Discussion

Ordering Membrane Hypothetical Proteins

About 26% of the *T. cruzi* genes code for putative membrane proteins (5,174 out of 19,673 protein-coding genes), of which 48% (2,499) are annotated as hypothetical proteins. Taking as a starting point all the 5,174 genes coding for putative membrane proteins available on TriTrypDB, an automated TMS prediction was performed and the proteins were classified according to the number of TMS, ranging from 1 to 37.

About 74% of the predicted proteins have one or two TMS, belonging to few multigene families of surface proteins including *trans*-sialidases, gp63, Tc85, mucin-associated surface proteins (MASP) and related proteins (El-Sayed et al. 2005a). In a preliminary analysis we found that many sequences within this group, annotated as "hypothetical proteins," had high-identity values with known soluble proteins. Therefore, proteins with one or two TMS were excluded from further analyses because the

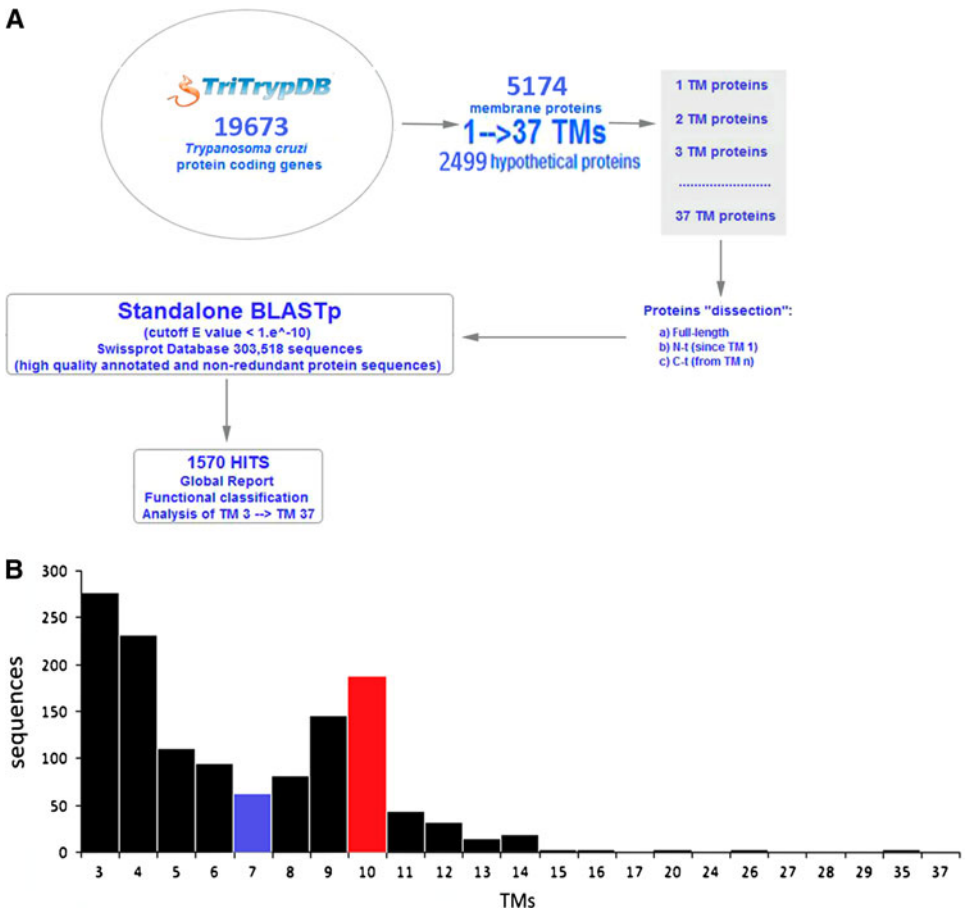

Fig. 1 **a** Schematic representation of the automated routine for functional classification of proteins. Input files are text reports obtained from TriTrypDB (http://tritrypdb.org/) containing the following data: gene ID, protein length, transmembrane span (TMS) count, sequence type (gene or pseudogene), predicted amino acid sequence and a table containing the TMS positions. Protein sequences have been divided into 37 groups according the TMS number, and then each group was subdivided in individual fasta files containing the full-length sequence and the N- and C-terminal domains, defined as the sequences before the first TMS and after the last TMS. Sequences were compared using the standalone BLAST and the Swissprot protein database. BLASTP version 2.2.24 was run under default parameters using a cut-off score for sequence hits of $E < 10^{-10}$. **b** Protein classification according to TMS count. Proteins containing 3–37 TMS were graphically represented as a function of sequence number

data obtained from these sequences were not reliable, an error in the prediction of one TMS leading to the inclusion of soluble proteins in this group (see Supplementary Table S1).

The distribution of membrane proteins is shown in Fig. 1b. Interestingly, the number of proteins in each group as a function of the number of TMS presents a regular pattern, which continuously decreases up to seven TMS, presenting a second "peak" at 10 TMS (Fig. 1b, blue and red bars, respectively). This pattern is probably related to protein function because loops are variable regions submitted to the selective pressure of the immune system of the hosts, but the TMS present a more restrictive degree of freedom in terms of variation during evolution due to the fact that they constitute the structural basis for protein function.

Predicting Functions of Membrane Proteins

The second part of the algorithm designed in this work is oriented to function prediction based on the subdivision of proteins in their N- and C-terminal domains for local sequence alignment using BLAST. All 1,570 proteins with BLAST hits under the cut-off score ($E < 10^{-10}$) were classified in 18 different groups according to their functions (Fig. 2a, b). Of the total analyzed, 918 proteins were above the cut-off E value and excluded, to maintain the stringency of the analysis. The largest protein group identified comprises different transporter proteins, including ion channels, amino acid permeases, ABC and MDR pumps (constituting about 50% of the membrane proteins analyzed) and enzymes related to lipid metabolism (18.8%). Amino acids and related transporters are those

A

Classification/TMs number	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	20	26	27	29	35	37	TOTAL	%
Lipid metabolism	8	22	10	5	8	6	3	4	2								2	1				71	18.8
Unclassified transporters	11	12	10	3	3	1	6	8	4	3												61	16.1
Ion transporters	2	8	1	2	2	5	2	2	8	1				1		1						35	9.3
ABC/MDR transporters	2	4	4	6	1		1	3	4	4	1	1	1	1								33	8.7
Surface proteins	24	3	1		1																	29	7.7
Unexpected hits	8	4	2	1	3	1		1	1	1					1							23	6.1
Amino Acid transporters		1			1	1		8	7	3												21	5.6
Unclassified non-transporter proteins	5	4	9	2																		20	5.3
Zinc transporters		1	1	2	2	3		1				1	1									12	3.2
Putative receptors	1	1	2		3		1		1	2	1											12	3.2
Carbohydrate transporters		2	1			2	3	1	1	2												12	3.2
F/P-type ATPases	2		4	3	1																	10	2.6
Lipid transporters			1		3	2				3												9	2.4
Carbohydrate metabolism	1	3		1			2	1														8	2.1
Protein kinases	2	1	2			1	1															7	1.9
Respiratory chain	2	1	1			1				1												6	1.6
Aquaporines				5																		5	1.3
Structural proteins																			1	2	1	4	1.1
TOTAL	68	67	49	30	28	23	19	29	28	20	2	2	2	2	1	1	2	1	1	2	1	378	100.0

B

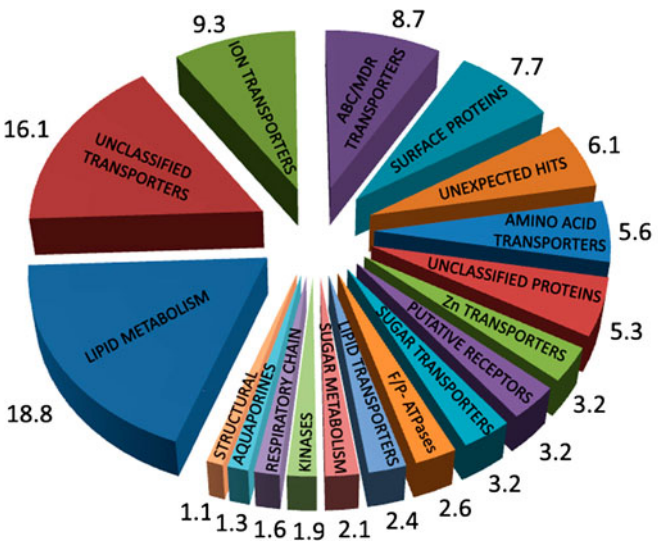

Fig. 2 Functional classification of membrane proteins. **a** Predicted protein sequences ordered according to TMS number and classified in 19 functional groups based on the corresponding BLAST hits report. *Gray rows* indicate groups analyzed in proof in 3.3. **b** Graphical representation of each functional group. A pie chart was constructed using the data shown in **a**. Numbers next to each triangle are the percentage of sequences corresponding to each functional group

sequences that constitute the "peak" at 9 or 10 TMS in Fig. 1. A special case are zinc transporters, which are interesting not for their function but for the large number of them found in the sequence analysis—about 25% of the protein involved in ion uptake in *T. cruzi* are putative zinc transporters. However, the physiological role of zinc in the parasites remains unknown. A group named "unexpected hits" contained proteins not previously found in this organism or unusual membrane proteins. For example, proteins related to nucleic acid remodeling, including those involved in DNA excision/repair process, DEAD-box ATP-dependent RNA helicases and endo- and exonucleases. These putative proteins are probably associated with the nuclear membrane.

Proteins Related to Signal-Transduction Processes

We specifically analyzed all the protein hits related to sensing or transducing extracellular signals. A total of 55 proteins were selected. The most interesting subgroups found are proteins involved in processes such as autophagy, programmed cell death, inositol 1,4,5-trisphosphate (IP3) receptors and components of the mitogen-activated protein kinase (MAPK) cascade (Supplementary Table S2). Finally, one group, belonging to the seven transmembrane receptors family (7TMR), is composed of adiponectin receptors. Adiponectin is an essential hormone secreted by adipocytes, which acts as an antidiabetic factor and is involved in metabolic pathways that regulate lipid metabolism such as fatty acid oxidation and glucose uptake throughout the AMP-activated protein kinase (AMPK) (Goldstein and Scalia 2004). Three different sequences belonging to the PAQR (progestin and adipoQ receptors) family were identified as putative adiponectin receptors. Considering that adipose tissue is one of the major sites of inflammation in Chagas disease and that a *T. cruzi* infection–associated decrease in adiponectin has been demonstrated (Nagajyothi et al. 2008), future studies on these

putative receptors can reveal new host–parasite interaction mechanisms.

Acknowledgements This work was supported by grants from the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET, PIP grant 0685 to C. A. P.), Agencia Nacional de Promoción Científica y Tecnológica (FONCYT PICT grant 2008-1209 to C. A. P.), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP grant 11/50631-1 to A. M. S.) and Instituto Nacional de Biologia Estrutural e Química Medicinal em Doenças Infecciosas (INBEQMeDI). C. A. P. is a career scientific investigator of CONICET (Argentina). The funders had no role in study design, data collection and analysis, the decision to publish or preparation of the manuscript.

References

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

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, Gardner MJ, Gingle A, Grant G, Harb OS, Heiges M, Hertz-Fowler C, Houston R, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Logan FJ, Miller JA, Mitra S, Myler PJ, Nayak V, Pennington C, Phan I, Pinney DF, Ramasamy G, Rogers MB, Roos DS, Ross C, Sivam D, Smith DF, Srinivasamoorthy G, Stoeckert CJ Jr, Subramanian S, Thibodeau R, Tivey A, Treatman C, Velarde G, Wang H (2010) TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res 38:D457–D462

Aufderheide AC, Salo W, Madden M, Streitz J, Buikstra J, Guhl F, Arriaza B, Renier C, Wittmers LE Jr, Fornaciari G, Allison M (2004) A 9,000-year record of Chagas' disease. Proc Natl Acad Sci USA 101:2034–2039

El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, Englund PT, Fazelina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler A, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Van Aken S, Vogt C, Ward PN, Wickstead B, Wortman J, White O, Fraser CM, Stuart KD, Andersson B (2005a) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 309:409–415

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran AN, Wortman JR, Alsmark UC, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N (2005b) Comparative genomics of trypanosomatid parasitic protozoa. Science 309:404–409

Goldstein BJ, Scalia R (2004) Adiponectin: a novel adipokine linking adipocytes and vascular function. J Clin Endocrinol Metab 89:2563–2568

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Nagajyothi F, Desruisseaux MS, Thiruvur N, Weiss LM, Braunstein VL, Albanese C, Teixeira MM, de Almeida CJ, Lisanti MP, Scherer PE, Tanowitz HB (2008) *Trypanosoma cruzi* infection of cultured adipocytes results in an inflammatory phenotype. Obesity (Silver Spring) 16:1992–1997

Parsons M, Ruben L (2000) Pathways involved in environmental sensing in trypanosomatids. Parasitol Today 16:56–62

Rassi A Jr, Rassi A, Marin-Neto JA (2010) Chagas disease. Lancet 375:1388–1402