FUNCTIONAL PROTEOMICS AND BIOINFORMATICS TO MONITOR GENE EXPRESSION, IDENTIFY VACCINE TARGETS AND DISCOVER NOVEL BIOCHEMICAL PATHWAYS IN TRYPANOSOMA CRUZI

by

JAMES ALEXANDER ATWOOD III

(Under the Direction of Ron Orlando)

ABSTRACT

Trypanosoma cruzi (*T. cruzi*) is a flagellated protozoan parasite endemic to much of Latin America and the causative agent of Chagas' disease. The functional annotation of the *T. cruzi* genome *in vivo* is best facilitated by measuring protein expression through proteomic analysis, which not only offers an understanding of stage specific protein expression, but also post-translational modifications. Presented here are proteomic and bioinformatic approaches, all of which center on the high-throughput analysis of *T. cruzi* gene expression.

Herein, we describe a heuristic method for organizing proteins identified at a specified false discovery rate using Mascot matched peptides. We call this method PROVALT. It was evaluated using Mascot identified peptides from a *Trypanosoma cruzi* epimastigote whole-cell lysate, which were separated by multidimensional liquid chromatography and analyzed by tandem mass spectrometry. PROVALT was shown to be superior to both the single peptide score and cumulative protein score methods.

The whole-organism, proteomic analysis of the four life-cycle stages of *T. cruzi* was

performed. Peptides mapping to 2784 proteins in 1168 protein groups from the annotated *T. cruzi* genome were identified across the four life-cycle stages. Evidence is presented for the expression of protein products from >1000 genes currently annotated as "hypothetical", including members of a gene family annotated as mucin-associated surface proteins (MASPs). Furthermore, this analysis revealed the apparent utilization of distinct energy sources by the individual parasite stages, including histidine for stages present in the insect vectors and fatty acids by intracellular amastigotes.

A method for the high-throughput identification of membrane associated N-linked glycoproteins from *Trypanosoma cruzi* is described. This analysis was based on enrichment of trypomastigote membrane proteins followed by capture of membrane associated N-linked glycoproteins by Concanavalin A affinity chromatography. Through stable isotope labeling of the glycan attachment sites with ^{18}O we unambiguously identified 19 glycopeptides which mapped to 17 glycoproteins, all of which were membrane associated. We also present the first evidence for the expression of 7 putative trypomastigote cell surface glycoproteins including GP-90, DGF-1, and a novel cysteine protease SCP1. We also discuss the implications of two ER localized glycoproteins identified in this analysis, STT3 and GRP94.

INDEX WORDS: False Discovery Rate, False Positive Rate, Glycomics, Glycoproteomics, Multidimensional Chromatography, N-glycosylation, Protein Expression Profiling, Proteomics, PROVALT, Trypanosoma cruzi

FUNCTIONAL PROTEOMICS AND BIOINFORMATICS TO MONITOR GENE EXPRESSION, IDENTIFY VACCINE TARGETS AND DISCOVER NOVEL BIOCHEMICAL PATHWAYS IN TRYPANOSOMA CRUZI

by

JAMES ALEXANDER ATWOOD III

B.S., The University of Georgia, 2001

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2005

FUNCTIONAL PROTEOMICS AND BIOINFORMATICS TO MONITOR GENE EXPRESSION, IDENTIFY VACCINE TARGETS AND DISCOVER NOVEL BIOCHEMICAL PATHWAYS IN TRYPANOSOMA CRUZI

by

JAMES ALEXANDER ATWOOD III

Major Professor: Ron Orlando

Committee: Jonathan Amster
James Anderson

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2005

DEDICATION

It is often said that the behind every successful man is a surprised woman. Thus I dedicate this dissertation to the three women who have most influenced me intellectually and creatively.

To My Grandmother:

Louise Knight Atwood.....for your quiet strength and devotion to the arts. The summers spent with you painting "happy little trees" were some of the happiest moments of my life. Thank you for teaching me to find the beauty in all things.

To My Mother:

Joan Lee Atwood.....for your generosity and unconditional love throughout my life. You sacrificed to show a young boy the world and I'll never forget it! You were right mom, "do what you love".

To My Wife:

Meredith Nesbitt Atwood.....for your compassion and steadfast support. Little did I know that saying "hello" would change my life so dramatically. From the weightlifter to the doctor you have and always will be my greatest inspiration.

ACKNOWLEDGEMENTS

I would like to extend my gratitude and appreciation to all the scientist and staff of the Complex Carbohydrate Research Center and the Center for Tropical and Emerging Global Diseases. My graduate experience has been enhanced through collaborations and intellectual discussions with numerous individuals from both centers and for this I thank you all. I am especially grateful to my mentors Ron Orlando and Rick Tarleton. You both taught me that science is not about the individual. Rather great scientific advances come from the combined efforts of many. I would also like to thank Carl Bergmann for his support and advice throughout the past 6 years. You are a champion for great science and a magnificent teacher. My deepest gratitude is also extended to a number of individuals who I have had the privilege of working with during my graduate studies; Kumar Kolli, Jeremi Johnson, Brent Weatherly, Amr Ghaleb, Gerardo Gutierrez-Sanchez, Gabre Kemp, Bryan Woosley, Cameron Cavola, and Jason Baker. Each of you has contributed significantly to the completion of this dissertation but more importantly throughout this experience you have all become wonderful friends. Lastly I would like to thank my entire family. I love you all.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

With the completion of the *Trypanosoma cruzi* (*T. cruzi*) genome a new and more challenging analytical project has arisen, the analysis of the proteome or proteomics (1). Proteomics, put simply, is the ability to systematically identify every protein expressed in a cell or tissue as well as to determine the salient properties of each protein (2). While the *T. cruzi* genome is relatively static, its proteome is extremely dynamic with protein expression differing based on developmental stage, environmental factors, and cell history. Although advances in genomic sequencing and mRNA-based analyses of gene expression have provided insight into the gene-phenotype relationship, an exact correlation between gene and protein expression has not been confirmed (3-6). Therefore, analyzing gene expression at the protein level is fundamentally important both for elucidation of gene regulation and for the development of therapeutics to combat Chagas' disease.

In chapter 3 we discuss the implementation of a high-throughput proteomics and bioinformatics platform for the identification of proteins from *T. cruzi* whole cell lysates. Based on the work by Yates and colleagues (6-10) a three dimensional liquid chromatographic separation was employed to resolve peptides prior to analysis by tandem mass spectrometry (MS/MS). The separation scheme was evaluated using proteins derived from *T. cruzi* epimastigote whole cell lysates. Both the methods for peptide separation and peptide identification by MS/MS are discussed in detail. An important step in the majority of "shotgun" proteomics experiments is the correlation of MS/MS spectra with peptide sequences derived from protein databases. As noted by Nesvizhskii (11, 12) this is not a straightforward process when analyzing proteins at the peptide level. With this in mind, chapter 3 also includes a discussion concerning the difficulties associated with computationally mapping MS/MS spectra

with protein sequences and proposes a new approach to statistically validate these protein identifications using software we call "PROVALT".

In chapter 4 we describe the high-throughput proteomic analysis of the four *T. cruzi* developmental stages. The proteomic and bioinformatic techniques described in chapter 3 were employed to identify proteins from the epimastigote, amastigote, trypomastigote, and metacyclic trypomastigote forms. This chapter focuses predominately on the major biological findings associated with the observed protein expression patterns between the four *T. cruzi* developmental stages. The chapter also contains a discussion of identified protein post-translational modifications, most notably on elongation factor 1-alpha. However, as noted in this chapter only post-translational modifications which occur with static mass shifts can be readily detected through this method. Protein modifications such as glycosylation, which are inherently important for *T. cruzi* development, are not identifiable via this proteomic approach (13, 14). Thus chapter 5 describes a method to identify N-linked glycoproteins from *T. cruzi* trypomastigotes. Both the methodology associated with this analysis and the biological implications of the glycoprotein identifications are discussed.

REFERENCES

1. Andersson et al. 2005. Science, In press.
2. Gygi, S., Aebersold, R., 2000. Curr. Opin. Chem. Biol. 4, 489-494.
3. Saborio, J., Hernandez, J., Narayanswami, S., Wrightsman, R., Palmer, E., Manning, J., 1989. J. Biol. Chem. 264, 4071-4075.
4. Minning, T., Bua, J., Garcia, G., McGraw, R., Tarleton, R., 2003. Mol. Biochem. Parasitol. 131, 55-64.
5. Diehl, S., Diehl, F., El-Sayed, N., Clayton, C., Hoheisal, J., 2002. Mol. Biochem. Parasitol. 123, 115-123.
6. Washburn, M., Wolters, D., Yates, J., 2001. Nat. Biotechnol. 19, 242-247.
7. Liu, H., Lin, D., Yates, J., 2002. Biotechniques 32, 898-902.
8. Link, A., Eng, J., Schieltz, D., Carmack, E., Mize, G., Morris, D., Garvik, B., Yates, J., 1999. Nat. Biotechnol. 17, 676-82.
9. Wolters, D., Washburn, M., Yates, J., 2001. Anal. Chem. 73, 5683-5690.
10. McDonald, W., Yates, J., 2002. Dis. Markers. 18, 99-105.
11. Nesvizhskii, A., Keller, A., Kolker, E., Aebersold, R. 2003. Anal. Chem. 75, 4646-4658.
12. Nesvizhskii, A., Aebersold, R. 2004. Drug. Discov. Today. 4, 173-181.
13. Scharfstein, J., Schmitz, V., Morandi, M., Capella, A., Lima, C., Morrot, A., Juliano, L., Muller-Esterl, W., 2000. J. Exp. Med., 192, 1289-1299.
14. Frasch, A., 2000. Parasitology Today. 16, 282-286.

CHAPTER 2

LITERATURE REVIEW

Trypanosoma cruzi: the causative agent of Chagas' disease

Chagas' disease currently afflicts ~18 million people and results in the loss of over US $1.2 billion/year in productivity (1,2). To date no vaccines are available for Chagas' disease and treatments have been limited to chemotherapeutics which are highly toxic and ineffective during the chronic stage of the disease (3). Furthermore, more than 100 million people are exposed to the causative agent of Chagas' disease, the protozoan parasite *Trypanosoma cruzi* (1). *Trypanosoma cruzi* (*T. cruzi*) is a protozoan parasite endemic to much of Latin America and a member of the kinetoplastid family (other members are *T. brucei* and *Leishmania*) (4). The life cycle of *T. cruzi* is complex, with multiple developmental stages persisting between a variety of mammalian host (including humans) and insect vectors. Metacyclic trypomastigotes in the insect hindgut initiate infection into the mammalian host via fecal contamination of mucus membranes or wound openings. The metacyclic trypomastigotes enter various cells and differentiate into aflagellated amastigotes which replicate in the host cell cytoplasm. These forms then give rise to flagellated trypomastigotes which are released into the blood stream of the mammalian host and either invade new cells or are ingested by the insect vector during the course of a blood meal. In the insect vector, non-replicative trypomastigotes migrate to the midgut and differentiate into replicative epimastigotes. Following multiple rounds of binary fission, epimastigotes transform into infective metacyclic trypomastigotes and migrate to the hindgut where they infect the mammalian host thus completing the life cycle. Throughout this developmental cycle, the parasite undergoes profound morphological changes. The parasite's static genome must encode several different proteomes that are responsible for maintaining stage specific functions such as host invasion and intracellular replication. Many of these stage regulated genes have been targeted with some success in vaccine and inhibitor studies (3,5,6). However, the number of

therapeutic targets analyzed to date has been limited by a lack of gene expression data *in vivo*. Thus methods which enable the functional annotation of the *T. cruzi* genome *in vivo* are a necessary prerequisite for vaccine and inhibitor discovery.

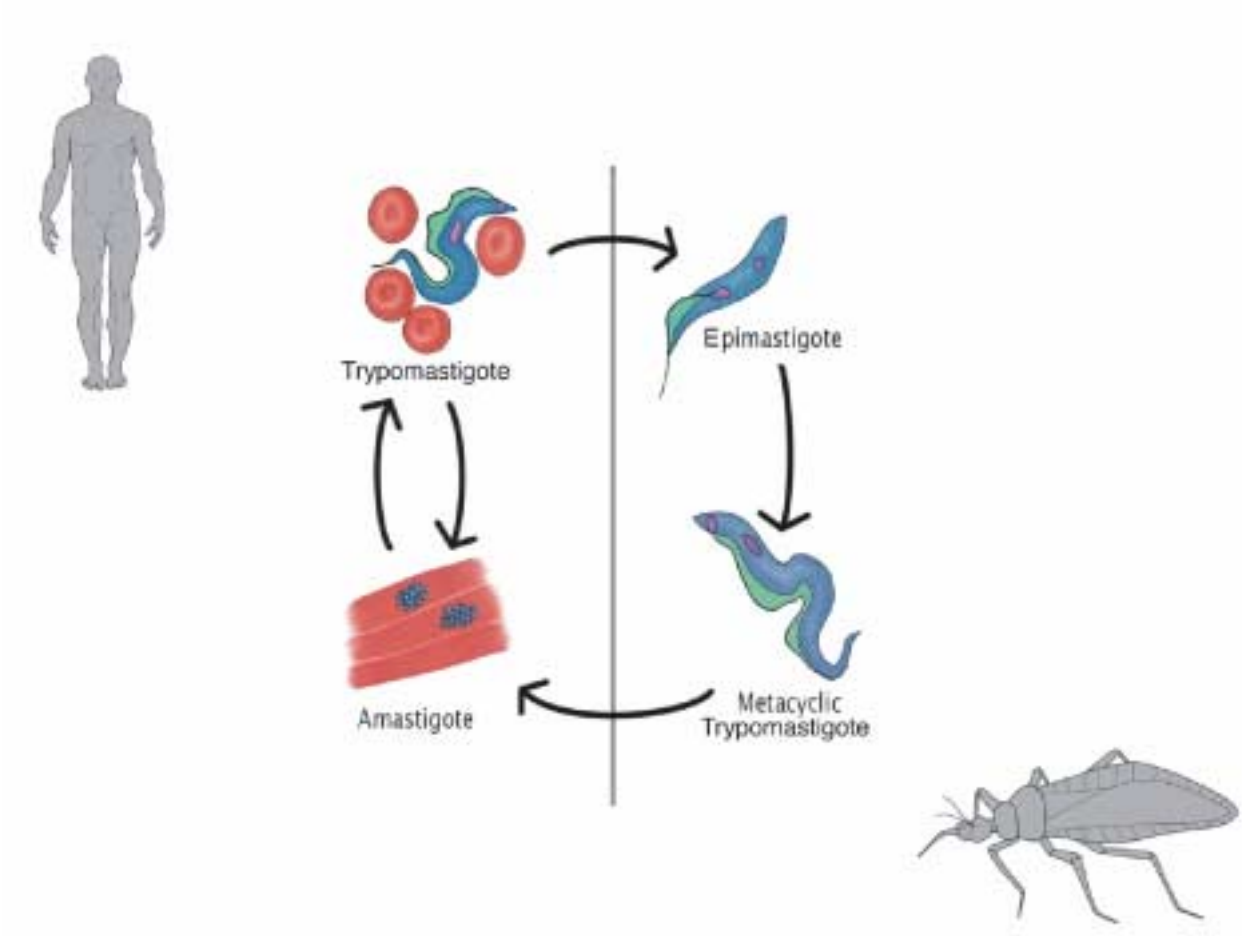

Figure 2.1 - Lifecycle of Trypanosoma cruzi

Protein expression profiling in kinetoplastids - 2D-PAGE

Historically, proteomics of kinetoplastids has involved the classic approach of two dimensional gel electrophoresis (2D-PAGE) as the chosen method for separation of complex protein mixtures (Fig 2.2) (8-15). Typically, proteins are extracted *in toto* from the parasite whole cell lysate and solubilized by the addition of various detergents. One advantage of 2D-PAGE is its compatibility with detergents and being that a large number of kinetoplastid proteins are membrane bound this approach is attractive. However, proteins which contain multiple transmembrane spanning regions and hence are very hydrophobic are often difficult to separate in the first dimension isoelectric focusing step due to poor solubility and extreme pI (16). In the first dimension isoelectric focusing, proteins are separated based on their relative charge in an electric field, migrating until their net charge is zero, at their isoelectric point. The proteins are then separated in the second dimension by size. Following separation the proteins are digested in situ and the peptides are extracted from the gel via dehydration with an organic solvent. Protein identification is facilitated through mass spectrometry (MS) and database searching (for review see 17). For gel based samples from kinetoplastids, two methods of mass spectrometric based protein identification have routinely been employed. The first involves the analysis of peptides by matrix assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF)(18). The experimental peptide m/z (mass/charge) values that were measured during the MS experiment are then compared to theoretical peptide masses derived from *in silico* enzymatic digestion of the specified protein database (peptide mass fingerprinting)(18). Protein identification is then determined based on the degree of overlap between the experimental and theoretical peptide masses.

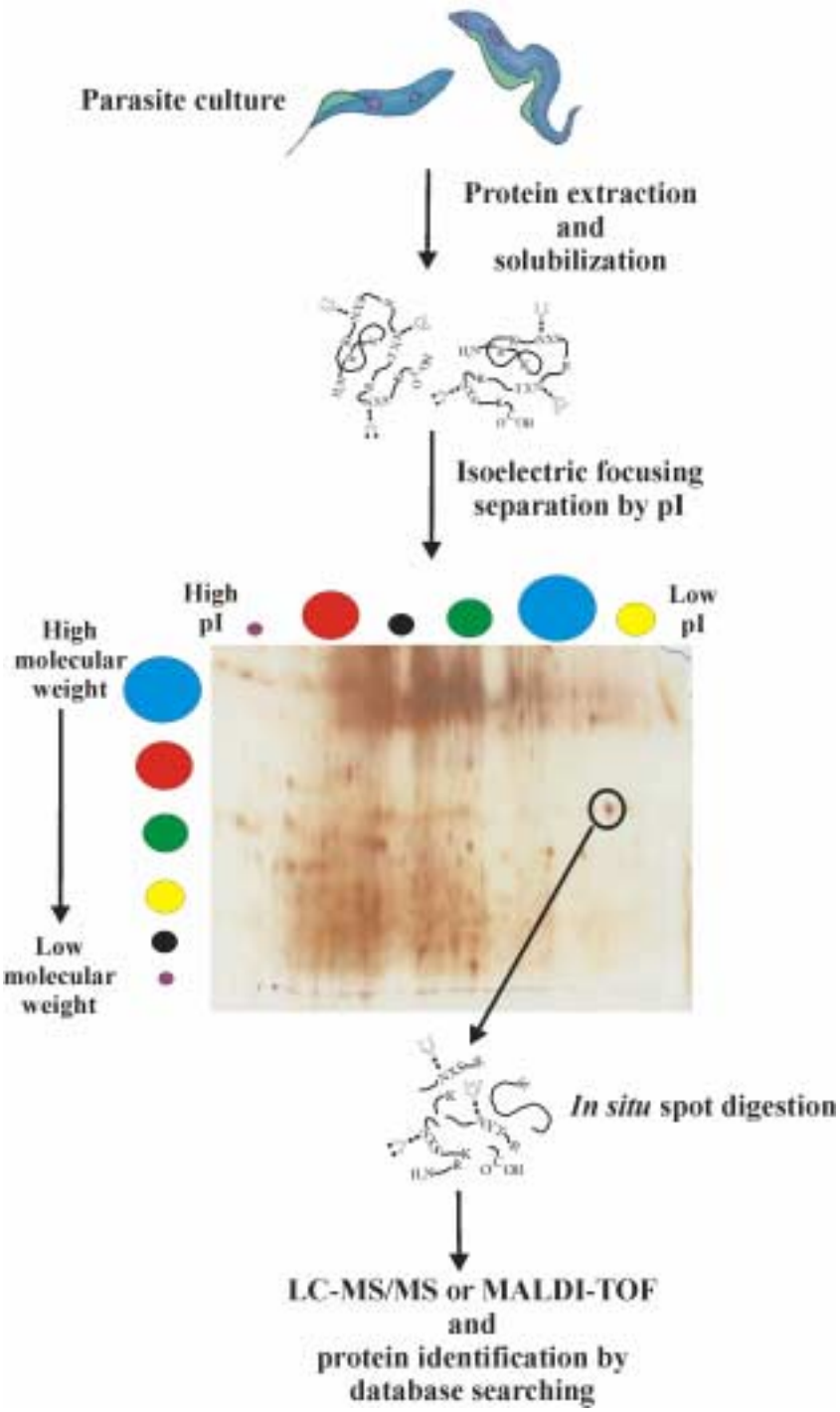

Figure 2.2 - Standard proteomic analysis of a complex protein mixture by 2D-PAGE

As an established technique to simultaneously separate and quantitate the protein profiles of cells and tissues, the 2D-PAGE method has been critical for describing the complexity of protein variation expressed by these protozoan parasites. Furthermore 2D-PAGE has allowed the rapid comparison of kinetoplastid gene regulation between developmental stages. In one study *Leishmania donovani* promastigotes (insect form) were compared with the mammalian intracellular amastigote stage (12). Like *T. cruzi*, *Leishmania* gene regulation between developmental stages has been poorly understood and in this analysis the authors were able to visualize as many as 2000 kinetoplastid proteins on a single silver stained 2-D gel. Furthermore, the extraction of peptides through in-gel digestion and the application of MALDI-TOF analysis followed by peptide mass fingerprinting allowed the identification of 41 different proteins. Surprisingly only 31 proteins were determined to be stage specific indicting a high degree of conserved gene expression between the two forms. The study also revealed the stage regulated expression of proteins which function in energy and amino acid metabolism, cytoskeletal assembly, and cell cycle control. Many of the findings were commensurate with the known gene regulation by *T. cruzi*, including the specific expression of paraflagellar rod components (19) in the promastigote forms representing the loss of the flagellum during conversion of *T. cruzi* trypomastigote to amastigote forms (20).

Another recent 2D-PAGE study analyzed the protein profiles of *T. cruzi* trypomastigote, epimastigote, and amastigote developmental forms (14, 15). It is expected that each developmental stage contains a unique subset of proteins responsible for the stage specific biological functions of the parasite. While a number of articles have used DNA microarrays successfully to study differential gene regulation for *T. cruzi* and related kinetoplastids, an exact correlation between gene and protein expression has not been confirmed in these organisms (20-

23). Furthermore, *T. cruzi* appears to regulate protein expression primarily post-transcriptionally via variations in mRNA stability and/or the translational efficiency of mRNAs (24). This limits the utility of gene expression profiling for monitoring stage-dependent changes in gene expression and makes proteomic analysis especially attractive for examining global changes in protein expression during development in *T. cruzi*. In the analysis by Paba *et al*, approximately 500 proteins were detected from 5×10^7 parasites in each developmental stage. Being that the *T. cruzi* genome is predicted to encode ~25,000 genes, these data indicted that the parasite expresses a large number of proteins at lower abundances than are readily detectable through 2D-PAGE analysis (25). The limited number of visualized proteins is due to the fact that *T. cruzi* expresses a number of very highly abundant proteins which include the tubulin and heat shock protein families. Thus a significant percentage of the total protein concentration is due to these high abundance proteins, therefore the detection of proteins in lower abundances is hindered due to the dynamic range of protein concentrations *in vivo*. Nevertheless, in similar fashion to the *Leishmania* proteome, *T. cruzi* was shown to express on average 30 of the 500 visualized proteins specifically in one developmental stage. Like *Leishmania,* conversion of flagellated trypomastigotes to aflagellate amastigotes occurs with a decreased expression of paraflagellar rod components. In addition, differences in the protein expression patterns between the two developmental stages was attributed to proteins responsible for cytoskeletal remodeling and metabolism. The most interesting finding was the increased expression in the mammalian stages of three proteins involved in the glycolysis, 2,3 bisphosphoglycerate mutase, enolase 2, and pyruvate kinase (15). Homologous energy sources are also utilized in the blood stream forms of *T. brucei* (26). However, in this analysis only 19 proteins were identified and all of the identifications resulted from proteins which were previously known to be expressed in high copy

number with *T. cruzi* cells (27, 28). While the number of kinetoplastid proteins which have been visualized by 2D-PAGE is substantial, the high-throughput identification of proteins through this approach is not practical because the proteins must by tediously digested and analyzed by mass spectrometry individually. Consequently, the largest number of kinetoplastid proteins identified by a single 2D-PAGE proteomic study remains at 49 (12). The relative insufficiency of identifications resulting from 2D-PAGE initiatives are a consequence of the intrinsic limitations of the technique. Complex samples like those produced from whole cell lysates, frequently exceed the resolution of 2D-PAGE and detection of low abundance proteins is often limited by a lack of dynamic range (10). As an alternative to 2D-PAGE, new technologies have emerged which have allowed the high throughput separation and identification of proteins without many of the drawbacks associated with gel electrophoresis (29-33).

Protein expression profiling in kinetoplastids - Multidimensional LC-MS/MS

Multidimensional liquid chromatography coupled with tandem mass spectrometry (2DLC-MS/MS) has become increasingly popular as a technique for separation and analysis of complex peptide mixtures (34-38). The effectiveness of this technique is attributed to the resolving power of two orthogonal types of liquid chromatography prior to peptide analysis by mass spectrometry. Typically, 2DLC-MS/MS experiments are subdivided into five stages (Fig. 2.3). First, the proteins are extracted from the biological material followed by enzymatic digestions to produce peptides. This results in a very complicated mixture of peptides. For example, if one assumes that there are 2,000 proteins present (as seen in the *Leishmania* 2D-PAGE proteome (12)) and that each of them leads to 10 peptides upon proteolysis, there will be 20,000 peptides, which is outside the complexity that can be handled by a single LC separation.

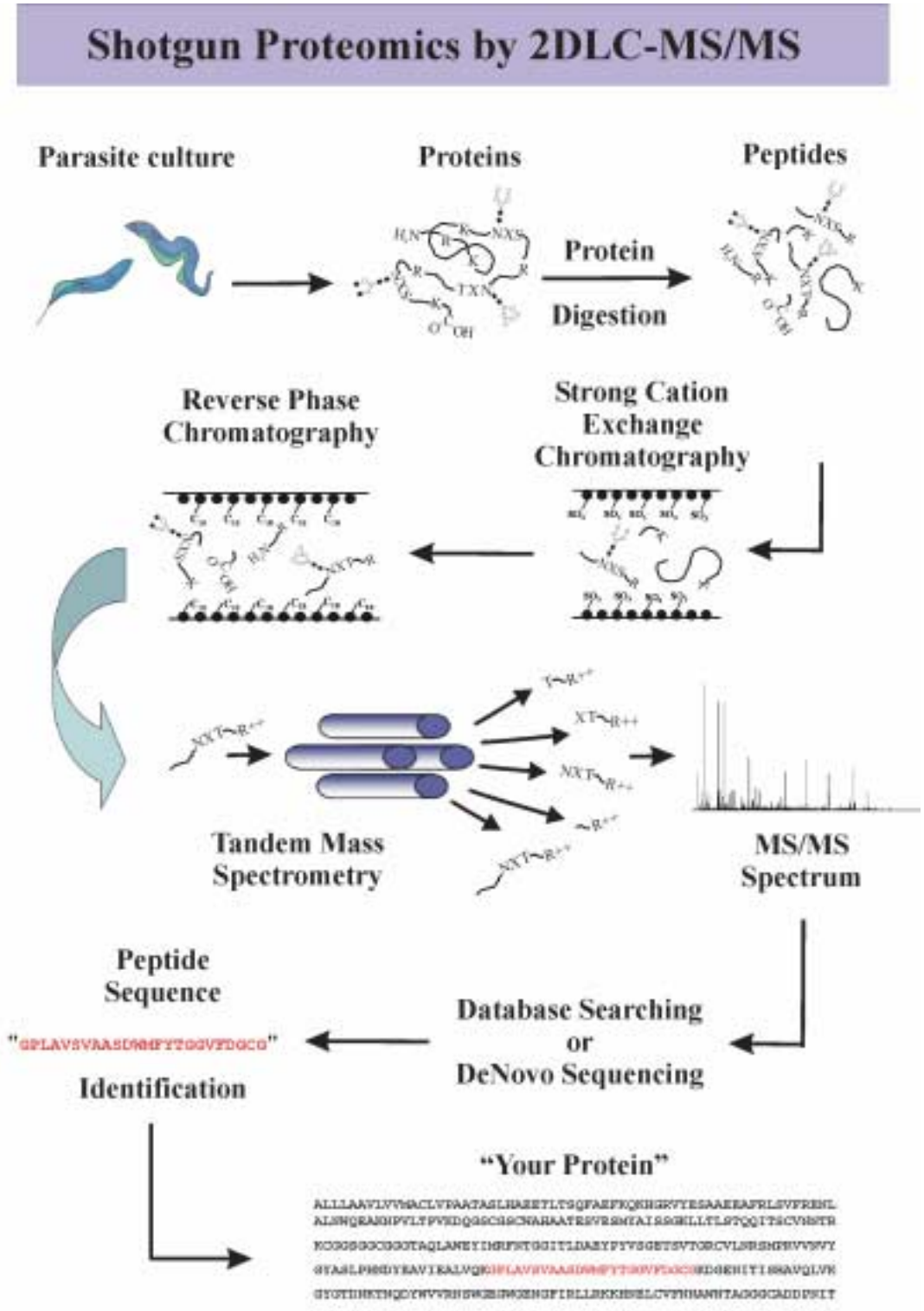

Figure 2.3 - The major steps in shotgun proteomics through 2DLC-MS/MS.

This degree of complexity dictates that multidimensional chromatographic separations are used prior to tandem mass spectrometry, since a single dimension simply does not have the peak capacity to resolve all of these components (39, 40). Thus 2DLC provides a significantly higher peak capacity, since the overall peak capacity is the product of the peak capacities of each orthogonal chromatographic separation performed (41). While many types of 2DLC have been employed for the separation of peptides and proteins prior to MS/MS analysis the most common to date is the coupling of strong cation exchange (SCX) and reverse phase (RP) (Fig 2.3) (34-38, 42-45). In this approach peptides are separated in the first dimension by strong cation exchange chromatography based on their relative charge in an acidic buffer. When peptides are placed into an acidic buffer (pH 3.0), the basic residues H, R, and K become positively charged, bind to the negatively charged column, and are eluted with an increasing salt concentration (46). Peptides are then separated in the second dimension based on their relative hydrophobicity through reverse phase chromatography. Generally this step is performed online so that the peptides elute directly into the mass spectrometer. The peptides are then analyzed by tandem mass spectrometry in which both the intact peptide mass and peptide fragment ion masses are sequentially analyzed (for review see 47, 48). The resulting MS/MS spectrum is then either sequenced de novo, yielding a complete amino acid sequence or sequence tag which is then searched against a protein database, or software algorithms are used to directly match experimental with theoretical MS/MS spectra generated from peptides in a protein database (for review see 49).

2DLC-MS/MS has also previously proven effective in the proteome analysis of *T. cruzi* (45). In this study, protein expression between the two mammalian stages (amastigote and trypomastigote) was quantified by isotope coded affinity tag technology (50). In this procedure

the cysteine containing peptides are differentially labeled, extracted from the peptide mixture via avidin chromatography, then identified and quantified by LC-MS/MS (50). In Paba *et al*, labeled peptides were analyzed by 2DLC-MS/MS resulting in the identification of 121 peptides which matched with high confidence to 41 proteins. By correlating the peak intensities between the light and heavy labeled peptides, relative abundances were calculated for each of the 41 proteins. While the relative expression for 29 proteins did not change between the two developmental stages the analysis did result in differential expression data for 12 proteins. As in the *T. cruzi* 2D-PAGE proteome (14, 15) many of these proteins were associated with the loss of the flagellum (PAR1 and PAR2), and an increase in cell differentiation (Poly zinc finger protein) following amastigogenesis. However, the scope of this study was limited by presence of several very high abundance proteins (heat shocks, histones, tubulins) which inhibited the detection of proteins in lower concentrations. This is certainly a problem because parasite specific proteins useful as targets for vaccine research may be present at 5-10 orders of magnitude less than these highly abundant proteins. Therefore, it is important, especially when working with kinetoplastids, to begin the proteomic analysis with sufficient starting material such that the low abundance proteins are of high enough concentration to be detected by the mass spectrometer following thorough peptide separations.

As discussed, a comprehensive understanding of *T. cruzi* gene expression has not been achieved to date. The vast majority of *T. cruzi* proteins identified by high-throughput proteomic strategies have been among the most highly expressed in the cell. With this in mind, the following three manuscripts all focus on the implementation of shotgun proteomics to further elucidate the regulation of *T. cruzi* gene expression and post-translational modifications throughout its developmental cycle.

REFERENCES

1. World Health Organization., 2002. WHO Tech. Rep. Ser. 905, 82-83.
2. Cubillos-Garzon, L., Casas, J., Morillo, C., Bautista, L., 2004. Am. Heart. J. 147, 412-417.
3. Urbina, J., 2002. Cur. Pharm. Des. 8, 287-295.
4. De Souza, W., 2002. Kinetoplastid Bio. Dis. 1, 1-21.
5. Costa, F., Franchin, G., Pereira-Chioccola, V., Ribeirao, M., Schenkman, S., Rodrigues, M., 1998. Vaccine. 16, 768–774.
6. Wizel, B., Garg, N., Tarleton, R., 1998. Infect. Immun. 66, 5073–5081.
7. Planelles, L., Thomas, M., Alonso, C., Lopez, M., 2001. Infect. Immun. 69, 6558–6563.
8. Huang, L., Jacob, R. J., Pegg, S. C., Baldwin, M. A., Wang, C. C., Burlingame, A. L., Babbitt, P. C., 2001. J. Biol. Chem. 276, 28327-39.
9. Huang, L., Shen, M., Chemushevich, I., Burlingame, A, L., Wang, C. C., Robertson, C. D., 1999. Mol. Biochem. Parasitol. 102, 211-223.
10. Hua, S., To, W. Y., Nguyen, T. T., Wong, M. L., Wang, C. C., 1996. Mol. Biochem. Parasitol. 78, 33-46.
11. van Deursen, F. J., Thorton, D. J., Matthews, K, R., 2003. *Mol. Biochem. Parasitol.* 128, 107-110.
12. Bente, M., Harder, S., Wiesgigl, M., Heukeshoven, J., Gelhaus, C., Kraus, E., Clos, J., Bruchhaus, J., 2003. Proteomics. 3, 1811-1829.
13. Gongora, R., Acestor, N., Quadroni, M., Fasel, N., Saravia, N. G., Walker, J., 2003. Biomedica. 23, 153-160.

14. Parodi-Talice, A., Duran, R., Arrambide, N., Prieto, V., Pineyro, M., Cayota, A., Cervenansky, C., Robello, C., 2004. Inter. J. Parasit. 34, 881-886.

15. Paba, J., Santana, J., Teixeira, A., Fontes, W., Sousa, M., Ricart, C., 2004. Proteomics. 4, 1052-1059.

16. Santoni, V., Molloy, M., Rabilloud, T., 2000. Electrophoresis. 21, 1054-1070.

17. Mann, M., Hendrickson, R., Pandey, A., 2001. Annu. Rev. Biochem. 70, 437-73.

18. Yates, J., 1998, J. Mass. Spec. 33, 1-19.

18. Beranova-Giorgianni, S., 2003. Trends. Anal. Chem. 22, 273-281.

19. Mishra, K., Holzer, T., Moore, L., LeBowitz, J., 2003. Eukaryot. Cell. 2, 1009-1017.

20. Saborio, J., Hernandez, J., Narayanswami, S., Wrightsman, R., Palmer, E., Manning, J., 1989. J. Biol. Chem. 264, 4071-4075.

21. Minning, T., Bua, J., Garcia, G., McGraw, R., Tarleton, R., 2003. Mol. Biochem. Parasitol. 131, 55-64.

22. Diehl, S., Diehl, F., El-Sayed, N., Clayton, C., Hoheisal, J., 2002. Mol. Biochem. Parasitol. 123, 115-123.

23. Saxena, A., Worthey, E., Yan, S., Leland, A., Stuart, K., Myler, P., 2003. Mol. Biochem. Parasitol. 129, 103-114.

24. Clayton, C., 1999. Parasitol. Today. 15, 372-378.

25. Andersson et al. 2005. Science, In press.

26. Besteiro, S., Barrett, M., Riviere, L., Bringaud, F., 2005. Trends. Parasitol. 4, 185-91.

27. Giambiagi-DeMarval, M., Souto-Pardron, T., Rondinelli, E., 1996. Exp. Parasitol. 83, 335-345.

28. Billaut-Mulot, O., Fernandez-Gomez, R., Loyens, M., Ouaissi, A., 1996. 174, 19-26.

29. Opiteck, J., Jorgenson, W., 1997. Anal. Chem. 69, 2283-2291.

30. Opiteck, J., Lewis, C., Jorgenson, W., Anderegg, J.,1997. Anal. Chem. 69, 1518-1524.

31. Link, J., Eng, J., Schieltz, M., Carmack, E., Mize, J., Morris, R., Garvik, M., Yates, J., 1999. Nat. Biotechnol. 17, 676-682.

32. Wall, B., Kachman, T., Gong, S., Parus, J., Long, W., Lubman, M., 2001. Rapid. Commun. Mass. Spectrom. 15, 1649-61.

33. Jensen, K., Pasa-Tolic, L., Peden, K., Martinovic, S., Lipton, S., Anderson, A., Tolic, N., Wong, K., Smith, D., 2000. Electrophoresis. 21, 1372-1380.

34. Florens, L., *et al*., 2002. Nature. 419, 520-526.

35. Peng, J., Elias, E., Thoreen, C., Licklider, L., Gygi, S., 2003. J. Proteome. Res. 2, 43-50.

36. Yan, W., Lee, H., Yi, C., Reiss, D., Shannon, P., Kwieciszewski, K., Coito, C., Li, J., Keller, A., Eng, J., Galitski, T., Goodlett, R., Aebersold, R., Katze, G., 2004. Gen. Bio. 5:R54.

37. Delahunty, C., Yates, J., 2005. Methods. 35, 248-255.

38. Aebersold, R., Mann, M. 2003. Nature. 422, 198-207.

39. Wang, H., Hanash, S., 2003. J. Chromatogr. B. 787, 11-18.

40. Issaq, H., Conrads, T., Janini, G., Veenstra, T., 2002. Electrophoresis. 23, 3048-3061.

41. Giddings, J., 1987. High. Resolut. Chromatogr. 10, 319.

42. Wolters, D., Washburn, M., Yates, J., 2001. Anal. Chem. 73, 5683-5690.

43. Xiang, R., Shi, Y., Dillon, D., Negin, B., Horvath, C., Wilkins, J., 2004. J. Proteome. Res. 3, 1278-1283.

44. Vollmer, M., Horth, P., Nagele, E., 2004. Anal. Chem. 76, 5180-5185.

45. Paba, J., Ricart, C., Fontes, W., Santana, J., Teixeira, A., Marchese, J., Wiliamson, B., Hunt, T., Karger, B., Sousa, M., 2004. 3, 517-524.

46. Alpert, A., Andrews, P., 1988. J. Chromatogr. 443, 85-96.

47. Michaud, A., Frank, A., Ding, C., Zhao, X., Douglas, D., 2005. J. Am. Soc. Mass. Spectrom.16, 835-49.

48. Chernushevich, I., Loboda, A., Thomson, B., 2001. J. Mass. Spectrom. 36, 849-65.

49. MacCoss, M., 2005. Curr. Opin. Chem. Biol. 9, 88-94.

50. Gygi, S., Rist, B., Gerber, S., Turecek, F., Gelb, M., Aebersold, R., 1999. Nat. Biotechnol. 17, 994-999.

CHAPTER 3

A HEURISTIC METHOD FOR ASSIGNING A FALSE DISCOVERY RATE FOR PROTEIN IDENTIFICATIONS FROM MASCOT DATABASE SEARCH RESULTS[1]

[1] Atwood, J., Weatherly, D., Minning, T., Cavola, C., Tarleton, R., and Orlando, R. 2005. *Molecular and Cellular Proteomics*. 4:762-772.

Reprinted here with permission of publisher.

ABSTRACT

Tandem mass spectrometry and database searching has emerged as a valuable technology for rapidly analyzing protein expression, localization, and post-translational modifications. The probability-based search engine Mascot has found widespread use as a tool to correlate tandem mass spectra with peptides in a sequence database. While the Mascot scoring algorithm provides a probability-based model for peptide identification, the independent peptide scores do not correlate with the significance of the proteins to which they match. Herein, we describe a heuristic method for organizing proteins identified at a specified false discovery rate using Mascot matched peptides. We call this method PROVALT, and it uses peptide matches from a random database to calculate false discovery rates for protein identifications and reduces a complex list of peptide matches to a nonredundant list of homologous protein groups. This method was evaluated using Mascot identified peptides from a *Trypanosoma cruzi* epimastigote whole-cell lysate, which were separated by multidimensional liquid chromatography and analyzed by tandem mass spectrometry. PROVALT was then compared with the two traditional methods of protein identification when using Mascot, the single peptide score and cumulative protein score methods, and was shown to be superior to both in regards to the number of proteins identified and the inclusion of lower scoring nonrandom peptide matches.

INTRODUCTION

As a complement to gene expression profiling, proteomics is the analysis of gene and cellular function at the protein level. The definition of proteomics has expanded to include not simply the proteins encoded by a genome but the analysis of protein isoforms, post-translational modifications, and protein-protein interactions (1-4). While two-dimensional gel electrophoresis and mass spectrometry have been the dominant techniques used in this field, liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) has emerged as a valuable technology for the analysis of complex protein mixtures (5). Typical LC-MS/MS experiments are subdivided into five stages. First, the proteins are extracted from the biological material followed by enzymatic digestions to produce peptides. The peptides are partitioned through multiple separations followed by analysis with LC-MS/MS. Two methods can then be employed to correlate an MS/MS spectrum with the peptide and protein from which it originated. The MS/MS spectrum is either sequenced de novo, yielding a complete amino acid sequence or sequence tag (6-13) which is then searched against a protein database, or software algorithms are used to directly match experimental with theoretical MS/MS spectra generated from peptides in a protein database (14-21).

All software designed for matching MS/MS spectra to peptide sequences function in a similar manner. The experimental peptide masses are first compared to theoretical peptide masses derived from *in silico* enzymatic digestion of the specified protein database. The subsets of theoretical peptides with similar masses (within a user defined mass tolerance) to the experimental peptides are fragmented *in silico* according to specific cleavage rules. These theoretical fragment ion masses are then compared with the masses of fragment ions from the experimental MS/MS spectra. While all search engines will match a theoretical peptide sequence

to an experimental MS/MS spectrum, every match is not always correct. Thus, all search algorithms attempt to assign scores indicating the degree of similarity between the experimental and theoretical MS/MS spectra.

Probability-based scoring algorithms attempt to accurately reflect the probability of a match being random by gathering information about the database itself and using this information in the score calculation. Of the probability-based search engines described in the literature, Mascot (14) is one of the most widely used. The Mascot probability model is based on the MOWSE (22) algorithm. The MOWSE algorithm functions by creating a matrix of weighting factors for the specified enzymatic cleavage, and sequence entries are clustered into cells formed by a distribution of intact protein and peptide fragment molecular weights. Therefore, each cell contains the frequency at which peptide molecular weights occur for a distribution of protein masses. The frequency factors are normalized and used to calculate the final score. Mascot reports a probability-based ion score for each peptide match, which indicates the statistical significance of that MS/MS spectral assignment. Following clustering of peptides to proteins, Mascot combines individual ion scores and reports a cumulative protein score. While the Mascot algorithm (www.matrixscience.com/help/results_help.html) establishes a threshold ion score under which there is less than 95% confidence in any individual peptide match, the algorithm attributes no such statistical significance to the cumulative protein scores.

Published proteomics initiatives using Mascot have established a variety of criteria necessary to distinguish between correct and random protein assignments. A general approach has been to report a protein identified if at least one peptide from that protein is matched at or above the threshold ion score (23-25). We refer to this as the single peptide score method (SPM). Another approach, the cumulative protein score method (CPM), has been to eliminate peptide

matches below a given ion score and utilize an empirically derived protein score to identify proteins (26, 27). These strategies hinge on the ability of Mascot to accurately reflect true probabilities associated with matching MS/MS spectra to peptide sequences. With this in mind, evaluations of the Mascot scoring algorithm have demonstrated that the ion score thresholds established for a match to be considered significant accurately reflect a 95% confidence level when searching data from peptide mass fingerprints (28) and tandem mass spectra generated on high precision quadrupole-time of flight (QTOF) instruments (17). However, in large scale proteomics projects, when thousands of peptides are identified, this level of confidence may be unsatisfactory, resulting in several hundred false positive peptide and/or protein identifications. To discriminate between correct and random peptide assignments from a Mascot search, several methods have been proposed. These include statistical models to evaluate peptide assignments using information gathered during the database search (29, 30) and filters which eliminate random matches based on the properties of the assigned peptides and experimental conditions (31-33). These methods, while allowing the removal of a large portion of the false positive peptide identifications, either ignore potentially valid lower scoring peptide matches or do not account for the false discovery rate (FDR) associated with the resulting protein identifications.

For most proteomics projects the eventual goal is to establish the repertoire of expressed proteins in the sample of interest (34). As noted by Nesvizhskii (35, 36) this is not a straightforward process when analyzing proteins at the peptide level. The statistical significance associated with matching MS/MS spectra to peptide sequences does not correlate with the likelihood that the consequent proteins are identified (35 - 37). This is ascribed to real peptide matches clustering to individual proteins while random matches occur with an equal distribution across the entire database (35, 36). To calculate an overall probability or expectation value for a

protein assignment, Nesvizhskii (35, 36), Sadygov (37), and Fenyo (30) have proposed combining individual probabilities or expectation values for potential peptide matches. Such methods allow the specification of expectation values (30) or probabilities (35-37) to the proteins rather than peptide assignments and enable the inclusion of potentially valid low scoring peptide matches in the final protein assignment (35-37). However, these methods require the recalculation of expectation values or probabilities of peptide assignments made by the database search software and provide no information about the overall proportion of incorrect matches in a subset of identified proteins.

Herein we detail PROVALT, a tool which organizes large proteomic datasets and calculates protein false discovery rates (*PRO-FDR*) using peptides identified by Mascot. PROVALT extracts peptide matches from multiple Mascot results files, eliminates peptide redundancy, and clusters peptides to their corresponding proteins. In addition, homologous proteins for which no distinguishing peptide has been identified are organized into protein groups. Pertinent information such as sample origin and experimental conditions are linked to each peptide and protein match as well. We also report the implementation of a statistical model within PROVALT: an algorithm which determines *PRO-FDRs* when using Mascot. This model is based on the implementation of a random database (32, 38, 39) and Mascot peptide probability scores to calculate expected *PRO-FDRs* for a minimal number of expressed proteins identified by Mascot matched peptides. As in prior reports, the random database served as the null hypothesis. Identifications resulting from the null hypothesis are considered to be random and the scores assigned to the random matches should follow a quasi-normal distribution with a false positive rate related to each score. PROVALT compares the score distributions obtained from searching the normal and random databases to calculate the false discovery rates associated with

each score threshold. The goal is to set score thresholds which identify as many real proteins as possible while encountering a minimal number of false positives protein identifications. It is important to note the difference between the false positive rate and false discovery rate as it applies to protein identification. The false positive rate is the rate at which truly null protein identifications are treated as significant while the false discovery rate is the rate at which significant protein identifications are actually null (40-42). For example, Peng *et al*, used a random database to identify 7,537 peptides at a false discovery rate of 1%. Thus, 75 of the identified peptides were likely random matches (38).

Rather than calculate error rates at the peptide level we demonstrate that the protein false discovery rate calculations employed by PROVALT provide a reasonable balance between the number of correct and incorrect protein assignments. Here we evaluate this tool using a dataset of ~50,000 MS/MS spectra generated from 100 LC-MS/MS analyses of peptides from *Trypanosoma cruzi* epimastigote whole-cell lysates, which were separated by multidimensional liquid chromatography. Using this dataset we evaluate the two predominant modes of protein identification when using Mascot, the single peptide score method and cumulative protein score method. We then compare the *PRO-FDRs* resulting from these traditional methods with the *PRO-FDRs* generated from the PROVALT analysis.

EXPERIMENTAL PROCEDURES

Cell culture and peptide preparation

T. cruzi (Brazil) epimastigotes were grown as previously described (43). Proteins were extracted from 5×10^7 parasites using Tri-Reagent (Sigma, St. Louis, MO) per the manufacturers' instructions. Proteins were reduced (8 M urea, 200 mM Tris-HCl, 40 mM DTT, pH 8.5) for 1hr

at 55ºC followed by carboxyamidomethylation with iodoacetamide (80 mM) for 30 min at room temperature. The protein solution was diluted to 6 M urea by addition of H_2O and digested overnight at 37ºC with endoproteinase Lys-C (1:100, Sigma). Following dilution to 1 M urea with 30 mM ammonium bicarbonate proteins were digested overnight at 37ºC with sequencing grade porcine trypsin (1:50, Promega, Madison, WI). The digest was lyophilized to dryness, reconstituted in 500 µl of 0.1% TFA, and filtered prior to separation.

First Dimension Reverse Phase Chromatography (RP)

First dimension separation was performed on an Agilent 1100 series workstation (Palo Alto, CA) configured with a 4.6 × 150 mm Jupiter C_{18} column (Phenomenex, Torrance, CA). Buffer A was H_2O/0.1% TFA, and buffer B was acetonitrile (ACN) /0.1% TFA. The peptides were loaded onto the column (flow rate 0.75 ml/min), desalted for 10 min at 5% buffer B, and then sequentially eluted during a 40 min linear gradient from 5 to 45% buffer B. Ten fractions were collected (Table 3.1A), frozen, and lyophilized to dryness.

Second Dimension Strong Cation Exchange Chromatography (SCX)

SCX was performed on a Hewlett Packard 1100 series workstation (Palo Alto, CA) equipped with a 1.0 × 150 mm polysulfoethyl A column (5 µm, 300 Å, PolyLC, Columbia, MD). The buffer solutions were 5% ACN/0.5% acetic acid (pH 3.0, SCX-A) and 175 mM ammonium acetate/5% ACN/0.5% acetic acid (pH 4.0, SCX-B). Each RP fraction was resuspended in 100 µl SCX-A, loaded at a flow rate of 75 µl/min, and washed with 100% SCX-A for 10 min. Ten fractions were collected at 5 min intervals during a 50 min linear gradient from 10 to 60% SCX-B. The fractions were dried by vacuum centrifugation and stored at -20ºC until analysis by LC-MS/MS.

RPLC-MS/MS

Each fraction was analyzed independently using a Waters CapLC (Milford, MA) interfaced directly to a QTOF-2 tandem mass spectrometer (Micromass, UK). Mobile phase A and B were H_2O/0.1% formic acid and ACN/0.1% formic acid, respectively. Each fraction was reconstituted in 5 µl of mobile phase A and loaded onto the column (PEPMAP, 15 cm × 180 µm, LC-Packings, Sunnyvale, CA) at a flow rate of 1 µl/min. In order to maximize separation, RP gradients were designed to correspond with the percentage of organic at which each peptide fraction eluted from the first dimension RP (Table 3.1).

MS/MS Parameters

The instrument was set to acquire a 1 sec MS scan from 500-1700 Da during which up to four precursor ions were selected for MS/MS analysis from 50-2000 Da for 2 sec. Precursor selection was based upon a threshold of 5 counts/sec for peptides with charge states of 1-4, and dynamic exclusion was enabled for 8 min. Raw mass spectra were processed into peak-list format prior to database searching.

Protein Sequence Database

Two sequence databases were constructed for our analyses. First, we created a representative database (normal) consisting of the approximately 25,000 *T. cruzi* gene annotations provided by *Trypanosoma cruzi* Sequencing Consortium (TSK-TSC) as well as possible contaminating proteins from *Bos taurus*, *Equus caballus*, *Homo sapien*, and proteases. A randomized database (random) was constructed by reversing the sequences in the normal database. The random database was used to establish accurate scoring thresholds for protein identification in the normal database.

Database Searching

Mascot (version 1.8) searches were performed with the following parameters: a specified trypsin enzymatic cleavage with 1 possible missed cleavage, peptide tolerance of 50 parts-per-million, fragment ion tolerance of 0.2 Da, variable modifications due to carbamylation (+ 43 Da), and carboxyamidomethylation (+ 57 Da).

PROVALT: Protein Organization and False Discovery Rate Calculations

The PROVALT method for identifying proteins at a specified *PRO-FDR* is outlined in (Fig 3.1)**.** Protein identification begins by extracting all Mascot matched peptides and corresponding ion scores from the normal and random Mascot search results. The results from each dataset are combined and filtered yielding nonredundant lists of peptides. The peptides are grouped as function of score thus forming score bins (B_i) containing all peptides equal to or exceeding each Mascot ion score (i) where i represents a specific ion score S ranging from M to N.

$$\begin{aligned} M &= \mathrm{Min}\left(S \middle| nPEP(S) \ > \ rPEP(S)\right) \\ N &= \mathrm{Min}\left(S \middle| rPEP(S) = 0\right) \\ nPEP(S) &= \#\text{ peptides in normal database} \geq S \\ rPEP(S) &= \#\text{ peptides in random database} \geq S \end{aligned} \qquad (1)$$

Peptides in each score bin are then clustered to their corresponding proteins. Proteins are selected based on the degree of peptide coverage c.

$$\begin{aligned} c &= \{C, C-1, ..., 1\} \\ C &= \text{User define maximum degree of peptide coverage} \end{aligned} \qquad (2)$$

Starting with C, a histogram is formed based on the frequency of protein identifications within each B_i for both the random and normal peptide matches (Fig 3.2). Protein identification false discovery rates are then calculated for each score bin as follows.

$$PRO\text{-}FDR_c(S) = \left(\frac{rPRO_c(S)}{nPRO_c(S)} \right) \times 100\%$$

$$nPRO_c(S) = \#\text{ proteins identified in normal w/ peptide coverage } c \text{ in } B_s \qquad (3)$$

$$rPRO_c(S) = \#\text{ proteins identified in random w/ peptide coverage } c \text{ in } B_s$$

PROVALT then determines threshold S_c (for $c = C$)

$$S_c = \text{Min}\left(S \mid PRO\text{-}FDR_c(S) \leq \text{Max } PRO\text{-}FDR, \text{for } S = \{M, M+1, \ldots N\}\right) \qquad (4)$$

$$\text{Max } PRO\text{-}FDR = \text{user defined maximum protein false discovery rate}$$

and thus the minimal Mascot ion score threshold necessary to achieve the specified *PRO-FDR* for the given degree of peptide coverage (*c)*. Peptides meeting these criteria are stored and the remaining peptides which were not matched to proteins are then grouped as a function of ion score (*S*) thus forming new score bins (B_i) containing all peptides equal to or exceeding each Mascot ion score (*i*). For the next degree of coverage ($c = C$-1) the distribution of ion scores will have a minimum (*M)* which is dependent on score threshold (S_c) determined for the previous degree of peptide coverage as follows.

$$M = S_{c+1} + 1 \text{ for } c < C$$

$$N = \text{Min}\left(S \mid rPEP(S) = 0\right) \qquad (5)$$

$$rPEP(S) = \#\text{ peptides in random database} \geq S$$

Previously unmatched peptides in each bin are again clustered to their corresponding proteins along with the peptides matched at the previous levels. For the next degree of peptide coverage (C-1) another histogram is formed and the *PRO-FDR* is then calculated for each score bin. S_c (for $c = C$-1) is calculated and the peptides meeting these criteria are stored. This process is repeated until c =1. For this dataset the ion score thresholds for each coverage level are displayed in Table 3.2.

RESULTS

Our proteomic analysis of epimastigotes of *Trypanosoma cruzi* generated approximately 50,000 spectra from 100 LC-MS/MS analyses of peptides from epimastigote whole-cell lysates. Whole cell lysates similar to those used in this study may contain greater than 10,000 proteins expressed at a 10^6 dynamic range of concentrations. Also, such complex samples are often comprised of large protein families and post-translationally modified proteins which complicate protein identification because many peptides are identified multiple times and/or match to several homologous proteins. While Mascot, which is one of the most widely used probability-based search engines, groups peptides to proteins and ranks protein identifications as a function of cumulative protein score, it does not combine results from multiple LC-MS/MS analyses. Our analysis of the epimastigote proteome generated 100 results files containing peptide identifications which needed to be combined to identify proteins, which made using Mascot in its native form impractical. While a single file could have been generated by concatenating all 100 peak-list files, we chose to search each file individually using the parameters described above. The reasoning for this was two fold: not only would this amount of data require robust and expensive computational resources if searched as a single file, it would also make identifying peptides present in different fractions impossible, as the visibility to the source fraction for each spectrum would be lost. These problems were eliminated by the integration of a Mascot results parser and peptide clustering algorithm into PROVALT.

Mascot Parsing Tools

The integration of the Mascot parser allows the extraction of relevant peptide information from multiple Mascot results files. The parsing tool functions with both dat and html versions of Mascot results. The inclusion of an html parser allows users to access the nonlicensed version of

Mascot (www.matrixscience.com), save multiple results files as html, and extract the peptide matches for later protein identification. The extracted information includes sample source, peptide sequence, ion score, precursor mass error, and fragment mass errors. The information is stored along with the peptide sequences and is later integrated with the protein identification. Following parsing, PROVALT eliminates redundant peptide identifications (but tracks the information associated with each valid occurrence), matches peptides to proteins, and then determines a list of unique protein groups. While redundant peptides are removed and only the highest probability peptide match is used for the subsequent protein identification, all peptide matches are reported. Peptide matches are considered redundant if they have identical sequences plus any modifications or resulted from precursor ions of multiple charge states. If a peptide was identified with and without a modification, PROVALT does not consider these redundant. All matches to a given peptide regardless of precursor ion charge state are reduced to a single identification and the highest ion score (*S*) is used in the *PRO-FDR* calculation. While spectra from multiple precursor ion charge states may exhibit distinct fragmentation patterns, these peptides are not considered as independent peptide matches by PROVALT and are not allowed to contribute to the peptide coverage (*c*) (44). We have adopted this approach because of the fact that during database searching precursor ions of multiple charge states will be converted to their singly charged precursor masses prior to comparison with the *in silico* derived peptides. The subset of *in silico* derived peptides passing within the precursor ion mass tolerance will essentially be identical for the singly charged and multiply charge species. Thus identical assignments made from peptides with different precursor ion charge states are not unequivocally independent events. Conversely, peptide assignments with alternate modifications are matched through different parent mass filters and are treated as independent.

Peptide Clustering Tools

Correlating peptides to proteins in a large-scale proteomics project is not a straightforward process; Nesvizhskii (36) and Rappsilber (45) noted that eukaryotic protein databases contain many homologous proteins which are often indistinguishable by a single peptide identification. Therefore, if a group of peptides corresponds to more than one protein it is impossible to determine which protein is actually expressed without performing additional experiments. In a manner similar to the organization described in Nesvizhskitt *et al*, PROVALT accounts for the presence of protein families by determining protein groups which represent all peptides used to identify a group of homologous proteins. The protein groups were determined as follows: If protein “A” and protein "B" both contained the peptide sequences "x" and "y" then the identification of peptides "x" and "y" would not allow proteins "A" and "B" to be distinguished from each other. Thus protein group 1 would be defined as peptides “x” and “y” and would contain both proteins “A” and “B”. If protein “C” contained the peptide sequences “x” and “y” and "z" then the identification of peptide “z” would allow protein “C” to be differentiated from both proteins “A” and “B”. Thus protein “C” would be placed in protein group 2. Following peptide clustering a protein which contains all the peptides in the protein group is used in the *PRO-FDR* calculation. For example, protein B would represent protein group 1 with a peptide coverage of $c = 2$ and the ion scores (S) of both peptides “x” and “y” would be used in the *PRO-FDR* calculation. While proteins A and B may both be expressed, for statistical reasons we considered the protein group as a single protein identification. If the representative protein (B) is determined to be identified above the defined *PRO-FDR* then the final report contains all of the homologous proteins (A and B), not just the representative protein. Our method assumes that if a protein is identified above a given error rate the constituent

peptides are also identified with the same confidence. Therefore, this method makes no statistical distinction between proteins in the same group that are identified by differing degrees of peptide coverage. While it may be assumed that a protein is more likely to be expressed if it contains a larger number of matched peptides it is indeterminate whether a subset of those peptides resulted from the presence of a homologous protein. As a result, the total number of protein groups identified represents the minimal number of expressed proteins.

Single Peptide Score Method

The most common approach for separating random from real protein identifications is the single peptide score method (SPM), in which a protein is identified if it contains at least a single peptide at or above the Mascot derived ion score threshold. Mascot defines a peptide match as significant if the probability of that event occurs by chance with a frequency of less than 5%. Thus a peptide with an ion score corresponding to at least an absolute probability ≤ 0.05 is considered a real match and used in the protein identification. To model this method the ion score thresholds which corresponded to each absolute probability needed to be determined over the entire dataset. This was accomplished by searching both the normal and random databases, forming a histogram of the frequency of peptide matches as a function of ion score, then applying the equation below to calculate the distribution of peptide false discovery rates (*PEP-FDR*).

$$PEP\text{-}FDR(S) = \left(\frac{rPEP(S)}{nPEP(S)} \right) \times 100\% \quad (5)$$

Figure 3.3 is the distribution of peptide matches which occurred at or above a range of ion scores for the normal and random database search results as well as the calculated *PEP-FDRs*. At an ion score of 27 or above, 5% of the peptides matching in the normal database would be random

identifications. Typically, when the SPM is applied, only peptide matches with ion scores equal to or exceeding 27 would be used to identify proteins.

Comparison of PROVALT and SPM

In order to compare the protein false discovery rates generated from PROVALT with the SPM, peptide matches above discrete ion scores were extracted from the normal and random database search results and clustered to proteins using the PROVALT clustering tool. A histogram was formed based on the frequency of protein identifications by at least one peptide at or above ion scores in the range of 22-50. *PRO-FDRs* were then calculated according to equation 3 and plotted along with the *PEP-FDRs* as a function ion score (Fig 3.4). Figure 3.4 demonstrates that an acceptable *PEP-FDR* does not necessarily correlate with an acceptable *PRO-FDR*, which is a serious drawback of using the *PEP-FDR* to identify proteins. In fact, the *PRO-FDR* grows much more rapidly than the *PEP-FDR* as the minimum ion score threshold decreases. For example, if peptides are selected for protein identification based on a *PEP-FDR* of 5%, the resulting *PRO-FDR* is 24%. This occurs because the peptide false discovery rate only represents the rate at which a peptide is randomly matched during a database search and does not account for the fact that random peptide matches do not cluster to individual proteins at the same rate as real peptide matches. This trend is shown in Figure 3.5 in which the ratio of peptides to proteins was plotted versus peptide score for both the normal and random database searches using the SPM. From Figure 3.5 it is evident that the ratio of peptides to proteins is much larger in the normal database search than it is in the random database search, proving the improbability that multiple random peptide matches will occur on a single protein. Thus a set of randomly identified peptides will result in far more protein identifications than an equal number of real peptide identifications.

In order to utilize the single peptide score method with high confidence in protein identifications, one must be highly stringent when choosing a minimum ion score threshold. To achieve a *PRO-FDR* of 1% for this dataset, a minimum ion score of 42 is necessary (a *PEP-FDR* of 0.2%). This threshold was chosen, and the peptides meeting this criteria were clustered to proteins. These results, along with those of using PROVALT at a 1% *PRO-FDR*, are displayed in Figures 3.6A and 3.6B. Using the PROVALT method, 1935 unique peptides were used to identify 444 unique proteins. To achieve the same *PRO-FDR* using the single peptide score method, 1064 unique peptides would match to 386 unique proteins. While our method results in over 15% more proteins, the increase in the number of peptides is nearly double. Clearly, the single peptide score method limits the number of peptides that can match to each protein by discarding all lower scoring peptides. However, this is not advantageous. As Figure 3.3 indicates, non-random peptide matches exist at lower scores, but the use of the single score threshold does not facilitate discriminatory inclusion of these peptides.

Cumulative Protein Score Method

The cumulative protein score method involves the extraction of all peptide matches above a given ion score and utilizes an empirically derived protein score (sum of peptide ion scores) to identify proteins. This method has two advantages. First, potentially nonrandom lower scoring peptides are used in the final protein identification. Second, individual peptide ion scores are summed to yield a non-probabilistic cumulative protein score. However, to date there has been no statistical basis for differentiating between random and nonrandom peptide matches below the ion score threshold supplied by Mascot, and furthermore, no statistical method for choosing a cumulative protein score has been published. In addition, for this method to be effectively

applied the peptides must be correctly clustered to their corresponding proteins first; consequently for unconcatenated peak lists the Mascot derived protein score is unusable.

Comparison of PROVALT and CPM

We evaluated the effectiveness of the cumulative score method by identifying proteins in both the normal and random databases and calculating false discovery rates based on cumulative protein score thresholds. We first extracted all top-ranking unique peptides with an ion score at or above 1 from both databases, clustered them to the corresponding proteins, and calculated a total protein score by summing the individual ion scores**.** Figure 3.7 is the frequency distribution of proteins which match at a range of protein score thresholds (from 20 to 120) for the normal and random databases as well as the calculated *PRO-FDRs*. As stated above, determining a protein score threshold has previously been largely empirical, and, for small datasets, Mascot reports as "significant" hits all proteins with a cumulative score above the ion score threshold. However, Figure 3.7 indicates that employing this method for this dataset would result in a *PRO-FDR* of 97% at a score of 27 (ion score determined from Figure 3.3). To achieve a *PRO-FDR* of 1%, a minimum protein score of 110 would be required. Figures 3.6A and 3.6B compare the number of proteins and peptides identified at a 1% *PRO-FDR* when using the CPM and PROVALT methods. While a high-confidence *PRO-FDR* for protein identifications is achieved using both, a vast number of nonrandom protein matches are discarded by the CPM (Fig 3.7). On the other hand, if a lower protein score threshold is used, the *PRO-FDR* would be unacceptably high.

Due to the prevalence of lower scoring random peptide matches contributing to the cumulative protein score, the CPM has been employed following removal of these lower scoring peptides prior to protein identification (26, 27). While this approach offers a good compromise,

as one increases the peptide score threshold required for protein identification the method degenerates into a single peptide score method, where a large number of nonrandom peptide matches are removed that are important for accurate protein identifications**.**

Comparison of PROVALT and modified CPM

Since the frequency of random peptide matches increases as the Mascot ion score decreases, the application of a low score cutoff to filter potentially random peptide matches has been proposed (26-27). However, this method ignores higher scoring potentially random matches which will positively contribute to the final protein false discovery rate calculation. For comparison of PROVALT with this version of the CPM, we applied several peptide ion score thresholds (1, 10, 20, 30, and 40) to filter peptides prior to protein identification with the CPM at a 1% *PRO-FDR*. Figure 3.8 compares the distribution of peptides at various peptide ion score ranges using our method and the cumulative score methods to identify proteins. From Figure 3.8 it is evident that our method is more discriminating in the inclusion of lower scoring peptides than the cumulative score method, which utilizes a higher number of lower scoring peptides for protein identification. In fact, when the cumulative score method is employed with peptide score thresholds of less than 30, the predominant number of peptides are within the lowest ion score range. Thus, protein identification with this method is mostly facilitated by inclusion of the peptides that are most likely random matches. Using PROVALT the largest number of peptides used for protein identification falls within the highest ion score range. The cumulative score method does not attribute any significance to the lower scoring peptides but rather assumes that their presence with higher scoring peptides on the same proteins makes them significant. However, because so few high scoring peptides exist in the random database, lower scoring peptides that match to the normal database have an unfair statistical advantage. PROVALT

utilizes only a stochastic peptide ion score threshold that is based on the assumption that random matches will occur with an even distribution across all proteins in the database, while real matches will cluster to the proteins that are actually expressed. Thus, score thresholds are calculated according to the probability that some number of peptides match above a given ion score to the same protein regardless of total protein score. This allows us to have high confidence in all of the peptides that are selected by our method and thus the proteins that they identify. Without high confidence in the lower scoring peptides, the cumulative score method provides essentially the same amount of quality information as the single peptide score method.

DISCUSSION

Recent technological advancements in the field of proteomics have facilitated the analysis of complex protein mixtures resulting in a dramatic increase in the number of published proteomes. Many of these projects have utilized the Mascot algorithm in order to match tandem mass spectra to peptides in a protein database. Although its use has been widespread, it remains unclear how Mascot should be applied to large scale proteomics projects. To date, three approaches have been offered: 1) using Mascot in its native form and employing either the single peptide score or cumulative protein score approach, 2) calculating protein false discovery rates as is employed by PROVALT, or 3) calculating an overall probability for protein identifications using the probabilities of individual peptide matches.

The main goal in using external tools such as PROVALT is to separate the valid from incorrect protein identifications, the difficulty of which increases with the size of the dataset. As has been shown, the problem with using either the single score threshold or cumulative protein score approach is balancing the tradeoffs between choosing high stringency that discards

potentially useful information and low stringency which does not discard the useless information. Thus it would be advantageous to determine a measure of confidence in protein identifications using all possible peptide contributions to that protein and select proteins based on a user-defined minimum confidence. Previous work has demonstrated the utility of the random database approach for calculating false discovery rates for peptide identifications (38). Such methods facilitate the calculation of the proportion of random peptide matches among all peptide matches deemed significant. However, the calculation of a peptide false discovery rate provides no information about the error rate of a specific protein identification. With the ultimate goal being the identification of proteins rather than peptides, a more thorough approach would be to define a protein false discovery rate. With this in mind we have demonstrated the use of PROVALT, a computational tool designed to work in conjunction with Mascot results files to provide a method for organizing large proteomic datasets and calculating false discovery rates for protein identifications.

PROVALT assigns a protein false discovery rate based on the distribution of proteins with minimum peptide coverage (c) identified in peptide ion score bins (B_s), information which can be determined for any proteomics project. However, special consideration should be given to the dataset size when applying this analysis because the statistical significance of the calculated error rates is diminished for small datasets. As has been the case with the use of peptide false discovery rates (38), the calculation of protein false discovery rates as performed by PROVALT is best suited for large datasets exceeding several hundred protein identifications. Otherwise, Mascot is best used in its native form due to the fact that manual verification of questionable identifications is feasible. The utility of PROVALT is as a high-throughput tool for large datasets

because it assigns a measure of statistical significance to all identified proteins, thus diminishing the need for manual verification.

In order to use PROVALT, the user must determine the maximum degree of peptide coverage (*C*) to use in the calculation of false discovery rates. The concept of a false discovery rate is based on the assumption that the null hypotheses (proteins identified in the random database) follow a quasi-normal distribution of minimal ion scores (S). Indeed, score distributions at increasing *c* may not necessarily look normal themselves, especially when the sample size is small. Thus, as the sample size (number of protein identifications) increases, for each *c* the score distribution will approach normality and the false discovery rate calculations will become more accurate. In practice determining *C* prior to PROVALT analysis will be dependent on the distribution of minimal ion scores *S* in the normal and random protein assignments. Upon assignment of beginning $c = C$ by the user, a version of figure 3.2 can be generated and visual inspection of the distribution will suggest whether the assumption of normality is reasonable. The user can increase or decrease the value of C until a statistically significant level of peptide coverage is determined for the normal database. It is sensible to assume that as the size of the dataset decreases, *C* will also decrease until the method degenerates into a single peptide score method at $C = 1$. Accordingly the native form of Mascot may be best suited for these situations.

Another approach, used in tools like Protein Prophet or PROT_PROBE, is to combine the conditional probabilities of individual peptide matches in order to calculate an overall probability for the protein identification (35-37). A protein identification is then considered significant if its individual probability exceeds a chosen threshold. The major difference between this method and that of PROVALT is that PROVALT calculates a protein false discovery rate (*PRO-FDR*) based

on minimum ion score thresholds and peptide coverage but does not calculate individual probabilities for each protein identification. It is important to note that the false discovery rate does not represent the probability that a feature is significant but rather the expected proportion of random matches resulting when proteins are identified using peptide score and coverage criteria. On the other hand, an individual protein probability does not represent the proportion of false discoveries which may result when a probability threshold is employed. Thus probability thresholds alone may not provide an assessment of overall significance for datasets that include a large number of protein identifications. Sadygov *et al*, observed that protein probabilities always approach 1 when calculated through a binomial distribution of peptide probabilities which do not account for database or dataset size (37). This is crucial because as the size of the dataset increases the frequency of peptide matches to a given protein will increase regardless of the quality of the peptide match, resulting in a number of proteins being identified by a subset of random peptides. In contrast, the exploitation of a random database allows the calculation of false discovery rates regardless of database search parameters, database size, or dataset size. Another apparent benefit of this approach is that the entire distribution of random peptide matches is observed and factored into the final protein false discovery rate calculation. As the number of random matches increases with dataset size the significance thresholds necessary to achieve a specified protein false discovery rate will also increase thus compensating for the frequency of higher scoring random matches. Ideally, however, one would prefer to have both protein probabilities and protein false discovery rate estimations. This would allow researchers to select a set of proteins based on a minimum false discovery rate but remove those which have an unacceptable individual protein probability, which is especially applicable for low peptide coverage members of high confidence protein groups.

In this work, the traditional methods of protein identification when using Mascot have been examined. Both the single peptide score and cumulative protein score strategies mentioned above propose the filtering of peptides in some manner prior to protein identification. We feel this is incorrect because such strategies ignore the phenomenon of peptide clustering and fail to assign any statistical significance to the thresholds used for the protein identifications. In addition, an accurate treatment of redundant protein and peptide identifications is crucial if a legitimate false discovery rate is to be determined and careful consideration must also be given to how peptides are partitioned among homologous proteins. The PROVALT algorithm, while designed to function in conjunction with Mascot, is independent of the scoring algorithm used for the peptide identification. This method reduces a complex list of peptide matches to a nonredundant list of proteins which have been identified at a user specified false discovery rate. Using peptide and protein identifications resulting from the proteomic analysis of *T. cruzi* epimastigotes, we compared the PROVALT method with the traditional methods of protein identification. The PROVALT method was shown to be superior to both methods by the number of proteins identified and in the inclusion of lower scoring non random peptide matches.

All software used in this study to parse Mascot results files, cluster peptides to proteins, and select proteins for identification at a specified false discovery rate is integrated into the software package PROVALT. This software will be made publicly available at http://kiwi.rcr.uga.edu/tcprot.

REFERENCES

1. Rep, M., Dekker, H., Vossen, J., de Boer, A., Houterman, P., Speijer, D., Back, J., de Koster, C., Cornelissen, B., 2002. Plant. Physiol. 130, 904-917.
2. Bendt, A., Burkovski, A., Schaffer, S., Bott, M., Farwick, M., Hermann, T. 2003. Proteomics. 3, 1637-1646.
3. Zhang, H., Li, X., Martin, D., Aeboersold, R. 2003. Nat. Biotechnol. 21, 660-666.
4. Winters, M., Day, R., 2003. J. Bacteriol. 185, 4268-4275.
5. Aebersold, R., Mann, M., 2003. Nature. 422, 198-207.
6. Taylor, J., Johnson, R., 1997. Rapid. Commun. Mass. Spectrom. 11, 1067–1075.
7. Taylor, J., Johnson, R., 2001. Anal. Chem. 73, 2594–2604.
8. Johnson, R., Taylor, J., 2002. Mol. Biotechnol. 22, 301–315.
9. Dancik, V., Addona, T., Clauser, K., Vath, J., Pevzner, P., 1999. J. Comput. Biol. 6, 327–342.
10. Chen, T., Kao, M., Tepel, M., Rush, J., Church, G., 2001 J. Comput. Biol. 8, 325–337.
11. Sunyaev, S., Liska, A., Golod, A., Shevchenko, A., 2003. Anal. Chem. 75, 1307–1315.
12. Mann, M., Wilm, M., 1994. Anal. Chem. 66, 4390-4399.
13. Tab, D., Saraf, A., Yates, J., 2003. Anal. Chem. 74, 5383-5392.
14. Perkins, D., Pappin, D., Creasy, D., Cottrell, J., 1999. Electrophoresis. 20, 3551-3567.
15. Eng, J., McCormack, A., Yates, J., 1994. J. Am. Soc. Mass. Spectrom. 5, 976-989.
16. Zhang, N., Aebersold, R., Schqikowski, B. 2002. Proteomics. 2, 1406-1412.
17. Colinge, J., Masselot, A., Giron, M., Dessingy, T., Magnin, J. 2003. Proteomics. 3, 1454-1463.
18. Bafna, V., Edwards, N. 2001. Bioinformatics. 17, S13-S21.

19. Sadygov, R., Yates. J., 2003. Anal. Chem. 75, 3792-3798.

20. Field, H., Fenyo, D., Beavis, R., 2002. Proteomics. 2, 36-47.

21. Hernandez, P., Gras, R., Frey, J., Appel, R., 2003. Proteomics. 6, 870-880.

22. Pappin, D., Hojrup, P., Bleasby, A., 1993. Curr. Biol. 3, 327-332.

23. Mawuenyega, K., Kaji, H., Yamuchi, Y., Shinkawa, T., Saito, H., Taoka, M., Takahashi, N., Isobe, T., 2003. J. Proteome. Res. 2, 23-35

24. Laukens, K., Deckers, P., Esmans, E., Van Onckelan, H., Witters, E., 2004. Proteomics. 4, 720-727.

25. O'Neil, K., Miller, F., Barder, T., Lubman, D., 2003. Proteomics. 3, 1256-1269.

26. Lee, C., Hsiao, H., Lin, C., Wu, S., Huang, S., Wu, C., Wang, A., Khoo, K., 2003. Proteomics. 3, 2472-2486.

27. Lasonder, E., Ishihama, Y., Andersen, J., Vermunt, A., Pain, A., Sauerwein, R., Eling, W., Hall, N., Waters, A., Stunnenberg, H., Mann, M., 2002. Nature. 419, 537-542.

28. Chamrad, D., Korting, G., Stuhler, K., Meyer, H., Klose, J., Bluggel, M. 2004. Proteomics. 4, 619-628.

29. Keller, A., Nesvizhskii, A., Kolker, E., Aebersold, R. 2002. Anal. Chem. 74, 5383-5392.

30. Fenyo, D., Beavis, R., 2003. Anal. Chem. 75, 768-774.

31. Petritis, K., Kangas, L., Ferguson, P., Anderson, G., Pasa-Tolic, L., Lipton, M., Auberry, K., Strittmatter, E., Shen, Y., Zhao, R., Smith, R., 2003. Anal. Chem. 75, 1039-1048.

32. Cargile, B., Bundy, J., Freeman, T., Stephenson, J., 2004. J. Proteome. Res. 3, 112-119.

33. Resing, K., Meyer-Arendt, K., Mendoza, A., Aveline-Wolf, L., Jonscher, K., Pierce, K., Old, W., Cheung, H., Russell, S., Wattawa, J., Goehle, G., Knight, R., Ahn, N., 2004. Anal. Chem. in press.

34. Lin, D., Tabb, D., Yates, J., 2003. Biochimica et Biophysica Acta. 1646, 1-10.

35. Nesvizhskii, A., Keller, A., Kolker, E., Aebersold, R., 2003. Anal. Chem. 75, 4646-4658.

36. Nesvizhskii, A., Aebersold, R. 2004. Drug. Discov. Today. 4, 173-181.

37. Sadygov, R., Liu, H., Yates, J., 2004. Anal. Chem. 76, 1664-1671.

38. Peng, J., Elias, J., Thoreen, C., Licklider, L., Gygi, S., 2003. J. Proteome. Res. 2, 43-50.

39. Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., Emili, A., 2003. Mol. Cell. Proteomics. 2, 96-106.

40. Benjamini, Y., Hochberg, Y., 1995. J. R. Stat. Soc. B. 64,479-498.

41. Storey, J., 2003. Ann. Statist. 31, 2013-2035.

42. Storey, J., Tibshirani, R., 2003. PNAS. 16, 9440-9445.

43. Rondinelli, E., Silva, R., Carvalho, J., de Almeida Soares, C., de Carvalho, E., de Castro F., 1988. Exp. Parasitol. 66, 197-204.

44. Sonsmann, G., Romer, A., Schomburg, D., 2002. J. Am. Soc. Mass. Spectrom. 12, 47-58.

45. Rappsilber, J., Mann, M., 2002. Trends. Biochem. Sci. 27, 74-78.

Table 3.1

Table I: Peptide Fractions and RP-LC

Fraction Number	1st Dimension RP-LC Percentage RP-B[a]	3rd Dimension RP-LC Percentage RP-B[b]
1	5 - 10%	2 - 12%
2	10 - 15%	6 - 17%
3	15 - 20%	11 - 22%
4	20 - 23%	16 - 25%
5	23 - 26%	19 - 28%
6	26 - 29%	22 - 31%
7	29 - 32%	25 - 34%
8	32 - 35%	28 - 37%
9	35 - 40%	31 - 42%
10	40 - 45%	36 - 47%

[a:] Range RP-B at which each fraction was collected during the 1st dimension reverse phase chromatography.

[b:] For the RP-LC-MS/MS analysis each fraction was subjected to a one hour linear gradient over the cooresponding range of RP-B.

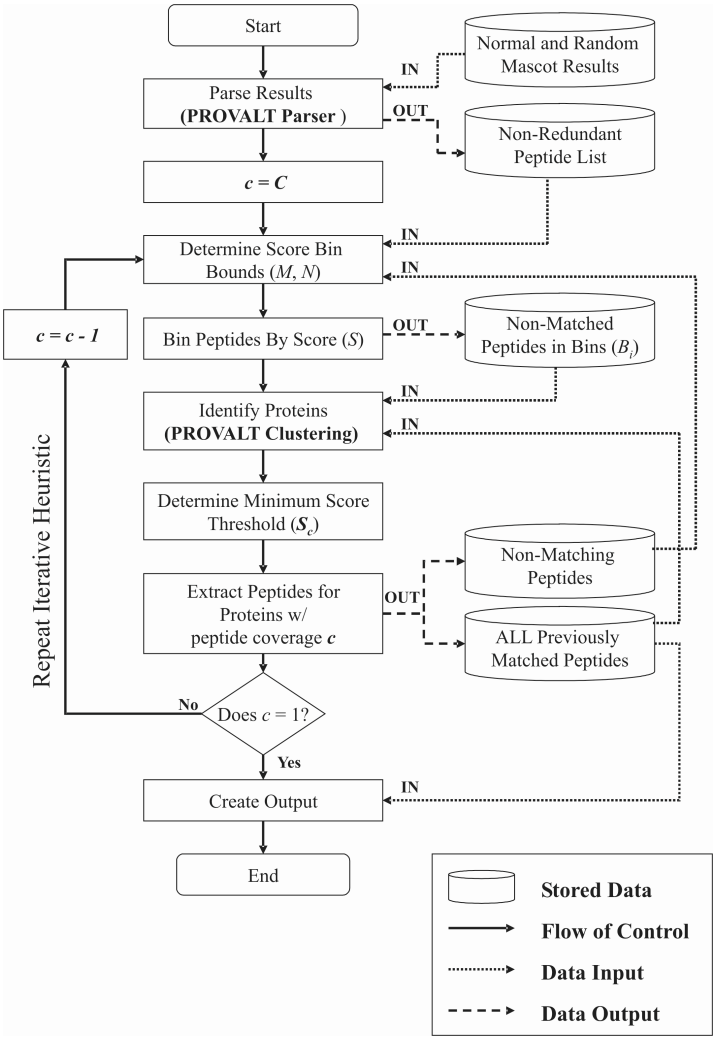

Figure 3.1 - The PROVALT Method for Protein Identification and Statistical Validation

Protein identification begins with the extraction of peptide matches from both normal and random Mascot database search results. Peptides are binned as a function of ion score and clustered to proteins to yield a minimal protein list. Proteins are selected based on degree of peptide coverage. The number of identified proteins between the random and normal database searches is compared as a function of peptide ion score to determine the minimal ion score and peptide coverage level necessary to achieve a 1% *PRO-FDR*. Matching proteins and peptides are output, and the process is iteratively repeated for each coverage level.

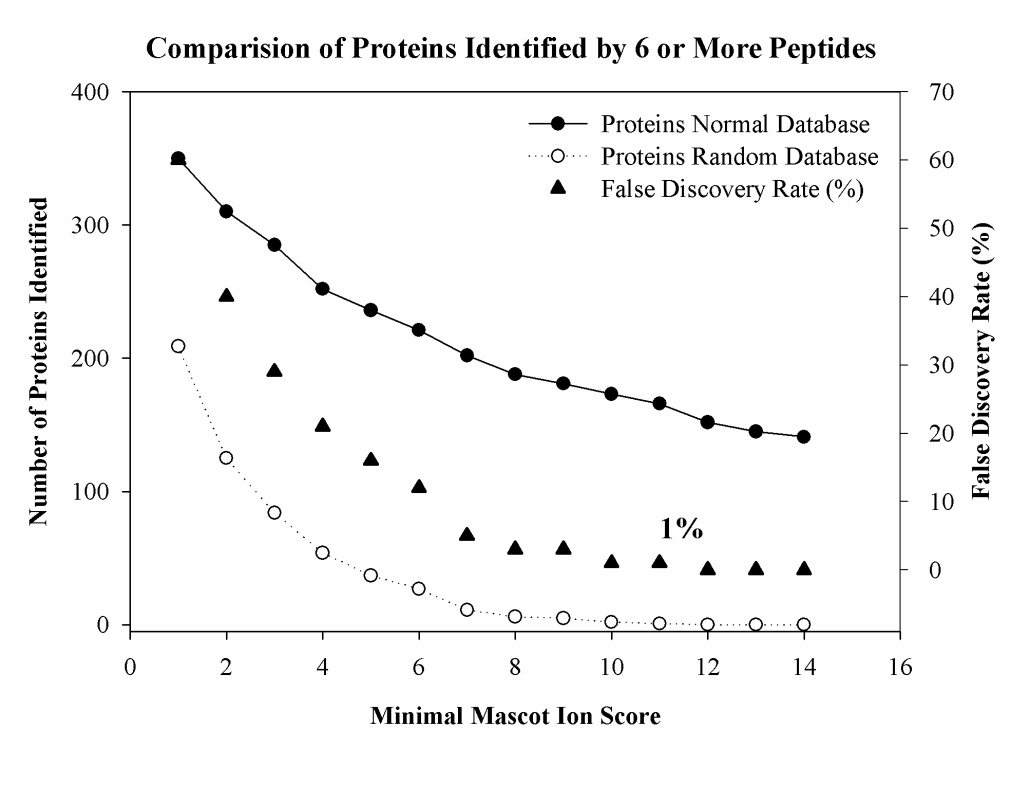

Figure 3.2 - Distribution of False Discovery Rates as a Function of Minimal Ion Score for Proteins Identified by 6 or More Peptides

PROVALT constructs a histogram which compares the number of proteins identified in both the normal and random database searches. This comparison is iteratively performed for each peptide coverage level. The above histogram compares the frequency of protein identifications for proteins containing at least 6 or more peptides at or above specific Mascot ion scores. Using equation 3 the expected percentage of false discoveries is calculated for each peptide bin, and a 1% false discovery rate is achieved if a protein is identified by 6 or more peptides with at least a minimal Mascot ion score of 11.

Table 3.2

Table II: PROVALT Calculated Ion Score Thresholds

Peptide Coverage[a]	Minimum ion score[b]
1	44
2	35
3	27
4	18
5	14
6	11

[a:] Minimum number of peptides which match to a protein following PROVALT clustering.

[b:] Minimum Mascot reported ion score for the peptide assignment to acheive a 1% protein false discovery rate.

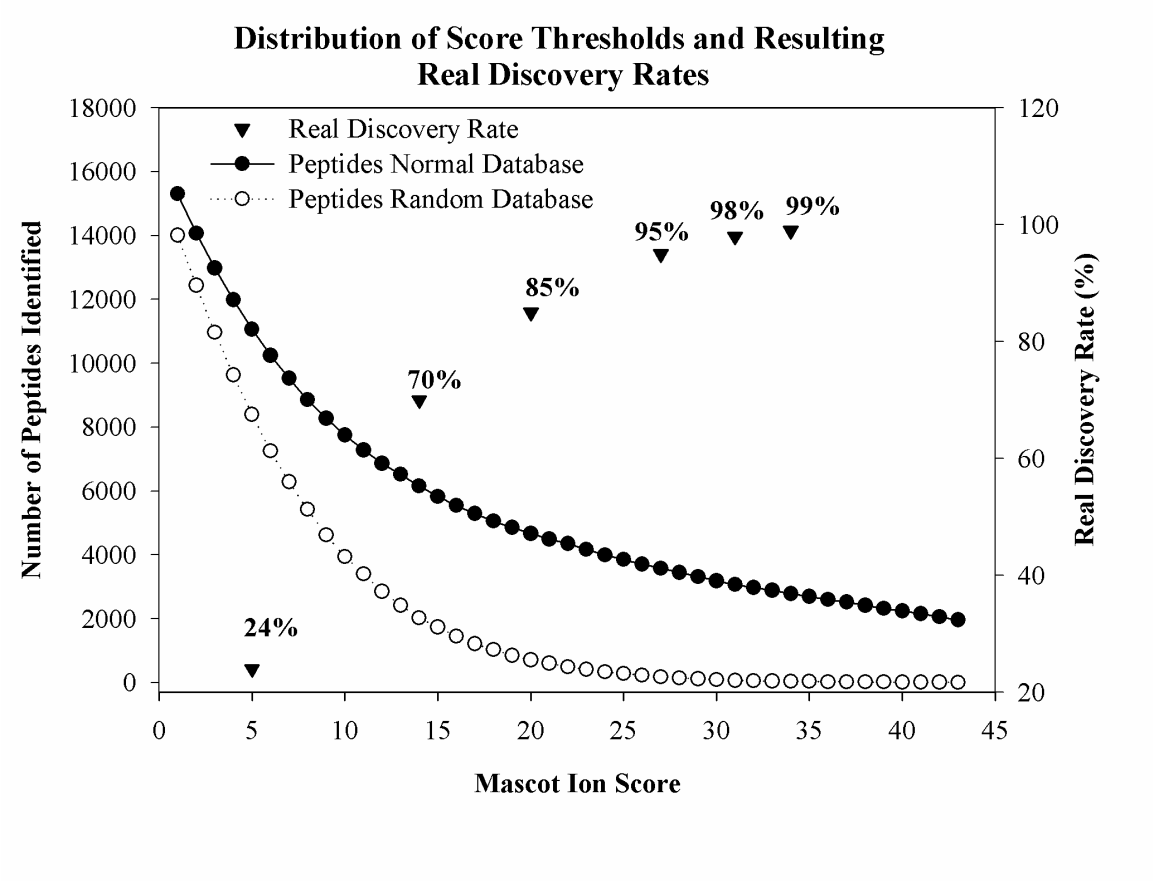

Figure 3.3 - Real Discovery Rates as a Function of Single Ion Score Thresholds

The frequency of peptide matches at or above Mascot ion score thresholds were plotted for both the random and normal database searches. Equation 5 was applied to calculate the real discovery rate (100% - *PEP-FDR*) for each threshold. To achieve a 1% *PEP-FDR* only peptides exceeding ion scores of the 34 would be selected to identify proteins. However, real matches may occur at lower scores.

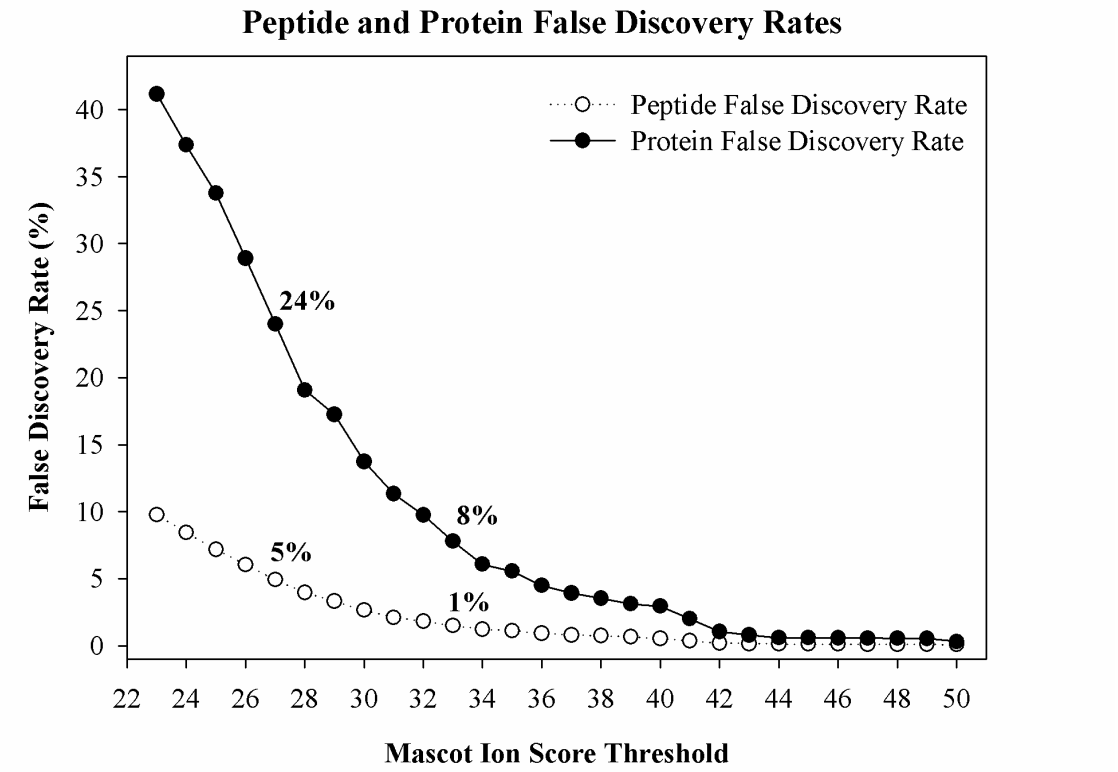

Figure 3.4 - Protein and Peptide False Discovery Rates as a Function of Mascot Ion Score Threshold

Using the single peptide score threshold method peptides were selected for protein identification if they exceeded specific Mascot ion scores. This method was applied at each Mascot ion score threshold and the resulting peptide and protein false discovery rates were plotted. From the figure above it is evident that a false discovery rate associated with peptide matches does not correspond to a false discovery rate for protein identifications.

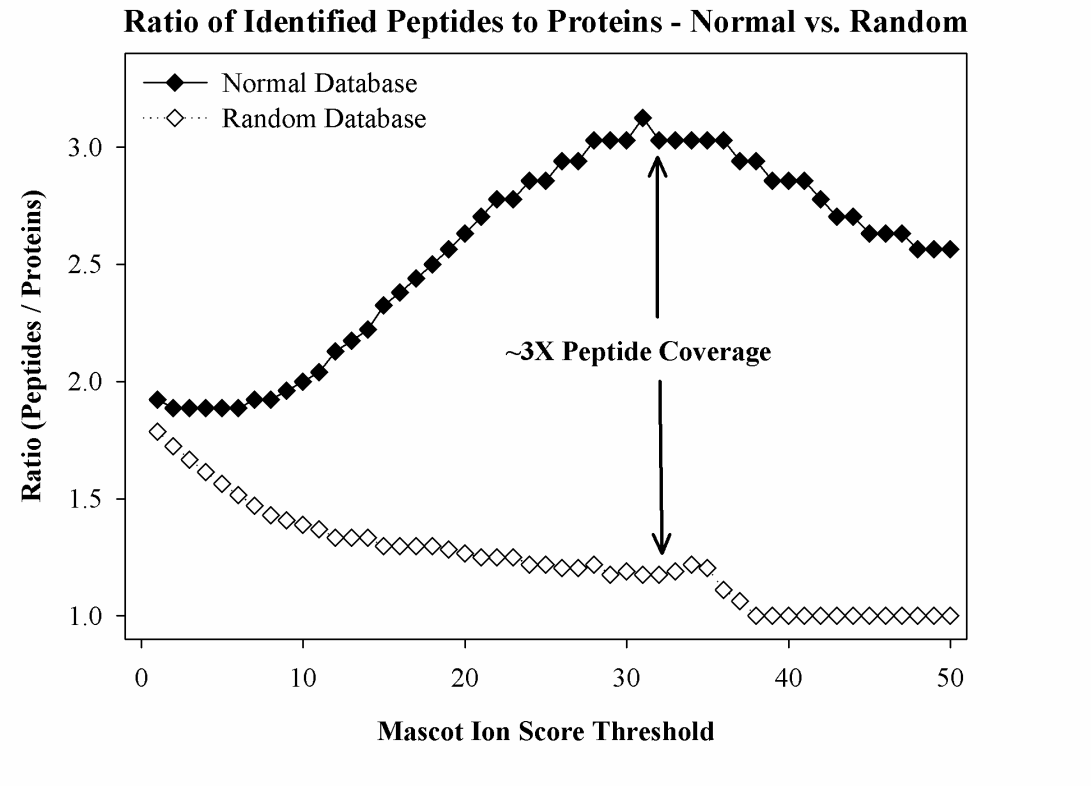

Figure 3.5 - Ratio of Peptides to Proteins Identified at a Distribution of Ion Score Thresholds

Proteins were identified by peptides at a distribution of peptide ion score thresholds in both the normal and random database searches. The ratio of peptides to proteins was calculated and plotted as a function of ion score. The ratio of peptides to proteins is 3X higher in the normal database over the random database indicating that random peptide matches do not cluster to proteins as do real peptide matches.

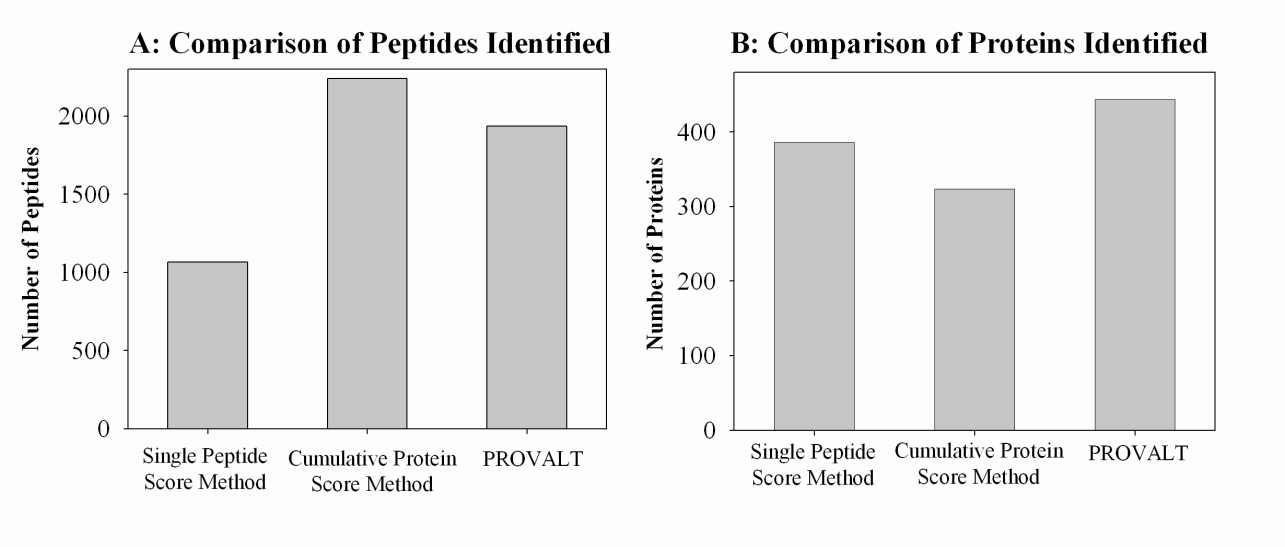

Figure 3.6A and 3.6B - Number of Proteins and Peptides Identified by PROVALT, SPM, and CPM

Using the PROVALT method of protein identification (1% *PRO-FDR*) the number of proteins identified was increased an average of 25% above the number identified by SPM and CPM. The PROVALT method also allowed the 83% more peptides than was identified by the SPM at a 1% *PRO-FDR*.

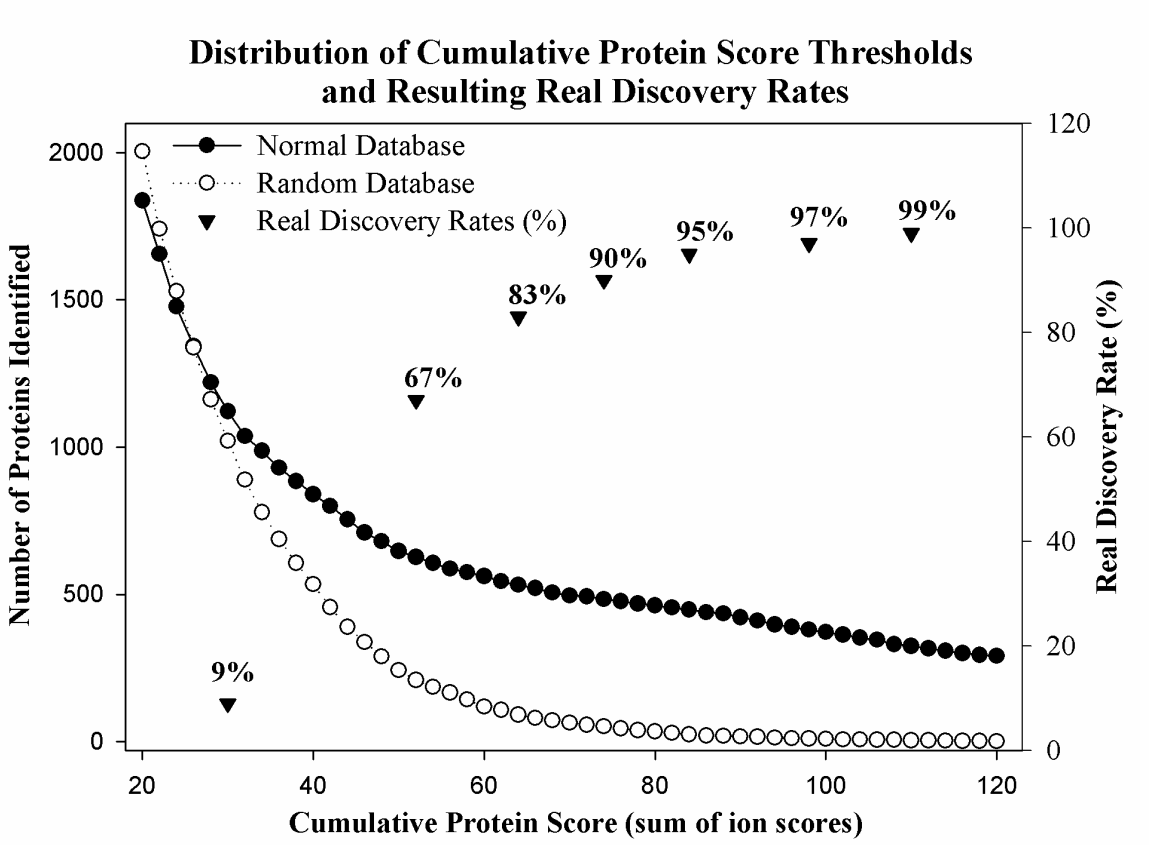

Figure 3.7 - Cumulative Score Method and Resulting Real Discovery Rates

Proteins were identified using the CSM at a distribution of cumulative protein score thresholds for both the normal and random database searches. The number of proteins identifications for each database was plotted along with the resulting real discovery rate at each cumulative protein score threshold. When the CSM is applied using the cumulative protein scores above, a cumulative score of 84 would be required to achieve a 5% false discovery rate and a large number of potentially real protein matches would be missed.

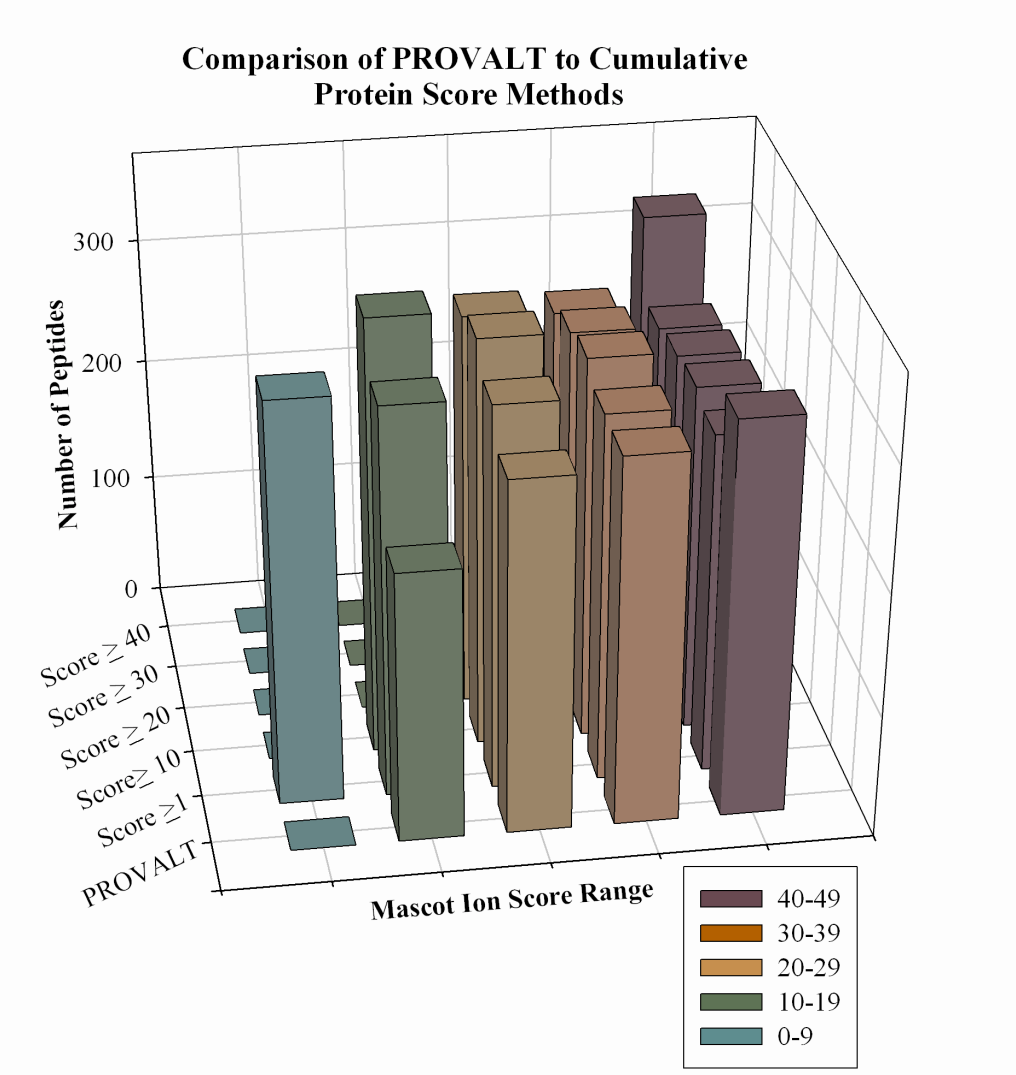

Figure 3.8 - Comparison of Peptides used in Protein Identifications from Cumulative Score Methods and the PROVALT Method

The cumulative protein score method (1% *PRO-FDR*) was employed after filtering peptides below given ion scores. The peptides used to identify proteins from each ion score filter were grouped into bins according to ion score. The distribution of peptide scores was then compared between each cumulative score method and the PROVALT (1% *PRO-FDR*) method. From the graph it is evident that the PROVALT method is more selective in the inclusion of lower scoring peptides.

CHAPTER 4

THE *TRYPANOSOMA CRUZI* PROTEOME[1]

[1] Atwood, J., Weatherly, D., Bundy, B., Minning, T., Cavola, C., Opperdoes, F., Orlando, R., Tarleton, R. 2005. Accepted by *Science*.

Reprinted here with permission of publisher.

ABSTRACT

To complement the sequencing of the 3 kinetoplastid genomes reported in this issue we have undertaken a whole-organism, proteomic analysis of the four life-cycle stages of *Trypanosoma cruzi*, the causative agent of Chagas disease. Peptides mapping to 2784 proteins in 1168 protein groups from the annotated *T. cruzi* genome were identified across the four life-cycle stages. Evidence is presented for the expression of protein products from >1000 genes currently annotated as "hypothetical" in the sequenced genome, including members of a newly defined gene family annotated as mucin-associated surface proteins (MASPs). This analysis reveals the apparent utilization of distinct energy sources by the individual parasite stages, including histidine for stages present in the insect vectors and fatty acids by intracellular amastigotes. Furthermore, by screening open reading frames (ORFs) from all accessible *T. cruzi* sequences, we have used the proteome data to identify > 70 genes or alleles missing from the annotated genome as well as modifications to a number of annotated genes.

INTRODUCTION

Trypanosoma cruzi exists in four morphologically and biologically distinct forms during its cycle of development in mammals and insects (Fig 4.1). Metacyclic trypomastigotes develop in the hind gut of triatomine insect vectors and initiate infection in a wide variety of animal species, including humans. Infection of host cells induces the conversion of trypomastigotes to replicative amastigote forms which reside in the host cell cytoplasm. Following multiple rounds of binary fission, the aflagellate amastigotes convert into flagellated trypomastigotes that burst from the host cell and circulate in the bloodstream. There the trypomastigotes can invade other host cells and thus spread the infection throughout the body. Alternatively, trypomastigotes acquired during a blood meal convert to epimastigote forms, which replicate in the insect gut before eventually differentiating into infective metacyclic trypomastigote forms. Drugs for the treatment of *T. cruzi* infection are inadequate and vaccines are lacking. Whole proteome analysis is a robust, high-throughput method to identify targets for vaccine and drug design. Like other trypanosomatids, *T. cruzi* appears to regulate protein expression primarily post-transcriptionally via variations in mRNA stability and/or the translational efficiency of mRNAs (1). This limits the utility of gene expression profiling analysis for monitoring stage-dependent changes in gene expression (2-5) and makes proteomic analysis especially attractive for examining global changes in protein expression during development in *T. cruzi*.

EXPERIMENTAL PROCEDURES

Parasites

Brazil strain *T. cruzi* trypomastigotes were grown in monolayers of Vero cells (ATCC no.

CCL-81) in RPMI supplemented with 5% horse serum as previously described (6). Emergent trypomastigotes were harvested daily and examined by light microscopy to determine the percentages of amastigotes and trypomastigotes. Only preparations containing > 95% trypomastigotes were used in the subsequent studies. Amastigotes (>95% pure) were prepared axenically from low pH-induced trypomastigotes as described previously (7). Amastigotes generated by this method are well-documented to be indistinguishable from intracellular amastigotes and have been widely used to study amastigote biology. However, it is possible that changes noted in the proteome of these artificially derived amastigotes might differ from that of amastigotes obtained from infected host cells. Epimastigotes were grown in Liver Infusion Tryptose media (LIT) as previously described (8). Cultures were harvested during mid-log phase by centrifugation at 3,000 X g for 10 m at room temperature. Metacyclic trypomastigotes were obtained from epimastigotes by axenic induction as previously described (9). The percentages of metacyclics were determined by microscopic examination of parasites stained with Dif-Quick (Baxter Diagnostics, McGaw Park, IL). Metacyclic trypomastigotes were purified from the resulting cultures by DEAE-Sephacel chromatography as described previously (10).

Whole and sub-cellular protein isolation

Proteins were isolated from 1-3 x 10^9 organisms/preparation using Tri-Reagent (Sigma). Whole-cell lysates from three epimastigote, two metacyclic trypomastigote, two amastigote, and two trypomastigote biological replicate preparations were generated and analyzed separately. In some cases, crude cytoplasmic and membrane preparations were obtained from amastigotes and trypomastigote by lysis in ice-cold lysis buffer (150 mM NaCl, 1.5 mM $MgCl_2$, 0.5% (v/v) NP-40, and 10 mM Tris-HCl, pH 8.0). Nuclei were removed by centrifugation at 2,000 X g at 4^{o}C

for 2 min., and membrane fractions were harvested from the post-nuclear supernatants by centrifugation at 12,000 X g at 4^{o}C for 5 min. Cytoplasmic fractions were obtained from the resulting post-membrane supernatant, and proteins were then isolated from each lysate with Tri-Reagent. Isolated proteins from each of the lysates were independently reduced, carboxyamidomethylated, digested with endoproteinase Lys-C, and digested with trypsin as previously described .

Peptide separation and MS/MS analysis

Peptide mixtures generated from the whole-cell lysates of *T. cruzi* epimastigotes, metacyclic trypomastigotes, amastigotes, and trypomastigotes were independently separated and analyzed as previously described (11). The strong cation exchange separation step was omitted in the preparation of the four sub-cellular lysates.

Protein sequence databases

Four sequence databases were constructed for these analyses. A representative database (normal) consisting of 23,095 *Trypanosoma cruzi* gene annotations provided by *Trypanosoma cruzi* Sequencing Consortium (TSK-TSC; version 3) was employed for the final protein identifications. A randomized database (random), which was created by inverting the sequences in the normal database, was used to establish accurate scoring thresholds for protein identification in the normal database. For removal of contaminant peptides, two databases were used: a composite database was created by combining the TSK-TSC database with 10,468 protein sequences from *Bos Taurus*, *Equus caballus*, *Homo sapien*, and proteases from the National Center for Biotechnology Information (NCBI) and a second database containing the TSK-TSC proteins plus 298,912 primate protein sequences from NCBI. The former database

was used to remove contaminant spectra matching peptides that may have been introduced during sample preparation and the latter to identify potential contaminants arising from the cultivation of *T. cruzi* in Vero green monkey kidney cells. Lastly, for the ORF analysis, a database of 817,000 open reading frames of at least 50 amino acids was constructed from a number of sequence sources. These include 48 large, partially assembled contigs from *T. cruzi* obtained from NCBI, consensus sequences constructed from raw sequence reads obtained from the TSK-TSC prior to the assembly of the *T. cruzi* genome, and the contigs used for the gene predictions made by the TSK-TSC. Unique peptides identified by spectra that failed to match proteins predicted by the annotated genome were clustered to the ORFs, and the new proteins were annotated and the spectra matching the new gene were manually verified. ORFan proteins identified by this method were annotated using BLAST homologies (GenBank NR and the *T. brucei*, *L. major*, and *T. cruzi* annotated genomes), protein domains (InterPro (12)), signal peptide motifs (SignalP, (13)), and GPI anchor addition motifs (http://129.194.185.165/dgpi/index_en.html, DGPI, Kronegg 1999).

Data processing and analysis

Peak-lists were first filtered to remove spectra originating from singly charged ions with parent ion masses <600 Da, and the remaining spectra were then submitted for database searching with Mascot (Matrix Science, Boston, MA). Mascot searches were limited to fully tryptic peptides to restrict the number of candidate peptides from the database which could match to each spectrum. Parent and fragment mass errors from identified peptides with Mascot scores exceeding 60 were used to perform linear recalibration of all spectra as previously described (14). Following recalibration, peak-lists were distributed into 4 bins as a function of maximum

parent mass error (50, 100, 150, 200 p.p.m.). Spectra within each bin were searched with Mascot using the following parameters: enzymatic cleavage with trypsin, 1 potential missed cleavage, peptide tolerances of 50, 100, 150 and 200 ppm, fragment ion tolerance of 0.2 Da, and variable modifications due to carbamylation (+ 43 Da) and carboxyamidomethylation (+ 57 Da). Spectra matching contaminant peptides were removed from the peak-lists and the database search was repeated against the normal and random databases.

Protein identification and validation

PROVALT (11) uses the confidence in individual peptide matches along with the number of peptides that match to proteins to identify high-confidence proteins in a high-throughput manner. To this end, PROVALT extracts peptide matches and corresponding ion scores from the normal and random Mascot results files and filters them to create a non-redundant list of peptides. The peptides in each list are then binned according to score, where each bin contains all peptides at or above the Mascot ion score that it represents. The peptides in the normal and random bins are then clustered to the proteins in their corresponding sequence databases. In cases where a peptide or a set of peptides map to more than one protein, and thus cannot be uniquely assigned to an individual protein, PROVALT groups proteins into "protein groups". In order to select protein groups with a false-discovery rate of <1%, PROVALT iteratively determines the score bin in the random database for which the number of identified protein groups meeting the specified minimum peptide coverage is <1% of that of the corresponding bin in the normal database. The peptide coverage value is decreased for each iteration. For this work, the peptide coverage levels and minimum score thresholds were as follows: 6 (or more) peptides with score > 14, 5 peptides with score >17, 4 peptides with score > 22, 3 peptides with

score > 28, 2 peptides with score > 35, and 1 peptide with score > 43.

Data access

Data on all peptides mapping to annotated genes is available on TcruziDB (http://tcruzidb.org). Raw data in either the original peak-list (.PKL) format or in mzData XML format (MIAPE standard) can be downloaded from (http://kiwi.rcr.uga.edu/tcprot/downloads.html). Complete lists of all peptides identified, pre-run queries identifying proteins expressed in specific life-cycle stages as well as tools to query and view these data are also available on (http://tcruzidb.org) and or (http://kiwi.rcr.uga.edu/tcprot/downloads.html).

RESULTS AND DISCUSSION

Metacyclic trypomastigotes, amastigotes, trypomastigotes and epimastigotes of *T. cruzi* were isolated, and proteins were extracted from whole-cell or sub-cellular lysates (Fig 4.2). Peptides generated by digestion of the whole-cell or sub-cellular lysates were independently separated and analyzed at least in duplicate by offline multidimensional liquid chromatography, online reverse phase liquid chromatography and tandem mass spectrometry (LC-MS/MS). A total of 602 tryptic peptide samples were analyzed, generating 139,147 tandem mass spectra. The breakdown in the number of spectra collected for each life-cycle stage is shown in Table 4.1 and demonstrates that, because of differences in protein recovery from the different life-cycle stages, trypomastigote and amastigote stages are under-sampled relative to metacyclic trypomastigotes and epimastigotes. Therefore conclusions on the stage restricted expression of proteins in over-sampled stages, and likewise their absence in under-sampled stages, should be considered

provisional.

To generate accurate representations of the *T. cruzi* proteomes, a number of steps were taken to minimize the possibility of incorrectly matching spectra to peptides in the sequence database (see Materials and Methods). The Mascot search engine was used to match spectra to peptides in a protein database composed of 23,095 *T. cruzi* gene predictions provided by the *T. cruzi* Sequencing Consortium (TSK-TSC; version 3). Peptides judged by Mascot as the best-fit for each spectrum were pooled for all fractions and then matched to proteins via the PROVALT parsing and clustering tool (*6*). In cases where a peptide or a set of peptides map to more than one protein, PROVALT groups proteins into "Protein Groups". In order to select only high-confidence proteins in a high-throughput manner, a randomized database was created by inverting the *T. cruzi* sequences, and the results obtained from the "normal" and reversed databases were compared. PROVALT was then used to identify cutoffs for minimum Mascot peptide scores and minimum peptide coverage (number of peptides) to yield a <1% protein group false-discovery rate.

Ultimately, a total of 5,720 unique peptides were matched with high confidence to 1168 protein groups containing 2784 total proteins (Table S2). The "protein groups" and "proteins" designations are necessary because in many cases it is not possible to uniquely assign individual or groups of peptides to a single protein. This approach is used to group protein isoforms (16, 17) and is particularly important in the case of *T. cruzi* because the genome contains multiple, non-identical copies of many genes, including a number of large gene families with hundreds of distinct members (Andersson, et. al. *T. cruzi* genome - this issue). In addition, the *T. cruzi* CL Brener strain used for the sequencing project is a hybrid of two genotypes and thus has multiple

distinct alleles for most genes. The 2784 proteins and 1168 protein groups represent the upper and lower limits, respectively, for the number of proteins confirmed to be expressed based upon this analysis. Slightly less than 25% (290) of the protein groups were identified by a single peptide match using a minimum Mascot score of 43 (corresponding to a peptide false-discovery rate of 0.09% based on comparison to the random database). Table 4.1 summarizes the proteins assigned to each life-cycle stage. Nearly 30% (838 of 2784) of the identified proteins, including most of the proteins previously documented or expected to be produced in the greatest abundance, were detected in all life-cycle stages. Shotgun proteome LC-MS/MS analysis as conducted herein does not detect changes in protein expression levels with the same precision as is possible using stable isotope labelling techniques. Nevertheless, it provides empirical evidence of protein expression and allows for high-throughput comparison of protein detection among the four life-cycle stages of *T. cruzi*, something that cannot be accomplished with current quantitative technologies. As others have done (15), we employed measures of peptide coverage, including total protein score, to indicate the relative abundance of proteins in the *T. cruzi* proteomes and to track relative changes in protein expression in the individual life-cycle stages. This approach provides provisional evidence for the relative abundance and the presence or absence of a particular protein in any given stage. For the relatively small subset of proteins in *T. cruzi* with known expression patterns (18-20), our results agree in virtually all cases.

Among the top scoring proteins in all four *T. cruzi* proteomes are many housekeeping proteins that are also among the highest ranking proteins in the best characterized eukaryotic proteome, yeast (21). However, many other highly abundant proteins in the *T. cruzi* proteome are either absent in the yeast genome (e.g. paraflagellar rod protein 3, 8152.t00002; flagellar

calcium-binding protein, 5387.t00002; I/6 autoantigen, putative, 7685.t00010; and 14-3-3 protein-like, 8730.t00013) or are expressed at very different relative levels in these two eukaryotes (e.g. d-isomer specific 2-hydroxyacid dehydrogenase-protein, 8304.t00012; malic enzyme, 7814.t00028; and alpha tubulin, 11788.t00001).

Table 4.2 summarizes some of the major protein groups and families identified in the *T. cruzi* proteome. These data reflect a combination of the relative abundance of the proteins comprising each group, the size of gene families and the ease with which certain proteins can be detected by LC-MS/MS analysis. The well-characterized trans-sialidase (ts) and mucin families highlight two limitations of shotgun proteomics: the lack of resolution due to shared peptides and the under-representation of highly glycosylated proteins. Peptides matching to 223 members of the ts family clustered into 50 protein groups were detected in one or more stages. In contrast, no peptides mapping to mucin family proteins were identified, presumably due to the high level of mucin glycosylation (22). Of the 2784 total proteins identified in this analysis, 1008 are from genes annotated as "hypothetical". Thus this analysis provides the first data validating these as *bona fide* genes in *T. cruzi*. Furthermore, over half of these hypothetical genes have orthologs in the *Leishmania major* and/or *Trypanosoma brucei* genomes.

The trypomastigote proteome and the expression of large gene families

T. cruzi trypomastigotes circulate in the blood where they are exposed to host immune effector molecules, including specific antibodies. Unlike the related African trypanosomes, *T. cruzi* trypomastigotes do not undergo antigenic variation but instead express on their surface multiple members of several large families of molecules; the best characterized of these are the mucins and trans-sialidases (ts) (23). Thirty of the 50 top-scoring proteins detected exclusively

in trypomastigotes are ts family members. Likewise, the amastigote and metacyclic stages appear to express subsets of ts molecules unique to each stage while no ts expression was detected in the epimastigote proteome (Fig 4.3). Trans-sialidase enzymatic activity is reportedly present in only a small subset of the >1000 ts proteins encoded in the *T. cruzi* genome and has been linked to the presence of Tyr342 in the catalytic N-terminal region and SAPA repeats in the c-terminus (23). As expected, peptides from the 15 ts genes with the highest sequence identity to the defined enzyme-active ts molecules, all of which have the Tyr342 and 6 of which have SAPA repeats, are represented in the total of 223 ts proteins detected in the *T. cruzi* proteomes. However, the production of a large number of non-enzymatic ts family members coincident with these ts enzymes may deflect immune responses away from the relatively few enzymatically active targets and/or may provide a pool of altered peptides that could antagonize T cell responses (24). The clear stage-specific production of subsets of ts genes within the three *T. cruzi* stages present in mammals has not been previously documented on a global level. Moreover, little is known about the biological significance of this restricted expression or the specific mechanism(s) by which it is achieved.

In addition to the ts and mucin families, the *T. cruzi* genome contains several other high copy, multi-gene families (Table 4.2). Herein we provide the first evidence for the expression of several mucin-associated surface proteins (MASP), a gene family first discovered as part of the sequencing and annotation effort (Andersson, et. al. *T. cruzi* genome - this issue). Members of this family were detected predominantly in the trypomastigote proteome and not in the amastigote and metacyclic trypomastigote proteomes. Like proteins from the other multi-gene families in *T. cruzi*, many MASP family members have predicted signal sequences and GPI

anchor addition sites and thus are likely to be surface-expressed. Nine MASP gene family proteins were identified in our analysis, each by only a single peptide match. This result suggests that either MASPs are not as abundantly expressed as the trans-sialidase proteins, or that, like the mucins, MASPs have extensive post-translational modifications which complicate their detection by shotgun proteomics. However, detection of the MASPs in the relatively under-sampled trypomastigote stage suggests that they are not minor constituents of the *T. cruzi* proteome.

Additional gene family members detected in the *T. cruzi* proteomes include those from the cysteine protease (detected in all life-cycle stages except trypomastigotes) and gp63 (detected in all stages except the amastigotes) families. Genes encoding the retrotransposon hot spot (RHS) proteins are plentiful in the *T. cruzi* and *T. brucei* genomes and were first identified in the latter as potential sites for insertions of retrotransposons. Although the function of the proteins encoded by RHS genes is not known, they were found to be constitutively expressed in *T. brucei* and to localize primarily to the nucleus (25). Here we show that the RHS proteins are expressed in *T. cruzi* from multiple loci and in all developmental stages (Fig 4.3). The RHS proteins are detected most prominently in the metacyclic forms, but this could be due to the greater overall sampling of this stage in our analysis.

Amastigote Proteome

Upon host cell invasion, trypomastigotes transform into aflagellate amastigotes, which replicate by binary fission in the host cell cytoplasm. Because the amastigote forms are localized intracellularly, relatively few studies have investigated the biology of this developmental stage. However, the transition from trypomastigote to amastigote can be stimulated extracelluarly by

simulating the low pH environment of the phagosomal/lysosomal compartment that *T. cruzi* initially encounters upon cell entry (26). This makes early time points in the transformation process to the amastigote stage amenable to transcriptome and proteome analysis. Amastigotes generated by this method are well-documented to be indistinguishable from intracellular amastigotes and have been widely used to study amastigote biology. However, it is possible that changes noted in the proteome of these artificially derived amastigotes might differ from that of amastigotes obtained from infected host cells. The results from this proteome analysis of amastigotes are in agreement (with one exception) with the restricted data set generated by comparison of trypomastigotes and early stage amastigotes using DNA microarray analysis (3)further supporting the quality of this analysis (Table S4).

In addition to the expression of a distinct subset of trans-sialidase-family genes, many of which are related to the amastigote surface protein 2 molecule previously reported to be preferentially expressed in amastigotes (27) (Fig 4.2), the transition of trypomastigotes to amastigotes also appears to be accompanied by a dramatic shift from carbohydrate to lipid dependent energy metabolism (Table S3). This conclusion is supported by the virtual absence of glucose transporters and the exclusive detection of enzymes for fatty acid beta-oxidation in amastigotes . The end-product of this beta-oxidation, acetyl coenzyme A, is further oxidized to carbon dioxide and water by the citric acid cycle for which most of the enzymes are abundant in amastigotes. Amastigotes are also likely to be dependent on gluconeogenesis for the synthesis of glycoproteins and glycoinositolphospholipids (GIPLs). Aspartate aminotransferases (4698.t00001, 4779.t00007) detected exclusively in amastigotes may be important in this

process. These proteins lack the mitochondrial targeting signal present on the aspartate aminotransferase expressed in all stages (6015.t00007) and thus likely reside in the cytoplasm. Mitochondrially produced oxaloacetate, after transamination, may be transported to the cytosol via an malate/aspartate shuttle and then converted by the cytosolic aspartate aminotransferase and a phosphoenol pyruvate carboxykinase into phosphoenol pyruvate, the substrate for gluconeogenesis.

In addition to several heat shock proteins and kinases, among the other proteins detected preferentially or exclusively in amastigotes are a group involved in ER to Golgi trafficking, including rab1 (4703.t00005), sec23 (8726.t00010), and sec31 (6890.t00029). The detection of this set of proteins involved in vesicular trafficking in amastigotes but not in the more highly sampled metacyclic and epimastigote stages suggests a more active trafficking process or the preferential use of selected rab and sec proteins in amastigotes (Table S3). We also extend the data on the selective expression in amastigotes and epimastigotes of several ABC transporters (7164.t00003, 8319.t00008) that are hypothesized to have a role in cargo selection and/or vesicular transport in trypanosomes (28). A putative lectin (6865.t00003) with homology to ERGIC, a protein involved in cargo selection in COPII vesicles, is also detected in trypomastigotes and amastigotes but not in metacyclic or epimastigote forms.

Insect Stages: Epimastigote and Metacyclic Trypomastigote Proteomes

Epimastigotes of *T. cruzi* undergo rapid expansion in the insect midgut before transformation into metacyclic trypomastigotes, the infective stage for mammals. In contrast to both *T. brucei* and *L. major*, the *T. cruzi* genome encodes enzymes capable of catalyzing the conversion of histidine to glutamate. The first two enzymes in this pathway, histidine ammonia-

lyase (6869.t00022) and urocanate hydratase (4881.t00011), are abundant in the insect stages but nearly undetectable in the mammalian stages (only a single spectrum matching histidine ammonia-lyase in amastigotes), consistent with the function of this pathway primarily in epimastigotes and metacyclic trypomastigotes. This expression pattern is interesting, given that histidine is the dominant free amino acid in both the excreta and hemolymph of *Rhodnius prolixus* (29, 30), a well-studied vector for *T. cruzi*. The abundance of histidine in this and other blood-feeding insects likely reflects the high histidine content of hemoglobin (31). Thus, *T. cruzi* epimastigotes seem uniquely adapted among the kinetoplastids to take advantage of this plentiful energy source in the gut of its insect vector. This is analogous to the use of proline as an energy source by *T. brucei* (32). To our knowledge, this is the first evidence for the use of histidine as an energy source in *T. cruzi*.

The transformation of epimastigotes to metacyclic trypomastigotes is accompanied by the production of a number of key enzymes and substrates important in antioxidant defense in *T. cruzi*. The H_2O_2 and peroxynitrite detoxifying enzymes ascorbate peroxidase (6846.t00006, 4731.t00003) (33) and the mitochondria-localized tryparedoxin peroxidase (8115.t00003) are both elevated following epimastigote to metacyclic conversion, as are tryparedoxin (5824.t00003), the substrate for tryparedoxin peroxidase, and the enzymes trypanothione synthase (8070.t00009, 7998.t00005) and iron superoxide dismutase (5781.t00004), responsible for synthesis of trypanothione and for the conversion of superoxide anion to hydrogen peroxides, respectively. These changes are consistent with a pre-adaptation of metacyclic forms to withstand the potential respiratory burst of phagocytic cells in the mammalian host. Enzymes of the pentose-phosphate shunt aid this process through the production of the NADPH required for

the reduction of trypanothione. Also noticeable in the transition of epimastigotes into metacyclic trypomastigotes is the substantial decrease in the representation of ribosomal proteins in the metacyclic proteome; 37 of the 50 highest scoring proteins in the epimastigote proteome that are not detected in this sampling of the metacyclic trypomastigote proteome are ribosomal proteins. A reduction in the capacity for protein production would be consistent with the stationary, non-replicating status of metacyclic trypomastigotes. Interestingly, DNA microarray analysis has also documented a substantial down-regulation of ribosomal protein expression in metacyclic forms in *L. major* (34).

Protein modifications

Protein function in cells is regulated not only by abundance but also by post-translational modifications, which can alter the localization, molecular interactions and activity of proteins. While complex and variable modifications such as glycosylation are difficult to detect by shotgun MS/MS analysis, modifications with static mass shifts such as acetylation, methylation and phosphorylation are readily detectable (35). In addition, deamidation, which can occur either as a post-translational modification or during sample preparation, can be monitored in high-throughput proteomics. Using the Mascot search engine, we matched spectra to peptides containing this limited set of modifications (Table S5). This search resulted in 8 new protein identifications and the detection of modifications on 81 previously identified proteins. Among those proteins with detected acetylations, methylations or phosphorylations are histones and histone-associated proteins, proteins involved in cytoskeletal structures, including alpha tubulin, beta tubulin, and actin, ribosomal proteins, elongation and initiation factors, heat shock proteins and metabolic enzymes, as well as a number of hypothetical proteins. In several cases, proteins detected with modifications in the *T. cruzi* proteome had previously been reported as similarly

modified in other species. However, only a few cases of such protein modifications have been documented in *T. cruzi*. For example, the triple methylation on the H3 histone has been documented at the same site in mouse histone H3 (36) but had not been reported in trypanosomes. In the case of elongation factor 1-alpha (eEF1A), these modifications were clearly stage-specific. Acetylation or methylation was detected in three distinct peptides from the replicative amastigote and epimastigote preparations but not in either of the non-replicating metacyclic or trypomastigote stages, consistent with proposed links between post-translational modifications of eEF1A, resultant changes in protein activity and cell proliferation (reviewed in (37, 38)).

ORFans

To identify genes potentially missed in the annotations provided by the *T. cruzi* sequencing consortium, a database of approximately 817,000 open reading frames (ORFs) of > 50 amino acids was constructed and screened using the Mascot search engine. Unique peptides identified by spectra that failed to match proteins predicted by the annotated genome were then clustered to the ORFs, and the new proteins were annotated. Finally, each annotation and the spectra matching the new gene were manually verified, yielding 79 new genes, new alleles or modifications to existing gene annotations (Table S6). In general, these analyses suggest that the prediction models and annotations by the TSK-TSC have been extremely efficient in accurately predicting genes. Sixty-six of the 79 ORFans are new alleles of annotated genes or corrections to existing annotations that may have been truncated, possibly due to assembly errors. In all cases, these new annotations map to the "coding" strand of DNA among genes which are presumably part of polycistronic units. This result is consistent with the model of kinetoplastid genes being clustered in large transcriptional units on the coding strand of DNA (39). Strand-

switch regions separate these clusters and allow for changing of the coding strand at sites of transcription initiation. Thus, although transcriptional activity on the "non-coding" DNA strand has been documented (40), the proteome does not provide evidence for translation of those alternative strand transcripts.

CONCLUSIONS

High-throughput proteome analyses are inherently incomplete, as the available methodologies do not have sufficient dynamic range to identify and quantify all proteins expressed in an organism. In this analysis, nearly 50% of all of the spectra matching to proteins mapped to the 67 most abundant protein groups. A higher number of lower abundance proteins can likely be revealed by depleting these highly abundant proteins prior to whole proteome analysis. Nevertheless, these analyses can serve as extremely useful tools for the generation and testing of hypotheses emerging from the knowledge of genome composition. Analysis of the proteomes of *T. cruzi* reveals the operation of several previously undocumented stage-specific pathways that could be appropriate targets for drug intervention. Among the most interesting of these is the proposed pathways for energy generation in amastigotes and epimastigotes. Additionally, the identification of the proteins expressed in abundance in trypomastigotes and amastigotes of *T. cruzi* provides a substantial new resource of candidates for vaccine development. Avirulent strains of *T. cruzi* produced by specific gene knockouts have potential as vaccines. Genes encoding proteins involved in fatty acid oxidation and in anti-oxidant defense, pathways that we would predict to be crucial for amastigote and metacyclic trypomastigote survival in mammalian hosts, are potential targets for gene deletion for the production of avirulent strains. This proteome analysis of *T. cruzi* also validates the high quality of the gene predictions generated by the *T. cruzi* genome sequencing consortium by confirming the expression of > 1000 hypothetical

genes and at the same time revealing <15 genes missed in the initial annotation. Although restricted to the analysis of *T. cruzi*, insight is provided into the unique aspects of other kinetoplastids by confirming the expression of conserved hypothetical genes that are also predicted in the genomes of *T. brucei* and *L. major*. This first-pass screening should also serve as a strong base for proteomic analyses of sub-cellular fractions and for high-accuracy quantitative exploration of stage-regulated protein expression (Table S3).

Supporting Online Material

http://www.sciencemag.org/

Tables S1, S3, S4, S5, S6

REFERENCES

1. C.E. Clayton, Embo J 21, 1881 (Apr 15, 2002).
2. N. S. Akopyants et al., Mol Biochem Parasitol 113, 337 (4/6, 2001).
3. T. A. Minning, J. Bua, G. A. Garcia, R. A. McGraw, R. L. Tarleton, Mol Biochem Parasitol 131, 55 (Sep, 2003).
4. S. Diehl, F. Diehl, N. M. El-Sayed, C. Clayton, J. D. Hoheisel, Mol Biochem Parasitol 123, 115 (Aug 28, 2002).
5. R. Duncan, Trends Parasitol 20, 211 (May, 2004).
6. M. M. Piras, R. Piras, D. Henriquez, S. Negri, Mol Biochem Parasitol 6, 67 (Aug, 1982).
7. S. Tomlinson, F. Vandekerckhove, U. Frevert, V. Nussenzweig, Parasitology 110 (Pt 5), 547 (Jun, 1995).
8. E. Rondinelli et al., Exp Parasitol 66, 197 (Aug, 1988).
9. D. Chao, D. G. Dusanic, Zhonghua Min Guo Wei Sheng Wu Ji Mian Yi Xue Za Zhi 17, 146 (Aug, 1984).
10. E. L. Isola, E. M. Lammel, O. Giovanniello, A. M. Katzin, S. M. Gonzalez Cappa, J Parasitol 72, 467 (Jun, 1986).
11. D. B. Weatherly et al., Mol Cell Proteomics 4, 762 (June, 2005).
12. N. J. Mulder et al., Nucleic Acids Res 31, 315 (Jan 1, 2003).
13. H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, Int J Neural Syst 8, 581 (Oct-Dec, 1997).
14. E. F. Strittmatter, N. Rodriguez, R. D. Smith, Anal Chem 75, 460 (Feb 1, 2003).
15. L. Florens et al., Nature 419, 520 (Oct 3, 2002).
16. A. I. Nesvizhskii, A. Keller, E. Kolker, R. Aebersold, Anal Chem 75, 4646 (Sep 1, 2003).

17. K. A. Resing et al., Anal Chem 76, 3556 (Jul 1, 2004).

18. J. Paba et al., Proteomics 4, 1052 (Apr, 2004).

19. J. Paba et al., J Proteome Res 3, 517 (May-Jun, 2004).

20. A. Parodi-Talice et al., Int J Parasitol 34, 881 (Jul, 2004).

21. S. Ghaemmaghami et al., Nature 425, 737 (Oct 16, 2003).

22. J. M. DiNoia, D. O. Sanchez, A. C. C. Frasch, J Biol Chem 270, 24146 (1995).

23. A. C. Frasch, Parasitol Today 16, 282 (Jul, 2000).

24. D. Martin, R. Tarleton, Immunol Rev 201, 304 (Oct, 2004).

25. F. Bringaud et al., Eukaryot Cell 1, 137 (Feb, 2002).

26. S. Tomlinson, F. Vandekerckhove, U. Frevert, V. Nussenzweig, Parasitology 110 (Pt 5), 547 (Jun, 1995).

27. H. P. Low, R. L. Tarleton, Mol. Biochem. Parasitol. 160, 1817 (1997).

28. C. Torres, F. J. Perez-Victoria, A. Parodi-Talice, S. Castanys, F. Gamarro, Mol Microbiol 54, 632 (Dec, 2004).

29. J. S. Harington, Parasitology 51, 309 (Dec, 1961).

30. J. S. Harington, Nature 178, 268 (Sep 4, 1956).

31. H. B. Vickery, J. Biol. Chem. 144, 719 (1942).

32. D. A. Evans, R. C. Brown, J Protozool 19, 686 (Nov, 1972).

33. S. R. Wilkinson, S. O. Obado, I. L. Mauricio, J. M. Kelly, Proc Natl Acad Sci U S A 99, 13453 (Oct 15, 2002).

34. R. Almeida et al., Mol Biochem Parasitol 136, 87 (Jul, 2004).

35. M. Mann, O. N. Jensen, Nat Biotechnol 21, 255 (Mar, 2003).

36. R. R. Cocklin, M. Wang, J Protein Chem 22, 327 (May, 2003).

37. S. Thornton, N. Anand, D. Purcell, J. Lee, J Mol Med 81, 536 (Sep, 2003).

38. S. Ejiri, Biosci Biotechnol Biochem 66, 1 (Jan, 2002).

39. S. Martinez-Calvillo et al., Mol Cell 11, 1291 (May, 2003).

40. E. A. Worthey et al., Nucleic Acids Res 31, 4201 (Jul 15, 2003).

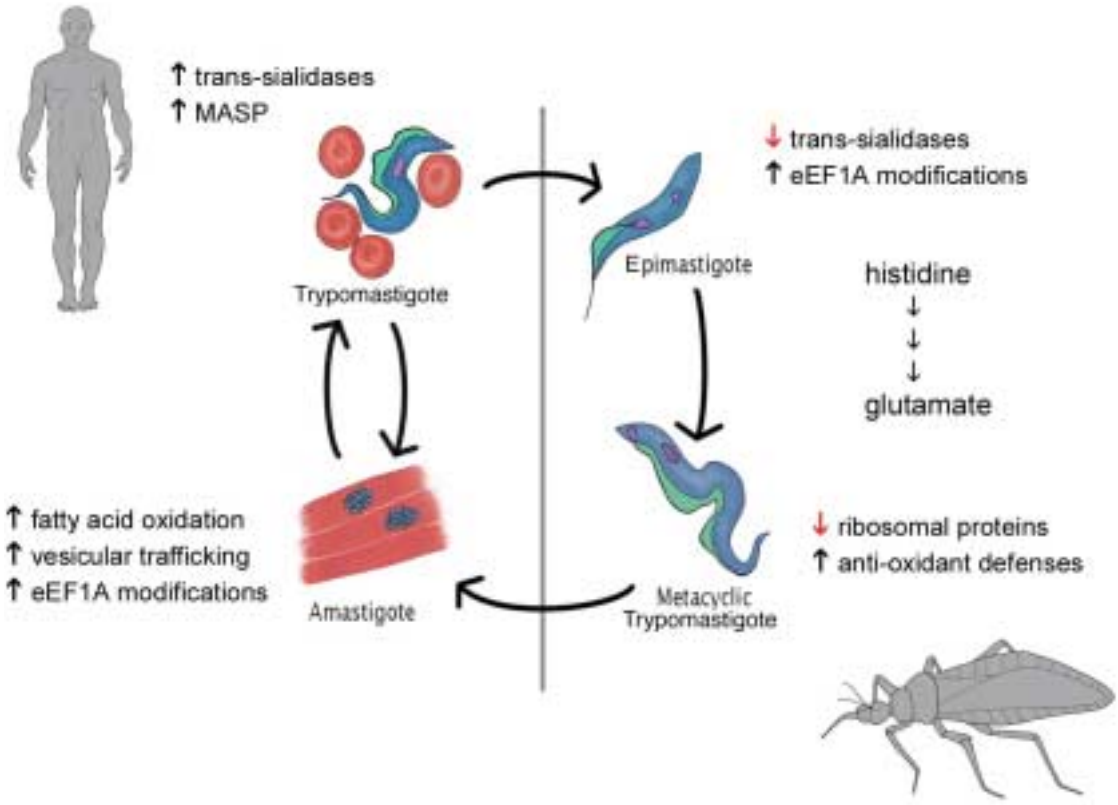

Figure 4.1 - Life cycle and summary of the major findings of proteome analysis in *T. cruzi*

T. cruzi trypomastigotes circulate in the blood of infected hosts, including humans, but must enter host cells (often times muscle cells) and convert to amastigote forms in order to replicate. Triatomine bug vectors become infected by ingesting trypomastigotes during the course of a bloodmeal on infected mammalian hosts. Conversion of the trypomastigotes into epimastigotes, replication of these epimastigotes, and their eventual transformation into metacyclic trypomastigotes, occurs in the insect gut. Metacyclic trypomastigotes initiate new infection in mammals when infected insects are ingested or by deposition of parasites in the feces, usually during a bloodmeal.

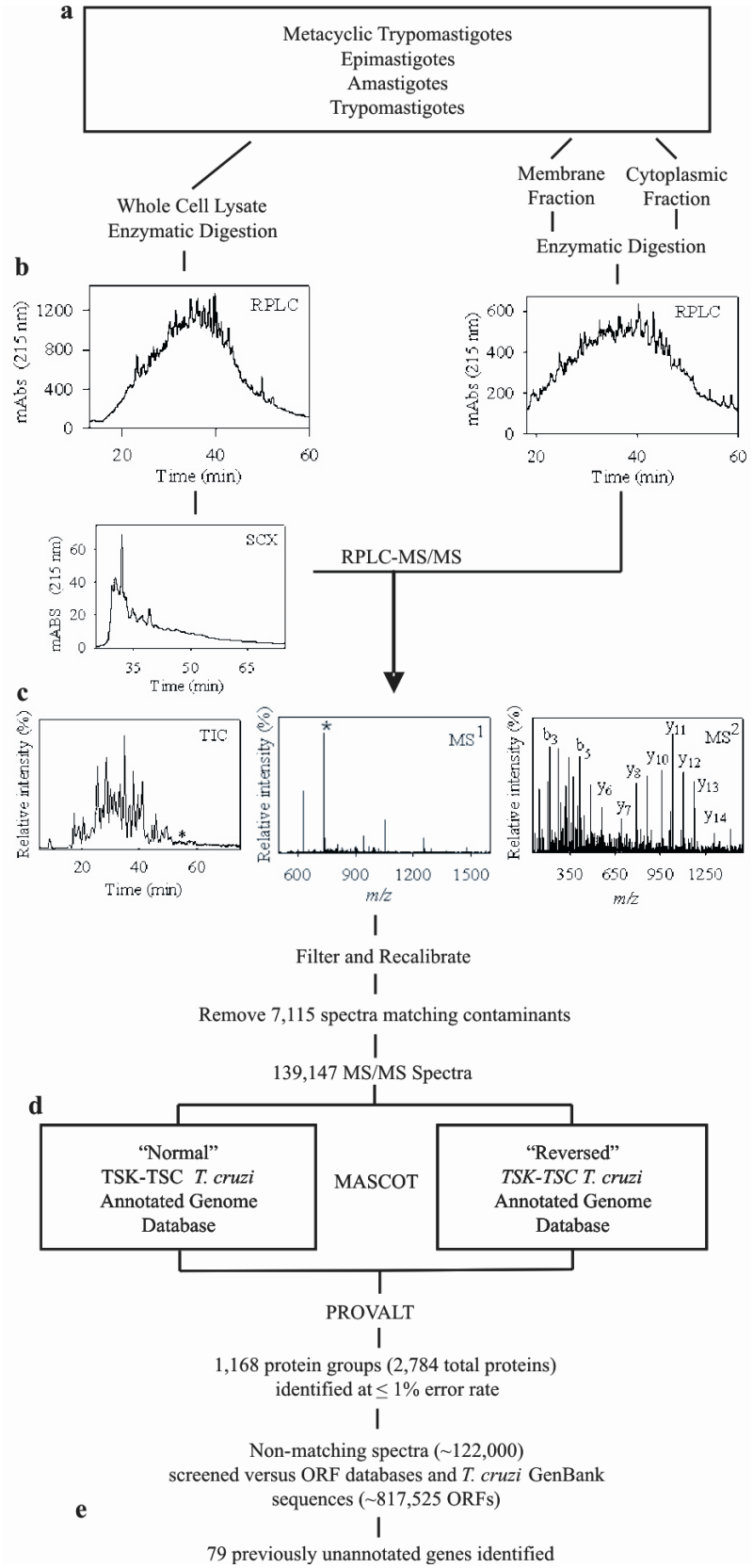
a
Metacyclic Trypomastigotes
Epimastigotes
Amastigotes
Trypomastigotes
Whole Cell Lysate
Enzymatic Digestion
Membrane Fraction
Cytoplasmic Fraction
Enzymatic Digestion
b
RPLC
mAbs (215 nm)
Time (min)
SCX
mABS (215 nm)
RPLC-MS/MS
c
TIC
Relative intensity (%)
MS1
m/z
MS2
Filter and Recalibrate
Remove 7,115 spectra matching contaminants
139,147 MS/MS Spectra
d
"Normal"
TSK-TSC T. cruzi
Annotated Genome
Database
MASCOT
"Reversed"
TSK-TSC T. cruzi
Annotated Genome
Database
PROVALT
1,168 protein groups (2,784 total proteins)
identified at ≤ 1% error rate
Non-matching spectra (~122,000)
screened versus ORF databases and T. cruzi GenBank
sequences (~817,525 ORFs)
e
79 previously unannotated genes identified

Figure 4.2 - Flow chart of protein isolation, separation, and analysis of four *Trypanosoma cruzi* developmental stages

a, Peptides were isolated from either whole-cell lysates or subcellular preparations from the four *T. cruzi* developmental stages. **b**, Chromatograms resulting from separation of peptides from a whole-cell lysate of 1 H 10^9 trypomastigotes (RPLC1, SCX) and cytoplasmic preparation of 1 H 10^9 amastigotes (RPLC2). **c**, All peptide fractions were analyzed independently by LC-MS/MS. The left panel is a total ion chromatogram from LC-MS/MS analysis of a single trypomastigote peptide fraction. A single mass spectrum (middle panel) measured at 55 min (TIC) contained a peptide with a mass of 759.37 Da, which was further analyzed by tandem mass spectrometry (right panel). **d**, Following filtering, recalibration and database searching the peptide at mass 759.37 was identified as peptide AAAEAAATATEAAEAAK from a newly annotated MASP protein family member. Iterative database searching against a normal and random database allowed identification of the MASP protein family member and 2,748 other proteins at less than 1% error rate. **e,** Non-matching spectra were used to identify an additional 79 unannotated genes form the *T. cruzi* ORF database.

Table 4.1

Table 1A **Protein group identification as a function of developmental stage**

Protein groups	Amastigote	Trypmastigote	Metacylic trypomastigote	Epimastigote
29	X	-	-	-
21	X	X	-	-
44	X	X	X	-
335	X	X	X	X
27	X	X	-	X
65	X	-	X	-
146	X	-	X	X
24	X	-	-	X
43	-	X	-	-
47	-	X	X	-
53	-	X	X	X
12	-	X	-	X
187	-	-	X	-
92	-	-	X	X
43	-	-	-	X
1168	691	582	969	732

Table 1B **Protein identification as a function of developmental stage**

Proteins	Amastigote	Trypmastigote	Metacylic trypomastigote	Epimastigote
49	X	-	-	-
41	X	X	-	-
161	X	X	X	-
838	X	X	X	X
84	X	X	-	X
110	X	-	X	-
538	X	-	X	X
50	X	-	-	X
125	-	X	-	-
122	-	X	X	-
93	-	X	X	X
22	-	X	-	X
315	-	-	X	-
162	-	-	X	X
74	-	-	-	X
2784	1871	1486	2339	1861

Table 4.2

Table 2. **Major protein families and functional classes**

Protein functional classes	Number identified proteins
Ribosomal	259
Proteasome/Ubiquitin	67
Heat Shock/Chaperonins	53
Translation/Transcription	49
Histones	36
Gene Families	
Trans-sialidase	226
RHS	156
GP63	21
Cruzipain/cysteine peptidase	16
MASP	9
Mucins	0
Hypothetical Genes	
Hypothetical	156
Hypothetical conserved	505
Hypothetical to be annotated	347

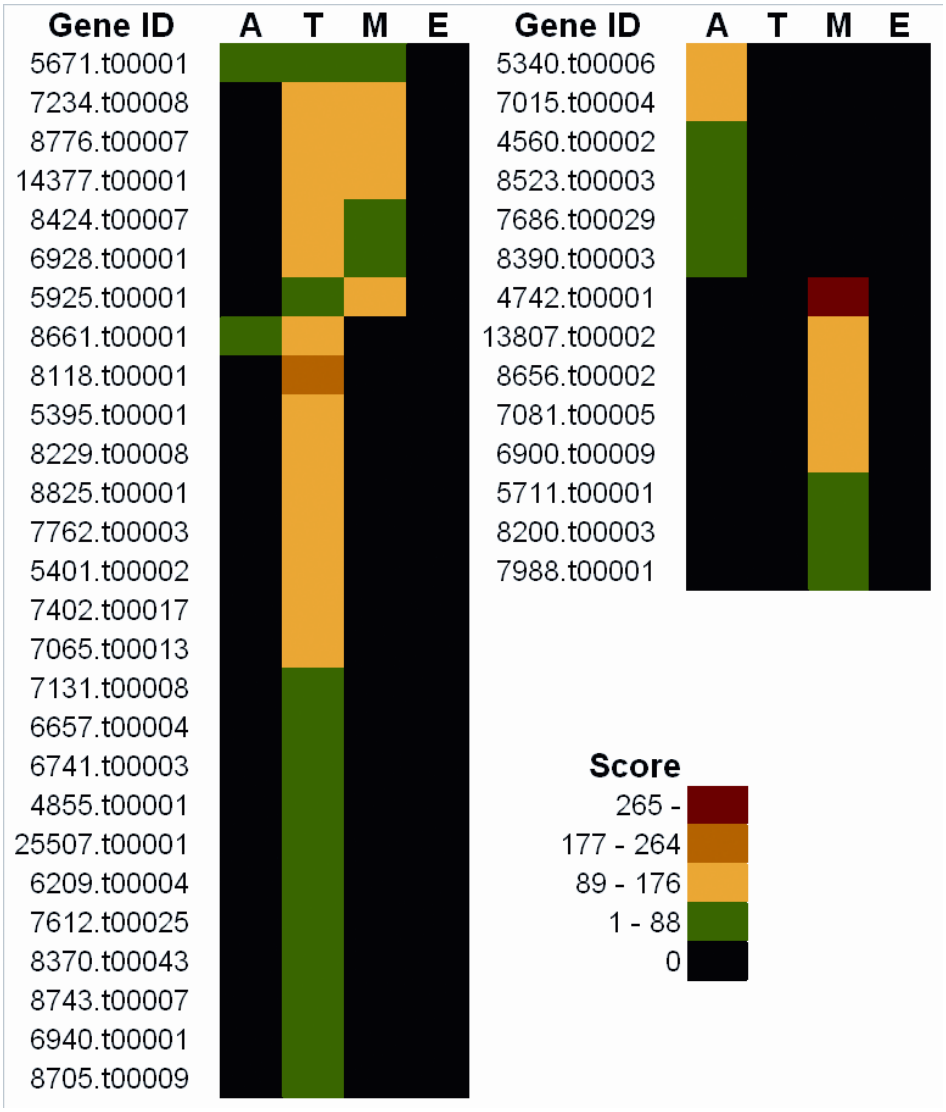

Figure 4.3 - Stage specific detection of trans-sialidase (ts) and retrotransposom hot spot (RHS) proteins. Cumulative protein scores based upon summing the Mascot scores for all high confidence peptides are used to display the stage-regulated expression of ts (a) and RHS (b) proteins detected in the proteomes. Only the top-scoring protein for each protein group is shown. Most ts proteins are detected exclusively in one stage while most RHS proteins are detected in multiple stages (A = amastigote, T= trypomastigote, M = metacyclic trypomastigote, E = epimastigote).

CHAPTER 5

IDENTIFICATION OF N-LINKED GLYCOPROTEINS FROM TRYPANOSOMA CRUZI TRYPOMASTIGOTES USING LECTIN AFFINITY AND STABLE ISOTOPE LABELING[1]

[1] Atwood, J., Minning, T., Weatherly, D., Alvarez-Manilla, G., Tarleton, R., Orlando, R.

To be submitted to *Proteomics*.

ABSTRACT

Here we describe the high-throughput identification of membrane associated N-linked glycoproteins from *Trypanosoma cruzi*, the causative agent of Chagas' disease. This analysis was based on enrichment of trypomastigote membrane proteins followed by capture of membrane associated N-linked glycoproteins by Concanavalin A affinity chromatography. Through stable isotope labeling of the glycan attachment sites with ^{18}O we unambiguously identified 19 glycopeptides which mapped to 17 glycoproteins, all of which were membrane associated. We also present the first evidence for the expression of 7 putative trypomastigote cell surface glycoproteins including GP-90, DGF-1, and a novel cysteine protease SCP1. Furthermore, we discuss the implications of two ER localized glycoproteins identified in this analysis, STT3 and GRP94.

INTRODUCTION

Chagas' disease currently afflicts ~18 million people and results in the loss of over US $1.2 billion/year in productivity (1,2). Furthermore, greater than 100 million people are exposed to the causative agent of Chagas' disease, the protozoan parasite *Trypanosoma cruzi* (1). *Trypanosoma cruzi* (*T. cruzi*) is a protozoan parasite endemic to much of Latin America. The life cycle of *T. cruzi* is complex, with multiple developmental stages persisting between a variety of mammalian host (including humans) and insect vectors. Metacyclic trypomastigotes in the insect hindgut initiate infection into the mammalian host via fecal contamination of mucus membranes or wound openings. The metacyclic trypomastigotes enter various cells and differentiate into aflagellated amastigotes which replicate in the host cell cytoplasm. These forms then give rise to flagellated trypomastigotes which are released into the blood stream of the mammalian host and either invade new cells or are ingested by the insect vector during the course of a blood meal. In the insect vector, non-replicative trypomastigotes migrate to the midgut and differentiate into replicative epimastigotes. Following multiple rounds of binary fission, epimastigotes transform into infective metacyclic trypomastigotes and migrate to the hindgut where they infect the mammalian host thus completing the life cycle.

To date no vaccines are available against *T. cruzi* and treatments for Chagas disease have been limited to chemotherapeutics which are highly toxic and ineffective during the chronic stage of the disease (3) Nevertheless, promising new therapeutic strategies have emerged which target proteins necessary for *T. cruzi* survival and pathogenesis within the mammalian host (3). Recent studies have also demonstrated that vaccination with specific *T. cruzi* surface antigens is capable of enhancing the survival of infected mice, suggesting that they may be useful vaccine candidates in humans as well (4-6). To date, the number of inhibitor and vaccine targets

analyzed has been limited by a lack of protein expression data, specifically in the mammalian stages. However, the recent completion of the *T. cruzi* genome and proteome provides a significant resource for the identification of novel inhibitor and vaccine targets (7,8). While the latter confirmed expression for greater than 1000 genes, traditional proteomics initiatives as applied in the *T. cruzi* proteome do not facilitate the identification of glycoproteins. This is unfortunate because increasing evidence indicates that *T. cruzi* glycoproteins mediate cellular recognition, host cell invasion, and immune evasion (9,40). Furthermore N-linked glycosylation has been shown to occur on proteins destined for secretion and for membrane incorporation on the parasite cell surface (9,10). Therefore this class of proteins is attractive as targets for therapeutic interventions.

In the present study we describe a targeted analysis of membrane associated N-linked glycoproteins from *T. cruzi* trypomastigotes. The approach is based on membrane protein enrichment followed by the capture of membrane associated N-linked glycoproteins by lectin affinity chromatography (11-13). Through stable isotope labeling of the glycan attachment sites with ^{18}O we provide the first evidence for the expression of many membrane bound glycoproteins. We also demonstrate that this approach facilitates the detection of several novel cell surface glycoproteins which may prove useful as vaccine or inhibitor targets against *T. cruzi* infection.

MATERIAL AND METHODS

Parasites preparation and membrane isolation

Brazil strain *T. cruzi* trypomastigotes were grown in monolayers of Vero cells (ATCC no. CCL-81) in RPMI supplemented with 5% horse serum as previously described (14). Emergent trypomastigotes were harvested daily and examined by light microscopy to determine the

percentages of trypomastigotes. Only preparations containing > 95% trypomastigotes were used in this study. 5x 10^8 cells were lysed in ice-cold lysis buffer (150 mM NaCl, 1.5 mM $MgCl_2$, 0.5% (v/v) NP-40, and 10 mM Tris-HCl, pH 8.0). The insoluble membrane fraction was harvested by centrifugation at 12,000 X g at 4°C for 5 min. The insoluble material was washed with 10ml of ice-cold wash buffer (100mM NaCl, 1mM EDTA, 1mM DTT, 20mM Tris-HCl, pH 7.5) for 10min on ice. The insoluble material was isolated by centrifugation at 20,000 X g for 10 min at 4°C. To produce the membrane enriched fraction, six replicate washes were performed to deplete the insoluble lysate of soluble proteins.

Membrane solubilization and protein digestion

The membrane enriched fraction was resuspended in 2ml of solubilization buffer (8 M urea, 100 mM Tris-HCl, 40 mM DTT, 0.1% RapiGest (Waters, Milford, MA, USA), pH 8.5). Protein reduction, carboxyamidomethylation, and enzymatic digestion were carried as previously described (15). Following digestion the RapiGest was precipitated by the addition of TFA to a final concentration of 0.5%, per the manufacturer's instructions. The precipitated RapiGest and remaining insoluble material were removed via centrifugation at 15,000 X g for 15min at room temperature. The supernatant was filtered over a 0.2μm filter (Millipore, Bedford, MA), frozen, lyophilized to dryness, then reconstituted in 500 μl of 0.1% TFA. The peptide mixture was desalted on an Agilent 1100 series workstation (Palo Alto, CA) configured with a 4.6 × 150 mm Jupiter C_{18} column (Phenomenex, Torrance, CA). Buffer A was H_2O/0.1% TFA, and buffer B was acetonitrile (ACN) /0.1% TFA. The peptides were loaded onto the column (flow rate 0.5 ml/min), desalted for 10 min at 0% buffer B then eluted with 80% buffer B. The desalted peptide

fraction was collected and dried by vacuum centrifugation. 1/10th of the membrane enriched fraction was analyzed directly by LC-MS/MS.

Lectin affinity chromatography

The desalted peptides where mixed with 300µl of binding buffer (20mM Tris-HCl pH 7.4, 0.5M NaCl, 1mM $MnCl_2$, 1mM $CaCl_2$) and passed over a spin column containing 300µl of Con A Sepharose (GE-Amersham Biosciences, Piscataway, NJ, USA) which had been washed with 1ml of binding buffer prior to addition of the peptide mixture. To elute non-glycopeptides, the Con A column was washed with 2ml of binding buffer. The bound glycopeptides were eluted from the column with 500µl of 0.5 M α-D-methylmannoside in binding buffer. The eluted glycopeptides were desalted as described above then dried by vacuum centrifugation. 1/10th of the desalted glycopeptide fraction (separated prior to drying) was analyzed directly by LC-MS/MS and the remainder was deglycosylated.

Deglycosylation of N-linked glycopeptides in $H_2{}^{18}O$

The dried glycopeptides were rehydrated in 20µl of 50mM ammonium bicarbonate in 95% $H_2{}^{18}O$ (Isotec through Sigma). PNGaseF (0.1 U) which had been suspended in $H_2{}^{18}O$ was added to the glycopeptides and deglycosylation was carried out overnight at 37°C. After deglycosylation the peptides were analyzed by LC-MS/MS.

LC-MS/MS

The membrane enriched fraction, glycopeptide fraction and deglycosylated peptides were analyzed independently on an Agilent 1100 capillary LC (Palo Alto, CA) interfaced directly to a

LTQ linear ion trap mass spectrometer (Thermo, San Jose, CA). Mobile phase A and B were H_2O/0.1% formic acid and ACN/0.1% formic acid, respectively. Each fraction was loaded onto the C18 column (15 cm × 150 μm, Grace Vydac) at a flow rate of 1 μl/min and peptides were eluted into the mass spectrometer during a 70 min linear gradient from 5-45% B. The instrument was set to acquire MS/MS spectra on the 9 most abundant precursor ions from each MS scan with a repeat count of 3 and repeat duration of 15 sec. Dynamic exclusion was enabled for 160 sec. Raw mass spectra were processed into .dta format, combined, and database searching was performed with Mascot 1.9 (Matrix Science, Boston, MA).

Protein Databases

Two sequence databases were constructed. First, we created a database (normal) consisting of 23,095 *T. cruzi* gene annotations provided by *Trypanosoma cruzi* Sequencing Consortium (TSK-TSC) combined with 1,786 possible contaminant proteins from *Equus caballus* (NCBI, www.ncbi.nih.gov). A randomized database (random) was then constructed by reversing the sequences in the normal database.

Database searching and protein identification

Database searches were performed against the normal and random databases using the following parameters: fully tryptic enzymatic cleavage with 1 possible missed cleavage, peptide tolerance of 500 parts-per-million, fragment ion tolerance of 0.6 Da, and a variable modification due to carboxyamidomethylation (+ 57 Da). For identification of deglycosylated ^{18}O labeled peptides the database search was performed using a variable modification of +3 Da on asparagine. Following database searching the peptide matches above discrete Mascot ion scores

were extracted from the normal and reverse databases search results. Protein identifications which resulted from a contaminant sequences were removed and the protein false-discovery rates (PRO-FDR) was calculated as previously described (15). For the identification of proteins following membrane enrichment, proteins were considered significant if were identified by a peptide exceeding a Mascot ion score of 44 (1% PRO-FDR). Glycoproteins were considered identified if they contained a glycopeptide which matched with a Mascot score above 39 (3% PRO-FDR) and contained an ^{18}O label on asparagines only present in the N-glycosylation consensus sequence (NXS/T).

Annotation of identified proteins

Gene ID's and annotations were acquired from T.cruziDB (http://tcruzidb.org (16)). Protein homologies were predicted by BLAST against NR (www.ncbi.nih.gov), the *T. brucei*, *L. major*, and *T. cruzi* annotated genomes. Transmembrane spanning domains were predicted with TMHMM 2.0 (17). Signal peptide motifs were predicted with SignalP3.0 (18). GPI anchor addition motifs were predicted with DGPI (http://129.194.185.165/dgpi/)(19)). For characterized proteins, subcellular localization was annotated using literature references within SWISSPROT (www.expasy.org) and predictions by PSORT (20). For hypothetical and uncharacterized proteins, subcellular localization was predicted by PSORT.

RESULTS AND DISCUSSION

Sample preparation - membrane enrichment

The goal of this study was to implement a method which would allow for the identification of a specific subset of novel glycoproteins which are likely significant for *T. cruzi*

development, infectivity, and survival within the host. The trypomastigote developmental stage provided a good starting point because it is present in blood stream of the mammalian host and is known to express a variety of membrane bound N-linked glycoproteins central in its pathogenicity (21). Thus we adopted a strategy in which the trypomastigote whole cell lysate was first depleted of soluble proteins prior to the lectin affinity chromatography (Fig.5.1). As others have done, enrichment of the membrane proteins was accomplished by repeatedly washing the insoluble portion of the lysate in a high salt buffer followed by centrifugation to pellet the membrane and subsequent removal of the supernatant (22). This procedure was performed for two reasons. First, the trypomastigote proteome is dominated by a subset of highly expressed soluble proteins such as heat shock proteins and dehydrogenases (8). Removal of these high abundance proteins reduces the sample complexity allowing for identification of proteins expressed at lower concentrations. Secondly, as other studies have noted, proteins which are highly abundant tend to exhibit "carry over" in the glycopeptide fraction following lectin affinity chromatography (23). This is most likely due to interactions between these proteins and the column packing material rather than non-specific binding by Con A. Thus by removing the soluble proteins we decreased the likely hood of "carry over" into the glycoprotein fraction following lectin affinity chromatography. Following membrane enrichment the membrane was solubilized by the addition of RapiGest and urea. Protein digestion was then carried out with endoproteinase Lys-C and trypsin. Unlike other detergents typically used for solubilizing membrane proteins, which are both difficult to remove and are not compatible with reverse phase separations, RapiGest is an acid cleavable detergent which was precipitated from the sample by acidification following digestion. The sample was then filtered to remove the detergent and any

non soluble membrane components then the soluble peptides were desalted by reverse phase chromatography.

Identification of proteins following membrane enrichment

To determine if the crude membrane enrichment had allowed for the depletion of high abundance soluble proteins, a portion of the desalted peptide sample was analyzed by LC-MS/MS. The resulting tandem mass spectra were searched using Mascot (24) against a normal and reversed *T. cruzi*/*Equus caballus* protein database. Mass spectra for which the best match was to a horse sequence were removed and the remaining *T. cruzi* matched peptides from both the normal and random databases were used to calculate the protein false-discovery rates (PRO-FDR) at individual Mascot ion scores as previously described (15). Only *T. cruzi* specific proteins which contained peptides exceeding ion scores of 44 (1% PEP-FDR) were considered significant. Table 5.1 shows the protein identifications from the membrane enriched fraction which were ranked by total protein score and compared to their rank in the trypomastigote whole proteome (8). Seven of the top eleven identifications (6998.t00004, 11788.t00001, 8152.t00002, 7083.t00002, 8485.t00013, 8623.t00012, 4937.t00013) in the membrane enriched fraction are components of the *T. cruzi* cytoskeleton and were also identified among the top 150 proteins in the trypomastigote proteome. This analysis also resulted in seven protein identifications which were not made in Atwood *et al*. Of these, three have predicted transmembrane spanning domains (7172.t00001, 8242.t00014, 6931.t00014) and four are components of the cytoskeleton (7407.t00012, 6287.t00004, 8197.t00006). Furthermore, the depletion of soluble proteins was clearly visible by the identification of only a single peptide from a carboxykinase (7378.t00009), dehydrogenase (7148.t00005), or heat shock proteins (7414.t00028, 8621.t00017); each of which

were among the 20 most abundant proteins detected in the trypomastigote proteome (8). Of the total 42 proteins identified, 12 (29%) were cytoskeletal, 10 (24%) were transmembrane or components of a transmembrane complexes, and 4 (10%) were associated with an organelle (Fig. 5.2). Only 5 (12%) were predicted to be in the cytoplasm, indicting that the membrane enrichment strategy was effective in depleting soluble non-membrane associated proteins.

Lectin affinity chromatography and N-glycosylation site mapping by ^{18}O labeling

N-glycosylation of membrane bound proteins is known to occur in *T. cruzi* (21). While many of these are components of the highly abundant glycoprotein layer on the cell surface others may be expressed at lower abundances but are equally significant in regards to the pathogenesis of Chagas' disease. However, membrane bound glycoproteins, especially if they are not highly expressed, are often not identified in traditional proteome analyses. One solution to this problem has been to capture glycoproteins by lectin affinity chromatography (23,25). Thus the glycosylated proteins are enriched relative to the highly abundant population of non-glycoproteins. With this in mind, we extracted the glycopeptides from the membrane enriched fraction using Con A lectin affinity chromatography. Con A was employed in this procedure for two reasons. First, Con A specifically binds high mannose, hybrid, or complex N-linked oligosaccharides containing α-D-mannopyranosyl or α-D-glucopyranosyl monosaccharides (26), glycans known to be present on *T. cruzi* membrane proteins. Furthermore, Con A has previously been used to isolate proteins from the parasite cell surface (12).

Following displacement of the bound glycopeptides from the lectin column a fraction of the eluent was directly analyzed by LC-MS/MS. As expected, this analysis confirmed that peptides containing N-linked oligosaccharides were present in the Con A bound fraction. One example is shown in figure 5.3. Two doubly charged glycopeptides (Fig. 5.3A) differing in mass

by one hexose (81 m/z) were subjected to collision induced dissociation (Fig. 5.3A and 5.3B). Figure 5.3B is the MS/MS spectrum which resulted from fragmentation of the glycopeptide at 1425.66 m/z. The fragmentation pattern clearly indicated that the glycopeptide at 1425.66 m/z contained an N-linked oligosaccharide with the structure $HexNac_2Hex_5$. The difference in mass correlating to a single hexose between the two glycopeptides at 1425.66 m/z and 1506.98 m/z (Fig.5.3A) was confirmed by the MS/MS spectrum in figure 5.3C which exhibited fragment ions consistent with the dissociation of a glycopeptide modified with the N-linked oligosaccharide $HexNac_2Hex_6$. While the stereochemistry of the glycan structure can not be determined by the MS or MS/MS spectra alone, the presence of fragment ions correlating to loss of HexNac (918.23 m/z, 917.77 m/z) and Hexose from the peptide bound oligosaccharide is indicative of a high mannose type glycosylation. In addition, knowledge that Con A preferentially binds high mannose type N-linked glycans allows us to assign the oligosaccharide structure in figure 5.3 as $GlcNac_2Man_{5-6}$. While evidence for the presence of high mannose type N-linked glycans were observed in the Con A bound fraction we were not able to detect glycopeptides with complex or hybrid type structures. This certainly does not mean other types of N-linked glycopeptides were not present in the Con A bound fraction. *T. cruzi* is known to express both complex and hybrid type N-linked oligosaccharides on proteins which are membrane associated and there presence in the Con A bound fraction would be expected. However, the glycopeptides containing high mannose type oligosaccharides are most likely present at a higher abundance in this preparation and thus were more readily detectable.

The remaining portion of the Con A bound glycopeptide fraction was then treated with PNGaseF in presence of $H_2{}^{18}O$ to remove the N-linked oligosaccharides and introduce a mass tag at the site of asparagine deamidation. PNGaseF catalyzed deglycosylation in $H_2{}^{16}O$ converts

formerly N-glycosylated asparagines to aspartic acids, a mass shift of 1Da. A 1Da mass shift is readily detectable by MS/MS and has previously been used to identify deglycosylated peptides. However, this method often results in false positive identifications due to incorrect precursor ion selection by the mass spectrometer or from identification of peptides containing asparagines which are naturally deamidated (23,27). This problem was avoided by performing the deglycosylation in $H_2{}^{18}O$, which introduces one stable ^{18}O into the deamidated asparagine (13,28). As figure 5.4 indicates, the ^{18}O label results in a mass shift of 3Da in the deamidated asparagine following deglycosylation. Thus the formerly N-linked glycopeptides were specifically identified by searching the MS/MS data using an asparagine modification of 3Da. Following database searching a number of criteria were employed to filter out identifications which did not result from peptides containing N-linked carbohydrates. Glycopeptides were only considered significant if they matched to a *T. cruzi* specific sequence above a score of 39 (3% PRO-FDR) and contained an ^{18}O label on an asparagine within the N-glycosylation consensus sequence NXS/T, where X is any amino acid other than proline. Table 5.2 shows the list of identified glycopeptides which bound to Con A. In total 19 unique glycopeptides were identified with high confidence and matched to 17 unique glycoproteins. Notably, fifteen glycoproteins were predicted to have at least one transmembrane spanning domain, signifying that the membrane enrichment was effective in depleting high abundance soluble proteins. Furthermore, thirteen were predicted to have signal peptides, signal anchors, or GPI anchors and were either known or predicted to be localized in the plasma membrane or the membrane of an organelle. Surprisingly, this is the first expression evidence for over 50% (n=10) of these glycoproteins and more than 25% (n=5) are novel putative cell surface glycoproteins.

Identification of T. cruzi organelle specific glycoproteins

With the exception of one hypothetical protein (7925.t00001) which is predicted to reside in the nucleus, the majority of identified organelle specific glycoproteins were predicted to localize in either the ER or the Golgi. In *T. cruzi* the ER and Golgi are the subcellular sites where proteins attain their proper folding and are modified by N-linked oligosaccharides (29). Subsequently, many of the glycoproteins identified in this analysis are involved in protein glycosylation (5150.t00008, 6196.t00003) or trafficking (7164.t00019, 8681.t00019) within these organelles. Additionally the identification of N-linked glycosylation sites within these proteins reflects the fact that protein linked oligosaccharides are involved in intercellular recognition phenomena (29). This is especially true for the oligosaccharyltransferase (OTase) complex, a multimeric enzyme located in the ER membrane which catalyzes the transfer of the oligosaccharide $Glc_3Man_9GlcNac_2$ to the asparagine residues of the nascent polypeptide chains (29). Recently, a glycosylated subunit (STT3) of the OTase complex has been shown to be directly involved in the substrate recognition process (30) and more specifically the mutation of a single N-glycosylation site on STT3 is lethal in *Saccharomyces cerevisiae* (31). The identification of the Con A bound glycopeptide ILAWWDYGYQITGIGNR from the *T. cruzi* STT3 is noteworthy because this peptide is highly conserved among other eukaryotic STT3 subunits and contains both the glycosylation site recognition domain (WWDYG) and an N-linked oligosaccharide (31). This would suggest that N-linked glycosylation on *T. cruzi* STT3 may also be imperative for the function of the OTase complex in *T. cruzi* as it is in other organisms.

In the ER, once proteins have been modified by enzymes such as STT3 they must obtain their appropriate tertiary or quaternary structures. This is facilitated by an abundant class of

soluble molecular chaperones, including GRP94 (32). In protozoan organisms, the expression of GRP94 homologues have been confirmed in *Leishmania* and *T. cruzi* (8, 32, 33), but the localization of GRP94 in the ER has only been demonstrated in *Leishmania* (33). However, this analysis does provide putative evidence that the GRP94 homologue (7164.t00019) is a soluble ER localized chaperone in *T. cruzi*. First, the GRP94 homologue identified here is homologous to both the *Leishmania* (55% sequence identity) and human GRP94's (38% sequence identity) (32,34) including the presence of a signal peptide, conserved HSP90 dimerization and nucleotide binding domains (35), and N-linked oligosaccharides (11,35). Second, the *T. cruzi* GRP 94 lacks a predicted transmembrane spanning domain or GPI-anchor addition site, indicating that if it was a peripheral membrane protein it would have been removed during the membrane enrichment. Furthermore, other identifications of ER localized GRP94's from pig and human have resulted from similar membrane preparations (11,36). However, unlike GRP94 in *Leishmania* and other eukaryotes, the *T. cruzi* GRP94 does not contain a known C-terminal ER targeting sequence *X*DEL or *X*DDL where the amino acid *X* is species specific (37). Rather *T. cruzi* GRP94 contains the C-terminal tetrapeptide AQDL. While the tetrapeptide AQDL is somewhat similar to other ER targeting sequences it has not been shown to target proteins to the ER in other organisms. Nevertheless, soluble ER proteins without C-terminal ER targeting sequences are known to occur in both *Leishmania* (38) and *T. cruzi* (39).

Identification of T. cruzi cell surface glycoproteins

For survival in the mammalian host T. *cruzi* must be capable of infecting and replicating within a variety of cell types while persisting in the presence of a potent and multifaceted immune response. *T. cruzi* infective trypomastigotes facilitate both immune evasion and host cell

entry through the expression of numerous heterogeneous cell surface glycoproteins. While the most abundant cell surface glycoproteins belong to the mucin and trans-sialidase families (40), a variety of less abundant cell surface enzymes, transporters, and receptors have recently been identified, some of which contain N-linked oligosaccharides and have been implicated in *T. cruzi* survival and cell invasion (41). However, with growing evidence that interactions between the parasite and the host cell are mediated by glycoproteins, the distribution and diversity of cell surface glycoproteins is still poorly understood. Thus methods which allow for the isolation and identification of novel cell surface glycoproteins or glycoprotein isoforms could provide information which leads to new antiparasitic targets. Through the selective enrichment of N-linked glycopeptides by Con A affinity chromatography we were able to identify a number of glycoproteins which are either known or predicted to be expressed on the surface of trypomastigotes.

Encoded by more than a thousand genes in the *Trypanosoma cruzi* genome the trans-sialadase family is one of the most abundant protein families expressed on the trypomastigote cell surface (7). Members of the trans-sialidase family are classified into four groups based on sequence homology and function. A number of genes in group II encode for glycoproteins which have recently been implicated in host cell recognition and adhesion (42,43). In this analysis we identified the glycopeptide SLLNESTIAAHK from gp90, a group II glycoprotein which is known to be GPI anchored to the cell surface of *T. cruzi* metacyclic trypomastigotes (10,44). While gp90 is known to contain high mannose type N-glycosylation, the sites of glycosylation have not been determined (45). The localization of the N-linked glycosylation site to Asn^{457} in gp90 is intriguing because the N-glycosylation consensus sequence is conserved within homologous peptides between gp82 and TC-85, two *T. cruzi* glycoproteins which share a high

degree of sequence similarity with gp90 but are known to bind distinct extracellular receptors on host cells (42,43). Unlike TC-85, previous reports have demonstrated that gp90 expression was strictly limited to metacyclic trypomastigotes in the *T. cruzi* G strain (46,47). However, our analysis indicated that gp90 was expressed in mammalian stage trypomastigotes, a finding commensurate with the identification of a 90kDa glycoprotein in trypomastigotes by Andrews *et al* (12) Nonetheless when taking into account that gp90 is encoded by a large gene family (48) and its expression in metacyclic trypomastigotes varies greatly between strains (49) it is not unreasonable that it is expressed in *T. cruzi* trypomastigotes from the Brazil strain. The expression of gp90 in trypomastigote forms is interesting being that in metacyclic trypomastigotes, levels of gp90 expression are inversely correlated with parasite infectivity of mammalian cells (49). This is potentially based on gp90 impairing binding of gp82 to receptors on the cell surface thus reducing the Ca^{2+} response required for parasite internalization (51). Thus modulatory control of parasite infectivity by gp90 may occur, albeit to lesser extent, in trypomastigote forms as well.

Like the tran-sialadase gene family, the genes encoding a probable cell surface protein called DGF-1 (Dispersed Gene Family-1) are plentiful in the *T. cruzi* genome (Cruzi Genome Reference). While DGF-1 and multiple DGF-1 gene homologues are abundantly transcribed in epimastigotes (51,52) the identification of a peptide from DGF-1 in this analysis is the first evidence that the protein is expressed. Although the function of DGF-1 is not known, sequence analysis predicts that the protein is expressed on the cell surface with 9 transmembrane spanning domains and it contains multiple epidermal growth factor (EGF)-like motifs which may indicate that DGF-1 is involved in binding of the parasite to the host cell surface (51). This is supported by recent evidence that proteins containing EGF-like domains are involved in the adhesion of

Toxoplasma gondii and Plasmodium vivax to the surface of host cells (53,54). Furthermore DGF-1 contains a 73-amino acid mucin-like region rich in threonine and serine residues which likely contain multiple O-linked oligosaccharides (51). Thus DGF-1 represents a class of proteins which would normally not be identified by conventional proteome analysis. This is highlighted by the fact that DGF-1 was not identified in the *T. cruzi* proteome (8) and its identification in this analysis emphasizes the effectiveness of discriminatory separation strategies such as lectin affinity chromatography for proteome analyses.

Another important class of cell surface and secreted enzymes identified here is the papain family of cysteine proteases which mediate intracellular growth and antigenicity (9,55,56). To date, the identification and functional characterization of *T. cruzi* papain cysteine proteases has been limited to the main *T. cruzi* lysosomal CP, cruzipain (57-59) and its highly homologous isoforms (60,61). Interest in cruzipain has been augmented by recent evidence that cruzipain is involved in the invasion of endothelial cells (9), cardiomyocytes (62) and human smooth muscle cells (63). Commensurate with these findings, the inhibition of cruzipain has been shown to arrest intracellular replication, impair host cell invasion and cure mice from experimental *T. cruzi* infection (55,64,65). In this analysis we identified two glycopeptides from a senescence-specific cysteine protease (SCP1, 7624.t00011). The identification is interesting since it provides the first evidence for the expression in a *T. cruzi* mammalian stage of a papain-like cysteine protease other than cruzipain or one of the cruzipain isoforms. While the papain-like cysteine proteases described to date in *T. cruzi* share >86% sequence identity with cruzipain (66), the cysteine protease identified in this analysis is more divergent, sharing less than 40% sequence identity with all other cruzipain isoforms (Figure 5.5). Nevertheless, SCP1 does share noteworthy similarities with cruzipain. Sequence alignment indicated that the active site residues

are well conserved and both enzymes contain N-glycosylation sites within their catalytic domains (Figure 5.5). Furthermore, the identification of a Con A bound N-linked glycan on Asn^{184} in SCP1 is consistent with the high mannose type glycosylation previously reported on Asn^{202} in cruzipain (67) (Fig. 5.5). Like various cruzipain isoforms (68-70), SCP1 is likely to be present on the surface of *T. cruzi* through either a GPI or a transmembrane domain. This is supported by the predication of a signal peptide in the protein sequence and the identification of the protein in the insoluble membrane preparation. The confirmed expression of SCP1 in *T. cruzi* trypomastigotes and the prediction that it is surface expressed is compelling in light of evidence that cruzipain and its isoforms are expressed at the lowest levels in trypomastigote forms (8,71) and they differ in both substrate specificity and susceptibility to inhibition (66). While the functional role of SCP1 has not been investigated its homology to cruzipain and its expression in trypomastigote forms may indicate a role in cell invasion.

Of the 17 glycoproteins identified 8 were annotated as "hypothetical" and had not been identified in any prior studies. Thus this analysis provides the first data validating these as expressed glycoproteins in *T. cruzi*. Furthermore, 4of these hypothetical genes (7912.t00006, 7414.t00005, 8369.t00013, 6532.t00010) have orthologs in the *Leishmania major* and/or *Trypanosoma brucei* genomes and are predicted by sequence analysis to be expressed on the cell surface. In addition these hypothetical proteins appear to be trypanosome specific being that they have no homology to any proteins in other organisms. While this class of proteins is interesting due to their likely surface expression and apparent parasite specificity, further studies will be necessary to draw conclusions regarding their function.

CONCLUDING REMARKS

In conclusion we presented an approach to identify membrane associated N-linked

glycoproteins from the trypomastigote developmental stage of the parasite *Trypanosoma cruzi*. We demonstrated that by repeated washing of the insoluble cell lysate the soluble high abundance proteins were effectively depleted. The application of Con A lectin affinity chromatography and stable isotope labeling with ^{18}O, facilitated the identification of 17 membrane associated glycoproteins of which 7 were predicted to be expressed on the parasite surface. The identification of many novel *T. cruzi* specific glycoproteins which were not identified in any previous analyses including the *T. cruzi* proteome indicated that by probing specific sub proteomes the dynamic range of detectable protein identifications can be dramatically increased. However, the glycoproteins identified in this analysis most certainly represent only a fraction of the *T. cruzi* glycoprotein expression. *T. cruzi* is known to express other types of N and O-linked oligosaccharides so the application of other lectins specific for different oligosaccharides will provide a better understanding of this complex glycoprotein distribution.

REFERENCES

1. World Health Organization., 2002. WHO Tech. Rep. Ser. 905, 82-83.

2. Cubillos-Garzon, L., Casas, J., Morillo, C., Bautista, L., 2004. Am. Heart. J. 147, 412-417.

3. Urbina, J., 2002. Cur. Pharm. Des. 8, 287-295.

4. Costa, F., Franchin, G., Pereira-Chioccola, V., Ribeirao, M., Schenkman, S., Rodrigues, M., 1998.Vaccine. 16, 768–774.

5. Wizel, B., Garg, N., Tarleton, R., 1998. Infect. Immun. 66, 5073–5081.

6. Planelles, L., Thomas, M., Alonso, C., Lopez, M., 2001. Infect. Immun. 69, 6558–6563.

7. Andersson, et. al., 2005. Science. In Press.

8. Atwood, J., Weatherly, D., Bundy, B., Minning, T., Cavola, C., Opperdoes, F., Orlando, R., Tarleton, R., 2005. Science. In Press.

9. Scharfstein, J., Schmitz, V., Morandi, M., Capella, A., Lima, C., Morrot, A., Juliano, L., Muller-Esterl, W., 2000. J. Exp. Med., 192, 1289-1299.

10. Cardoso de Almeida, M., Heise, N., 1993. Biol. Res. 26 285–312.

11. Zhang, H., Li, X, Martin, D., Aebersold, R., 2003. Nat. Biotech. 21, 660-666.

12. Andrews, N., Katzin, A., Colli, W., 1984. Eur. J. Biochem. 140, 599-604.

13. Kaji, H., Saito, H., Yamauchi, Y., Shinkawa, T., Taoka, M., Hirabayashi, J., Kasai, K., Takahashi, N., Isobe, T., 2003. Nat. Biotech. 21, 667-672.

14. Piras, M., Piras, R., Henriquez., D., Negri, S., 1982. Mol. Biochem. Parasitol. 6, 67-81.

15. Weatherly B, Atwood J, Minning T, Cavola C, Tarleton R, Orlando R. *Mol. Cell. Prot.* 2005; 4: 762.

16. Luchtan, M., Warade, C., Weatherly, D., Degrave, W., Tarleton, R., Kissinger, J., 2004. Nucleic. Acid. Res. 32, D344-346.

17. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E., 2001. J. Mol. Biol. 305, 567-580.

18. Bendtsen, J., Nielsen, H., von Heijne, G., Brunak, S., 2004. J. Mol. Biol. 340, 783-95.

19. Kronegg, J., Buloz., 1999. http://129.194.185.165/dgpi/index_en.html

20. Nakai, K., Horton, P., 1999. Trends Biochem. Sci. 24, 34-36.

21. McConville, M., Mullin, K., Ilgoutz, S., Teasdale, R., 2002. Micro. Mol. Bio. Rev. 66, 122-154.

22. Wei, J., Yu, W., Jones, A., Oeller, P., Keller, M., Woodnutt, G., Short, J., 2005. J. Prot. Res. 4, 801-808.

23. Brunkenborg, J., Pilch, B., Podtelejnikov, A., Wisniewski, J., 2004. 4, 454-465.

24. Perkins D, Pappin D, Creasy D, Cottrell, S. *Electrophoresis*. 1999; 20: 3551.

25. Xiong, L., Andrews, D., Regnier, F., 2003. J. Proteome. Res. 2, 618-625.

26. Baenziger, J., Fiete, D., 1979. J. Bio. Chem. 254, 2400-2407.

27. Atwood, J., Sahoo, S., Alvarez-Manilla, G., Weatherly, D., Kolli, K., Orlando, R., York, W., 2005. Rapid. Comm. Mass. Spec. Submitted.

28. Gonzalez, J., Takao, T., Hori, H., Besada, V., Rodriquez, R., Padron, G., Shimonishi, Y., 1992. Anal. Biochem. 205, 151-158.

29. Parodi, A., 2000. Biochem. J. 348, 1-13.

30. Wacker, M., Linton, D., Hitchen, P., Nita-Lazar, M., Haslam, S., North, S., Panico, M., Morris, H., Dell, A., Wren, B., Aebi, M., 2002. Science. 298, 1790-1793.

31. Li, G., Yan, Q., Nita-Lazar, A., Haltiwanger, R., Lennarz, W., 2005. J. Bio. Chem. 280, 1864-1871.

32. Larreta, R., Soto, M., Alonso, C., Requena, J., 2000. Exp. Parasitol. 96, 108-115.

33. Descoteaux, A., Avila, H., Zhang, K., Turco, S., Beverley, S., 2002. EMBO. 21, 4458-4469.

34. Maki, R.., Old, L., Srivastava, P., 1990. Proc. Natl. Acad. Sci. USA, 87, 5658-5662.

35. Argon, Y., Simen, B., 1999. Semin. Cell. Dev. Biol. 10, 495-505.

36. Wearsch et al 1996.

37. Bangs, J., Brouch, E., Ransom, D., Roggy, J., 1996. J. Bio. Chem. 271, 18387-18393.

38. Padilla, A., Noiva, r., Lee, N., Krishna Mohan, K. Nakhasi, H., Debrabant, A., 2003. J. Biol. Chem. 278, 1872-1878.

39. Conte, I., Labriola, C., Cazzulo, J., Docampo, R., Parodi, A., 2003. Mol. Bio. Cell. 14, 3529-3540.

40. Frasch, A., 2000. Parasitology Today. 16, 282-286.

41. Burleigh, B., Woolsey, A., 2002. 4, 701-711.

42. Giordano, R., Fouts, D., Tewari, D., Colli, W., Manning, J., Alves, M., 1999. J. Bio. Chem. 274, 3461-3468.

43. Manque, P., Neira, I., Atayde, V., Cordero, E., Ferreira, A., da Silveira, J., Ramirez, M., Yoshida, N., 2003. Infect. Immun. 71, 1561-1565.

44. Schechter, M., Flint, J., Voller, A., Guhl, F., Marinkelle, C., Miles, M., 1983. Lancet. 322, 939-941.

45. Yoshida, N., Bianco, S., Araguth, M., Russo, M., Gonzalez, J., 1990. Mol. Biochem. Paratitol. 39, 39-46.

46. Teixeira, M., Yoshida, N., 1986. Mol. Biochem. Parasitol. 18, 271-282.

47. Franco, F., Paranhos-Bacalla, G., Yamauchi, L., Yoshida, N., Franco da Silveira, J., 1993. Infect. Immun. 61, 4196-4201.

48. do Carmo, M., dos Santos, M., Cano, M., Araya, J., Yoshida, N., da Silveira, J., 2002. Mol. Biochem. Parasitol. 125, 201-206.

49. Ruiz, R., Favoreto, S., Dorta, M., Oshiro, M., Ferreira, A., Manque, P., Yoshida, N., 1998. Biochem. J. 330, 505-511.

50. Malaga, S., Yoshida, N., 2001. Infect. Immun. 61, 4196-4201.

51. Winker, P., Murto-Dovales, A., Goldenberg, S., 1992. Mol. Biochem. Parasitol. 55, 217-220.

52. Kim, D., Chiurillo, M., El-Sayed, N., Jones, K., Santos, M., Porcile, P., Andersson, B., Myler, P., da Silveira, J., Ramirez, J., 2005. Gene. 346, 153-161.

53. Garcia-Reguet, N., Lebrun, M., Fourmaux, N., Mercereau-Puijalon, O., Mann, T., Beckers, J., Samyn, B., van Beeumen, J., Bout, D., Dubremetz, J., 2000. Cell. Mico. 2, 353-364.

54. Han, H., Park, S., Kim, S., Hwang, S., Han, J., Traicoff, J., Kho, W., Chung, J., 2004. Biochem. Biophys. Res. Comm. 320, 563-570.

55. Engel, J., Doyle, P., Hsieh, I, McKerrow, K., 1998. J. Exp. Med. 188, 725-734.

56. Scharfstein, J.,Schechter, M., Senna, M., Peralta, J., Mendonca-Previato, L., Miles, M., 1986. J. Immunol. 137, 1336-1341.

57. Cazzulo, J., Couso, R., Raimondi, A., Wernstedt, C., Hellman, U., 1989. Mol. Biochem. Parasitol. 33, 33-41.

58. Murta, A., Persechini, P., Souto-Padron, T., De Souza, W., Guimaraes, J., Scharfstein, J., 1990. Mol. Biochem. Parasitol. 43, 27-38.

59. Eakin AE, Mills R., Harth, G., McKerrow, J., Craik, C., 1992. J. Biol. Chem. 267, 7411-20.

60. Campetella, O., Henricksson, J., Aslun, L., Frasch, A., Pettersson, U., Cazzulo, J., 1992. Mol. Biochem. Parasitol. 50, 225-234.

61. Lima, A., Tessier, D., Thomas, D., Scharfstein, J., Storer, A., Vernet, T., 1994. Mol. Biochem. Parasitol. 67, 333-338.

62. Todorov, A., Andrade, D., Pesquero, J., Araujo, R., Bader, M., Stewart, J., Gera, L., Muller-Esterl, W., Morandi, V., Goldenberg, R., Neto, H., Scharfstein, J., 2003. FASEB J. 7, 73-75.

63. Aparicio, I., Scharfstein, M., Lima, A., 2004. Infect. Immun. 72, 5892-5902.

64. Meirelles, M., Juliano, L., Carmona, e., Silva, S., Costa, E., Murta, A., Scharfstein, J., 1992. Mol. Biochem. Parasitol. 52, 175-184.

65. Harth, G., Andrews, N., Mills, A., Engel, J., Smith, R., McKerrow, K., 1993. Mol. Biochem. Parasitol. 58, 17-27.

66. Lima, A., Reis, F, Serveau, C., Lalmanach, G., Juliano, L., Ménard, R., Vernet, T., Thomas, D., Storer, A., Scharfstein, J., 2001. *Mol. Biochem. Parasitol.* 114, 41-52.

67. Parodi, A., Labriola, C., Cazzulo, J., 1995. Mol Biochem Parasitol. 69, 247-55.

68. Nascimento, A., Souza, W., 1996. Biol. Cell. 86, 53-58.

69. Fresno, M., Hernandez-Munaim, C., de Diego, J., Rivas, L., Scharfstein, J., Bonay, P., 1994. Braz. J. Med. Biol. Res. 27, 431-437.

70. Parussini, F., Duschak, V., Cazzulo, J., 1998. Cell. Mol. Biol. 44, 513-519.

71. Tomas, A., Kelly, J., 1996. Mol. Biochem. Parsitol. 76, 91-103.

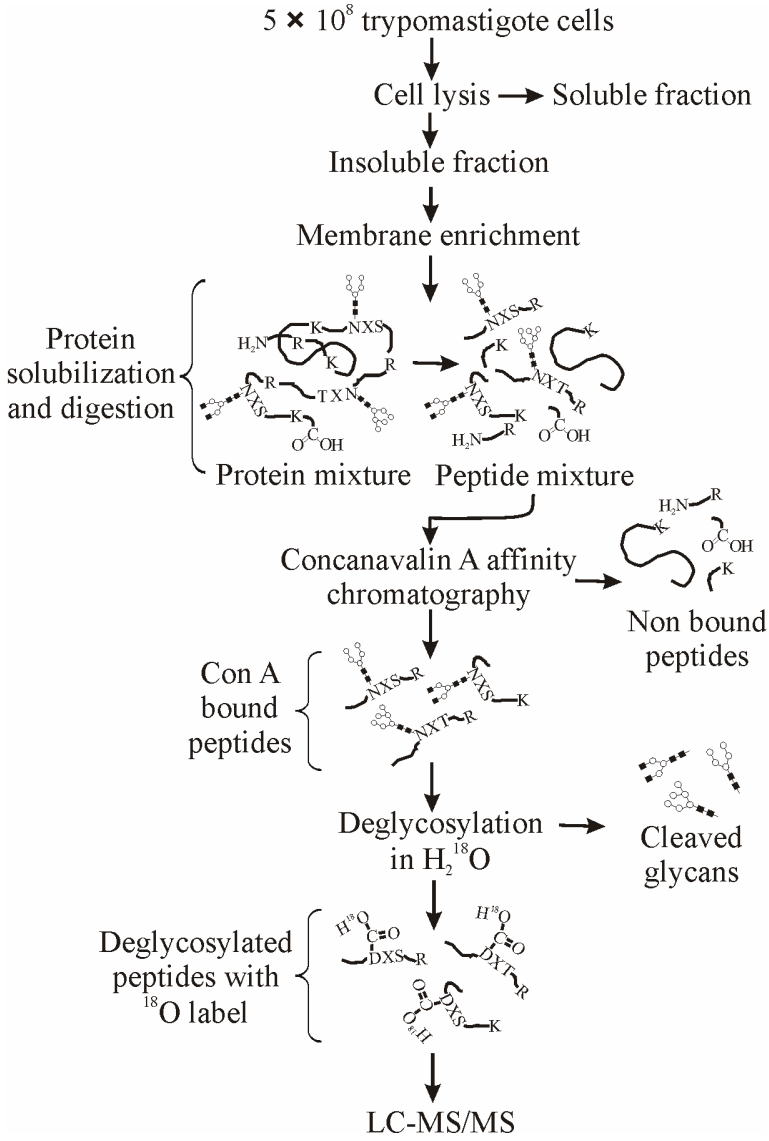

Figure 5.1 - Schematic of procedure for the analysis of membrane associated N-linked glycoproteins from *Trypanosoma cruzi* trypomastigotes.

Table 5.1

Table 1

Identified *Trypanosoma cruzi* proteins by LC-MS/MS following membrane enrichment

Gene ID	Protein name	Membrane enriched			Whole Trypomastigote Proteome			No. of TMSD
		Rank	No. of peptides identified	Protein score	Rank	No. of peptides identified	Protein score	
6998.t00004	Beta tubulin	1	9	598	2	19	1128	-
11788.t00001	Alpha tubulin	2	9	490	1	26	1504	-
7172.t00001	Gim5A protein, putative	3	3	250	-	0	-	1
8152.t00002	Paraflagellar rod protein 3	4	4	247	30	10	493	-
8045.t00018	ATPase beta subunit	5	3	181	25	10	546	-
7083.t00002	paraflagellar rod component	6	2	176	76	3	239	-
8741.t00001	Histone H2B	7	2	171	45	7	352	-
8485.t00013	Kinetoplastid membrane protein	8	3	171	106	7	198	-
8623.t00012	Paraflagellar rod component 2	9	3	163	12	14	707	-
7637.t00003	Histone H4	10	3	158	106	19	594	-
4937.t00013	ToIT2	11	1	110	149	3	157	1
5442.t00003	Hypothetical	12	2	108	93	4	217	-
8242.t00014	Hypothetical	13	1	98	-	0	-	1
4893.t00008	ATP synthase F1 subunit gamma protein	14	1	94	397	1	49	-
7739.t00048	60S ribosomal protein L23	15	1	91	251	2	91	-
8515.t00004	60S ribosomal protein L7a	16	1	79	234	4	127	-
6741.t00004	Histone H3	17	2	83	50	10	329	-
8417.t00008	Histone H2A	18	2	80	157	3	149	-
7009.t00004	Glucose-regulated protein 78	19	1	81	10	5	995	-
5568.t00006	60S ribosomal protein L2	20	1	79	284	2	80	-
7407.t00012	Hypothetical	21	1	77	-	0	-	-
6287.t00004	Microtubule-associated protein Gb4	22	1	76	-	0	-	-
8197.t00006	Hypothetical	23	1	75	-	0	-	-
5738.t00003	Mitochondrial phosphate transporter	24	1	75	338	1	62	2
7694.t00017	Hypothetical	25	1	74	-	0	-	-
6866.t00017	40S ribosomal protein S6	26	1	74	284	3	81	-
7414.t00028	Chaperonin heat-shock protein 60	27	1	73	7	15	863	-
5784.t00012	Kinetoplast DNA-associated protein	28	1	71	59	3	286	2
7685.t00010	I/6 autoantigen	29	1	68	33	9	457	-
5420.t00003	ATP-dependent DEAD-box RNA helicase	30	1	66	114	5	192	-
7678.t00007	Hypothetical	31	1	61	206	2	114	-
6931.t00014	Hypothetical	32	1	60	-	0	-	10
8046.t00014	Calpain-like cysteine peptidase	33	1	69	-	0	-	-
7634.t00001	60S ribosomal protein L11	34	1	57	43	1	38	-
8621.t00017	heat-shock protein 70	35	1	54	3	26	1045	-
6890.t00027	60S ribosomal protein L28	36	1	53	129	4	175	-
6172.t00013	Vesicle-associated membrane protein	37	1	53	273	2	83	1
4940.t00004	Fructose-bisphosphate aldolase, glycosomal	38	1	52	39	9	411	-
7269.t00002	Ribosomal protein L21E	39	1	51	478	2	39	-
7378.t00009	Glycosomal phosphoenolpyruvate carboxykinase	40	1	49	20	17	593	-
7148.t00005	Glyceraldehyde 3-phosphate dehydrogenase	41	1	44	10	17	778	-
8538.t00015	Proteasome regulatory non-ATP-ase subunit 2	42	1	44	520	1	28	-

The protein accession numbers are available from TcruziDB (Luchtan et al., 2004). Protein identifications resulting from the LC-MS/MS analysis of the trypomastigote membrane enriched preparation were ranked by protein score (sum of peptide ion scores) then compared with the distribution of protein identifications collected in the trypomastigote whole proteome analysis (Atwood et al., 2005). TMSD indicates transmembrane spanning domain predicted by TMHMM 2.0 (Krogh et al., 2001).

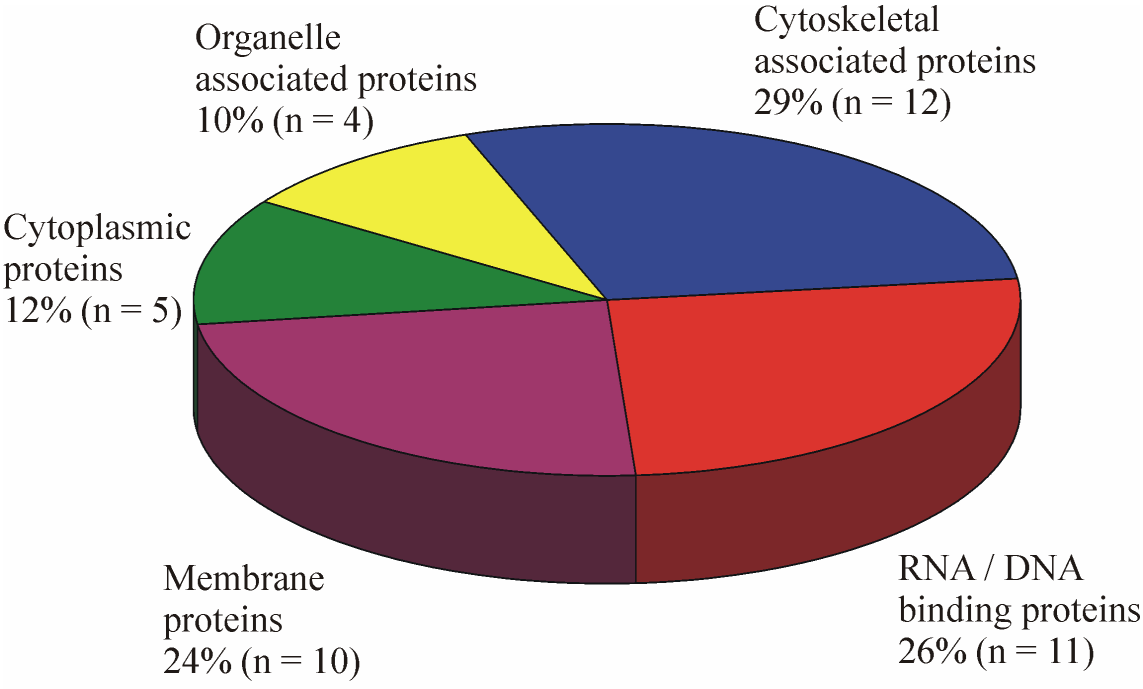

Figure 5.2 - Subcellular localization of proteins identified by LC-MS/MS of the membrane enriched fraction

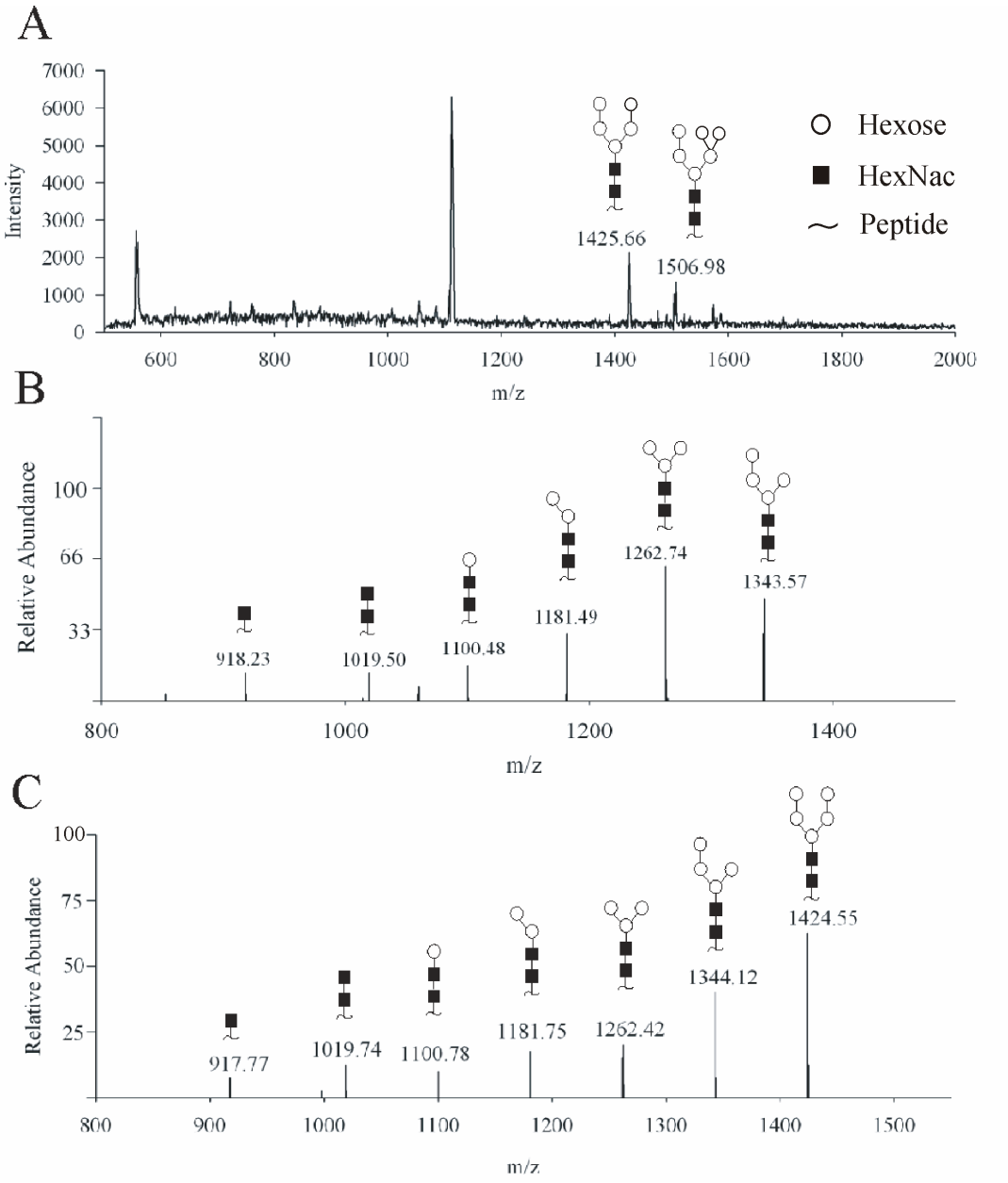

Figure 5.3 - LC-MS/MS of Con A bound glycopeptides

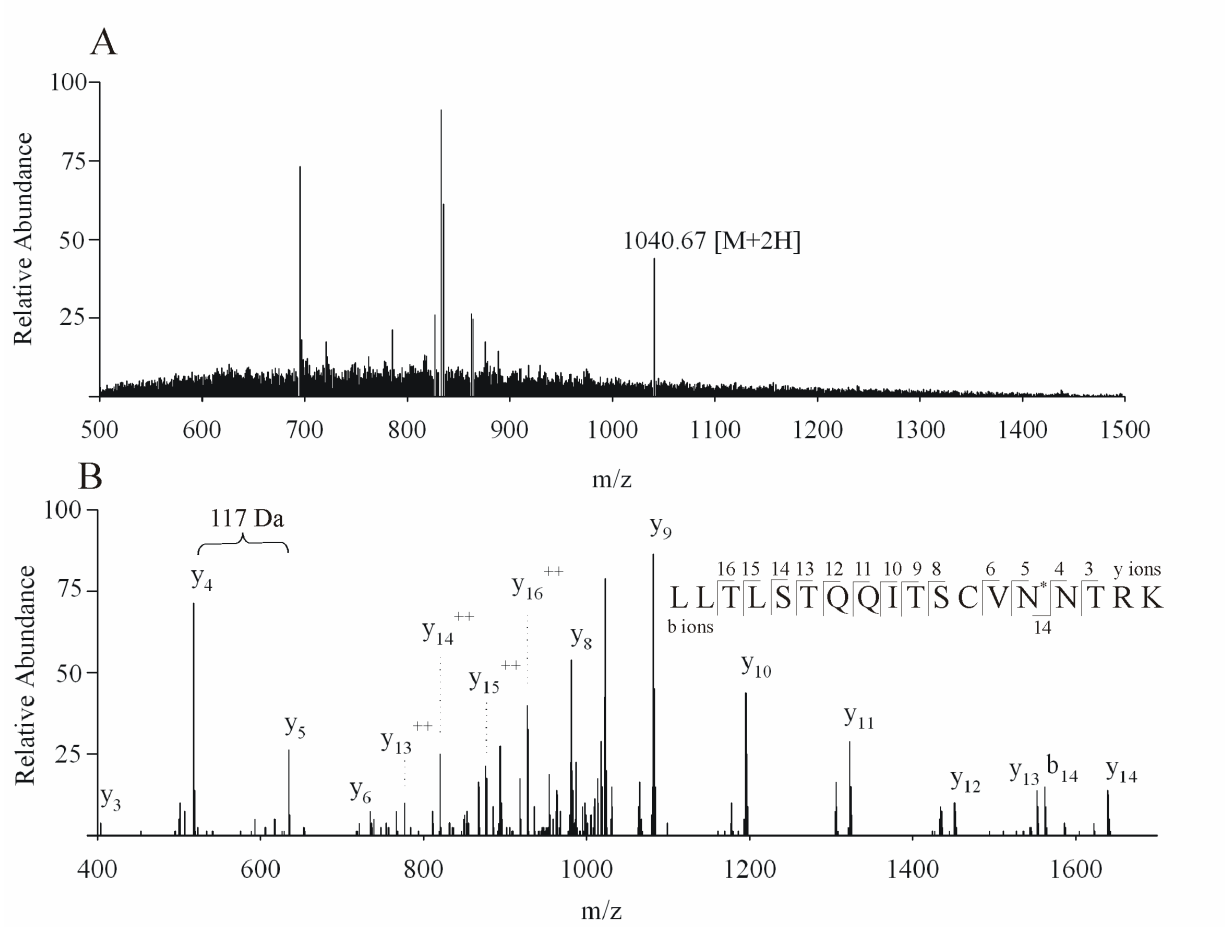

Figure 5.4 - MS and MS/MS spectra of a formerly N-linked glycopeptide from SCP1

Table 5.2

Table 2

Concanavalin A bound N-linked glycoproteins identified in *Trypanosoma cruzi* trypomastigotes

Gene ID	Protein name[a]	Identified peptide sequence[b]	Score	No. of TMSD[c]	SP[d] / SA[e] (*P* value) GPI-C[f] / GPI-A[f]	Subcellular location[g] (*P* value)
7624.t00011	senescence-specific cysteine protease	(K)LLTLSTQQITSCVN*NTRK(C)	91	1	SP (1.0) / GPI-C	PM (0.46)
		(K)LLTLSTQQITSCVN*NTR(K)	69	-	-	-
8681.t00018	Golgi/lysosome glycoprotein	(K)SN*STILSDIGILSVEGSR(H)	86	3	-	G (0.4)
		(K)SLKLTFN*ESK(I)	42	-	-	-
4976.t00001	hypothetical	(R)LLEDFVSGPLDYFN*STAER(V)	107	3	SA (0.83)	ER (0.68) / PM (0.64)
7164.t00019	lipophosphoglycan (GRP94 homolog)[#]	(K)SN*GTLVSLQEYTDR(M)	95	-	SP (0.84)	OS (0.8) / ER (0.1)
6196.t00003	hypothetical protein	(R)NVLN*STQPVVTEDEAHR(L)	84	1	SA (1.0)	PM (0.81) / G (0.3)
6532.t00010	hypothetical protein	(R)HTN*ATELAVAVR(Q)	81	10	-	PM (0.6)
5150.t00008	oligosaccharyl transferase subunit (STT3)[#]	(R)ILAWWDYGYQITGIGN*R(T)	78	8	-	PM (0.6) / ER (0.3)
5967.t00008	hypothetical protein (LEM3 family)	(R)KLICN*ATDFSK(G)	73	1	SA (0.98)	PM (0.1)
7149.t00034	hypothetical	(K)SNPIVGEN*STLLLCNIR(Y)	67	9	GPI-C	PM (0.6)
7414.t00005	hypothetical protein	(R)IFAVSGNGLLN*HTLGEK(C)	66	1	SP (1.0) / GPI-C	PM (0.46)
8320.t00007	HSP-70 like protein[#]	(R)TAIVN*NTLGGR(A)	64	1	SP (0.56)	PM (0.9)
8369.t00013	hypothetical protein[#]	(K)KLELLPTSN*NTR(L)	60	-	SA (0.56)	PM (0.46)
6906.t00004	GP90[#]	(R)SLLN*ESTIAAHK(G)	56	3	SA (0.71) / GPI-A	PM (0.60)
8364.t00009	C-8 sterol isomerase putative	(R)VQHGVN*ATPEQVIEK(V)	54	1	SP (0.9)	OS (0.82) / ER (0.1)
7993.t00001	probable cell surface protein (DGF-1)	(R)IVVQN*VSLR(N)	55	9	GPI-C	PM (0.6)
7912.t00006	hypothetical protein conserved	(R)APIPLSSIN*HTAGTISSK(M)	50	1	SP (0.9) / GPI-C	MIM (0.83)/ PM (0.6)
7925.t00001	hypothetical protein conserved	(R)GN*ISVFSVNPR(G)	39	3	-	N (0.98)

Gene identifications from TcruziDB (Luchtan et al., 2004); [a] # indicates the protein was identified in the *T. cruzi* trypomastigote proteome (Atwood et al., 2005); [b] Peptide sequence not enclosed by parentheses was identified by MS/MS; * indicates N-glycosylation site; [c]Transmembrane spanning domain predicted by TMHMM 2.0 (Krogh et al., 2001); [de]Signal peptide and anchor predicted by SignalP 3.0 (Nielsen et al., 1997); [f]GPI cleavage site and/or anchor predicted by DGPI (Kronegg et al., 1999) [g] Protein localization in the golgi (G), endoplasmic reticulum (ER), lysosome (L), nucleus (N), plasma membrane (PM), peroxisome (PER), and outside (OS) was predicted by PSORT (Nakai., Horton., 1999). *P* value is the probability score between 0 and 1.

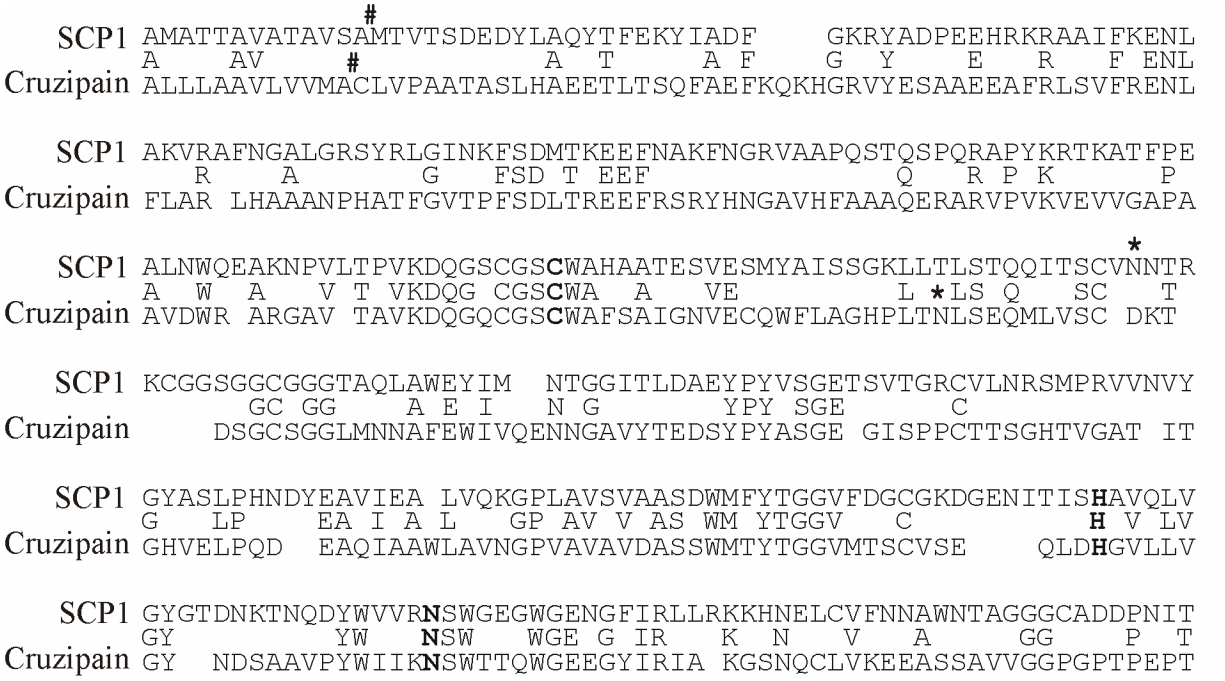

Figure 5.5 - BLAST sequence analysis between native cruzipain and SCP1

CHAPTER 6

CONCLUSIONS

Chapter 3:

Here the proteome of the *T. cruzi* epimastigote developmental stage was analyzed by 3DLC-MS/MS. The resulting data were analyzed by the traditional methods of protein identification when using Mascot, the single peptide score (SPM) and cumulative protein score (CPM) strategies. The PROVALT algorithm was described and evaluated against the SPM and CPM using the data produced from the analysis of the epimastigote proteome. PROVALT was shown to reduce a complex list of peptide matches to a nonredundant list of proteins groups. Being that *T. cruzi* expresses large number of gene families resulting in multiple protein isoforms accurate clustering of proteins in this analysis was crucial. Furthermore, PROVALT was shown to facilitate the accurate calculation of protein false-discovery rates, therefore reducing random protein identifications which often result from implantation of the SPM and CPM. Thus the rigorous application of bioinformatics tools like PROVALT allow for the filtering of large proteomic datasets such that only non-random high quality data are reported.

Chapter 4:

The tools developed in chapter 3 were applied in the protein expression analysis of the four developmental stages of the *T. cruzi* lifecycle. Analysis of the proteomes of *T. cruzi* revealed the operation of several previously undocumented stage-specific pathways that could be appropriate targets for drug intervention. Among these was the apparent metabolism of histidine in the vector stages and fatty acids in intracellular amastigotes. In addition this dataset provided confirmation of expression for over 1000 hypothetical genes and allowed for the identification of many newly assigned stage specific proteins which may also be good targets for therapeutic strategies.

Chapter 5:

In chapter five the *T. cruzi* proteome is extended to include the analysis of N-linked glycoproteins from the mammalian stage trypomastigotes. Glycoproteins represent a protein class which is not easily analyzed by traditional proteomics methods thus this chapter outlines an approach for their analysis in a high-throughput fashion. We demonstrate that membrane bound glycoproteins can be effectively enriched through isolation of the parasite membrane and extraction of glycoproteins via lectin affinity chromatography and their identification by stable isotope labeling. The analysis revealed several novel glycoproteins of which many were predicted to be expressed on the surface of the parasite. This class of protein is interesting because they are likely to come in contact with the mammalian immune response and thus are promising as vaccine targets.