

Overview

One of the difficulties in analyzing the results of shotgun proteomics experiments is the differentiation of protein isoforms from the identified peptides. In this study we demonstrate how to effectively resolve protein isoforms in complex proteomic analyses by combining molecular weight information, protein grouping, peptide filtering, and set logic.

Introduction

In bottom up proteomics experiments, identified peptides must be mapped to protein sequences for reporting of protein identifications. However differentiating between protein isoforms is complicated by the fact that peptides are analyzed rather than intact proteins. Thus if a peptide is shared between two proteins then, without additional information, it is impossible to distinguish which protein is actually expressed or if both proteins are expressed. This problem arises when the sequence database of the organism of study contains many similar (but unique) proteins, either because the genome contains related families of proteins or because the database was constructed from different sequenced isolates and thus contains many protein isoforms.

Case Study

The organism *Trypanosoma cruzi* (*T. cruzi*) is the causative agent of human Chagas disease and with over 30% of its genome being comprised of multi-copy gene families, the proteome is highly redundant and difficult to discern by traditional proteomic methods (Figure 1). The proteome of *T. cruzi* is dominated by the presence of multiple large protein families, including members of the trans-sialidase (TS) superfamily, surface glycoprotein gp63 protease (gp63), glycosyltransferases, and to a lesser extent kinesins, and elongation factors.

While *T. cruzi* represents a worst case scenario, the problem of identifying protein isoforms is certainly not unique to this organism. Many organisms are known to contain multiple protein families and as protein sequencing efforts continue to expand public database are becoming cluttered with many copies of the same protein with minor changes in amino acid composition. For example, a search of NCBI for "HIV Envelope Protein" yields 12,523 protein sequences.

Figure 1. *T. cruzi* Genome Map

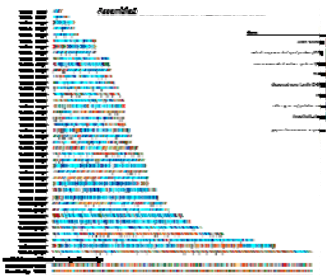


Figure 1. Assembled Chromosomes of *T. cruzi*. The *T. cruzi* proteome is dominated by members of several large gene families. The largest and most abundant being cell surface proteins such as trans-sialidases, mucins, and mucin associated membrane proteins (MASP).

Experimental

Biological Sample Selection

CL-Brenner strain (Genome strain) trypomastigotes were selected for this study because they are known to express an abundant array of large protein family members and this lifecycle stage is of particular importance due to its presence in the host blood stream and interaction with the host's immune system.

Membrane Enrichment

Being that many of the large protein families are expressed on the parasites cell surface, three types of membrane enrichment strategies were employed^{1,2}.

Sucrose Gradient Centrifugation

Membrane proteins were enriched using the sucrose gradient centrifugation as previously described with minor modifications.¹ Briefly cells were suspended in 3 mL of ice-cold lysis buffer (10 mM HEPES, 1 mM EDTA, pH 7.2) containing protease inhibitors and then homogenized by 25 strokes of a 7 mL dounce homogenizer. An equal amount of sucrose buffer (10 mM HEPES, 1mM EDTA, 500 mM sucrose, pH 7.2) was added with additional 25 strokes of the homogenizer. The samples were centrifuged (6,000g) for 10 min at 4°C to pellet cellular debris. The supernatant was collected and centrifuged at 150,000g. The membrane pellet was washed with Na₂CO₃ (pH 11.3).

Detergent Resistance

Detergent resistant lipid raft membranes were isolated using previous methods with minor modifications.² Cells were suspended in 3 mL of ice-cold lysis buffer (10 mM HEPES, 1 mM EDTA, pH 7.2) containing protease inhibitors and 1% (w/v) Triton X-100. The solution was homogenized with 50 strokes on the dounce homogenizer. The sample was centrifuged at 6,000g to pellet down cellular debris, the supernatant was collected and centrifuged at 150,000g. Membrane pellet was washed twice with 1% (w/v) Triton X-100 solution.

Detergent Resistance + Sucrose Gradient Centrifugation

Similar to detergent resistant method, cells were treated with same lysis buffer containing 1% Triton X-100. An equal amount of sucrose buffer was added. The mixed solution was homogenized with 50 strokes on the dounce homogenizer. The sample was centrifuged (6,000g) for 10 min at 4°C to pellet cellular debris. The supernatant was collected and centrifuged at 150,000g. Membrane pellet was washed with Na₂CO₃ (pH 11.3).

Peptide Isolation and Analysis

Proteins from each membrane preparation was separated by 1D-SDS-PAGE. Gel lanes were excised into 20 bands each. Proteins in each gel band were in-gel digested with trypsin and the peptides were then separated by reverse phase liquid chromatography and analyzed by tandem mass spectrometry on a linear ion trap (LTQ) mass spectrometer (Thermo).

Data Analysis Workflow

The ultimate goal in this analysis was to identify as many protein family members as possible. Toward this goal, subcellular fractionation was combined with 1D gel electrophoresis. Since many of the large protein family members are known to localize in the various membranes (many in the plasma membrane) of trypomastigotes this approach offered the best chance to achieve increased proteome coverage of these proteins.

While protein separation prior to mass spectrometry reduces the overall complexity of each sample, a large number of MS data files are produced. In this study, each membrane preparation led to 20 MS analyses for a total of 60 LC-MS/MS runs. The following steps were performed to analyze this data set.

The native MS data was transformed to mzXML format using ReAdW then peak list were created with mxXML2Other. The peak list were then searched using Mascot (v 1.9) against a target and decoy (reversed) database composed of 23,095 *T. cruzi* gene annotations. This resulted in 120 database search results.

ProteoIQ was utilized to combine, statistically validate, and identify protein groups using the database search results derived from Mascot. In the first analysis, database search results from all membrane preparations were combined into a single project to create a master list of protein groups at a protein FDR of 1.0%.

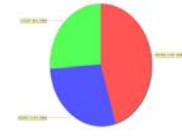
To determine the number of "unique" protein identifications, peptides are then restricted to being present in only one protein within a group.

In ProteoIQ, database search results were also grouped according to gel band and membrane preparation. This allowed protein within groups to be resolved based on comparing experimental and theoretical molecular weights.

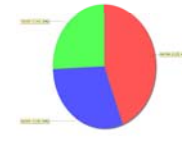
Results Overview

A total of 2601 proteins were identified including validation of expression for greater than 700 large gene family members.

Protein Group Distribution



Protein Distribution



● Sucrose
● Detergent
● Detergent + Sucrose

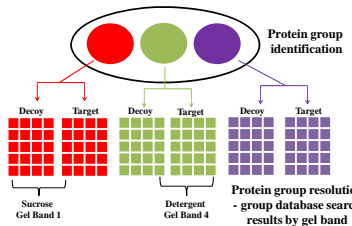
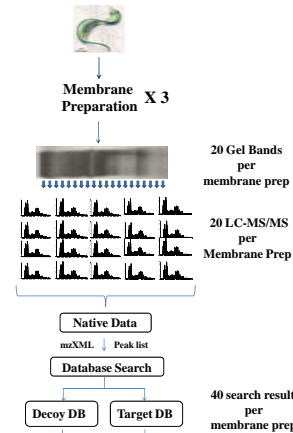
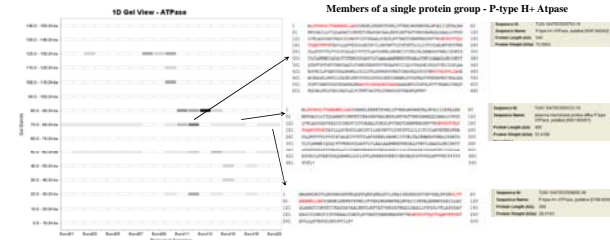


Table 1. Identification of proteins and comparison to genome

Protein name	# proteins	# of genes in genome
Trans-sialidase	487	1430
GP63	54	425
MASP	13	1377
Mucins	4	758
RHS	39	752
Glycosyltransferase	58	139
ATPase	41	99
Ribosomal	70	525
Elongation factors	16	22

Virtual 1D-Gels to Resolve Individual Protein Groups

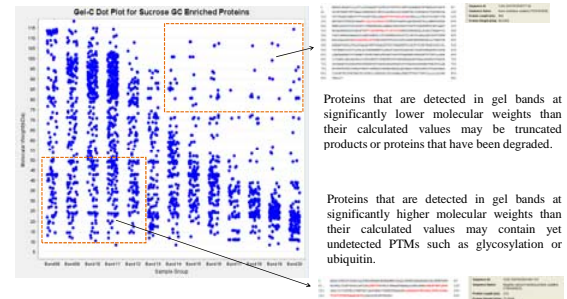


The virtual 1D gel view groups protein expression (in the form of spectral counts) by predicted molecular weight range. Since the groups are composed of individual gel bands protein calculated molecular weights can be compared with experimentally determined elution profiles from the 1D gel. The darkness of each gel section indicates the number of spectral counts.

In the example above, three P-Type H+ ATPases are identified by a shared set of peptides. Since all of the peptides are shared across all proteins it is impossible to assign a single protein as "identified". Rather all three proteins are placed into a single protein group. Utilizing the gel elution information we can see that this protein group contains peptide identifications in multiple gel bands at decreasing experimental molecular weights. This evidence indicates that each of the three P-Type ATPases within the single group are likely being expressed.

Resolving Protein Groups in Complex Datasets

In a Gel-C analysis it is not uncommon to observe a single protein or protein group eluting in multiple gel bands. While the virtual 1D-Gel view presented above allows you to detect protein expression based on spectral counting combined with experimental and observed molecular weights, the Gel-C dot plot can be used to determine proteins that have are in disagreement between their observed and calculated masses.



Conclusions

Here we demonstrate an approach to resolve protein isoforms based on combining shotgun proteomic results with molecular weight information and set logic. A membrane proteomic study of *T. cruzi* trypomastigotes provided a unique set of large protein family members to assess the feasibility of this approach. We anticipate that this approach will find applicability in the proteomic analyses of other organisms and will assist in resolving protein groups arising from redundant database entries.

References

- Seyfried et al., *Cancer Letters* 2008, 263, 243-252
- Radeva et al., *Biochem. J.* 2004, 380, 219-230

Acknowledgments

This work was funded in part by the NIH grant P01 AI-44979 to R.L. Tarleton and 1R41GM082525-01 to BioInquire, LLC.