

Eukaryotic Signature Proteins: Guides to modern eukaryotic parasites

Jian Han^{1,2}
jian272@hotmail.com

Lesley Collins^{1,2}
l.j.collins@massey.ac.nz

Biggs, Patrick^{1,2}
p.biggs@massey.ac.nz

Penny, David^{1,2}
d.penny@massey.ac.nz

¹ Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand.

² The Institute for Molecular BioSciences, Massey University, Palmerston North, New Zealand.

Abstract

Eukaryotic signature proteins (ESPs) are proteins that are found in every eukaryotic proteome but have no significant homology to proteins in Archaea and Bacteria. We have now calculated ESP datasets for *Giardia lamblia*, *Plasmodium falciparum*, *Trichomonas vaginalis* and human (the host). The ESP datasets and their databases form the ground work for future research about the parasites in how they maintain their proteomes, and understanding mechanism of protein loss. Future research will look for essential proteins involved in the growth of the parasites, and investigate potential drug targets.

Keywords: Eukaryotic signature protein, Giardia, Plasmodium, Trichomonas, MySQL database

1 Introduction

Eukaryotic signature proteins are signature proteins that delineate the Eukarya from the Archaea and Bacteria. They have no homologues in prokaryotic genomes, but their homologues are present in all main branches of eukaryotes and they are involved in the most core functions of a eukaryote. Previously Hartman *et al.*[2] calculated ESP datasets for Giardia, however their calculations were based on yeast proteins and the number of organisms with genomes sequenced were few at the time. We have calculated new ESP datasets for *Giardia lamblia*, and two other protist parasites (*Plasmodium falciparum*, *Trichomonas vaginalis*), as well as for humans (the host). By comparing the parasitic ESP datasets to that of the host, we observe trends in the loss and acquisition of proteins in parasites. Future research could lead to the identification of proteins from both hosts and parasites that are essential for parasites' growth and metabolism, and potential new drug targets can be unveiled for treating parasite infections.

2 Method and Results

Only analyses performed on *Giardia lamblia* are presented on this poster abstract, database and workflow design also apply to the other three organisms.

2.1 ESP datasets

Giardia ESP datasets were collected using the following procedure:

From all 6500 annotated Giardia proteins, first we removed those that have homologues in any of the 16 bacterial and 9 archaeal species, then we removed proteins do not have homologues in *Drosophila melanogaster*, *Caenorhabditis elegans* (animals), *Arabidopsis thaliana*, *Oryza sativa* (plants), *Saccharomyces cerevisiae*, *Eremothecium gossypii* and *Yarrowia lipolytica* (fungi). Lastly we screened against human and mouse, and removed proteins which do not have homologues in human and mouse.

We have yielded 267 ESPs, these include 262 distinctive proteins (four ESPs possess multiple gene copies in the genome). These 267 ESPs were divided into seven groups according to their function based on their description (Table 1). Besides the ESPs, we have also generated eight other sets of *Giardia* proteins (e.g. Giardia unique proteins) which will be useful in future for parasite metabolism studies.

Table 1. Giardia ESP data listed by protein group.

Protein group	Number of ESPs
Membrane	37
Cytoskeleton	43
Signaling system	63
Nucleus	38
Protein synthesis and breakdown	11
Unknown	40
Hypothetical protein	26
Total	267

2.2 Database design

MySQL databases were used for the storage of our ESP data. The advantage of using such databases is that they allow fast and organised information retrieval, and easier updating when newer parasitic genomes/proteomes become available. In addition the relational database management system allows large volumes of information to be efficiently stored.

The Giardia database is illustrated by Figure 1 (created using MySQL workbench). Each box corresponds to a table, the lines connecting tables show the relationship between the columns of the two tables, whether it is a one to one or one to many relationship, and the captions indicate which two columns correspond to each other.

Our MySQL databases enabled the storage of a variety of information: the Giardia_source table contained detailed information of Giardia proteins from NCBI in a tabular format; BLAST_results table stored all information of our path filters, and enabled us to track down precisely when individual proteins were excluded from the datasets. MySQL functional tools allowed the comparison of multiple datasets. The use of databases will greatly assist future analyses on the selected parasitic metabolic pathways as the complexities of metabolic interactions can only be examined using an integrative approach, which our databases enable.

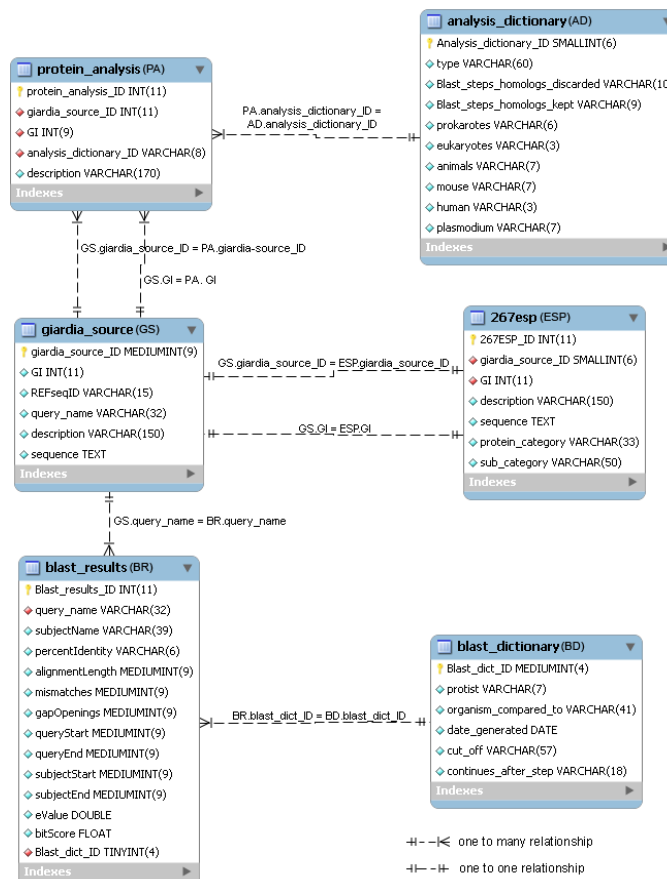


Figure 1. Giardia database layout.

3 Discussion

Hartman *et al.* collected of Giardia ESPs in 2001 and obtained 347 ESPs for Giardia [2]. We compared our set of ESP against Hartman *et al.*'s dataset. The results showed out of our 267 Giardia ESPs, 208 proteins had homologues in Hartman *et al.*'s set, and 59 did not. The main cause of this variation is the difference in the methods the ESPs are calculated. Hartman *et al.* started with the yeast proteome because the Giardia genome was very poorly annotated at that time, and performed BLAST search with these yeast proteins against 44 prokaryotes, then four eukaryotes and lastly Giardia; whereas we used the more straightforward approach and started our BLAST searches with Giardia proteins. There is also a difference in the threshold used for deciphering whether the hits were considered as homologues, we used a less stringent cut-off because the Giardia proteins are more divergent than yeast proteins when they are aligned with proteins from other eukaryotes [1].

With our study we are not restricted to highly researched model organisms or predefined proteomic datasets but can undertake ESP analysis on more medically or evolutionary important species. The ESP calculation is just the preliminary work. We are currently performing a Giardia small RNA Solexa run and this data will be integrated into the ESP analysis. Metabolic pathways are presently being chosen for detailed investigations. These investigations will enable close analysis of protein and protein-protein interactions and possible differences between host and parasite proteomes. In future, our research will move towards a medical perspective, and we will focus on parasitic metabolism, observing the trend of their loss and gain of proteins, with the ultimate aim of enabling drug target selection.

References

- [1] Gillin, F.D., D.S. Reiner, and J.M. McCaffery, Cell biology of the primitive eukaryote Giardia lamblia. *Annual Review of Microbiology*, 50: p. 679-705,1996.
- [2] Hartman, H. and A. Fedorov, The origin of the eukaryotic cell: A genomic investigation. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3): p. 1420-1425, 2002.