

A Markov Model Based Gene Discrimination Approach in Trypanosomes

Allison Griggs

Rochester Institute of Technology

Department of Biological Sciences

85 Lomb Memorial Drive, Rochester, NY 14623

adg9398@rit.edu

Shuba Gopal

shuba@bioinformatics.rit.edu

The trypanosomes are a class of eukaryotic parasites that diverged from *Saccharomyces cerevisiae* about 800 million years ago. Many possible gene structures are present within these genomes but most appear to be non-functional and do not code for proteins. Initial analyses of these genomes suggest that over 70% of the putative genes have no biological function.¹

To provide confirmation for genes that are likely to be biologically relevant but which lack other sources of evidence (sequence homology based evidence is limited because of the evolutionary distance from these organisms to their better-studied counterparts), we developed a method that takes advantage of unusual signals in immediately upstream regions of known trypanosome genes. This signal is required for mRNA maturation prior to translation. Our method uses Markov models and linear discriminant analysis to compare these upstream regions with coding regions and identifies coding regions most likely to be truly functional. We have been able to identify true coding regions in *Trypanosoma brucei* with 93% accuracy (96% sensitivity and 90% specificity). The related organism, *Leishmania major*, is 300 million years diverged from *T. brucei* yet our approach is able to identify true coding regions with a similar accuracy 91% (91% sensitivity and 92% specificity).

Our approach significantly improves on existing methods. Current approaches have an error discovery rate² of 0.21 [1]; the approach presented here has an error discovery rate of 0.08. Our success in these organisms suggests such an approach may be applicable to other organisms that share aspects of trypanosome biology.

cation in the *Leishmania* genome project. *BMC Bioinformatics*, 4:23, 2003. Web: <http://www.biomedcentral.com/1471-2105/4/23>.

References

- [1] G. Aggarwal, E. Worthey, P. McDonagh, and P. Myler. Importing statistical measures into Artemis enhances gene identi-

¹ Based on public annotations of Chromosome I of *Trypanosoma brucei* and *Leishmania major*.

² Error discovery rate is calculated as the total false positives and false negatives over the total ORFs analyzed (FP+FN/ORFs).