BMC Bioinformatics

BioMed Central

Research article

Importing statistical measures into Artemis enhances gene identification in the *Leishmania* genome project

Gautam Aggarwal[1], EA Worthey[1], Paul D McDonagh[2] and Peter J Myler*[1,3]

Address: [1]Seattle Biomedical Research Institute 4 Nickerson Street, Seattle, WA 98109, USA, [2]Immunex Corporation, 51 University Street, Seattle, WA 98101, USA and [3]Departments of Pathobiology and Medical Education and Biomedical Informatics, University of Washington, Seattle, WA 98195, USA

Email: Gautam Aggarwal - gaggarwal@sbri.org; EA Worthey - lworthey@sbri.org; Paul D McDonagh - pmcdonagh@rii.com; Peter J Myler* - peter.myler@sbri.org

* Corresponding author

Published: 7 June 2003

BMC Bioinformatics 2003, **4**:23

Received: 19 February 2003
Accepted: 7 June 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/23

Abstract

Background: Seattle Biomedical Research Institute (SBRI) as part of the *Leishmania* Genome Network (LGN) is sequencing chromosomes of the trypanosomatid protozoan species *Leishmania major*. At SBRI, chromosomal sequence is annotated using a combination of trained and untrained non-consensus gene-prediction algorithms with ARTEMIS, an annotation platform with rich and user-friendly interfaces.

Results: Here we describe a methodology used to import results from three different protein-coding gene-prediction algorithms (GLIMMER, TESTCODE and GENESCAN) into the ARTEMIS sequence viewer and annotation tool. Comparison of these methods, along with the CODONUSAGE algorithm built into ARTEMIS, shows the importance of combining methods to more accurately annotate the *L. major* genomic sequence.

Conclusion: An improvised and powerful tool for gene prediction has been developed by importing data from widely-used algorithms into an existing annotation platform. This approach is especially fruitful in the *Leishmania* genome project where there is large proportion of novel genes requiring manual annotation.

Background

At Seattle Biomedical Research Institute (SBRI), we are involved, as part of the *Leishmania* Genome Network (LGN), in the sequencing and annotation of the trypanosomatid protozoan species *L. major* Friedlin (LmjF). Following DNA sequence determination, putative protein-coding regions within the sequence are predicted and functionally classified. Although trypanosomatids are eukaryotes, their gene structure is more similar to that of prokaryotes; they have essentially no introns and small intergenic regions. Two small LmjF chromosomes (chr1 and chr3) have been completely sequenced and annotated. The 79 protein-coding genes predicted from chr1 are organized in two large divergent polycistronic gene clusters of 29 and 50 genes, on the "bottom" and "top" DNA strains, respectively [1]; while chr3 contains two convergent polycistronic clusters of 65 and 29 genes, with a single divergent gene at one telomere and a single tRNA between the two large clusters [2].

Presently, a large number of methods exist for *in silico* prediction of coding regions [3–7]. These computational methods use a range of underlying statistical properties of the coding regions and can be generally classified as

consensus (signal sensors) and non-consensus (content sensors) [8,9]. The non-consensus methods can be further classified as trained, which require unbiased sets of coding regions, and untrained, which use statistical properties to discriminate between coding and non-coding regions. Although non-consensus methods have been very successful in identifying genes in most of the sequencing projects, currently none have 100% specificity and sensitivity. In the absence of such a method, the use of a combination of methods is next best option [10–13]. Since LmjF genes do not contain introns, and the signal sequences for trans-splicing and polyadenylation are poorly defined, consensus methods have little utility for *Leishmania* gene prediction. In addition, ~70% of the genes have no significant homology to existing genes in sequence databases, so extrinsic content sensing methods are of limited use; leaving only intrinsic content sensing methods for possible use in gene prediction. Given that the number of experimentally confirmed gene prediction in *Leishmania* is currently small, and many methods use similar statistical approaches [4], the choice of two trained methods (GLIMMER[14] and CODONUSAGE[15]) and two untrained methods (TESTCODE[16], and GENESCAN[17]) which rely on unrelated statistical measures should provide substantial power for gene prediction in LmjF.

The freely available JAVA-based software package ARTEMIS[18] was designed specifically as an annotation platform and has a user-friendly graphical interface. It simplifies time-consuming processes such as inter-file format conversion, BLAST analysis [19], and provides a convenient environment for viewing the gene structure and organization of large DNA segments. Here we describe a method for importing data from GLIMMER, TESTCODE, and into GENESCAN into ARTEMIS, to enhance gene prediction and annotation.

Results and Discussion

We have developed a partially automated process for prediction and annotation of LmjF protein-coding genes in which the gene predictions from GLIMMER and the statistical outputs from TESTCODE and GENESCAN are imported into ARTEMIS (see additional file 1), where they can be viewed graphically alongside the CODONUSAGE statistics already built into ARTEMIS. Figure 1 shows a panel containing results from each of the four gene-prediction methods for a typical LmjF sequence. The predictions from GLIMMER are imported as CDS features and displayed as colored rectangles in the panel showing ORFs (the vertical bars are the stop codons) in all six reading frames. The window scans from TESTCODE, GENESCAN and CODONUSAGE are displayed graphically in panels above the GLIMMER predictions. The thresholds used to indicate likely protein-coding ORFs for TESTCODE and GENESCAN are 4.0 and 9.7, respectively. This allows visual comparison of the four gene prediction methods and manual alteration of the GLIMMER-predicted CDS features if necessary. The reliance on multiple gene prediction methods increases confidence in the predictions.

In Table 1, we show a comparison of the results of automated gene prediction using the four different programs with the manual annotations for three completely sequenced chromosomes (chr1, chr3 and chr4) from LmjF. The False Positive rate for each individual method was quite high, with GLIMMER being significantly worse than the others. Most of the False Positives were due to prediction of genes on the wrong coding strand. All methods, with the exception of TESTCODE, showed a low number of False Negatives. The poor performance of TESTCODE was largely due to use of a high cut-off value (9.7) for the average Fickett statistic of the whole ORF, rather than smaller windows. Thus, individually, each of the automated programs had high Error Discovery Rates (fraction of incorrect predictions made for expected predictions, Table 1), ranging from 0.77 for GENESCAN to 1.96 for GLIMMER.

Combination of the programs improved the Error Discovery Rate, especially in terms of false positives (Table 2). When only ORFs predicted by all four programs are considered, the false positive rate was <1%, but the false negative rate was almost 50%. By including ORFs predicted by only three of the four programs, the false negative rate was dramatically lowered to 10%, but the false positive rate rose to >10%. Further relaxation of stringency (two of four programs) resulted in a substantial increase in false positives (78%), with only modest decrease in false negatives (~5%). Thus, the Error Discovery Rate is least (21%) by considering the consensus prediction of three out of four programs. The use of two trained (GLIMMER and CODON USAGE), and two non-trained (TESTCODE and GENESCAN) algorithms reduced false positives and false negatives.

Conclusions

The semi-automated comparative analysis clear shows that some degree of manual annotation is still necessary in projects where there is large proportion of novel genes. The manual annotation is time consuming and labor intensive. The ARTEMIS desktop environment, with importation of trained and non-trained non-consensus gene-prediction algorithms, facilitates easy comparison of the results and allows the user to make more-informed decisions for calling protein-coding genes. Thus, this improvised and powerful software, developed using already existing gene identification methods and annotation platform, is extremely helpful for whole genome sequencing projects.

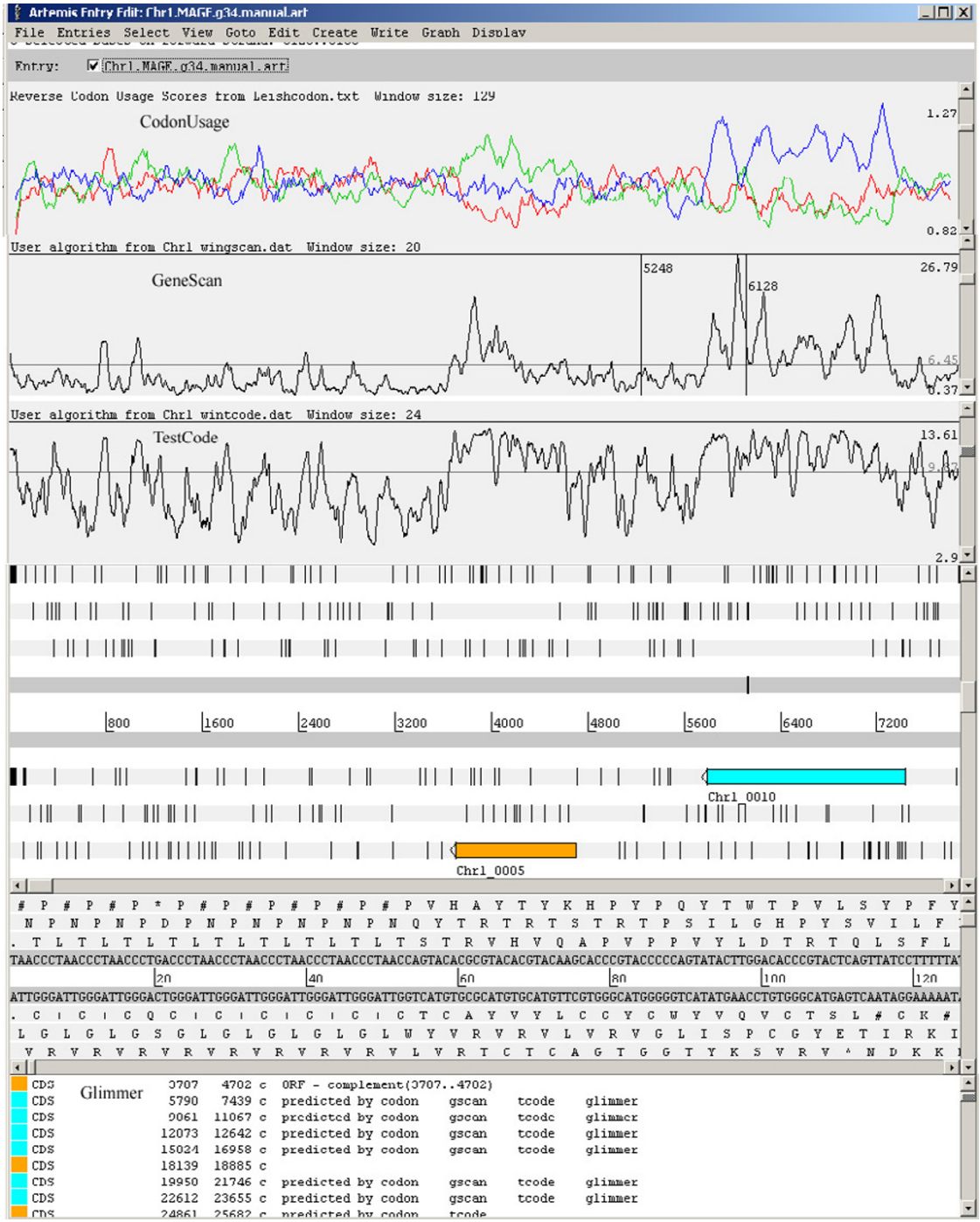

Figure 1

This panel of ARTEMIS shows the comparison of four different methods used at SBRI for sequence annotation: a) CODONUSAGE b) GENESCAN c) TESTCODE and d) GLIMMER. The CODONUSAGE panel shows results for the three reading frames (shown by different colors) of the top strand; those from the bottom strand are not shown. The panel immediately following the TESTCODE panel displays the position of all stop codons (with vertical lines) in all six reading frames. The vertical scales in the top three panels refer to the value of the statistic calculated by the corresponding algorithm. The predictions of GLIMMER appear as blue boxes in this panel. The horizontal scale in the center of this panel indicates the nucleotide coordinates of the sequence for this and the three upper panels (and is adjustable on the right hand scroll bar). The bottom panel displays the translated amino acids in six different reading frames. The horizontal scale refers to the nucleotide coordinates for the sequence within this panel.

Table 1: Automated gene prediction[a] in *Leishmania major*

	Annotated CDS[b]	GLIMMER		GENESCAN		TESTCODE		CODONUSAGE	
		FP[c]	FN[d]	FP	FN	FP	FN	FP	FN
Chr1	79	131	0	61	1	68	33	75	4
Chr3	94(1)	116	1	57	5	119	51	108	8
Chr4	123	328	1	97	6	130	56	139	9
Total	**295**	**575**	**2**	**215**	**12**	**317**	**180**	**322**	**21**
EDR[e]		1.96		0.77		1.68		1.16	

[a] All possible ORFs (*i.e.* starting with an ATG and ending with TAA, TAG or TGA) of >300 bp in the three chromosome sequence were scored by each of the programs. GLIMMER predictions (for ORFs > 100 amino acids, with default settings) were taken straight from the trained software. For GENESCAN and TESTCODE, ORFs were considered to be positive if the average score for the ORF exceeded a threshold of 4.0 and 9.7, respectively. For overlapping ORFs on the same strand, that with the highest score was chosen. In case of CODONUSAGE, ORFs were predicted as coding when the average in-frame score was higher than the two out-of-frame scores. [b] The number of CDS of more than 300 bp in GenBank Accession numbers AE001274 (chr1), AC125735 (chr3), AL389894 and AL139794 (chr4). The number of annotated CDS of <300 bp are shown in parentheses. [c] False positives [d] False negatives [e] Error Discovery Rate (EDR) = (FN+FP)/(CDS)

Table 2: Automated gene prediction by combination of different methods.

Chr	Annotated CDS	4 methods		3 methods		2 methods	
		FP	FN	FP	FN	FP	FN
Chr1	79	0	34	13	5	65	1
Chr3	90	1	50	7	10	50	5
Chr4	123	1	58	14	13	109	6
Total	**295**	**2**	**142**	**34**	**28**	**224**	**12**
EDR		0.49		0.21		0.80	

Methods

GLIMMER 2.0 http://www.tigr.org/software/glimmer/ [14] was trained using predicted protein-coding genes from LmjF chr1 [1] (manual annotations based on TESTCODE and CODON USAGE) and chr4 (manual annotations using HEXAMER and CODON USAGE: A. Ivens, personal communication) using the default settings. The trained GLIMMER was run on LmjF sequence using the default setting with a minimum gene length of 75 amino acids and output was parsed into an EMBL-formatted feature table file. This data were imported into ARTEMIS 4.0 (installed on Intel-based Linux or Windows 2000 machines) using the "Read Features Into" option of the "File" menu. This allows the GLIMMER-predicted genes to be displayed as CDS Features. The TESTCODE[16], GENESCANhttp://202.41.10.146/public_htmlnew/gs.htm[17] and CODONUSAGE[15] algorithms were re-coded in C++ and the statistical results collected in text files with single value for each sliding window (100 nt windows, sliding by onent increments). These TESTCODE and GENESCAN data were imported into ARTEMIShttp://www.sanger.ac.uk/Software/Artemis/[18] using the "Add User Plot" option of the "Display" menu, and displayed graphically. This procedure can be used to import other sliding window methods. The CODON USAGE bias statistics, which has been coded as part of ARTEMIS, is calculated for the three reading frames of each DNA strand and displayed in different colors using the "Add Usage Plot" option of the "Display" menu to import *Leishmania* CODON USAGE tables. Figure 1 shows a panel containing results from each of the four gene-prediction methods for a typical LmjF sequence.

For automated GENESCAN, TESTCODE and CODONUSAGE predictions, genes were called only for those ORFs larger than 100 amino acids with mean scores (over the entire ORF) above thresholds of 4.0, 9.7, and 0, respectively. For overlapping ORFs (on the same or opposite strands), the one with the highest signal was used.

Authors' contributions

GA re-coded the TESTCODE, GENESCAN and **CODONUSAGE** algorithms in C++ for UNIX environment and performed the automated combined prediction analysis. PDM coded the wrapper for parsing the GLIMMER predictions. All authors read and approved the final manuscript.

Additional material

Additional File 1

This is a zip file that contains one perl script (glimmer_atremis.pl), two (testcode_unix and testcode_win.exe) executable files and a readme.txt file describing the details of usage and other information relevant to the programs.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-23-S1.zip]

Acknowledgements

The authors thank Kim Rutherford (Wellcome Trust Sanger Institute) for the help and useful discussion. This work was supported by NIH grant AI40599.

References

1. Myler PJ, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, Magness C, Rickell E, Sisk E and Sunkin S: ***Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes** *Proc Natl Acad Sci U S A* 1999, **96:**2902-2906.
2. Worthey E, Aggarwal G, Cawthra J, Fazelinia G, Fu G, Hassebrock M, Hixson G, Ivens AC, Kiser P and Marsolini F: ***Leishmania major* chromosome 3 contains two long "convergent" polycistronic gene clusters separated by a tRNA gene** *Nucl Acids Res* .
3. Claverie JM: **Computational methods for the identification of genes in vertebrate genomic sequences** *Hum Mol Genet* 1997, **6:**1735-1744.
4. Fickett JW: **The gene identification problem: an overview for developers** *Computers Chem* 1996, **20:**103-118.
5. Guigo R: **DNA composition, codon usage and exon prediction** *In Genetics Databases* Edited by: *Bishop M. San Diego: Academic Press, Inc*; 1999:53-80.
6. Jones J, Field JK and Risk JM: **A comparative guide to gene prediction tools for the bioinformatics amateur** *Int J Oncol* 2002, **20:**697-705.
7. Mathe C, Sagot MF, Schiex T and Rouze P: **Current methods of gene prediction, their strengths and weaknesses** *Nucl Acids Res* 2002, **30:**4103-4117.
8. Stormo GD: **Gene-finding approaches for eukaryotes** *Genome Res* 2000, **10:**394-397.
9. Burge CB and Karlin S: **Finding the genes in genomic DNA** *Curr Opin Struct Biol* 1998, **8:**346-354.
10. Aggarwal G and Ramaswamy R: ***Ab initio* gene identification: prokaryote genome annotation with Genescan and Glimmer** *J Biosci* 2002, **27:**7-14.
11. Yada T, Takagi T, Totoki Y, Sakaki Y and Takaeda Y: **DIGIT: a novel gene finding program by combining gene-finders** *Pac Symp Biocomput* 2003:375-387.
12. Pavloviç V, Garg A and Kasif S: **A Bayesian framework for combining gene predictions** *Bioinformatics* 2002, **18:**19-27.
13. Howe KL, Chothia T and Durbin R: **GAZE: a generic framework for the integration of gene-prediction data by dynamic programming** *Genome Res* 2002, **12:**1418-1427.
14. Delcher AL, Harmon D, Kasif S, White O and Salzberg SL: **Improved microbial gene identification with GLIMMER** *Nucl Acids Res* 1999, **27:**4636-4641.
15. Staden R and McLachlan AD: **Codon preference and its use in identifying protein coding regions in long DNA sequences** *Nucl Acids Res* 1982, **10:**141-156.
16. Fickett JW: **Recognition of protein coding regions in DNA sequences** *Nucl Acids Res* 1982, **10:**5303-5318.
17. Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S and Ramaswamy R: **Prediction of probable genes by Fourier analysis of genomic sequences** *Comput Appl Biosci* 1997, **13:**263-270.
18. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A and Barrell B: **Artemis: sequence visualisation and annotation** *Bioinformatics* 2000, **16:**944-945.
19. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ: **Basic local alignment search tool** *J Mol Biol* 1990, **215:**403-410.