Automatic annotation of organellar genomes with DOGMA

Stacia K. Wyman
stacia@cs.utexas.edu
Department of Computer Sciences
University of Texas at Austin
Austin, TX 78712

Jeffrey L. Boore
jlboore@lbl.gov
DOE Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598

Robert K. Jansen
jansen@mail.utexas.edu.
Section of Integrative Biology
University of Texas at Austin,
Austin, TX 78712

Abstract

Dual Organellar GenoMe Annotator (DOGMA) automates the annotation of extra-nuclear organellar (chloroplast and animal mitochondrial) genomes. It is a web-based package that allows the use of comparative BLAST searches to identify and annotate genes in a genome. DOGMA presents a list of putative genes to the user in a graphical format for viewing and editing. Annotations are stored on our password-protected server. Complete annotations can be extracted for direct submission to GenBank. Furthermore, intergenic regions of specified length can be extracted, as well as the nucleotide sequences and amino acid sequences of the genes.
URL: *http://phylocluster.biosci.utexas.edu/dogma/*
Keywords: *annotation, organelles, chloroplasts, mitochondria.*

1 Introduction

The comparison of complete organellar genome sequences is becoming increasingly important for reconstructing the evolutionary relationships of organisms [2, 3, 7, 8], for studying population structure and history [11], including those of humans [6], for identifying forensic materials [10], and for understanding the inheritance of certain human diseases [12]. Identifying and annotating genes is currently a time consuming and error fraught process and, with the input of high-throughput genome sequencing centers, is becoming the rate-limiting step in the production of complete chloroplast and mitochondrial genome sequences. For extra-nuclear organellar genomes, gene content and function is largely known, and annotation involves locating and identifying the set of known genes, and clearly, an automated and accurate method such as DOGMA is an invaluable tool. We also may be able to use this program as a model on which to base methods for automating annotation of other genomes.

DOGMA is a web-based annotation package that takes as input a file containing the complete nucleotide sequence of an animal mitochondrial or chloroplast genome in Fasta format. The genome is BLASTed against our custom databases constructed from all the genes from a set of animal mitochondrial and green plant chloroplast genomes. DOGMA constructs a list of genes from the BLAST output, and graphically displays the list of genes to the user for annotation. The putative genes are laid out on a number line, and when the gene is selected, a detailed view of the gene's sequence and BLAST hits is displayed. The user can then choose a start and stop codon for each protein coding gene, and a begin and end position for each transfer RNA (tRNA) and ribosomal RNA (rRNA) in the genome. Annotations are stored on our password-protected server so they can be retrieved and edited. When complete, the annotation may be retrieved in Sequin format for direct submission to GenBank. Additionally, intergenic regions of specified length can be extracted, as well as the nucleotide and amino acid sequences of the genes.

2 Background

Organelles are membrane-bound structures in the cell that carry out various functions. Two organelles, chloroplasts and mitochondria, have circular, double-stranded chromosomes with an almost completely known set of genes.

Animal mitochondrial genomes Animal mitochondrial genomes typically are about 15,000 basepairs (bp) in length and contain 37 genes: 13 protein coding genes, 22 transfer RNAs (tRNAs) and 2 ribosomal RNAs (rRNAs) [1]. Gene content is mostly fixed, though the gene order can be highly rearranged. Duplications or deletions of genes are rare, most genes do not overlap (though there are some well-identified exceptions), and genes do not contain introns. At the time of writing, there were 467 complete, annotated, animal mitochondrial genomes in GenBank.

Chloroplast genomes Chloroplast genomes, on the other hand, are usually about 150,000 bp (but can be as long as 220,000 bp) and contain 110-130 genes [9]. There are 4 ribosomal RNA genes, about 30 transfer RNAs and about 80 protein coding genes. Introns are infrequent in chloroplast genomes, occurring in *Nicotiana* in 20 genes. Chloroplast genomes contain 4 distinct regions. Two of the re-

gions (IRA and IRB) involve a large inverted repeat. The other two regions are the large and small single-copy regions. In general, gene content and order are highly conserved [9], although in some groups numerous structural rearrangements have been identified [4]. Some genes can contain large introns or even contain tRNA genes *within* the intron. There are currently 26 complete plant chloroplast genomes in GenBank and 18 of these are green plants.

3 Details

Databases We created custom databases for select chloroplast and all animal mitochondrial genomes. For the animal mitochondrial database, we downloaded the complete genomes for 243 organisms (the total number in GenBank at the time) and extracted the annotated genes to compile a database for each individual protein coding gene. Each database contains the amino acid sequence for a specific gene from each of the genomes in which it appears. There are databases for each of the 13 protein coding genes, plus one for each of the two rRNA genes.

For chloroplast genomes, we created the database from 16 complete genomes of green plants. They include *Adiantum*, *Arabidopsis*, *Chlorella*, *Epifagus*, *Lotus*, *Marchantia*, *Mesostigma*, *Nephroselmis*, *Nicotiana*, *Oenothera*, *Oryza*, *Pinus*, *Psilotum*, *Spinacia*, *Triticum*, and *Zea*. Database files were created for 98 chloroplast protein coding genes (with two entries for the each of the trans-spliced pieces of rps12). We did not include open reading frames (ORFs) in the database, however, hypothetical chloroplast reading frames (ycfs) were included. There are 4 rRNA nucleotide sequence database files and 35 nucleotide sequence database files for tRNA genes. Transfer RNA database files were divided up based on anticodon. Compilation of the chloroplast gene databases required correcting many errors in the Genbank entries.

Identifying protein coding genes Protein coding genes are identified in the input genome based upon conservation of sequence similarity with that gene in other genomes in the database. The input nucleotide sequence is translated to amino acids and then BLASTed against the database for each gene. Various BLAST parameters may be set by the user based upon their data. Once DOGMA has identified the putative protein coding genes, the user must select start and stop codons for each gene as part of the annotation process. For each gene in the genome, the program displays to the user the nucleotide sequence for the gene from the input genome with the translation to amino acids, along with the amino acid sequences from the BLAST hits. For genes containing introns, DOGMA will identify exon boundaries based upon the BLAST output. DOGMA compiles a list of possible exons for the user to put together in the annotation. This has proven to work quite well, however, genes with very small exons (less than 5 amino acids) may be difficult to annotate with DOGMA because they can be missed by BLAST.

Identifying tRNAs In chloroplast genomes, the nucleotide sequences for tRNAs are highly conserved. We have found that sequence similarity is sufficient for detection of tRNAs in chloroplast genomes. Databases of nucleotide sequences are used for searching with BLAST. The tRNA databases are divided up based on anticodons and several databases may exist for one amino acid if there is more than one anticodon. In this way, the anticodon of the tRNA is annotated for the user.

Transfer RNAs have almost no sequence similarity in animal mitochondrial genomes and therefore, sequence similarity is not a sufficient criterion to locate the genes, and BLAST cannot be used. Animal mt tRNAs must be identified based on conservation of basepairing in the secondary structure. This is a difficult task, but we have found that methods based on hidden Markov models work best [13] and DOGMA uses COVE [5] to identify sequences and then uses our own program to fold the putative tRNAs.

Identifying rRNAs Ribosomal RNAs can be detected through BLAST searches for sequence similarity for both mitochondrial and chloroplast genomes. For mt genomes, the BLAST parameters (such as gap penalty or percent identity) must be adjusted since there can be significant degradation in the quality of the match in the middle of the rRNAs.

4 Web-based Display and Editing Tool

The package contains a web-based display and editing tool (Figure 1). The tool consists of three panels. The main (middle) panel displays all the genes along a number line. The genes are color coded by strand and gene type and labeled with the gene name. When a gene in the middle panel is selected, details about that gene appear in the top panel. DOGMA displays both strands of the genome's nucleotide sequence, with the translation to amino acids lined up above the nucleotides and the BLAST hits lined up above that. The potential start and stop codons for the gene appear as links. To choose a stop codon, the user simply clicks on the codon. The taxon names are displayed to the far right and when clicked, the gene's amino acid sequence from the database is shown in a separate window. When the gene name is clicked, the original blast output is displayed in a separate window. For annotation of rRNAs and chloroplast tRNAs, DOGMA functions similarly to the protein coding genes, except that nucleotide sequences are displayed rather than amino acid sequences, and the user chooses the start and end of the gene. Animal mitochondrial tRNAs are notoriously difficult to annotate. DOGMA uses Eddy and Durbin's COVE software to identify a list of putative tRNAs. When a tRNA is clicked in the middle panel, a list of the possible tRNAs, with its secondary structure is shown in the top panel. The user can choose the tRNA based on

the quality of the secondary structure folding and its COVE score.

The bottom panel of DOGMA consists of a set of buttons for performing different tasks. The user can add or delete genes, extract intergenic sequences, extract amino acid or nucleotide sequences for genes, or get a summary of the genes or extract the Sequin format for direct submission of an annotation to GenBank.

When a researcher first uses the software, they create a user id and password which keeps their (perhaps unpublished) data private from other users of the software. Users may also save and retrieve existing annotations. This is especially important for annotating chloroplast genomes because there are so many genes.

5 Future Work

In the future, plant mitochondrial genomes will be added to DOGMA, as well as private custom databases for individuals. Researchers only interested in a subset of the database can identify the genomes they are interested in for comparison. This will also allow people interested in phylogenetic reconstruction to identify evolutionary similarity among genomes in anticipation of alignment of the whole genomes. It will also allow users to use their own unpublished data for annotation.

Future work also includes adding a feature to identify the gene order of a genome with respect to another reference genome. This will allow the genome to be automatically added to a phylogenetic reconstruction by gene order.

Acknowledgments SKW was supported by National Science Foundation IGERT grant 0114387. RKJ was supported by National Science Foundation grant DEB0120709. Part of this work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, in the University of California, Lawrence Berkeley National Laboratory, under contract No. DE-AC03-76SF00098.

References

[1] J. Boore. Animal mitochondrial genomes. *Nucleic Acids Research*, 27:1767–1780, 1999.

[2] J. Boore and W. Brown. Big trees from little genomes: Mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev.*, 8(6):668–674, 1998.

[3] Y. Cao, M. Fujiwara, M. Nikaido, N. Okada, and M. Hasegawa. Interordinal relationships and timesecale of eutherian evolution as inferred from mitochondrial genome data. *Gene*, 259:149–158, 2000.

[4] M. Cosner, R. Jansen, J. Palmer, and S. Downie. The highly rearranged chloroplast genome of *trachelium caeruleum* (campanulaceae): Multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Current Genetics*, 31:419–429, 1997.

[5] S. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22:2079–2088, 1994.

[6] M. Ingman, H. Kaessmann, S. Pääbo, and U. Gyllensten. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408:708–713, 2001.

[7] M. Martin, T. Rujan, T. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny. Evolutionary analysis of arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences*, 99:122246–12251, 2002.

[8] M. Miya, A. Kawaguchi, and M. Nishida. Mitogenomic exploration of higher teleostean phylogenies: A case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial DNA sequences. *Mol. Biol. Evol.*, 18:1993–2009, 2001.

[9] J. Palmer. Plastid chromosomes: structure and evolution. *Cell Culture and Somatic Cell Genetics of Plants*, 7A:5–53, 1991.

[10] T. Parsons and M. Coble. Increasing forensic discrimination of mitochondrial DNA testing through the analysis of the entire mitochondrial DNA genome. *Croatian Med. J.*, 42:304–309, 2001.

[11] D. Rand. The unites of selection on mitochondrial DNA. *Ann. Rev. Ecol. Syst.*, 32:415–448, 2001.

[12] D. Wallace. Mitochondrial diseases in man and mouse. *Science*, 283:482–488, 1999.

[13] S. Wyman and J. Boore. Annotating animal mitochondrial trnas: an experimental evaluation of four methods. *Proc. European Conf. on Computational Biology*, pages 44–46, 2003.

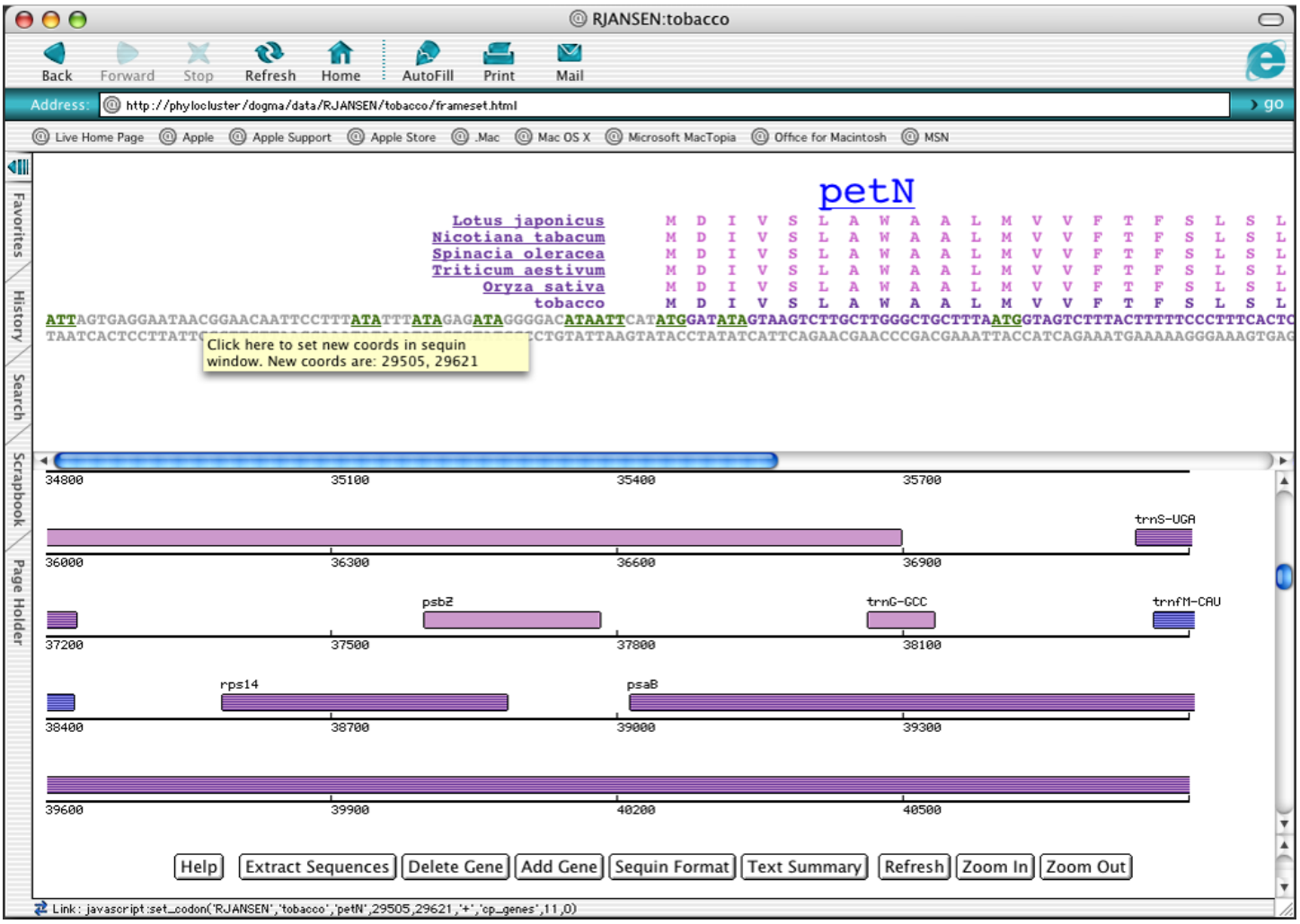

Figure 1. The main DOGMA annotation window showing the three panels.