BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 17 2004, pages 3252–3255
doi:10.1093/bioinformatics/bth352

Automatic annotation of organellar genomes with DOGMA

Stacia K. Wyman[1,*], Robert K. Jansen[2] and Jeffrey L. Boore[3]

[1]Department of Computer Sciences and [2]Section of Integrative Biology, Institute of Cellular and Molecular Biology, University of Texas at Austin, Austin, TX 78712, USA and [3]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

Received on April 6, 2004; revised on May 5, 2004; accepted on May 25, 2004
Advance Access publication June 4, 2004

ABSTRACT

Summary: The Dual Organellar GenoMe Annotator (DOGMA) automates the annotation of organellar (plant chloroplast and animal mitochondrial) genomes. It is a Web-based package that allows the use of BLAST searches against a custom database, and conservation of basepairing in the secondary structure of animal mitochondrial tRNAs to identify and annotate genes. DOGMA provides a graphical user interface for viewing and editing annotations. Annotations are stored on our password-protected server to enable repeated sessions of working on the same genome. Finished annotations can be extracted for direct submission to GenBank.

Availability: http://phylocluster.biosci.utexas.edu/dogma/

Contact: stacia@cs.utexas.edu

Supplementary information: Detailed documentation and tutorials for annotating both animal mitochondrial and plant chloroplast genomes can be found on the DOGMA home page.

1 INTRODUCTION

The comparison of complete organellar genome sequences is becoming increasingly important for many tasks, including reconstructing the evolutionary relationships of organisms (Boore and Brown, 1998; Cao *et al*., 2000; Martin *et al*., 2002; Miya *et al*., 2001), and understanding the inheritance of certain human diseases (Wallace, 1999). In the past, annotating organellar genomes has been a time-consuming and error-fraught process and, with the input of high-throughput genome sequencing centers, has been the rate-limiting step in the production of complete chloroplast and mitochondrial genome sequences. DOGMA is the first tool which automates this process.

DOGMA is a Web-based annotation package. It takes as input a file containing the complete nucleotide sequence of an animal mitochondrial or plant chloroplast genome in the FASTA format. For protein coding genes, the genome is translated in all six reading frames and queried against our custom amino acid sequence databases using BLAST (Altschul *et al*., 1990). Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are queried against a nucleotide sequence database. The databases were constructed from a set of annotated 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. When a gene is selected, a detailed view of the gene's nucleotide and amino acid sequence and BLAST hits is displayed. Because the putative genes are located using sequence similarity with genes in other genomes, and because BLAST is not explicitly looking for a start and stop codon for the protein coding genes, the start and stop codons must be chosen by the user. The start and end positions for each tRNA and rRNA also must be verified. Annotations are stored on our password-protected server so that they can be retrieved and edited. When complete, the annotation may be retrieved in Sequin format for direct submission to GenBank. DOGMA also allows the user to extract sub-sequences of the genome (including intergenic regions, introns, amino acid sequences of protein coding genes, etc.) for further analysis.

2 DOGMA DATABASES

Chloroplasts and mitochondria typically have circular, double-stranded chromosomes that vary little in gene content. This similarity in gene content allows us to use a comparative approach for annotations. Genes which have been previously identified in closely-related taxa can be used to identify those genes in newly sequenced genomes.

Animal mitochondrial genomes. Animal mitochondrial genomes typically are ∼15 000 bp in length and contain 37 genes: 13 protein coding genes, 22 tRNA and 2 rRNA genes (Boore, 1999). 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 (except for some cnidarians). The animal

[*]To whom correspondence should be addressed.

mitochondrial genome databases were compiled from the 367 annotated genomes in GenBank. Each database contains the amino acid sequence for a specific gene from each of the genomes in which it appears. There is a database for each of the 13 protein coding genes, plus one for each of the two rRNA genes.

Chloroplast genomes. Chloroplast genomes, on the other hand, are usually ∼150 000 bp (but can be as long as 220 000 bp) and contain 110–130 genes. There are 4 ribosomal RNA genes, ∼30 transfer RNAs and ∼80 protein coding genes. Introns are infrequent in chloroplast genomes, occurring in 20 genes in *Nicotiana*. In general, gene content and order are highly conserved (Palmer, 1991), although in some groups numerous structural rearrangements have been identified (Cosner *et al*., 1997). For chloroplast genomes, we created databases for genes 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 databases and 35 tRNA nucleotide sequence databases.

Gene nomenclature. GenBank is fraught with errors in annotation. These errors include typos, incorrect sequences and gene names, and inconsistencies in naming conventions. We have endeavored to clarify this issue by correcting and standardizing all gene names for both plant chloroplast and animal mitochondrial genomes. The naming conventions follow those set forth by Martin *et al*. (2002) for plastid genes and used by Boore (2000) for mitochondrial genes. In each case, these are generally the same as the gene names established for their bacterial homologs.

3 IDENTIFYING GENES

Protein coding genes. Protein coding genes are identified in the input genome based upon conservation of sequence similarity to genes in other genomes in the database. The input nucleotide sequence is queried in all six reading frames against the amino acid sequence database for each gene using BLASTX. Various BLAST parameters (such as E-value and number of hits returned) may be set by the user. Once DOGMA has identified the putative protein coding genes, the user then selects start and stop codons for each gene. 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 hit boundaries which must then be verified by the user. This has proven to work quite well; however, genes with very small exons (two or three amino acids) will be missed by BLAST and they must be located by hand. These exons occur in three well-documented genes in chloroplast genomes (petB, petD and rpl16), and the user will know to look for them when there is no start or stop codon for one of these three genes.

Identifying tRNAs. In chloroplast genomes, the nucleotide sequences for tRNAs are highly conserved, and we have found that sequence similarity is sufficient for their detection. Databases of nucleotide sequences for chloroplast tRNAs are used for searching with BLASTN. DOGMA identifies the anticodon for each tRNA based on the database entry.

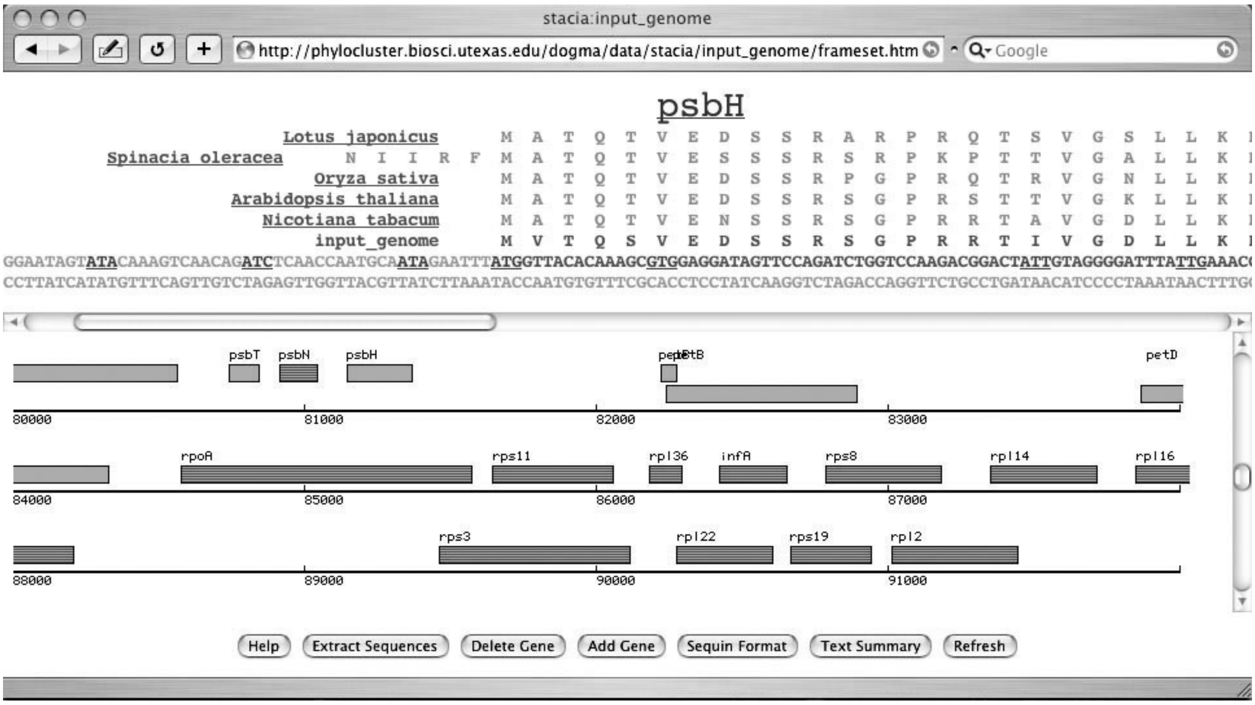
Transfer RNAs diverge rapidly in sequence in animal mitochondrial genomes and therefore, sequence similarity is not a sufficient criterion to locate the genes. They must be identified based on conservation of basepairing in the cloverleaf-shaped secondary structure. This is a difficult task, and we have found that methods based on hidden Markov models can do well (Wyman and Boore, 2003) and DOGMA uses the COVE (Eddy and Durbin, 1994) program. COVE identifies candidate sequences of the tRNA genes based on secondary structure, but does not give a putative folding, and so DOGMA then uses a custom program to infer the stem and loop folding of the secondary structure.

Identifying rRNAs. Ribosomal RNAs can be detected through BLAST searches for sequence similarity for both mitochondrial and chloroplast genomes. For mitochondrial genomes, the BLAST parameters (such as gap penalty or percent identity) must be optimized since some portions of the rRNA genes can be highly diverged.

4 WEB-BASED DISPLAY AND EDITING TOOL

DOGMA is a Web-based display and editing tool. On first use, a researcher creates a userid and password which keeps their (perhaps unpublished) data private from other users of the software. Users may also save and retrieve existing annotations.

The tool consists of three panels (Fig. 1). The main (middle) panel displays all the putative genes, and 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 for annotating that gene appear in the top panel and a new window appears for recording the annotation information for that gene for input to Sequin (NCBI's software for submitting annotations to GenBank). DOGMA displays the nucleotide sequence for both strands, with the translation to amino acids lined up above the nucleotides, and the amino acid sequences for the BLAST hits in the other taxa above that. All the potential in-frame start and stop codons for a gene appear as links and the user simply clicks on the codon to select it as the start or stop codon of the gene. For annotation of rRNAs and chloroplast tRNAs, DOGMA functions similarly to the protein coding genes, except that nucleotide sequences are displayed rather



stacia:input_genome
http://phylocluster.biosci.utexas.edu/dogma/data/stacia/input_genome/frameset.htm
Google
psbH

```
            Lotus japonicus           M  A  T  Q  T  V  E  D  S  S  R  A  R  P  R  Q  T  S  V  G  S  L  L  K
         Spinacia oleracea   N  I  I  R  F  M  A  T  Q  T  V  E  S  S  S  R  S  R  P  K  P  T  T  V  G  A  L  L  K
              Oryza sativa            M  A  T  Q  T  V  E  D  S  S  R  P  G  P  R  Q  T  R  V  G  N  L  L  K
      Arabidopsis thaliana            M  A  T  Q  T  V  E  D  S  S  R  S  G  P  R  S  T  T  V  G  K  L  L  K
         Nicotiana tabacum            M  A  T  Q  T  V  E  N  S  S  R  S  G  P  R  R  T  A  V  G  D  L  L  K
              input_genome            M  V  T  Q  S  V  E  D  S  S  R  S  G  P  R  R  T  I  V  G  D  L  L  K
GGAATAGTATACAAAGTCAACAGATCTCAACCAATGCAATAGAATTTATGGTTACACAAAGCGTGGAGGATAGTTCCAGATCTGGTCCAAGACGGACTATTGTAGGGGATTTATTGAAAC
CCTTATCATATGTTTCAGTTGTCTAGAGTTGGTTACGTTATCTTAAATACCAATGTGTTTCGCACCTCCTATCAAGGTCTAGACCAGGTTCTGCCTGATAACATCCCCTAAATAACTTTG
```

Fig. 1. The DOGMA annotation window showing details of the psbH gene in a chloroplast genome.

than amino acid sequences, and the user chooses the start and end of the gene. The user can also view the putative secondary structure of the tRNAs.

Animal mitochondrial tRNAs are notoriously difficult to annotate. DOGMA uses Eddy and Durbin's COVE software to identify a list of putative tRNA sequences and then tries to infer the secondary structure. When a tRNA is selected in the middle panel, a list of the possible tRNAs for that amino acid, with schematic drawings of its secondary structure, are shown in the top panel. The user can choose the tRNA based on the quality of the secondary structure folding and its COVE score.

5 FUTURE WORK

In the future, the chloroplast database will be expanded to include more taxa. The mitochondrial genomes of plants, fungi and protists will also be added to DOGMA, as well as private custom databases for individuals. Researchers only interested in a subset of the database will be able to identify the genomes they are interested in for comparison. This will also allow people interested in phylogenetic reconstruction to identify evolutionary similarity in anticipation of alignment of the whole genomes. It will also allow users to use their own unpublished data for annotation. We plan to construct a searchable database of folded tRNA structures for all organellar genomes as well as adding the capability to DOGMA of searching for tRNAs using a variety of types of methods. Future versions of DOGMA will additionally address the difficult issue of RNA editing of start and stop codons. There are also plans for including an ORF finder as a subroutine so putative new genes can be identified.

ACKNOWLEDGEMENTS

S.K.W. was supported by National Science Foundation IGERT grant 0114387. R.K.J. was supported by National Science Foundation grant DEB0120709. Part of this work was performed under the auspices of the US 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

Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Boore,J. (1999) Animal mitochondrial genomes. *Nucleic Acids Res.*, **27**, 1767–1780.

Boore,J. (2000) The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals. In Sankoff,D. and Nadeau,J. (eds), *Comparative Genomics*. Vol. 1, Kluwer Academic Publisher, Dordrecht, The Netherlands, pp. 133–147.

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

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

Cosner,M., Jansen,R., Palmer,J. and Downie,S. (1997) The highly rearranged chloroplast genome of *Trachelium caeruleum* (campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Curr. Genet.*, **31**, 419–429.

Eddy,S. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.

Martin,M., Rujan,T., Richly,T., Hansen,A., Cornelsen,S., Lins,T., Leister,D., Stoebe,B., Hasegawa,M. and Penny,D. (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl Acad. Sci., USA*, **99**, 12246–12251.

Miya,M., Kawaguchi,A. and Nishida,M. (2001) 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.

Palmer,J. (1991) Plastid chromosomes: structure and evolution. *Cell Culture Somatic Cell Genet. Plants*, **7A**, 5–53.

Wallace,D. (1999) Mitochondrial diseases in man and mouse. *Science*, **283**, 482–488.

Wyman,S. and Boore,J. (2003) Annotating animal mitochondrial tRNAs: an experimental evaluation of four methods. In *Proceedings of European Conference on Computer. Biology*, Local Proceedings, Self Published, Paris, France, pp. 44–46.