Automatic Annotation of Mitochondrial Genomes in Fungi

A. Salamov, I. Grigoriev
DOE Joint Genome Institute, Walnut Creek, CA 94598, US

The sizeable fraction of fungal mitochondrial protein-coding genes contain introns of type I or (rarely) of type II, which presents a challenge for their correct prediction. We have developed the annotation pipeline, which for the first time allows the computational prediction of such types of genes. When tested on 82 genomes from GenBank, the algorithm has the accuracy of 91%/88% (sensitivity/specificity) at nucleotide level, and 84%/79% on the exon/ORF level.

Keywords: *automatic annotation, intron type I and II, mitochondrial genome*

The revolution in genome sequencing technologies made it possible to cost effectively sequence thousands of eukaryotic genomes, including those from the kingdom of fungi. The fungi, estimated to include as many as 1.5 million of diverse species (Stajich et al., 2009), may have an important impact in solving problems related to energy and environment (Grigoriev et al., 2012).

Mitochondria are organelles that are present in almost all eukaryotes, including fungi, with the main role of production of energy for cell needs through oxidative phosphorylation or the citric acid cycle (Gray et al., 1999). Mitochondrial genomes are also playing an important role in phylogenetic and population genetics studies (Gissi et al., 2008).

While currently there are known several automated pipelines for annotation of eukaryotic genomes, such as JGI's pipeline (Martinez et al., 2010) and Broad Institute's fungal/eukaryotic annotation pipeline (Haas et al., 2011), to the best of our knowledge, no such automatic pipeline exist for annotation of mitochondrial genomes.

There are however some tools that facilitate the annotation process of organellar genomes, like for example DOGMA (Wyman et al., 2004), which is a web-based server for manually editing and annotating genes based on BLAST searches. But it does not produce automatically the list of predicted gene models, including genes, which contain introns.

The main challenge for automatic annotation of fungal mitochondrial genomes stems from the fact, that sizeable fraction of protein-coding genes contains introns, which are of different type, than the spliceosomal introns, predominantly present in eukaryotic nuclear genes.

Most gene finding algorithms were developed for predicting genes with spliceosomal introns, which have a strongly conserved consensus sequences at splice sites. In contrast, most introns in fungal mitochondria are predominantly of group I type, with rare occurrence also of group II introns (Lang et al., 2007).

It was the fungal mitochondrial genomes, where the group I introns were first characterized in early 1980s (Michel et al., 1982; Waring et al., 1982). Subsequently they also were found in some nuclear rRNA genes and some bacterial and plastid genomes. Group II introns on the other hand are most prevalent in plant mitochondrial genomes (Lang et al., 2007).

Some of group I and group II introns are mobile and move into intronless genes by mechanism called intron homing (Lang et al., 2007). Group I introns using for that purpose enzymes belonging to a very diverse family, termed LAGLIDADG homing endonucleases (Belfort and Perlman, 2005). ORFs, encoding homing endonucleases often reside inside introns itself, further complicating the computational prediction of genes containing such introns.

Although currently there are no software exist for prediction of protein-coding genes with group I or II introns, the number of tools were developed for prediction of introns itself.

As the RNA secondary structure is frequently conserved in these introns, most methods exploit that feature in their algorithms. For example RFAM database has 2 covariance models, based on training set of aligned RNA sequences for group I (RF00028/Intron_gpI) and group II (RF00029/Intron_gpII) introns respectively. (Griffiths-Jones et al., 2005). However we had found that the sensitivity of these models is too low to be used alone in the annotation pipeline (see below). Low sensitivity of RFAM models may stem back from the fact, that they consider only core features of group I and II introns, while it is known that their secondary structure is quite variable at peripheral parts (Michel and Westhof, 1990). Another popular tool frequently used for finding introns is RNAweasel (Lang et al, 2007). RNAweasel utilize RNA primary and secondary structure profiles, which are computed from training sets of RNA sequence alignments and user-defined secondary structure information. However it is available only through web-server

20 *E-mail: asaf.salamov@gmail.com

(http://megasun.bch.umontreal.ca/RNAweasel/), and not as a stand-alone tool, so we not tested it extensively. Besides as was noted by authors it may suffer from the same problems as RFAM models; to rectify that, they added new tools at their web server for manually editing and visualization of alignments.

ANALYSIS OF ANNOTATED FUNGAL MITOCHONDRIAL GENOMES FROM GENBANK

There are currently (as of Dec.2011) 82 complete fungal mitochondrial genomes deposited at the GenBank. Some characteristics of genomes are presented in Table 1. Genome size of fungal mitochondria ranges from 19 kb to 109 kb, with average size of 44 kb, which is about twice larger than for typical animal genome (~15-20 kb) (Boore, 1999). Genomes are preferentially AT-rich, with average GC content of 26% (Fig. 1). Typical mitochondrial genome contains about 19 protein-coding genes, 24 tRNAs and 2 rRNAs. The most often present protein families include different subunits of NADH dehydrogenase, ATP synthase and cytochrome C oxidase. About 10-15% of protein-encoding genes contain introns. The length of introns varies wildly from 15 bp to ~15 kb, with the bulk of introns having length in the range of 1 kb – 2k b (Fig. 2).

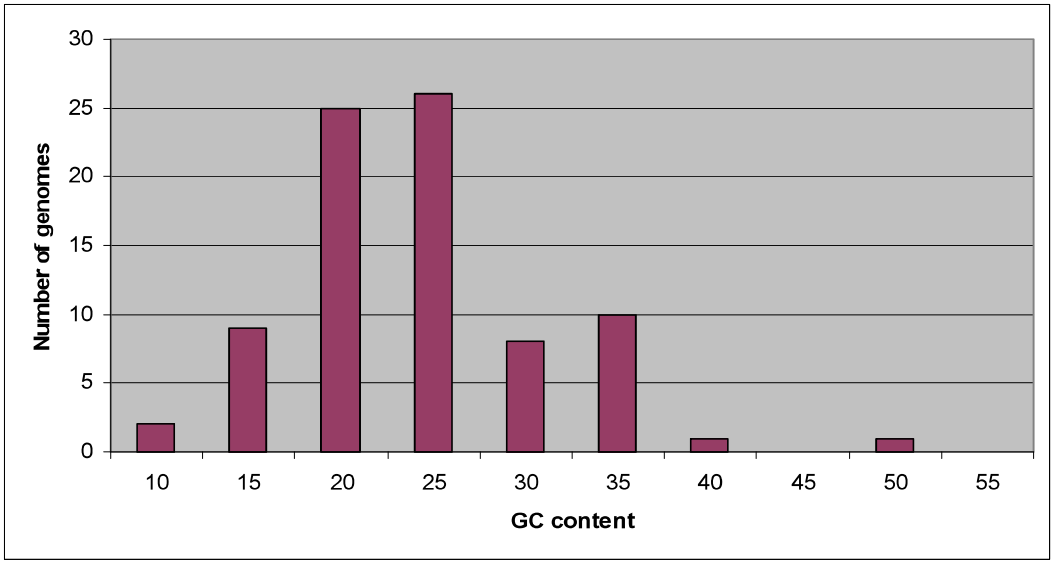

Fig. 1. The histogram of GC content across the 82 mitochondrial genomes from GenBank.

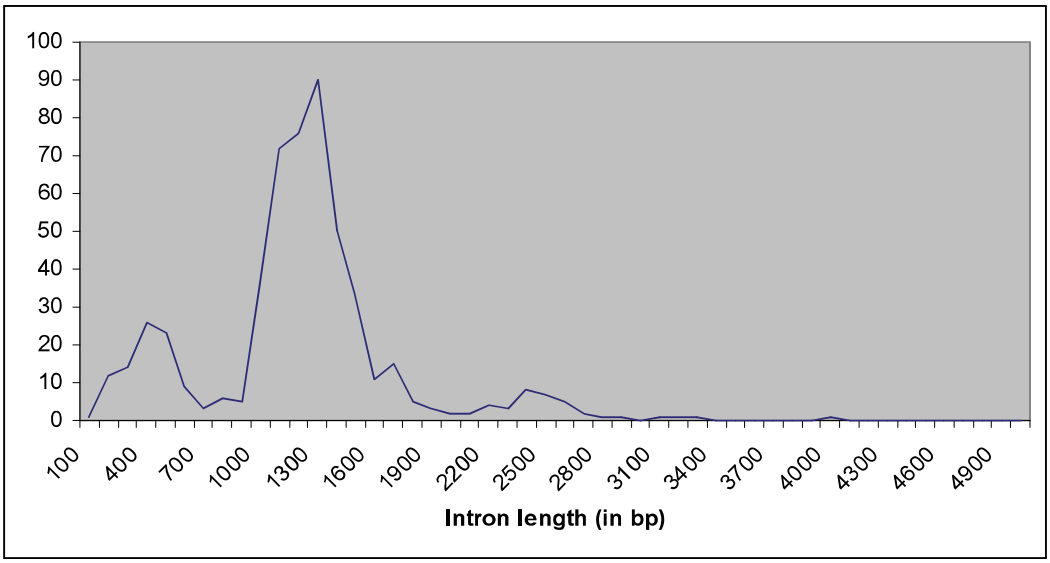

Figure 2. Intron length distribution in fungal mitochondrial genomes based on 82 genomes from GenBank.

Table 1. Characteristics of 82 fungal mitochondrial genomes from GenBank

	Average	Minimum	Maximum
Size of genome (kb)	44.1	18.8	109.1
GC content (%)	26	11	53
Number of protein-coding genes	19.4	8	89
Percent of intron-containing genes (%)	12.8	0	48
Number of exons per gene	3.7	1	17
Length of introns (bp)	1221	15	14968
Number of tRNAs	23.8	7	30
Number of rRNAS	2	0	4

Because 60 out of 82 genomes have intron-containing genes, we may assume that for most of them at least some kind of manual curation was conducted. At the same time significant fraction of introns (~30%) have no exact boundaries, with labels like '>' or '<'.

To test the accuracy of RFAM models, we searched for introns of group I and group II using 2 corresponding covariance models. In total just 40 introns were predicted in all 82 genomes (30 of group I type and 10 of group II), and from them 30 were partially overlapped with annotated introns. So while specificity of predictions was relatively high (75%), the sensitivity was unacceptably low, around 5%, even assuming that not all of annotated introns were correct.

AUTOMATIC ANNOTATION PIPELINE

We have developed an automated pipeline for predicting genes in fungal mitochondrial genomes. As the fungal mitochondria has two types of genes, prokaryotic-like single exon ORFs and genes interrupted by group I introns, we used 2 methods for predicting protein-coding genes in their genomes. The first method was similar to prokaryotic genes finder algorthms, like Fgenesb (Solovyev and Salamov, 2011), with using translation code and protein-coding potential specific for mitochondrial genomes. The second method utilizes homology-based approach to map intron-containing genes. Analysis of intron-containing genes shows that many of them are from a relatively few families, such as different subunits of NADH-dehydrogenases (NAD), ATP-synthases (ATP) and cytochrome oxidases (COX). We collected all the available mitochondrial proteins from GenBank and also from collection curated by F.Lang from U.Montreal (http://www.bch.umontreal.ca/People/lang/FMGP/proteins.html), removing redundant sequences.

Because all the current homology-based methods, like Genewise (Birney et al., 2004) were developed for prediction of genes with spliceosomal introns, they are not quite suitable for predicting mitochondrial intron-containing genes. So instead we used 'protmap' program from Softberry (www.softberry.com), which maps proteins to genome without consideration of splice site consensuses and then using custom-made Perl script refined the boundaries, preserving the reading frame. When 2 or more gene models were overlapped by their coding regions, the one with longest ORF was chosen, while preserving ORFs predicted entirely within an intron of another gene.

tRNAs genes were predicted using tRNAscan-SE with organellar option (Lowe and Eddy, 1997). Ribosomal RNAs present in mitochondria are usually short and not predictable by HMM-based methods, like RNAammer (Lagesen et al., 2007). So instead they were predicted using BLASTN against rRNA database.

We tested the pipeline accuracy on the set mitochondrial genomes from GenBank.

Because tRNA and rRNA genes usually were annotated by the same method as we were using in the pipeline, we restricted our analysis to the estimating of accuracy of predicted protein-coding genes.

Average accuracies on nucleotide and exon/ORF level based on 82 genomes are presented in Table 2. The accuracy was estimated by jackknife method, i.e. for each genome, gene prediction methods not used the information about annotated genes from that genome. In particular annotated proteins in predicted genome were not used by homology-based gene prediction. Because for about 30% of annotated introns in GenBank files their coordinates were not defined exactly, we considered predicted exons to be correct, when at least one boundary of it coincided with annotated exons. While overall accuracy at nucleotide and ORF level is acceptable, the accuracy for intron-

containing genes both on sensitivity and specifity level was around 50%. Without considering exact boundaries, at least 63% of all annotated introns were predicted because we cannot be sure in the accuracy of all annotated intron-containing genes, and taking into account the difficulty of prediction (f.e. availability of nested genes), these results can also be considered acceptable. We also should mention that this is the first algorithm that attempts to automatically predict protein-coded genes with type I or type II introns, so we cannot compare it with the analogous programs.

Table 2. The accuracy of gene finding on 82 mitochondrial genomes from GenBank

	Sn	Sp
Nucleotide level	91%	86%
Exons/ORF level (all genes)	84%	79%
Exons level (intron-containin genes)	50%	49%

The accuracy was estimated by jackknife method, i.e. for each genome, gene prediction methods not used data from that genome.
Sn – sensitivity = TP/TP + FN, Sp – specificity = TP/TP + FP, where TP – true positives, FN – false negatives, and FP – false positives

With appropriate modifications our pipeline can also be used for annotation of mitochondrial and organellar genomes in other eukaryotic phyla, like plants and protists.

ACKNOWLEGMENTS

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231.

REFERENCES

Belfort M., Perlman P.S. (2005) Mechanisms of intron mobility. J. Biol. Chem. **270:** 30237-30240.

Birney E., Clamp M., Durbin R. (2004) GeneWise and genomewise. Genome Research, **14:** 988-995.

Boore J.L. (1999) Animal mitochondrial genomes. Nucl. Acid Res. **27:** 1767-1780.

Gissi G., Iannelli F., Pesole G. (2008) Evolution of the mitochondirial genome of Metazoa as exemplified by comparison of congeneric species. Heredity **101:** 301-320.

Gray M.W., Burger G., Lang B.F. (1999) Mitochondrial evolution. Science **283:** 1476 - 1481.

Griffiths-Jones S., Moxon S., Marshall M. et al. (2005) Rfam: Annotating Non-Coding RNAs in Complete Genomes. Nucl. Acid Res. **33:** D121-D141.

Grigoriev I.V., Nordberg H., Shabalov I. et al. (2012) The Genome Portal of the Department of Energy Joint Genome Institute. Nucl. Acids Res. **40:** D26-D32.

Haas B.J., Zeng Q., Pearson M.D., Cuomo C.A., Wortman J.R. (2011) Approaches to Fungal Genome Annotation. Mycology **2:** 118-141.

Michel F. et al. (1982) Comparison of fungal mitochondrial introns reveals extensive homologues in RNA secondary structure. Biochimie **64:** 867-881.

Michel F., Westhof E. (1990) Modeling of the three-dimensional architecture group I catalytic introns based on comparative sequence analysis. J. Mol. Biol. **216:** 585-610.

Lang B.F., Laforest M.-J., Burger G. (2007) Mitochondrial introns: a critical view. Trend. Gen. **23:** 119-125.

Lagesen K., Hallin P., Rodland E.A. et al (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucl. Acid Res. **35:** 3100-3108.

Lowe T.M., Eddy S.R. (1997) tRNAscan-SEL a program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acid Res., **25:** 955-964.

Martinez D., Grigoriev I., Salamov A. (2010) Annotation of protein-coding genes in fungal genomes. Appl. Comput. Mathem. **9:** 55-65.

Solovyev V., Salamov A. (2011) Automatic annotation of microbial genomes and metagenomic sequences. In: Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies" (Li R.W., ed.): 61-78.

Stajich J. (2009) The Fungi. Curr. Biol. **29:** 840-845.

Waring R.B. et al. (1982) Internal structure of a mitochondrial intron of *Aspergillus nidulans*. Proc. Natl. Acad. Sci. USA **79:** 6332-6336.

Wyman S., Jansen R.K., Boore J.L. (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics **20:** 3252-3255.

A. Salamov, I. Grigoriyev

Göbələklərdə Mitoxondrial Genomlarının Avtomatlaşdırılmış Annotasiyası

Göbələklərin zülal kodlaşdıran mitoxondrial genlərinin xeyli hissəsi I və (nadir hallarda) II tip intronlara malikdirlər ki, bu da onların korrekt şəkildə öncədən prognozlaşdırılmasına əngəl yaradır. İlk dəfə bizim tərəfimizdən bu tip genlərin kompüterlə öncə proqnozlaşdırılmasına imkan verən annotasiya məlumat sistemi işlənib hazırlanmışdır. GenBank-da yerləşən 82 genomun testləşdirilməsi zamanı alqoritm nukletid səviyyəsində 91%/88% (həssaslıq/spesifiklik) və ekzon/ORF səviyyəsində 84%/79% dəqiqliyə malik olmuşdur.

А. Саламов, И. Григорьев

Автоматическая Аннотация Митохондриальных Геномов Грибов

Значительная часть митохондриальных белок-кодирующих генов грибов содержат интроны I типа или (редко) II типа, что представляет собой проблему для их точного предсказания. Нами впервые была разработана информационная система аннотирования, которая обеспечивает компьютерное предсказание такого рода генов. При тестировании 82 геномов из GenBank-а, на нуклеотидном уровне алгоритм имел точность 91%/88% (чувствительность/специфичность), а на уровне экзон/ОРС - 84%/79%.