Mitochondrial Genome Annotation

Protein Genes

Marwa Al Arab[1,2]

[1]Institute of Bioinformatics
University of Leipzig

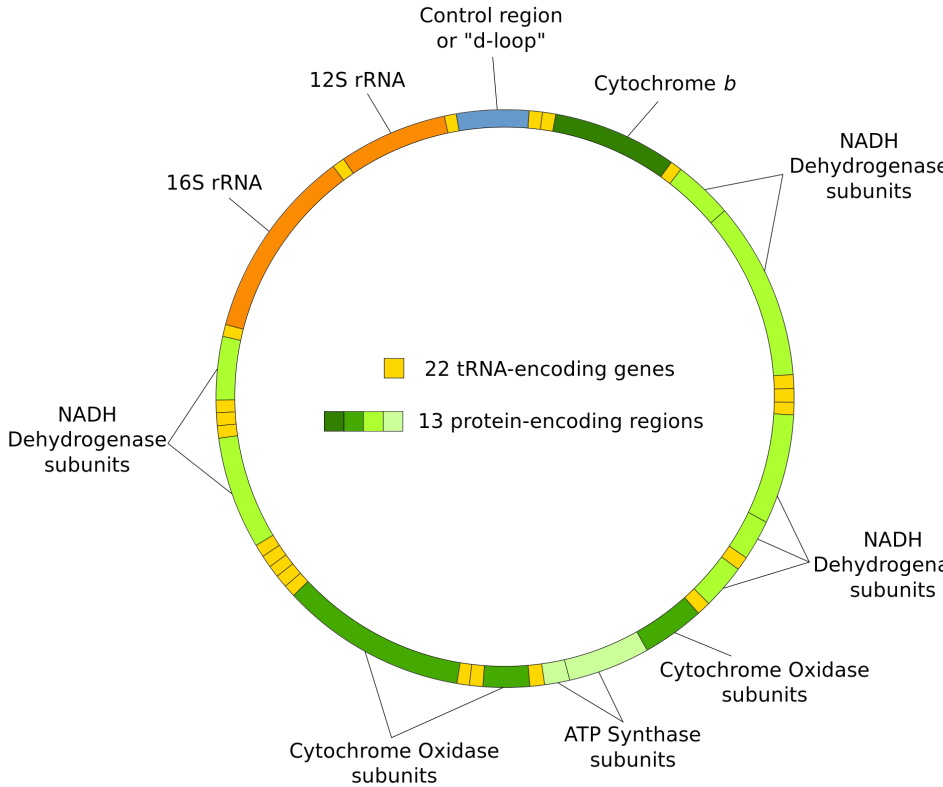
[2]Department of Bioinformatics
Lebanese University

TBI Bled 2015

Outline

Mitochondrial DNA

- Circular molecule located in mitochondria within eurkaryotic cells.
- transform energy to a form used by the cells
- Length about 16500 nucleotide.
- 13 protein coding genes
- 22 trna genes
- 2 rrna



[wikipedia]

Problem

- Refseq is the most used repository for mitochondrial genome annotation
- Refseq suffers form several inconsistencies and errors in annotation
  - Missing or incorrect strand
  - Confusing trnL1/trnL2
  - trnS1/trnS2
  - Inconsistencies in gene names(Bernt et. al 2012)
- Problem in developing automated analysis for mitochondrial data

Objective

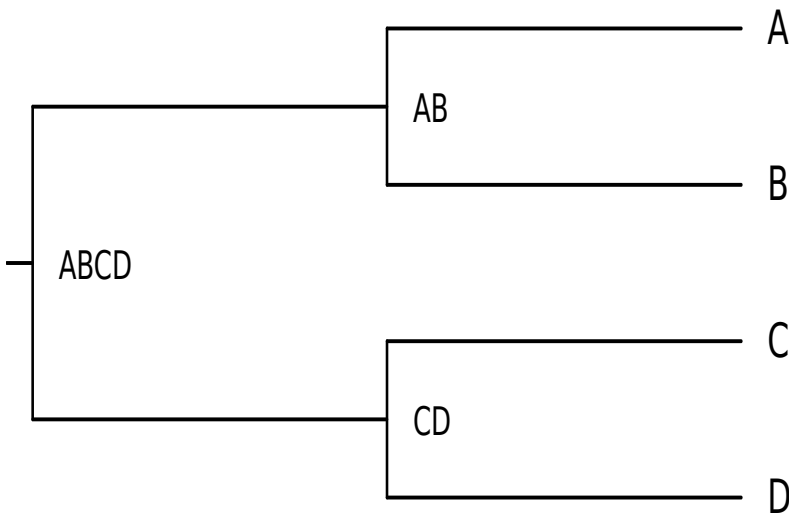
- Develop an automated pipeline for mtdna annotation by refining taxon specific hmm and covariance models .

Tools

- **HMMER** an implementation of profile HMMs (Sean Eddy and his group).
  - **hmmbuild** build a model from multiple sequences
  - **hmmalign** algin a model to sequences
  - **hmmsearch** search a model in sequences database
  - **hmmscan** search a genome in models database

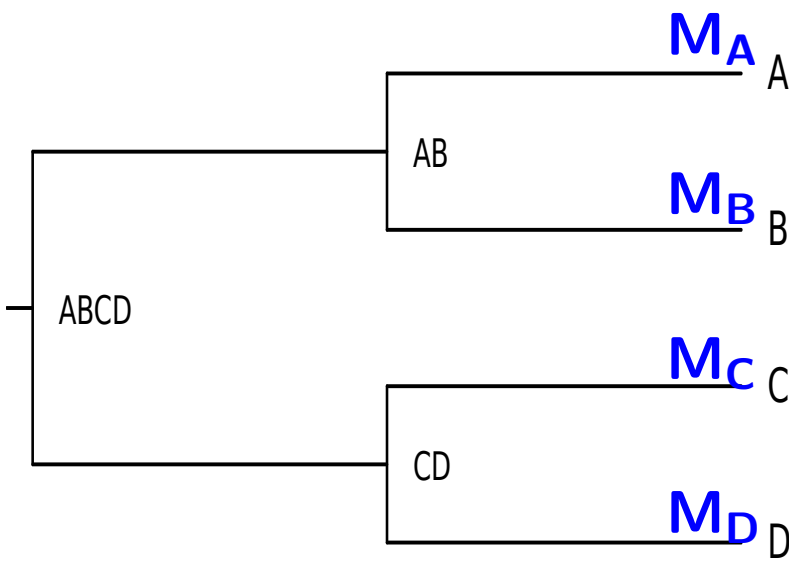
Training Data

- Build taxon specific models along the phylogenetic tree nodes
- Step 1: Build protein models for leaf sequences
- Step 2: Recursively lift up the model with best score
- Step 3: Build the models database



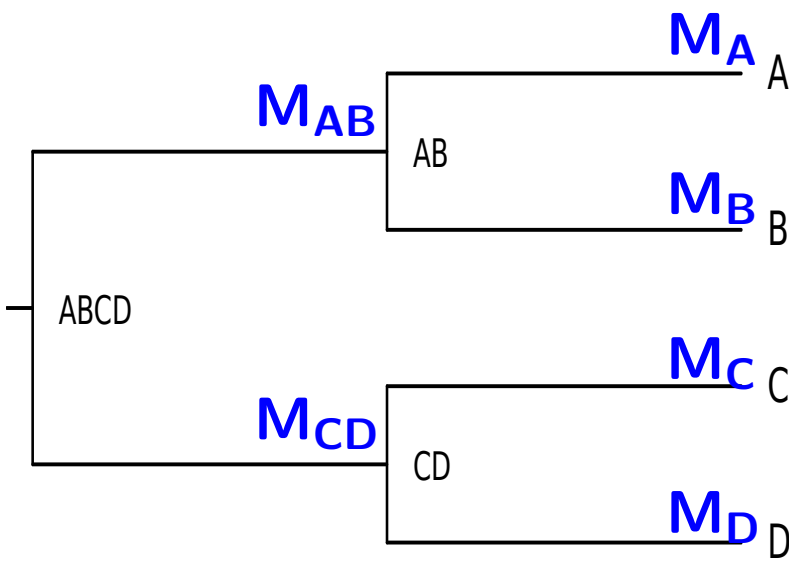
Training Data

- Build taxon specific models along the phylogenetic tree nodes
- Step 1: Build protein models for leaf sequences
- Step 2: Recursively lift up the model with best score
- Step 3: Build the models database



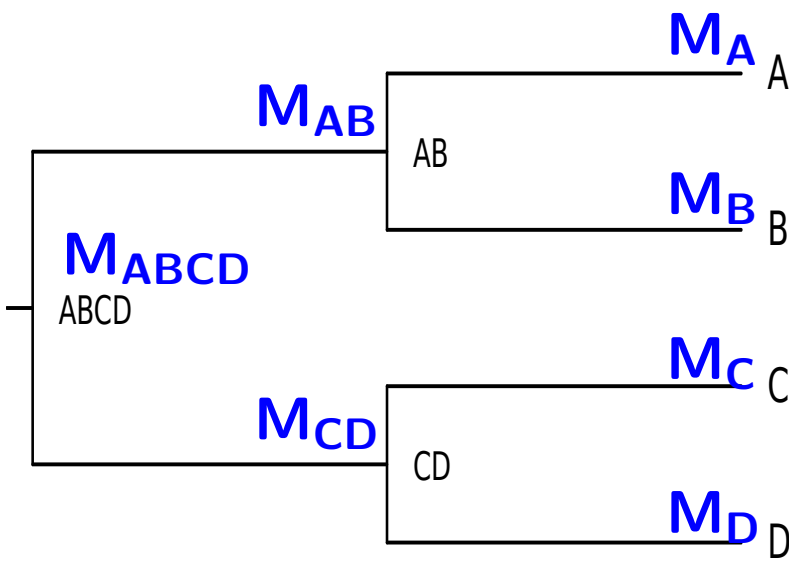
Training Data

- Build taxon specific models along the phylogenetic tree nodes
- Step 1: Build protein models for leaf sequences
- Step 2: Recursively lift up the model with best score
- Step 3: Build the models database

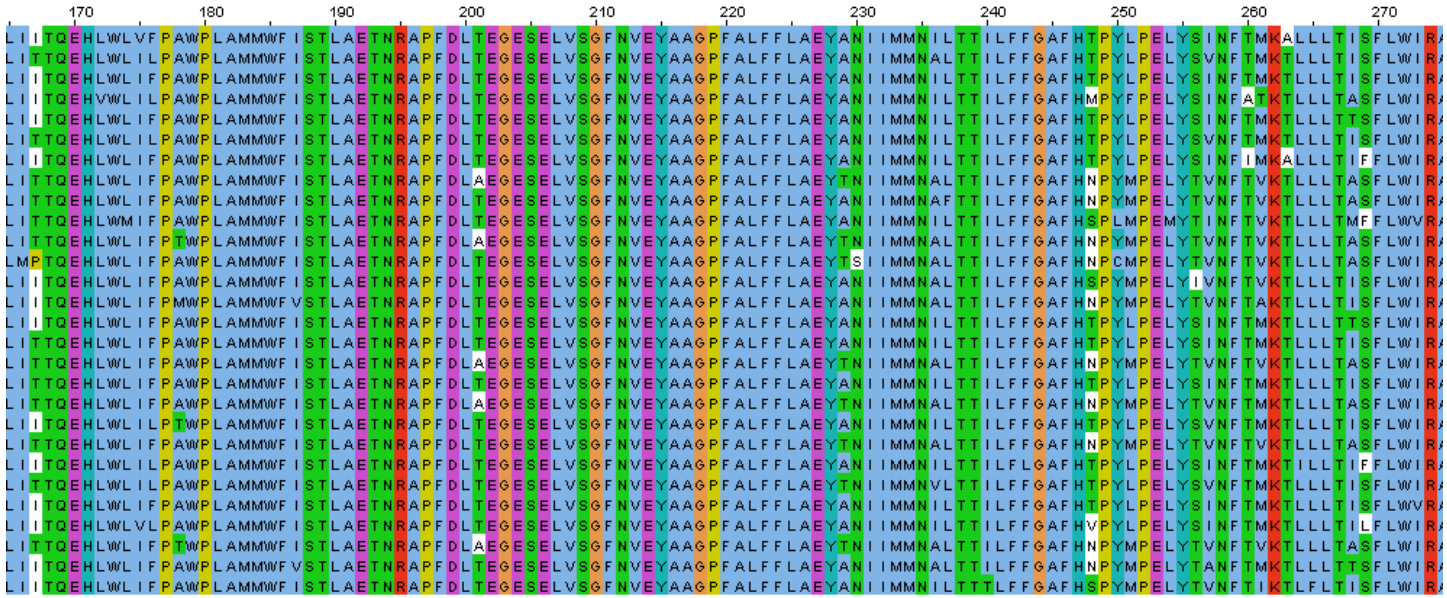


Training Data

- Build taxon specific models along the phylogenetic tree nodes
- Step 1: Build protein models for leaf sequences
- Step 2: Recursively lift up the model with best score
- Step 3: Build the models database

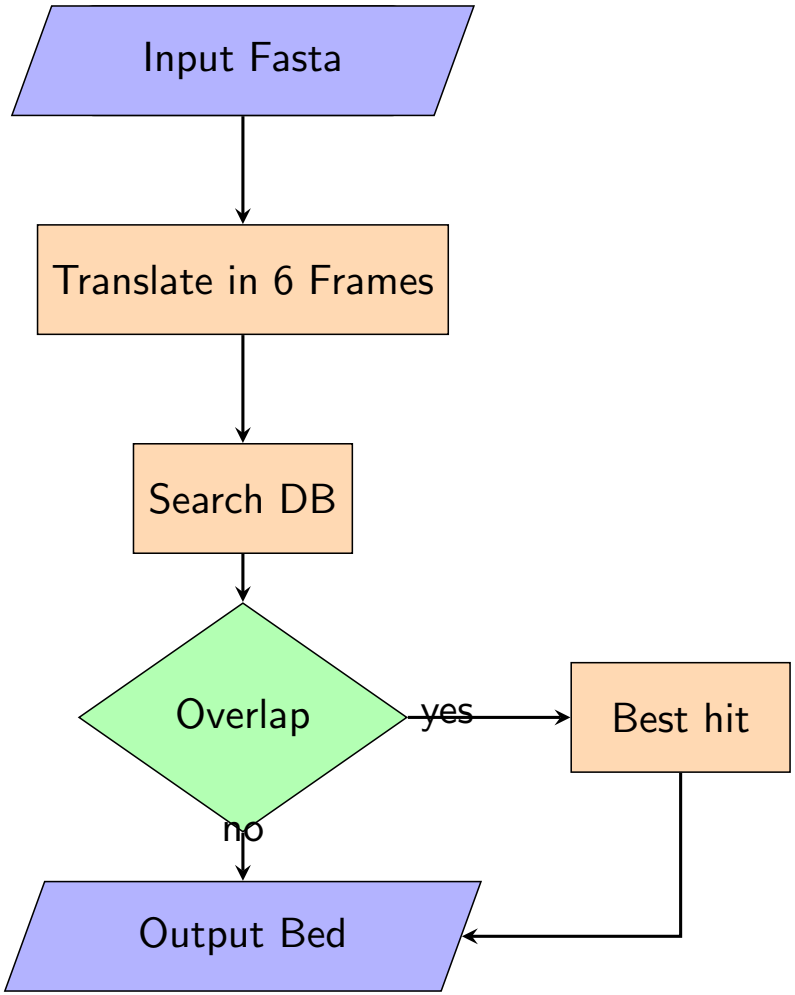


Example: Part of nad1 Alignment



Annotation

- Taxonomic level database



Results

- Train 3843 mt genome sequence from refseq63
- Test on 925 genome which are newly annotated in refseq69
- Scan against level models database

Phylum and Class Models

	equal	Δ±	FN	FP
atp6	923 (0.99)	0 (0.00)	0 (0.00)	4 (0.00)
cox3	920 (0.99)	0 (0.00)	2 (0.00)	6 (0.01)
nad3	915 (0.94)	0 (0.00)	59 (0.06)	2 (0.00)
nad4l	907 (0.97)	0 (0.00)	7 (0.01)	10 (0.01)
nad4	886 (0.99)	0 (0.00)	5 (0.01)	2 (0.00)
nad5	911 (0.99)	0 (0.00)	5 (0.01)	2 (0.00)
nad6	1104 (0.92)	0 (0.00)	54 (0.05)	8 (0.01)
cob	886 (0.98)	0 (0.00)	14 (0.02)	5 (0.01)
nad1	920 (1.00)	0 (0.00)	1 (0.00)	3 (0.00)
nad2	919 (0.95)	0 (0.00)	0 (0.00)	8 (0.01)
cox2	922 (1.00)	0 (0.00)	0 (0.00)	3 (0.00)
atp8	870 (0.65)	0 (0.00)	19 (0.01)	124 (0.09)
cox1	924 (0.99)	0 (0.00)	2 (0.00)	3 (0.00)
gene	12 007 (0.94)	0 (0.00)	168 (0.01)	180 (0.01)

Phylum models

Phylum and Class Models

	equal	Δ±	FN	FP
atp6	923 (0.99)	0 (0.00)	0 (0.00)	4 (0.00)
cox3	920 (0.99)	0 (0.00)	2 (0.00)	6 (0.01)
nad3	915 (0.94)	0 (0.00)	59 (0.06)	2 (0.00)
nad4l	907 (0.97)	0 (0.00)	7 (0.01)	10 (0.01)
nad4	886 (0.99)	0 (0.00)	5 (0.01)	2 (0.00)
nad5	911 (0.99)	0 (0.00)	5 (0.01)	2 (0.00)
nad6	1104 (0.92)	0 (0.00)	54 (0.05)	8 (0.01)
cob	886 (0.98)	0 (0.00)	14 (0.02)	5 (0.01)
nad1	920 (1.00)	0 (0.00)	1 (0.00)	3 (0.00)
nad2	919 (0.95)	0 (0.00)	0 (0.00)	8 (0.01)
cox2	922 (1.00)	0 (0.00)	0 (0.00)	3 (0.00)
atp8	870 (0.65)	0 (0.00)	19 (0.01)	124 (0.09)
cox1	924 (0.99)	0 (0.00)	2 (0.00)	3 (0.00)
gene	12 007 (0.94)	0 (0.00)	168 (0.01)	180 (0.01)

Phylum models

	equal	Δ±	FN	FP
atp6	910 (0.99)	0 (0.00)	0 (0.00)	4 (0.00)
cox3	921 (0.99)	0 (0.00)	3 (0.00)	5 (0.01)
nad3	914 (0.94)	0 (0.00)	61 (0.06)	2 (0.00)
nad4l	909 (0.97)	0 (0.00)	8 (0.01)	9 (0.01)
nad4	890 (0.99)	0 (0.00)	9 (0.01)	3 (0.00)
nad5	901 (0.98)	0 (0.00)	19 (0.02)	2 (0.00)
nad6	1062 (0.96)	0 (0.00)	9 (0.01)	8 (0.01)
cob	919 (0.99)	0 (0.00)	0 (0.00)	6 (0.01)
nad1	920 (1.00)	0 (0.00)	0 (0.00)	3 (0.00)
nad2	920 (0.97)	0 (0.00)	0 (0.00)	6 (0.01)
cox2	926 (0.99)	0 (0.00)	0 (0.00)	3 (0.00)
atp8	878 (0.67)	0 (0.00)	13 (0.01)	90 (0.07)
cox1	923 (0.99)	0 (0.00)	2 (0.00)	3 (0.00)
gene	11 993 (0.95)	0 (0.00)	124 (0.01)	144 (0.01)

Class models

Phylum and Class Models

$$Sn = \frac{TP}{TP+FN} \qquad Sp = \frac{TP}{TP+FP}$$

	equal	Δ±	FN	FP
atp6	923 (0.99)	0 (0.00)	0 (0.00)	4 (0.00)
cox3	920 (0.99)	0 (0.00)	2 (0.00)	6 (0.01)
nad3	915 (0.94)	0 (0.00)	59 (0.06)	2 (0.00)
nad4l	907 (0.97)	0 (0.00)	7 (0.01)	10 (0.01)
nad4	886 (0.99)	0 (0.00)	5 (0.01)	2 (0.00)
nad5	911 (0.99)	0 (0.00)	5 (0.01)	2 (0.00)
nad6	1104 (0.92)	0 (0.00)	54 (0.05)	8 (0.01)
cob	886 (0.98)	0 (0.00)	14 (0.02)	5 (0.01)
nad1	920 (1.00)	0 (0.00)	1 (0.00)	3 (0.00)
nad2	919 (0.95)	0 (0.00)	0 (0.00)	8 (0.01)
cox2	922 (1.00)	0 (0.00)	0 (0.00)	3 (0.00)
atp8	870 (0.65)	0 (0.00)	19 (0.01)	124 (0.09)
cox1	924 (0.99)	0 (0.00)	2 (0.00)	3 (0.00)
gene	12 007 (0.94)	0 (0.00)	168 (0.01)	180 (0.01)

Phylum models

	equal	Δ±	FN	FP
atp6	910 (0.99)	0 (0.00)	0 (0.00)	4 (0.00)
cox3	921 (0.99)	0 (0.00)	3 (0.00)	5 (0.01)
nad3	914 (0.94)	0 (0.00)	61 (0.06)	2 (0.00)
nad4l	909 (0.97)	0 (0.00)	8 (0.01)	9 (0.01)
nad4	890 (0.99)	0 (0.00)	9 (0.01)	3 (0.00)
nad5	901 (0.98)	0 (0.00)	19 (0.02)	2 (0.00)
nad6	1062 (0.96)	0 (0.00)	9 (0.01)	8 (0.01)
cob	919 (0.99)	0 (0.00)	0 (0.00)	6 (0.01)
nad1	920 (1.00)	0 (0.00)	0 (0.00)	3 (0.00)
nad2	920 (0.97)	0 (0.00)	0 (0.00)	6 (0.01)
cox2	926 (0.99)	0 (0.00)	0 (0.00)	3 (0.00)
atp8	878 (0.67)	0 (0.00)	13 (0.01)	90 (0.07)
cox1	923 (0.99)	0 (0.00)	2 (0.00)	3 (0.00)
gene	11 993 (0.95)	0 (0.00)	124 (0.01)	144 (0.01)

Class models

Phylum and Class Models

$$Sn = \frac{TP}{TP+FN} \qquad Sp = \frac{TP}{TP+FP}$$

	equal	Δ±	FN	FP
atp6	923 (0.99)	0 (0.00)	0 (0.00)	4 (0.00)
cox3	920 (0.99)	0 (0.00)	2 (0.00)	6 (0.01)
nad3	915 (0.94)	0 (0.00)	59 (0.06)	2 (0.00)
nad4l	907 (0.97)	0 (0.00)	7 (0.01)	10 (0.01)
nad4	886 (0.99)	0 (0.00)	5 (0.01)	2 (0.00)
nad5	911 (0.99)	0 (0.00)	5 (0.01)	2 (0.00)
nad6	1104 (0.92)	0 (0.00)	54 (0.05)	8 (0.01)
cob	886 (0.98)	0 (0.00)	14 (0.02)	5 (0.01)
nad1	920 (1.00)	0 (0.00)	1 (0.00)	3 (0.00)
nad2	919 (0.95)	0 (0.00)	0 (0.00)	8 (0.01)
cox2	922 (1.00)	0 (0.00)	0 (0.00)	3 (0.00)
atp8	870 (0.65)	0 (0.00)	19 (0.01)	124 (0.09)
cox1	924 (0.99)	0 (0.00)	2 (0.00)	3 (0.00)
gene	12 007 (0.94)	0 (0.00)	168 (0.01)	180 (0.01)

Phylum models

	equal	Δ±	FN	FP
atp6	910 (0.99)	0 (0.00)	0 (0.00)	4 (0.00)
cox3	921 (0.99)	0 (0.00)	3 (0.00)	5 (0.01)
nad3	914 (0.94)	0 (0.00)	61 (0.06)	2 (0.00)
nad4l	909 (0.97)	0 (0.00)	8 (0.01)	9 (0.01)
nad4	890 (0.99)	0 (0.00)	9 (0.01)	3 (0.00)
nad5	901 (0.98)	0 (0.00)	19 (0.02)	2 (0.00)
nad6	1062 (0.96)	0 (0.00)	9 (0.01)	8 (0.01)
cob	919 (0.99)	0 (0.00)	0 (0.00)	6 (0.01)
nad1	920 (1.00)	0 (0.00)	0 (0.00)	3 (0.00)
nad2	920 (0.97)	0 (0.00)	0 (0.00)	6 (0.01)
cox2	926 (0.99)	0 (0.00)	0 (0.00)	3 (0.00)
atp8	878 (0.67)	0 (0.00)	13 (0.01)	90 (0.07)
cox1	923 (0.99)	0 (0.00)	2 (0.00)	3 (0.00)
gene	11 993 (0.95)	0 (0.00)	124 (0.01)	144 (0.01)

Class models

$Sn = 0.986$ $Sp = 0.985$ $Sn = 0.989$ $Sp = 0.988$

Conclusion and Outlook

- An automated pipeline to annotate protein coding genes in mtDNA by refining taxon specific hmm
- Outlook
  - Find the best level database and best parameters to minimize FN and maximize TP
  - Analyse the results in deep to improve the refseq annotation
  - Apply on tRna

Thanks to

- Matthias Bernt
- Christian Honer
- Frank Juhling
- Abdullah Sahyoun
- Peter Stadler