Mitochondrial Genome Annotation Protein Genes

Marwa Al Arab^{1,2}

¹Institute of Bioinformatics University of Leipzig

²Department of Bioinformatics Lebanese University

TBI Bled 2015

< ∃ →

Outline

Introduction

Mitochondrial DNA Problem

Materials and Methods

Tools Training Annotation

Results

문 🛌 문

Mitochondrial DNA Problem

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



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

Mitochondrial DNA Problem

Objective

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



< E

Tools Training Annotation

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



Tools **Training** Annotation

- 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



Tools **Training** Annotation

- 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



Tools Training Annotation

- 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



Tools **Training** Annotation

- 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



Tools Training Annotation

Example: Part of nad1 Alignment

170	180	190	200	210	220	230	240	250	260	270
LITOCHLWLV	FPAWPLAMMWF	ISTLAETNRAP	FDLTEGESEL	VSOFNVEYA	AGPFALFFLA	EYANTIMM	ILTTILFFOA	FHTPYLPELYS	INFTMKALLL	ISFLWIR
LITTOEHLWLI	L PAWPLAMMWF	ISTLAETN <mark>R</mark> ap	FDL <mark>TEGES</mark> EL	VSOF NVEYA	A <mark>GP</mark> FALFFLA	A <mark>ey</mark> an Limmn	AL <mark>TT</mark> ILFF <mark>0</mark> A	FHTPYLPELYS	VNFTMKTLLLT	ISFLWIRA
LIITQEHLWLI	F PAWPLAMMWF	ISTLAETNRAP	FDLTEGESEL	VSGF NVEYA	A <mark>G P</mark> FALFFLA	EYAN LIMMN	ILTTILFF GA	FHTPYLPELYS	INFIMETLLLT	ISFLWIR/
LIITQEHVWLI	L <mark>PAWP</mark> LAMMWF	I STLAETN <mark>R</mark> AP	FDLTEGESEL	VSGF NVEYA	A <mark>G P</mark> FALFFLA	EYAN LIMMN	IL TTILFF GA	FHMPYFPELYS	INFATELLET	ASFLWIR/
LITQEHLWLI	F PAWPLAMMWF	ISTLAETNBAP	FDLTEGESEL	VSGENVEYA	ia <mark>g p</mark> ealeel <i>i</i>	EYAN LIMMN	ILTTILFF OA	FHTPYLPELYS	INFTMKTLLLT	TSFLWIR/
LITTREHLWLI	LPAWPLAMMWF	ISTLAETNRAP	FDLTEGESEL	VSGENVEYA	AGPEALEEL	EYANTIMM	ILTTILFFOA	FHTPYLPELYS	VNFIMETULLI	ISFLWIR/
LITTOEHLWLI	F PAWPLAMMWF	STLAUTNBAP	FDLTEGESEL	USOF NUEYA	AGPEALFELA	AN I I MM	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	FHIPYLPELYS	INFIMEALLL	TE FLAMTER
	F PAUPLAMMOF	OTLACTNAAP	COL ACOCOCC		ACTALIFIES		ACTILLE		VAP TV TULL	
TTOCHLWM		I STLACTNRAP		VEGENVEYA	AGPEALEEL		II TTILEEGA		INCTUNTING	MEELWOVE
I TTREHIWI I	F P THIP LAMMINE	ISTIGETNBOR	FDLAEGESEL	VEGENVEYA	AGPEALEEL	FYTNILMM	ALTTILFEGA	EHNPYMPELYT	VNETVETILLT	ASELWIR
LMPTQEHLWLT	FRAWPLAMMWE	STLAETNBAP	FDLTEGESEL	VSGENVEYA	AGPEALEELA	EYTSILMM	ALTTILFEGA	FHNPCMPELYT	VNETVETLLLT	ASFLWIR
LITOENLWLT	F PAWPLAMMWF	ISTLAETNRAP	FDLTEGESEL	VSOFNVEYA	ACPEALFELA	EYANTIMM	ALTTILFFOA	FHSPYMPELYI	VNETVKTULLT	ISFLWIRA
LIITQEHLWLI	F <mark>PMWP</mark> LAMMWF	VSTLAETN <mark>R</mark> AP	F D L <mark>T E O E S</mark> E L	VSOF NVEYA	A <mark>GP</mark> FALFFLA	EYAN LIMMN	AL <mark>TT</mark> ILFF <mark>0</mark> A	FHNPYMPELYT	VNFTAKTULLT	ISFLWIRA
LITOEHLWLI	F PAWPLAMMWF	ISTLAETN <mark>R</mark> ap	FDL <mark>TEGES</mark> EL	VSGFNVEYA	A <mark>GP</mark> FALFFLA	A <mark>ey</mark> an Limmn	ILTTILFF GA	FHTPYLPELYS	INFTMKTLLLT	TSFLWIRA
LITTOEHLWLI	F PAWP LAMMWF	I STLAETN <mark>R</mark> AP	FDLTEGESEL	VSGF NVEYA	A <mark>GP</mark> FALFFLA	EYAN LIMMN	IL TTILFF GA	FHTPYLPELYS	INFTMKTLLLT	ISFLWIR/
LITTOEHLWLI	F PAWPLAMMWF	ISTLAETNBAP	FDLAEGESEL	VSGENVEYA	A <mark>G P</mark> FALFFLA	EYTN I I MMN	ALTTILFF <mark>O</mark> A	F HNPYMPELYT	VNFTVKTELET	ASELWIRA
LITTREHLWLI	F PAWP LAMMWF	ISTLAETNBAP	FDLTEGESEL	VSGENVEYA	AGPEALEEL	EYANTIMM	ILTTILFFGA	FHIPYLPELYS	INFTMETLULT	ISFLWIR/
LITTOEHLWLI	FPAWPLAMMWF	STLAETNRAP	FDLAEGESEL	USOF NUEYA	AGPEALFELA	EYTNIIMM	ALTILFFOA	FHNPYMPELYT	VNFTMETLLLT	ASFLWIR
		DTLACTNAAP			AGFFALFFL	CAN LINIMA			VAP THE TULL	
LITOCHUWLI		I STLACTNRAP		VOGE NUE YA	AGPEALEEL		IL TTUELOA	CHTOMI DELVO	INCTONTICLE	
TTREHLWLL	PAMPLAMMIE	ISTIGETNRAP	FDITEGESEL	VEGENVEYA	AGPEALEELA	FYTNIIMMN	VITTILEEGA	FHTPYLPELYS	VNETMETLLLT	I SEL WIR
LITCENLWLT	FRAWPLAMMWE	STLAETNBAP	FDLTEGESEL	VSGENVEYA	AGPEALEELA	EYANTIMM	ILTTILFEGA	FHTPYLPELYS	INFTMETLLLT	ISFLWVR/
LITTOEHLWLV	L PAWP LAMMWF	ISTLAETNBAP	FDLTEGESEL	VSGENVEYA	AGPEALEEL	EYANTIMM	ILTTILFFOA	FHVPYLPELYS	INFTMETULUT	ILFLWIR/
LITTOENLWLI	FPTWPLAMMWF	ISTLAETNRAP	FDLAEGESEL	VSOFNVEYA	AGPFALFFLA	EYTNIIMMN	ALTTILFF GA	FHNPYMPELYT	VNFTVKTUUUT	ASFLWIRA
LIITQEHLWLI	F <mark>P</mark> AW <mark>P</mark> L AMMWF '	VSTLAETN <mark>R</mark> ap	FDL <mark>TEGES</mark> EL	VSOF NVEYA	A <mark>GP</mark> FALFFLA	A <mark>ey</mark> an Limmn	AL <mark>TTI</mark> LFF <mark>Ø</mark> A	F H N P Y M P E L Y T	ANFTMKTLLLT	TSFLWIRA
LIITQEHLWLI	FPAWPLAMMWF	ISTLAETNRAP	FDLTEGESEL	VSGENVEYA	AGPFALFFLA	EYANIIMM	ILTTLFF	FHSPYMPELYT	VNFTIKTLFLF	ISFLWIR/

æ

・聞き ・ ほき・ ・ ほき

Introduction Materials and Methods Results Conclusion and Outlook Training Annotation

Annotation



Results

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

< E.

Phylum and Class Models

	equal	$\Delta \pm$	FN	\mathbf{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	12007~(0.94)	0 (0.00)	168(0.01)	180(0.01)

Phylum models

æ

'문▶' ★ 문▶

Phylum and Class Models

	equal	$\Delta \pm$	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)
$\cos 2$	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)
$\cos 1$	924~(0.99)	0 (0.00)	2(0.00)	3(0.00)
gene	12007~(0.94)	0 (0.00)	$168\ (0.01)$	$180\ (0.01)$

	equal	$\Delta \pm$	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)
$\cos 2$	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)
$\cos 1$	923~(0.99)	0(0.00)	2(0.00)	3(0.00)
gene	11993(0.95)	0 (0.00)	$124\ (0.01)$	144(0.01)

Phylum models

Class models

・聞き ・ ヨト ・ ヨトー

æ

Phylum and Class Models

$Sn = \frac{TP}{TP+FN}$	$Sp = \frac{TP}{TP+FP}$
IP+FN	-r IP+F

	equal	$\Delta \pm$	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)
$\cos 2$	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)
$\cos 1$	924~(0.99)	0 (0.00)	2(0.00)	3(0.00)
gene	12007(0.94)	0 (0.00)	168(0.01)	180 (0.01)

	equal	$\Delta \pm$	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)
$\cos 2$	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)
$\cos 1$	923 (0.99)	0 (0.00)	2(0.00)	3 (0.00)
gene	11993(0.95)	0(0.00)	$124\ (0.01)$	144(0.01)

Phylum models

Class models

□ ▶ ▲ 臣 ▶ ▲ 臣 ▶

æ

Phylum and Class Models

$Sn = \frac{11}{TP + FN}$	$Sp = \frac{1}{TP+FF}$
P+PN	υρ ΤΡ

	equal	$\Delta \pm$	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)
$\cos 2$	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)
$\cos 1$	924 (0.99)	0 (0.00)	2(0.00)	3(0.00)
gene	12007(0.94)	0 (0.00)	168(0.01)	180(0.01)

	equal	$\Delta \pm$	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)
$\cos 2$	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)
$\cos 1$	923 (0.99)	0 (0.00)	2(0.00)	3 (0.00)
gene	11993(0.95)	0(0.00)	$124\ (0.01)$	$144\ (0.01)$

Phylum models

Class models

Sn = 0.986

Sn = 0.989 Sp = 0.988

Marwa Al Arab

Mitochondrial Genome Annotation

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