An Evolutionary Hypothesis and Computational Identification of Insertional RNA Editing Sites

Ralf Bundschuh

Department of Physics, Ohio State University, KITP

Collaborators:

Tsunglin Liu, UCSB, Hayoun Lee, University of Rochester,
Jonatha Gott, Neeta Parimi, Case Western Reserve University,
Christina Ainsley, COSI Columbus

Outline:

- Introduction to RNA editing
- An evolutionary model for codon position bias
- How to find insertional editing sites
- Conclusions and outlook

supported by:
National Science Foundation (RB), National Institutes of Health (JG)

- Central dogma: DNA $\overset{\text{exact copy}}{\longrightarrow}$ RNA $\overset{\text{genetic code}}{\longrightarrow}$ protein
- RNA editing: RNA gets edited before it is translated
- Example: mitochondrion of *Physarum polycephalum*
  - most prevalent editing event: C insertion
  - e.g., a piece of *nad7*:

```
DNA     ...CAGAATTGCGATCCACATAT GGGCTTCTACAT GAGGTACTGAAAAACTTATAGAACATAAGAATTTCTTACAATCT  TCCTTATTTTGAT GTCTTGAT...
mRNA    ...CAGAAUUGCGAUCCACAUAUCGGGCUUCUACAUCGAGGUACUGAAAAACUUAUAGAACAUAAGAAUUUCUUACAAUCUCUUCCUUAUUUUGAUCGUCUUGAU...
protein ... Q  N  C  D  P  H  I  G  L  L  H  R  G  T  E  K  L  I  E  H  K  N  F  L  Q  S  L  P  Y  F  D  R  L  D ...
```

  - other editing events: U insertion, dinucleotide insertions, C→U conversion
  - Editing is frequent: one insertion per 25 bases on average

- Other types of RNA editing occur in all kinds of organisms: humans, plant organelles, nematodes, kinetoplastids, viruses
- Some RNA editing is implied in viral defense.
- Some RNA editing is directed by guide RNAs.
- Some editing enzymes have been identified.
- Main issues in general:
  - What is the mechanism of RNA editing?
  - How are editing sites recognized?
  - What is the biological function of RNA editing?

Specifically in *Physarum polycephalum*:

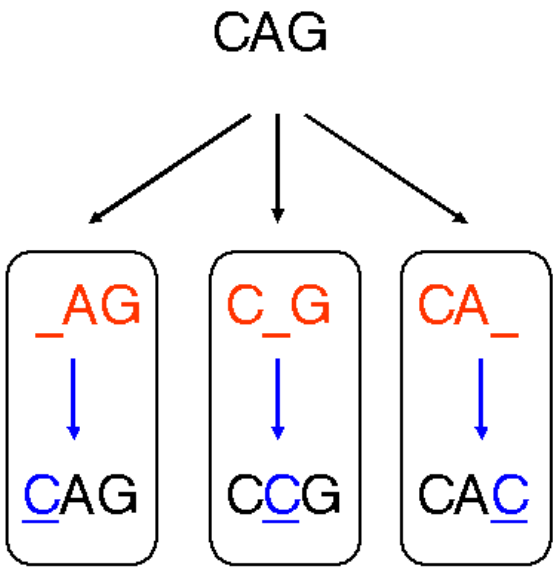
- Editing is extremely reliable
- Editing occurs co-transcriptionally
- All known mitochondrial protein coding genes are edited
- Nearly all mitochondrial stable RNA genes are edited
- Nothing is known about the actual editing mechanism
- Nothing is known about the recognition of editing sites
- Nothing is known about the biological function
- 497 editing sites known → later part of the talk
- 227 unambiguous C insertions in protein coding regions known

- Sort unambiguous C-insertions by codon positions
- Codon positions for editing sites in coding sequences

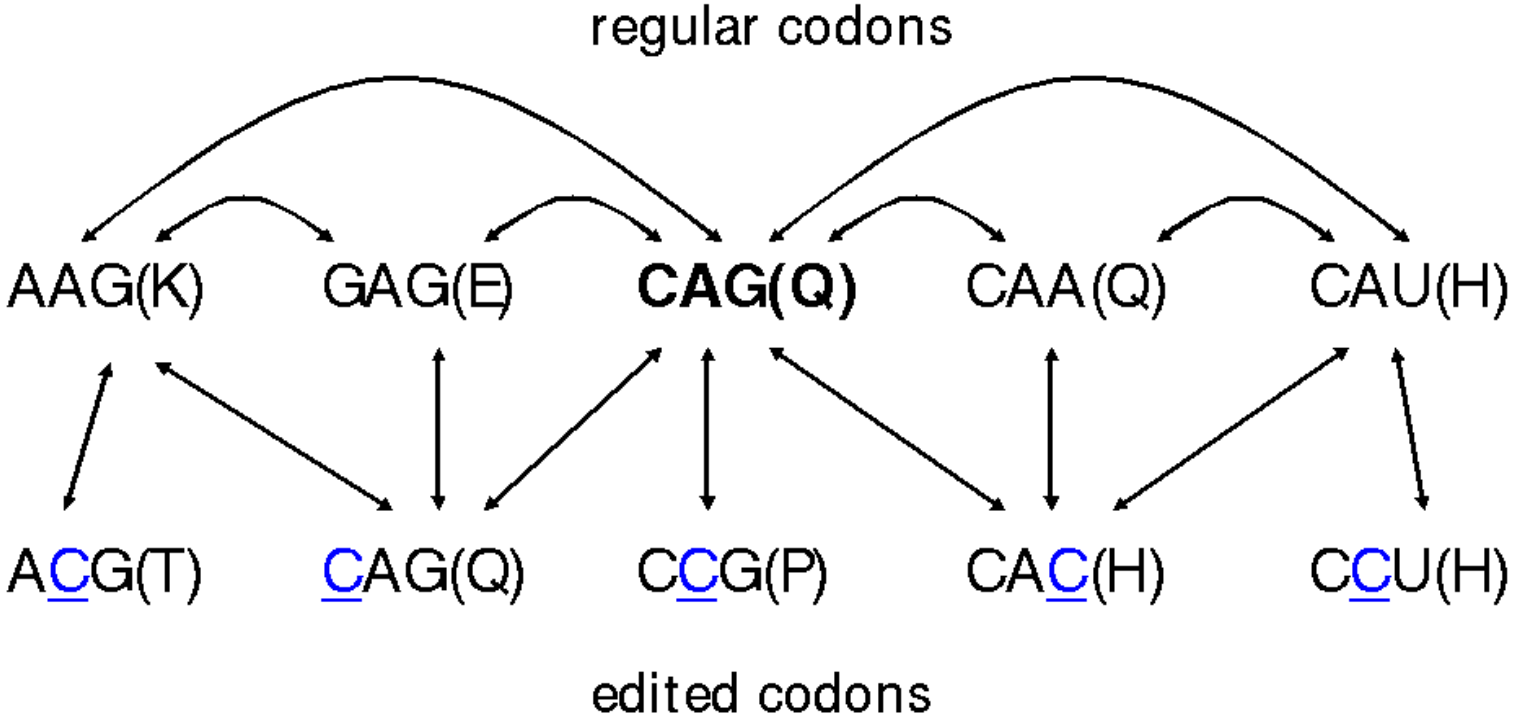
codon position	1	2	3
number	58	24	145
percentage	26%	11%	64%

- Codon bias surprising since RNA editing is co-transcriptional

- Can we understand the codon preference?
- Simple evolutionary model:
  - No codon preference in editing machinery
  - Base deletion occurs during sequence evolution
  - Sometimes base deletion can be rescued by editing
  - Results in effective replacement of original base by C
  - Fitness of new sequence depends only on amino acid sequence

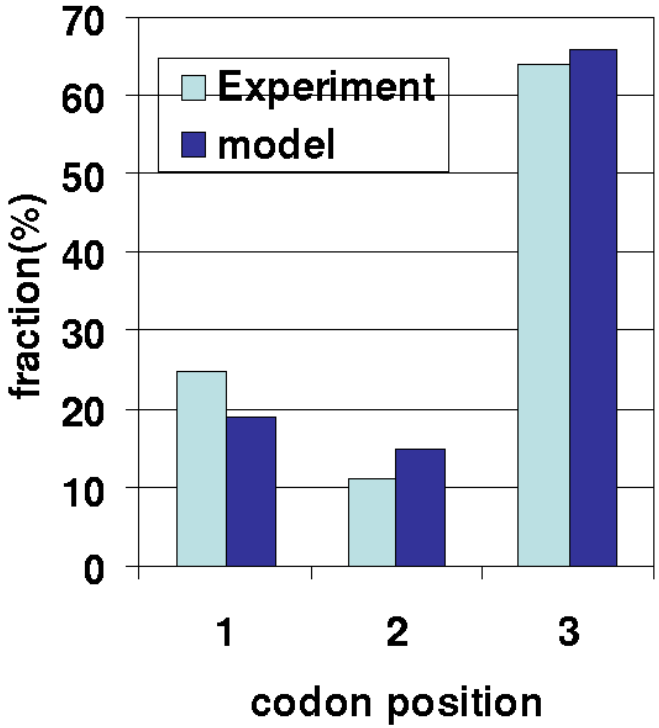


- Include mutations and insertions: complete evolution model



- Fitness given by similarity of amino acid to original amino acid according to BLOSUM62 similarity matrix
- Know states, transitions, and fitness
  $\Rightarrow$ can use Eigen theory to determine stationary state
- Average over all original codons

- Result:



- Insensitive to parameter choice

Note: our model implies that there is no other reason to choose the positions of most editing site but to "fix" the amino acid sequence

- Consistent with "cheap" editing
- Recent unpublished data from several organisms confirms random acquisition and loss of editing sites in myxomycetes

- How do we know the editing sites?
- Need to sequence both the genomic DNA and the RNAs
- Genomic DNA fully sequenced for *Physarum polycephalum* Takano *et al.*, 2001
- Sequencing RNAs is hard
  - need to know where genes are
  - need primers
  - primers need to be complementary to edited RNA
- Situation for mitochondrion of *Physarum polycephalum*:
  - six protein coding genes with experimentally determined editing sites in GenBank
  - a handful of genes identified but editing sites not known
  - several unidentified open reading frames
  - four typical mitochondrial genes apparently missing
  - Compare to *Dictyostelium discoideum*: 44 genes known
- Experimental determination of editing sites difficult
  $\Rightarrow$ computational prediction to be confirmed by experiment

- Main idea: use protein sequences from other organisms
- Pick gene to predict editing sites of, e.g., nad7
- Pick protein for this gene from another species, e.g., *Neisseria menigitidis*
- Find all related protein sequences out of GenBank
  $\longrightarrow$ 510 sequences for nad7
- Look at each position in multiple alignment

Neisseria meningitidis	...VRADPHIGLLHRGTEKLAETKT-YLQALPYMDRLD...
Drosophila melanogaster	...MRADPHIGLLHRGTEKLIEYKT-YTQALPYFDRLD...
Synechococcus sp.	...VDCEPVIGYLHRGMEKIAENRT-NVMFVPYVSRMD...
Buchnera aphidicola	...VDCVPDIGYHHRGAEKMAERQS-WHSYIPYTDRIE...
Chloroflexus aurantiacus	...VNVAPDVGYLHTGIEKTMESKT-YQKAVVLTDRMD...
Escherichia coli	...IDADYRLFYVHRGMEKLAETRMGYNEVTFLSDRVC...
Rhodospirillum rubrum	...IRNAVSTGTMWRGIELILKGRD-PRDAWAFTQRIC...
$\vdots$	$\vdots$

42 54

- Extract probabilities $p_i(a)$ to find amino acid a at position i

$i \setminus a$	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
$\vdots$																				
42	0.05	0.01	0.02	0.02	0.005	0.01	0.02	0.68	0.007	0.009	0.02	0.02	0.006	0.008	0.02	0.04	0.02	0.004	0.007	0.01
$\vdots$																				
54	0.07	0.09	0.14	0.05	0.005	0.04	0.04	0.04	0.07	0.02	0.03	0.03	0.009	0.03	0.02	0.09	0.05	0.007	0.15	0.02
$\vdots$																				

Editing site prediction:

- Start with genomic sequence

```
...CAGAATTGCGATCCACATATGGGCTTCTACATGAGGTACTGAAAAACTTATAGAACATAAGAATTTCTTACAATCTTCCTTATTTTGATGTCTTGAT...
```

- Insert C's and translate

```
...CAGAATTGCGACTCCACATATGGGCTTCTACATGACGGTACTGAAAAACTTATCAGAACATACAGAATTTCTCTACAATCTTCCTTATTTTGCATGTCTTGCAT...
   Q  N  C  D  S  T  Y  G  L  L  H  D  G  T  E  K  L  I  R  T  Y  R  I  S  L  Q  S  S  L  F  C  M  S  C
```

- Calculate probability

$$p(\ldots QNCDSTYGL\ldots) =$$
$$= \quad \ldots p_{35}(Q)p_{36}(N)p_{37}(C)p_{38}(D)p_{39}(S)p_{40}(T)p_{41}(Y)p_{42}(G)p_{43}(L)\ldots$$

- Defines "energy landscape" over space of 2^N discrete states
- Identify ground state $\longrightarrow$ prediction of editing sites

Identification of editing sites IV

- Use transfer matrix approach:
- Genomic sequence $b_1 \ldots b_N$; protein model: $p_i(a)$ for $i = 1, \ldots, M$
- Define $P_{i,j}$ as the probability of the most probable editing configuration ending at model position i and genomic position j
- Without editing:

$$P_{i,j} = p_i(aa[b_j - 2, b_j - 1, b_j])P_{i-1,j-3}$$

- With editing:

$$P_{i,j} = \max \left\{ \begin{array}{l} p_i(aa[b_j - 2, b_j - 1, b_j])P_{i-1,j-3} \\ p_i(aa[C, b_j - 1, b_j])P_{i-1,j-2} \\ p_i(aa[b_j - 1, C, b_j])P_{i-1,j-2} \\ p_i(aa[b_j - 1, b_j, C])P_{i-1,j-2} \end{array} \right\}$$

$\Rightarrow$ $O(NM)$ algorithm

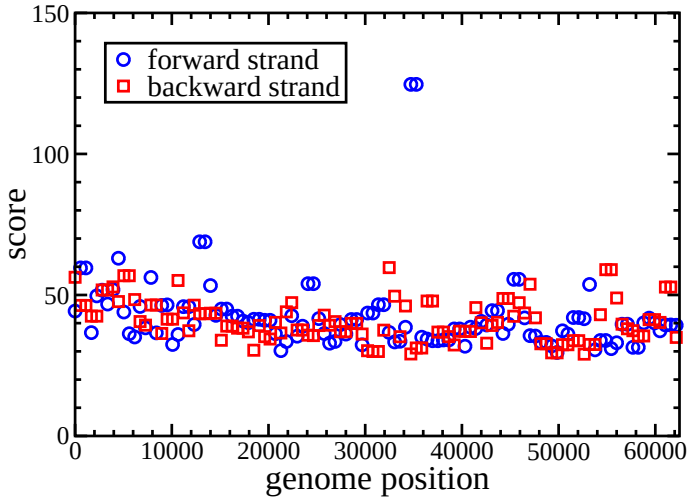
- In reality include amino acid insertions and deletions, local similarities, and sequence context

- Check performance on known genes:

gene	amino acids	C insertions	off by			
			1	2	3	≥ 4
nad7	92%	$116/171 = 68\%$	9	12	7	28
cox1	93%	$112/159 = 70\%$	8	15	8	27
cox3	81%	$134/181 = 74\%$	9	14	9	55
cytb	93%	$118/172 = 68\%$	11	11	6	15
atp	93%	$106/152 = 70\%$	7	8	4	15
pL	93%	$144/199 = 72\%$	10	18	9	38
total	92%	$122/173 = 71\%$	12	9	8	22

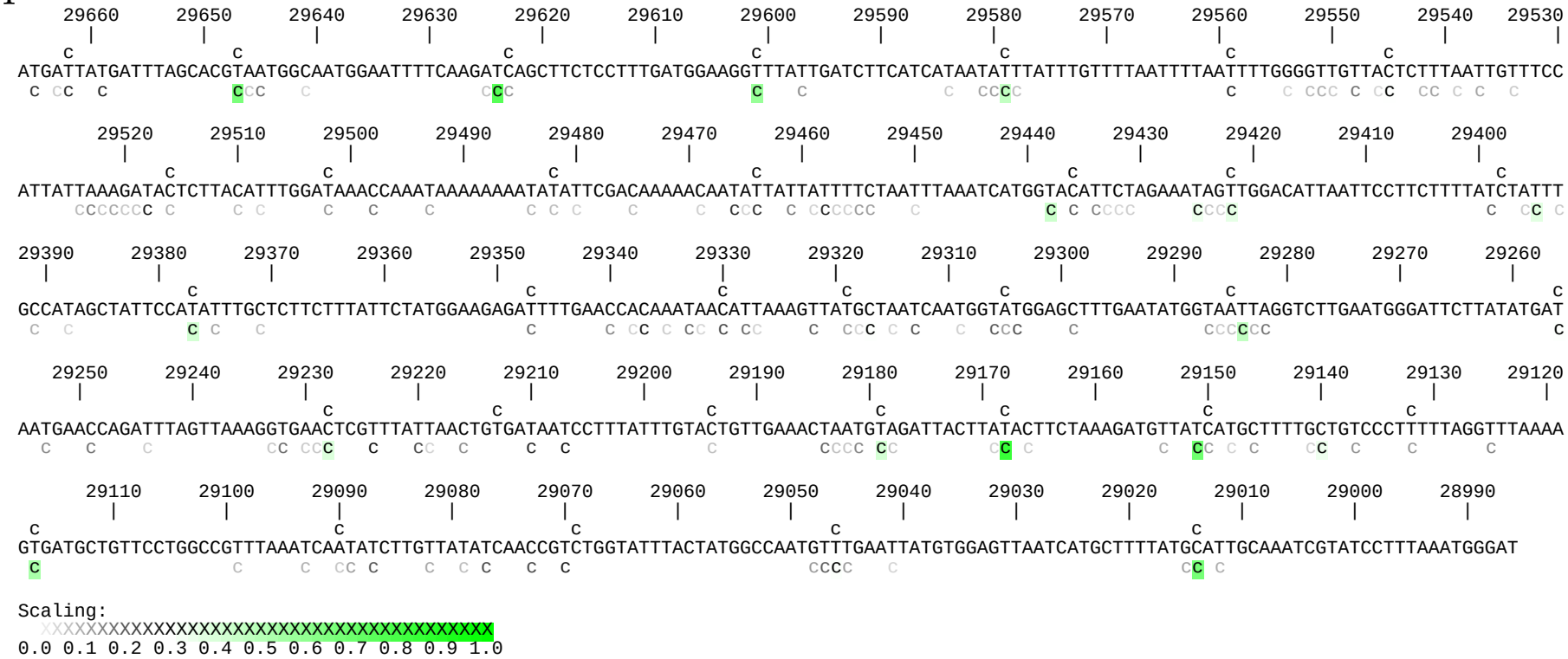
Real test: Finding new genes

- Search for missing genes nad2, nad4L, nad6, and atp8
- These genes could not be found by traditional gene finding
- Step 1: find location
  - Pick a gene from the list
  - Build PIE model for this gene from protein sequences of other organisms
  - Cut genome into short overlapping pieces (length 1200 bases)
  - Apply PIE to every piece of the genome
  - PIE predicts best way to insert C's in each piece plus goodness measure
  - Identify position of gene in genome by maximum in goodness measure

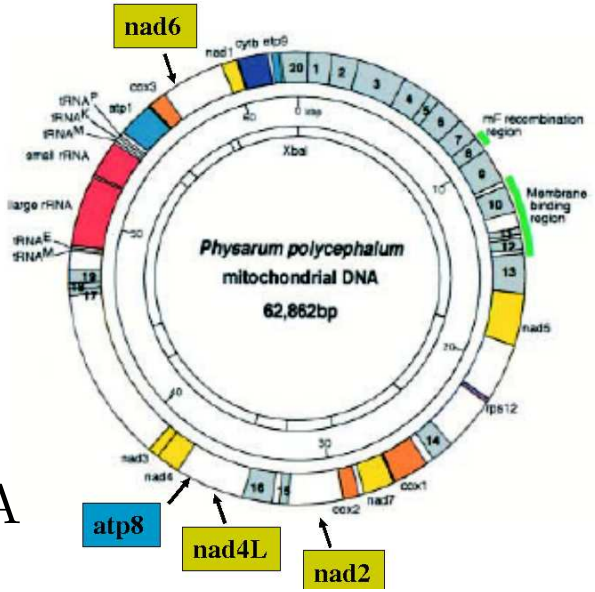


Step 2: primer design

- Primer has to be complementary to mRNA sequence
  but: Do not know mRNA sequence
- Use PIE to predict editing site positions ⇒ know mRNA sequence
  but: PIE makes mistakes
- Assign reliability measure to PIE's predictions by calculating probabilities in Boltzmann ensemble



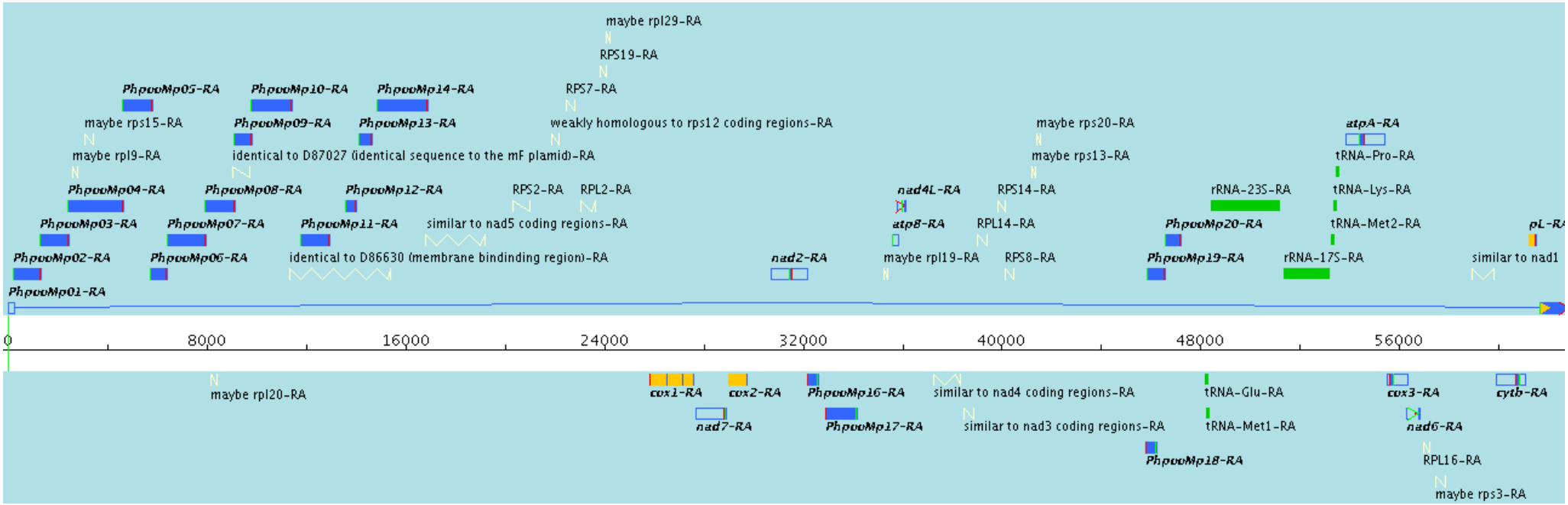
- Use to select primers



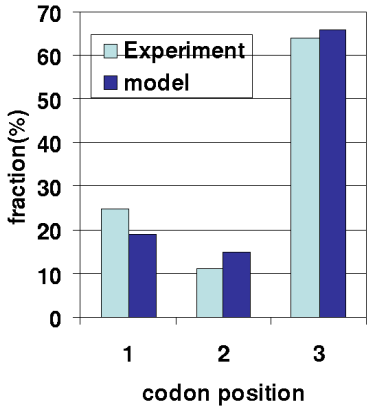
- Location of all four genes found
- All but one primer worked
- All four genes confirmed by sequencing of mRNA
- New editing type in *Physarum*: deletional RNA editing
- Total increase in known editing sites by 50%

	Previous coding	Total coding	Stable RNA	Previous total	total
Editing sites	250	390	107	357	497
C insertion	222	353	97	319	450
Unambiguous	140	227	66	206	293

- Systematically search for all known mitochondrial genes
- Find 11 genes beyond the four experimentally verified ones
- Find 8 more candidates with lower statistical significance

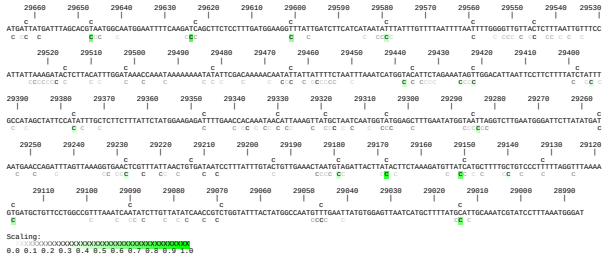


- In total increased number of predicted genes from 11 to 26–34
- Still have to be verified experimentally



Conclusions:

- Simple evolutionary model can explain codon bias
- Editing sites seem to be randomly acquired and lost
- RNA editing sites of known proteins can be computationally predicted with reasonable accuracy



Future directions:

- Comparative analysis of several organisms with editing
- Verify full genome predictions experimentally