Detection of ~~non~~coding RNAs by comparative sequence analysis

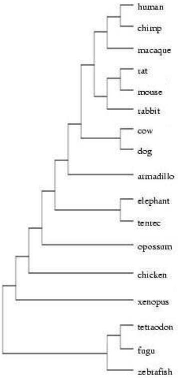
A mRNA model for `RNAz`

Stefan Washietl
Institute for Theoretical Chemistry
University of Vienna

Bled, February 2006

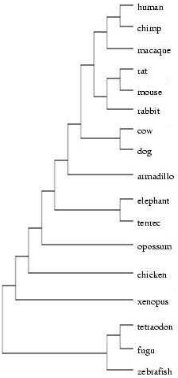
The challenge of comparative genomics

```
Mouse     ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Cow       ACGGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGCG-CC
Dog       ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Rat       ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Rhesus    ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Chimp     ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Human     ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Elephant  ACTGCTGGGCCTGTACTAGAGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Tenrec    ACTGCTGGGCTTGTACTAGAGGGTGTGCTGATGGGTACTGGGGGGTG-CT
Armadillo ACTGCTGGG-CTGCATCAGGGGGTGTGCTGTCGGGTACTGGGGAGTG-CC
Opossum   ACTGCTGAGCTTGCACCAAATGATGCGCTGTCGGGTACTGAGGGGTG-CT
Chicken   ATTGCTGCGCCTGTACCAAGTGGTGCGCTGTGGGGTACTGGGGGCTG-CC
Frog      AGTGTTGGGCTTGCACCAAGTGATGTGCTGTAGGGTACTGGGCGTTA-CT
Fugu      ACTGTTGCGTCTGCACCAAGTGATGCGCTGTCGGGAACTGTGGCGTG-GC
Tetraodon ACTGCTGCGTCTGCACCAGGTGATGCGCTGTCGGGAACTGCGGCGTG-GC
Zebrafish ATGGCTGCATGTGGCCCAGATGAT----TGACAGATGATGTCAGATGTGT
              *  * **    **    *   * *    **   *    **
```



The challenge of comparative genomics

```
Mouse     ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Cow       ACGGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGCG-CC
Dog       ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Rat       ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Rhesus    ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Chimp     ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Human     ACTGCTGGGCCTGGACCAGGGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Elephant  ACTGCTGGGCCTGTACTAGAGGGTGTGCTGTCGGGTACTGGGGGGTG-CT
Tenrec    ACTGCTGGGCTTGTACTAGAGGGTGTGCTGATGGGTACTGGGGGGTG-CT
Armadillo ACTGCTGGG-CTGCATCAGGGGGTGTGCTGTCGGGTACTGGGGAGTG-CC
Opossum   ACTGCTGAGCTTGCACCAAATGATGCGCTGTCGGGTACTGAGGGGTG-CT
Chicken   ATTGCTGCGCCTGTACCAAGTGGTGCGCTGTGGGGTACTGGGGGCTG-CC
Frog      AGTGTTGGGCTTGCACCAAGTGATGTGCTGTAGGGTACTGGGCGTTA-CT
Fugu      ACTGTTGCGTCTGCACCAAGTGATGCGCTGTCGGGAACTGTGGCGTG-GC
Tetraodon ACTGCTGCGTCTGCACCAGGTGATGCGCTGTCGGGAACTGCGGCGTG-GC
Zebrafish ATGGCTGCATGTGGCCCAGATGAT----TGACAGATGATGTCAGATGTGT
              *  * **    **    *   * *    **   *    **
```



- Protein coding?
- ncRNA?
- Regulatory or other functional element?

Outline

- Motivation
- Review of available methods
- A simple new scoring scheme
  - Shuffling
  - Exact
- Benchmark of some available and the new method
- Significance measure
- Currently only pairwise, ungapped global case without stop codons: Hofstadter's law

Outline

- Motivation
- Review of available methods
- A simple new scoring scheme
  - Shuffling
  - Exact
- Benchmark of some available and the new method
- Significance measure
- Currently only pairwise, ungapped global case without stop codons: Hofstadter's law

It always takes longer than you expect, even when you take into account Hofstadter's Law

Motivation

- Why a coding model in `RNAz`?
  - Get rid of the “false positives” in mRNAs
  - Increase the information content of the output

Motivation

- Why a coding model in `RNAz`?
  - Get rid of the “false positives” in mRNAs
  - Increase the information content of the output
- Why yet another protein gene finder: Sturgeon’s law

Motivation

- Why a coding model in `RNAz`?
  - Get rid of the “false positives” in mRNAs
  - Increase the information content of the output

- Why yet another protein gene finder: Sturgeon’s law

90% of everything is crap

Motivation

- Why a coding model in `RNAz`?
  - Get rid of the "false positives" in mRNAs
  - Increase the information content of the output

- Why yet another protein gene finder: Sturgeon's law

90% of everything is crap

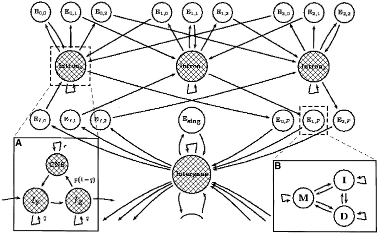
- Limitations of current coding potential detection approaches
  - Limited to pairwise alignments
  - Simplified models which do not include all available information
  - *Ad hoc* scores, poor statistics

Requirements

- Lightweight
- General
- Accurate
- Robust statistics
- Fast

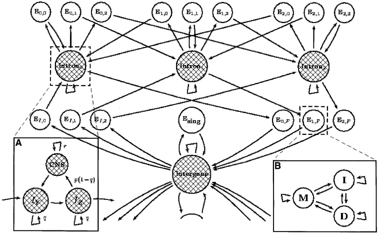
Plenty of Protein gene finders

- Full featured gene prediction
  - Genscan, Twinscan, N-Scan
  - SLAM
  - SGP2
  - Exoniphy



Plenty of Protein gene finders

- Full featured gene prediction
  - Genscan, Twinscan, N-Scan
  - SLAM
  - SGP2
  - Exoniphy
- Detection of coding potential
  - ETOPE (Ka/Ks ratio test)
  - CSTfinder
  - CRITICA
  - QRNA



K_a/K_s ratio test

1. Count synonymous and non-synonymous **sites** in both sequences.
2. Count synonymous and non-synonymous **differences**
3. Correct the observed differences and estimate the ratio of synonymous (K_s) and non-synonymous (K_a) **substitiutions per site**:
4. $K_a/K_s < 1 \Rightarrow$ purifying evolution

Nei & Gojobori *Mol. Biol. Evol.* **3**:418 (1986), Nekrutenko *et al. Nucl. Acids. Res.* **31**:3564 (2003)

K_a/K_s ratio test

1. Count synonymous and non-synonymous **sites** in both sequences.
2. Count synonymous and non-synonymous **differences**
3. Correct the observed differences and estimate the ratio of synonymous (K_s) and non-synonymous (K_a) **substitiutions per site**:
4. $K_a/K_s < 1 \Rightarrow$ purifying evolution

+ Properly normalized score

– Only considers synonymous changes (no conservative changes)

Nei & Gojobori *Mol. Biol. Evol.* **3**:418 (1986), Nekrutenko *et al. Nucl. Acids. Res.* **31**:3564 (2003)

CRITICA

- Scoring scheme based on theoretical considerations
  - Positive score for synonymous substitutions
  - Negative score for non-synonymous substitutions
- Also includes non-comperative score (di-nucleotide model)

Badger & Olsen *Mol. Biol. Evol.* **16**:512 (1999)

CRITICA

- Scoring scheme based on theoretical considerations
  - Positive score for synonymous substitutions
  - Negative score for non-synonymous substitutions
- Also includes non-comperative score (di-nucleotide model)

+ reasonable statistics

− Focused on bacteria, hard to use, no amino acid similarity

Badger & Olsen *Mol. Biol. Evol.* **16**:512 (1999)

CSTfinder

- Scans blast hits of ESTs for coding potential
- Defines Coding potential score:

$$CPS = (\frac{100}{N})(\frac{N_S + 1}{N_A + 1}) \sum_{i=1}^{N} s(c_i^A, c_i^B)$$

N	...	number of codon pairs
N_S,N_A	...	number of synonymous, non-synonymous pairs
c_i^A	...	codon number i in sequence A
$s(c_i^A, c_i^B)$	...	similarity of encoded amino acids

Mignone *et al. Nucl. Acids Res.* **31**:4639 (2003)

CSTfinder

- Scans blast hits of ESTs for coding potential
- Defines Coding potential score:

$$CPS = (\frac{100}{N})(\frac{N_S + 1}{N_A + 1}) \sum_{i=1}^{N} s(c_i^A, c_i^B)$$

N	...	number of codon pairs
N_S,N_A	...	number of synonymous, non-synonymous pairs
c_i^A	...	codon number i in sequence A
$s(c_i^A, c_i^B)$	...	similarity of encoded amino acids

+ considers amino acid similarity

− as *ad hoc* as it can be, no normalization, “Vaporware”

Mignone *et al. Nucl. Acids Res.* **31**:4639 (2003)

QRNA

- 3 pair hidden Markov models/SCFGs: Coding, RNA, other

$$P^{COD}(a_1a_2a_3, b_1b_2b_3) \approx P(a_1a_2a_3|A)P(b_1b_2b_3|B)P(A, B)$$

$a, b \in \mathcal{A} =$\{A,G,C,T\}, $A, B \in$ \{amino acids\}

Rivas & Eddy *BMC Bioinformatics* **2**:8 (2001)

QRNA

- 3 pair hidden Markov models/SCFGs: Coding, RNA, other

$$P^{COD}(a_1 a_2 a_3, b_1 b_2 b_3) \approx P(a_1 a_2 a_3 | A) P(b_1 b_2 b_3 | B) P(A, B)$$

$a, b \in \mathcal{A} =$\{A,G,C,T\}, $A, B \in$ \{amino acids\}

$$P(COD|\text{alignment}) = \frac{P(\text{alignment}|\text{COD})P(\text{COD})}{\sum_{\text{Models}} P(\text{alignment}|\text{Model})P(\text{Model}}$$

$$\text{Score} = \frac{P(\text{COD}|\text{alignment})}{P(\text{OTH}|\text{alignment})}$$

Rivas & Eddy *BMC Bioinformatics* **2**:8 (2001)

QRNA

- 3 pair hidden Markov models/SCFGs: Coding, RNA, other

$$P^{COD}(a_1a_2a_3, b_1b_2b_3) \approx P(a_1a_2a_3|A)P(b_1b_2b_3|B)P(A,B)$$

$a, b \in \mathcal{A} =$\{A,G,C,T\}, $A, B \in$ \{amino acids\}

$$P(COD|\text{alignment}) = \frac{P(\text{alignment}|\text{COD})P(\text{COD})}{\sum_{\text{Models}} P(\text{alignment}|\text{Model})P(\text{Model}}$$

$$\text{Score} = \frac{P(\text{COD}|\text{alignment})}{P(\text{OTH}|\text{alignment})}$$

+ considers amino acid similarity, elegant solution, can deal with frameshifts and local search

– no P value, independence assumption of codons and amino acids

Rivas & Eddy *BMC Bioinformatics* **2**:8 (2001)

A simple pairwise similarity score

Definitions

Alignment $\overline{AB}$ of sequence A and B:

$$\begin{array}{c} A: c_1^A c_2^A \ldots c_n^A \\ B: c_1^B c_2^B \ldots c_n^B \end{array}$$

L	$\ldots$	length in codons
$f_{\{A,G,C,T\}}$	$\ldots$	background frequency of nucleotides
ID	$\ldots$	pairwise identity
$d(c^A, c^B)$	$\ldots$	Hamming distance of two codons (e.g. $d(AGC, AGT) = 1$)
$s(c^A, c^B)$	$\ldots$	similarity of encoded amino acids (e.g. BLOSUM Matrix)

A simple pairwise similarity score

Normalizing with shuffling

- Unnormalized score

$$\widetilde{S}_{\overline{AB}} = \sum_{\substack{i=1 \\ d(c_i^A, c_i^B)>0}}^{L} s(c_i^A c_i^B)$$

- Shuffle columns: $\overline{AB}_{\text{random}}$

$$S_{\overline{AB}} = \widetilde{S}_{\overline{AB}} - \widetilde{S}_{\overline{AB}_{\text{random}}}$$

A simple pairwise similarity score

Exact normalization

- Calculate the *expected* score for pairs with 1,2 and 3 differences. e.g.:

$$\langle s_{d=1} \rangle = \frac{N^{\text{comb}}}{N^{\text{comb}}_{d=1}} \sum_{\substack{a,b,c,d,e,f \in \mathcal{A} \\ d(abc,def)=1}} s(c_{abc}, c_{def}) \prod_{i=a,b,c,d,e,f} f_i$$

- Correct each *observed* score by the *expected* score

$$S_{\overline{AB}} = \sum_{\substack{i=1 \\ d(c_i^A, c_i^B)>0}}^{L} s(c_i^A c_i^B) - \langle s_{d=d(c_i^A, c_i^B)} \rangle$$

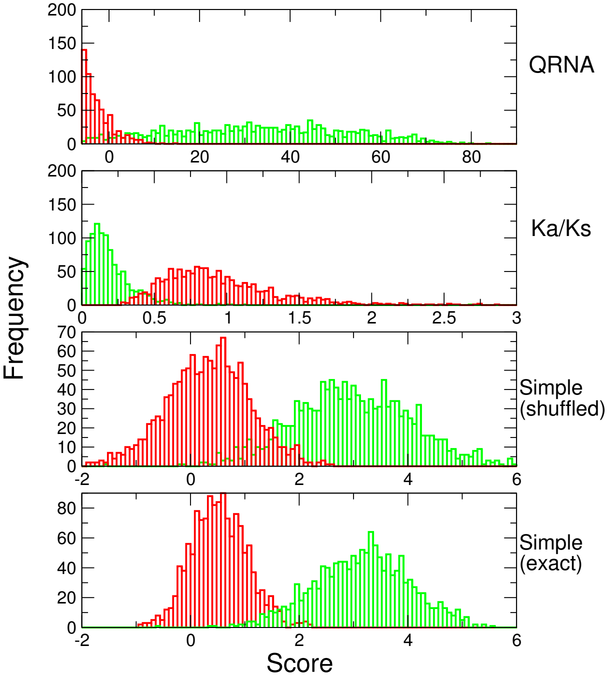
Test Set

- UCSC `Multiz` alignments (13-way)
- Extract mouse RefSeq genes from chromosome 1 and 10
- Take only “correct” genes which start with M and have exactly one stop codon on the last position.
- Select slices of different length (50–150 nts) and pairwise identity (60%–100%)
- Random control: Shuffle sequences, remove stop codons

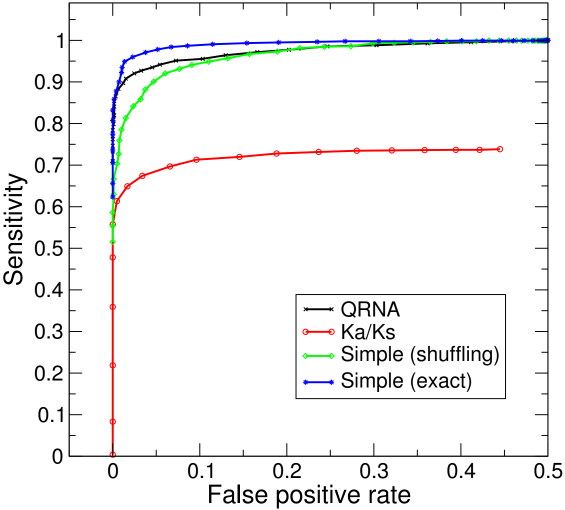
$\Rightarrow \approx$ **7000 positive and negative examples**

Score distribution of native and random alignments

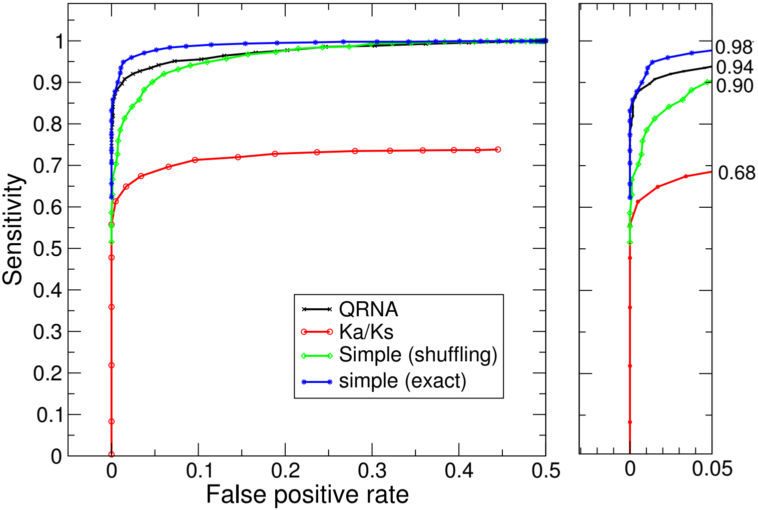
- 60%<ID<85%
- $L = 150$ nts



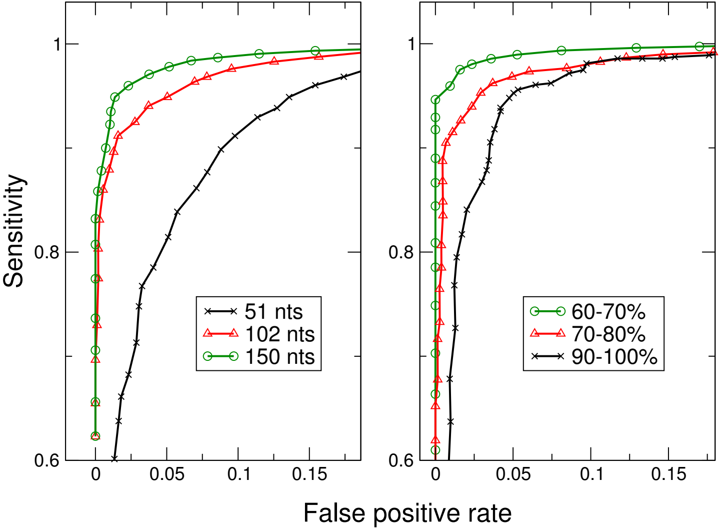
Comparison of methods (ROCs)



Comparison of methods (ROCs)



Dependence of length and sequence divergence



Estimating statistical significance

- Calculate the mean and variance of all sequences for a given (expected) base composition and pairwise identity. Assume normal distribution and calculate the P value.

$$\langle S \rangle_{ID,L} = L \sum_{a,b,c,d,e,f \in \mathcal{A}} s(c_{abc}, c_{def}) \prod_{i=a..f} (f_i) m_{d(abc,def)} \frac{N^{\mathrm{comb}}}{N^{\mathrm{comb}}_{d=d(abc,def)}}$$

Estimating statistical significance

- Calculate the mean and variance of all sequences for a given (expected) base composition and pairwise identity. Assume normal distribution and calculate the P value.

$$\langle S \rangle_{ID,L} = L \sum_{a,b,c,d,e,f \in \mathcal{A}} s(c_{abc}, c_{def}) \prod_{i=a..f} (f_i) m_{d(abc,def)} \frac{N^{\text{comb}}}{N^{\text{comb}}_{d=d(abc,def)}}$$

$$m_{d=0} = ID^3$$
$$m_{d=1} = ID^2(1-ID)\cdot 3$$
$$m_{d=2} = ID(1-ID)^2\cdot 3$$
$$m_{d=3} = (1-ID)^3$$

Estimating statistical significance

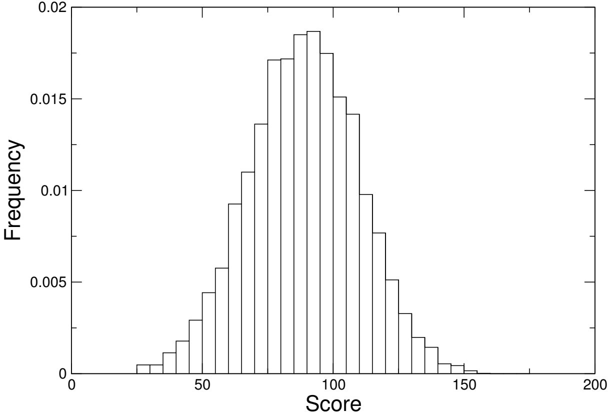
- Calculate the mean and variance of all sequences for a given (expected) base composition and pairwise identity. Assume normal distribution and calculate the P value.

$$\langle S \rangle_{ID,L} = L \underbrace{\sum_{a,b,c,d,e,f \in \mathcal{A}} s(c_{abc}, c_{def}) \underbrace{\prod_{i=a..f} (f_i) m_{d(abc,def)} \frac{N^{\text{comb}}}{N^{\text{comb}}_{d=d(abc,def)}}}_{K}}_{M}$$

$$m_{d=0} = ID^3$$
$$m_{d=1} = ID^2(1-ID) \cdot 3$$
$$m_{d=2} = ID(1-ID)^2 \cdot 3$$
$$m_{d=3} = (1-ID)^3$$

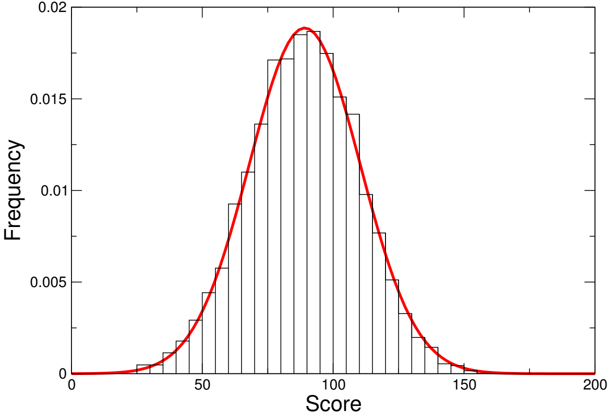
$$var(s)_{ID,L} = \sum_{a,b,c,d,e,f \in \mathcal{A}} (s(c_{abc}, c_{def})^2 K) - M^2$$

Sampled vs. calculated scores



- 10,000 alignments sampled with Markov method (black bars)

Sampled vs. calculated scores



- 10,000 alignments sampled with Markov method (black bars)
- Calculated distribution (red line)

Conclusions and outlook

- Comparative detection of coding potential is a useful feature
- Available methods are not perfect
- Considering amino acid similarity significantly improves accuracy compared to simply counting synonymous substitutions
- A simple and properly normalized score outperforms any other tested methods.
- The score allows direct calculation of a *P*-Value.

Conclusions and outlook

- Comparative detection of coding potential is a useful feature
- Available methods are not perfect
- Considering amino acid similarity significantly improves accuracy compared to simply counting synonymous substitutions
- A simple and properly normalized score outperforms any other tested methods.
- The score allows direct calculation of a *P*-Value.
- Include
  - stop codons
  - gaps (frameshifts)
  - local search?
- Extension to multiple alignments