Maximum Likelihood Estimation for RNA homology search

Bled 2009

Peter Menzel

Bioinformatics Group, IBHV, Faculty of Life Sciences, Copenhagen University
Professur für Bioinformatik, Institut für Informatik, Universität Leipzig

20.02.2009

RNA Homology Search

- RFAM 9.1 contains 1371 families of non-coding RNAs
- Tons of newly sequenced genomes can be expected in next years
- ncRNA annotation of already sequenced genomes is still sparse in many cases.
- What is the phylogenic distribution of a certain ncRNA family?
- Derive model for a family based on known sequences, search for homologs.

Search methods

- sequence based: `blastn`
- sequence + structure, automatic model learning: covariance models (`infernal`, `RaveNnA`), `erpin`
- sequence + structure, descriptor-based programs: `RNAMotif`, `rnabob`, `PatSearch`, ...
  - manually model properties of a family, (usually) very fast

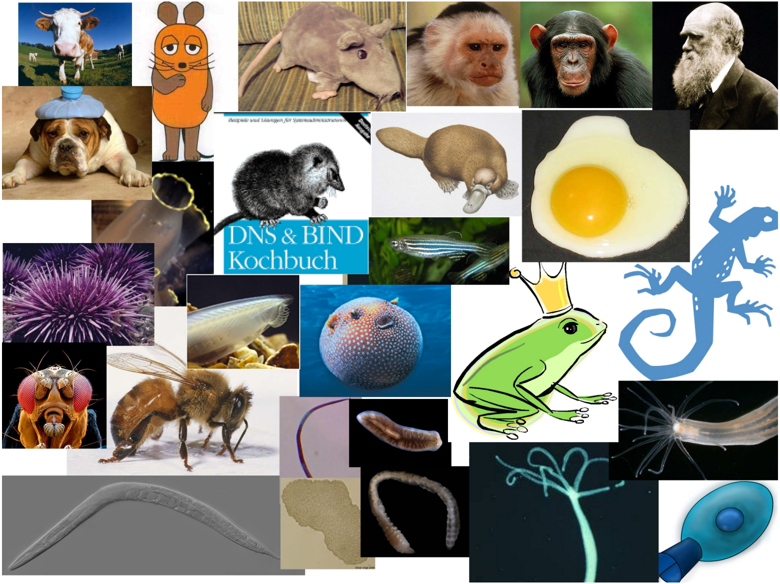
Comparison

- difficulties in ncRNA homolgy search: structure conservation, length variation, frequent small indels

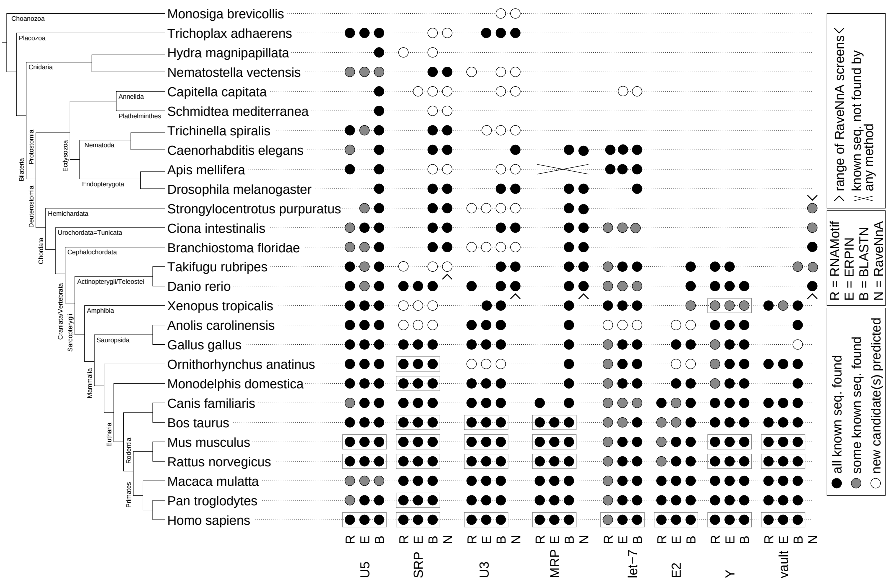
Little experiment

- How well do those methods generalize?
- 8 ncRNA families, 28 genomes
- `RNAMotif`, `blastn`, `erpin`, `RaveNnA`
- Phylogenetic restricted training set (structure annotated alignment)
- Derive models
- Search all genomes
- RNAMotif: manually derive descriptors from alignment, iteratively search genomes and modify descriptors (3 rounds)

Range of animal genomes



Results


Choanozoa
Placozoa
Cnidaria
Annelida
Plathelminthes
Nematoda
Endopterygota
Ecdysozoa
Protostomia
Bilateria
Deuterostomia
Hemichardata
Urochordata=Tunicata
Cephalochordata
Chordata
Actinopterygii/Teleostei
Amphibia
Craniata/Vertebrata
Sarcopterygii
Sauropsida
Mammalia
Eutharia
Rodentia
Primates
Monosiga brevicollis
Trichoplax adhaerens
Hydra magnipapillata
Nematostella vectensis
Capitella capitata
Schmidtea mediterranea
Trichinella spiralis
Caenorhabditis elegans
Apis mellifera
Drosophila melanogaster
Strongylocentrotus purpuratus
Ciona intestinalis
Branchiostoma floridae
Takifugu rubripes
Danio rerio
Xenopus tropicalis
Anolis carolinensis
Gallus gallus
Ornithorhynchus anatinus
Monodelphis domestica
Canis familiaris
Bos taurus
Mus musculus
Rattus norvegicus
Macaca mulatta
Pan troglodytes
Homo sapiens
R E B
U5
R E B N
SRP
R E B N
U3
R E B N
MRP
R E B
let−7
R E B
E2
R E B
Y
R E B N
vault
> range of RaveNnA screens
╳ known seq. not found by any method
R = RNAMotif
E = ERPIN
B = BLASTN
N = RaveNnA
● all known seq. found
● some known seq. found
○ new candidate(s) predicted

Automatic descriptor design

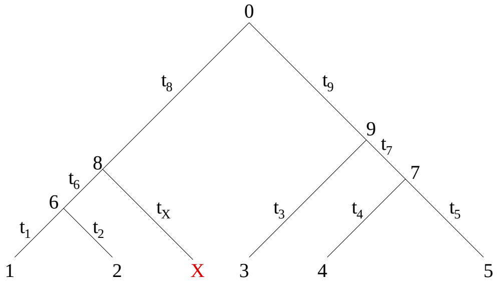
- Current problems in model design: Manually constructing descriptors for a specific ncRNA family is very hard for non-experts
- Tradeoff between specificity and sensitivity, small changes in the descriptor can dramatically affect number of hits
- What we want to have: Given a structure-annotated sequence alignment, automatically build the best descriptor for that family
- Even better: Build a descriptor aimed for searching in one target species

Targeted descriptor design

Idea

Which sequence and structure can be expected in a certain species?

- Knowing the relations between species, build a model optimized for a target organism
- Trade sensitivity off for specificity



Framework

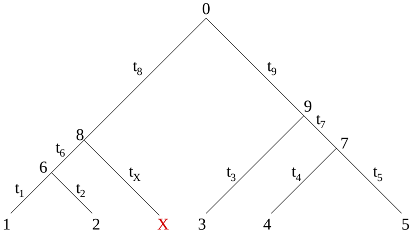
- Assume different evolutionary rates for different parts of the RNA molecule, e.g. loop regions are known to evolve faster than stems.
- Given a phylogenetic tree and a multiple alignment, calculate a mutation rate μ for each paired and unpaired column which maximizes the likelihood of the tree
- Then calculate probabilities for base occurences for a target species in the tree.
- High probability columns contribute to the model
- Derive descriptor (e.g. RNAMotif)

Framework

```
# STOCKHOLM

Seq1            CCAUCCGUAAGGGCGGUUGG
#=GR Seq1 SS    (((.((((....)))).)))
Seq2            GCAUCCGUAAGGGCGGUUGC
#=GR Seq2 SS    ((..((((....)))).)))
Seq3            GCAUCCGGAAGGGCGGUAGC
#=GR Seq3 SS    ((..(((......)))..))
Seq4            GCAUCCGGAAGGGGGGUAGC
#=GR Seq4 SS    ((..((........))..))
Seq5            GCAUC-----------UAGC
#=GR Seq5 SS    ((...-----------..))
#=GC SS_cons    ((..((((....))))..))

SeqX            ????????????????????
#=GC SeqX SS    ????????????????????
```



Algorithm

1. Delete leaf X (target species) from tree T
2. Optimize a mutation rate μ for each paired and unpaired column for maximizing the likelihood of the tree (root 0)

$$\hat{\mu} = \operatorname*{argmax}_{\mu} L(\mu)$$

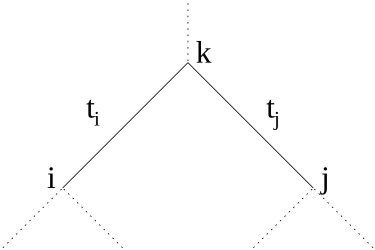
$$L(\mu) = \sum_{s_0} \pi_{s_0} L_{s_0}(\mu)$$

3. Add X to T and root T at X
4. Using the estimated $\hat{\mu}$, recalculate likelihoods for all states s_0 of the root node 0

Algorithm

- Tree likelihood is calculated by post-order traversal of the tree (leafs are evaluated first)
- For interior nodes k

$$L_{s_k}(\mu) = \left(\sum_{s_i} P_{s_k s_i}(t_i, \mu) L_{s_i}(\mu)\right) \cdot \left(\sum_{s_j} P_{s_k s_j}(t_j, \mu) L_{s_j}(\mu)\right)$$

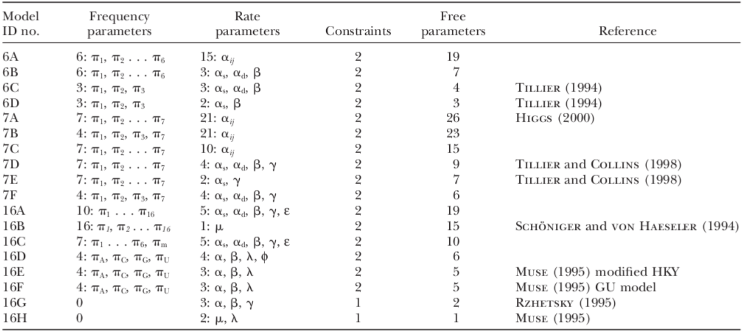


Algorithm

- Leaf nodes: $L_{s_k}(\mu) = 1$ if state s_k is in alignment, otherwise $L_{s_k}(\mu) = 0$
- Transition matrix $P_{s_k s_i}(t_i, \mu)$ gives probability for changing s_i into s_k in time t_i given mutation rate μ
- derived from a rate matrix Q containing instentanious substitution rates: $P_{xy}(t, \mu) = [e^{t\mu Q}]_{xy}$
- Q is a substitution rate model, unpaired and paired models, empirical models based on substitution rates derived from alignments (like RIBOSUM)

Problems

- Gaps not contained in most models, but in most of the alignments



Model ID no.	Frequency parameters	Rate parameters	Constraints	Free parameters	Reference
6A	6: $\pi_1, \pi_2 \dots \pi_6$	15: α_{ij}	2	19	
6B	6: $\pi_1, \pi_2 \dots \pi_6$	3: $\alpha_s, \alpha_d, \beta$	2	7	
6C	3: π_1, π_2, π_3	3: $\alpha_s, \alpha_d, \beta$	2	4	Tillier (1994)
6D	3: π_1, π_2, π_3	2: α_s, β	2	3	Tillier (1994)
7A	7: $\pi_1, \pi_2 \dots \pi_7$	21: α_{ij}	2	26	Higgs (2000)
7B	4: $\pi_1, \pi_2, \pi_3, \pi_7$	21: α_{ij}	2	23	
7C	7: $\pi_1, \pi_2 \dots \pi_7$	10: α_{ij}	2	15	
7D	7: $\pi_1, \pi_2 \dots \pi_7$	4: $\alpha_s, \alpha_d, \beta, \gamma$	2	9	Tillier and Collins (1998)
7E	7: $\pi_1, \pi_2 \dots \pi_7$	2: α_s, γ	2	7	Tillier and Collins (1998)
7F	4: $\pi_1, \pi_2, \pi_3, \pi_7$	4: $\alpha_s, \alpha_d, \beta, \gamma$	2	6	
16A	10: $\pi_1 \dots \pi_{16}$	5: $\alpha_s, \alpha_d, \beta, \gamma, \varepsilon$	2	19	
16B	16: $\pi_{1}, \pi_{2} \dots \pi_{16}$	1: μ	2	15	Schöniger and von Haeseler (1994)
16C	7: $\pi_1 \dots \pi_6, \pi_m$	5: $\alpha_s, \alpha_d, \beta, \gamma, \varepsilon$	2	10	
16D	4: $\pi_A, \pi_C, \pi_G, \pi_U$	4: $\alpha, \beta, \lambda, \phi$	2	6	
16E	4: $\pi_A, \pi_C, \pi_G, \pi_U$	3: α, β, λ	2	5	Muse (1995) modified HKY
16F	4: $\pi_A, \pi_C, \pi_G, \pi_U$	3: α, β, λ	2	5	Muse (1995) GU model
16G	0	3: α, β, γ	1	2	Rzhetsky (1995)
16H	0	2: μ, λ	1	1	Muse (1995)

Acknowledgements

- Peter and Jan
- Stefan W., Mario, Mandy, Stefan S. and other people in Leipzig and Copenhagen