Genomic identification of Structural RNAs using phylo-SCFGs

Jakob Skou Pedersen
Bioinformatics Center, University of Copenhagen
Structural RNA identification problem

ncRNA: any transcribed region with functional structure.

ncRNAs co-transcribed with protein-coding genes as:

- Independently transcribed ncRNAs
- As co-transcribed with protein-coding genes
Highly diverse set

- miRNA
- tRNA
- RNase P
- Xist (~20 kb long)

Single sequence signal:
- Lack of common nucleotide biases
- Lack of common sequence motifs
Evolutionary signal

Signal

Unprecedented comparative data
phylogenetic models

Continuous time Markov chain acting on branches of phylogenetic tree.

Features:
- Nucleotide biases
- Patterns of substitution
- Evolutionary sequence correlations
- Correlated changes (multi nucleotide models)

Substitution rates of Markov chain

Alignment

\[ P(\theta_1, \ldots, \theta_n | \xi, \eta) \quad \text{comp}(TQ) \quad \text{GAG} \]
models

- Single-nucleotide model
  - 4x4 rate matrix
  - Marginal average of di-nucleotide matrix
  - Fast substitution rate

<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>C</td>
</tr>
</tbody>
</table>

- Single-nucleotide model
  - 16x16 rate matrix
  - Learned from data
  - Favors pairing di-nucleotides
  - Slow substitution rate
EvoFold SCFGs

Structural model:

Non-structural model:

[Diagram showing RNA secondary structure with nucleotides T, G, C, A, and transitions between structural and non-structural states]
Structure derivation

begin structural (S)

stem pair (P)

bifurcation (B)

intermediate (S)

emit (S)

loop & bulge (L)

end (E)

Structure

GAGCTTGGCTTTTGGGCAGC
leotide c model

Di-nucleotide phylogenetic model

fold .(((((((((....))))))))
human GAGCTTGCTTTTGCGCAGCT
chimp. GAGCTTGCTTTTGCGCAGCT
mouse GAGTCTTGTATTTGGCGCAGCT
rat AAGCTTACTTACGTAGCT
dog GAGCATAACTAAGGTGGCT
chicken GGGCTTACGCTGGTGCGCC
z. fish GGGCTTACAATTTGTGGCC
p. fish GGGCTTAAAAATTGGGCC
Single nucleotide phylogenetic model

\[
\text{score} = \log \left( \frac{P(x|\phi_{str})}{P(x|\phi_{null})} \right)
\]

Di-nucleotide phylogenetic model

---

to fold

human
chimp.
mouse
rat
dog
chicken
z. fish
p. fish

---

fold

T - T
G
C - G
G - C
T - A
C - G
Algorithms and training

Algorithms: Traditional SCFG algorithms (CYK and insideoutside) combined with Felsenstein 81.

Complexities:

- Space: $O(mn^3)$
- Time: $O(n^2)$

Training of EvoFold:

Rfam structures mapped onto genomic alignment.
vertebrates & drosophilids

Input: conserved segments

Output: sub-folds

Sensitivity: 43%

Performance: 43%
Experimentally studied vertebrate cases

HAR1

Expression in developing neocortex

HAR1 F AS

Reelin

RA3 A-to-I RNA editing substrate
High confidence subset from Drosophila screen

Selection criteria:
- Min. two compensatory substitutions
- \#compensatory subs > 2 x \#contradictory subs

Predictions (394 total)

Genomic background

- 65% (8.2%)
- 40% (4.0%)
- 53% (5.3%)
UTR structures

A high fraction of UTR predictions on transcribed strand (5’UTR: 80% & 3’UTR: 90%).

Significant enrichment of genes regulatory roles.

Gene involved in biogenesis and assembly of the ribosome
(by homology to RPL24)
Acknowledgments

EvoFold: Gill Bejerano (Stanford), Adam Siepel (Cornell), Kate Rosenbloom (UCSC), Kirstin Lindblad-Toh (Broad), Eric S. Lander (Broad), Jim Kent (UCSC), Webb Miller (Penn State), and David Haussler (UCSC)

HAR1 study: Katherine S. Pollard (UCSC), Sofie R. Salama (UCSC), Nelle Lambert (ULB), Marie-Alexandra Lambot (ULB), Sandra Coppens (ULB), Sol Katzman (UCSC), Ryan King (UCSC), Courtney Onodera (USCS), Adam Siepel (Cornell), Andrew D. Kern (UCSC), Colette Dehay (Lyon), Haller Igel (UCSC), Manuel Ares, Jr (UCSC), Pierre Vanderhaeghen (ULB)

ABRA3 study: Johan Ohlson, Marie Ohman (Stockholm University), and David Haussler (UCSC)

Fly screen: Manolis Kellis (MIT), David Haussler (UCSC), Drosophila Sequencing and Analysis Consortium
Structure predictions

New case of A-to-I RNA editing

Mouse Brain
Intronic hairpin in RDE flanked by A-to-I edited exons
Spen function: Transcription co-factor and involved in neuronal cell fate, survival, and axon guidance. It has three RNA recognition motifs.
localization element in Orb

Stau

Orb
Hairpin in highly expressed intergenic region
Structure can be extended

RNAfold str.  EvoFold str.

Alignment & full structure
a. Intrinsic structure (103 cases)

[Diagram showing RNA structure with nucleotide sequences labeled from 1 to 41]