Prediction of structural non-coding RNAs by comparative sequence analysis

Stefan Washietl

Institute for Theoretical Chemistry
University of Vienna

Vienna, October 31st 2005
Outline

1. Introduction to non-coding RNAs and motivation of this work
2. New algorithms for detection of structural non-coding RNAs
3. A large scale screen of the human genome
4. Other applications
Non-coding RNAs

Non coding RNAs (“RNA genes”) are transcripts that exert their function as RNA without being translated to protein.

“Classical” examples:

- Protein expression: transfer RNA, ribosomal RNA
- Pre-mRNA splicing: spliceosomal RNAs
- tRNA maturation: Ribonuclease P
- Protein export: Signal recognition particle RNA

New abundant classes of small non-coding RNAs: microRNAs, snoRNAs

Many other examples are currently emerging in all organisms studied.
1. A vast amount of genomic data is available
Motivation

1. A vast amount of genomic data is available
2. There are fewer protein coding genes than expected
4. The transcriptional map of the human genome is much more complex than expected.
Non-coding RNAs?

4. The transcriptional map of the human genome is much more complex than expected.
Computational identification of non-coding RNAs

- Based on *a priori* knowledge: find members of known families
  - Sequence similarity alone: BLASTN
  - Sequence and additional motif information: specialized programs for e.g. tRNA or snoRNAs
- *De novo* prediction: find new genes and families
  - Unlike protein coding genes (ORFs, codon bias, . . .) ncRNAs lack strong statistical signals in primary sequence
  - The function of many ncRNA depend on a defined secondary structure

Can secondary structure predictions be used for ncRNA detection?
Significance of predicted RNA secondary structures: z-score statistics

Has a natural occurring RNA sequence a lower MFE than random sequences of the same size and base composition?

1. Calculate native MFE $m$.
2. Calculate mean $\mu$ and standard deviation $\sigma$ of MFES of a large number of shuffled random sequences.
3. Express significance in standard deviations from the mean as z-score

$$z = \frac{m - \mu}{\sigma}$$

Negative z-scores indicate that the native RNA is more stable than the random RNAs.
z-scores for 579 tRNAs

- Only 2% below a z-score threshold of $-4$.
- Native sequences are not clearly separated from the random bulk.
Consensus folding using RNAalifold

- RNAalifold uses the same algorithms and energy parameters as RNAfold
- Energy contributions of the single sequences are averaged
- Covariance information (e.g. compensatory mutations) is incorporated in the energy model.
- It calculates a consensus MFE consisting of an energy term and a covariance term:

```plaintext
((((((...........)))))((((...........)))))......(((...........)))))
GTTTCCGTAGTGTAGCGTTATCACATTGCCTACACACGCAGAAAGGTCCCCCGGTTCGATCCCCGGGCAGAAACA
GTTTCCGTAGTGTAGTTATACGTTGCCTAAGACGCAGAAAGGTCCCCGGTTCGAAACCGGGCGGAACAA
GTTTTCTCGTAGTGTAGTGTATACGTTGCTTCACACGCAAGGTCCCCGGTTGAAACGGGCAGAAACAA

**** ********** ********** * ** * ***** ********************** ********** *****
(-24.76 = -23.43 + -1.33)```
z-scores of consensus MFEs for tRNA alignments

- ▶ Alifoldz: Additional information from aligned sequences shifts MFE predictions towards significant levels.
The structure conservation index

The SCI is an efficient and convenient measure for secondary structure conservation.
The significance of a predicted MFE structure can be expressed as a z-score which is normalized w.r.t. sequence length and base composition.

Traditionally, z-scores are sampled by time-consuming random shuffling.

The shuffling can be replaced by a Support Vector Machine regression calculation which is of the same accuracy.
SVM classification based on both scores

- Both scores separate native ncRNAs from controls in two dimensions.
SVM classification based on both scores

- Both scores separate native ncRNAs from controls in two dimensions.
- A support vector machine is used for classification.
SVM classification based on both scores

- Both scores separate native ncRNAs from controls in two dimensions.
- A support vector machine is used for classification.
- RNAz: more accurate and faster than any other available programs.
Screening the human genome

- Large scale comparative screen of mammals/vertebrates
- \( \approx 5\% \) of the best conserved non-coding regions
- \( \rightarrow 438,788 \) alignments covering 82.64 MB (2.88% of the genome)
Statistics of detected structures

(a) Pie charts showing the distribution of structural RNA, estimated false positives, and other conserved noncoding elements at different confidence levels ($P > 0.5$ and $P > 0.9$).

(b) Bar graph illustrating the number of structural elements across different confidence levels for different categories (4 Mammals, 4 Mammals + chicken, All vertebrates).

(c) Pie charts for microRNA, H/ACA, and C/D snoRNA categories, showing detected ($P > 0.9$), detected ($0.5 < P < 0.9$), not detected, and not in input set.

(d) Pie chart breakdown of structural elements based on known gene, intron of coding region, 3'-UTR (exon or intron), 5'-UTR (exon or intron), and distances from nearest gene.
Novel structural RNAs of known classes: mirRNAs and H/ACA snoRNAs
Novel structures of unknown function
Other applications: Cyanobacterial ncRNAs

I. Axmann, P. Kensche et al. (*Genome Biol.* 6:R73, 2005) identified and characterized 7 novel ncRNAs in cyanobacteria using Alifoldz.
Other applications: Benchmarking alignment programs on structural RNAs

The SCI can be used to assess the quality of an alignment of a structural RNA (P. Gardner, A. Wilm & S. Washietl *Nucleic Acids Res.* **33**:2433, 2005).
Other applications

- Prediction of putative miRNA precursors in the miRNAMap (Hsu et al., submitted)
Summary and Conclusions

- *De novo* ncRNA prediction is notoriously difficult.
- Single sequence methods are of limited statistical significance.
- Comparative approaches dramatically improve accuracy.
- RNAz is an accurate and efficient approach for predicting ncRNAs.
- RNAz used for the first comprehensive annotation of conserved RNA secondary structures in the human genome.
- The data provides a strong basis for further computational and experimental studies.
- The programs and methods presented here were successfully used in a variety of other applications.
Acknowledgements

- Peter F. Stadler (Univ. Leipzig)
- Ivo Hofacker, Peter Schuster (Univ. Vienna)
- Paul Gardner (Univ. Copenhagen), Andreas Wilm (Univ. Düsseldorf)
- Alexander Hüttenhofer, Melanie Lukasser (Univ. Innsbruck)
- All other colleagues from Leipzig and Vienna