Prediction of noncoding RNAs with RNAz

John Dzmil, III
Steve Griesmer
Philip Murillo

April 4, 2007
What is non-coding RNA (ncRNA)?

- RNA molecules that are not translated into proteins
- Size range from 20 to 1000’s of nucleotides in length
- Significantly gained scientific interest since 1990’s
  - Originally thought as intermediates or accessories in protein biosynthesis
    - Little was known of their importance
    - Majority of research and funding towards protein coding RNA (messenger RNA)
  - Improved scientific methods and sequencing techniques
    - Led to the discovery of novel functions
    - Led to further classifications of RNA
  - Discovery of ten of thousands of ncRNA expressed in human cells
    - more ncRNA’s expressed in human cells than protein coding RNA’s.
Function of ncRNA?

- Structural, regulatory and catalytic molecules of protein biosynthesis
- Maturation of mRNA, tRNA and rRNA
- X-chromosome inactivation in mammals
- Gene regulation
Types of ncRNA

- **Transfer RNA (tRNA)**
  - ~73 – 93 nucleotides in length
  - **Function**
    - Transfer specific amino acid to ribosomal site during protein synthesis (translation)
  - **Specialized L-shape structure**
    - Allows tRNA to “dock” onto ribosomal site for amino acid transfer
Types of ncRNA (cont.)

- **Ribosomal RNA (rRNA)**
  - Primary constituent of ribosomes
    - Ribosomes primary role is to assemble polypeptides from amino acids (translation)
    - Ribosomal proteins combined with rRNA to create ribosome
  - Make up the majority of RNA found within a typical cell

- **Small nuclear RNA (snRNA)**
  - Located in nucleus of eukaryotic cells
  - Function
    - RNA splicing
    - Regulation of transcription factors
    - Maintaining telomeres
Types of ncRNA (cont.)

- **Small Nucleolar RNA (snoRNA)**
  - Located in the nucleolus
    - Ribosomes primary role is to assemble polypeptides from amino acids (translation)
    - Ribosomal proteins combined with rRNA to create ribosome
  - Function
    - Enhance functionality of mature RNA
      - chemical modifications to rRNA and other RNA genes (ex. methylation)

- **Micro RNA**
  - ~20 – 23 nucleotides in length
  - Single stranded
    - Complimentary to one or more messenger RNA (mRNA)
  - Function
    - Regulates gene expression
      - anneals itself to mRNA inhibiting translation
Why is it hard to predict non-coding RNA?

- Unlike protein coding genes, functional RNAs lack statistical signals for reliable detection from primary sequences.
- There is no protein product for which the ncRNAs are coding:
  - No evolutionary constraints on protein product
  - Constraints come in secondary RNA structure
    - Can be conserved even with substantial changes to primary DNA sequence.
How do ncRNA prediction programs overcome this problem?

- QRNA – uses pairwise alignment, but low reliability
- MSARI – uses multiple sequence alignments of 10-15 sequences with high sequence diversity; highly accurate
- RNAz – combines sequence alignment of 2-4 sequences with measures of:
  - Structural conservation
  - Thermodynamic stability
RNAz

- Predicts noncoding RNA sequences
- Relies on two features of structural noncoding RNAs:
  - Thermodynamic stability
  - Secondary structure conservation
- Uses comparative sequence analysis of 2-4 sequences
- Builds on other RNA programs to accomplish goal:
  - RNAFOLD – folding single sequences
  - RNAALIFOLD – consensus folding of aligned sequences
  - LIBSVM – support vector machine (SVM) learning
Thermodynamic stability

- Measure mean free energy (MFE)
- Compares mean free energy of given sequence to random sequences of same length and base composition
- Z-score calculated as:

\[ z = \frac{(m - \mu)}{\sigma} \]

where \( \mu \) and \( \sigma \) are the mean and standard deviations of the random sequences, respectively.

- Negative \( z \) scores indicate that a sequence is more stable than expected by chance.
Structural conservation

- Uses RNAalifold
  - Like RNAfold except augmented with covariance information
- For covariance information, compensatory mutations (e.g. a CG pair mutates to a UA pair) and consistent mutations (e.g. AU mutates to GU) give a bonus of energy while inconsistent mutations (e.g. CG mutates to CA) yield a penalty of energy
- Results in consensus MFE $E_A$.
- RNAz compares $E_A$ to average MFE of individual sequences ($E_{avg}$)
- Structural conservation index calculated as:

  $$SCI = \frac{E_A}{E_{avg}}$$

- SCI high $\Rightarrow$ sequences fold together equally well as fold individually
- SCI low $\Rightarrow$ no consensus fold
Combining z and SCI scores

- Z- and SCI scores used to classify the alignment as “structural noncoding RNA” or “other” using Support Vector Machine (SVM) learning algorithm
- Trained using a large set of well-known noncoding RNA sequences
RNAz: Input and Output

- **Input requires aligned sequences in ClustalW or MAF formats**
- **Output provides:**
  - Properties of sequences (number of sequences and base pairs, reading direction, pairwise identity)
  - Thermodynamic scores (MFE for sequences and consensus, energy contribution, covariance contribution, z-scores)
  - Secondary structure conservation (structure conservation index)
  - Classification prediction (SVM decision value, class probability, prediction)
  - Predicted secondary structure of each sequence and consensus for whole alignment
Example: Iron Response Element (IRE) RNA Input

CLUSTAL W (1.83) multiple sequence alignment

sacCer1
GCCTTGTGGCGCAATCGGTAGCGCGTATGACTCTTAATCATAAAGGTTAGGGGTTTCGAGC
sacBay
GCCTTGTGGCGCAATCGGTAGCGCGTATGACTCTTAATCATAAAGGTTAGGGGTTTCGAGC
sacKlu
GCCTTGTGGCGCAATCGGTAGCGCGTATGACTCTTAATCATAAAGGCTAGGGGTTTCGAGC
sacCas
GCTTCAGTAGCTCAGTCGGAGAGCGTCAGTCTCATAATCTGAAGGTCGAGAGTTTCGAGC
** * * ** **** ** ***** * *** ****** **** * *********

sacCer1
CCCCTACAGGGCT
sacBay
CCCCTACAGGGCT
sacKlu
CCCCTACAGGGCT
sacCas
CTCCCCTGGAGCA
* ** * **
Example: Iron Response Element (IRE) RNA Output
IRE RNA Structures Using RNA Fold

Mouse

Fugu

Rat

Zebrafish

RNAFOLD: MFE = -19.66 kcal/mol
MFE = -19.70 kcal/mol
MFE = -19.44 kcal/mol
MFE = -22.94 kcal/mol

Average MFE = -20.43 (vs. -19.23 for output of RNAz)
Consensus Folding via RNAALIFOLD

$\text{MFE} = \text{E}_A = -17.76 \text{ kcal/mol}$

$\text{SCI} = \frac{\text{E}_A}{\text{E}_{\text{avg}}} = \frac{-17.76}{-19.23} = 0.92$

Fold together equally well as individually
Classification of Z scores and SCI using SVM

- **Z score** = -3.24
- **SCI** = 0.92

Green = high probability of structural ncRNA

Red = low probability of structural ncRNA

High probability of structural noncoding RNA
3 Algorithms in RNAz

- Calculation of z-score
- Calculation of SCI
- SVM for classification of consensus as “structural noncoding RNA” or “other”

We will explain each of these algorithms in turn
Calculation of z-score

- Generated synthetic combinations of different length and base composition
  - 50 – 400 nucleotides in steps of 50 (8 sizes)
  - GC/AT, A/T, G/C ratios of sequences ranging from 0.25 to 0.75 in steps of 0.05 (11 percentages per ratio type)
  - 10,648 combinations (= 8 x 11 x 11 x 11)
- For each combination, generate 1000 random sequences and calculated mean and standard deviation of MFE
- Used SVM library LIBSVM to train 2 regression models for mean and standard deviation ($\mu$ and $\sigma$) rather than using random sampling. Verified accuracy by comparison of SVM algorithm and sampling.
- Z score calculation:

\[
z = \frac{(MFE - \mu)}{\sigma}
\]

where $\mu$ is the mean of sequences with a given length and base composition and sigma is the standard deviation
Accuracy of using SVM for Z-score Calculation

- Comparison of z scores through two methods:
  - Sampling
    - 100 sequences from random locations in human genome
    - 100 known ncRNAs from Rfam database
  - Using SVM regression model
- SVM model eliminates need for extensive sampling
Calculation of SCI

SCI calculation:

\[ \text{SCI} = \frac{E_A}{E_{\text{avg}}} \]

where \( E_A \) is the consensus MFE of the aligned sequences and \( E_{\text{avg}} \) is the average MFE of the individual sequences

\( E_A \) calculated through RNAALIFOLD
Support Vector Machines

- Support Vector Machines provide a means of classifying data into different classes or categories.
- Binary classifier separates data into two separate classes.
- Goal: Find hyperplane with the maximum margin that separates two classes of data.
  - Reduces impact of changes in underlying model.
  - Minimizes false positives.

![Diagram of Support Vector Machines](attachment:image.png)
Each value represented by tuple $(x_i, y_i)$ ($l = 1, 2$ in this example) where $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ corresponds to the attribute set for the $i$th value. $y_i$ can either be 1 or -1 to denote the binary choice.

Decision boundary of linear classifier has form:

$$w \cdot x + b = 0$$

where $w$ and $b$ are parameters in the model.

For test value $z$:

$$y = \begin{cases} 
1, & \text{if } w \cdot z + b \geq 0 \\
-1, & \text{if } w \cdot z + b < 0
\end{cases}$$
Training with SVM

Train model with data that has already been classified

- For this presentation, this means known ncRNA and known non-ncRNA.
- For a linear model, the training data is used to set \( w \) and \( b \) (after scaling) such that:

\[
\min f(w) = ||w||^2 / 2 \text{ subject to } y_i(w \cdot z_i + b) \geq 1, i = 1, 2, \ldots, N
\]

- \( w \cdot z + b \geq 1 \) if \( y_i = 1 \) (i.e., for known ncRNA),
- \( w \cdot z + b < 1 \) if \( y_i = -1 \) (i.e., for known non-ncRNA)
- Must also maximize the margin:

  - Equivalent to:

\[
\min f(w) = ||w||^2 / 2 \text{ subject to } y_i(w \cdot z_i + b) \geq 1, i = 1, 2, \ldots, N
\]
Two Additional SVM Issues

Two additional SVM issues need explanation for this paper:

1. What if training data not outside of margin because of noise in the training data?
2. What if two classes cannot be separated by a line?

To handle the first issue, positive slack variables are added into the constraints of the $f(w)$ optimization such that:

$$
\min_w f(w) = \frac{||w||^2}{2} + C(\sum_{i=1}^{N} \xi_i) \text{ subject to } y_i(w \cdot z_i + b) \geq 1 - \xi_i, \ i = 1, 2, ..., N
$$

where $C$ and $k$ represent penalties for misclassifying training instances.

To handle the second issue, we transform the data from its original space to a transformed space with a mapping function $\Phi(x)$ where there is a linear hyperplane between the two datasets. This mapping has the property:

$$
K(u, v) = \Phi(u) \cdot \Phi(v) = (u \cdot v + 1)^2
$$

where $K$ is a kernel function.

Only certain kernel functions can be used. Some common ones include:

- Polynomial: $K(x, x) = (\gamma x^T x + r)^d, \ \gamma > 0,$
- Radial basis function: $K(x, x) = \exp(-\gamma ||x - x||^2), \ \gamma > 0,$
- Sigmoid $K(x, x) = \tanh(\gamma x^T x + r)$
Back to Paper: Classification SVM

- Binary classification SVM trained to classify alignments as “RNA” or “other”
- Classification parameters were:
  - Mean of MFE z scores of the individual sequences
  - SCI
  - Mean pairwise identity
  - Number of sequences in the alignment
- Training data
  - All classes of ncRNA with exception of tmRNAs and U70 small nucleolar RNAs
  - For each native alignment, included one randomized version
- Testing
  - Generated models from all classes, leaving out one class at a time
  - Alignments with mean pairwise identities between 50-100%
- Kernel function
  - Radial basis function $K(x,x) = \exp(-\gamma \|x - x\|^2)$, with $\gamma = 2$
  - Slack penalty variable $C = 32$
Resulting ncRNA Classification

- Alignments of tRNAs and 5S rRNAs with 2-4 sequences per alignment and mean pairwise identities between 60-90%
- Green circles – native alignments
- Red crosses – shuffled random controls
- Background color indicates RNA class probability in z-SCI plane
Results of RNAz

- At cutoff of classification probability (P) of 0.9 over 12 ncRNA types:
  - Average sensitivity = 72.27%
  - Average specificity = 98.93%
- Results varied by ncRNA type:
  - U70 snoRNA – stable but not well conserved
  - tmRNA – conserved, but not stable
- Scan of Comparative Regulatory Genomics (CORG) database:
  - 89 ncRNA regions with P > 0.5
  - 11 known ncRNAs; 78 unknown
  - Hits in 5’ UTRs of protein coding genes, introns, unannotated regions
References

- Hsu, C-W., Chang, C-C., and C-J. Lin. “A Practical Guide to Support Vector Classification.”

- Tan, P-N., Steinback, M., and V. Kumar. 2005. Introduction to Data Mining.


- Washietl, S. 2006. “RNAz 1.0: Predicting structural non-coding RNAs.” Dept. of Theoretical Chemistry, University of Vienna.