

## A Learning Framework for Detecting Remote Non-Coding RNA Homologies

Keyur Desai<sup>1</sup>, John R. Deller Jr.<sup>2</sup> and Hayder Radha<sup>3</sup>

<sup>1</sup>desaikey,<sup>2</sup>deller,<sup>3</sup>radha@egr.msu.edu, Michigan State University

Some genes produce transcripts that function directly in a catalytic, regulatory or a structural role in the cell. These non-coding RNAs (ncRNAs) are prevalent in all living organisms and their number scales directly with the complexity of the organism. Recent years have witnessed a great increase in the size of databases that provide trusted ncRNA candidate sequences. For example, the popular RNA family (RFAM) database [1] has evolved from 25 families in version 1.0 (release date August 2002) to 503 families in version 7.0 (release date March 2005). This entails two important tasks: (i) Finding homologues of a query sequence within the database (homology detection problem); (ii) Deciding which RNA family a query sequence belongs to (classification problem). Interestingly, what makes an ncRNA gene a functional transcript is the phenomenon of base-pairing, which gives rise to its secondary structure (SS). The SS dictates to a large extent the three dimensional structure as well as function of such a molecule. This makes the widely successful techniques such as global and local alignments, which take into account just the primary sequence similarity, of little worth in ncRNA sequence analysis. One option is to model an ncRNA sequence as a realization of some stochastic source. In this case much explored generative models like hidden Markov models (HMMs) become inapplicable, because base-pairing gives rise to a zigzag looking mutual information profile - difficult to model via HMM. Fortunately the stochastic context-free grammars (SCFG), which are one step higher in the Chomsky hierarchy of grammars than the HMMs, can model ncRNA sequences very well. As an extension of these SCFGs, [2] proposed the profile-SCFGs. These profile-SCFGs (also known as covariance models (CMs)) are successfully used in the RFAM database to score a query sequence across each of the RFAM family. There are two issues with existing usage of SCFG within RFAM: (i) This is a purely generative model and only positive examples belonging to a certain RFAM family are used to train the profile-SCFG of that particular family; (ii) A query sequence has to be scored by all the CMs in the database to decide the homology/closest family. The first issue poses a question related to the accuracy of classification/homology detection, especially when a family has only a small number of training sequences. And the second question is related to the computational complexity: as the scoring of any sequence using SCFG is an  $O(L^3)$  ( $L$  is the length of the query), the search procedure is time consuming for a database with many families.

To address both issues in a single go we propose a learning framework, which combines the generative power of SCFG and the discriminative power of support vector machines (SVM) type classifiers that can take into account both positive and negative training examples. We use the SCFG as a generative model common to all the sequences belonging to various functional RNA families. Our hypothesis is: for a properly trained SCFG all RNA sequences belonging to the same functional family take similar posterior distribution of state paths. Based on this we develop a similarity measure which takes into account the posterior probability of two consecutive states across all possible SCFG state paths and in some sense form a vector of sufficient statistics. This similarity measure is incorporated in the multi-class SVM to provide the discriminative power across various functional RNA families. At some level our approach is motivated by the development presented in [3]. To check the efficacy of the proposed approach we first filter the *seed sequences* of RFAM database (trusted ncRNA sequences collected and grouped by human efforts) for  $\geq 70\%$  sequence identity within the family. This simulates the moderate to weak homology detection scenario. The results of K-fold and leave-one-out cross-validations on the filtered database consistently show the classification accuracy of  $\geq 90\%$  for most of the families. This motivates us to draw the attention of researchers interested in RNA sequence analysis towards approaches that combine both generative and discriminative models. A tool-suite based on the proposed approach will be made available at <http://www.egr.msu.edu/~desaikey>.

[1] <http://www.sanger.ac.uk/Software/Rfam/>.

[2] S. R. Eddy, R. Durbin, Nucleic Acids Research, 22:2079-2088, 1994.

[3] T. Kin, K. Tsuda and K. Asai, "Marginalized Kernels for RNA Sequence Data Analysis", Proc. Genome Info., 2002.