A Learning Framework for Detecting Remote Non-Coding RNA Homologues

Keyur Desai[1], J. R. Deller Jr.[2], and Hayder Radha[3]

[1]desaikey@egr.msu.edu, [2]deller@egr.msu.edu, [3]radha@egr.msu.edu

Department of Electrical & Computer Engineering, Michigan State University

1. Introduction

Motivation

➢Non-coding RNA (ncRNA) genes produce a functional RNA instead of a translated protein.

➢State-of-the-art ncRNA sequence analysis tools like profile stochastic context-free grammar (profile-SCFG) are based purely on generative model approach.

➢Can we combine SCFG with discriminative methods like support vector machine (SVM) in order to increase overall accuracy and/or speed of classification?

Short-comings of profile-SCFG

1. Families with a small number of training sequences can have low accuracy.

2. To detect homology, a query sequence must be scored by all profile-SCFG in the database: for databases with many families, search procedure becomes time consuming because scoring is an O(L^3) (L is the length of the query) process.

Proposed framework

➢We use just one SCFG for the whole database.

➢For a well-trained SCFG, all RNA sequences belonging to the same functional family should exhibit a similar posterior distribution of state paths: based on this hypothesis we develop a feature extractor which takes into account the posterior probability of two consecutive states across all possible SCFG state paths.

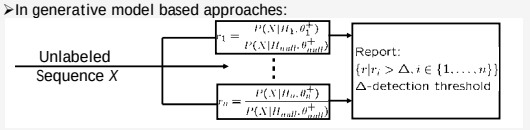
➢Extracted feature vectors are combined with a multi-class SVM to derive decision boundaries.

➢ For each query sequence, feature extraction, which is an O(L^3) process, is performed just once.

3. Framework of Discriminative Learning for ncRNA

1. Evolutionary-patterns and likelihood

➢In generative model based approaches:



➢Shortcomings of this approach:

1. Only positive training examples are used.
2. Improved model does not guarantee better accuracy unless the null model is improved.
3. For complex graphical models like SCFG, the quality of trained parameters may be poor: Expectation-Maximization (EM) is guaranteed only to converge to a local maxima.

2. Discriminative methods

➢Decision boundary is estimated using both positive and negative training examples.

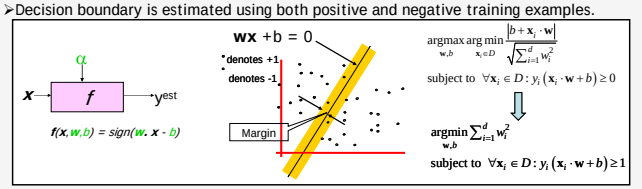


Figure 2: SVM: basic objective, geometric interpretation and mathematical optimization

➢Kernel trick (dot product in high-dimension space) adds non-linearity.

➢Symbolic objects like biological sequences need to be mapped to a numeric feature space using some mapping $f: X \to \mathbb{R}^D$

3. Posteriori State Distribution (PSD) Kernel

➢Let the posteriori probability of a SCFG passing through the non-terminal V at parse index (i,j) be denoted by : $P(W(i,j)=V|x,\Theta)$.

➢Then the joint a posteriori probability of SCFG passing through non-terminals V_1 at (i,j) and V_2 at $(i+\delta, j-\delta')$, where $(\delta, \delta') \in \{0, 1\}$ is:

$$\xi_{V_1,V_2}(i,j)|_{(\delta,\delta')} = P(W(i,j)=V_1, W(i+\delta, j-\delta')=V_2|x,\Theta)$$

➢This can be efficiently computed using the inside-outside probabilities as:

$$\xi_{V_1,V_2}(i,j)|_{(\delta,\delta')} = \frac{1}{P(x|\Theta)} O_{V_1}(i,j) t_{V_1}(V_2) I_{V_2}(i+\delta, j-\delta')$$

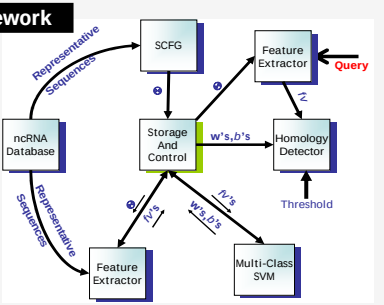
$$O_V(i,j) = P(S \to x[1,i-1] V x[j+1,n]|\Theta);$$

$$I_V(i,j) = P(V \to x[i,j]|\Theta); t_{V_1}(V_2) = P(V_2|V_1)$$

➢The extracted feature vector is:

$$f_D = \{\xi_{V_1,V_2}(i,j)|_{(\delta,\delta')}; i = 1:L, j = i+1:L\}(\delta,\delta') \in \{(0,1),(1,0),(1,1)\}$$

4. The framework



2. ncRNA and SCFG

Statistical characteristics of ncRNA

➢A transcribed ncRNA folds to form secondary structure (SS).

➢Its subsequences that participate in base pairing are palindromes (strings that read the same backwards) with common base pairs.

➢A good generative model for ncRNAs should contain only palindrome sequences in its "maximal probability sequence set" .

➢SCFG in the Chomsky hierarchy of transformation grammars are best suited for this task.

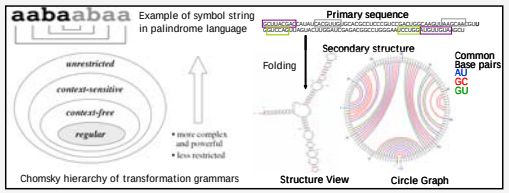


Figure 1: Folding of ncRNAs and its implications

Stochastic Context-Free Grammar

➢ CFG is a set G= {V(non-terminals), T(terminals), P(production rules)}.

➢ In SCFG, we associate probability with each CFG production rule.

➢ Three SCFG problems:

1. Find optimal state path: $\hat{\pi} = \arg\max_{\pi} P(x,\pi|G,\Theta)$
2. Find P(sequence|G): $P(x|G,\Theta) = \sum_{\pi} P(x,\pi|G,\Theta)$
3. Find optimal parameters given example sequences (Training):

$$P^{+}_{u\to\beta} = \frac{\sum_{j=1}^{K}\sum_{\pi} P(\pi|x_j,\Theta) n(\beta,u,\pi,x_j)}{\sum_{j=1}^{K}\sum_{\pi} P(\pi|x_j,\Theta) n(u,\pi,x_j)}$$

4. Results for RFAM Database

Datasets

➢Seed sequence families of RNA family (RFAM) database were filtered for <= 70% identity.

➢3 RFAM families with avg. sequence length <80 and no. of sequences > 15 were extracted –RF05 57 sequences, RF29 15 sequences, and RF31 19 sequences.

➢These 3 families were divided in equal size test and train sets– 44 sequences test set and 47 sequences train set.

Results for profile-SCFG

➢For each family of training set, a multiple sequence alignment was constructed using its family profile-SCFG (obtained from RFAM covariance model library).

➢This alignment was used to train a new profile-SCFG for the same family.

➢Newly trained profiles were used to detect corresponding family homologs in the test set.

	RF0005	RF00029	RF00031
Number of true homologs	28	7	9
Number of identified homologs	28(100%)	5(%71)	1(11%)

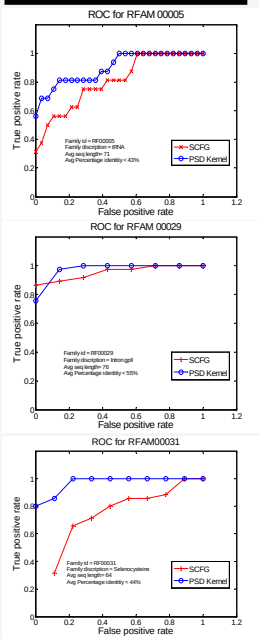
Results for SCFG and PSD Kernel

➢For each family of training set, an SCFG (BPLR topology as in reference 1) was trained using corresponding training sequences.

➢For PSD kernel method, a common SCFG (BPLR topology) for the whole training set was trained.

➢Trained SCFG was used to extract PSD feature vectors from the training sequences; these feature vectors were used to train a multi-class SVM (one-against-all).

Receiver operating curves



5. Discussion and Summary

Discussions

➢In the presented framework of discriminative learning, the role of SCFG is to provide features based on hidden process of sequence generation.

➢In contrast to the Fisher kernel[2] and Marginalized kernel[1], which average the posteriori distribution of hidden states, the PSD kernel captures the position information.

➢Although huge in size, the feature vectors generated by the PSD kernel are sparse in nature, and hence easily manageable by state-of-the-art SVM implementations.

➢The proposed framework does not require multiple sequence alignment.

➢For each query sequence, the SCFG is used only once-to extract the feature vector; hence the search time is appreciably less than that of likelihood based approaches.

Summary

➢We have developed a discriminative framework for ncRNA homology detection. The framework combines a newly proposed Posteriori State Distribution Kernel with a multi-class support vector machine. The method improves the speed and accuracy of ncRNA homology search procedure.

References

1. T. Kin, K. Tsuda, and K. Asai, "**Marginalized Kernels for RNA Sequence Data Analysis**",In *Proceedings of Genome Information*, 2002.
2. T. Jaakkola, M. Diekhans, and D. Haussler, "**Using the fisher kernel method to detect remote protein homologies**",In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, 1999.

Acknowledgement: This research is supported by Michigan State University incubator grant. Visit www.egr.msu.edu/~desaikey for the availability of the tool-suite and further details