Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM)

Sumeet Agarwal, Candida Vaz, Alok Bhattacharya, Ashwin Srinivasan

What are microRNA?

- A type of non-coding RNA, thought to play a part in the regulation of gene expression.
- Mature form is single-stranded, typically 21-23 nucleotides long.
- Has a longer DNA precursor (70-120 nucleotides long), which is transcribed to RNA and folds up to form a stem-loop structure, which then undergoes enzymatic cleavage to form the mature sequence.

The Problem

- To identify the miRNA-precursor (pre-miRNA) sequences, i.e. those portions of the genome which code for miRNA.
- Depends not only on primary but also on secondary structure, i.e. the hairpin-like stem-loop structure formed on base-pairing.
- Requires some kind of grammar/model to represent the secondary structure and model the inherent long-range dependencies.

Example

hsa-let-7a-1

Primary sequence:

ugggaugagguaguagguuguauaguuuuagggucacacccaccacugggagauaacuauaca
aucuacugucuuuccua

Stem-loop secondary structure:

```
     u   gu                uuagggucacac
uggga gag  aguagguuguauaguu            c
||||| |||  ||||||||||||||||            c
auccu uuc  ucaucuaacauaucaa            a
     -   ug                uagagggucacc
```

Mature sequence: ugagguaguagguuguauaguu

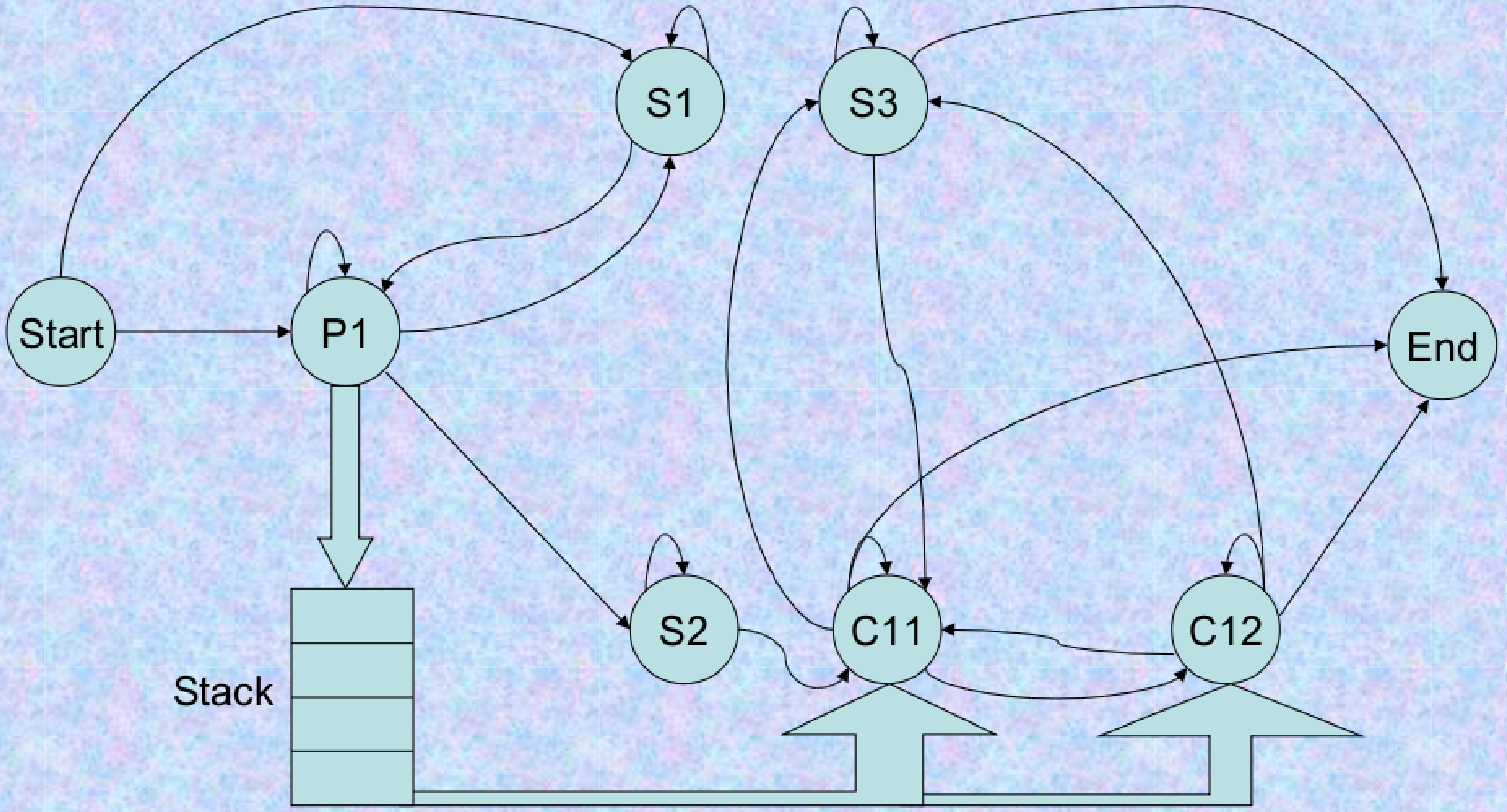
Methodology

- Approaches like HMMs and SCFGs have previously been used with some success for biological sequence identification.
- Due to the apparent context-sensitive nature of the grammar required to generate the secondary structure, these don’t work very well for pre-miRNA identification.
- Context-Sensitive HMMs have recently been proposed as a possible solution.

Context-Sensitive HMMs

- An extension of HMMs, they have certain ‘paired states’, linked by a memory store such as a stack or a queue.
- One state writes its emissions to the store, while the other reads from it; the subsequent emission and transition depends on what has been read.
- Fairly efficient dynamic programming algorithms exist for alignment and scoring, based on the corresponding HMM algorithms (Viterbi, Forward-Backward).

CSHMM Structure



Trace

State	Emits	Stack	Structure
P1	u	u	u
P1	g	ug	ug
S1	a	ug	a ug
P1	c	ugc	a ug c
S2	c	ugc	a ug c c
C11	g	ug	a ug c |c g

Trace (contd.)

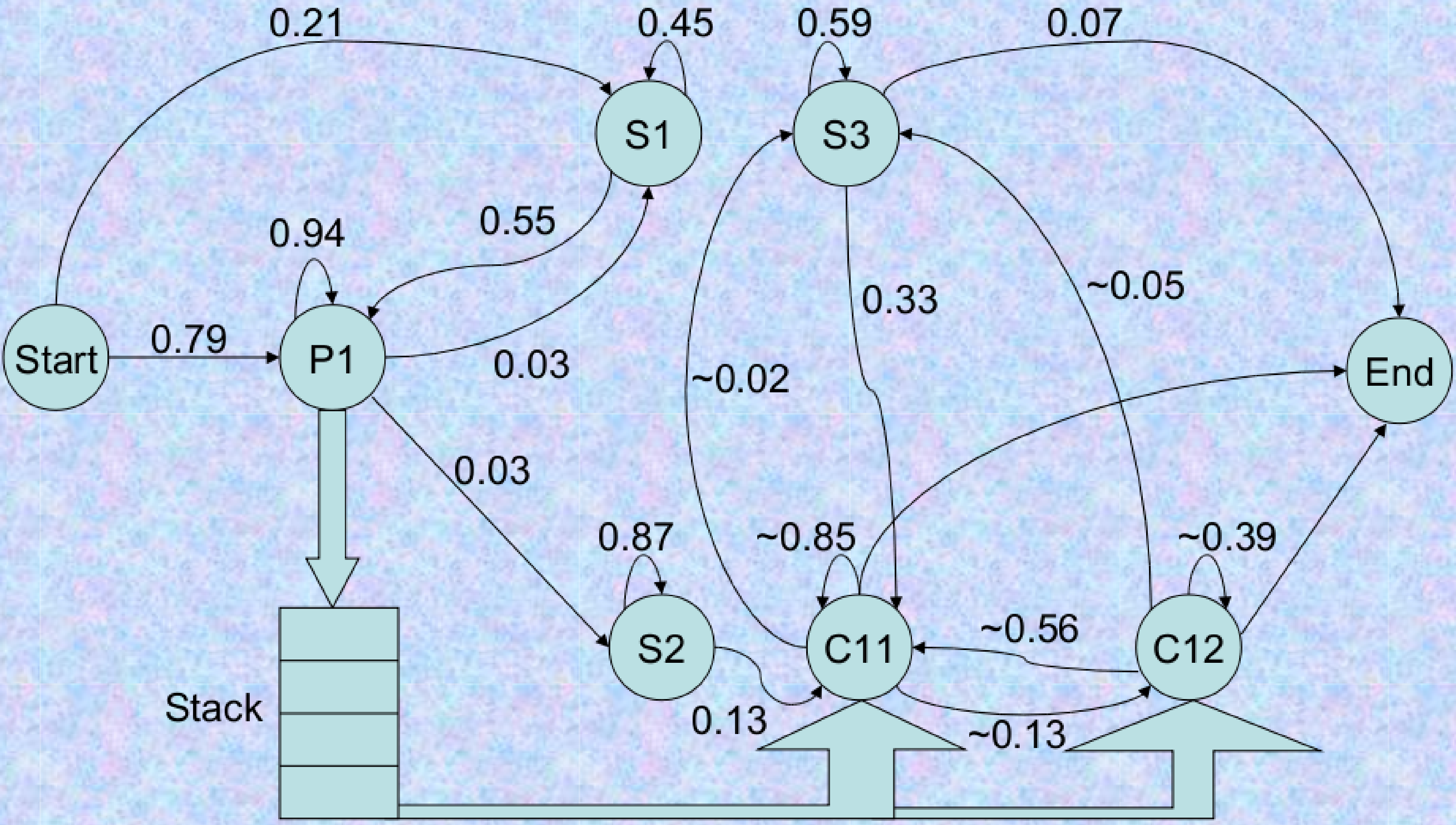
State	Emits	Stack	Structure
C11	c	u	a ug c | |c c g -
C12	c		u a g c | |c c g c -

Learning the Parameters

- Known miRNA secondary-structures were used to learn the emission and transition probabilities of the various states in the model.
- The sequences were read in their stem-loop form, and the number of emission and transition events of different kinds from all the states were counted in order to estimate the probabilities.
- There is also an algorithm to learn from the primary sequences only (analogous to Baum-Welch), but it takes much longer and gives slightly poorer results.

The CSHMM with some estimated parameters

(~ denotes values averaged over the 4 possible stack symbols)



Predicting Secondary Structure

- Once the CSHMM parameters had been estimated, the alignment algorithm could be used to find the most likely state sequence for a given input.
- This gave the most likely secondary structure for a given linear RNA sequence, assuming it to be a miRNA-precursor.
- Secondary structure features could then be extracted from the sequence, such as number of nucleotides in bulges, size of loop etc.

Classification

- Based on the secondary structure features, it is possible to build a classifier to identify miRNA-precursors. Alternatively, this can be done using just the CSHMM score.
- We use a simple decision tree (J48) with just a single feature, the likelihood score (i.e., a threshold is set on the score) to discriminate between precursors and non-precursors. We tried including other features but did not observe any significant improvement in performance.

Training the Classifier

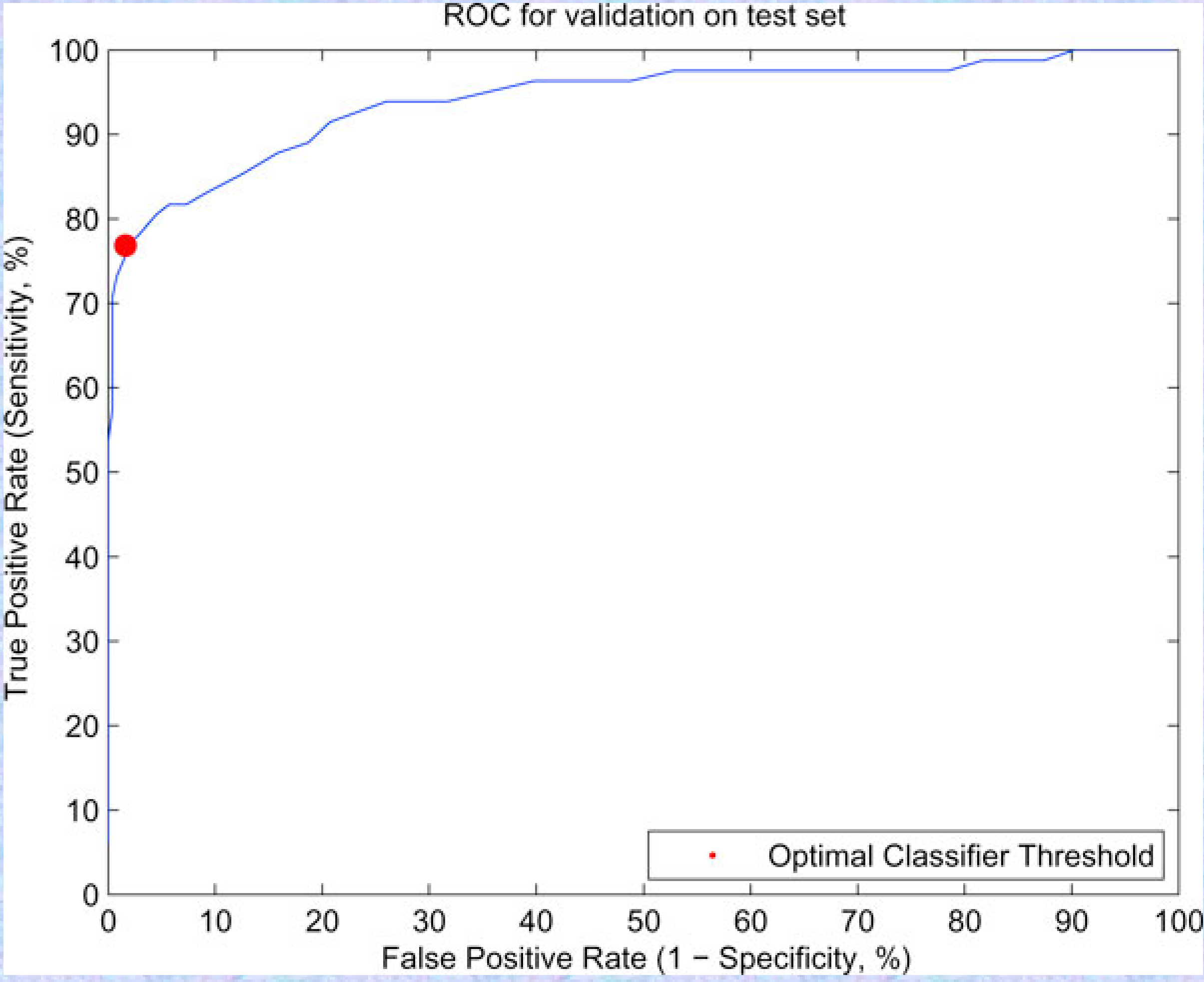
- Experimentally known human miRNA-precursors were used as positive examples. For negatives, pseudo-hairpins from human genes were used. First the CSHMM was trained on the positives only; subsequently likelihood scores were obtained for both positives and negatives based on the CSHMM alignment, and these were used to train the classifier (i.e., set a threshold).
- Five-fold cross-validation, as well as a held-out test set, were used to determine the predictive accuracy of the constructed classifier.

Cross-validation Results

		Actual		
		miRNA	non-miRNA	
Predicted	miRNA	170 (60.67)	12 (121.33)	**182**
	non-miRNA	30 (139.33)	388 (278.67)	**418**
		200	**400**	**600**

5-fold cross-validation performance of the CSHMM using a human miRNA dataset. The number in parentheses below each entry is the expected value of the entry under the hypothesis that the actual class is independent of the predicted one. Estimates of predictive accuracy, sensitivity and specificity from this table are 0.93 (93%), 0.85 (85%) and 0.97 (97%) respectively.

Hold-out Validation Results



ROC curve for validation on held-out test set (82 positives, 246 negatives)

Identifying novel pre-miRNA

- We used our classifier to scan human chromosome 19: 70 out of 80 known miRNAs were identified. 18,188 additional hairpins with high scores were subjected to post-prediction filters based on ESTs and Drosha sites, resulting in 49 final predictions (most of these 49 were also predicted by two other classifiers).
- Mature miRNA derived from one of the predicted sequences was experimentally detected, verifying the prediction.
- We also carried out analysis of small RNA sequences obtained via deep sequencing of human peripheral blood mononuclear cells. 308 novel miRNAs were predicted; a similar number were obtained with an independent classifier.

Conclusions

- A very simple classifier is constructed that shows a sensitivity of about 85% along with a specificity of about 97-98% on human miRNA sequences.
- Good discrimination performance is obtained with the use of only a likelihood score threshold, i.e. the CSHMM is able to capture structural information to a fair extent.
- When used to try and identify novel pre-miRNA in a genome, the technique is useful in narrowing down the number of possible precursors to be examined experimentally.

Acknowledgments

- **IBM IRL**

  Sachindra Joshi, Ganesh Ramakrishnan
- **JNU**

  Sonika Tyagi
- **IIT Delhi**

  Prof. Sachin Maheshwari
- **Department of Biotechnology, Govt. of India**