Using Base Pairing Probabilities for MiRNA Recognition

Yet Another SVM for MiRNA Recognition: yasMiR

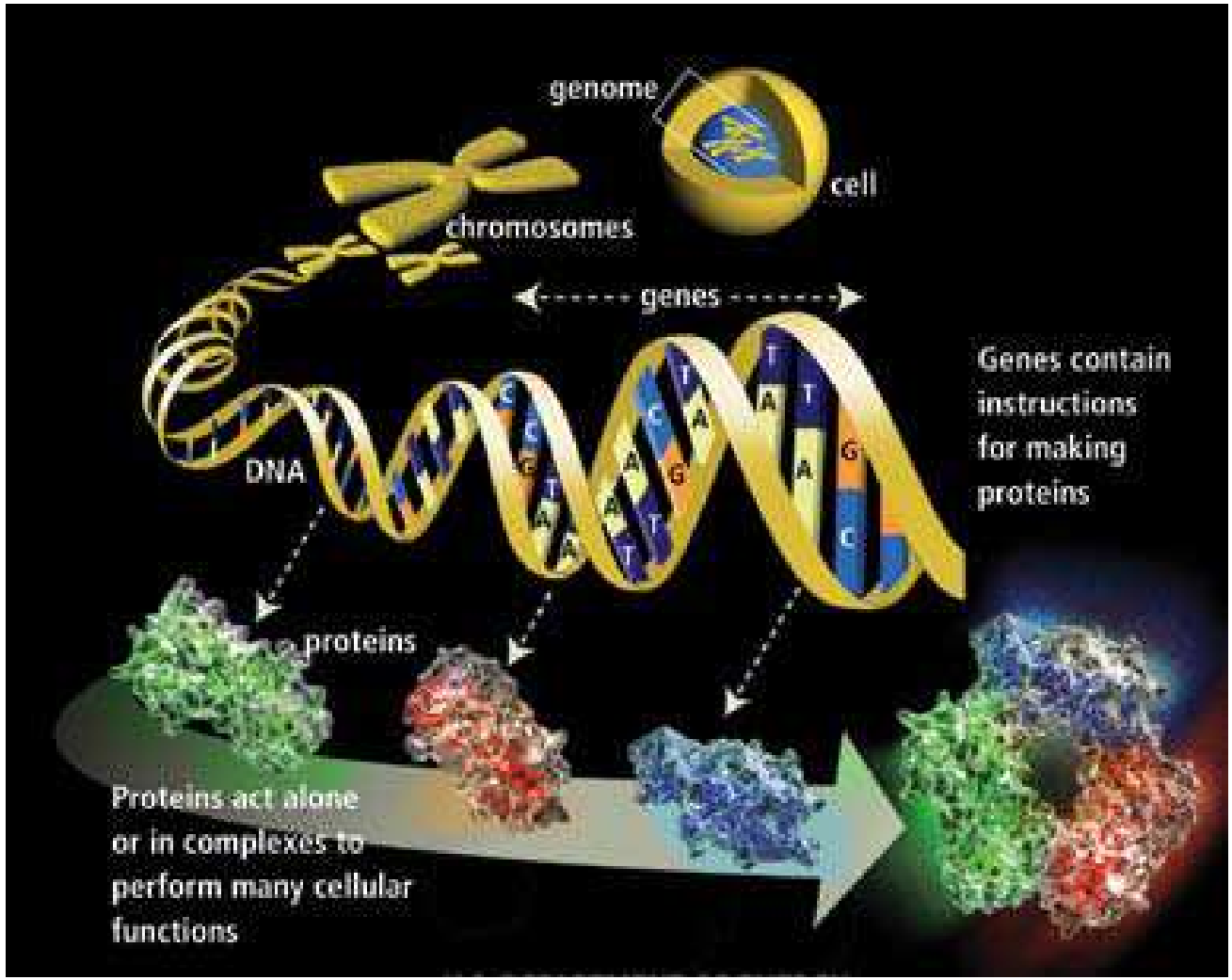
Daniel Pasailă, Irina Mohorianu, Liviu Ciortuz
Department of Computer Science
"Al. I. Cuza" University, Iaşi, Romania

PLAN

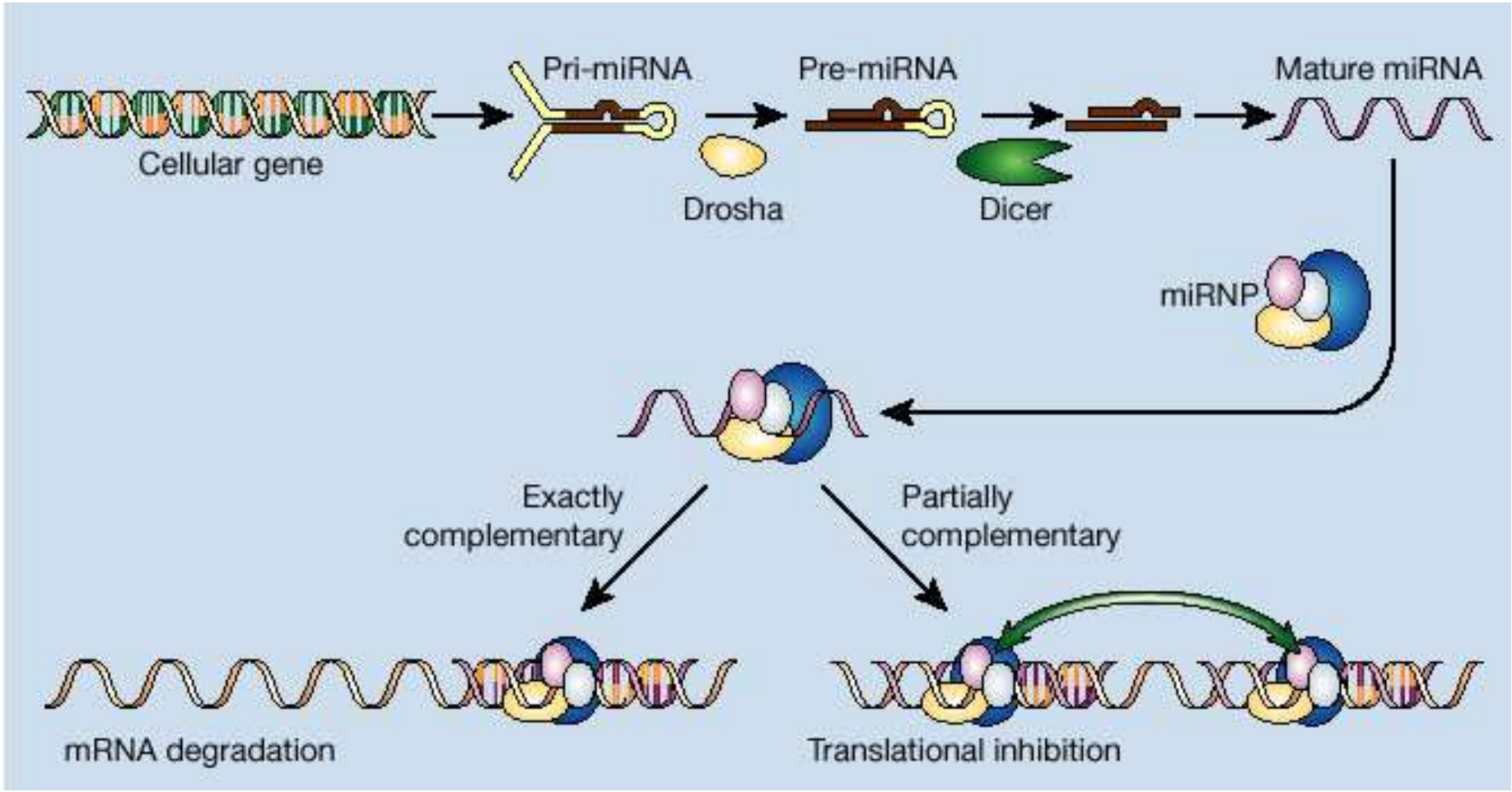
- microRNAs and SVMs
- our approach: using base-pairing probabilities and pivots
- yasMiR features
- tests and comparisons with other systems and classifiers
- conclusions

The Central Dogma of Molecular Biology

From "Genomics and its impact on science and society: The Human Genome Project and beyond", US Department of Energy, Genome Research Programs

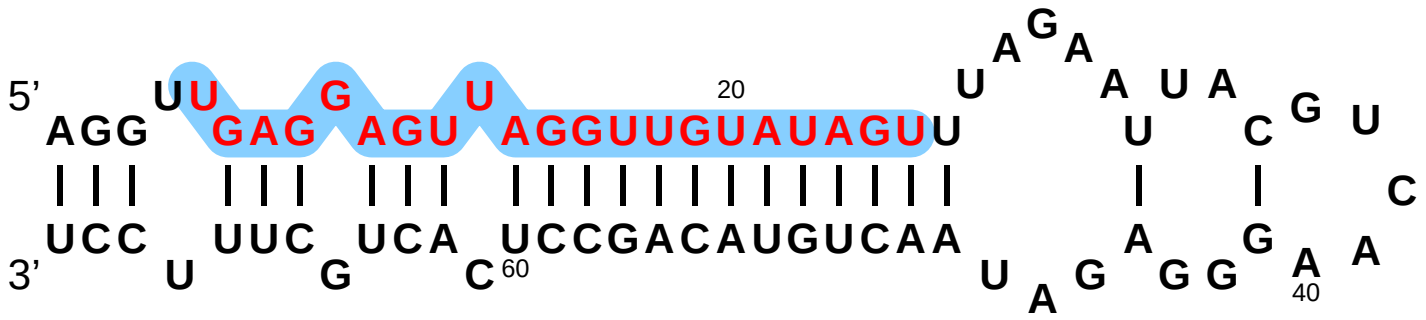


miRNA in the RNA interference process



From D. Novina and P. Sharp, *The RNAi Revolution*, Nature 430:161-164, 2004.

A pre-miRNA example: *hsa-let-7a-2*




```
AGGUUGAGGUAGUAGGUUGUAUAGUUUAGAAUUACAUCAAGGGAGAUAACUGUACAGCCUCCUAGCUUUCCU
(((..(((.(((.((((((((((((.....(..(.....)..)...)))))))))))))).))).))).)))
ppp..ppp.ppp.pppppppppppppp.....p..p.....p..p...pppppppppppppp.ppp.ppp.ppp
```

SVMs for microRNA Identification

Sewer et al. (Switzerland)	2005	miR-*abela*
Xue et al. (China)	2005	Triplet-SVM
Jiang et al. (S. Korea)	2007	MiPred
Zheng et al. (Singapore)	2006	miREncoding
Szafranski et al. (SUA)	2006	DIANA-microH
Helvik et al. (Norway)	2006	Microprocessor SVM & miRNA SVM
Hertel et al. (Germany)	2006	RNAmicro
Sakakibara et al. (Japan)	2007	stem kernel
Ng et al. (Singapore)	2007	miPred

Base-pairing probabilities

Definition: $p_{ij} = \sum_{S_\alpha \in \mathcal{S}} P(S_\alpha)\, \delta_{ij}^{\alpha}$, where

$\mathcal{S}$ is the set of all possible secondary structures for the given RNA sequence, and

$$\delta_{ij}^{\alpha} = \begin{cases} 1 & \text{if the nucleotides } i \text{ and } j \text{ form a base-pair in the structure } S_\alpha \\ 0 & \text{otherwise.} \end{cases}$$

Note: $P(S_\alpha)$, the probability of the structure $S_\alpha \in \mathcal{S}$ follows a Boltzmann distribution:

$$P(S_\alpha) = \frac{e^{-MFE_\alpha/(R \cdot T)}}{Z}$$

with

$Z = \sum_{S_\alpha \in \mathcal{S}} e^{-MFE_\alpha/(R \cdot T)}$,

$R = 8.31451\ \text{J mol}^{-1}K^{-1}$ (a molar gas constant), and

$T = 310.15\text{K}$ (37° C).

Note: The probabilities p_{ij} are efficiently computed using McCaskill's algorithm (1990).

The non-null components of the arrays $PF[i,0]$ and $PF[i,1]$ computed for *hsa-let-7a-2*, using base-pairing probabilities.

1	2	3	6	7	8	9	10	11	12	14	15
.54	.98	1	.96	.99	1	.01	1	1	.99	.99	1
16	17	18	19	20	21	22	23	24	25	26	27
1	1	1	1	1	1	1	1	1	.92	.87	.17
28	29	30	31	32	33	34	35	36	37	38	
.22	.10	.01	.06	.56	.32	.01	.50	.22	.32	.31	

33	34	35	37	38	39	40	41	42	43	44	45	46
.01	.01	.08	.01	.01	.01	.04	.46	.14	.26	.47	.31	.33
47	48	49	50	51	52	53	54	55	56	57	58	59
.51	.94	.99	1	1	1	1	1	1	1	1	1	1
60	62	63	64	65	66	67	68	69	70	71	72	
.99	.99	1	.99	.01	1	1	.96	.01	.92	1	.60	

A similarity measure for two RNAs based on their pattern ("profile") of base-pairing (Meireles, 2006)

For every nucleotide i compute the probability of i forming a base pairing upstream, downstream, or not forming a base pairing at all:

$$\boldsymbol{PF}[i,0] = \sum_{j>i} p_{ij} \qquad \boldsymbol{PF}[i,1] = \sum_{j<i} p_{ij} \qquad \boldsymbol{PF}[i,2] = 1 - \boldsymbol{PF}[i,0] - \boldsymbol{PF}[i,1]$$

The similarity measure is the global alignment score of two profiles, calculated using the Needleman-Wunsch algorithm.

We use zero gap penalties, and as match score the inner product of the two profile vectors associated to the corresponding positions in the input sequences:

$$S[i,j] = \boldsymbol{max} \begin{cases} S[i-1,j] \\ S[i,j-1] \\ S[i-1,j-1] + \sum_{k=0}^{2} \boldsymbol{PF}[i,k] \cdot \boldsymbol{PF}[j,k] \end{cases}$$

yasMiR profile-based features

We will construct a set of RNA sequences that we call **pivots.**

Then, the **profile alignment scores** of a given (training or testing) pre-miRNA with all the pivot sequences will be included in the pre-miRNA's feature vector.

We **conjecture** that the way in which the pre-miRNA base-pairing profiles align to the profiles of pivot sequences can be successfully used as a **discriminative factor** in classifying real vs. pseudo pre-miRNAs.

Remarks on pivots

In the developing phase of our system, we used pseudo-miRNAs and pre-miRNAs as pivots, but we saw that the prediction accuracy didn't significantly change when we used randomly generated RNA sequences.

Also, we noticed that about 50−200 pivots were needed to achieve best performance.

The length of the used pivot sequences seemed to affect the result. In practice we noticed that sequences of 45-65 nucleotides were most appropriate.

Triplet probabilistic patterns

For any 3-mer there are $8 = 2^3$ possible structure patterns: 'ppp', 'pp.', 'p.', 'p..', '.pp', '.p.', '..p', and '...'.

Further on, if we consider the middle nucleotide (A, C, G or U) in a 3-mer, there will be $32 = 8 \times 4$ possible combinations.

Given a pre-miRNA, we will compute the probability of every such combination occurring inside the sequence.

Example: The probability for the pattern 'p.p' to occur for a certain position i inside the given RNA sequence, is:

$$(1 - \boldsymbol{PNP}[i-1]) \cdot \boldsymbol{PNP}[i] \cdot (1 - \boldsymbol{PNP}[i+1])$$

where $\boldsymbol{PNP}[i]$ is the probability of base i being unpaired: $\boldsymbol{PNP}[i] = \boldsymbol{PF}[2]$.

yasMiR non-profile-based features (I)

- **32 features, each one representing the probability that nucleotide a appears in the middle position of occurrences of pattern j:**

$$Pn[a,j] = \frac{\sum_{S[i]=a} Pt[i,j]}{cnt(a)/L}$$

where $S[1..L]$ is the current sequence, $\boldsymbol{Pt}[i,j]$ stores the probability that the 3-mer centered of the i-th nucleotide has the pattern j, and $cnt(a)$ denotes the number of nucleotides of type a in the sequence.

- **12 features, one for each pair of distinct nucleotides (a,b):**
the sum of the base-pair probabilities for all the corresponding positions in the sequence:

$$\sum_{S[i]=a,S[j]=b} p_{ij}$$

yasMiR non-profile-based features (II)

- **the overall non base-pairing probability:**

$$\sum_{i=1}^{L} PNP[i]/L$$

- **4 features: the non base-pairing probability for every nucleotide** $a \in \{A, C, G, U\}$**:**

$$\sum_{S[i]=a} PNP[i]/cnt(a)$$

- **the mean base pair distance in the equilibrium state of the given RNA (a measure of the structural diversity), computed by the *mean_bp_dist* function in the Vienna RNA package, also using base pairing probabilities.**

yasMiR non-profile-based features (III) not using base pairing probabilities

- **the folding *minimum free energy*, obtained using the *fold* function in the Vienna RNA package**
- **4 features: the *average frequency* for each nucleotide $a \in \{A, C, G, U\}$ in the current sequence, calculated as $cnt(a)/L$**
- **16 features: the *average dinucleotide frequency* (one for each dimer ab).**

Comparison of yasMiR with Triplet-SVM

Test	**yasMiR accuracy(%)**	**Triplet-SVM accuracy(%)**
TE-C: Human pre-miRNAs	**96.6** (29/30)	93.3
TE-C: Pseudo pre-miRNAs	**96.5** (965/1000)	88.1
UPDATED	**92.3** (36/39)	**92.3**
CROSS-SPECIES	**95.4** (554/581)	90.9
CONSERVED-HAIRPIN	**93.5** (2287/2444)	89.0

The results for Triplet-SVM are taken from [Xue et al., 2005].
In paranthesis: the ratio of correctly classified instances.

Detailed comparison of yasMiR with Triplet-SVM: accuracy on the CROSS-SPECIES dataset

Test	**yasMiR accuracy(%)**	**Triplet-SVM accuracy(%)**
Mus musculusi	**97.2** (35/36)	94.4
Rattus norvegicus	**84.0** (21/25)	80.0
Callus Gallus	**100.0** (13/13)	84.6
Dnio Rerio	**83.3** (5/6)	66.7
Caenorhabditis briggsae	**100.0** (73/73)	95.9
Caenorhabditis elegans	**92.7** (102/110)	86.4
Drosophila pseudoobscura	**94.3** (67/71)	90.1
Drosophila melanogaster	**95.7** (68/71)	91.5
Oryza sativa	**96.8** (93/96)	94.8
Arabidopsis thaliana	**97.3** (73/75)	92.0
Epstein Barr Virus	80.0 (4/5)	**100.0**
Total	**95.35** (554/581)	90.9

Comparison of yasMiR with miPred and Triplet-SVM

Test	yasMiR accuracy(%) se.(%) sp.(%)	miPred accuracy(%) se.(%) sp.(%)	Triplet-SVM accuracy(%) se.(%) sp.(%)
TE-H	**93.77** **87.80** 96.74	93.50 84.55 **97.97**	87.96 73.15 93.57
IE-NH	94.11 90.35 95.99	**95.64** **92.08** **97.42**	86.15 86.15 96.27
IE-NC	**82.75**	68.68	78.37
IE-M	**100**	87.09	0

The results for miPred and Triplet-SVM are taken from [Ng and Mishra, 2007].

Note: Only accuracy is given for IE-NC and IE-M since these datasets are made only of non miRNAs; in such a case, specificity is equal to accuracy, and sensitivity is null.

Comparing the predictive accuracy (%) of RF and SVM using yasMiR features

- on test datasets from Triplet-SVM

Test	RF		SVM
	without feat. selection	with feat. selection	with feat. selection
TE-C	61.1	93.2	94.4
UPDATED	94.9	89.7	**97.4**
CROSS-SPECIES	**96.1**	89.5	89.8
CONSERVED-HAIRPIN	92.6	89.6	91.0

- on test datasets from miPred

Test	RF		SVM
	without feature sel.	with feature sel.	with feature sel.
TE-H	92.14	92.14	91.86
IE-NH	93.82	92.72	91.87
IE-NC	63.46	63.30	88.31
IE-M	74.19	16.12	100

Prediction results of yasMiR on miPred's test datasets

using 200 pivots
selected via clustering from a pool of 2000 randomly generated pivots

Test	SVM accuracy(%) sens.(%) spec.(%)	RF accuracy(%) sens.(%) spec.(%)
TE-H	92.55 83.74 97.34	91.69 83.74 96.01
IE-NH	93.37 86.36 96.88	93.67 89.66 95.68
IE-NC	**91.11**	63.77
IE-M	**100**	19.35

using 88 pivots
selected via PCA and *varSelRF* from the 200 pivots obtained by clusterisation

Test	SVM accuracy(%) sens.(%) spec.(%)	RF accuracy(%) sens.(%) spec.(%)
TE-H	92.68 83.74 97.15	91.06 82.11 95.53
IE-NH	93.57 89.0 95.86	94.07 92.23 94.99
IE-NC	**93.11**	63.11
IE-M	**100**	19.35

Replacing the probabilistic triplet features in yasMiR with their non-probabilistic counterpart: The effect on Triplet-SVM datasets, using 100 pivots

Test	**yasMiR accuracy(%)**	**yasMiR′ accuracy(%)**
TE-C: Human pre-miRNAs	**100** (30/30)	96.67 (29/30)
TE-C: Pseudo pre-miRNAs	**96.20** (962/1000)	95.90 (952/1000)
UPDATED	**94.87** (37/39)	**94.87** (37/39)
CROSS-SPECIES	95.18 (553/581)	**95.87** (557/581)
CONSERVED-HAIRPIN	**94.23** (2303/2444)	93.09 (2275/2444)

In paranthesis: the ratio of correctly classified instances.

Conclusions

- We showed that the base-pairing probabilities combined with some other, simple statistical measures lead a SVM to achieve high pre-miRNA prediction accuracy rates, comparable to the best published miRNA classification results up to our knowledge.
- The RF classifier is a not good enough candidate to replace SVM for miRNA identification using our set of features.
- One of the advantages of our approach is that it makes no use of so-called normalised features which are based on sequence shuffling (as for instance miPred does), which is a sensitive issue from the biological point of view, and also makes our approach much less time consuming.