Support Vector Machines for microRNA Identification

Liviu Ciortuz, CS Department, University of Iasi, Romania

Plan

0. Related work
1. RNA Interference; microRNAs
2. RNA Features
3. Support Vector Machines;
   other Machine Learning issues
4. SVMs for MicroRNA identification
5. Research directions / Future work

0. Related work: Non-SVM systems for miRNA identification

using sequence alignment systems (e.g. BLASTN):

- miRScan [Lim et al, 2003] worked on the *C. elegans* and *H. sapiens* genomes
- miRseeker [Lai et al, 2003] on *D. melanogaster*
- miRfinder [Bonnet et al, 2004] on *A. thaliana* and *O. sativa*

adding secondary structure alignment:

- [Legendre et al, 2005] used ERPIN, a secondary structure alignment tool (along with WU-BLAST), to work on miRNA registry 2.2
- miRAlign [Wang et al, 2005] worked on animal pre-miRNAs from miRNA registry 5.0 except *C. elegans* and *C. briggsae*, using RNAfoster for secondary structure alignment.

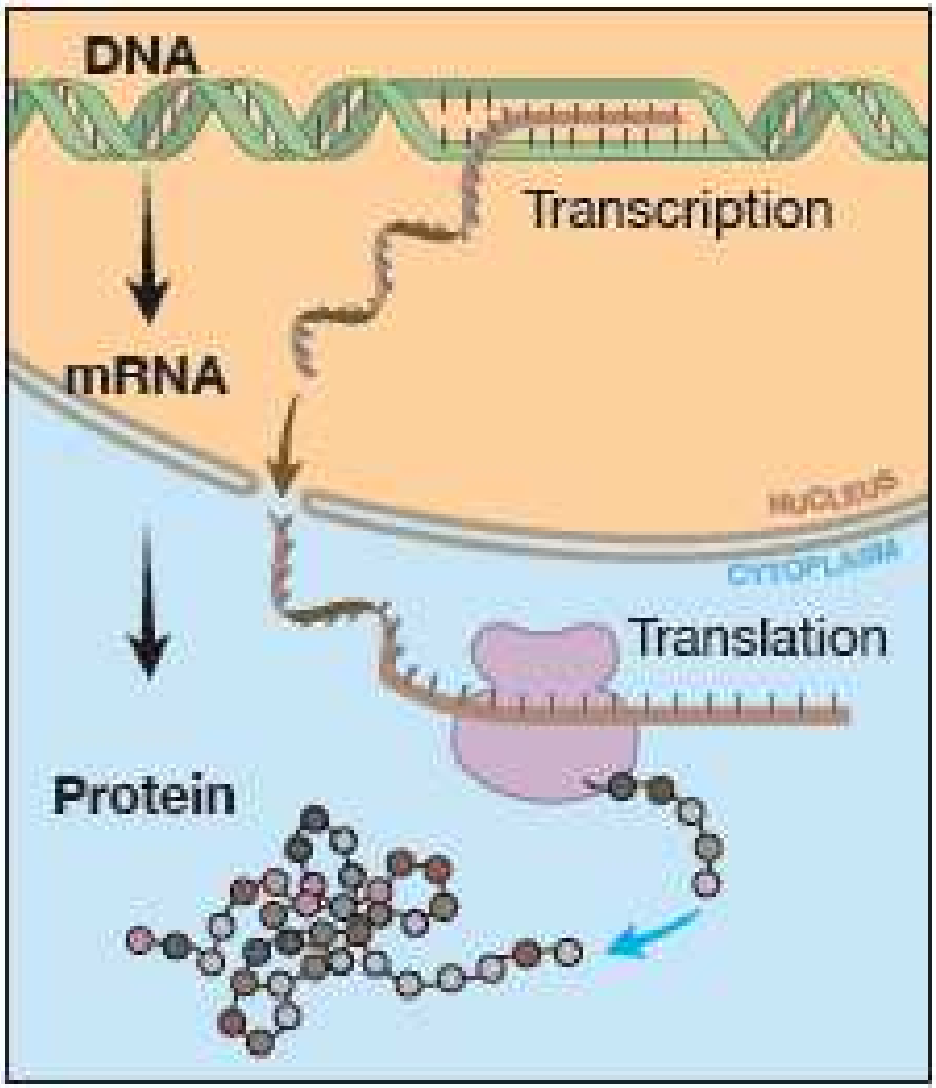
Non-SVM systems for miRNA identification (cont'd)

non-SVM machine learning systems for miRNA identification:

- **proMIR** [Nam et al, 2005] uses a Hidden Markov Model,
- **BayesMIRfinder** [Yousef et al, 2006] is based on the naive Bayes classifier
- [Shu et al, 2008] uses clustering (the k-NN algorithm) to learn how to distinguish
  - between different categories of non-coding RNAs,
  - between real miRNAs and pseudo-miRNAs obtained through shuffling.
- **MiRank** [Xu et al, 2008], uses a ranking algorithm based on Markov *random walks*, a stochastic process defined on weighted finite state graphs.

1. RNA Interference

Remember the Central Dogma of molecular biology:
DNA → RNA → proteins

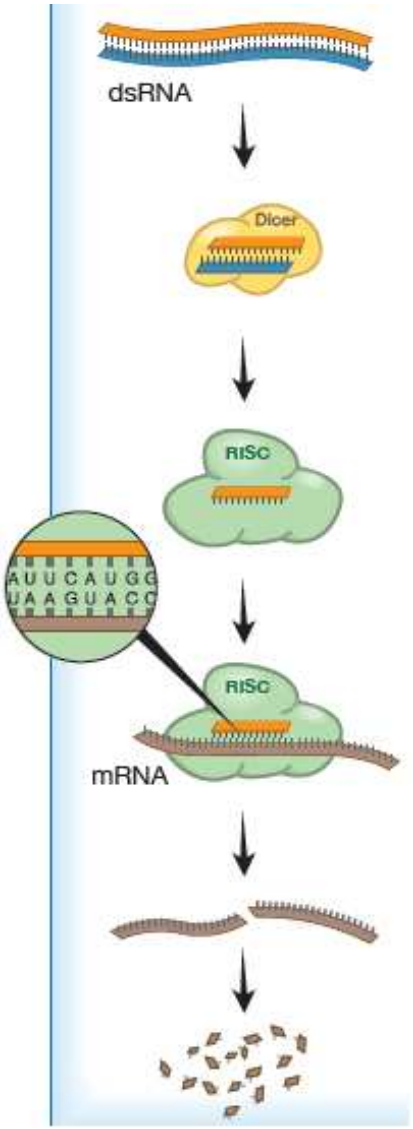


A remarcable exception to the Central Dogma

RNA-mediated interference (RNAi):
a natural process that uses small double-stranded RNA molecules (dsRNA) to control — and turn off — gene expression.

Recommended reading:
Bertil Daneholt, "RNA Interference", Advanced Information on The Nobel Prize in Physiology or Medicin 2006.

Note: this drawing and the next two ones are from the above cited paper.



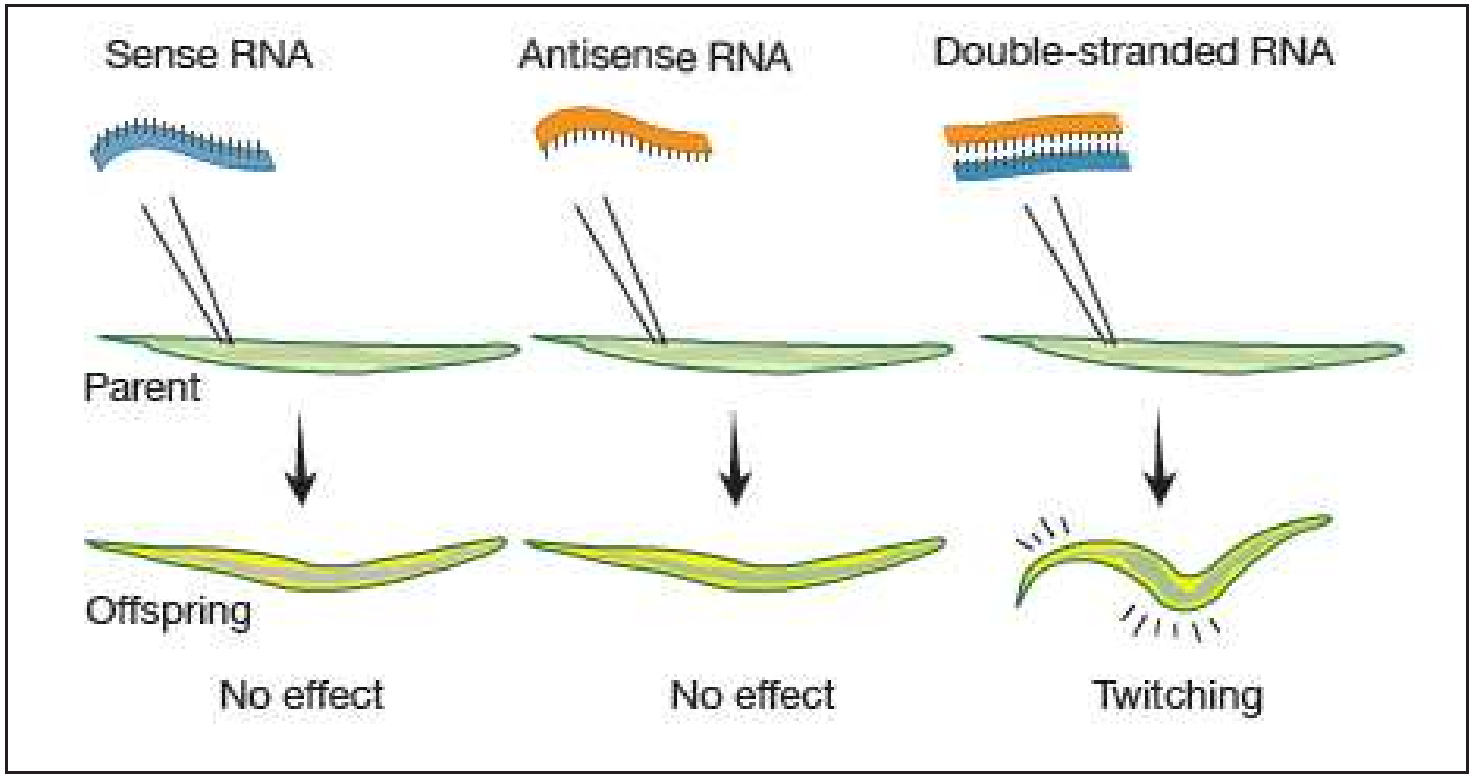
Nobel Prize for Physiology or Medicine, 2006

Awarded to Prof. Andrew Fire (Stanford University) and Prof. Craig Mello (University of Massachusetts),

for the elucidation of the RNA interference phenomenon,

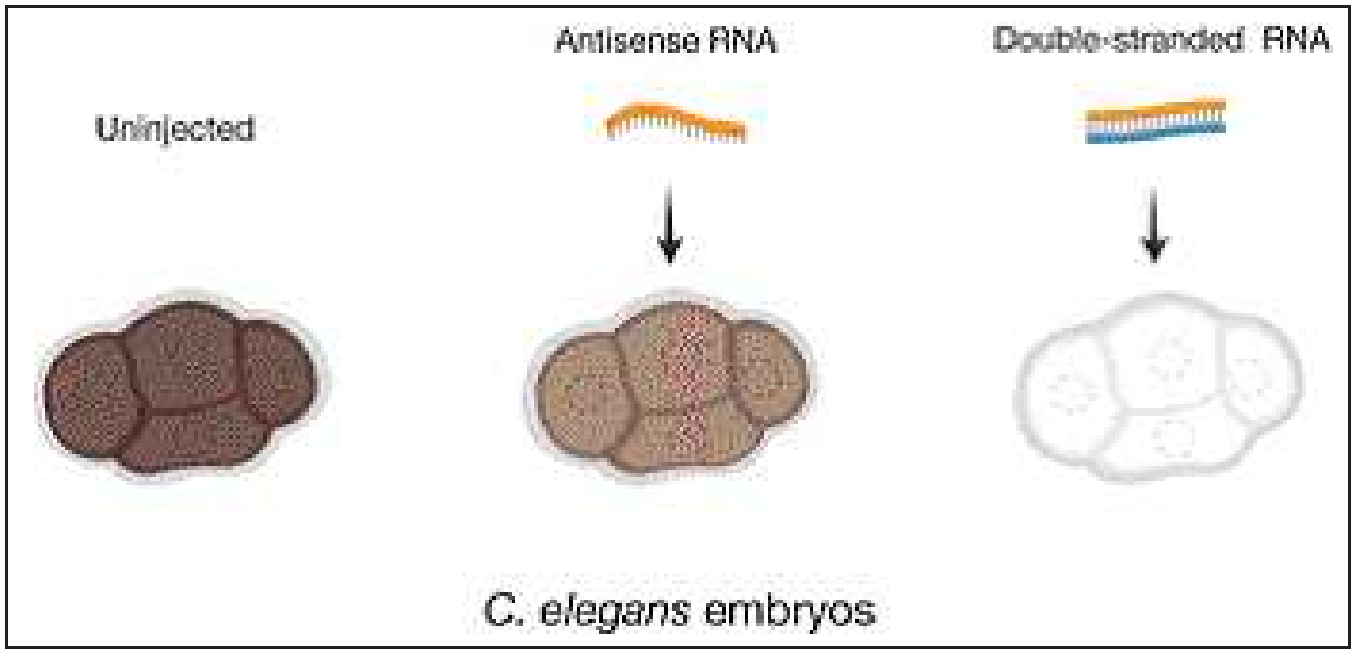
as described in the 1998 paper "Potent and specific genetic interference by double-stranded RNA in *Caernohabditis Elegans*" (Nature 391:806-811).

Fire & Mello experiences (I)



Phenotypic effect after injection of single-stranded or double-stranded *unc-22* RNA into the gonad of *C. elegans*.
Decrease in the activity of the *unc-22* gene is known to produce severe twitching movements.

Fire & Mello experiences (II)



The effect on *mex-3* mRNA content in *C. elegans* embryos after injection of single-stranded or double-stranded *mex-3* RNA into the gonad of *C. elegans*.

***mex-3* mRNA is abundant in the gonad and early embryos.**

The extent of colour reflects the amount of mRNA present.

RNAi explained *co-suppression of gene expression,* a phenomenon discovered in the early 1990s

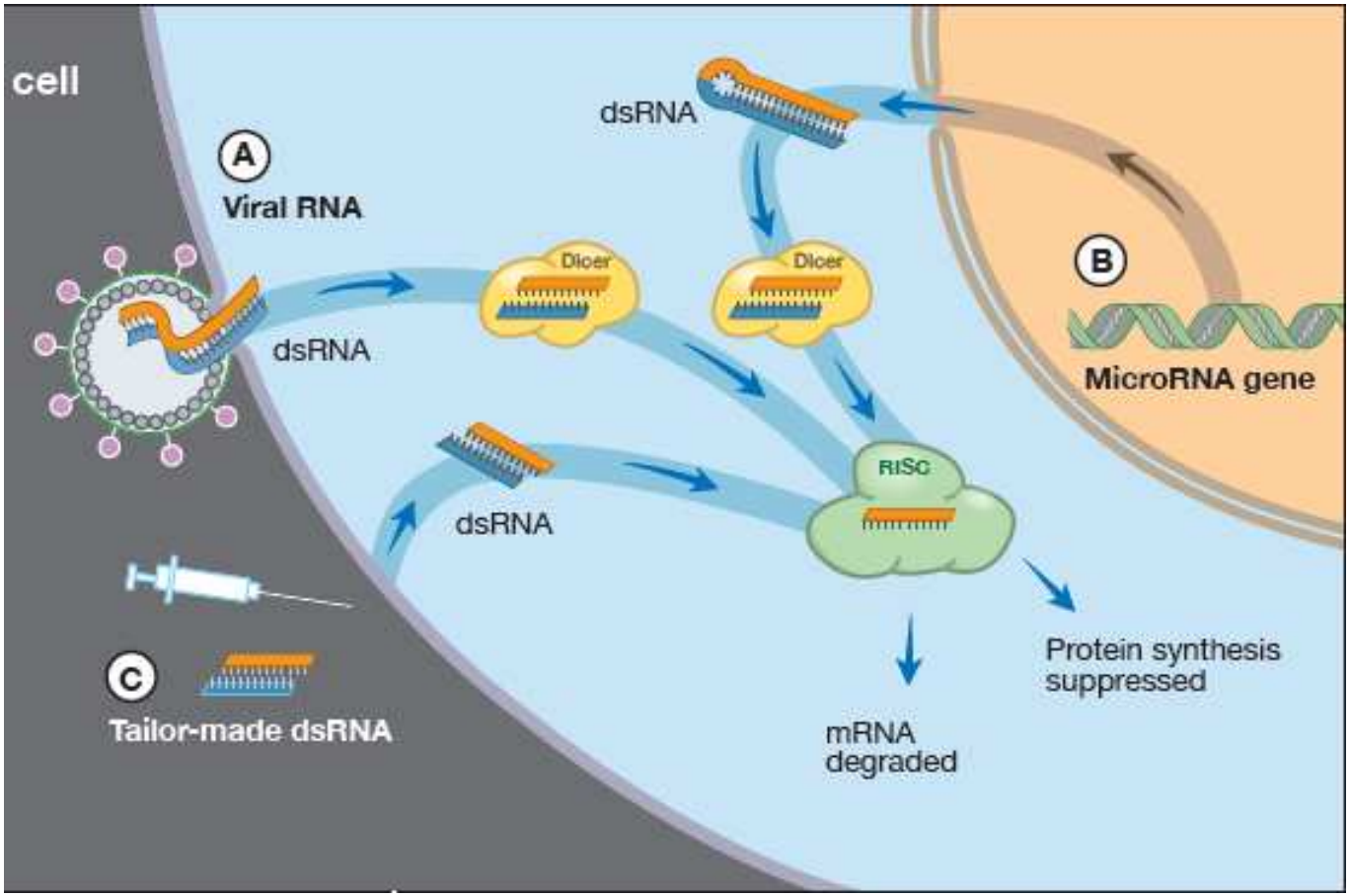
In an attempt to alter flower colors in petunias, researchers introduced additional copies of a gene encoding chalcone synthase, a key enzyme for flower pigmentation into petunia plants. The overexpressed gene instead produced less pigmented, fully or partially white flowers, indicating that the activity of chalcone synthase decreased substantially.

The left plant is wild type. The right plants contain transgenes that induce suppression of both transgene and endogeneous gene expression, giving rise to the unpigmented white areas of the flower. (From http://en.wikipedia.org/wiki/RNA_interference.)

RNAi implications

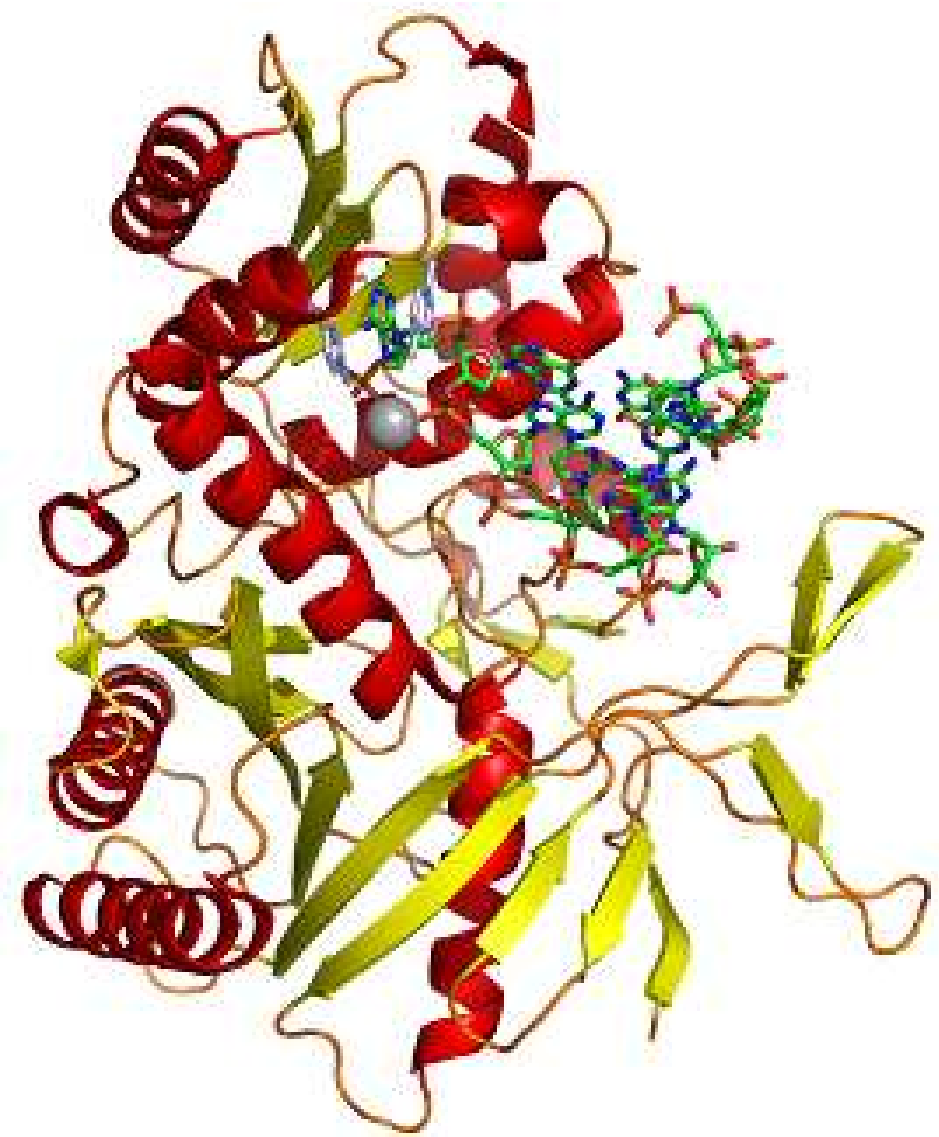
- **transcription regulation:**
  RNAi participates in the control of the amount of certain mRNA produced in the cell.
- **protection from viruses:**
  RNAi blocks the multiplication of viral RNA, and as such plays an import part in the organism's immune system.
- RNAi may serve to **identify the function of virtually any gene,** by knocking down/out the corresponding mRNA. In recent projects, entire libraries of short interfering RNAs (siRNAs) are created, aiming to silence every one gene of a chosen model organism.
- **therapeutically:**
  RNAi may help researchers design drugs for cancer, tumors, HIV, and other diseases.

RNA interference, a wider view



From D. Bertil Daneholt, "RNA interferation".
Advanced Information on the Nobel Prize in Physiology or Medicin 2006.
Karolinska Institutet, Sweden, 2006.

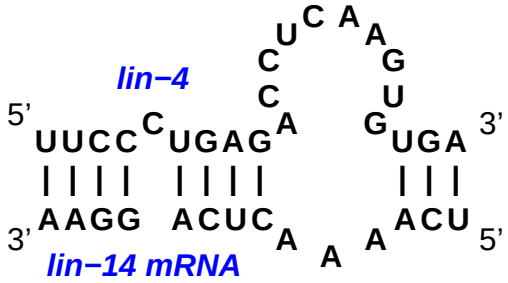
A double-stranded RNA attached to the PIWI domain of an argonaute protein in the RISC complex



From
http://en.wikipedia.org/wiki/RNA_interference
at 03.08.2007.

The first miRNA discovered: lin-4.

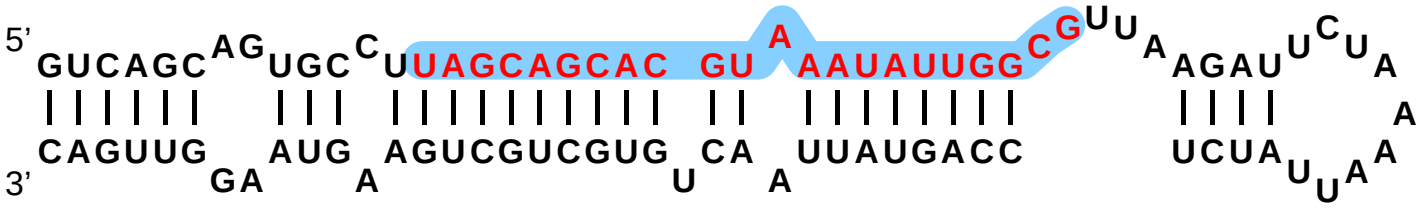
It regulates the lin-14 mRNA, a nuclear protein that controls larval development in *C. elegans*.



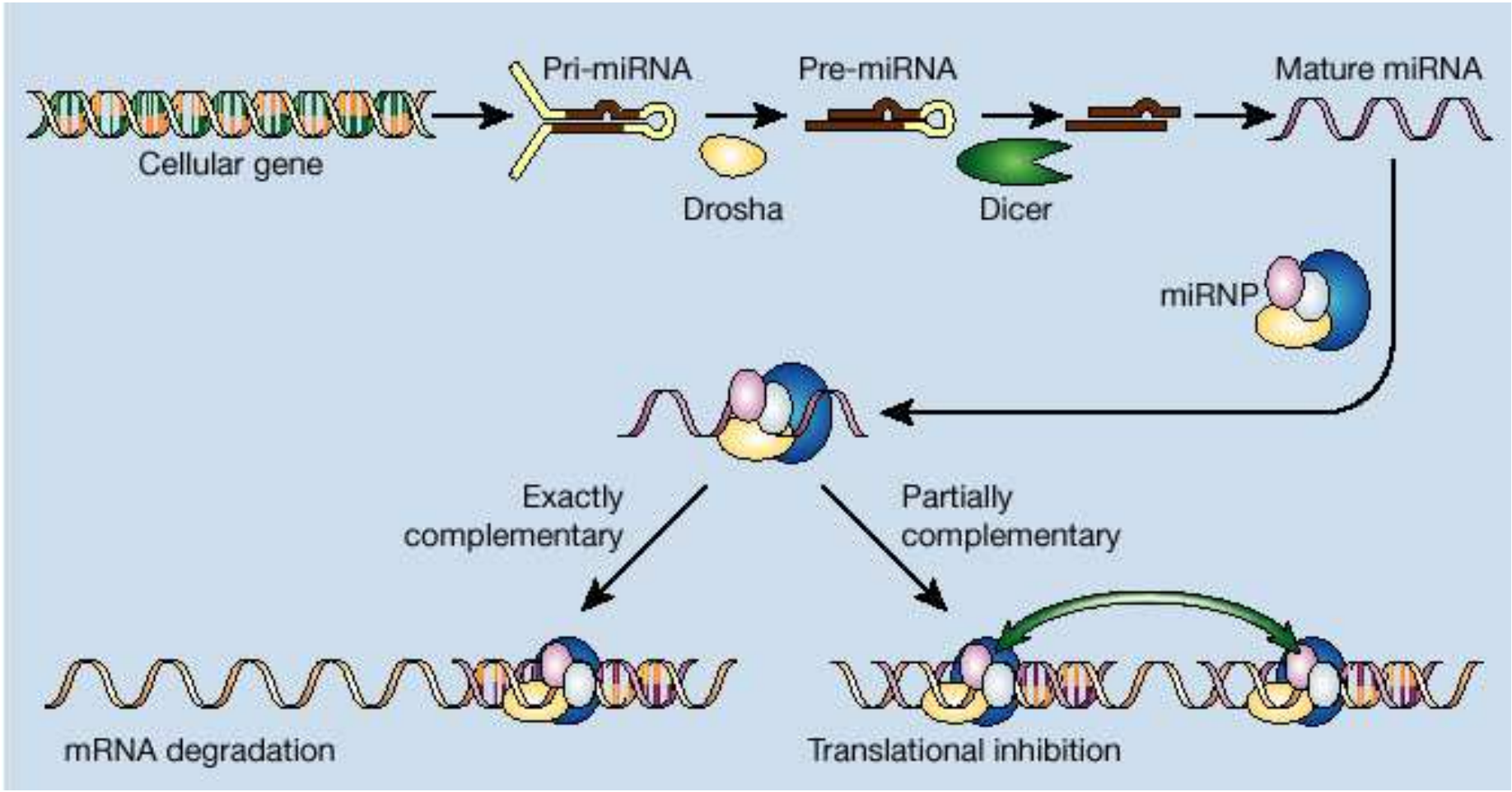
From P. Bengert and T. Dandekar, *Current efforts in the analysis of RNAi and RNAi target genes*, Briefings in Bioinformatics, Henry Stewart Publications, 6(1):72-85, 2005.

The stem-loop structure of human precursory miRNA mir-16.

Together with its companion mir-15a, they both have been proved to be deleted or downregulated in more than two thirds of cases of chronic lymphocytic leukemia. (The mature miRNA is shaded.)

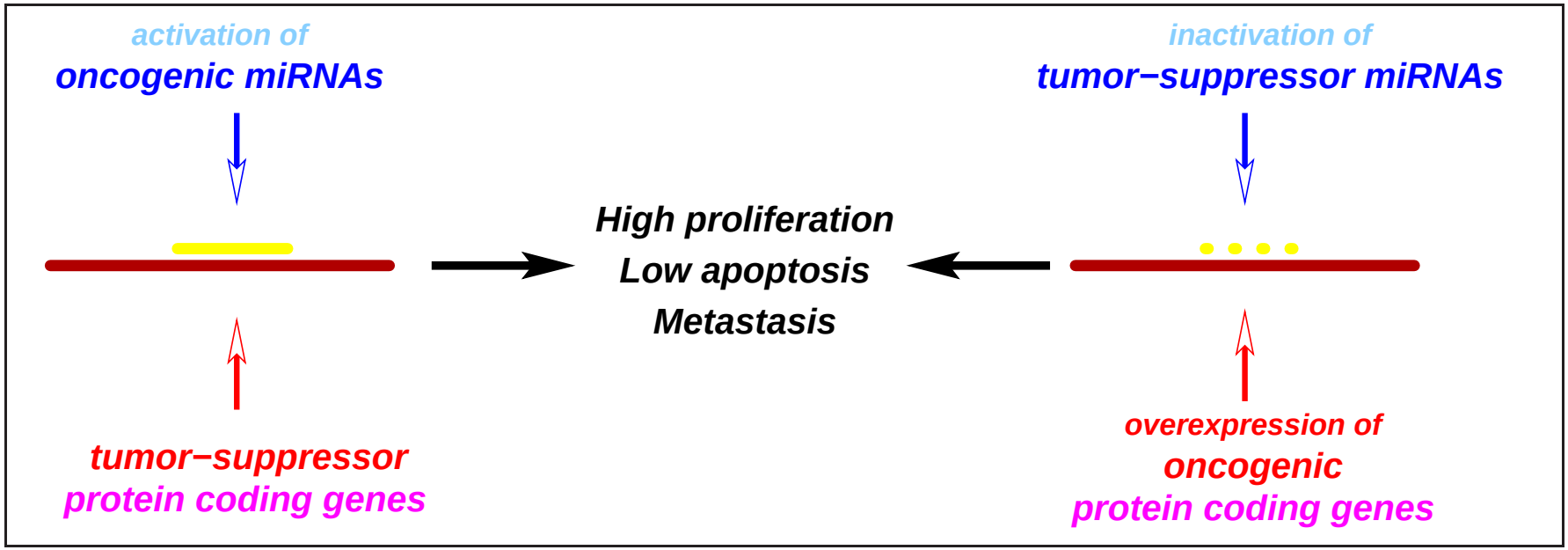


miRNA in the RNA interference process



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

The miRNA – cancer connection



Inspired by G.A. Călin, C.M. Croce, *MicroRNA–cancer connection: The beginning of a new tale*, Cancer Research, 66:(15), 2006, pp. 7390-7394.

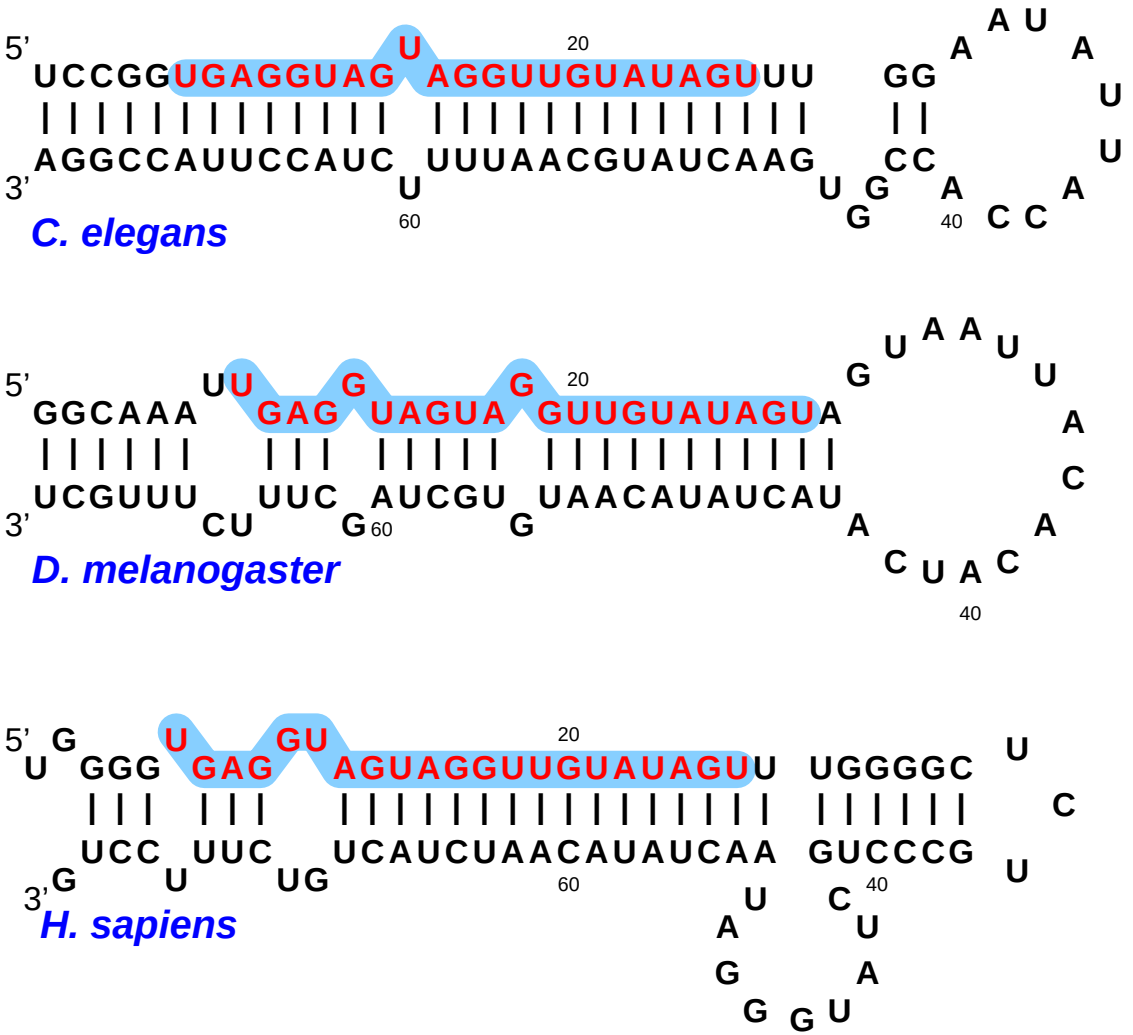
Specificities of miRNAs

- **Primary miRNAs can be located in**
  - **introns of protein-coding regions,**
  - **exons and introns of non-coding regions,**
  - **intergenic regions.**
- **MiRNAs tend to be situated in clusters, within a few kilobases. The miRNAs situated in a same cluster can be transcribed together.**
- **A highly conserved motif (with consensus CTCCGCCC for *C. elegans* and *C. briggsae*) may be present within 200bp upstream the miRNA clusters.**
- **The stem-loop structure of a pre-miRNA should have a low free energy level in order to be stable.**

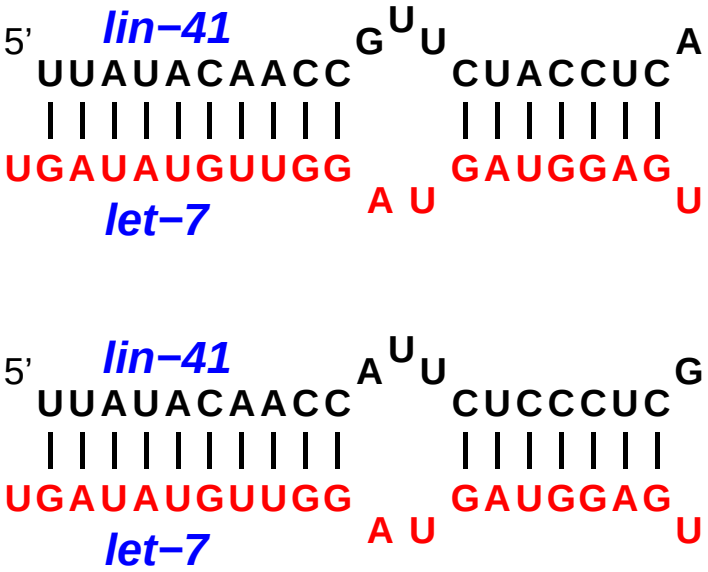
Specificities of miRNAs (Cont'd)

- Many miRNAs are conserved across closely related species (but there are only few universal miRNAs), therefore many prediction methods for miRNAs use genome comparisons.
  - The degree of conservation between orthologuos miRNAs is higher on the mature miRNA subsequence than on the flanking regions; loops are even less conserved.
- Conservation of miRNA sequences (also its length and structure) is lower for plants than it is for animals. In viruses, miRNA conservation is very low. Therefore miRNA prediction methods usually are applied/tuned to one of these three classes of organisms.
  - Identification of MiRNA target sites is easy to be done for plants (once miRNA genes and their mature subsequence are known) but is more complicated for animals due to the fact that usually there is an imperfect complementarity between miRNA mature sequences and their targets.

Example: A conserved microRNA: *let-7*

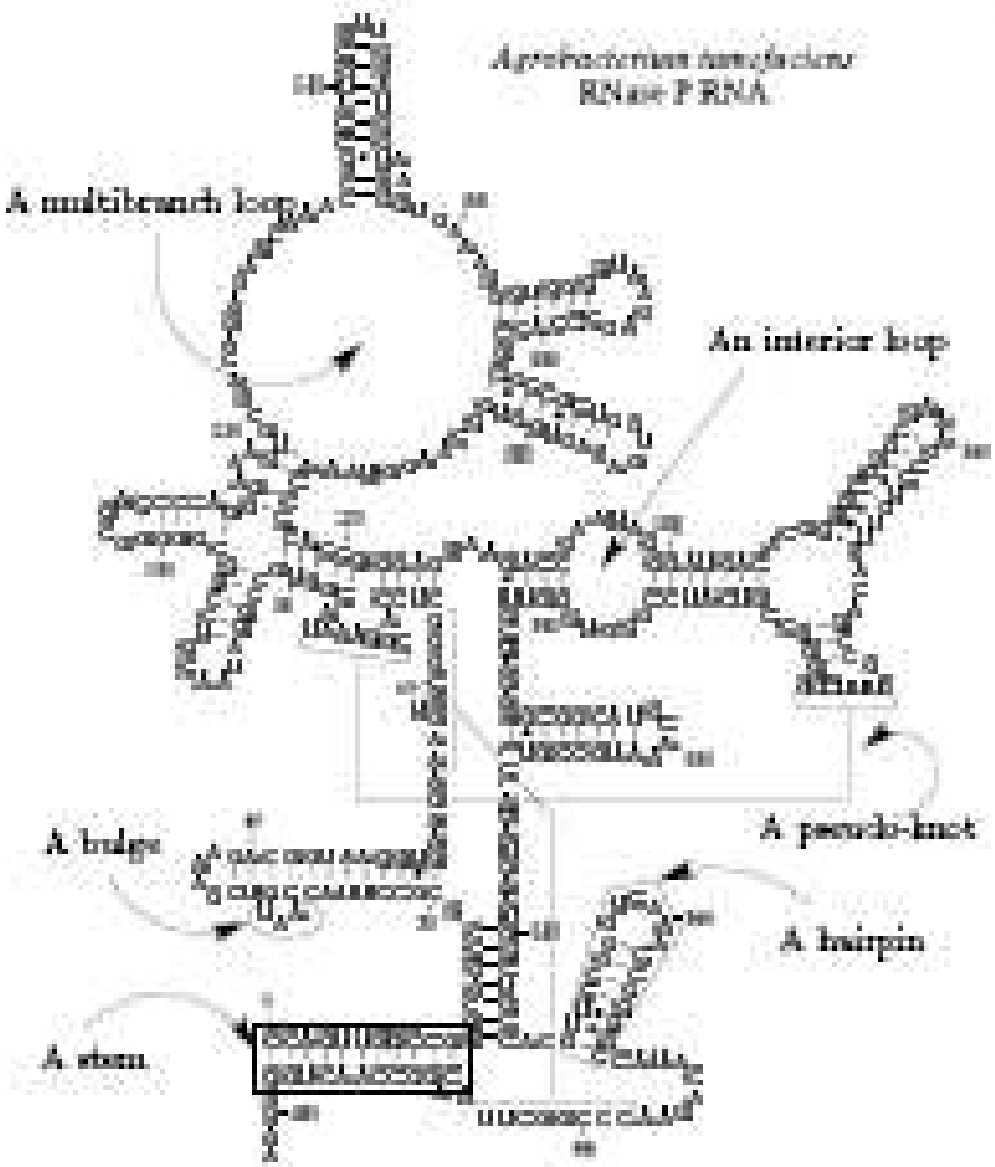


Example: Two targets sites of mature *let-7* miRNA on *lin-41* mRNA in *C. elegans*



2. RNA features

RNA secondary structure elements



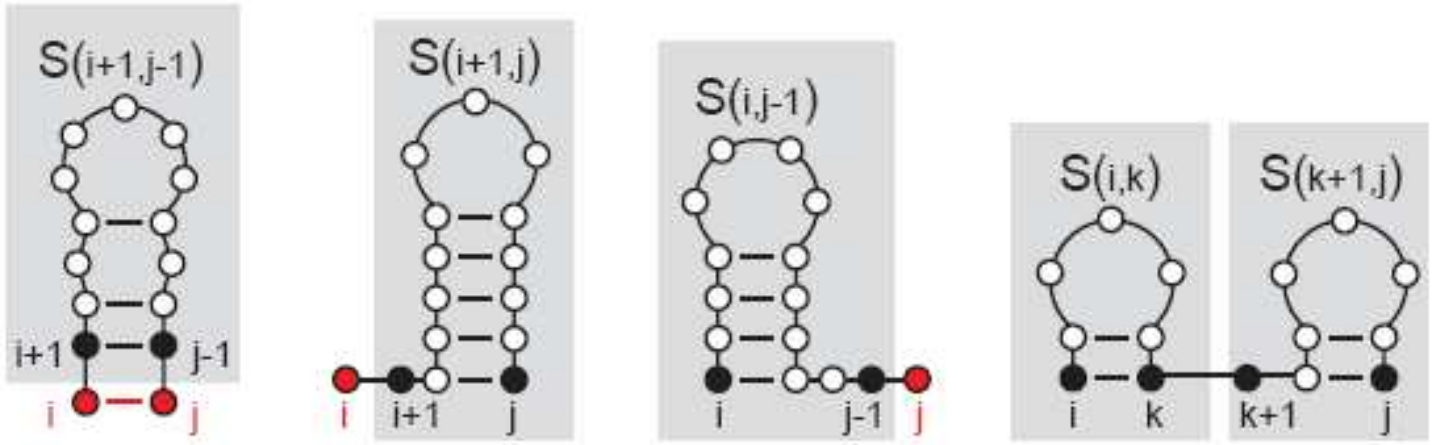
From "Efficient drwaing of RNA secondary structure", D. Auber, M. Delest, J.-P. Domenger, S. Dulucq, Journal of Graph Algorithms and Applications, 10(2):329-351 (2006).

2.1 A simple algorithm for RNA folding: Nussinov (1978)

Initialization: $S(i, i-1) = 0$ **for** $i = 2$ **to** L **and** $S(i, i) = 0$ **for** $i = 1$ **to** L

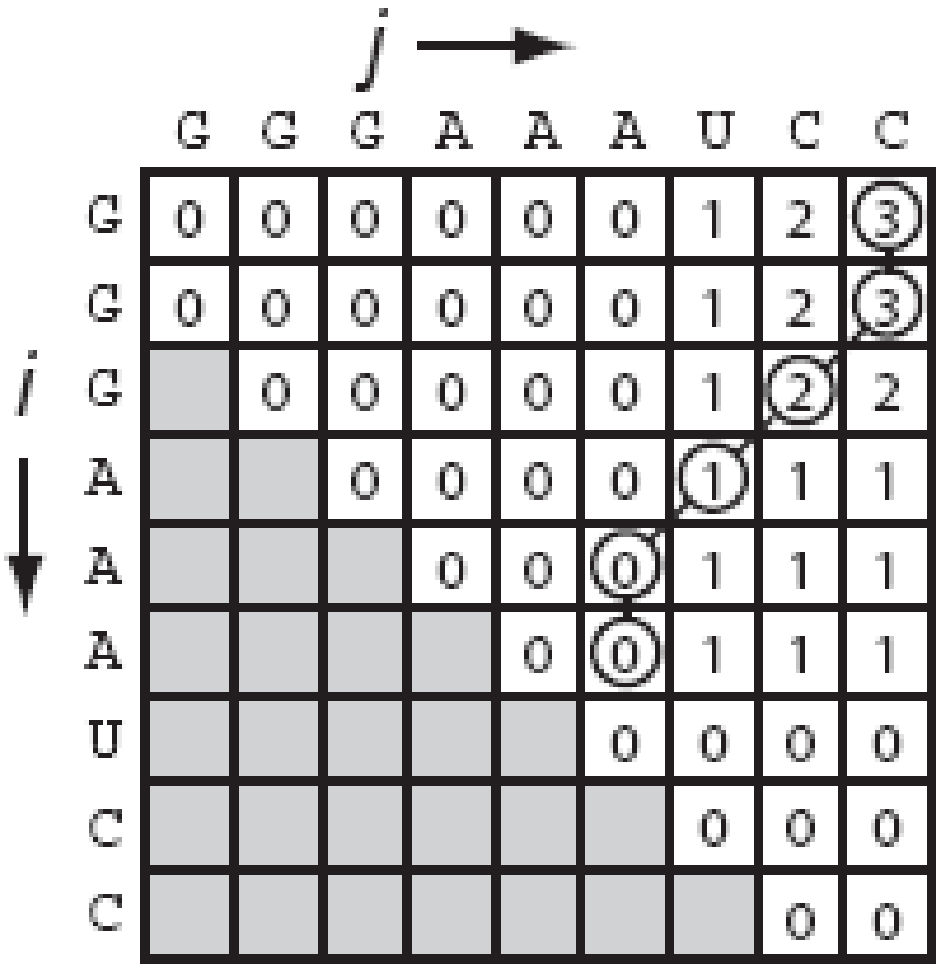
Recurrence:
$$S(i,j) = \mathbf{max} \begin{cases} S(i+1, j-1) + 1 \text{ if } [\, i,j \text{ base pair }] \\ S(i+1, j) \\ S(i, j-1) \\ \mathbf{max}_{i<k<j}\{S(i,k) + S(k+1, j)\} \end{cases}$$

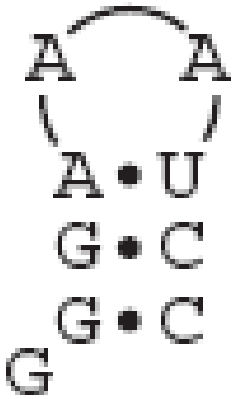
Output: $S(1, L)$



1. *i,j* pair; 2. *i* unpaired; 3. *j* unpaired; 4. bifurcation.

Nussinov algorithm: exemplification

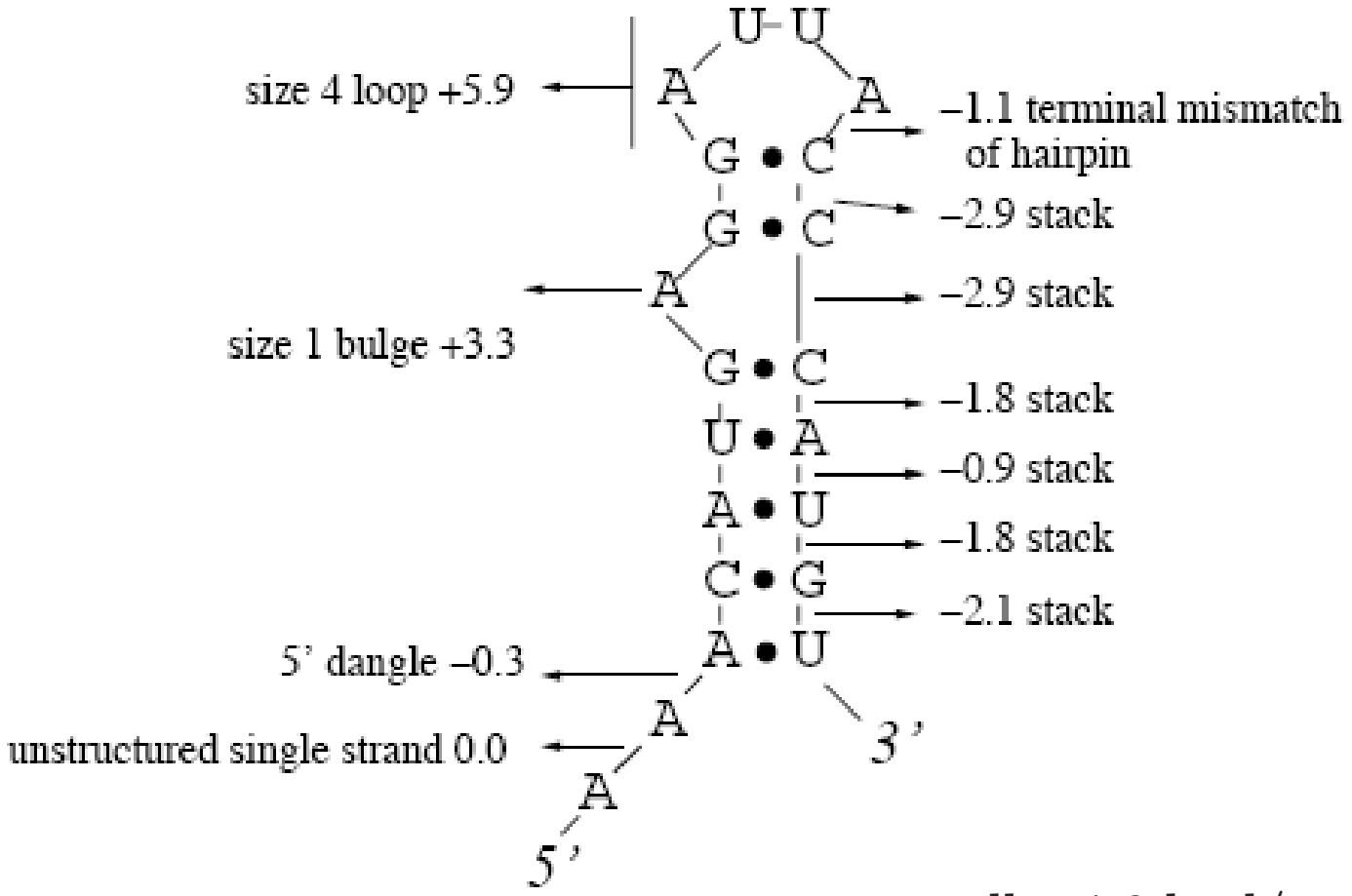




2.2 Computing the Minimum Free Energy (MFE) for RNAs

An example from [Durbin et al., 1999]

Note: For this example, the so-called "Freier's rules" were used [Turner et al, 1987]; they constitute a successor of Zuker's initial algorithm (1981).



overall: -4.6 kcal/mol

Predicted free-energy values (kcal/mol at $37°C$)

for base pair stacking

	A/U	**C/G**	**G/C**	**U/A**	**G/U**	**U/G**
A/U	−0.9	−1.8	−2.3	−1.1	−1.1	−0.8
C/G	−1.7	−2.9	−3.4	−2.3	−2.1	−1.4
G/C	−2.1	−2.0	−2.9	−1.8	−1.9	−1.2
U/A	−0.9	−1.7	−2.1	−0.9	−1.0	−0.5
G/U	−0.5	−1.2	−1.4	−0.8	−0.4	−0.2
U/G	−1.0	−1.9	−2.1	−1.1	−1.5	−0.4

for predicted RNA secondary structures, by size of loop

size	**internal loop**	**bulge**	**hairpin**
1	.	3.9	.
2	4.1	3.1	.
3	5.1	3.5	4.1
4	4.9	4.2	4.9
5	5.3	4.8	4.4
10	6.3	5.5	5.3
15	6.7	6.0	5.8
20	7.0	6.3	6.1
25	7.2	6.5	6.3
30	7.4	6.7	6.5

Remarks:

1. **The optimal folding of an RNA sequence corresponds to its minimum (level of) free energy.**
2. **We will not deal here with pseudo-knots.**

Notations

Given the sequence $x_1, \ldots, x_L$, we denote

$W(i,j)$: MFE of all non-empty foldings of the subsequence $x_i, \ldots, x_j$

$V(i,j)$: MFE of all non-empty foldings of the subsequence $x_i, \ldots, x_j$, containing the base pair (i,j)

***eh*(i,j): the energy of the hairpin closed by the pair (i,j)**

***es*(i,j): the energy of the stacked pair (i,j) and $(i+1, j-1)$**

***ebi*(i,j,i',j'): the energy of the bulge or interior loop that is closed by (i,j), with the pair (i',j') *accessible* from (i,j) (i.e., there is no base pair (k,l) such that $i < k < i' < l < j$ or $i < k < j' < l < j$).**

Zuker algorithm (1981)

Initialization: $W(i,j) = V(i,j) = \infty$ **for all** i, j **with** $j - 4 < i < j$**.**

Recurrence: **for all** i, j **with** $1 \leq i < j \leq L$

$$V(i,j) = \mathbf{min} \begin{cases} \boldsymbol{eh}(i,j) \\ \boldsymbol{es}(i,j) + V(i+1,j-1) \\ \boldsymbol{VBI}(i,j) \\ \boldsymbol{VM}(i,j) \end{cases} \qquad W(i,j) = \mathbf{min} \begin{cases} V(i,j) \\ W(i+1,j) \\ W(i,j-1) \\ \mathbf{min}_{i<k<j}\{W(i,k) + W(k+1,j)\} \end{cases}$$

$$\boldsymbol{VBI}(i,j) = \mathbf{min}_{\substack{i < i' < j' < j \\ i' - i + j - j' > 2}} \{\boldsymbol{ebi}(i,j,i',j') + V(i',j')\}$$

$$\boldsymbol{VM}(i,j) = \mathbf{min}_{i<k<j-1}\{W(i+1,k) + W(k+1,j-1)\} + a$$

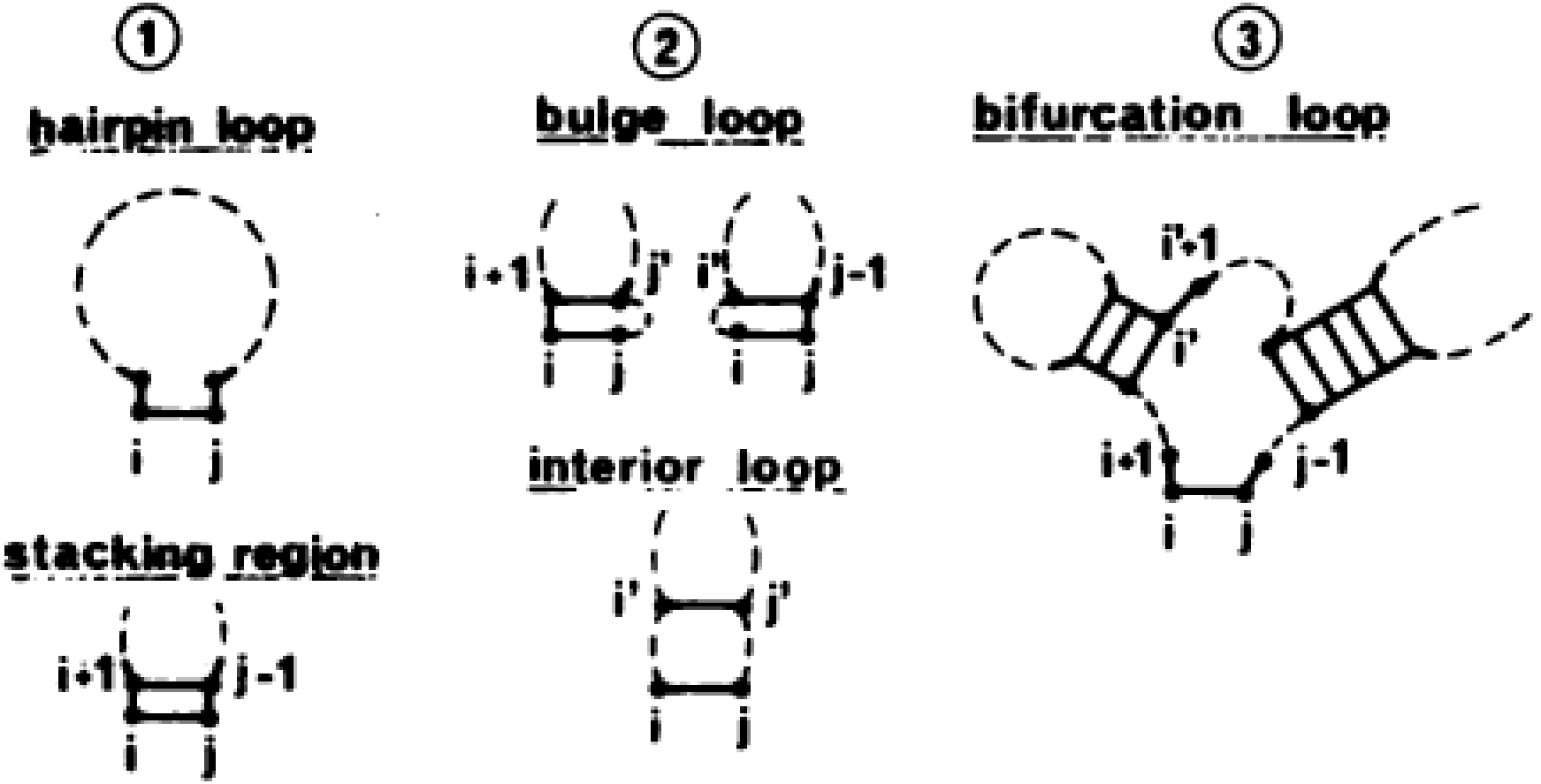
where a **is a constant energy contribution to close the multi-loop**

(more generally: e**(multi-loop)** $= a + b \times k' + c \times k$ **where** a, b, c **are constants and** k' **is the number of unpaired bases in the multi-loop)**

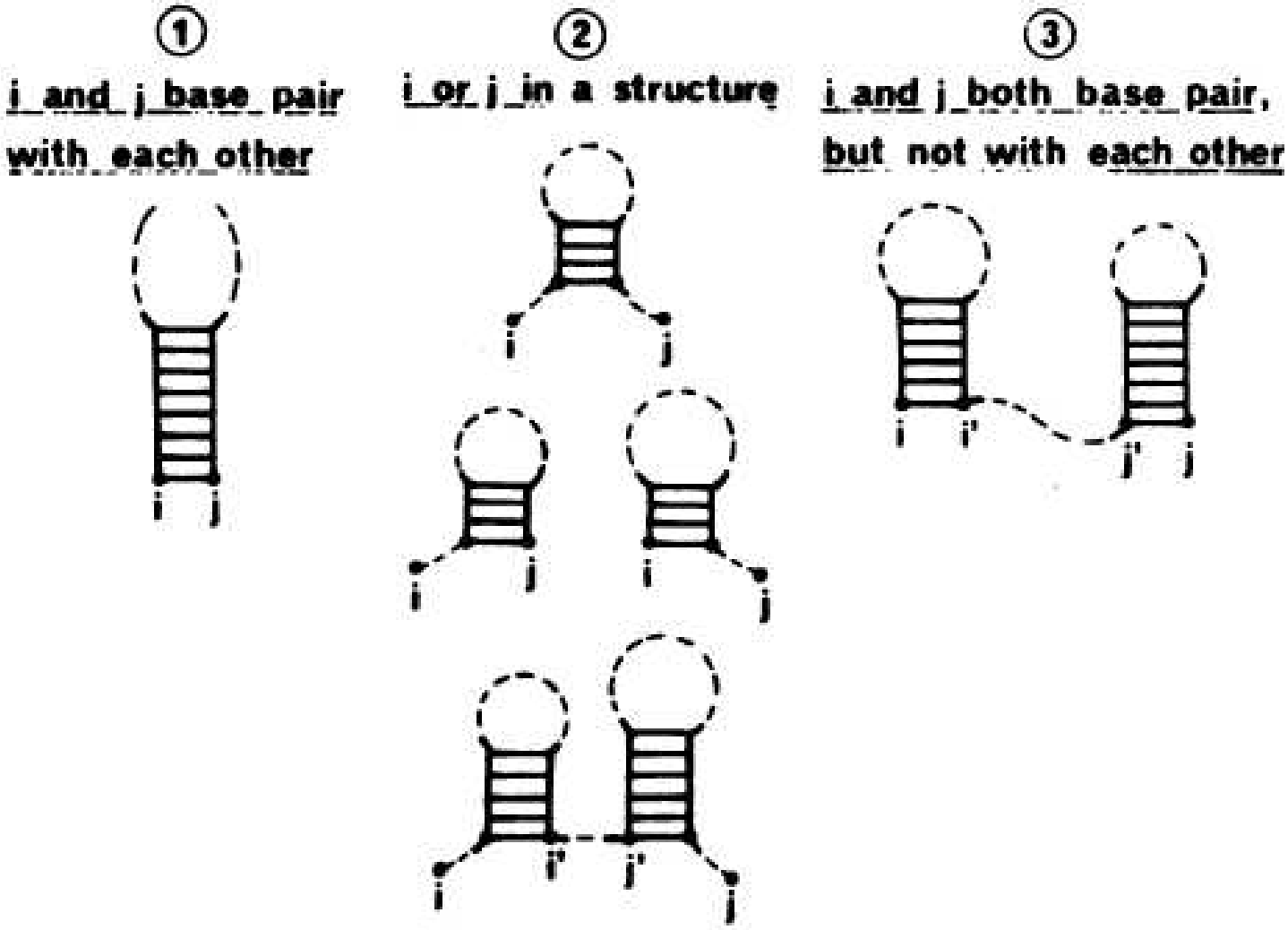
Complexity: $\mathcal{O}(L^4)$

(W : $\mathcal{O}(L^3)$, V : $\mathcal{O}(L^2)$, ***VBI***: $\mathcal{O}(L^4)$, ***VM***: $\mathcal{O}(L^3)$)

Illustrating the computation MFE for RNAs: $V(i,j)$



Illustrating the computation MFE for RNAs: $W(i,j)$



Subsequent refinements

Zuker implemented his algorithm as the *mfold* program and server. Later, various refinements have been added to the algorithm. For instance:

- **apart from the terms *eh* and *ebi* used in the computation of V, the *mfold* program uses stacking energies for the mismatched pairs additional the to stem closing base pairs.**
- **similarly, for bulges made of only one base, the stacking contribution of closing base pairs is added;**
- **there is a penalty for grossly asymmetric interior loops;**
- **an extra term is added for loops containing more than 30 bases: 1.75 RT *ln*(*size*/30), where $R = 8.31451$ J mol$^{-1}K^{-1}$ is the molar universal gas constant, and T is the absolute temperature.**

Zuker's algorithm was also impelmented by the *RNAfold* program, which is part of the Vienna RNA package and server.

2.3 Other RNA Folding Measures

[Freyhult et al., 2005]

Adjusted MFE:

$$dG(x) = \frac{MFE(x)}{L}$$

where $L =$ *length*(x).
It removes the bias that a long sequence tends to have a lower MFE.

MFE Index 1:

the ratio between $dG(x)$ and the percentage of the G+C content in the sequence x.

MFE Index 2:

$$\frac{dG(x)}{S}$$

where S is the number of stems in x that have more than three contiguous base-pairs.

Z-score — the number of standard deviations by which $MFE(x)$ differs from the mean MFE of $X_{\boldsymbol{shuffled}}(x)$, a set of shuffled sequences having the same dinucleotide composition as x:

$$Z(x) = \frac{MFE(x) - E(MFE(x') \ : \ x' \in X_{\boldsymbol{shuffled}}(x))}{\sigma(MFE(x') \ : \ x' \in X_{\boldsymbol{shuffled}}(x))}$$

P-value:

$$\frac{|\ \{x' \in X_{\boldsymbol{shuffled}}(x) \ : \ MFE(x') < MFE(x)\}\ |}{|\ X_{\boldsymbol{shuffled}}(x)\ |}$$

Note: See the **Altschul-Erikson** algorithm (1985) for sequence shuffling.

Adjusted base-pairing propensity: $dP(x)$

the average number of base pairs in the secondary structure of x.
It removes the bias that longer sequences tend to have more base-pairs.

Adjusted Shannon entropy:

$$dQ(x) = \frac{-\sum_{i<j} p_{ij} \log_2(p_{ij})}{L}$$

where p_{ij} is the probability that (x_i, x_j) is a base-pair in x:

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

and

$\mathcal{S}(x)$ is the set of all secondary structures corresponding to x;

δ_{ij}^{α} is 1 if x_i and x_j is a base-pair in the structure S_α, and 0 otherwise;

the probability of $S_\alpha \in \mathcal{S}(x)$ follows a Boltzmann distribution:

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

with

$Z = \sum_{S_\alpha \in \mathcal{S}(x)} e^{-MFE_\alpha/RT}$,
$R =$ **8.31451 Jmol$^{-1}K^{-1}$ (a molar gas constant), and**
$T =$ **310.15K (37° C)**

Note: Low values of dQ indicate that i. one or a few base-pairs are dominant in the RNA's structure(s), or ii. there are no base-pairs at all.

Adjusted base-pair distance (or *ensemble diversity*):

$$dD(x) = \frac{\frac{1}{2}\sum_{S_\alpha, S_\beta \in \mathcal{S}(x)} P(S_\alpha)P(S_\beta)d_{BP}(S_\alpha, S_\beta)}{L}$$

where $d_{BP}(S_\alpha, S_\beta)$, the base-pair distance between two structures S_α and S_β of the sequence x, is defined as the number of base-pairs not shared by the structures S_α and S_β:

$$d_{BP}(S_\alpha, S_\beta) = \mid S_\alpha \cup S_\beta \mid - \mid S_\alpha \cap S_\beta \mid = \mid S_\alpha \mid + \mid S_\beta \mid -2 \mid S_\alpha \cap S_\beta \mid .$$

Because $\mid S_\alpha \mid = \sum_{i<j} \delta_{ij}^{\alpha}$, we get $d_{BP}(S_\alpha, S_\beta) = \sum_{i<j}(\delta_{ij}^{\alpha} + \delta_{ij}^{\beta} - 2\delta_{ij}^{\alpha}\delta_{ij}^{\beta})$, and following a quite straightforward calculus (see [Freyhult et al., 2005]) we arrive at a simpler form for dD:

$$dD(x) = \frac{\sum_{i<j} (p_{ij} - p_{ij}^2)}{L}$$

Note: The probabilities p_{ij} are efficiently computed using the algorithm presented in [McCaskill, 1990].

2.4 A similarity measure for the RNA secondary structure

In order to approximate the topology of an RNA, [Gan et al., 2003] proposed the following notions:

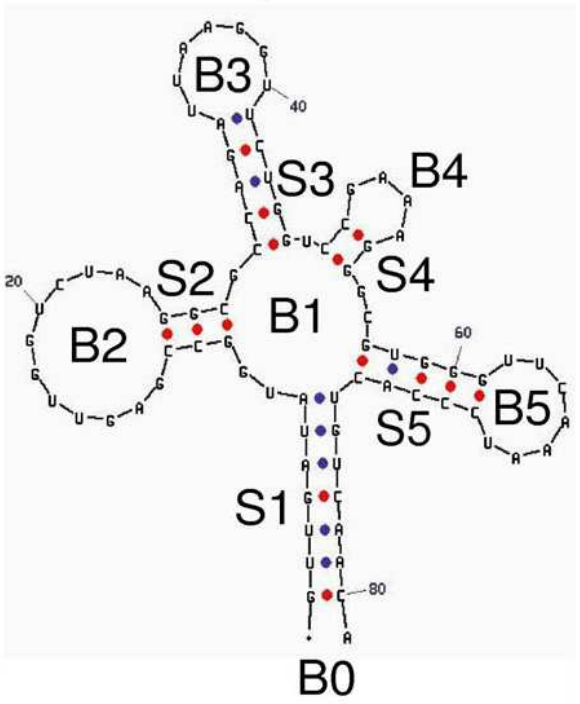
- tree graph for an RNA without pseudo-knots
  - each bulge, hairpin loop or wobble ("internal loop") constitutes a vertex
  - the 3' and 5' ends of a stem are assigned (together) a vertex;
  - a multi-loop ("junction") is a vertex;
- dual graph for an RNA with of without pseudo-knots
  - a vertex is a double stranded stem;
  - an edge is a single strand that connects secondary structure elements (bulges, wobbles, loops, multi-loops and stems).

Note: It is possible that two distinct RNAs map onto the same (tree and respectively dual) graph.

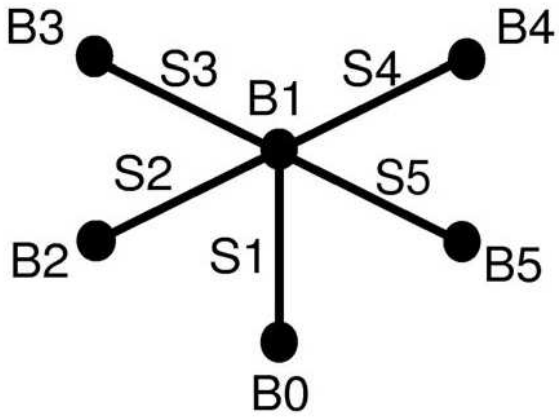
Tree graphs and dual graphs: Exemplification

A tRNA (leu) from [Fera et al., 2004]

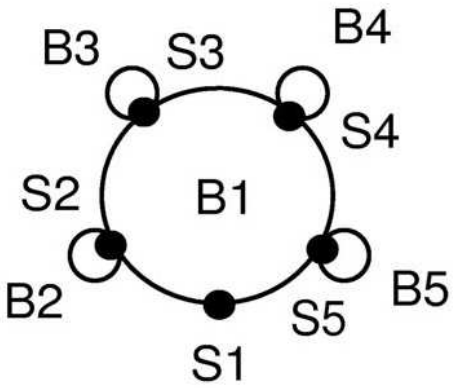
Secondary structure



Tree graph representation



Dual graph representation



A similarity measure for the RNA secondary structure (Cont'd)

Spectral techniques in graph theory [Mohar, 1991] can serve to quantitatively characterize the tree graphs and dual graphs assigned to RNAs.

Let G be an unoriented graph, possibly having loops and multiple edges.

Notations:

- **$A(G)$ is the *adjiacency matrix* of the graph G: a_{uv} is the number of edges between vertices u and v;**
- **$D(G)$ is the *degree matrix* of G: $d_{uv} = 0$ for $u \neq v$, and $d_{uu} = \sum_v a_{uv}$;**
- **$L(G) = D(G) - A(G)$ is called the Laplacian matrix of the graph G;**
- **$L(G)X - \lambda X$ is named the characteristic polynomial of the matrix $L(G)$. Its roots $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ are called the Laplacian eigenvalues of G, where $n = \mid V(G) \mid$ denotes the number of vertices in G. The tuple $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is called the spectrum of G; it can be shown that it is independent of the labelings of the graph vertices.**

It can be proved that

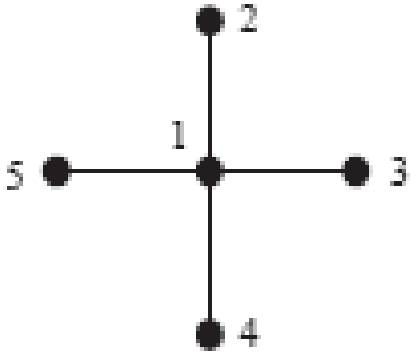
$\lambda_1 = 0$ and $\lambda_2 > 0$ if and only if the graph G is connected, and graphs with resambling topologies have closed λ_2 values.

Thus λ_2 can be used as a measure of similarity between graphs; some authors call it graph connectivity.

Computing eigenvalues for a tree graph: Exemplification

(from [Gan et al, 2004])

Tree Graph for tRNA (NDB: TRNA12)
(Randomly Labeled)



Adjacency Matrix (A)

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Diagonal Matrix (D)

$$\begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Laplacian Matrix ($L=D-A$)

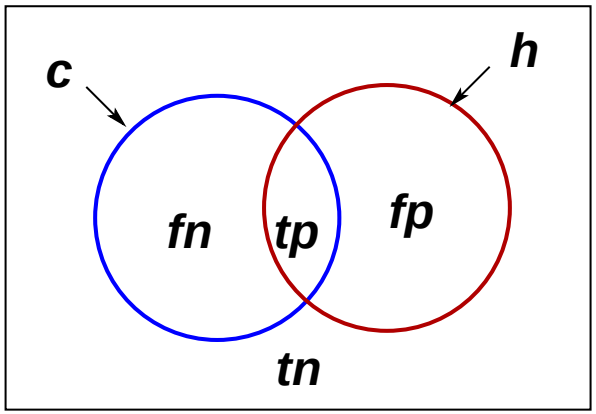
$$\begin{bmatrix} 4 & -1 & -1 & -1 & -1 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Laplacian Eigenvalues

$\lambda_1=0, \lambda_2=1, \lambda_3=1, \lambda_4=1, \lambda_5=5$

3. Machine Learning (ML) issues

3.1 Evaluation measures in Machine Learning



tp – true positives
fp – false positives
tn – true negatives
fn – false negatives

accuracy: $Acc = \frac{tp + tn}{tp + tn + fp + fn}$

precision: $P = \frac{tp}{tp + fp}$

recall (or: sensitivity): $R = \frac{tp}{tp + fn}$

specificity: $Sp = \frac{tn}{tn + fp}$

follout: $= \frac{fp}{tn + fp}$

F-measure: $F = \frac{2\ P \times R}{P+R}$

Mathew's Correlation Coefficient:

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp) \times (tn + fn) \times (tp + fn) \times (tn + fp)}}$$

3.2 Support Vector Machines (SVMs)

3.2.1 SVMs: The linear case

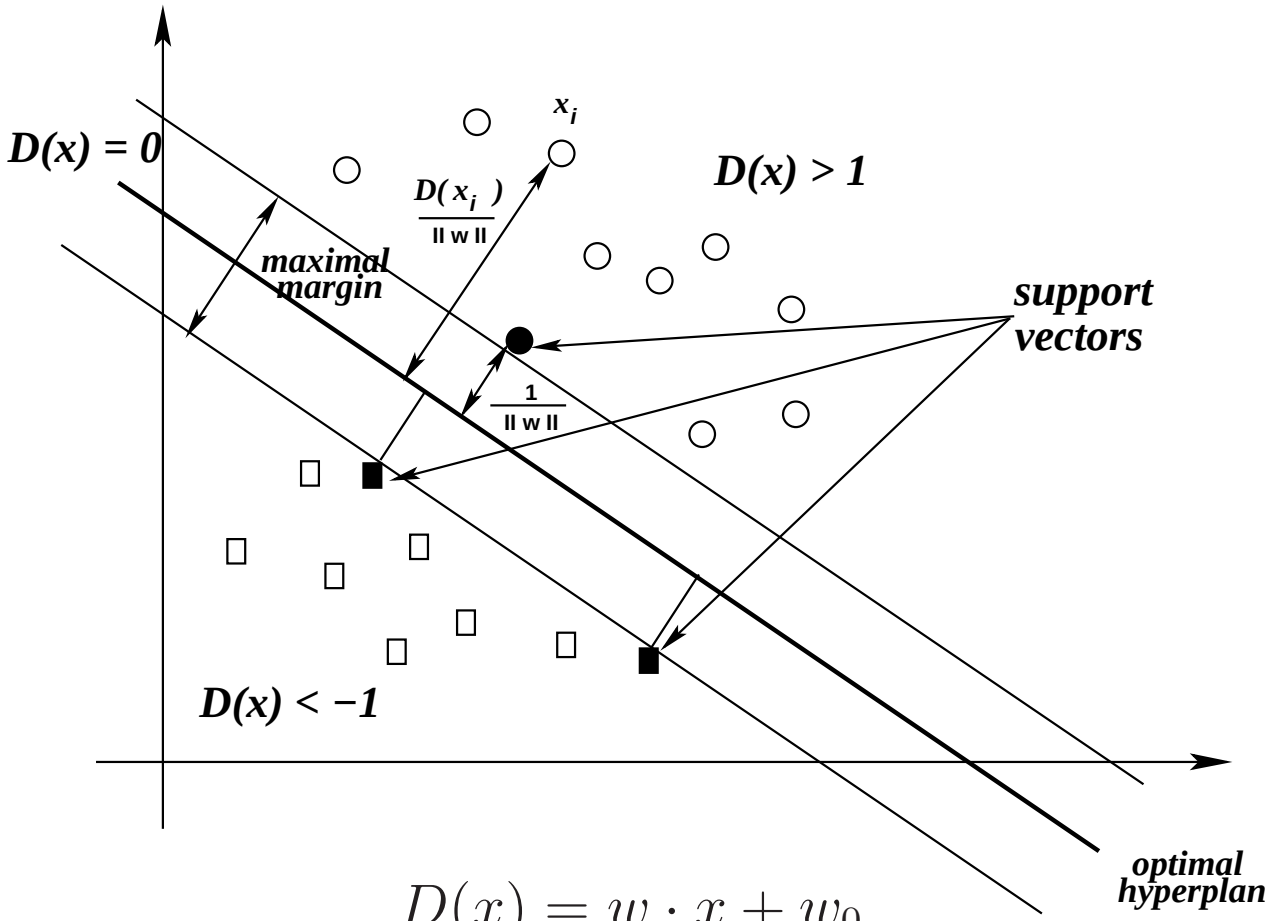
Formalisation:

Let S be a set of points $x_i \in R^d$ with $i = 1, \ldots, m$.
Each point x_i belongs to either of two classes, with label $y_i \in \{-1, +1\}$.
The set S is linear separable if there are $w \in R^d$ and $w_0 \in R$ such that

$$y_i(w \cdot x_i + w_0) \geq 1 \qquad i = 1, \ldots, m$$

The pair (w, w_0) defines the hyperplane of equation $w \cdot x + w_0 = 0$, named the separating hyperplane.

The optimal separating hyperplane



$$D(x) = w \cdot x + w_0$$

Linear SVMs

The Primal Form:

$$\begin{array}{ll} \textbf{\textit{minimize}} & \frac{1}{2}||w||^2 \\ \textbf{\textit{subject to}} & y_i(w \cdot x_i + w_0) \geq 1 \textbf{ for } i = 1, \ldots, m \end{array}$$

Note: This is a constrained quadratic problem with $d+1$ parameters. It can be solved by optimisation methods if d is not very big (10^3).

The Dual Form:

$$\begin{array}{ll} \textbf{\textit{maximize}} & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \; x_i \cdot x_j \\ \textbf{\textit{subject to}} & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & \alpha_i \geq 0, i = 1, \ldots, m \end{array}$$

The link between the optimal solutions of the primal and the dual form:

$$\overline{w} = \sum_{i=1}^{m} \overline{\alpha}_i y_i x_i$$

$$\overline{\alpha}_i(y_i(\overline{w} \cdot x_i + \overline{w}_0) - 1) = 0 \textbf{ for any } i = 1, \ldots, m$$

Linear SVMs with Soft Margin

If the set S is not linearly separable — or one simply ignores whether or not S is linearly separable —, the previous analysis can be generalised by introducing m non-negative variables ξ_i, for $i = 1, \ldots, m$ such that $y_i(w \cdot x_i + w_0) \geq 1 - \xi_i$, for $i = 1, \ldots, m$

The primal form:

$$\begin{array}{ll} \textbf{\textit{minimize}} & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i \\ \textbf{\textit{subject to}} & y_i(w \cdot x_i + w_0) \geq 1 - \xi_i \text{ for } i = 1, \ldots, m \\ & \xi_i \geq 0 \text{ for } i = 1, \ldots, m \end{array}$$

The associated dual form:

$$\begin{array}{ll} \textbf{\textit{maximize}} & \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \; x_i \cdot x_j \\ \textbf{\textit{subject to}} & \sum_{i=1}^{m} y_i\alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, i = 1, \ldots, m \end{array}$$

$\overline{w} = \sum_{i=1}^{m} \overline{\alpha}_i y_i x_i$

As before: $\overline{\alpha}_i(y_i(\overline{w} \cdot x_i + \overline{w}_0) - 1 + \overline{\xi}_i) = 0$

$(C - \overline{\alpha}_i)\,\overline{\xi}_i = 0$

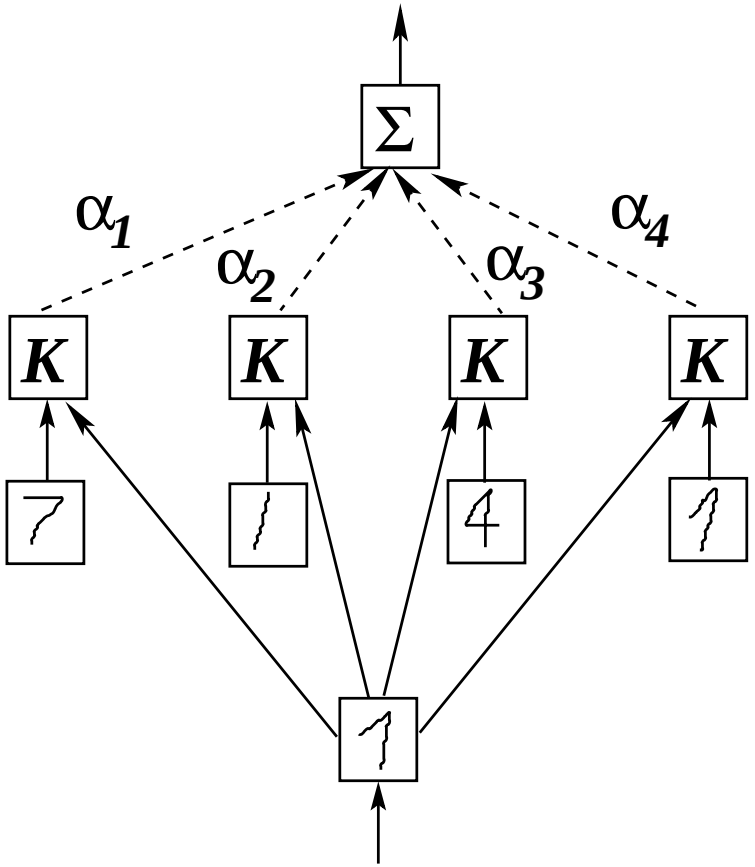
The role of the regularizing parameter C

large $C \Rightarrow$ minimize the number of misclassified points

small $C \Rightarrow$ maximize the minimum distance $1/||w||$

3.2.2 Non-linear SVMs and Kernel Functions

illustrated for the the problem of hand-written character recognition



Output: $\boldsymbol{sign}(\sum_i \alpha_i y_i K(x_i, x) + w_0)$

Comparison: $K(x_i, x)$

Support vectors: $x_1, x_2, x_3, \ldots$

Input: x

3.3 Feature selection: An information theory-based approach

Basic notions:

Let X and Y be two random variables.

- **The entropy of Y:**

  $H(Y) = -\sum_y P(Y=y) \log P(Y=y)$
  rewritten for convenience as $-\sum_y p(y) \log p(y) = E(\log \frac{1}{p(y)})$

  $H(Y)$ **describes the diversity of (the values taken by)** Y**:**
  the greater the diversity of Y**, the larger the value** $H(Y)$**.**

- **The mutual information between X and Y:**

  $I(X;Y) = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y)$
  with $H(Y \mid X) = -\sum_x \sum_y p(x,y) \log p(y \mid x)$**.**

  $I(X;Y)$ **characterises the relation between** X **and** Y**:**
  the stronger the relation, the larger the value of $I(X;Y)$**.**

  $I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)\ p(y|z)}$.

Descrete Function Learning (DFL) algorithm

[Zheng and Kwoh, 2005]

The theoretical setup

Theorem: (Cover and Thomas, 1991, "Elements of Information Theory"):

$I(X;Y) = H(Y)$ **implies that** Y **is a function of** X**.**

It is immediate that $I(X;Y) = H(Y)$ **is equivalent with** $H(Y \mid X) = 0$ **i.e., there is no more diversity of** Y **if** X **has been known.**

Generalisation: Let $X_1, X_2, \ldots, X_n$ and Y be random variables;
if $I(X_1, X_2, \ldots, X_n; Y) = H(Y)$ **then** Y **is a function of** $X_1, X_2, \ldots, X_n$**.**

The proof uses the following chain rules:

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, X_2, \ldots, X_{n-1})$$

$$I(X_1, X_2, \ldots, X_n; Y) =$$
$$I(X_1; Y) + I(X_2; Y \mid X_1) + \ldots + I(X_n; Y \mid X_1, X_2, \ldots, X_{n-1})$$

This generalisation is the basis of the DF Learning algorithm.

DFL: the algorithm

Let us consider a set of training instances characterised by $X_1, X_2, \ldots, X_n$ as input (categorical) attributes and Y, the output (i.e. class) attribute.

We aim to find the input attributes that contribute most to the class distinction.

Algorithm:

$V = \{X_1, X_2, \ldots, X_n\},\ U_0 = \emptyset,\ s = 1$

do

$A_s = \text{argmax}_{X_i \in V \setminus U_{s-1}}\ I(U_{s-1}, X_i; Y)$

$U_s = U_{s-1} \cup \{A_s\}$

s = s+1

until $I(U_s; Y) = H(Y)$

Improvements:

The 'until' condition can be replaced with either

$H(Y) - I(U_s; Y) < \epsilon$

or

$s > K$

with ϵ and K used as parameters of the DFL (modified) algorithm.

3.4 Ensemble Learning: a very brief introduction

There exist two well-known meta-learning techniques that aggregate *classification trees*:

Boosting [Shapire et al., 1998]:

When constructing a new tree, the data points that have been incorrectly predicted by earlier trees are given some extra wight, thus forcing the learner to concentrate successively on more and more difficult cases.
In the end, a weighted vote is taken for prediction.

Bagging [Breiman, 1996]:

New trees do not depend on earlier trees; each tree is independently constructed using a boostrap sample (i.e. sampling with replacing) of the data set.
The final classification is done via simple majority voting.

Random Forests (RF)

[Breiman, 2001]

RF extends bagging with and additional layer of randomness:

random feature selection:

While in standard classification trees each node is split using the best split among all variables, in RF each node is split using the best among a subset of features randomly chosen at that node.

RF uses only two parameters:

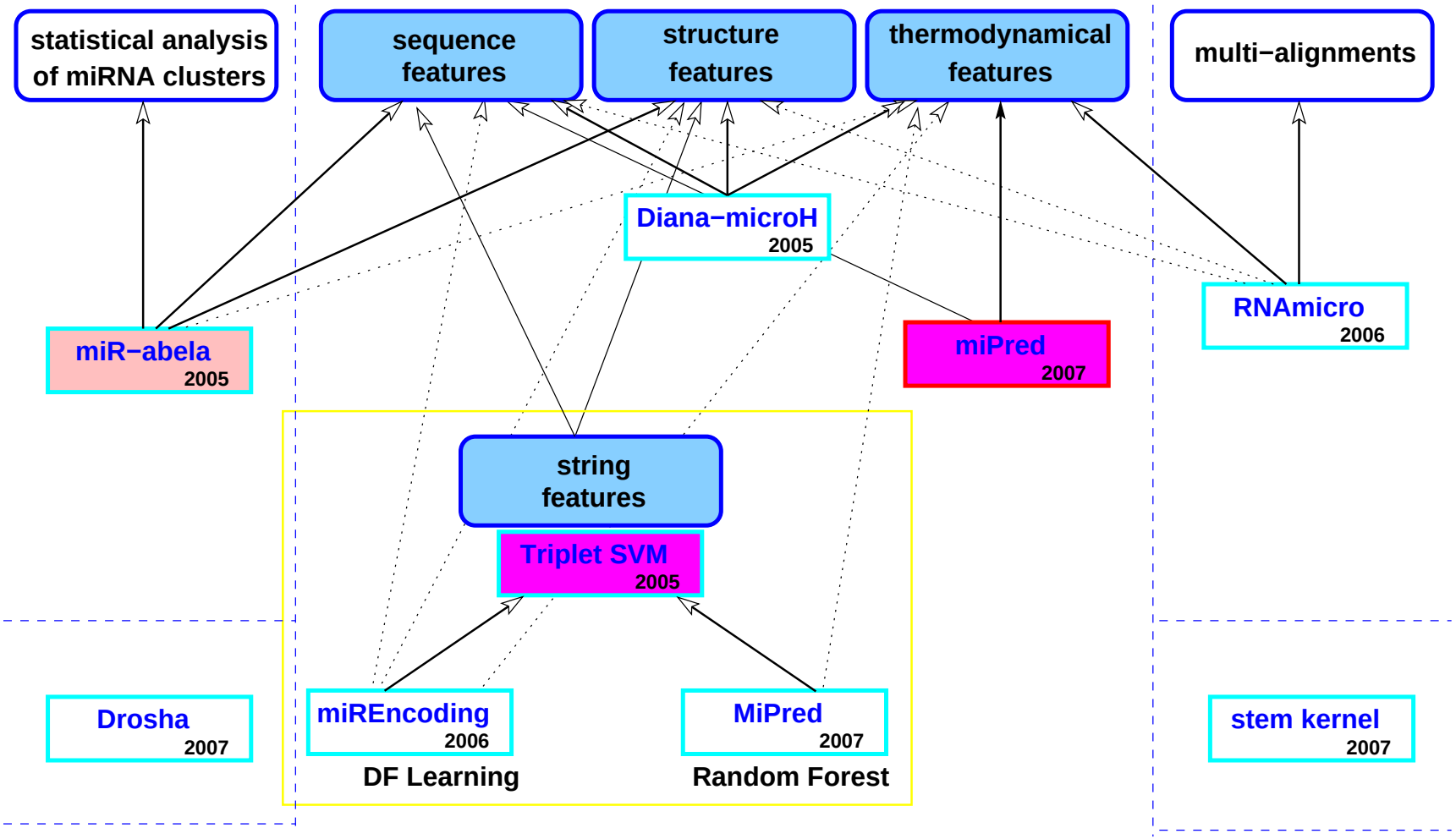
- the number of variables in the random subset at each node ($m_{\boldsymbol{try}}$)
- the number of trees in the forest ($n_{\boldsymbol{tree}}$).

This somehow counter-intuitive strategy is robust against overfitting, and it compares well to other machine learning techniques (SVMs, neural networks, discriminat analysis etc).

4. 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

An overview of SVMs for miRNA identification



4.1 miR-*abela* SVM

[Sewer et al., 2005]

Types of features:

(16) features over the entire hairpin structure

(10) features over the longest "symmetrical" region of the stem, i.e. the longest region without any asymmetrical loops

(11) features over the relaxed symmetrical region, i.e. the longest region in which the difference between the 3' and 5' component of assymetrical loops is not larger than Δl, a parameter

(3) features over all windows of lengths equal to l_m, the (assumed) length of mature miRNA; l_m is the second parameter used for tuning the miR-*abela* classifier.

Features over

the entire hairpin structure	1	free energy of folding
	2	length of the longest simple stem
	3	length of the hairpin loop
	4	length of the longest perfect stem
	5	number of nucleotides in symmetrical loops
	6	number of nucleotides in asymmetrical loops
	7	average distance between internal loops
	8	average size of symmetrical loops
	9	average size of asymmetrical loops
	10-13	proportion of A/C/G/U nucleotides in the stem
	14-16	proportion of A-U/C-G/G-U base pairs in the stem
the longest "symmetrical" region of the stem	17	length
	18	distance from the hairpin loop
	19	number of nucleotides in internal loops
	20-23	proportion of A/C/G/U nucleotides
	24-26	proportion of A-U/C-G/G-U base pairs
the relaxed symmetrical region	27	length
	28	distance from the hairpin loop
	29	number of nucleotides in symmetrical internal loops
	30	number of nucleotides in asymmetrical internal loops
	31-34	proportion of A/C/G/U nucleotides
	35-37	proportion of A-U/C-G/G-U base pairs
all windows of lengths l_m, the (assumed) length of mature miRNA	38	maximum number of base pairs
	39	minimum number of nucleotides in asymmetrical loops
	40	minimum asymmetry over the internal loops in this region

miR-*abela*: Performances

miR-*abela* was trained on 178 human pre-miRNAs as positive examples and 5395 randomly chosen sequences (from genomic regions, tRNA, rRNA and mRNA) as negative examples.

miR-*abela*'s output on 8 human pathogenic viruses was validated via laboratory investigations:

– out of 32 pre-miRNA predictions made by miR-*abela*, 13 were confirmed by the cloning study.

– similarly, 68 out of 260 predictions of new pre-miRNAs made by miR-*abela* were experimentally confirmed for the human, mouse and rat genomes.

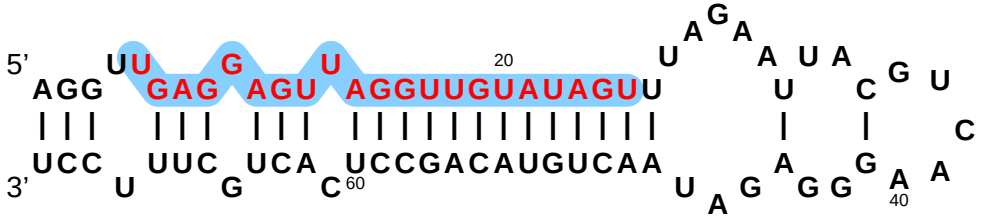
Note: In order to guide the experimental work, the miR-*abela*'s authors have developed a statistical model for estimating the number of pre-miRNAs in a given genomic sequence, using the scores assigned by miR-*abela* SVM to the "robust" candidate pre-miRNAs found in that region.

4.2 Triplet-SVM

[Xue et al, 2005]

Uses string features that combine first and second level structure informations on 3-mers.

Example: hsa-let-7a-2




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

There are seven 3-mers for which *i.* the middle position is occupied by the nucleotide G, and *ii.* all three positions are paired. Therefore the feature Gppp, which represents the pattern $\begin{bmatrix} \texttt{G} \\ \texttt{ppp} \end{bmatrix}$, will have the value 7.

Triplet-SVM: Performances

Triplet-SVM was trained on human pre-miRNAs from the miRNA Registry database [Griffiths-Jones, 2004] and pseudo pre-miRNAs from the NCBI RefSeq database [Pruitt & Maglott, 2001].

It achieved

– around 90% accuracy in distinguishing real from pseudo pre-miRNA hairpins in the human genome and

– up to 90% precision in identifying pre-miRNAS form other 11 species, including *C. briggsae*, *C. elegans*, *D. pseudoobscura*, *D. melanogaster*, *Oryza sativa*, *A. thaliana* and the *Epstein Barr* virus.

Note: *Pseudo pre-miRNA* hairpins are defined as RNA hairpins whose stem length and minimum free energy are in the range of those exhibited by the genuine, miRNAs.

Triplet-SVM: Training dataset

- **TR-C**

  +: 163 pre-miRNAs, randomly selected from the 193 human pre-miRNAs in miRBase 5.0 (207-193: multiple loops)

  −: 168 pseudo pre-miRNAs, randomly selected from those 8494 in the CODING dataset (see next slide)

Constructing the CODING dataset

1. extract protein coding sequences (CDSs) from those human genes registered in the RefSeq database that have no known alternative splice events
2. join these CDSs together and extract non-overlaping segments, keeping the distribution of their length identical to that of human pre-miRNAs
3. use the RNAfold program to predict from the RNA Vienna package to predict the secondary structure of the previously extracted segments
4. criteria for selecting pseudo pre-miRNAs:
   - minimum 18 base pairings on the stem (including GU wobble pairs);
   - maximum -18 kcal/mol free energy;
   - no multiple loops.

Triplet-SVM: Test datasets

- **TE-C**
  (+) 30 (193-163) human pre-miRNAs from miRBase 5.0: 93.3% acc.
  (−) 1000 pseudo pre-miRNAs, randomly selected from the 8494-168 in the CODING dataset: 88.1% acc., 93.3% sensitivity, 88.1% specificity
- **UPDATED**
  (+) 39 human pre-miRNAs, newly reported when Triplet-SVM was completed: 92.3% acc./sensit.
- **CROSS-SPECIES**
  (+) 581 pre-miRNAs from 11 species (excluding all those homologous to human pre-miRNAs): 90.9% acc./sensit.
- **CONSERVED-HAIRPIN**
  (−) 2444 pseudo pre-miRNAs on the human chromosome 19, between positions 56,000,001 and 57,000,001 (includes 3 pre-miRNAs): 89.0% acc./spec.

Two refinements of Triplet-SVM

miREncoding SVM [Zheng et al, 2006]

- added 11 (*global*) features:

–GC content,
–sequence length,
–length basepair ratio,
–number of paired bases,
–central loop length,
–symmetric difference
(i.e. the difference of length of the two arms)
–number of bulges,
–(average) bulge size,
–number of tails,
–(average) tail size,
–free energy per nucleotide.

- tried to improve the classification performance using the DFL feature selection algorithm to determine the *essential attributes*.

MiPred SVM [Jiang et al, 2007]

- added 2 *thermodynamical features*:

–MFE,
–P-value.

- replaced the SVM with the Random Forests ensamble learning algorithm.

- achieved nearly 10% greater overall accuracy compared to Triplet-SVM on a new test dataset.

miREncoding SVM

Trained and tested on the same datasets as Triplet-SVM, miREncoding obtained an overall 4% accuracy gain over Triplet-SVM, and reported a specificity of 93.3% at 92% sensitivity.

The miREncoding's authors showed that

using only four most "essential" features determined with the DFL algorithm, namely

Appp, G.pp, the length basepair ratio, and the energy per nucleotide,

the classification results obtained with the C4.5, kNN and RIPPER algorithms are significantly improved.

However, in general miREncoding SVM performs better when using all attributes.

In several cases, the performances of C4.5, kNN and RIPPER on the essential (DFL-selected) feature set are better than those obtained by the SVM on the full feature set.

MiPred: Datasets

Training:

- **TR-C (same as Triplet-SVM)**

  RF (Out Of Bag estimation):
  96.68% acc., 95.09% sensitivity, 98.21% specificity

Test:

- **(+) 263 (426-163) from miRBase 8.2 (462-426 pre-miRNAs with multiple loops)**
  (−) 265 pseudo pre-miRNAs randomly chosen from those 8494 in the CODING dataset (see Triplet-SVM, the TR-C training data set)

  RF vs Triplet-SVM:
  91.29% vs 83.90% acc., 89.35% vs 79.47% se., 93.21% vs 88.30% sp.
- **(+) 41 pre-miRNAs from miRBase 9.1 \ miRBase 8.2**
  100% acc. (vs 46.34% of miR-*abela*)

4.3 Microprocessor & miRNA SVM

[Helvik et al, 2007]

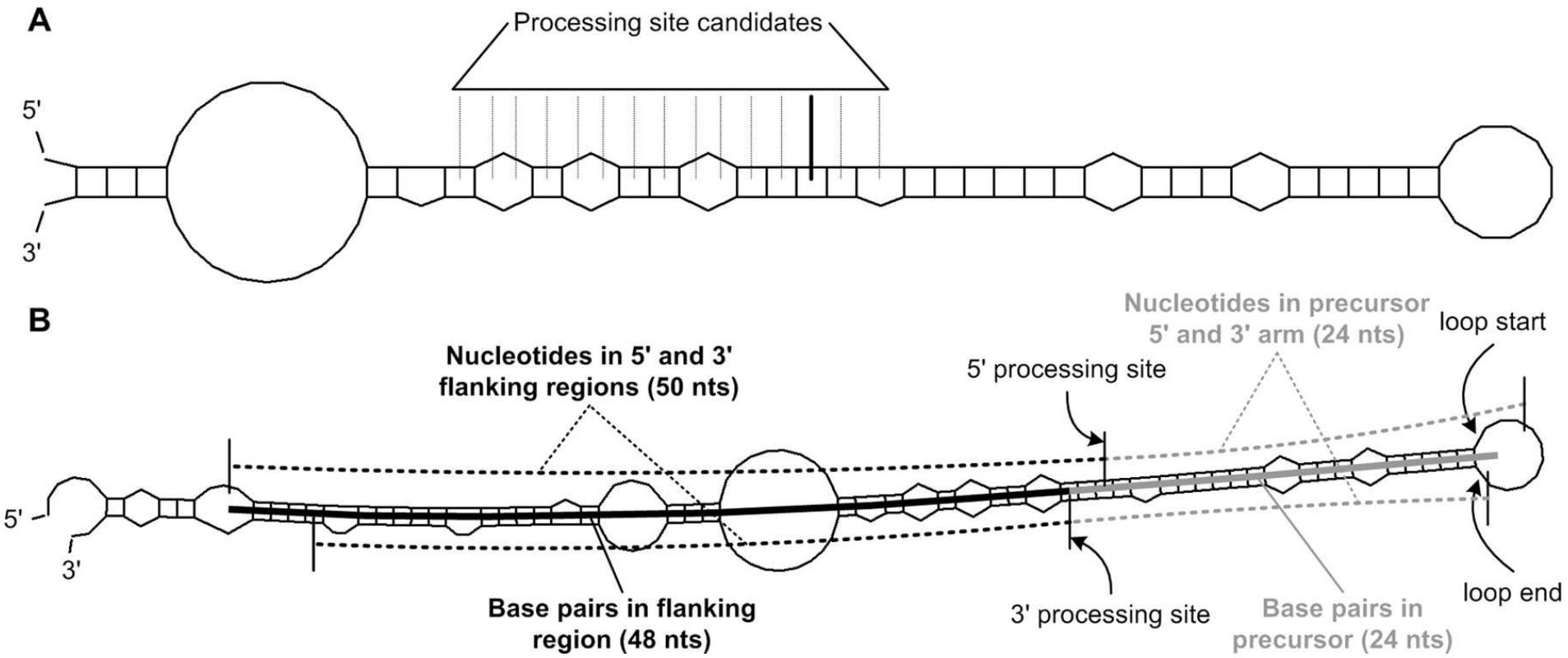
Microprocessor SVM:

designed for the recognition of **Drosha cutting sites** on sequences that are presumed to extend pre-miRNA sequences.

For a given hairpin, Microprocessor SVM proposes a bunch of **processing site candidates** for the Drosha processor. For each candidate site, a high number of **features** (242) are computed. These features register local, (including very low-level) detailed informations on the regions up (24nt) and down (50nt) the candidate site.

Trained on miRNAs from miRBase 8.0, and **tested** via 10-fold cross validation, Microprocessor SVM successfully identified 50% of the Drosha processing sites. Moreover, in 90% of the cases, the positions predicted by Microprocessor SVM are within 2nt of the true site.

A human pre-miRNA sequence (hsa-mir-23a), extended with the flanking regions processed by Microprocessor SVM



Acknowledgement: From [Helvik, Snove, and Saetrom, 2007].

miRNA SVM:

designed for the identification of pre-miRNAs.

Features:

– the features of the best predicted Drosha cutting site among those computed by Microprocessor SVM, and

– seven other features that gather statistics on all Drosha candidate sites considered by Microprocessor SVM for that pre-miRNA.

Training was made on pre-miRNAs from miRBase 8.0 plus 3000 random genomic hairpins.

Tests done via cross-validation made the authors conclude that

– its performance is close to those of other miRNA classification systems (Triplet-SVM, miR-*abela*, and ProMiR [Nam, 2005]);

– in general, the validation of newly proposed (extended) pre-miRNAs should include a check on whether they exhibit or not Drosha cutting sites. Indeed, their work pointed to several entries that seem to have been mistakenly added to the miRBase repository.

microprocessor & miRNA SVM features:

1	precursor length
2	loop size
3	distance from the 5' processing site to the loop start
4	(48x4) nucleotide occurrences at each position in the 24nt regions of the precursor 5' and 3' arms
5	(24) base-pair information of each nucleotide for the 24nt at the precursor base
6	(4) nucleotide frequencies in the two regions in feat. 4
7	number of base pairs in feat. 5
8	(100x4) nucleotide occurrences at each position in the 50nt 5' and 3' flanking regions
9	(48) base-pair information of each nucleotide for the 48nt in the flanking region outside the precursor
10	(4) nucleotide frequencies in the two regions in feat. 8
11	number of base pairs for the 15nt immediately flanking the precursor
12	number of base pairs in the region in feat. 9
13	number of potential processing sites
14	score of the best processing site
15	average score for all potential processing sites
16	standard deviation for all potential processing sites
17	difference between feat. 14 and 15
18	distance between the three top scoring processing sites
19	number of local maximums in the processing site score distribution

Explaining some terms used in the previous feature list:

candidate Drosha processing site:

the 5' end of a 50-80nt sequence centered around a stem loop; (the 3' end is determined by a 2nt overhang wrt the 5' end)

position specific base-pair information (BP_x):

BP_x is 0, 0.5, or 1 if respectively none, one, or both of the nucleotides on the position x upstream of the 5' processing site and $x-2$ downstream of the 3' processing site are base-paired with a nucleotide in the opposite strand

4.4 RNAmicro SVM

[Hertel & Stadler, 2006]

RNAmicro was constructed with the aim to find those miRNAs that have conserved sequence and secondary structures.

Therefore it works on alignments instead of sequences, as the other SVMs here presented do.

```
human  a a G A C u u c g G a U C u G G c G a c a C C C
mouse  u a C A C u u c g G a U G a C A c C a a a G U G
 worm  a g G U C u u c g G c A C g G G c A c c a U U C
  fly  c c A A C u u c g G a U U u U G c U a c c A U A
  orc  a a G C C u u c g G a G C g G G c G u a a C U C
struc. . . ( ( ( . . . . ) . ) ) . ( ( . ( . . . ) ) )
```

RNAmicro: Datasets

The positive examples on which RNAmicro was trained were 295 alignments that have been built starting from the miRNA registry 6.0, using homologous sequences.

The negative examples were first generated from the positive alignments by doing shuffling until the consensus structure yielded a hairpin structure; 483 alignments of tRNAs were further added to the set of negative examples.

RNAz [Washietl et al, 2005] is an SVM-based system that identifies noncodant RNAs using multiple alignments.

RNAmicro was tested by applying it as a further filter to the output provided by RNAz for several genome-wide surveys, including *C. elegans*, *C. intestinalis*, and *H. sapiens*.

RNAmicro: Features

1. **the stem length for the miRNA candidate alignment**
2. **the loop length**
3. **the G+C content**
4. **$\overline{MFE}$, the mean of the minimum folding energy *MFE***
5. **the mean of the z-scores,**
6. **the mean of the *adjusted MFE*,**
7. **the mean of *MFE* index 1,**
8. **the *structure conservation index*, defined as the ratio of $\overline{MFE}$ and the energy of the consensus secondary structure.**

9-11. **the average *column-wise entropy* for the 5' and 3' sides of the stem and also for the the loop; it is defined as**

$$S(\xi) = -\frac{1}{len(\xi)} \Sigma_{i \in \xi} \, \Sigma_{\alpha} \, p_{i,\alpha} \, ln \, p_{i,\alpha}$$

where $p_{i,\alpha}$ is the frequency of the nucleotide α (one of A, C, G, U) at the sequence position i

12. **S_{min}, the minimum of the column-wise entropy computed (as above) for 23nt windows on the stem**

4.5 miPred SVM
[Ng & Mishra, 2007]

Features:

- dinucleotide frequencies (16 features)
- G+C ratio
- folding features (6 features):
  - dG – adjusted MFE
  - MFEI_1 – MFE index 1 (see [Zhang et al., 2006])
  - MFEI_2 – MFE index 2
  - dQ – adjusted Shannon entropy
  - dD – adjusted base-pair distance (see [Freyhult et al., 2005])
  - dP – adjusted base pairing propensity (see Schultes at al., 1999)
- dF – a topological descriptor: the degree of compactness (see [Fera et al., 2004], [Gran et al., 2004])
- zG, zP, zD, zQ, zF: normalized versions of dG, dP, dD, dQ, dF respectively, just as the Z-score is a normalized version of MFE (5 features).

miPred: Training datasets

- **TR-H**

  +: 200 human pre-miRNAs from miRBase 8.2

  −: 400 pseudo pre-miRNAs randomly selected from the CODING dataset

  Results:

  accuracy at 5-fold cross-validation: 93.5%

  area under the ROC curve: 0.9833.

miPred: Test datasets

- **TE-H**
  (+) 123 (323-200) human pre-miRNAs from miRBase 8.2
  (−) 246 pseudo pre-miRNAs randomly chosen from those 8494 in the CODING dataset (see Triplet-SVM, the TR-C training data set)

  93.50% acc., 84.55% sensitivity, 97.97% specificity
  (Triplet-SVM: 87.96% acc., 73.15% sensitivity, 93.57% specificity)
- **IE-NH**
  (+) 1918 pre-miRNAs from 40 non-human species from miRBase 8.2
  (−) 3836 pseudo pre-miRNAs

  95.64% acc., 92.08% sensitivity, 97.42% specificity
  (Triplet-SVM: 86.15% acc., 86.15% sensitivity, 96.27% specificity)
- **IE-NC:**
  (−) 12387 ncRNAs from the Rfam 7.0 database: 68.68% specificity (Triplet-SVM: 78.37%)
- **IE-M:**
  (−) 31 mRNAs from GenBank: 27/31 specificity (Triplet-SVM: 0%)

Remark: On four complete viral genomes — ***E.Barr virus***, ***K.sarcoma-associated herpesvirus***, ***M.γ-herpesvirus 68 strain WUMS*** and ***H.cytomegalovirus strain AD169*** — and seven other full genomes, miPred's sensitivity is 100%(!) while its specificity is >93.75%.

Remark: Empirically it is shown that six features ensure most of miPred's discriminative power: $\mathbf{MFEI}_1$, zG, dP, zP, zQ, dG.

4.6 Other SVMs for miRNA prediction

DIANA-microH [Szafranski et al, 2006]

Features:

- the minimum free energy,
- the number of based pairs,
- central loop length,
- GC content,
- the *stem linearity*,
  defined as the largest possible section of the stem subregion that is likely to form a mostly double-stranded conformation
- the *arm conservation*,
  an evolutionary based feature, computed using human vs. rat or human vs. mouse sequence comparisons.

Trained on the miRNAs from the human from miRBase as positive examples, and pseudo-hairpins from the RefSeq database as negative examples, the authors claimed a 98.6% accuracy on a test set made of 45+ and 243− hairpins.

5. Research directions / Future work

- Test strategies for automatic learning of kernel functions to be used in connection with SVMs here presented.
  - In particular, test InfoBoosted GP [Gîrdea and Ciortuz, 2007] on Triplet-SVM (and its extensions), miR-*abela* and miPred.
- Find out (meta-)learning algorithms (other than RF) capable of better results than SVMs, and test them on the miRNA identification task.

  See for instance MDO, the Margin Distribution Optimisation algorithm ([Sebe et al, 2006], ch. 3 and 6) that has been proved to perform better than both Boosting and SVM on certain UCI data sets.
  - In particular, test RF on the feature sets specific to other SVMs (than MiPred) here presented.
- Explore different feature selection algorithms that would eventually work well in connection with SVMs (see [Chen and Lin, 2004]).
  - In particular, test the effect of DFL algorithm on feature sets of the SVMs here presented (other than miREncoding).

Research directions / Future work (Cont'd)

- Verify the claim of [Helvik et al, 2006] that identifying the Drosha cutting site (the output of Microprocessor SVM) significantly improves the quality of the SVMs for miRNA identification.
  - Apply DFL (and/or other feature selection algorithms) on Microprocessor SVM's feature.
- Make a direct comparison of the as many as possible of the mirRNA identification SVMs on up-to-date data sets (derived from miRBase).
- See whether features on randomised sequences could be replaced with other features withouit loss of classification performance. (This would be most interesting for miPred.)
- Make the connection with the problem of identifying miRNA target sites or other classification problems for non-coding RNAs.

Student Projects (2008)