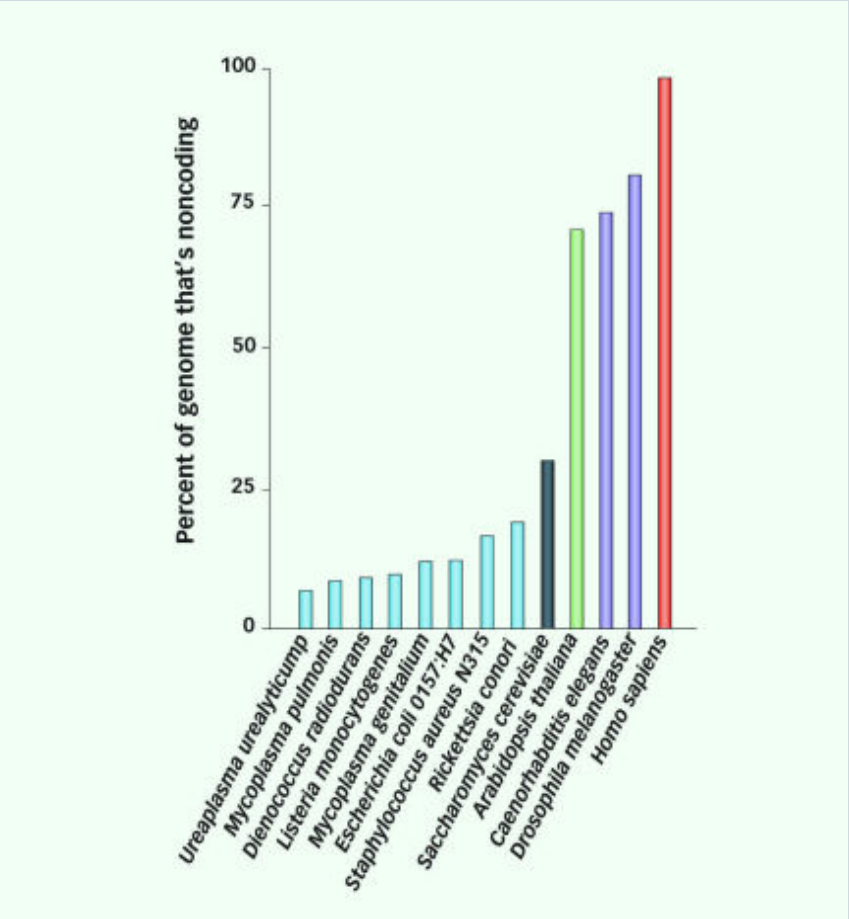
Lecture Thu 10.11.

HIDDEN MARKOV MODELS & RNA GENES

Non-coding DNA

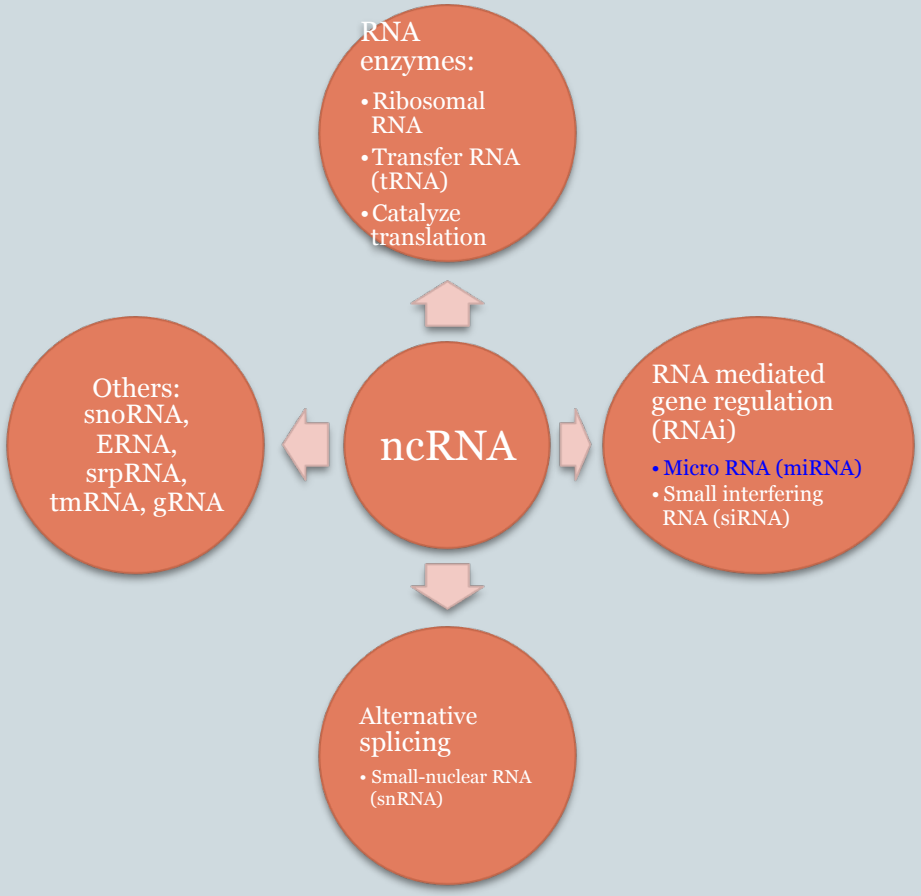
- Non-coding DNA include all segments of the genome that does not get translated to proteins
- In higher organisms, most of the DNA is non-coding
  - In humans, over 98% of the genome is non-coding



Types of non-coding DNA

Noncoding functional RNA, RNA genes	•Functional RNA molecules that are not translated into protein.
Introns	•Regions inside the coding region that are not transcribed into mRNA •Common in higher organisms
Regulatory elements	•Binding sites of special proteins called transcription factors •Typically within in the promotor region of the gene or within the introns •Carry important function
Pseudogenes	•Genes that have lost their protein coding ability •Thought to be non-functional
Repeat sequences	•Simple repeats, CpG islands •DNA satellites •Mobile sequences (transposons) •Possible role in epigenetics
'Junk DNA'	•DNA with no function •Open question: How much of that is there?

Functions of non-coding RNA



Enabling technology: whole transcriptome profiling

- Modern high-throughput measurement technology allows one to observe all expressed genomic content at the same time
  - Tiling arrays are based on microarray technology (binding of mRNA to pre-designed probes)
  - RNA-seq is based on sequencing the transcriptome
- Both give an unbiased view to the transcriptome, hence useful for ncRNA studies

Enabling technologies: ncRNA databases

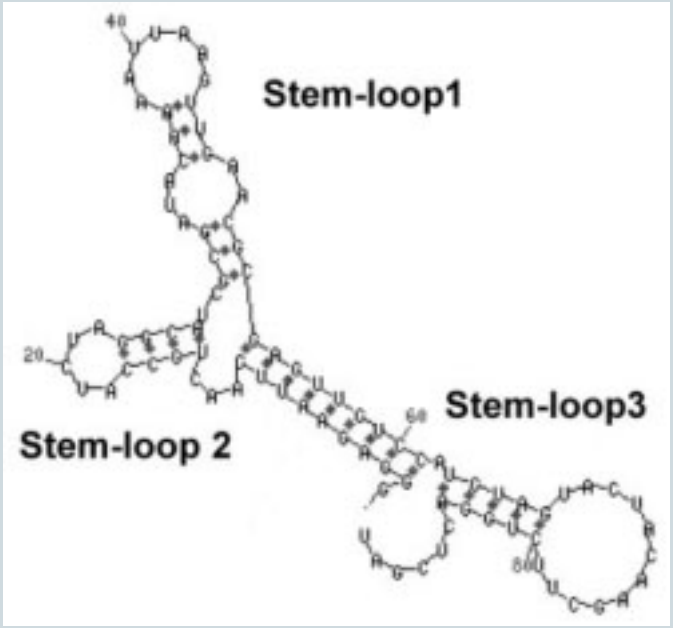
- Online databases that integrate and store data from different studies are a key piece of bioinformatics infrastructure
  - Given a candidate ncRNA gene a database search may reveal its function or at least clues of it
- Development of bioinformatics methods and tools also relies on curated datasets

Discovering the function of RNA genes

- For RNA genes we can profile their expression
  - Tiling arrays, RNA-seq
- However expression of a sequence does not directly reveal its function
  - Differential expression studies can reveal association of the expression to phenotypes
  - Indirect evidence can be obtained from functions of co-expressed sequences
  - Danger of “guilt by association”
- If a homologous sequence with known function can be found, one can transfer the annotation
  - This is easier with protein coding RNA as the databases are more comprehensive

From sequence…to structure… to function

- 3D structure of RNA is thought to determine the function
  - hard to predict from sequence
  - prediction of secondary structure (local loops) as a intermediate problem
- RNA secodary structure prediction is a well-established field of bioinformatics



Predicted RNA secondary structures

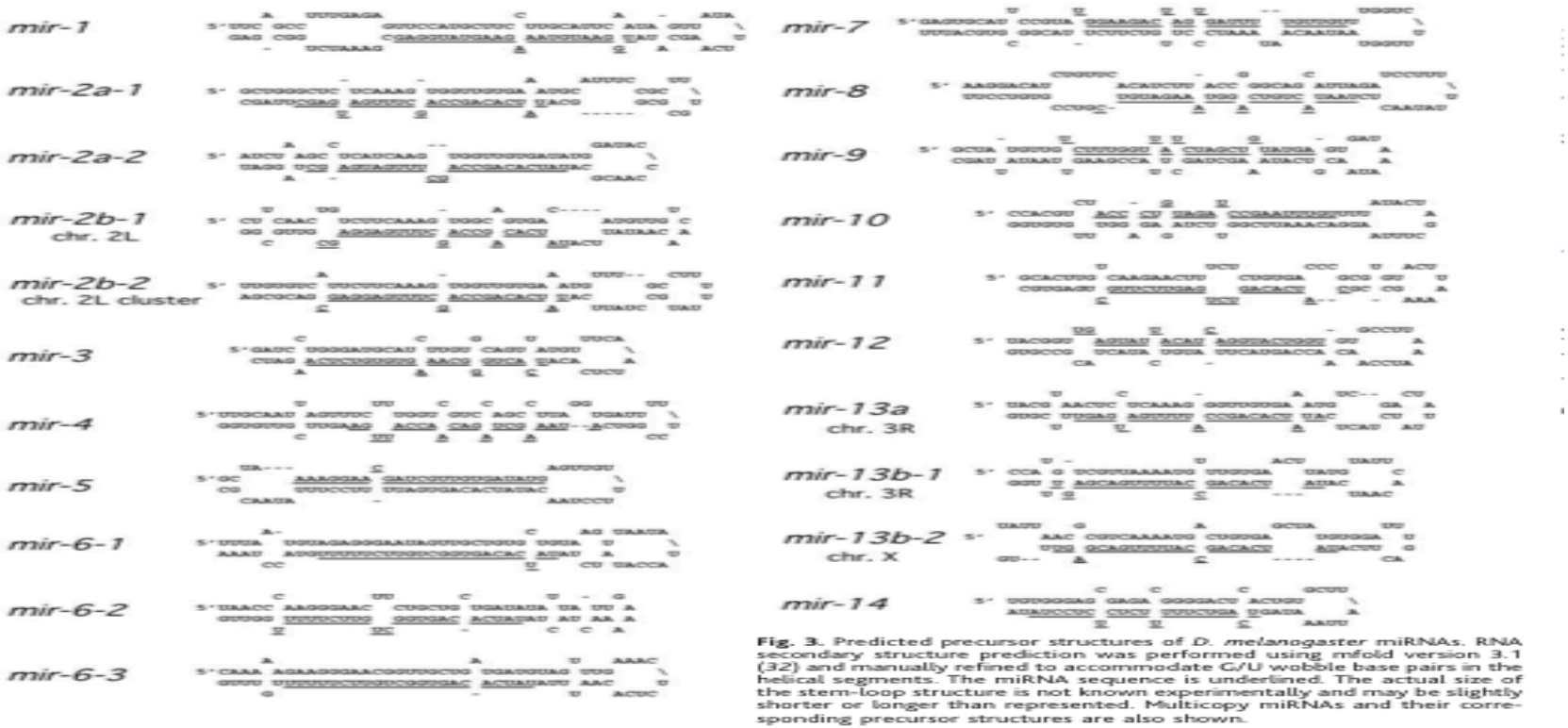
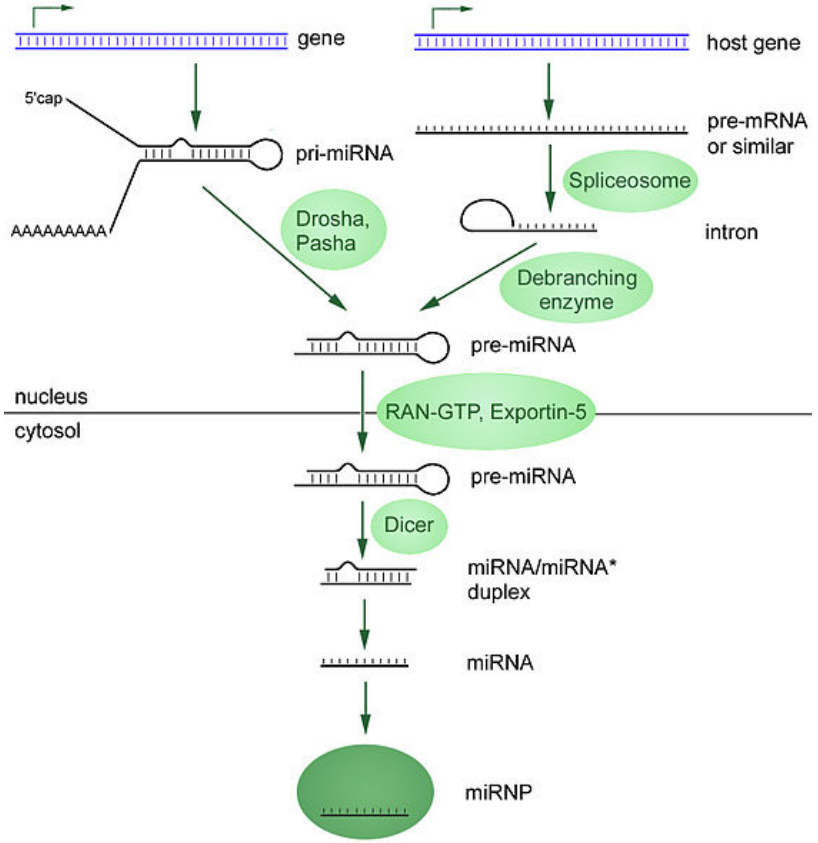


Fig. 3. Predicted precursor structures of *D. melanogaster* miRNAs. RNA secondary structure prediction was performed using mfold version 3.1 (32) and manually refined to accommodate G/U wobble base pairs in the helical segments. The miRNA sequence is underlined. The actual size of the stem-loop structure is not known experimentally and may be slightly shorter or longer than represented. Multicopy miRNAs and their corresponding precursor structures are also shown.

Identification of Novel Genes Coding for Small Expressed RNAs
Mariana Lagos-Quintana, *et al. Science* 294, 853 (2001);

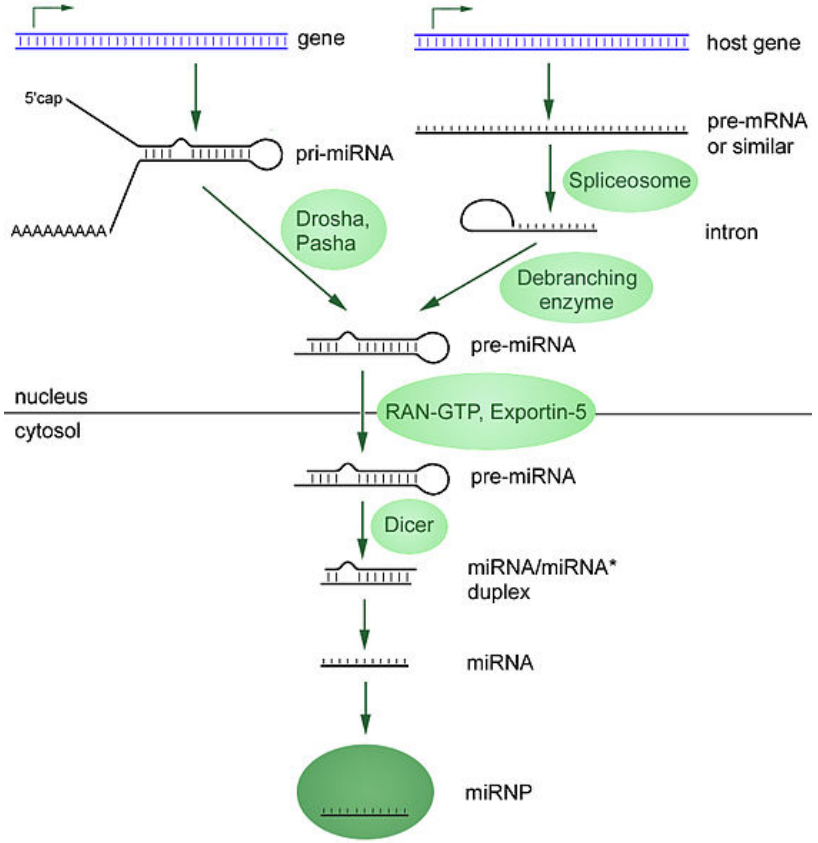
Modelling MicroRNA

- A **microRNA** (**miRNA**) is a short RNA molecule (avg 22 nt)
- miRNAs bind to complementary sequences on mRNA
- Usually results in repression of translation



Modelling MicroRNA

- MicroRNA animation
- Two problems of interest
  - microRNA gene finding - locate microRNA genes from the genome
  - microRNA target prediction
- Many techniques exist, we will look how HMMs can help
  - Pair HMM

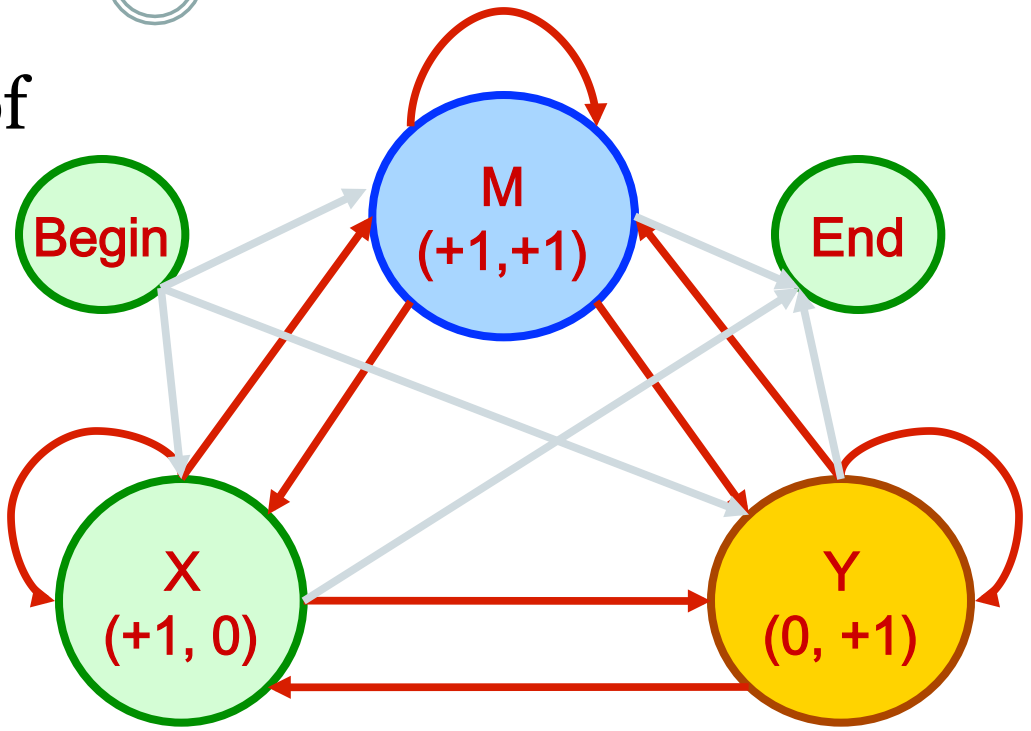


Hidden Markov Models for Sequence Alignment

- So far, we have used HMMs to detect certain regions from single a sequence
- HMMs can also be used for sequence alignment tasks
  - Pair-HMM can be used to find high-scoring alignments between two sequences, allowing gaps
  - Profile-HMM can be used to model a multiple alignment of a set of sequences
- Probabilistic alternative to combinatorial pattern matching algorithms (e.g. edit distance minimization)

Pair HMM

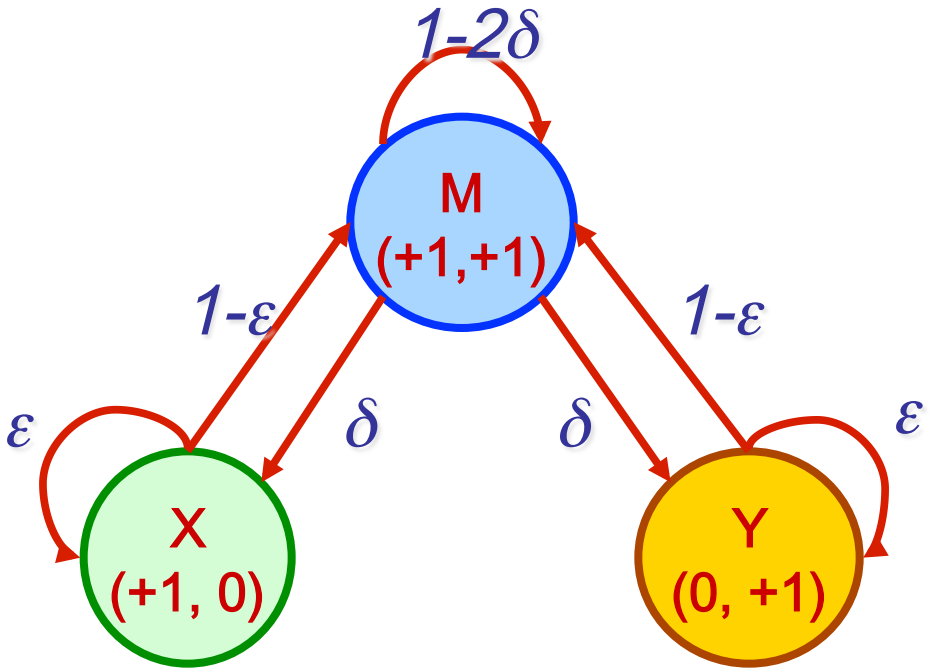
- Pair HMM consist of
  - Begin and End state which do not emit symbols
  - Three normal states
    - M (match)
    - X (gap in Y)
    - Y (gap in X)




```
X TAG-CTATCAC--GACCGC-GGTCGATTTGCCCGACC
Y -AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
  XMMYMMMMMMMYYMMMMMMYMMMMMMMXXMMMMMXXX
```

Pair HMM - Transitions

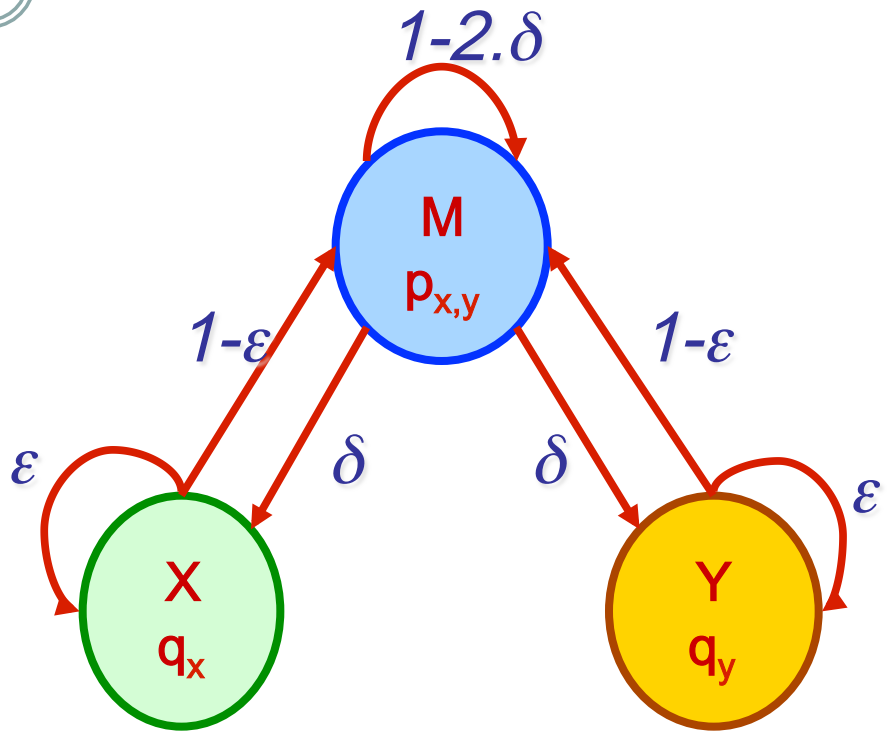
- Transition from M to X (resp. Y) opens a gap in Y (resp. X), transition back to M closes the gap
  - δ ~ open gap probability
  - ε ~ extend gap probability




```
X TAG-CTATCAC--GACCGC-GGTCGATTTGCCCGACC
Y -AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
  XMMYMMMMMMMYYMMMMMMYMMMMMMMXXMMMMMXXX
```

Pair HMM - Emissions

- State M: emit (b,b') with probability $e_M(b,b')$
- State X: emit (b,-) against a gap with probability $e_x(b)$
- State Y: emit (-,b') with probability $e_Y(b')$




```
X TAG-CTATCAC--GACCGC-GGTCGATTTGCCCGACC
Y -AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
  XMMYMMMMMMMYYMMMMMMYMMMMMMMXXMMMMMXXX
```

Pair HMMs – Finding Optimal Alignment

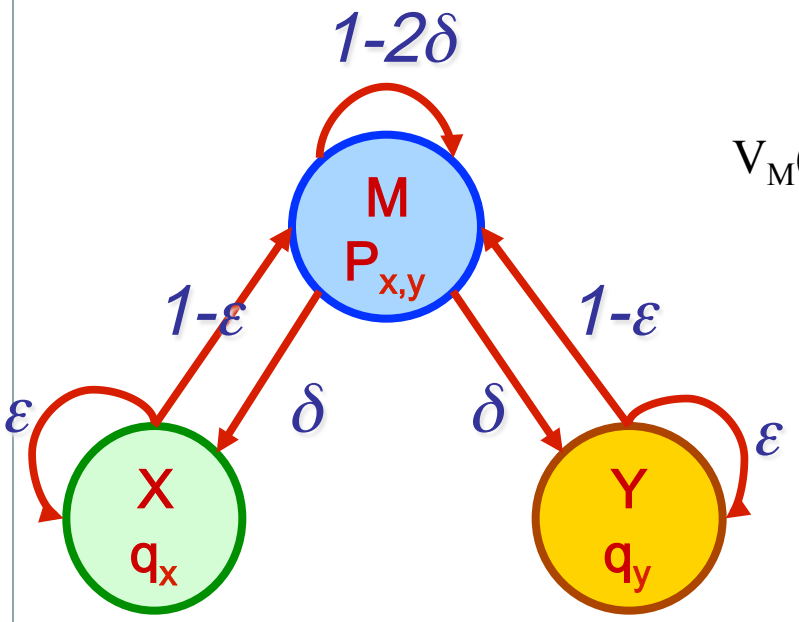
- A state sequence π from begin to end state that emits x and y gives an alignment for them
  - Transition and emission probabilites give the probability of the alignment
- The best alignment of two sequences corresponds to the most probable state sequence

$$\pi^* = \text{argmax}_\pi P(x, y, \pi)$$

- Can be computed by the Viterbi algorithm

```
X TAG-CTATCAC--GACCGC-GGTCGATTTGCCCGACC
Y -AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
  XMMYMMMMMMMMYYMMMMMMMYMMMMMMMXXMMMMMXXX
```

Viterbi for pair-HMMs



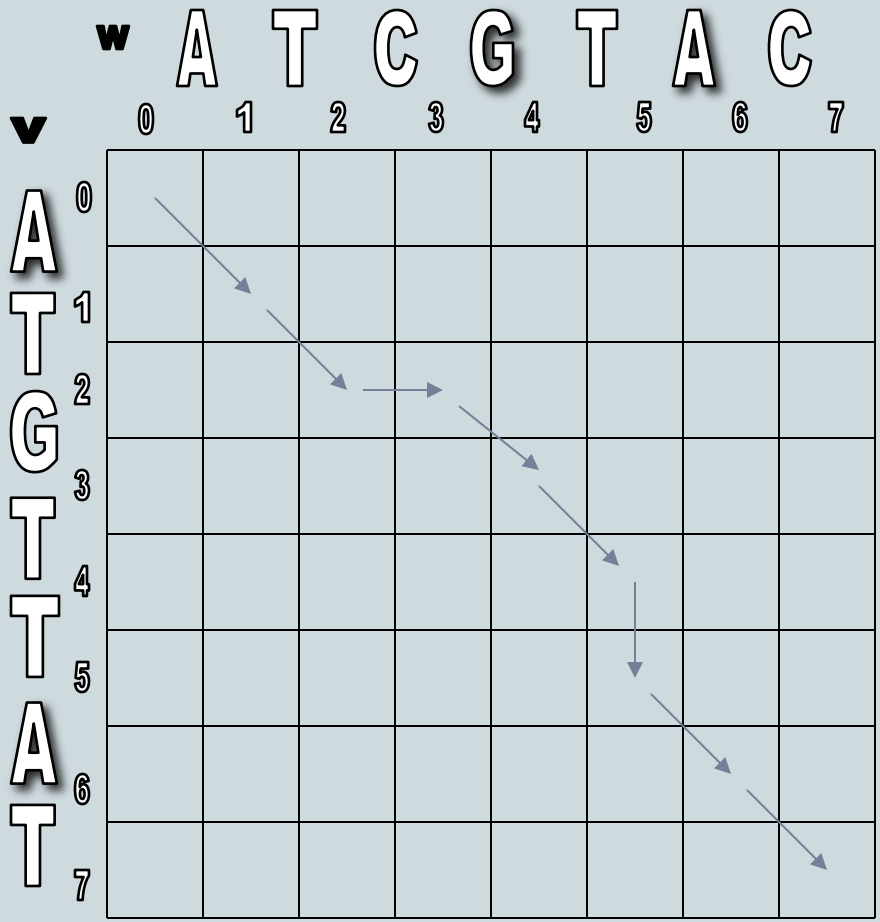
$$V_M(i, j) = e_M(x_i, y_j) \max \begin{cases} (1 - 2\delta)\, V_M(i - 1, j - 1) \\ (1 - \varepsilon)\, V_X(i - 1, j - 1) \\ (1 - \varepsilon)\, V_Y(i - 1, j - 1) \end{cases}$$

$$V_X(i, j) = e_X(x_i) \max \begin{cases} \delta\, V_M(i - 1, j) \\ \varepsilon\, V_X(i - 1, j) \end{cases}$$

$$V_Y(i, j) = e_Y(y_j) \max \begin{cases} \delta\, V_M(i, j - 1) \\ \varepsilon\, V_Y(i, j - 1) \end{cases}$$

```
X TAG-CTATCAC--GACCGC-GGTCGATTTGCCCGACC
Y -AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
  XMMYMMMMMMMYYMMMMMMMYMMMMMMMXXMMMMMXXX
```

Pair-HMM as sequence aligner

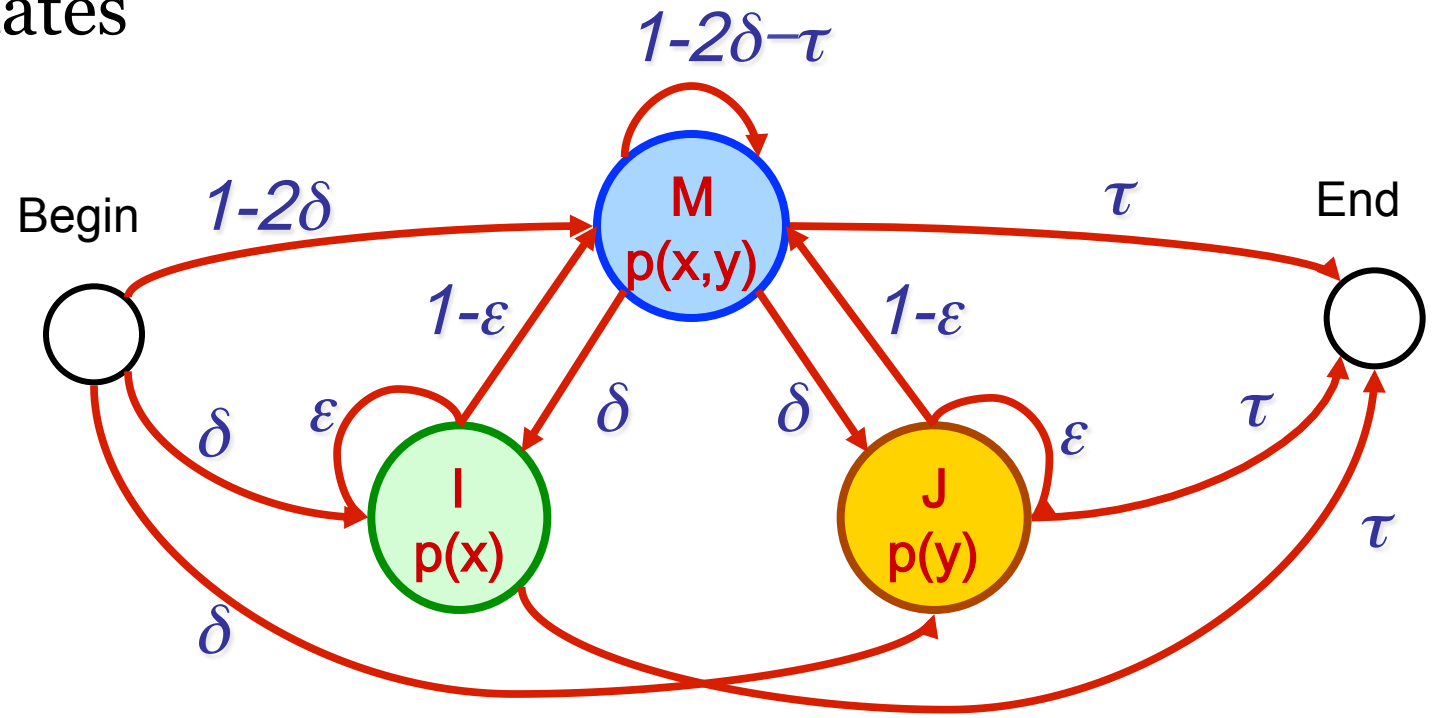


- pair HMM can be seen as an analogy to edit distance –based sequence alignment
- Instead of minimizing cost of edit operations (insert, delete, match) we maximize their probability

A T – G T T A T
A T C G T – A C
M M Y M M X M M

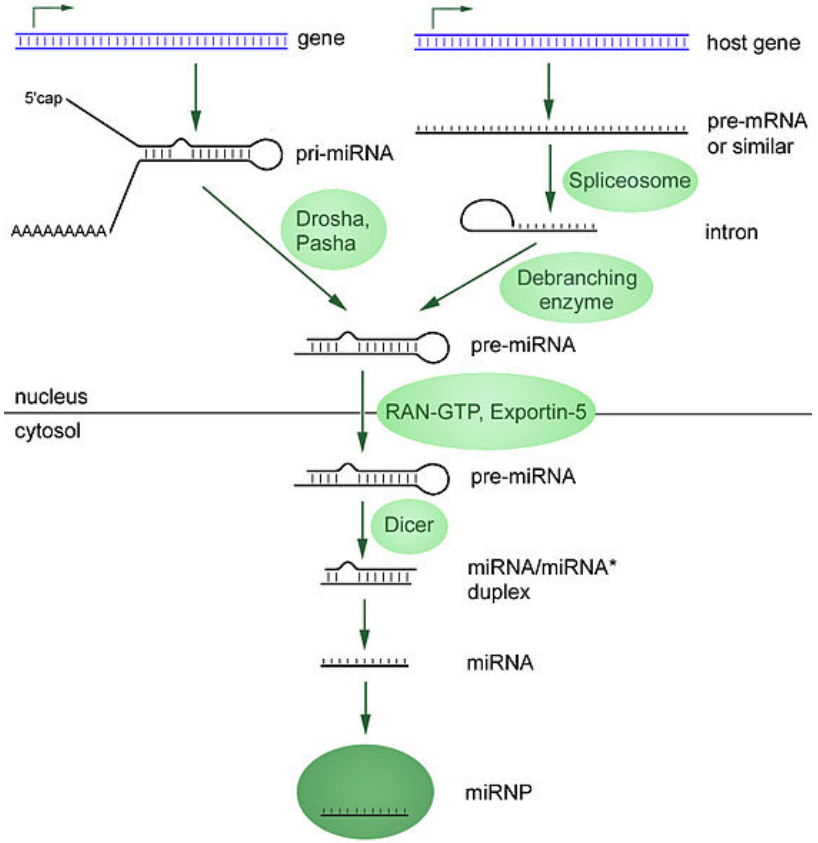
Full model

- The complete model should also contain the transitions between the begin, end and normal states



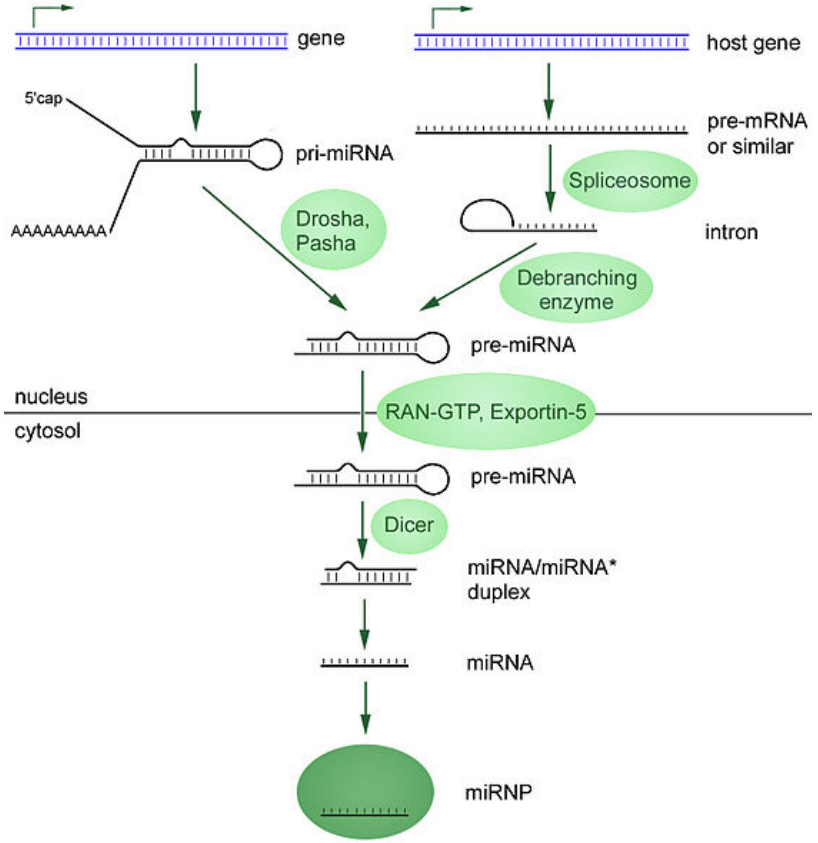
1st Problem: predicting microRNA genes

- Main approches for finding miRNA genes:
  - Via expression (RNA-seq); what if our gene is not higly expressed
  - Via homology to known miRNAs; but how to find new miRNA genes?
  - Ab initio prediction from sequence; how can we get accurate predictions?



Ab initio prediction of microRNA genes

- Challenges for HMMs
  - miRNA genes are short
  - no codon structure to help modelling
  - hard to make an accurate HMM based on that
- ProMIR system
  - Takes advantage of the secondary structure of the RNA



Nam, et al. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Research, 2005, 33 (11), 3570-3581

Pair HMM in ProMiR

- States: two components
  - Match (M), mismatch (U), insertion (I), deletion (D)
  - Inside (+) or outside the miRNA region (-)
  - Total of 8 states: M+,U+,I+,D+,M-,U-,I-,D-
- Emissions (“.” denotes gap):
  - A-U,U-A,G-C,C-G,*U-G,G-U* in match state
  - .-A,.-U,.-G,.-C in deletion state
  - A-.,U-.,G-.,C-. in insertion state
  - All other pairs can be emitted in mismatch state

Pair HMM modelling of miRNA

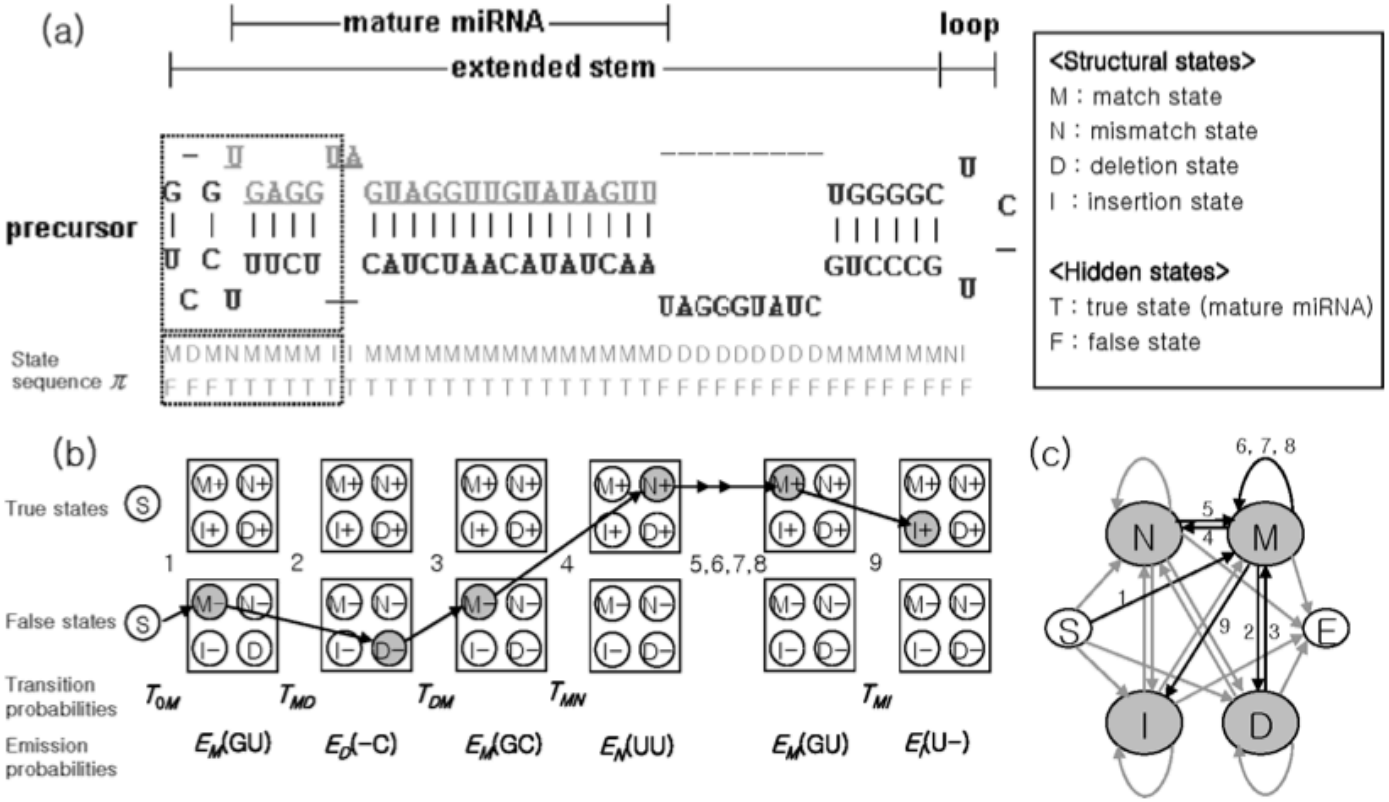


Figure 1. Pairwise representation of stem–loop structures and state sequences of pre-miRNAs, where the state of each pair includes structural information and mature miRNA region information (hidden states). (**a**) The structure of the pre-miRNA. (**b**) The transition and emission scheme of the structural states and the hidden states for pairwise sequence in the dotted rectangle shown in (a). T_{0M}, T_{DM}, T_{MN} and T_{MI} are transition probabilities. E_M(GU), E_D(−C), E_M(GC), E_N(UU), E_M(GU) and E_I(U−) are emission probabilities. (**c**) The four-state finite state automaton. Finally, the probability of the pairwise sequence is assigned by multiplication of the transition probabilities and the emission probabilities.

Training data for ProMiR

- Need to have a collection of RNA secondary structures from miRNA and other genes
- Positive data: 81 5’ strand, 55 3’ strand known human miRNAs
  - The true miRNA region will give the T/F labeling
- Negative data: 1000 extended stem–loop structures randomly extracted from human chromosomes
  - This is really *pseudo-negative* data: something that is likely to not to be a miRNA
- Stem–loop structures were predicted using the Vienna RNA software package

ProMIR Pipeline

a) Predict RNA extended stem-loop structures
b) Match to database of expressed sequence tags (EST)
c) pre-miRNA Scoring by pairwise-HMM
d) In silico verification:
   - free energy calculations (MFE)
   - negative evidence: BLAST match to known non-miRNA
   - presence of known conservation patterns

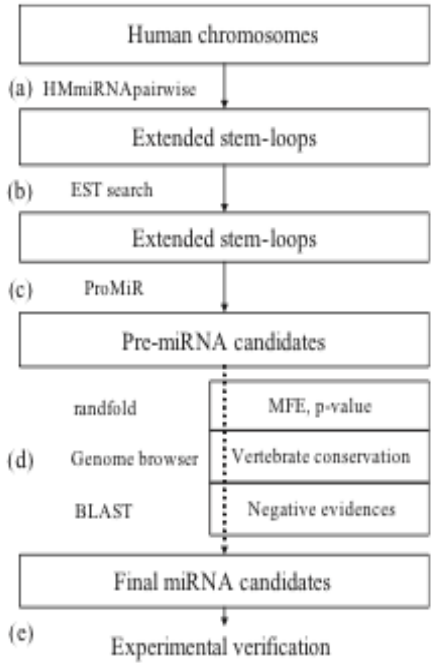
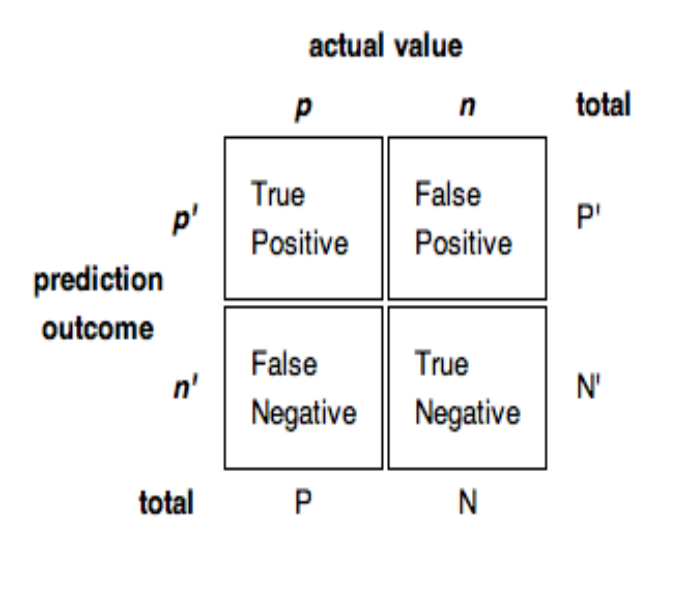


Figure 3. Flow chart for human miRNA gene finding. (a) The program, HMmiRNApairwise using an RNAfold algorithm extracts extended stem–loops with several criteria described in the Supplementary Material; (b) human EST database search; (c) ProMiR predicts pre-miRNA candidates, the region of mature miRNA and the location of a functional strand; (d) screening by additional evidence—MFE values, vertebrate conservations and negative evidences; (e) experimental verification.

Evaluation metrics cheat sheet

Confusion matrix



Evaluation metrics

- Accuracy: $ACC = (TP + TN) / (TP + TN + FP + FN)$
- Precision/Positive predictive value: $PPV = TP / (TP + FP)$
- Recall/Sensitivity/True positive rate: $TPR = TP / (TP + FN)$
- Specificity/True negative rate: $SPC = TN / (FP + TN)$
- False positive rate: $FPR = FP / (FP + TN)$
- False discovery rate: $FDR = FP / (FP + TP)$
- Negative predictive value: $NPV = TN/(TN + FN)$

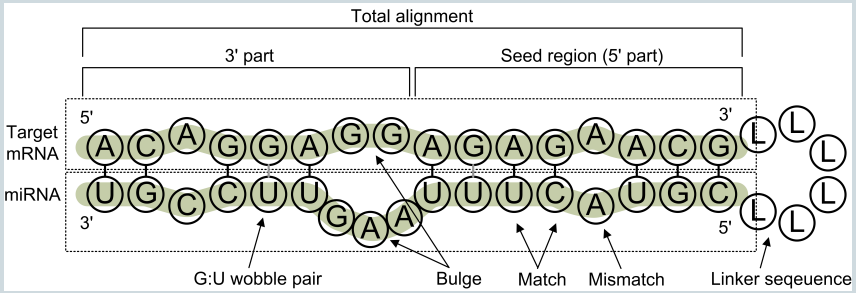
Prediction results for ProMiR

	Training data	Sensitivity	Specificity
HMMer	10	0.03	1.00
	30	0.00	0.00
	50	0.00	0.00
	68	0.00	0.00
INFERNAL	30	0.68 (0.00)[a]	0.50 (0.00)
	50	0.91 (0.00)	0.30 (0.00)
	68	0.94 (0.00)	0.18 (0.00)
Conservation[b]	68	0.34	0.87
esRCSG	50	0.36 (0.67)[c]	0.96 (0.89)
ProMiR	68	0.69	0.94
	5-fold cross validation	0.73	0.96

[a]Results by sequential and structural multiple alignment.

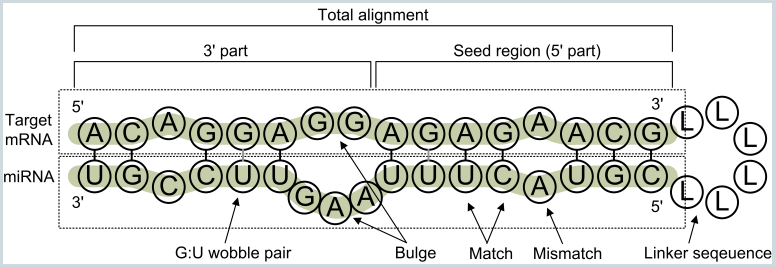
2nd problem: predicting miRNA targets

- miRNAs bind the mRNA transcripts to regulate (typically stop) their translation proteins
- The binding is established via complementary base pairing (A-U,C-G)
- The base pairing does not need to be perfect
  - Wobble pairing (Non-watson-crick base pairing)
  - Mismatches
  - Insertions



2nd problem: predicting miRNA targets

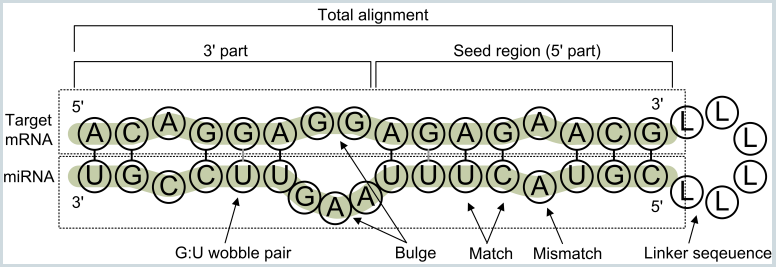
- Given a miRNA sequence, can we predict which mRNA transcripts it will bind?
- Gao et al. position their method as a post-processing tool aiming to decrease the false positive rate (FP) of the primary prediction tools



Gao et al. MicroRNA target prediction based on second-order Hidden Markov Model. Front. Biol. 2010, 5(2): 171–179

2nd problem: predicting miRNA targets

- Gaso et al's tool is a Pair-HMM representing the alignment of miRNA-mRNA
- Second order HMM-model: transition probability a_{jkl} to state l depends on two previous states j and k



Gao et al. MicroRNA target prediction based on second-order Hidden Markov Model. Front. Biol. 2010, 5(2): 171–179

HMM structure

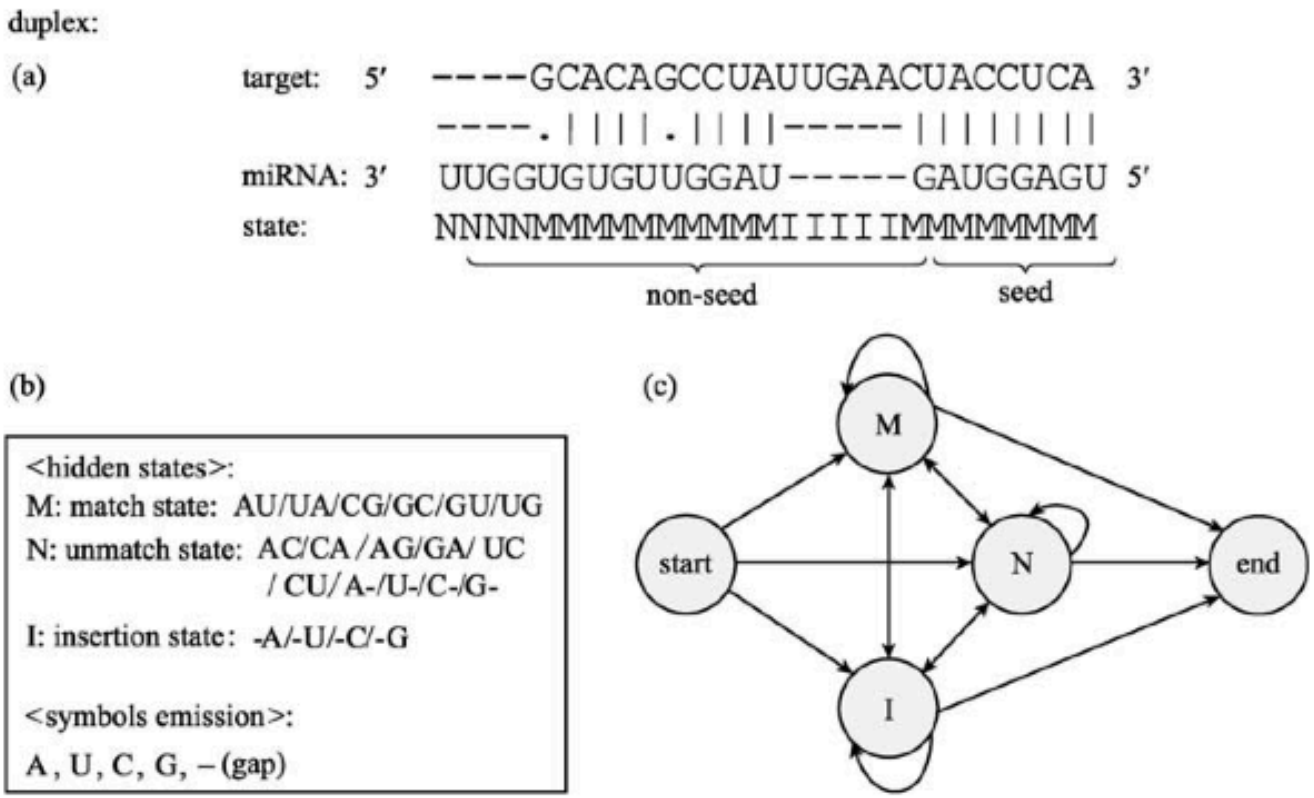


Fig. 1 MiRNA:target duplex and the definition of Hidden Markov Model (HMM). (a) MiRNA:target duplex and hidden states. (b) The definition of hidden states and symbols emission in HMM. (c) The three-hidden-state finite state automaton.

Training data for the pair HMM

- Positive data: 244 known miRNA-target pairs from Tarbase, including worm, fruit fly, zebrafish, rat,mouse and human sequences from Tarbase
- Negative data: 49 (only!) pairs that are believed not to interact: 22 from Tarbase, rest collected from scientific papers
- Two HMMs are built, one from each dataset
  - "True Target Binding Site" model
  - "False Target Binding Site" model
  - Higher scoring model "wins"

Pipeline

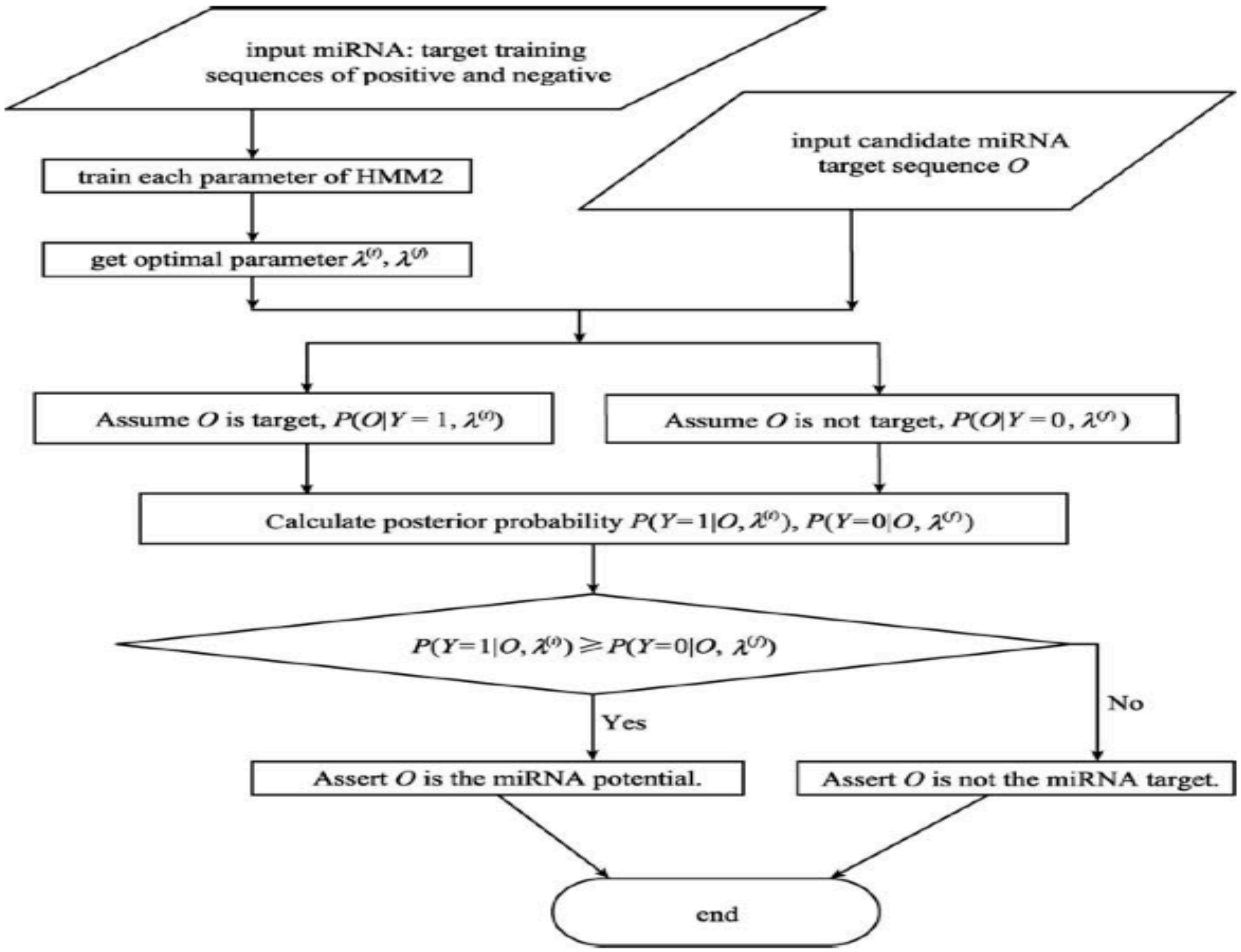
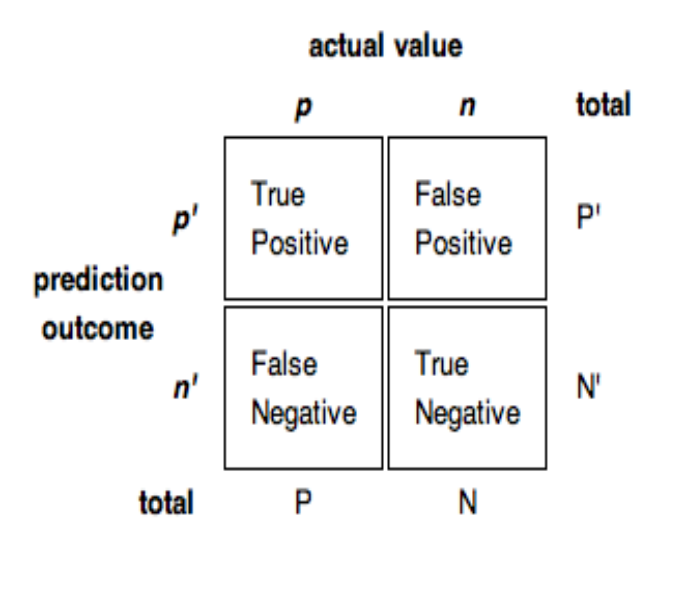


Fig. 2 Flowchart of HMM2 training and recognizing miRNA target

Evaluation metrics cheat sheet

Confusion matrix



Evaluation metrics

- Accuracy: $ACC = (TP + TN) / (TP + TN + FP + FN)$
- Precision/Positive predictive value: $PPV = TP / (TP + FP)$
- Recall/Sensitivity/True positive rate: $TPR = TP / (TP + FN)$
- Specificity/True negative rate: $SPC = TN / (FP + TN)$
- False positive rate: $FPR = FP / (FP + TN)$
- False discovery rate: $FDR = FP / (FP + TP)$
- Negative predictive value: $NPV = TN/(TN + FN)$

Prediction results

TP	FN	Se/%	TN	FP	Sp/%	ACC/%
177	67	72.54	27	22	55.10	69.62

TP stands for correctly predicted positive miRNA:target pairs; FN stands for wrongly predicted positive miRNA:target pairs; TN stands for correctly predicted negative miRNA:target pairs; FP stands for wrongly predicted negative miRNA:target pairs. Se: the sensitivity; Sp: specificity; ACC: classified accuracy.

Table 3 Prediction results of among different positive data and 49 negative

number of positive	positive set			negative set			
	TP	FN	Se/%	TN	FP	Sp/%	ACC/%
30	18	12	60.00	38	11	77.55	70.89
50	32	18	64.00	38	11	77.55	70.71
75	49	26	65.33	35	14	71.43	67.74
100	68	32	68.00	32	17	65.31	67.11
150	106	42	70.70	28	21	57.14	67.73
200	144	56	72.00	28	21	57.14	69.08
230	166	64	72.17	28	21	57.14	69.53

TP stands for correctly predicted positive miRNA:target pairs; FN stands for wrongly predicted positive miRNA:target pairs; TN stands for correctly predicted negative miRNA:target pairs; FP stands for wrongly predicted negative miRNA:target pairs. Se: the sensitivity; Sp: specificity; ACC: classified accuracy.

Table 4 Prediction results of 195 positive and 38 negative

positive set			negative set			
TP	FN	Se/%	TN	FP	Sp/%	ACC/%
162	33	83.08	23	15	60.53	79.40