Master of Molecular Medicine

Module IV: Functional Genomics :
RNA structure & gene prediction

Uwe Ohler

Berlin Institute for Medical Systems Biology

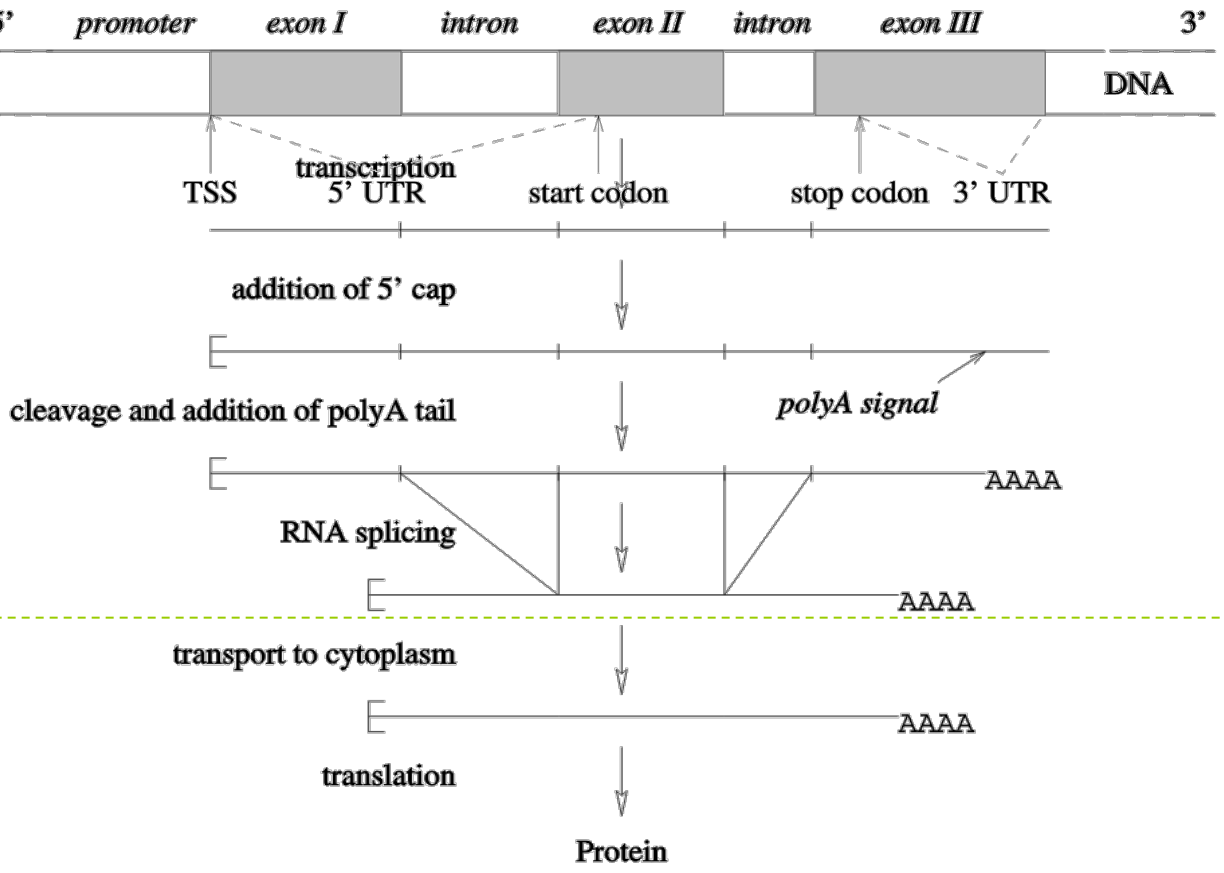
Max Delbruck Center/Humboldt University

uwe.ohler@mdc-berlin.de

Overview

- The RNA world
- RNA folding/secondary structure prediction
  - Nussinov algorithm: maximal base pairing
  - Zuker algorithm: minimal free energy
  - Probabilistic interpretation
- Prediction of non-coding RNAs
  - Comparative genomics
  - Deep sequencing

Steps in eukaryotic gene regulation



The RNA world

- RNA is thought to be the “original” molecule of life, predating DNA as the so-called “ancient RNA world”
- RNA in modern organisms was thought to be
  - only an intermediary product: the ***messenger RNA***
  - a structural component: ***rRNA***
  - involved in translation: ***tRNA***
  - but not to have an active functional role: the *Central Dogma*
- The “modern RNA world” recognizes that **RNA molecules** have a variety of important **functions**
  - challenging the central dogma and notion of “genes”

PERSPECTIVES: MOLECULAR BIOLOGY

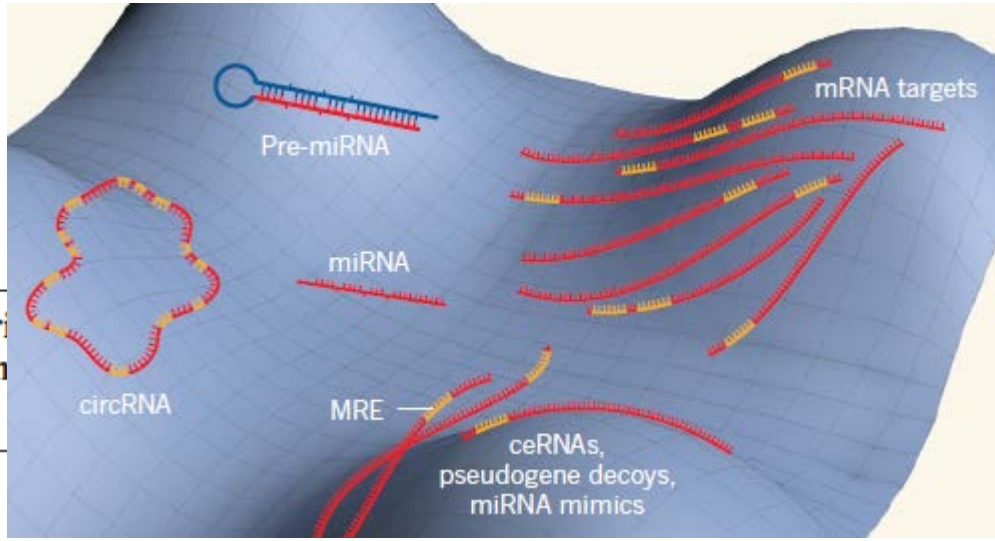
Glimpses of a Tiny RNA World

Gary Ruvkun
Science **294**, 797 (2001);
DOI: 10.1126/science.1066315

Circles reshape the RNA world

The versatility of RNA seems limitless. The latest surpri
RNAs, which are found to counteract the function of an
regulatory RNA — the microRNAs.

Nature Mar 21, 2013



Post-transcriptional control

- Increased appreciation for **RNA regulatory mechanisms**
  - Processing (e.g. polyadenylation, splicing)
  - Export/Localization
  - Stability
  - Translation
- Driven by realization of **importance of regulatory non-coding RNAs**
  - Caution: mechanism not completely universal across kingdoms

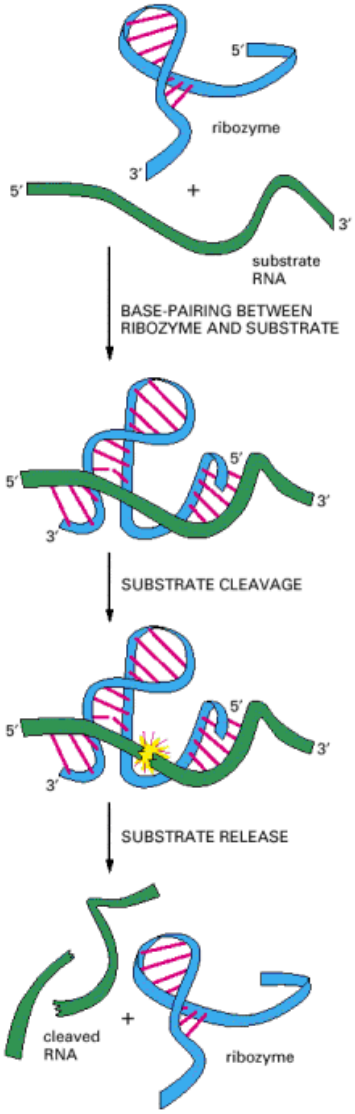
Classes of small functional RNAs

- small nuclear RNAs (snRNAs) -- Spliceosome
  - Recognize the splice sites / branch point
- Small nucleolar RNAs (snoRNAs) -- Modification
  - Lead to changes in the sequence of r/sn/m?RNAs
- Micro RNAs (miRNAs)
  - Translation repression/degradation of target mRNAs
- piRNAs (Piwi-associated RNAs)
  - Silencing of transposable elements in the germline
- Emerging picture:
  targeting of specific other RNAs or DNA regions

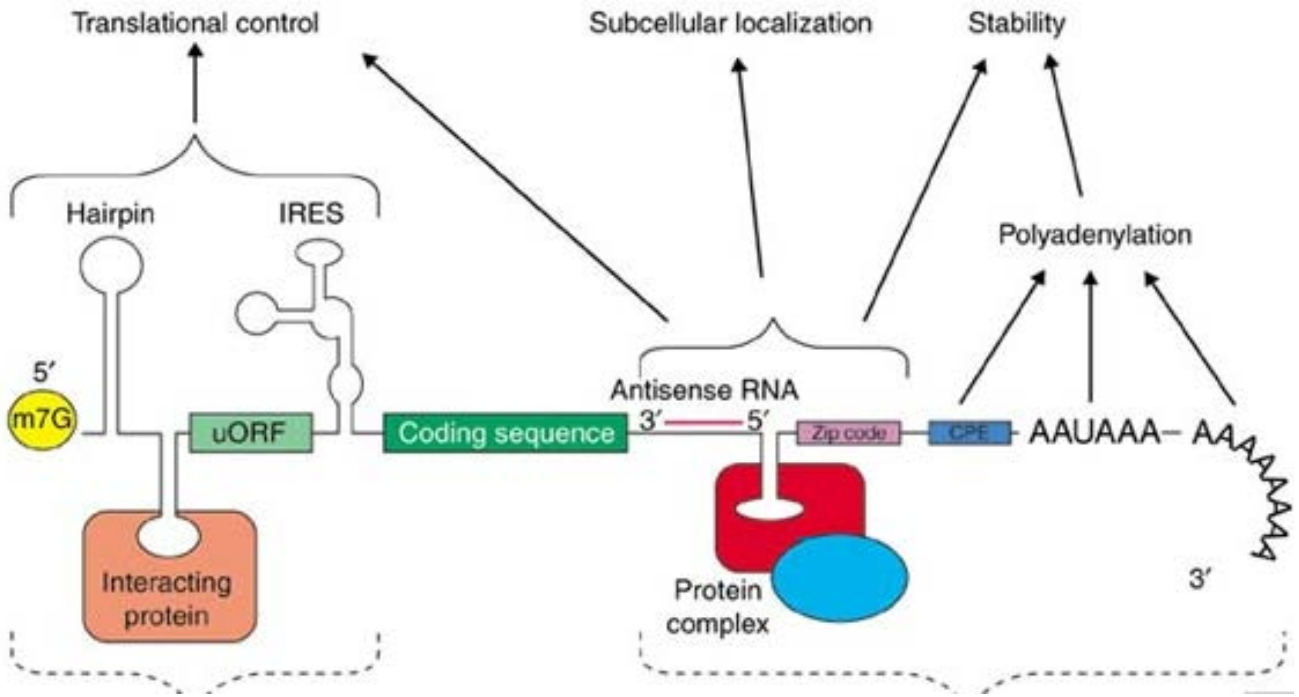
Large functional RNAs

Many different roles:

- Ribozymes --- RNA-based enzymes
  - Proof of the ancient RNA world?
- Part of RNP (ribonucleo-protein) complexes, e.g. signal recognition particle for protein export or
  polycomb complex (chromatin repression)
- lincRNAs (long intervening RNAs)
  - Xist --- the Goliath with 17kB
  - X chromosome inactivation trigger
- “the hidden pervasive transcriptome”



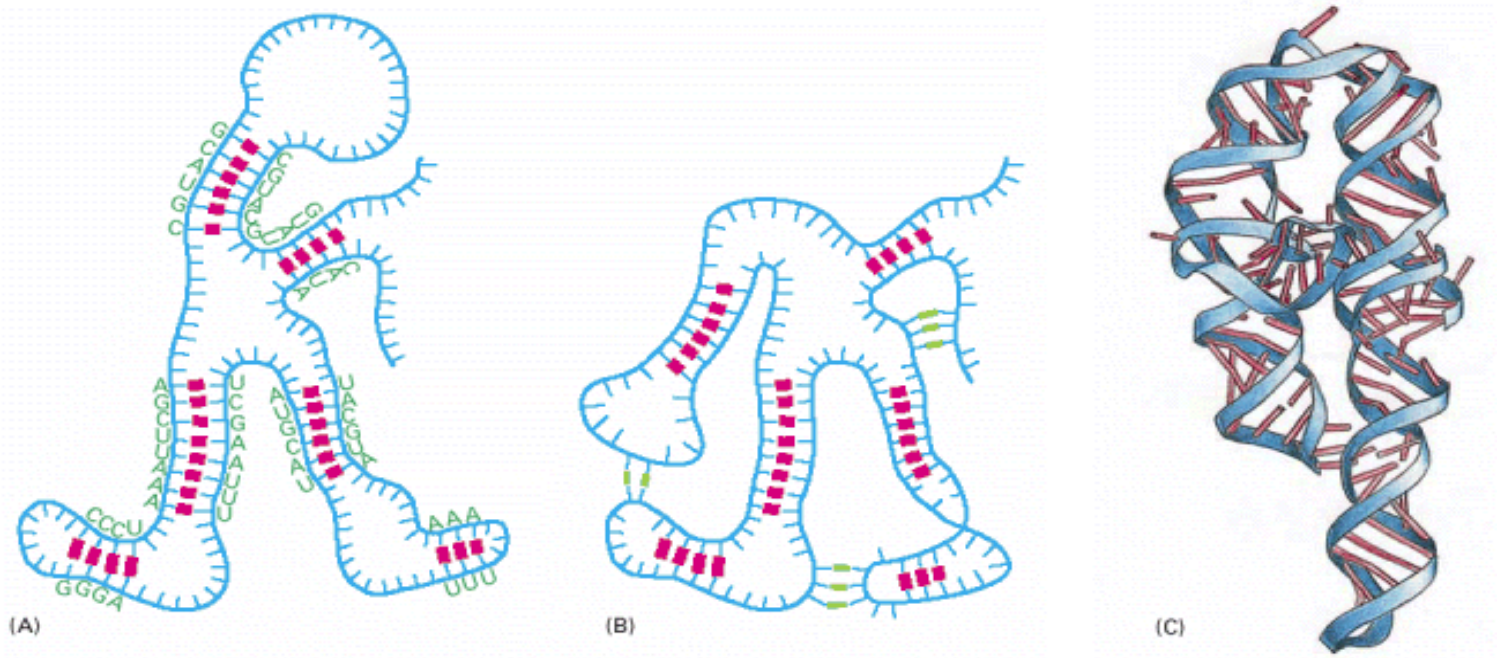
The view from cis



Structure is important

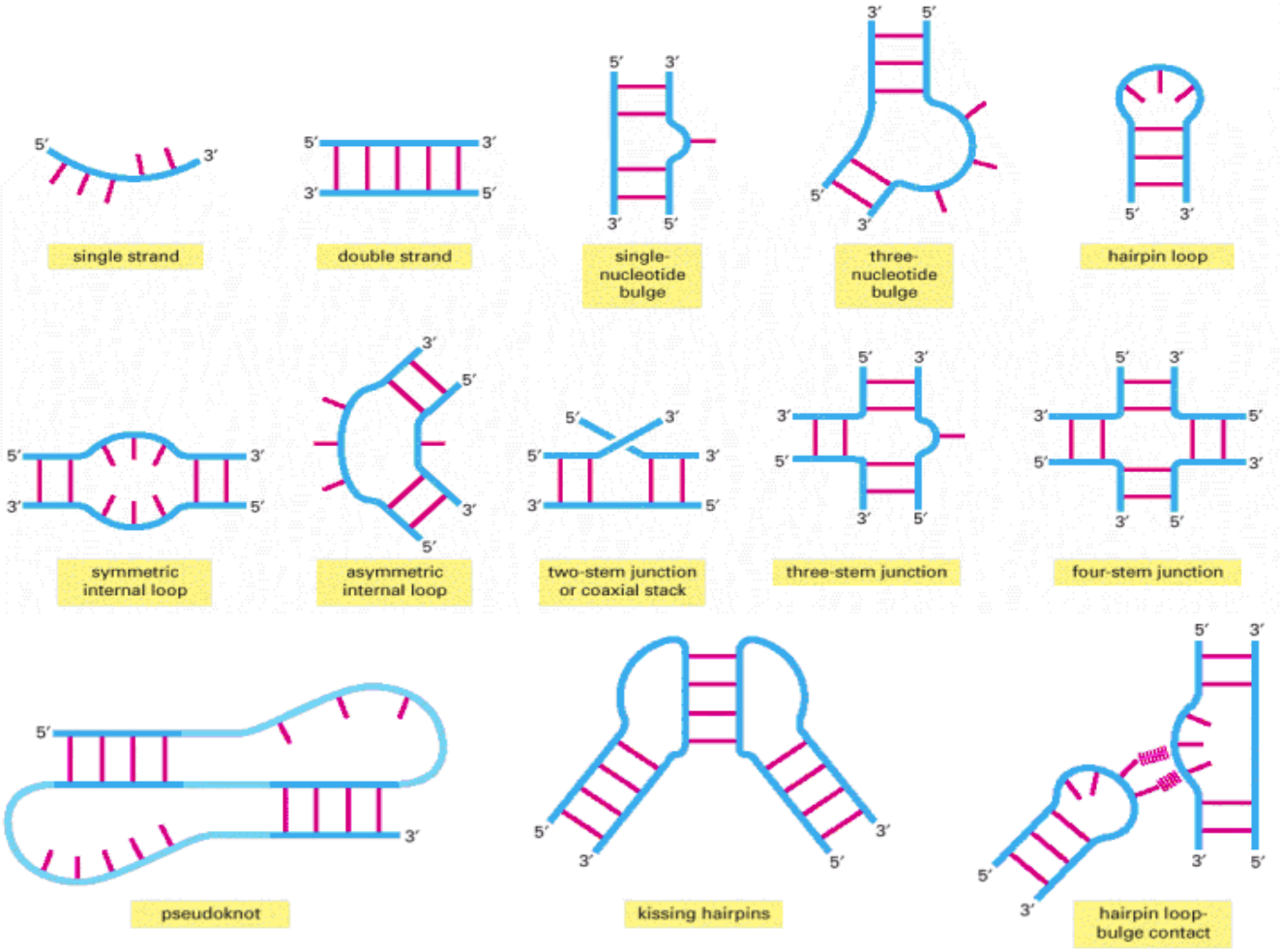
- RNA is a single-stranded molecule, and can fold back onto itself: **secondary structure**
  - G:C > A:U > G:U (Q: consequence of G:U?)
- RNAs from the same functional class often have similar secondary structure but not primary sequence
- Apart from independent RNA transcripts, secondary structure often plays a **role in *cis***
  - Splicing ⟺ recognition of splice sites
  - Riboswitches ⟺ obstruction of start codon
  - Coding sequence ⟺ efficiency of translation
  - RNA editing ⟺ change of coding sequence

From primary to tertiary structure



“Tertiary” structure sometimes refers to interactions based on the secondary structure, not the 3-D structure

RNA secondary & tertiary structure



Computational problems in RNA biology

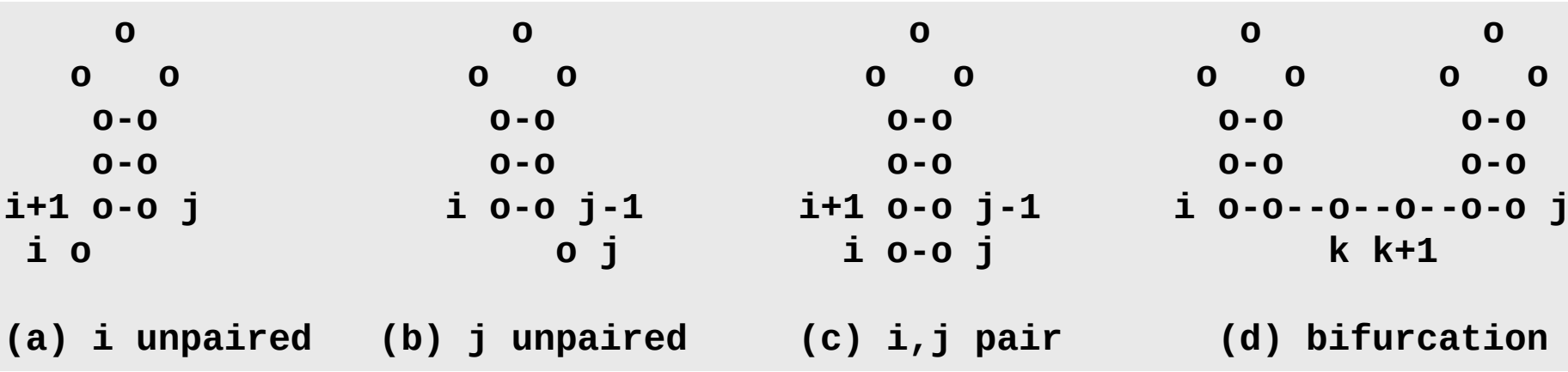
- RNA structure prediction: single molecule
  - Energy optimization
  - Probabilistic interpretation
- RNA family modeling: multiple sequences
  - Stochastic context-free grammars
  - Covariance models
- Prediction of trans-acting RNA genes/factors (microRNAs, lncRNAs…)
  and cis-acting regulatory RNA elements
  (e.g. miRNA target sites, riboswitches)

Nussinov algorithm: idea

- Premise: The more nucleotides are paired in a structure, the more stable is the structure
- Simple idea: Find the secondary structure with the highest possible number of pairs
- Naïve approach --- enumeration (have fun…)
- Instead (and I'm sure you've seen this before): Dynamic Programming !!
  - Align the sequence to itself
  - Count C:G, A:U, G:U as one, singletons as zero
  - Compute global alignment
  - But…

Nussinov algorithm: operations

- Position (*i*,*j*) in the alignment:
  - Best substructure from *i* to *j*
  - Fill matrix up to (1,*N*) and we are done
- At each position in the matrix **W**, we maximize over *four* basic cases



  - Consequence of (*d*) on the complexity?
    Runtime: $O(N^3)$

Nussinov algorithm: traceback

- We potentially have *nested* substructures, so we need to use a “stack” for traceback

```
init: push (1,N)
repeat
    pop (i,j); if (i>=j) continue;
      // done in this substructure
    else if (W(i+1,j) = W(i,j)) push (i+1,j);
      // unpaired
    else if (W(i,j-1) = W(i,j)) push (i,j-1);
      // unpaired
    else if (W(i+1,j-1)+s(i,j) = W(i,j)) push(i+1, j-1);
      // base pair
    else for (k=i+1 to j-1)
      // split substructure
        if W(i,k) + W(k+1,j) = W(i,j)
        push (k+1, j); push (i,k); break;
```

Folding energy parameters

- Simply counting a match as one and a mismatch as zero is not very close to reality

- Instead, stacking energy parameters have been (and continue to be) estimated
  - Decrease in free energy by stacking one pair of nucleotides on top of the previous pair
    - Means: Dependency on neighbor [“Markov order 1”]
  - Increase by various kinds and lengths of unpaired sequences: bulges, internal, terminal/hairpin loops
    - http://www.bioinfo.rpi.edu/~zukerm/cgi-bin/efiles-3.0.cgi
  - Incorporate qualitative restrictions, e.g., minimum hairpin loop size

Parameter examples

Stacking energy in stem, X:Y following A:U

```
5' --> 3' AX
3' <-- 5' UY
.       .       .       -0.90
.       .       -2.20   .
.       -2.10   .       -0.60
-1.10   .       -1.40   .
```

Terminal mismatch in hairpin loop, X:Y following A:U

```
5' --> 3' AX
3' <-- 5' UY
-0.30   -0.50   -0.30   -0.30
-0.10   -0.20   -1.50   -0.20
-1.10   -1.20   -0.20   0.20
-0.30   -0.30   -0.60   -1.10
```

Zuker algorithm: idea

- Start from DP maximal base pairing algorithm, but use free energy parameters instead
  - Becomes quickly complicated:
    - Two matrices needed: overall best energy, paired energy at *i*,*j* (similar to insertion/deletion in local sequence alignment)
    - Tracking of different types of unpaired regions
    - Size restrictions of unpaired/paired regions
  - Extensions allow to find suboptimal structures
    - Current assessment: For only about ~60-70% of structures, the minimal energy structure (i.e., the base pairs) is the correct one *according to the current parameter estimates*
    - Modifications to standard DP algorithm allow to predict all or a set of suboptimal structures within x% of the optimal one

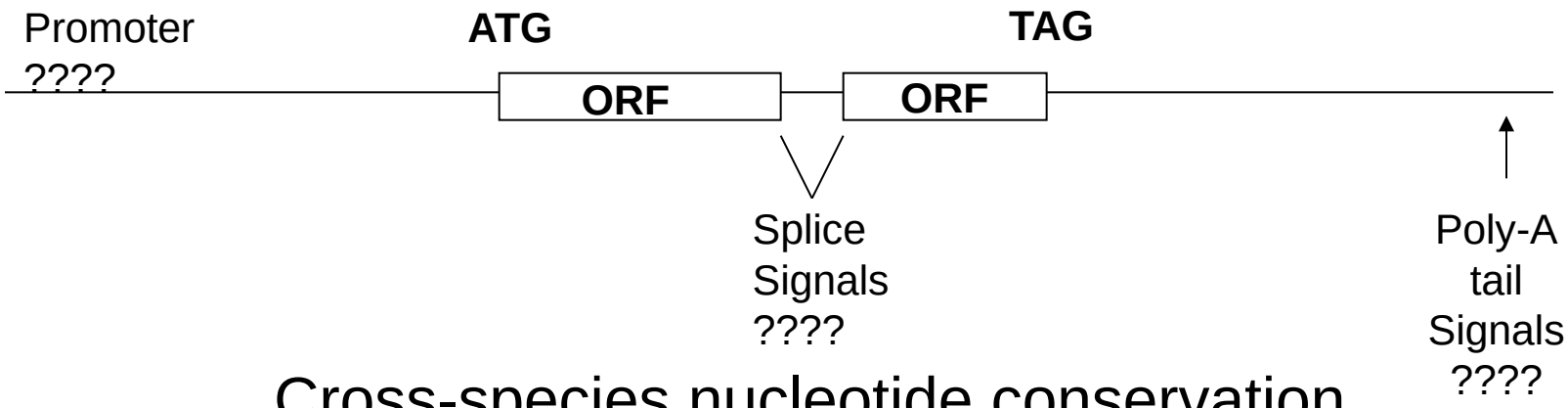
Zuker & Stiegler (1981), **Zuker** (1989)

References

- Vienna RNA package (Ivo Hofacker)
  - http://www.tbi.univie.ac.at/RNA
- MFOLD (Michael Zuker)
  - http://mfold.rit.albany.edu
- RFAM database of RNA (gene) families
  - http://rfam.sanger.ac.uk

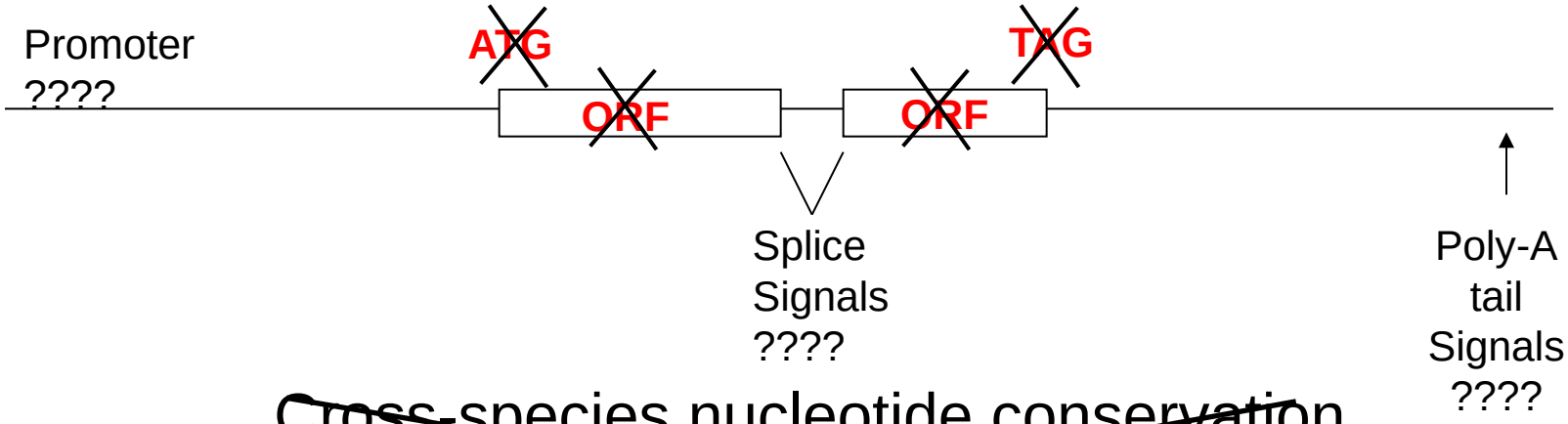
- Durbin et al, *Probabilistic sequence analysis* (chapter 10)
- Mount, *Bioinformatics 2nd edition* (chapter 8)

Protein-Coding Gene Prediction



Cross-species nucleotide conservation
Mutation bias (synonymous changes)
CpG Islands

ncRNA-Gene Prediction



~~Cross-species nucleotide conservation~~
~~Mutation bias (synonymous changes)~~
~~CpG Islands~~

So what else is there????

More systematic ncRNA gene finding

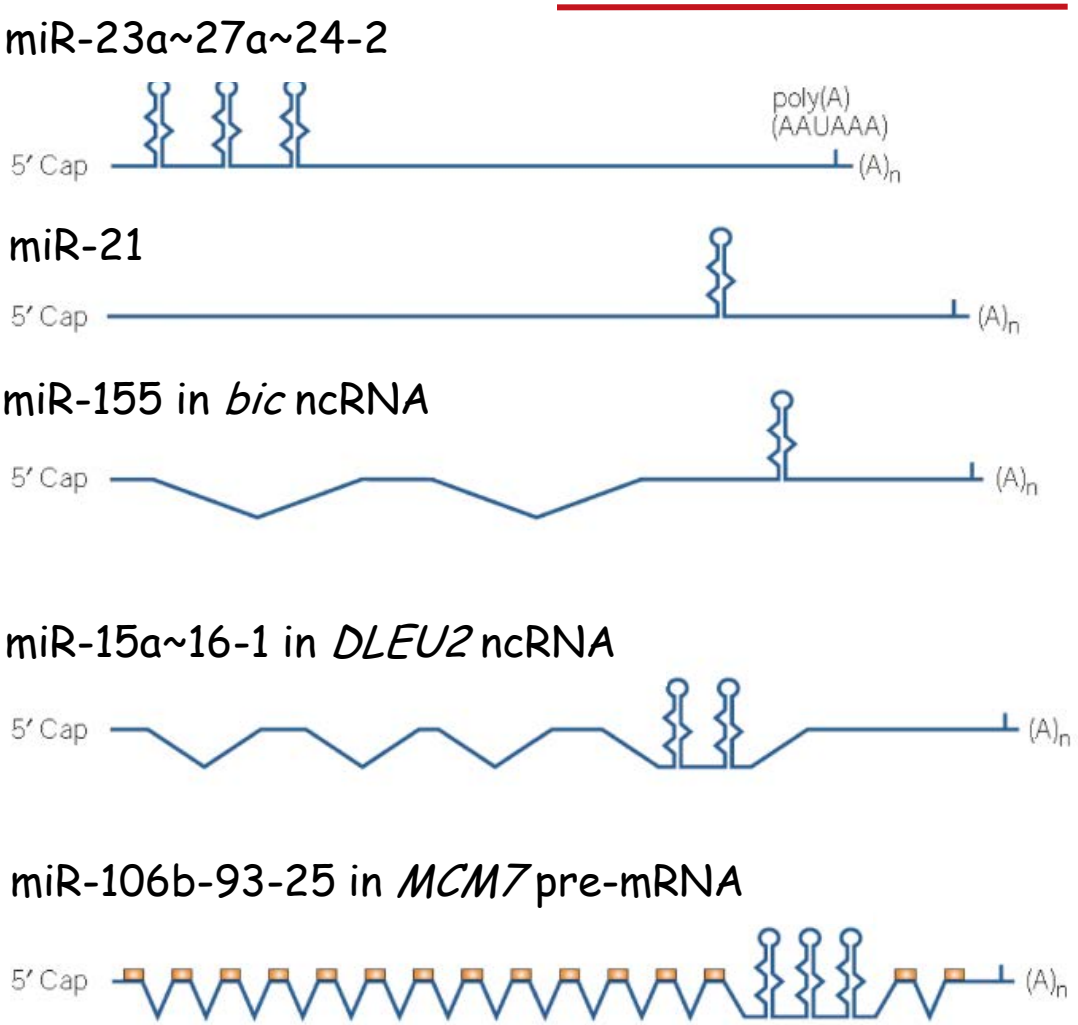
- In an ideal world, we would like to **predict RNA genes** independent of their function (just like protein coding genes)
- Bad news first: A good and fancy secondary structure does **not** imply a functional RNA
  - Large enough foldbacks occur frequently by chance
- Remedy I: **class-specific features**
- Remedy II: use **comparative algorithms**
  - Require either a formal model or *ad hoc* filtering
- Remedy III: deep RNA **sequencing**

Non-coding *regulatory* genes

- Prominent and increasingly well understood case: **miRNAs**
  - Small regulatory RNAs which repress target genes
  - ~50% of human genes are targeted
- To build a successful predictor, we need to understand the biogenesis of miRNAs:
  - Primary transcripts (several kb; nucleus; RNA pol II)
  - Precursor foldbacks (70 nt; nucleus; Drosha)
  - Mature miRNA (20-25 nt; cytoplasm; Dicer)
- Parallel to protein coding genes, with many different processing steps

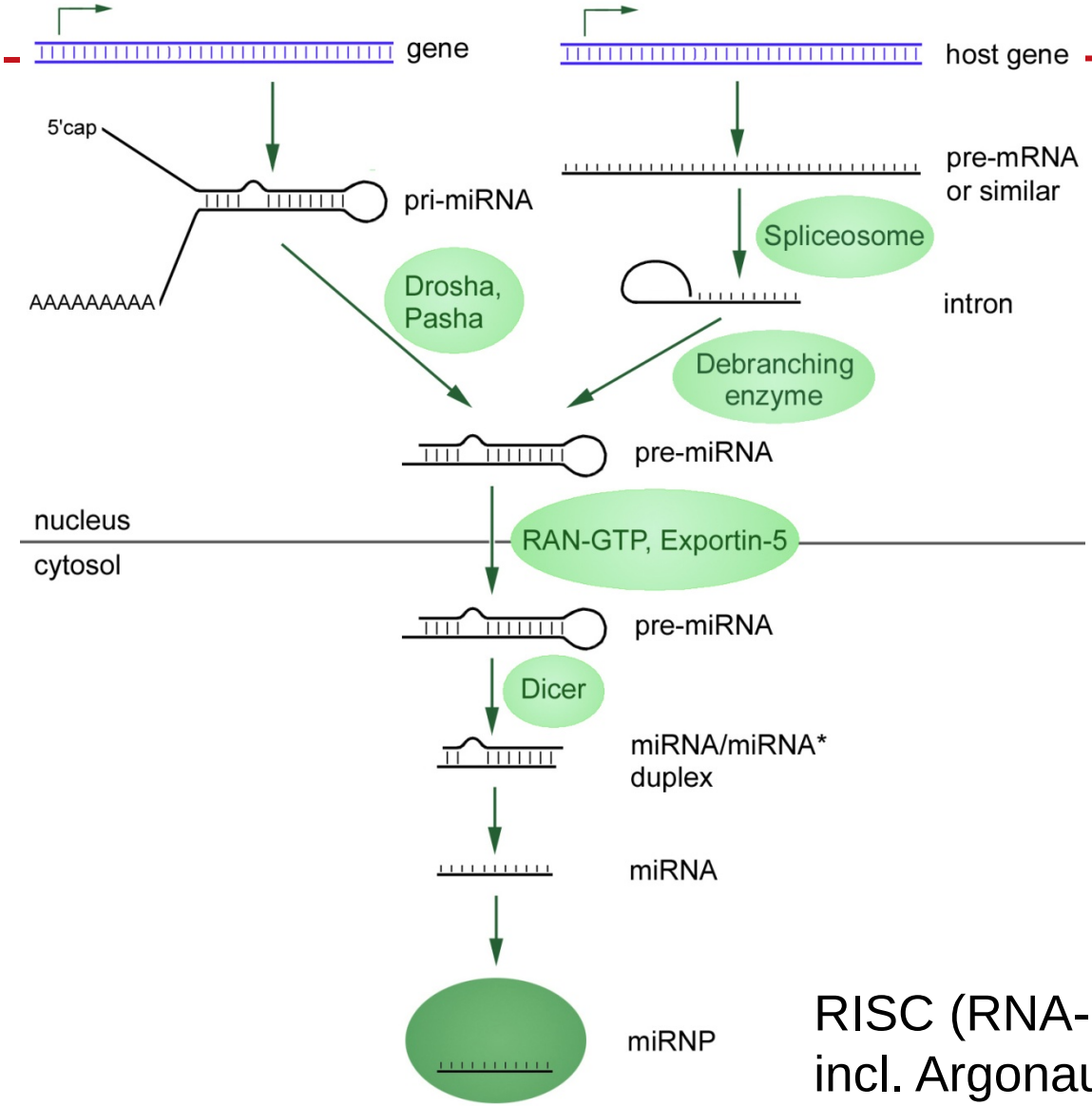
Location of miRNA genes

- miRNAs come in a variety of disguises
  - Can be independently transcribed
  - Can be intron/exon of a non-coding transcript
  - Can be part of an intron of a protein-coding gene
  - Can come in clusters



Kim (2005)

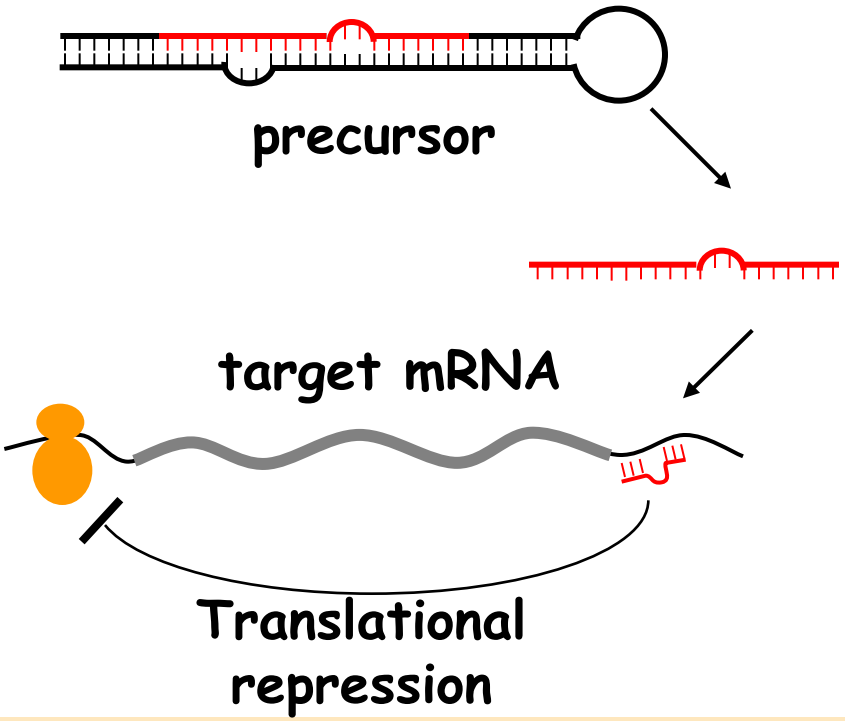
MicroRNA biogenesis



RISC (RNA-induced silencing complex), incl. Argonaute (AGO) proteins

What is the function of microRNAs?

- Animal miRNAs target protein coding genes through complementary sequence regions in their 3′ UTR
- Part of a protein complex (RISC); miR defines target
- “Natural counterpart” to siRNAs/RNAi (RISC complex)



- One miRNA influences many target genes
- One miRNA can have several target sites in one UTR
- One UTR can have multiple miRNA targets
- miRNA genes and targets are often conserved

Identification of miRNA genes: conserved foldbacks

Example: original `miRscan` (Lim et al 2003)

1. Scan *C.elegans* genome for potential RNA hairpin structures
   - Fold every 110 base segment in genome using RNAfold (Vienna RNA software package, Hofacker *et al*)
2. Identify hairpins with homology to *C.briggsae* shotgun traces
   - BLAST cutoff E1.8; RNAfold *C.briggsae* sequence
3. Align *C.elegans* and *C.briggsae* hairpins
   - Pair must have certain secondary structure similarity
4. Classify foldback into miRNA/no miRNA using features representative of miRNAs

Excursion: Classification

- Many problems in molecular biology can be approached by computational methods, in particular classification
  - Finding/locating genes
  - Determining protein domains
  - Cancer diagnosis using microarrays
  - Inference of regulatory networks
- With the availability of large-scale data, we have the ability to do this
  - One needs examples to build models for different classes

Classification

- Representation:
  - Samples/objects from particular problem domain
  - Objects represented by specific *features/attributes*
- Supervised: Class labels are known
  - We have objects from several *classes* and want to distinguish between them
    - Simple: assign class label to whole sample
    - Complex: parse sample into different classes
- Unsupervised/clustering: Class unknown
  - Determine meaningful groupings of the samples

Classification systems

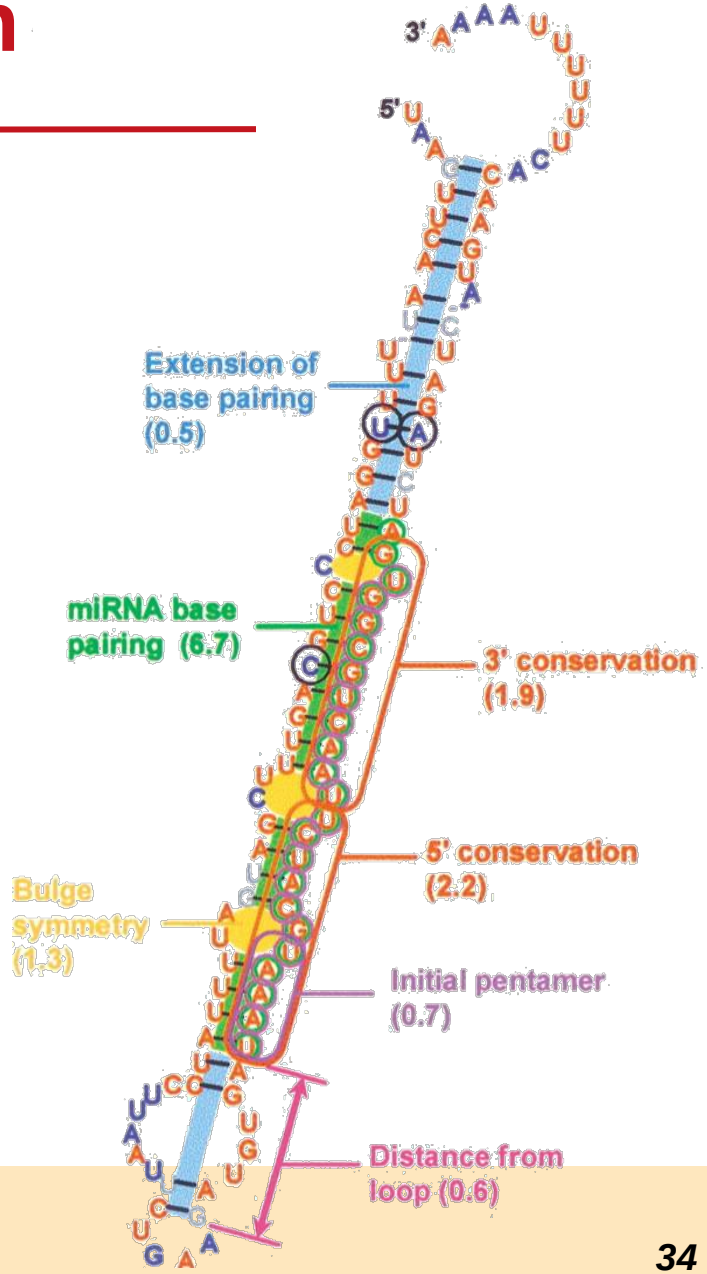
- Many classification systems consist of the following components
  - *Preprocessing:*
    - Noise removal (e.g. different filters)
    - Discretization
    - Normalization (to standardize input data)
  - *Feature extraction:*
    - Compute values from the (analog) input data
    - Categorical (e.g. male/female) or numerical (e.g. size; discrete or continuous)
    - Dimensionality reduction
  - *Classification*

An example you know: PWMs

- A weight matrix is a *model* of related sequences (eg, transcription factor binding sites)
  - The model represents our *knowledge* about the sequences in form of *parameters* (here, the relative frequencies or scores for nucleotides at different PWM positions)
  - The parameters are *estimated* using a *representative* set of examples (positives and negatives/background)
  - Using the log-odds scores, we evaluate the probabilities of two *competing* models: binding site vs background genomic content
- What are possible parameters for a miRNA classifier?

Example: miRNA prediction

- miRNAs are excised from precursor foldbacks/hairpins
- Real precursor foldbacks have distinct *features*
- Classifier can distinguish real miRNAs from ubiquitous foldbacks of similar size

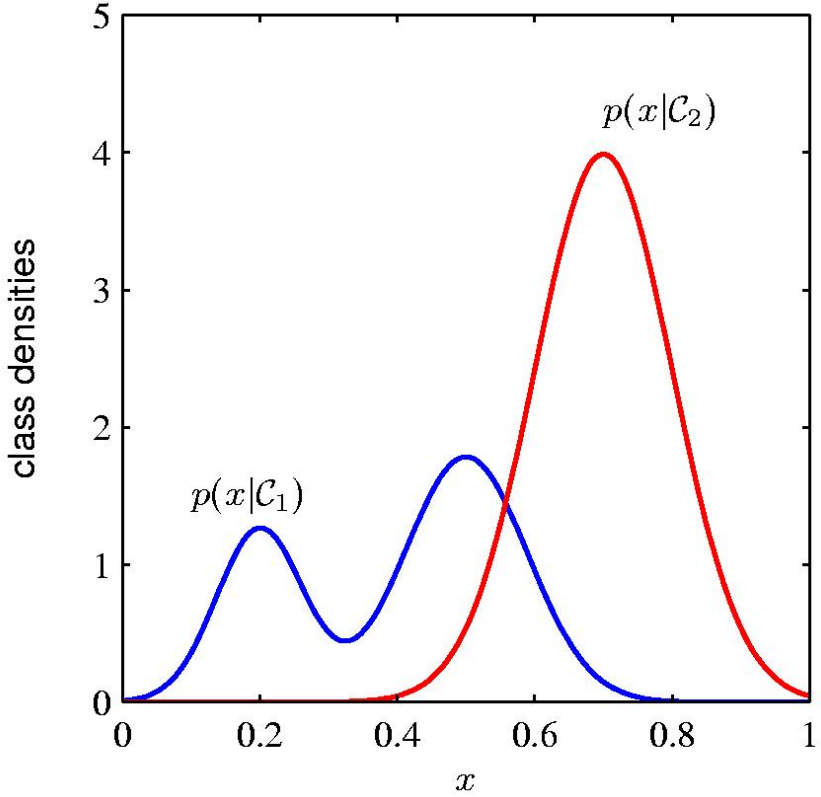


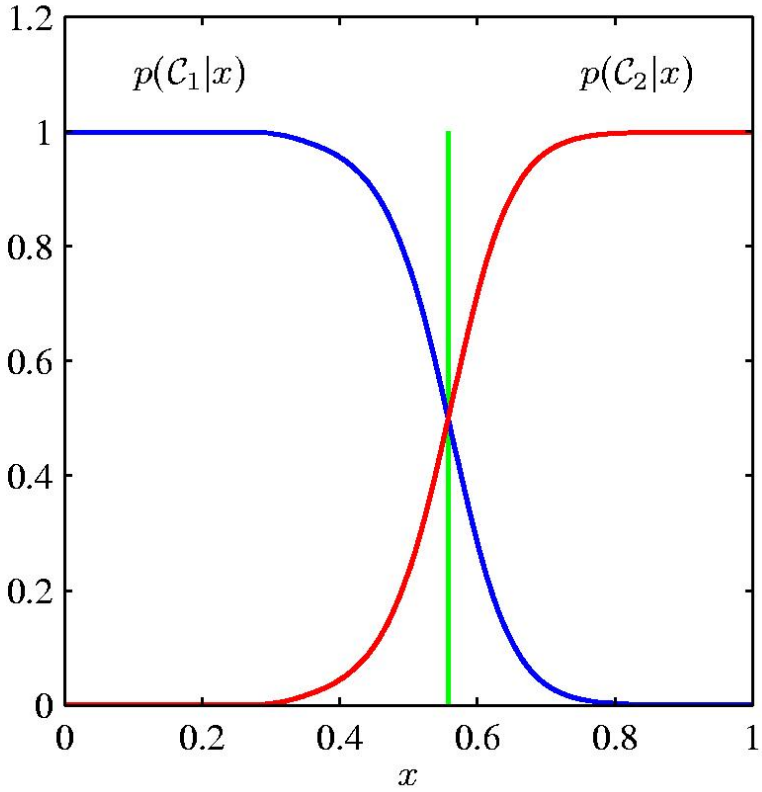
Lim *et al* (2002)

Probability-distribution based approach

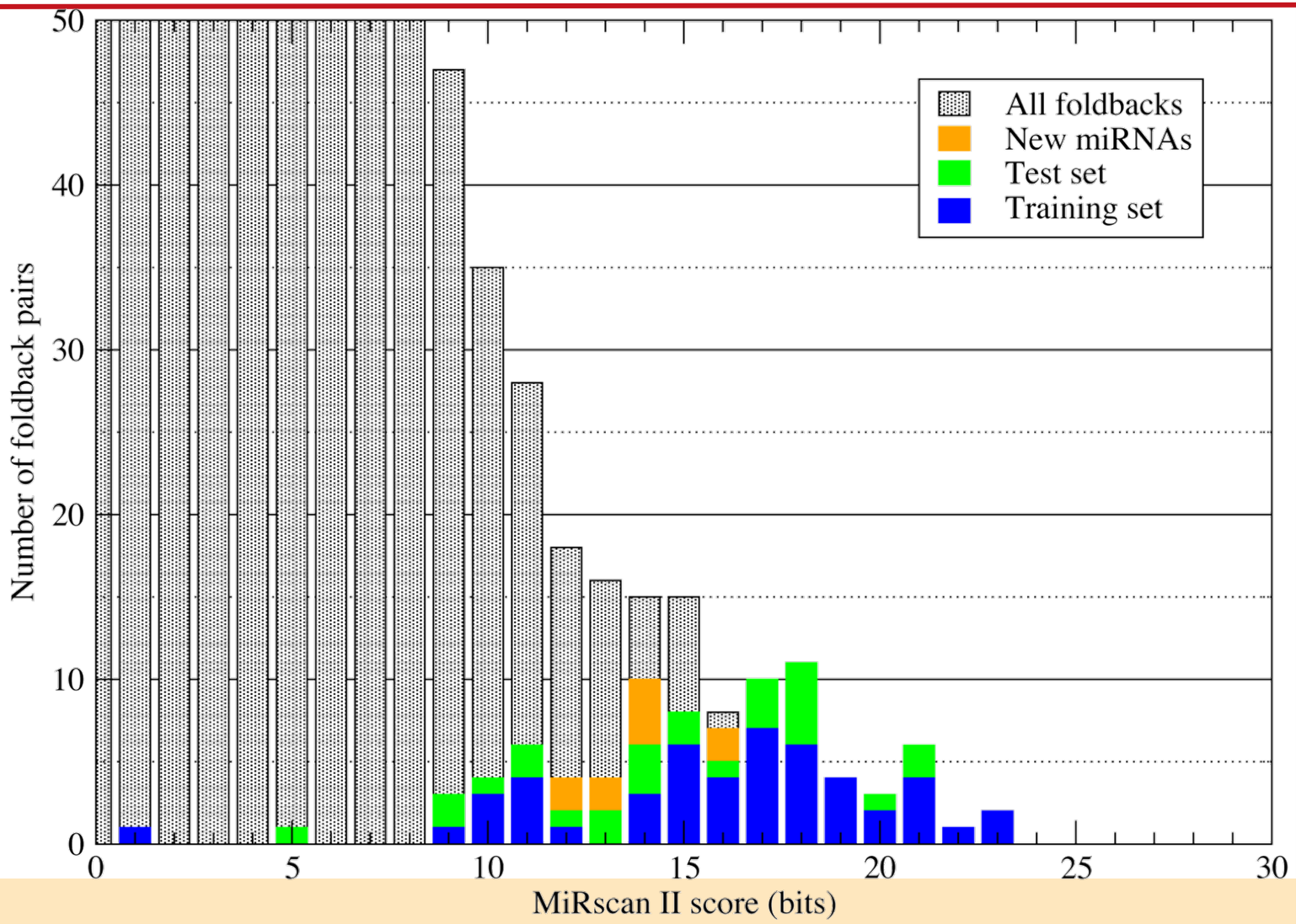
- 1. **Training**: Estimate distributions for the feature values for each class from training data: the *models*
  - Discrete (e.g. histogram) or continuous (e.g. Gaussian) distributions
- 2. **Classification**: Determine the probability/likelihood for unseen *test* data based on their features
  - Decision rule: Class with highest posterior probability wins (*Bayes classifier*)
- If features are independent, i.e. uncorrelated: *naïve* Bayes
  - Separate distribution for each feature;
  - probabilities of individual features can simply be multiplied

Class-specific density vs. posterior





Back to miRs: Predictions and validations

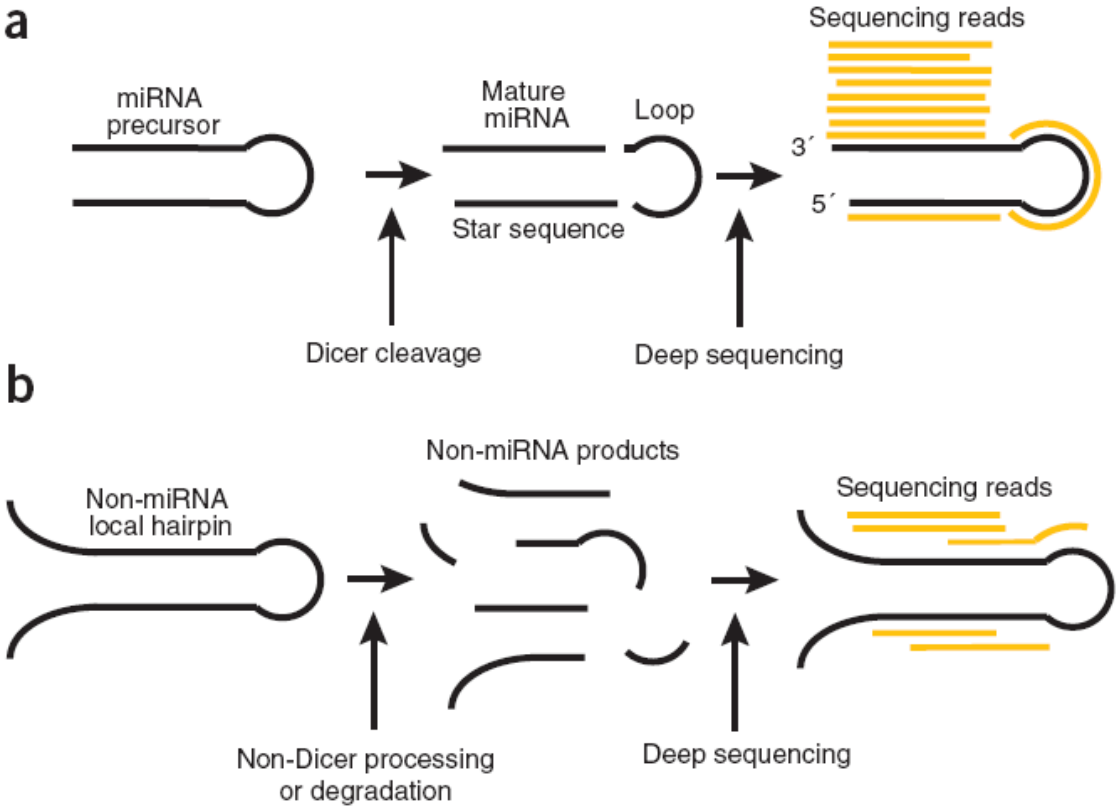


Principles to predict specific classes of ncRNA

- RNAs from the same class show specific *primary* sequence features
- RNAs from the same class show similar *secondary* structure
- Other features: length, genomic context, conservation patterns
  - Common problem: Availability of training/test data
- Combine these in a model, search for highly probable regions in the genome
  - Due to the complexity of structure prediction, this is often done in a sliding window

Current approaches: deep sequencing

- mirDeep: Simplified picture of short reads mapping to real miRs and spurious foldbacks [Friedlander et al 2008]



miRdeep (II)

- Exploit features from deep sequencing
  - Align reads, discard multiple matches
  - Cluster reads within 30nt, extract two candidates of 110 nt length, fold
    - extend on both sides: miRs can be on either 5’ or 3’ arm
  - Score (naïve Bayes)
    - # reads for miR candidate (position with most reads)
    - Presence of miR* (aligned position offset by 2nt)
    - Loop: region in between
    - Precursor MFE
    - Conservation of seed region

miRdeep (III)

- Results: *C elegans* and human