Sequence/String Alignment Using HMMs

Gaurav Sharma

University of Rochester

Motivation

- What does the following sentence say?
  - It ws a rny dy*

String Alignment Problem

- Example: "It ws a rny dy*" vs. "It was a rainy day"
- Conventional Hamming distance is large
  - Hamming distance (assuming blanks extend shorter string) $= 13$
  - yet we are able to understand the message in this string

$\mathbf{x}_2 \longrightarrow$	I	t	-	w	s	-	a	-	r	n	y	-	d	y	*	-	-	-
$\mathbf{x}_1 \longrightarrow$	I	t	-	w	a	s	-	a	-	r	a	i	n	y	-	d	a	y
Diff $\longrightarrow$	0	0	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1

- Intuition? How are we effectively correcting the errors in the string?
- String edit distance instead of Hamming distance
  - smallest number of insertions, deletions and substitutions that will transform one string to another
  - How should this be computed efficiently?

String Edit Distance Computation: Example

- Formulation
  - Strings x_1 and x_2 with lengths M and N, respectively
  - ${}_{n_1}^{n_2}x \equiv$ subsequence of symbols from index n_1 to n_2 in sequence x
  - $d(m, n) =$ min. # of edits req. to match ${}_1^m x_1$ and ${}_1^n x_2$
- Example:
  - $x_1 =$ "It was a rainy dy*"
  - $x_2 =$ "It ws a rny dy*"

String Edit Distance: Traceback

- $d(m, n) =$ min. # of edits req. to match ${}_1^m x_1$ and ${}_1^n x_2$

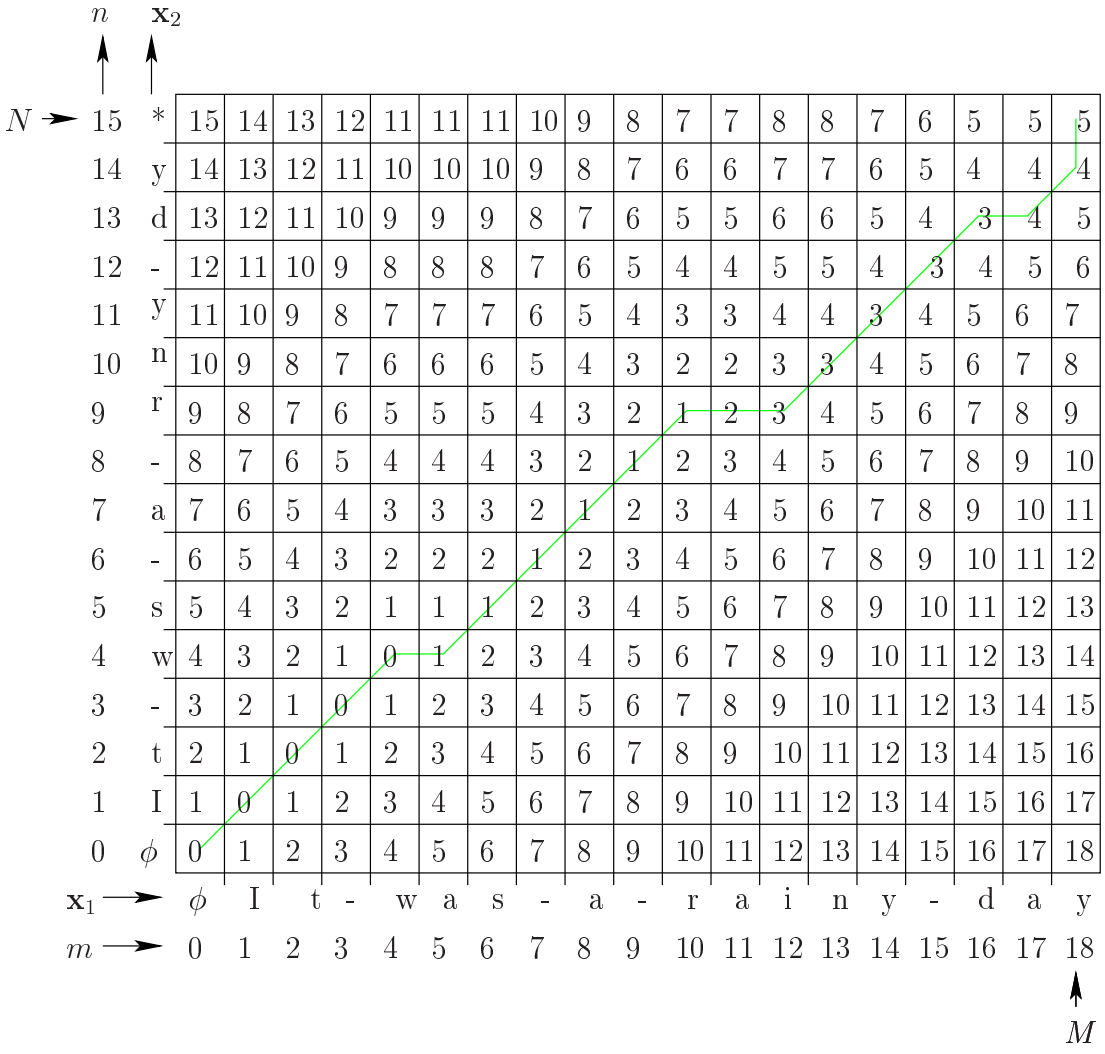


Figure: Dynamic programming computation of string edit distance example: traceback.

String Edit Distance: General Recursion

- General recursion

$d(m, n) \stackrel{\text{def}}{=}$ min. # of edits req. to match ${}_1^m\mathrm{x}_1$ and ${}_1^n\mathrm{x}_2$

$$= \min \begin{cases} d(m-1, n) + 1 & \text{insert in sequence 1} \\ d(m, n-1) + 1 & \text{insert in sequence 2} \\ d(m-1, n-1) + \delta({}_m\mathrm{x}_1, {}_n\mathrm{x}_2) & \text{aligned} \end{cases}$$

where, for any characters a and b,

$$\delta(a, b) = \begin{cases} 0 & a = b \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

String Alignment Problem

- String edit distance example for "It ws a rny dy*" vs. "It was a rainy day"
  - String edit distance is 5
  - Contrast with Hamming distance of 13
- Another instance of dynamic programming
- Reminiscent of which HMM algorithm?
  - Problem 2: MAP sequence estimation.
  - What's different here?
  - Heuristic framework as opposed to a formal probabilistic model
- Can we cast in a formal probabilistic framework?
  - How would that be more useful?

HMMs for Sequence/String Alignment

- What is the natural latent variable/process for sequence alignment?
  - State of "alignment" between the two sequences
  - Possible states?
  - Align (ALN), Insertion in sequence 1 (INS1), Insertion in sequence 2 (INS2)
  - What about deletions?
- Underlying Markov process of sequence alignment states
- What are the emissions?
  - Should correspond to observed strings
  - One string? Both strings?
  - Pair of elements, one per string
    - Observed strings may not be of same length
    - Dummy "gap" symbol
    - Emission possibilities are state dependent

Pairwise Sequence Alignment HMM

- Will describe in the context of aligning RNA sequences
  - symbols correspond to nucleotides, 4 -letter alphabet $\{A, U, G, C\}$
  - Readily generalizes to other scenarios
- Sequences x_1 and x_2, ${}_{n_1}\mathrm{x}$ denotes the n_1^{th} nucleotide of x
- ${}_{n_1}^{n_2}\mathrm{x} \equiv$ subsequence of nucleotides from index n_1 to n_2 in sequence x
- Alignment between two sequences specified by sequence of states from the set

$$M = \{\mathrm{ALN}, \mathrm{INS1}, \mathrm{INS2}\} \tag{2}$$

Alignment Example

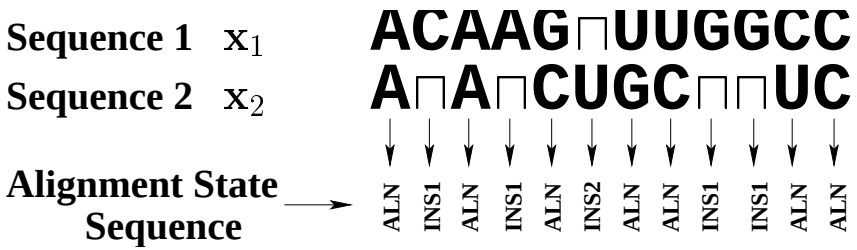
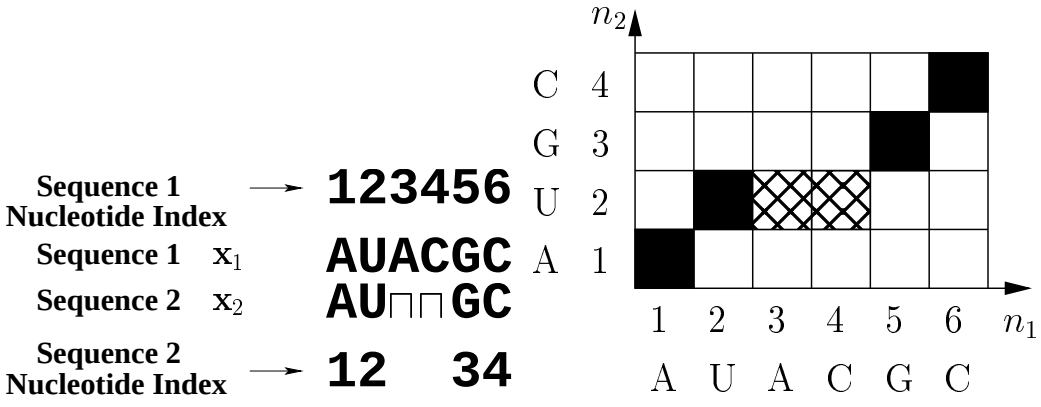


Figure: A sample sequence alignment and corresponding state sequence.

Alignment Representation Using Coincidence Map

- Example alignment and correponding coincidence map
  - Recall string edit distance computation



(a) Sample Alignment (b) Coincidence Map

Figure: Example illustrating *co-incidence*. (a) A sample alignment of two sequences, where inserted locations in a sequence are shown with a gap ⊓ in the other sequence in the corresponding location. (b) The set of *co-incident* position pairs is depicted. Coordinates corresponding to the co-incident position pairs are indicated by black (aligned) or cross-hatched (insertion) blocks.

Alignment HMM

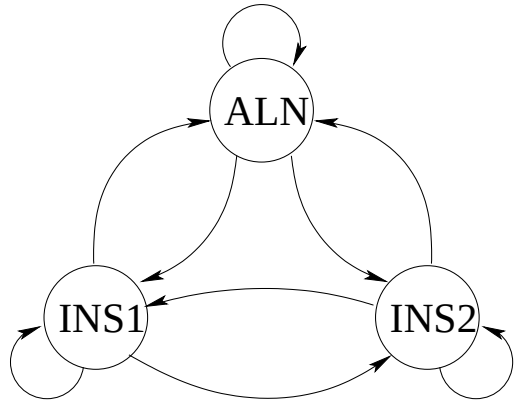


Figure: State transition diagram for the (hidden) Markov Process determining alignment between the sequences. The three states ALN, INS1, and INS2 represent alignment, insertion in sequence 1, and insertion in sequence 2, respectively.

Trellis for Alignment HMM

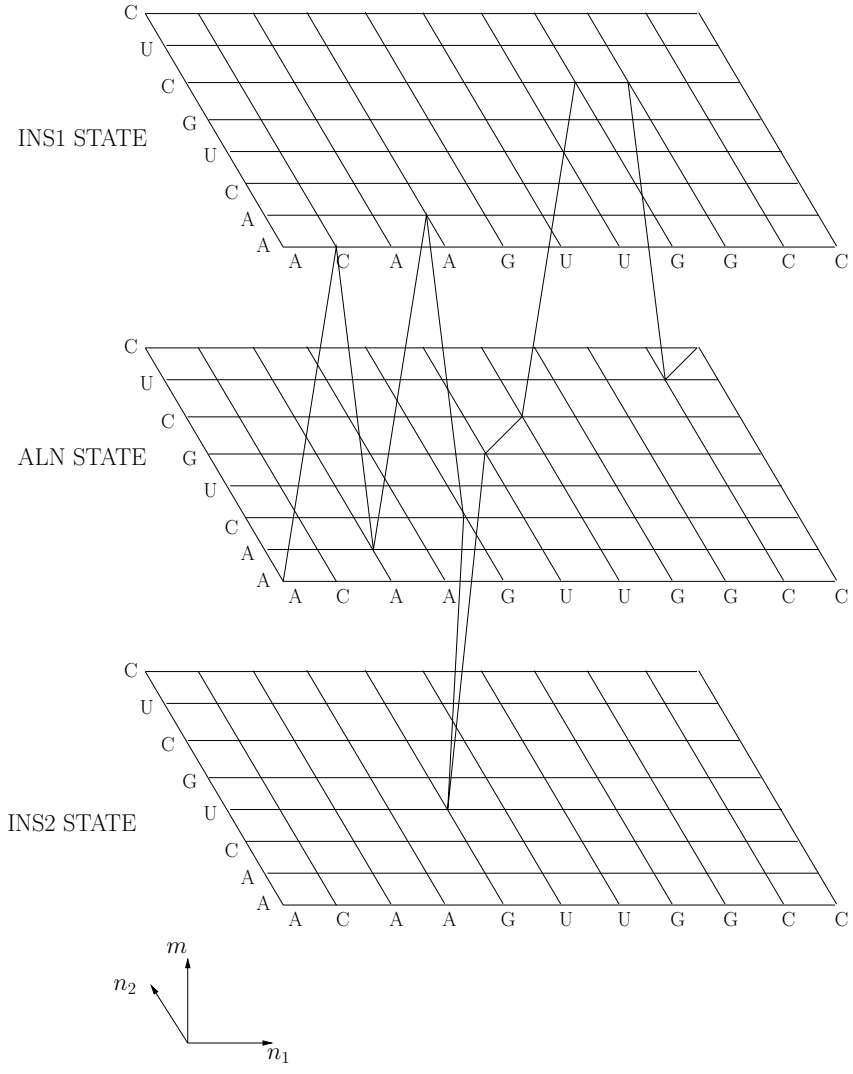


Figure: Trellis illustrating an alignment path.

Use of Posterior Probabilities for an Alignment Envelope

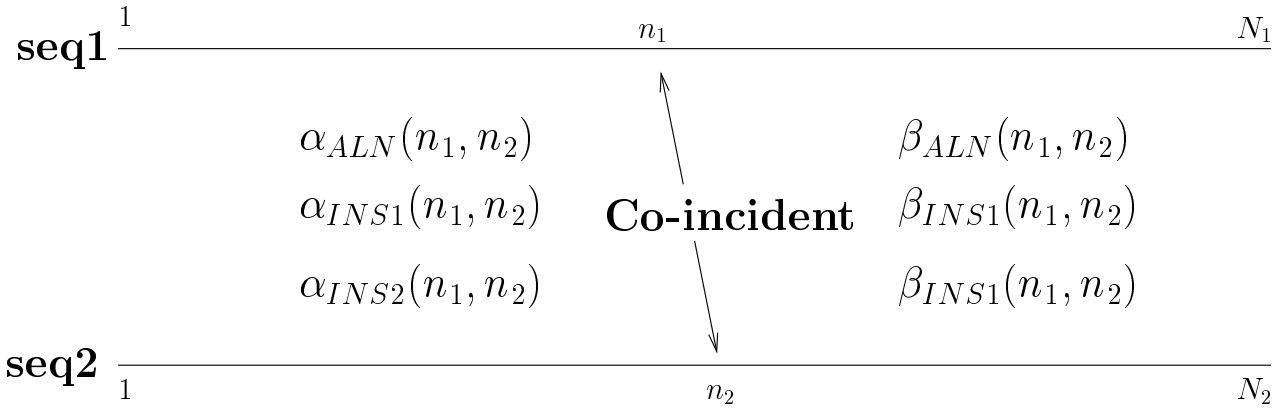


Figure: Illustration of alignment of nucleotide at n_1 in 1^{st} sequence and nucleotide at n_2 in 2^{nd} sequence and how forward and backward variables are related to alignment of n_1 and n_2. Forward variable keeps track of events before and up to position (n_1, n_2) and backward variable keeps track of events after position (n_1, n_2)

Forward-Backward Terms for Alignment HMM

- Forward term

$\alpha_m(n_1, n_2)$ represents the probability that the subsequences ${}_1^{n_1}x_1$ and ${}_1^{n_2}x_2$ of the first and second sequence, respectively, are produced and the nucleotide positions n_1 and n_2 are in the state m

- Backward term

backward variable $\beta_m(n_1, n_2)$ represents the probability that subsequences ${}_{n_1+1}^{N_1}x_1$ and ${}_{n_2+1}^{N_2}x_2$ are observed given that the n_1^{th} and n_2^{th} nucleotide positions are in state m

Forward-Backward Terms for Alignment HMM

- Forward term

$$\alpha_m(n_1, n_2) = \mathrm{P}(S_m(n_1, n_2), \ {}^{n_1}_{1}\mathsf{x}_1, \ {}^{n_2}_{1}\mathsf{x}_2), \tag{3}$$

- Backward term

$$\beta_m(n_1, n_2) = \mathrm{P}({}^{N_1}_{n_1+1}\mathsf{x}_1, \ {}^{N_2}_{n_2+1}\mathsf{x}_2 \mid S_m(n_1, n_2)), \tag{4}$$

Forward Recursions for Alignment HMM

$$
\begin{aligned}
\alpha_{ALN}(n_1, n_2) &= \sum_{m \in M} \tau(m, ALN)\, \gamma_{ALN}({}_{n_1}\mathsf{x}_1, {}_{n_2}\mathsf{x}_2)\alpha_m(n_1 - 1, n_2 - 1) \\
\alpha_{INS1}(n_1, n_2) &= \sum_{m \in M} \tau(m, INS1)\, \gamma_{INS1}({}_{n_1}\mathsf{x}_1, \sqcap)\alpha_m(n_1 - 1, n_2) \qquad (5) \\
\alpha_{INS2}(n_1, n_2) &= \sum_{m \in M} \tau(m, INS2)\, \gamma_{INS2}(\sqcap, {}_{n_2}\mathsf{x}_2)\alpha_m(n_1, n_2 - 1)
\end{aligned}
$$

Backward Recursions for Alignment HMM

$$
\begin{aligned}
\beta_m(n_1, n_2) \;=\; & \tau(m, INS1)\, \gamma_{INS1}({}_{n_1+1}\mathsf{x}_1, \sqcap)\beta_{INS1}(n_1+1,\ \ n_2) \\
& +\tau(m, ALN)\, \gamma_{ALN}({}_{n_1+1}\mathsf{x}_1, {}_{n_2+1}\mathsf{x}_2)\cdot \\
& \beta_{ALN}(n_1+1,\ \ n_2+1) \\
& +\tau(m, INS2)\, \gamma_{INS2}(\sqcap, {}_{n_2+1}\mathsf{x}_2)\beta_{INS2}(n_1,\ \ n_2+1)
\end{aligned}
$$

Posterior Probabilities for Alignment States

- The posterior probability in terms of forward and backward variables

$$P(n_1 \leftrightarrow n_2 \mid \mathrm{x}_1, \ \mathrm{x}_2) = \frac{\sum_m \alpha_m(n_1, n_2)\beta_m(n_1, n_2)}{P(\mathrm{x}_1, \ \mathrm{x}_2)} \tag{6}$$

- Recall general HMM discussion
  - Probabilities of coincidence similarly computed

Posterior Probabilities for Alignment States

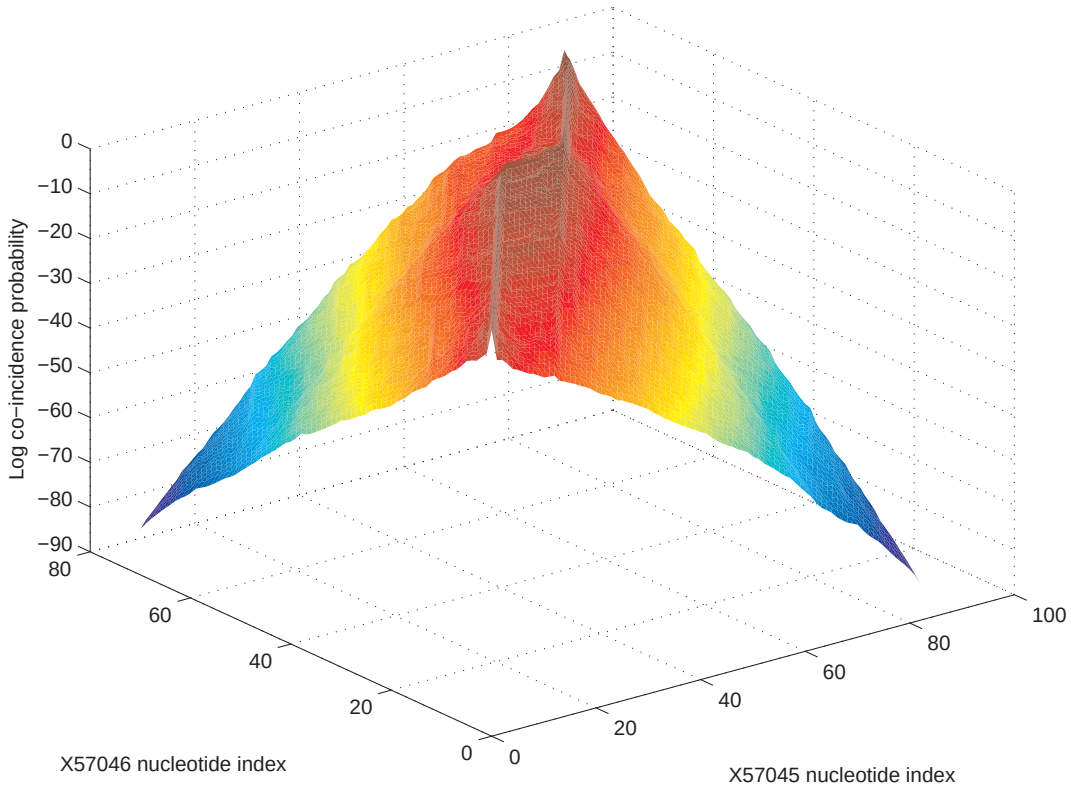


Figure: Logarithm of posterior probabilities for co-incidences of nucleotide positions for a pair of sequences in a surface plot representation.

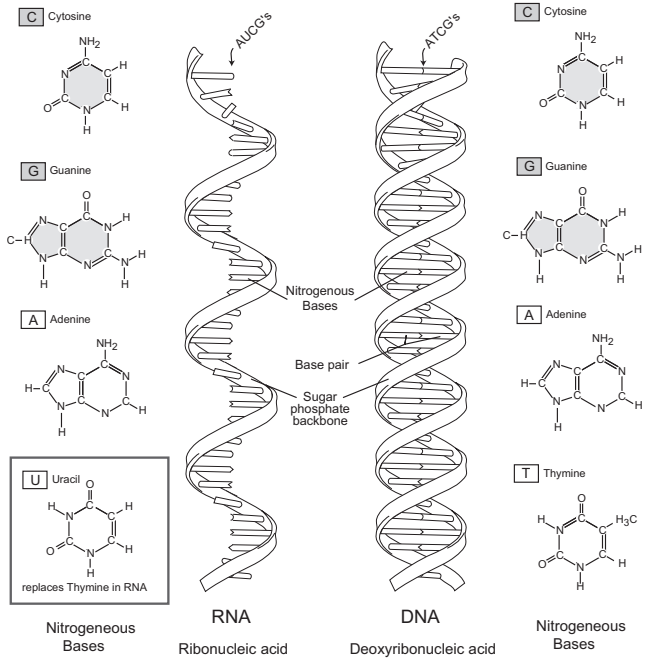
Application: RNA Structural Alignment

- Principled heuristic for "pruning computation" in joint RNA secondary structure prediction and alignment [6]
- Background: RNA structural alignment
  - Joint alignment and secondary structure prediction for homologous RNA sequences

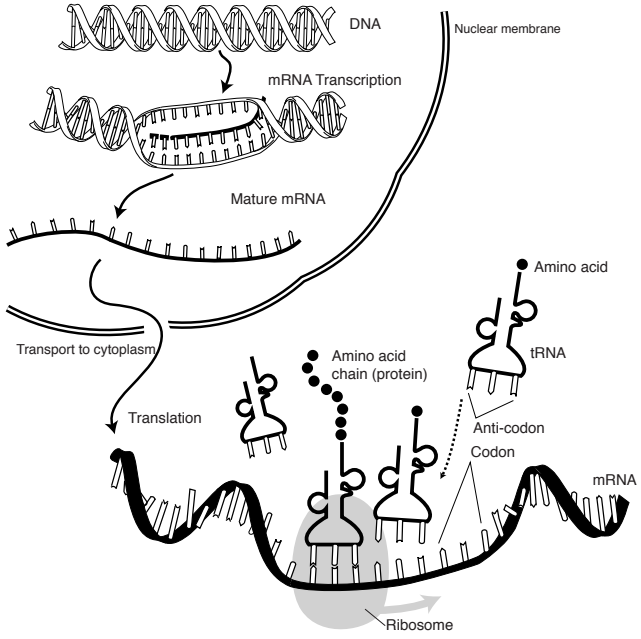
RNA: Ribonucleic Acid

- Nucleic Acid of long chain of units named nucleotides: Nitrogenous Base, Ribose sugar, Phosphate
- Adjacent nucleotides linked together by strong (covalent) phosphodiester bonds between sugar and phosphate
- Adjacent nucleotides linked together by strong (covalent) phosphodiester bonds between sugar and phosphate
- Information encoded with 4 different types of nucleotides differentiated by base content: Adenine, Guanine, Cytosine, Uracil

- RNA and DNA
  - `http://www.genome.gov`



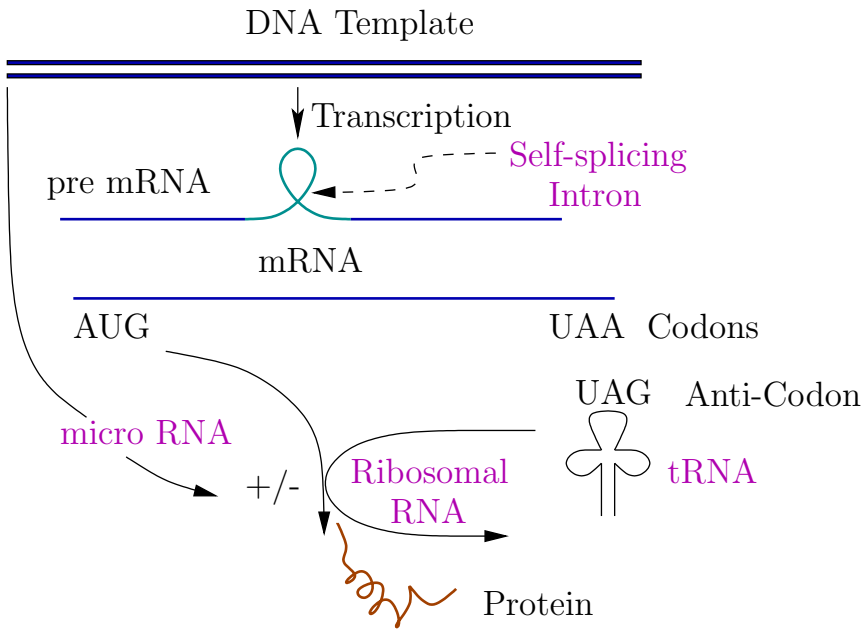
The Central Dogma



- Genetic information flows unidirectionally:
  - DNA $\rightarrow$ RNA $\rightarrow$ Protein
- RNA plays a passive role
  - Transient copy created for protein synthesis

`http://www.genome.gov`

RNA an Active Player: ncRNAs



- ncRNAs

- ncRNAs: play direct functional roles in cellular processes
  - w/o translation to protein $\Longrightarrow$ "noncoding"
- Increasing numbers (being) discovered
- 1989 Nobel Prize in Chemistry: Ribozymes
  - Thomas Cech and Sidney Altman
- 2006 Nobel Prize in Physiology/Medicine: siRNA
  - Andrew Fire and Craig Mello

Noncoding RNAs (ncRNAs): Examples

- Commonly known ncRNAs
  - Protein synthesis: tRNA, rRNA
- RNA modification: snoRNAs
- Up/Down regulation of gene expression
  - Regulation of transcription
  - siRNA/miRNA post transcription regulation silencing of genes
- piRNAs regulation of retroransposons
- RNA Splicing (autocatalysis)
- Many more: ...
- RNA Genomes (Many viruses including HIV and SIV)
- ncRNAs and diseases
  - Abnormal expression for ncRNAs observed in cancerous cells
  - Prader-Willi Syndrome (over-eating and learning disabilities)
  - Autism, Alzheimer's, ...

Noncoding RNAs (ncRNAs)

- RNA molecules that directly play functional roles in cellular processes
  - Do not code for protein synthesis $\implies$ "noncoding"
- Structure determines function in noncoding roles
- Determination of structure is of significant interest
  - Further understanding of ncRNA function
  - Enhances understanding of cellular processes and interactions
  - Provides targets for drug design

Computational Prediction of RNA Structure

- Structure determines function in noncoding roles
- Experimental determination of structure is challenging
  - X-ray Crystallography
  - Crystallization difficult and expensive
- Computational estimation of structure is of significant interest
  - Understanding ncRNA function in cellular processes and interactions
  - Genome understanding: structure based ncRNA gene search
  - Therapeutics: targets for drug design

RNA Structure Hierarchy [11]

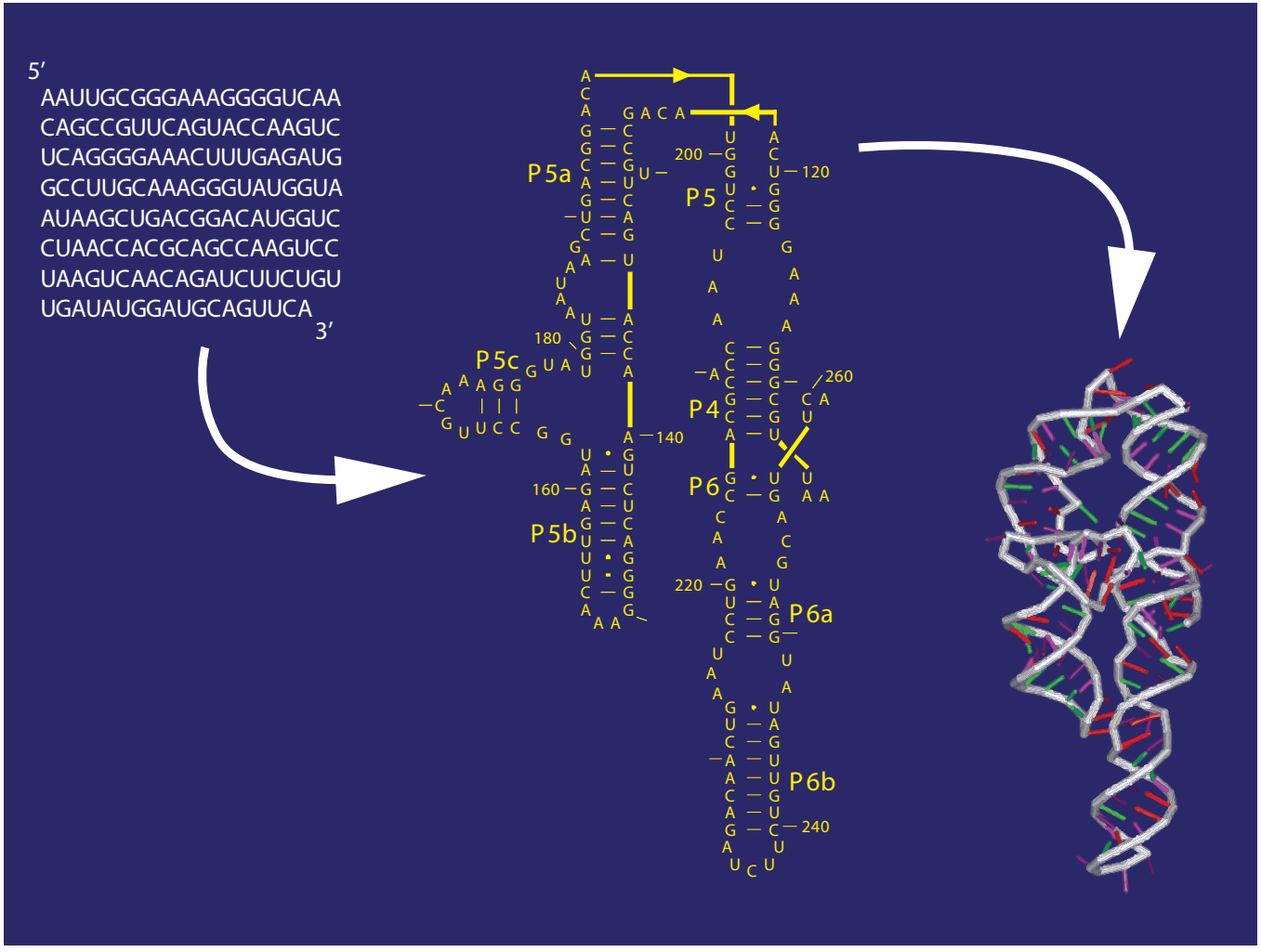


Figure: Hierarchy of RNA structure formation [12, 3, 4]

RNA Secondary Structure

- Folding of RNA linear molecular chain onto itself with base pairing rules
- Formation of hydrogen bonds between nucleotides
  - Canonical base pairs
    - A can pair with U
    - G can pair with C and U
    - G-U pair called non Watson-Crick pair
- Greater variety of structures than the DNA double helix

RNA Secondary Structure Elements

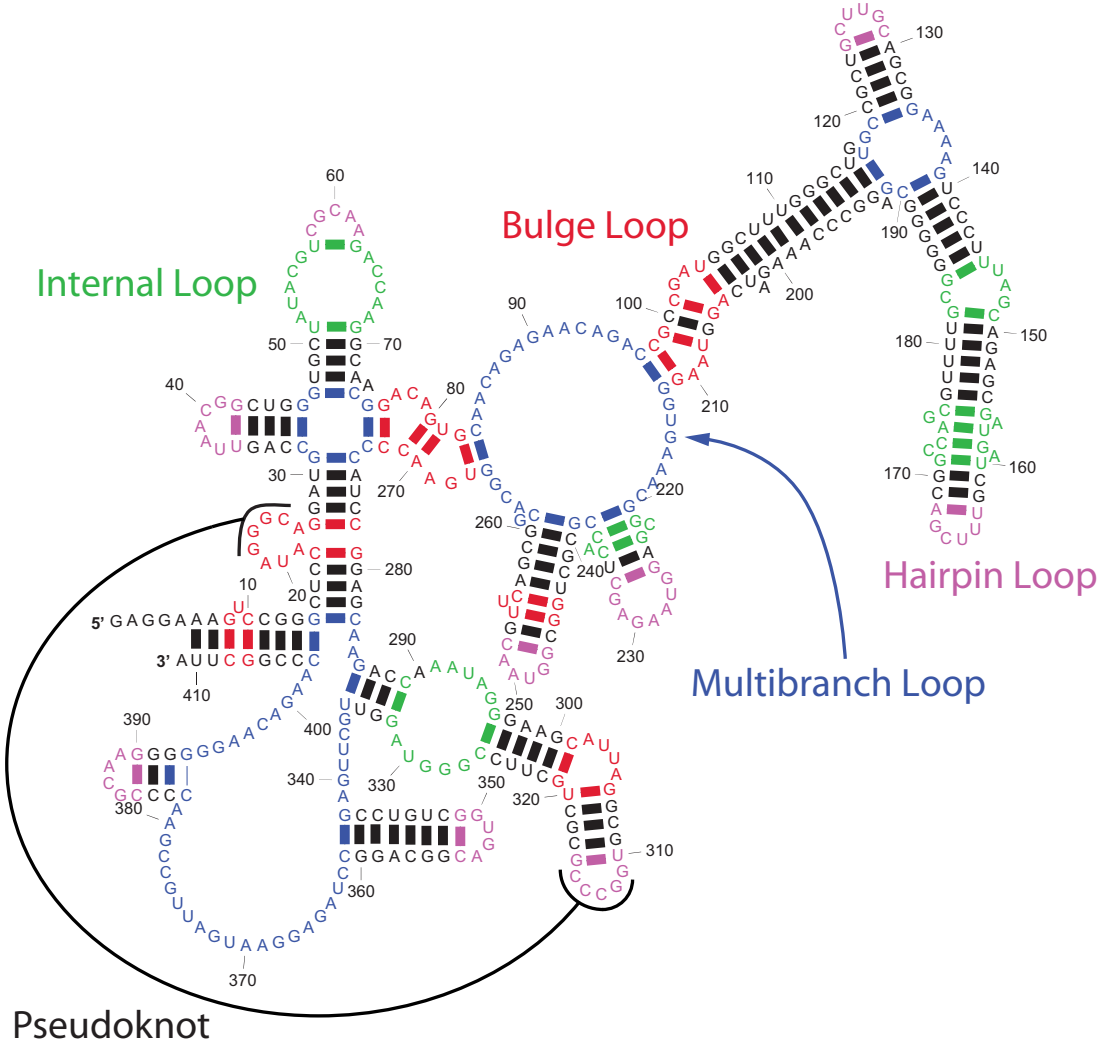
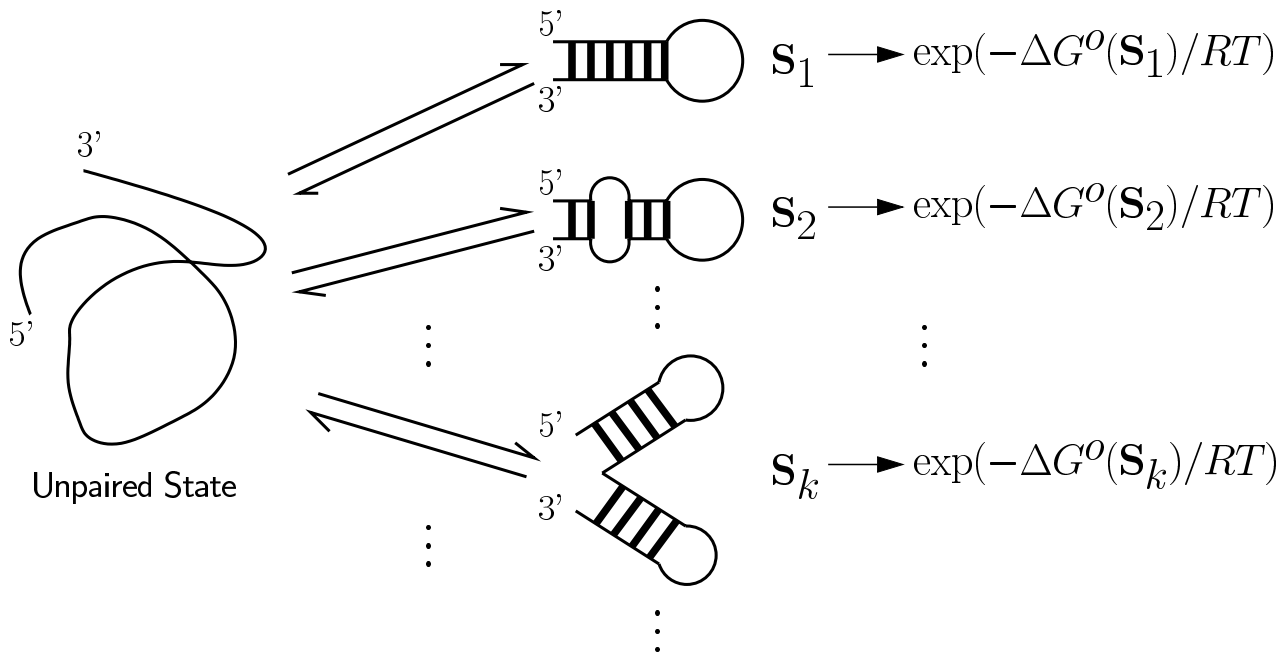


Figure: Structural Elements of LGW17 sequence from RNAse P Database [1]

RNA Structure: Thermodynamics

- Equilibrium: Boltzmann Distribution of structures



- Lower $\Delta G^o(\mathrm{S}_k)$, higher the probability of S_k
- Most likely structure $\rightarrow$ Minimization of free energy

Modeling RNA Thermodynamics: Nearest Neighbor Model

- Nearest neighbor model [13, 8]
  - Computational model for free energy change of RNA structure
  - Experimentally determined free energy terms for each nearest neighbor interaction in secondary structure
    - Loop decomposition

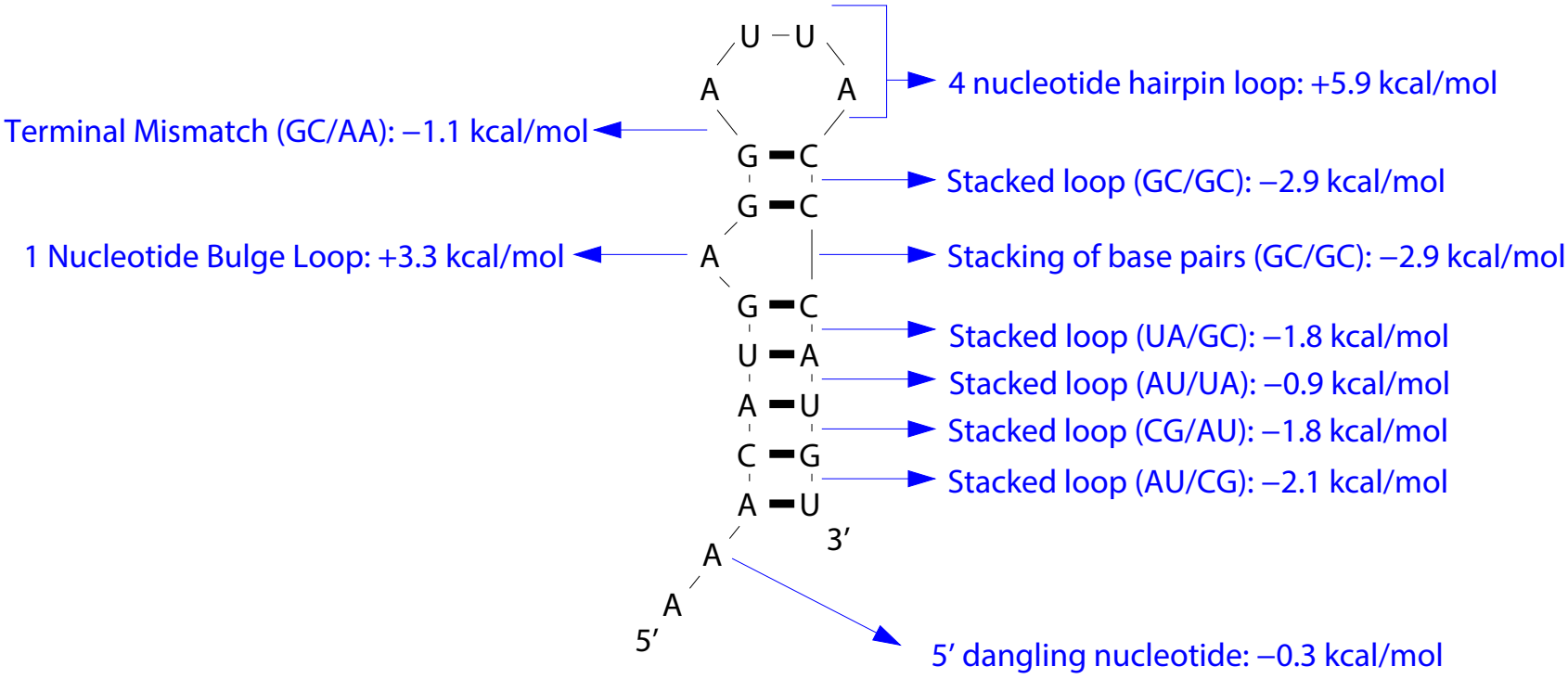
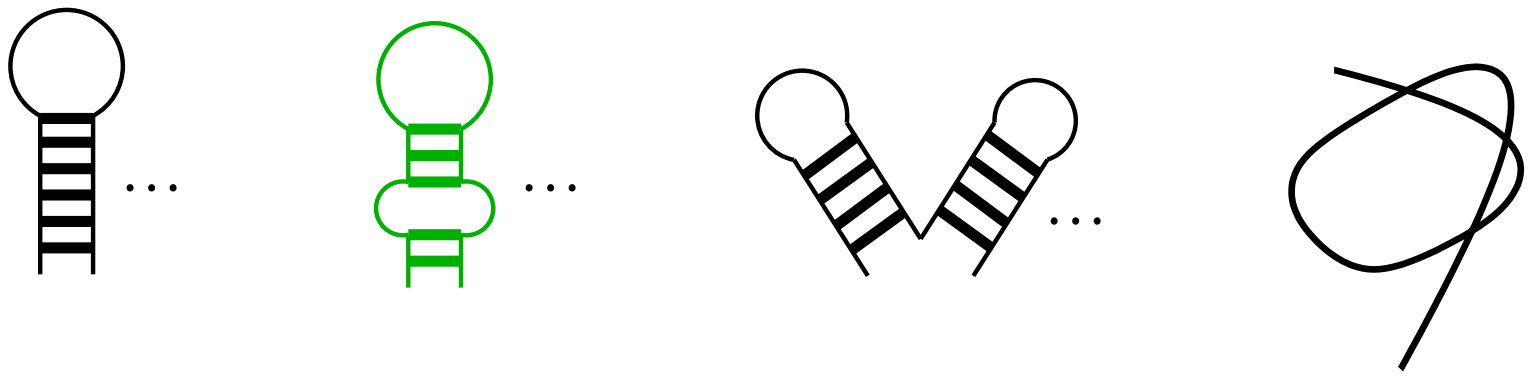


Figure: Total free energy change is summation of all nearest neighbor energies [5]

RNA ML Decoding of Structure: Single Sequence

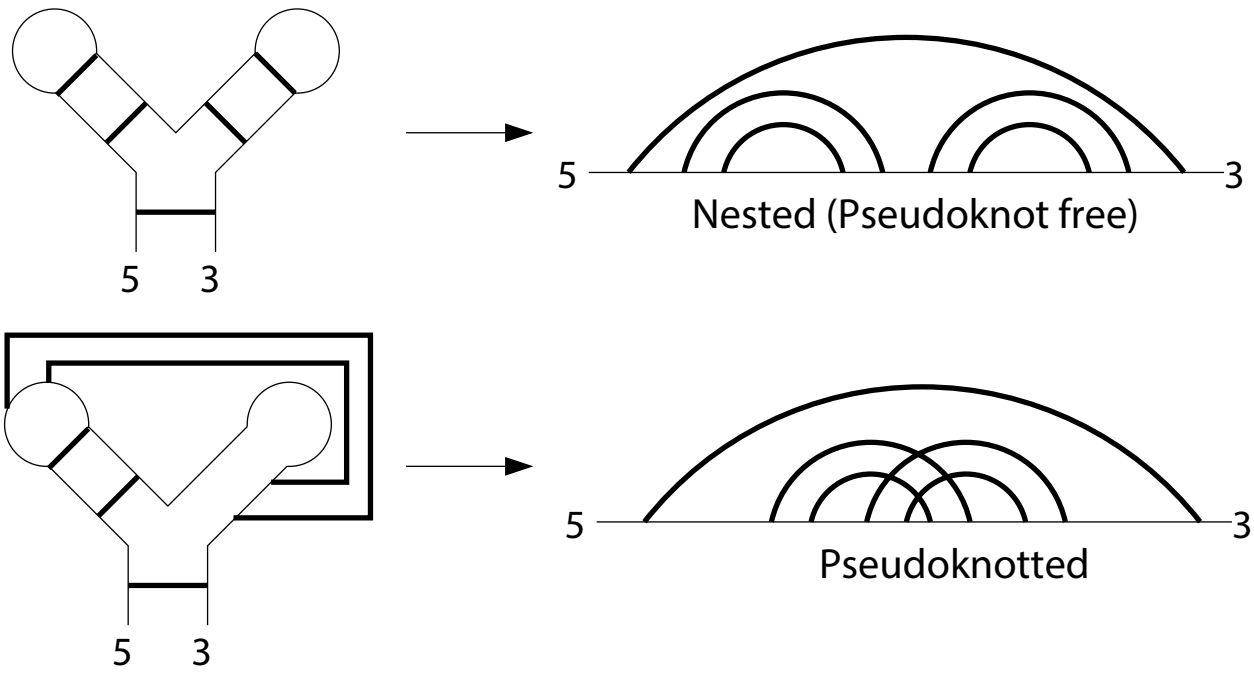
- Most likely or minimum free energy structure, given sequence



- Dynamic Programming MFold [14] $O(N^3)$ complexity

Min. Free Energy Structure Prediction: Dynamic Programming

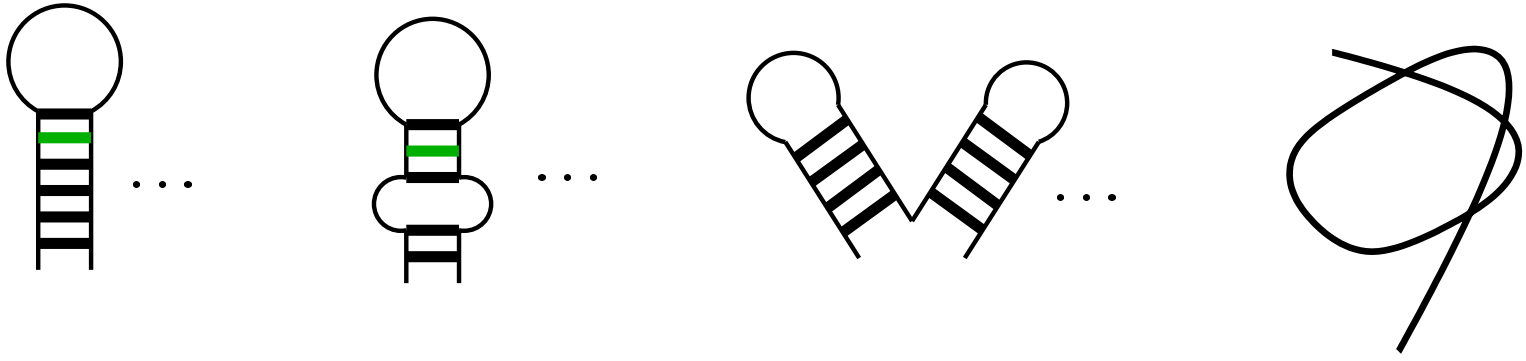
- Pseudoknot free structures
  - Separates computation into inside and outside segments



- MFold [14]: Free energy minimization
  - 73% accuracy in prediction of base pairs
- Estimate base pairing probabilities: McCaskill's Algorithm [9]

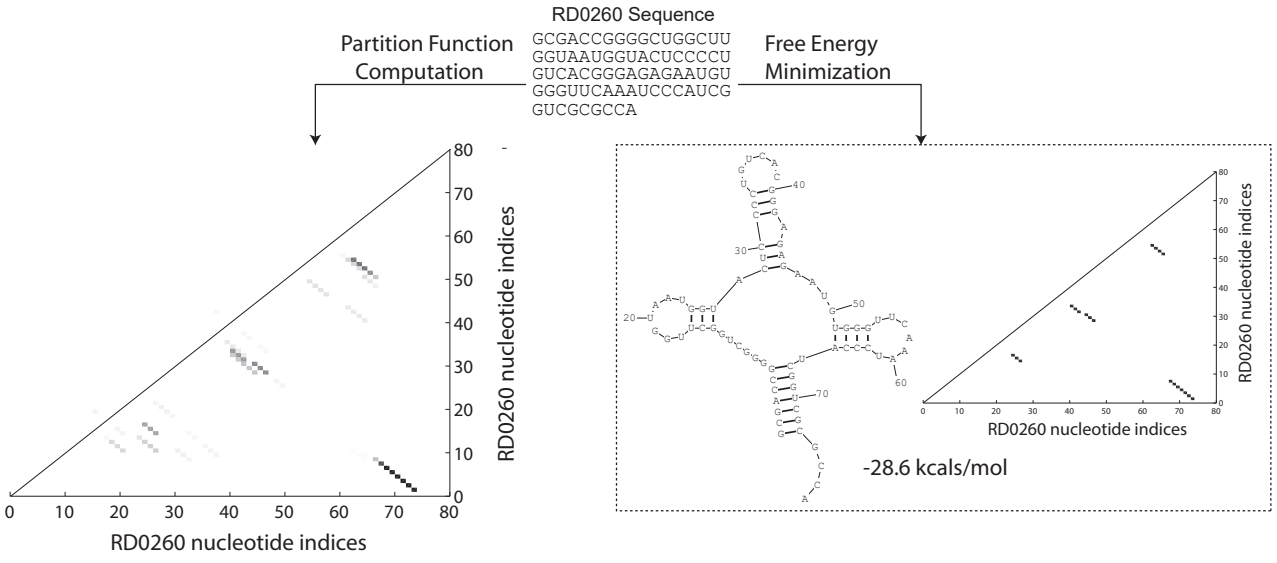
RNA MAP Decoding of Structure

- Posterior probability of base pairing, given sequence



- Dynamic Programming [9], MFold, RNAfold ($O(N^3)$ in time, $O(N^2)$ in space)
- Localized probabilistic information

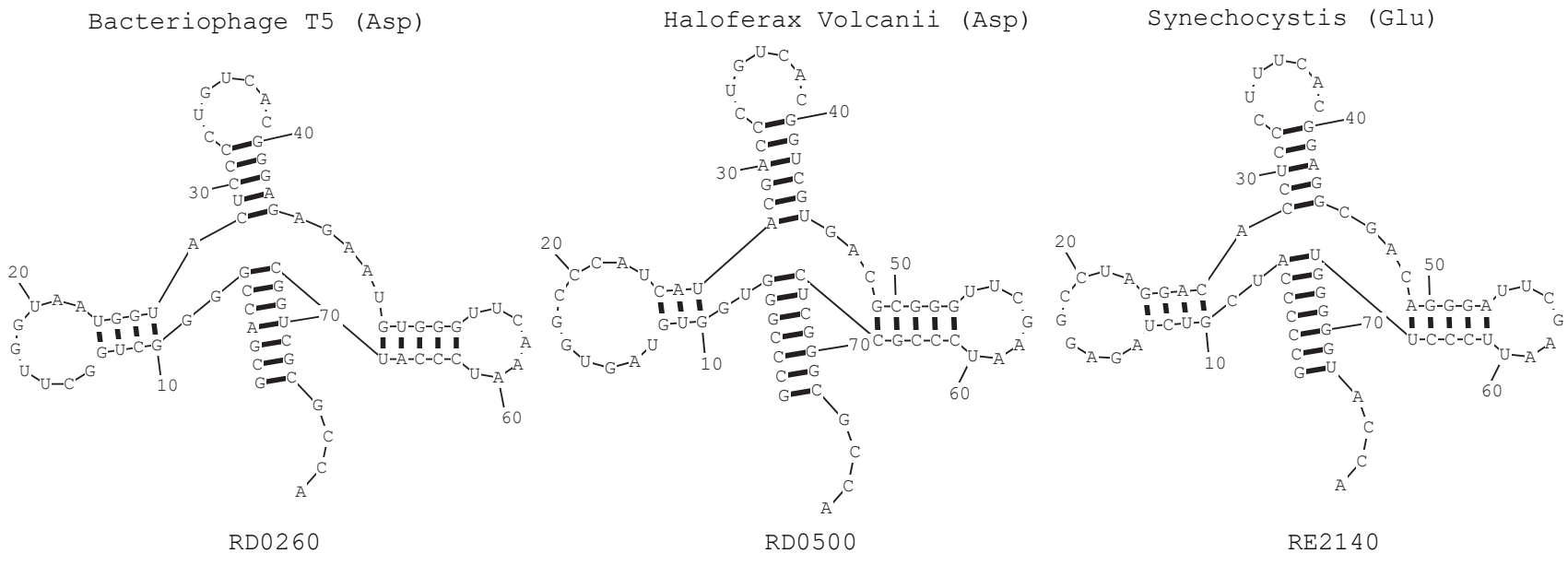
RNA Structure Prediction (Single Sequence)



- Free energy minimization: "Hard" Prediction
  - Single prediction structure
- Base pairing probabilities: "Soft" Prediction
  - Thresholding may yield pseudo-knotted structures
  - Maximum Expected Accuracy Structure Prediction [2, 7]

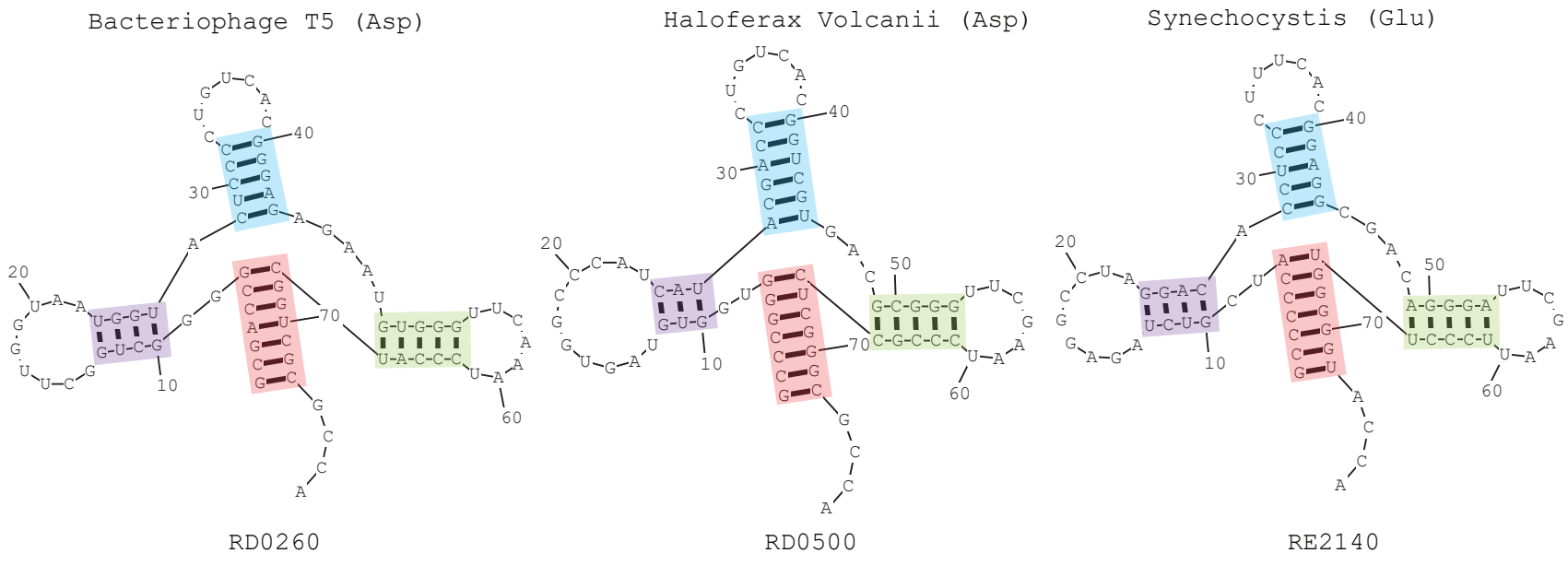
Structure Prediction for Multiple Sequences: Homologous ncRNAs

- Homologous ncRNAs
  - Share evolutionary ancestor
  - Serve same function
  - Structural similarity in terms of topology of structures

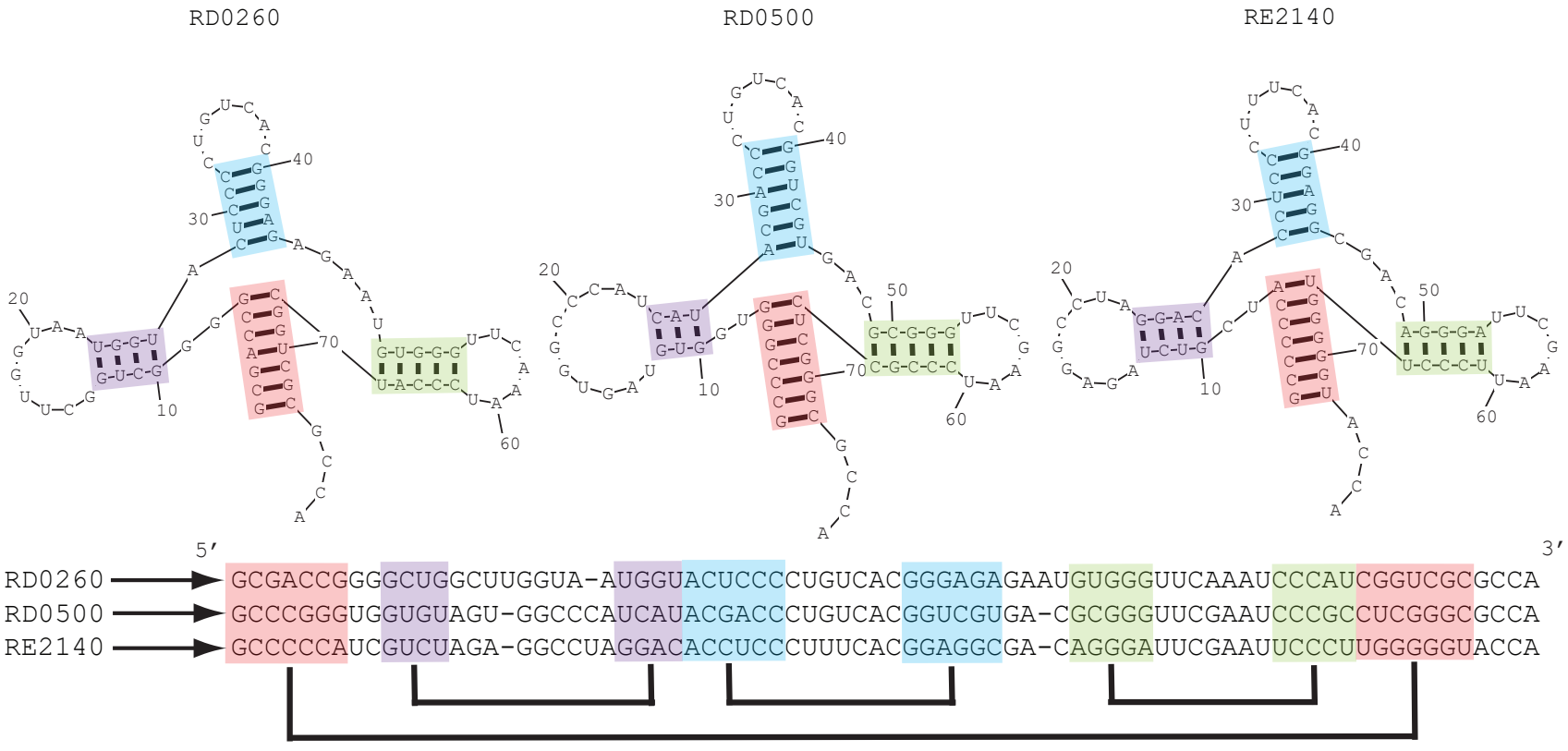


Structure Prediction for Multiple Sequences: Homologous ncRNAs

- Homologous ncRNAs
  - Share evolutionary ancestor
  - Serve same function
  - Structural similarity in terms of topology of structures



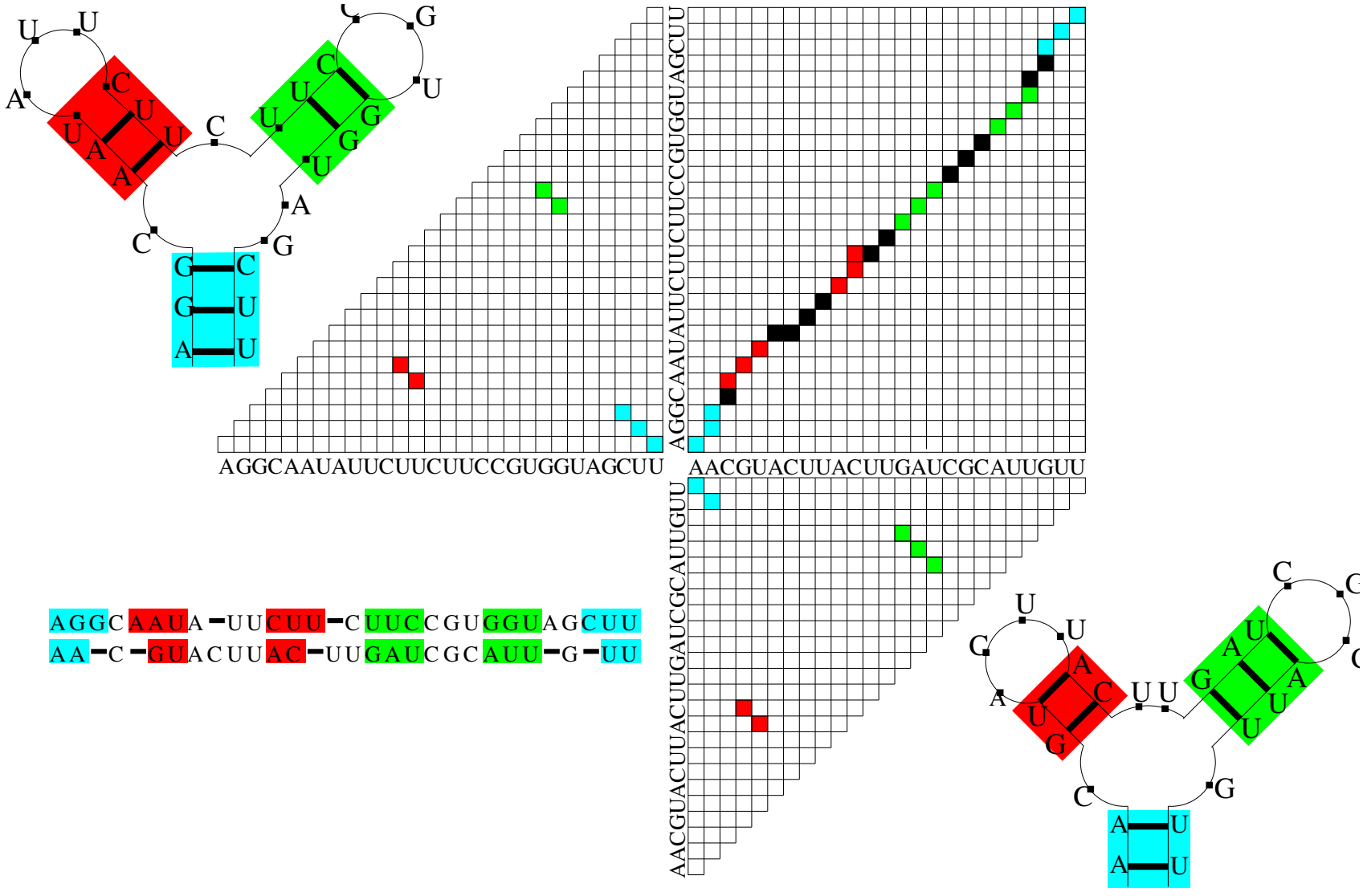
Structure Prediction for Multiple Sequences: Homologous ncRNAs



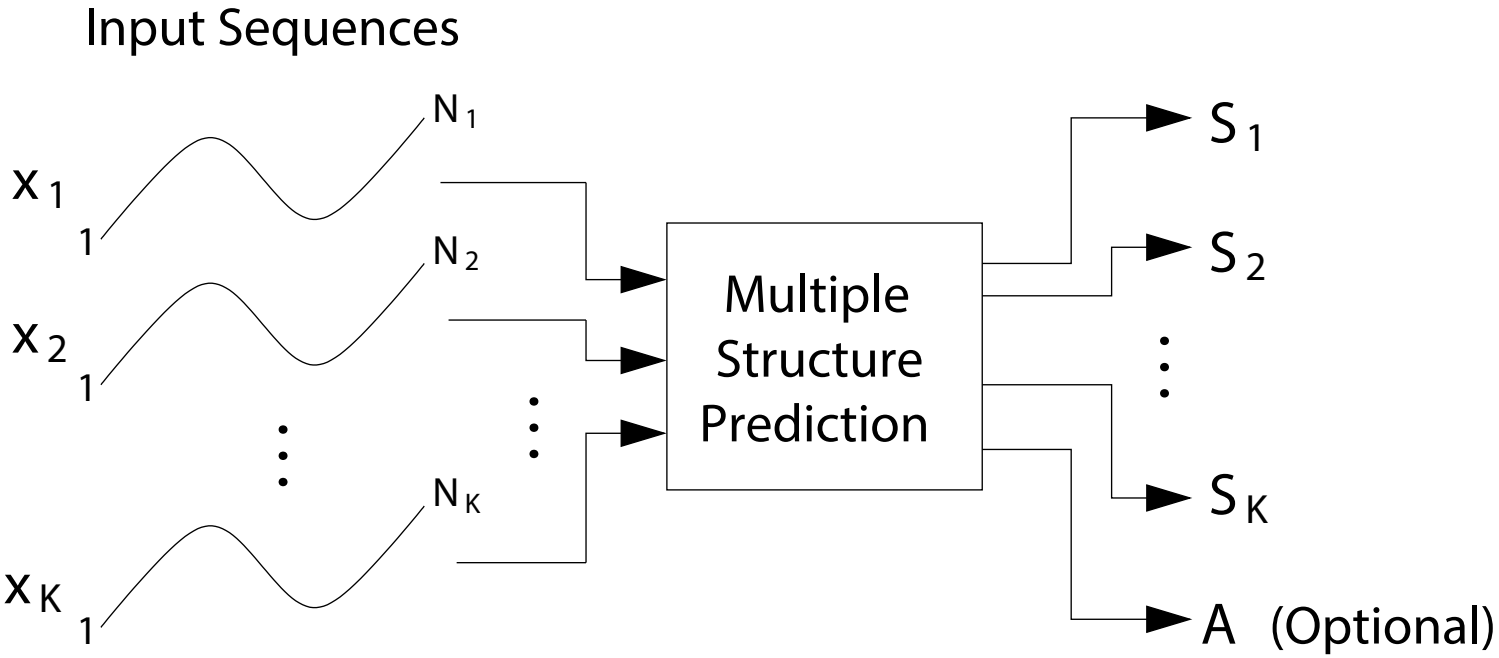
- "Common" structures and conforming sequence alignment
- Joint estimation can harness comparative structure and sequence information across homologs

RNA Structural Alignment Representation

- Conforming structures and alignment



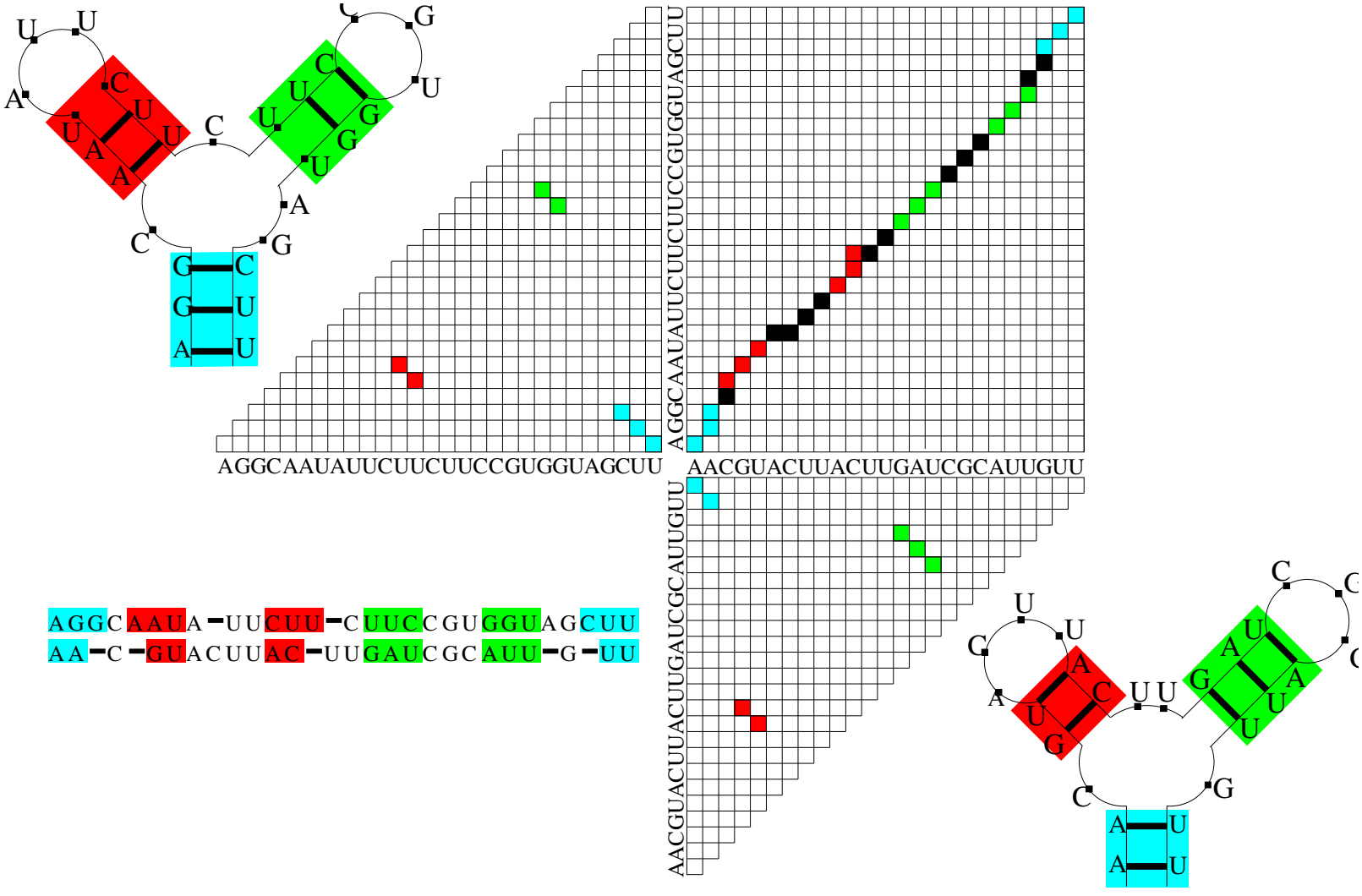
Multiple Sequence RNA Structure Prediction



- Sankoff's dynamic programming algorithm [10]
  - Simultaneous folding (pseudo-knot free) and alignment of K sequences
  - Time (Memory) complexity: $O(N^{3K})$ $(O(N^{2K}))$
  - Computationally infeasible even for short sequences and $K = 2$ w/o cutting corners

Corner Cutting for RNA Structural Alignment

- Limit search space
  - Our focus: restricted space of allowable alignments



Banded Constraint Corner Cutting

- Filter out space of aligments that deviate from a diagonal band of width L

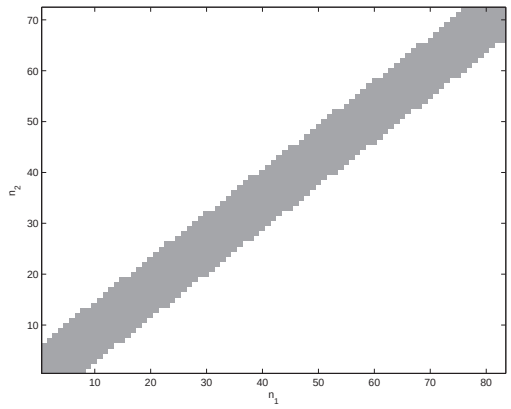


Figure: Diagonal band constraint set.

- Challenge hard to estimate width L of diagonal band
- For homologous sequences with long inserts, require large L, computationally demanding
- An un-informed heuristic

Alternative: Smarter Constraint Using Posterior Probabilities

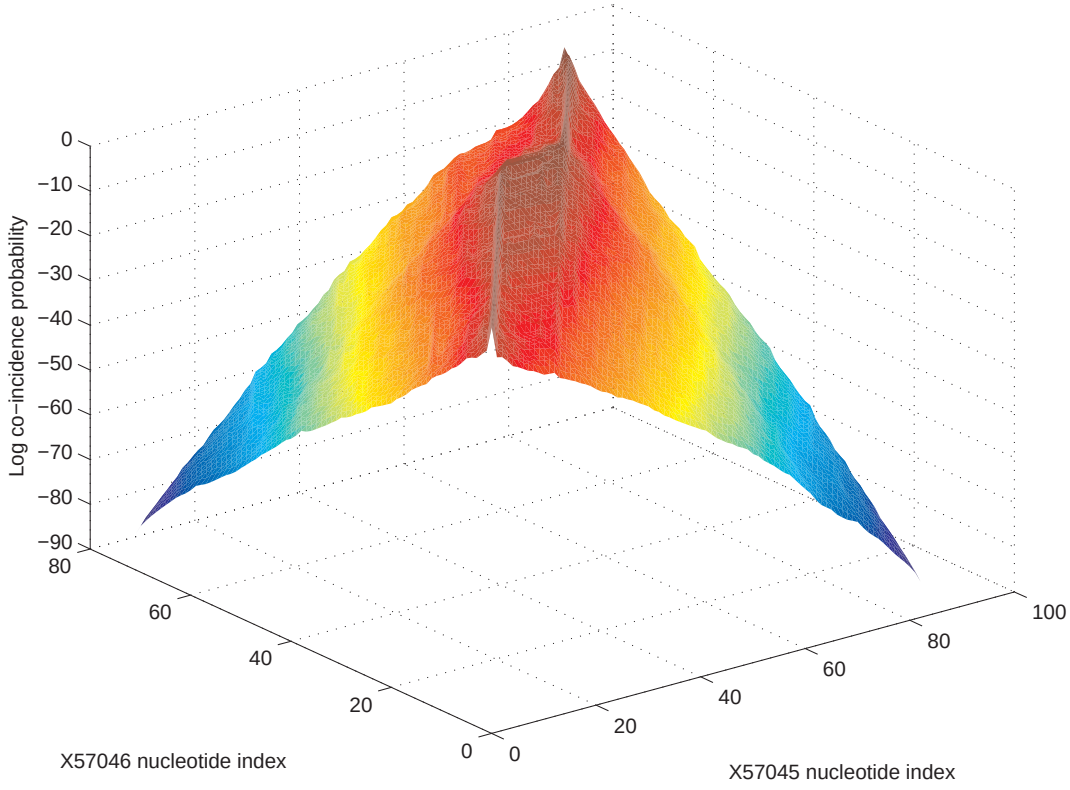
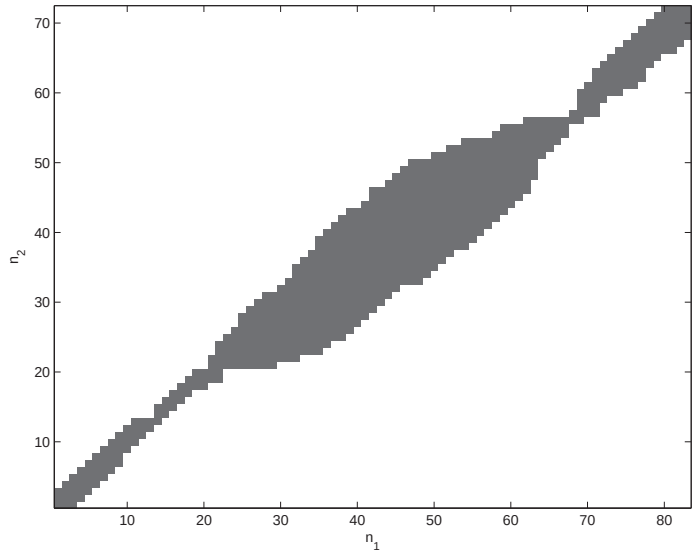


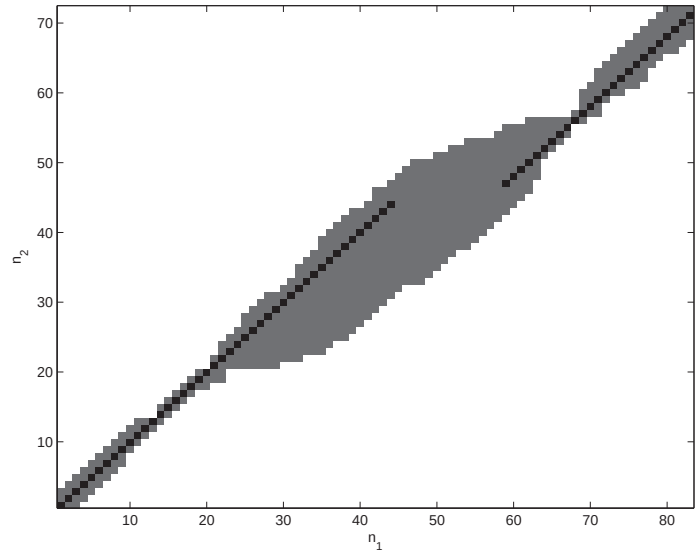
Figure: Logarithm of posterior probabilities for co-incidences of nucleotide positions for a pair of sequences in a surface plot representation.

- Exclude probabilities below a small threshold
  - Eliminate regions of alignment space that are anyway improbable

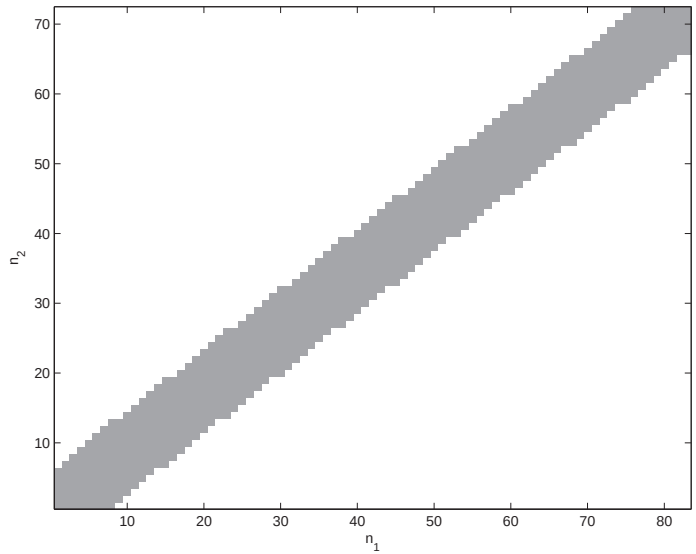
Alignment Constraint Set: Probabilistic vs. Banded



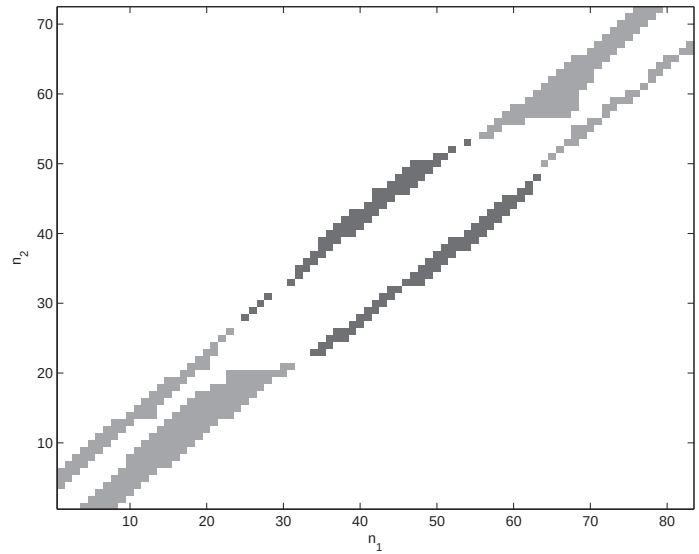
(a) Probabilistic constraint set



(b) Prob. constr. & true align



(c) Banded (M) constraint set



(d) Diff. prob. and banded

Use of Posterior Probabilities for an Alignment Envelope

		tRNA		5S RNA	
		New constraint	M constraint	New constraint	M constraint
Memory	avg	10.988	10.960	12.377	14.306
	min	9.36	10.528	10.176	12.664
	max	13.592	12.040	17.436	14.840
Timing	avg	9.98	9.39	34.38	73.07
	min	1.0	4.0	20.0	36.0
	max	55.0	29.0	234.0	111.0

Table: Average, minimum, maximum run times (in seconds) and memory (in megabytes) requirement results of proposed probabilistic alignment constraint (New constraint) and the previously employed M constraint. A dual-core AMD Opteron®-270 2.0 GHz system with 4 GBytes of main memory running Linux Fedora Core 4 was utilized for the timing experiments.

Use of Posterior Probabilities for an Alignment Envelope

	Dynalign		**Single Sequence Prediction**
	New constraint	***M* constraint**	
Sensitivity	0.907	0.905	0.739
PPV	0.821	0.817	0.647

Table: Structural prediction accuracy of Dynalign with M constraint, new constraint and single prediction over 2000 random 5S RNA pairs.

Benefits of Probabilistic alignment constraints [6]

- Reduces time and memory: adaptively
  - On average 2-fold decrease in run-time with marginal increase in accuracy
  - Contrast with prior banded constraints: "principled heuristic"
    - Concentrates computation where required and eliminates where it is likely unnecessary

Summary HMMs for String Alignment

- Example illustrates how nuances of HMM formulation vary from application to application
  - particularly, latent states allows us to address different number of elements in the two strings
- Probabilistic model using HMMs enables richer inference than purely heuristic string edit distance
  - Specific example, posterior probabilities for two positions in the two strings being aligned
  - Useful, when alignment is part of a larger problem that is computationally expensive
    - Can pre-filter and rule out extremely improbable alignment possibilities, improving computational efficiency
  - Also useful for iterative schemes, one of which we will see at a high level

References I


J. W. Brown. "The Ribonuclease P Database". In: *Nucleic Acids Res.* 27.1 (1999), p. 314.


Chuong B. Do, Daniel A. Woods, and Serafim Batzoglou. "CONTRAfold: RNA secondary structure prediction without physics-based models". In: *Bioinformatics* 22.14 (2006), pp. 90–98.


J. A. Doudna and T. R. Cech. "The chemical repertoire of natural ribozymes". In: *Nature* 418.6894 (2002), pp. 222–228.


J.A. Doudna and J.H. Cate. "RNA structure: crystal clear?" In: *Current Opinions in Structural Biology* 7 (1997), pp. 310–316.


R. Durbin et al. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press, 1999. ISBN: 0521629713.

References II

- A. O. Harmanci, G. Sharma, and D. H. Mathews. "Efficient Pairwise RNA Structure Prediction Using Probabilistic Alignment Constraints in Dynalign". In: *BMC Bioinformatics* 8 (2007), p. 130. DOI: 10.1186/1471-2105-8-130.
- Zhi John Lu, Jason W. Gloor, and David H. Mathews. "Improved RNA secondary structure prediction by maximizing expected pair accuracy". In: *RNA* 15.10 (2009), pp. 1805–1813.
- D. H. Mathews et al. "Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure". In: *J. Mol. Biol.* 288.5 (1999), pp. 911–940.
- J. S. McCaskill. "The equilibrium partition function and base pair binding probabilities for RNA secondary structure". In: *Biopolymers* 29.6-7 (1990), pp. 1105–1119.

References III


D. Sankoff. "Simultaneous Solution of RNA Folding, Alignment and Protosequence Problems". In: *SIAM J. App. Math.* 45.5 (1985), pp. 810–825.


I. Tinoco Jr. and C. Bustamante. "How RNA folds". In: *J Mol Biol* 293.2 (1999), pp. 271–281.


Richard B. Waring and R. Wayne Davies. "Assessment of a model for intron RNA secondary structure relevant to RNA self-splicing – a review". In: *Gene* 28.3 (1984), pp. 277–291.


T. Xia et al. "Thermodynamic Parameters for an Expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs". In: *Biochemistry* 37.42 (1998), pp. 14719–14735.


M. Zuker. "Computer prediction of RNA structure". In: *Methods Enzymol.* 180 (1989), pp. 262–288.