Probabilistic decoding in communications and bioinformatics: A turbo approach

Gaurav Sharma

Department of Electrical and Computer Engineering
Department of Computer Science
Department of Biostatistics and Computational Biology

UNIVERSITY of ROCHESTER

Outline

- Turbo-decoding in Communications: A Quick Review

Outline

- Turbo-decoding in Communications: A Quick Review
- RNA Structure Analysis: Motivation and Background
  - RNA, noncoding RNA, RNA structure and its significance
  - RNA structure prediction
    - Single/Multiple sequence methods

Outline

- Turbo-decoding in Communications: A Quick Review
- RNA Structure Analysis: Motivation and Background
  - RNA, noncoding RNA, RNA structure and its significance
  - RNA structure prediction
    - Single/Multiple sequence methods
- Turbo-decoding RNA secondary structure
  - Iterative probabilistic decoding of structures of multiple homologs: TurboFold
- Turbo-decoding: RNA vs communications

Outline

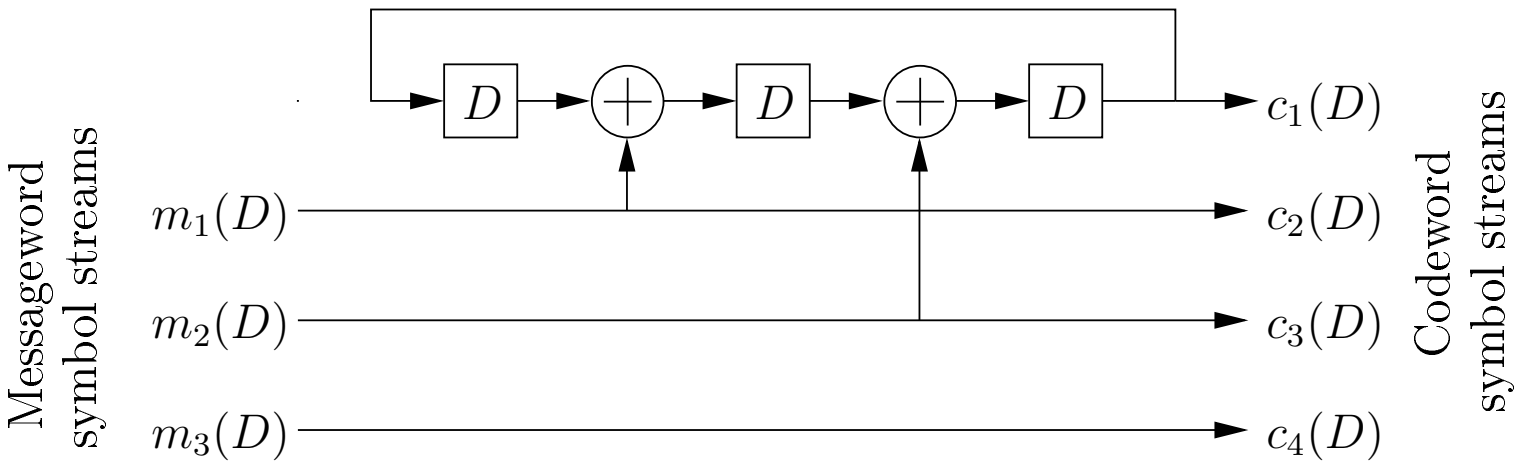
- Turbo-decoding in Communications: A Quick Review
- RNA Structure Analysis: Motivation and Background
  - RNA, noncoding RNA, RNA structure and its significance
  - RNA structure prediction
    - Single/Multiple sequence methods
- Turbo-decoding RNA secondary structure
  - Iterative probabilistic decoding of structures of multiple homologs: TurboFold
- Turbo-decoding: RNA vs communications
- Conclusions

Outline

- Turbo-decoding in Communications: A Quick Review
- RNA Structure Analysis: Motivation and Background
  - RNA, noncoding RNA, RNA structure and its significance
  - RNA structure prediction
    - Single/Multiple sequence methods
- Turbo-decoding RNA secondary structure
  - Iterative probabilistic decoding of structures of multiple homologs: TurboFold
- Turbo-decoding: RNA vs communications
- Conclusions
- Ongoing related work

Convolutional Codes

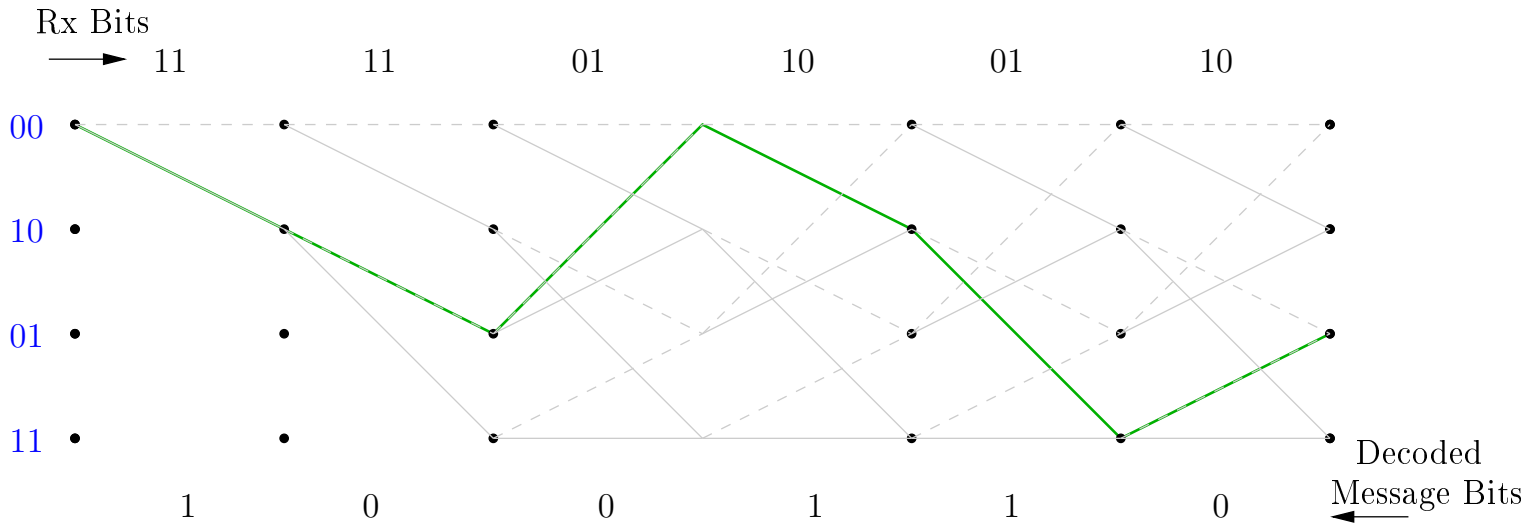
- Encoder



- Finite state machine
  - Output and next state are functions of current state and inputs

ML Decoding: Convolutional Code

- Convolutional code structure constrains possibilities to a trellis



- ML Decoding: Most likely path through the trellis given the received information

Turbo Decoding in Communications

- An encoder construction

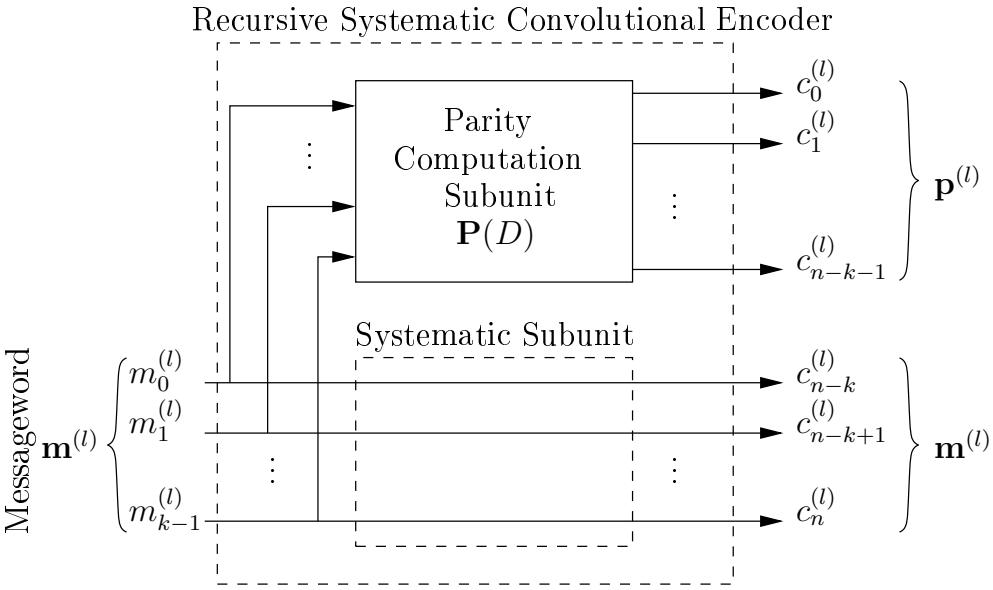


Figure: Systematic convolutional encoder (recursive)

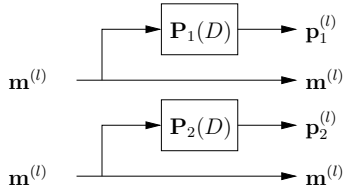


Figure: Two encoders.

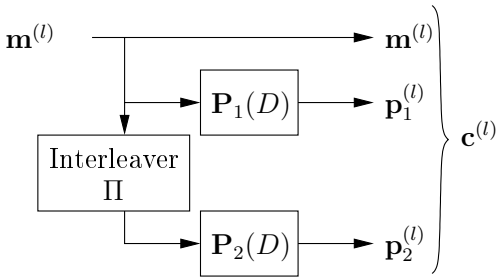


Figure: Parallel concatenation.

Turbo Decoding in Communications

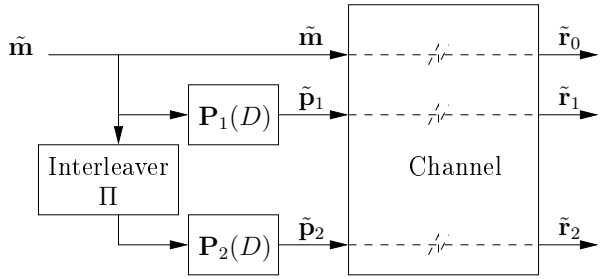


Figure: Encoder + channel

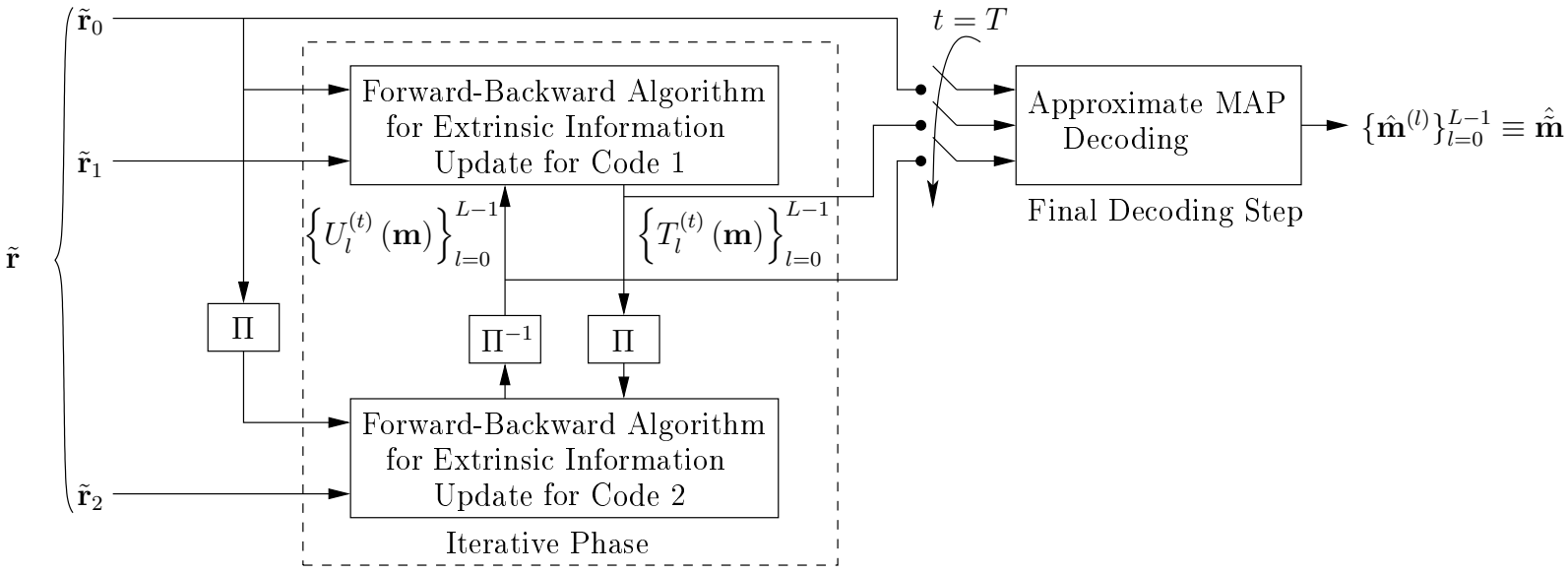
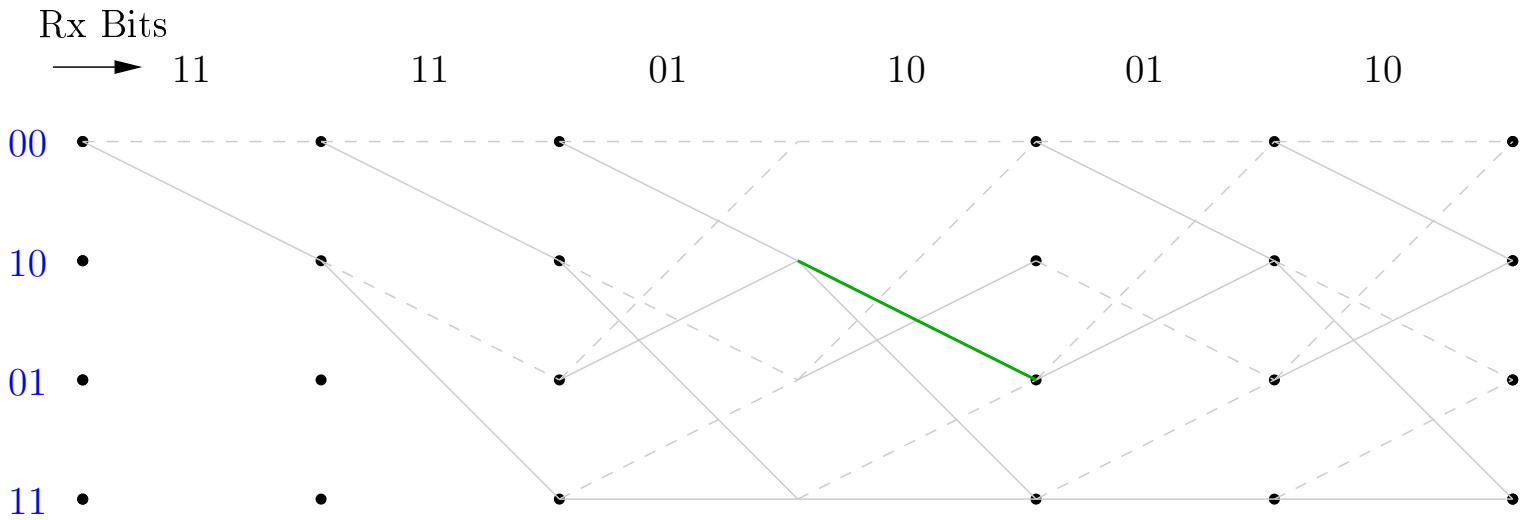


Figure: Iterative Decoder

Symbol-wise MAP Decoding: Convolutional Code

- Convolutional code structure constrains possibilities to a trellis



- Most probable value of a bit for all possible paths through the trellis, given the received information
- Localized probabilistic information

Turbo Decoding in Communications: Observations

- Multiple encodings of same message information
- Joint (optimal) decoding desirable
  - Exact joint decoding $\approx$ exponential complexity
  - Computationally Efficient Decoding: Iterative approximation (belief propagation)
    - Localized MAP probabilistic formulation
    - Decomposition into loosely coupled individual decodings + information exchange at each iteration
    - Linear complexity in length of data
    - Pseudo-prior interpretation

Turbo Decoding in Communications: Observations

- Multiple encodings of same message information
- Joint (optimal) decoding desirable
  - Exact joint decoding $\approx$ exponential complexity
  - Computationally Efficient Decoding: Iterative approximation (belief propagation)
    - Localized MAP probabilistic formulation
    - Decomposition into loosely coupled individual decodings + information exchange at each iteration
    - Linear complexity in length of data
    - Pseudo-prior interpretation

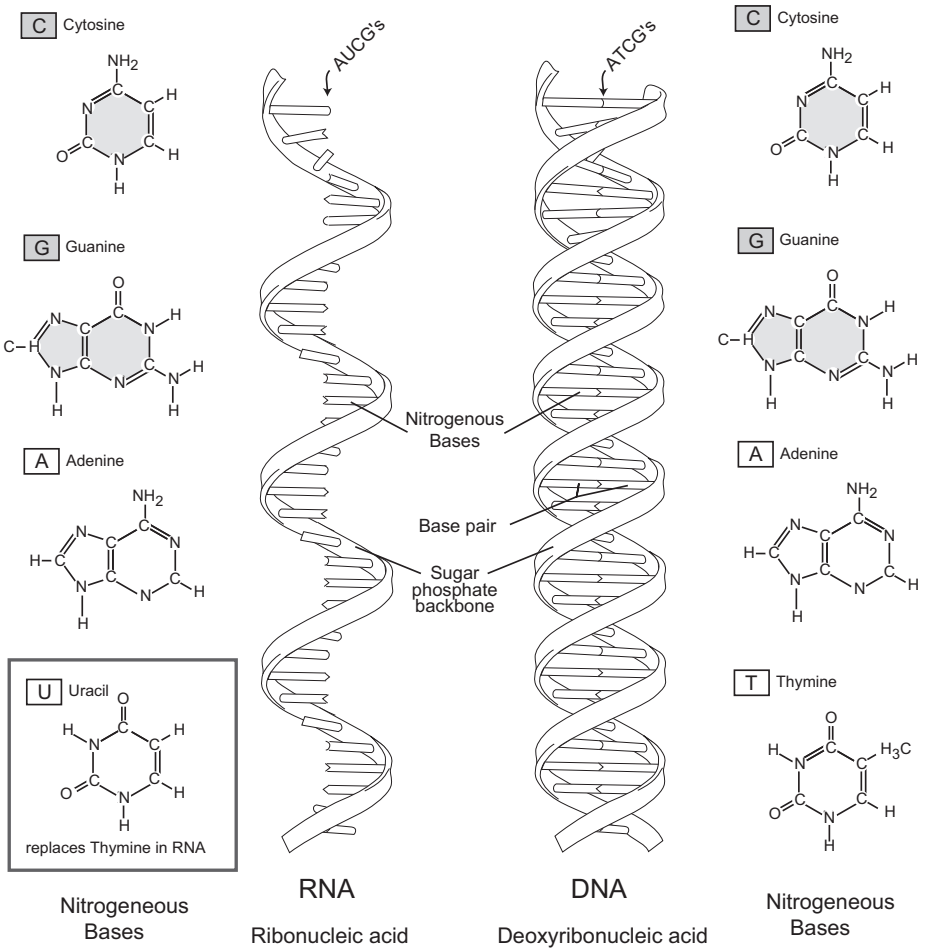
RNA?

What does this have to do with RNA?

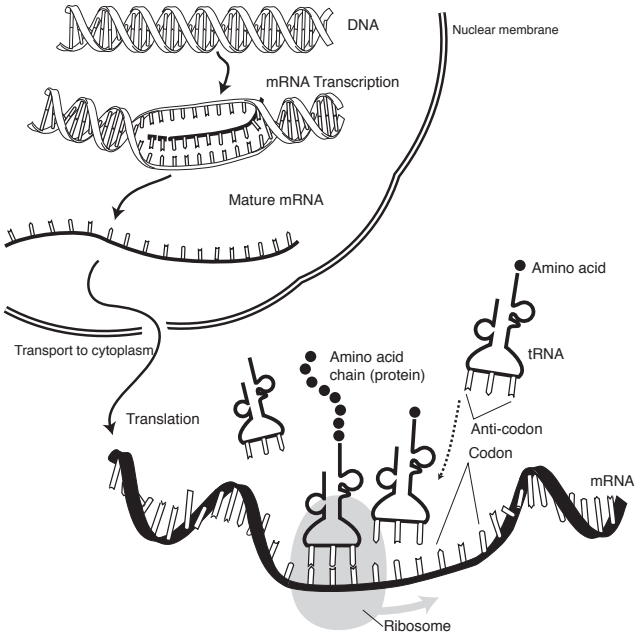
RNA: Ribonucleic Acid

- 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
- Information encoded with 4 different types of nucleotides differentiated by base content: Adenine, Guanine, Cytosine, Uracil

http://www.genome.gov



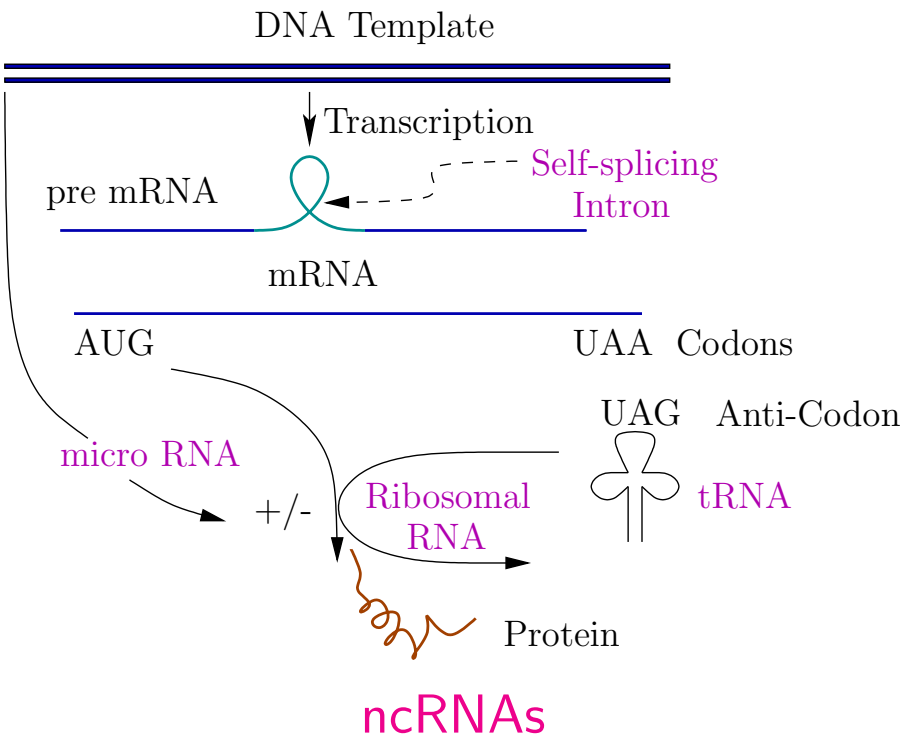
The Central Dogma



- 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: 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 $\Longrightarrow$ "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

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 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

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 [Tinoco and Bustamante, 1999]

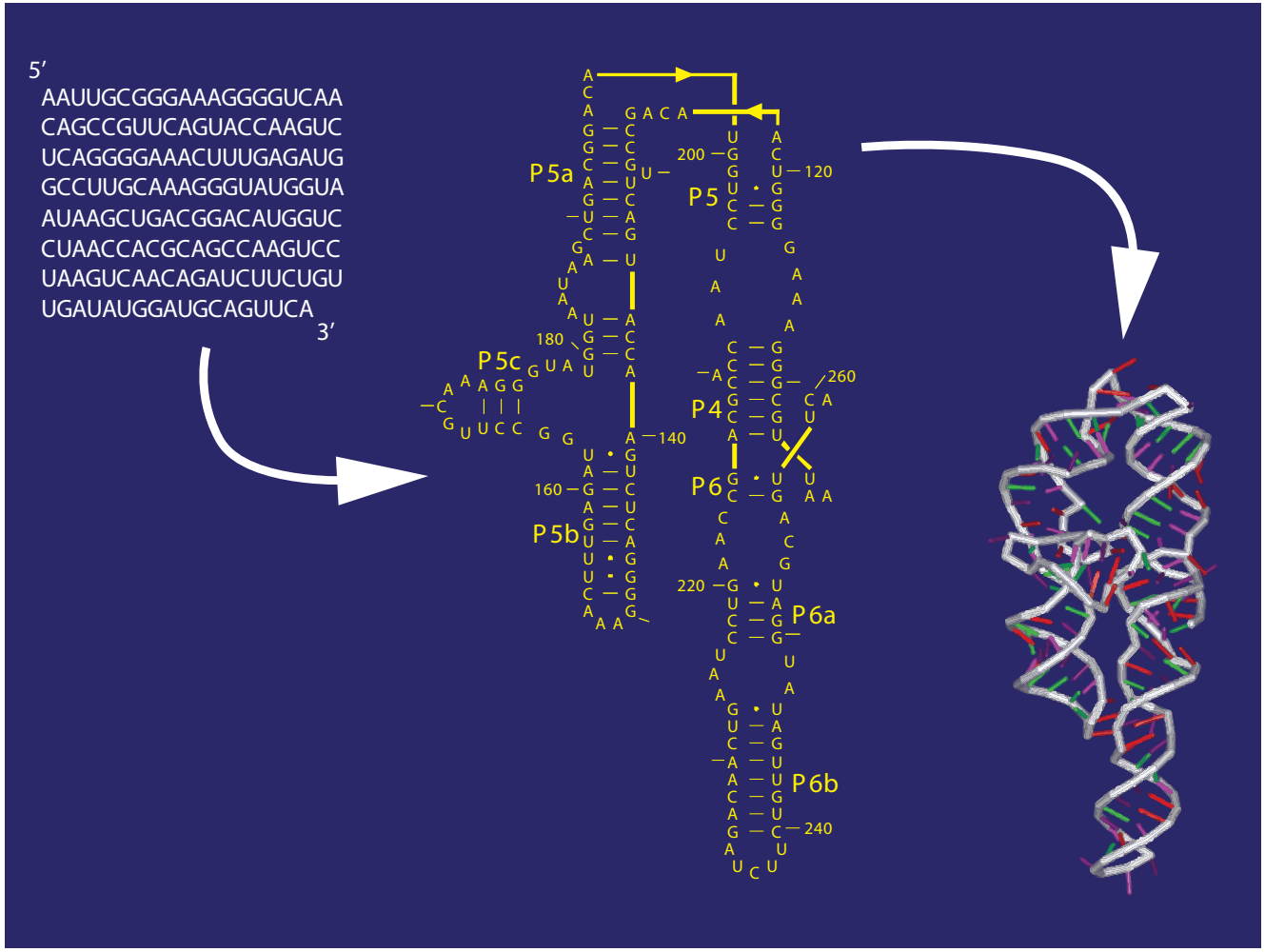


Figure: Hierarchy of RNA structure formation [Waring and Davies, 1984, Doudna and Cech, 2002, Doudna and Cate, 1997]

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

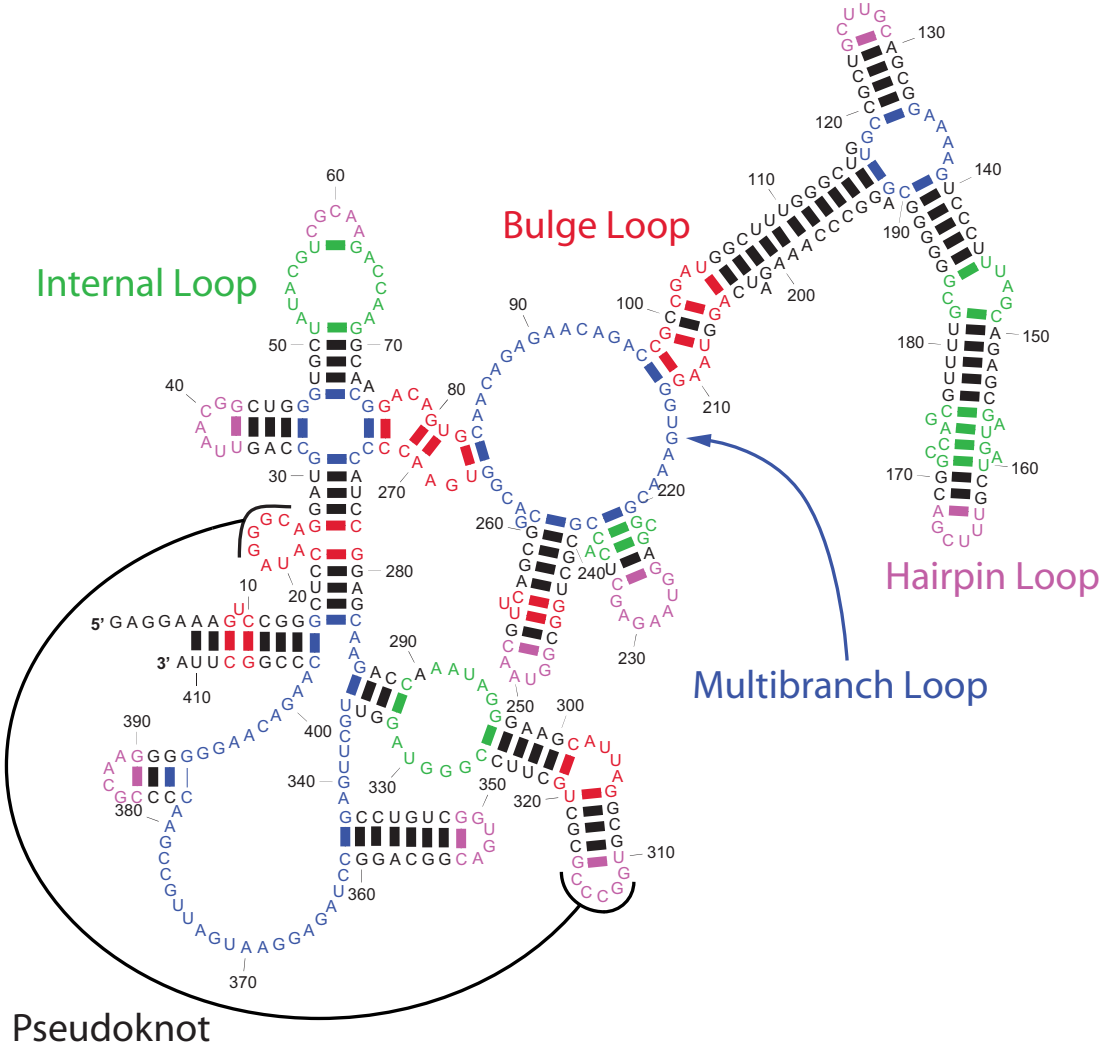
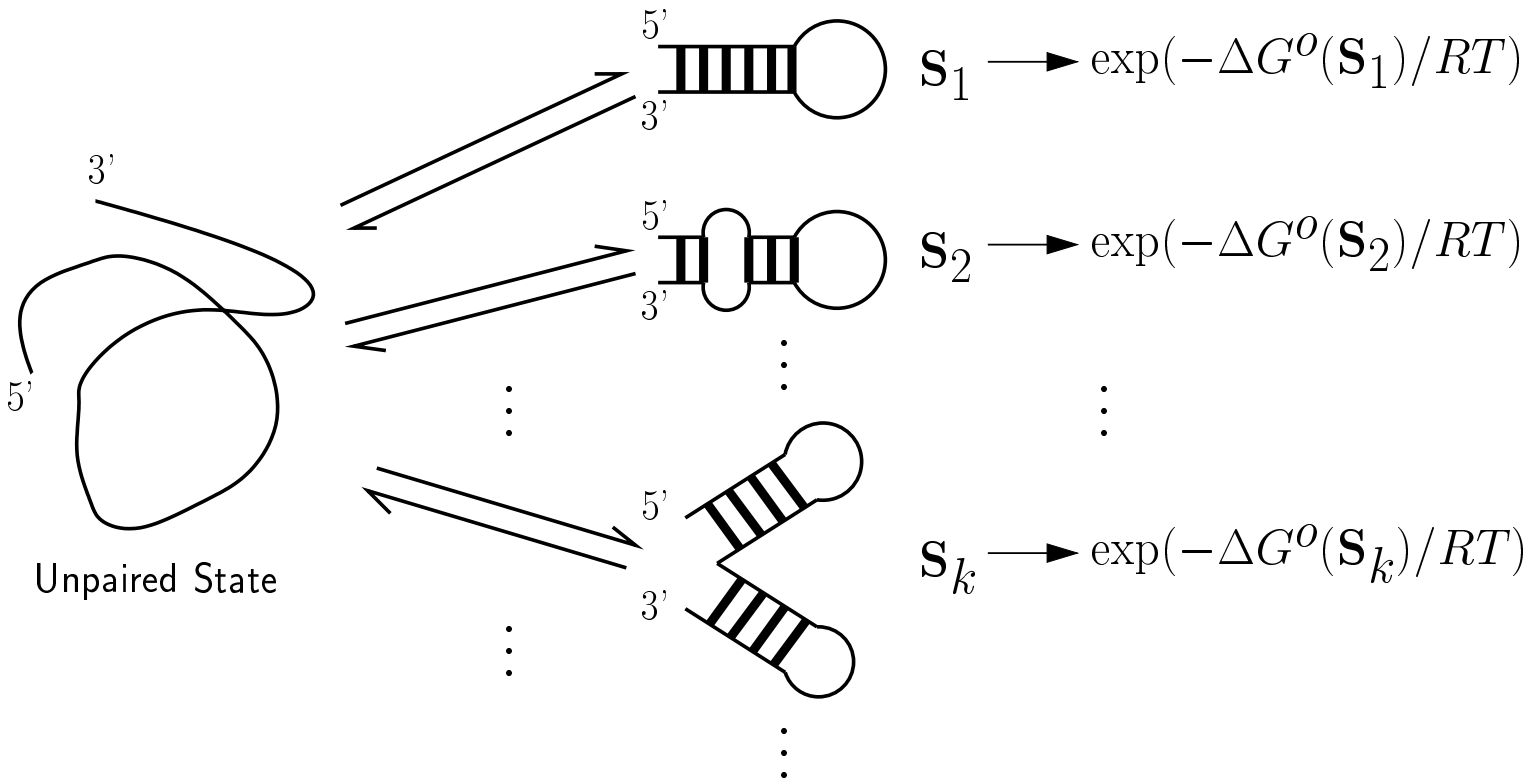


Figure: Structural Elements of LGW17 sequence from RNAse P Database [Brown, 1999]

RNA Structure: Thermodynamics

- Equilibrium: Boltzmann Distribution of structures



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

Modeling RNA Thermodynamics: Nearest neighbor model

- Nearest neighbor model [Xia et al., 1998, Mathews et al., 1999]
  - Computational model for free energy change of RNA structure
  - Experimentally determined free energy terms for each nearest neighbor interaction in secondary structure
    - Loop decomposition

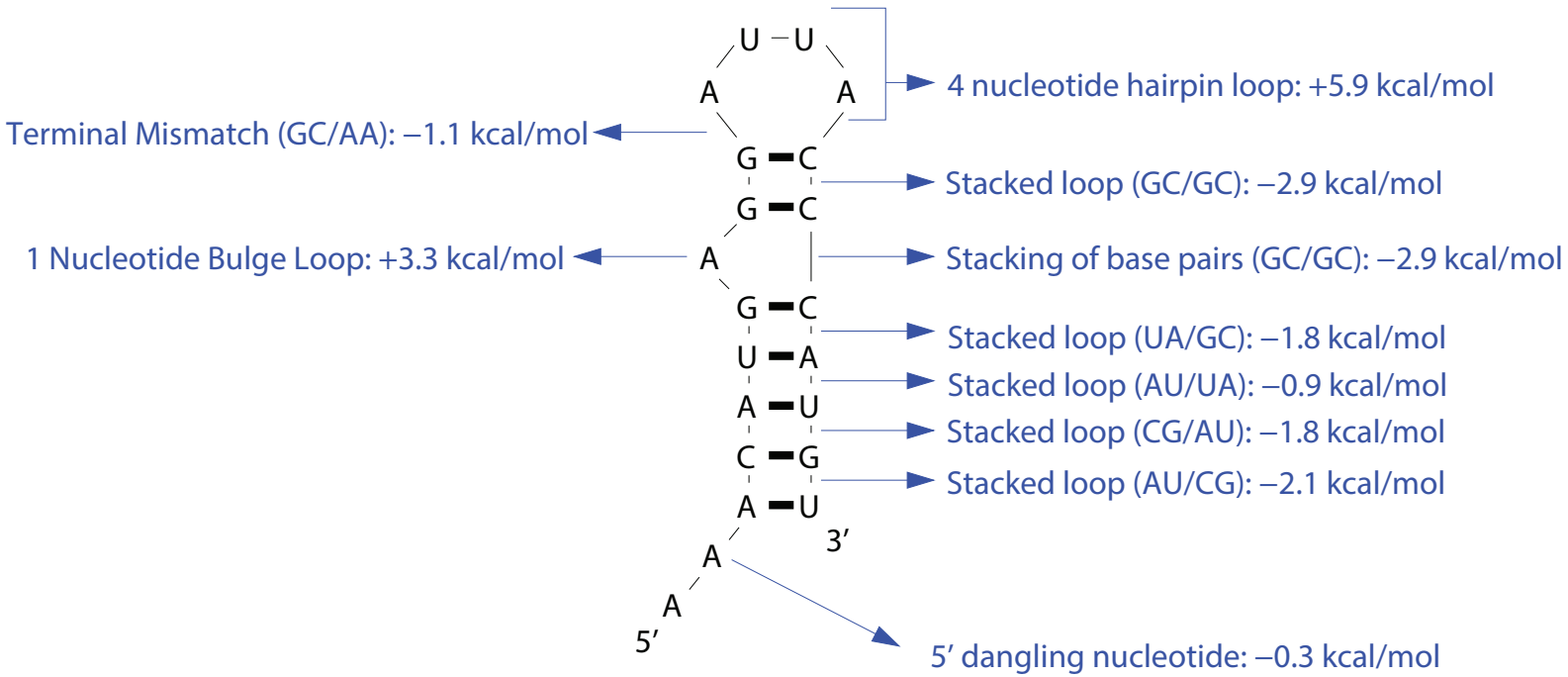


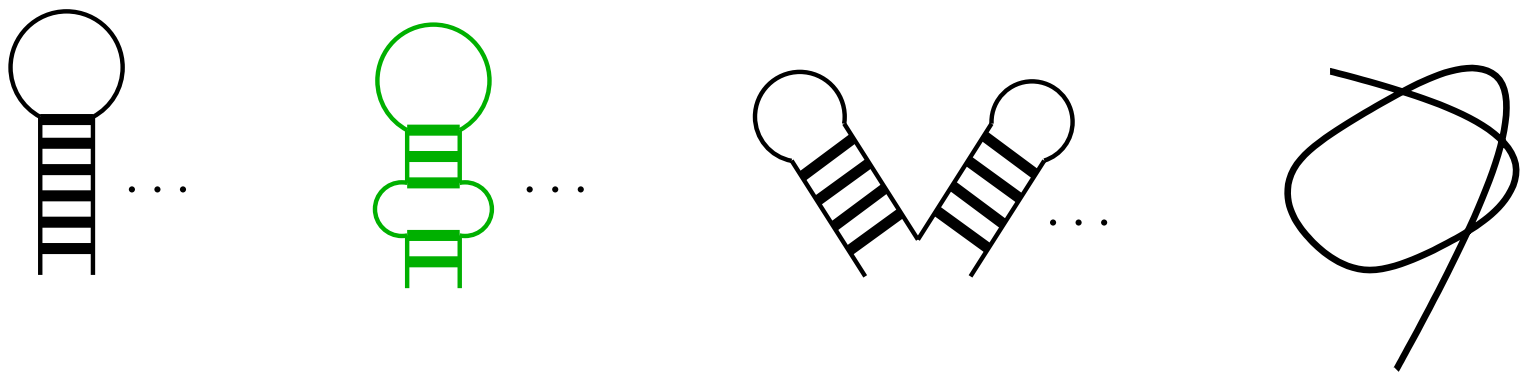
Figure: Total free energy change is summation of all nearest neighbor energies [Durbin et al., 1999]

Nearest (BP) Neighbor Model for RNA Secondary Structure

- The thermodynamic nearest neighbor model is an example of a probabilistic model
- The model is non-generative
  - The model does not directly lead to a method for generating instances of secondary structures
  - Unlike the SCFG model we discussed for secondary structures (and HMM model for alignments)
- The model is useful for inference nonetheless
  - Given an RNA sequence, from the model, we can determine
    - most likely secondary structure (min model free energy)
    - probability that nts at positions i and j are paired
- There exist several other such non-generative models: Conditional random fields, Ising model, ...

RNA ML Decoding of Structure: Single Sequence

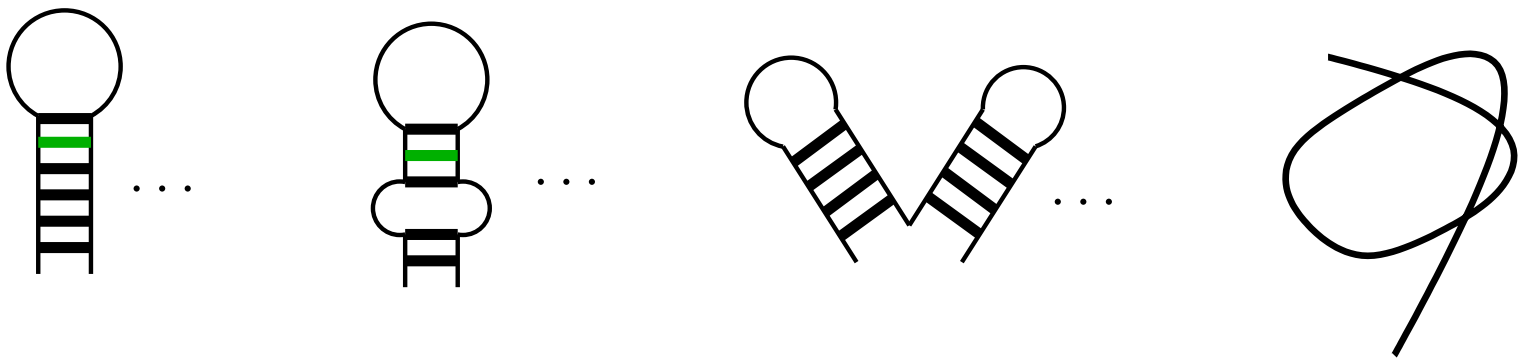
- Most likely or minimum free energy structure, given sequence



- Dynamic Programming MFold [Zuker, 1989] $O(N^3)$ complexity

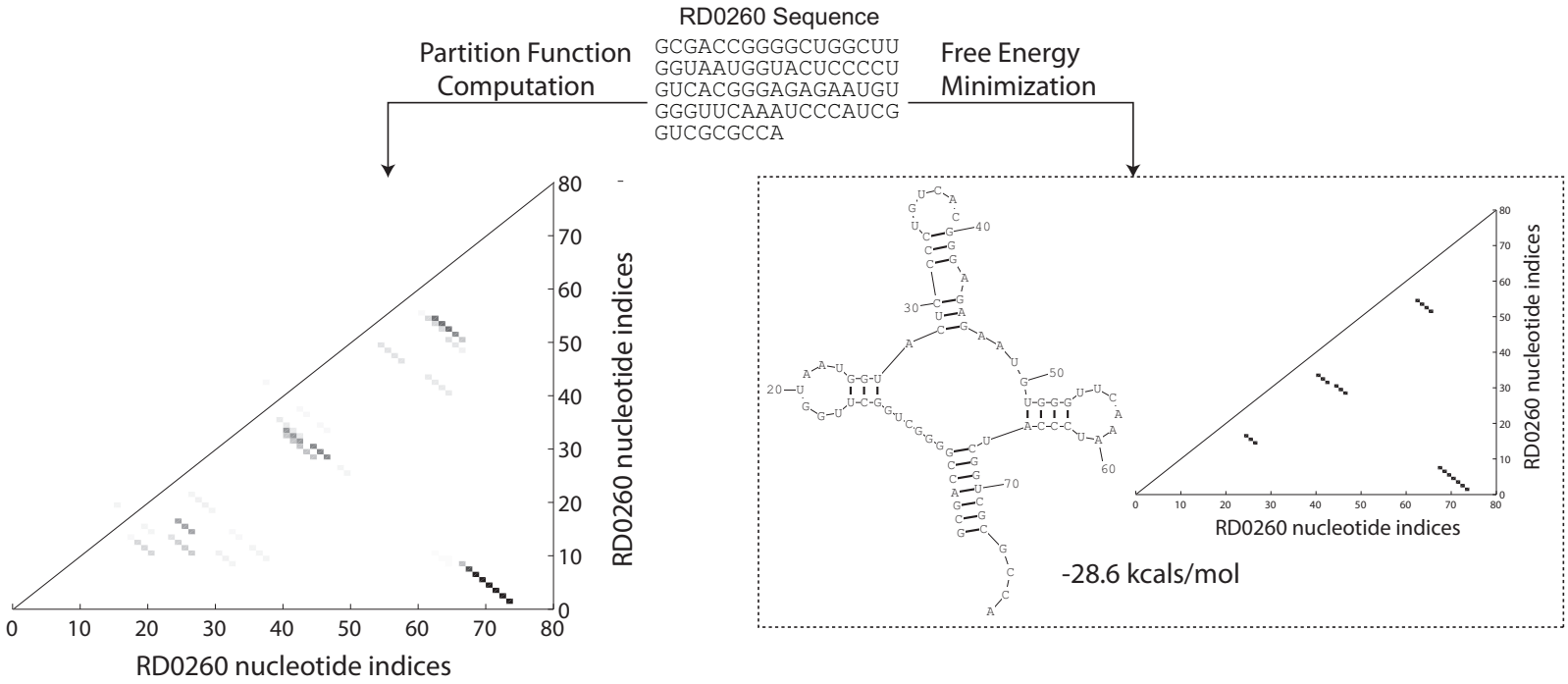
RNA MAP Decoding of Structure

- Posterior probability of base pairing, given sequence



- Dynamic Programming [McCaskill, 1990], MFold, RNAfold ($O(N^3)$ in time, $O(N^2)$ in space)
- Localized probabilistic information

RNA Structure Prediction (Single Sequence)

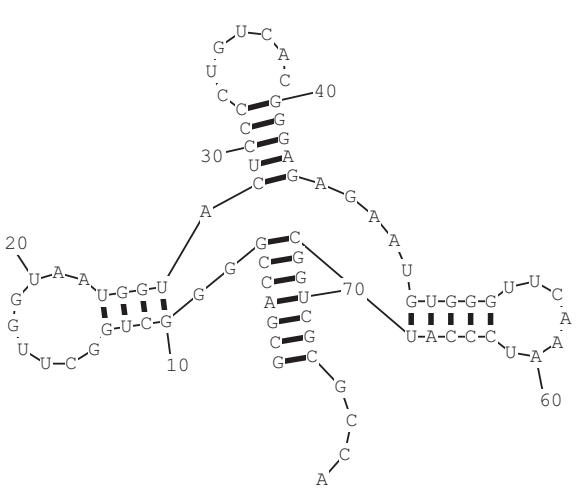


- Free energy minimization: "Hard" Prediction
  - Single prediction structure
- Base pairing probabilities: "Soft" Prediction
  - Thresholding may yield pseudo-knotted structures
  - Maximum Expected Accuracy Structure Prediction, [Do et al., 2006, Lu et al., 2009]

Structure Prediction for Multiple Sequences: Homologous ncRNAs

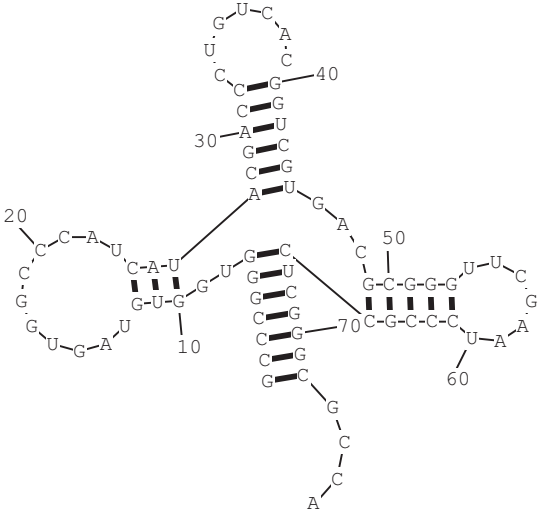
- Homologous ncRNAs
  - Share evolutionary ancestor
  - Serve same function
  - Structural similarity in terms of topology of structures

Bacteriophage T5 (Asp)



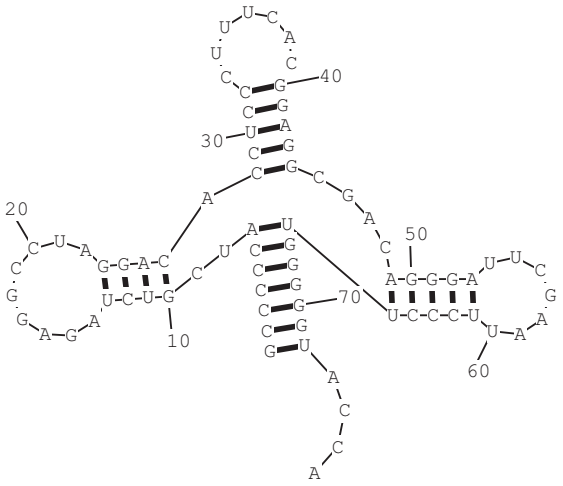
RD0260

Haloferax Volcanii (Asp)



RD0500

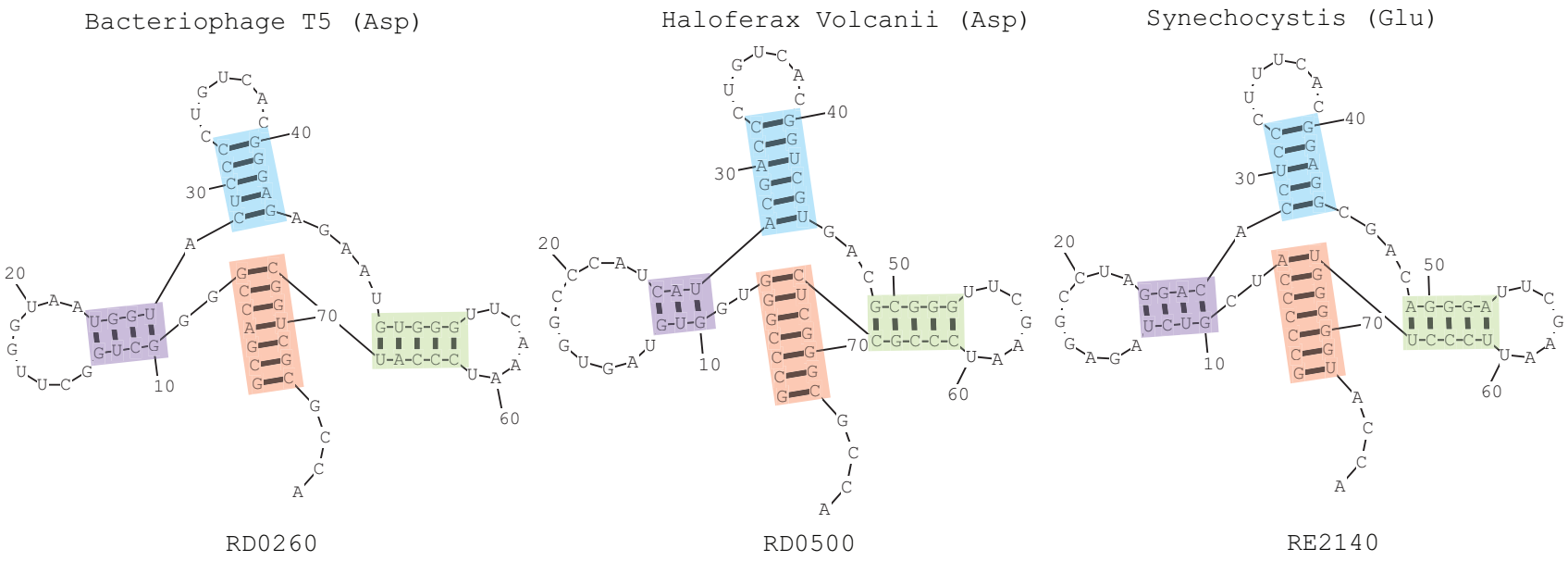
Synechocystis (Glu)



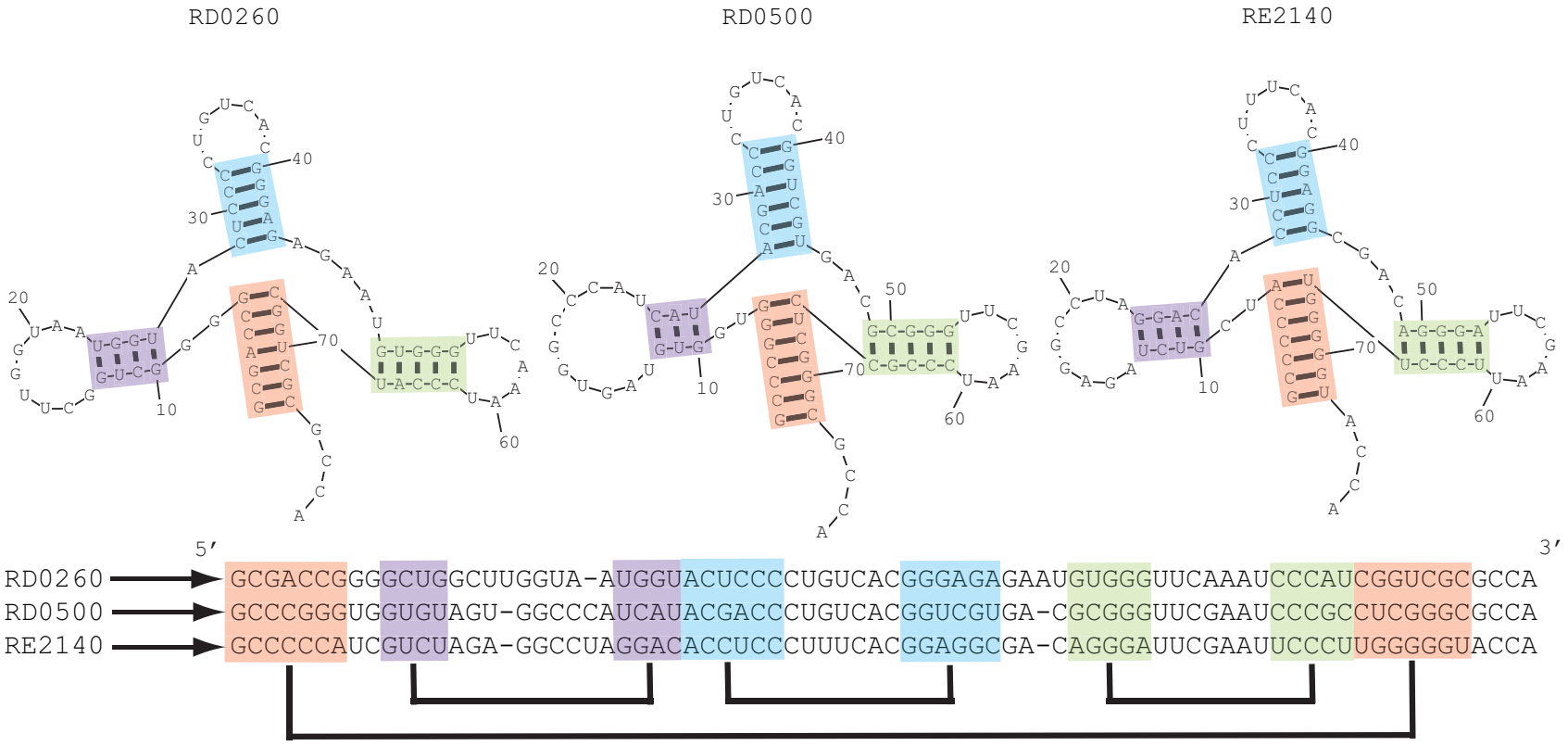
RE2140

Structure Prediction for Multiple Sequences: Homologous ncRNAs

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

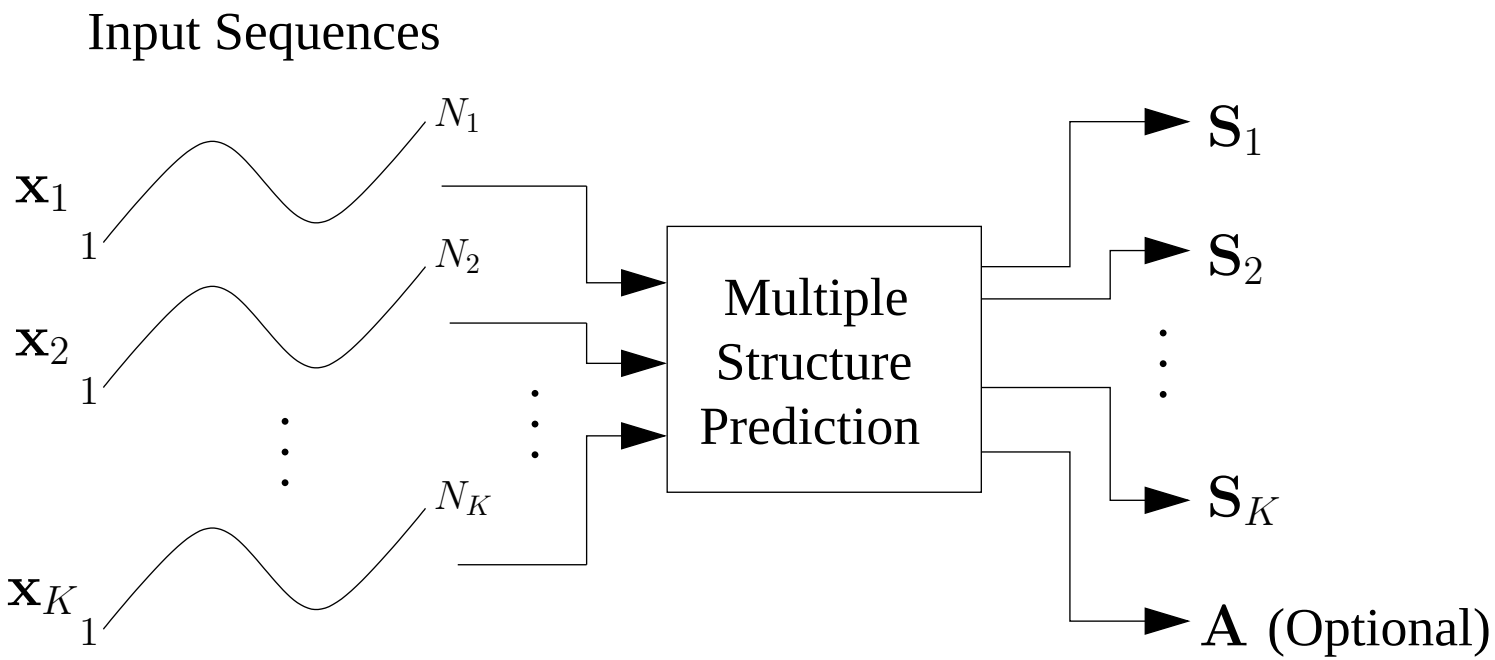


Structure Prediction for Multiple Sequences: Homologous ncRNAs

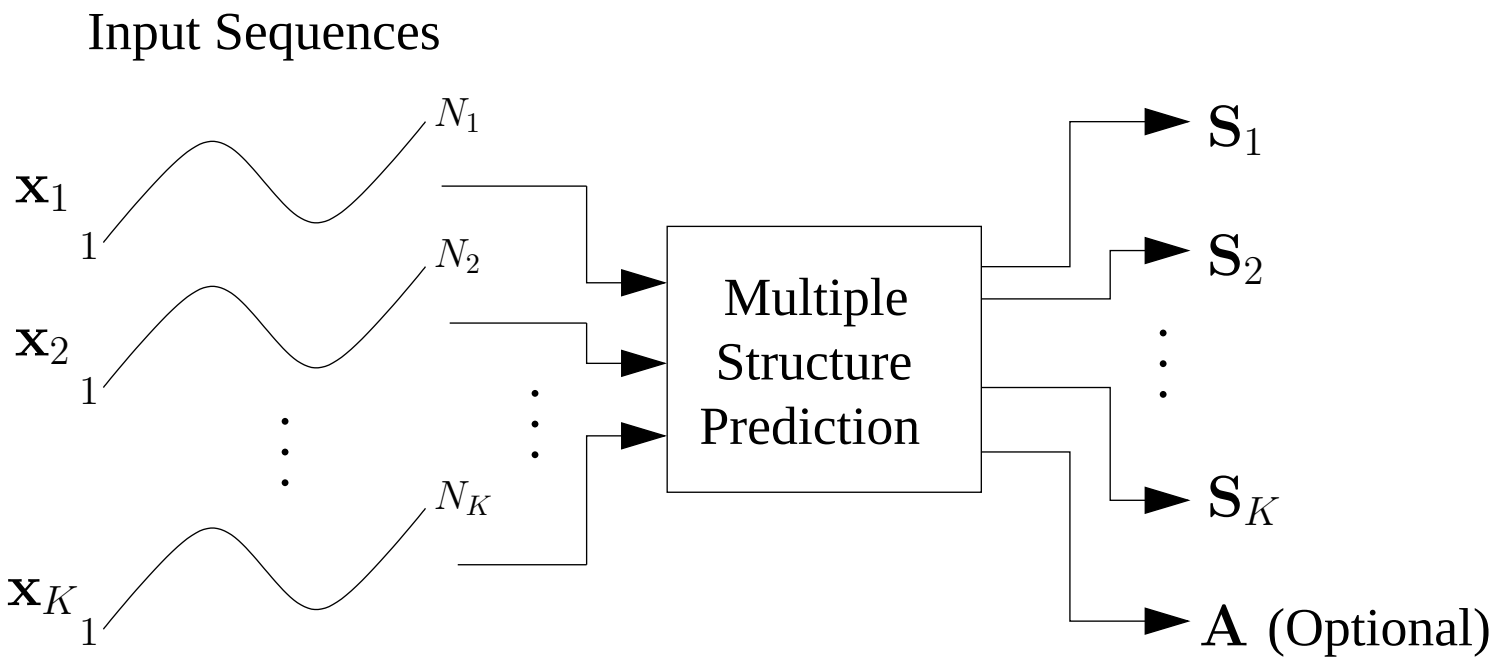


- "Common" structures and conforming sequence alignment
- Joint estimation can harness comparative structure and sequence information across homologs

Multiple Sequence RNA Structure Prediction



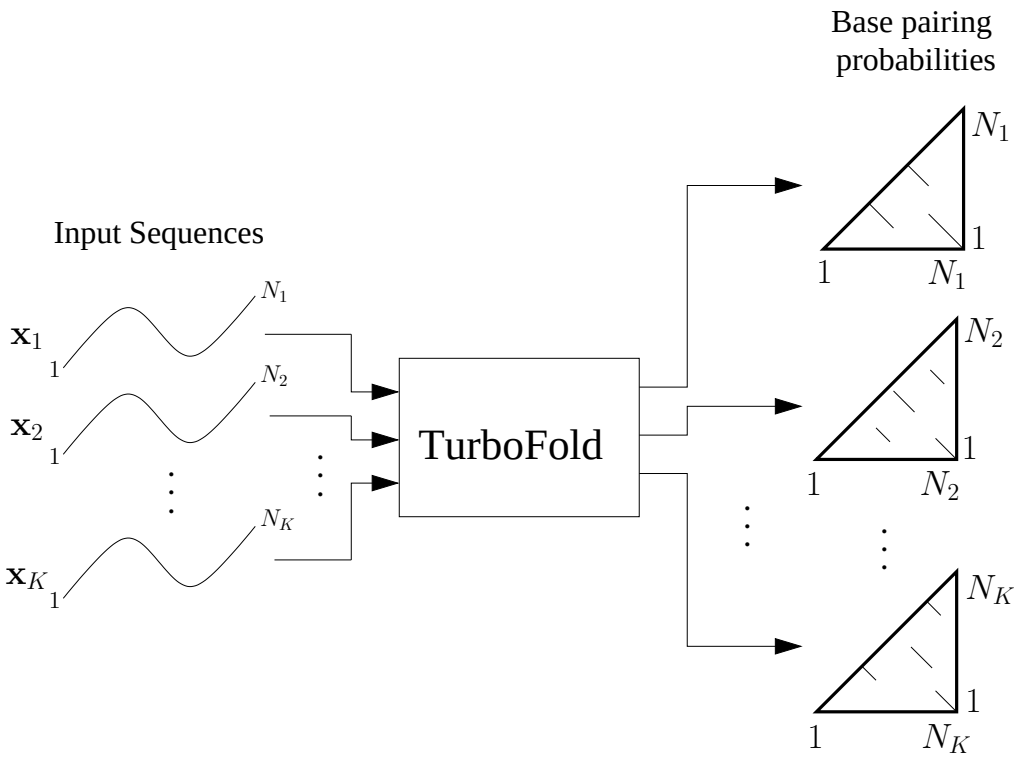
Multiple Sequence RNA Structure Prediction



- Sankoff's dynamic programming algorithm [Sankoff, 1985]
  - 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

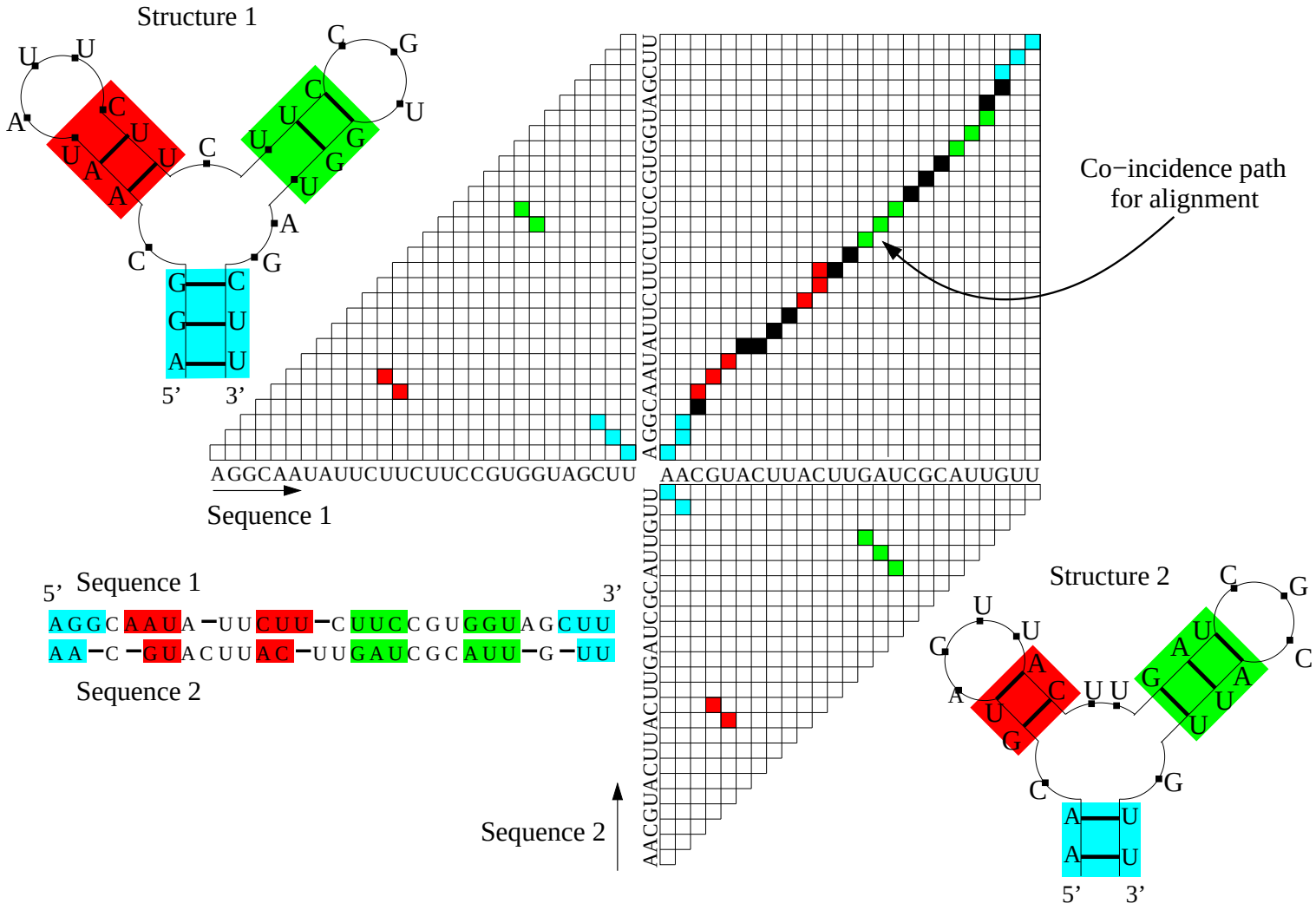
Turbo-Decoding RNA Secondary Structure

- Goal: Performance similar ("better") than joint estimation, complexity similar to single sequence computation.
- Probabilistic formulation of folding and alignment
  - Base pairing probabilities, posterior alignment probabilities
- Iteratively update each using information from other
- TurboFold [Harmanci et al., 2007, 2011].



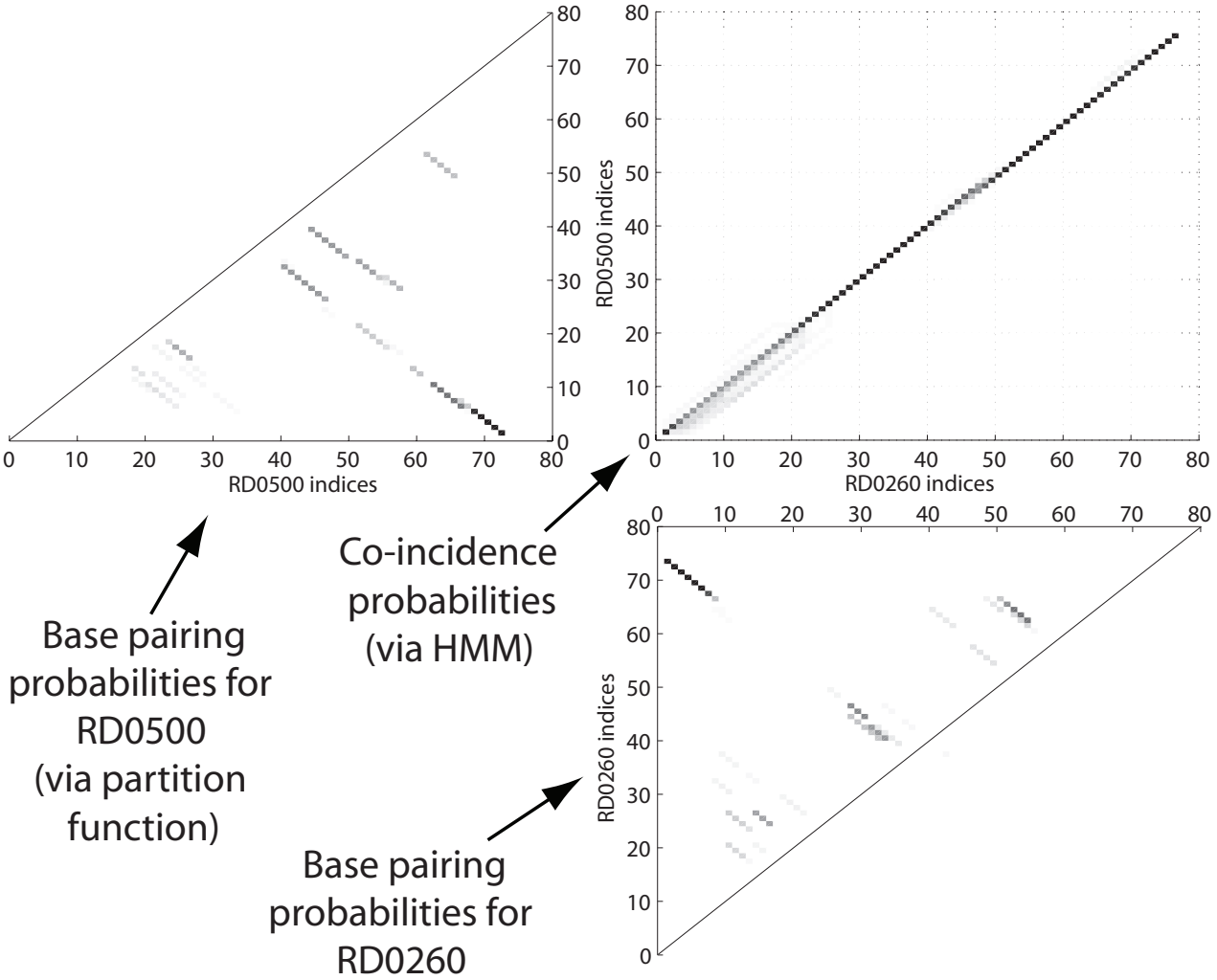
Structural Alignment: Joint Representation of Structures and Alignment

- Two sequence case



Decoupled Probabilistic Representation for RD0260, RD0500 Structural Alignment

- Formulate in probabilistic framework and separate the folding/alignment representations



TurboFold

Given K homologous RNA sequences, each sequence contains some information about folding of every other sequence.

TurboFold

Given K homologous RNA sequences, each sequence contains some information about folding of every other sequence.

- Extrinsic information for a sequence
  - The information about folding of a sequence which is computed using base pairing probabilities of other sequences
  - Thermodynamic model + Alignment model

TurboFold

Given K homologous RNA sequences, each sequence contains some information about folding of every other sequence.

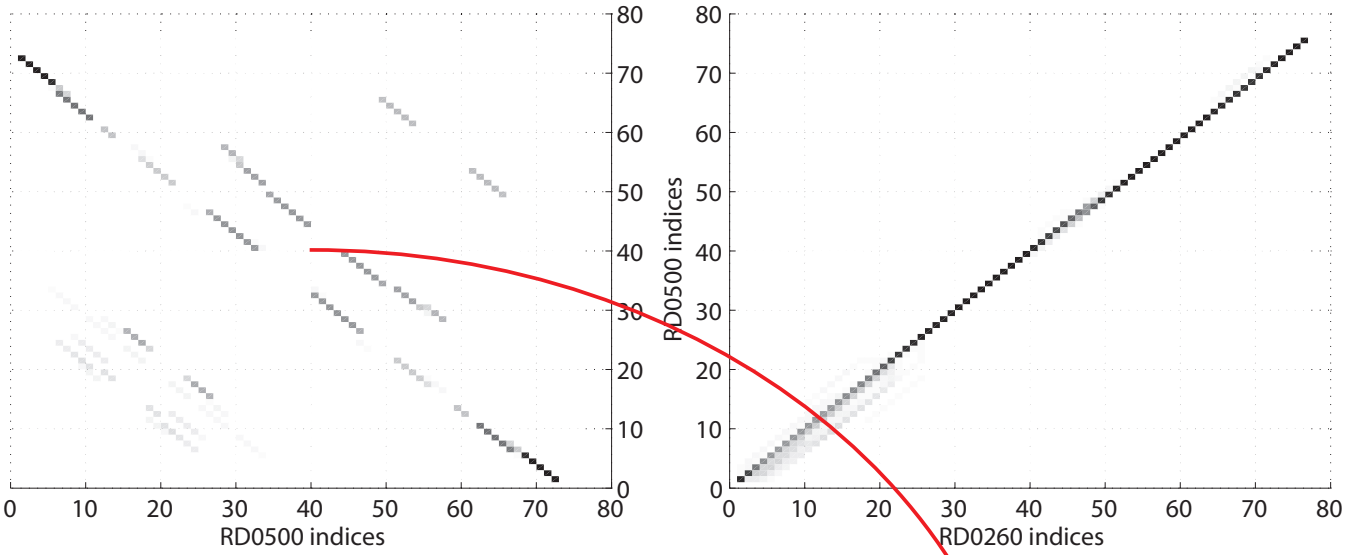
- Extrinsic information for a sequence
  - The information about folding of a sequence which is computed using base pairing probabilities of other sequences
  - Thermodynamic model + Alignment model
- Base pairing probabilities of a sequence (Intrinsic Information)
  - From sequence itself
  - Thermodynamic model

TurboFold

Given K homologous RNA sequences, each sequence contains some information about folding of every other sequence.

- Extrinsic information for a sequence
  - The information about folding of a sequence which is computed using base pairing probabilities of other sequences
  - Thermodynamic model + Alignment model
- Base pairing probabilities of a sequence (Intrinsic Information)
  - From sequence itself
  - Thermodynamic model
- Iterative updates:
  - Compute extrinsic information using base pairing probabilities and alignment co-incidence probabilities
  - Update base pairing probabilities using updated extrinsic information
  - Update extrinsic information using updated base pairing probabilities
  - ...

Extrinsic Information for Base Pairing for RD0260



Extrinsic (RD0260)

$$ {}_{p}^{0}\tilde{\mathbf{\Pi}}^{\text{RD0260}} $$

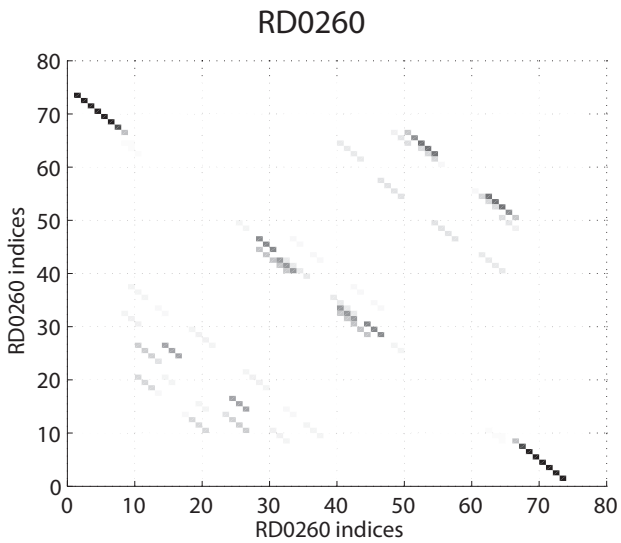
$$ \left({}_{c}^{0}\mathbf{\Pi}^{\text{RD0500, RD0260}}\right)^{\mathbf{T}} \times $$

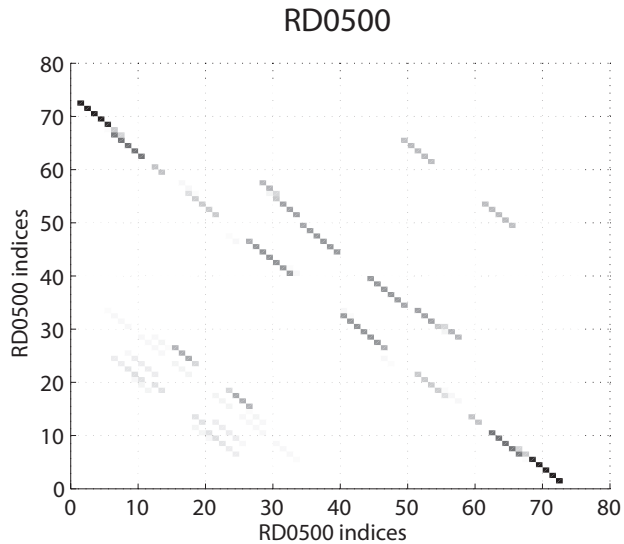
$$ {}_{p}^{0}\mathbf{\Pi}^{\text{RD0500}} \times $$

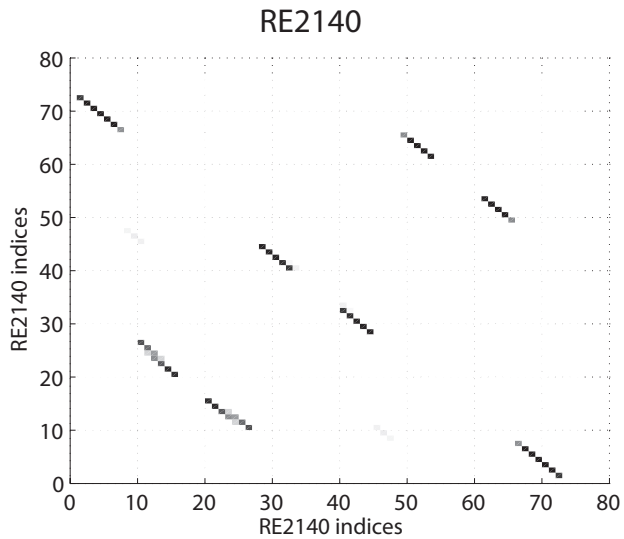
$$ {}_{c}^{0}\mathbf{\Pi}^{\text{RD0500, RD0260}} \times $$

$$ \left(1 - \psi_{\text{RD0500, RD0260}}\right) $$

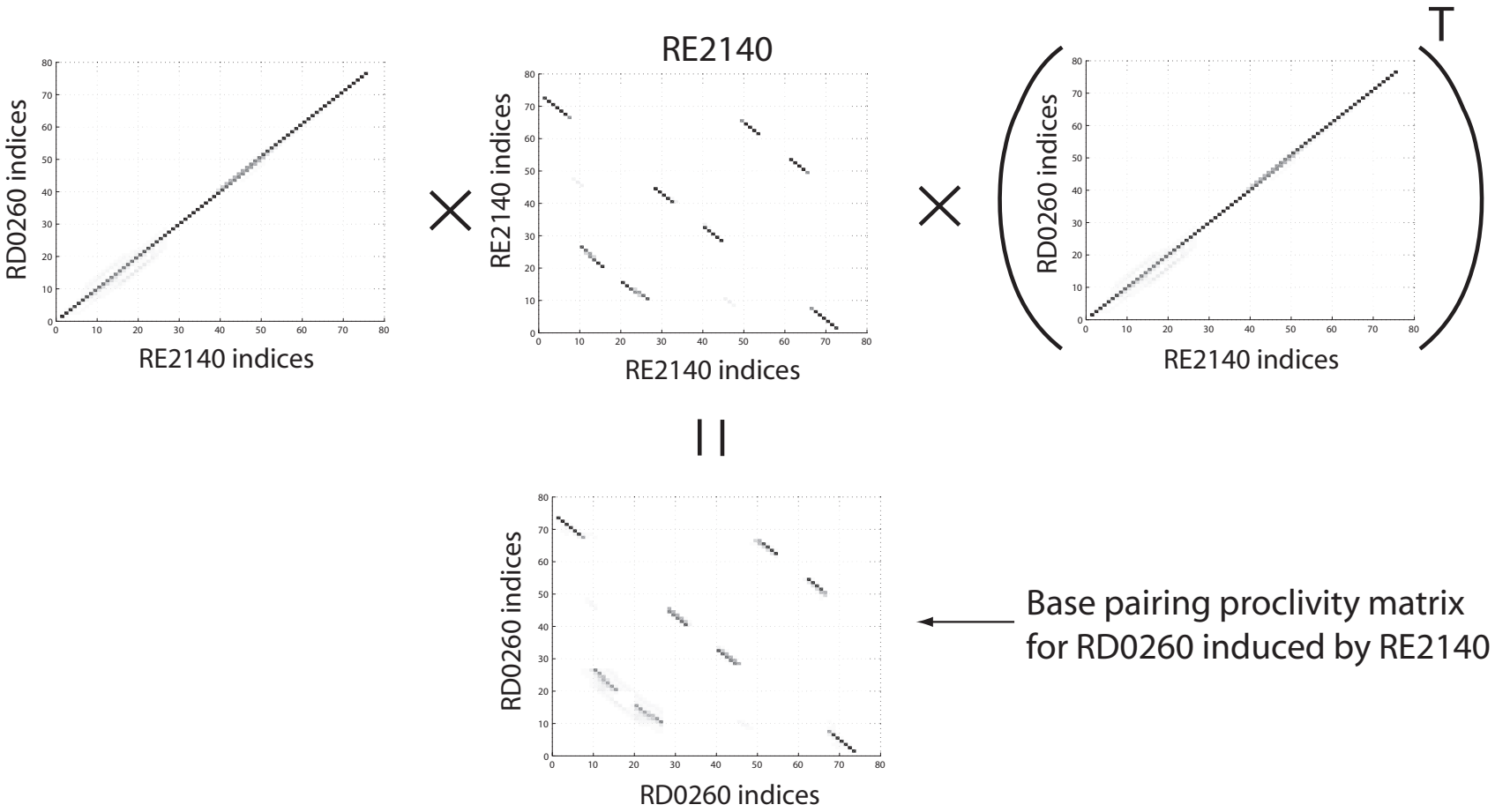
3 Sequences


RD0260
RD0260 indices
RD0260 indices
0
10
20
30
40
50
60
70
80


RD0500
RD0500 indices
RD0500 indices
0
10
20
30
40
50
60
70
80


RE2140
RE2140 indices
RE2140 indices
0
10
20
30
40
50
60
70
80

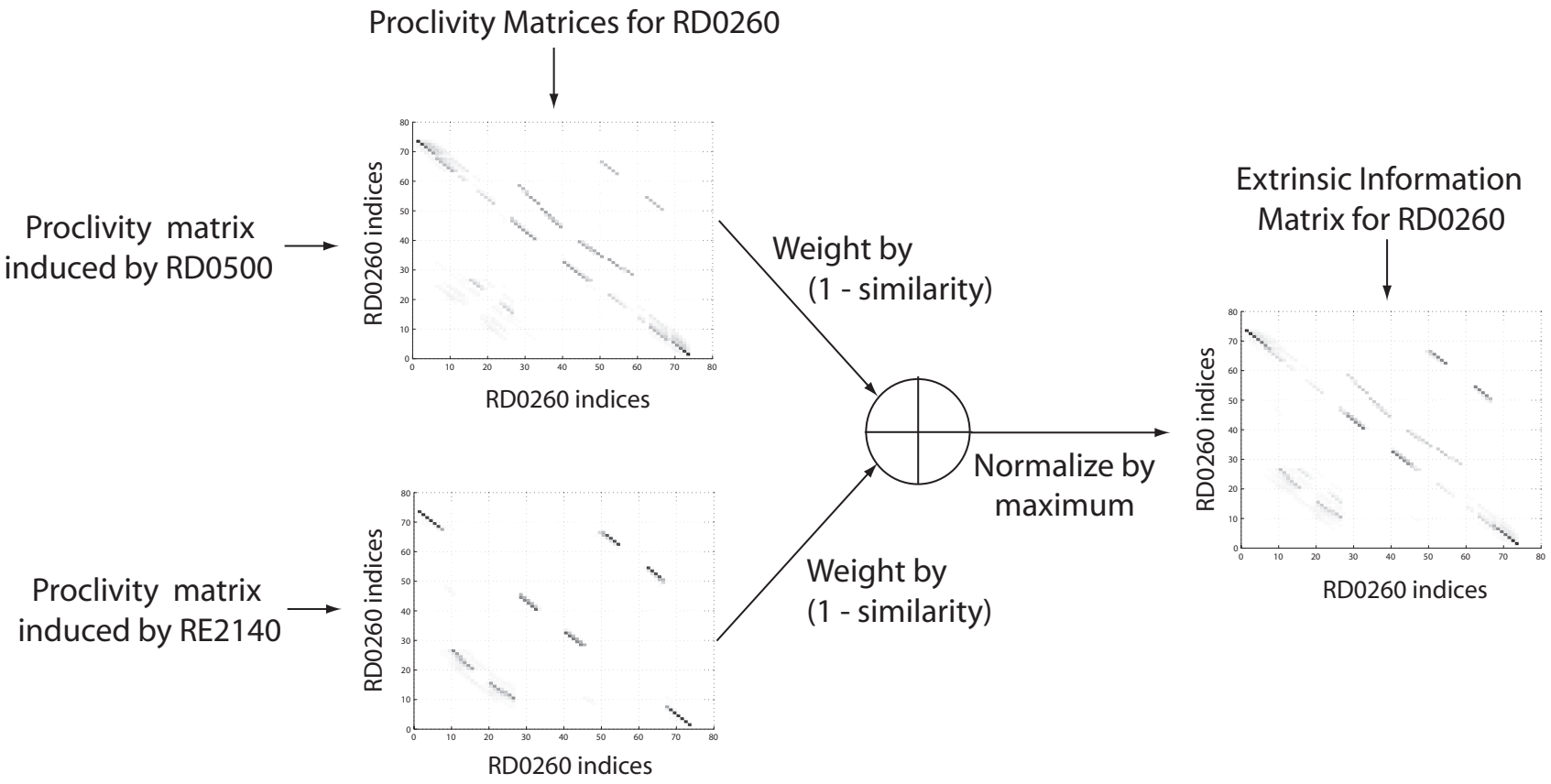
Base Pairing *Proclivity* Matrix for RD0260 Induced by RE2140



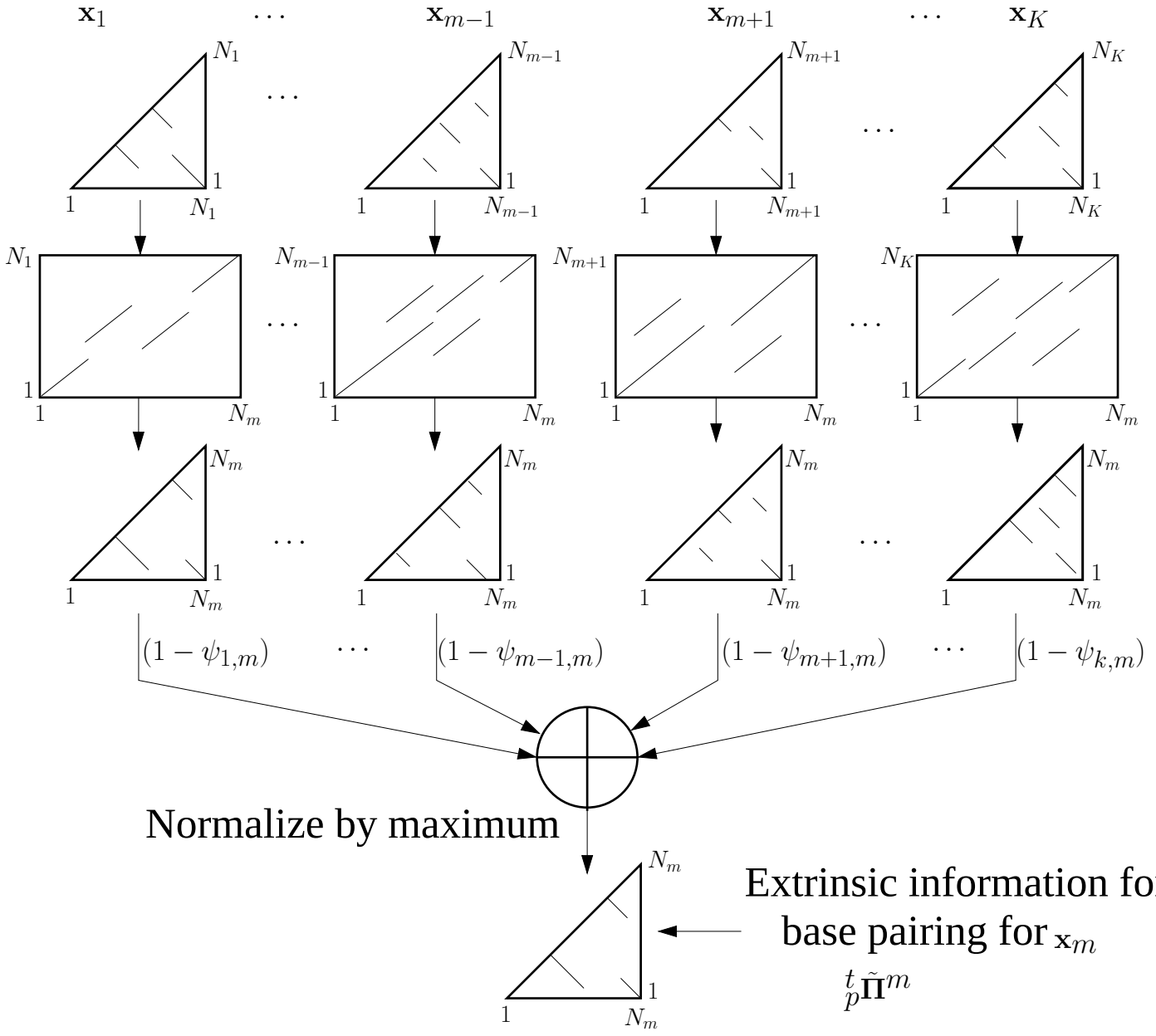
- Information in RE2140 about folding of RD0260

$$ {}^{t}_{p}\tilde{\mathbf{\Pi}}^{(s\to m)} = {}_{c}\mathbf{\Pi}^{(m,s)} \; {}^{t-1}_{p}\mathbf{\Pi}^{s} \; \left({}_{c}\mathbf{\Pi}^{(m,s)}\right)^{\mathrm{T}} \qquad (1) $$

3 Sequences: Extrinsic information Computation



K Sequences: Extrinsic Information Computation for $\mathbf{x}_m$



Base pairing probabilities

$\{{}_p^{t-1}\mathbf{\Pi}^s\}_{s\in\mathcal{N}\setminus m}$

Co−incidence probabilities

$\{{}_c^t\mathbf{\Pi}^{(s,m)}\}_{s\in\mathcal{N}\setminus m}$

Induced base pairing proclivity matrices for $\mathbf{x}_m$

$\{{}_p^t\tilde{\mathbf{\Pi}}^{(s\to m)}\}_{s\in\mathcal{N}\setminus m}$

Base Pairing Probability Computation Using Extrinsic Information

- Modified Boltzmann distribution of secondary structures:

$$P(\mathbf{S}) \propto \exp\left(-\frac{\Delta\tilde{G}(\mathbf{S})}{RT}\right)$$

Base Pairing Probability Computation Using Extrinsic Information

- Modified Boltzmann distribution of secondary structures:

$$P(\mathbf{S}) \propto \exp\left(-\frac{\Delta\tilde{G}(\mathbf{S})}{RT}\right)$$

where

$$\Delta\tilde{G}(\mathbf{S}) = \Delta G^o(\mathbf{S}) - \gamma \sum_{(i,j)\in\mathbf{S}} \log(\tilde{\pi}(i,j))$$

is the *modified* free energy change for structure $\mathbf{S}$.

- $\tilde{\pi}(i,j)$: Extrinsic information for pairing of nucleotides at indices i and j
- γ: Weight of extrinsic information on modified free energy relative to $\Delta G^o(\mathbf{S})$

Extrinsic information introduced via a pseudo free energy for each base pair

Base Pairing Probability Computation Using Extrinsic Information

- Modified Boltzmann distribution of secondary structures:

$$P(\mathbf{S}) \propto \exp\left(-\frac{\Delta\tilde{G}(\mathbf{S})}{RT}\right) \tag{2}$$

where

$$\Delta\tilde{G}(\mathbf{S}) = \Delta G^{o}(\mathbf{S}) - \gamma \sum_{(i,j)\in\mathbf{S}} \log(\tilde{\pi}(i,j)) \tag{3}$$

Base Pairing Probability Computation Using Extrinsic Information

- Modified Boltzmann distribution of secondary structures:

$$P(\mathbf{S}) \propto \exp\left(-\frac{\Delta\tilde{G}(\mathbf{S})}{RT}\right) \tag{2}$$

where

$$\Delta\tilde{G}(\mathbf{S}) = \Delta G^o(\mathbf{S}) - \gamma \sum_{(i,j)\in\mathbf{S}} \log(\tilde{\pi}(i,j)) \tag{3}$$

Replace (3) in (2):

$$P(\mathbf{S}) \propto \underbrace{\exp(-\frac{\Delta G^o(\mathbf{S})}{RT})}_{\substack{\text{Boltzmann distribution} \\ \text{proportionality term}}} \underbrace{\left(\prod_{(i,j)\in\mathbf{S}} (\tilde{\pi}(i,j))^{\gamma/RT}\right)}_{\substack{\text{Extrinsic} \\ \text{information}}}$$

Base Pairing Probability Computation Using Extrinsic Information

- Modified Boltzmann distribution of secondary structures:

$$P(\mathbf{S}) \propto \exp\left(-\frac{\Delta\tilde{G}(\mathbf{S})}{RT}\right) \tag{2}$$

where

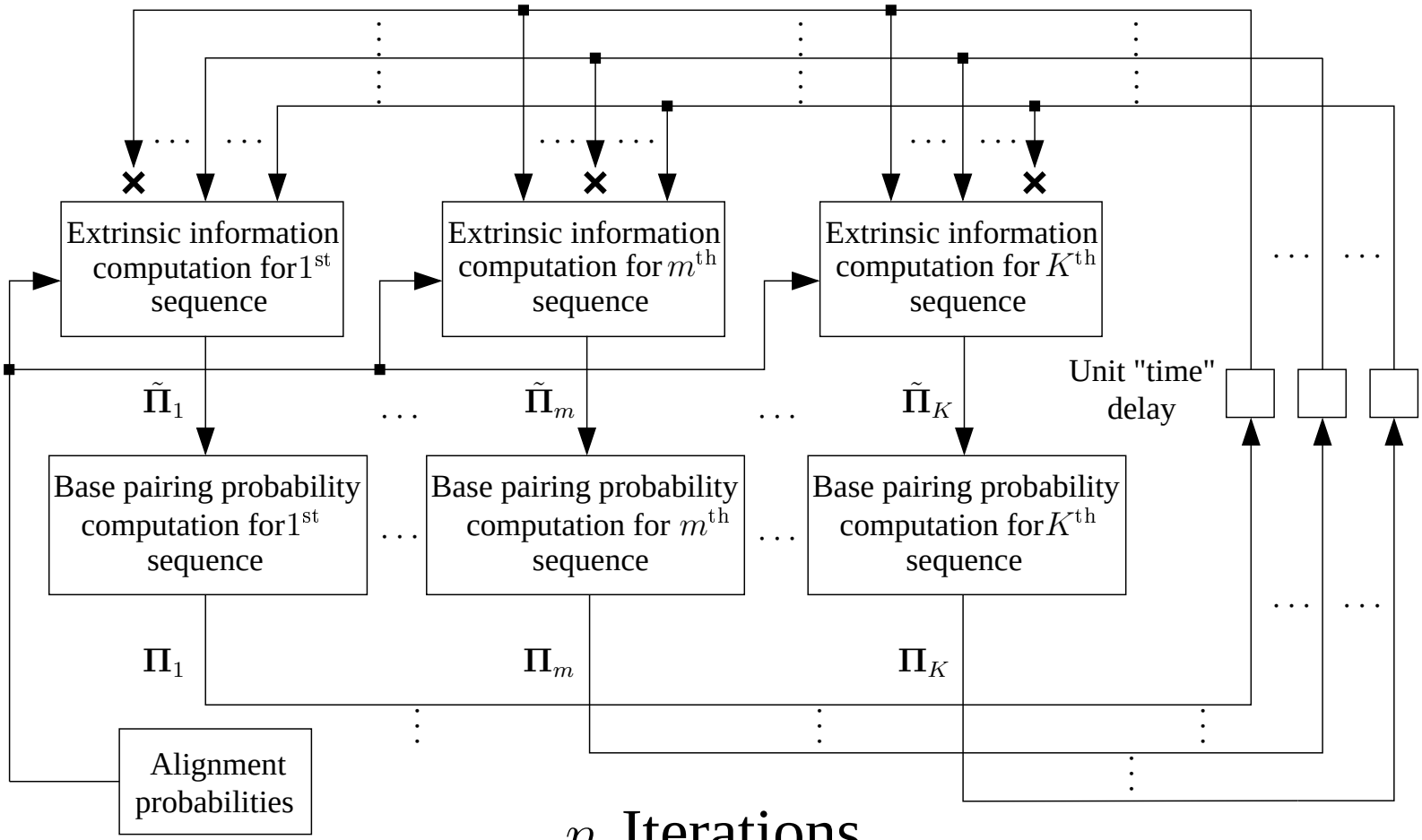
$$\Delta\tilde{G}(\mathbf{S}) = \Delta G^o(\mathbf{S}) - \gamma \sum_{(i,j)\in\mathbf{S}} \log(\tilde{\pi}(i,j)) \tag{3}$$

Replace (3) in (2):

$$P(\mathbf{S}) \propto \underbrace{\exp(-\frac{\Delta G^o(\mathbf{S})}{RT})}_{\substack{\text{Boltzmann distribution} \\ \text{proportionality term}}} \underbrace{\left(\prod_{(i,j)\in\mathbf{S}} (\tilde{\pi}(i,j))^{\gamma/RT}\right)}_{\substack{\text{Extrinsic} \\ \text{information}}}$$

- Base pair (i,j) has a pseudo prior probability of $(\tilde{\pi}(i,j))^{\gamma/RT}$ due to extrinsic information.

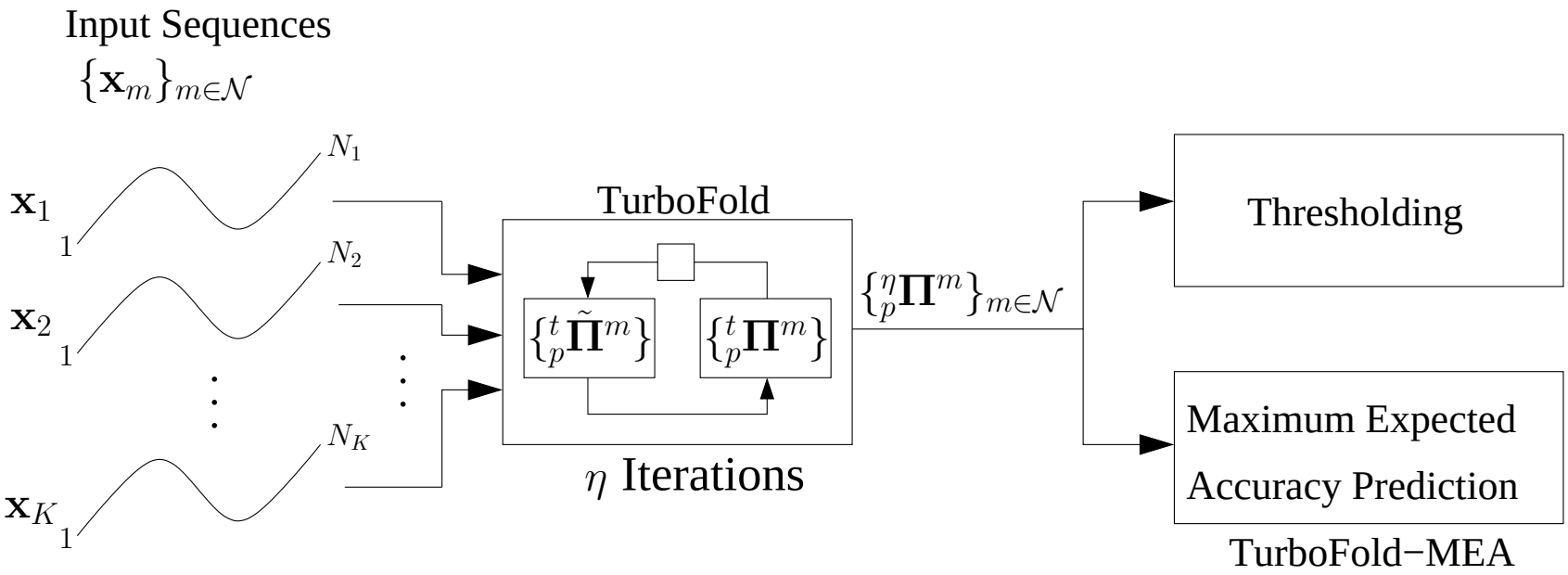
TurboFold: Iterative Updates



- For low K, per iteration complexity is comparable to single sequence structure prediction
- Benefits from comparative analysis

TurboFold Structure Prediction Overview

- Obtain base pairing probabilities after η iterations, then predict structures
  - Significant base pairs
  - Maximum expected accuracy (MEA) structures



Structure Prediction

- Structure for $\mathbf{x}_m$ composed of base pairs with probabilities greater than P_{thresh}:

$$\mathbf{S}_m^* = \{(i,j) \ni {}_p^{\eta}\pi^m(i,j) > P_{\text{thresh}}\} \tag{4}$$

TurboFold: Computation Complexity

- Initialization
  - Computation of co-incidence matrices: $O(K^2N^2)$
  - Computation of sequence similarities: $O(K^2N^2)$
- Iterations
  - Extrinsic information computation: $O(\eta K^2 d^2 N^2)$
  - Base pairing probability computation: $O(\eta KUN^3)$
- Structure prediction
  - Thresholding: $O(KN^2)$
  - MEA prediction: $O(KN^3)$

Compare to Sankoff's algorithm: $O(N^3(U^2d)^K)$

Evaluating Accuracy of Estimates

- Sensitivity: Ratio of number of correctly predicted base pairs to the total number of base pairs in the **known** structure

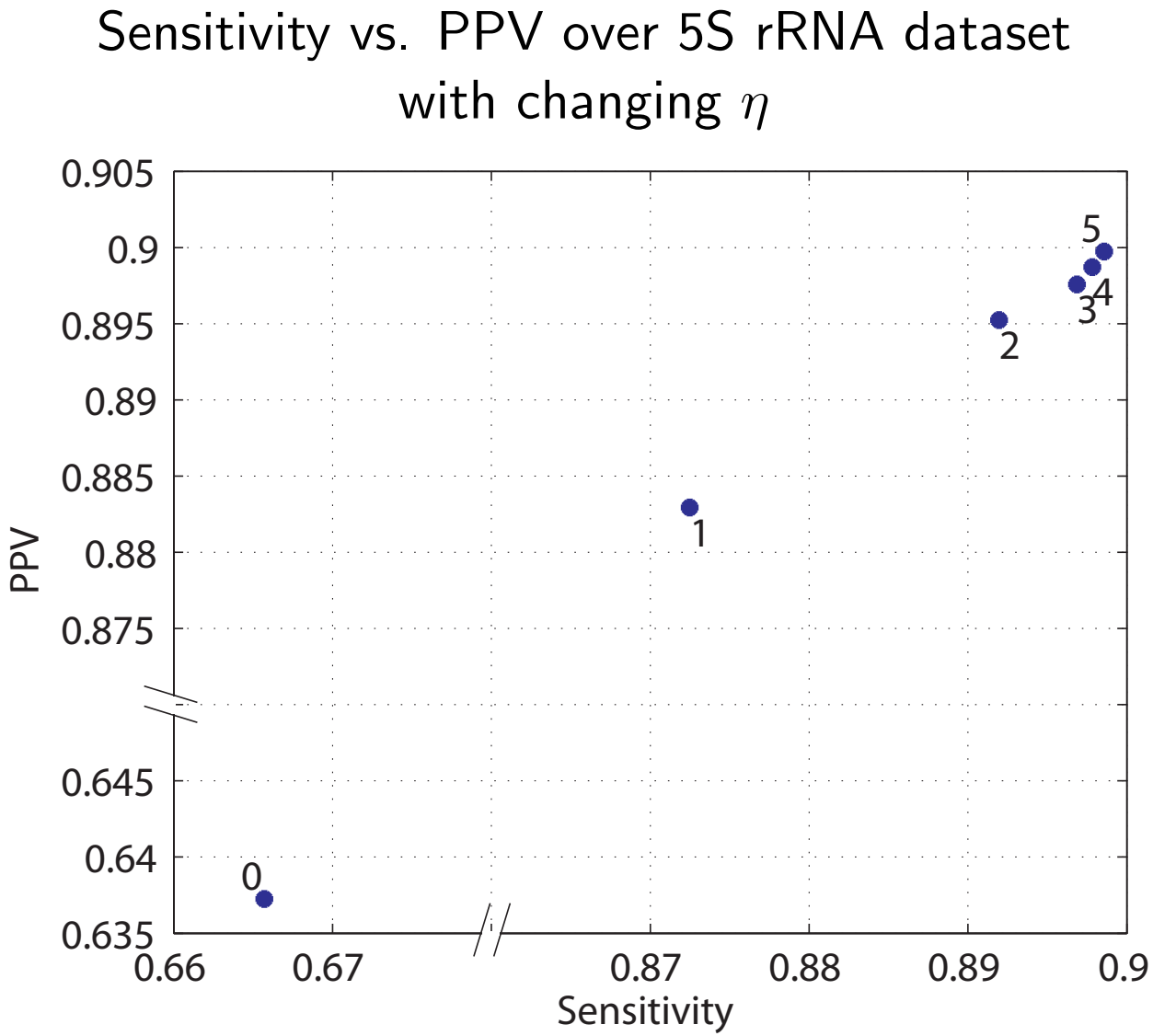
$$\frac{\text{True Positive}}{\text{True Positive} + \text{False } \mathbf{Negative}}$$

  - Recall
- Positive Predictive Value(PPV): Ratio of number of correctly predicted base pairs to the total number of base pairs in the **predicted** structure

$$\frac{\text{True Positive}}{\text{True Positive} + \text{False } \mathbf{Positive}}$$

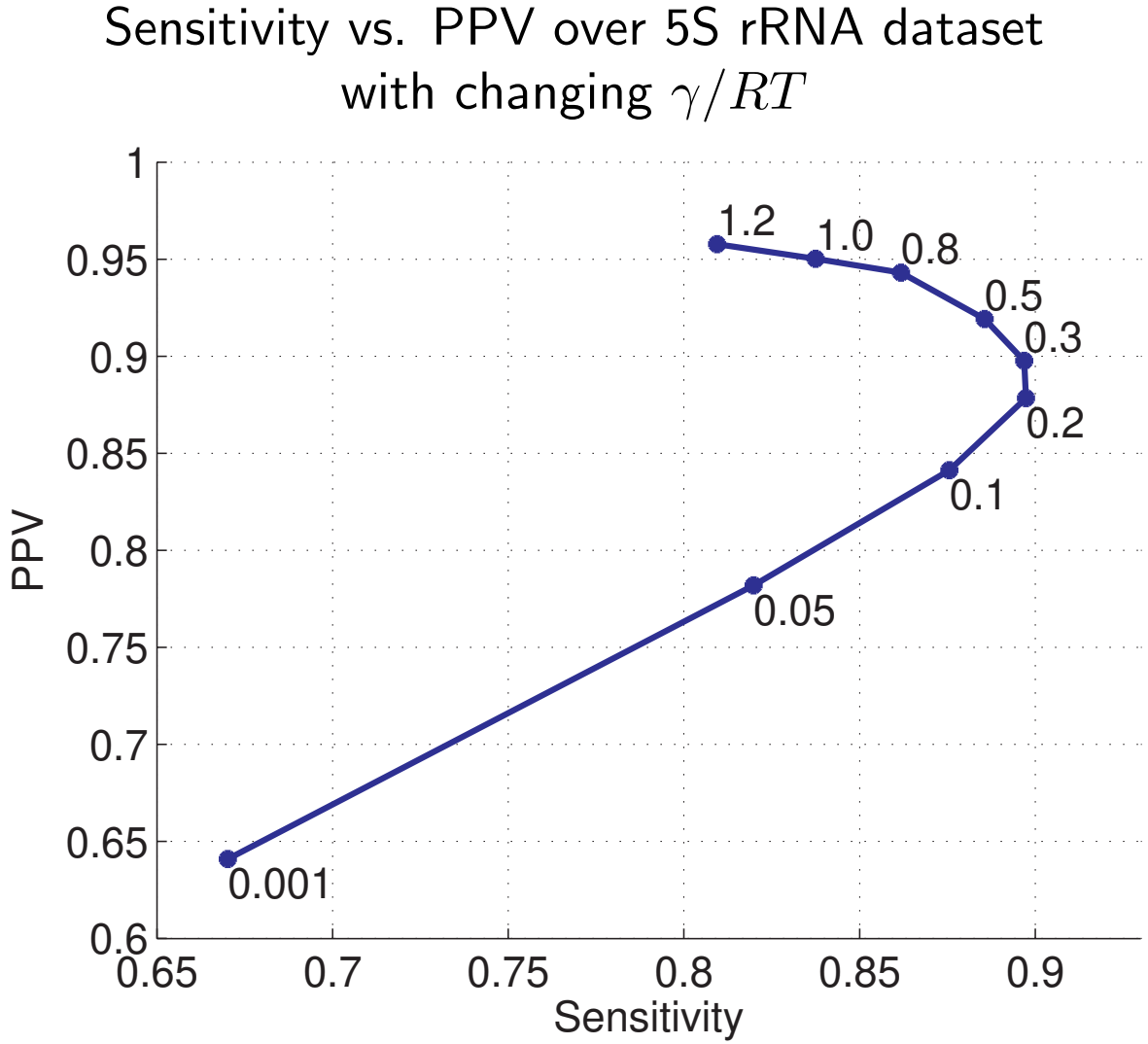
  - Precision

Parameter Selection: Number of iterations, η



- $\eta = 3$ is used in TurboFold

Parameter Selection: Weight of Extrinsic Information, γ



- $\gamma = 0.3RT$ is used in TurboFold

Benchmarking Experiments: Datasets

- Randomly choose 200 RNase P, 400 5S rRNA, 400 SRP, and 400 tRNA sequences and divide into K combinations
  - Choose and divide for $K = 2, \ldots, 10$
- Yields 36 datasets

The datasets have significant diversity:

- RNase Ps: 336 nucleotides, 50% average pairwise identity
- tmRNA: 366 nucleotides, 45% average pairwise identity
- telomerase RNA: 445 nucleotides, 54% average pairwise identity
- SRPs: 187 nucleotides, 42% average pairwise identity
- tRNAs: 77 nucleotides, 47% average pairwise identity
- 5S rRNAs: 119 nucleotides, 63% average pairwise identity

Benchmarking Experiments: Datasets

- Randomly choose 200 RNase P, 400 5S rRNA, 400 SRP, and 400 tRNA sequences and divide into K combinations
  - Choose and divide for $K = 2, \ldots, 10$
- Yields 36 datasets

The datasets have significant diversity:

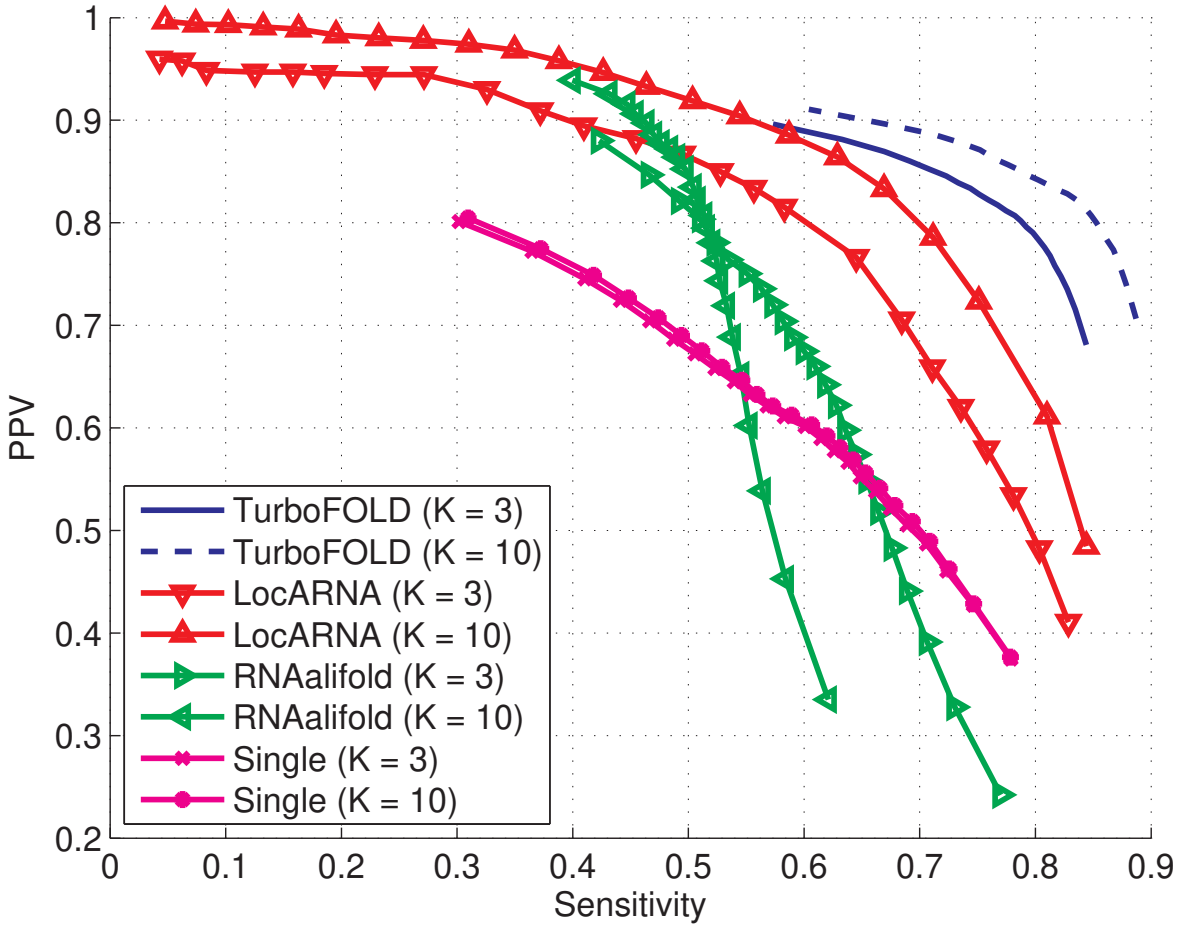
- RNase Ps: 336 nucleotides, 50% average pairwise identity
- tmRNA: 366 nucleotides, 45% average pairwise identity
- telomerase RNA: 445 nucleotides, 54% average pairwise identity
- SRPs: 187 nucleotides, 42% average pairwise identity
- tRNAs: 77 nucleotides, 47% average pairwise identity
- 5S rRNAs: 119 nucleotides, 63% average pairwise identity

Benchmarking Experiments

- TurboFold is benchmarked against methods that estimate base pairing probabilities:
  - LocARNA [Will et al., 2007]
  - RNAalifold [Bernhart et al., 2008]
  - Single sequence partition function [Mathews, 2004]
- The set of base pairs with estimated probabilities higher than P_{thresh} are scored
- Plotted sensitivity versus PPV while varying P_{thresh} between 0 and 1 with step size of 0.04

Benchmarking Experiments

Sensitivity vs PPV ROC curves for TurboFold vs three alternative methods



Run Time Requirements

- Run time requirements over 50 RNase P sequence datasets

	Runtime (seconds) for		
	$K = 3$	$K = 5$	$K = 10$
TurboFold	136.75	277.9	517.0
LocARNA	746.44	2815.9	11395.8
RNAalifold	0.2	0.3	0.6

Table: Time requirements (in seconds) for the methods.

- TurboFold scales slower with increase in K

Conclusions

- TurboFold: A multiple sequence structure prediction method
  - Lowers Complexity with iterative combination of intrinsic and extrinsic information for folding
    - Intrinsic information: From sequence via thermodynamic folding model (nearest neighbor model)
    - Extrinsic information: From other sequences
- TurboFold accuracy: close to or higher than the simultaneous folding and alignment methods
- Details: BMC Bioinformatics article Harmanci et al. [2011].
- Connections to coding theory in digital communications

Turbo Decoding: RNA vs Communications

- Multiple encodings of same information
  - Nature/Man
- Joint (optimal) decoding desirable
  - Exact joint decoding $\approx$ exponential complexity
  - Iterative approximation (belief propagation)
    - Localized MAP probabilistic formulation (base pairing/symbol probs.)
    - Decomposition into loosely coupled individual decodings + information exchange at each iteration
    - Linear/polynomial complexity in length of data
    - Pseudo-prior interpretation

Ongoing Related Work

- Moving beyond TurboFold
  - Alignment probability updates based on structures
  - Better handling of dependencies
  - Domain insertions/deletions
  - Linear time approximation using beam search Li et al. [2021]
- Connecting with experiments
  - Incorporating experimental information (e.g. SHAPE) in structural alignments
  - Postulating mechanisms and experimental validation (HIV)

Acknowledgments

- Collaborators at UR:
  - Arif O. Harmanci, UR ECE, currently faculty at Univ. of Houston
  - David H. Mathews, Department of Biochemistry and Biophysics
- Research support:
  - National Institutes of Health (NIH) (Award # GM097334-01)
  - Center for Research Computing, University of Rochester

Thank you

Questions?

S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008.

J. W. Brown. The Ribonuclease P database. *Nucleic Acids Res.*, 27(1):314, Jan. 1999.

Chuong B. Do, Daniel A. Woods, and Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):90–98, 2006.

J. A. Doudna and T. R. Cech. The chemical repertoire of natural ribozymes. *Nature*, 418(6894):222–228, 2002.

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

R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1999. ISBN 0521629713.

A. Ozgun Harmanci, Gaurav Sharma, and David H. Mathews. Toward turbo decoding of RNA secondary structure. In *Proc. IEEE Intl. Conf. Acoustics Speech and Sig. Proc.*, volume I, pages 365–368, Apr. 2007. doi: 10.1109/ICASSP.2007.366692.

A. Ozgun Harmanci, Gaurav Sharma, and David H. Mathews. TurboFold: iterative probabilistic estimation of secondary structures for multiple RNA sequences. *BMC Bioinformatics*, 12: 108, Apr. 2011. doi: 10.1186/1471-2105-12-108. early access available online, April 20, 2011.

Sizhen Li, He Zhang, Liang Zhang, Kaibo Liu, Boxiang Liu, David H. Mathews, and Liang Huang. LinearTurboFold: Fast folding and alignment for rna homologs with applications to coronavirus, 2021. URL `https://www.biorxiv.org/content/early/2021/06/24/2020.11.` bioRxiv preprint.

Zhi John Lu, Jason W. Gloor, and David H. Mathews. Improved

RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, 15(10):1805–1813, 2009.

D. H. Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):1178–1190, 2004.

D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J. Mol. Biol.*, 288(5):911–940, 1999.

J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, Nov. 1990.

D. Sankoff. Simultaneous solution of RNA folding, alignment and protosequence problems. *SIAM J. App. Math.*, 45(5):810–825, Oct. 1985.

I. Tinoco, Jr. and C. Bustamante. How RNA folds. *J Mol Biol*, 293(2):271–281, 1999.

Richard B. Waring and R. Wayne Davies. Assessment of a model for intron rna secondary structure relevant to RNA self-splicing – a review. *Gene*, 28(3):277–291, 1984.

Sebastian Will, Kristin Reiche, Ivo L. Hofacker, Peter F. Stadler, and Rolf Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3(4):680–691, Apr. 2007.

T. Xia, J. SantaLucia, Jr., R Kierzek, S. J. Schroeder, X Jiao, C Cox, and Douglas Henry Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs. *Biochemistry*, 37(42): 14719–14735, 1998.

M. Zuker. Computer prediction of RNA structure. *Methods Enzymol.*, 180:262–288, 1989.