CS681: Advanced Topics in Computational Biology

Week 9 Lectures 2-3

Can Alkan

EA224

calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

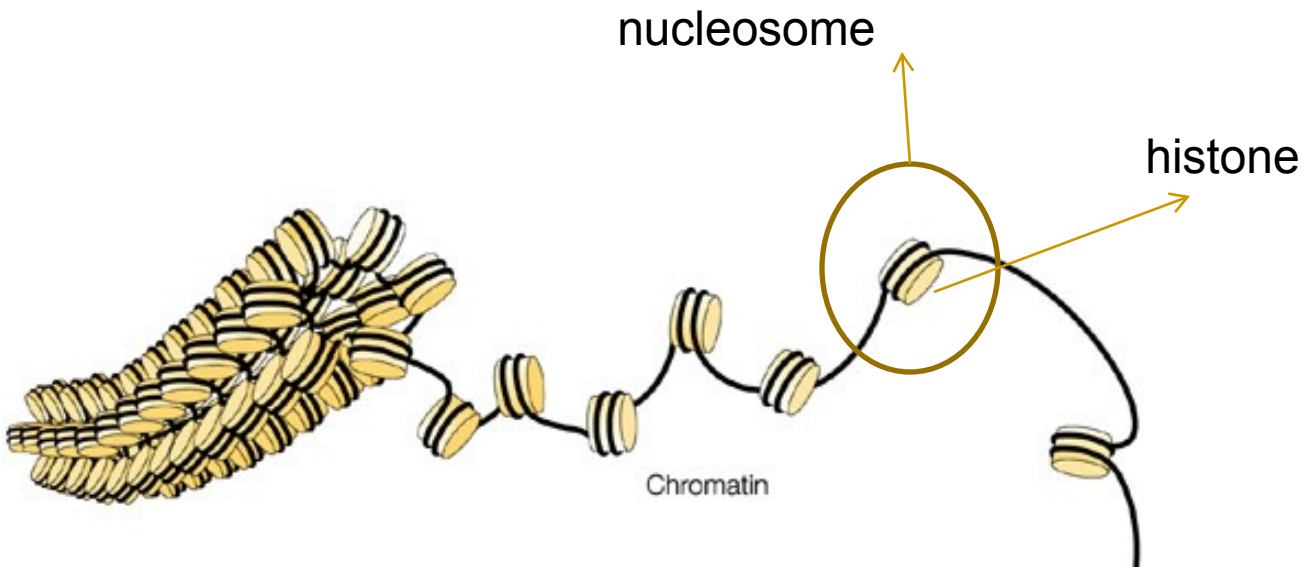
Epigenetics

- Epigenetics: study of all meiotically and mitotically heritable changes in gene expression that are not coded in the DNA sequence itself
  - DNA methylation
  - RNA associated silencing
  - Histone modification

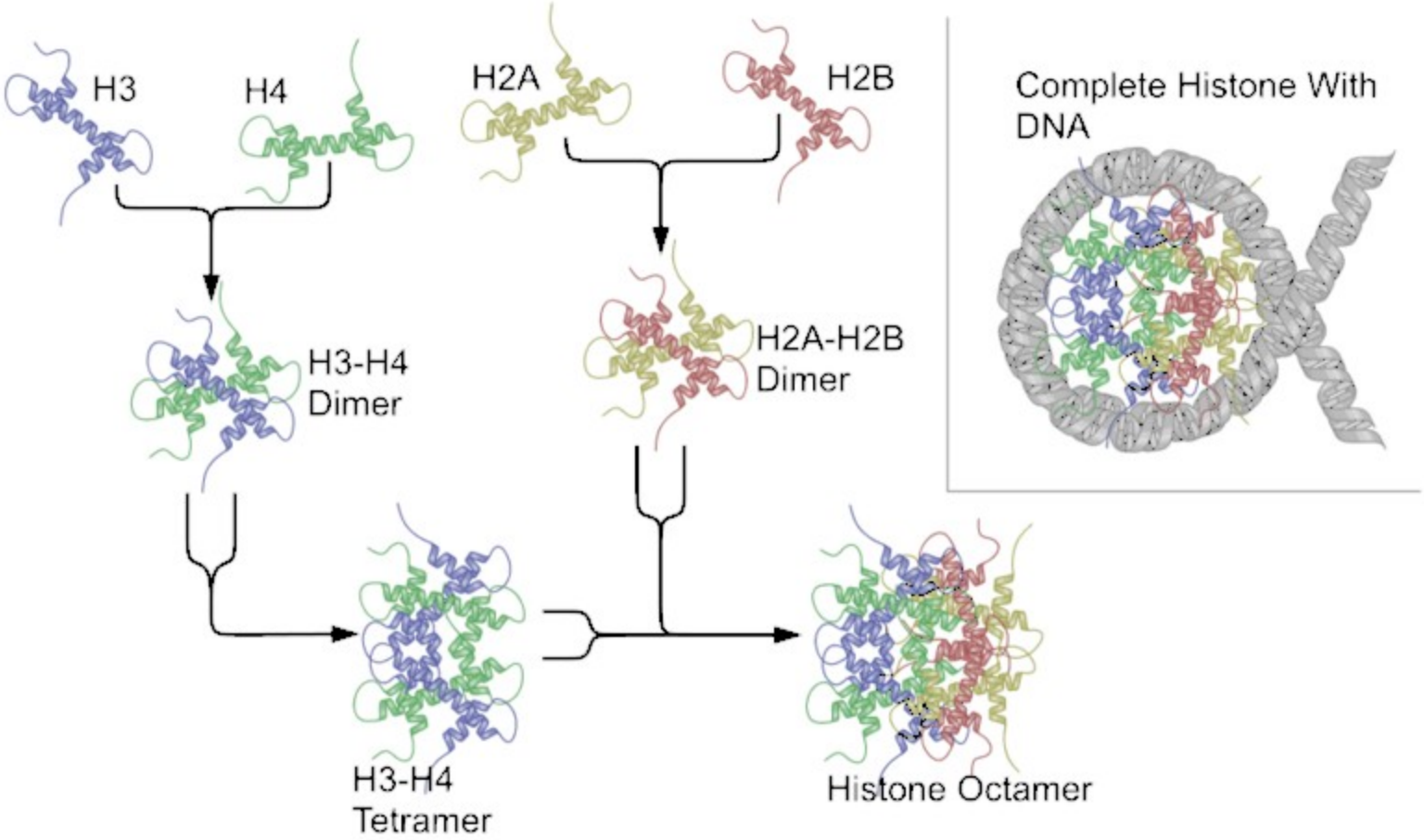
Nature 2004;429:457-73

Histones

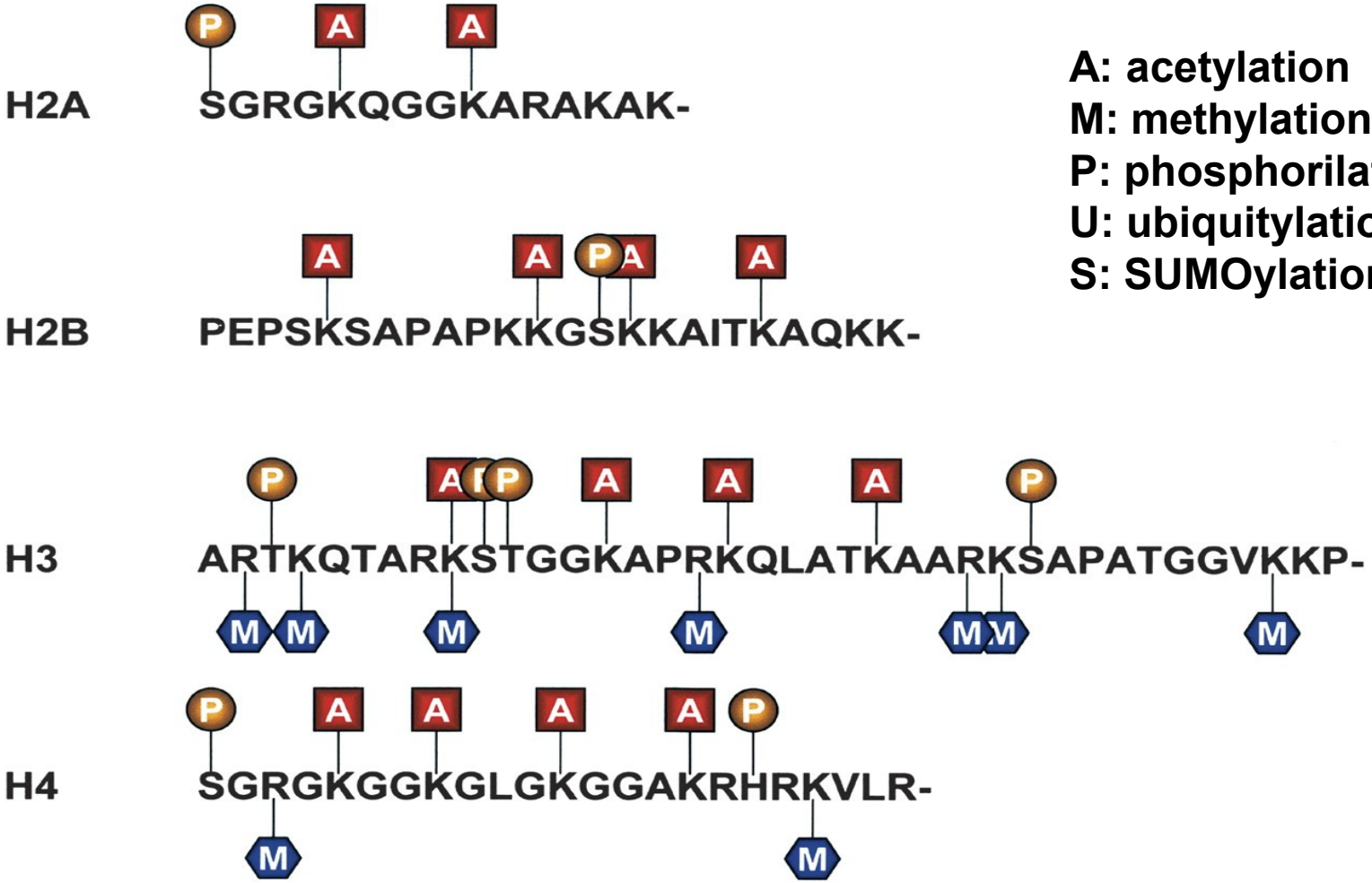
- Proteins in eukaryotic cells that package DNA into nucleosomes



Histone structure



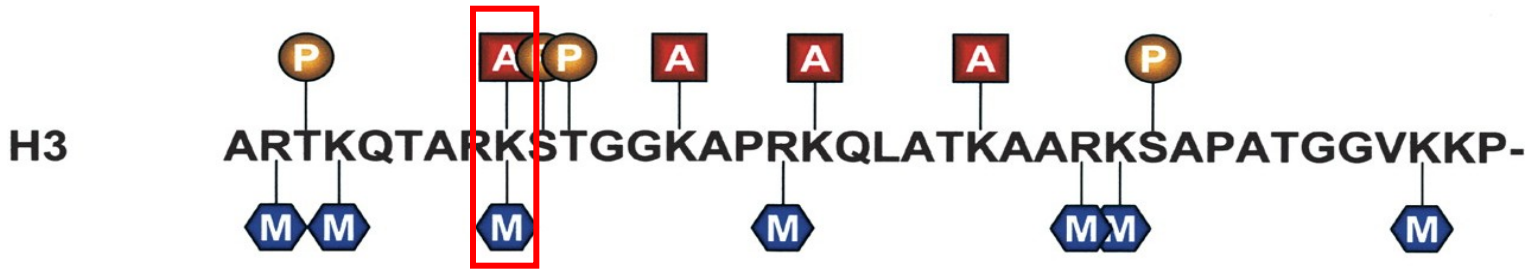
Histone modifications



A: acetylation
M: methylation
P: phosphorilation
U: ubiquitylation
S: SUMOylation

Lund and Lohuizen Genes Dev 2004

Histone modifications



- **Gene activation correlated with H3-K9 acetylation**
- **Gene silencing associated with H3-K9 methylation**

Histone Modifications and Human Diseases

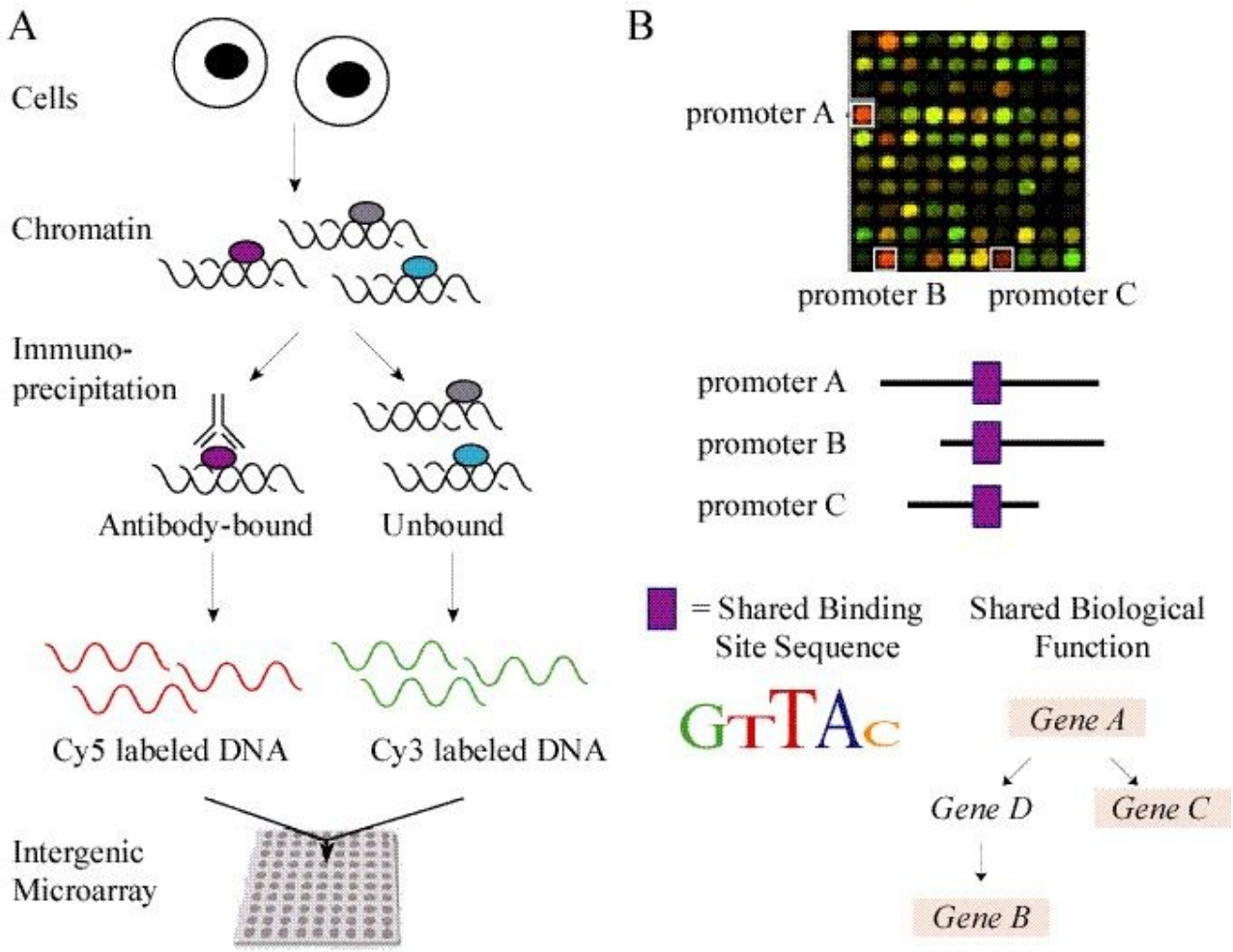
Coffin-Lowry syndrome is a rare genetic disorder characterized by mental retardation and abnormalities of the head and facial and other areas. It is caused by mutations in the RSK2 gene (histone phosphorylation) and is inherited as an X-linked dominant genetic trait. Males are usually more severely affected than females.

Rubinstein-Taybi syndrome is characterized by short stature, moderate to severe intellectual disability, distinctive facial features, and broad thumbs and first toes. It is caused by mutations in CREB-binding protein (histone acetylation)

Detection of histone modifications

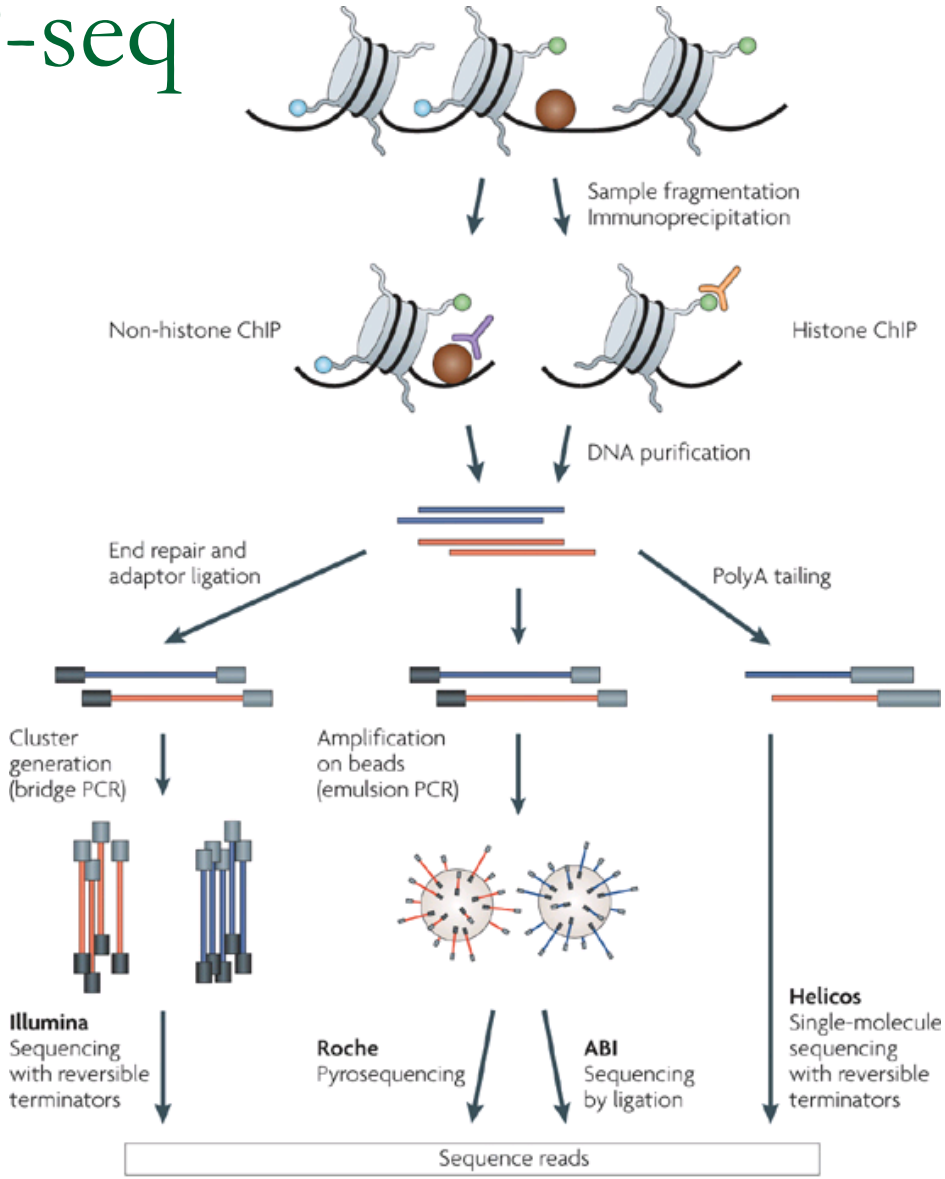
- ChIP: chromatin immunoprecipitation
  - Similar to MeDIP assay
  - Proteins are used to enrich for DNA that are packaged by modified histones
  - Collect, then
  - ChIP-on-chip: analyze with microarray
  - ChIP-seq: sequence

ChIP-chip



Wong and Chang, Journal of Investigative Dermatology, 2005

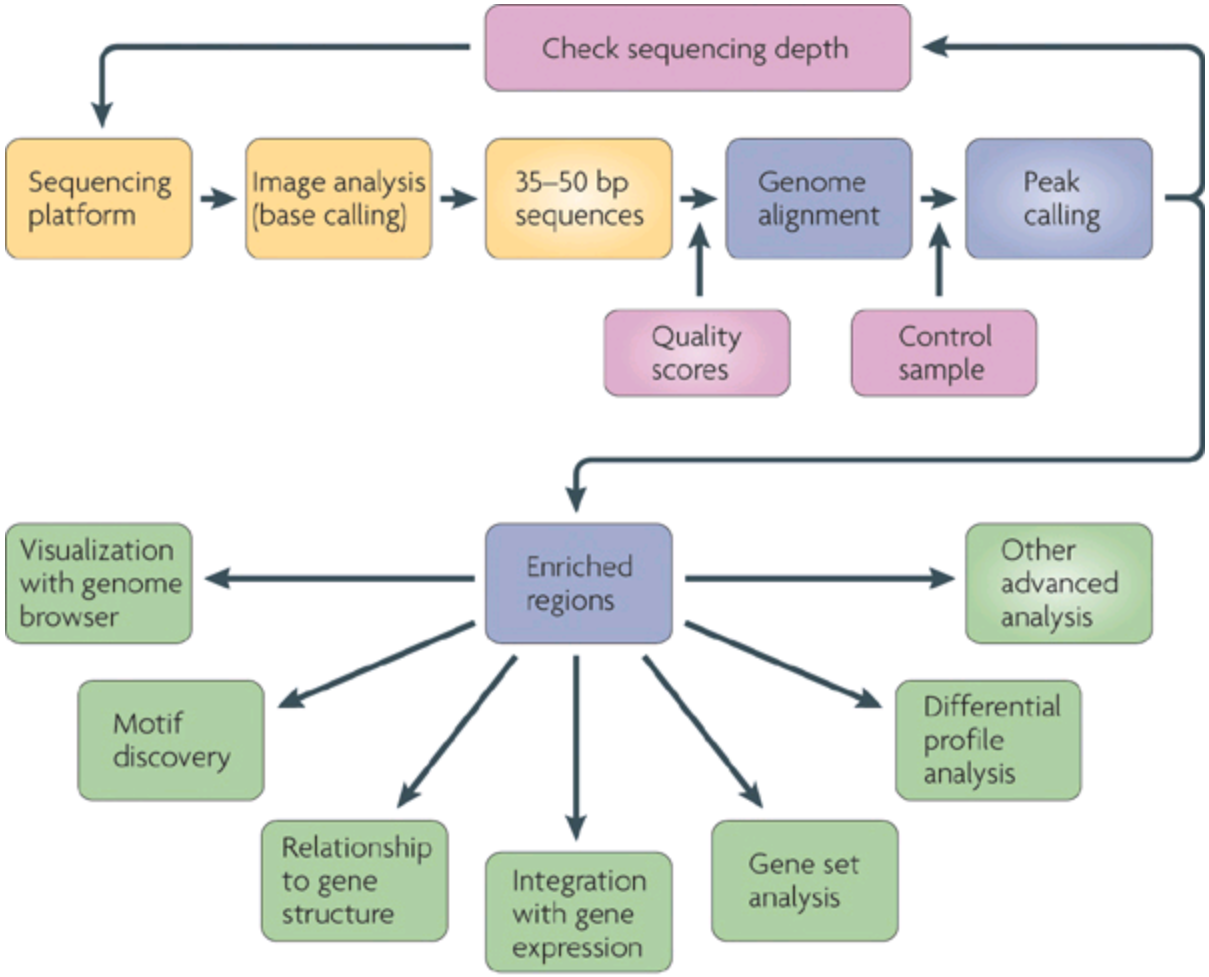
ChIP-seq



Nature Reviews | Genetics

Peter J Park, Nat Rev Genet, 2009

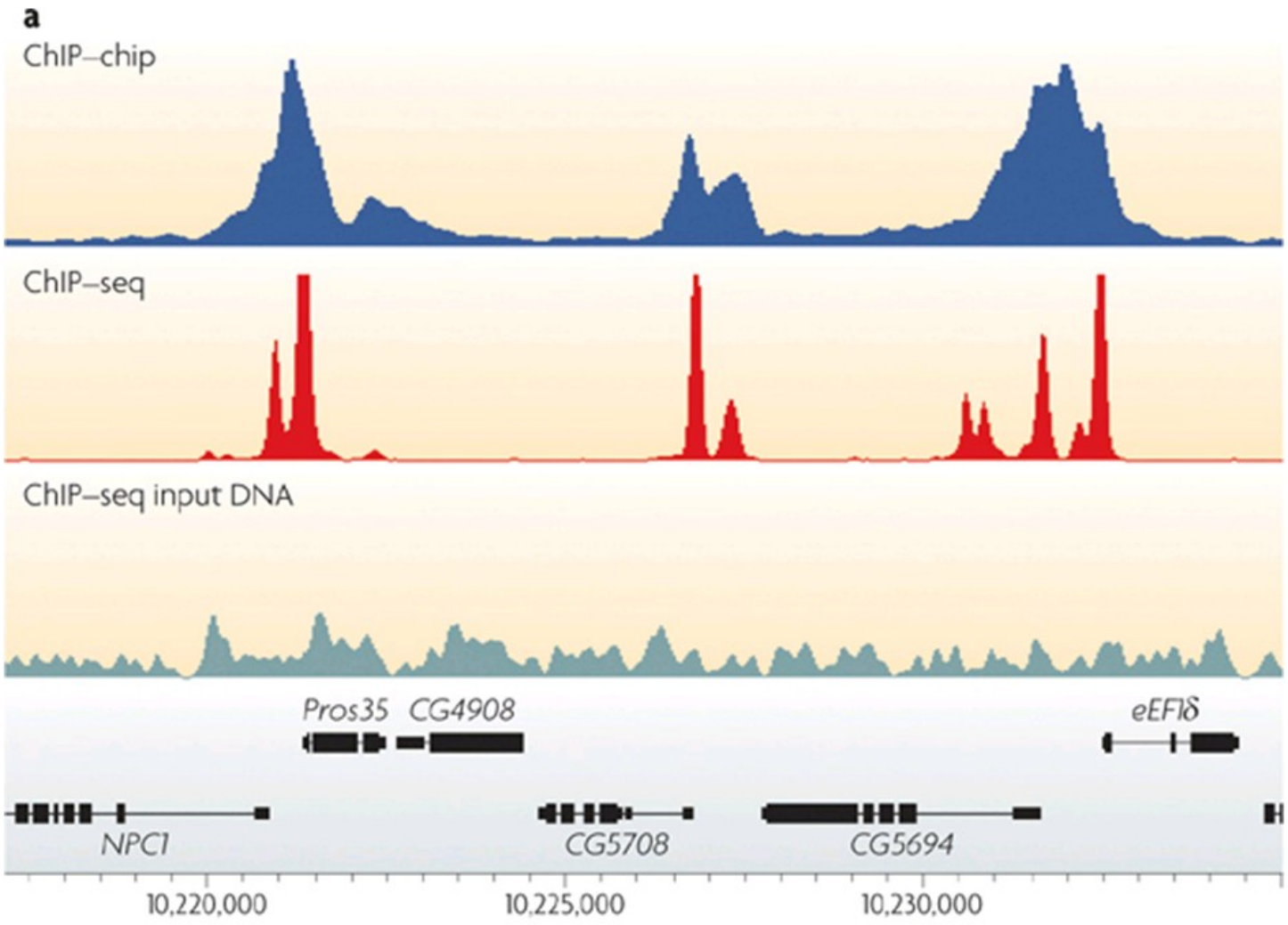
ChIP-seq



Nature Reviews | Genetics

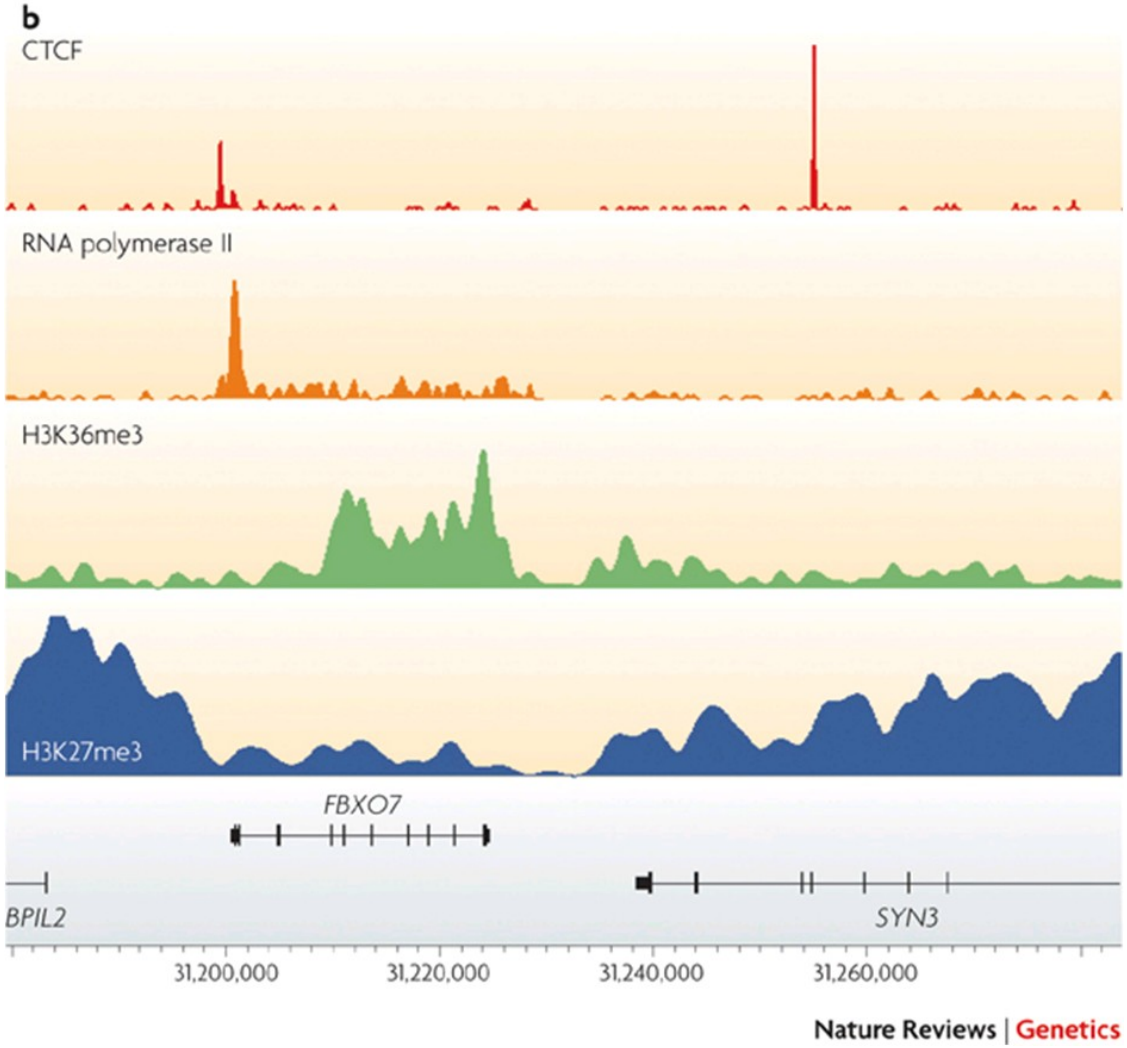
Peter J Park, Nat Rev Genet, 2009

Peaks: ChIP-chip vs ChIP-seq



Peter J Park, Nat Rev Genet, 2009

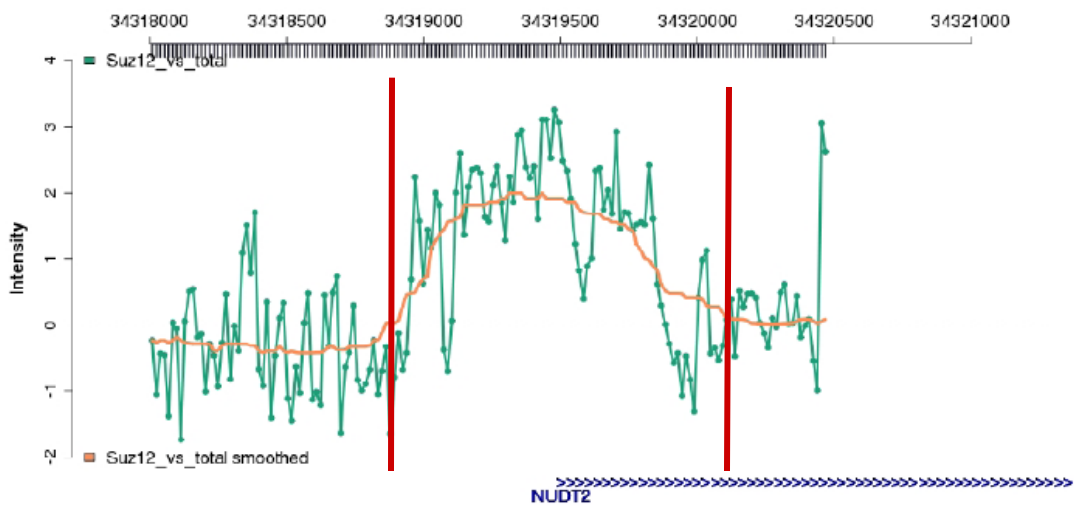
Peaks: ChIP-seq



Peter J Park, Nat Rev Genet, 2009

Peak calling

- Segmentation algorithms
  - HMMseg, etc.
  - Dynamic Bayesian Network based segmentation:
    - Segway (Hoffman et al., Nat Methods, 2012)
- Poisson models and binomial distribution
  - PeakSeq (Rozowsky et al., Nat Biotech, 2009)



RNA FOLDING

RNA folding

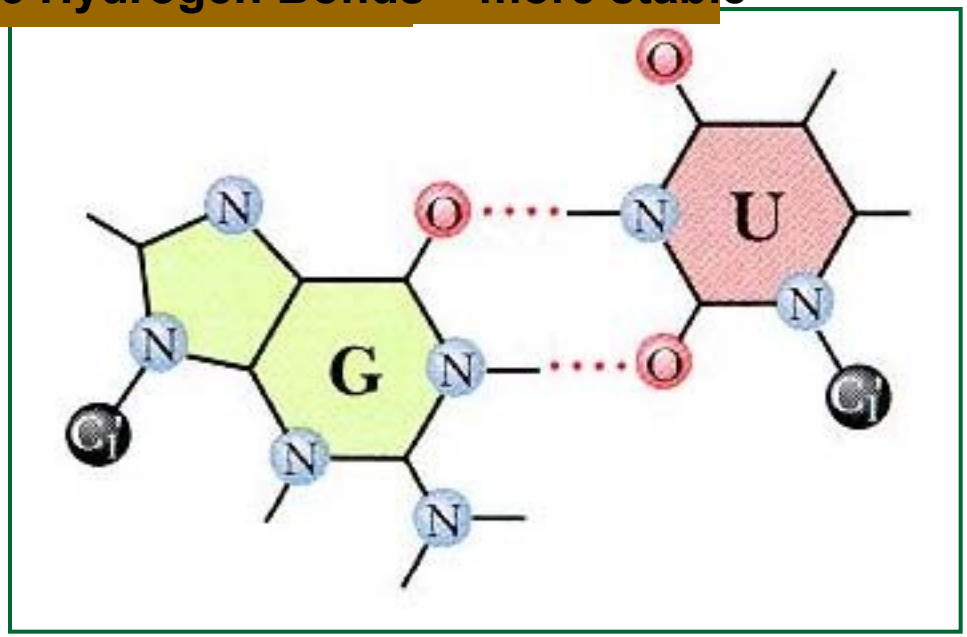
- Prediction of secondary structure of an RNA given its sequence
- General problem is NP-hard due to “difficult” substructures, like pseudoknots
- Most existing algorithms require too much memory ($\geq O(n^2)$), and run time ($\geq O(n^3)$) thus limited to smaller RNA sequences

RNA Basics

- RNA bases A,C,G,U
- Canonical Base Pairs
  - A-U
  - G-C
  - G-U

    "wobble" pairing
  - Bases can only pair with **one** other base.

3 Hydrogen Bonds – more stable

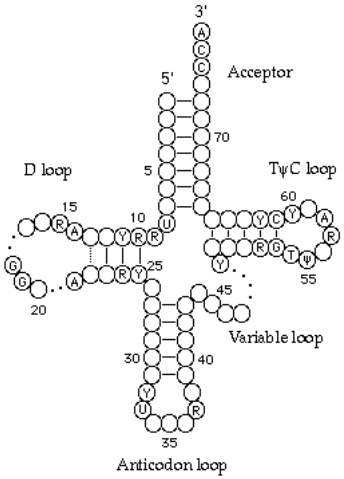


http://www.bioalgorithms.info/

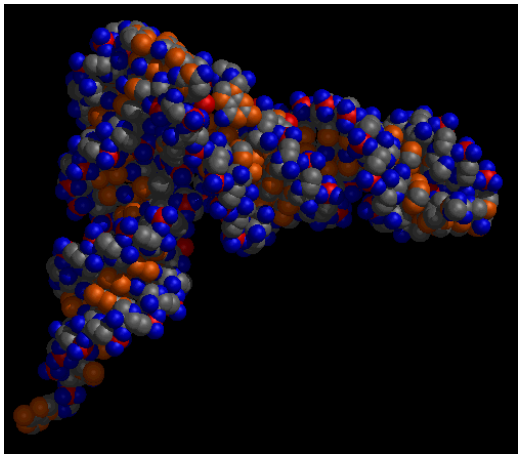
RNA Structural Levels

AAUCG…CUUCUUCCA

Primary



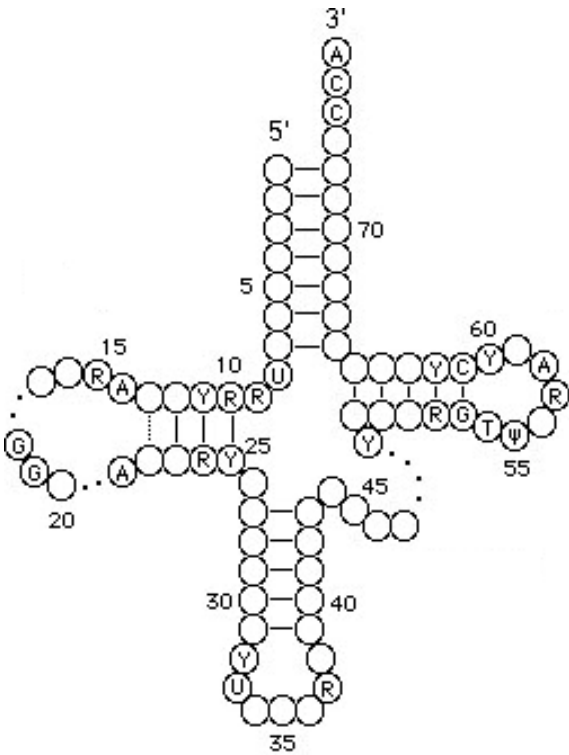
Secondary



Tertiary

RNA Basics

- transfer RNA (tRNA)
- messenger RNA (mRNA)
- ribosomal RNA (rRNA)
- small interfering RNA (siRNA)
- micro RNA (miRNA)
- small nucleolar RNA (snoRNA)



http://www.genetics.wustl.edu/eddy/tRNAscan-SE/

RNA families

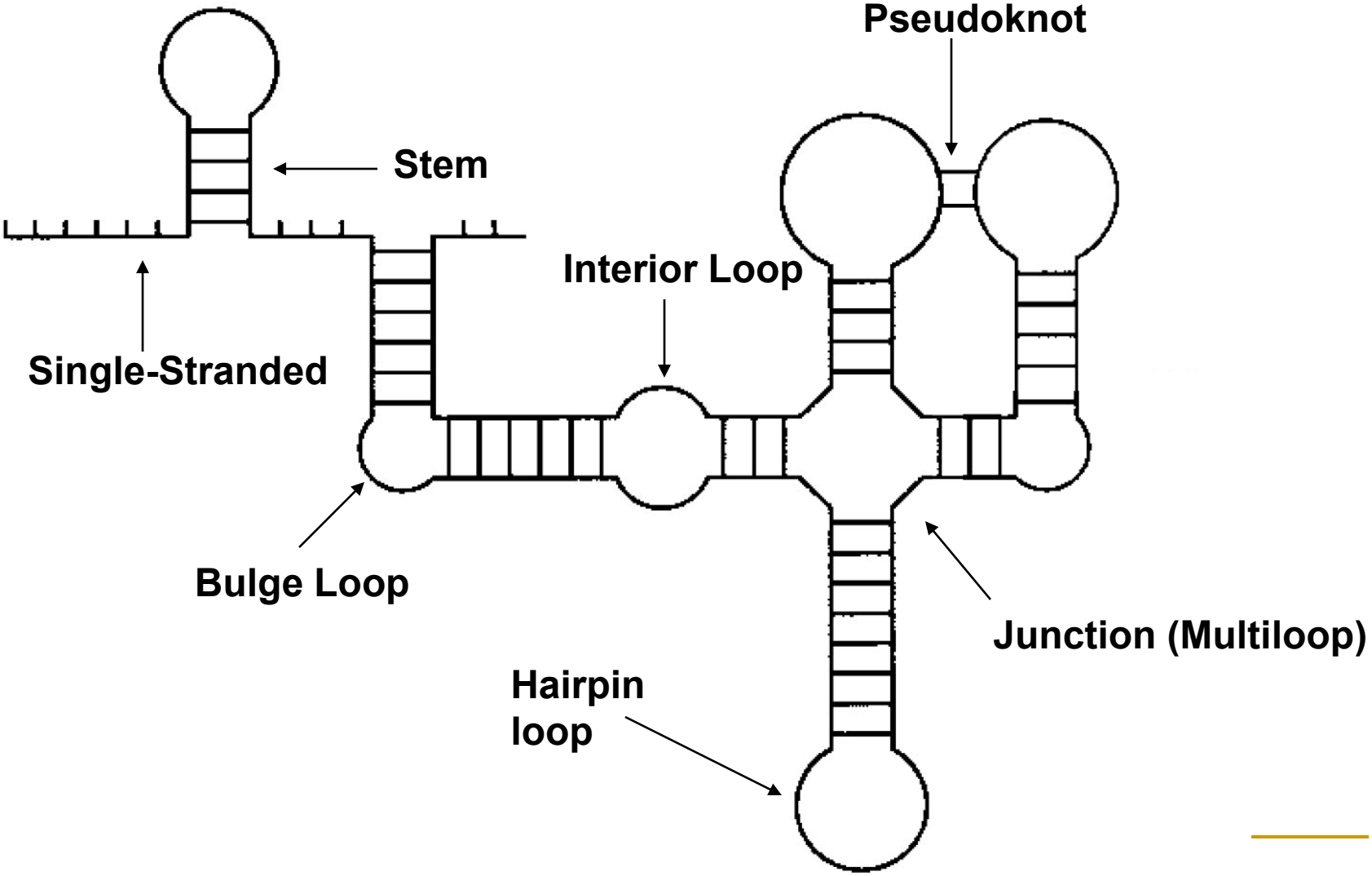
- Rfam : General non-coding RNA database

(most of the data is taken from specific databases)

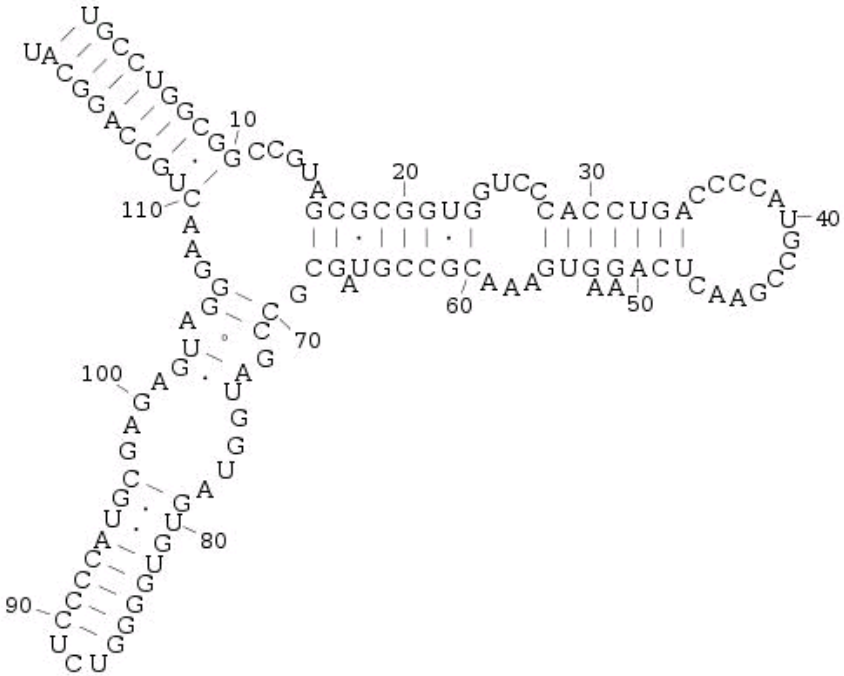
http://www.sanger.ac.uk/Software/Rfam/

Includes many families of non coding RNAs and functional Motifs, as well as their alignement and their secondary structures

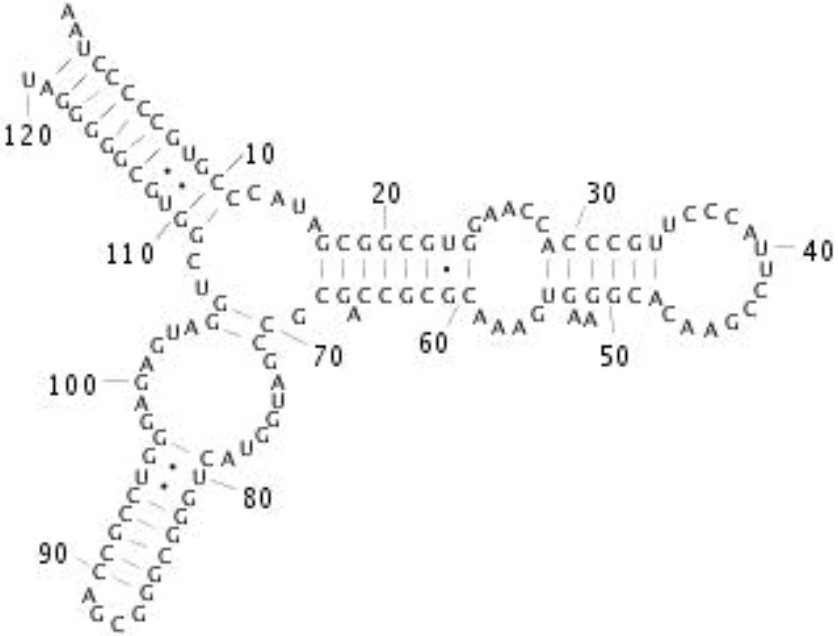
RNA Secondary Structure



Example: 5S rRNA

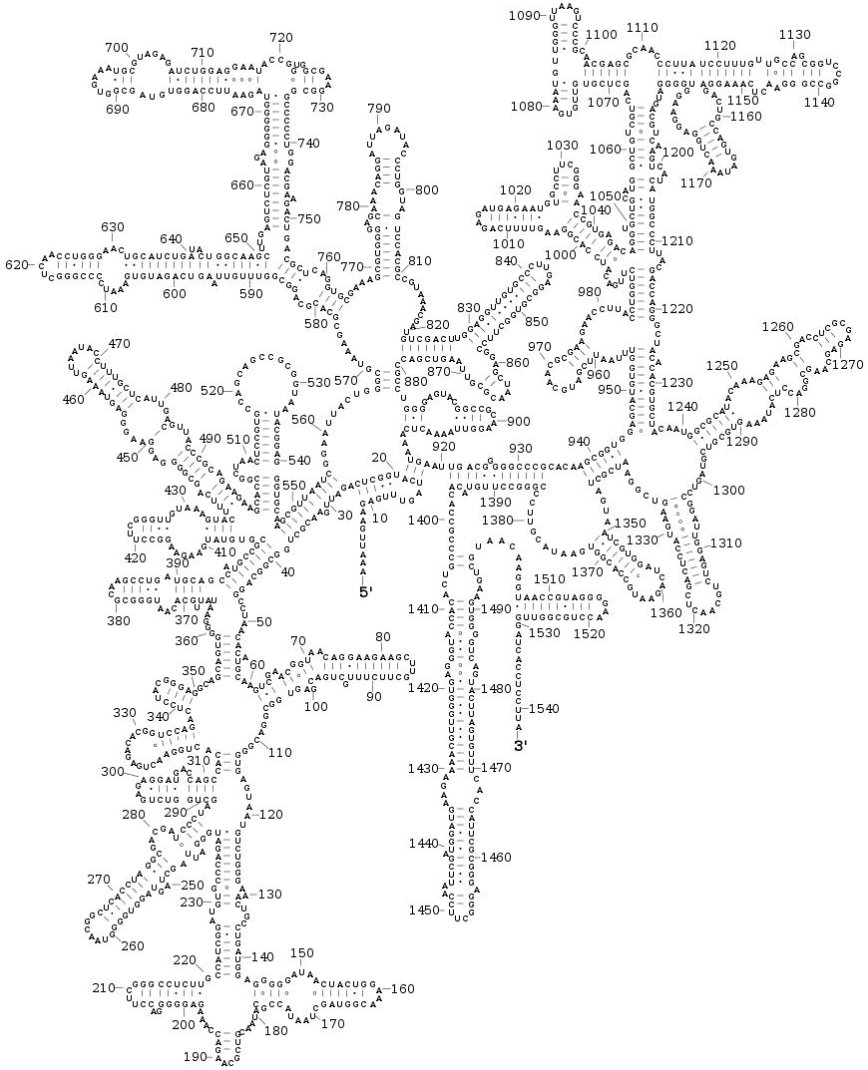


E. coli 5S
120 bases



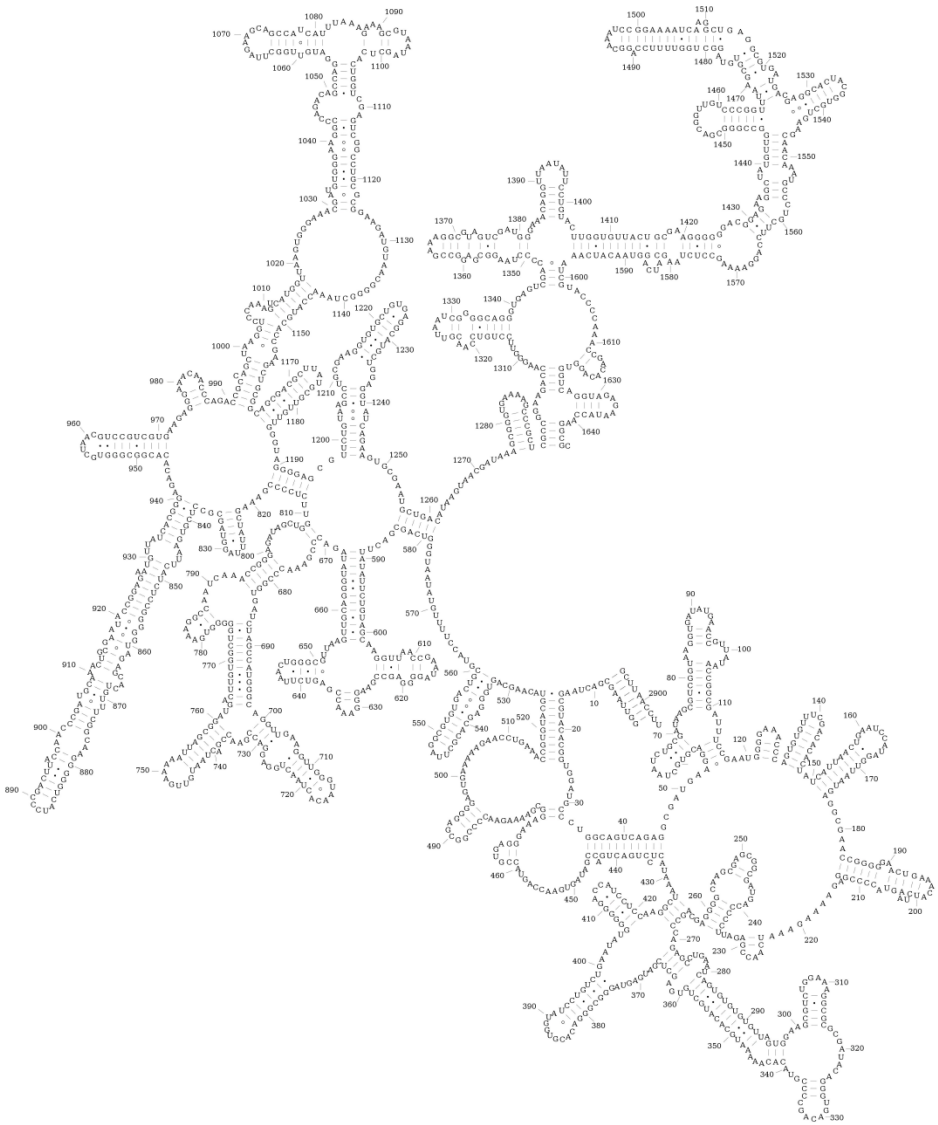
T. thermophilus 5S
120 bases

Example: E. coli 16S rRNA



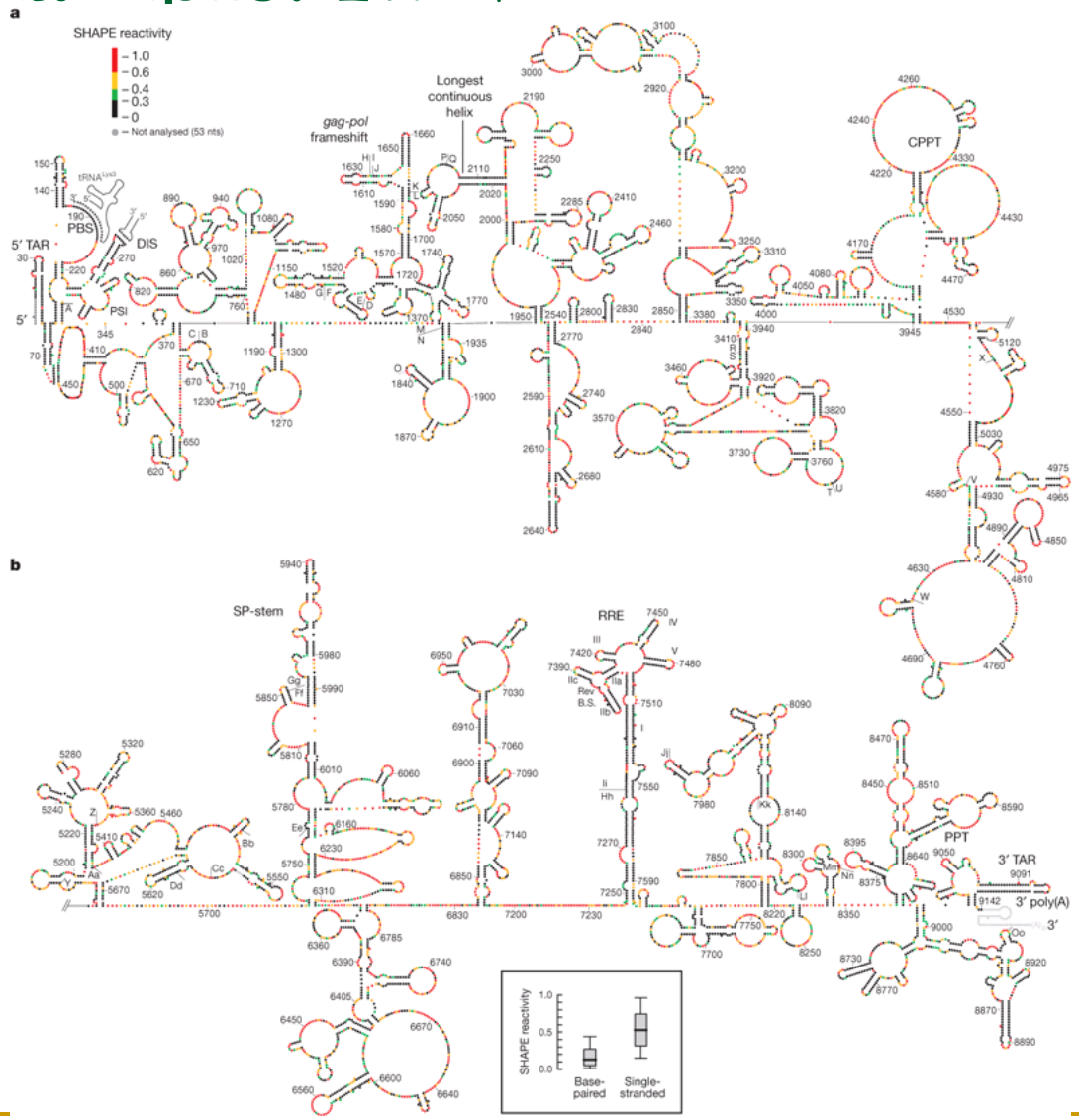
1542 bases

Example: E. coli 23S rRNA



2904 bases

Example: HIV

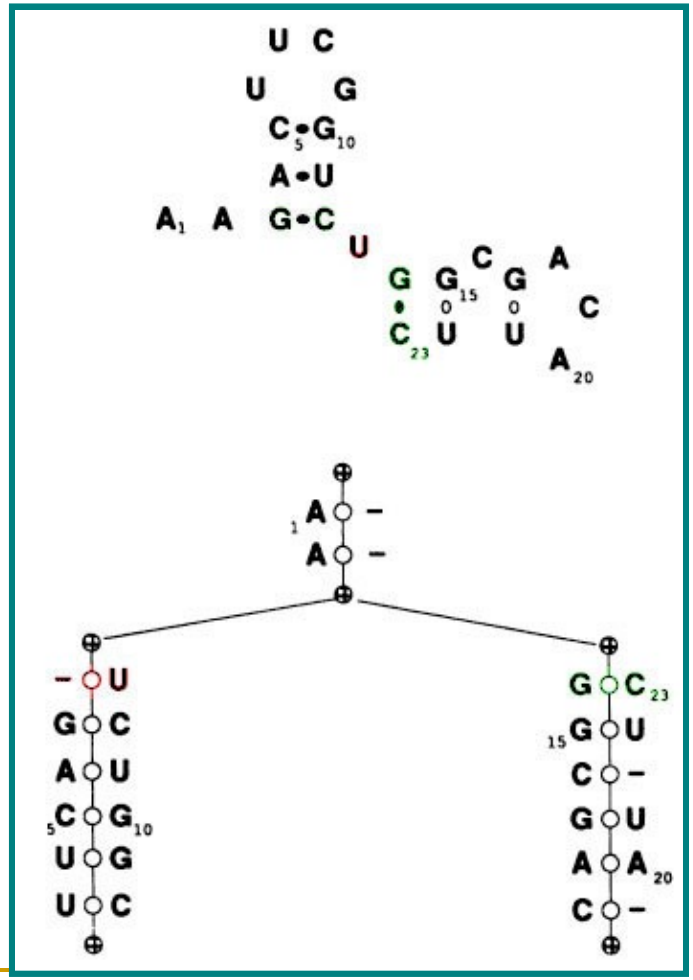


9173 bases

Watts et al., Nature, 2009

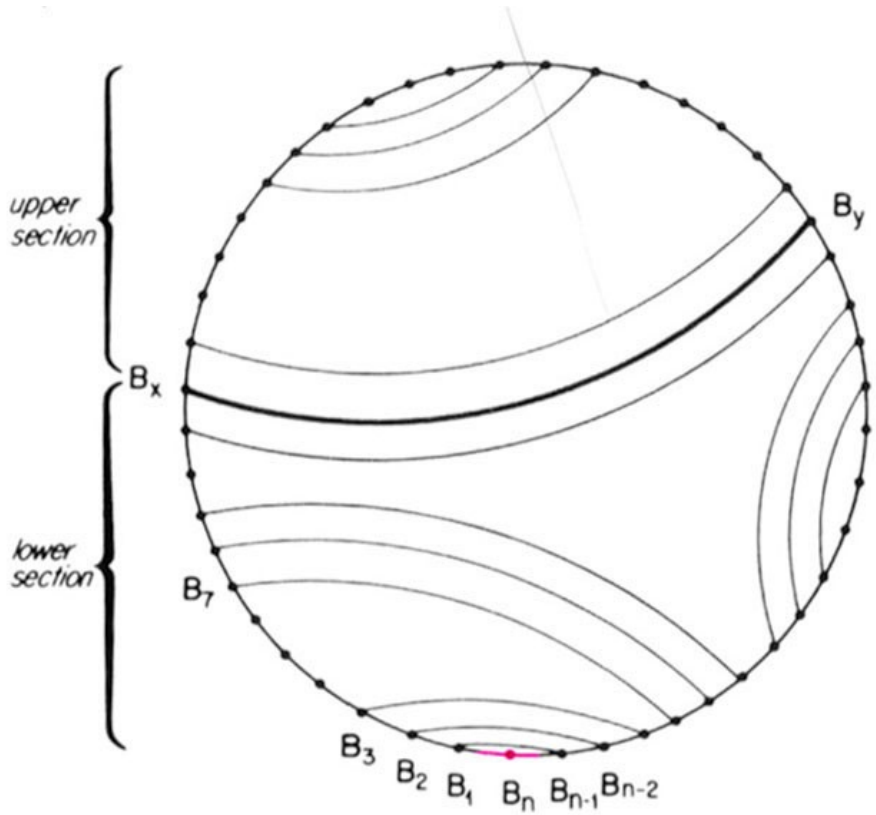
Binary Tree Representation of RNA Secondary Structure

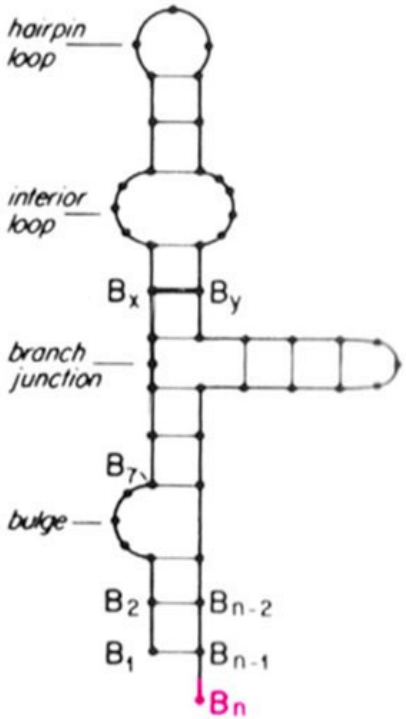
- Representation of RNA structure using Binary tree
- Nodes represent
  - Base pair if two bases are shown
  - Loop if base and “gap” (dash) are shown
- Pseudoknots still not represented
- Tree does not permit varying sequences
  - Mismatches
  - Insertions & Deletions



Images – Eddy et al.

Circular Representation

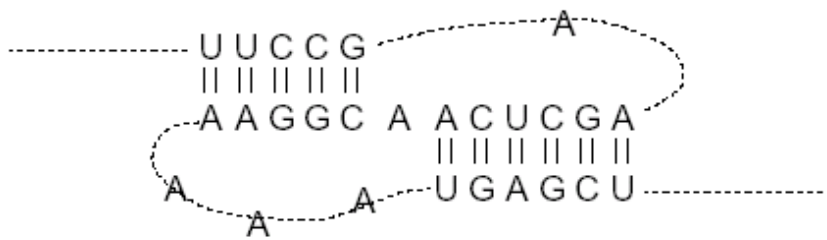


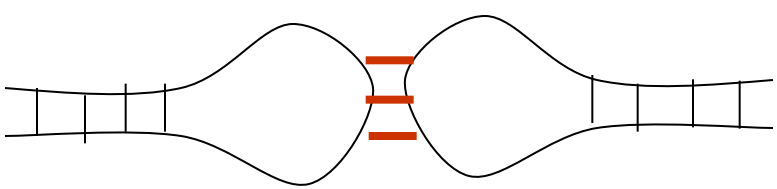


Images – David Mount

Examples of known interactions of RNA secondary structural elements

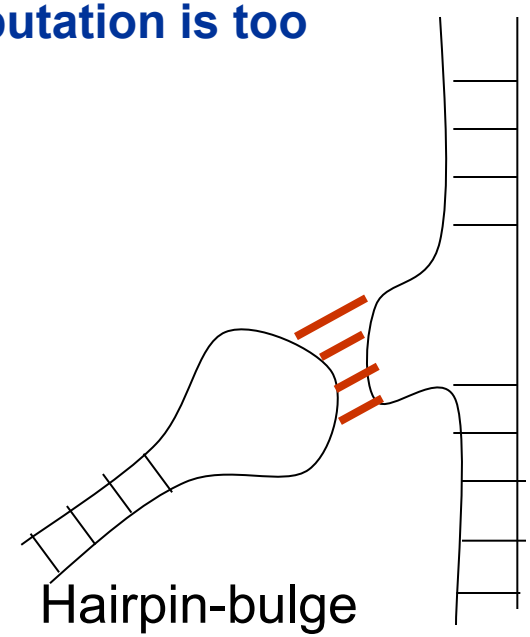
Pseudoknot





Kissing hairpins

These patterns are excluded from the prediction schemes as their computation is too intensive.



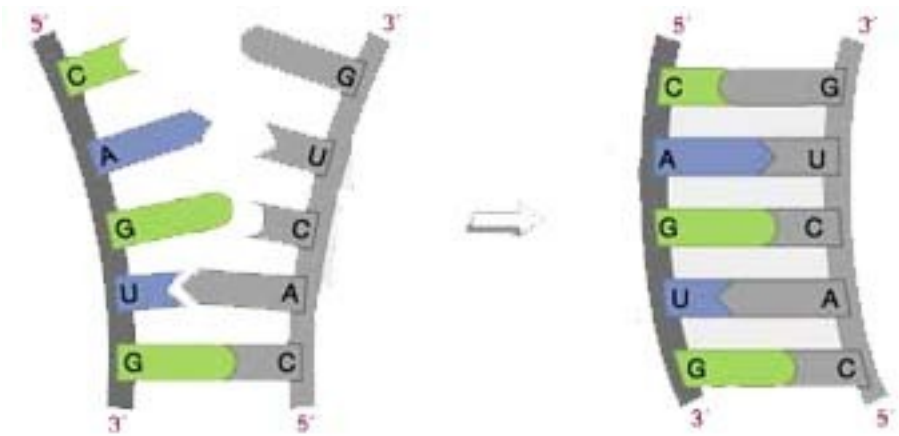
Hairpin-bulge contact

Predicting RNA secondary structure

- Base pair maximization
- Minimum free energy (most common)
  - Fold, Mfold (Zuker & Stiegler)
  - RNAfold (Hofacker)
- Multiple sequence alignment
  - Use known structure of RNA with similar sequence
- Covariance
- Stochastic Context-Free Grammars

Sequence Alignment as a method to determine structure

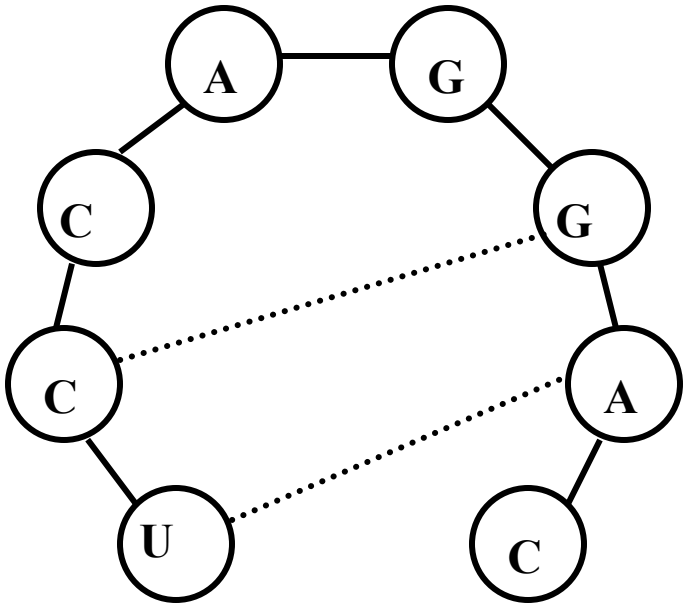
- Bases pair in order to form backbones and determine the secondary structure
- Aligning bases based on their ability to pair with each other gives an algorithmic approach to determining the optimal structure

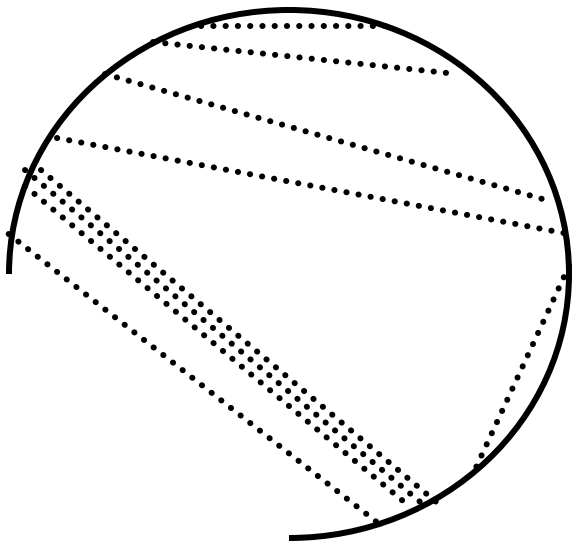


Simplifying Assumptions

- RNA folds into one minimum free-energy structure.
- There are no knots (base pairs never cross).
- The energy of a particular base pair in a double stranded regions is sequence independent
  - Neighbors do not influence the energy.
- Was solved by dynamic programming, Zuker and Stiegler 1981

Base Pair Maximization





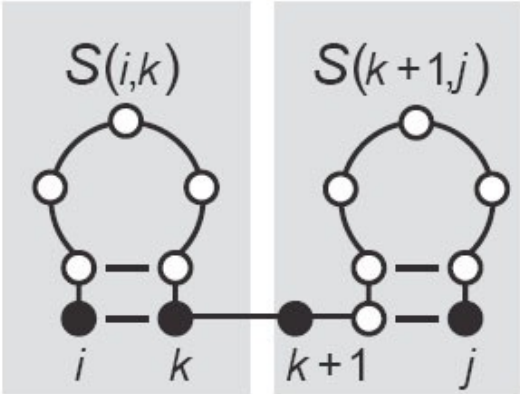
Zuker (1981) *Nucleic Acids Research* 9(1) 133-149

Base Pair Maximization – Dynamic Programming Algorithm

S(i,j) is the folding of the subsequence of the RNA strand from index i to index j which results in the highest number of base pairs

Maximizing Base Pair

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



Base Pair Maximization – Dynamic Programming Algorithm

- **Alignment Method**
  - **Align RNA strand to itself**
  - **Score increases for feasible base pairs**
- **Each score independent of overall structure**
- **Bifurcation adds extra dimension**

S(i + 1, j)

i \ j	G	G	G	A	A	A	U	C	C
G	0	0	0	0	0	0	1	2	2
G	0	0	0	0	0	0	1	2	3
G		0	0	0	0	0	1	2	2
A			0	0	0	0	1	1	1
A				0	0	0	1	1	1
A					0	0	1	1	1
U						0	0	0	0
C							0	0	0
C								0	0

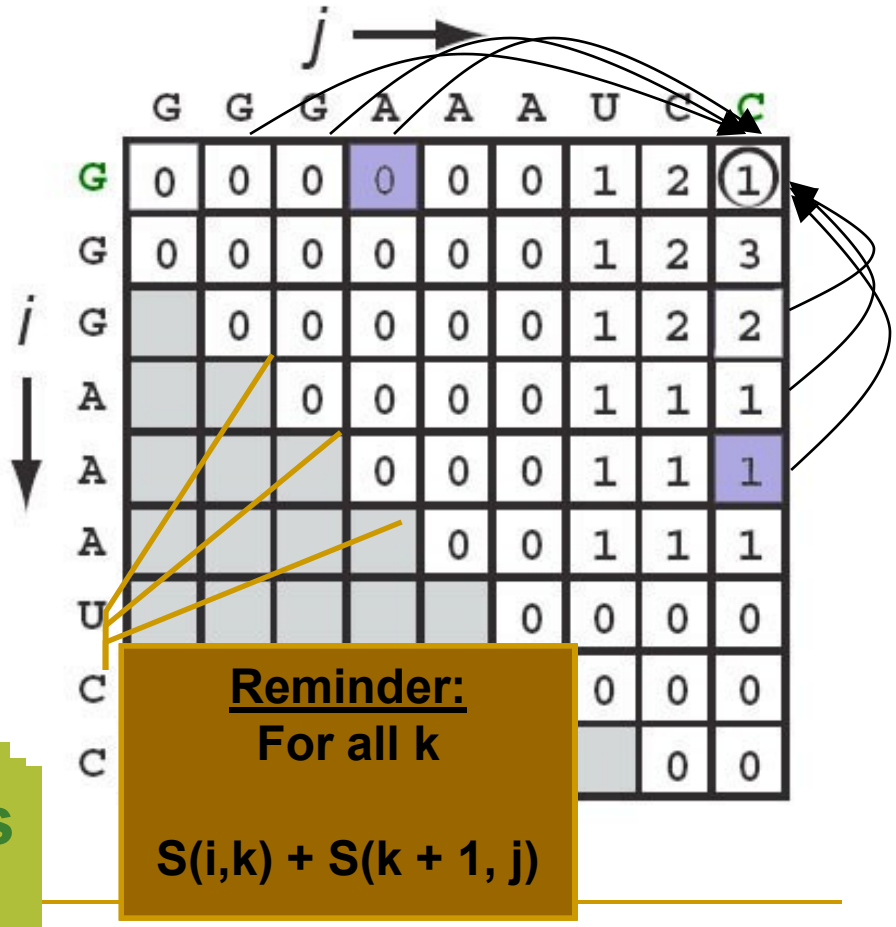
S(i + 1, j – 1) +1

Dynamic Programming – possible paths

Images – Sean Eddy

Base Pair Maximization – Dynamic Programming Algorithm

- **Alignment Method**
  - **Align RNA strand to itself**
  - **Score increases for feasible base pairs**
- **Each score independent of overall structure**
- **Bifurcation adds extra dimension**



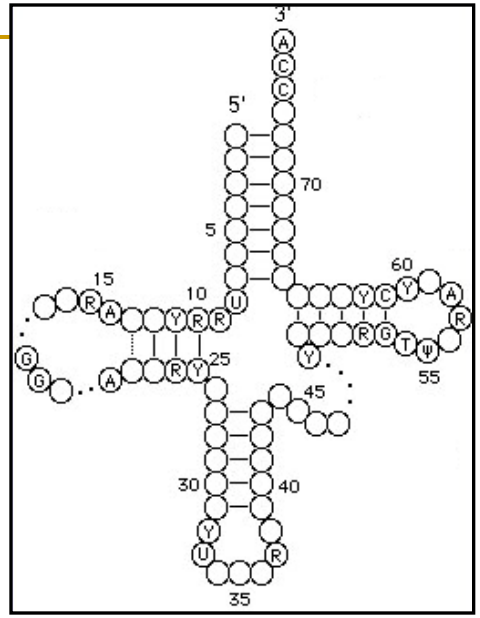
Bifurcation – add values for all k

Images – Sean Ed

Base Pair Maximization - Drawbacks

- Base pair maximization will not necessarily lead to the most stable structure
  - May create structure with many interior loops or hairpins which are energetically unfavorable
- Comparable to aligning sequences with scattered matches – not biologically reasonable

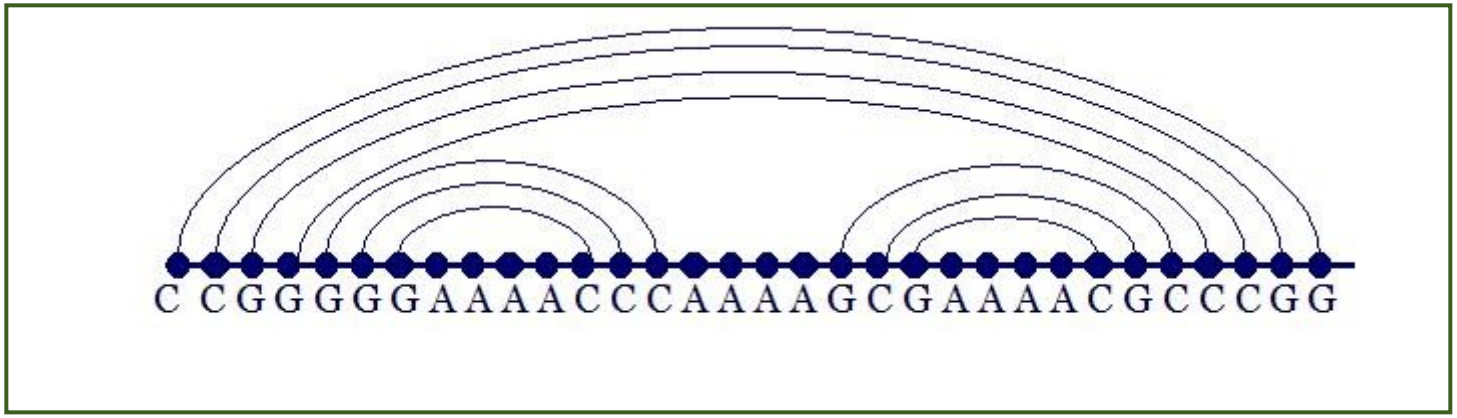
Energy Minimization



- Thermodynamic Stability
  - Estimated using experimental techniques
  - Theory : Most Stable is the Most likely
- No Pseudoknots due to algorithm limitations
- Uses Dynamic Programming alignment technique
- Attempts to maximize the score taking into account thermodynamics
- MFOLD and ViennaRNA

Free energy model

Free energy of a structure is the sum of all interactions energies

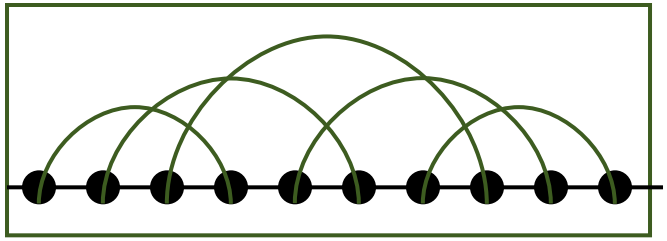


Free Energy(E) = E(CG)+E(CG)+…..

Each interaction energy can be calculated thermodynamically

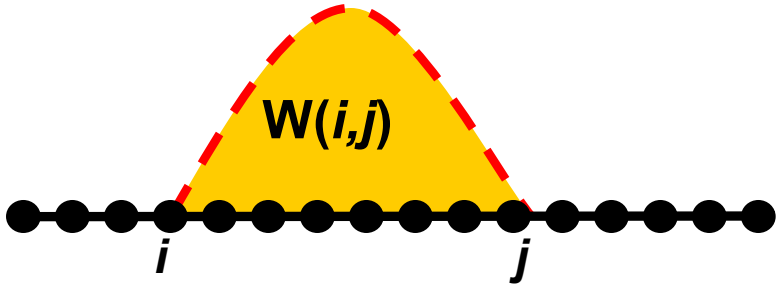
Why is MFE secondary structure prediction hard?

- MFE structure can be found by calculating free energy of all possible structures
- BUT the number of potential structures grows exponentially with the number, n, of bases



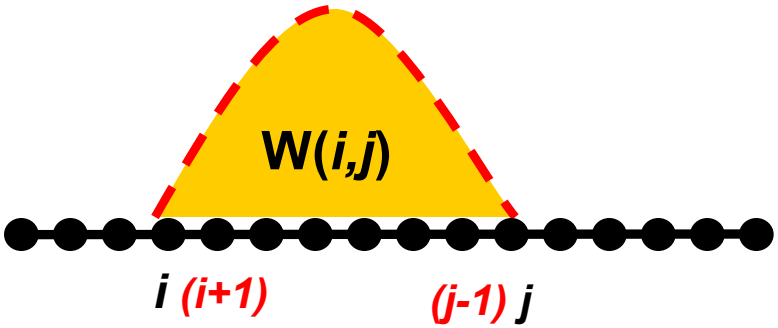
RNA folding with Dynamic programming (Zuker and Stiegler)

- **W(i,j):** MFE structure of substrand from i to j



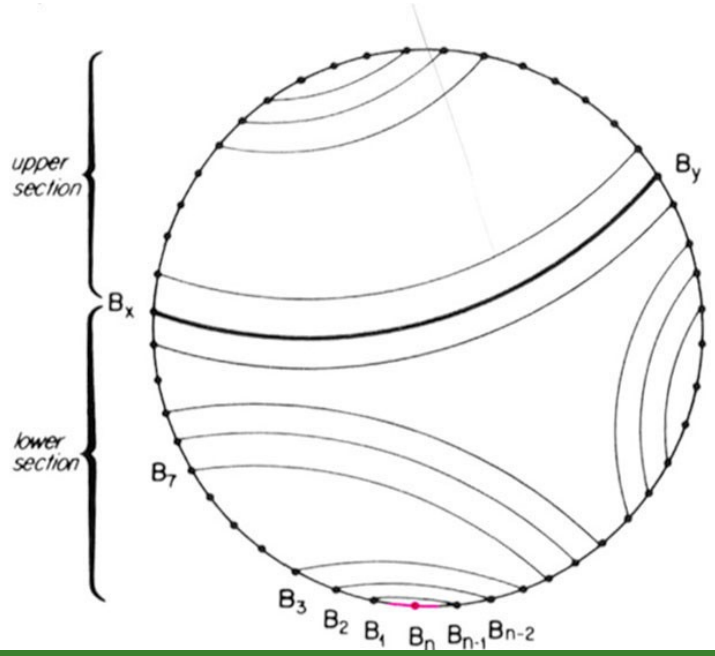
RNA folding with dynamic programming

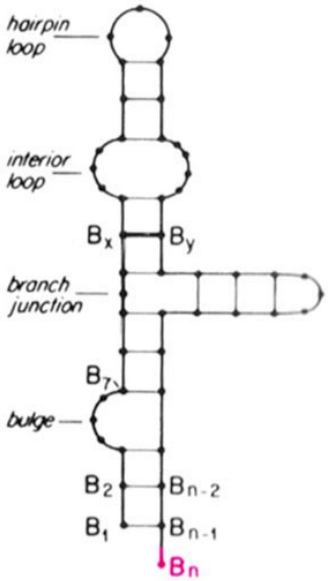
- Assume a function W(i,j) which is the MFE for the sequence starting at i and ending at j (i<j)



- Define scores, for example base pair (CG) =-1 non-pair(CA)=1 (we want a negative score)
- Consider 4 possibilities:
  - i,j are a base pair, added to the structure for *i+1..j-1*
  - i is unpaired, added to the structure for *i+1..j*
  - j is unpaired, added to the structure for *i..j-1*
  - i,j are paired, but not to each other;
- Choose the minimal energy

Energy Minimization Results



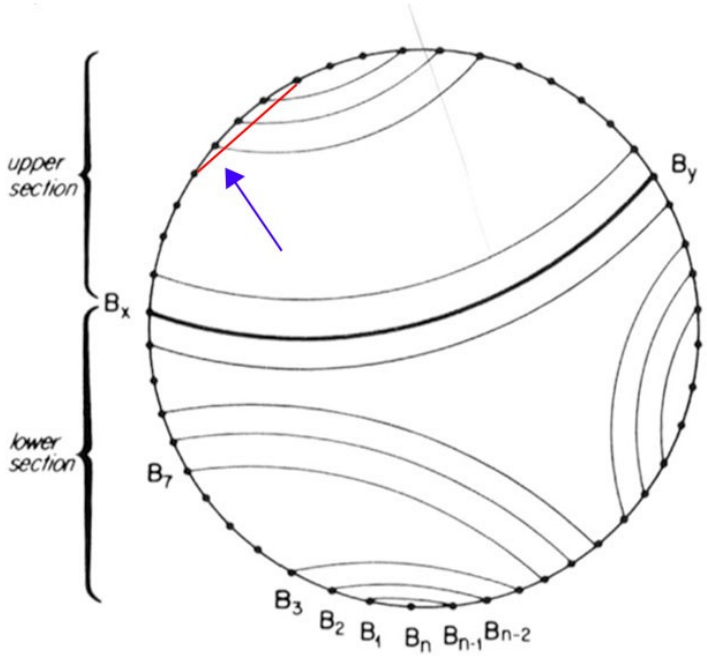


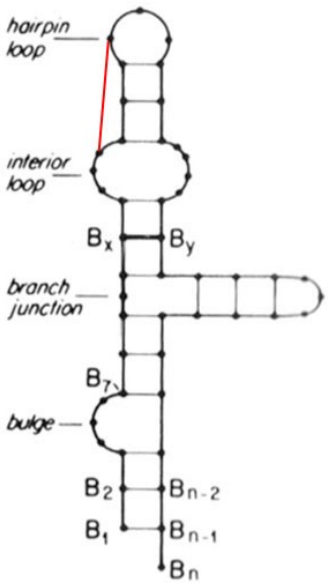
- **All loops must have at least 3 bases in them**
  - **Equivalent to having 3 base pairs between all arcs**

Exception: Location where the beginning and end of RNA come together in circularized representation

Images – David Mount

Trouble with Pseudoknots



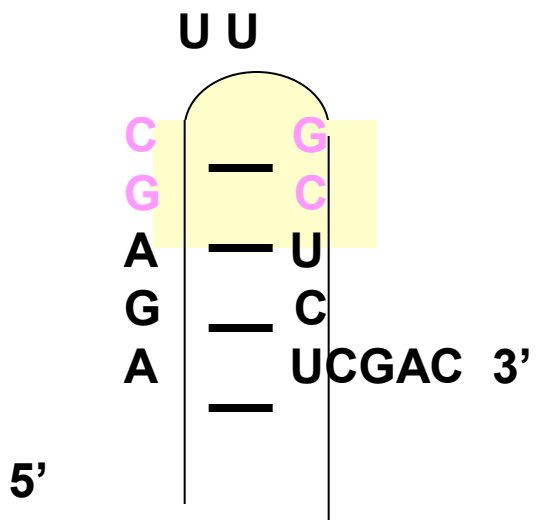


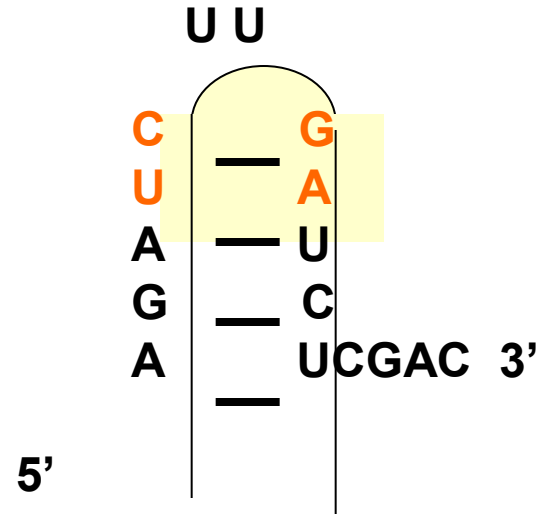
- Pseudoknots cause a breakdown in the Dynamic Programming Algorithm.
- In order to form a pseudoknot, checks must be made to ensure base is not already paired – this breaks down the recurrence relations

Images – David Mount

Sequence dependent free-energy

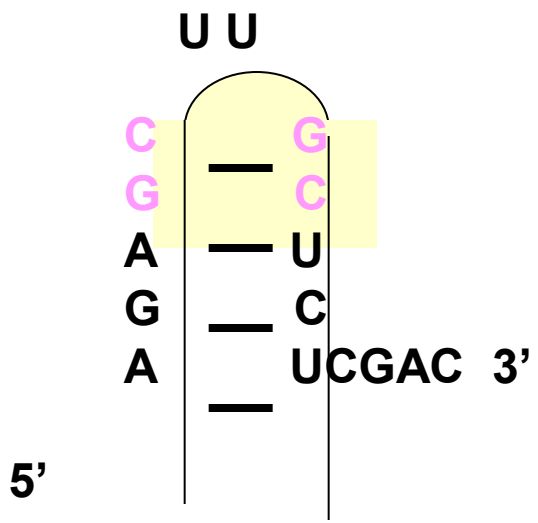
Nearest Neighbor Model

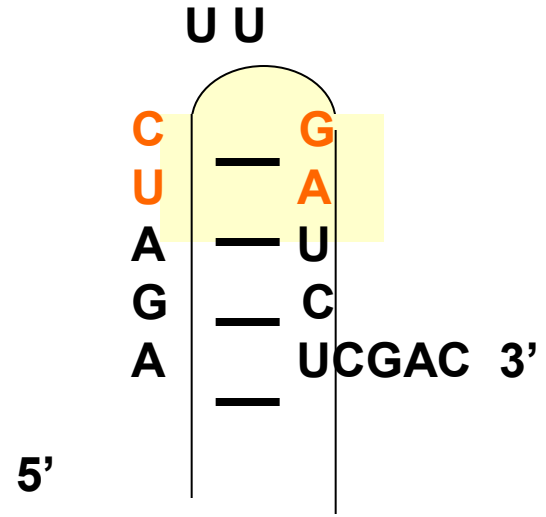




Energy is influenced by the previous base pair (not by the base pairs further down).

Sequence dependent free-energy values of the base pairs





These energies are estimated experimentally from small synthetic RNAs.

Example values:

GC	GC	GC	GC
AU	GC	CG	UA
-2.3	-2.9	-3.4	-2.1

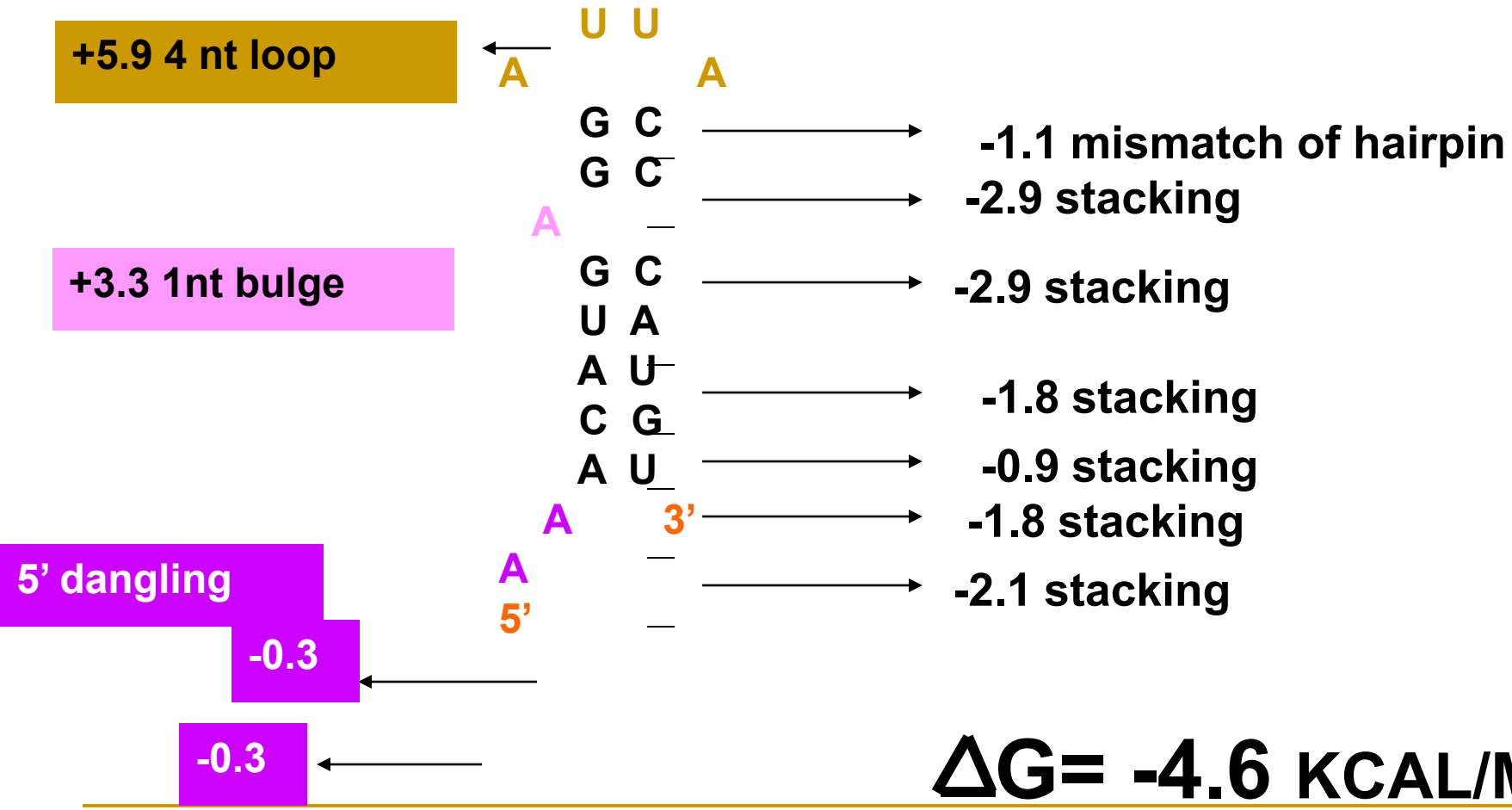
Adding Complexity to Energy Calculations

- Stacking energy - Assign negative energies to these *between base pair* regions.
  - Energy is influenced by the previous base pair (not by the base pairs further down).
  - These energies are estimated experimentally from small synthetic RNAs.
- Positive energy - added for destabilizing regions such as bulges, loops, etc.
- More than one structure can be predicted

Mfold

- Positive energy - added for destabilizing regions such as bulges, loops, etc.
- More than one structure can be predicted

Free energy computation

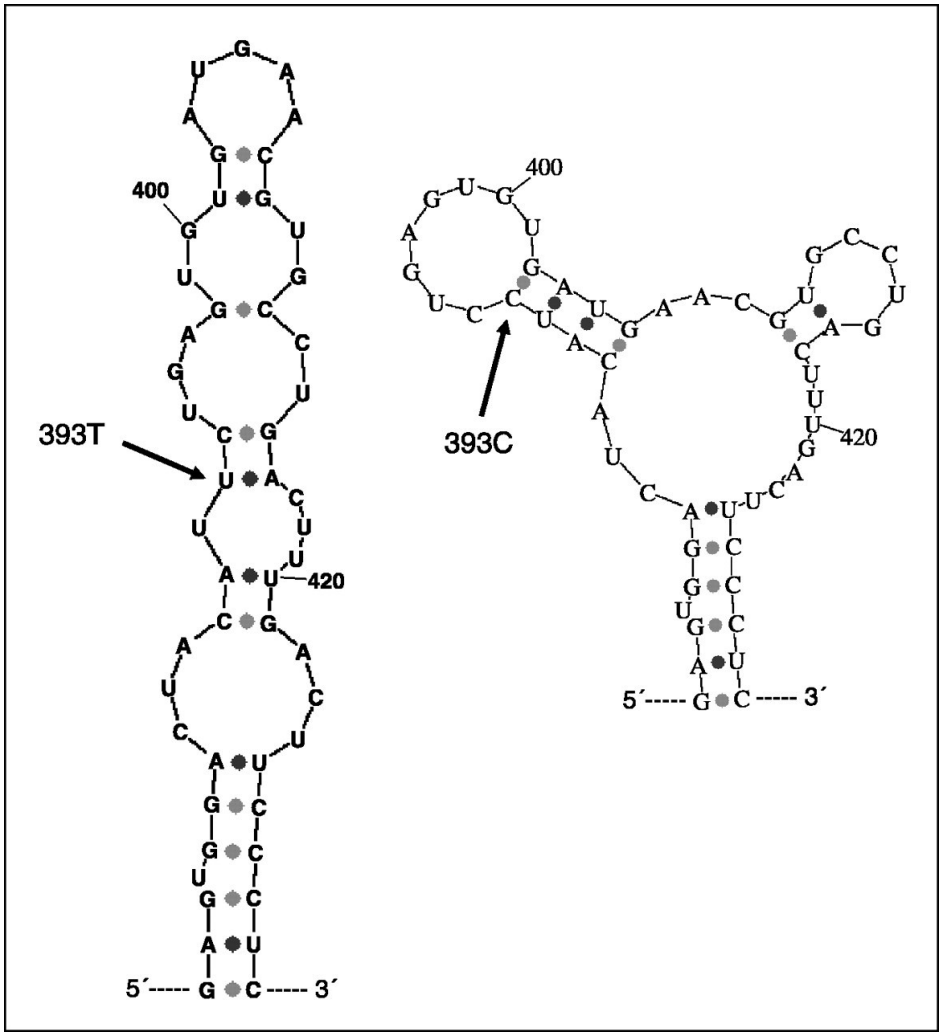


ΔG= -4.6 KCAL/MOL

Mfold

- Positive energy - added for destabilizing regions such as bulges, loops, etc.
- More than one structure can be predicted

More than one structure can be predicted for the same RNA



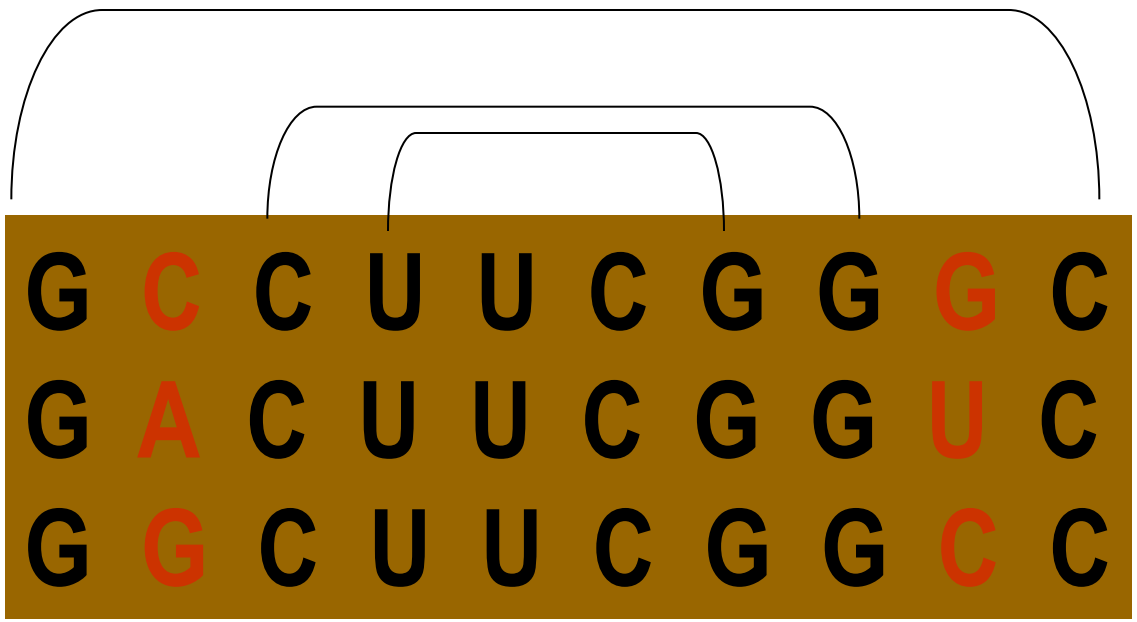
Frey U H et al. Clin Cancer Res 2005;11:5071-5077

Energy Minimization Drawbacks

- Compute only one optimal structure
- Usual drawbacks of purely mathematical approaches
  - Similar difficulties in other algorithms
    - Protein structure
    - Exon finding

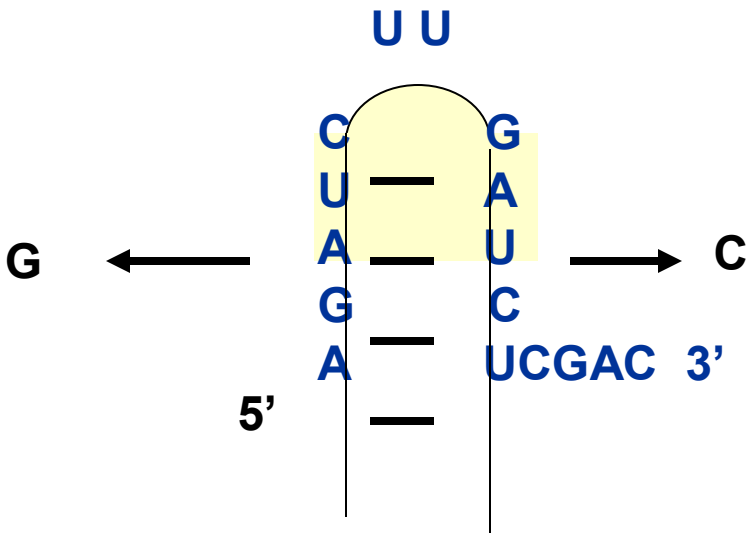
RNA fold prediction based on Multiple Alignment

Information from multiple sequence alignment (MSA) can help to predict the probability of positions i,j to be base-paired.

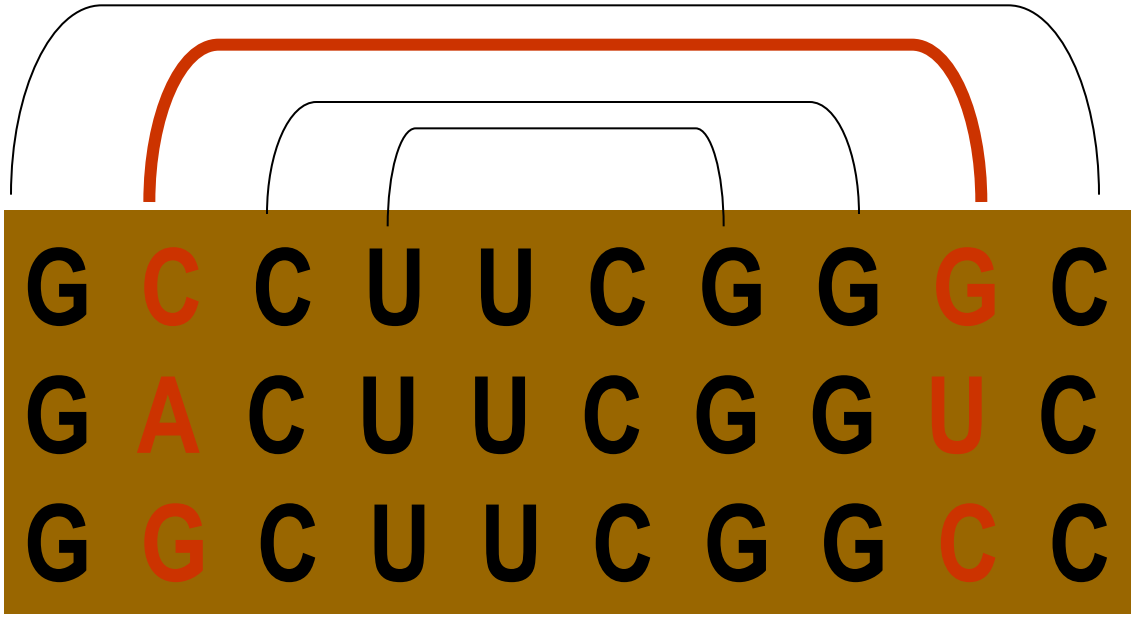


Compensatory Substitutions

Mutations that maintain the secondary structure can help predict the fold



RNA secondary structure can be revealed by identification of compensatory mutations



Insight from Multiple Alignment

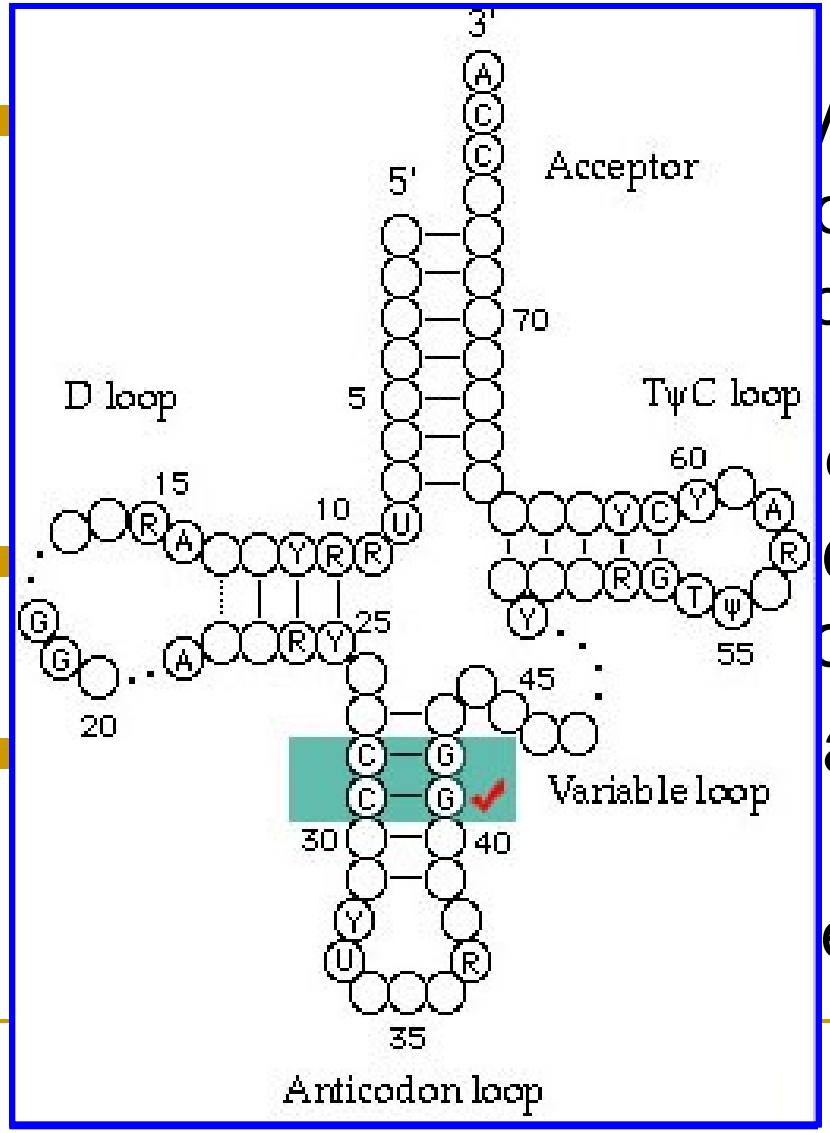
Information from multiple sequence alignment (MSA) can help to predict the probability of positions i,j to be base-paired.

- Conservation – no additional information
- Consistent mutations (GC→ GU) – support stem
- Inconsistent mutations – does not support stem.
- Compensatory mutations – support stem.

RNAalifold

- Predicts the consensus secondary structure for a set of aligned RNA sequences by using modified dynamic programming algorithm that add alignment information to the standard energy model
- Improvement in prediction accuracy

Alternative Algorithms - Covariaton



- y-based method
  quences that are important
  bin
  pa
  co
- d
  covarying base pairs
- ased on results

e Grammar

Covariation ensures ability to base pair is maintained and RNA structure is conserved

Covariance Model

- HMM which permits flexible alignment to an RNA structure –
  - emission and transition probabilities
- Model trees based on finite number of states
  - Match states – sequence conforms to the model:
    - MATP – State in which bases are paired in the model and sequence
    - MATL & MATR – State in which either right or left bulges in the sequence and the model
  - Deletion – State in which there is deletion in the sequence when compared to the model
  - Insertion – State in which there is an insertion relative to model
- Transitions have probabilities
  - Varying probability – Enter insertion, remain in current state, etc
  - Bifurcation – no probability, describes path

Covariance Model (CM) Training Algorithm

- S(i,j) = Score at indices i and j in RNA when aligned to the Covariance Model

$$S(i,j) = \max \begin{cases} S(i+1,j-1) + \mathrm{M}(i,j) \\ S(i+1,j) \\ S(i,j-1) \\ \max_{i<k<j} S(i,k) + S(k+1,j) \end{cases}$$

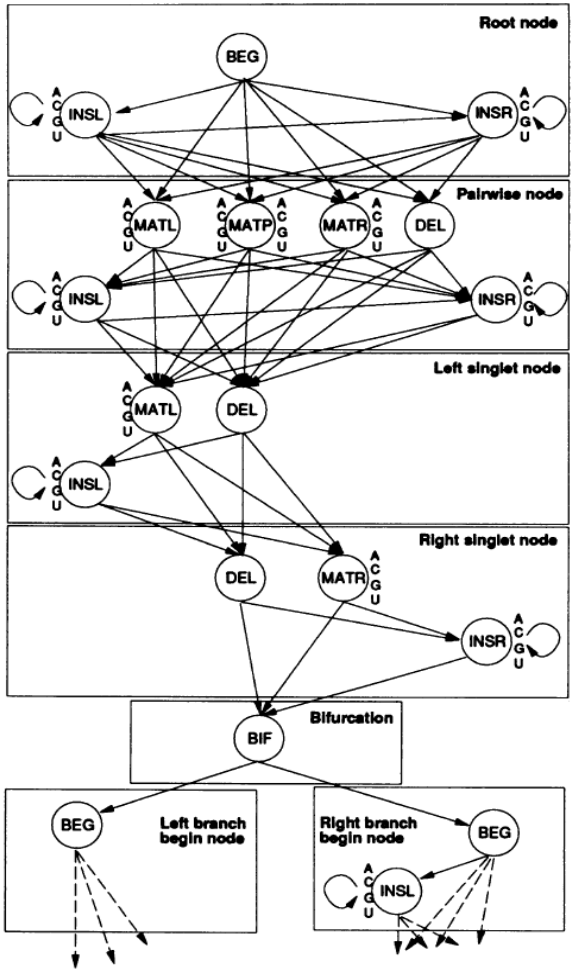
$$M_{i,j} = \sum_{x_i,x_j} f_{x_i x_j} \log_2 \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}}$$

Frequency of seeing the symbols (A, C, G, T) together in locations i and j depending on symbol.

Independent frequency of seeing the symbols (A, C, G, T) in locations i or j depending on symbol.

- Frequencies obtained by aligning model to “training data” – consists of sample sequences
  - Reflect values which optimize alignment of sequences to model

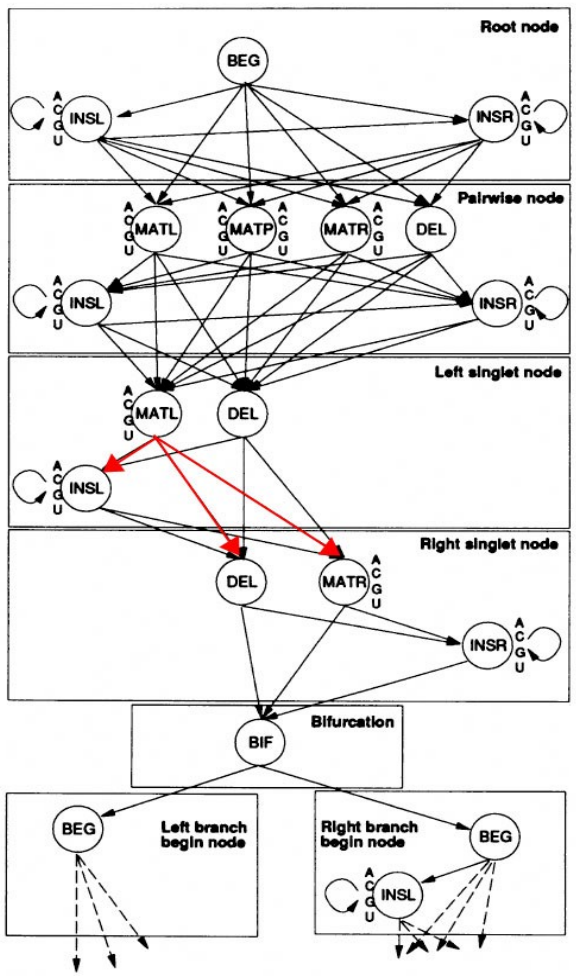
Alignment to CM Algorithm



- Calculate the probability score of aligning RNA to CM
- Three dimensional matrix – $O(n^3)$
  - Align sequence to given subtrees in CM
  - For each subsequence calculate all possible states
- Subtrees evolve from Bifurcations
  - For simplicity Left singlet is default

Images – Eddy et al.

Alignment to CM Algorithm



- For each calculation take into account the
  - Transition (T) to next state
  - Emission probability (P) in the state as

determined by training data

Deletion – does not have an emission probability (P) associated with it

$$S_{i,j,y}(y = BIFURC) = \max_{i-1<=mid<=j} [S_{i,mid,y_{left}} + S_{mid+1,j,y_{right}}]$$

Covariance Model Drawbacks

- Needs to be well trained
- Not suitable for searches of large RNA
  - Structural complexity of large RNA cannot be modeled
  - Runtime
  - Memory requirements