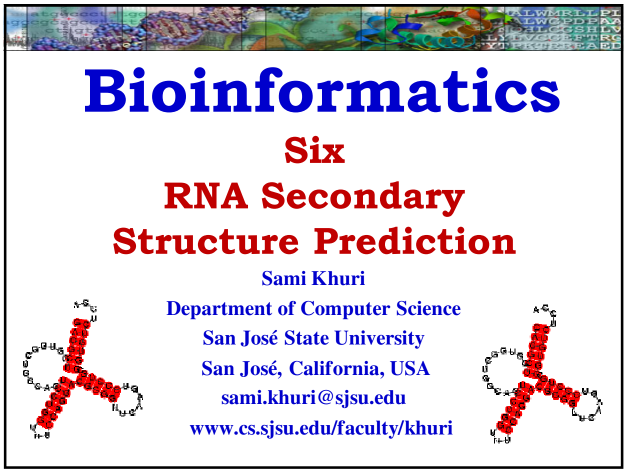
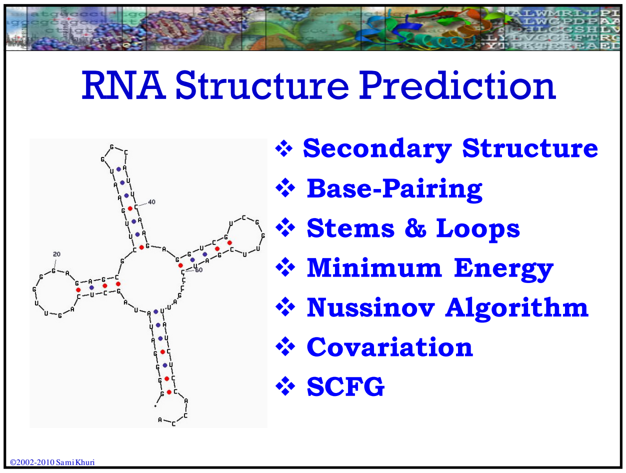


# Bioinformatics

## Six
## RNA Secondary Structure Prediction

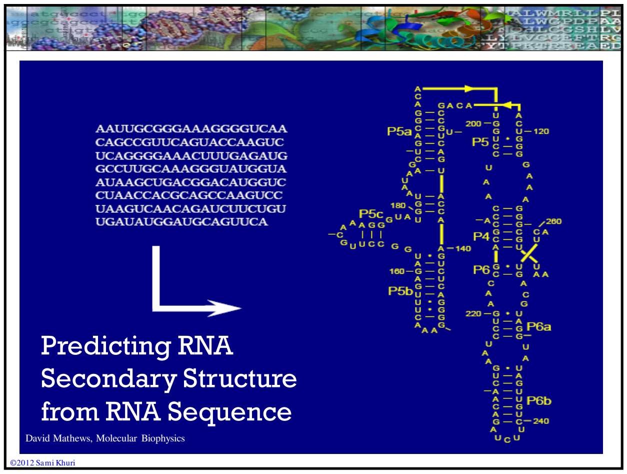
Sami Khuri
Department of Computer Science
San José State University
San José, California, USA
sami.khuri@sjsu.edu
www.cs.sjsu.edu/faculty/khuri



## RNA Structure Prediction

- ❖ Secondary Structure
- ❖ Base-Pairing
- ❖ Stems & Loops
- ❖ Minimum Energy
- ❖ Nussinov Algorithm
- ❖ Covariation
- ❖ SCFG

©2002-2010 Sami Khuri

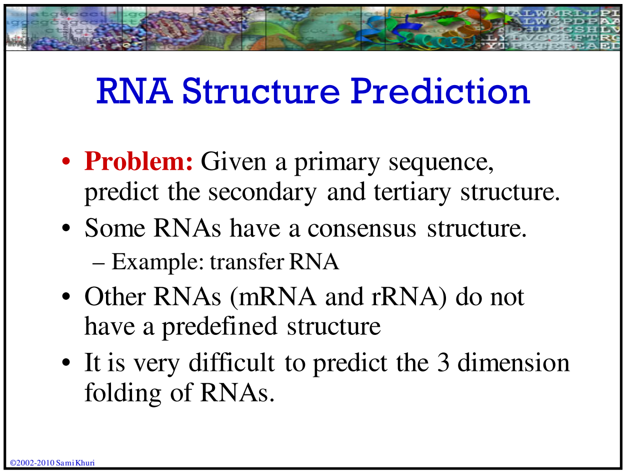


AAUUGCGGGAAAGGGGUCAA
CAGCCGUUCAGUACCAAGUC
UCAGGGGAAACUUUGAGAUG
GCCUUGCAAAGGGUAUGGUA
AUAAGCUGACGGACAUGGUC
CUAACCACGCAGCCAAGUCC
UAAGUCAACAGAUCUUCUGU
UGAUAUGGAUGCAGUUCA

Predicting RNA Secondary Structure from RNA Sequence

David Mathews, Molecular Biophysics

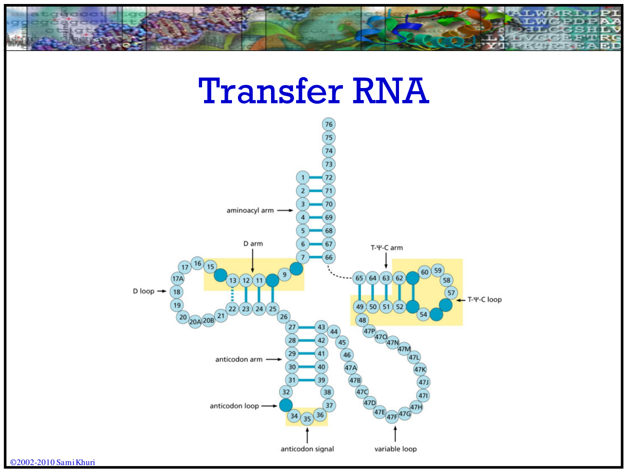
©2012 Sami Khuri



## RNA Structure Prediction

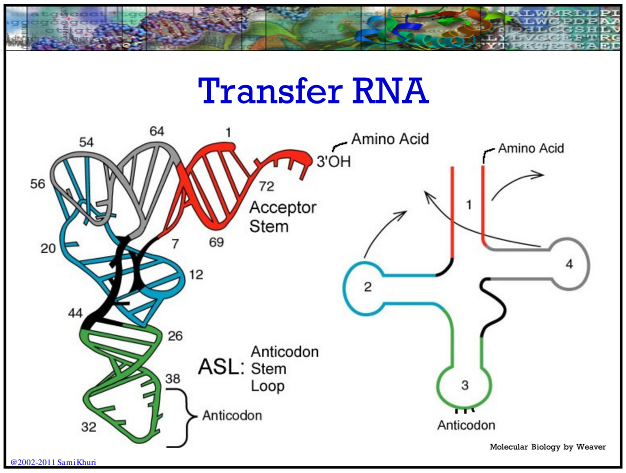
- **Problem:** Given a primary sequence, predict the secondary and tertiary structure.
- Some RNAs have a consensus structure.
  - Example: transfer RNA
- Other RNAs (mRNA and rRNA) do not have a predefined structure
- It is very difficult to predict the 3 dimension folding of RNAs.

©2002-2010 Sami Khuri



## Transfer RNA

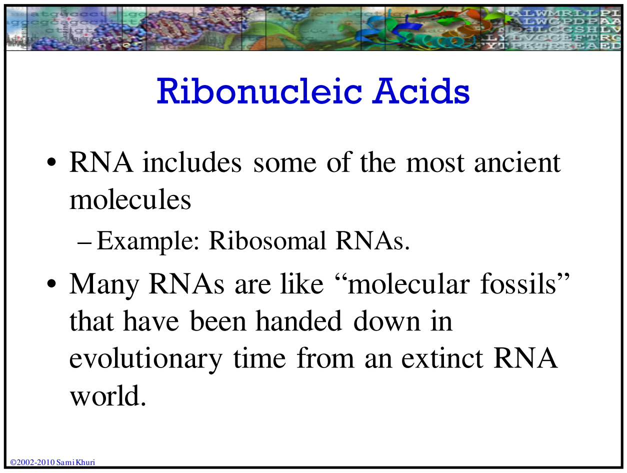
©2002-2010 Sami Khuri



## Transfer RNA

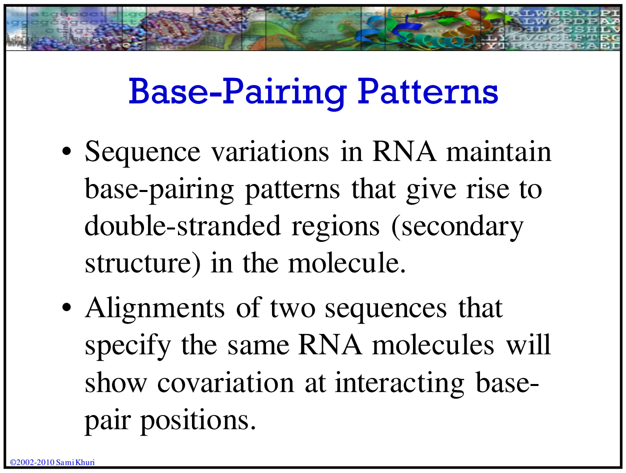
Molecular Biology by Weaver

@2002-2011 Sami Khuri



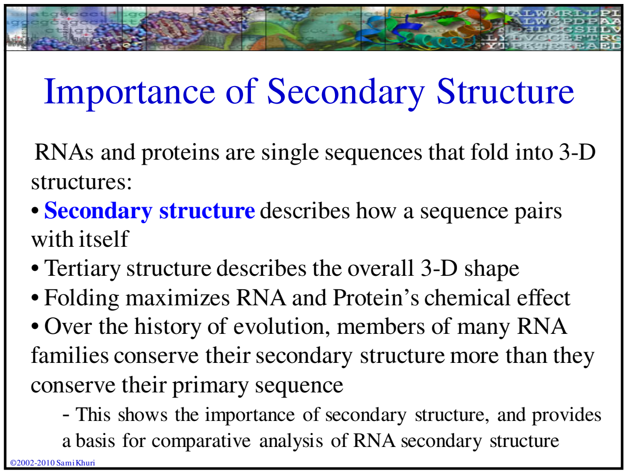
## Ribonucleic Acids

- RNA includes some of the most ancient molecules
  - Example: Ribosomal RNAs.
- Many RNAs are like “molecular fossils” that have been handed down in evolutionary time from an extinct RNA world.



## Base-Pairing Patterns

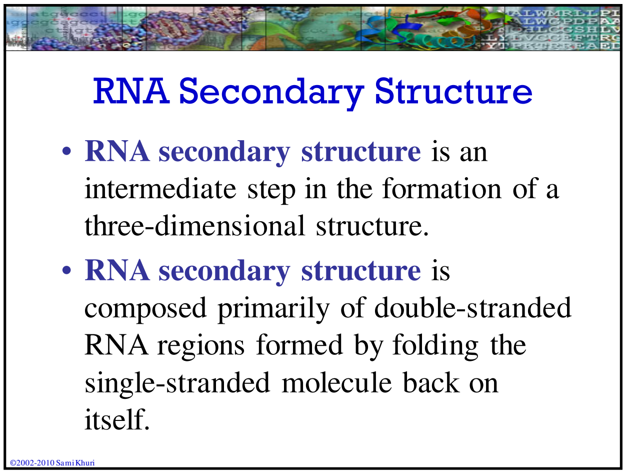
- Sequence variations in RNA maintain base-pairing patterns that give rise to double-stranded regions (secondary structure) in the molecule.
- Alignments of two sequences that specify the same RNA molecules will show covariation at interacting base-pair positions.



## Importance of Secondary Structure

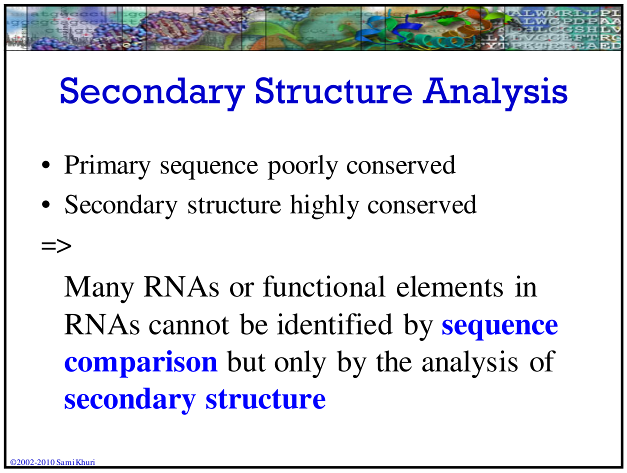
RNAs and proteins are single sequences that fold into 3-D structures:

- **Secondary structure** describes how a sequence pairs with itself
- Tertiary structure describes the overall 3-D shape
- Folding maximizes RNA and Protein’s chemical effect
- Over the history of evolution, members of many RNA families conserve their secondary structure more than they conserve their primary sequence
  - This shows the importance of secondary structure, and provides a basis for comparative analysis of RNA secondary structure



## RNA Secondary Structure

- **RNA secondary structure** is an intermediate step in the formation of a three-dimensional structure.
- **RNA secondary structure** is composed primarily of double-stranded RNA regions formed by folding the single-stranded molecule back on itself.

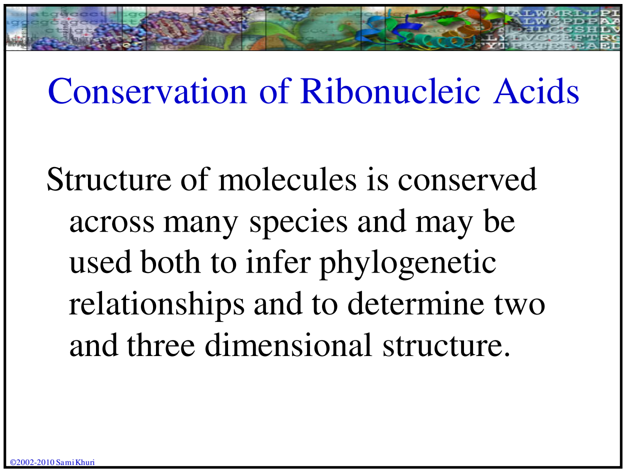


## Secondary Structure Analysis

- Primary sequence poorly conserved
- Secondary structure highly conserved

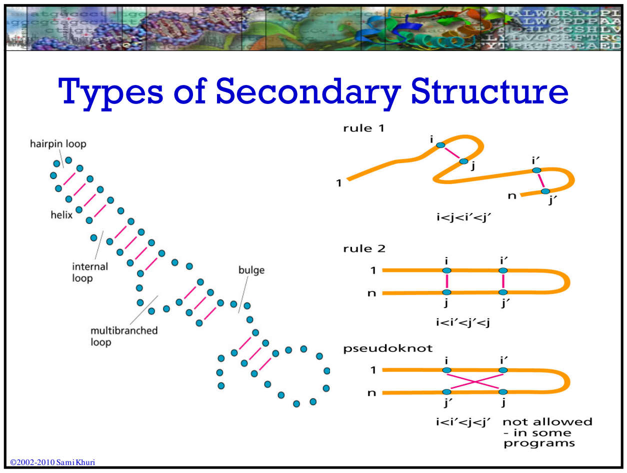
=>

Many RNAs or functional elements in RNAs cannot be identified by **sequence comparison** but only by the analysis of **secondary structure**



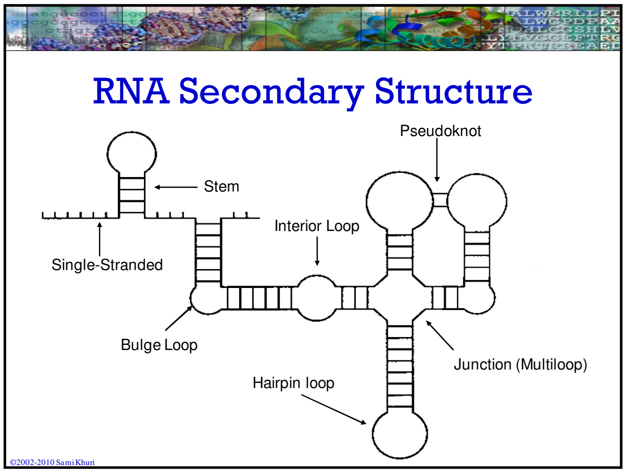
## Conservation of Ribonucleic Acids

Structure of molecules is conserved across many species and may be used both to infer phylogenetic relationships and to determine two and three dimensional structure.



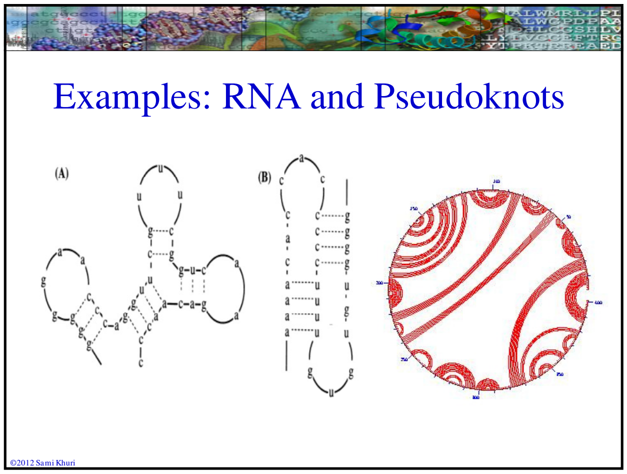
## Types of Secondary Structure

©2002-2010 Sami Khuri



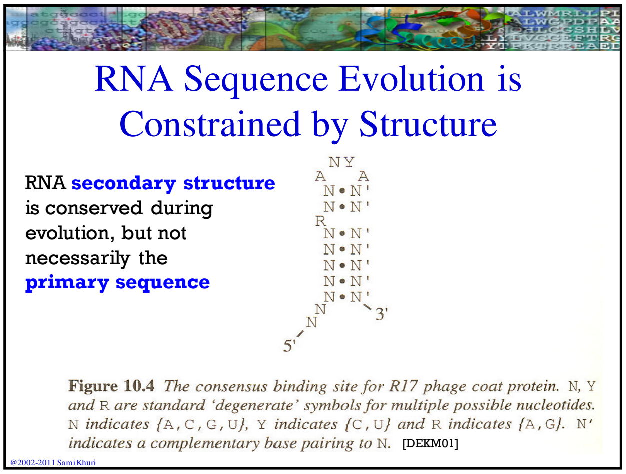
## RNA Secondary Structure

©2002-2010 Sami Khuri



## Examples: RNA and Pseudoknots

©2012 Sami Khuri

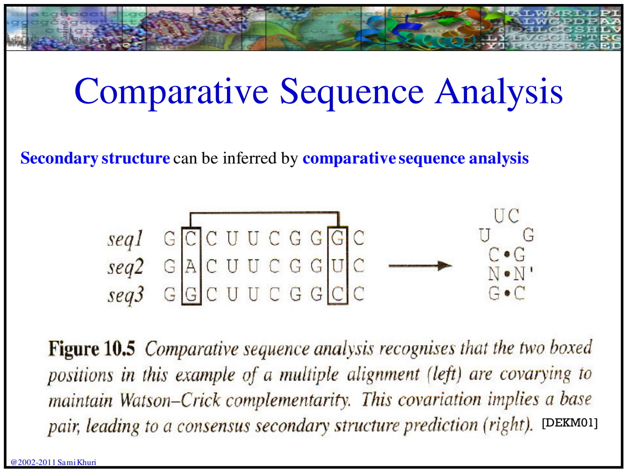


## RNA Sequence Evolution is Constrained by Structure

RNA **secondary structure** is conserved during evolution, but not necessarily the **primary sequence**

**Figure 10.4** *The consensus binding site for R17 phage coat protein.* N, Y *and* R *are standard 'degenerate' symbols for multiple possible nucleotides.* N *indicates {*A, C, G, U*},* Y *indicates {*C, U*} and* R *indicates {*A, G*}.* N' *indicates a complementary base pairing to* N. [DEKM01]

@2002-2011 Sami Khuri

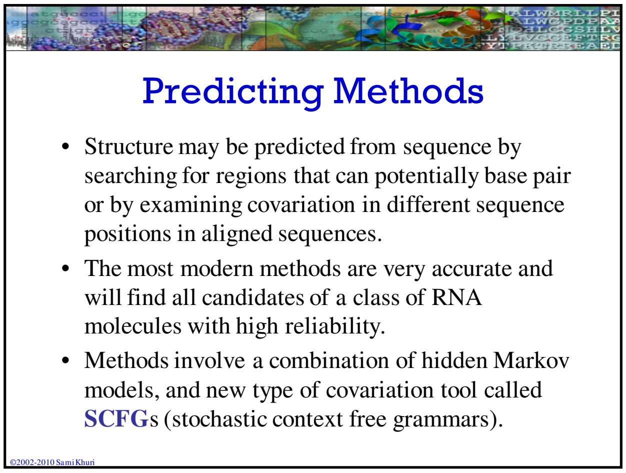


## Comparative Sequence Analysis

**Secondary structure** can be inferred by **comparative sequence analysis**

**Figure 10.5** *Comparative sequence analysis recognises that the two boxed positions in this example of a multiple alignment (left) are covarying to maintain Watson–Crick complementarity. This covariation implies a base pair, leading to a consensus secondary structure prediction (right).* [DEKM01]

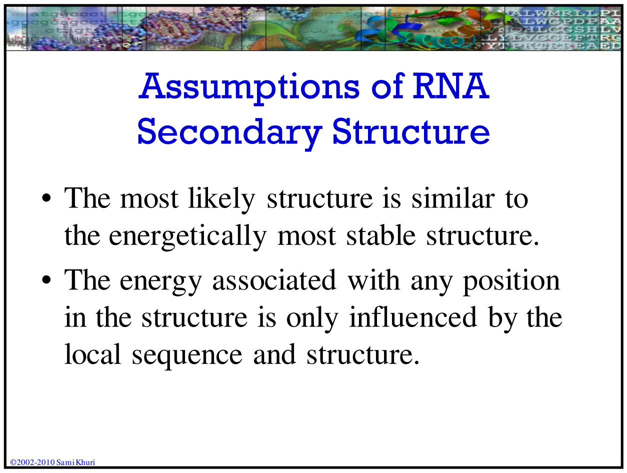
@2002-2011 Sami Khuri



## Predicting Methods

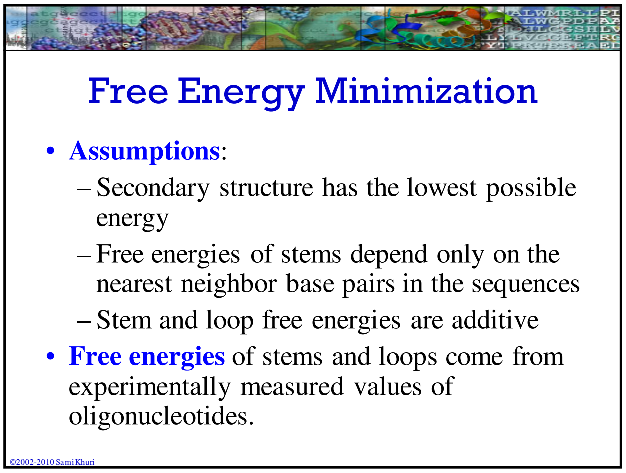
- Structure may be predicted from sequence by searching for regions that can potentially base pair or by examining covariation in different sequence positions in aligned sequences.
- The most modern methods are very accurate and will find all candidates of a class of RNA molecules with high reliability.
- Methods involve a combination of hidden Markov models, and new type of covariation tool called **SCFG**s (stochastic context free grammars).

©2002-2010 Sami Khuri



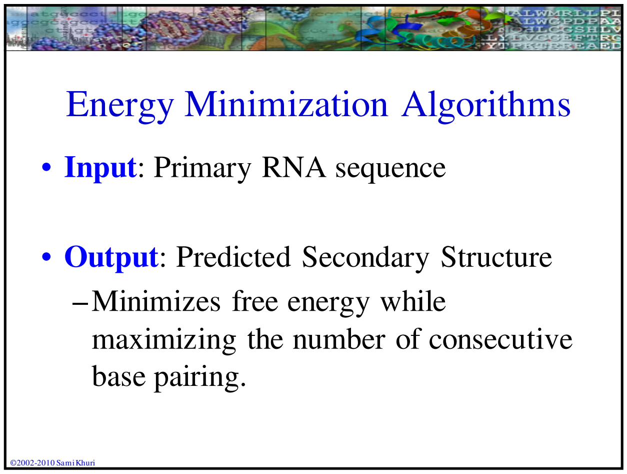
## Assumptions of RNA Secondary Structure

- The most likely structure is similar to the energetically most stable structure.
- The energy associated with any position in the structure is only influenced by the local sequence and structure.



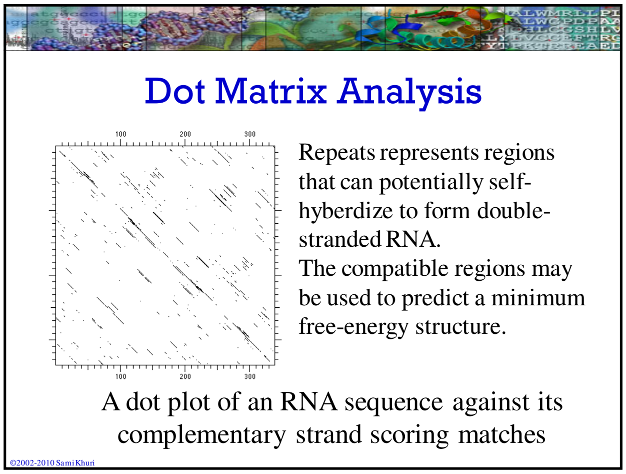
## Free Energy Minimization

- **Assumptions**:
  - Secondary structure has the lowest possible energy
  - Free energies of stems depend only on the nearest neighbor base pairs in the sequences
  - Stem and loop free energies are additive
- **Free energies** of stems and loops come from experimentally measured values of oligonucleotides.



## Energy Minimization Algorithms

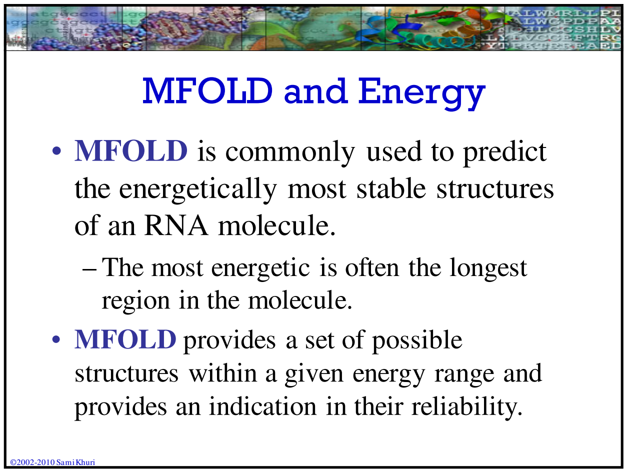
- **Input**: Primary RNA sequence
- **Output**: Predicted Secondary Structure
  - Minimizes free energy while maximizing the number of consecutive base pairing.



## Dot Matrix Analysis

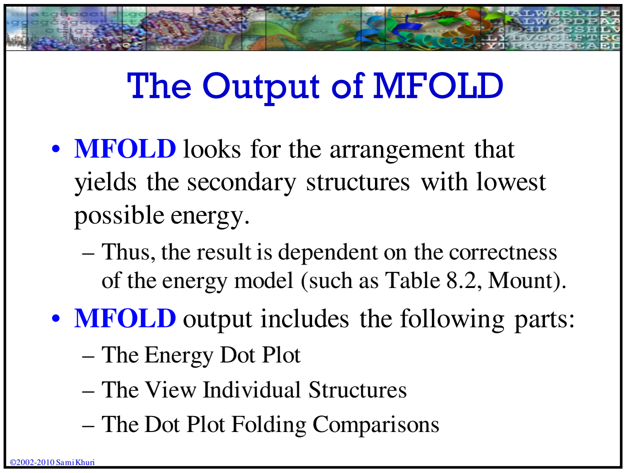
Repeats represents regions that can potentially self-hyberdize to form double-stranded RNA.
The compatible regions may be used to predict a minimum free-energy structure.

A dot plot of an RNA sequence against its complementary strand scoring matches



## MFOLD and Energy

- **MFOLD** is commonly used to predict the energetically most stable structures of an RNA molecule.
  - The most energetic is often the longest region in the molecule.
- **MFOLD** provides a set of possible structures within a given energy range and provides an indication in their reliability.



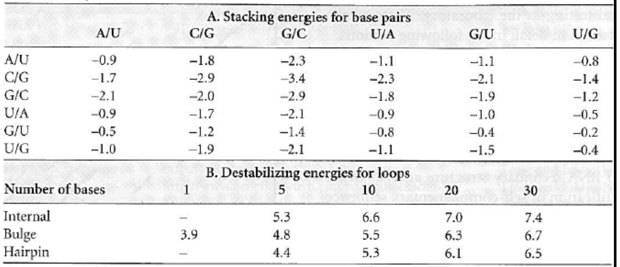
## The Output of MFOLD

- **MFOLD** looks for the arrangement that yields the secondary structures with lowest possible energy.
  - Thus, the result is dependent on the correctness of the energy model (such as Table 8.2, Mount).
- **MFOLD** output includes the following parts:
  - The Energy Dot Plot
  - The View Individual Structures
  - The Dot Plot Folding Comparisons

## Obtaining Minimal Energies

- Plot sequences across the page and also down the left side of the page.
- Look for rows of complementary matches.
- Use the table of predicted free-energy values (kcal/mole at 37 degrees Celsius) for base pairs to add up the stacking energies.

TABLE 8.2. *Predicted free-energy values (kcal/mole at 37°C) for base pairs and other features of predicted RNA secondary structures*



| A. Stacking energies for base pairs | | | | | | |
|---|---|---|---|---|---|---|
| | A/U | C/G | G/C | U/A | G/U | U/G |
| A/U | –0.9 | –1.8 | –2.3 | –1.1 | –1.1 | –0.8 |
| C/G | –1.7 | –2.9 | –3.4 | –2.3 | –2.1 | –1.4 |
| G/C | –2.1 | –2.0 | –2.9 | –1.8 | –1.9 | –1.2 |
| U/A | –0.9 | –1.7 | –2.1 | –0.9 | –1.0 | –0.5 |
| G/U | –0.5 | –1.2 | –1.4 | –0.8 | –0.4 | –0.2 |
| U/G | –1.0 | –1.9 | –2.1 | –1.1 | –1.5 | –0.4 |

| B. Destabilizing energies for loops | | | | | |
|---|---|---|---|---|---|
| Number of bases | 1 | 5 | 10 | 20 | 30 |
| Internal | – | 5.3 | 6.6 | 7.0 | 7.4 |
| Bulge | 3.9 | 4.8 | 5.5 | 6.3 | 6.7 |
| Hairpin | – | 4.4 | 5.3 | 6.1 | 6.5 |

(*A*) Stacking energy in double-stranded region when the base pair listed in left column is followed by the base pair listed in top row. C/G followed by U/A is therefore the dinucleotide 5′ CU 3′ paired to 5′ AG 3′. (*B*) Destabilizing energies associated with loops. Hairpin loops occur at the end of a double-stranded region, internal loops are unpaired regions flanked by paired regions, and a bulge loop is a bulge of one strand in an otherwise paired region (Fig. 8.2). An updated and more detailed list of energy parameters may be found at the Web site of M. Zuker (http://bioinfo.math.rpi.edu/~zuker/rna/energy/). From Turner and Sugimoto (1988); Serra and Turner (1995).

## Dynamic Programming

- Add back energies to accommodate destabilizing structures like bulge loops, hairpins.
- The entire matrix is scanned with a dynamic programming algorithm to find the most energetic structure.
- Note that there are no elements of tertiary structure in this analysis.

## Free Energy Calculating

| 5' | A | C | G | U | 3' |
|---|---|---|---|---|---|
| A | | | | U/A -1.8 + (-3.4) + (-1.8) = - 7.0 | |
| C | | | G/C -1.8 + (-3.4) | | |
| G | | C/G -1.8 | | | |
| U | A/U 0 | | | | |
| 3‘ | | | | | |

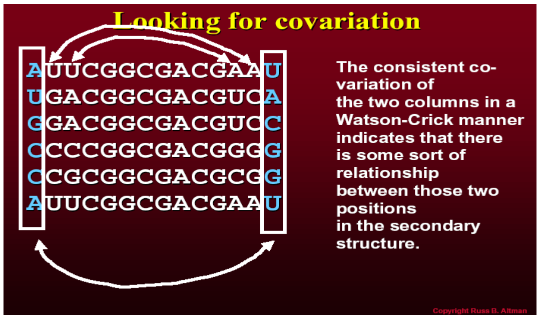
Bioinformatics by David Mount

The diagonal **A/U, C/G, G/C, U/A** is a potential double stranded region with energy -7.0 kcal/mole.

## Covariant Analysis

- **Covariant analysis** uses a set of homologous, aligned sequences to identify evolutionary conserved structures and to identify covarying residues in the sequence
  - Need many sequences
  - Longer sequences can be used
- **Assumption**:
  - Secondary structure is more conserved than primary structure

## Looking for Covariation



Looking for covariation

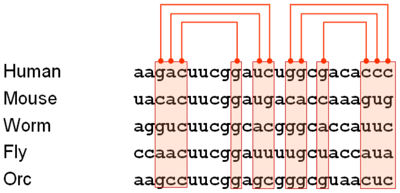
AUUCGGCGACGAAU
UGACGGCGACGUCA
GGACGGCGACGUCC
CCCCGGCGACGGGG
CCGCGGCGACGCGG
AUUCGGCGACGAAU

The consistent co-variation of the two columns in a Watson-Crick manner indicates that there is some sort of relationship between those two positions in the secondary structure.

Copyright Russ B. Altman

## MSA and RNA Folding

Given K homologous aligned RNA sequences:



| | |
|---|---|
| Human | aagacuucggaucuggcgacaccc |
| Mouse | uacacuucggaugacaccaaagug |
| Worm | aggucuucggcacgggcaccauuc |
| Fly | ccaacuucggauuuugcuaccaua |
| Orc | aagccuucggagcgggcguaacuc |

If i^th and j^th positions are always base paired and covary, then they are likely to be paired

## Covariation Analysis of tRNA

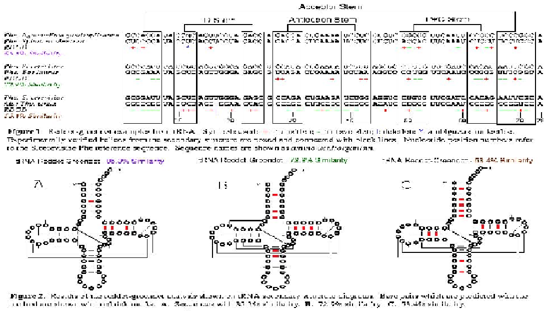


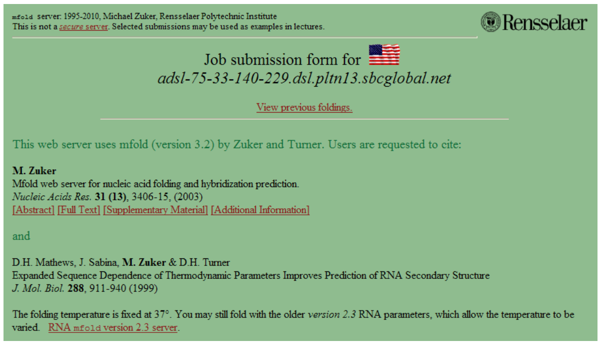
Figure 1.

Figure 2.

## SCFG For Modeling RNA

- Use both, **covariational** and **energy minimization** methods together generally yield very good results.
- **Stochastic Context Free Grammars** (SCFG) can help define base interactions in specific classes of RNA molecules and sequence variations at those positions.

## MFold



mfold server: 1995-2010, Michael Zuker, Rensselaer Polytechnic Institute
This is not a secure server. Selected submissions may be used as examples in lectures.

Rensselaer

Job submission form for
*adsl-75-33-140-229.dsl.pltn13.sbcglobal.net*

View previous foldings.

This web server uses mfold (version 3.2) by Zuker and Turner. Users are requested to cite:

**M. Zuker**
Mfold web server for nucleic acid folding and hybridization prediction.
*Nucleic Acids Res.* **31 (13)**, 3406-15, (2003)
[Abstract] [Full Text] [Supplementary Material] [Additional Information]

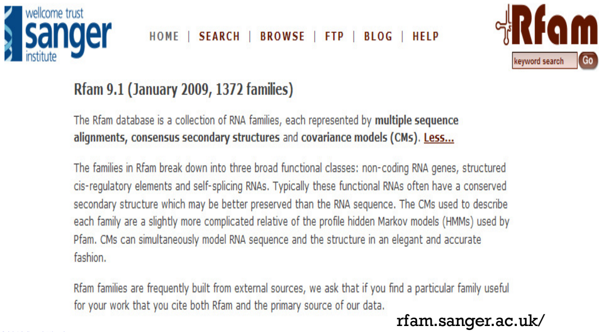
and

D.H. Mathews, J. Sabina, **M. Zuker** & D.H. Turner
Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure
*J. Mol. Biol.* **288**, 911-940 (1999)

The folding temperature is fixed at 37°. You may still fold with the older *version 2.3* RNA parameters, which allow the temperature to be varied. RNA mfold version 2.3 server.

mfold.bioinfo.rpi.edu/cgi-bin/rna-form1.cgi

## Rfam at Sanger Institue


wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | BLOG | HELP
Rfam
keyword search
Go

Rfam 9.1 (January 2009, 1372 families)

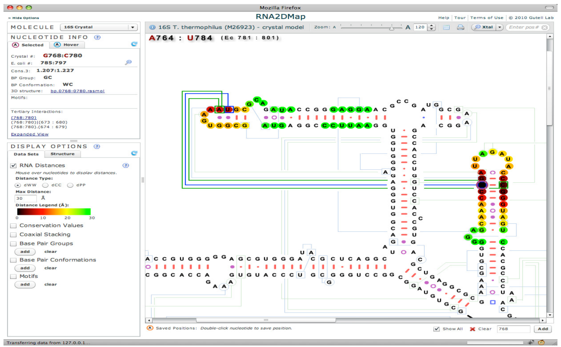
The Rfam database is a collection of RNA families, each represented by **multiple sequence alignments, consensus secondary structures** and **covariance models (CMs)**. Less...

The families in Rfam break down into three broad functional classes: non-coding RNA genes, structured cis-regulatory elements and self-splicing RNAs. Typically these functional RNAs often have a conserved secondary structure which may be better preserved than the RNA sequence. The CMs used to describe each family are a slightly more complicated relative of the profile hidden Markov models (HMMs) used by Pfam. CMs can simultaneously model RNA sequence and the structure in an elegant and accurate fashion.

Rfam families are frequently built from external sources, we ask that if you find a particular family useful for your work that you cite both Rfam and the primary source of our data.

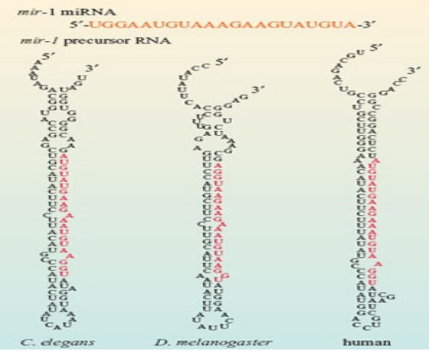
rfam.sanger.ac.uk/

## RNA2DMap



University of Texas at Austin

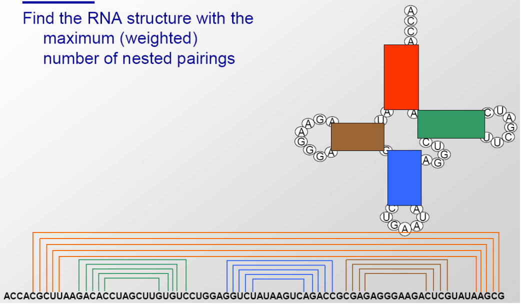
## Mir-1 miRNA Family



## Nussinov's Algorithm

**Problem:**

Find the RNA structure with the maximum (weighted) number of nested pairings


ACCACGCUUAAGACACCUAGCUUGUGUCCUGGAGGUCUAUAAGUCAGACCGCGAGAGGGAAGACUCGUAUAAGCG

Algorithms for Computational Biology by Manolis Kellis (MIT)

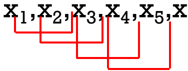
## Dynamic Programming Approach

- Solve problem for all subproblems of size 1:
  - The solution is zero
- Iteratively, knowing the solution of all problems of size less than k, compute the solution of all problems of size k.

## Optimally Solving Subproblems

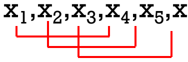
- Input X = $x_1,x_2,x_3,x_4,x_5,x_6,\ldots,x_n$
- Solve subproblems of size 2:

$x_1,x_2,x_3,x_4,x_5,x_6,\ldots,x_n$



- Solve subproblems of size 3:

$x_1,x_2,x_3,x_4,x_5,x_6,\ldots,x_n$



- Continue until all sizes are studied

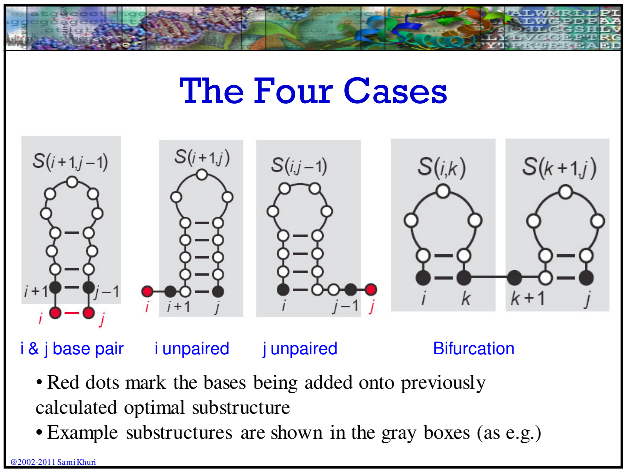
## Nussinov: Base Pair Maximization

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

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

## Recursive Nature of S(i,j)

- To identify the structure with the maximum number of base pairs, the scoring system rewards +1 for a base pair and 0 for anything else.
- The optimal score, S(i,j), of a subsequence of the RNA from position i to position j, can be defined recursively in terms of optimal scores of smaller subsequences.

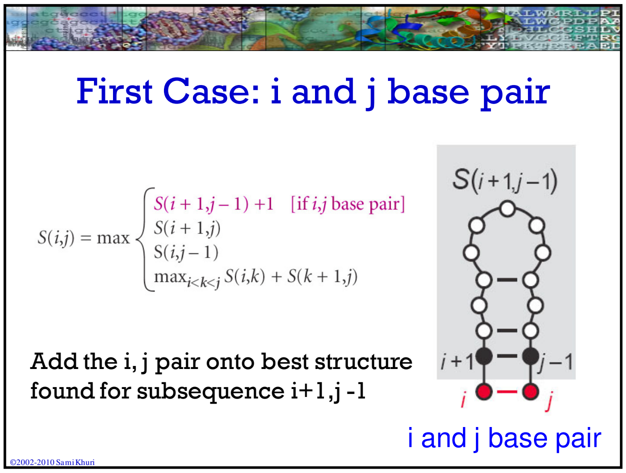


## The Four Cases

i & j base pair　　i unpaired　　j unpaired　　Bifurcation

- Red dots mark the bases being added onto previously calculated optimal substructure
- Example substructures are shown in the gray boxes (as e.g.)

©2002-2011 Sami Khuri



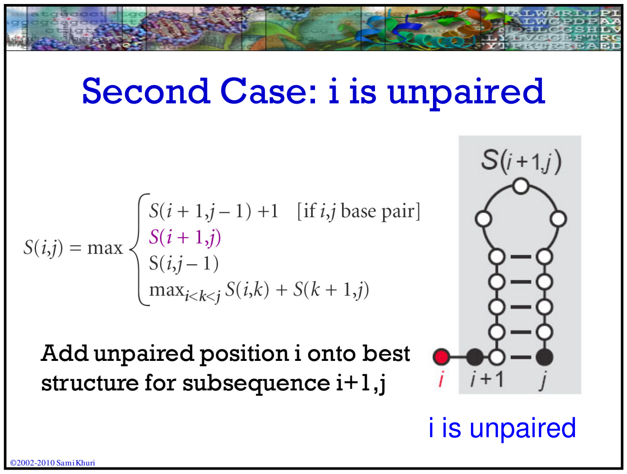
## First Case: i and j base pair

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

Add the i, j pair onto best structure found for subsequence i+1, j -1

i and j base pair

©2002-2010 Sami Khuri



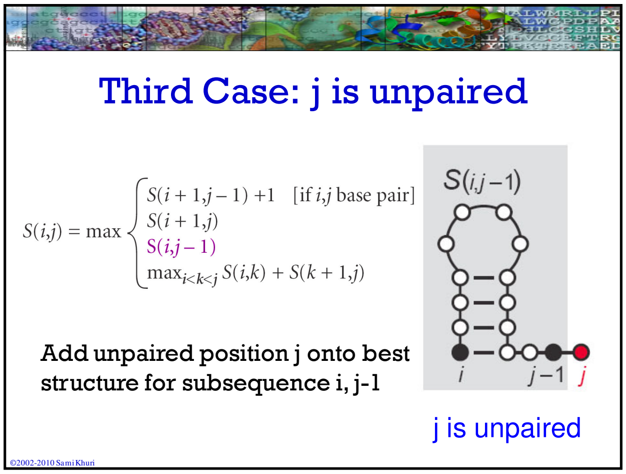
## Second Case: i is unpaired

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

Add unpaired position i onto best structure for subsequence i+1,j

i is unpaired

©2002-2010 Sami Khuri



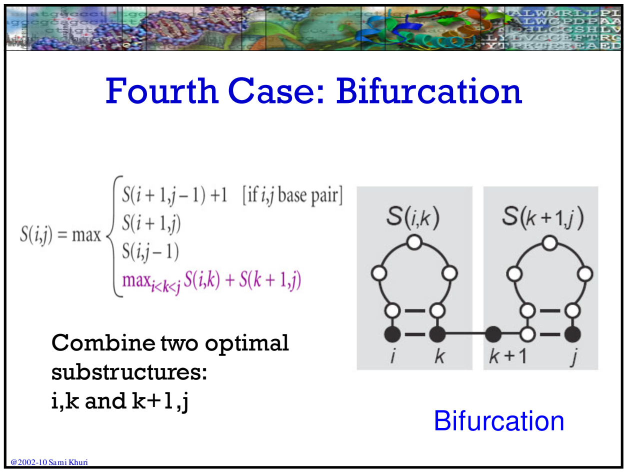
## Third Case: j is unpaired

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

Add unpaired position j onto best structure for subsequence i, j-1

j is unpaired

©2002-2010 Sami Khuri



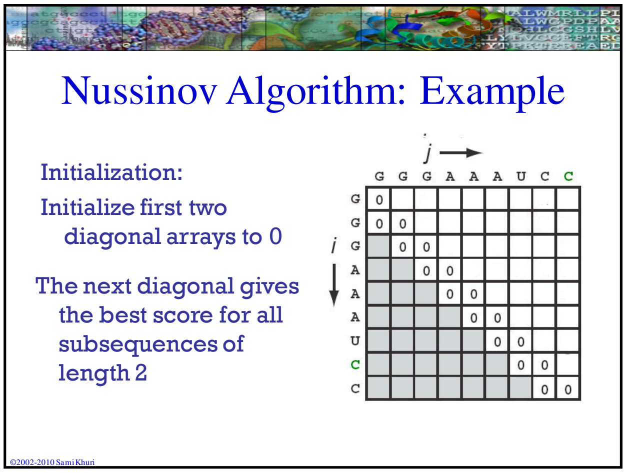
## Fourth Case: Bifurcation

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

Combine two optimal substructures:
i,k and k+1,j

Bifurcation

©2002-10 Sami Khuri



## Nussinov Algorithm: Example

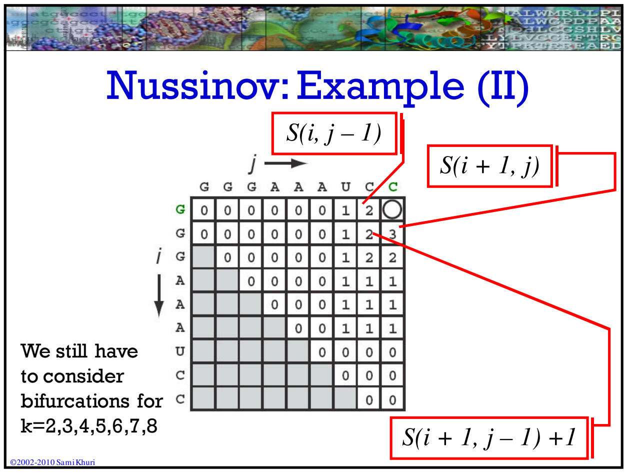
Initialization:

Initialize first two diagonal arrays to 0

The next diagonal gives the best score for all subsequences of length 2

| | G | G | G | A | A | A | U | C | C |
|---|---|---|---|---|---|---|---|---|---|
| G | 0 | | | | | | | | |
| G | 0 | 0 | | | | | | | |
| G | | 0 | 0 | | | | | | |
| A | | | 0 | 0 | | | | | |
| A | | | | 0 | 0 | | | | |
| A | | | | | 0 | 0 | | | |
| U | | | | | | 0 | 0 | | |
| C | | | | | | | 0 | 0 | |
| C | | | | | | | | 0 | 0 |

©2002-2010 Sami Khuri



## Nussinov: Example (II)

*S(i, j – 1)*

*S(i + 1, j)*

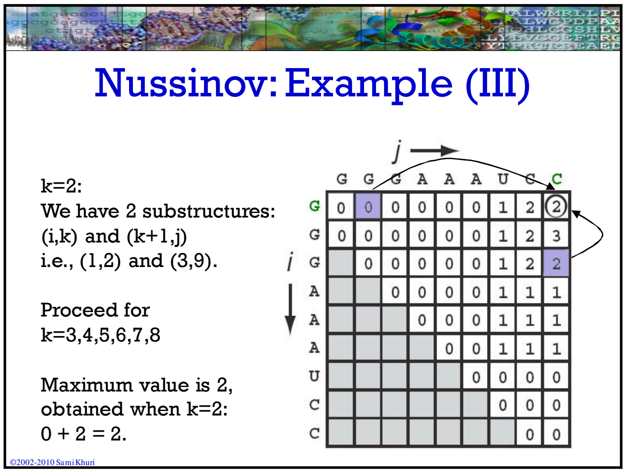
j →

|   | G | G | G | A | A | A | U | C | C |
|---|---|---|---|---|---|---|---|---|---|
| G | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 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 |

i ↓

We still have to consider bifurcations for k=2,3,4,5,6,7,8

*S(i + 1, j – 1) +1*



## Nussinov: Example (III)

k=2:
We have 2 substructures:
(i,k) and (k+1,j)
i.e., (1,2) and (3,9).

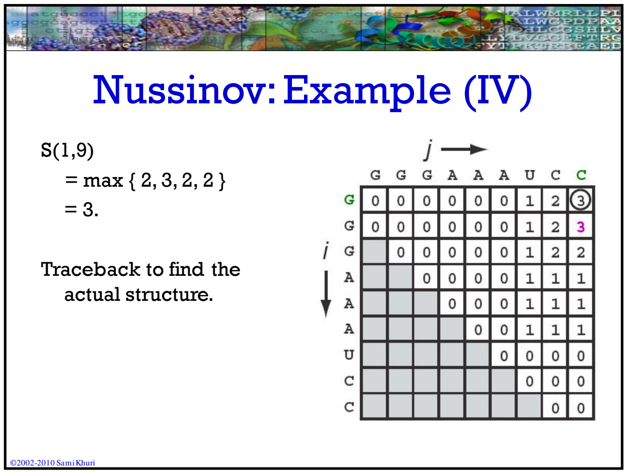
Proceed for k=3,4,5,6,7,8

Maximum value is 2, obtained when k=2:
0 + 2 = 2.

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 |

i ↓



## Nussinov: Example (IV)

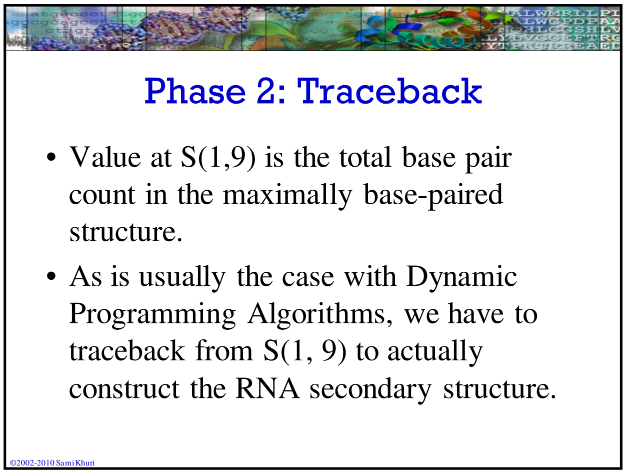
S(1,9)
= max { 2, 3, 2, 2 }
= 3.

Traceback to find the actual structure.

j →

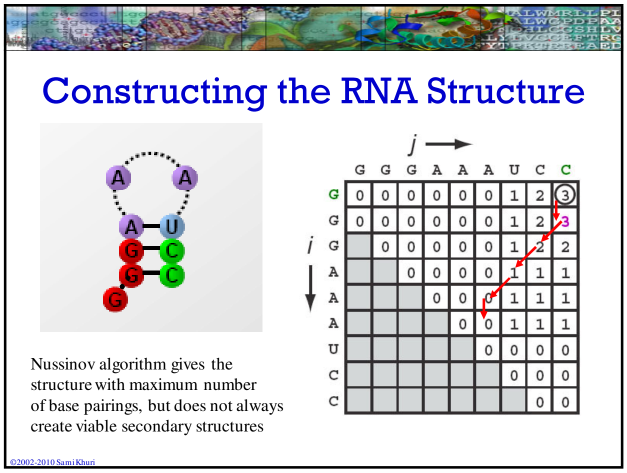
|   | G | G | G | A | A | A | U | C | C |
|---|---|---|---|---|---|---|---|---|---|
| G | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| G |   | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 2 |
| A |   |   | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| A |   |   |   | 0 | 0 | 0 | 1 | 1 | 1 |
| A |   |   |   |   | 0 | 0 | 1 | 1 | 1 |
| U |   |   |   |   |   | 0 | 0 | 0 | 0 |
| C |   |   |   |   |   |   | 0 | 0 | 0 |
| C |   |   |   |   |   |   |   | 0 | 0 |

i ↓



## Phase 2: Traceback

- Value at S(1,9) is the total base pair count in the maximally base-paired structure.
- As is usually the case with Dynamic Programming Algorithms, we have to traceback from S(1, 9) to actually construct the RNA secondary structure.



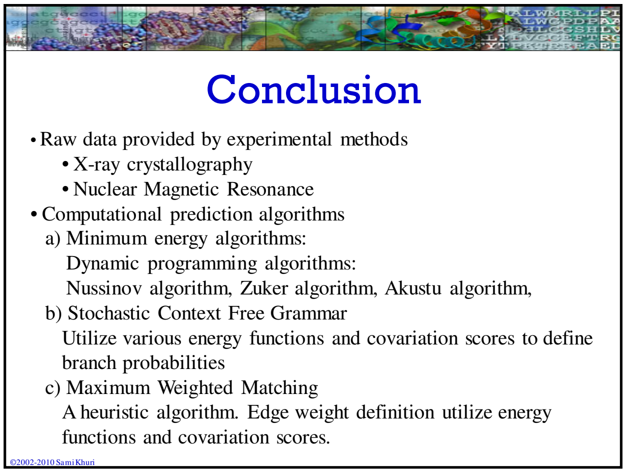
## Constructing the RNA Structure

Nussinov algorithm gives the structure with maximum number of base pairings, but does not always create viable secondary structures

j →

|   | G | G | G | A | A | A | U | C | C |
|---|---|---|---|---|---|---|---|---|---|
| G | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| 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 |

i ↓



## Conclusion

- Raw data provided by experimental methods
  - X-ray crystallography
  - Nuclear Magnetic Resonance
- Computational prediction algorithms
  - a) Minimum energy algorithms:
    Dynamic programming algorithms:
    Nussinov algorithm, Zuker algorithm, Akustu algorithm,
  - b) Stochastic Context Free Grammar
    Utilize various energy functions and covariation scores to define branch probabilities
  - c) Maximum Weighted Matching
    A heuristic algorithm. Edge weight definition utilize energy functions and covariation scores.