http://creativecommons.org/licenses/by-sa/2.0/

creative commons

COMMONS DEED

Attribution-ShareAlike 2.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works
- to make commercial use of the work

Under the following conditions:

Attribution. You must give the original author credit.

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

RNA Structure Prediction

Prof:Rui Alves
ralves@cmb.udl.es
973702406
Dept Ciencies Mediques Basiques,
1st Floor, Room 1.08
Website of the Course: http://web.udl.es/usuaris/pg193845/Courses/Bioinformatics_2007/
Course: http://10.100.14.36/Student_Server/

RNA functions

Storage/transfer of genetic information

- **Genomes**
  - many viruses have RNA genomes
    - single-stranded (ssRNA)
      - e.g., retroviruses (HIV)
    - double-stranded (dsRNA)

- **Transfer of genetic information**
  - mRNA = "coding RNA" - encodes proteins

RNA functions

Structural

- e.g., rRNA, which is a major structural component of ribosomes

  BUT - its role is *not* just structural, also:

Catalytic

RNA in the ribosome has *peptidyltransferase* activity

- Enzymatic activity responsible for peptide bond formation between amino acids in growing peptide chain
- Also, many small RNAs are enzymes

"ribozymes"

RNA functions

Regulatory

Recently discovered important new roles for RNAs

In normal cells:

- in "defense" - esp. in plants
- in normal development

e.g., siRNAs, miRNA

As tools:

- for gene therapy or to modify gene expression
  - RNAi
  - RNA aptamers

RNA types & functions

Types of RNAs	Primary Function(s)
mRNA - messenger	translation (protein synthesis) regulatory
rRNA - ribosomal	translation (protein synthesis) <catalytic>
t-RNA - transfer	translation (protein synthesis)
hnRNA - heterogeneous nuclear	precursors & intermediates of mature mRNAs & other RNAs
scRNA - small cytoplasmic	signal recognition particle (SRP) tRNA processing <catalytic>
snRNA - small nuclear snoRNA - small nucleolar	mRNA processing, poly A addition <catalytic> rRNA processing/maturation/methylation
regulatory RNAs (siRNA, miRNA, etc.)	regulation of transcription and translation, other??

miRNA Challenges for Computational Biology

- Find the genes encoding microRNAs
- Predict their regulatory targets

Computational Prediction of MicroRNA Genes & Targets

- Integrate miRNAs into gene regulatory pathways & networks

Need to modify traditional paradigm of "transcriptional control" primarily by protein-DNA interactions to include miRNA regulatory mechanisms!

- Predict RNA structure

Outline

- RNA primary structure
- Small RNA prediction
- RNA secondary structure & prediction
- RNA tertiary structure & prediction

Hierarchical organization of RNA molecules

Primary structure:

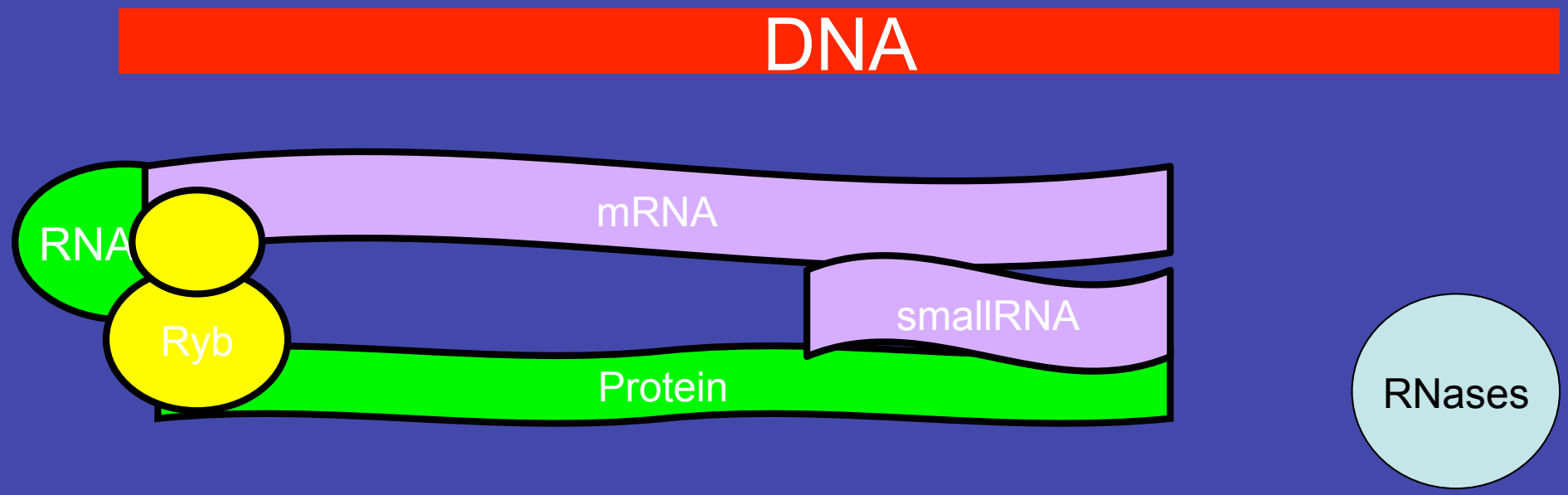
- 5' to 3' list of covalently linked nucleotides, named by the attached base

- Commonly represented by a string S over the alphabet $\Sigma=\{A,C,G,U\}$

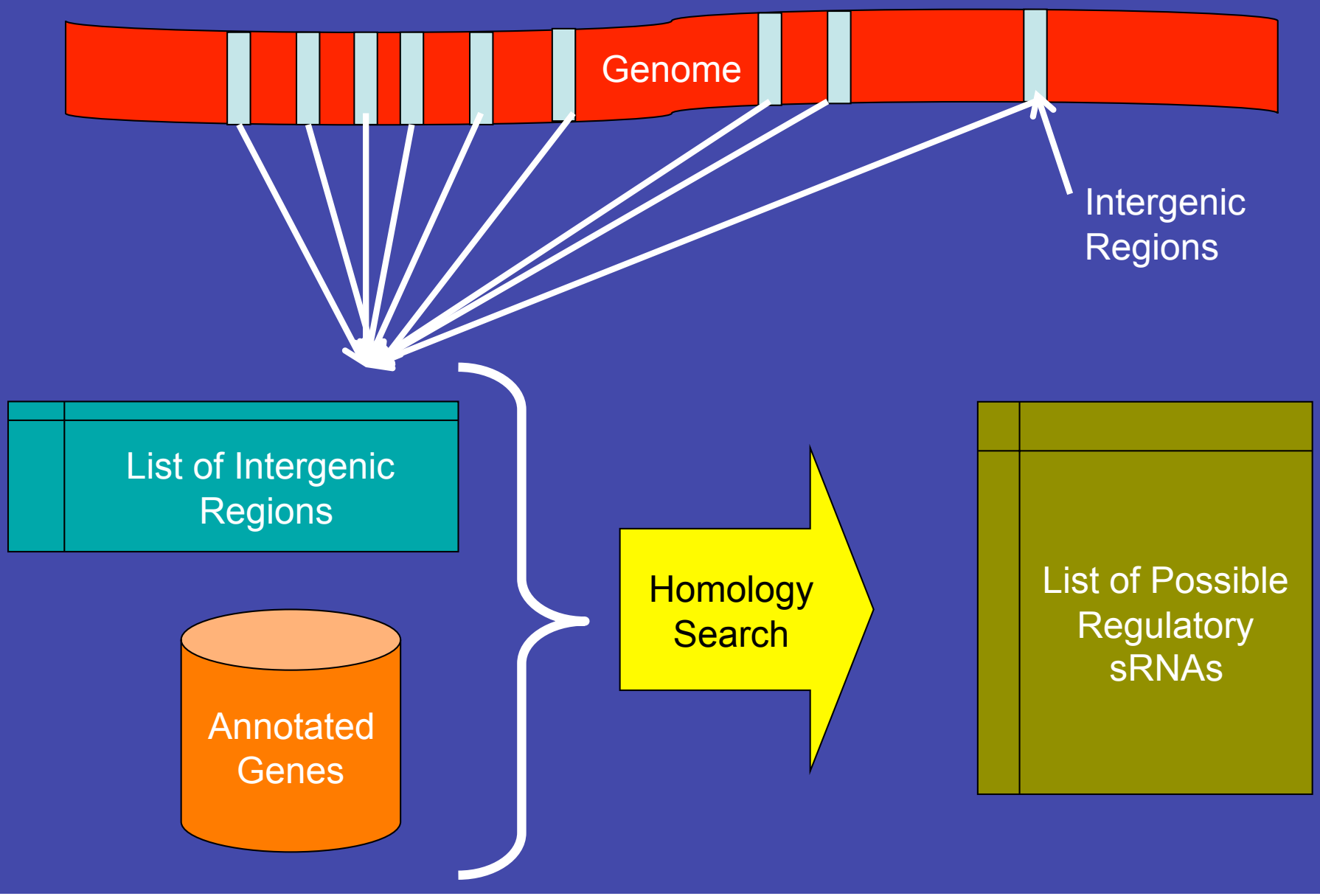
Outline

- RNA primary structure
- Small RNA prediction
- RNA secondary structure & prediction
- RNA tertiary structure & prediction

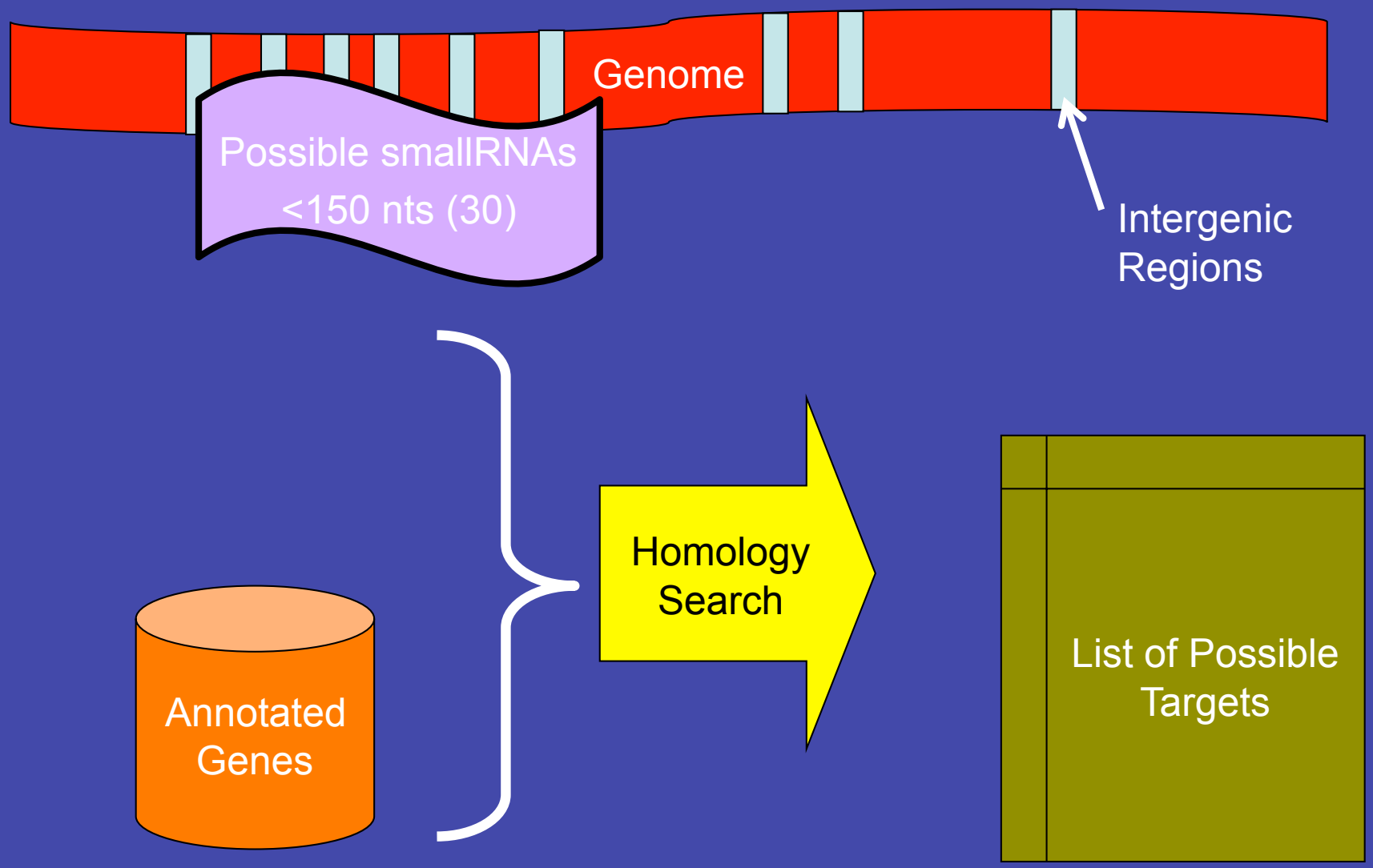
mRNAi



How to predict possible small RNAs



How to predict possible targets for small RNAs



Outline

- RNA primary structure
- Small RNA prediction
- RNA secondary structure & prediction
- RNA tertiary structure & prediction

Hierarchical organization of RNA molecules

Primary structure:

5' to 3' list of covalently linked nucleotides, named by the attached base
Commonly represented by a string S over the alphabet $\Sigma=\{A,C,G,U\}$

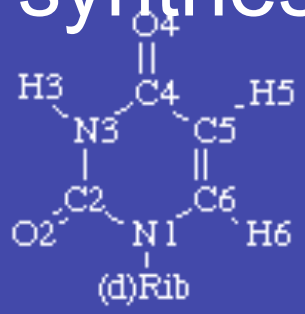
Secondary Structure

List of base pairs, denoted by i•j for a pairing between the i-th and j-th Nucleotides, r_i and r_j, where $i<j$ by convention.

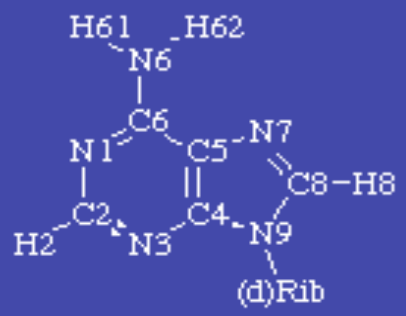
Helices are inferred when two or more base pairs occur adjacent to one another

RNA synthesis and fold

- RNA immediately starts to fold when it is synthesized

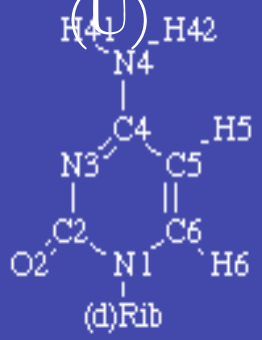


Uracyl
(U)

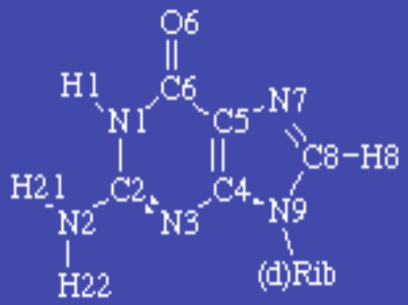


Adenine
(A)

Wobble Base Pairing



Cytosine
(C)

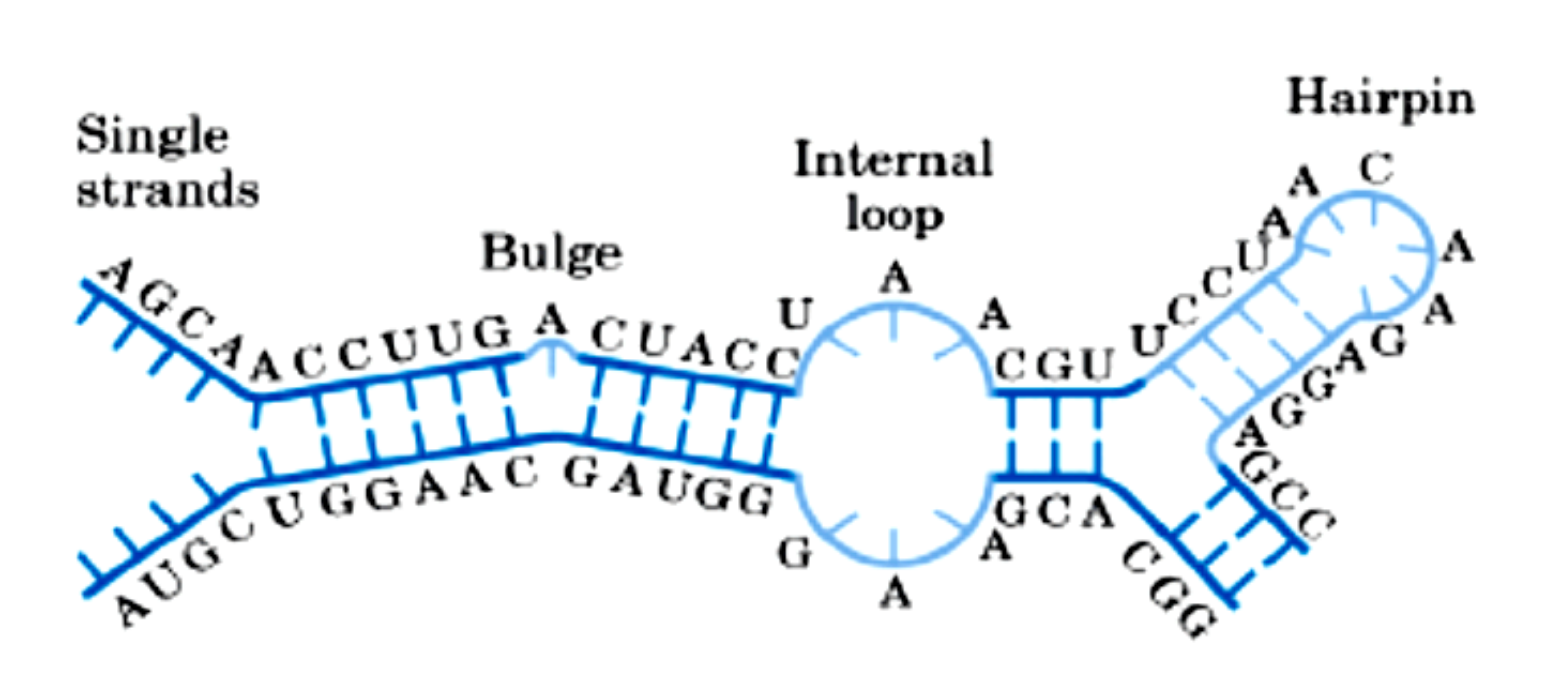


Guanine
(G)

RNA secondary structures

Single stranded bases within a stem are called a bulge of bulge loop if the single stranded bases are on only one side of the stem.

If single stranded bases interrupt both sides of a stem, they are called an internal (interior) loop.

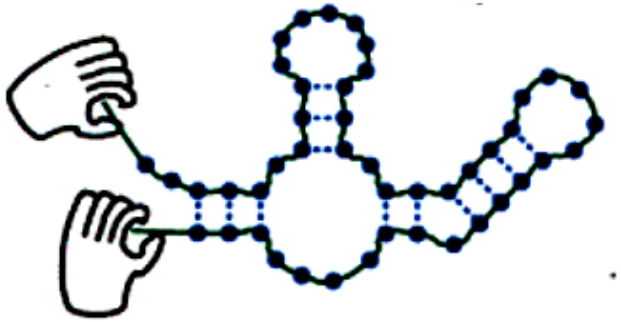


RNA secondary structure representation

- Grammatically correct string of parentheses

```
..(((.(((.....))).((((((....)))).)).....)))
AGCUACGGAGCGAUCUCCGAGCUUUCGAGAAAGCCUCUAUUAGC
```

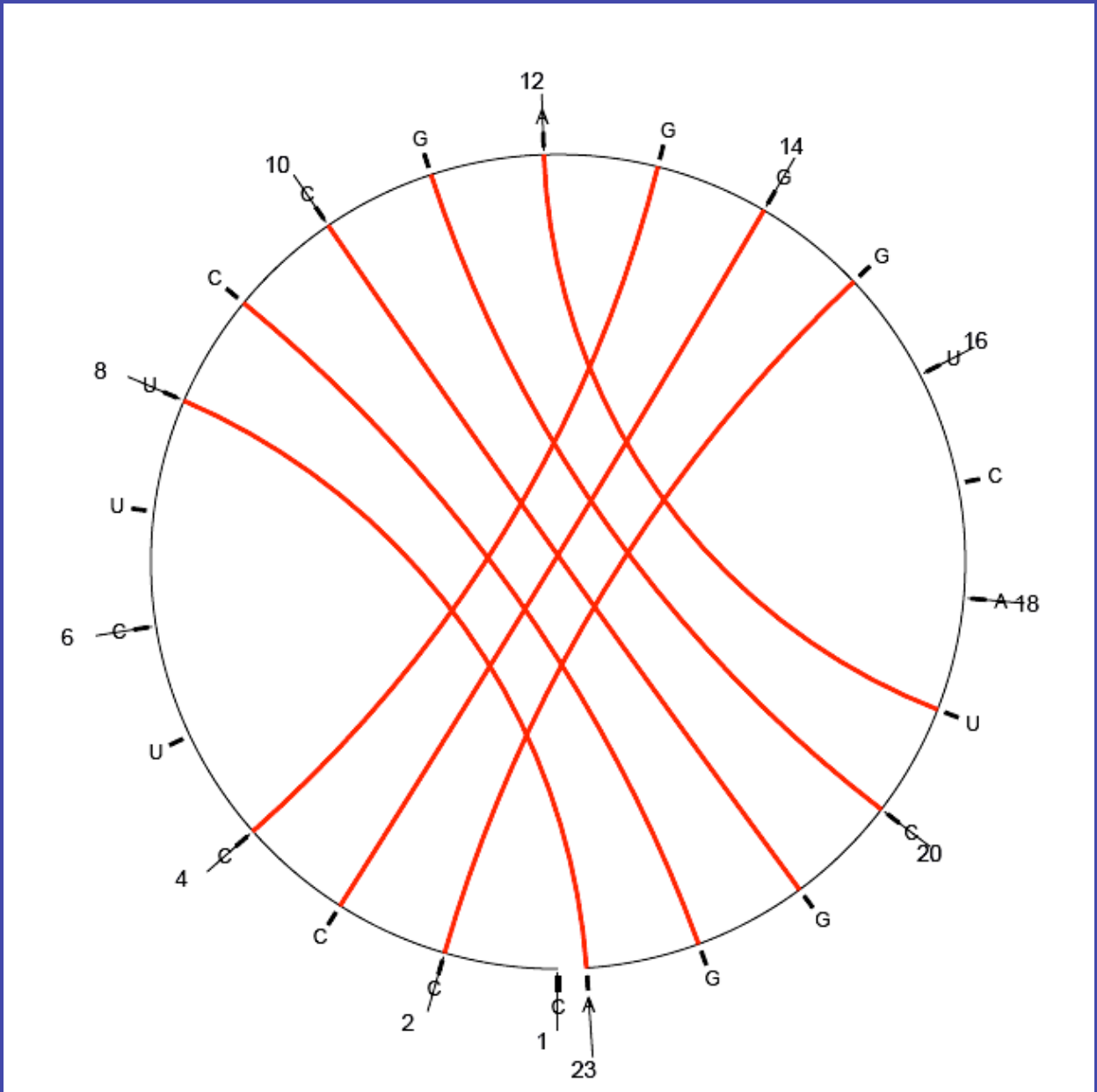
- Planar graph



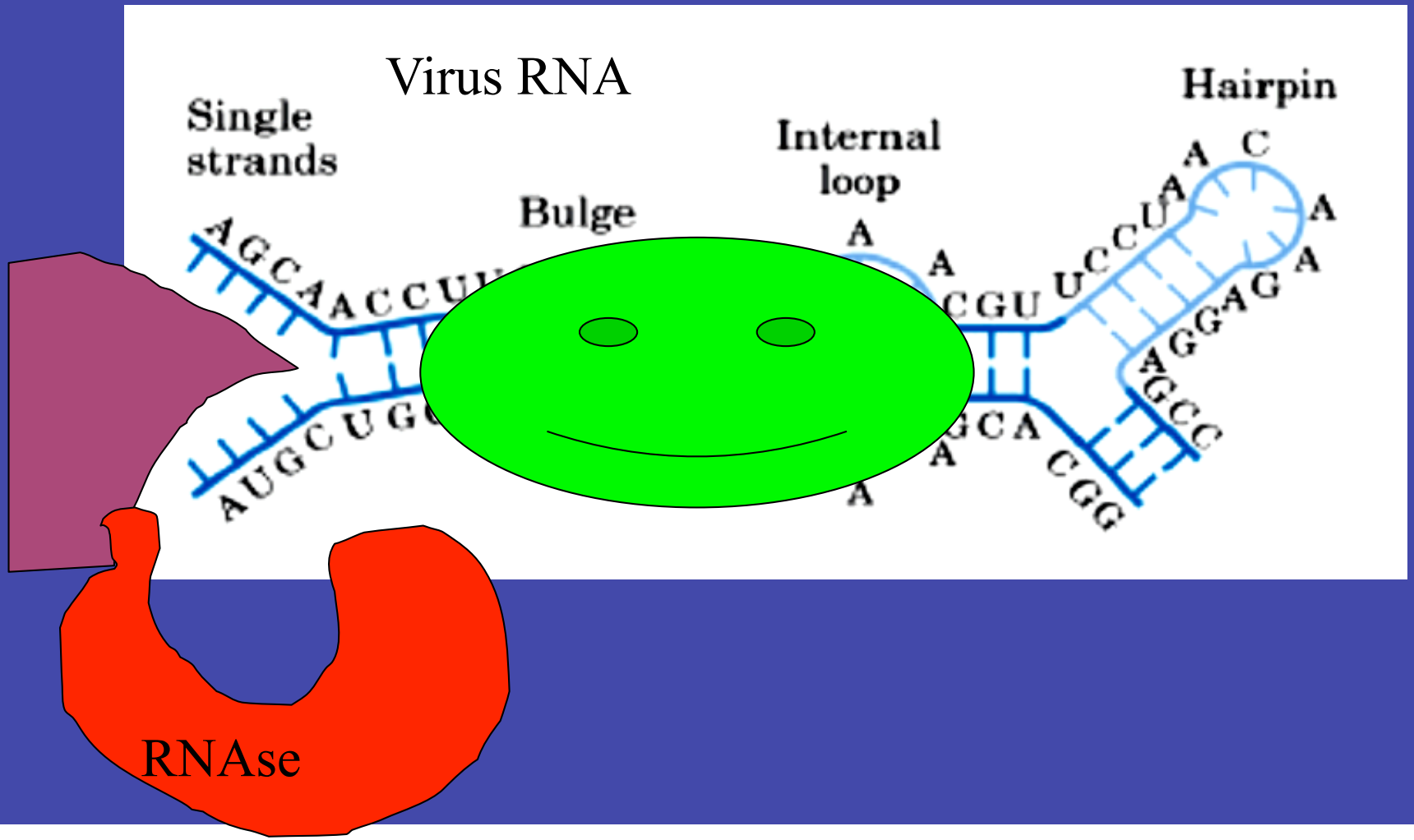
- Arch diagram

- Mountain diagram

Circular representation of RNA



Why predicting RNA secondary structures ?



Existing computational methods for RNA structure prediction

- Comparative methods using sequence homology (ab initio)
  - By examining a set of homologous sequence along with their covarying position, we can predict interactions between non adjacent positions in the sequence, such as base pairs, triples, etc.

Existing computational methods for RNA structure prediction

- Minimum energy predictive methods (ab initio)
  - Try to compute the RNA structure solely based on its nucleotide contents by minimizing the free energy of the predicted structure.

Existing computational methods for RNA structure prediction

- Structural Inference Methods (“homology modelling”)
  - –Given a sequence with a known structure, we infer the structure of another sequence known to be similar to the first one by maximizing some similarity function

RNA structure prediction

Two primary methods for ab initio RNA secondary structure prediction:

-**Co-variation analysis** (comparative sequence analysis)

. Takes into account conserved patterns of basepairs during evolution (more than 2 sequences)

-**Minimum free-energy method**

. Determine structure of complementary regions that are energetically stable

Quantitative Measure of Co-variation

- Maximize some function of covariation of nucleotides in a multiple alignment of RNAs
- Why?
- If two nucleotides change together from AU to GC they are likely to be a pair and the pair should be important for the RNA function

Co-variation

```
Escherichia coli        CACACUGGAA(CUGAGACACG)GUCCAGACUCC
Hildenbrandia rubra     GAGAGGGAGC(CUGAGAAACG)GCUACCACAUC
Banqia fuscopurpurea    GAGAGGGAGC(CUGAGAAAUG)GCUACCACAUC
Rhodochaete parvula     GAGAGGGAGC(CUGAGAAACG)GCUACCACAUC
Cordyceps kanzashiana   GAGAAGGAGC(CUGAGAGACG)GCUACUACAUC
Stichococcus bacillaris GAGAGGGAGC(CUGAGAAACG)GCUACCACAUC
Graphiola phoenicis     GAGAGGGAGC(CUGAGAAACG)GCUACCACAUC
```

	G	C	U	A
i	5/7	1/7	0	1/7
j	1/7	5/7	1/7	0

	G	C	U	A
G	0	0.6	-0.4	0
C	0.6	0	0	0
U	-0.4	0	0	0.4
A	0	0	0.4	0

Computing RNA secondary structure: Minimum free-energy method

- ***Working hypothesis:***

  The native secondary structure of a RNA molecule is the one with the minimum free energy

- ***Restrictions:***
  - ***No knots***
  - ***No close base pairs***
  - ***Base pairs: A-U, C-G and G-U***

Computing RNA secondary structure: Minimum free-energy method

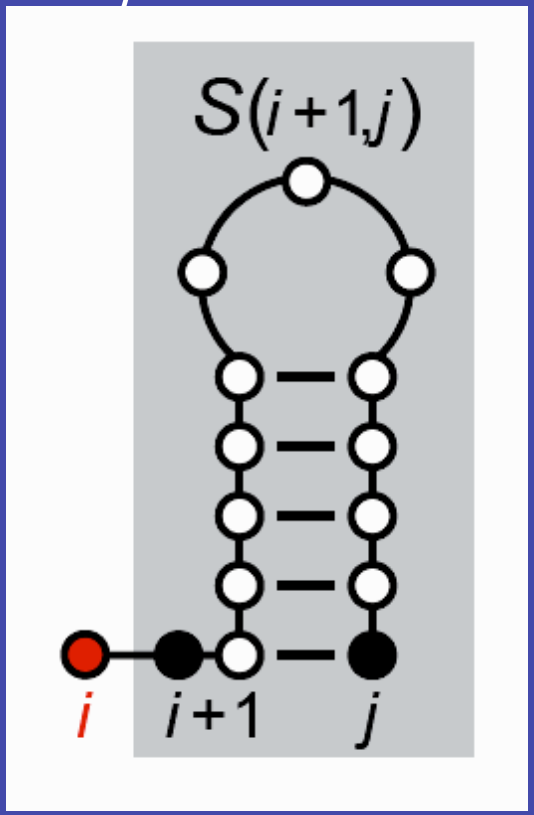
- *Tinoco-Uhlenbeck postulate:*
  - *Assumption: the free energy of each base pair is independent of all the other pairs and the loop structures*
  - *Consequence: the total free energy of an RNA is the sum of all of the base pair free energies*

Independent Base Pairs Approach

- Use solution for smaller strings to find solutions for larger strings
- This is precisely the basic principle behind dynamic programming algorithms!

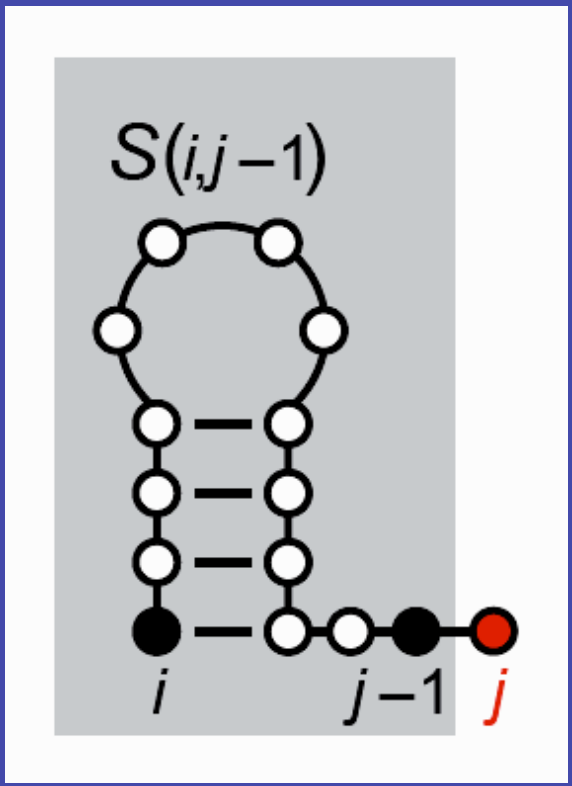
RNA folding: Dynamic Programming

There are only four possible ways that a secondary structure of nested base pair can be constructed on a RNA strand from position i to j:



1. i is unpaired, added on to a structure for i+1…j

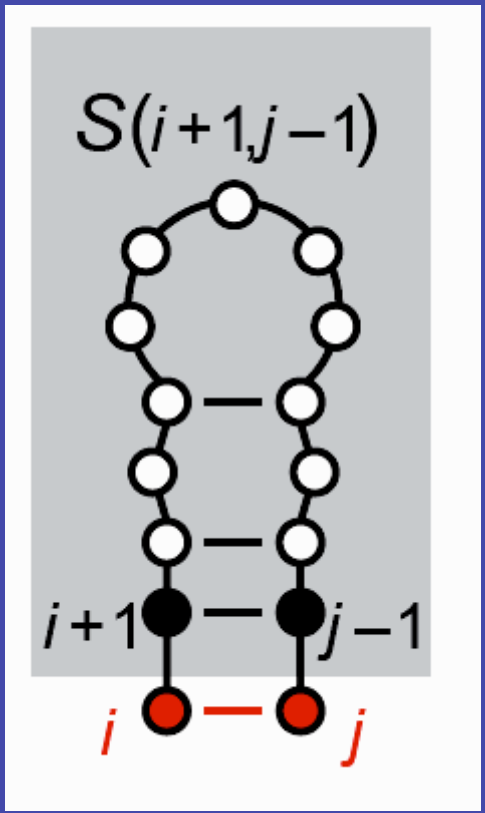
   S(i,j) = S(i+1,j)

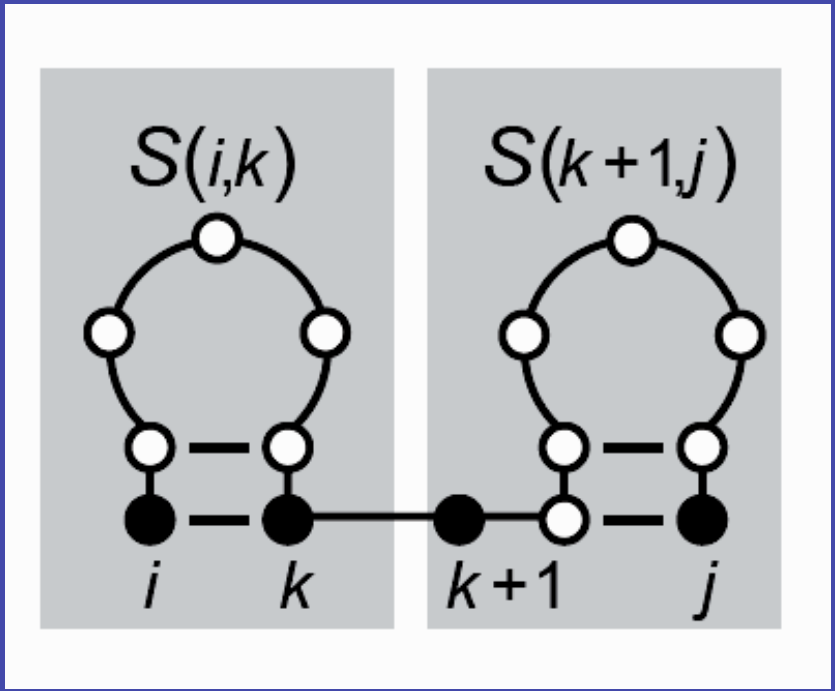


2. j is unpaired, added on to a structure for i…j-1

   S(i,j) = S(i,j-1)

RNA folding: Dynamic Programming





3. i j paired, added on to a structure for i+1…j-1
   $S(i,j) = S(i+1,j-1)+e(r_i,r_j)$

4. i j paired, but not to each other; the structure for i…j adds together structures for 2 sub regions, i…k and k+1…j
   $S(i,j) = \max_{i<k<j} \{S(i,k)+S(k+1,j)\}$

RNA folding: Dynamic Programming

Because there are only four cases, the optimal score S(i,j) is just the maximum of the four possibilities:

$$S(i,j) = \max \begin{cases} S(i+1,j) & r_i \ unpaired \\ S(i,j-1) & r_j \ unpaired \\ S(i+1,j-1) + e(r_i, r_j) & i,j \ base \ pair \\ \max\limits_{i<k<j} \{S(i,k) + S(k+1,j)\} & i,j \ paired, but \ not \ to \ each \ other \end{cases}$$

To compute this efficiently, we need to make sure that the scores for the smaller sub-regions have already been calculated

RNA folding: Dynamic Programming

Notes:

$S(i,j) = 0$ if $j-i < 4$: do not allow "close" base pairs

Reasonable values of e are -3, -2, and -1 kcal/mole for GC, AU and GU, respectively. In the DP procedure, we use 3, 2, 1

Build upper triangular part of DP matrix:

- *start with diagonal – all 0*
- *works outward on larger and larger regions*
- *ends with $S(1,n)$*

Traceback starts with $S(1,n)$, and finds optimal path that leads there.

Initialisation:

No close basepairs

j →

i ↓

		1 A	2 U	3 A	4 C	5 C	6 C	7 U	8 G	9 U	10 G	11 G	12 U	13 A	14 U
1	A	0	0	0	0										
2	U		0	0	0	0									
3	A			0	0	0	0								
4	C				0	0	0	0							
5	C					0	0	0	0						
6	C						0	0	0	0					
7	U							0	0	0	0				
8	G								0	0	0	0			
9	U									0	0	0	0		
10	G										0	0	0	0	
11	G											0	0	0	0
12	U												0	0	0
13	A													0	0
14	U														0

GC 3

AU 2

GU 1

Initialisation:

No close basepairs

j →

i ↓

		1 A	2 U	3 A	4 C	5 C	6 C	7 U	8 G	9 U	10 G	11 G	12 U	13 A	14 U
1	A	0	0	0	0										
2	U		0	0	0	0									
3	A			0	0	0	0								
4	C				0	0	0	0							
5	C					0	0	0	0						
6	C						0	0	0	0					
7	U							0	0	0	0				
8	G								0	0	0	0			
9	U									0	0	0	0		
10	G										0	0	0	0	
11	G											0	0	0	0
12	U												0	0	0
13	A													0	0
14	U														0

GC 3

AU 2

GU 1

Propagation:

C5….G11 :

C5 unpaired:
S(6,11) = 3

G11 unpaired:
S(5,10)=3

C5-G11 paired
S(6,10)+e(C,G)=6

C5 paired, G11 paired:
S(5,6)+S(7,11)=1
S(5,7)+S(8,11)=0
S(5,8)+S(9,11)=0
S(5,9)+S(10,11)=0

j

		1 A	2 U	3 A	4 C	5 C	6 C	7 U	8 G	9 U	10 G	11 G	12 U	13 A	14 U
1	A	0	0	0	0	0									
2	U		0	0	0	0	0								
3	A			0	0	0	0	2							
4	C				0	0	0	0	3						
5	C					0	0	0	0	0					
6	C						0	0	0	0	3				
7	U							0	0	0	0	1			
8	G								0	0	0	0	1		
9	U									0	0	0	0	2	
10	G										0	0	0	0	1
11	G											0	0	0	0
12	U												0	0	0
13	A													0	0
14	U														0

GC 3

AU 2

GU 1

Propagation:

j →

i ↓

		1 A	2 U	3 A	4 C	5 C	6 C	7 U	8 G	9 U	10 G	11 G	12 U	13 A	14 U
1	A	0	0	0	0	0	0	2	3	5	6	6	8	10	12
2	U		0	0	0	0	0	2	3	5	6	6	8	10	10
3	A			0	0	0	0	2	3	5	5	6	8	8	8
4	C				0	0	0	0	3	3	3	6	6	6	6
5	C					0	0	0	0	0	3	6	6	6	6
6	C						0	0	0	0	3	3	3	3	3
7	U							0	0	0	0	1	1	3	3
8	G								0	0	0	0	1	2	2
9	U									0	0	0	0	2	2
10	G										0	0	0	0	1
11	G											0	0	0	0
12	U												0	0	0
13	A													0	0
14	U														0

Traceback:

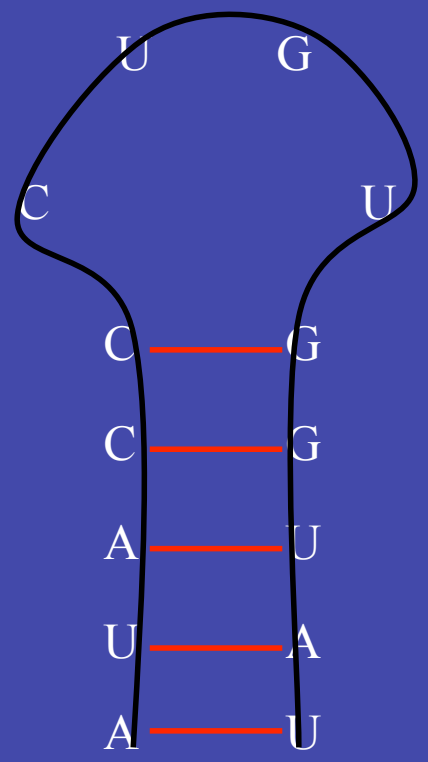
j →

i ↓

	A	U	A	C	C	C	U	G	U	G	G	U	A	U
A	0	0	0	0	0	0	2	3	5	6	6	8	10	12
U		0	0	0	0	0	2	3	5	6	6	8	10	10
A			0	0	0	0	2	3	5	5	6	8	8	8
C				0	0	0	0	3	3	3	6	6	6	6
C					0	0	0	0	0	3	6	6	6	6
C						0	0	0	0	3	3	3	3	3
U							0	0	0	0	1	1	3	3
G								0	0	0	0	1	2	2
U									0	0	0	0	2	2
G										0	0	0	0	1
G											0	0	0	0
U												0	0	0
A													0	0
U														0

Final prediction

AUACCCUGUGGUAU



Total free energy: -12 kcal/mol

Some notes

- Computational complexity: N^3
- Does not work with tertiary structure features (would invalidate DP algorithm)

Other methods

- Base pair partition functions
  - Calculate energy of all configurations
  - Lowest energy is the prediction
- Statistical sampling
  - Randomly generating structure with probability distribution = energy function distribution
    - This makes it more likely that lowest energy structure is found
- Sub-optimal sampling

RNA homology structure prediction

- Molecules with similar functions and different nucleotide sequences will form similar structures
- Correctly identifies high percentage of secondary structure pairings and a smaller number of tertiary interactions
- Primarily a manual method

How well do these methods perform?

- **Energy minimization** (via dynamic programming)
  - **73% avg. prediction accuracy** - single sequence
- **2) Comparative sequence analysis**
  - **97% avg. prediction accuracy** - multiple sequences (e.g., highly conserved rRNAs)
  - **much lower** if sequence conservation is lower &/or fewer sequences are available for alignment
- **3) Combined -** recent developments:
- combine thermodynamics & co-variation
- & experimental constraints? **IMPROVED RESULTS**

Outline

- RNA primary structure
- Small RNA prediction
- RNA secondary structure & prediction
- RNA tertiary structure & prediction

Hierarchical organization of RNA molecules

Primary structure:

5' to 3' list of covalently linked nucleotides, named by the attached base

Secondary Structure

List of base pairs, denoted by i•j for a pairing between the i-th and j-th Nucleotides, r_i and r_j, where $i<j$ by convention.
Pairing mostly occur as A•U and G•C (Watson Crick), and G •U (wobble)

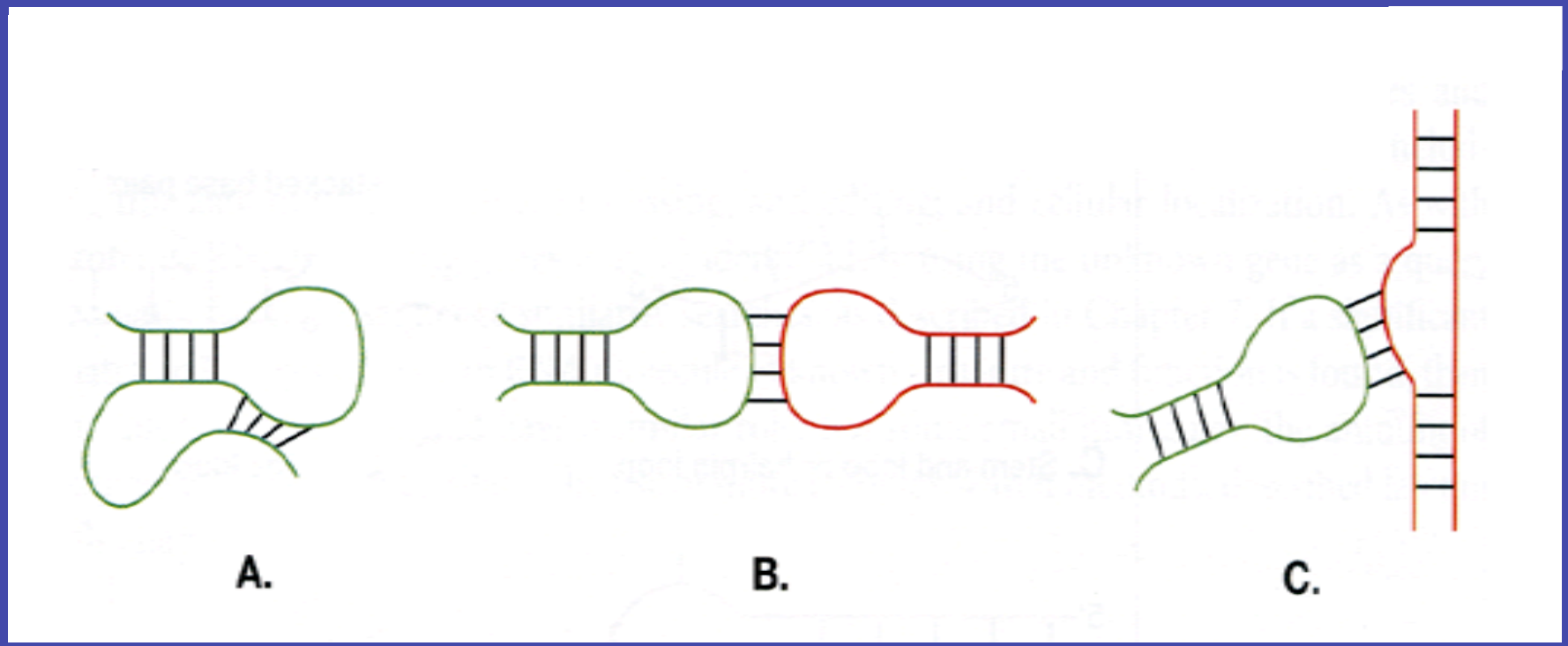
Helices are inferred when two or more base pairs occur adjacent to one another

Tertiary structure:

List of interactions between secondary structure features

RNA "tertiary interactions"

In addition to secondary structural interactions in RNA, there are also tertiary interactions, including: (A) pseudoknots, (B) kissing hairpins and (C) hairpin-bulge contact.



Pseudoknot **Kissing hairpins** **Hairpin-bulge**

Do not obey "parentheses rule"

RNA structure prediction strategies

Tertiary structure prediction

Requires "craft" & significant user input & insight

1) **Extensive *comparative sequence analysis* to predict tertiary contacts (co-variation)**
   e.g., **MANIP** - *Westhof*
2) **Use *experimental data to constrain* model building**
   e.g., **MC-CYM** - *Major*
3) ***Homology modeling* using sequence alignment & reference tertiary structure (not many of these!)**
4) **Low resolution *molecular mechanics***
   e.g., **yammp** - *Harvey*

Summary

- RNA primary structure
- RNA secondary structure & prediction
- RNA tertiary structure & prediction

Prediction programs

- ILM (web server)

http://cic.cs.wustl.edu/RNA/

- MFOLD (Zuker) (web server)

http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1.cgi

- Genebee (both comparative + energy model) (web server)

  http://www.genebee.msu.edu/services/rna2_reduced.html

- Vienna RNA package

  http://www.tbi.univie.ac.at/~ivo/RNA/

- Mc-Sym (Computer Science approach)

  http://www-lbit.iro.umontreal.ca/mcsym

Useful web sites on RNA

- *Comparative RNA web site*
  http://www.rna.icmb.utexas.edu/
- *RNA world*
  http://www.imb-jena.de/RNA.html
- *RNA page by Michael Suker*
  http://www.bioinfo.rpi.edu/~zukerm/rna/
- *RNA structure database*
  http://www.rnabase.org/
  http://ndbserver.rutgers.edu/ (nucleic acid database)
  http://prion.bchs.uh.edu/bp_type/ (non canonical bases)
- *RNA structure classification*
  http://scor.berkeley.edu/
- *RNA visualisation*
  http://ndbserver.rutgers.edu/services/download/index.html#rnaview
  http://rutchem.rutgers.edu/~xiangjun/3DNA/

To Do

Take the Ribosomal RNAs annotated in your genome and predict their structure using one of the servers above

Small RNA prediction

Prof:Rui Alves
ralves@cmb.udl.es
973702406
Dept Ciencies Mediques Basiques,
1st Floor, Room 1.08

Prediction programs

- ILM (web server)

http://cic.cs.wustl.edu/RNA/

- MFOLD (Zuker) (web server)

http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1.cgi

- Genebee (both comparative + energy model) (web server)

  http://www.genebee.msu.edu/services/rna2_reduced.html

- Vienna RNA package

  http://www.tbi.univie.ac.at/~ivo/RNA/

- Mc-Sym (Computer Science approach)

  http://www-lbit.iro.umontreal.ca/mcsym

Useful Primers on microRNA

- *Comparative RNA web site*
  http://www.rna.icmb.utexas.edu/
- *RNA world*
  http://www.imb-jena.de/RNA.html
- *RNA page by Michael Suker*
  http://www.bioinfo.rpi.edu/~zukerm/rna/
- *RNA structure database*
  http://www.rnabase.org/
  http://ndbserver.rutgers.edu/ (nucleic acid database)
  http://prion.bchs.uh.edu/bp_type/ (non canonical bases)
- *RNA structure classification*
  http://scor.berkeley.edu/
- *RNA visualisation*
  http://ndbserver.rutgers.edu/services/download/index.html#rnaview
  http://rutchem.rutgers.edu/~xiangjun/3DNA/

Useful web sites on micro RNA prediction

http://en.wikipedia.org/wiki/MicroRNA

http://cbit.snu.ac.kr/~ProMiR2/

http://miracle.igib.res.in/miracle/

Useful web sites on micro RNA target prediction

http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/

http://www.microrna.org/

http://pictar.bio.nyu.edu/

http://mirna.imbb.forth.gr/microinspector/

http://cbit.snu.ac.kr/~miTarget/

http://tiger.dbs.nus.edu.sg/microtar/

http://cbcsrv.watson.ibm.com/rna22.html

To Do

- Take the intergenic regions of the M. xanthus genome and got to http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi
- Predict possible small RNAs
- Look for a server where you can then look for targets for the small RNAs