RNA Structure Prediction

Jianlin Cheng, PhD

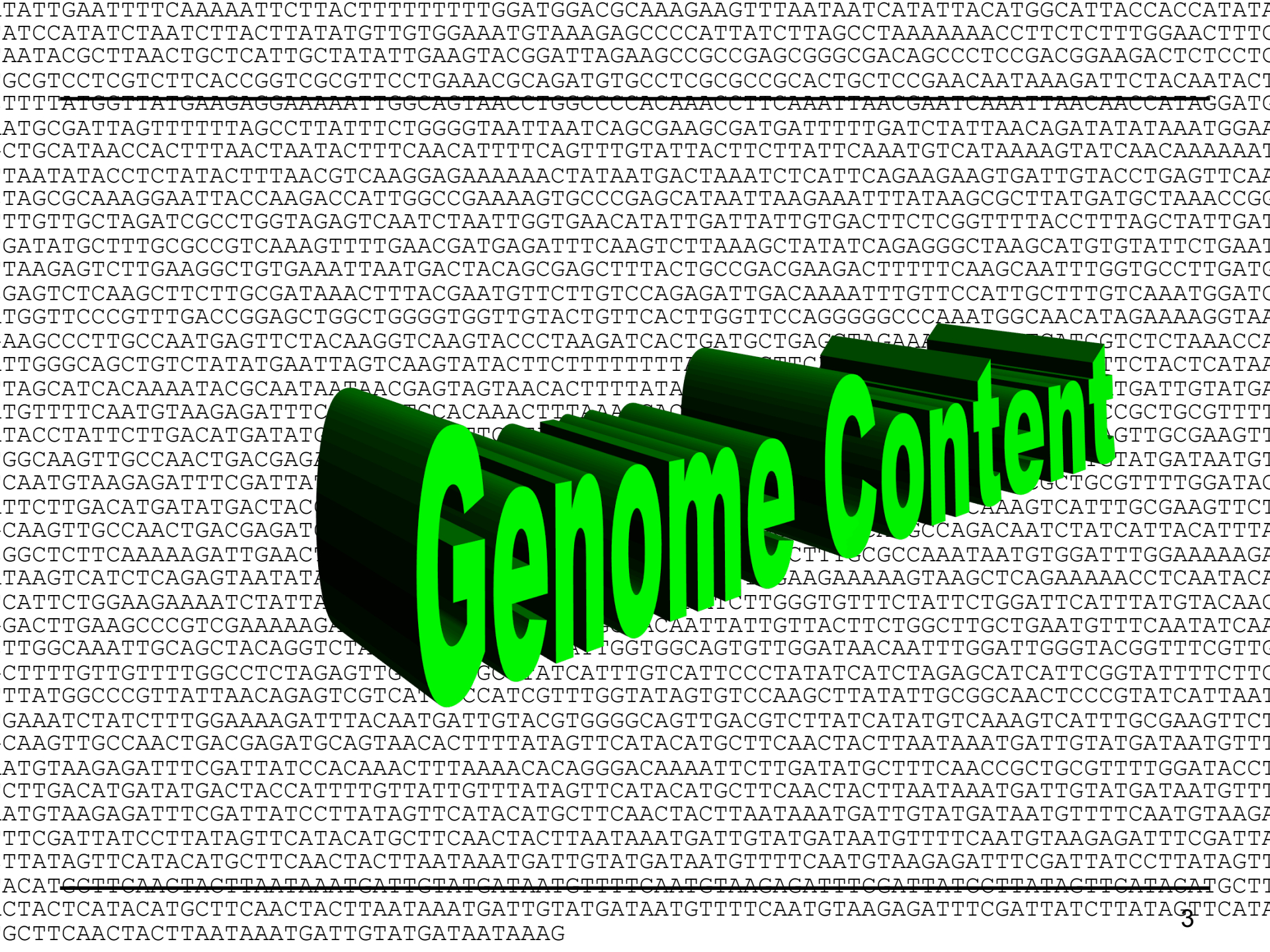
Associate Professor
Computer Science Department
University of Missouri, Columbia

2013

Based on Gill Bejerano's CS173 at Stanford and Cheng's lecture

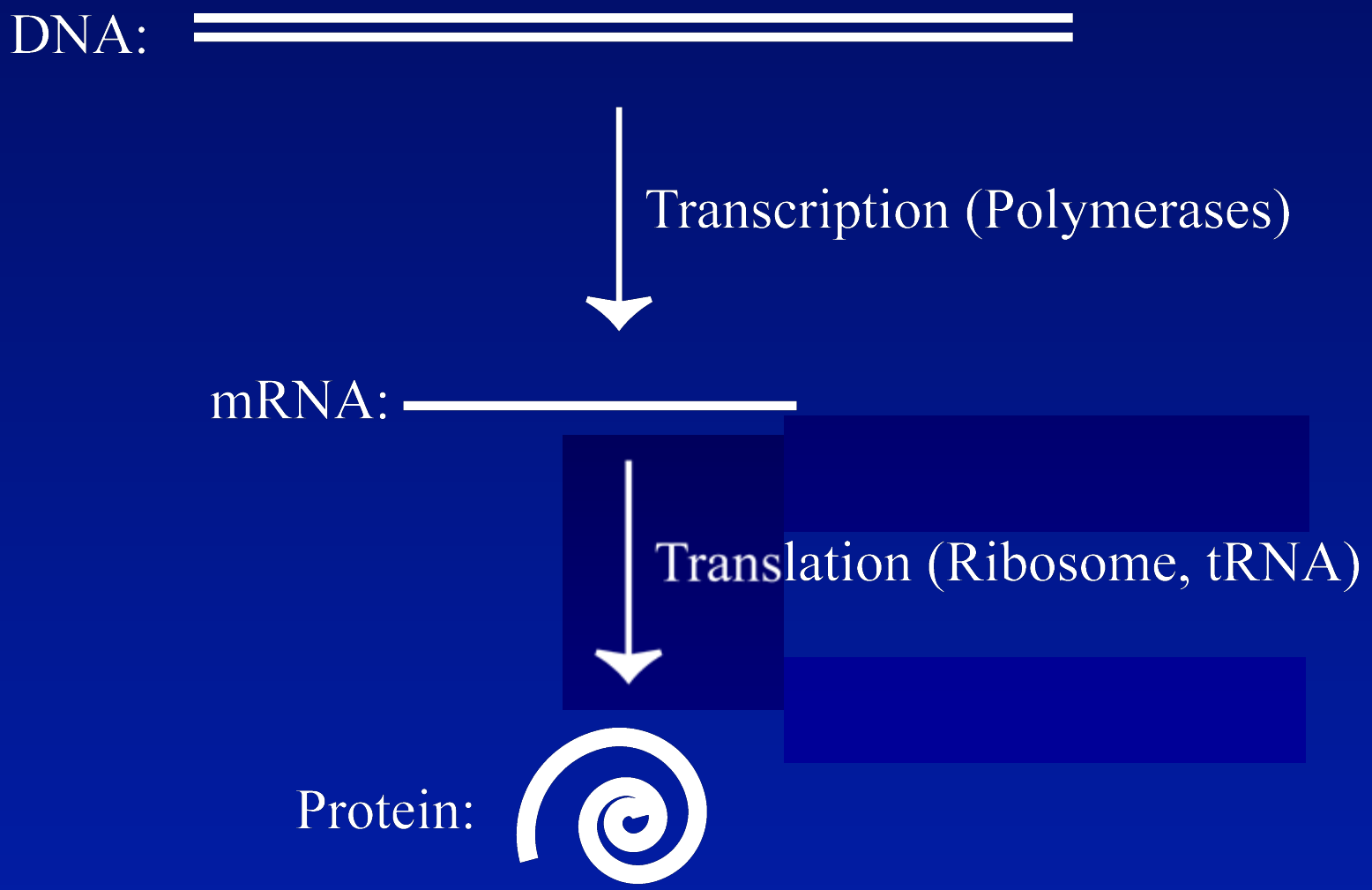
Outline

- Introduction to non coding RNAs
- RNA secondary structure prediction
- RNA tertiary structure prediction

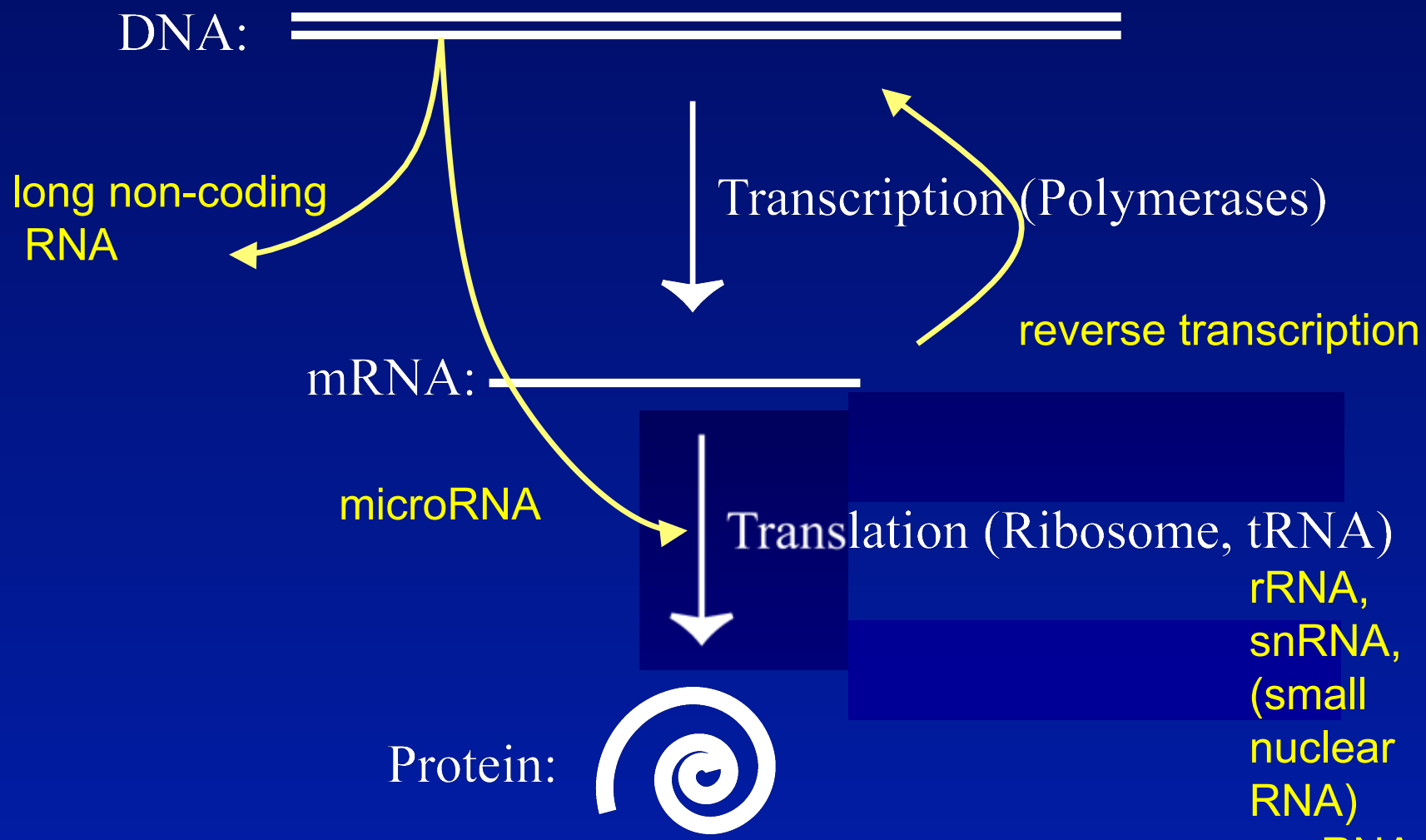

Genome Content

“non coding” RNAs (ncRNA)

Central Dogma of Biology:



Active forms of “non coding” RNA



What is ncRNA?

- Non-coding RNA (ncRNA) is an RNA that functions without being translated to a protein.
- Known roles for ncRNAs:
  - RNA catalyzes excision/ligation in introns.
  - RNA catalyzes the maturation of tRNA.
  - RNA catalyzes peptide bond formation.
  - RNA is a required subunit in telomerase.
  - RNA plays roles in immunity and development (RNAi).
  - RNA plays a role in dosage compensation.
  - RNA plays a role in carbon storage.
  - RNA is a major subunit in the SRP, which is important in protein trafficking.
  - RNA guides RNA modification.

– RNA can do so many different functions, it is thought in the beginning there was an RNA World, where RNA was both the information carrier and active molecule.

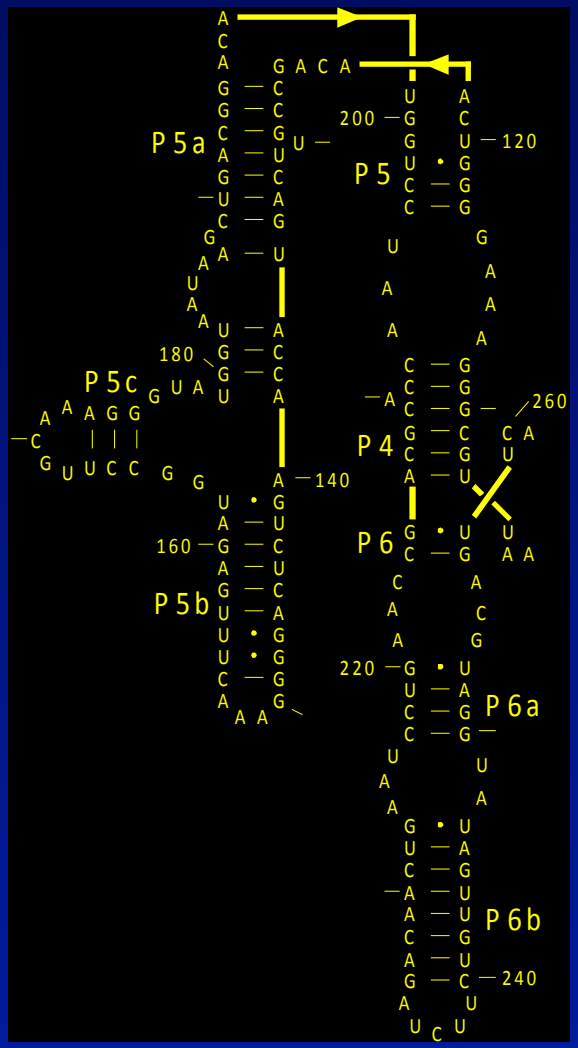
“non coding” RNAs (ncRNA)

Small structural RNAs (ssRNA)

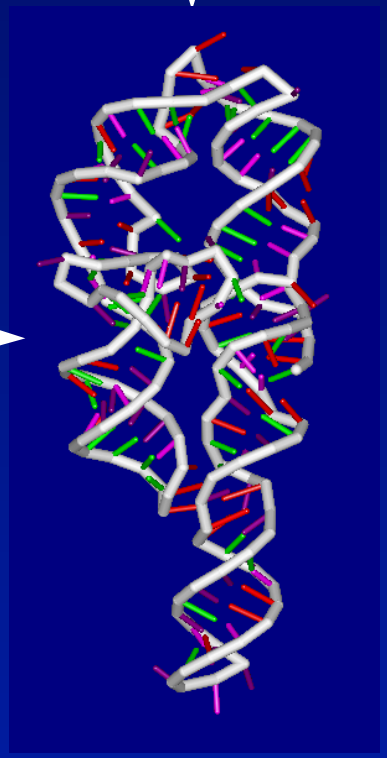
ssRNA Folds into Secondary and 3D Structures

AAUUGCGGGAAAGGGGUCAA
CAGCCGUUCAGUACCAAGUC
UCAGGGGAAACUUUGAGAUG
GCCUUGCAAAGGGUAUGGUA
AUAAGCUGACGGACAUGGUC
CUAACCACGCAGCCAAGUCC
UAAGUCAACAGAUCUUCUGU
UGAUAUGGAUGCAGUUCA

We would like to predict them from sequence.

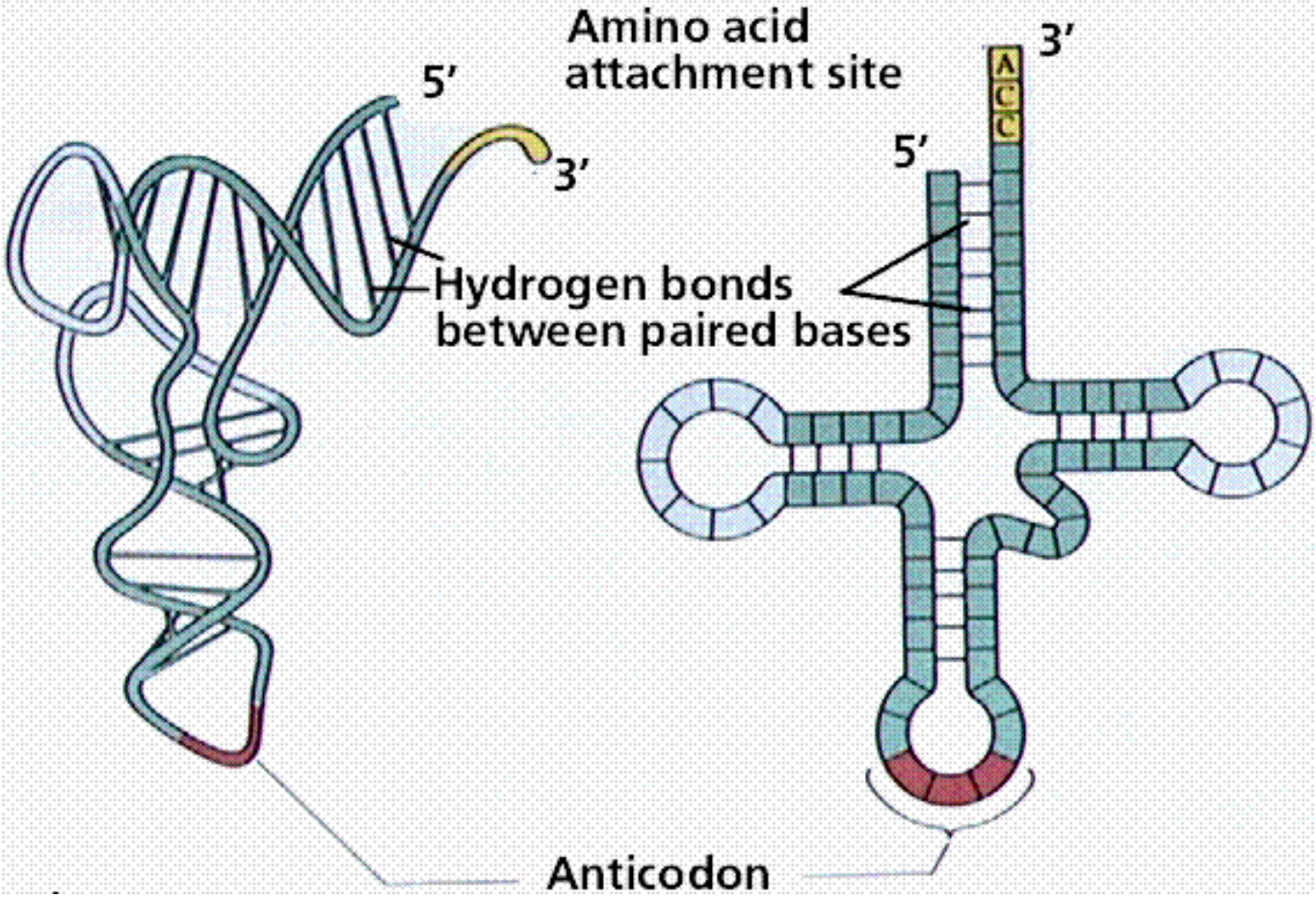


Waring & Davies. (1984) *Gene 28*: 277.

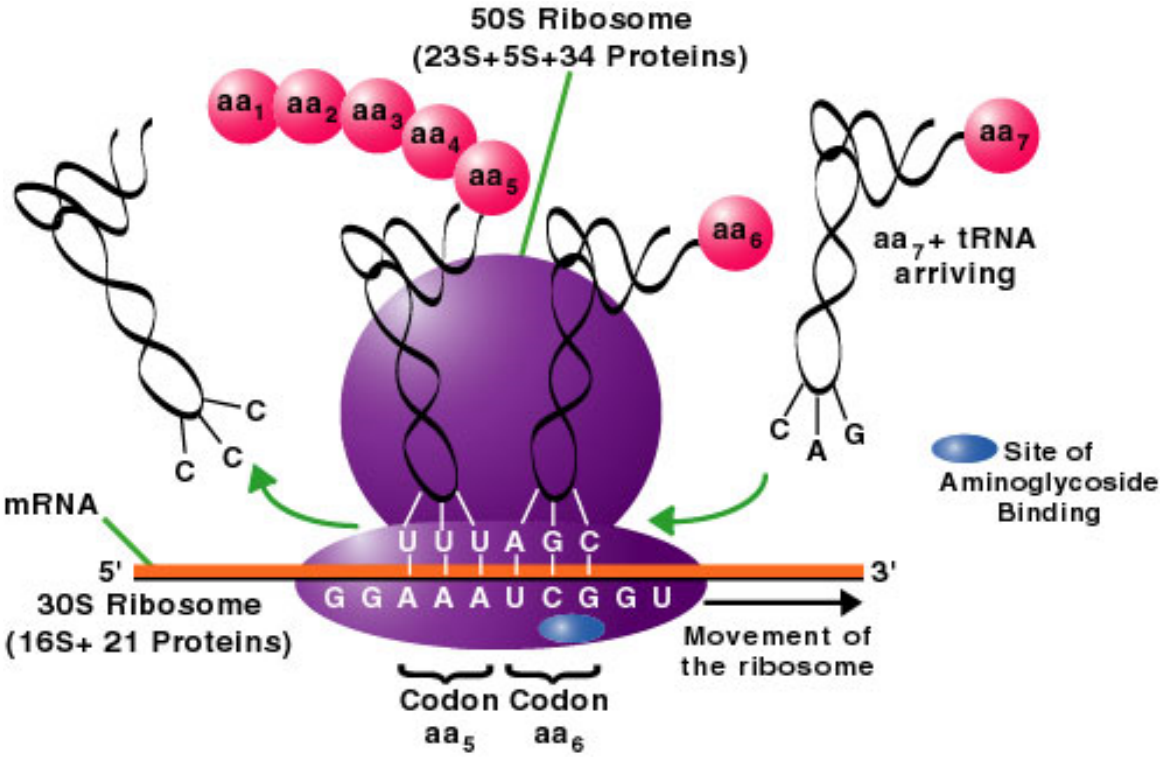


Cate, *et al.* (Cech & Doudna). (1996) *Science 273*:1678.

For example, tRNA

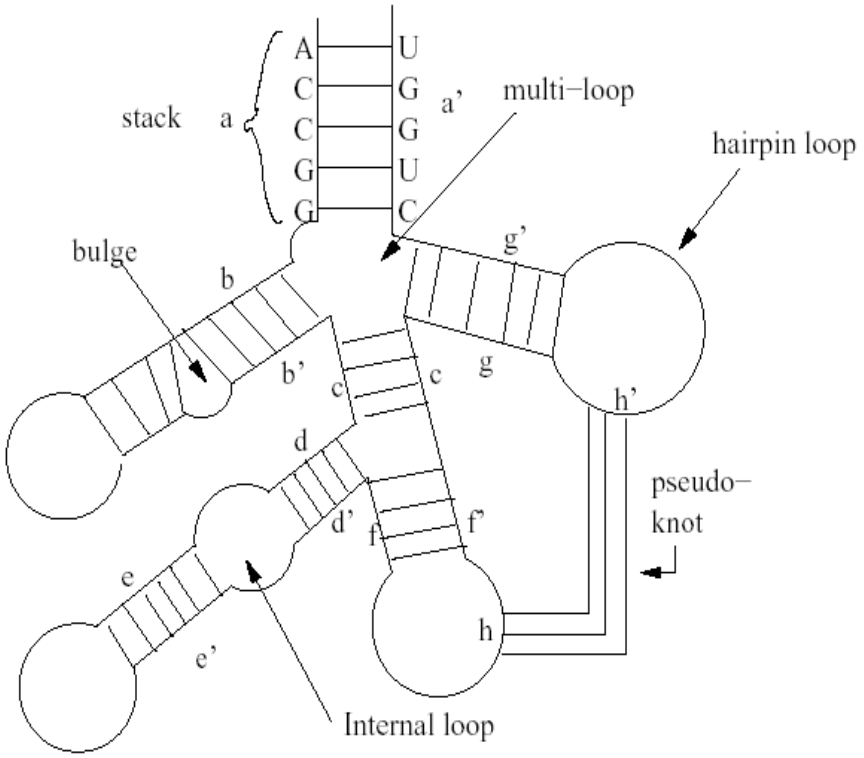


tRNA Activity



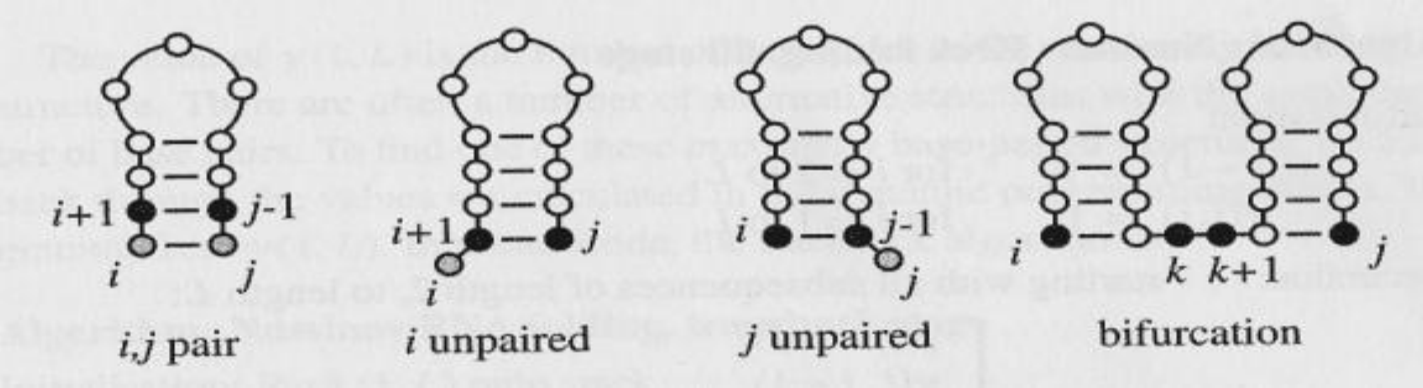
ssRNA structure rules

- Canonical basepairs:
  - Watson-Crick basepairs:
    - G – C
    - A – U
  - Wobble basepair:
    - G - U
- Stacks: continuous nested basepairs. (energetically favorable)
- Non-basepaired loops:
  - Hairpin loop
  - Bulge
  - Internal loop
  - Multiloop
- Pseudo-knots



Ab initio RNA structure prediction: lots of Dynamic Programming

- Objective: Maximizing the number of base pairs (Nussinov *et al*, 1978)

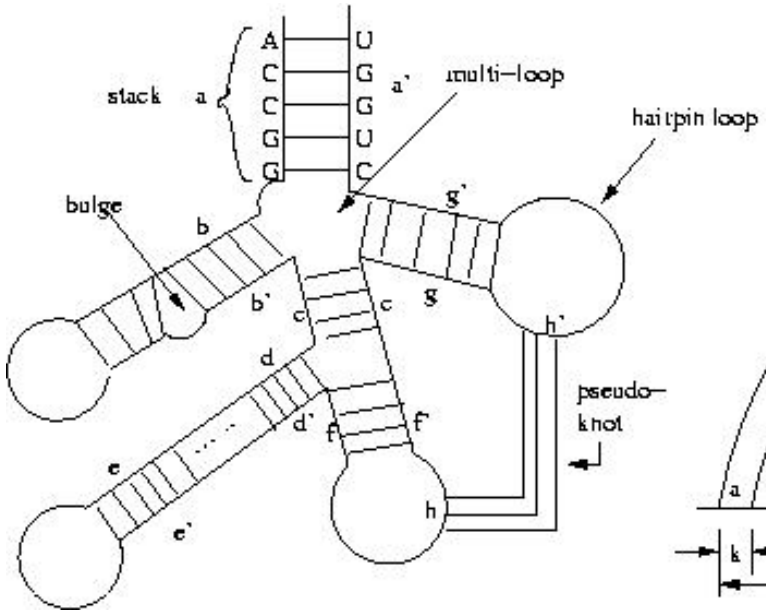


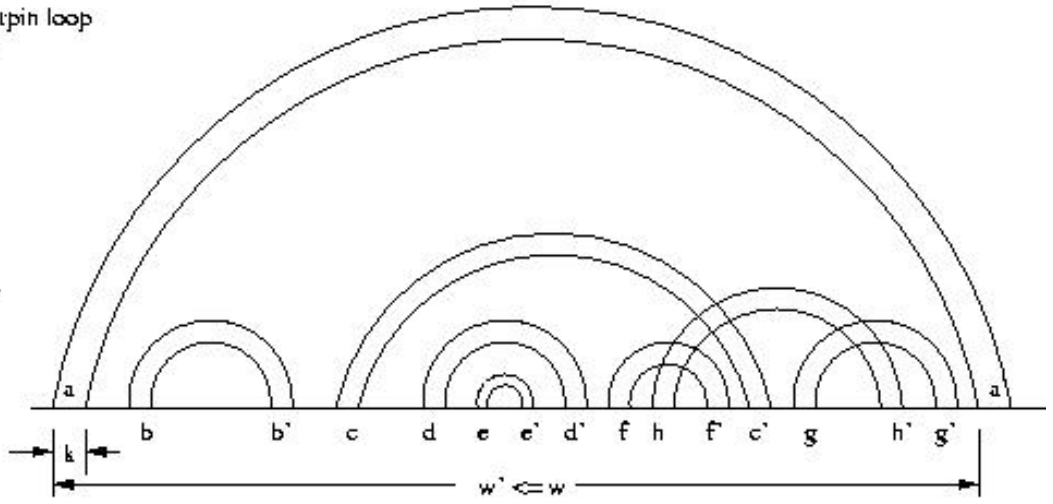
$$C_{i,j} = \max\{\, C_{i+1,j-1} + \delta(i,j),\; C_{i,j-1},\; C_{i+1,j},\; \max_{i \le k < j}\{C_{i,k} + C_{k+1,j}\} \,\}$$

simple model:
$\delta(i, j) = 1$ if allowed
fancier model:
GC > AU > GU

RNA structure

- Base-pairing defines a secondary structure. The base-pairing is usually non-crossing.





Stochastic context-free grammar

S → aSu

S → cSg

S → gSc

S → uSa

S → a

S → c

S → g

S → u

S → SS

S ➔ aSu

➔ acSgu

➔ accSggu

➔ accuSaggu

➔ accuSSaggu

➔ accugScSaggu

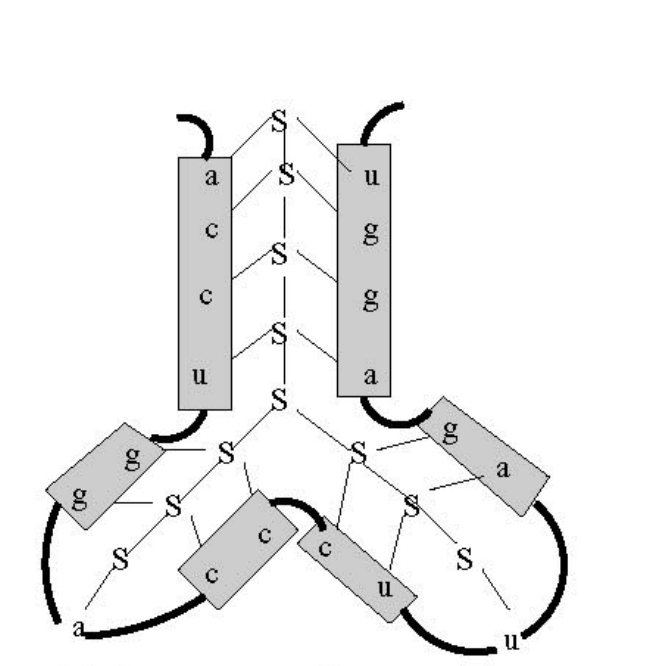
➔ accuggSccSaggu

➔ accuggaccSaggu

➔ accuggacccSgaggu

➔ accuggacccuSagaggu

➔ accuggacccuuagaggu



1. A CFG 2. A derivation of "accuggacccuuagaggu" 3. Corresponding structure

Cool algorithmics. Unfortunately…

Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs

*Elena Rivas and Sean R. Eddy**

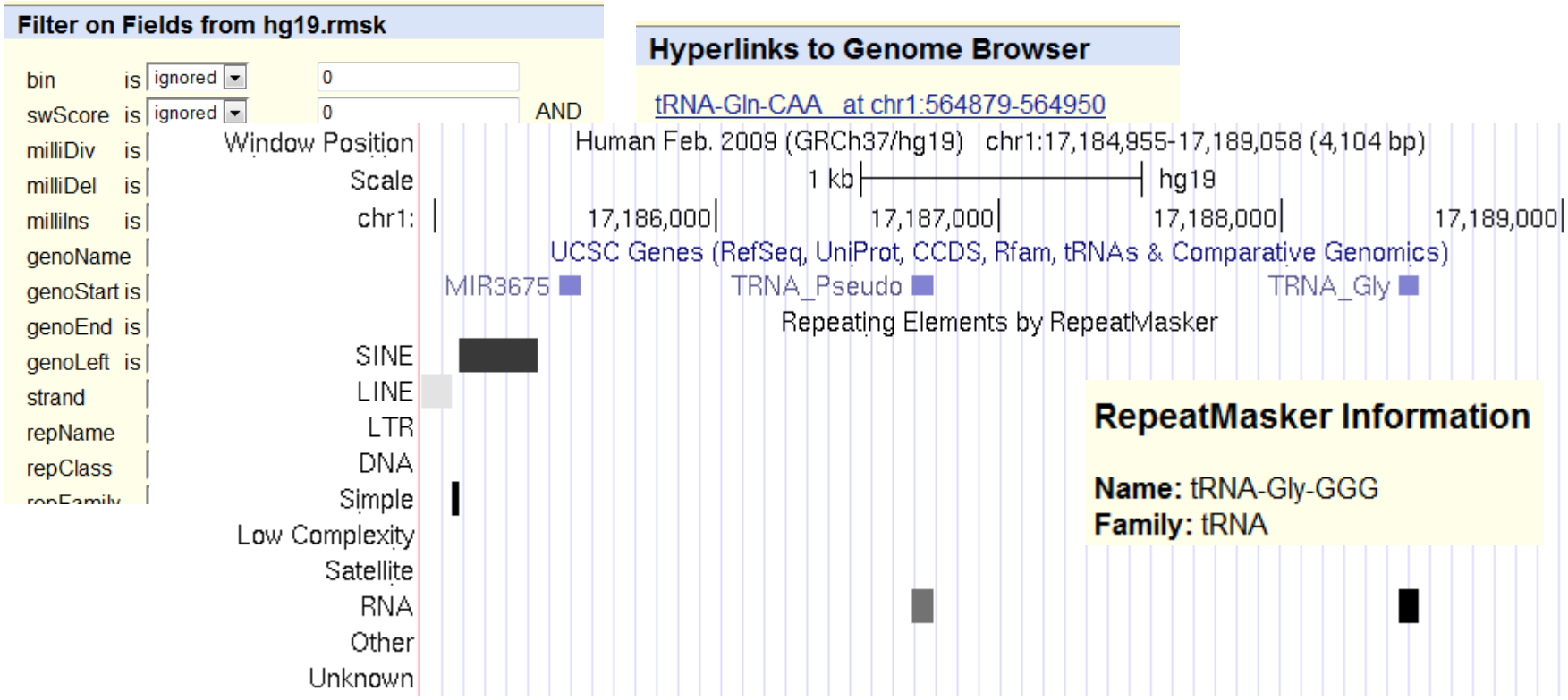
Department of Genetics, Washington University, St. Louis, MO 63110, USA

Received on August 4, 1999; revised on December 15, 1999; accepted on December 21, 1999

- Random DNA (with high GC content) often folds into low-energy structures.
- We will mention powerful newer methods later on.

ssRNA transcription

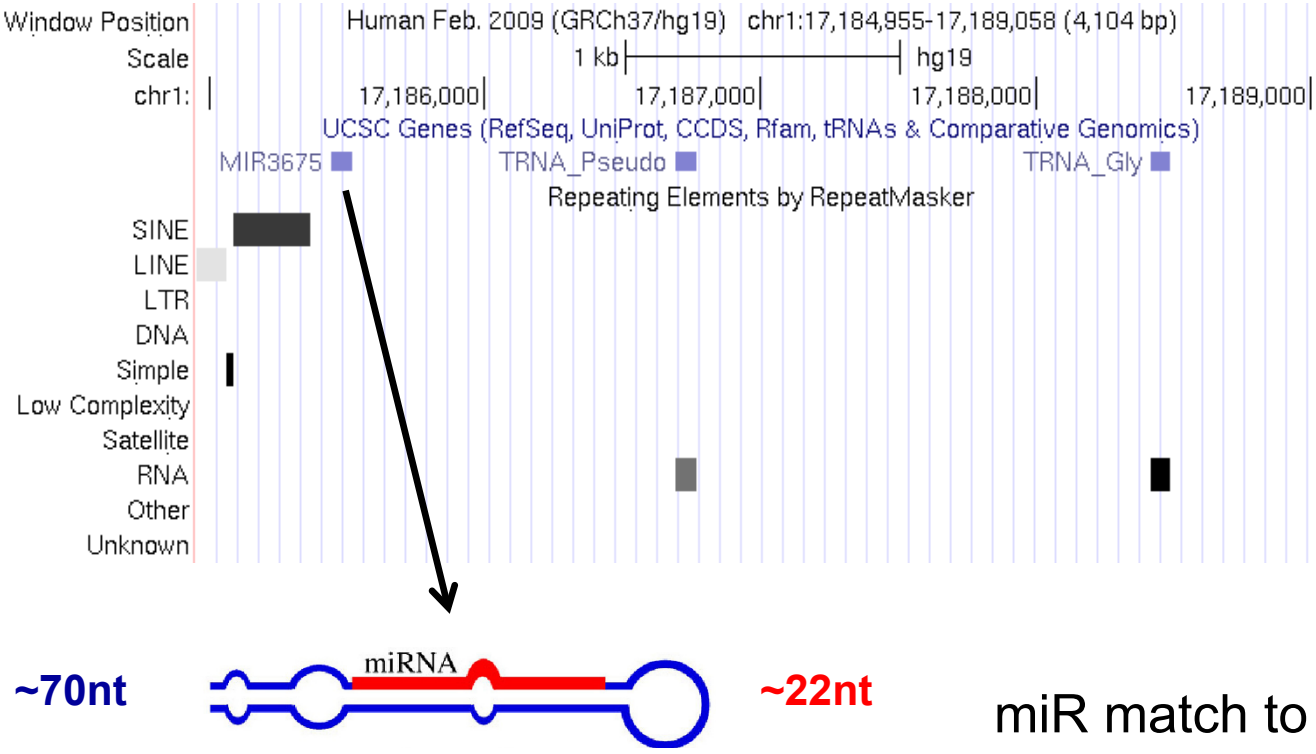
- ssRNAs like tRNAs are usually encoded by short "non coding" genes, that transcribe independently.
- Found in both the UCSC "known genes" track, and as a subtrack of the RepeatMasker track



“non coding” RNAs (ncRNA)

microRNAs (miRNA/miR)

MicroRNA (miR)

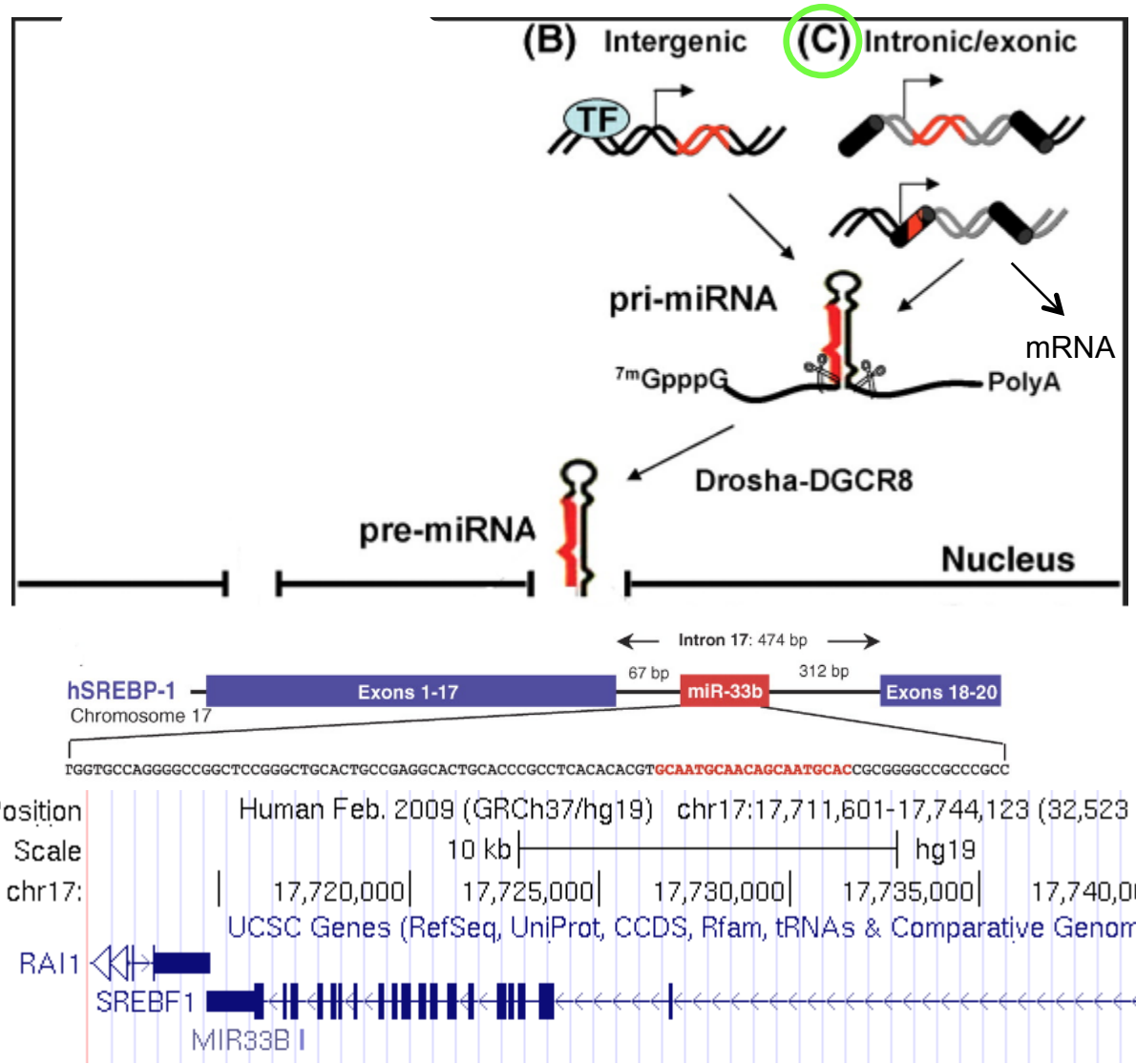


miR match to target mRNA is quite loose.

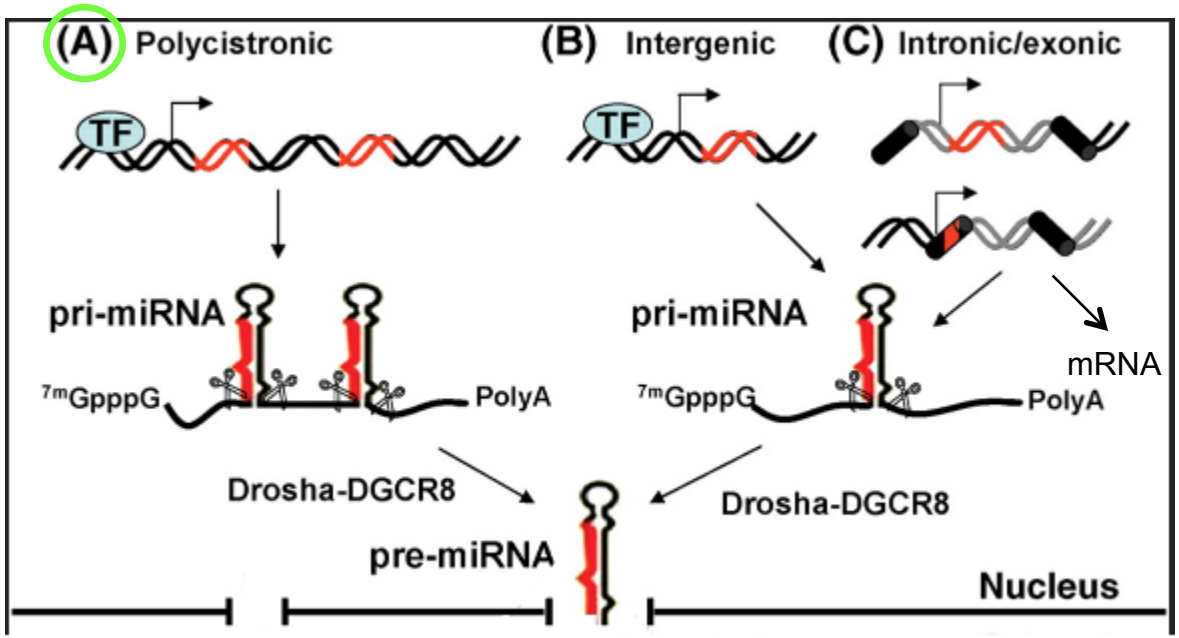
```
3' uagcgccaaauauggUUUACUUA 5' has-miR-579
   ||: |  |||||:  ||||||:|
5' atttctttttatggaAAATGAGT 3' LR1G3
     Out-seed        Seed
```

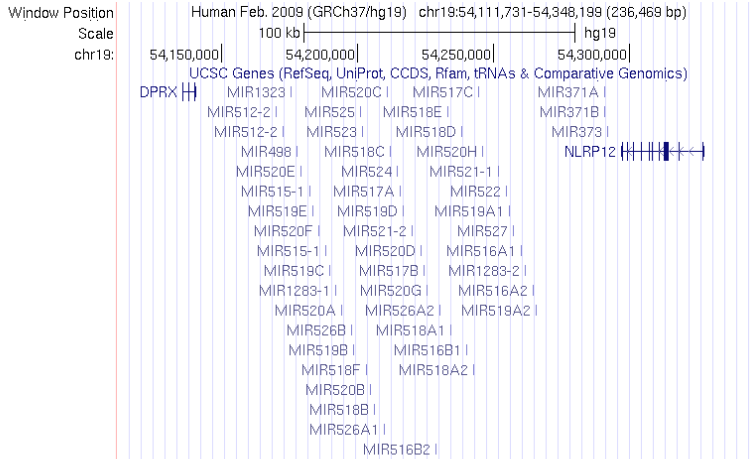
→ a single miR can regulate the expression of hundreds of genes.

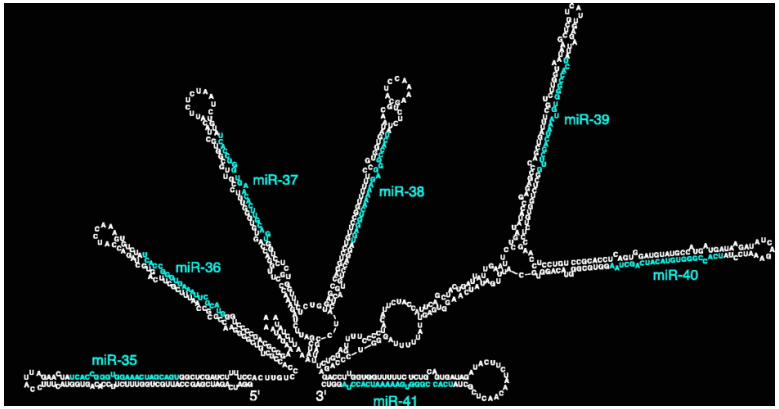
MicroRNA Transcription



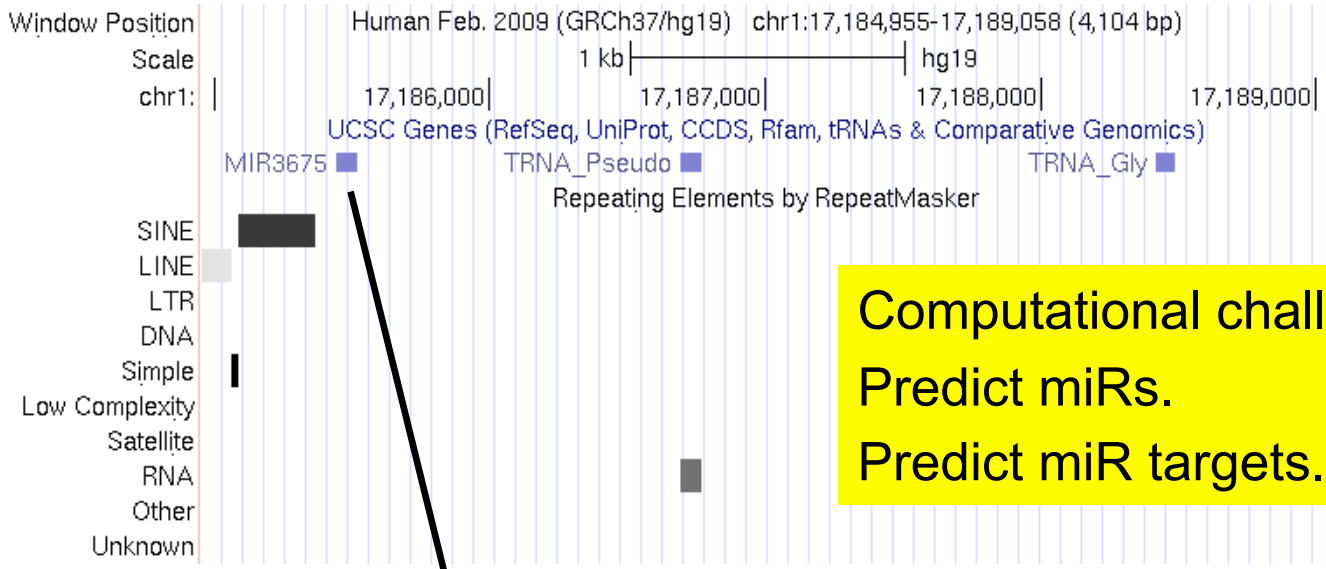
MicroRNA Transcription







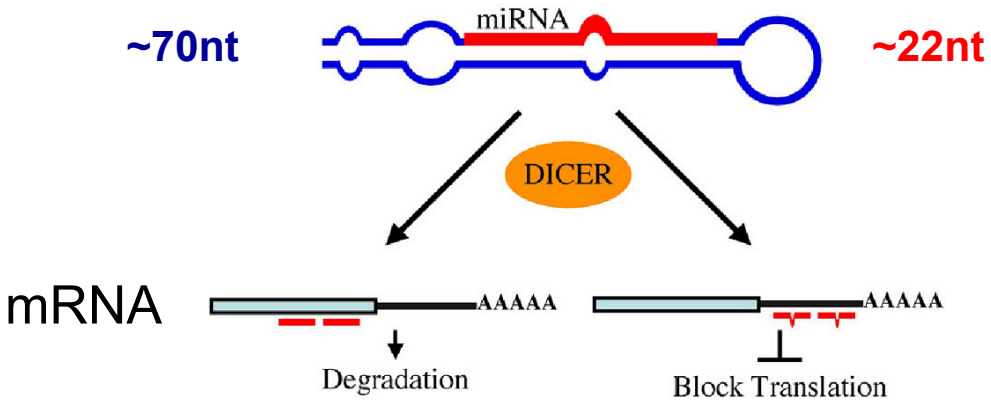
MicroRNA (miR)



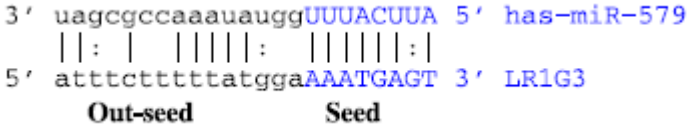
Computational challenges:

Predict miRs.

Predict miR targets.



miR match to target mRNA is quite loose.



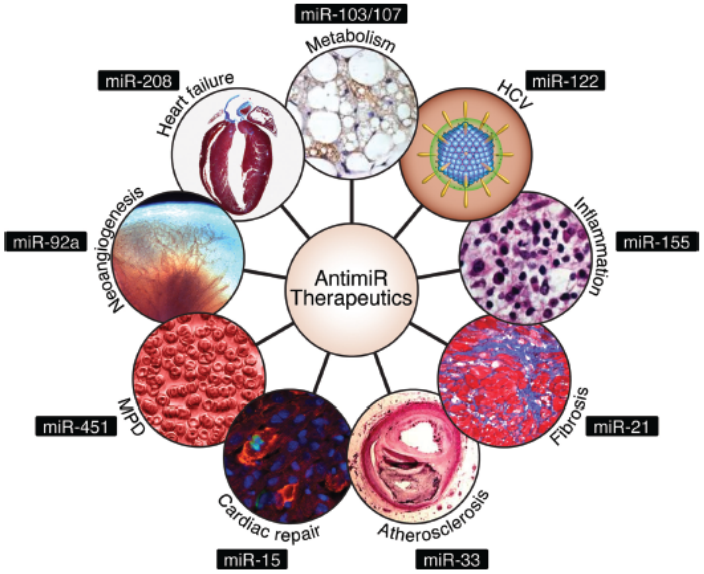
→ a single miR can regulate the expression of hundreds of genes.

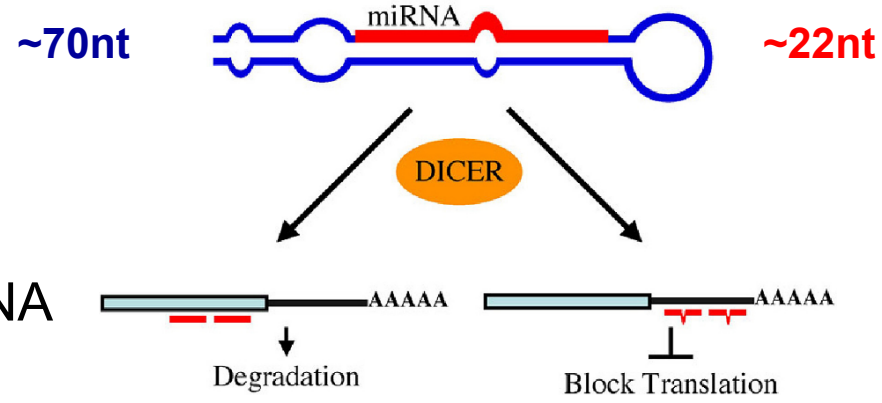
MicroRNA Therapeutics

Idea: bolster/inhibit miR production to broadly modulate protein production

Hope: “right” the good guys and/or “wrong” the bad guys

Challenge: and not vice versa.





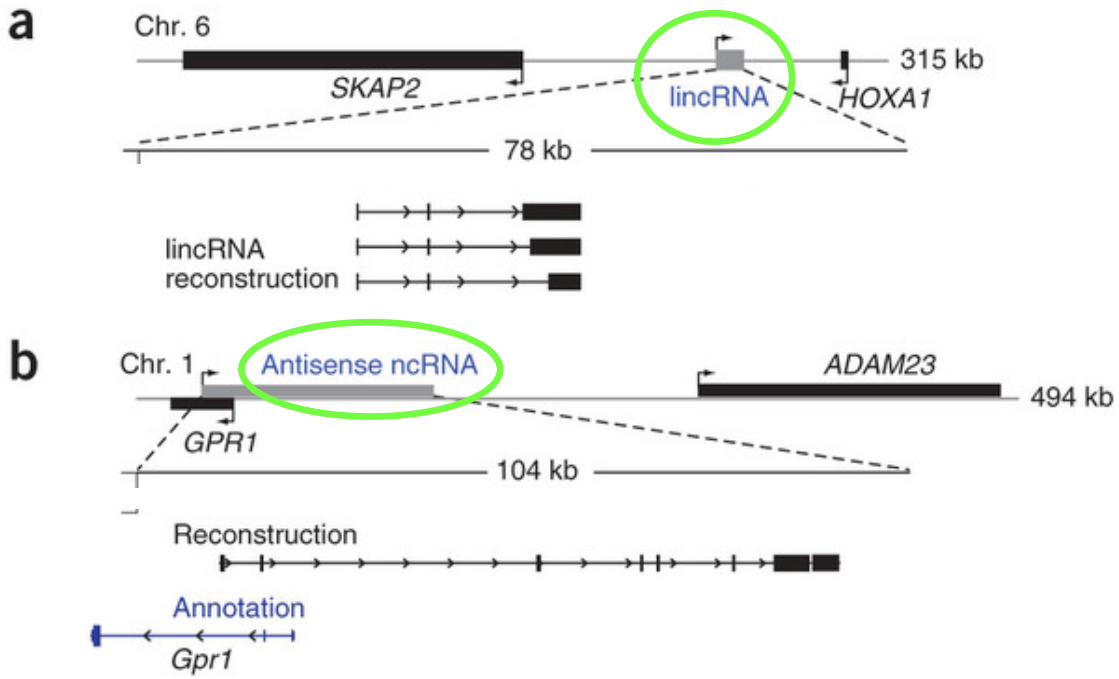
miR match to target mRNA is quite loose.

```
3' uagcgccaaauauggUUUACUUA 5' has-miR-579
   ||: |  |||||:  |||||||:|
5' atttctttttatggaAAATGAGT 3' LR1G3
     Out-seed        Seed
```

→ a single miR can regulate the expression of hundreds of genes.

Other Non Coding Transcripts

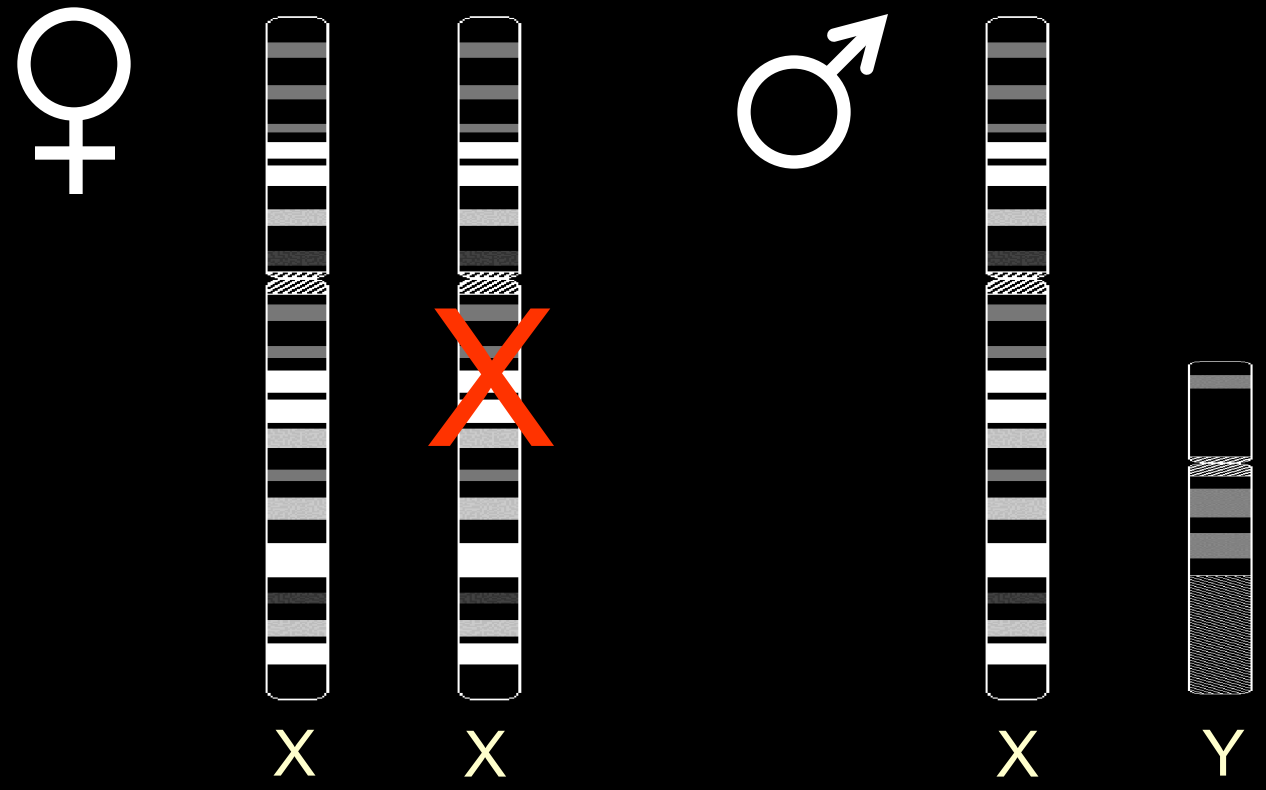
lncRNAs (long non coding RNAs)



Don't seem to fold into clear structures (or only a sub-region does).
Diverse roles only now starting to be understood.

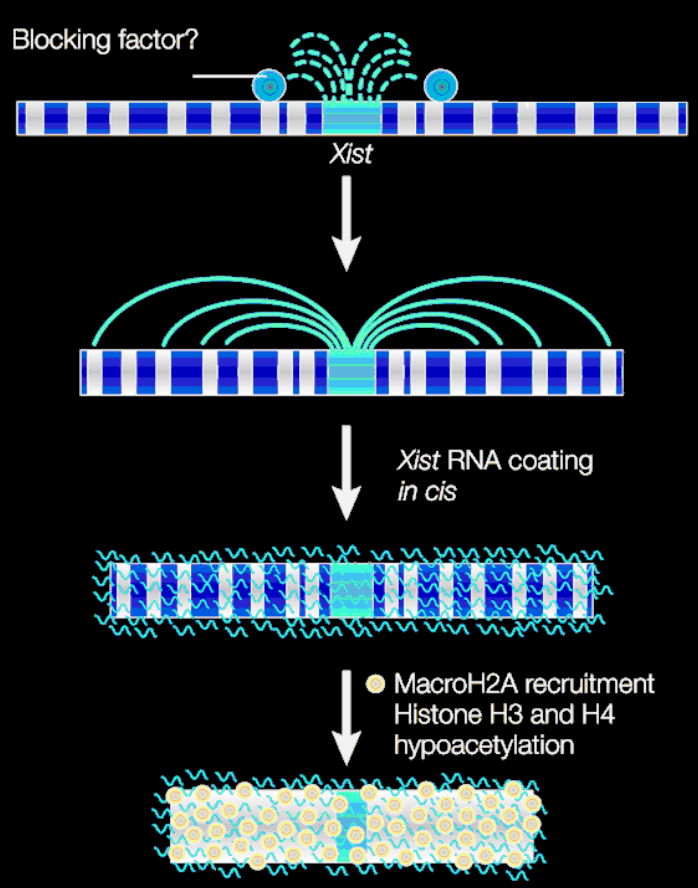
→ Hard to detect or predict function computationally (currently)

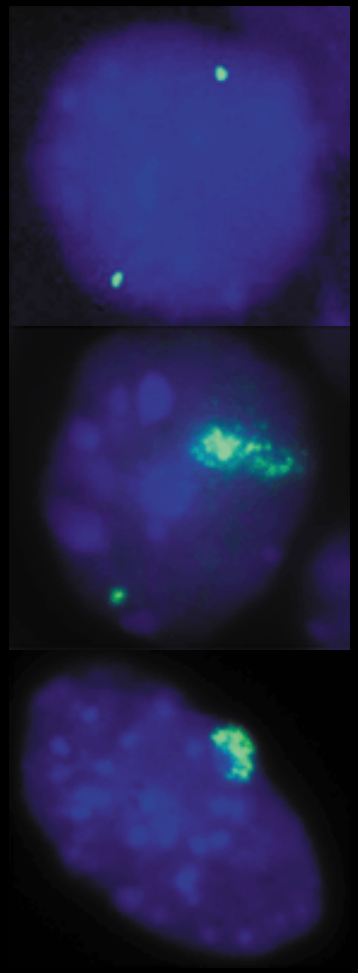
X chromosome inactivation in mammals



Dosage compensation

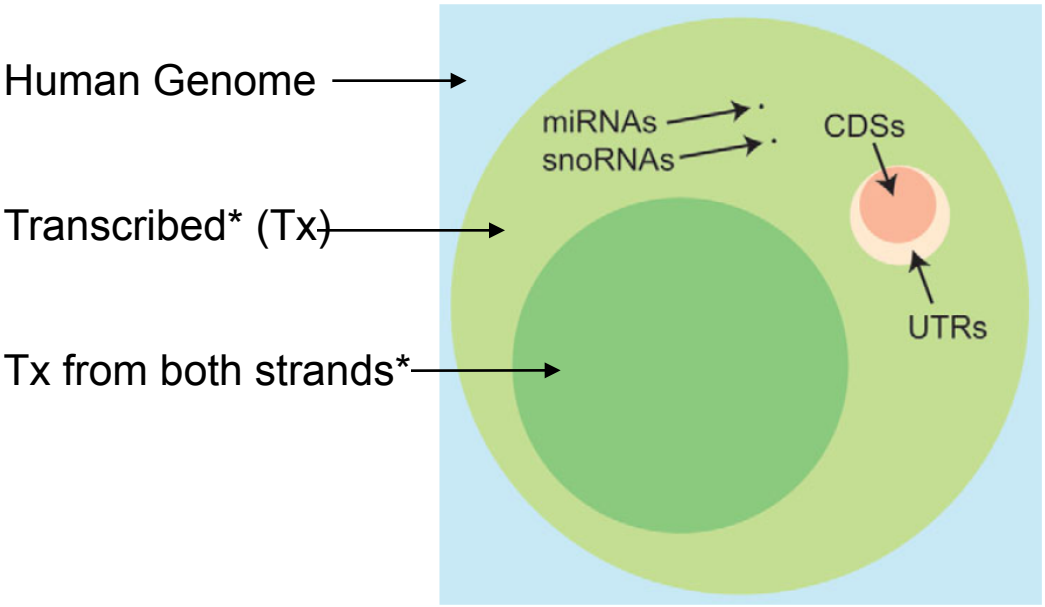
Xist – X inactive-specific transcript





Avner and Heard, Nat. Rev. Genetics 2001 2(1):59-67

Transcripts, transcripts *everywhere*



* True size of set unknown

Or are they?

Most "Dark Matter" Transcripts Are Associated With Known Genes

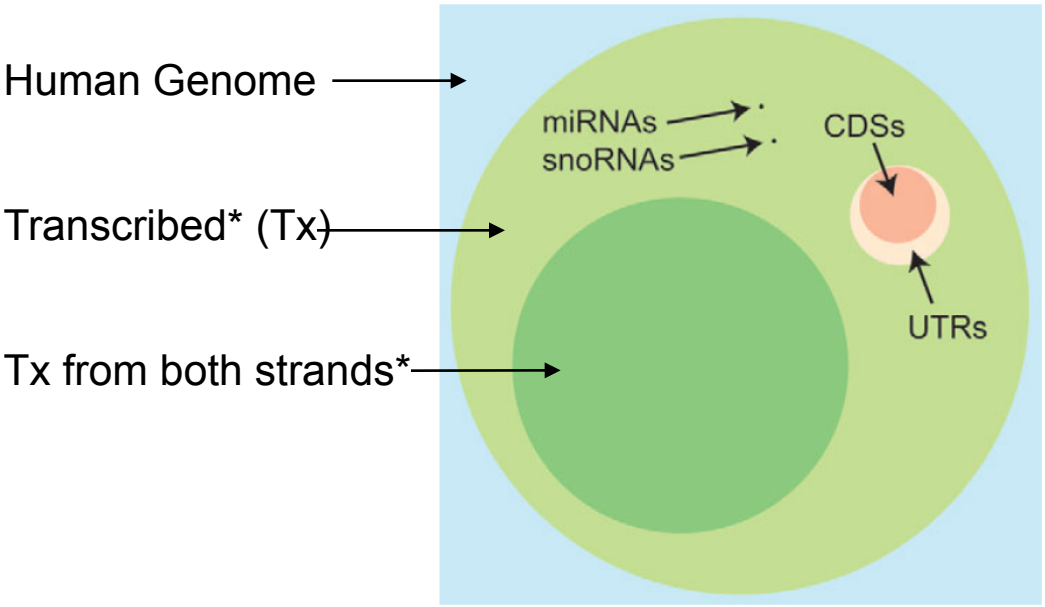
Harm van Bakel[1], Corey Nislow[1,2], Benjamin J. Blencowe[1,2], Timothy R. Hughes[1,2]*

1 Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, **2** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

Abstract

A series of reports over the last few years have indicated that a much larger portion of the mammalian genome is transcribed than can be accounted for by currently annotated genes, but the quantity and nature of these additional transcripts remains unclear. Here, we have used data from single- and paired-end RNA-Seq and tiling arrays to assess the quantity and composition of transcripts in PolyA+ RNA from human and mouse tissues. Relative to tiling arrays, RNA-Seq identifies many fewer transcribed regions ("seqfrags") outside known exons and ncRNAs. Most nonexonic seqfrags are in introns, raising the possibility that they are fragments of pre-mRNAs. The chromosomal locations of the majority of intergenic seqfrags in RNA-Seq data are near known genes, consistent with alternative cleavage and polyadenylation site usage, promoter- and terminator-associated transcripts, or new alternative exons; indeed, reads that bridge splice sites identified 4,544 new exons, affecting 3,554 genes. Most of the remaining seqfrags correspond to either single reads that display characteristics of random sampling from a low-level background or several thousand small transcripts (median length = 111 bp) present at higher levels, which also tend to display sequence conservation and originate from regions with open chromatin. We conclude that, while there are bona fide new intergenic transcripts, their number and abundance is generally low in comparison to known exons, and the genome is not as pervasively transcribed as previously reported.

The million dollar question

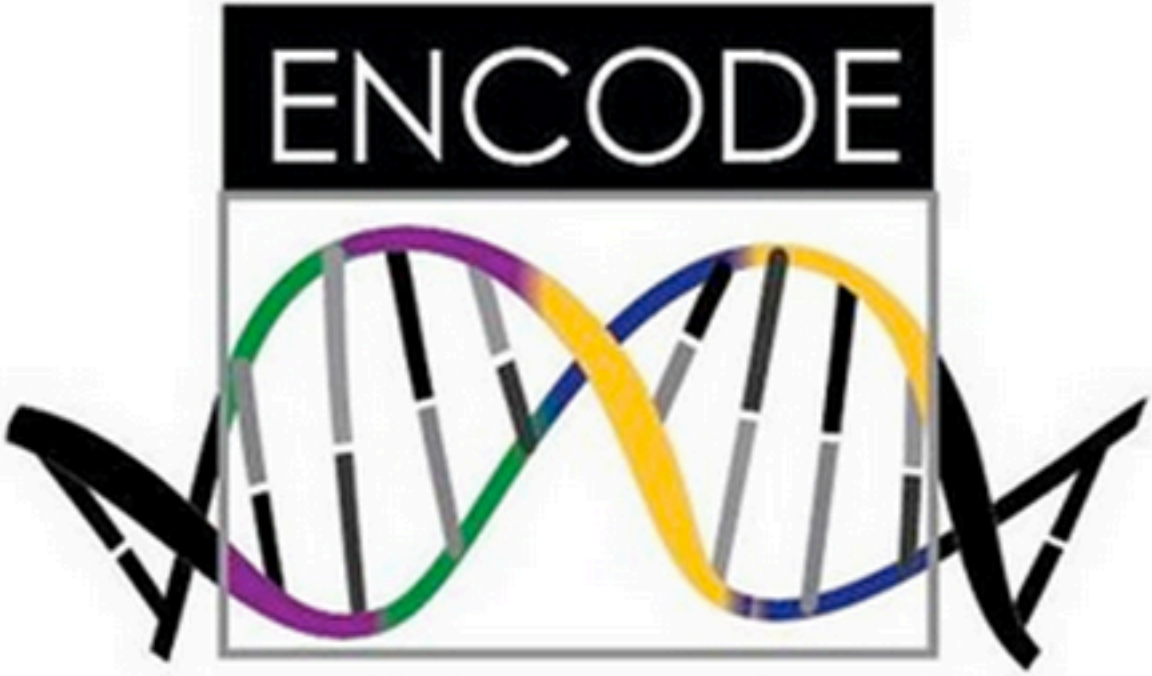


Leaky tx?

Functional?

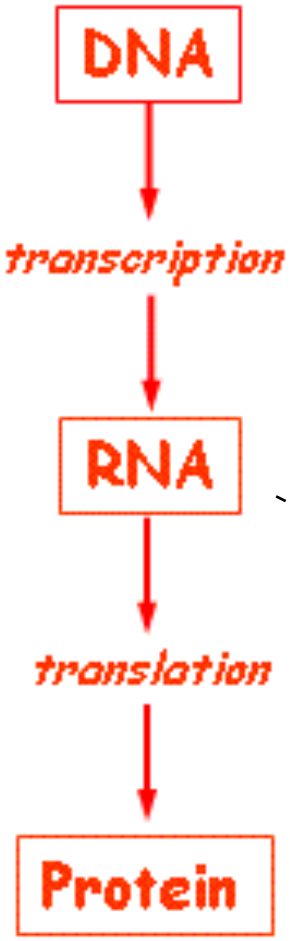
* True size of set unknown

National Human Genome Research Institute

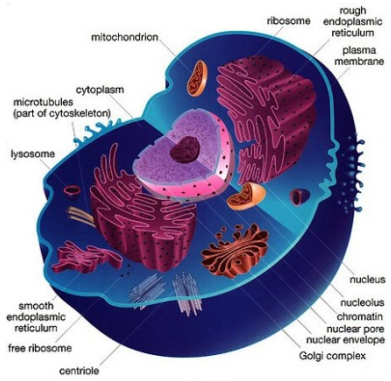


With an additional $30.3 million in funding from the NIH, the ENCODE project will focus on more data types and more human cells and tissues. Source: NIH.gov

Coding and non-coding gene production



To change its behavior a cell can change the repertoire of genes and ncRNAs it makes.



The cell is constantly making new proteins and ncRNAs.

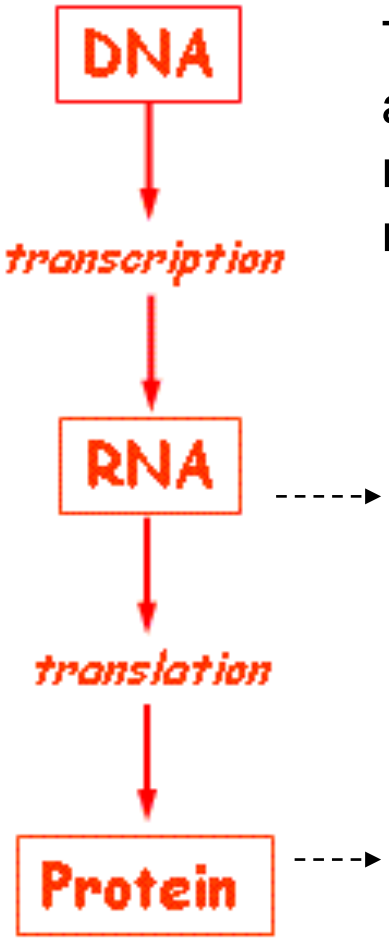
These perform their function for a while,

And are then degraded.

Newly made coding and non coding gene products take their place.

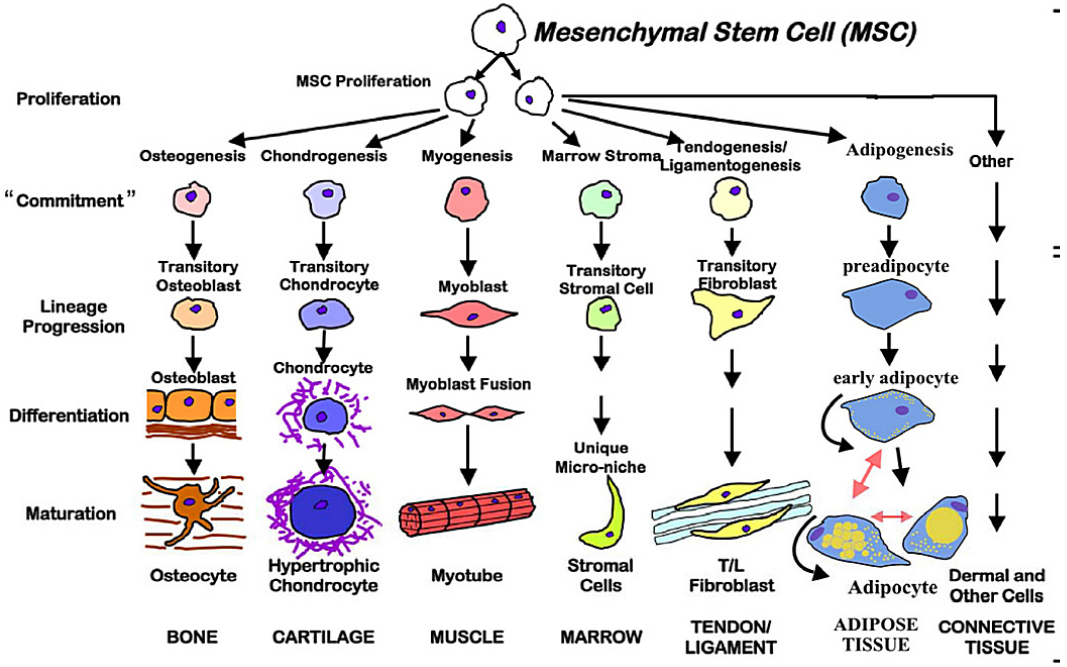
The picture within a cell is constantly “refreshing”.

Cell differentiation

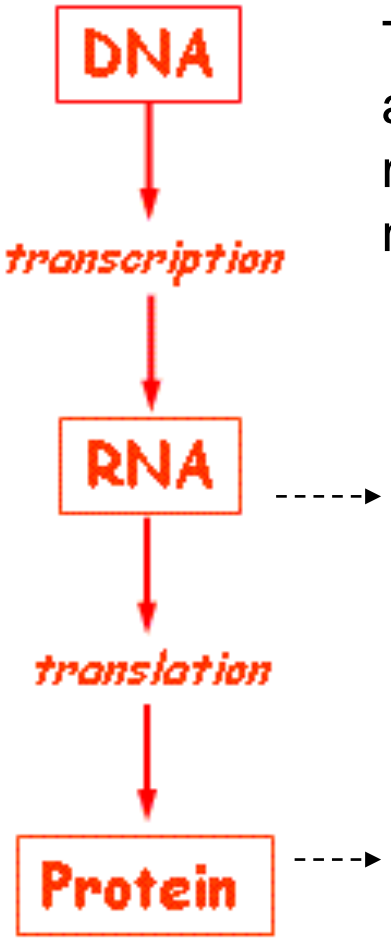


To change its behavior a cell can change the repertoire of genes and ncRNAs it makes.

That is exactly what happens when cells differentiate during development from stem cells to their different final fates.

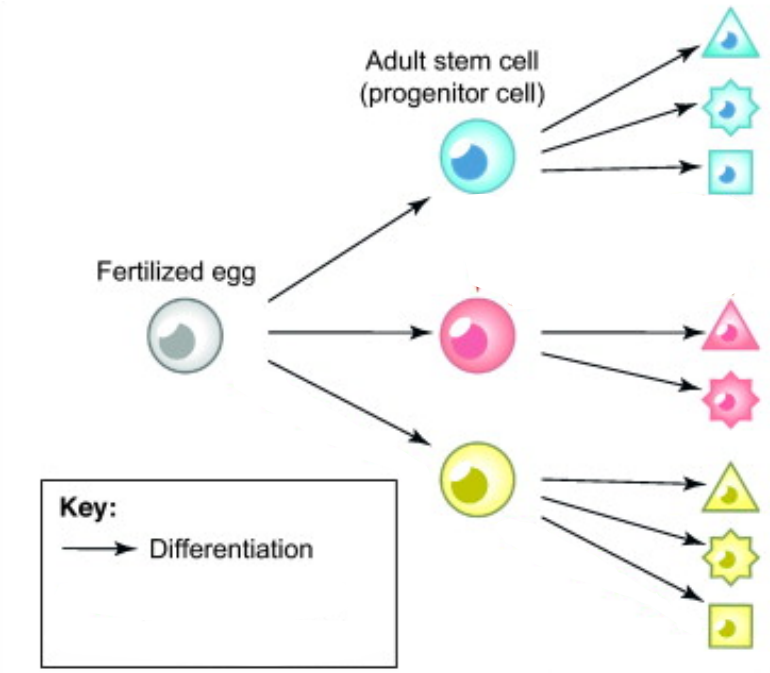


Human manipulation of cell fate

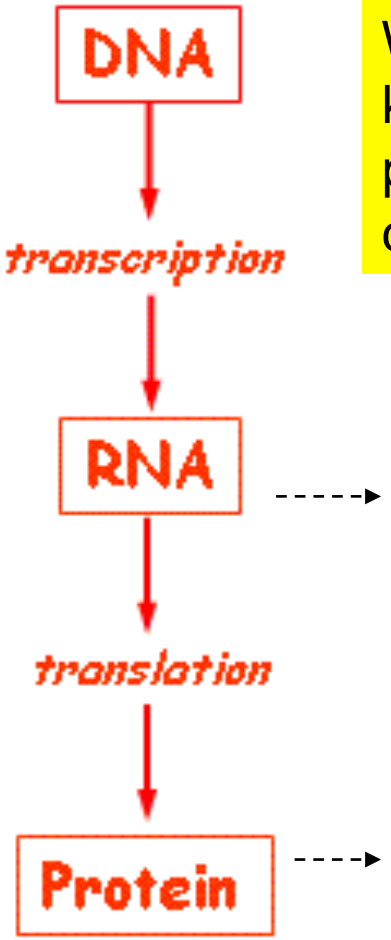


To change its behavior a cell can change the repertoire of genes and ncRNAs it makes.

We have learned (in a dish) to:
1 control differentiation
2 reverse differentiation
3 hop between different states

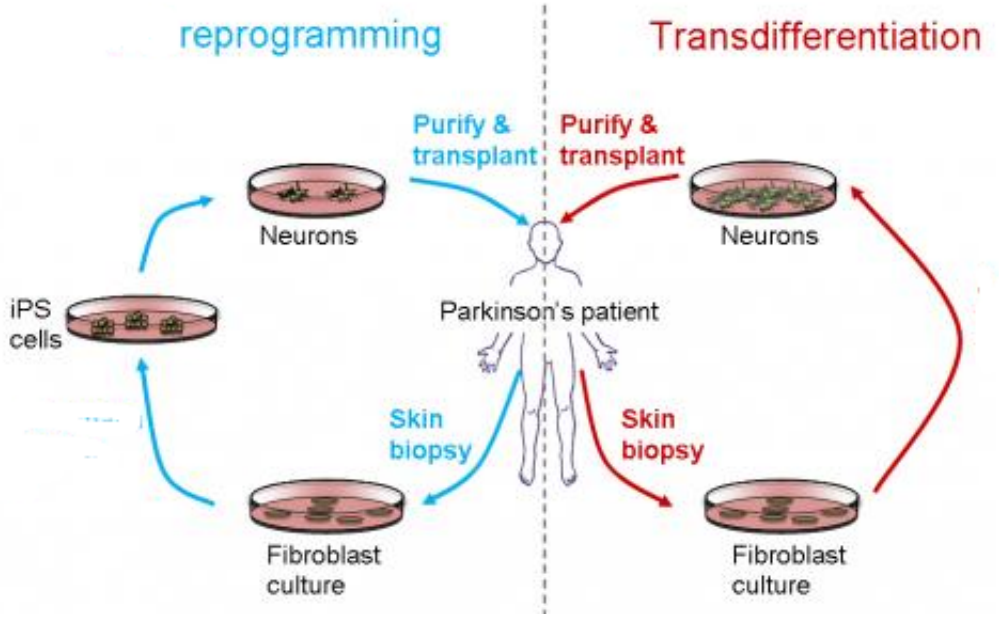


Cell replacement therapies



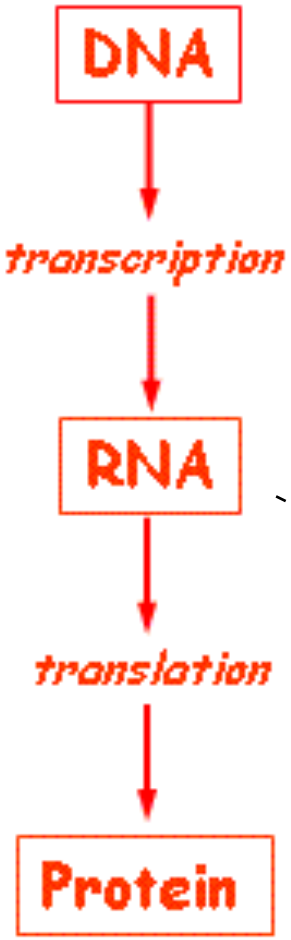
We want to use this knowledge to provide a patient with healthy self cells of a needed type.

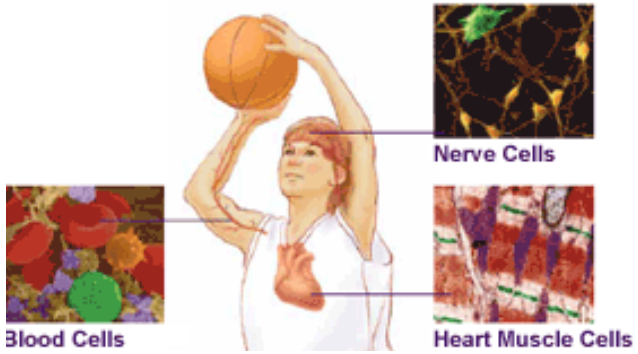
We have learned (in a dish) to:
1 control differentiation
2 reverse differentiation
3 hop between different states

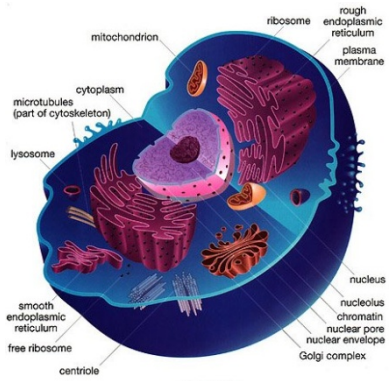


(iPS = induced pluripotent stem cells)

How does this happen?







Different cells in our body hold copies of (essentially) the same genome.

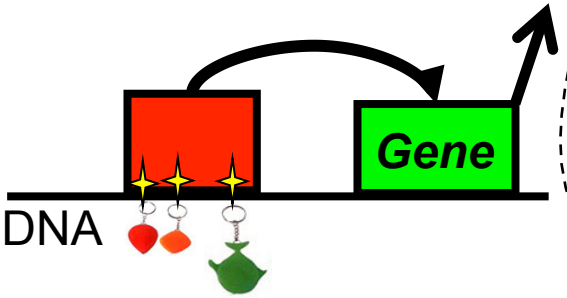
Yet they express *very* different repertoires of proteins and non-coding RNAs.

How do cells do it?

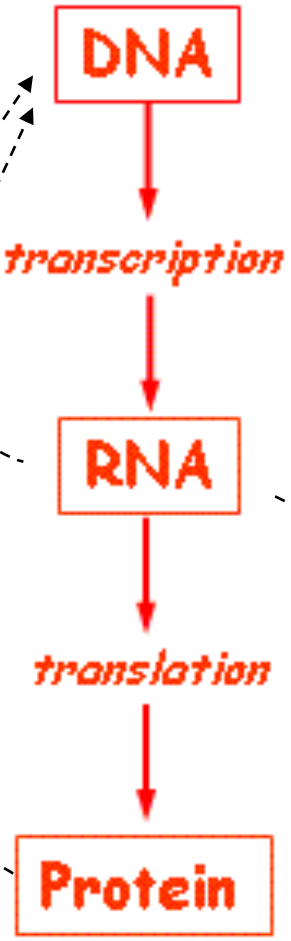
A: like they do everything else: using their proteins & ncRNAs…

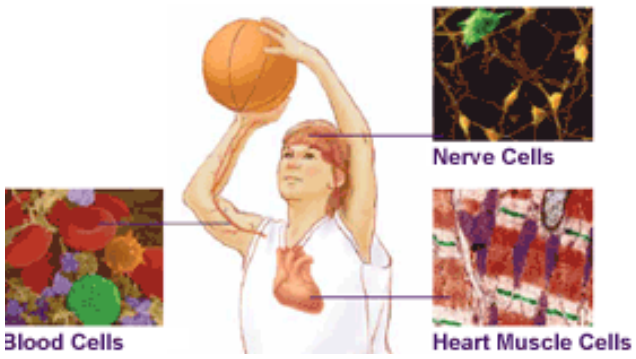
Gene Regulation

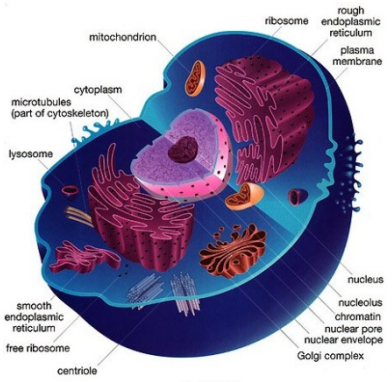
Some proteins and non coding RNAs go “back” to bind DNA near genes, turning these genes on and off.



Proteins







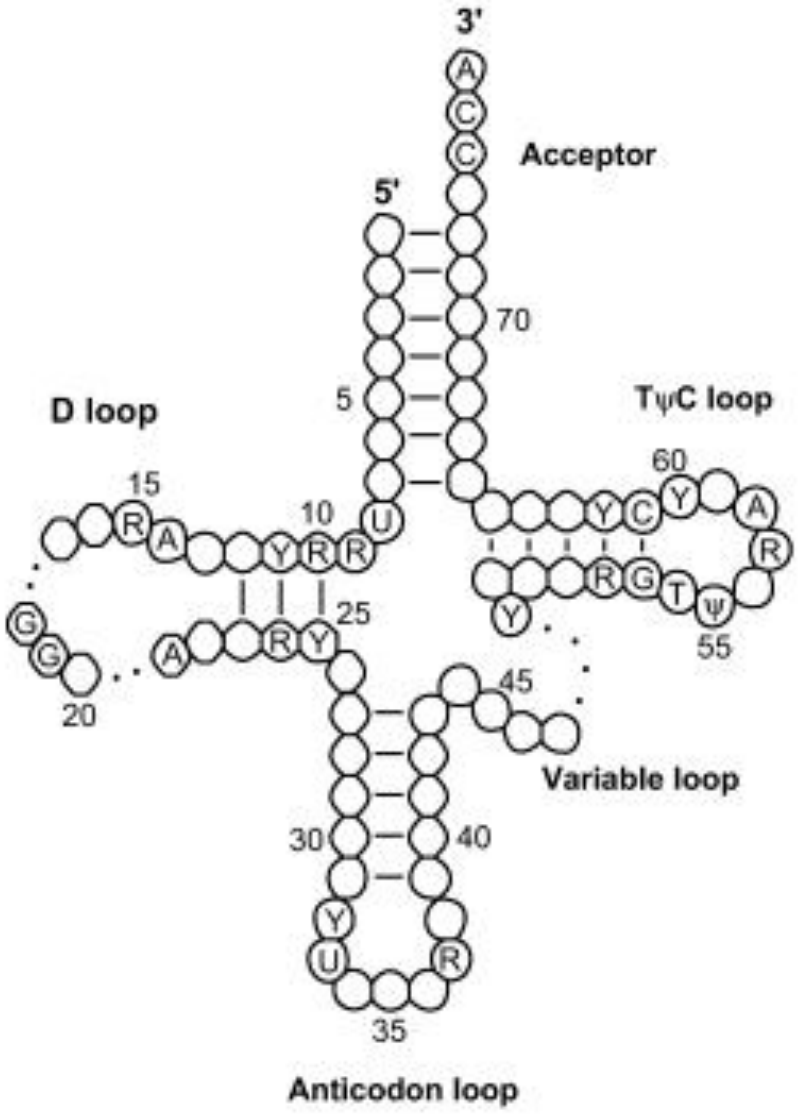
To be continued…

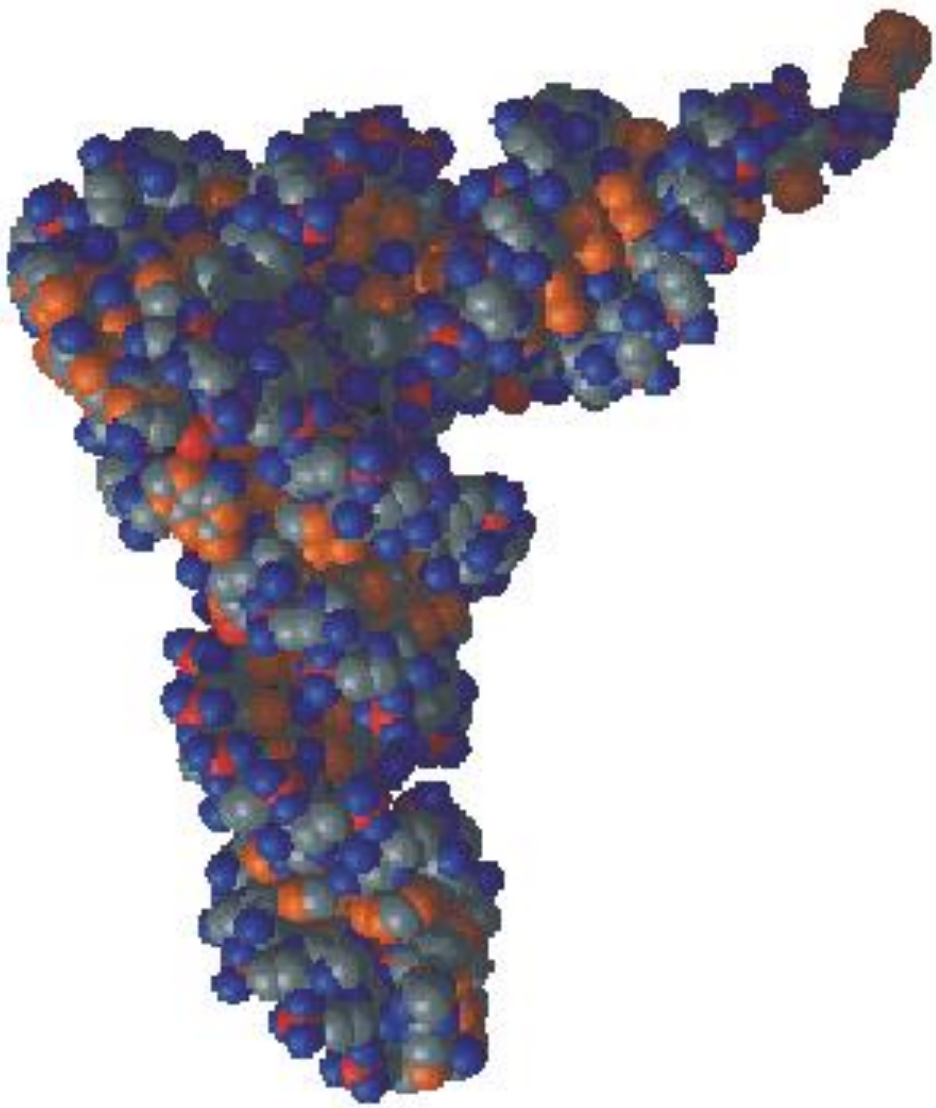
Review

- Central dogma recap
  - Genes, proteins and non coding RNAs
- RNA world hypothesis
- Small structural RNAs
  - Sequence, structure, function
  - Structure prediction
  - Transcription mode
- MicroRNAs
  - Functions
  - Modes of transcription
- lncRNAs
  - Xist
- Genome wide (and context wide) transcription
  - How much?
  - To what goals?
- Gene transcription and cell identity
  - Cell differentiation
  - Human manipulation of cell fates
- Gene regulation control

Algorithms

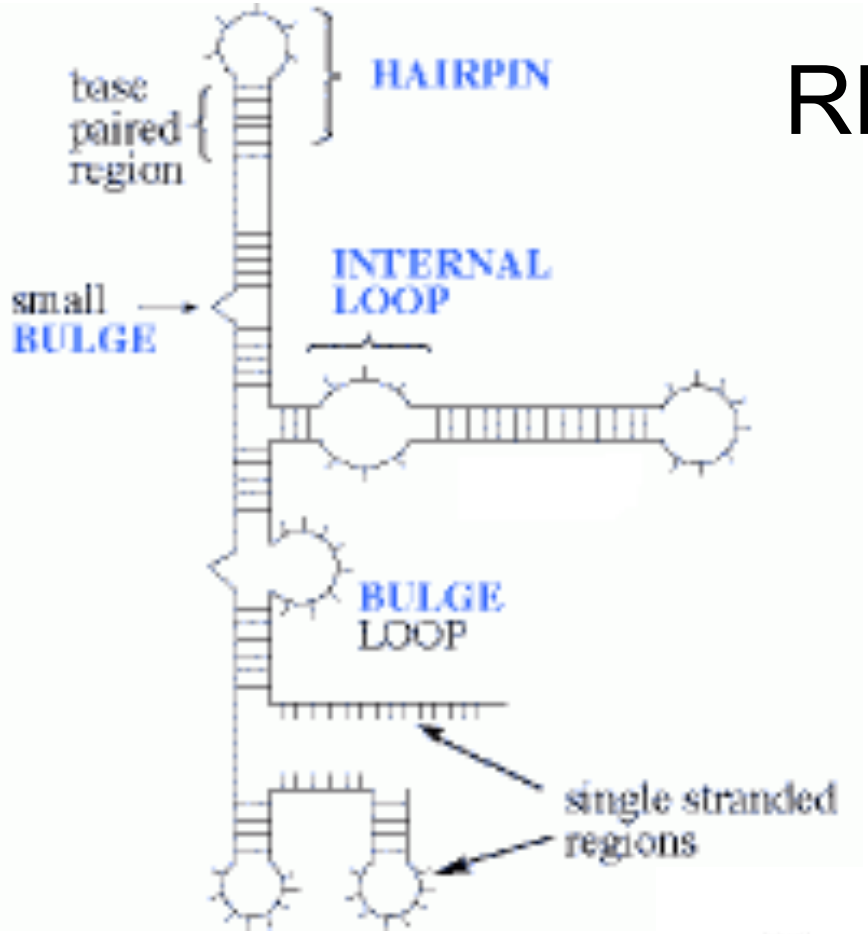
Typical transfer RNA structure

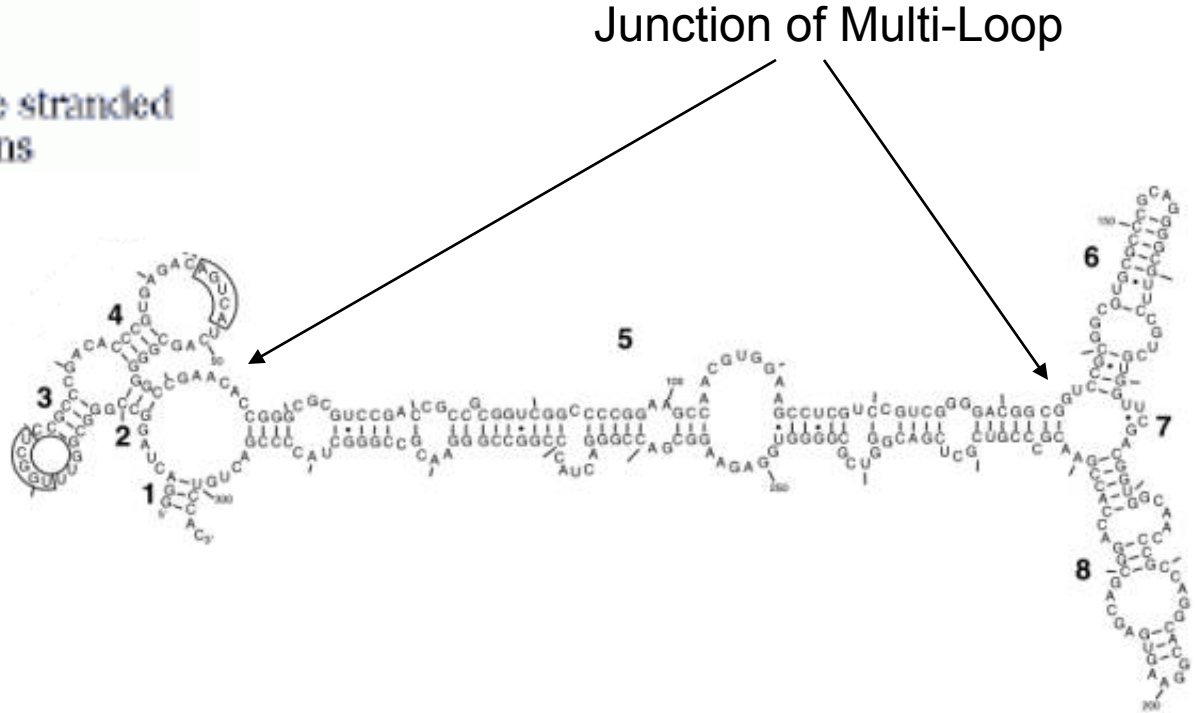




Jaco de Ridder, 2004

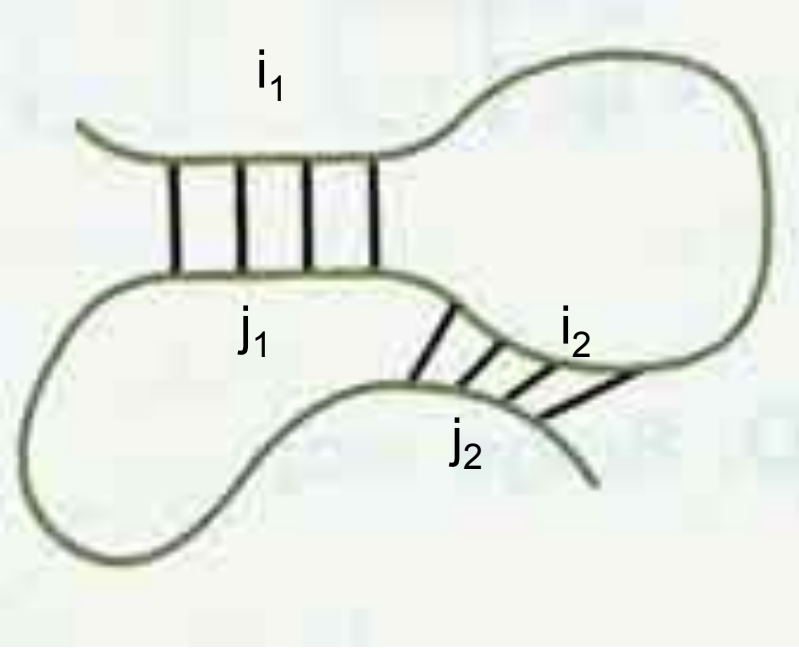
RNA Secondary Structure





Tertiary structure elements: Pseudoknots

Pseudoknot: interaction of bases inside a loop with bases outside the loop



$$i_1 < i_2 < j_1 < j_2$$

RNA Secondary Structure Stability

Structure stability is dependent upon:

1) The number of GC versus AU and GU base pairs (Higher energy bonds form more stable structures)

2) The number of base pairs in a stem region (longer stems results in more bonds)

3) The number of base pairs in a hairpin loop region (formation of loops with more than 10 or less than 5 bases requires more energy)

4) The number of unpaired bases, whether interior loops or bulges (unpaired bases decrease the stability of the structure)

Energy Table for Secondary Structure

- Free-energy values (kcal/mole at 37°C) are as follows:

	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

	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

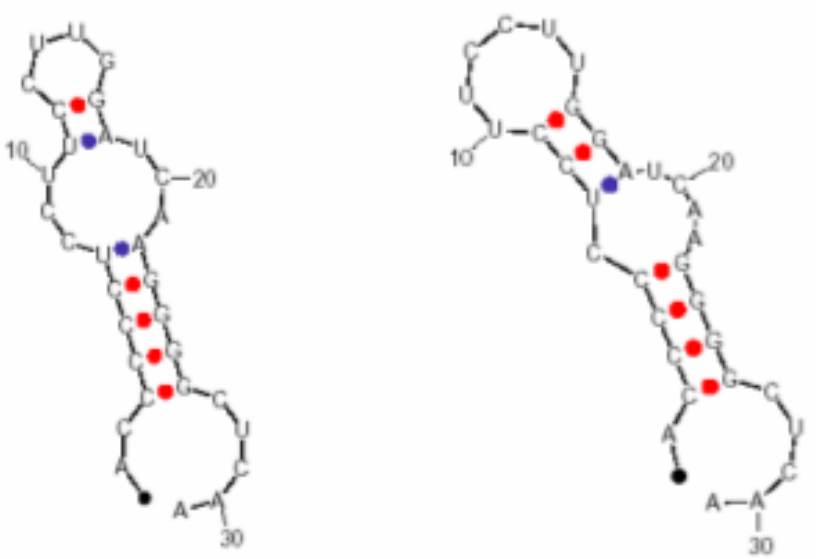
Sacha Baginsky, 2005

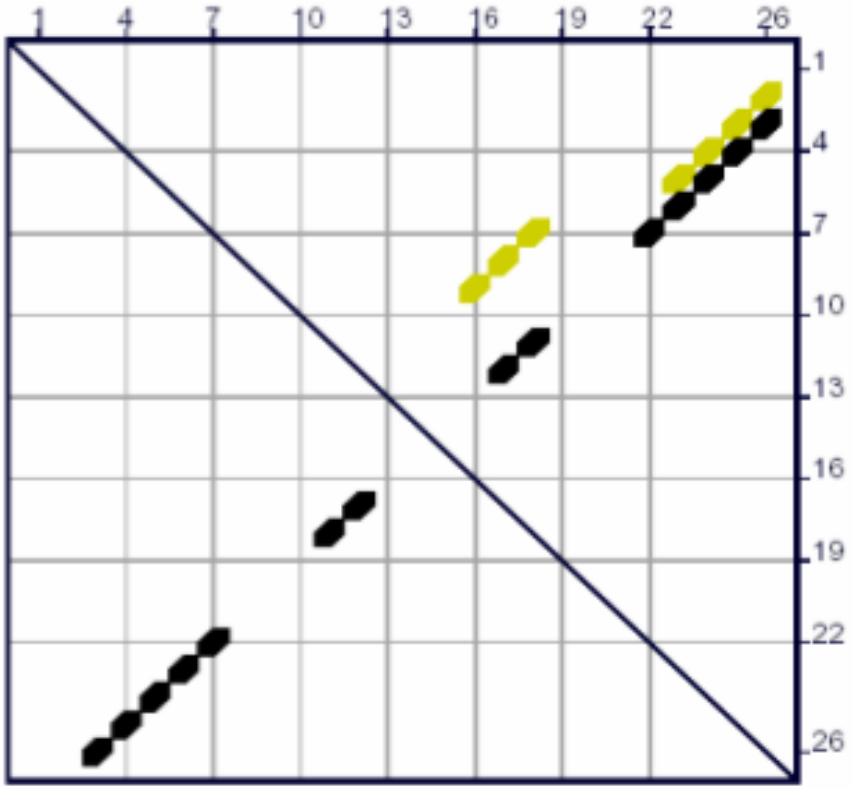
RNA Secondary Structure Prediction Approaches

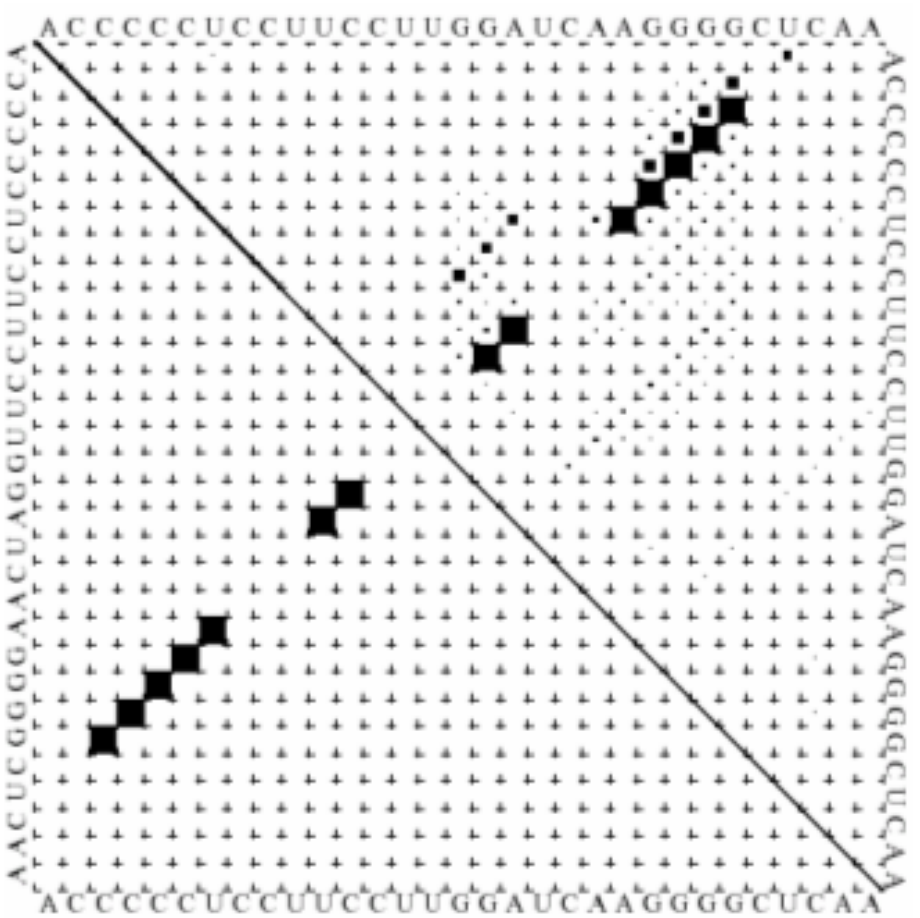
- **Minimum energy**: Look for folds with the lowest free energy (most stable) folds. The fold with more negative free energy, is more stable. The free energy of a fold is the addition of free energy of all motifs found in the structure. Require estimation of energy terms.
- **Comparative Method**: uses multiple sequence alignments of homologous sequences to find conserved regions and covariant base pairs (most trusted if there is enough data)
- Most methods predict secondary structure. Not successful for tertiary structure prediction, which is determined by X-ray and NMR.

Prediction Assumption of Energy-Based Method

- The most likely structure is similar to the energetically most stable structure
- Energy associated with any position in the structure is only influenced by local sequence and structure (previous pair, not next pair)
- No knots.







Jaco de Ridder, 2004

Dynamic programming algorithm

• Recursive definition of the optimal score (We use a simplified scoring function.)

•Initialization of optimal scoring matrix

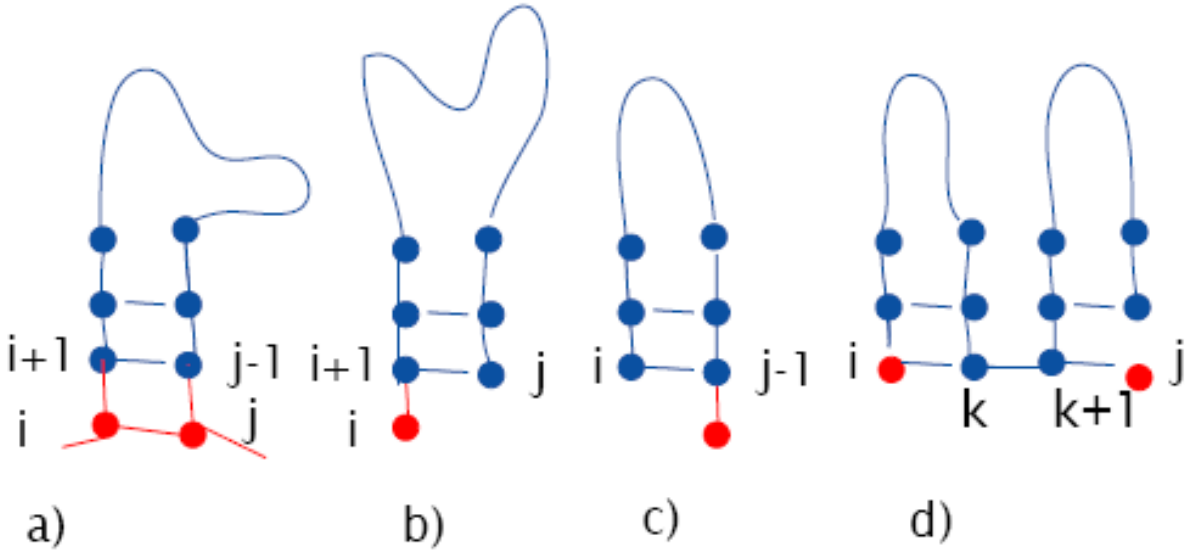
• Bottom-up approach to fill the scoring matrix (bottom-up because smallest subproblems are solved first). Run from diagonal to diagonal

• Traceback of the matrix to recover the global, optimal solution

Dynamic programming approach

Let $E(i,j)$ = minimum energy for subchain starting at i and ending at j

$\alpha(ri,rj)$ = energy of pair ri, rj (rj = base at position j)



a) i,j is paired $E(i,j) = E(i+1,j-1) + \alpha(ri,rj)$

b) i is unpaired $E(i,j) = E(i+1,j)$

c) j is unpaired $E(i,j) = E(i,j-1)$

d) bifurcation $E(i,j) = E(i,k)+E(k+1,j)$

Sacha Baginsky, 2005

Dynamic Programming Algorithm for RNA Secondary Structure Prediction

- Given a RNA sequence: $x_1,x_2,x_3,\ldots,x_L$ and a scoring function a(x,y)
- **Initialization**: E[i,i-1] = 0, E[i,i] = 0
- **Recursion**:

```
for d = 1,2,3,4,…,L-1
{
        for ( i = 1; i + d <= L; i++)
        {
                j = i + d;
                E[i,j] = min {      E[i+1,j],
                                    E[i,j-1],
                                    E[i+1,j-1] + a(r_i,r_j)
                                    min_{i<k<j} ( E[i,k] + E[k+1,j] )
                                }
        }
}
```

Note: i is always smaller than j.

Dynamic Programming Algorithm for RNA Secondary Structure Prediction

- Given a RNA sequence: $x_1,x_2,x_3,\ldots,x_L$ and a scoring function a(x,y)
- **Initialization**: E[i,i-1] = 0, E[i,i] = 0
- **Recursion**:

```
for d = 1,2,3,4,…,L-1
{
        for ( i = 1; i + d <= L; i++)
        {
                j = i + d;
                E[i,j] = min {      E[i+1,j],
                                    E[i,j-1],
                                    E[i+1,j-1] + a(r_i,r_j)
                                    min_{i<k<j} ( E[i,k] + E[k+1,j] )
                                }
        }
}
```

Time Complexity?

Note: i is always smaller than j.

A Simple Example

Input: GGAAAUCC

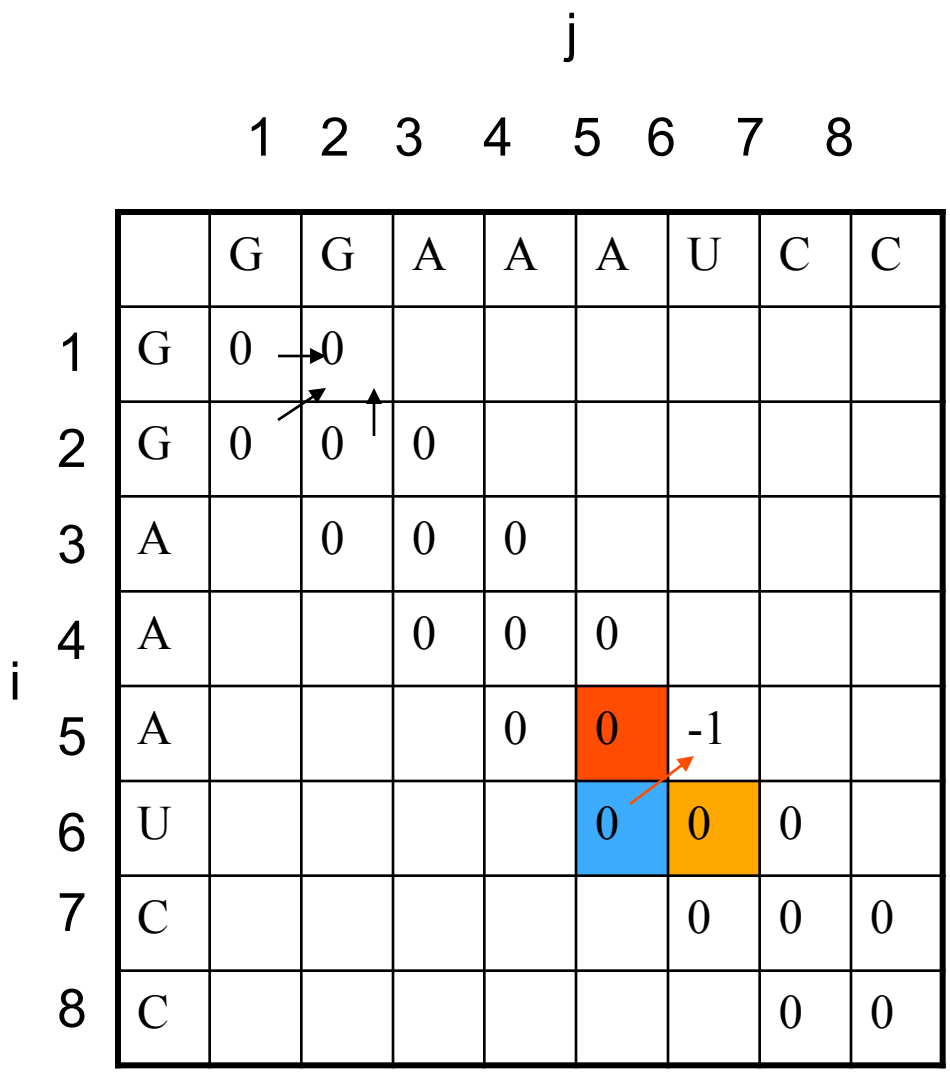
Scoring Function: $a(r_i,r_j) = -1$ if r_i and r_j form a Watson-Crick base pair. Otherwise, 0.

i \ j		1 G	2 G	3 A	4 A	5 A	6 U	7 C	8 C
1	G	0							
2	G	0	0						
3	A		0	0					
4	A			0	0				
5	A				0	0			
6	U					0	0		
7	C						0	0	
8	C							0	0

Initialization

E[i,i-1] = 0, E[i,i] = 0

Fill matrix from diagonal to diagonal



E[1,2] = min(
E(1,2-1),
E(1+1,2),
E[1+1,2-1]+a(1,2)
~~$\min_k(E[1,k]+E[k+1,2]$~~
)
= 0

E[5,6] = min(
E[5,5],
E[6,5]+a(5,6),
E[6,6])
~~$\min_k(E[5,k]+E[k,6])$~~
)
= -1

(may set to 0 if two Adjacent ones can't be paired?)

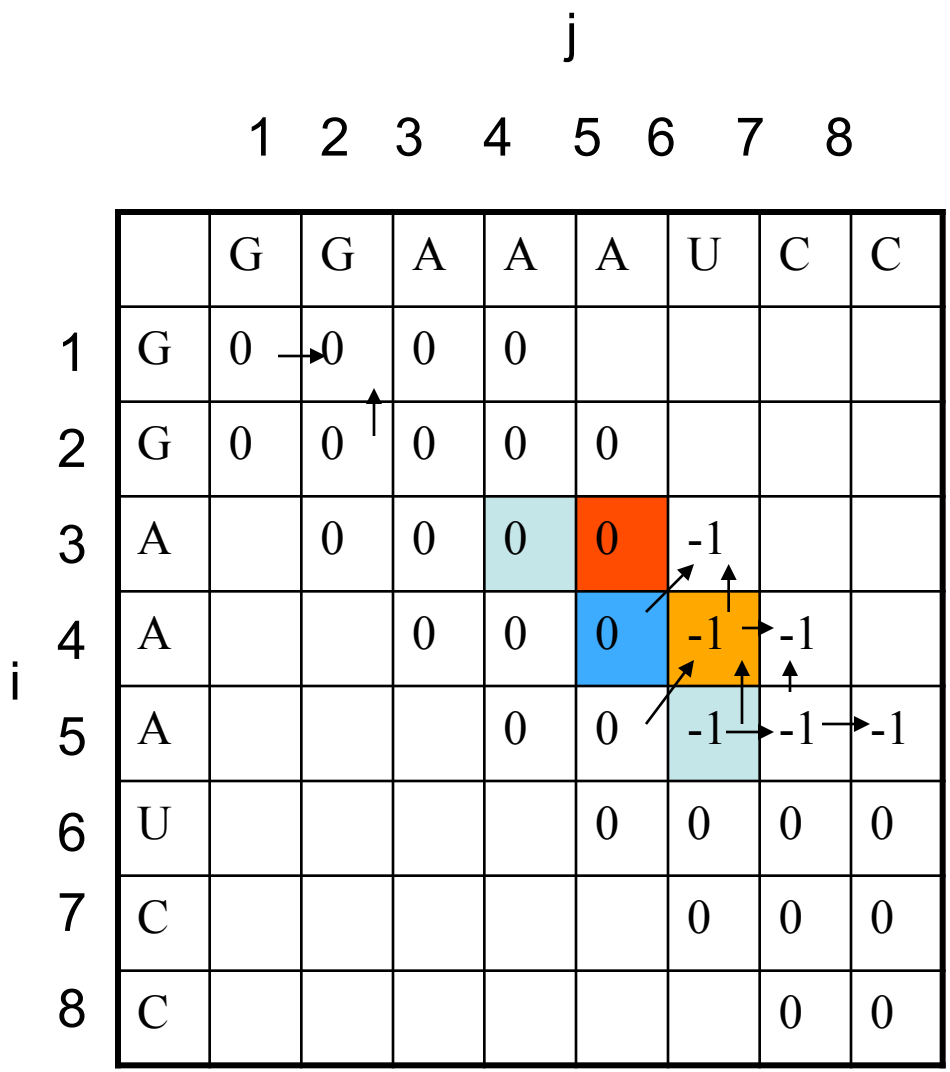
j

i		1 G	2 G	3 A	4 A	5 A	6 U	7 C	8 C
1	G	0	0	0					
2	G	0	0	0	0				
3	A		0	0	0	0			
4	A			0	0	0	-1		
5	A				0	0	-1	-1	
6	U					0	0	0	0
7	C						0	0	0
8	C							0	0

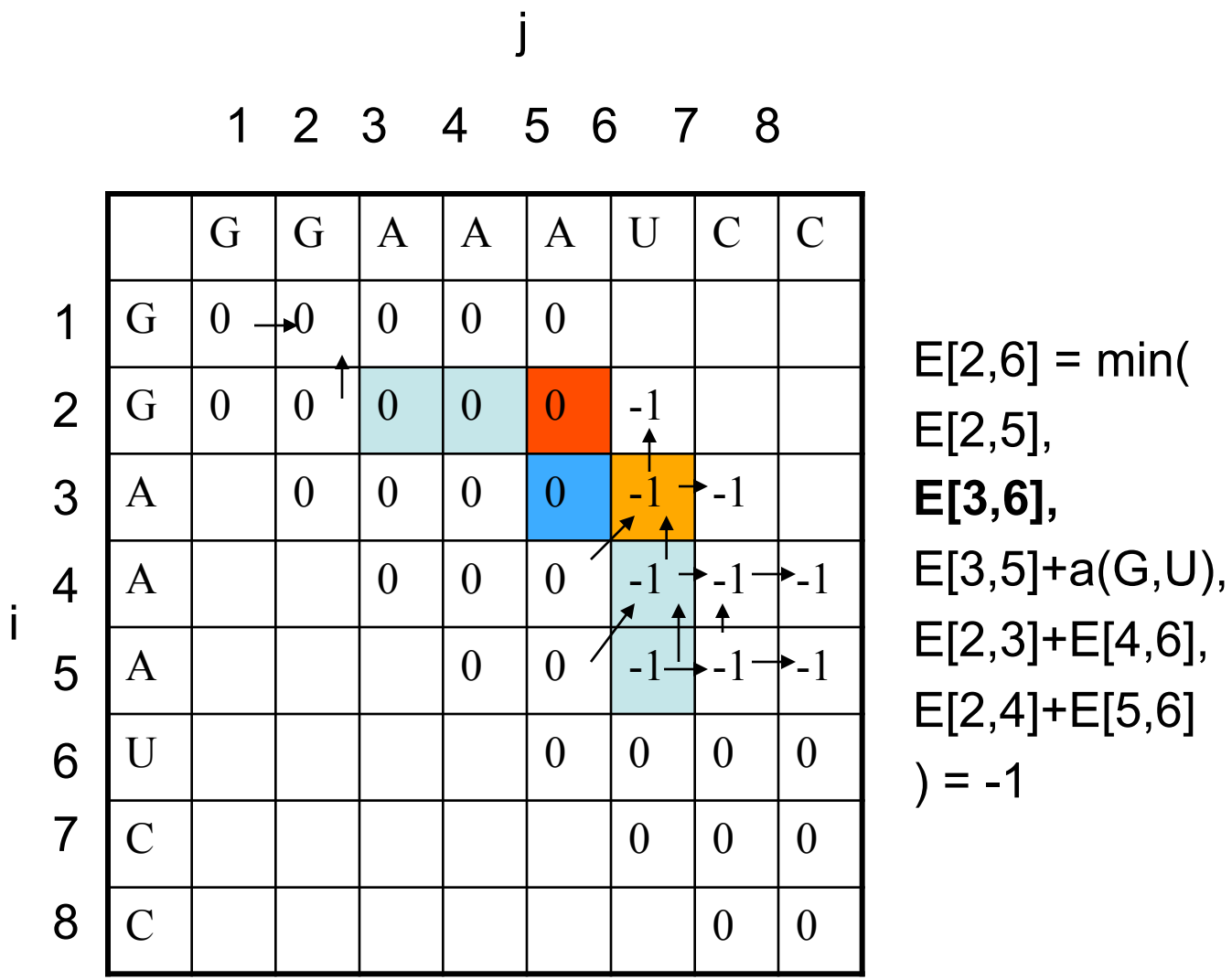
E[1,3] = min(
E(1,2),
E(2,3),
E[2,2] + a(G,A)
)
= 0

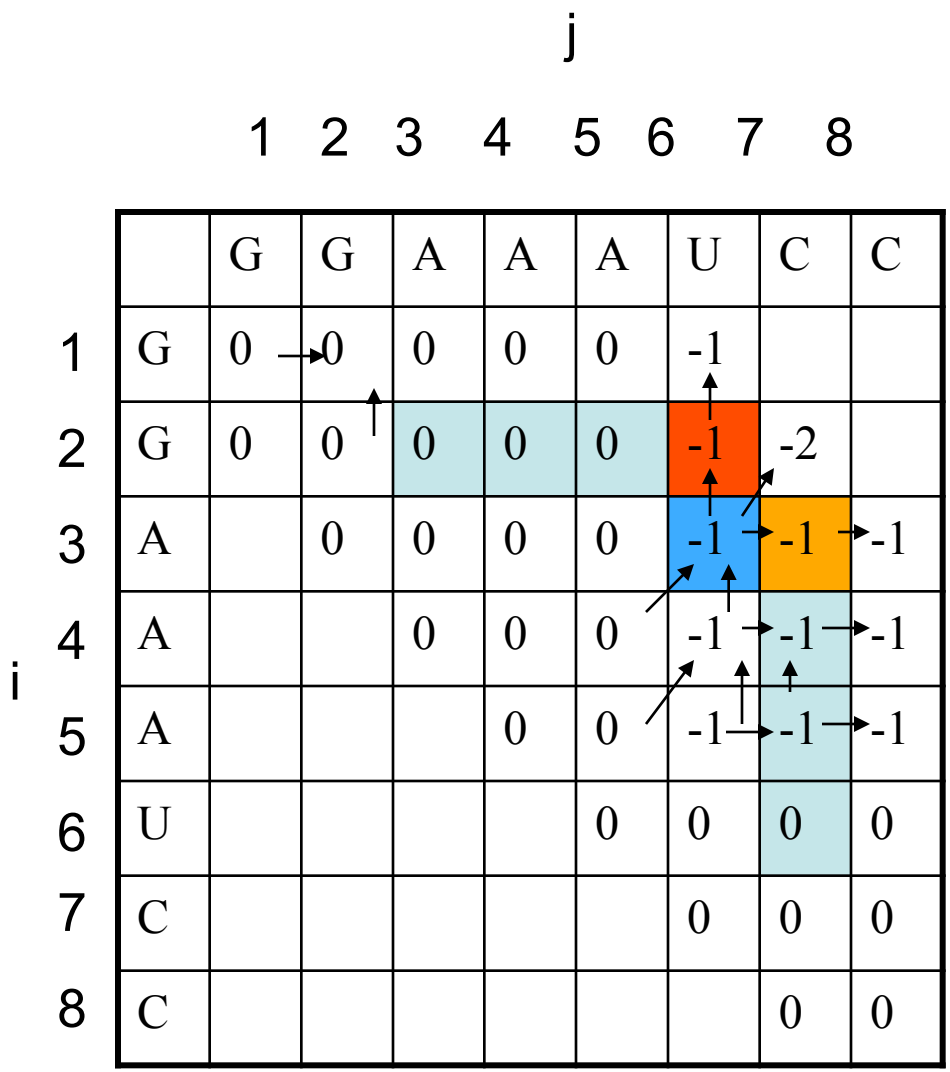
Any valid k?

E[4,6] = min (
E[4,5],
E[5,6],
E[5,5] + a(A,U)
)
= -1



E[3,6] = min(
E[3,5],
E[4,6],
E[4,5]+a(A,U),
E[3,4]+E[5,6]
)
= -1


j
1 2 3 4 5 6 7 8
i
1 2 3 4 5 6 7 8
G G A A A U C C
G 0 0 0 0 0
G 0 0 0 0 0 -1
A 0 0 0 0 0 -1 -1
A 0 0 0 0 -1 -1 -1
A 0 0 0 -1 -1 -1
U 0 0 0 0
C 0 0 0
C 0 0
E[2,6] = min(
E[2,5],
E[3,6],
E[3,5]+a(G,U),
E[2,3]+E[4,6],
E[2,4]+E[5,6]
) = -1



E[2,7] = min (

E[2,6],

E[3,7],

E[3,6]+a(G,C),

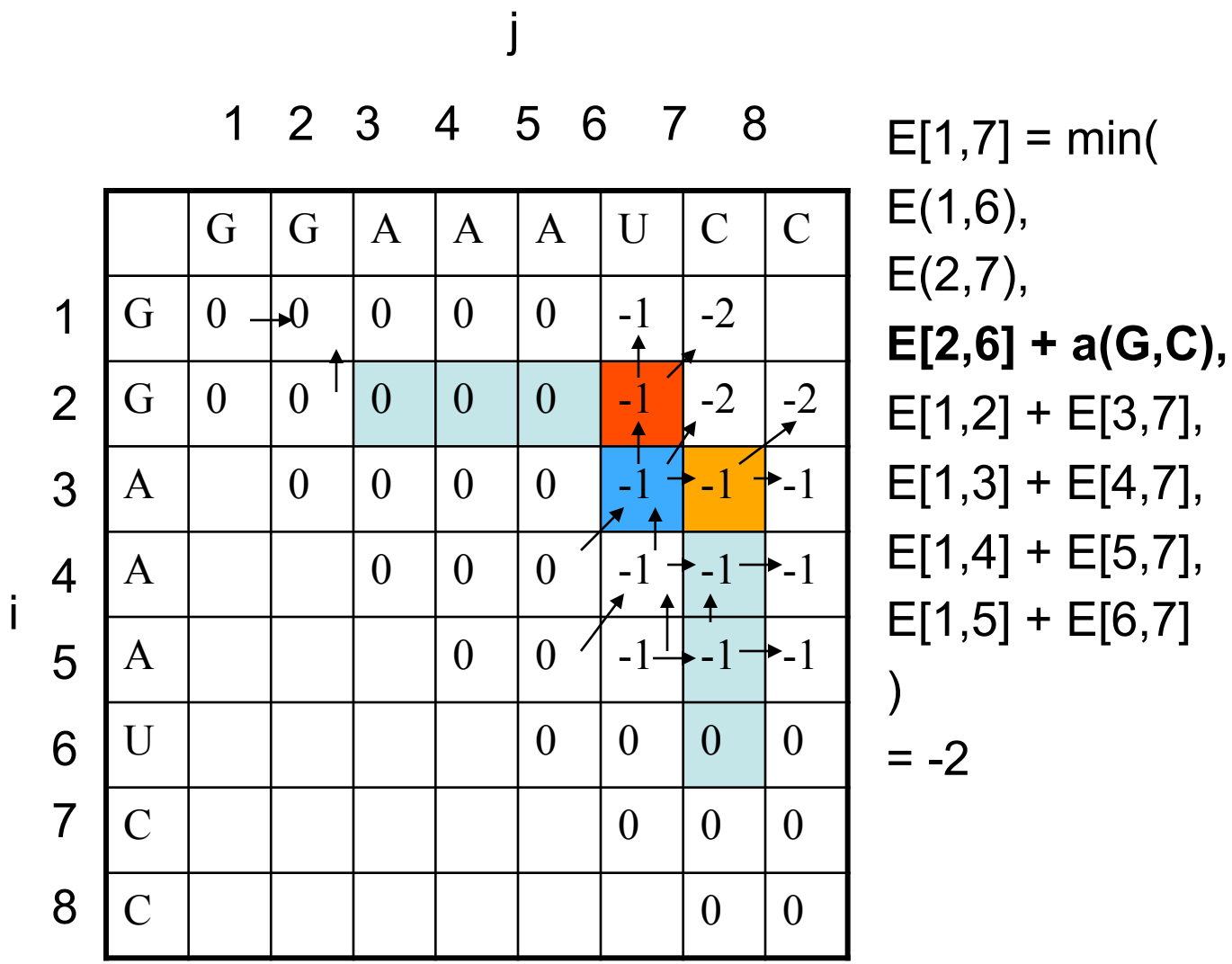
E[2,3] + E[4,7],

E[2,4] + E[5,7],

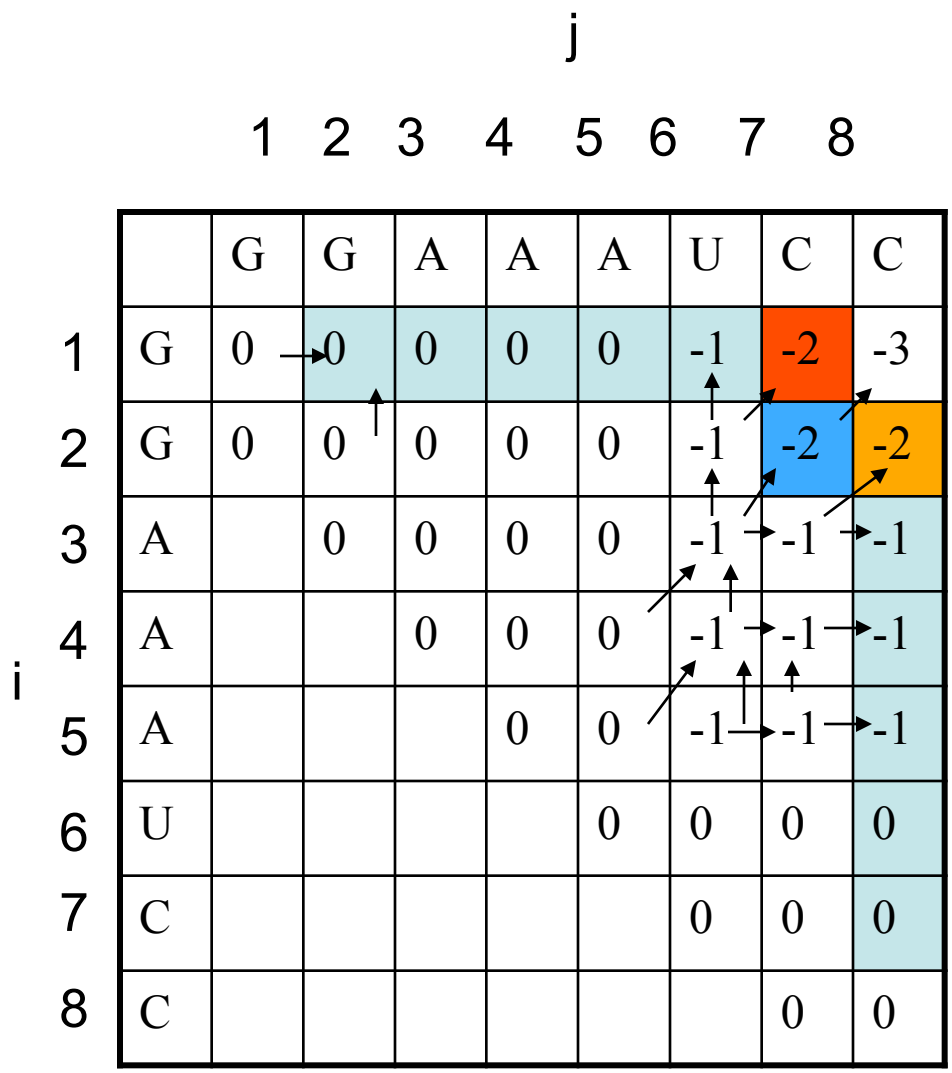
E[2,5] + E[6,7]

)

= -2



E[1,7] = min(
E(1,6),
E(2,7),
E[2,6] + a(G,C),
E[1,2] + E[3,7],
E[1,3] + E[4,7],
E[1,4] + E[5,7],
E[1,5] + E[6,7]
)
= -2



Best score:

E[1,8] = E[2,7] + a(G,C)

= -3

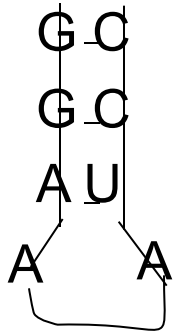
Time Complexity?

Trace Back

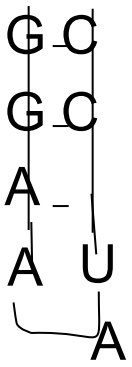
i \ j		1 G	2 G	3 A	4 A	5 A	6 U	7 C	8 C
1	G	0	0	0	0	0	-1	-2	-3
2	G	0	0	0	0	0	-1	-2	-2
3	A		0	0	0	0	-1	-1	-1
4	A			0	0	0	-1	-1	-1
5	A				0	0	-1	-1	-1
6	U					0	0	0	0
7	C						0	0	0
8	C							0	0

Best score: E[1,8] = -3

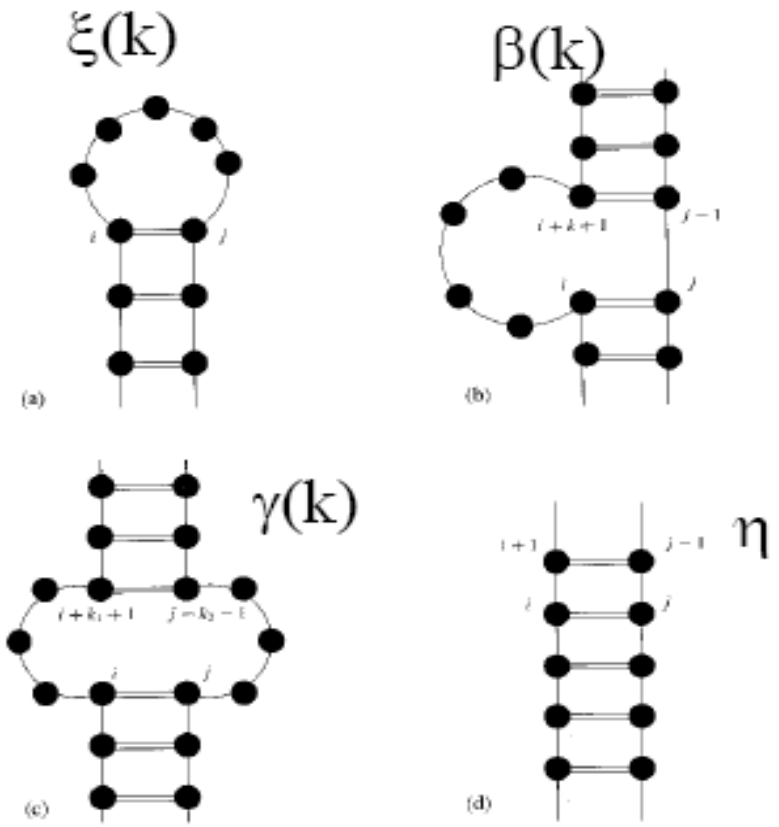
Path 1



Path 2



Even more realistic energy function



Loops have destabilizing effect structure (d) should have lower energy that (b).

Destabilizing contribution of loops should depend on the loop length (k).

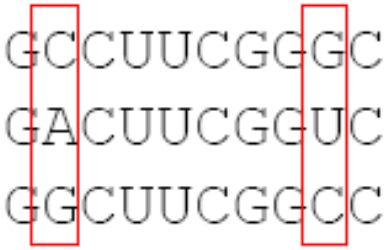
Stacking has additional stabilizing contribution η.

So in reality, more realistic energy function that considers different loops are needed. But the basic idea of dynamic programming is still applied.

Sacha Baginsky, 2005

Covariance method

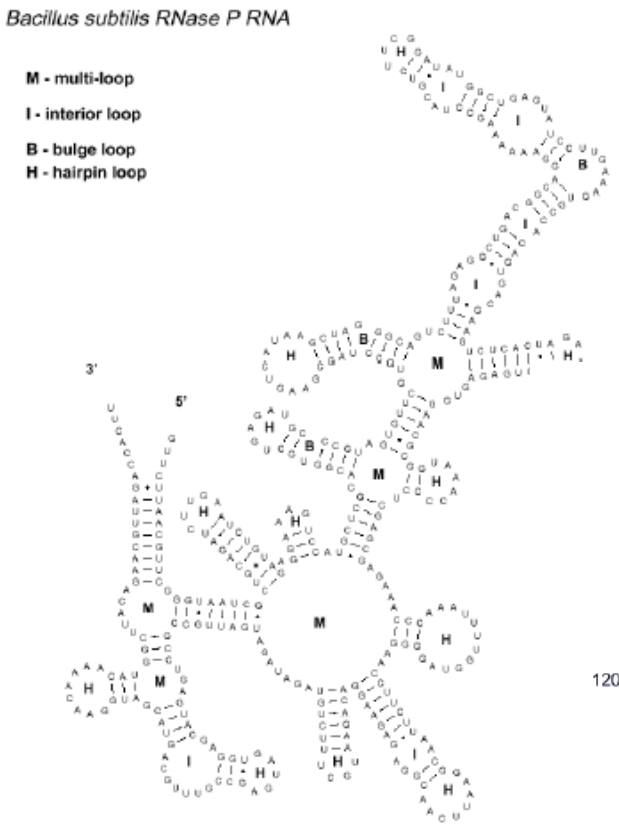
In a correct multiple alignment RNAs, conserved base pairs are often revealed by the presence of frequent correlated compensatory mutations.

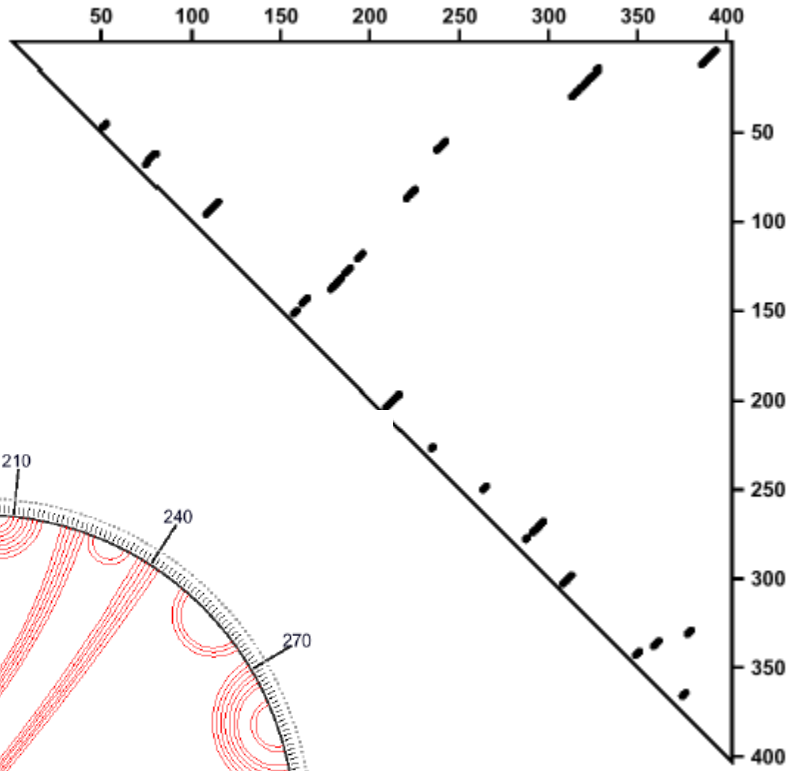


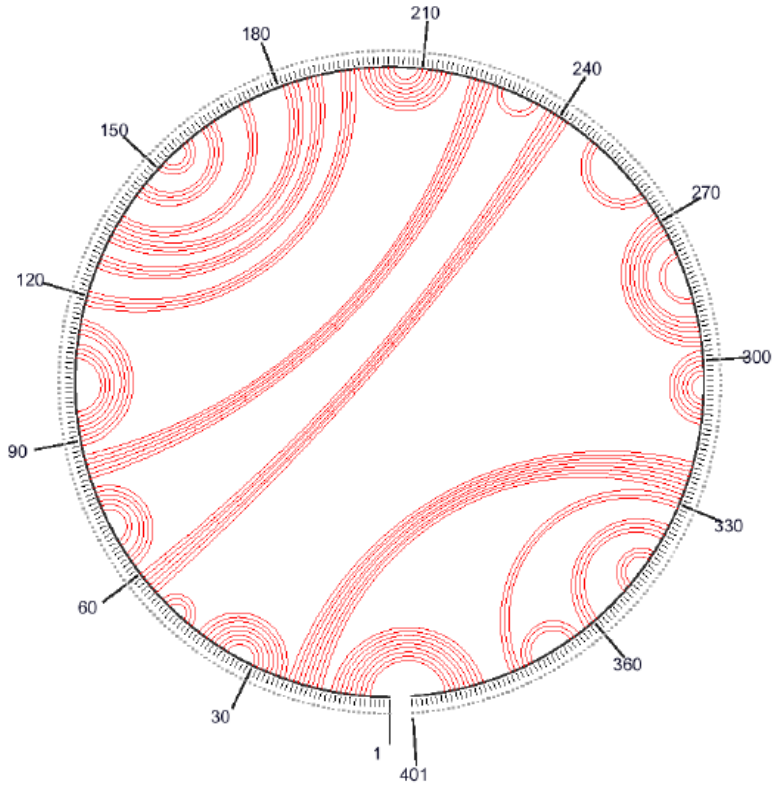
Two boxed positions are covarying to maintain Watson-Crick complementary. This covariation implies a base pair which may then be extended in both directions.

More information: www.rna.icmb.utexas.edu/METHODS/menu.html

Representation of RNA secondary structures

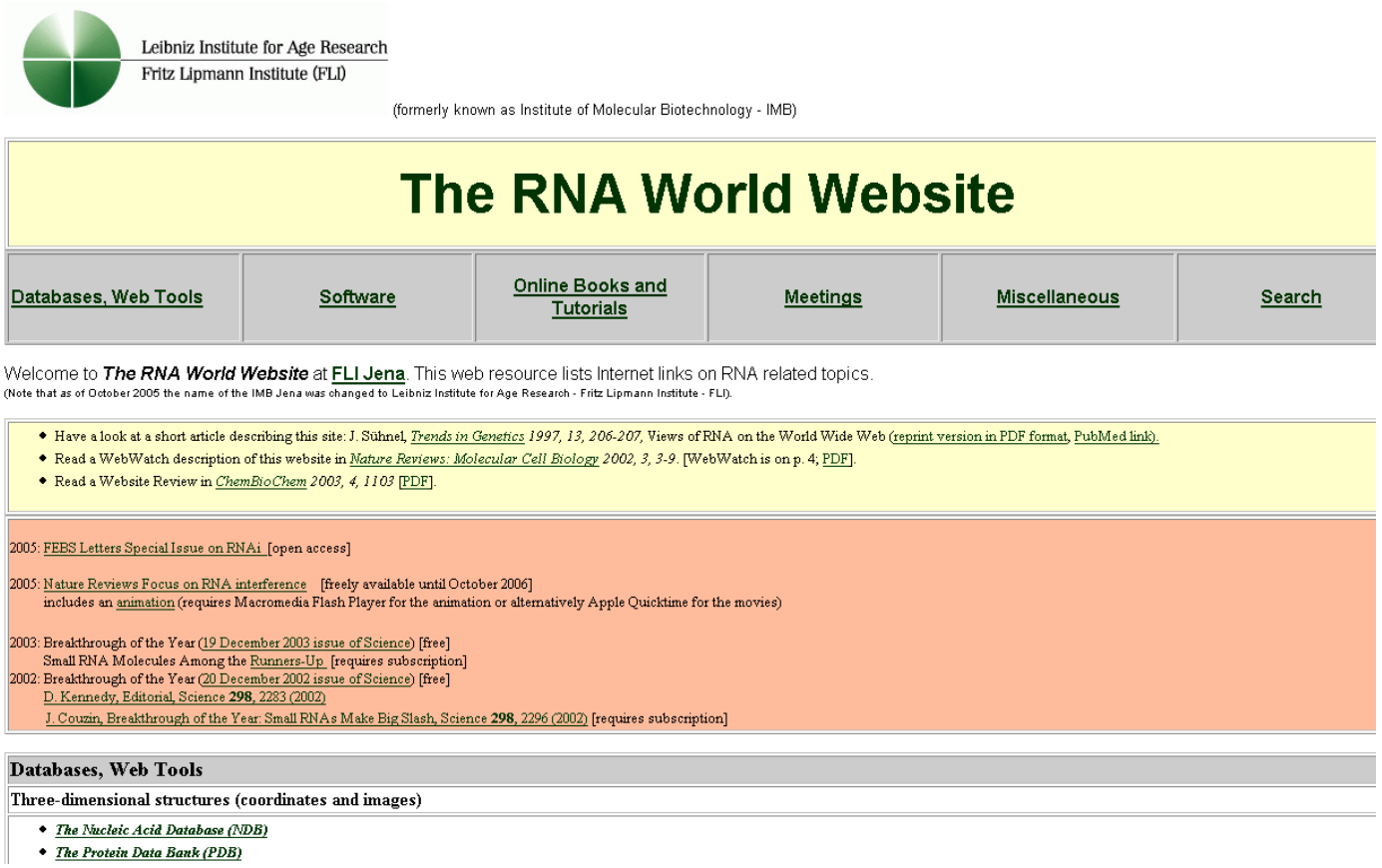






RNA Resources

Web: http://www.imb-jena.de/RNA.html



Leibniz Institute for Age Research
Fritz Lipmann Institute (FLI)

(formerly known as Institute of Molecular Biotechnology - IMB)

The RNA World Website

Databases, Web Tools	Software	Online Books and Tutorials	Meetings	Miscellaneous	Search

Welcome to ***The RNA World Website*** at FLI Jena. This web resource lists Internet links on RNA related topics.
(Note that as of October 2005 the name of the IMB Jena was changed to Leibniz Institute for Age Research - Fritz Lipmann Institute - FLI).

- Have a look at a short article describing this site: J. Sühnel, *Trends in Genetics 1997, 13, 206-207*, Views of RNA on the World Wide Web (reprint version in PDF format, PubMed link).
- Read a WebWatch description of this website in *Nature Reviews: Molecular Cell Biology 2002, 3, 3-9*. [WebWatch is on p. 4; PDF].
- Read a Website Review in *ChemBioChem 2003, 4, 1103* [PDF].

2005: FEBS Letters Special Issue on RNAi [open access]

2005: Nature Reviews Focus on RNA interference [freely available until October 2006]
includes an animation (requires Macromedia Flash Player for the animation or alternatively Apple Quicktime for the movies)

2003: Breakthrough of the Year (19 December 2003 issue of Science) [free]
Small RNA Molecules Among the Runners-Up [requires subscription]
2002: Breakthrough of the Year (20 December 2002 issue of Science) [free]
D. Kennedy, Editorial, Science **298**, 2283 (2002)
J. Couzin, Breakthrough of the Year: Small RNAs Make Big Slash, Science **298**, 2296 (2002) [requires subscription]

Databases, Web Tools

Three-dimensional structures (coordinates and images)

- ***The Nucleic Acid Database (NDB)***
- ***The Protein Data Bank (PDB)***

RNA Folding Software

- Vienna: http://www.tbi.univie.ac.at/RNA/
- MFold: http://www.bioinfo.rpi.edu/applications/mfold/rna/form1.cgi
- AliFold: http://rna.tbi.univie.ac.at/cgi-bin/alifold.cgi (use aligned sequences)
- Genebee: http://www.genebee.msu.su/services/rna2_reduced.html (alignment)

Summary of the Course
Topic 1 – Template-Based Modeling

- Template-Based Protein Modeling
- Probability of features conditioned on template
- X, Y, Z representation
- Maximum Likelihood
- Gradient Descent
- Conjugate Gradient Descent

Topic 2 – Template-Free Modeling

- Template-Free Protein Modeling
- Prior probability: Prob(structure)
- Prob (sequence | structure): $P(aa_i \mid E_i)$ or $P(aa_i, aa_j \mid E_i, E_j)$
- Maximum a Posterior
- Fragment assembly
- Simulated Annealing

Topic 3 – Protein Docking

- Protein Docking
- Shape complementarity
- Geometric approach
- Energy-based approach
- Flexible docking
- Evaluation methods
- Fourier Transformation
- Geometric graph method
- Energy optimization
- Side chain and backbone flexibility

Topic 4 - Genome Structure Modeling

- 3D Genome Structure
- Chromosome Conformation Capturing
- Hi-C technique
- Inter- / intra chromosome contact map
- Chromosomal translocation
- Interaction frequency and distance conversion
- Markov Chain Monte Carlo method
- Initialization
- Probability of a conformation
- Conformation Perturbation
- Mixing detection
- Evaluation

Topic 5 – RNA Structure Modeling

- Message RNA
- Non-coding RNA
- No-coding structural RNA (e.g. tRNA)
- microRNA
- Non-coding long RNA
- RNA Secondary Structure modeling
- Energy or scoring function
- Dynamic programming
- Context free grammar approach

Discussion of Final Presentation and Final Report

- Presentation
- Power point slides
- Introduction
- All four projects
- Conclusion

- Report
- Title, abstract (0.5 page)
- Introduction (~1 page)
- Methods (?)
- Results and discussions (?)
- Conclusion (~0.5 page)
- Bibliography
- May 14

Other Related Courses

- Structural Bioinformatics by Dr. Korkin
- Computational Optimization Methods by me.