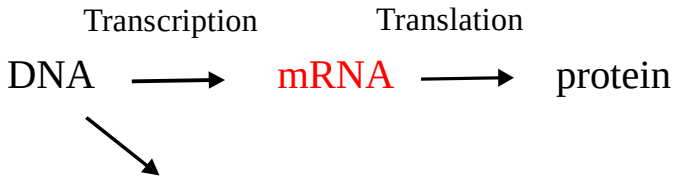
RNA bioinformatics

Two major problems are addressed:

* How do we find RNA genes computationally?
* How do we predict the secondary structure of RNAs?

Tore Samuelsson Oct 2005



Non-coding RNAs (ncRNAs)

tRNA	(adaptor in translation)
rRNA	(component of ribosome)
snRNA	(plays important role in splicing)
snoRNA	(takes part in processing of rRNA)
RNAse P	(ribozyme processing tRNA precursors)
SRP RNA	(component of signal recognition particle)
and others	

The RNA world

Early in the evolution of life RNA molecules played an important role, and they were to a large extent independent of protein and DNA.

RNA molecules have two important properties to make this possible:

* they are able to carry genetic information
* they are able to catalyse chemical reactions

Many RNAs today are “molecular fossils” from the RNA world.

For instance, a number of important cellular processes such as

splicing &
protein synthesis

are catalyzed by RNA molecules (ribozymes)

RNA molecules, unlike DNA molecules, are single-stranded

DNA

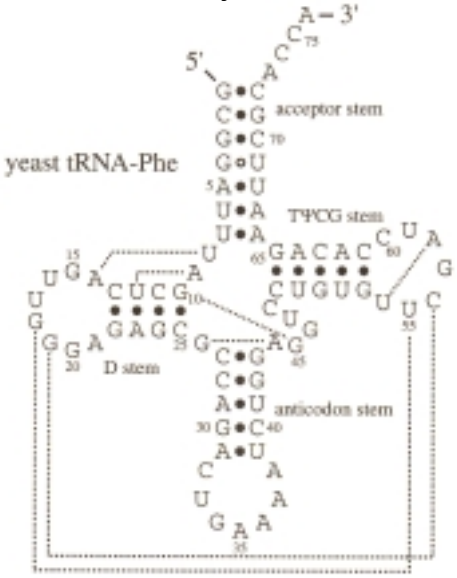
5′ GCCAAGGTTCGAAA 3′
3′ CGGTTCCAAGCTTT 5′

RNA

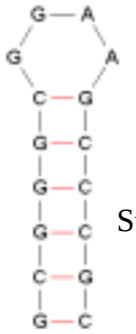
5′ GCCAAGGUUCGAAA 3′

… but through internal base-pairing helical structures occur like those in the double-stranded DNA .

Each family of ncRNA typically adopts a characteristic secondary structure



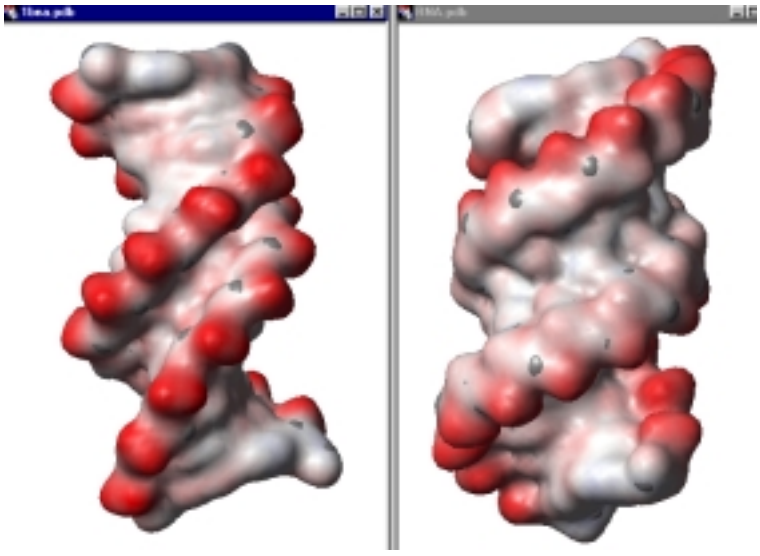
Hairpin

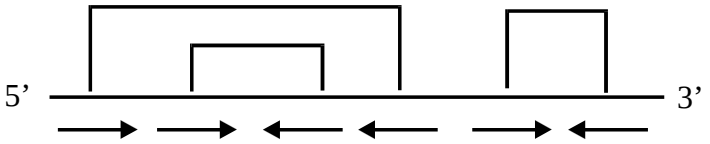


Loop

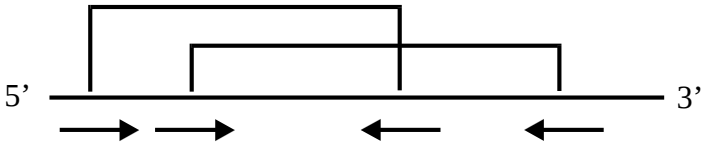
Stem

DNA **RNA**





“Normal” nested base-pairing



Pseudoknot

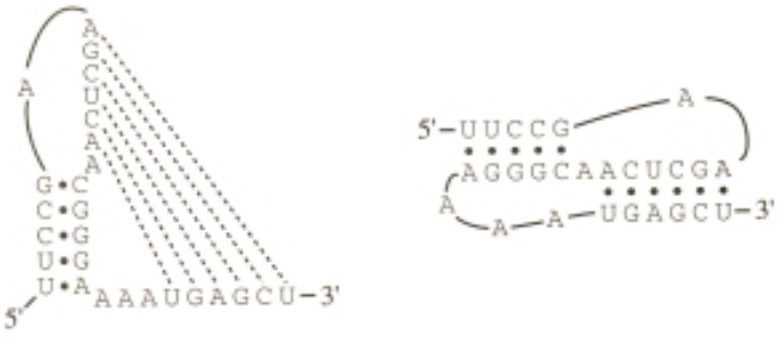
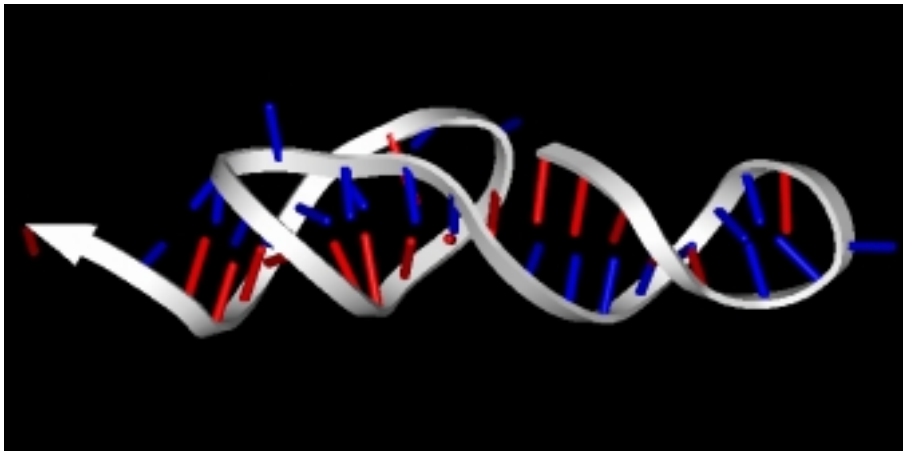
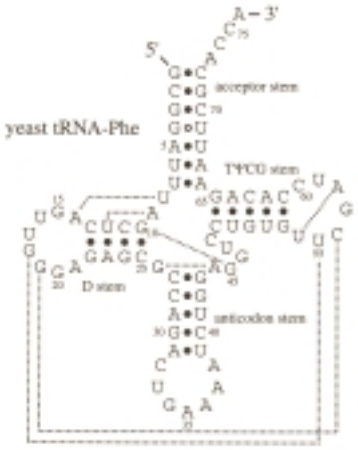
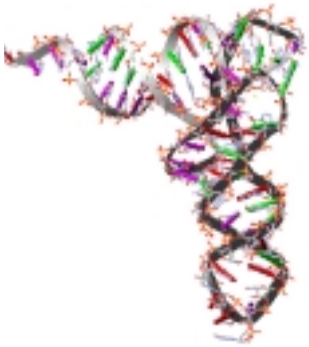


Figure 10.3 *Base pairs between a loop and positions outside the enclosing stem are called a pseudoknot (left). Another representation of the same pseudoknot is shown on the right. In three-dimensional space, the two stems can stack coaxially and mimic a contiguous A-form helix. This particular example is an artificially selected RNA inhibitor of the human immunodeficiency virus reverse transcriptase [Tuerk, MacDougal & Gold 1992].*

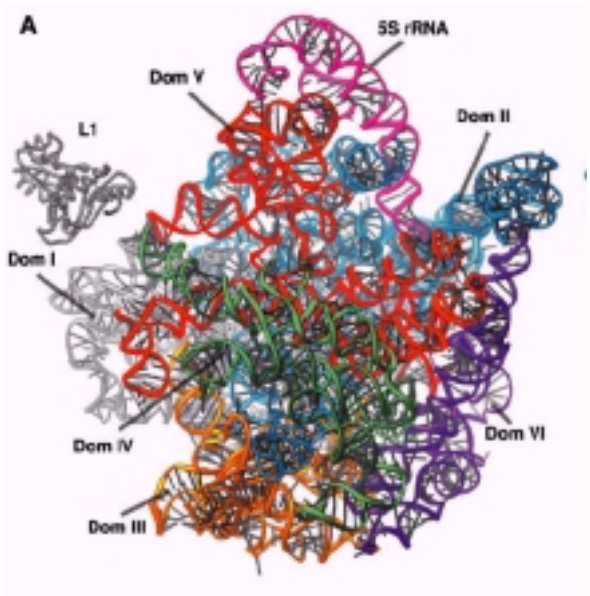
Pseudoknot 3D structure

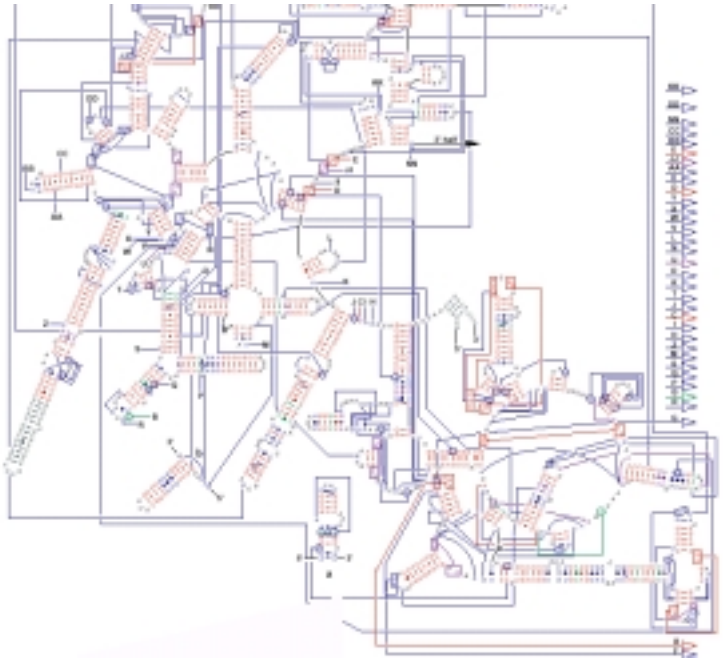






50S subunit of ribosome
RNA 2900 nucleotides





Also mRNAs carry functional secondary structures

1) RNA secondary structure as a transcription termination signal

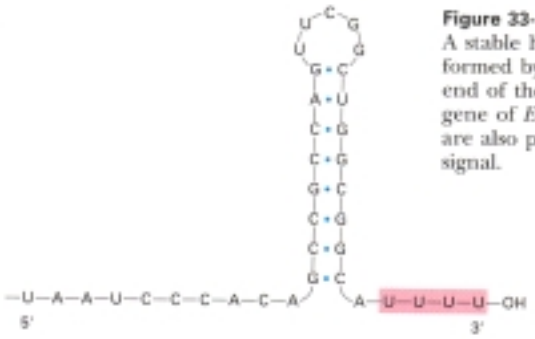
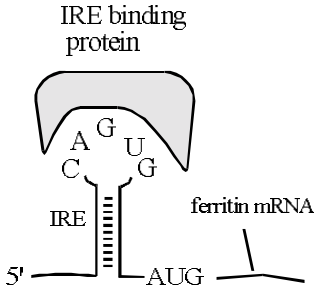


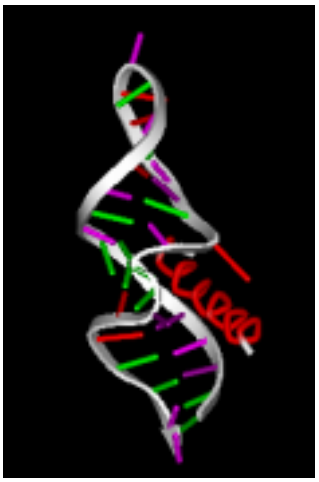
Figure 33-13
A stable hairpin structure can be formed by the base sequence of the 3′ end of the mRNA transcript of the *trp* gene of *E. coli*. The four U residues are also part of the termination signal.

2) RNA secondary structure as a regulator of translation

When the IRE binding protein is bound to the IRE, translation initiation is prevented



RNA secondary structure as a regulator of gene expression in HIV - The RRE / Rev complex



How do we identify RNA genes ?

Predicting protein and RNA genes

Protein

Identifiable primary sequence signals such as
promoters, coding sequences,
polyadenylation sites, exon/intron signals

Methods based on homology to previously known seqs

```
A MVAKQRIRMANEKHSKNITQRGNVAKTSRNAPEEKASVGPWLLALFIFVVC
  :. :::. .::.. ::::.:::::::. . : :.:  ..:::.::.:::: 
B MAPKQRMTLANKQFSKNVNNRGNVAKSLKPA-EDKYPAAPWLIGLFVFVVC
```

RNA

? No identifiable primary sequence
signal that is shared by all RNA genes

? Poor sequence homology

The primary sequence is poorly conserved
in many RNAs

=>

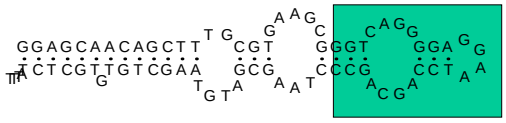
Many RNAs or functional elements in RNA
cannot be identified by sequence alignment methods

-->

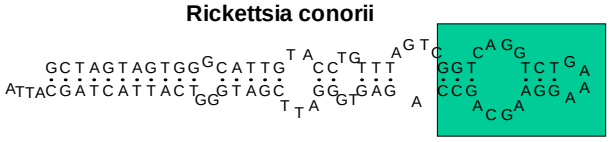
Methods for RNA identification rely on analysis of both primary sequence and secondary structure.

Example of RNA homologs with limited primary sequence homology -
- bacterial SRP RNAs

Streptococcus mutans



Rickettsia conorii



```
Ricket                  GCTAGTAGTGG.GCATTGTAC..CT....GTTTAGTCGGTCAGGT 38
                              || |||    ||  ||    ||  ||  |||||||
Strept                        GGAGCA....ACAGCTTTGCGTGAAGCGGGTCAGGG 32

Ricket          CTGAAAGGAAGCAGCC...AGAGTGGGATTCGATG.GGTC.ATTA 78
                  | ||   |||||||   || |  | |  |  || | || |||
Strept          GAGGAATCCAGCAGCCCTAAGCGATGTAAGCTGTGTGCTCTATTT 77

Ricket          CTAGCATTA                                     87

Strept
```

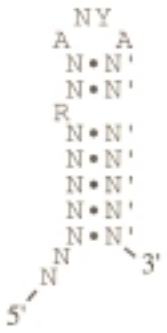


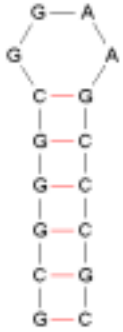
Figure 10.4 *The consensus binding site for R17 phage coat protein.* N, Y *and* R *are standard 'degenerate' symbols for multiple possible nucleotides.* N *indicates* {A, C, G, U}. Y *indicates* {C, U} *and* R *indicates* {A, G}. N' *indicates a complementary base pairing to* N.

What methods may be used to identify RNA genes that encode RNAs that belong to known families?

- Pattern matching
- Covariance models
- Mfold (prediction of secondary structure)

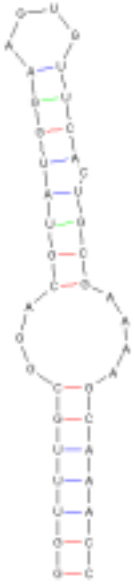
Specifying patterns in PatScan

```
p1=5...7
GGAA
~p1
```

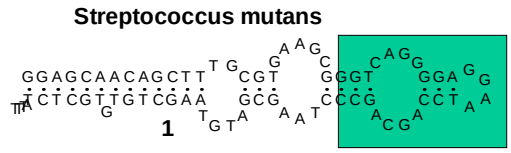


```
r1={au,ua,gc,cg,gu,ug}
p1=6...7
GGG [1,0,0]
p2=8...9
4...5
r1~p2[1,0,1]
3...4
~p1
```

```
GGTTTGC GGA CGTATGGA
AGTG TTCACTGCG AAAA
GCAAACC
```



Using pattern matching to find RNA genes




```
      YYAGG    G
--NNN      NNN N
--NNN      NNN R
     GACGA    A
```

```
p1=3...3
YYAGG
p2=3...3
GNRA
~p2
AGCAC
~p3
```

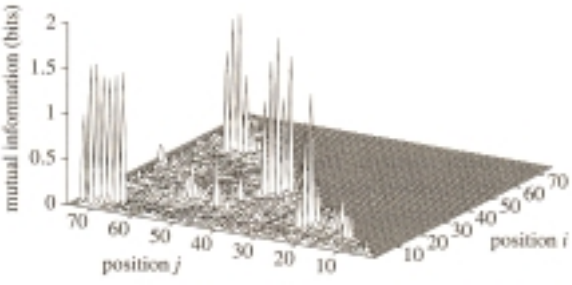
Disadvantage : Yes/no answer - will miss genes that have only one mismatch with respect to the pattern

What methods may be used to identify RNA genes that encode RNAs that belong to known families?

- Pattern matching
- Covariance models
- Mfold (prediction of secondary structure)

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).*



Intuitively, M_{ij} tells us how much information we get about the identity of the residue in one position if we are told the identity of the residue in the other position. In the case of a base pair with no sequence constraints, we get 2 bits

$$M_{ij} = \sum_{x_i, x_j} f_{x_i x_j} \log_2 \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}}.$$

f_{x_i} is the frequency of one of the four bases (A, C, G, U) observed in column i. $f_{x_i x_j}$ is the joint (pairwise) frequency of one of the sixteen possible base pairs observed in columns i and j. M_{ij} measures how much the joint frequency distribution deviates from the distribution that is expected if the two columns vary independently. For the four-letter RNA alphabet, M_{ij} varies between 0 and 2 bits. M_{ij} is maximal if i and j individually appear completely random ($f_i = f_j = 0.25$), but i and j are perfectly correlated, for instance in a Watson–Crick base pair.

Intuitively, M_{ij} tells us how much information we get about the identity of the residue in one position if we are told the identity of the residue in the other position. In the case of a base pair with no sequence constraints, we get 2 bits

Mutual information example

Consider the multiple alignment:

GGCC
GCCG
GACU
GUCA

The Mij expression (Durbin p. 266) gives:

Example 1)
For columns 2 and 4:

For the pair GC : 0.25 * log2(0.25/(0.25*0.25)) = 0.5
For the pair CG : 0.25 * log2(0.25/(0.25*0.25)) = 0.5
For the pair AU : 0.25 * log2(0.25/(0.25*0.25)) = 0.5
For the pair UA : 0.25 * log2(0.25/(0.25*0.25)) = 0.5

The sum gets 0.5 * 4 = 2 bits , which means that we have the highest possible mutual information between columns 2 and 4.

Example 2)
For columns 1 and 3:

There is only pair , GC :
1 * log2 (1/(1*1)) = 0 , i.e we have no mutual information at all.

RNA genes that belong to an already known RNA family can efficiently be identified using ‘covariance models’

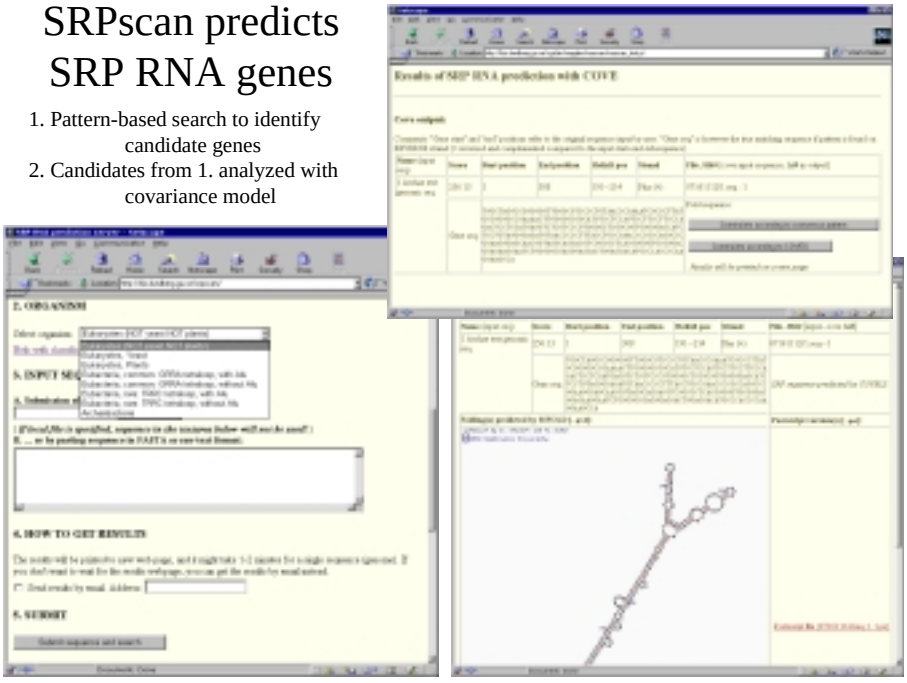
Name	No. seed	No. full	Average Length	Average %id	Description
5_8S_rRNA	64	12366	151.9	76	5.8S ribosomal RNA
5S_rRNA	606	2990	115.5	58	5S ribosomal RNA
6S	7	18	168.6	74	6S / SsrS RNA
CsrB	6	14	351.8	75	CsrB/RsmB RNA family
DsrA	3	11	86.0	93	DsrA RNA
GcvB	4	7	113.0	98	GcvB RNA
Hammerhead	68	305	35.9	68	Hammerhead ribozyme
Histone3	66	808	26.0	78	Histone 3' UTR stem-loop
Intron_gpI	30	2698	312.9	40	Group I catalytic intron
Intron_gpII	117	2963	77.7	54	Group II catalytic intron
let-7	12	28	79.7	69	let-7 microRNA precursor
MicF	9	15	91.3	93	MicF RNA
OxyS	6	12	109.3	95	OxyS RNA
RNase_MRP	25	39	265.1	52	RNase MRP
RNaseP_bact_a	226	379	318.0	62	Bacterial RNase P class A
RNaseP_bact_b	24	46	328.2	65	Bacterial RNase P class B
RNaseP_nuc	47	67	290.3	41	Nuclear RNase P
RprA	8	10	108.2	84	RprA RNA
RRE	65	486	227.9	95	HIV Rev response element
SECIS	65	217	64.3	43	Selenocysteine insertion sequence
Spot_42	4	12	117.9	95	Spot 42 RNA


Accession number: RF00013 [Data: SEED]
6S / SsrS RNA
The E. coli 6S RNA is 184 nucleotides long and forms a hairpin type structure. Although one of the earliest discovered RNAs its function has only recently been described. 6S RNA associates with RNA polymerase in a highly specific manner. 6S RNA represses expression from a sigma70-dependent promoter during stationary phase. It has been suggested that the RNA modulates sigma70-holoenzyme activity [1].
Consensus secondary structure for family 6S. Click the picture for an enlarged image, or here for more information.
Alignment
Seed (7) Full (10)
Format Coloured Blocked alignment
Get alignment
Help relating to Rfam alignments here
Member sequences
Seed (7) Full (10)
View Members
Species Distribution
Tree depth Show all levels
View Species Tree


Rfam
RNA families database of alignments and CMs
Sanger Institute
seed alignment for 6S

SRPscan predicts SRP RNA genes

1. Pattern-based search to identify candidate genes
2. Candidates from 1. analyzed with covariance model

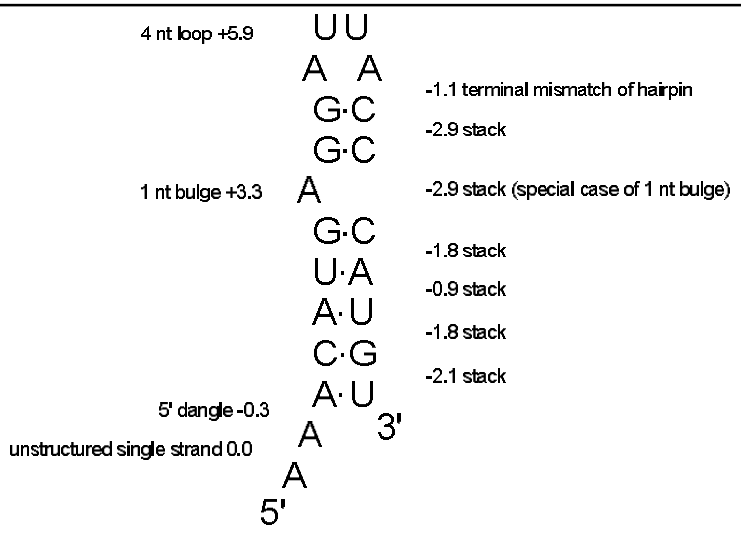


Prediction of RNA secondary structure

The simplest approach to predicting the secondary structure of RNA molecules is to find the configuration with the greatest number of paired bases. Testing and scoring each possible structure is numerically impossible. We therefore must turn to dynamic programming for an efficient solution.

Nussinov 1978
Durbin et al pp 269-272

Mfold predicts optimal and suboptimal secondary structures for an RNA (or DNA) molecule using the energy minimization method of Zuker



Overall ΔG = -4.6 kcal/mol

ΔG calculation for an RNA stem loop