Introduction to RNA Bioinformatics

Marcel Turcotte
turcotte@site.uottawa.ca

School of Information Technology and Engineering – SITE
800 King Edward, Ottawa, Ontario, Canada K1N 6N5
STE 5-003

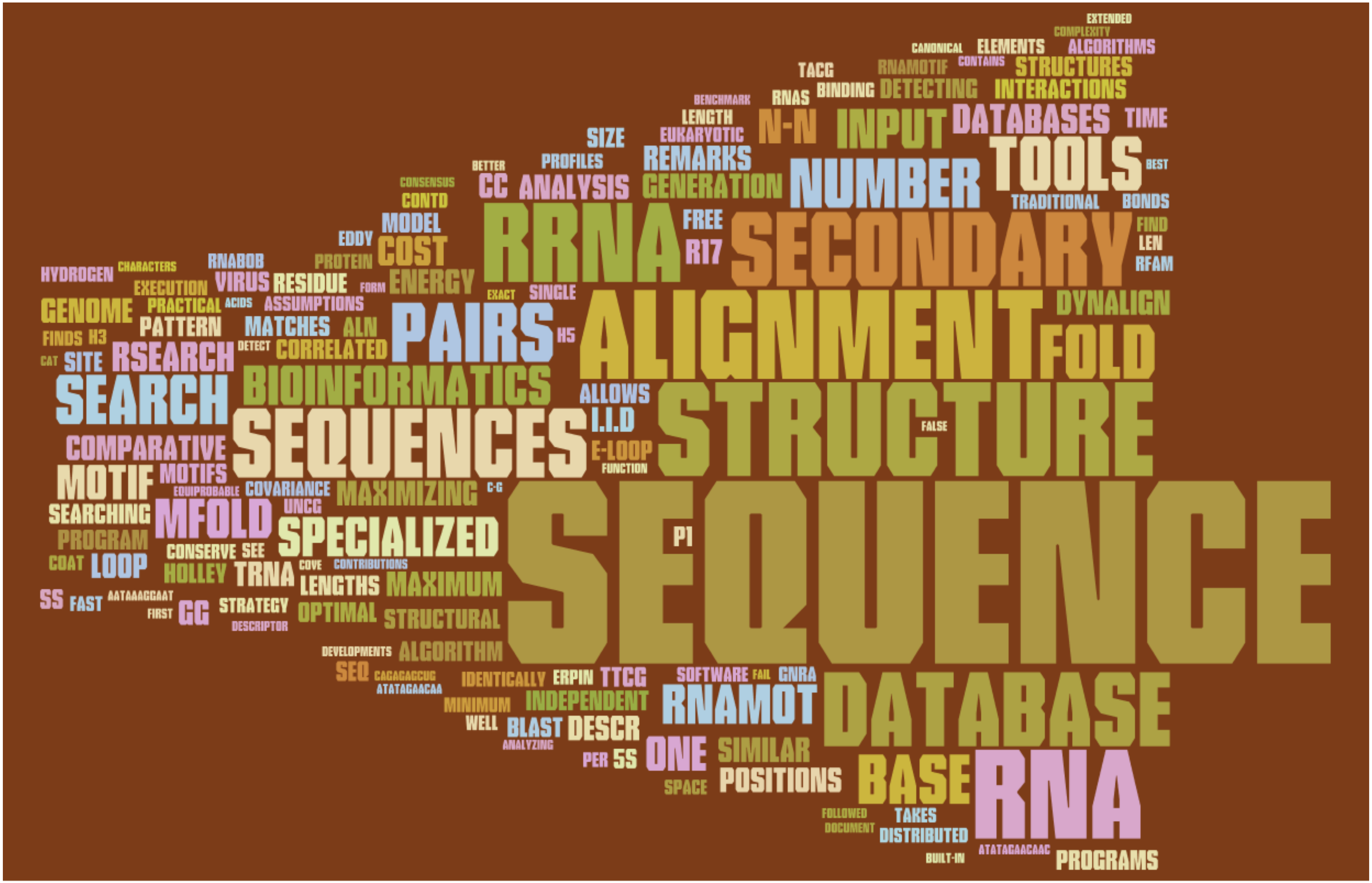
bio.site.uottawa.ca

uOttawa

L'Université canadienne
Canada's university

Université d'Ottawa | University of Ottawa

www.uOttawa.ca


SEQUENCE
STRUCTURE
ALIGNMENT
SECONDARY
RRNA
SEQUENCES
DATABASE
RNA
NUMBER
TOOLS
PAIRS
BIOINFORMATICS
SEARCH
SPECIALIZED
MFOLD
RSEARCH
FOLD
INPUT
DATABASES
BASE
RNAMOT
COMPARATIVE
MOTIF
GENERATION
REMARKS
N-N
STRUCTURES
INTERACTIONS
DETECTING
ALGORITHMS
ELEMENTS
DYNALIGN
GENOME
COST
ENERGY
MAXIMIZING
MAXIMUM
OPTIMAL
STRUCTURAL
ALGORITHM
INDEPENDENT
DESCR
BLAST
ONE
SIMILAR
POSITIONS
DISTRIBUTED
PROGRAMS
TRNA
LOOP
HOLLEY
SEARCHING
PROGRAM
MOTIFS
COVARIANCE
MATCHES
CORRELATED
RESIDUE
VIRUS
ASSUMPTIONS
PATTERN
PRACTICAL
EXECUTION
HYDROGEN
CHARACTERS
RNABOB
PROTEIN
MODEL
EDDY
CONTD
CONSENSUS
CC
ANALYSIS
PROFILES
BETTER
SIZE
EUKARYOTIC
LENGTH
BENCHMARK
RNAS
BINDING
TACG
RNAMOTIF
CONTAINS
CANONICAL
COMPLEXITY
EXTENDED
TIME
BEST
TRADITIONAL
BONDS
FIND
LEN
RFAM
FREE
R17
SINGLE
EXACT
FORM
ACIDS
ALN
DETECT
H5
ALLOWS
I.I.D
E-LOOP
FUNCTION
FALSE
PI
C-G
UNCG
EQUIPROBABLE
CONSERVE
SEE
COVE
CONTRIBUTIONS
LENGTHS
COAT
SS
FAST
AATAAAGGAAT
FIRST
GG
STRATEGY
DESCRIPTOR
DEVELOPMENTS
SEQ
CAGAGACCUG
ATATAGAACAA
IDENTICALLY
ERPIN
TTCG
SOFTWARE
FAIL
GNRA
MINIMUM
WELL
ANALYZING
PER
5S
SPACE
FOLLOWED
DOCUMENT
TAKES
BUILT-IN
ATATAGAACAAC
FINDS
H3
CAT
SITE

uOttawa

Objectives

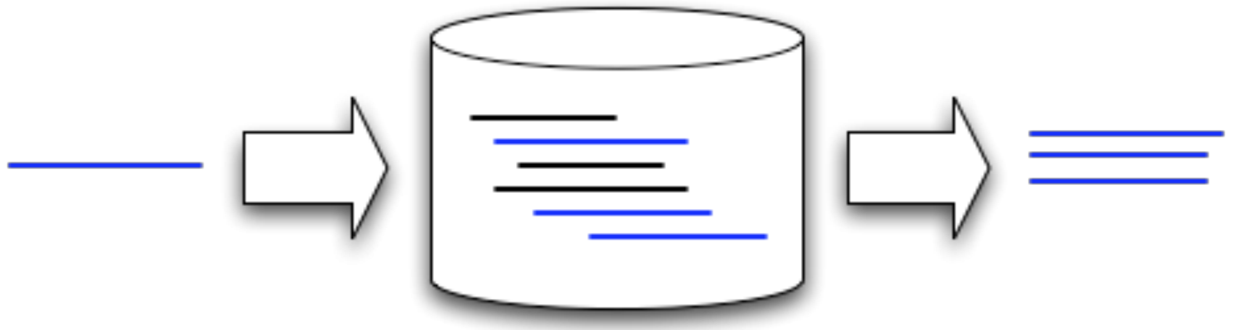
- Understanding the limitations of traditional bioinformatics tools w.r.t. RNA molecules
- Providing an overview of the bioinformatics tools that are specific to RNA research

Bioinformatics

- **Database search**, in the form of sequence comparison, is the workhorse of bioinformatics
- "Basic Local Alignment Search Tool (**BLAST**) is one of the most heavily used sequence analysis tools available in the public domain"
- In 2004, on average, NCBI was running **140,000 blast runs per weekday**, on a farm consisting of 200 CPUs (running Linux)
- In 2008, "BLAST is the most popular bioinformatics tool and is used to run millions of queries each day"

Database search

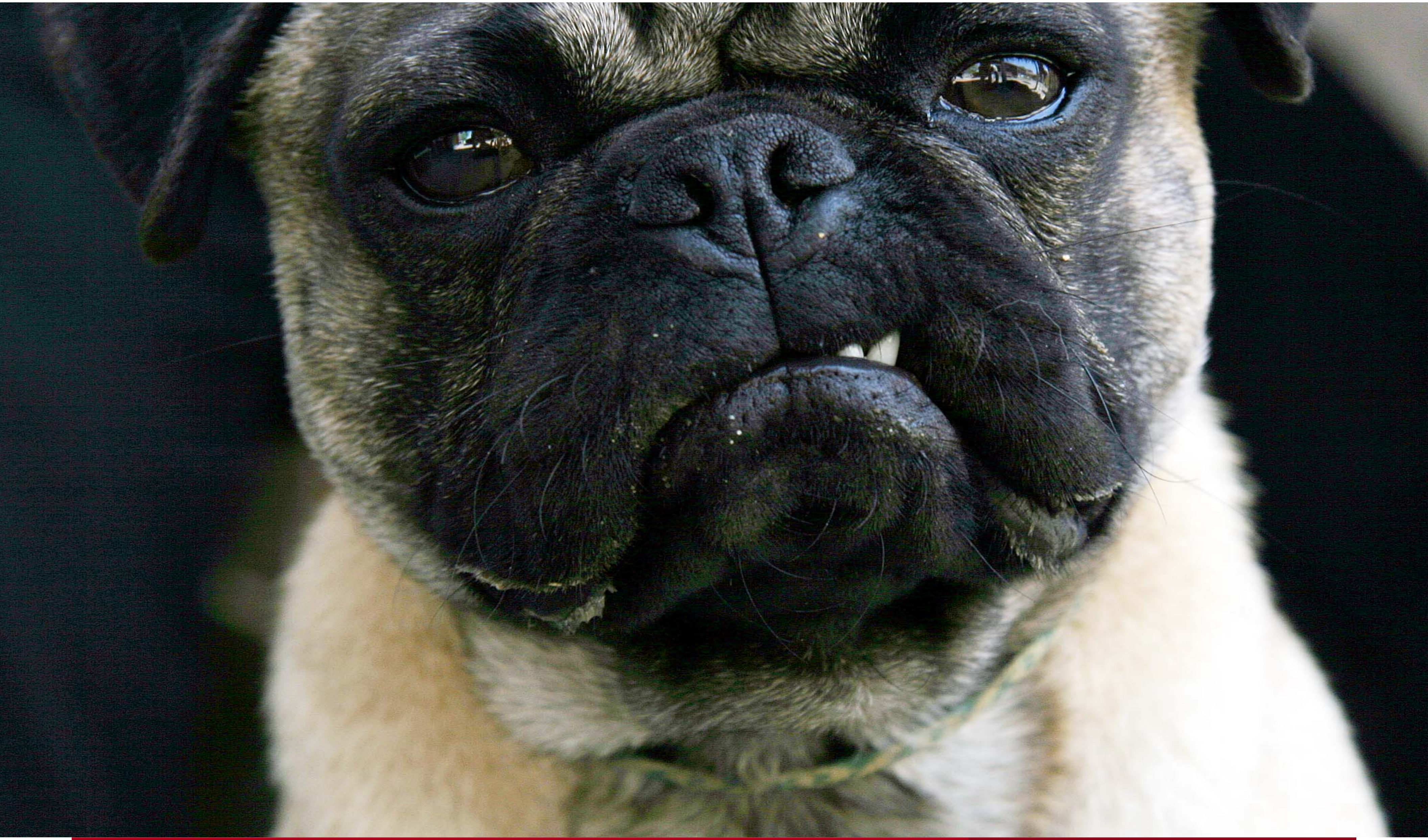
Find all GenBank gene's that are similar to *Clostridium botulinum*'s toxin gene



```
>gi|27867582(fragment of the known Clostridium botuninum toxin gene)
GTGAATCAGCACCTGGACTTTCAGATGAAAAATTAAATTTAACTATCCAAAATGATGCTT
ATATACCAAAATATGATTCTAATGGAACAAGTGATATAGAACAACATGATGTTAATGAAC
TTAATGTATTTTTCTATTTAGATGCACAGAAAGTGCCCGAAGGTGAAAATAATGTCAATC
TCACCTCTTCAATTGATACAGCATTATTAGAACAACCTAAAATATATACATTTTTTTCAT
CAGAATTTATTAATAATGTCAATAAACCTGTGCAAGCAGC
```

Result of a database search

```
>gi|49138|emb|X68262.1|CBBONTF  C.barati gene for type F neurotoxin

Length=4073 Score = 81.8 bits (41),  Expect = 1e-12
Identities = 99/121 (82.82%), Gaps = 2/121 (0.02%)
Strand=Plus/Plus

Query  48    CAAAATGATGCTTATATACCAAAATATGATTCTAATGGAACAAGTGATATAGAACAACAT  107
             ||||||||| ||||  | |||||||||||||||||||| |||||||| ||| |  || ||
Sbjct  1712  CAAAATGATTCTTACGTTCCAAAATATGATTCTAATGGTACAAGTGAAATAAA-GAATAT  1771

Query  108   GATGTTAATGAACTTAATGTATTTTTCTATTTAGATGCACAGAAAGTGCC-GAAGGTGAA  167
               |||| || |||| |||||||||||||||||| ||||||| ||||  || |||||||||
Sbjct  1772  ACTGTTGATAAACTAAATGTATTTTTCTATTTATATGCACAAAAAGCTCCTGAAGGTGAA  1831

Query  168   A  168             |
Sbjct  1832  A  1832

…
```



How does it work?

uOttawa

Pairwise Sequence Alignment (Algorithm)

- **An optimal alignment is obtained by extending**:
  - An optimal alignment with one more residue from each sequence (**match** or mismatch);
  - An optimal alignment with one residue from the first sequence and a gap symbol (**deletion**);
  - An optimal alignment with one residue from the second sequence and a gap symbol (**insertion**).

Algorithm

Alignment cost **aln**(ATATAGAACAAC, AATAAAGGAAT) is

The maximum of:

aln(ATATAGAACAA, AATAAAGGAA) + cost of substituting C by T

```
ATATAGAACAA C
AATAAAGGAA  T
```

aln(ATATAGAACAA, AATAAAGGAAT) + cost of deleting C

```
ATATAGAACAA C
AATAAAGGAAT -
```

aln(ATATAGAACAAC, AATAAAGGA) + cost of inserting T

```
ATATAGAACAAC -
AATAAAGGAA   T
```

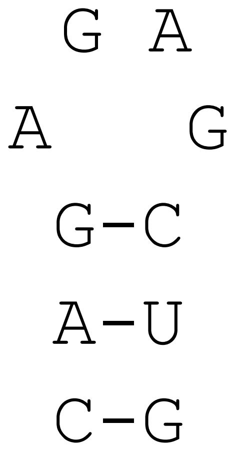
Molecular Sequence Alignment Assumptions

- *i.i.d.*
- Positions along the sequence are **independent and identically distributed**
- Independence is necessary for the development of efficient exact algorithms (Smith-Waterman) or heuristics (such as BLAST)
- The **execution time** of the exact algorithms grows proportionally to the **product** of the **size of the database times the size of input sequence**

RNA Sequence Alignment

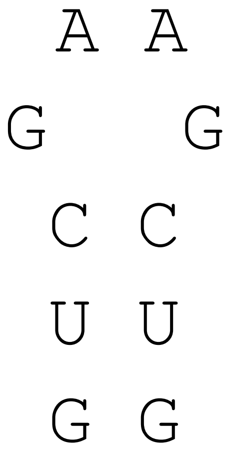
```
1    GUCGAGAGAC
     *****
2    GUCGAAGCUG
          *****
3    CAGAGAGCUG
```

1 and 2 are 50% identical (similarly for 2 and 3), however, 1 and 3 don’t seem to have anything in common



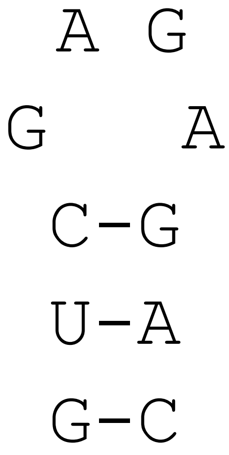
CAGAGAGCUG

1



GUCGAAGCUG

2



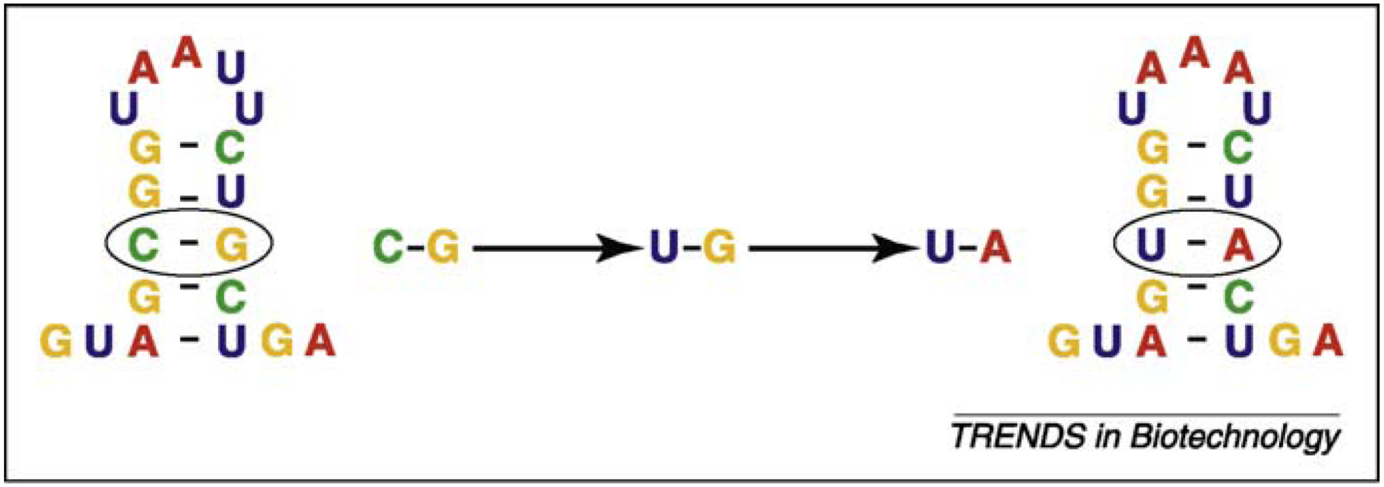
GUCGAGAGAC

3

Yes, but sequences 1 and 3 share the same secondary structure!

Caveat

- RNAs conserve secondary structure interactions more than they conserve their sequence
- Traditional bioinformatics tools, assuming that positions are independent, perform poorly



TRENDS in Biotechnology

Paradigms

1. Inference
2. Searching

Bias

- **Secondary structure** plays an important role in the elements that are sought

Time and space complexity

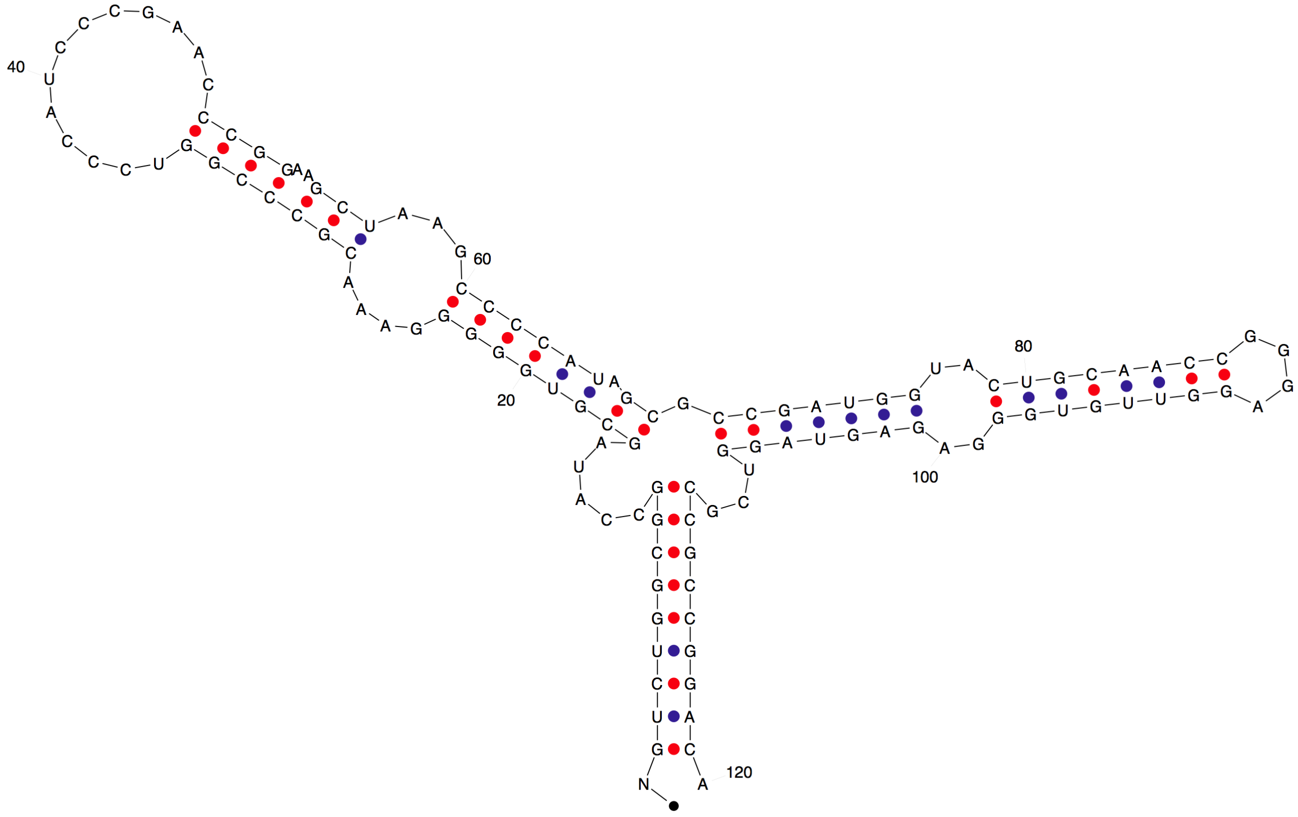
- Should we worry about the time and space complexity of the methods?
- After all, we can always buy a faster computer, right?
- Computer scientists use mathematical approaches to analyze the execution time and memory requirements

Time and space complexity

- Some algorithms require a **linear** amount of resources
- Some require **polynomial** amounts of resources
- Some always require **exponential** resources, these are **NP-hard**

Part I: Inference

Stems, hairpins, interior loops, bulges, and multi-branch loops



Definitions

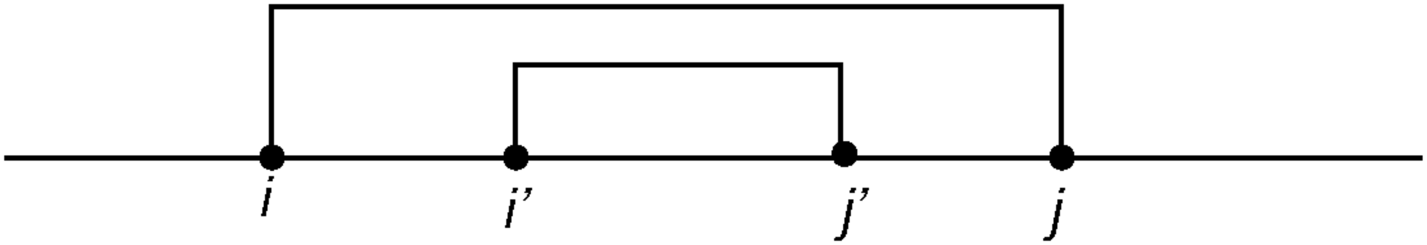
Given an RNA **sequence** $S = s_1, s_2, \ldots, s_n$ where s_i is the i^{th} nucleotide.

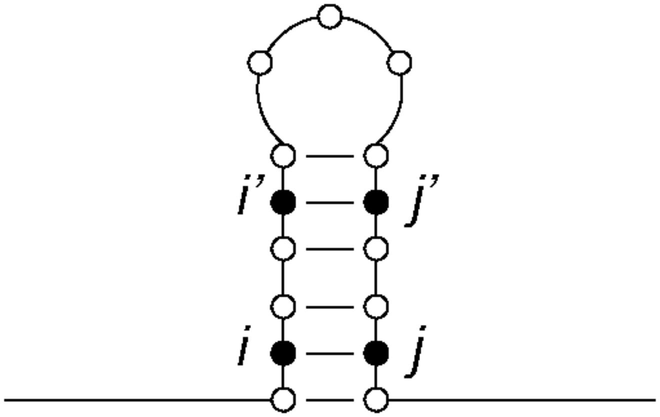
A **secondary structure** is an ordered list of pairs, *i.j,*

$1 \leq i < j \leq n$ such that:

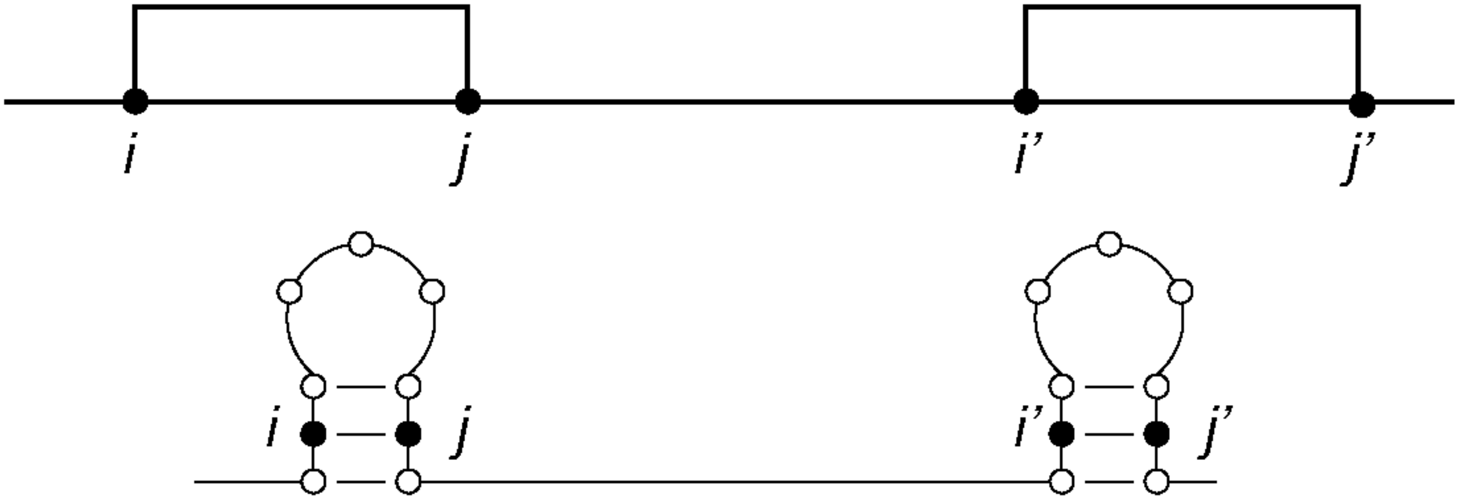
- j – i $\geq$ 4
- Given *i.j* and *i′.j′*, two base pairs, then either:
  - $i = i'$ and $j = j'$ (they are the same)
  - $i < j < i' < j'$ (*i.j* precedes *i′.j′*)
  - $i < i' < j' < j$ (*i.j* includes *i′.j′*)
  - $i < i' < j < j'$ (pseudoknot)

i < *i’* < *j’* < *j* (*i.j* includes *i’.j’*)

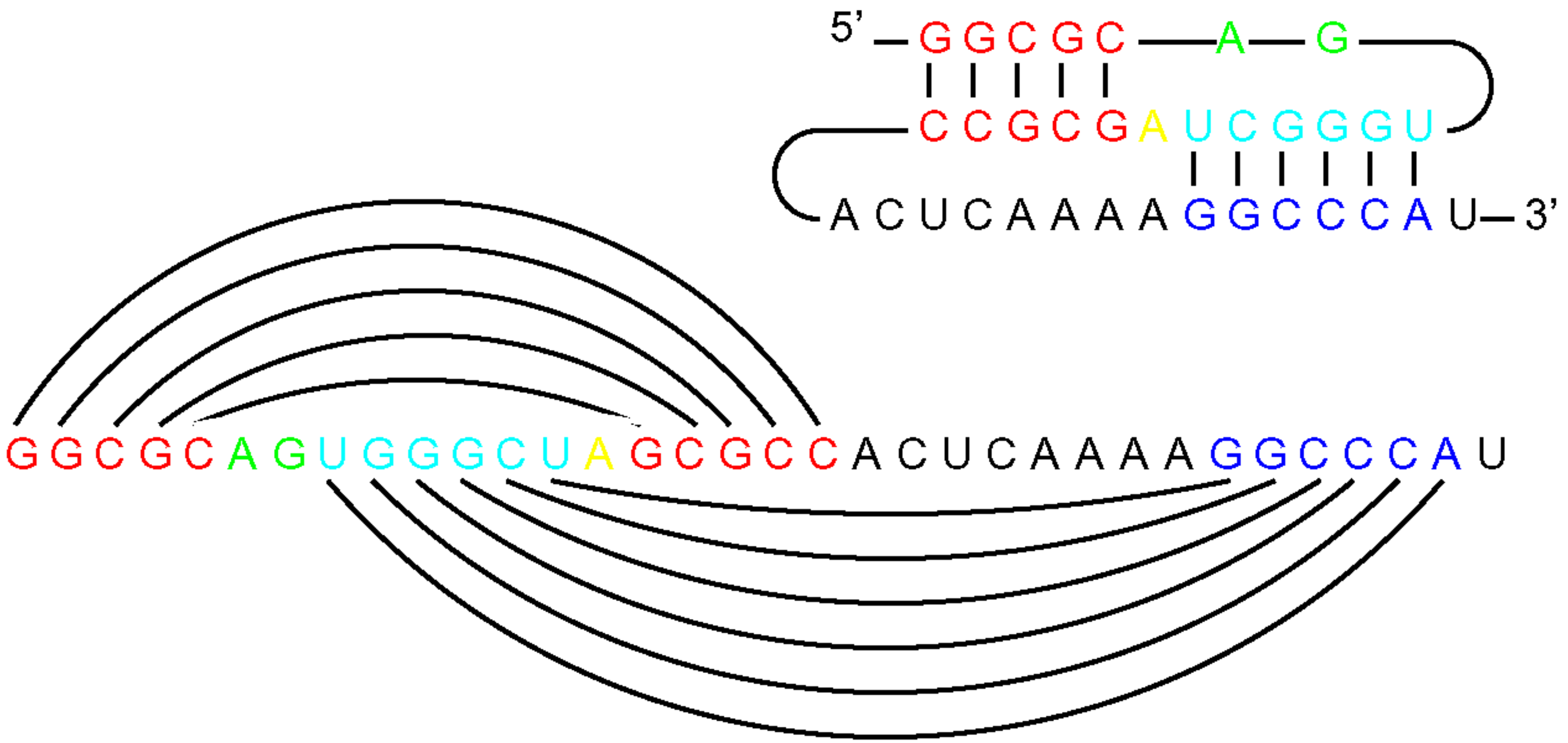


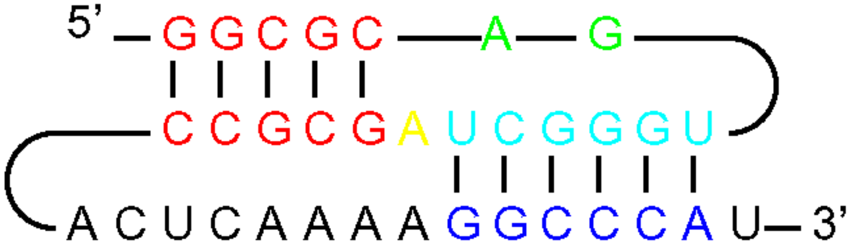


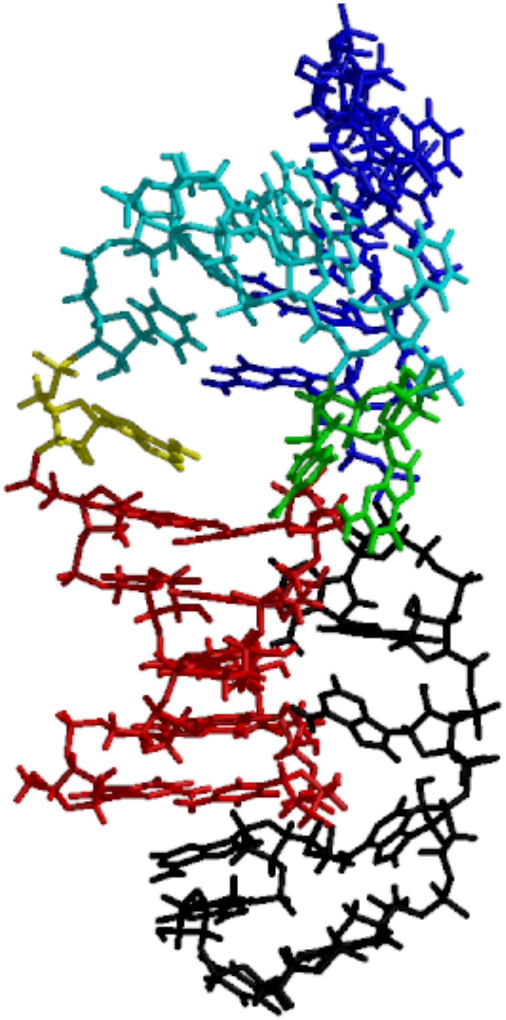
$i < j < i' < j'$ ($i.j$ precedes $i'.j'$)

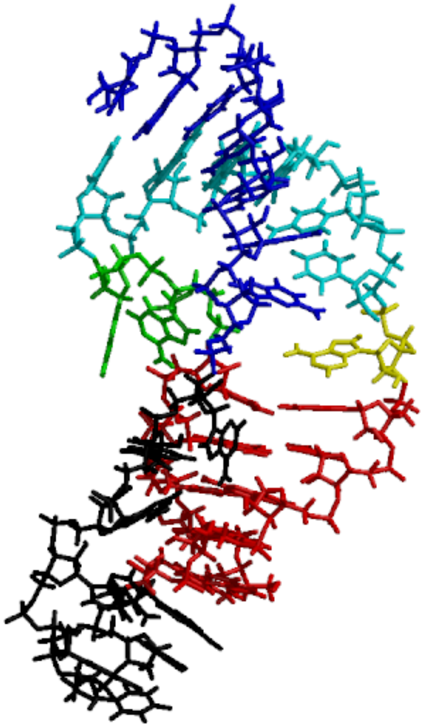


i < i' < j < j' (pseudoknot)



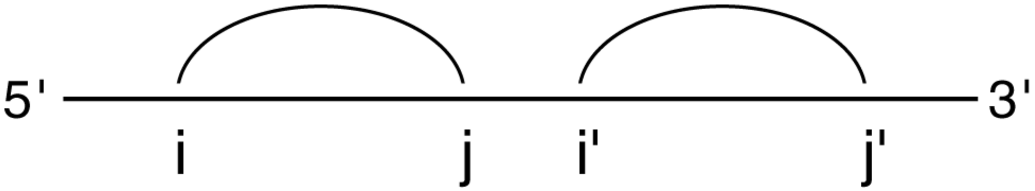

5’ GGCGC A G
CCGCGAUCGGGU
ACUCAAAAGGCCCAU 3’



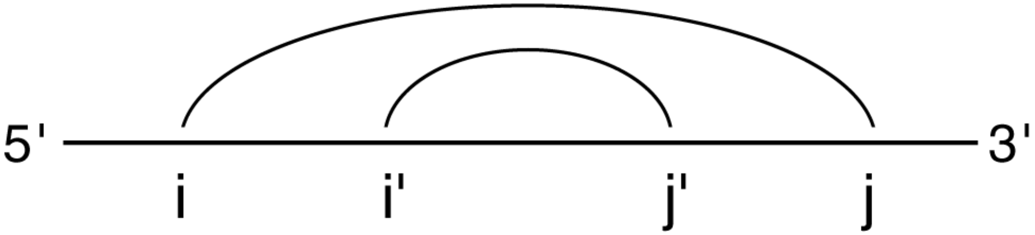


The three cases

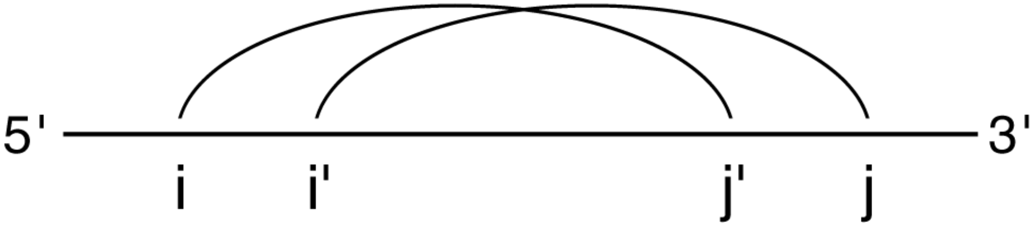
a. $i < j < i' < j'$



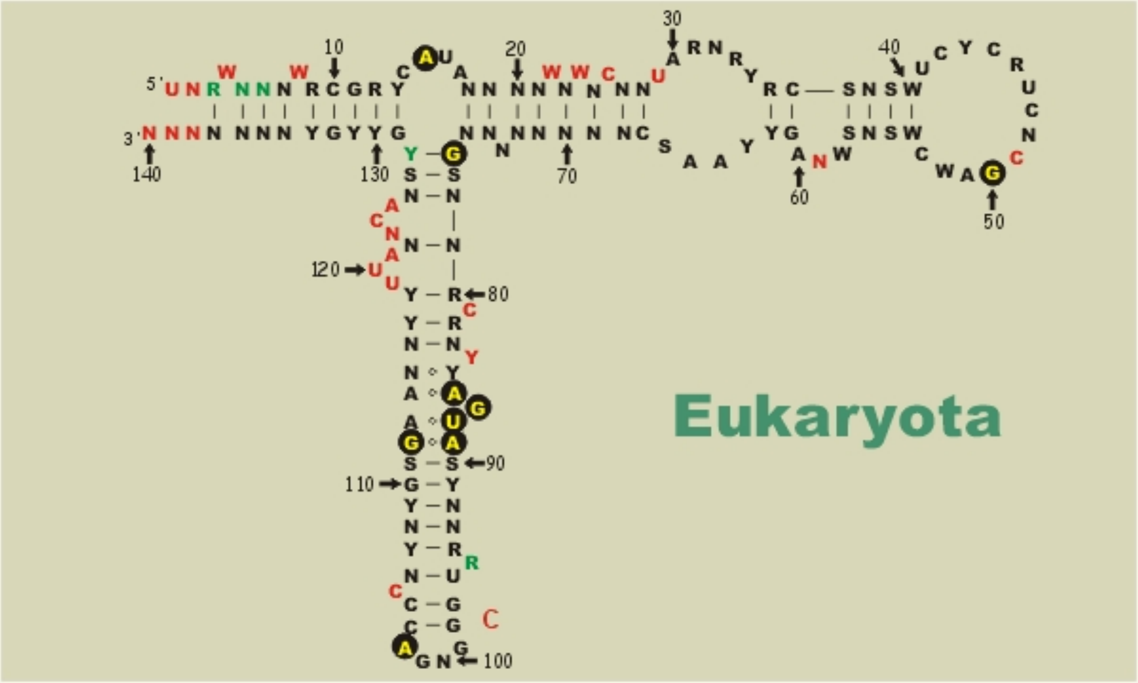
b. $i < i' < j' < j$



c. $i < i' < j < j'$

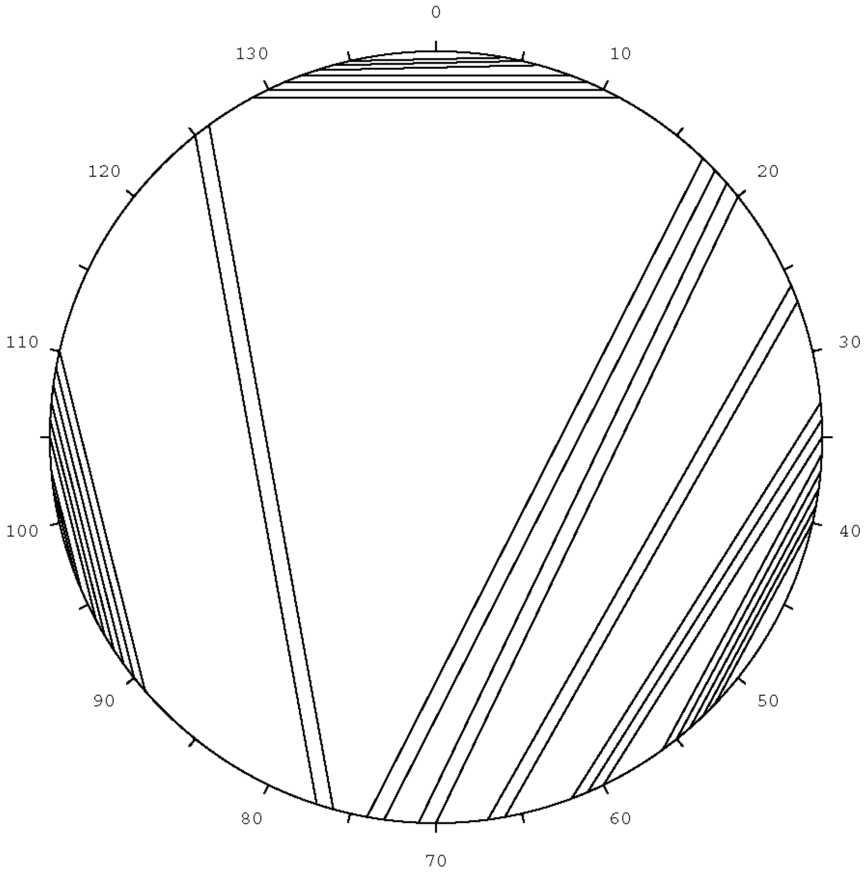


5S rRNA

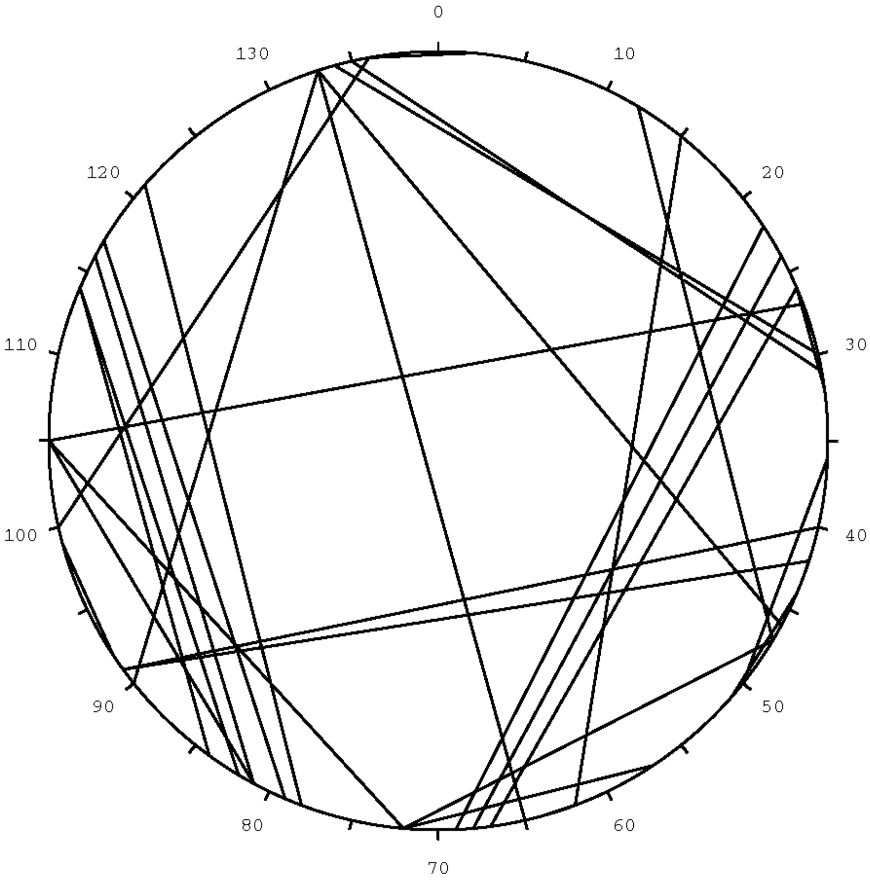


from http://rose.man.poznan.pl/5SData/

Eukaryotic 5S RNA sequences secondary structure interactions



Eukaryotic 5S RNA sequences (possible 3D interactions)



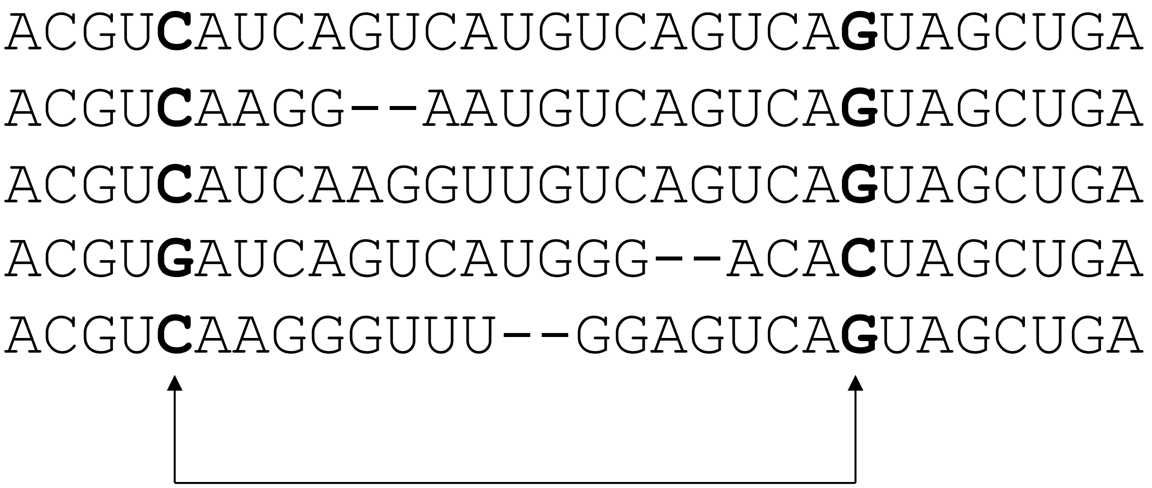
Secondary Structure Determination

- X ray crystallography, N.M.R.
- Chemical and enzymatic probing, cross-linking
- **Comparative sequence analysis**
- **Minimum free energy (MFE) methods**
- **Comparative sequence analysis + MFE**

Comparative Sequence Analysis

"Today, comparative analysis has become the method of choice for establishing higher-order structure for large RNA"

Pace, Thomas, Woese (1999) In The RNA World. Cold Spring Harbor.



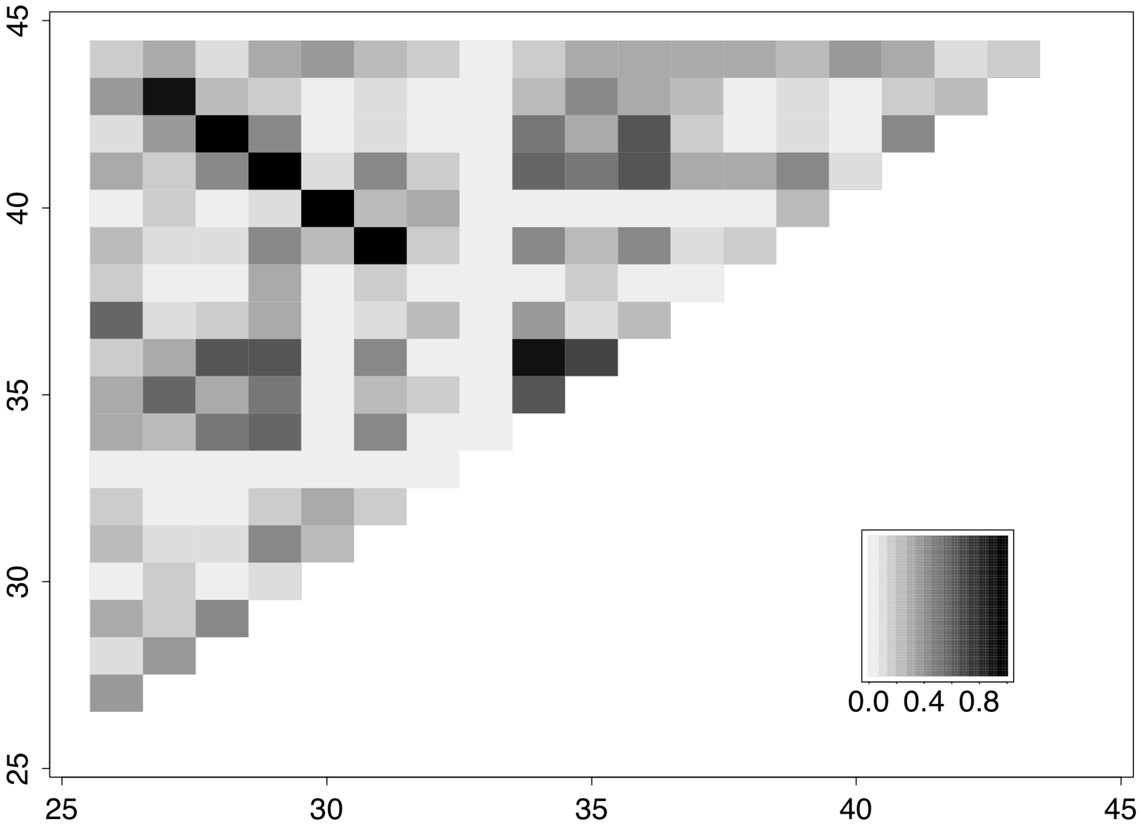
ACGU**C**AUCAGUCAUGUCAGUCA**G**UAGCUGA

ACGU**C**AAGG--AAUGUCAGUCA**G**UAGCUGA

ACGU**C**AUCAAGGUUGUCAGUCA**G**UAGCUGA

ACGU**G**AUCAGUCAUGGG--ACA**C**UAGCUGA

ACGU**C**AAGGGUUU--GGAGUCA**G**UAGCUGA



Saccharomyces cerevisiae	...**CCAGA**CUGAAGA**UCUGG**...
Spiroplasma meliferum	**CCUGC**CUUGCAC**GCAGG**
Mycoplasma capricolum	**CCUCC**CUGUCAC**GGAGG**
Mycoplasma mycoides	**CACGG**UUUUCAU**CCGUG**
Spiroplasma meliferum	**UUUGA**UUGAAGC**UCAAA**
Streptomyces lividans	**ACGĢC**CUGÇAAA**GÇCGU**
	30 35 40

Comparative Sequence Analysis

- Starts with the alignment of a set of homologous sequences (computer-assisted, but manually refined)
- Detecting correlated pairs
- Analyzing correlated pairs:
  - Parallel chords implies helices
  - Others are tertiary structure interactions

Detecting Correlated Pairs

- Chi-square test of independence
- **Mutual information**

$$M(I,J) = H(I) + H(J) - H(I,J)$$

where

$$H(I,J) = -\sum_{\alpha\beta} P(i=\alpha, j=\beta)\log P(i=\alpha, j=\beta)$$

$$H(I) = -\sum_{\alpha} P(i=\alpha)\log P(i=\alpha)$$

Analyzing Correlated Pairs

- Detecting secondary structure elements:
  - Mostly canonical base pairs (Watson-Crick)
  - Parallel (i:j, i+1:j-1)
  - Wobble (G:U) and A:G are occurring frequently
- Non-canonical (isosteric)
- Detecting tertiary structures (including pseudoknot)
- Tetraloop: UNCG, CUYG, GMRA (GNRA)
- Base-triples

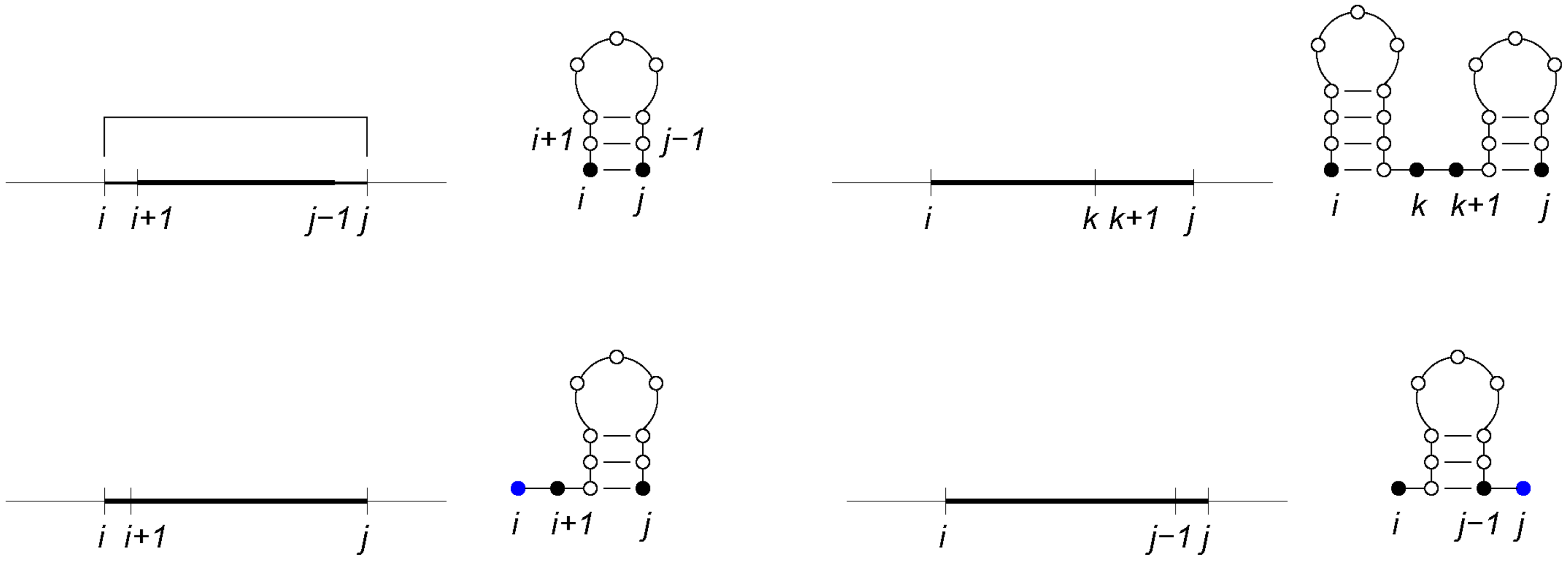
What are the main difficulties?

- Needs an alignment, but sequence alignment techniques are not well adapted for RNA sequences
- To produce a high quality alignment, the sequences should be similar
- If the sequences are similar, there will be few observed compensatory changes

RNA folding

- How to search the space of all possible secondary structures?
- How to select the best structure?
  - Maximizing the number of base-pairs (Nussinov)
  - Maximizing the number of hydrogen bonds
  - Minimizing the free energy (Zuker/*mfold*)

What is the maximum number of base pairs that can be formed for the segment *i .. j*?



Putting it all together

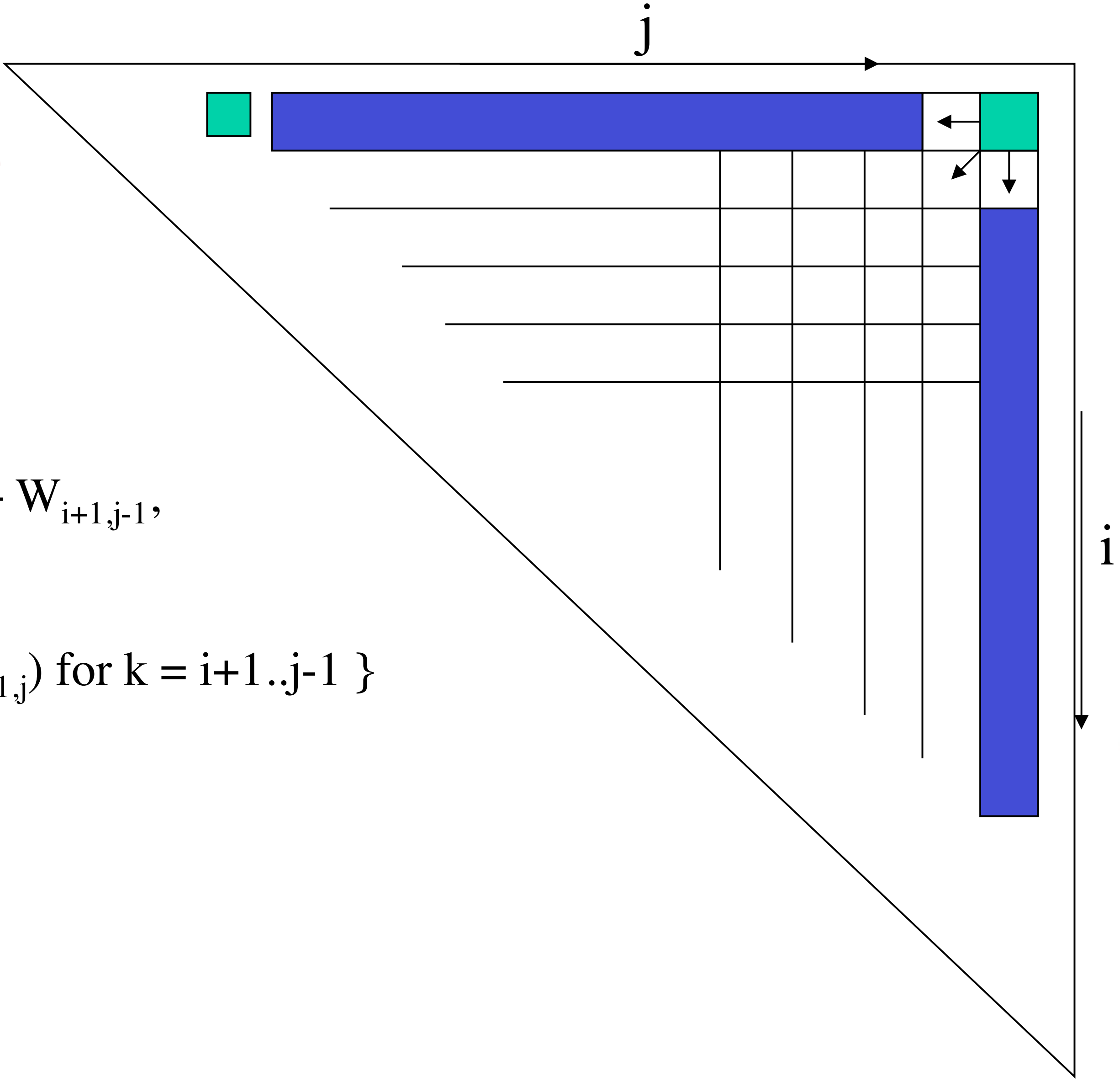
- We know that for $j\text{-}i \leq 4$ **fold(*s,i,j*) = 0**
- Otherwise, **fold(*s,i,j*)** is the maximum of
  - **1 + fold(*s,i+1,j-1*)** if *s(i)* and *s(j)* form a canonical base pair;
  - **fold(*s,i+1,j*);**
  - **fold(*s,i,j-1*);**
  - **fold(*s,i,k*) + fold(*s,k+1,j*)** for some *k* s.t. $i \leq k \leq j$.
- The answer we're looking for is **fold(*s,1,n*)**.

Remarks

- The proposed algorithm is not practical, it requires an **exponential** number of calls to **fold(*s,i,j*)**
- However, there is a maximum of $n \times n$ distinct values of **fold(*s,i,j*)**
- This suggests a caching strategy (tabular computation)

Filling the DP table

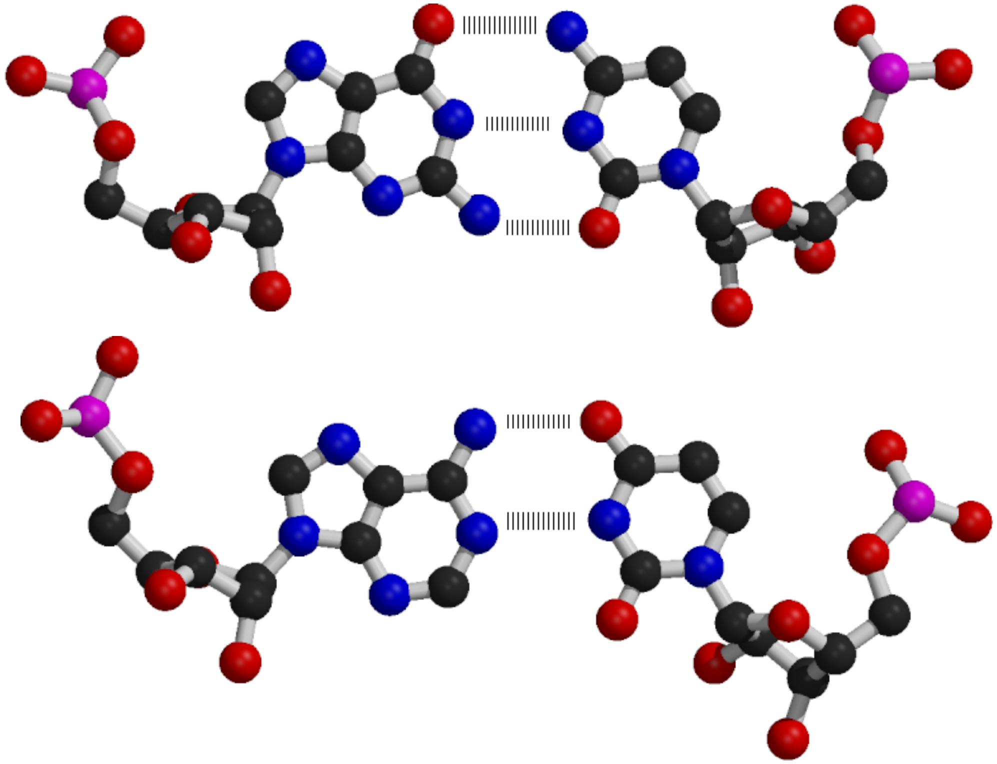
$$W_{ij} = \max\{\delta(s(i),s(j)) + W_{i+1,j-1},\; W_{i+1,j},\; W_{i,j-1},\; (W_{i,k} + W_{k+1,j}) \text{ for } k = i+1..j-1\ \}$$



Maximizing the number of base pairs is not a good strategy

Maximizing the number of hydrogen bonds: A better cost function?

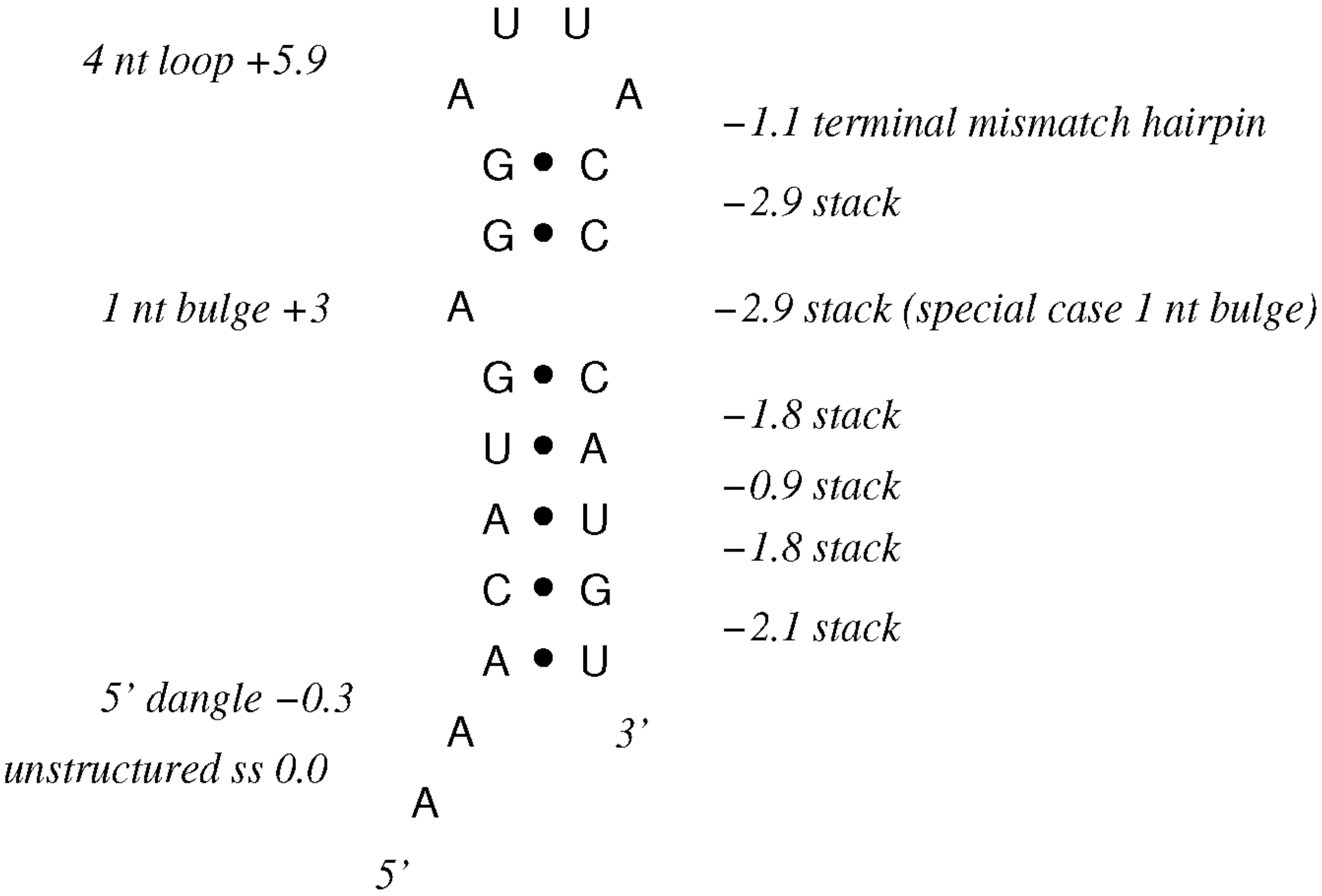
+ 3 for a G:C base pair
+2 for an A:U
+1 for a Wobble (G.U)



Better cost functions

- It turns out that maximizing the number of base pairs, or the number of hydrogen bonds, is not what Nature has favored
- The **stacking** contributions from the interface between neighboring base pairs seem to be preferred

$\Delta G = -4.9$ kcal/mol



U U

4 nt loop +5.9

A A

−1.1 terminal mismatch hairpin

G • C

−2.9 stack

G • C

1 nt bulge +3

A

−2.9 stack (special case 1 nt bulge)

G • C

−1.8 stack

U • A

−0.9 stack

A • U

−1.8 stack

C • G

−2.1 stack

A • U

5’ dangle −0.3

A

3’

unstructured ss 0.0

A

5’

www.bioinfo.rpi.edu/~zukerm/rna/energy

From Durbin *et al* (1998) Cambridge Press.

MFOLD

- Sophisticated energy minimization program developed by **Mike Zuker**
- Finds the structure with the minimum equilibrium free energy (ΔG), as approximated by **neighboring base pair contributions**
- **Takes into account:** stacking, hairpin loop lengths, bulge loop lengths, interior loop lengths, multi-branch loop lengths, single dangling nucleotides and terminal mismatches on stems

MFOLD and PKNOTS (Implementation)

- MFOLD does not include pseudoknots
- MFOLD and the dynamic programming algorithm is in $O(N^3)$
- PKNOTS is an implementation of the dynamic programming that includes pseudoknots
- PKNOTS with pseudoknots is in $O(N^6)$

Some recent developments

- Dynalign is an algorithm that **simultaneously align two RNA sequences and finds a common secondary structure** with minimum free energy:
  $\Delta G_1 + \Delta G_2 + \Delta G_{gap}$ *(number of gaps)*
- Computationally intensive! $O(M^3 N^3)$, where N is the length of the shortest sequence and M is maximum insertion size

Practical Remarks

- MFOLD was benchmark on a set of 955 structures of 700 nt or less:
    - Before 1999, 64% of the known base pairs were correctly predicted
    - 1999+, **73%**
- Dynalign (a standalone program)
    - 13 tRNAs: Dynalign = **86.1%**, MFOLD = 59.7 %
    - 7 5S rRNA: Dynalign = **86.4%**, MFOLD = 47.8 %

Further extensions

- **eXtended Dynalign** takes three input sequences and produces 1) alignment as well as 2) a consensus secondary structure
- **Profile-Dynalign** takes as input an arbitrarily large number of input sequences, applies a **progressive alignment strategy** akin to CLUSTAL and produces 1) a multiple sequence alignment as well as 2) a consensus secondary structure

eXtended and Profile-Dynalign

- See PDF document.

Practical Remarks (contd)

- MFOLD requires a single sequence;
- MFOLD allows for constraints;
- MFOLD reports sub-optimal solutions;

Seed

- See PDF document.

Part II

- **Database search**
  - Traditional bioinformatics tools
  - Specialized tools
  - Specialized databases

S

- See Backhofen’s Garfield the fat and old cat vs Garfield the cat and the old hat

Important Observations

- Many RNAs conserve their (secondary) structure more than their sequence
- Consequently, sequence alignment techniques (such as blast) fail to detect homologues
- More sophisticated tools are required

R17 virus coat protein binding site

```
   N Y
  A   A
   N-N'
   N-N'
  R
   N-N'
   N-N
   N-N'
   N-N'
   N-N'
  N   3'
 N
5'
```

IUPAC ambiguity codes

R = [GA]

Y = [CT]

M = [AC]

K = [GT]

S = [GC]

W = [AT]

N = [ACGT]

D = [^C]

H = [^G]

V = [^T]

N' is the complement of N

i.i.d. sequence model

- Under the assumptions that positions are independent and identically distributed (*i.i.d.*), and all 4 nucleotide types are equiprobable;
- i.e. the sequence motif NNNNNNNRNNANYANNNNNNN;
- The probability that a random sequence matches the **sequence** motif of the R17 coat protein binding site is,

$$\left(\frac{1}{2}\right)^2 \times \left(\frac{1}{4}\right)^2 \times 1^{17} = \frac{1}{64} = 0.015625$$

- You would expect 56 hits in the 3,569 nts of the R17 virus genome.

i.i.d. structural model

- Under the assumptions that positions are independent, **except for paired positions**, and identically distributed (*i.i.d.*), and all 4 nucleotide types are equiprobable;
- The probability that a randomly selected sequence matches the secondary structure motif of the R17 virus coat protein binding site is,

$$\left(\frac{1}{4}\right)^{7}\times\left(\frac{1}{4}\right)^{2}\times\left(\frac{1}{2}\right)^{2}=\left(\frac{1}{2}\right)^{20}\approx 9.5\times 10^{-7}$$

  - Would occur 0.003 times by chance in R17 virus genome.

Searching for Structural Motifs

- General purpose tools
  - Generation 1: pattern
  - Generation 2: built-in scoring mechanisms
  - Generation 3: built-in covariance model
  - Future: automatic inference
- Specialized programs
  - tRNA-scan-SE
  - snoRNA

Searching for Structural Motifs: A first generation of algorithms

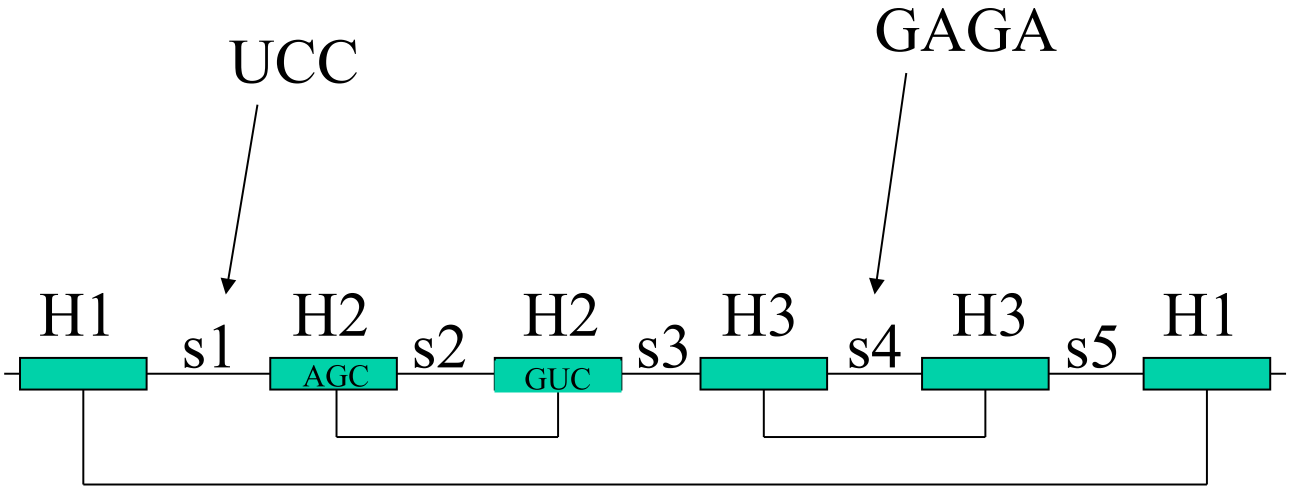
The input of general motif search procedures, such as RNAMOT or RNABOB, requires a description of the motif in terms of its secondary and tertiary structure: the **descriptor** or **pattern**

RNAMOT Descriptor

H1 s1 H2 s2 H2 s3 H3 s4 H3 s5 H1

H1 3:5 0
H2 4:5 1 AGC:GCU
H3 4:5 1
S1 3:6 UCC
S2 5:7
S3 0:3
S4 5:8 GAGA
S5 3:5

R H2 H3 H1
M 1



RNAMOT execution

- RNAMOT -s -s mydb.fa -d mystery.mot

```
--- HUM7SLR1 Human 7SL RNA pseudogene, clone p7L30.1. --- (110 bases)
|SCO:  201.40|POS:6-56|MIS: 0|WOB: 0|
|CAGCU|GAUGCU|AGCU|GAUGCU|AGCU|-|GAUCG|UAGCUAGU|CGAUC|CGU|AGCUG|
…
```

RNAMOT Descriptor

H1 s1 H2 s2 H2 s3 H3 s4 H3 s5 H1 ← Secondary structure description

Length range

H1 3:5 0 ← Number of allowed mismatches

H2 4:5 1 AGC:GCU ← Sequence pattern

H3 4:5 1

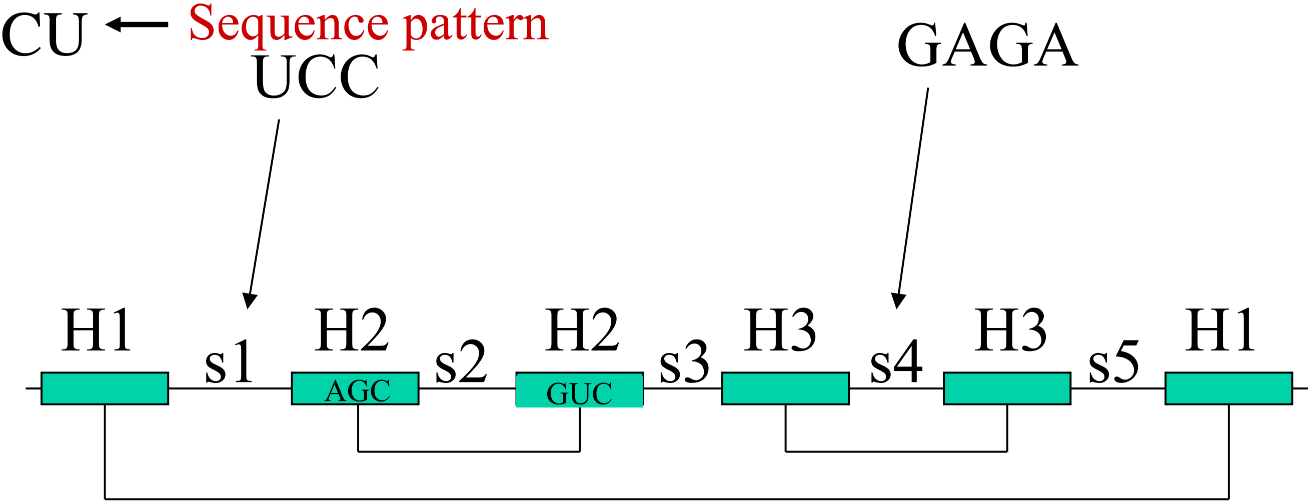
S1 3:6 UCC

S2 5:7

S3 0:3

S4 5:8 GAGA

S5 3:5



R H2 H3 H1 ← Search order information

M 1 ← Total number of mismatches

Similar tools

- RNABOB
  http://www.genetics.wustl.edu/eddy/software/
- PatScan
  - http://www-unix.mcs.anl.gov/compbio/PatScan/
  - scan_for_matches (stand alone program)
  - p1=4…7 3…8 ~p1
    (p1 contains 4 to 7 characters, it is followed by 3 to 8 characters, followed by the reverse complement of p1)

Remarks

- These computer programs are practical and can be applied to large data-sets

- One of the major difficulties arises from the **subjectivity in deriving the best descriptor** for a family of sequences

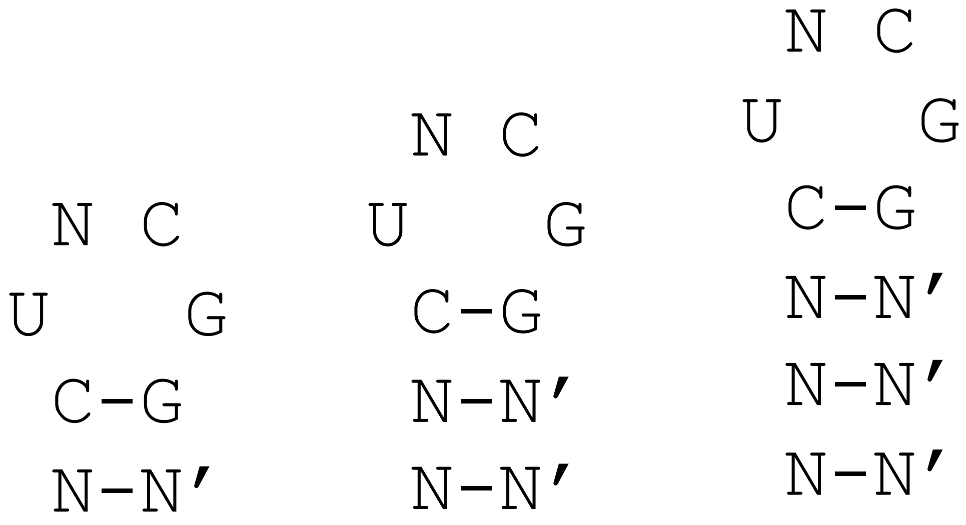
Second Generation of Pattern Matching Engines

- 10+ years after RNAMOT was published, RNAMOTIF was released;
- It has all the functionalities of RNAMOT + the ability for the user to define a scoring function!
- It also features a powerful scripting language.

- Macke *et al*. (2001) *Nuc. Acids. Res.* **29(22)**: 4724-4735.

UNCG loop

```
descr
  h5(minlen=2,maxlen=4,seq="C$")
    ss(len=4,seq="UNCG")
  h3(seq="^G")
```



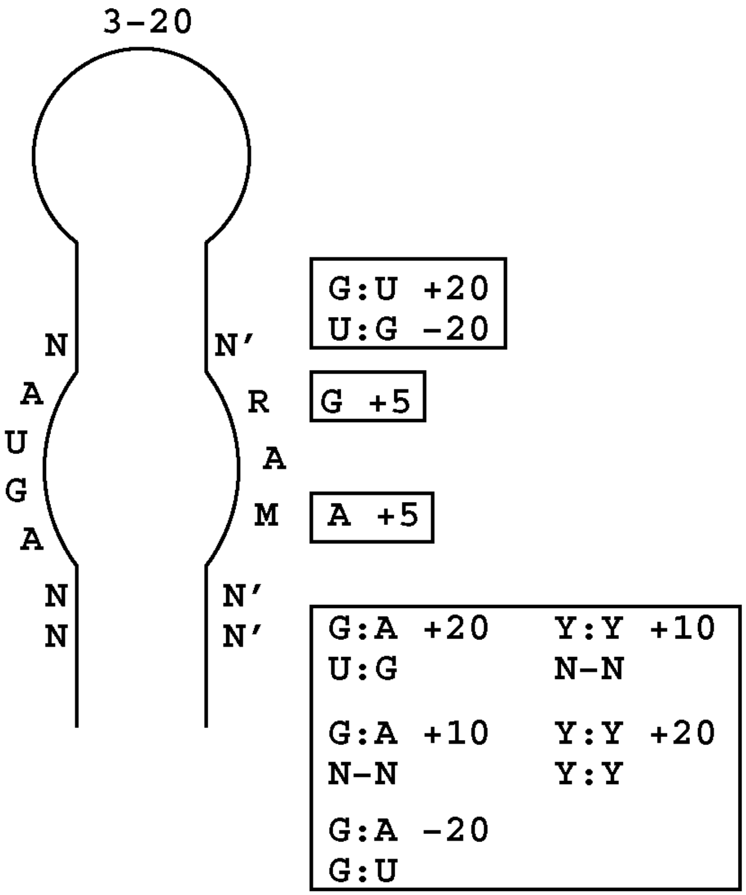
```
$ rnamotif -descr uncg.descr 16S_E_Coli.fa
uncg.descr: complete descr length: min/max = 8/12
#RM scored
#RM descr h5 ss h3
>rRNA
rRNA                0.000 0      206    8 cc ttcg gg
>rRNA
rRNA                0.000 0      339   12 ctcc tacg ggag
>rRNA
rRNA                0.000 0      340   10 tcc tacg gga
>rRNA
rRNA                0.000 0      341    8 cc tacg gg
>rRNA
rRNA                0.000 0      418    8 cc ttcg gg
>rRNA
rRNA                0.000 0     1027    8 cc ttcg gg
>rRNA
rRNA                0.000 0     1448    8 cc ttcg gg
```

GNRA

Allowing for Wobble (GU) base pairs

```
Parms
   wc +=gu;
descr
   h5( len=3 )
   ss( len=4,seq="GNRA” )
   h3
```

E-loop



G:U +20
U:G –20

G +5

A +5

G:A +20
U:G

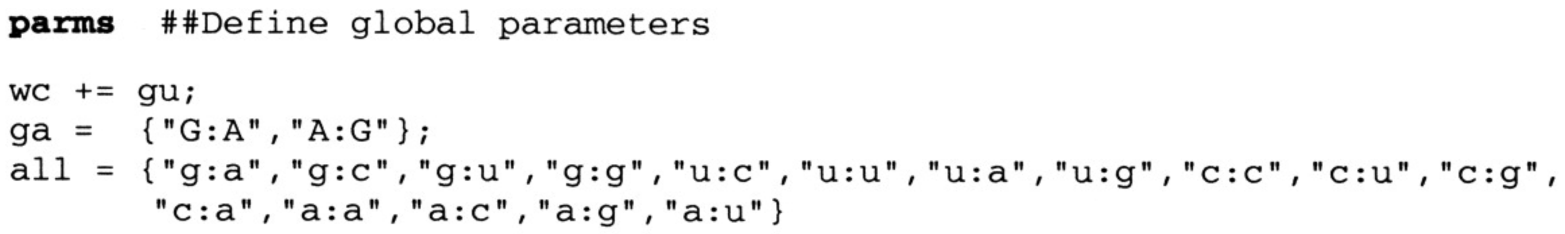
G:A +10
N–N

G:A –20
G:U

Y:Y +10
N–N

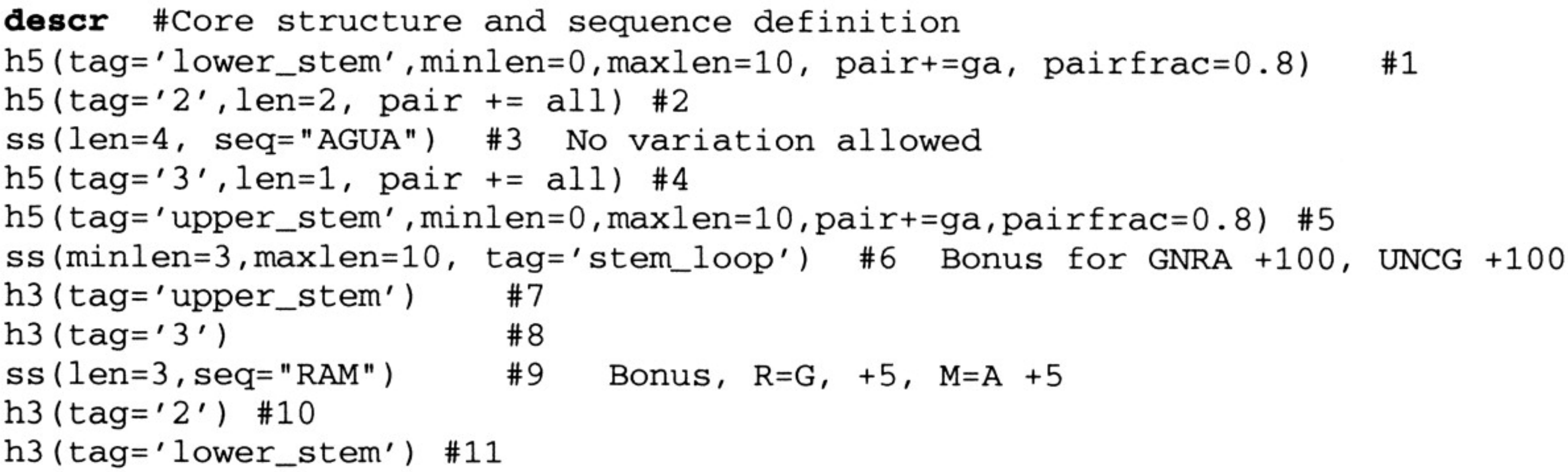
Y:Y +20
Y:Y

E-loop: defining new base pairs



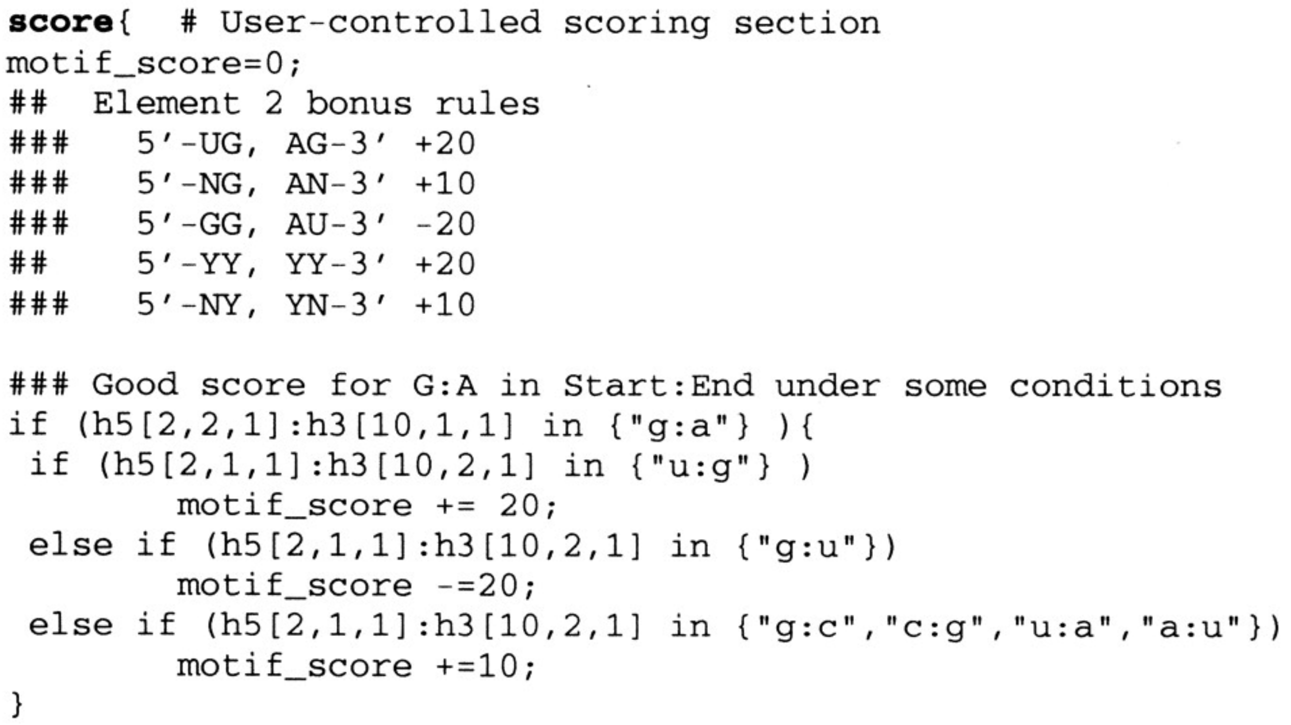
```
parms  ##Define global parameters

wc += gu;
ga =  {"G:A","A:G"};
all = {"g:a","g:c","g:u","g:g","u:c","u:u","u:a","u:g","c:c","c:u","c:g",
       "c:a","a:a","a:c","a:g","a:u"}
```

E-loop: pattern description



```
descr  #Core structure and sequence definition
h5(tag=’lower_stem’,minlen=0,maxlen=10, pair+=ga, pairfrac=0.8)   #1
h5(tag=’2’,len=2, pair += all) #2
ss(len=4, seq="AGUA")  #3  No variation allowed
h5(tag=’3’,len=1, pair += all) #4
h5(tag=’upper_stem’,minlen=0,maxlen=10,pair+=ga,pairfrac=0.8) #5
ss(minlen=3,maxlen=10, tag=’stem_loop’)  #6  Bonus for GNRA +100, UNCG +100
h3(tag=’upper_stem’)    #7
h3(tag=’3’)             #8
ss(len=3,seq="RAM")     #9   Bonus, R=G, +5, M=A +5
h3(tag=’2’) #10
h3(tag=’lower_stem’) #11
```

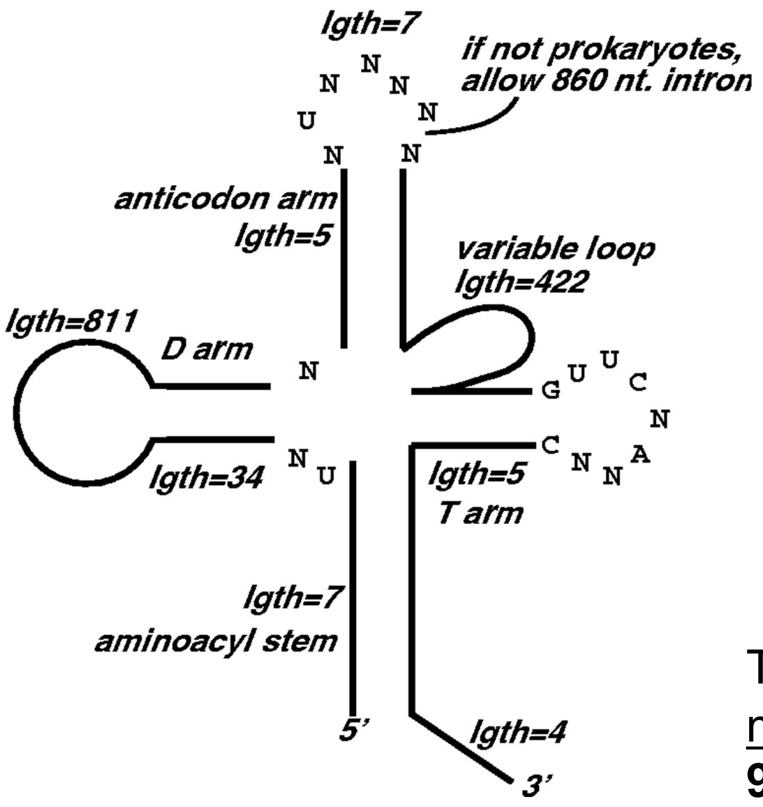
E-loop: score



```
score{   # User-controlled scoring section
motif_score=0;
##   Element 2 bonus rules
###    5'-UG, AG-3' +20
###    5'-NG, AN-3' +10
###    5'-GG, AU-3' -20
##     5'-YY, YY-3' +20
###    5'-NY, YN-3' +10

### Good score for G:A in Start:End under some conditions
if (h5[2,2,1]:h3[10,1,1] in {"g:a"} ){
 if (h5[2,1,1]:h3[10,2,1] in {"u:g"} )
        motif_score += 20;
 else if (h5[2,1,1]:h3[10,2,1] in {"g:u"})
        motif_score -=20;
 else if (h5[2,1,1]:h3[10,2,1] in {"g:c","c:g","u:a","a:u"})
        motif_score +=10;
}
else if( h5[2,2,1]:h3[10,1,1] in {"u:u","u:c","c:u","c:c"} ){
 if (h5[2,1,1]:h3[10,2,1] in {"u:u","u:c","c:u","c:c"})
        motif_score +=20;
 else if (h5[2,1,1]:h3[10,2,1] in {"g:c","c:g","u:a","a:u"})
        motif_score +=10;
}

##  Element 4  bonus rules
## Bonus GU +20, Penalty UG -20
if (h5[4,1,1]:h3[8,1,1] in {"g:u"})
        motif_score +=20;
else if (h5[4,1,1]:h3[8,1,1] in {"u:g"})
        motif_score -=20;


### Element 9 bonus rules
### Bonus M=A +5

if ( ss[9,3,1] =~ "a")
        motif_score +=5;

### Bonus R=G +5
if ( ss[9,1,1] =~ "g")
        motif_score +=5;

###Reject poor matches to the E-loop descriptor
if (motif_score < 0)
        REJECT;
SCORE = motif_score;
        }
```

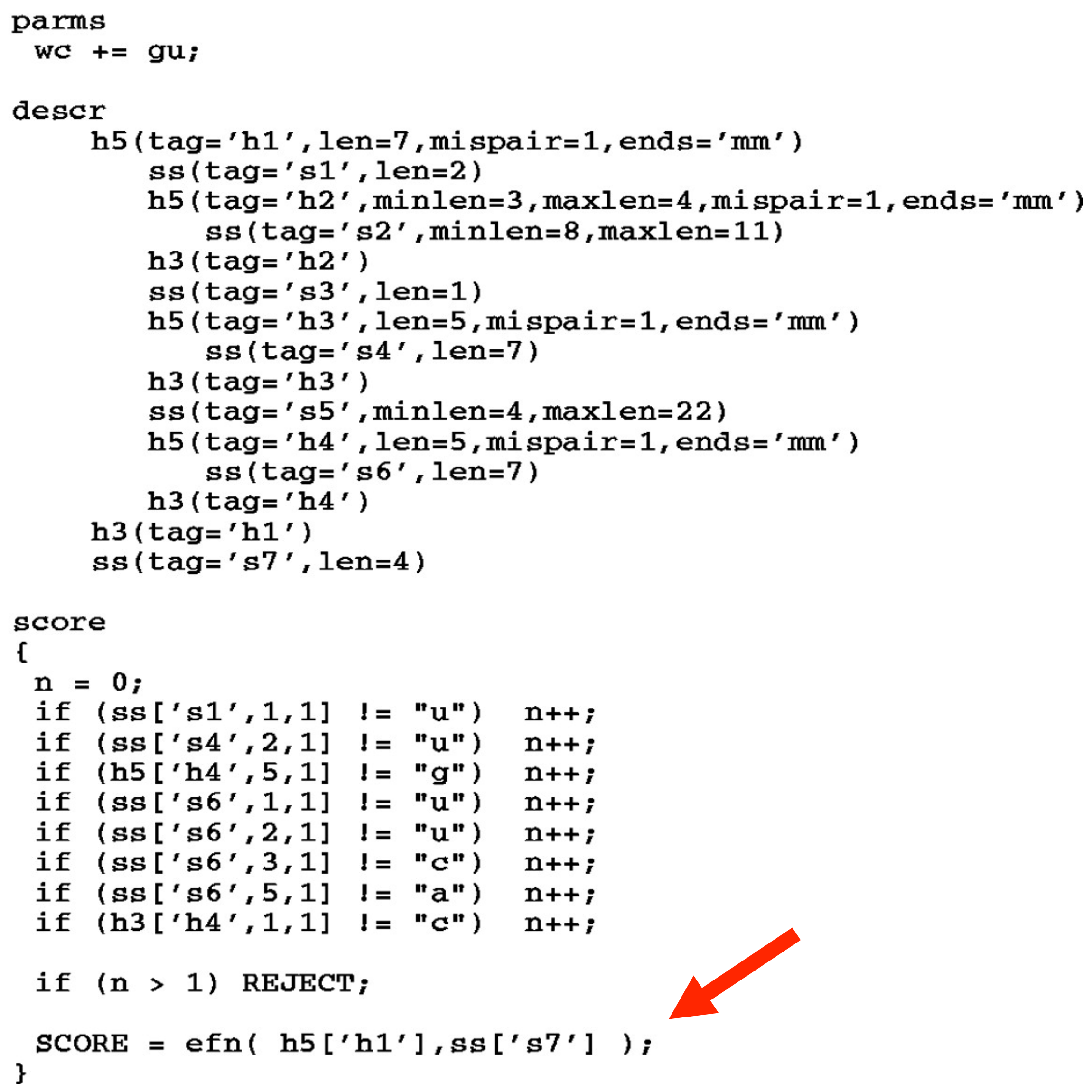
tRNA

A.



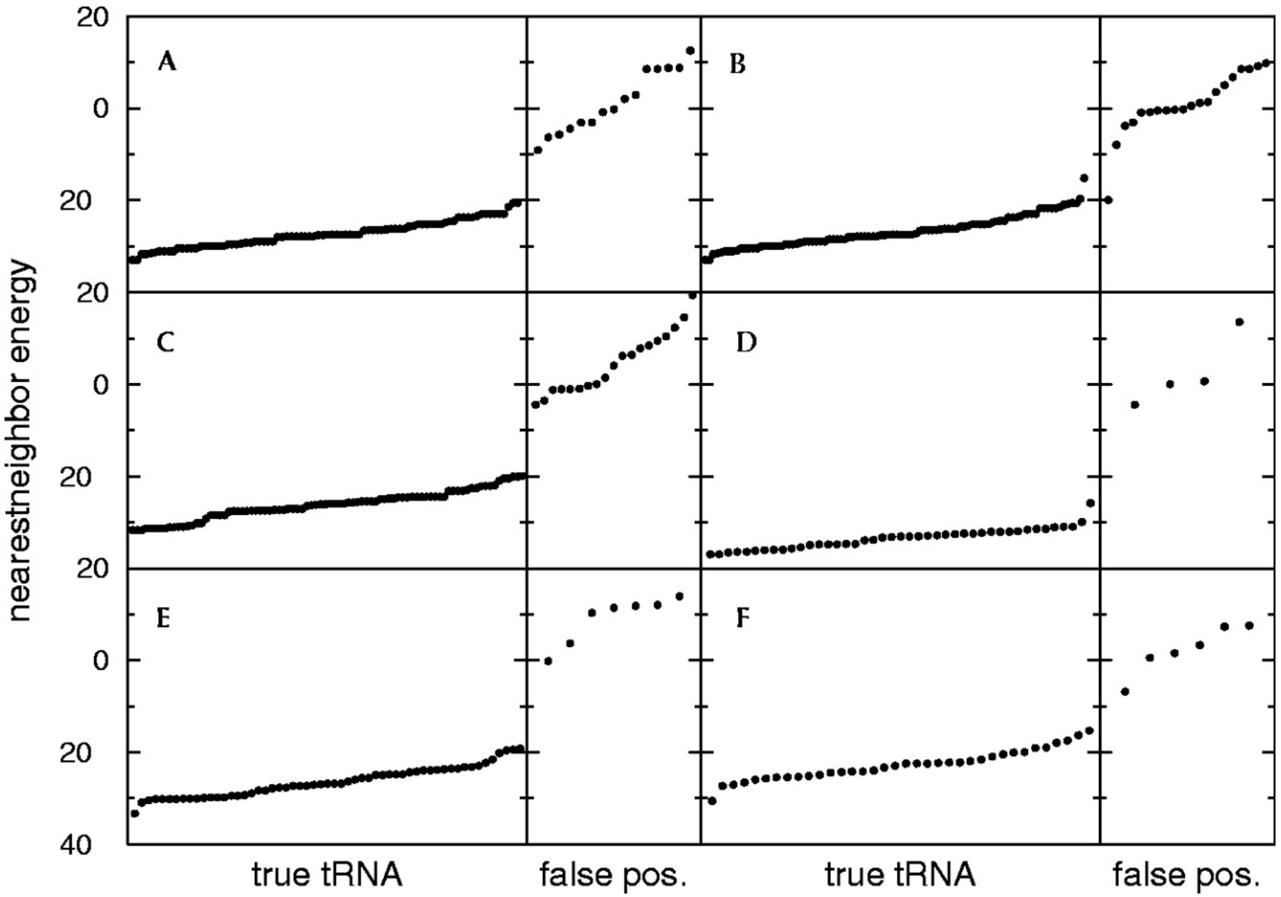
Tsui, Macke and Case (2003) A novel method for finding tRNA genes. *RNA* **9**:507-517.

B.



```
parms
 wc += gu;

descr
    h5(tag='h1',len=7,mispair=1,ends='mm')
       ss(tag='s1',len=2)
       h5(tag='h2',minlen=3,maxlen=4,mispair=1,ends='mm')
          ss(tag='s2',minlen=8,maxlen=11)
       h3(tag='h2')
       ss(tag='s3',len=1)
       h5(tag='h3',len=5,mispair=1,ends='mm')
          ss(tag='s4',len=7)
       h3(tag='h3')
       ss(tag='s5',minlen=4,maxlen=22)
       h5(tag='h4',len=5,mispair=1,ends='mm')
          ss(tag='s6',len=7)
       h3(tag='h4')
    h3(tag='h1')
    ss(tag='s7',len=4)

score
{
 n = 0;
 if (ss['s1',1,1] != "u")  n++;
 if (ss['s4',2,1] != "u")  n++;
 if (h5['h4',5,1] != "g")  n++;
 if (ss['s6',1,1] != "u")  n++;
 if (ss['s6',2,1] != "u")  n++;
 if (ss['s6',3,1] != "c")  n++;
 if (ss['s6',5,1] != "a")  n++;
 if (h3['h4',1,1] != "c")  n++;

 if (n > 1) REJECT;

 SCORE = efn( h5['h1'],ss['s7'] );
}
```

RNA “threading”



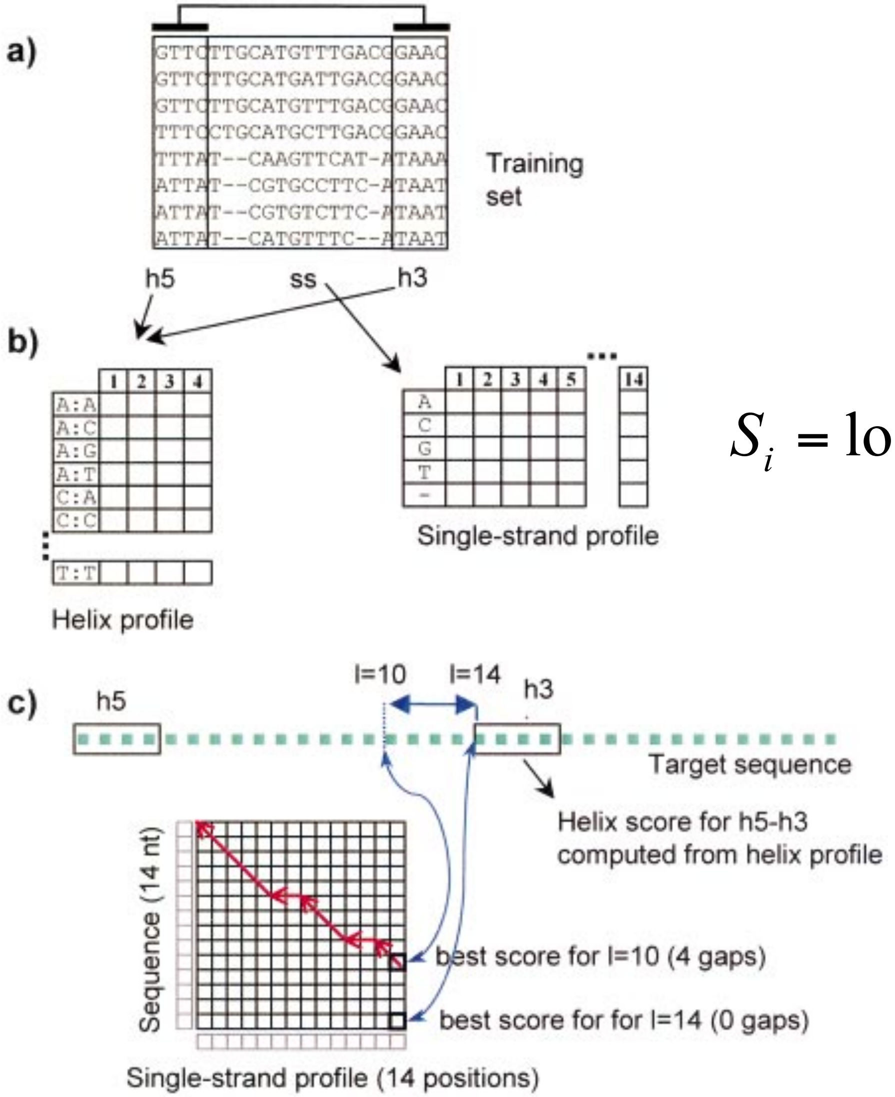
Recent Software Developments

- Profiles
  - ERPIN (Gautheret & Lambert, 2001)
- Stochastic Context-Free Grammars (SCFG)
  - Cove (Eddy & Durbin, 1994)
  - Rfam

ERPIN

- **Problem**: Pattern matchers, such as RNAMOT, are “hit of fail”;
- The solution to this problem for proteins has been to use profiles, which are a probabilistic representation of the sequence;
- ERPIN generalizes this idea to “structural” profiles.

$$S_{i:j} = \log \frac{O_{i:j}}{E_i \times E_j}$$



$$S_i = \log \frac{O_i}{E_i}$$

Gautheret & Lambert (2001) *JMB* **313**, 103-101.

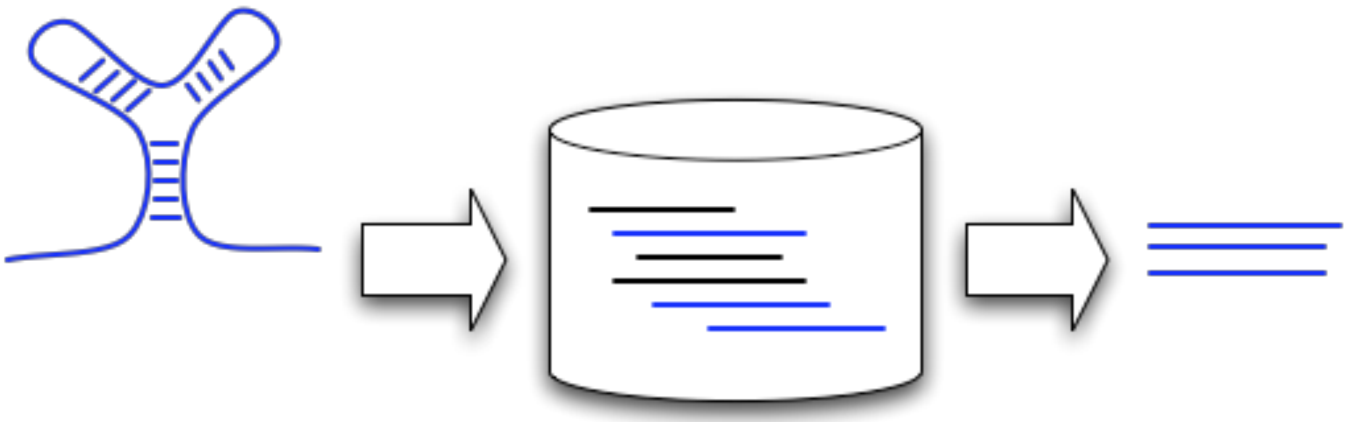
Remarks

- Limitation: gaps are not allowed in helical regions;
- Initial version only allows searching for one hairpin (Hp), one helix (Hx), one strand (St) or two helices (H2);
- Fast enough to scan entire genomes;
- Iterative search; *à la* PSI-BLAST;
- tRNA benchmark: sensitivity = 95%, 0.2 false positive per *E.coli* genome

RSEARCH

- R.J. Klein and S.R. Eddy (2003) RSEARCH: Finding homologs of single structured RNA sequences. BMC Bioinformatics 2003, 4:44 (doi:10.1186/1471-2105-4-44)
- **Input**: an RNA sequence and its secondary structure
- **Output**: similar RNAs on the basis of both primary sequence and secondary structure

RSEARCH (contd)



RSEARCH Input

```
# STOCKHOLM 1.0

#=GS Holley DE tRNA-Ala that Holley sequenced from Yeast genome

Holley
   GGGCGTGTGGCGTAGTCGGTAGCGCGCTCCCTTAGCATGGGAGAGGtCTCCGGTTCGATTCCGGACTCGTCCA
#=GR Holley SS
   (((((.(..((((........)))).(((((.......))))).....(((((.......)))))).))))).

//
```

RSEARCH (contd)

- RIBOSUM substitution matrices (analogous to residue substitution scores such as PAM and BLOSUM but for base pairs)
- Reports the statistical significance of all the matches
- Execution time is $O(NM^3)$ where N is the size of the database and M is the length of the input sequence
- **"(…) a typical single search of a metazoan genome may take a few thousand CPU hours"**

Specialized Programs: tRNAs

- tRNAscan-SE
  - tRNAscan and EufindtRNA identify candidates that are subsequently analyse by Cove.
  - 1 false positive per 15 billion nt
  - Detect 99% of true tRNA
  - www.genetics.wustl.edu/eddy/tRNAscan-SE/
  - rna.wustl.edu/GtRDB/ (Genomic tRNA database)
- FAStRNA (El-Mabrouk and Lisacek)
- tRNAscan (Fichant & Burks, 1991)

Specialized Programs: others

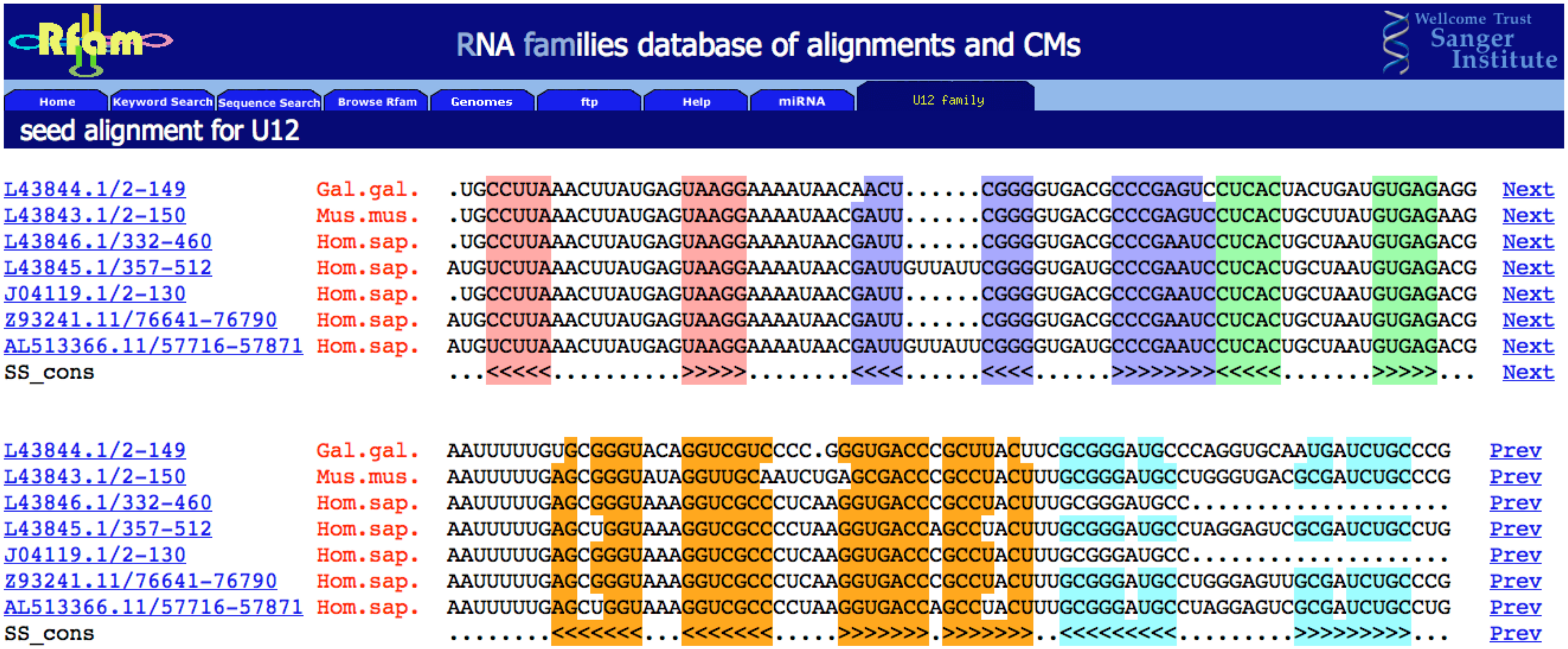
- tmRNA genes
  - BRUCE
  - Laslett, Canback, Andersson (2002) *NAR* **30**, 344903453.

Database search: summary

- Specialized programs: high specificity/sensitivity, fast
- SCFG-based approaches (such as INFERNAL): good specificity/sensitivity, work best if some sequence conservation is observed, slooow
- General motif searching tools (such as RNABOB): fast, writing descriptors is an art

RNA Motif Databases: Rfam

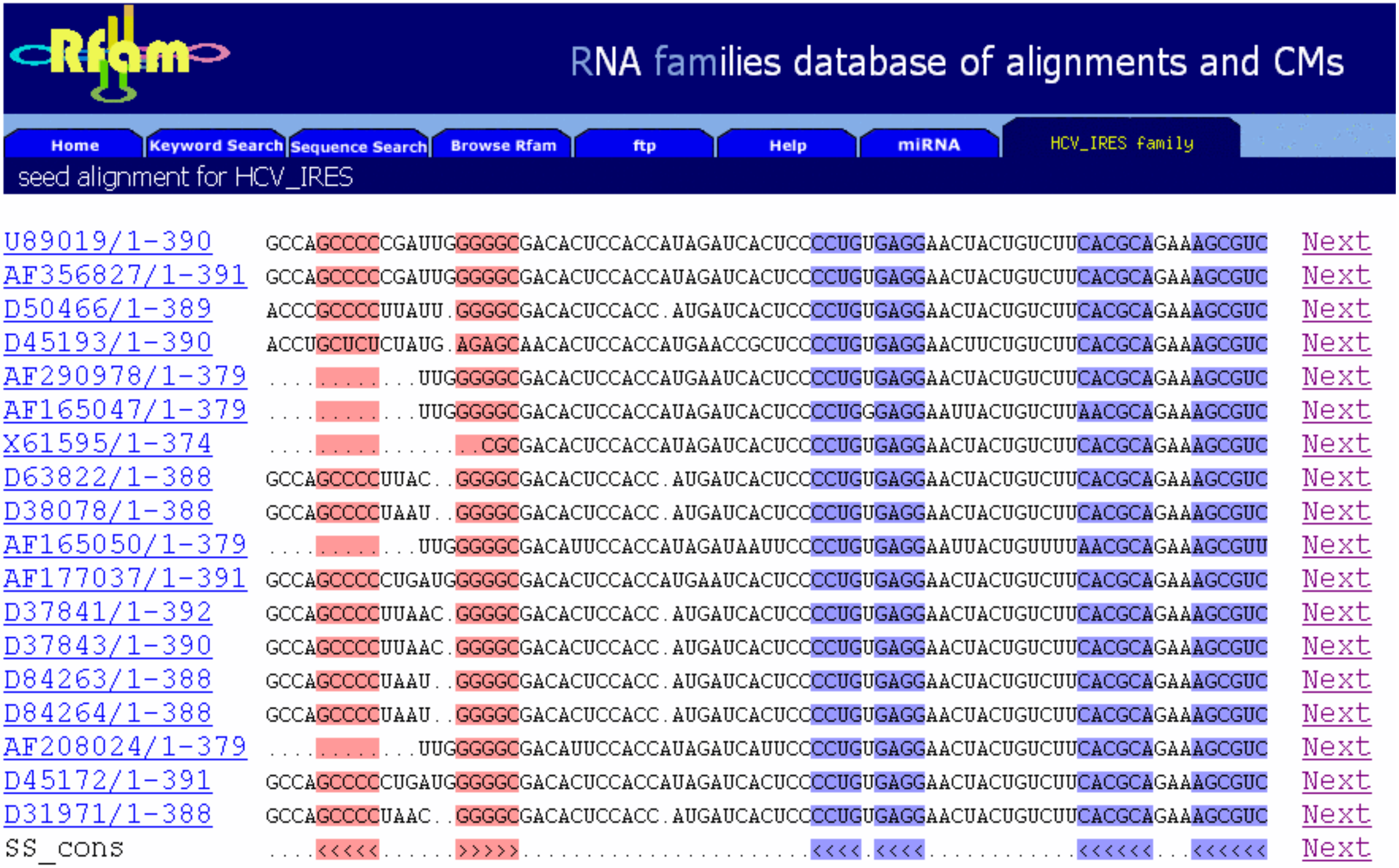
- A database of **multiple sequence alignments** and **covariance models**
- Rfam 9.1 contains 1372 families
- Search a query sequence to find instances of known motifs
- rfam.wustl.edu/ (database)
- infernal.wustl.edu/ (software)



RNA families database of alignments and CMs

seed alignment for U12

```
L43844.1/2-149           Gal.gal.  .UGCCUUAAACUUAUGAGUAAGGAAAAUAACAACU......CGGGGUGACGCCCGAGUCCUCACUACUGAUGUGAGAGG   Next
L43843.1/2-150           Mus.mus.  .UGCCUUAAACUUAUGAGUAAGGAAAAUAACGAUU......CGGGGUGACGCCCGAGUCCUCACUGCUUAUGUGAGAAG   Next
L43846.1/332-460         Hom.sap.  .UGCCUUAAACUUAUGAGUAAGGAAAAUAACGAUU......CGGGGUGACGCCCGAAUCCUCACUGCUAAUGUGAGACG   Next
L43845.1/357-512         Hom.sap.  AUGUCUUAAACUUAUGAGUAAGGAAAAUAACGAUUGUUAUUCGGGGUGAUGCCCGAAUCCUCACUGCUAAUGUGAGACG   Next
J04119.1/2-130           Hom.sap.  .UGCCUUAAACUUAUGAGUAAGGAAAAUAACGAUU......CGGGGUGACGCCCGAAUCCUCACUGCUAAUGUGAGACG   Next
Z93241.11/76641-76790    Hom.sap.  AUGCCUUAAACUUAUGAGUAAGGAAAAUAACGAUU......CGGGGUGACGCCCGAAUCCUCACUGCUAAUGUGAGACG   Next
AL513366.11/57716-57871  Hom.sap.  AUGUCUUAAACUUAUGAGUAAGGAAAAUAACGAUUGUUAUUCGGGGUGAUGCCCGAAUCCUCACUGCUAAUGUGAGACG   Next
SS_cons                            ...<<<<<..........>>>>>........<<<<......<<<<......>>>>>>>><<<<<.......>>>>>...   Next

L43844.1/2-149           Gal.gal.  AAUUUUUGUGCGGGUACAGGUCGUCCC.GGGUGACCCGCUUACUUCGCGGGAUGCCCAGGUGCAAUGAUCUGCCCG   Prev
L43843.1/2-150           Mus.mus.  AAUUUUUGAGCGGGUAUAGGUUGCAAUCUGAGCGACCCGCCUACUUUGCGGGAUGCCUGGGUGACGCGAUCUGCCCG   Prev
L43846.1/332-460         Hom.sap.  AAUUUUUGAGCGGGUAAAGGUCGCCCUCAAGGUGACCCGCCUACUUUGCGGGAUGCC....................   Prev
L43845.1/357-512         Hom.sap.  AAUUUUUGAGCUGGUAAAGGUCGCCCCUAAGGUGACCAGCCUACUUUGCGGGAUGCCUAGGAGUCGCGAUCUGCCUG   Prev
J04119.1/2-130           Hom.sap.  AAUUUUUGAGCGGGUAAAGGUCGCCCUCAAGGUGACCCGCCUACUUUGCGGGAUGCC....................   Prev
Z93241.11/76641-76790    Hom.sap.  AAUUUUUGAGCGGGUAAAGGUCGCCCUCAAGGUGACCCGCCUACUUUGCGGGAUGCCUGGGAGUUGCGAUCUGCCCG   Prev
AL513366.11/57716-57871  Hom.sap.  AAUUUUUGAGCUGGUAAAGGUCGCCCCUAAGGUGACCAGCCUACUUUGCGGGAUGCCUAGGAGUCGCGAUCUGCCUG   Prev
SS_cons                            ........<<<<<<<...<<<<<<<.....>>>>>>>.>>>>>>>..<<<<<<<<<.........>>>>>>>>>...   Prev
```



seed alignment for HCV_IRES

```
U89019/1-390     GCCAGCCCCCGAUUGGGGGCGACACUCCACCAUAGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
AF356827/1-391   GCCAGCCCCCGAUUGGGGGCGACACUCCACCAUAGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
D50466/1-389     ACCCGCCCCUUAUU.GGGGCGACACUCCACC.AUGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
D45193/1-390     ACCUGCUCUCUAUG.AGAGCAACACUCCACCAUGAACCGCUCCCCUGUGAGGAACUUCUGUCUUCACGCAGAAAGCGUC   Next
AF290978/1-379   ............UUGGGGGCGACACUCCACCAUGAAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
AF165047/1-379   ............UUGGGGGCGACACUCCACCAUAGAUCACUCCCCUGGGAGGAAUUACUGUCUUAACGCAGAAAGCGUC   Next
X61595/1-374     .................CGCGACACUCCACCAUAGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
D63822/1-388     GCCAGCCCCUUAC..GGGGCGACACUCCACC.AUGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
D38078/1-388     GCCAGCCCCUAAU..GGGGCGACACUCCACC.AUGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
AF165050/1-379   ............UUGGGGGCGACAUUCCACCAUAGAUAAUUCCCCUGUGAGGAAUUACUGUUUUAACGCAGAAAGCGUU   Next
AF177037/1-391   GCCAGCCCCCUGAUGGGGGCGACACUCCACCAUGAAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
D37841/1-392     GCCAGCCCCUUAAC.GGGGCGACACUCCACC.AUGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
D37843/1-390     GCCAGCCCCUUAAC.GGGGCGACACUCCACC.AUGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
D84263/1-388     GCCAGCCCCUAAU..GGGGCGACACUCCACC.AUGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
D84264/1-388     GCCAGCCCCUAAU..GGGGCGACACUCCACC.AUGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
AF208024/1-379   ............UUGGGGGCGACAUUCCACCAUAGAUCAUUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
D45172/1-391     GCCAGCCCCCUGAUGGGGGCGACACUCCACCAUAGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
D31971/1-388     GCCAGCCCCUAAC..GGGGCGACACUCCACC.AUGAUCACUCCCCUGUGAGGAACUACUGUCUUCACGCAGAAAGCGUC   Next
SS_cons          ....<<<<<......>>>>>.......................<<<<.<<<<............<<<<<<...<<<<<<   Next
```

RNA Motif Databases: UTRdb and UTRsite

Pesole G., Liuni S., Grillo G., Licciulli F., Mignone F., Gissi C., and Saccone C. - "*UTRdb and UTRsite: specialized database of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs.Update 2002"*. Nucleic Acids Res (2002), 30(1):335-340.

http://bighost.area.ba.cnr.it/BIG/UTRHome/

Specialized Motif Databases

- Methylation Guide snoRNA Database
  - snoscan (Lowe & Eddy, 1999)
  - http://rna.wustl.edu/snoRNAdb/
- tRNA databases
  - rna.wustl.edu/GtRDB/
- European Large Subunit Ribosomal RNA Database
- SRP database
- uRNA database
- Comparative RNA Web
- …

Summary

- Sequence alignment methods are not appropriate for comparing divergent RNA sequences
- Tools such as RNAMOT, RNABOB and RNAMOTIF allows to describe and find RNA structure motifs in sequence databases
- RSEARCH finds all the sequences having a similar sequence and secondary structure to that of an input sequence and structure
- Homologous sequences and structures can be represented as a covariance model. The software program INFERNAL allows to find all the sequences that are likely to share the same overall fold (secondary structure)

Rocket science is for kids
Bioinformatics is for scientists
CLC bio

uOttawa