BNF 5106. Bioinformatics

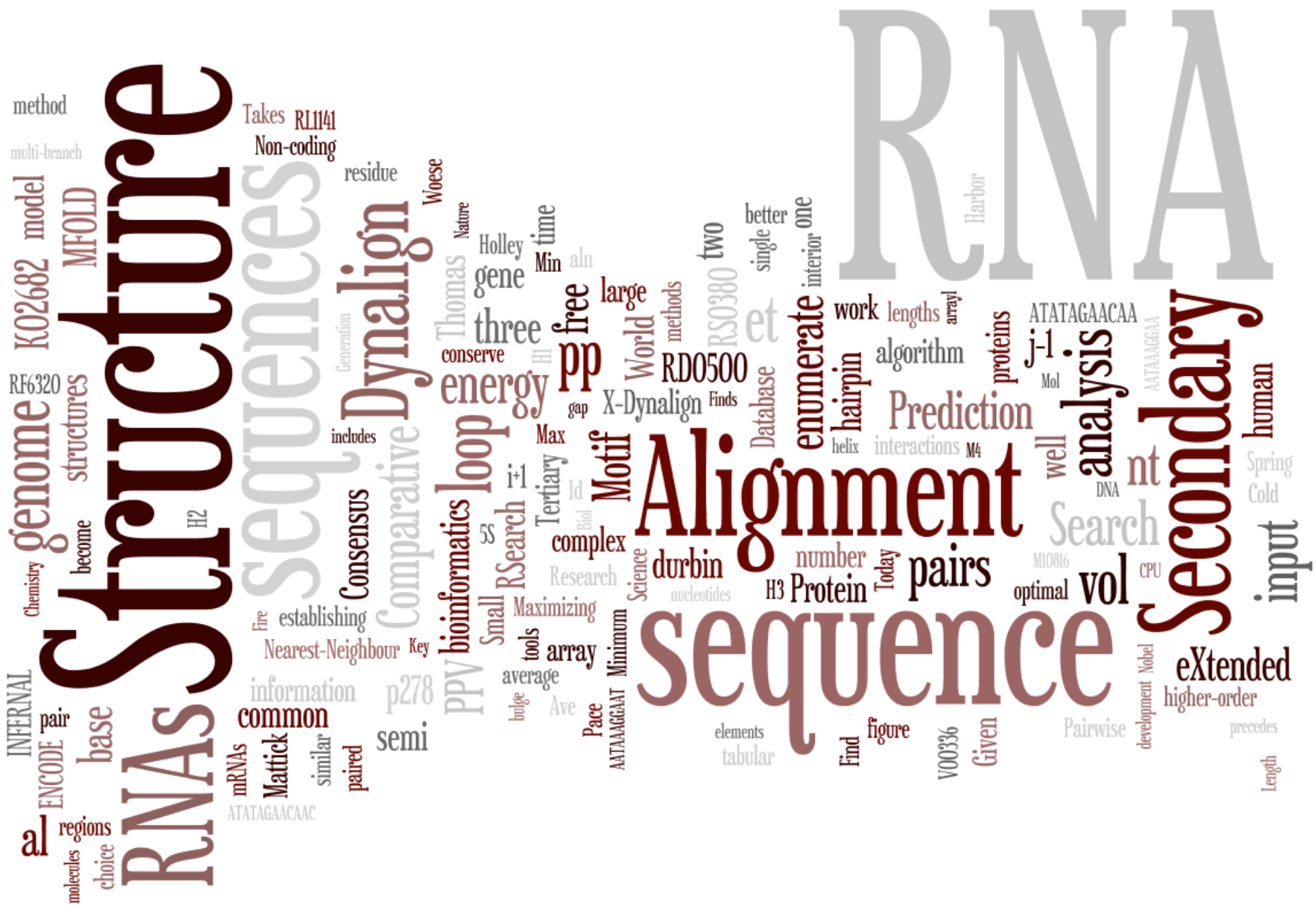
RNA Structure **Inference** & Database **Search**

Marcel Turcotte

uOttawa

School of Electrical Engineering and Computer Science (EECS)
University of Ottawa

Version November 4, 2019



http://www.site.uottawa.ca/~turcotte/teaching/bnf-5106/

1. **Preamble**
   - Outline
   - About Me
   - Take home message
2. **Introduction**
   - Key Discoveries
   - RNA Continent
   - Challenges for Traditional Bioinformatics Tools
3. **Inference**
   - Definitions
   - Comparative Sequence Analysis
   - Thermodynamics
   - Consensus
4. **Search**
   - First Generation: RNAMOT, RNABOB
   - Second Generation: RNAMOTIF
   - Third Generation: RSearch, INFERNAL

About Me

- 1989-95 Ph.D. **Université Montréal**
  Lapalme (CS), Cedergren (Biochemistry)
  RNA Tertiary Structure Prediction (MC-Sym)

About Me

- 1989-95 Ph.D. **Université Montréal**
  Lapalme (CS), Cedergren (Biochemistry)
  RNA Tertiary Structure Prediction (MC-Sym)
- 1995-97 Postdoc **University of Florida**
  Benner (Chemistry)
  Protein Secondary Structure Prediction

About Me

- 1989-95 Ph.D. **Université Montréal**
  Lapalme (CS), Cedergren (Biochemistry)
  RNA Tertiary Structure Prediction (MC-Sym)
- 1995-97 Postdoc **University of Florida**
  Benner (Chemistry)
  Protein Secondary Structure Prediction
- 1997-2000 Postdoc **Imperial Cancer Research Fund/UK**
  Sternberg (Biomolecular Modelling)
  Protein Fold Signature Discover & Machine Learning

About Me

- 1989-95 Ph.D. **Université Montréal**
  Lapalme (CS), Cedergren (Biochemistry)
  RNA Tertiary Structure Prediction (MC-Sym)
- 1995-97 Postdoc **University of Florida**
  Benner (Chemistry)
  Protein Secondary Structure Prediction
- 1997-2000 Postdoc **Imperial Cancer Research Fund/UK**
  Sternberg (Biomolecular Modelling)
  Protein Fold Signature Discover & Machine Learning
- 2000- Professor (Full) **University of Ottawa (EECS)**
  RNA Secondary Structure Prediction, Motif Inference, and Pattern Matching (eXtended-Dynalign, Profile-Dynalign, Seed), ACSEA, ModuleInducer, RiboFSM, MotifGP

Take home **message**

- With RNAs, **base pair patterns** are more preserved than sequence

Take home **message**

- With RNAs, **base pair patterns** are more preserved than sequence
- Consequently, **traditional bioinformatics tools** are generally not well adapted to RNA research

The **other** take home **message**

- “It is impossible to understand the biology of multicellular organisms without appreciation of the roles that small RNAs play.”

Neilson, J. R., & Sharp, P. A. (2008). Small RNA regulators of gene expression. *Cell*, **134**(6), 899–902. http://doi.org/10.1016/j.cell.2008.09.006

RNA **Catalyses** Reaction

- Ribozymes (**RNA enzymes**) discovery, early 1980s
- The Nobel Prize in Chemistry 1989

Thomas R. **Cech** (Colorado)

Sidney **Altman** [a][b] (Yale)

---

[a] Born in Montréal

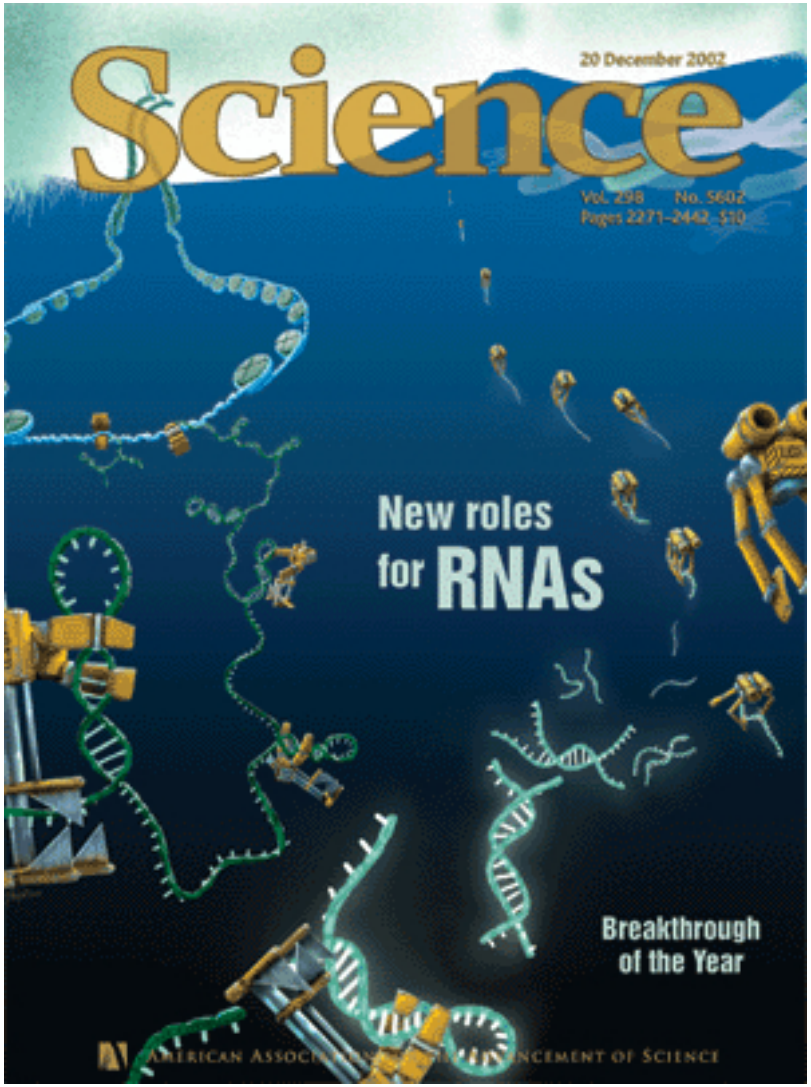
[b] Guest member uOttawa OISB

RNA **World**

- 1986, **RNA World hypothesis**
- RNA has the ability to store information, as DNA does
- RNA has the ability to catalyze reactions, as proteins do
- RNA is an ideal candidate for an earlier simple form of life

Walter **Gilbert**
(Nobel Prize in Chemistry 1980)

Small RNAs as 2002 Science Breakthrough



"Researchers are discovering that small RNA molecules play a surprising variety of key roles in cells. They can **inhibit translation of messenger RNA into protein**, **cause degradation of other messenger RNAs**, and **even initiate complete silencing of gene expression** from the genome."

RNA **Controls** Gene Expression

- The Nobel Prize in Physiology or Medicine 2006
- **RNA interference**, gene silencing by double-stranded RNA
- An other key protein function

Andrew Z. **Fire**
(Stanford)

Craig C. **Mello**
(Massachusetts Medical School)

The Nobel Prize in Chemistry 2009

“for studies of the structure and function of the ribosome”

Venkatraman **Ramakrishnan**

MRC Laboratory of Molecular Biology

Cambridge, United Kingdom

Thomas A. **Steitz**

Yale University, Howard Hughes

Medical Institute New Haven, CT, USA

Ada E. **Yonath**

Weizmann Institute of Science

Rehovot, Israel

Non coding RNA and cancer

- Huang, P. et al. lncRNA profile study reveals the mRNAs and lncRNAs associated with docetaxel resistance in breast cancer cells. Sci Rep 8, 17970 (2018).
- Li, Y. et al. Extracellular Vesicles Long RNA Sequencing Reveals Abundant mRNA, circRNA, and lncRNA in Human Blood as Potential Biomarkers for Cancer Diagnosis. Clin. Chem. 65, 798–808 (2019).
- Gao, Y. et al. Lnc2Cancer v2.0: updated database of experimentally supported long non-coding RNAs in human cancers. Nucleic Acids Res 47, D1028–D1033 (2019).
- Xu, S., Kong, D., Chen, Q., Ping, Y. & Da Pang. Oncogenic long noncoding RNA landscape in breast cancer. Molecular Cancer 2017 16:1 16, 129 (2017).

Nobel Laureates in RNA biology

- The Wikipedia page belows lists **31 Nobel laureates in RNA biology**:
  - `https://en.wikipedia.org/wiki/History_of_RNA_biology`

Non-Protein-Coding RNA (**ncRNA**) in Protein Synthesis

tRNA: transfer RNAs are adapter molecules that recognize mRNA codons and carry a specific amino acid

rRNA: ribosomal RNAs account for 2/3 of the molecular mass of the ribosome, which is a large RNA/protein complex responsible for translating genomic information (stored in mRNAs) into proteins

Cech, T. R., & Steitz, J. A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. **Cell**, **157**(1), 77–94. http://doi.org/10.1016/j.cell.2014.03.008

Non-coding RNAs

miRNA: microRNAs modulate the development in C. elegans, Drosophila, and mammals ($\sim$20 nt)

snRNA: small nuclear RNAs are involved in splicing of eukaryotic mRNAs ($\sim$200 nt)

snoRNA: small nucleolar RNA direct nucleotide modifications in rRNAs ($\sim$100 nt)

gRNA: guide RNAs play an important role in editing of certain mRNAs in trypanosomes ($\sim$ 70 nt)

Non-Coding RNAs (contd)

tmRNA: have the combined features of tRNAs and mRNAs and plays a role in translation regulation in bacterial genomes ($\sim$400 nt)

SRP: (signal recognition particle RNA-protein complex) directs newly synthesized proteins through the endoplasmic reticulum

M1 RNA: is the catalytic part of Ribonuclease P in bacteria, involves in the maturation of pre-tRNA ($\sim$375 nt)

TERC: telomerase RNA is an integral part of telomerase enzyme that serves as a template for the synthesis of the telomeres ($\sim$450 nt)

...

Rfam database

- **Rfam** 14.1 (January 2019) contains **3,016** RNA families
- For each family
  - Multiple sequence alignment (seed, full)
  - Consensus secondary structure (from literature or predicted)
  - Covariance model
- rfam.org
- Kalvari, I. et al. Non-Coding RNA Analysis Using the Rfam Database. *Curr Protoc Bioinformatics* **62**, e51 (2018).
- Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46**, D335–D342 (2018).

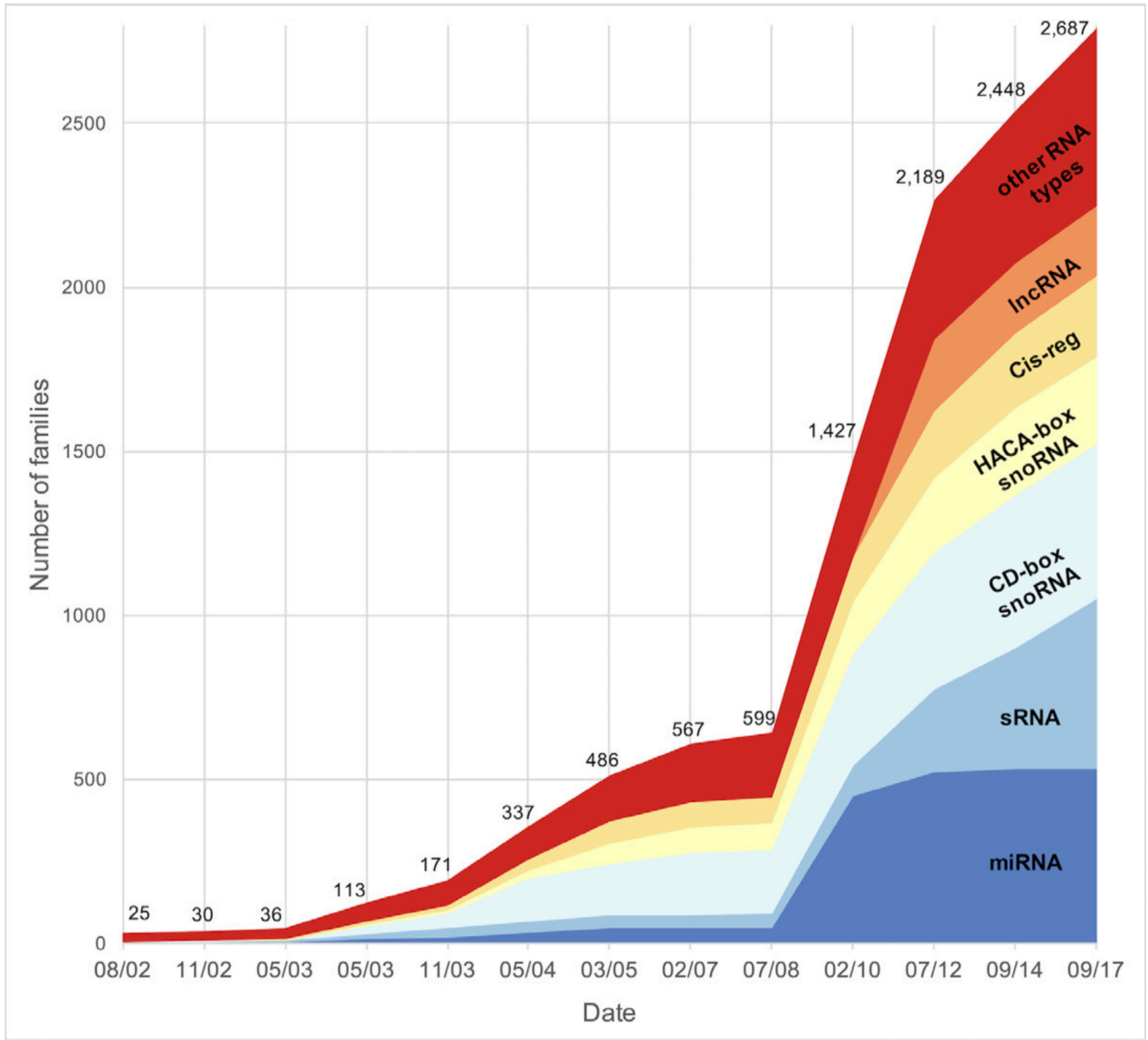


Figure 1. Growth in the number of RNA families grouped by RNA type in major database releases. The *other RNA types* group includes types with less than 50 families, such as rRNA, tRNA, snRNA or riboswitches.

RNACentral — rnacentral.org

- The RNAcentral Consortium. (2017). RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Research*, **45**(D1), D128–D134. http://doi.org/10.1093/nar/gkw1008
- Petrov, A. I., Kay, S. J. E., Gibson, R., Kulesha, E., Staines, D., Bruford, E. A., et al. (2015). RNAcentral: An international database of ncRNA sequences. *Nucleic Acids Research*, **43**(D1), D123–D129. http://doi.org/10.1093l/narlgku991
- Bateman, A., Agrawal, S., Birney, E., Bruford, E. A., Bujnicki, J. M., Cochrane, G., et al. (2011). RNAcentral: A vision for an international database of RNA sequences. *RNA*, **17**(11), 1941–1946. http://doi.org/10.1261/rna.2750811

ENCODE

- "The human genome is pervasively transcribed, such that **the majority of its bases are associated with at least one primary transcript** (…)"

Birney et al. *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature (2007) vol. **447** (7146) pp. 799-816

How Many **Non-Coding RNAs**?

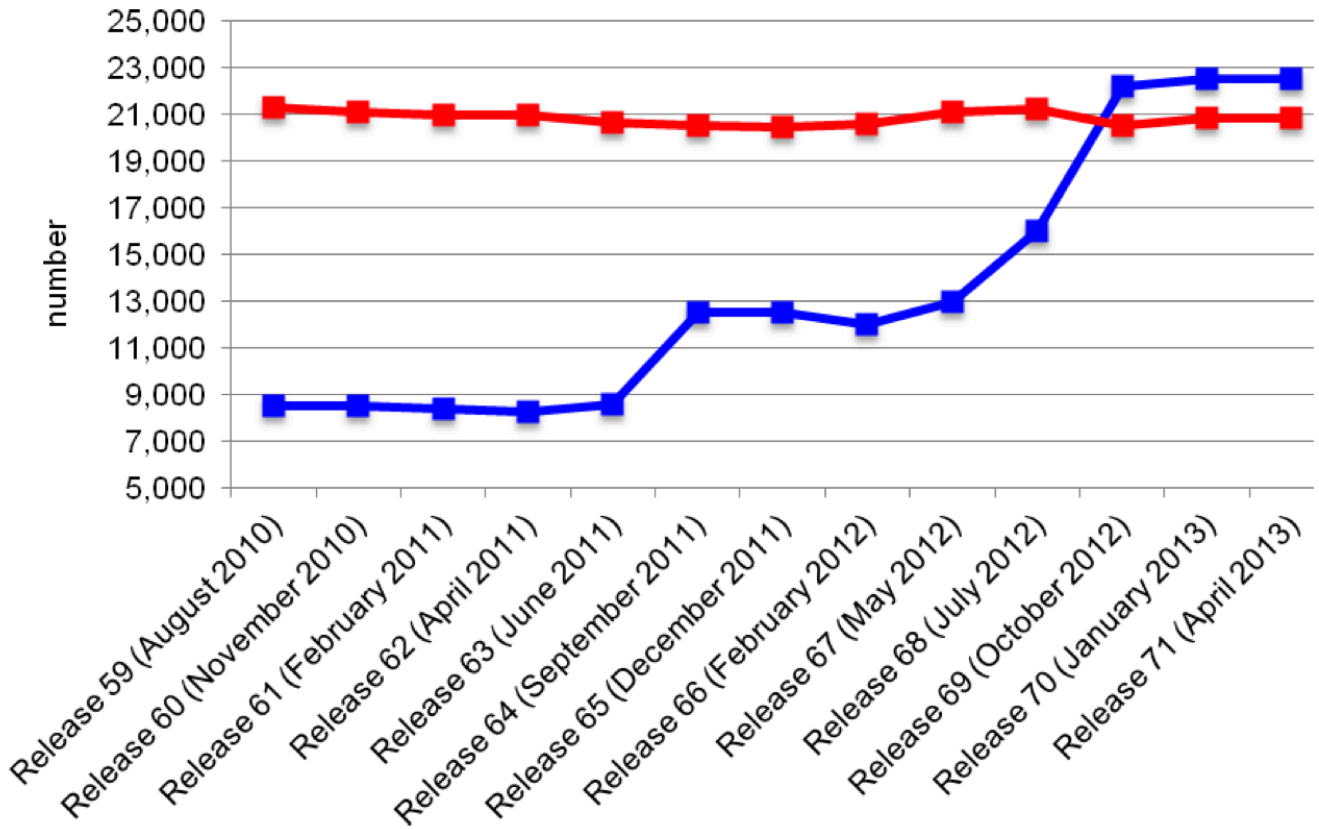
- **48,479** candidates in the human genome (**EvoFold**)
  Pedersen et al. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol (2006) vol. 2 (4) pp. e33
- Studies based on the **ENCODE** data set
  - **3,267 RNAz**, **3,134 EvolFold**
    Washietl et al. Structured RNAs in the ENCODE selected regions of the human genome. Genome Res (2007) vol. 17 (6) pp. 852-64
  - **4,933 CMfinder**
    Torarinsson et al. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. Genome Res (2008) vol. 18 (2) pp. 242-51

How Many **Non-Coding RNAs**?

- Deveson, I. W., Hardwick, S. A., Mercer, T. R., & Mattick, J. S. (2017). The Dimensions, Dynamics, and Relevance of the Mammalian Noncoding Transcriptome. *Trends in Genetics*, **33**(7), 464–478. `http://doi.org/10.1016/j.tig.2017.04.004`

Protein versus **ncRNA** annotations

Figure 4. Number of non-coding and protein-coding genes annotated over the last Ensembl releases. The *x*-axis indicates the number and the date of the release. The vertical axis reports the number of ncRNA (blue line) and protein-coding genes (red line).



- Bussotti,G. et al. (2013) Detecting and comparing non-coding RNAs in the high-throughput era. *Int J Mol Sci,* 14, 15423-15458.

Action **Mechanisms**

- direct **base-pairing** with RNA or DNA target: snoRNAs, miRNAs
- **mimic the structure** of other nucleic acids (or proteins?): tmRNA, some snRNAs, IRES
- **catalyst**: RNAs P

John S. **Mattick**

- Around 500 publications, **66,740 citations**!
- Over **150 co-authors**, h-index $=$ 116 (Google Scholar)
- **Garvan Institute of Medical Research**, Australia, Sydney

John S. Mattick (contd)

- Morris,K.V. and Mattick,J.S. (2014) **The rise of regulatory RNA**. Nat Rev Genet, 15, 423–437.

"it seems that RNA is the **computational engine** of cell biology, developmental biology, brain function and perhaps even evolution itself. The complexity and interconnectedness of these systems should not be cause for concern but rather the motivation for **exploring the vast unknown universe of RNA regulation, without which we will not understand biology**."

Smith et al. (2013) **Widespread purifying selection on RNA structure in mammals**. Nucleic Acids Res, 41, 8220-8236.
Amaral,P.P. et al. (2008) **The eukaryotic genome as an RNA machine**. Science, 319, 1787–1789.
Mattick et al. **Non-coding RNA**. Hum Mol Genet (2006) vol. 15 Spec No 1 pp. R17-29
Carninci et al. **The transcriptional landscape of the mammalian genome**. Science (2005) vol. 309 (5740) pp. 1559-63
Mattick. **RNA regulation: a new genetics?**. Nat Rev Genet (2004) vol. 5 (4) pp. 316-23
Mattick. **Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms.** Bioessays (2003) vol. 25 (10) pp. 930-9
Mattick et al. **The evolution of controlled multitasked gene networks**: the role of introns and other noncoding RNAs in the development of complex organisms. Mol Biol Evol (2001) vol. 18 (9) pp. 1611-30

Fascinating **RNAs**

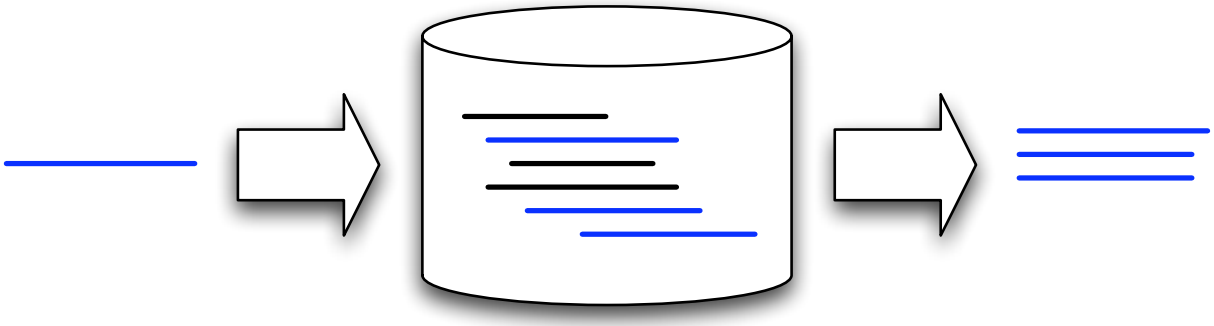
- **Versatile** molecules that can **carry information**, as DNA does, and perform **catalytic functions**, as proteins do

Fascinating **RNAs**

- **Versatile** molecules that can **carry information**, as DNA does, and perform **catalytic functions**, as proteins do
- Seem to be governed by simpler laws, as a result RNA analysis is a big **bioinformatics success** (see Gutell's work on predicting secondary and tertiary interactions, and Major's work on predicting tertiary structure)

Database Search Problem

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



```
>gi|27867582(fragment of the known Clostridium botuninum toxin gene)
GTGAATCAGCACCTGGACTTTCAGATGAAAAATTAAATTTAACTATCCAAAATGATGCTTATATACCAAAATATGATTCTAATGGAACAA
GTGATATAGAACAACATGATGTTAATGAACTTAATGTATTTTTCTATTTAGATGCACAGAAAGTGCCCGAAGGTGAAAATAATGTCAATC
TCACCTCTTCAATTGATACAGCATTATTAGAACAACCTAAAATATATACATTTTTTTCATCAGAATTTATTAATAATGTCAATAAACCTG
TGCAAGCAGC
```

Database Search Problem

```
>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?

Pairwise Sequence **Alignment**

An **optimal pairwise 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**)

Pairwise Sequence **Alignment**

`aln( ATATAGAACAAC, AATAAAGGAAT )` is the maximum of:

Pairwise Sequence **Alignment**

`aln( ATATAGAACAAC, AATAAAGGAAT )` is the maximum of:

- `aln( ATATAGAACAA, AATAAAGGAA )` + sub(C,T)

```
ATATAGAACAA C
AATAAAGGAA  T
```

Pairwise Sequence **Alignment**

`aln( ATATAGAACAAC, AATAAAGGAAT )` is the maximum of:

- `aln( ATATAGAACAA, AATAAAGGAA )` + sub(C,T)

```
ATATAGAACAA C
AATAAAGGAA  T
```

- `aln( ATATAGAACAA, AATAAAGGAAT )` + del(C)

```
ATATAGAACAA C
AATAAAGGAAT -
```

Pairwise Sequence **Alignment**

`aln( ATATAGAACAAC, AATAAAGGAAT )` is the maximum of:

- `aln( ATATAGAACAA, AATAAAGGAA )` + sub(C,T)

```
ATATAGAACAA C
AATAAAGGAA  T
```

- `aln( ATATAGAACAA, AATAAAGGAAT )` + del(C)

```
ATATAGAACAA C
AATAAAGGAAT -
```

- `aln( ATATAGAACAAC, AATAAAGGA )` + ins(T)

```
ATATAGAACAAC -
AATAAAGGAA   T
```

Assumptions

Assumptions

- Positions along the sequence are independent and **identically distributed** *i.i.d.*

Assumptions

- Positions along the sequence are independent and **identically distributed** *i.i.d.*
- **Independence** is necessary for the development of efficient exact (Smith-Waterman) or heuristics (such as BLAST) algorithms

Assumptions

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

RNA Sequence Alignment (**Toy Example**)

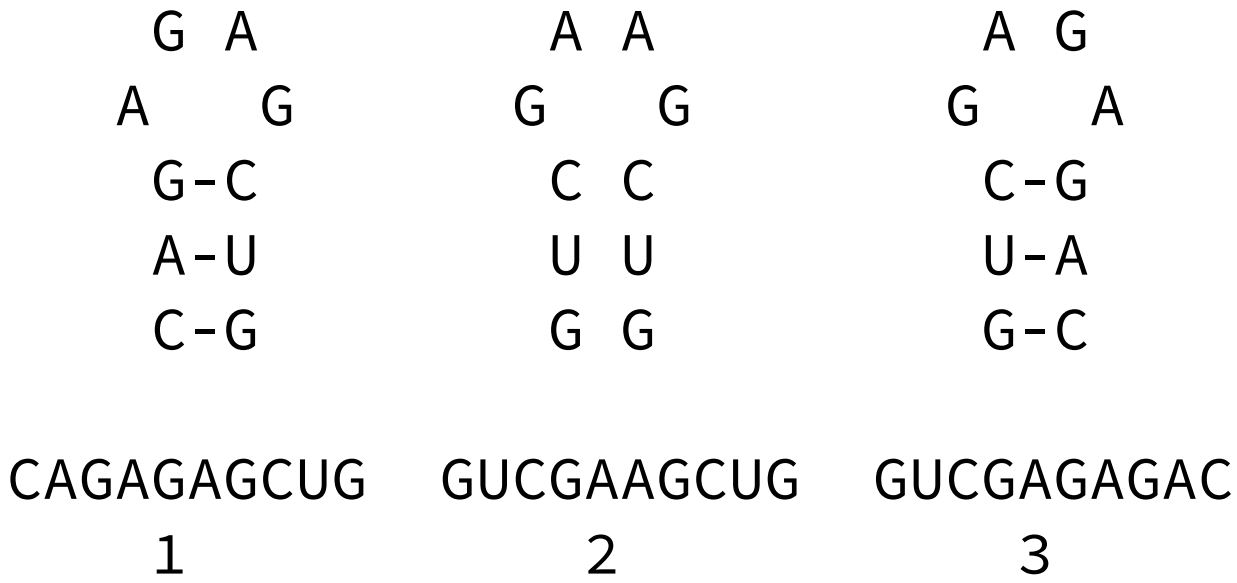
```
1   GUCGAGAGAC
    !!!!!
2   GUCGAAGCUG
          !!!!!
3   CAGAGAGCUG
```

RNA Sequence Alignment (**Toy Example**)

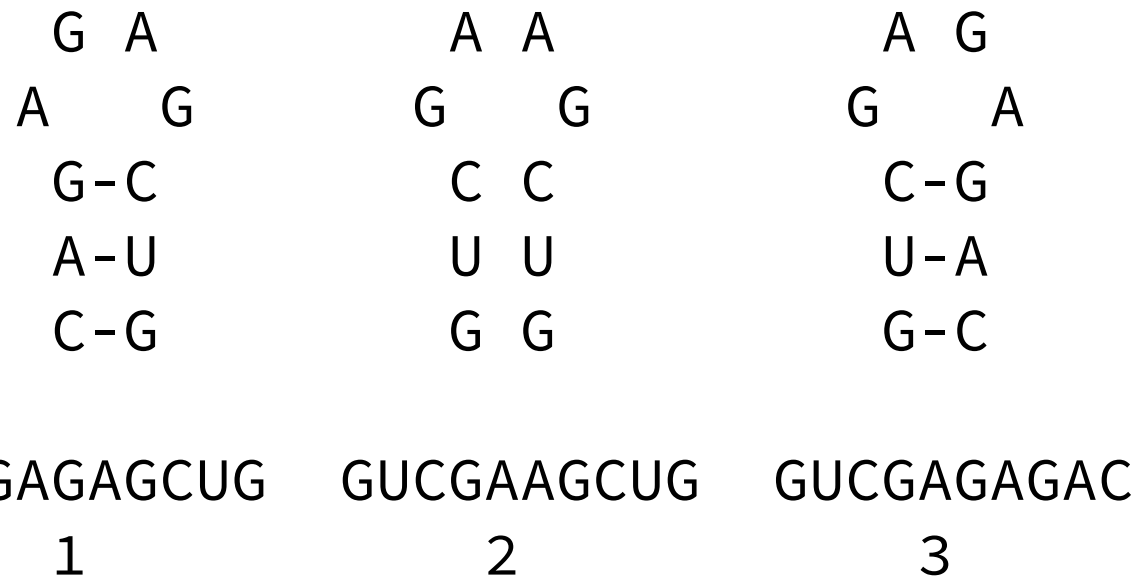
```
1    GUCGAGAGAC
     ! ! ! ! !
2    GUCGAAGCUG
           ! ! ! ! !
3    CAGAGAGCUG
```

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

RNA Sequence Alignment (**Toy Example**)

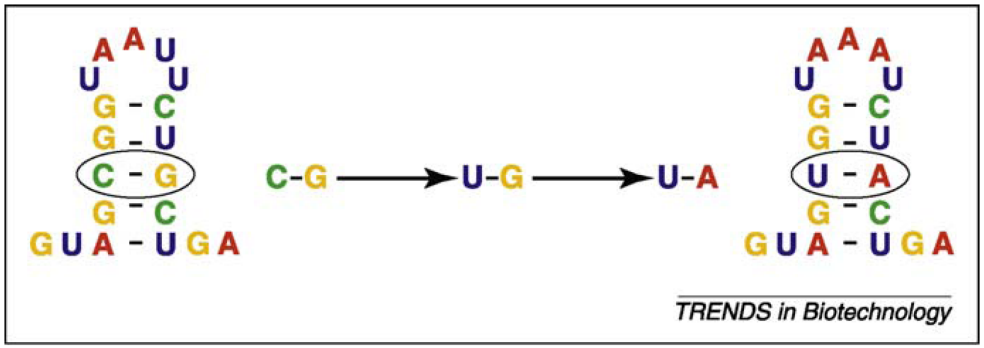


RNA Sequence Alignment (**Toy Example**)



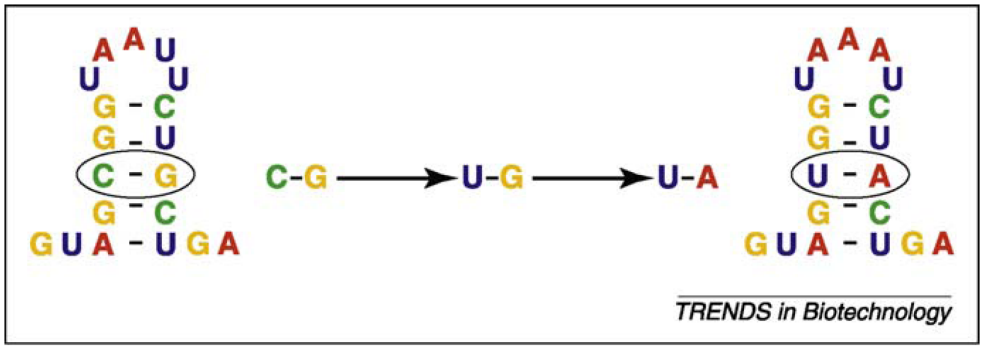
Yes, but sequences 1 and 3 share the same secondary structure!

Caveat



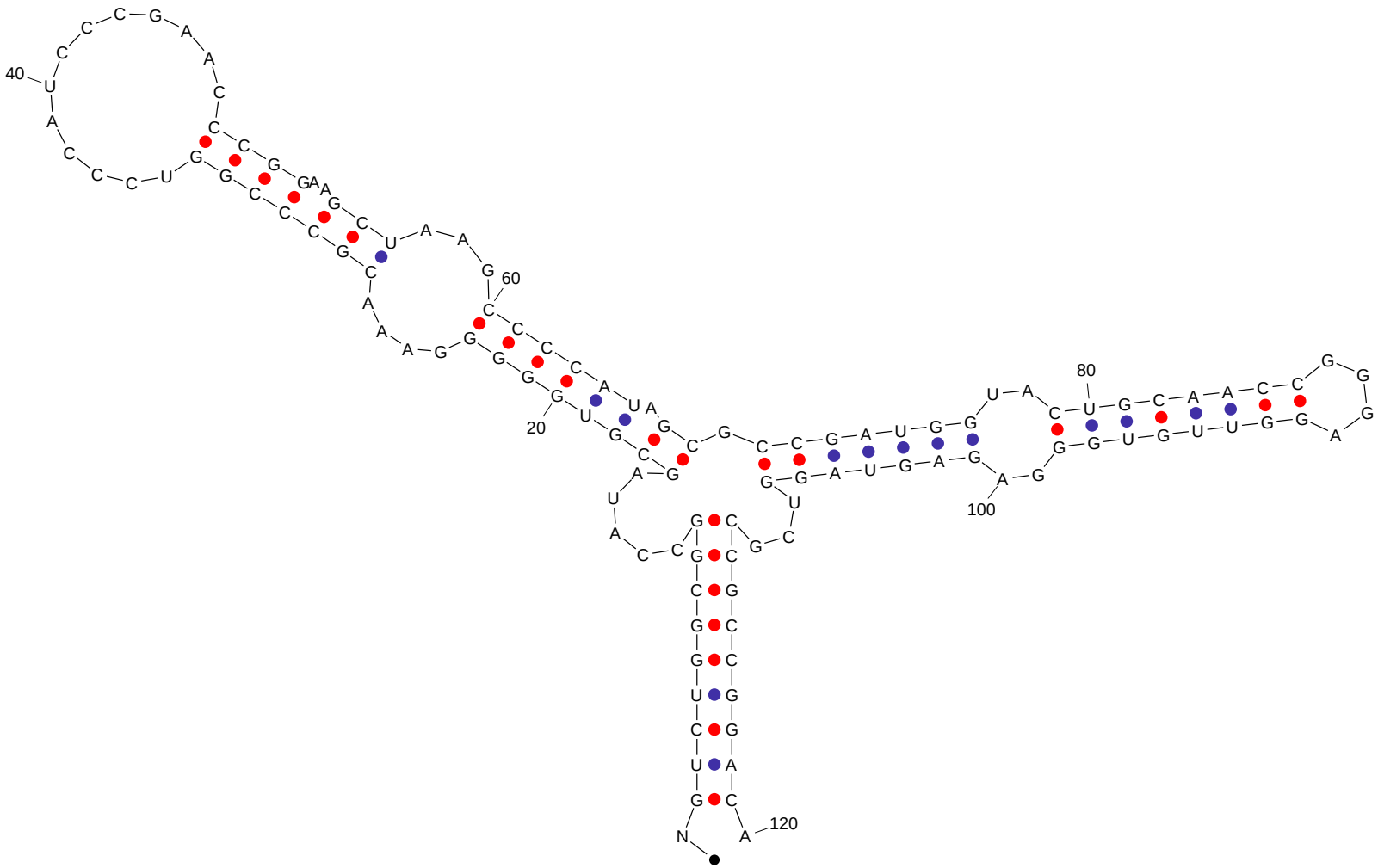
- **RNAs conserve secondary structure interactions more than they conserve their sequence**

Caveat



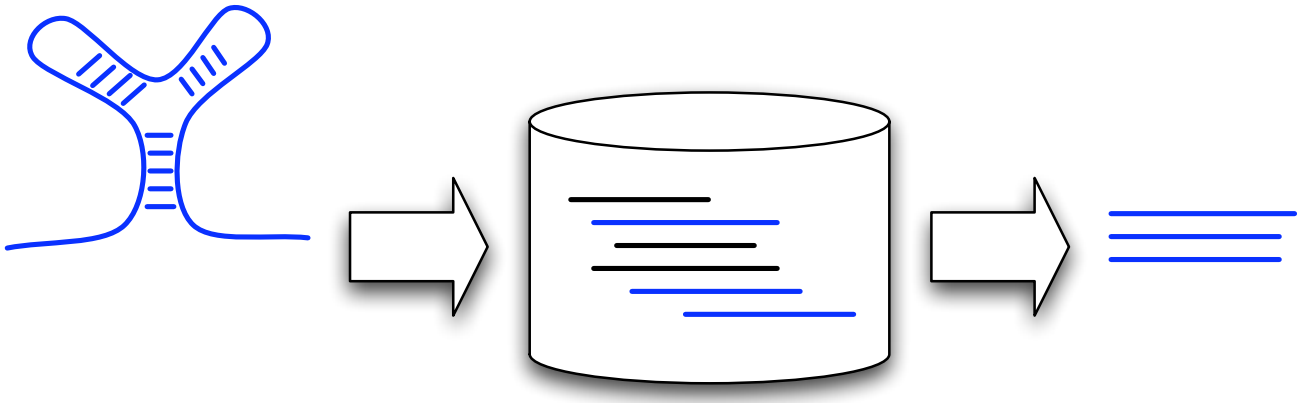
- **RNAs conserve secondary structure interactions more than they conserve their sequence**
- Traditional bioinformatics tools, **assuming that positions are independent**, perform poorly

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



Database Search Problem

Find all sequences containing a user specified motif or **all the sequences that can be folded into a user specified structure**



Secondary Structure Prediction

- X ray **crystallography** and **N.M.R.**

Secondary Structure Prediction

- X ray **crystallography** and **N.M.R.**
- Chemical and enzymatic **probing**, **cross-linking**

Secondary Structure Prediction

- X ray **crystallography** and **N.M.R.**
- Chemical and enzymatic **probing**, **cross-linking**
- **Comparative sequence analysis**

Secondary Structure Prediction

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

Secondary Structure Prediction

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

Given an **RNA sequence** $S = s_1, s_2 \ldots s_n$, where s_i is the ith nucleotide. A **secondary structure** is an ordered list of pairs, $i.j$, $1 \leq i < j \leq n$ such that:

Given an **RNA sequence** $S = s_1, s_2 \ldots s_n$, where s_i is the ith nucleotide. A **secondary structure** is an ordered list of pairs, $i.j$, $1 \leq i < j \leq n$ such that:

- $j - i \geq c$, where $c = 4$ for instance

Given an **RNA sequence** $S = s_1, s_2 \ldots s_n$, where s_i is the ith nucleotide. A **secondary structure** is an ordered list of pairs, $i.j$, $1 \leq i < j \leq n$ such that:

- $j - i \geq c$, where $c = 4$ for instance
- Given $i.j$ and $i'.j'$, two base pairs, then either:

Given an **RNA sequence** $S = s_1, s_2 \ldots s_n$, where s_i is the ith nucleotide. A **secondary structure** is an ordered list of pairs, $i.j$, $1 \leq i < j \leq n$ such that:

- $j - i \geq c$, where $c = 4$ for instance
- Given $i.j$ and $i'.j'$, two base pairs, then either:
  - $i = i'$ and $j = j'$ (they are the same)

Given an **RNA sequence** $S = s_1, s_2 \ldots s_n$, where s_i is the ith nucleotide. A **secondary structure** is an ordered list of pairs, $i.j$, $1 \leq i < j \leq n$ such that:

- $j - i \geq c$, where $c = 4$ for instance
- 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 $i'.j'$)

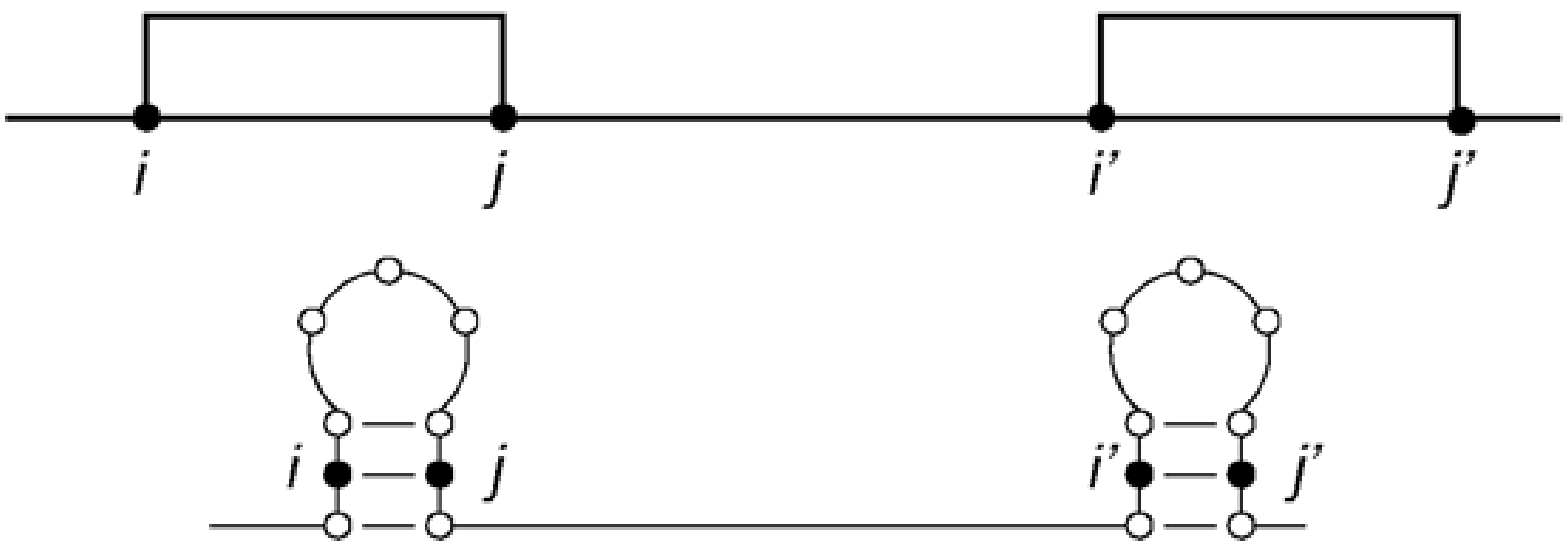
Given an **RNA sequence** $S = s_1, s_2 \ldots s_n$, where s_i is the ith nucleotide. A **secondary structure** is an ordered list of pairs, $i.j$, $1 \leq i < j \leq n$ such that:

- $j - i \geq c$, where $c = 4$ for instance
- 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 $i'.j'$)

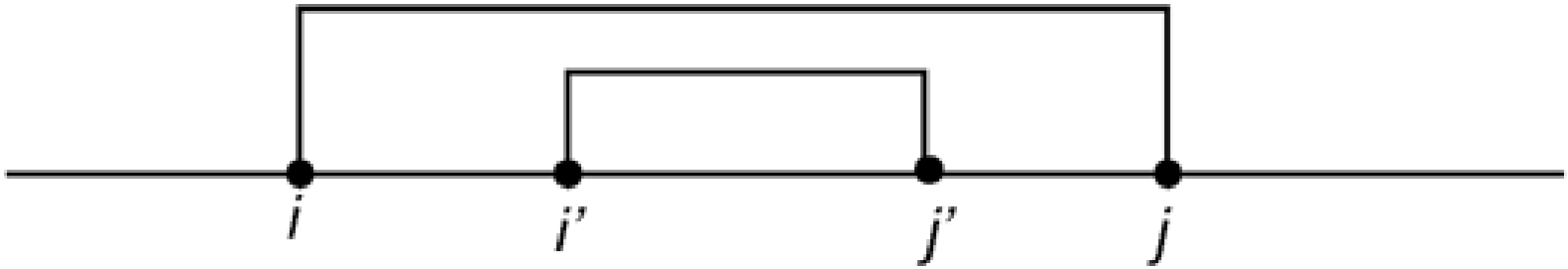
Given an **RNA sequence** $S = s_1, s_2 \ldots s_n$, where s_i is the ith nucleotide. A **secondary structure** is an ordered list of pairs, $i.j$, $1 \leq i < j \leq n$ such that:

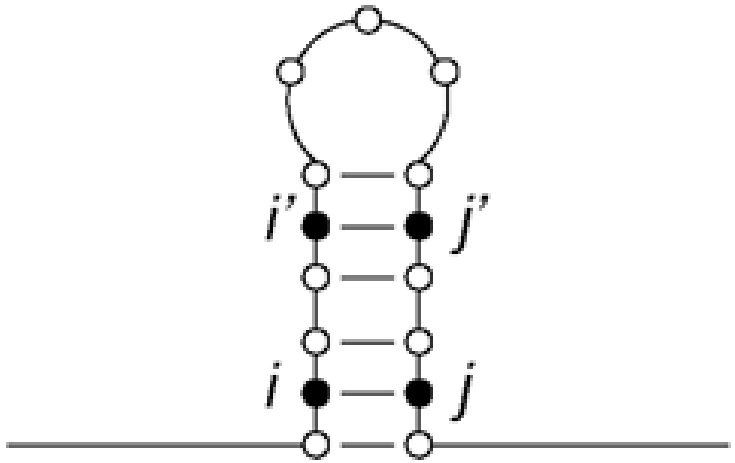
- $j - i \geq c$, where $c = 4$ for instance
- Given $i.j$ and $i'.j'$, two base pairs, then either:
  - $i = i'$ and $j = j'$ (they are the same)
  - $i < j < i' < j'$ ($i.j$ precedes $i'.j'$)
  - $i < i' < j' < j$ ($i.j$ includes $i'.j'$)
  - $i < i' < j < j'$ (pseudoknot)

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



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





Problems

- Reporting **sub-optimal** structures (MFOLD, SFOLD)

Problems

- Reporting **sub-optimal** structures (MFOLD, SFOLD)
- **Partition function** and the McCaskill's calculation of P_{ij}'s

Problems

- Reporting **sub-optimal** structures (MFOLD, SFOLD)
- **Partition function** and the McCaskill's calculation of P_{ij}'s
- Folding **kinetics**, identifying ribo-switches

Problems

- Reporting **sub-optimal** structures (MFOLD, SFOLD)
- **Partition function** and the McCaskill's calculation of P_{ij}'s
- Folding **kinetics**, identifying ribo-switches
- MFE for secondary structure for **interacting** RNA molecules

Problems

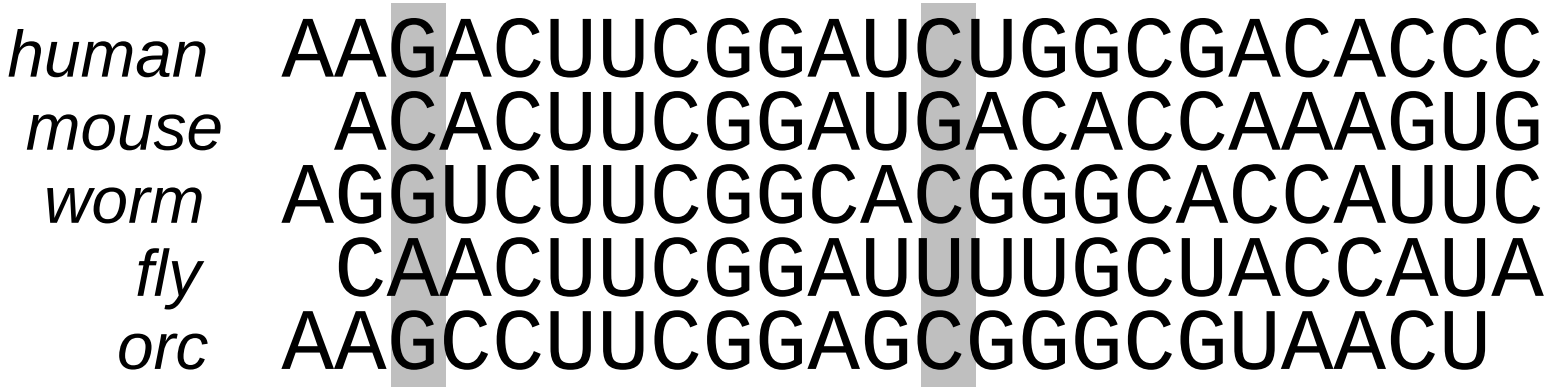
- Reporting **sub-optimal** structures (MFOLD, SFOLD)
- **Partition function** and the McCaskill's calculation of P_{ij}'s
- Folding **kinetics**, identifying ribo-switches
- MFE for secondary structure for **interacting** RNA molecules
- **Partition function** for secondary structure for **interacting** RNA molecules

Problems

- Reporting **sub-optimal** structures (MFOLD, SFOLD)
- **Partition function** and the McCaskill's calculation of P_{ij}'s
- Folding **kinetics**, identifying ribo-switches
- MFE for secondary structure for **interacting** RNA molecules
- **Partition function** for secondary structure for **interacting** RNA molecules
- Non-protein-coding **gene identification** (EvoFold, RNAz…)

human	AAGACUUCGGAUCUGGCGACACCC
mouse	ACACUUCGGAUGACACCAAAGUG
worm	AGGUCUUCGGCACGGGCACCAUUC
fly	CAACUUCGGAUUUUGCUACCAUA
orc	AAGCCUUCGGAGCGGGCGUAACU

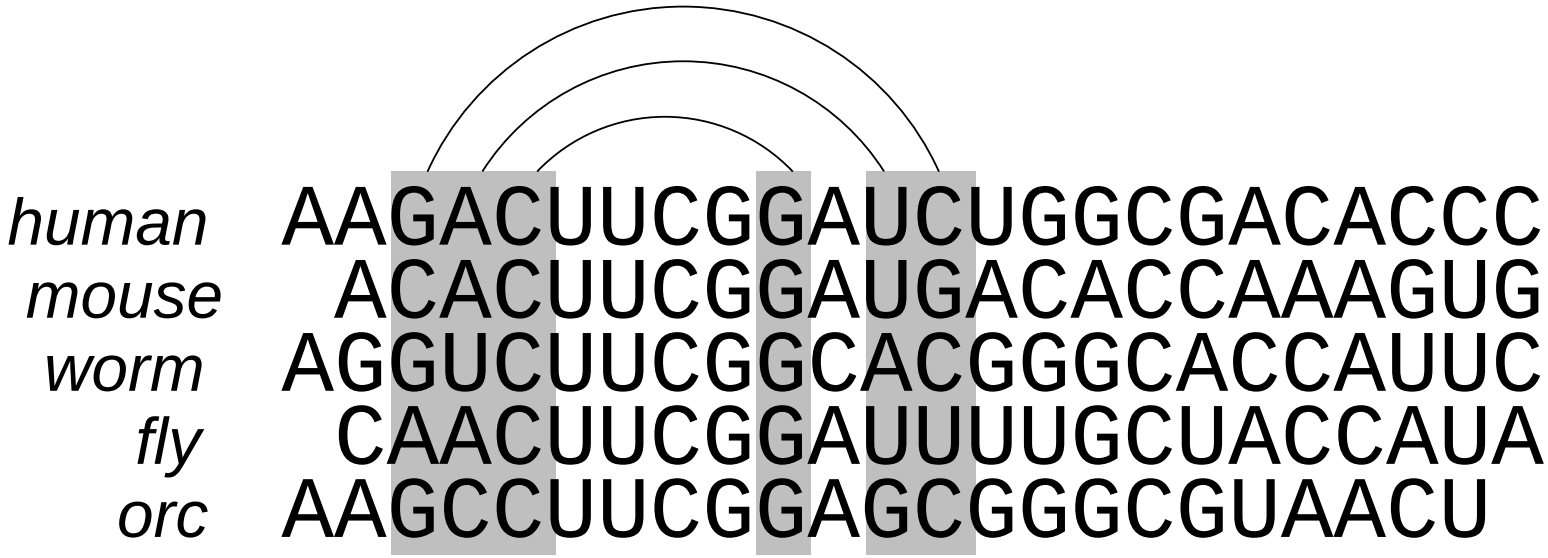
"Today, comparative analysis has become the method of choice for establishing higher-order structure for large RNA." Pace, Thomas, Woese (1999) In *The RNA World*. Cold Spring Harbor.



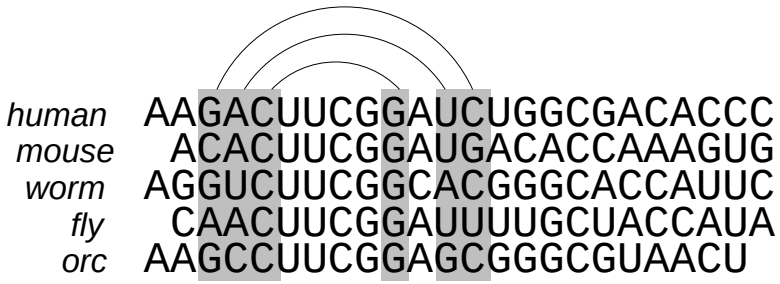
"Today, comparative analysis has become the method of choice for establishing higher-order structure for large RNA." Pace, Thomas, Woese (1999) In *The RNA World*. Cold Spring Harbor.

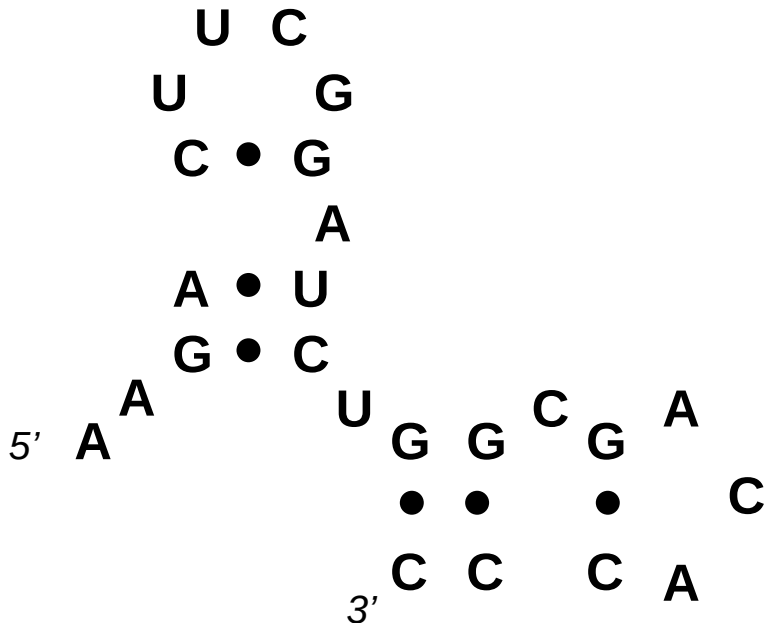
```
human   AAGACUUCGGAUCUGGCGACACCC
mouse     ACACUUCGGAUGACACCAAAGUG
worm    AGGUCUUCGGCACGGGCACCAUUC
fly       CAACUUCGGAUUUUGCUACCAUA
orc     AAGCCUUCGGAGCGGGCGUAACU
```

"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.

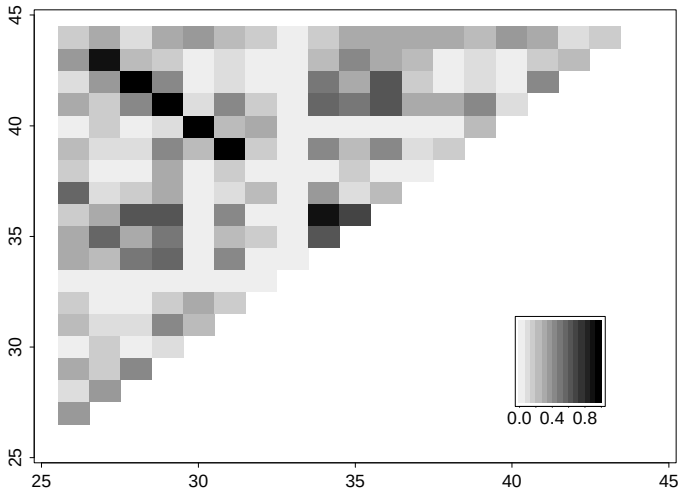


"Today, comparative analysis has become the method of choice for establishing higher-order structure for large RNA." Pace, Thomas, Woese (1999) In *The RNA World*. Cold Spring Harbor.


human AAGACUUCGGAUCUGGCGACACCC
mouse ACACUUCGGAUGACACCAAAGUG
worm AGGUCUUCGGCACGGGCACCAUUC
fly CAACUUCGGAUUUUGCUACCAUA
orc AAGCCUUCGGAGCGGGCGUAACU


5'
3'

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 **ACGGC**CUGCAAA**GCCGU**
30 35 40



- Starts with the **alignment** of a set of homologous sequences

- Starts with the **alignment** of a set of homologous sequences
- Detecting **correlated pairs** of sites

- Starts with the **alignment** of a set of homologous sequences
- Detecting **correlated pairs** of sites
  - **Parallel chords implies helices** (stems)

- Starts with the **alignment** of a set of homologous sequences
- Detecting **correlated pairs** of sites
  - **Parallel chords implies helices** (stems)
  - 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 $H(I) = -\sum_\alpha P(i=\alpha) \log P(i=\alpha)$
    and $H(I,J) = -\sum_{\alpha\beta} P(i=\alpha, j=\beta) \log P(i=\alpha, j=\beta)$

Accuracy of comparative analysis on rRNAs

- Late **1970**'s, comparative sequence analysis
- 16S $\sim$ 1500 nt long, 23S $\sim$ 3000 nt long
- 4.3×10^{393} and 6.3×10^{740} possible secondary structures
- **2000**, high-resolution crystal structures of rRNAs produced
- Gutell et al. The accuracy of ribosomal RNA comparative structure models. Curr Opin Struct Biol (2002) vol. 12 (3) pp. 301-10
- "**97–98% of the base pairings predicted with covariation analysis are indeed present in the 16S and 23S rRNA crystal structures**"

What are the **main difficulties**?

What are the **main difficulties**?

- **Needs an alignment**, but sequence alignment techniques are not well adapted for RNA sequences

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**

What are the **main difficulties**?

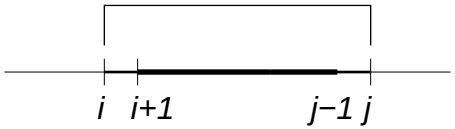
- **Needs an alignment**, but sequence alignment techniques are not well adapted for RNA sequences
- To produce a high quality alignment, the **sequences should be similar**
- If the sequences are similar, there will be **few observed compensatory changes**

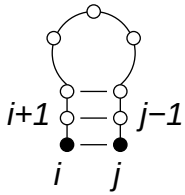
RNA **Folding**

- **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)

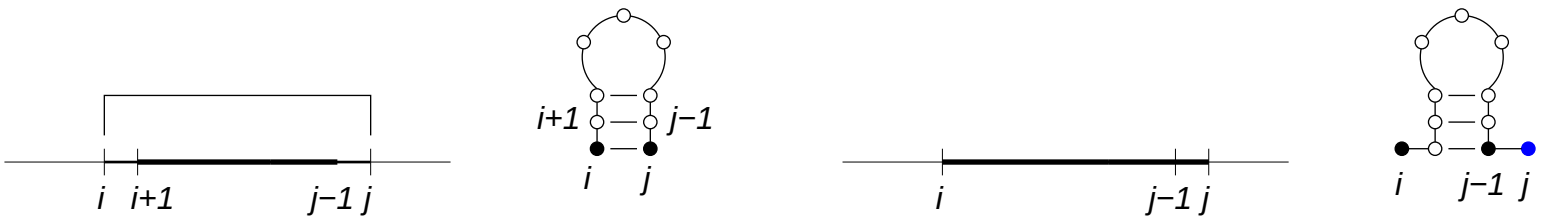
Maximizing the **number of pairs** for the segment $i..j$

Maximizing the **number of pairs** for the segment $i..j$

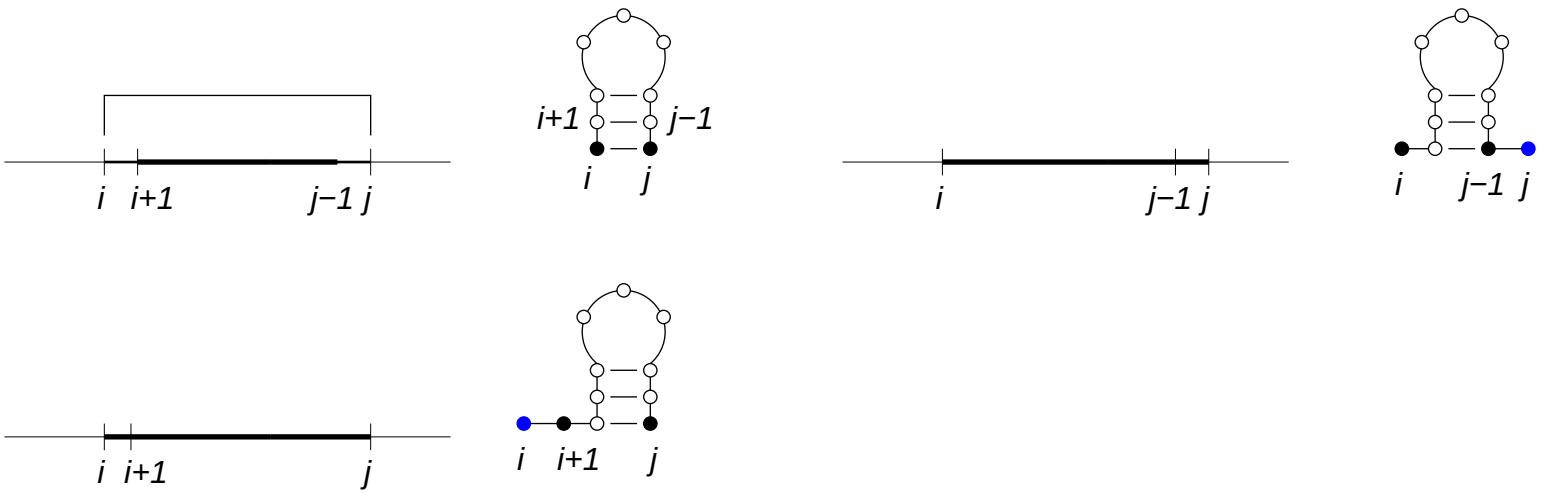




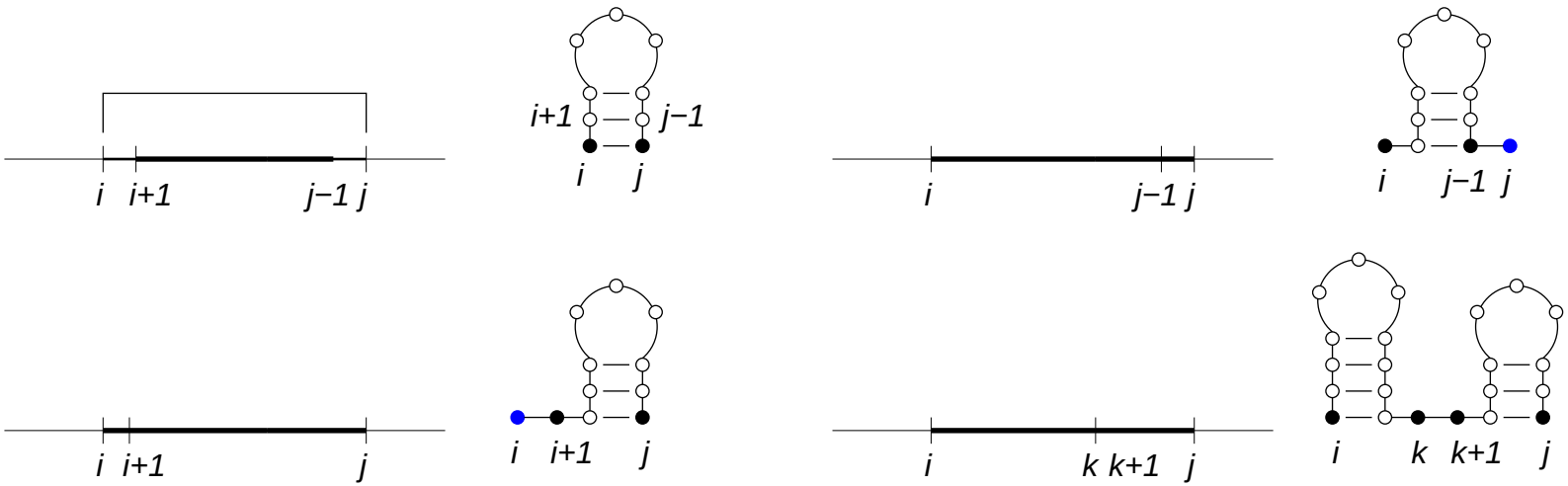
Maximizing the **number of pairs** for the segment $i..j$



Maximizing the **number of pairs** for the segment $i..j$



Maximizing the **number of pairs** for the segment $i..j$



Nussinov

Initialization:

$$\gamma(i, i+k) \quad = \quad 0 \quad \text{for k} = 0 \text{ to } 3 \text{ and for i} = 1 \text{ to n} - \text{k}.$$

Recurrence:

$$\gamma(i,j) = \max \begin{cases} \gamma(i+1, j-1) + \delta(i,j); \\ \gamma(i+1, j); \\ \gamma(i, j-1); \\ \max_{i<k<(j-1)} [\gamma(i,k) + \gamma(k+1, j)]. \end{cases}$$

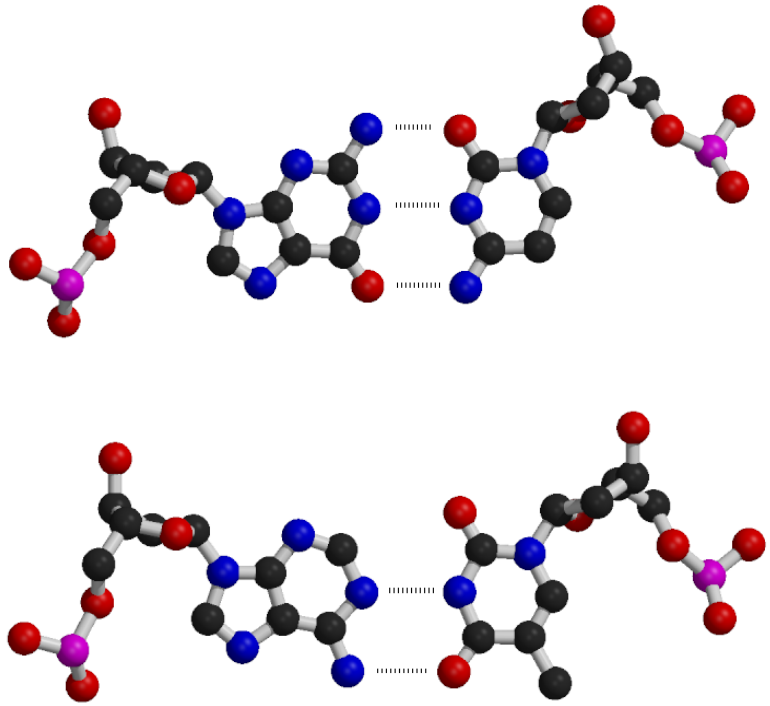
Matching score:

$$\delta(i,j) = \begin{cases} 1, \text{if } a_i : a_j \in \{A:U, U:A, G:C, C:G\} \bigcup \{G:U, U:G\}; \\ 0, \text{otherwise}. \end{cases}$$

Nature **does not** use this strategy!

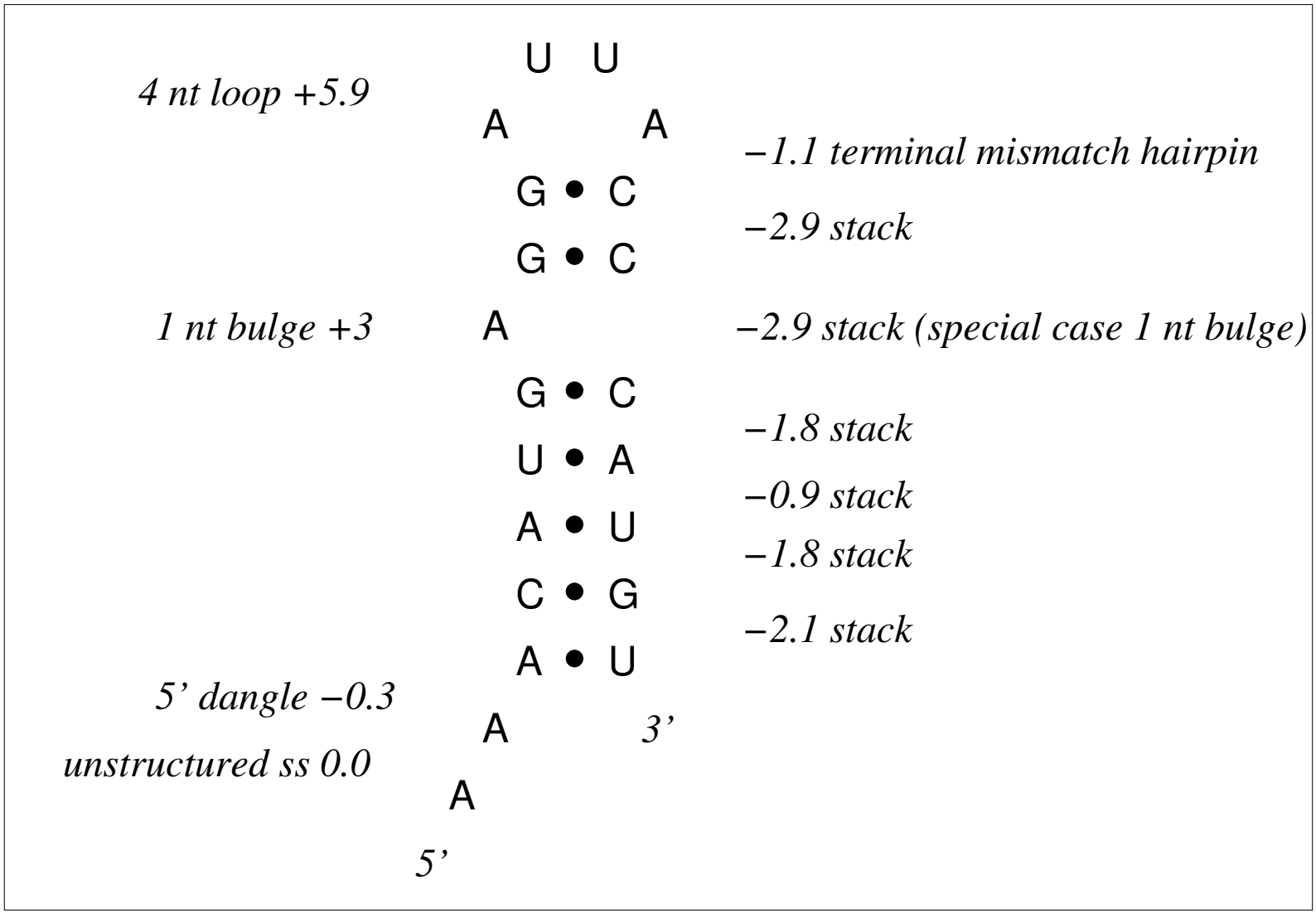
How about **maximizing the number of hydrogen bonds**?

+3 for G.C pairs
+2 for A.U pairs
+1 for G.U pairs



Nature **does not** use this strategy either!

Nearest-neighbor model



Simplified **Zuker** (MFOLD) algorithm

$$W(i,j) = \min \begin{cases} W(i+1,j), \\ W(i,j-1), \\ V(i,j), \\ \min_{i \leq k < j}[W(i,k) + W(k+1,j)]. \end{cases}$$

where V models a segment such that i and j form a base pair.

$$V(i,j) = \min \begin{cases} V_1(i,j), & \text{hairpin closed by } i \bullet j \\ V_2(i,j), & \text{helix extension, bulge, interior loop} \\ V_3(i,j), & \text{multiple loop} \end{cases}$$

Simplified **Zuker** (MFOLD) algorithm

V_2 extending a helix, bulge, or interior loop:

$$V_2(i,j) = \min_{i<i'<j'<j} [e(\text{motif}) + V(i',j')]$$

V_3 multi-branch loop structures:

$$V_3(i,j) = \min_{i+1<k<j-1} [e(\text{motif}) + W(i+1,k) + W(k+1,j-1)]$$

MFOLD

Volume 9 Number 1 1981 Nucleic Acids Research

Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information

Michael Zuker and Patrick Stiegler[+]

Division of Biological Sciences, National Research Council of Canada, Ottawa K1A 0R6, Canada

MFOLD

- Sophisticated **energy minimization** program
- Developed by **Mike Zuker** (NRC/Ottawa, now at RPI)
- 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
- Takes $\mathcal{O}(n^2)$ space and $\mathcal{O}(n^3)$ time

Performance of the Nearest-Neighbour Model (for a single sequence)

- The nearest-neighbour model works reasonably well for small RNAs, **69 %** and **71 % PPV** (positive predictive value) for the tRNA and 5S rRNA, which are approximately 80 and 120 nucleotides long, respectively.

K. J. Doshi, J. J. Cannone C. W. Cobaugh, et R. R. Gutell (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* **5**(1):105.

Observations

- RNAs **conserve secondary structure interactions** more than they conserve their sequence

Observations

- RNAs **conserve secondary structure interactions** more than they conserve their sequence
- The nearest-neighbour model **performs well on average** but fails for certain sequences

Observations

- RNAs **conserve secondary structure interactions** more than they conserve their sequence
- The nearest-neighbour model **performs well on average** but fails for certain sequences
- Single-sequence methods can be **generalised** to determine a consensus structure for more than one sequence

Observations

- RNAs **conserve secondary structure interactions** more than they conserve their sequence
- The nearest-neighbour model **performs well on average** but fails for certain sequences
- Single-sequence methods can be **generalised** to determine a consensus structure for more than one sequence
- **As the number of input sequences increases**, it becomes unlikely that the nearest-neighbour model simultaneously fails for all of them

eXtended Dynalign

- David Sankoff (1985) **Simultaneous solution of RNA folding, alignment and protosequence problems**. SIAM J. Appl. Math. **45**(5):810–825.

eXtended Dynalign

- David Sankoff (1985) **Simultaneous solution of RNA folding, alignment and protosequence problems**. SIAM J. Appl. Math. **45**(5):810–825.
- Objective function is a linear combination of the free energy of each sequence given the common secondary structure;

eXtended Dynalign

- David Sankoff (1985) **Simultaneous solution of RNA folding, alignment and protosequence problems**. SIAM J. Appl. Math. **45**(5):810–825.
- Objective function is a linear combination of the free energy of each sequence given the common secondary structure;
- D.H. Mathews et D.H. Turner (2002) **Dynalign: An Algorithm for Finding the Secondary Structure Common to Two RNA Sequences**. J. Mol. Biol. **317**:191–203.

eXtended Dynalign

- David Sankoff (1985) **Simultaneous solution of RNA folding, alignment and protosequence problems**. SIAM J. Appl. Math. **45**(5):810–825.
- Objective function is a linear combination of the free energy of each sequence given the common secondary structure;
- D.H. Mathews et D.H. Turner (2002) **Dynalign: An Algorithm for Finding the Secondary Structure Common to Two RNA Sequences**. J. Mol. Biol. **317**:191–203.
- We extended this work for three sequences.

Idea

- The **objective function** is a linear combination of the free energy of each sequence given the common structure;

$$\Delta G^\circ_{\text{total}} = \Delta G^\circ_{\text{seq 1}} + \Delta G^\circ_{\text{seq 2}} + \Delta G^\circ_{\text{seq 3}} + \Delta G^\circ_{\text{insertions}}$$

- No terms for substitutions;
- Solved by **dynamic programming**: constructing an alignment and a common secondary structure for $S_1[i,j]$, $S_2[k,l]$ and $S_3[m,n]$, from the smallest to the largest segment.

Idea

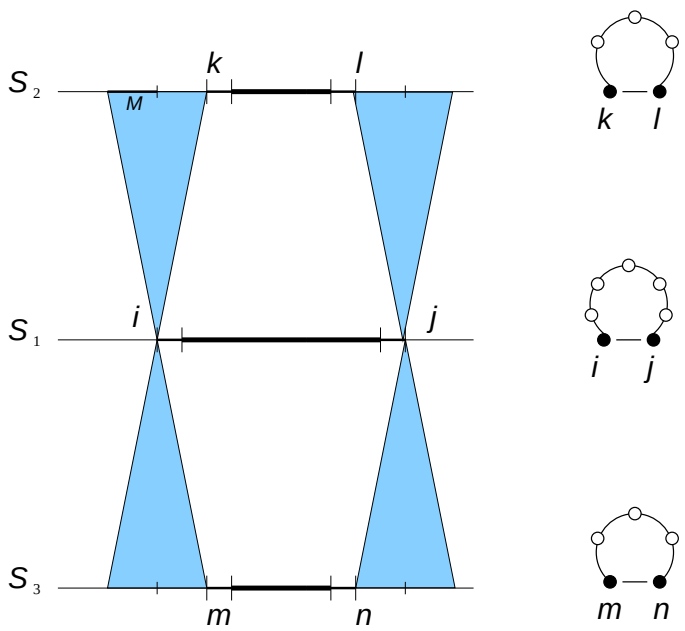
```
Score= -578

GCCCGGGTGGTGTAGTGGCCCATCATACGACCCTGTCACGGTCG-TGACGCGGGTTCAAATCCCGCCTCGGGCGCCA
GTCGCAATGGTG-TAGTTGGGAGCATGACAGACTGAAGATCTGTTGGTCATCGGTTCGATCCCGGTTTGTGACACCA
GCCCCCAUCGUCUAGAGGCCUAGGACACCUCCCUUUCACGGAGG-CGACAGGGAUUCGAAUUCCCUUGGGGGUACCA

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

eXtended Dynalign

The recurrence equations describing the free energy are somewhat complex. There are 140 cases: $V_1, V_2, V_{3_{1-64}}, W_1, W_2, W_{3_{1-64}}, W_{9_{1-8}}$. Let S_1, S_2 and S_3, be three RNA sequences.

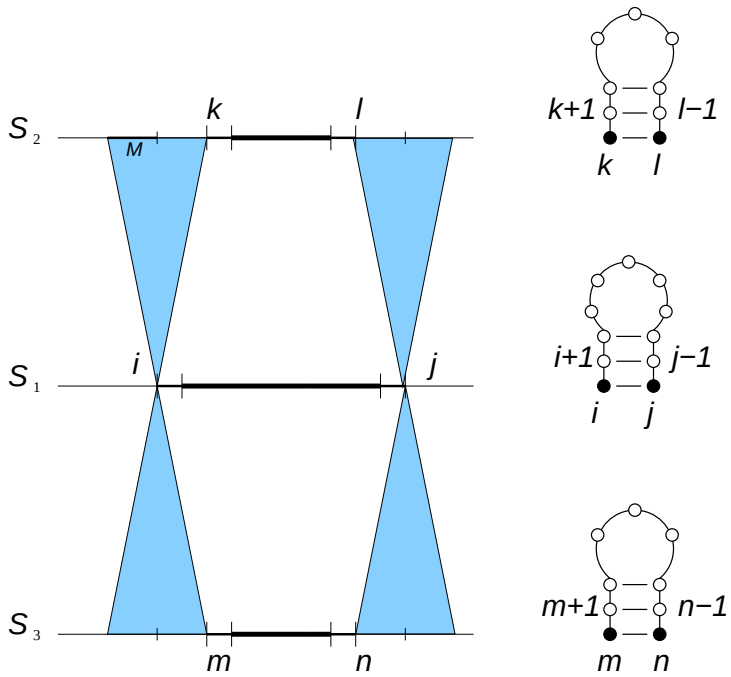
- $W(i,j;k,l;m,n)$ represents the some of the free energy of $S_1[i,j]$, given the common structure, $S_2[k,l]$ given the common secondary structure and $S_3[m,n]$;
- $V(i,j;k,l;m,n)$ is defined similarly to W but also imposes constraints such that i is paired with j, k is paired with l, and m is paired with m;
- $W9$ represents the free energy for a prefix alignment of $S_1[1,j]$, $S_2[1,l]$ and $S_3[1,n]$.

Hairpin loop closed by a base-pair: $V_1(i,j;k,l;m,n)$



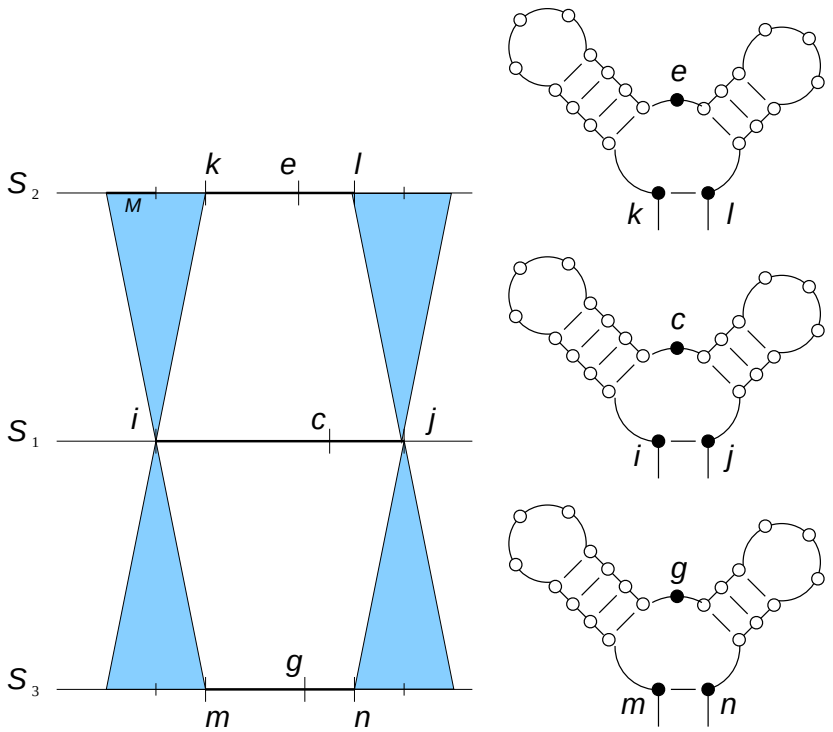
$$\Delta G^\circ_{\text{hairpin}}(i,j)+\Delta G^\circ_{\text{hairpin}}(k,l)+\Delta G^\circ_{\text{hairpin}}(m,n)+\Delta G^\circ_{\text{gap}}(\text{no. of gaps})$$

Helix Extension: $V_{2.1}(i,j;k,l;m,n)$



$$V(i+1,j-1;k+1,l-1;m+1,n-1)+\Delta G^{\circ}_{\text{motif}_1}+\Delta G^{\circ}_{\text{motif}_2}+\Delta G^{\circ}_{\text{motif}_3}$$

Multibranch Loop: $V_{3.1}(i,j;k,l;m,n)$



$$W(i,c;k,e;m,g)+W(c+1,j;e+1,l;g+1,n)+\Delta G^{\circ}_{\text{motif}_1}+\Delta G^{\circ}_{\text{motif}_2}+\Delta G^{\circ}_{\text{motif}_3}$$

Summary

- A base pair is predicted only if it simultaneously occurs in all three sequences
- The algorithm finds a consensus structure
- An alignment is produced as a byproduct, it is reliable only in the base paired regions, as no substitution scores are used

MFOLD: tRNAs

Id	Sensitivity	PPV	MCC
RD0260	33.3	29.2	31.2
RD0500	47.6	43.5	45.5
RD4800	42.9	56.2	49.1
RE2140	95.2	87	91
RE6781	33.3	28	30.6
RF6320	0	0	0
RL0503	0	0	0
RL1141	40	43.5	41.7
RS0380	52	56.5	54.2
RS1141	19.2	25	21.9

MFOLD: 5S rRNAs

Id	Sensitivity	PPV	MCC
AJ131594	23.7	60	37.7
AJ251080	26.3	45.5	34.6
D11460	15.8	37.5	24.3
K02682	20.5	40	28.6
M10816	31.6	70.6	47.2
M16532	10.3	21.1	14.7
M25591	26.3	45.5	34.6
V00336	37.5	65.2	49.5
X02024	15.8	37.5	24.3
X02627	38.5	68.2	51.2
X04585	0	0	0
X08000	0	0	0
X08002	0	0	0

Are **three** input sequences better than **two**?

1. The worse prediction (minimum accuracy) **should be more accurate**;
2. Use of three input sequences should **improve the average accuracy**;
3. Average **coverage should be less**.

Masoumi, B. and Turcotte, M. (2005) Simultaneous alignment and structure prediction of three RNA sequences. *Int. J. Bioinformatics Research and Applications.* Vol. 1, No. 2, pp. 230-245

Beeta Masoumi and Marcel Turcotte. Simultaneous alignment and structure prediction of RNAs: Are three input sequences better than two? In S. V. Sunderam et al., editor, *2005 International Conference on Computational Science (ICCS 2005)*, Lecture Notes in Computer Science 3515, pages 936-943, Atlanta, USA, May 22-25 2005.

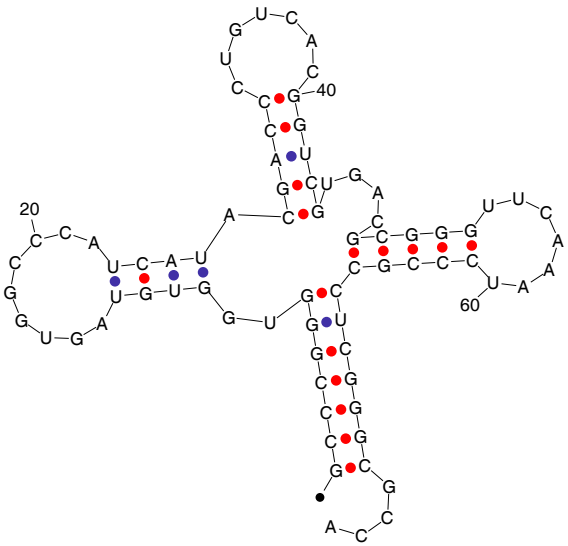
PPV: tRNA Dataset

Id	N_{xd}	N_d	Min_{xd}	Min_d	Max_{xd}	Max_d	Ave_{xd}	Ave_d
RD0260	4	5	100	80	100	100	100.0	96.0
RD0500	4	5	76	45	100	100	82.2	80.8
RD4800	5	5	100	80	100	100	100.0	96.0
RE2140	2	4	100	100	100	100	100.0	100.0
RE6781	2	4	100	77	100	100	100.0	94.3
RF6320	4	5	95	45	100	100	96.4	89.1
RL0503	1	2	100	100	100	100	100.0	100.0
RL1141	2	3	100	70	100	100	100.0	90.3
RS0380	1	2	100	83	100	87	100.0	85.2
RS1141	2	3	100	70	100	100	100.0	90.3

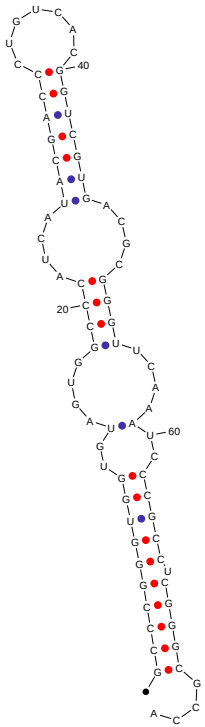
xd stands for eXtended Dynalign, *d* stands for Dynalign.

X-Dynalign 96.8 ± 7.6 vs Dynalign 92.1 ± 14.6.

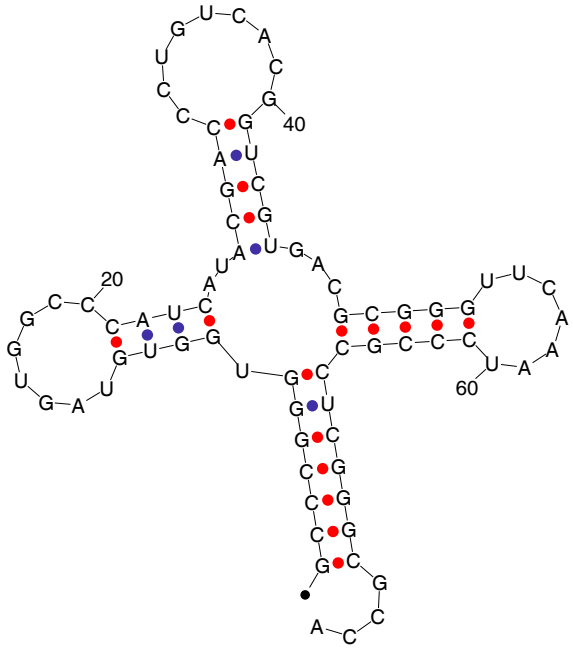
eXtended-Dynalign reproduces the clover-leaf structure



(a) RD0500

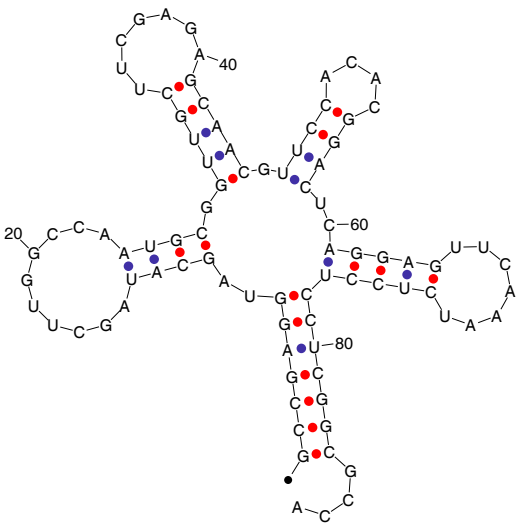


(b) Dynalign

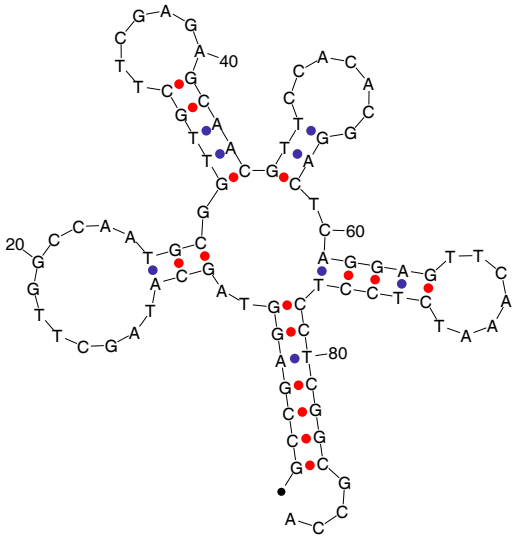


(c) X-Dynalign

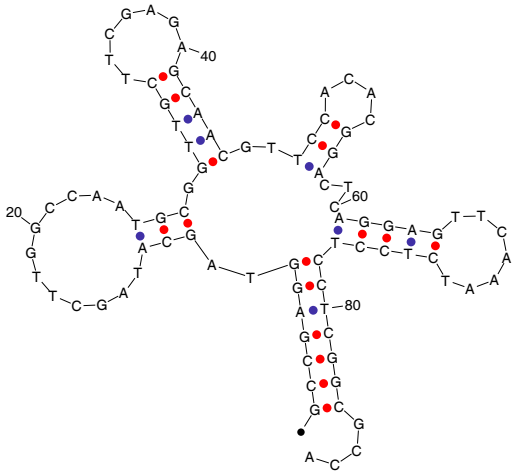
Fine **detail**s are better reproduced as well



(a) RS0380



(b) Dynalign



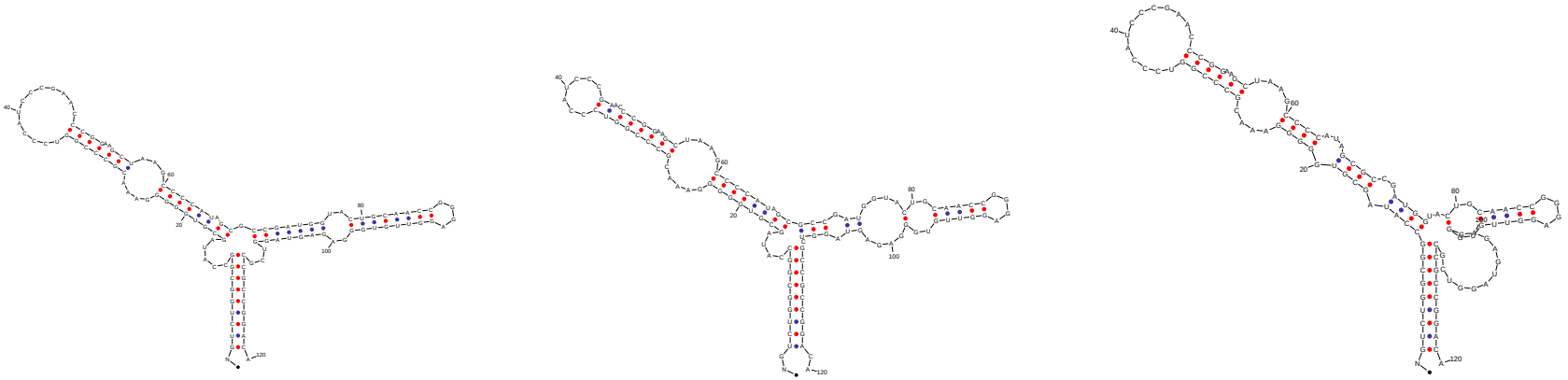
(c) X-Dynalign

PPV: 5S rRNA

Id	N_{xd}	N_d	Min_{xd}	Min_d	Max_{xd}	Max_d	Ave_{xd}	Ave_d
AJ131594	2	3	100	91	100	100	100.0	94.5
AJ251080	6	5	88	82	90	86	90.3	84.8
D11460	6	5	87	66	87	88	87.6	79.4
K02682	8	9	63	88	100	97	89.1	92.0
M10816	3	4	90	85	90	88	90.7	87.8
M16532	1	2	94	77	94	85	94.1	81.8
M25591	6	5	87	82	90	86	89.8	84.8
V00336	3	4	75	65	100	100	91.9	91.4
X02024	9	6	88	82	90	88	90.1	85.8
X02627	1	2	100	92	100	100	100.0	96.0
X04585	2	3	72	68	94	93	83.4	82.7
X08000	5	5	90	88	90	90	90.6	89.4
X08002	5	5	90	88	90	90	90.6	89.4

X-Dynalign 90.3 ± 5.8, Dynalign = 87.7 ± 7.4.

(K02682,V00336,X04585), PPV = 63%



Reference, Dynalign and X-Dynalign structures for the 5S rRNA K02682

Pros: eXtended Dynalign

- The **mean PPV** is **higher**
- Better worse case scenario
- The average sensitivity is slightly degraded. However, for the majority of the sequences the minimum sensibility is higher for eXtended Dynalign
- Some subtle details, such as the variable loop of some tRNAs, are well reproduced

Cons: eXtended Dynalign

- $\mathcal{O}(|S_1|^2 M^4)$ space, $\mathcal{O}(|S_1|^3 M^6)$ time
- Severe constraint M, $M \leq 6$
- Up to **two weeks of CPU time** for some sequences*
- Length limited to some 150 nucleotides

*Sun Fire V20z, AMD Opteron 2.2 GHz

Summary

- **Comparative sequence analysis**, gold standard, tedious, partially automated

Summary

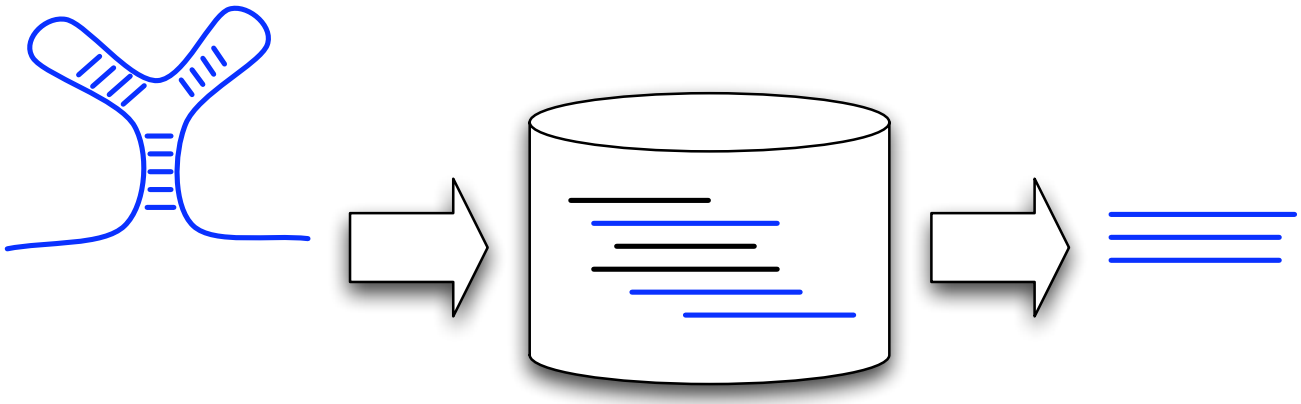
- **Comparative sequence analysis**, gold standard, tedious, partially automated
- Single sequence methods work best for **short sequences**, less than 100 nt, but still the quality of the results vary greatly

Summary

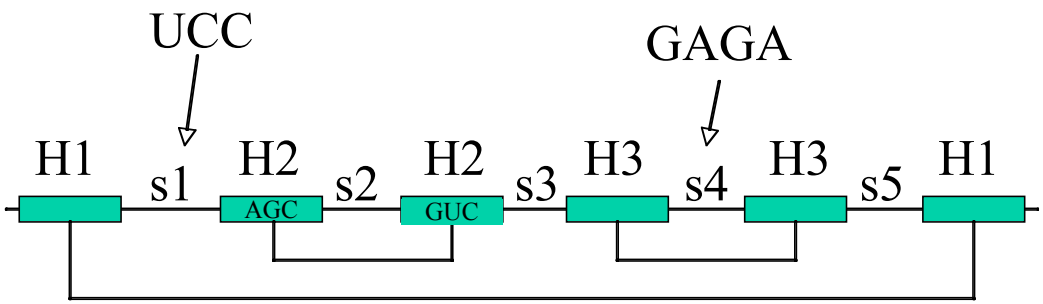
- **Comparative sequence analysis**, gold standard, tedious, partially automated
- Single sequence methods work best for **short sequences**, less than 100 nt, but still the quality of the results vary greatly
- **Consensus approaches produce good results**, but are CPU intensive

Database Search Problem

Find all sequences containing a user specified motif or all the sequences that can be folded into a user specified structure



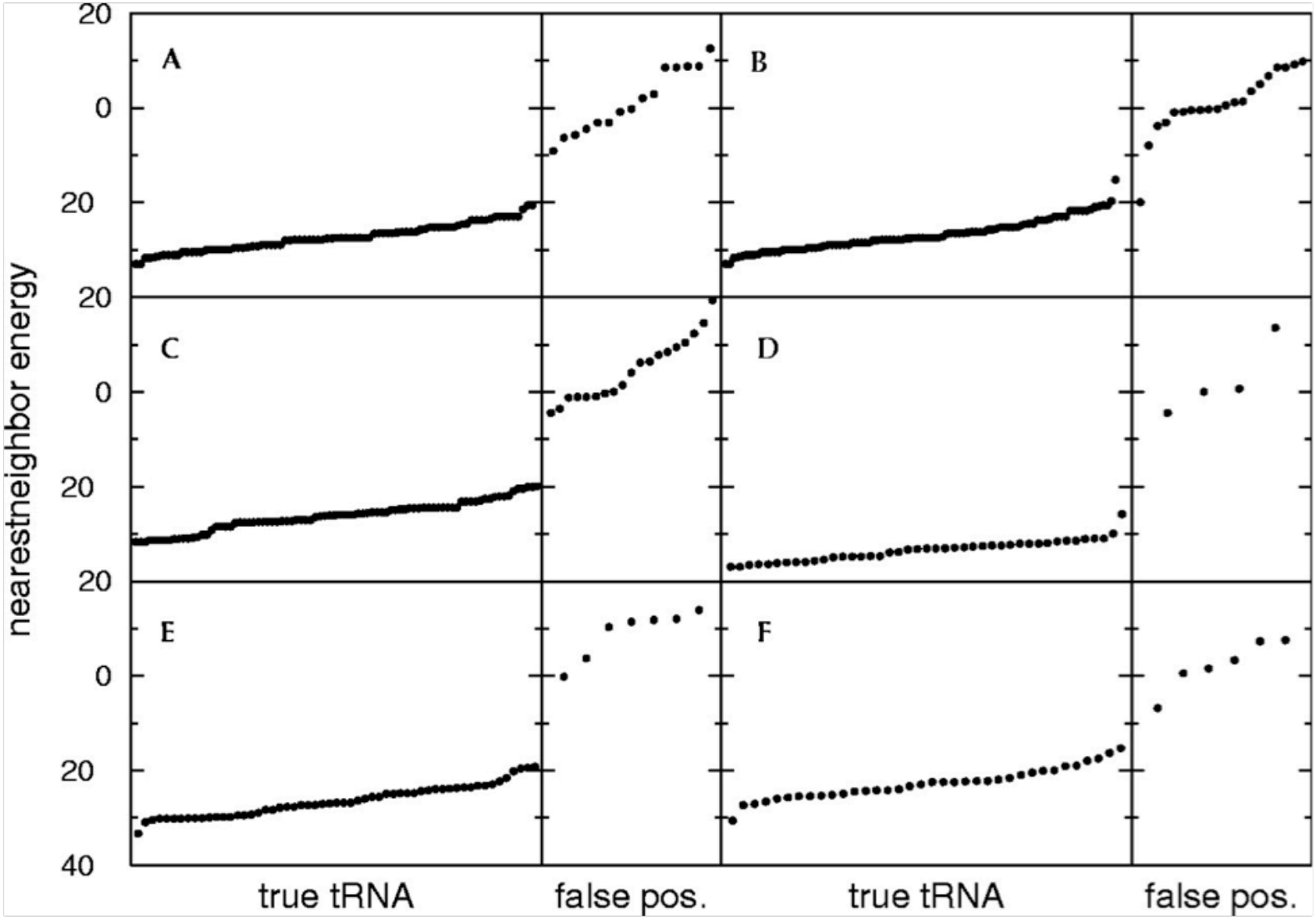
```
H1 s1 H2 s2 H2 s3 H3 s4 H3 s5 H1

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

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

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'] );
}
```


nearestneighbor energy
20
0
20
20
0
20
20
0
20
40
A
B
C
D
E
F
true tRNA
false pos.
true tRNA
false pos.

- **RSearch** and **INFERNAL** are principled approaches

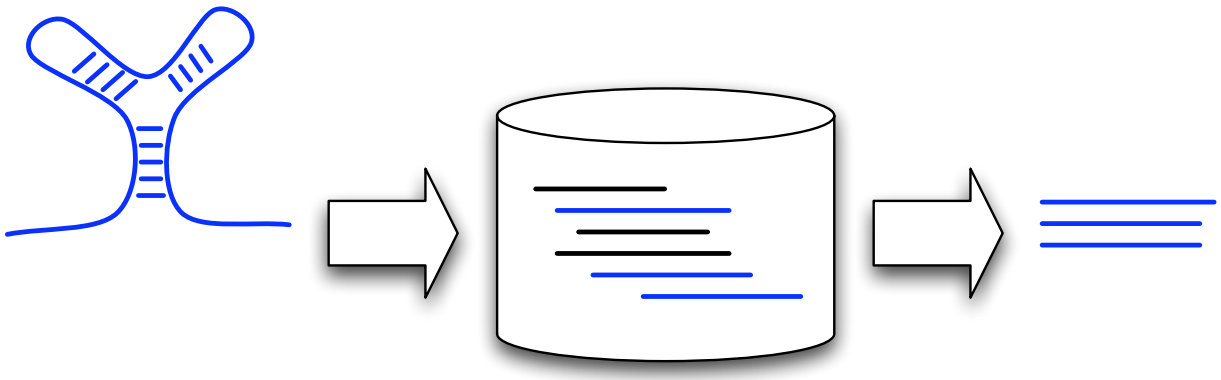
- **RSearch** and **INFERNAL** are principled approaches
- Based on **CYK** (Cocke-Younger-Kasami) algorithm for parsing context-free grammars

- **RSearch** and **INFERNAL** are principled approaches
- Based on **CYK** (Cocke-Younger-Kasami) algorithm for parsing context-free grammars
- **Solid statistical foundation**

- **RSearch** and **INFERNAL** are principled approaches
- Based on **CYK** (Cocke-Younger-Kasami) algorithm for parsing context-free grammars
- **Solid statistical foundation**
- **RSearch** takes as input a secondary and a secondary structure

- **RSearch** and **INFERNAL** are principled approaches
- Based on **CYK** (Cocke-Younger-Kasami) algorithm for parsing context-free grammars
- **Solid statistical foundation**
- **RSearch** takes as input a secondary and a secondary structure
- **INFERNAL** takes as input an MSA and a consensus structure

RSearch



```
# STOCKHOLM 1.0

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

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

RSearch

- **RIBOSUM** substitution matrices
  (analogous to residue substitution scores such as PAM and BLOSUM, but for base pairs)

RSearch

- **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

RSearch

- **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 $\mathcal{O}(NM^3)$ where N is the size of the database and M is the length of the input sequence

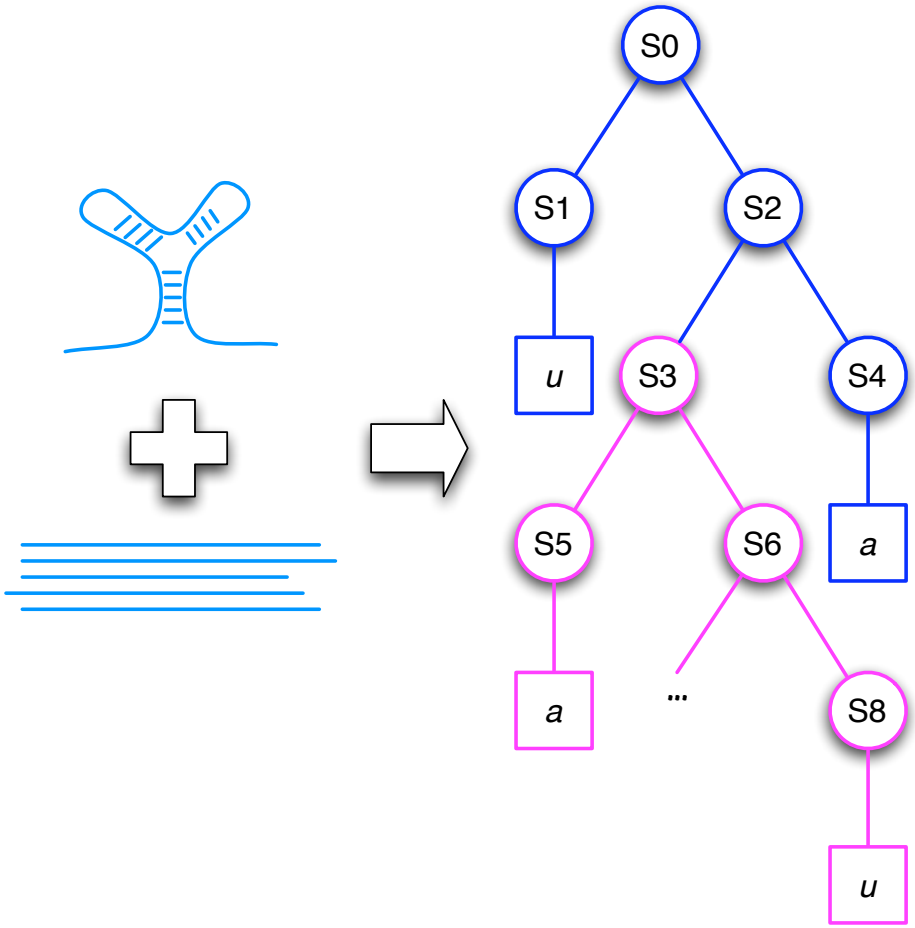
RSearch

- **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 $\mathcal{O}(NM^3)$ 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**"

INFERNAL

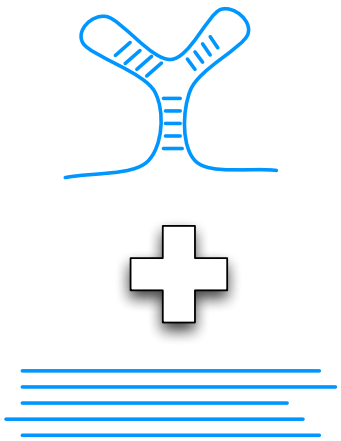
- Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**(22), 2933–2935. `http://doi.org/10.1093/bioinformatics/btt509`
- Also based on CYK, but uses an MSA as input
- MSA + Structure $\Rightarrow$ Covariance Model

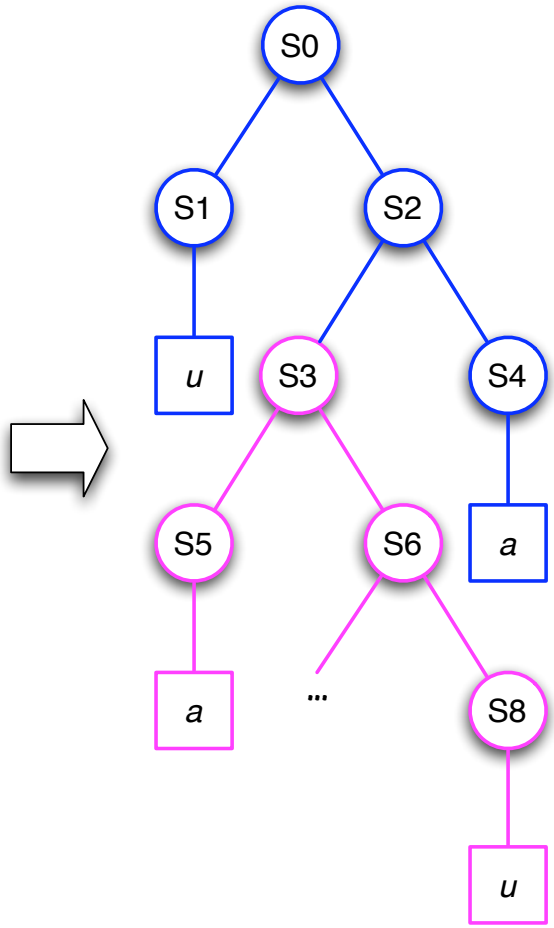
INFERNAL/Rfam Covariance Models



INFERNAL/Rfam Covariance Models

```
# STOCKHOLM 1.0

#=GC SS_cons   <<<<..>>>>
seq1           GGAGAUCUCC
seq2           GGGGAUCCCC
seq3           UGGGAACCCA
seq4           GGGGAUCCCU
seq5           GGGGAACCCC
//
```





Specialized Programs

- tRNAscan-SE
  - tRNAscan and EufindtRNA identify candidates that are subsequently analyse by Cove (INFERNAL)
  - 1 false positive per 15 billion nt
  - Detect 99% of true tRNA
  - `http://lowelab.ucsc.edu/tRNAscan-SE/`

Summary

- **Sequence alignment** methods are (generally) not appropriate for **RNA**

Summary

- **Sequence alignment** methods are (generally) not appropriate for **RNA**
- Tools such as **RNAMOT**, **RNABOB** and **RNAMOTIF** allows to describe and find RNA structure motifs in sequence databases

Summary

- **Sequence alignment** methods are (generally) not appropriate for **RNA**
- 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;

Summary

- **Sequence alignment** methods are (generally) not appropriate for **RNA**
- 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)

References

Can Alkan, Emre Karakoc, Joseph H Nadeau, S Cenk Sahinalp, and Kaizhong Zhang.
RNA-RNA interaction prediction and antisense RNA target search.
J Comput Biol, 13(2):267–82, Mar 2006.

Mirela Andronescu, Zhi Chuan Zhang, and Anne Condon.
Secondary structure prediction of interacting RNA molecules.
J Mol Biol, 345(5):987–1001, Feb 2005.

Alex Bateman, Shipra Agrawal, Ewan Birney, Elspeth A Bruford, Janusz M Bujnicki, Guy Cochrane, James R Cole, Marcel E Dinger, Anton J Enright, Paul P Gardner, Daniel Gautheret, Sam Griffiths-Jones, Jen Harrow, Javier Herrero, Ian H Holmes, Hsien-Da Huang, Krystyna A Kelly, Paul Kersey, Ana Kozomara, Todd M Lowe, Manja Marz, Simon Moxon, Kim D Pruitt, Tore

References (cont.)

Samuelsson, Peter F Stadler, Albert J Vilella, Jan-Hinnerk Vogel, Kelly P Williams, Mathew W Wright, and Christian Zwieb.
RNAcentral: A vision for an international database of RNA sequences.
RNA (New York, NY), 17(11):1941–1946, November 2011.

Ho-Lin Chen, Anne Condon, and Hosna Jabbari.
An O(n(5)) algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids.
J Comput Biol, 16(6):803–15, Jun 2009.

Hamidreza Chitsaz, Raheleh Salari, S Cenk Sahinalp, and Rolf Backofen.
A partition function algorithm for interacting nucleic acid strands.
Bioinformatics, 25(12):i365–73, Jun 2009.

References (cont.)

Y. Ding and C. E. Lawrence.
A bayesian statistical algorithm for rna secondary structure prediction.
Computers & Chemistry, pages 387–400, 1999.

Paul P Gardner, Jennifer Daub, John G Tate, Eric P Nawrocki, Diana L Kolbe, Stinus Lindgreen, Adam C Wilkinson, Robert D Finn, Sam Griffiths-Jones, Sean R Eddy, and Alex Bateman.
Rfam: updates to the RNA families database.
Nucleic Acids Research, 37(Database issue):D136–40, Jan 2009.

References (cont.)

Ioanna Kalvari, Joanna Argasinska, Natalia Quinones-Olvera, Eric P Nawrocki, Elena Rivas, Sean R Eddy, Alex Bateman, Robert D Finn, and Anton I Petrov.
Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families.
Nucleic acids research, 46(D1):D335–D342, 2018.

Ioanna Kalvari, Eric P Nawrocki, Joanna Argasinska, Natalia Quinones-Olvera, Robert D Finn, Alex Bateman, and Anton I Petrov.
Non-Coding RNA Analysis Using the Rfam Database.
Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], 62(1):e51, June 2018.

References (cont.)


Robert J Klein and Sean R Eddy.
RSEARCH: finding homologs of single structured RNA sequences.
BMC Bioinformatics, 4:44, Sep 2003.


J S McCaskill.
The equilibrium partition function and base pair binding probabilities for RNA secondary structure.
Biopolymers, 29(6-7):1105–19, Jan 1990.


E Nawrocki, D Kolbe, and S Eddy.
Infernal 1.0: inference of rna alignments.
Bioinformatics, Mar 2009.

References (cont.)

Jakob Skou Pedersen, Gill Bejerano, Adam Siepel, Kate Rosenbloom, Kerstin Lindblad-Toh, Eric S Lander, W James Kent, Webb Miller, and David Haussler.
Identification and classification of conserved rna secondary structures in the human genome.
PLoS Comput Biol, 2(4):e33, Apr 2006.

Anton I Petrov, Simon J E Kay, Richard Gibson, Eugene Kulesha, Dan Staines, Elspeth A Bruford, Mathew W Wright, Sarah Burge, Robert D Finn, Paul J Kersey, Guy Cochrane, Alex Bateman, Sam Griffiths-Jones, Jennifer Harrow, Patricia P Chan, Todd M Lowe, Christian W Zwieb, Jacek Wower, Kelly P Williams, Corey M Hudson, Robin Gutell, Michael B Clark, Marcel Dinger, Xiu Cheng Quek, Janusz M Bujnicki, Nam-Hai Chua, Jun Liu, Huan Wang, Geir Skogerbø, Yi Zhao, Runsheng Chen, Weimin Zhu, James R

References (cont.)

Cole, Benli Chai, Hsien-Da Huang, His Yuan Huang, J Michael Cherry, Artemis Hatzigeorgiou, and Kim D Pruitt.
RNAcentral: An international database of ncRNA sequences.
Nucleic acids research, 43(D1):D123–D129, January 2015.

David Sankoff.
Simultaneous solution of RNA folding, alignment and protosequence problems.
SIAM J. Appl. Math., 45(5):810–825, 1985.

The RNAcentral Consortium.
RNAcentral: a comprehensive database of non-coding RNA sequences.
Nucleic acids research, 45(D1):D128–D134, January 2017.

References (cont.)

M. Zuker.
On finding all suboptimal foldings of an RNA molecule.
Science, 244:48–52, 1989.

M. Zuker and P. Stiegler.
Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.
Nucl. Acids Res., 9:133–148, 1981.

Michael Zuker and David Sankoff.
RNA secondary structure and their prediction.
Bulletin of Mathematical Biology, 46(4):591–621, 1984.

Please don’t print these lecture notes unless you really need to!