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)
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.”

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}\text{Born in Montréal}
^{b}\text{Guest member uOttawa OISB}
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)
“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.”
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


The Wikipedia page below lists **31 Nobel laureates in 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

Non-coding RNAs

**miRNA:** microRNAs modulate the development in C. elegans, Drosophila, and mammals (~20 nt)

**snRNA:** small nuclear RNAs are involved in splicing of eukaryotic mRNAs (~200 nt)

**snoRNA:** small nucleolar RNA direct nucleotide modifications in rRNAs (~100 nt)

**gRNA:** guide RNAs play an important role in editing of certain mRNAs in trypanosomes (~ 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 (~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 (~375 nt)

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

…
**Rfam database**

- For each family
  - Multiple sequence alignment (seed, full)
  - Consensus secondary structure (from literature or predicted)
  - Covariance model

- rfam.org
**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.


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

How Many **Non-Coding RNAs**?

- **48,479** candidates in the human genome (**EvoFold**)
- Studies based on the **ENCODE** data set
  - **3,267 RNAz, 3,134 EvolFold**
  - **4,933 CMfinder**
How Many **Non-Coding RNAs?**

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).

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)


“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.”
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 botulinum toxin gene)
GTGAATCAGCACCCTGACTTTTCAGATGAAAATTTAATTAACTATCCAAAATGATGCTTTATATACCAATATGATTCTAATGGAAACAAGTGATATAGAAACATGATGTTAATGAACTTAATGTATTTTTCTATTTAGATGCACAGAAAGTGCCCGAAGGTGAAAATAATGTCAATCTCACCTCTTCAATTGATACAGCATTATTAGAAACAACCTAAATATATACATTTTTTTTTTATCAGAATTTATTAAATAATGTCAATAAACCTGTGCAAGCAGGC
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  CAAAATGATTCTTACGTCCAATATATGATTCTAATGTTACAAGTGAAATAAA-GAATAT  1771

Query  108  GATGTTAATGAACTTAATGTATTTTTCTATTTAGATGCACAGAAAGTGCC-GAAGGTGAA  167
Sbjct  1772  ACTGTTGATAAACTTAATGTATTTTTCTATTTATGACAGAAAGTGCC-GAAGGTGAA  1831

Query  168  A  168
Sbjct  1832  A  1832

...
How does it work?
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

\[
\text{aln}( \text{ATATAGAACAAC, AATAAAGGAAT} ) \text{ is the maximum of:}
\]
Pairwise Sequence Alignment

\[ \text{aln}( \text{ATATAGAACAAC}, \text{AATAAAGGAAT} ) \] is the maximum of:

- \[ \text{aln}( \text{ATATAGAACAA}, \text{AATAAAGGAA} ) + \text{sub}(\text{C},\text{T}) \]
- \[ \text{aln}( \text{ATATAGAACAAC}, \text{AATAAAGGA} ) + \text{del}(\text{C}) \]
- \[ \text{aln}( \text{ATATAGAACAAC}, \text{AATAAAGGAA} ) + \text{ins}(\text{C}) \]

ATATAGAACAAC  C
AATAAAGGAA  T
Pairwise Sequence Alignment

\[ \text{aln}( \text{ATATAGAACAAC, AATAAAGGAAT} ) \text{ is the maximum of:} \]

\[ \text{aln}( \text{ATATAGAACAA, AATAAAAGGAA} ) + \text{sub}(C,T) \]
\[ \text{ATATAGAACAA} \quad \text{C} \]
\[ \text{AATAAAAGGAA} \quad \text{T} \]

\[ \text{aln}( \text{ATATAGAACAA, AATAAAAGGAAT} ) + \text{del}(C) \]
\[ \text{ATATAGAACAA} \quad \text{C} \]
\[ \text{AATAAAAGGAAT} \quad - \]
Pairwise Sequence Alignment

\[ \text{aln}(\text{ATATAGAACAAC}, \text{AATAAAGGAAT}) \text{ is the maximum of:} \]
\[ \text{aln}(\text{ATATAGAACAA}, \text{AATAAAGGAA}) + \text{sub}(C,T) \]
\[ \begin{array}{c}
\text{ATATAGAACAA} \quad \text{C} \\
\text{AATAAAGGAA} \quad \text{T}
\end{array} \]
\[ \text{aln}(\text{ATATAGAACAA}, \text{AATAAAGGAAT}) + \text{del}(C) \]
\[ \begin{array}{c}
\text{ATATAGAACAA} \quad \text{C} \\
\text{AATAAAGGAAT} \quad \text{−}
\end{array} \]
\[ \text{aln}(\text{ATATAGAACAAC}, \text{AATAAAGGA}) + \text{ins}(T) \]
\[ \begin{array}{c}
\text{ATATAGAACAAC} \quad \text{−} \\
\text{AATAAAGGA} \quad \text{T}
\end{array} \]
Assumptions

Positional 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.
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)**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>G-C</td>
<td>C-C</td>
<td>C-G</td>
<td></td>
</tr>
<tr>
<td>A-U</td>
<td>U-U</td>
<td>U-A</td>
<td></td>
</tr>
<tr>
<td>C-G</td>
<td>G-G</td>
<td>G-C</td>
<td></td>
</tr>
</tbody>
</table>

CAGAGAGCUG       GUCAAGGCUAGCUG       GUCGAGAGAC

Yes, but sequences 1 and 3 share the same secondary structure!
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 $i$th 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 $i$th nucleotide. A secondary structure is an ordered list of pairs, $i,j$, $1 \leq i < j \leq n$ such that:

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

- $j - i \geq c$, 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 $i$th nucleotide. A **secondary structure** is an ordered list of pairs, $i.j$, $1 \leq i < j \leq n$ such that:

- $j - i \geq c$, 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 $i$th nucleotide. A **secondary structure** is an ordered list of pairs, $i.j$, $1 \leq i < j \leq n$ such that:

- $j - i \geq c$, 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 \( i \)th nucleotide. A **secondary structure** is an ordered list of pairs, \( i.j \), \( 1 \leq i < j \leq n \) such that:

- \( j - i \geq c \), 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 $i$th nucleotide. A secondary structure is an ordered list of pairs, $i.j$, $1 \leq i < j \leq n$ such that:

- $j - i \geq c$, 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 $i'.j'$)
$i < i' < j' < j$ \text{(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...)
<table>
<thead>
<tr>
<th>species</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>AAGACUUCGGAUUCUGGCGGACACCC</td>
</tr>
<tr>
<td>mouse</td>
<td>ACACUUCGGAUGACACCAAAGUG</td>
</tr>
<tr>
<td>worm</td>
<td>AGGUCUUUCGGGCACGGGCACCAUUC</td>
</tr>
<tr>
<td>fly</td>
<td>CAACUUCGGAUUUUGCUACCAUA</td>
</tr>
<tr>
<td>orc</td>
<td>AAGCCUUUCGGAGCGGGGCGUAACU</td>
</tr>
</tbody>
</table>

human: AAGACUUCGGAUCUGGCGACACCC
mouse: ACACUUCGGAUGACACCAAAAGUG
fly: AGGUCUUCGGCAGCGCCACCAUUCC
worm: CAACUUCGGAUUUGCUACCAUA
orc: AGGCUUCGGAUCUGGCGACACCC

AAGCCUUCGGAGCGGGCUGUAACU
ACACUUCGGAUGACACCAAAGUG
CAACUUCGGAUUUUGCUACCAUA
AGGUCUUCGGCAGCGCCACCAUUCC

5' A A U A G C A C A C C C
3' G G G A C G G G G G C A C A U C U
Saccharomyces cerevisiae
Spiroplasma meliferum
Mycoplasma capricolum
Mycoplasma mycoides
Spiroplasma meliferum
Streptomyces lividans

...CCAGACUGAAGAUCUGG...
CCUCCCUUGCAUGCGAGG
CCUCCCUUGCAUGCGAGG
CACGCUGUGUCAUCCGUG
UUUGAUUGAA GCUCAAA
ACGGCUCGCAAAAGCCGU

Marcel Turcotte

BNF 5106. Bioinformatics
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 ~ 1500 nt long, 23S ~ 3000 nt long
- $4.3 \times 10^{393}$ and $6.3 \times 10^{740}$ possible secondary structures
- 2000, high-resolution crystal structures of rRNAs produced
- “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**.
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 $i..j$
Initialization:

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

Recurrence:

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

Matching score:

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

- 4 nt loop +5.9
- 1 nt bulge +3
- 5’ dangle −0.3
- unstructured ss 0.0

−1.1 terminal mismatch hairpin
−2.9 stack
−2.9 stack (special case 1 nt bulge)
−1.8 stack
−0.9 stack
−1.8 stack
−2.1 stack

Marcel Turcotte
BNF 5106. Bioinformatics
Simplified Zuker (MFOLD) algorithm

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

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

\[ V(i, j) = \min \left\{ \begin{array}{l}
V_1(i, j), \quad \text{hairpin closed by } i \cdot j \\
V_2(i, j), \quad \text{helix extension, bulge, interior loop} \\
V_3(i, j), \quad \text{multiple loop}
\end{array} \right\} \]
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:

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

Developed by **Mike Zuker** (NRC/Ottawa, now at RPI)

Finds the structure with the **minimum equilibrium free energy** ($\Delta G$), 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 $O(n^2)$ space and $O(n^3)$ time
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.

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

eXtended Dynalign


- Objective function is a linear combination of the free energy of each sequence given the common secondary structure;

- Objective function is a linear combination of the free energy of each sequence given the common secondary structure;

eXtended Dynalign


- Objective function is a linear combination of the free energy of each sequence given the common secondary structure;


- 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.
Score= -578

GCCCGGGTGGTGATAGTGCCCATCATACGACCCTGTACGTTTGTTCAATCCCGCCTCGGCGCCA
GTCGCAATGGT-GTAGTTGGGAGCATGACAGACTGAAGATCTGTTGGTCATCGGTTCGATCCCGTTGTGACACCA
GCCCGCAUCGUCUAGAGCCUAGGACACCUCUUUCACGGAGG-CGACAGGGAUUCGAAUUCUCCUUGGGGUACCA
((((..((...........))))..((((........))))).(((...........)))))....
The recurrence equations describing the free energy are somewhat complex. There are 140 cases: $V_1, V_2, V_{31-64}, W_1, W_2, W_{31-64}, W_{91-8}$. 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 $S_3[1, n]$. 

Marcel Turcotte

BNF 5106. Bioinformatics
Hairpin loop closed by a base-pair: $V_1(i, j; k, l; m, n)$

$$\Delta G_{\text{hairpin}}^\circ(i, j) + \Delta G_{\text{hairpin}}^\circ(k, l) + \Delta G_{\text{hairpin}}^\circ(m, n) + \Delta G_{\text{gap}}^\circ(\text{no. of gaps})$$
Helix Extension: $V_{2.1}(i, j; k, l; m, n)$

$$V(i+1, j-1; k+1, l-1; m+1, n-1) + \Delta G_{\text{motif}_1} + \Delta G_{\text{motif}_2} + \Delta G_{\text{motif}_3}$$
Multibranch Loop: $V_{3.1}(i, j; k, l; m, n)$

\[
W(i, c; k, e; m, g) + W(c+1, j; e+1, l; g+1, n) + \Delta G_{\text{motif}_1} + \Delta G_{\text{motif}_2} + \Delta G_{\text{motif}_3}
\]
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

<table>
<thead>
<tr>
<th>Id</th>
<th>Sensitivity</th>
<th>PPV</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD0260</td>
<td>33.3</td>
<td>29.2</td>
<td>31.2</td>
</tr>
<tr>
<td>RD0500</td>
<td>47.6</td>
<td>43.5</td>
<td>45.5</td>
</tr>
<tr>
<td>RD4800</td>
<td>42.9</td>
<td>56.2</td>
<td>49.1</td>
</tr>
<tr>
<td>RE2140</td>
<td>95.2</td>
<td>87</td>
<td>91</td>
</tr>
<tr>
<td>RE6781</td>
<td>33.3</td>
<td>28</td>
<td>30.6</td>
</tr>
<tr>
<td>RF6320</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RL0503</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RL1141</td>
<td>40</td>
<td>43.5</td>
<td>41.7</td>
</tr>
<tr>
<td>RS0380</td>
<td>52</td>
<td>56.5</td>
<td>54.2</td>
</tr>
<tr>
<td>RS1141</td>
<td>19.2</td>
<td>25</td>
<td>21.9</td>
</tr>
</tbody>
</table>
# MFOLD: 5S rRNAs

<table>
<thead>
<tr>
<th>Id</th>
<th>Sensitivity</th>
<th>PPV</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AJ131594</td>
<td>23.7</td>
<td>60</td>
<td>37.7</td>
</tr>
<tr>
<td>AJ251080</td>
<td>26.3</td>
<td>45.5</td>
<td>34.6</td>
</tr>
<tr>
<td>D11460</td>
<td>15.8</td>
<td>37.5</td>
<td>24.3</td>
</tr>
<tr>
<td>K02682</td>
<td>20.5</td>
<td>40</td>
<td>28.6</td>
</tr>
<tr>
<td>M10816</td>
<td>31.6</td>
<td>70.6</td>
<td>47.2</td>
</tr>
<tr>
<td>M16532</td>
<td>10.3</td>
<td>21.1</td>
<td>14.7</td>
</tr>
<tr>
<td>M25591</td>
<td>26.3</td>
<td>45.5</td>
<td>34.6</td>
</tr>
<tr>
<td>V00336</td>
<td>37.5</td>
<td>65.2</td>
<td>49.5</td>
</tr>
<tr>
<td>X02024</td>
<td>15.8</td>
<td>37.5</td>
<td>24.3</td>
</tr>
<tr>
<td>X02627</td>
<td>38.5</td>
<td>68.2</td>
<td>51.2</td>
</tr>
<tr>
<td>X04585</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X08000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X08002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
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**.

---


**PPV: tRNA Dataset**

<table>
<thead>
<tr>
<th>Id</th>
<th>N_{xd}</th>
<th>N_{d}</th>
<th>Min_{xd}</th>
<th>Min_{d}</th>
<th>Max_{xd}</th>
<th>Max_{d}</th>
<th>Ave_{xd}</th>
<th>Ave_{d}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD0260</td>
<td>4</td>
<td>5</td>
<td>100</td>
<td>80</td>
<td>100</td>
<td>100</td>
<td>100.0</td>
<td>96.0</td>
</tr>
<tr>
<td>RD0500</td>
<td>4</td>
<td>5</td>
<td>76</td>
<td>45</td>
<td>100</td>
<td>100</td>
<td>82.2</td>
<td>80.8</td>
</tr>
<tr>
<td>RD4800</td>
<td>5</td>
<td>5</td>
<td>100</td>
<td>80</td>
<td>100</td>
<td>100</td>
<td>100.0</td>
<td>96.0</td>
</tr>
<tr>
<td>RE2140</td>
<td>2</td>
<td>4</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>RE6781</td>
<td>2</td>
<td>4</td>
<td>100</td>
<td>77</td>
<td>100</td>
<td>100</td>
<td>100.0</td>
<td>94.3</td>
</tr>
<tr>
<td>RF6320</td>
<td>4</td>
<td>5</td>
<td>95</td>
<td>45</td>
<td>100</td>
<td>100</td>
<td>96.4</td>
<td>89.1</td>
</tr>
<tr>
<td>RL0503</td>
<td>1</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>RL1141</td>
<td>2</td>
<td>3</td>
<td>100</td>
<td>70</td>
<td>100</td>
<td>100</td>
<td>100.0</td>
<td>90.3</td>
</tr>
<tr>
<td>RS0380</td>
<td>1</td>
<td>2</td>
<td>100</td>
<td>83</td>
<td>100</td>
<td>87</td>
<td>100.0</td>
<td>85.2</td>
</tr>
<tr>
<td>RS1141</td>
<td>2</td>
<td>3</td>
<td>100</td>
<td>70</td>
<td>100</td>
<td>100</td>
<td>100.0</td>
<td>90.3</td>
</tr>
</tbody>
</table>

{:}

\(x_{d}\) stands for eXtended Dynalign, \(d\) stands for Dynalign.

\(X\)-Dynalign \(96.8 \pm 7.6\) vs Dynalign \(92.1 \pm 14.6\).
eXtended-Dynalign reproduces the clover-leaf structure

(a) RD0500  (b) Dynalign  (c) X-Dynalign
Fine details are better reproduced as well

(a) RS0380  (b) Dynalign  (c) X-Dynalign
<table>
<thead>
<tr>
<th>Id</th>
<th>$N_{xd}$</th>
<th>$N_d$</th>
<th>$\text{Min}_{xd}$</th>
<th>$\text{Min}_d$</th>
<th>$\text{Max}_{xd}$</th>
<th>$\text{Max}_d$</th>
<th>$\text{Ave}_{xd}$</th>
<th>$\text{Ave}_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AJ131594</td>
<td>2</td>
<td>3</td>
<td>100</td>
<td>91</td>
<td>100</td>
<td>100</td>
<td>100.0</td>
<td>94.5</td>
</tr>
<tr>
<td>AJ251080</td>
<td>6</td>
<td>5</td>
<td>88</td>
<td>82</td>
<td>90</td>
<td>86</td>
<td>90.3</td>
<td>84.8</td>
</tr>
<tr>
<td>D11460</td>
<td>6</td>
<td>5</td>
<td>87</td>
<td>66</td>
<td>87</td>
<td>88</td>
<td>87.6</td>
<td>79.4</td>
</tr>
<tr>
<td>K02682</td>
<td>8</td>
<td>9</td>
<td>63</td>
<td>88</td>
<td>100</td>
<td>97</td>
<td>89.1</td>
<td>92.0</td>
</tr>
<tr>
<td>M10816</td>
<td>3</td>
<td>4</td>
<td>90</td>
<td>85</td>
<td>90</td>
<td>88</td>
<td>90.7</td>
<td>87.8</td>
</tr>
<tr>
<td>M16532</td>
<td>1</td>
<td>2</td>
<td>94</td>
<td>77</td>
<td>94</td>
<td>85</td>
<td>94.1</td>
<td>81.8</td>
</tr>
<tr>
<td>M25591</td>
<td>6</td>
<td>5</td>
<td>87</td>
<td>82</td>
<td>90</td>
<td>86</td>
<td>89.8</td>
<td>84.8</td>
</tr>
<tr>
<td>V00336</td>
<td>3</td>
<td>4</td>
<td>75</td>
<td>65</td>
<td>100</td>
<td>100</td>
<td>91.9</td>
<td>91.4</td>
</tr>
<tr>
<td>X02024</td>
<td>9</td>
<td>6</td>
<td>88</td>
<td>82</td>
<td>90</td>
<td>88</td>
<td>90.1</td>
<td>85.8</td>
</tr>
<tr>
<td>X02627</td>
<td>1</td>
<td>2</td>
<td>100</td>
<td>92</td>
<td>100</td>
<td>100</td>
<td>100.0</td>
<td>96.0</td>
</tr>
<tr>
<td>X04585</td>
<td>2</td>
<td>3</td>
<td>72</td>
<td>68</td>
<td>94</td>
<td>93</td>
<td>83.4</td>
<td>82.7</td>
</tr>
<tr>
<td>X08000</td>
<td>5</td>
<td>5</td>
<td>90</td>
<td>88</td>
<td>90</td>
<td>90</td>
<td>90.6</td>
<td>89.4</td>
</tr>
<tr>
<td>X08002</td>
<td>5</td>
<td>5</td>
<td>90</td>
<td>88</td>
<td>90</td>
<td>90</td>
<td>90.6</td>
<td>89.4</td>
</tr>
</tbody>
</table>

X-Dynalign $90.3 \pm 5.8$, Dynalign $= 87.7 \pm 7.4$. 
Reference, Dynalign and X-Dynalign structures for the 5S rRNA K02682

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

Introduction

Inference

Search

Marcel Turcotte

BNF 5106. Bioinformatics
Cons: eXtended Dynalign

- $O(|S_1|^2M^4)$ space, $O(|S_1|^3M^6)$ 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
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 bases)
|SCO: 201.40|POS:6-56|MIS: 0|WOB: 0|
|CAGCU|GAUGCU|AGCU|GAUGCU|AGCU|-|GAUCG|UAGCUAGU|CGAUC|CGU|AGCUG|
...
B.

```c
params
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'] );
}
```
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.
# STOCKHOLM 1.0

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

Holley

#=GR Holley SS

GGCGTGTTGGCGTAGCGGCGCTCCCTTAGCATGGAGAGGtCTCCGGTGATTCCGGAACGTCCA

(((((.(........)))))(((..........)))·(((..........)))·(((..........)))·(((..........)))·(((..........))))·

//
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 $O(NM^3)$ 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 $O(NM^3)$ 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”

- 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
//
**tRNAscan-SE**

- tRNAscan and EufindtRNA identify candidates that are subsequently analysed by Cove (INFERNAL)
- 1 false positive per 15 billion nt
- Detect 99% of true tRNA
- [http://lowelab.ucsc.edu/tRNAscan-SE/](http://lowelab.ucsc.edu/tRNAscan-SE/)
Sequence alignment methods are (generally) not appropriate for RNA.
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


Y. Ding and C. E. Lawrence.  
A bayesian statistical algorithm for rna secondary structure prediction.  

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.  
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.

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.
References (cont.)

Robert J Klein and Sean R Eddy.  
RSEARCH: finding homologs of single structured RNA sequences.  

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.)


M. Zuker.
On finding all suboptimal foldings of an RNA molecule.

M. Zuker and P. Stiegler.
Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.

Michael Zuker and David Sankoff.
RNA secondary structure and their prediction.
Please don’t print these lecture notes unless you really need to!