RNA Secondary Structure Prediction: taking conservation into account
1. נשים ובראשן בע 请求 של סכום (סמכות)
   כל רעיון של סמכויות

2. סכום של חומרים
   התוכן המרשים המ isp
   כלharma של קווים

3. שיטה להכנת התוכן
   כל קטע של הפקה
   כל משל של זה軟
Main approaches to RNA secondary structure prediction

- **Energy minimization (Single-strand Folding)**
  - does not require prior sequence alignment
  - require estimation of energy terms contributing to secondary structure (could be based on parameter-learning)

- **Comparative structure analysis**
  - Using sequence alignment to find conserved residues and covariant base pairs.
  - most trusted
Main approaches to RNA secondary structure prediction

- **Energy minimization (Single-strand Folding)**
  - does not require prior sequence alignment
  - require estimation of energy terms contributing to secondary structure (could be based on parameter-learning)

- **Comparative structure analysis**
  - Using sequence alignment to find conserved residues and covariant base pairs.
  - most trusted
ContextFold:

RECOMB 2011 Best Paper Award
Main approaches to RNA secondary structure prediction

- Energy minimization (Single-strand Folding)
  - does not require prior sequence alignment
  - require estimation of energy terms contributing to secondary structure (could be based on parameter-learning)

- Comparative structure analysis
  - Using sequence alignment to find conserved residues and covariant base pairs.
  - most trusted
Main approaches to RNA secondary structure prediction

- **Energy minimization (Single-strand Folding)**
  - does not require prior sequence alignment
  - require estimation of energy terms contributing to secondary structure (could be based on parameter-learning)

- **Comparative structure analysis**
  - Using sequence alignment to find conserved residues and covariant base pairs.
  - most trusted
Witness 3: Compensatory Mutations (in stems)

\[ \text{G-U} \rightarrow \text{U-A} \]

\textbf{Lactobacillus acidophilus}

\textbf{Lactobacillus delbrueckii}
Witness 3: Compensatory Mutations (in stems)

Lactobacillus acidophilus

\[
\begin{array}{c}
\text{G-C} \quad \text{C-G}
\end{array}
\]

Lactobacillus delbrueckii
Another source of information: Conserved Structure

- RNA more conserved by structure than by sequence. Conserved base-pairs should be taken more seriously than unconserved ones, even though the energy is the same.
Inferring Structure By Comparative Sequence Analysis

- most reliable computational method for determining RNA secondary structure

- consider the example from Durbin, et al., p 266
Covariation

When two columns are not identical in their contents, but they change together in a regular manner, they bear on the potential establishment of a relationship between them. The relationship between the columns fulfills the conditions required for the establishment of a relationship between them in the context of STEM. Therefore, it is reasonable to hypothesize that they represent A C G T G G A G A A C G G
A C C C T A A A G G G G A
T A T A G C A A T T A T C
C G G A T T A G T T C C G
G A T T G G A C G A A T A
G G G C T A A A T G C C A
Ron Unger – Bar-Ilan University 2009
Inferring Structure By Comparative Sequence Analysis

- first step is to calculate a multiple sequence alignment

- Requires sequences be similar enough so that they can be initially aligned

- Sequences should be dissimilar enough for covarying substitutions to be detected
Escherichia coli
Hildenbrandia rubra
Bangia fuscopurpurea
Rhodochaete parvula
Cordyceps kanzashi
Stichococcus bacillarii
Graphiola phoenicis
Single Strand Folding

input: A - U
      G - C connected by edge

AACAG C U G

output

AGC u

Multiple Strand Folding

input: each pair of columns connected by edge

output
Mutual Information

\[ M_{ij} = \sum_{x_i, x_j} f_{x_i x_j} \log_2 \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}} \]

- \( f_{xi} \): frequency of a base in column \( i \)
- \( f_{xixj} \): joint (pairwise) frequency of a co-appearance of a pair between columns \( i \) and \( j \)
- Information ranges from 0 and 2 bits
- If \( i \) and \( j \) are uncorrelated, mutual information is 0
תקני לינוב ערך שנמצא בקצירה המודגש בחלק מ meanwhile. בהתקפה שלdur.jpg, בביתה של כל חודש הוצאת לשונית הпущית (סמש החרוב”).

:**א.2.1**

הראות את התוכנית המושפעת نحو שוף המדהיהUSIC 1כית (כלומר, רענו הירוסים בקצירה).

$$H(m, n) = \frac{1}{n} \log_2 \left( \frac{n}{m} \right)$$

**א.2.2**

הראות את התוכנית המושפעת של проверת כל שפים הпущенית (סמש החרוב’). בממדיה אחד הпущенית בממדיה שעון (סמש החרוב’). בממדיה אחד הпущенית בממדיה שעון (סמש החרוב’).

**א.2.3.2.1**

$$f(m, n) = \frac{1}{n} \log_2 \left( \frac{n}{m} \right)$$
Single Strand Folding

input: $\text{A-U, G-C connected by edge}$

output

Multiple Strand Folding

input: $\text{each pair of columns connected by edge}$

output
Folding a multiple alignment instead of a single sequence

\[ M_{ij} \]

\[
L(i, j) = \max_{i \leq q \leq j} \left( \text{max} \left( \begin{array}{c}
\text{A} & \text{C} & \text{G} & \text{T} & \text{G} & \text{G} & \text{A} & \text{G} & \text{A} & \text{A} & \text{C} & \text{G} & \text{G} \\
\text{A} & \text{C} & \text{C} & \text{C} & \text{T} & \text{A} & \text{A} & \text{A} & \text{G} & \text{G} & \text{G} & \text{G} & \text{A} \\
\text{T} & \text{A} & \text{T} & \text{A} & \text{G} & \text{C} & \text{A} & \text{A} & \text{T} & \text{T} & \text{T} & \text{T} & \text{C} \\
\text{C} & \text{G} & \text{G} & \text{A} & \text{T} & \text{T} & \text{A} & \text{A} & \text{G} & \text{T} & \text{T} & \text{T} & \text{C} \\
\text{G} & \text{A} & \text{T} & \text{T} & \text{G} & \text{G} & \text{A} & \text{C} & \text{G} & \text{A} & \text{A} & \text{T} & \text{A} \\
\text{G} & \text{G} & \text{G} & \text{C} & \text{T} & \text{A} & \text{A} & \text{A} & \text{T} & \text{G} & \text{C} & \text{C} & \text{A} \\
\end{array} \right) \right) \right)
\]
Time Complexity of MIXY

- Multiple alignment (NP Hard)
- All-pairs mutual information for m sequences of length n each: $O(n^2 \, m)$
- Nussinov-Jacobson algorithm $O(n^3)$
Figure 5.12. Covariation found in tRNA sequences reveals base interactions in tRNA secondary and tertiary structure. (A) Alignment of tRNA sequences showing regions of interacting base pairs. (+) Transition; (−) transversions; (−) deletion; (∗) ambiguous nucleotide. (B) Diagram of tRNA structure illustrating base–base interactions revealed by a covariance analysis. Adapted from the Web site of R. Gutell at http://www.rna.icmb.utexas.edu.
Mutual Information Plot
Mutual Information Plot

[Diagram of yeast tRNA-Phe with labeled stems and nucleotide positions]
Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome

Stefan Washietl¹, Ivo L Hofacker¹, Melanie Lukasser², Alexander Hüttenhofer² & Peter F Stadler³,⁴.

Nature Biotechnology Volume 23  Number 11  November 2005

לחפש אזורים שמורים ב 4 גנؤمن (אדם, עכבר, חולדת כלב) שאינםמקודדים ולחלבונים יש لهم את הפוטנציאל ليוצר מבנה שאיננו יציב.
Figure 3 Selected examples of candidates for novel structural RNAs detected with $P > 0.9$. Predicted consensus structures with annotation of consistent and compensatory mutations are shown. Circles indicate variable positions in stems, colors indicate the number of different types of base pairs that support stabilizing selection on the structure. (a,b) These structures conserved across all vertebrates can be unambiguously identified as microRNA precursors on the basis of several characteristic features: (i) a stable hairpin consensus structure; (ii) the sequence of one arm of the stem is highly conserved over 22 nt (the putative mature miRNA); (iii) the opposite stem is also conserved but not that strictly; (iv) the loop sequence is diverged due to the absence of functional constraints in this region; (v) compensatory, or at least consistent, mutations are found in the outer parts of the stem where only structure but not sequence is important for function. See also the alignment in Figure 1d illustrating these typical microRNA features. (c,d) Candidates for novel H/ACA snoRNAs identified by secondary structure and primary sequence motifs. Both candidates fold into the typical bipartite hairpin secondary structure. We observe H-box motifs ANANNA in the hinge regions and ACA motifs in the tail regions. (e–h) Novel structural RNA elements. The sequence of structure e has similarity to a transcript in the Chr. 7 set of RNAdb and is conserved in mammals. We found more than 50 conserved secondary structures throughout the genome with sequence similarity to this transcript. Within these hits, we could identify this structural motif seven times by visual inspection. The structures f–h are conserved across all vertebrates and have particularly strong RNAz signals. Structure f is located near an intron/exon boundary and EST data suggests alternative splicing events in this region. Structure h is also located in an intron of a coding gene. It is an extremely stable stem-loop structure, which is longer than a typical microRNA precursor and also shows a different mutational signature. Additional RNAz hits in the close vicinity suggest that this is a local substructure of a longer RNA. Genomic locations of all examples (based on hg17 assembly): (a) chr.20:33,041,857 (intron of a myosine protein gene, AB040945); (b) chr.15:43,512,536 (UTR region of FOAP-11, AF228422); (c) chr.9:92,134,300 (intron of isoleucine-tRNA synthetase, D28473); (d) chr.16:2,786,411 (near a pseudogene of ribosomal protein 27A, flanked by LINE elements); (e) chr.12:74,595,654 (intergenic); (f) chr.22:18,488,478 (in intron of RAN binding protein 1, D38076); (g) chr.8:57,457,661 (intergenic); (h) chr.5:32,415,412, (intron of zinc-finger RNA binding protein, AJ314790).