The Tedious Task of Finding Common RNA Sequence Structure Properties
AND
Lattice Models and Energy Landscape

Rolf Backofen

Lehrstuhl für Bioinformatik
Albert-Ludwigs-Universität Freiburg

13. April 2011
RNA

The rise and rise of RNA
Jun 14th 2007
From The Economist print edition

IT IS beginning to dawn on biologists that they may have got it wrong. Not completely wrong, but wrong enough to be embarrassing. For half a century their subject had been built around the relation between two sorts of chemical. Proteins, in the form of enzymes, hormones and so on, made things happen. DNA, in the form of genes, contained the instructions for making proteins. Other molecules were involved, of course. Sugars and fats were abundant (too abundant, in some people). And various vitamins and minerals made an appearance, as well. Oh, and there was also a curious chemical called RNA, which looked a bit like DNA but wasn't. It obediently carried genetic information from DNA in the nucleus to the places in the cell where proteins are made, rounded up the amino-acid units out of which those proteins are constructed, and was found in the protein factories themselves....
Overview

- Motivation
- sequence motif with secondary structure properties.
- finding the structure: RNA sequence/structure alignment
- RNA classification: clustering RNA into structural classes
- barrier trees and HP-lattice models
RNA-Molecules

- RNA:
  - bonds = secondary structure
  - hierarchical folding
  - secondary structure first

- properties
  - before: simple “transport element”
    DNA $\rightarrow$ RNA $\rightarrow$ protein
  - now: many functions
    - ribozyme: RNA-enzymes
    - non-coding RNAs, regulation etc.
    - RNA: scientific breakthrough 2002
    - Nobel prize for medicine 2006: RNAi
How many possible ncRNA out there? (Cont)

In humans:

- approx. 1% of genome encodes protein
- approx. 80-90% of genome is transcribed
  - at least 98% of transcribed RNA is non-coding
In humans: approx. 1% of genome encodes protein. Approx. 80-90% of genome is transcribed, so at least 98% of transcribed RNA is non-coding.
How many possible ncRNA out there? (Cont)

In humans:
- approx. 1% of genome encodes protein
- approx. 80-90% of genome is transcribed $\Rightarrow$ at least 98% of transcribed RNA is non-coding
How many possible ncRNA out there? (Cont)

In humans:
- approx. 1% of genome encodes protein
- approx. 80-90% of genome is transcribed ⇒ at least 98% of transcribed RNA is non-coding
Riboswitches:  
- cis-acting RNA-elements included in the mRNA
- can detect different metabolites.
- direct regulation of associated mRNA
Riboswitches: cis-acting RNA-elements included in the mRNA can detect different metabolites. direct regulation of associated mRNA
Riboswitches:
- cis-acting RNA-elements included in the mRNA
- can detect different metabolites.
- direct regulation of associated mRNA
Examples I: 6S-RNA

Population size: few bacteria vs. many bacteria

Cell state: DNA, RNA, and transcription rate

Feedback loop: pRNA and 6S RNA
Viroids

- small RNA pathogens infecting plants (240 – 400 nt)
- first identified: Potato spindle tuber viroid (PSTVd)
- viroids are pure RNA, no protein, no capsule
- smallest known self-replicating unit
Viroid Replication with Hammerhead

- viroid: circular + strand
- host RNA-polymerase generates a longer − strand by going through the circular genome more than once
- new, longer plus strand is then synthesized by the host RNA polymerase
- split into viroid-units by self-cleavage through Hammerhead ribozyme
Viroid Replication with Hammerhead

- viroid: circular + strand
- host RNA-polymerase generates a longer − strand by going through the circular genome more than once
- new, longer plus strand is then synthesized by the host RNA polymerase
- split into viroid-units by self-cleavage through Hammerhead ribozyme
Viroid Replication with Hammerhead

- rev. copy in host cell
- copy by host cell
- hammer-head
- circularization

- viroid: circular + strand
- host RNA-polymerase generates a longer − strand by going through the circular genome more than once
- new, longer plus strand is then synthesized by the host RNA polymerase
- split into viroid-units by self-cleavage through Hammerhead ribozyme
Example: Ribozymes

- Ribozymes: RNA enzyme

- Hammerhead-Ribozyme:
  - detected as site-specific self-cleavage unit
  - many variants with different specificity generated
  - requires only two metal atom

Lilley et al. ChemBio2002
Function is determined by **sequence** and **structure**

Next generation sequencing technologies allow high-throughput data collection of **sequence** information

...but high-throughput **structure** determination is (still) mostly done algorithmically
Overview

- Motivation
  - sequence motif with secondary structure properties.
  - finding the structure: RNA sequence/structure alignment
  - RNA classification: clustering RNA into structural classes
  - barrier trees and HP-lattice models
mRNA is Loaded with Proteins
mRNA is Loaded with Proteins
one important feature missing: secondary structure
example: fibronectin EDA exon

Buratti et al. Mol and Cell Bio. 24(3) 2004
MEMERIS

- often: additional knowledge
  RNA: some splice factors prefer single-stranded sites
  TFs: distances to TATA box, structural contexts, ...

integration into EM?

prior probabilities on start positions

[Hiller et al. NAR 2006]
Experimental Testing

inserts with known splicing motifs (TAGGGT, hnRNP A1)

[Hiller et al. PLoS Genetics 2007]
Experimental Testing

⇒ secondary structure of ESE/ESS affects splicing

[Hiller et al. PLoS Genetics 2007]
Overview

- Motivation
- sequence motif with secondary structure properties.
- finding the structure: RNA sequence/structure alignment
- RNA classification: clustering RNA into structural classes
- barrier trees and HP-lattice models
RNA Secondary Structure Prediction

- Turner energy-model: free energies for *loops*
- Efficient calculation of minimal free energy (MFE) structure

Mouse tRNA-ALA:

```
G
G
G
G
G
U
A
UA
GCUC
AGG
G
G
U A
G A G C
A U
U
U
G
A
C
U
G
C
A G
A
U
C
A
A
G
A
G
G
U
C
C
CU
G
G
U
U
C
A
A
A
U
C
C
A
GG
U
G
C
C
C
C
C
U
```
RNA Secondary Structure Prediction

- Turner energy-model: free energies for *loops*
- efficient calculation of minimal free energy (MFE) structure
- problem: MFE is often wrong
- Mouse tRNA-ALA:
- Mouse tRNA-CYS:
RNA Secondary Structure Prediction

- Turner energy-model: free energies for loops
- Efficient calculation of minimal free energy (MFE) structure
- Problem: MFE is often wrong
- Mouse tRNA-ALA:
- Mouse tRNA-CYS:
RNA Secondary Structure Prediction

- Turner energy-model: free energies for *loops*
- efficient calculation of minimal free energy (MFE) structure
- problem: MFE is often wrong
- Mouse tRNA-ALA:

Mouse tRNA-CYS:
RNA Secondary Structure Prediction

- Turner energy-model: free energies for loops
- efficient calculation of minimal free energy (MFE) structure
- problem: MFE is often wrong
- Mouse tRNA-ALA:

Mouse tRNA-CYS:

\[
\Delta G^{\text{II}}(S_i, S_{\text{mfe}})
\]

\[
\delta_{BP}(S_i, S_{\text{mfe}})
\]
Comparative RNA Analysis

Adopted from: [Gardner & Giegerich BMC 2004]

Consensus:

Consensus structure:

Plan A

Consensus

Structure

Alignment

FOLD

Plan B

ALIGN and FOLD simultaneously

[Sankoff 85]

Plan C

ALIGN and FOLD simultaneously

Consensus:

Consensus structure:

Adopted from:

[Gardner & Giegerich BMC 2004]
Comparative RNA Analysis

**Plan A**
- ALIGN single sequences
- FOLD alignment

**Plan B**
- A:
- B:

**Plan C**
- FOLD single sequences
- ALIGN sequence AND structure
- Plan B: ALIGN and FOLD simultaneously
  - [Sankoff 85]
- Plan A: consensus structure

**Alignment**
- A:
- B:

**Consensus Structure**
- A:
- B:

- adopted from: [Gardner & Giegerich BMC 2004]
Comparative RNA Analysis

Plan C

ALIGN sequence AND
FOLD single sequences

consensus:
consensus structure:

Plan B
ALIGN and FOLD simultaneoulsy
[Sankoff 85]

consensus:
consensus structure:

Plan A
consensus
structure

FOLD

single
sequences

A:
B:
single

A:
B:

A:
B:

A:
B:

Tree representation of RNA

- two representations of RNA secondary structure

- bonds between complementary nucleotides

- edit operation on trees
Zhang and Sascha’s Method

- **Associated Recursion Equation**
  - \( \delta(\emptyset, \emptyset) = 0 \)
  - \( \delta(F, \emptyset) = \delta(F - r_F, \emptyset) + c_{\text{del}}(r_F) \)
  - \( \delta(\emptyset, G) = \delta(\emptyset, G - r_G) + c_{\text{del}}(r_G) \)
  - \( \delta(F, G) = \min \left\{ \begin{align*}
    &\delta(F - r_F, G) + c_{\text{del}}(r_F), \\
    &\delta(F, G - r_G) + c_{\text{del}}(r_G), \\
    &\delta(R_F^\circ, R_G^\circ) + \delta(F - R_F, G - R_G) + c_{\text{match}}(r_F, r_G)
  \end{align*} \right\} \)

- \( F^\circ \) is the special case of \( F - r_F \) for \( F \) rooted
- \( R_F \) denotes the rightmost child in \( F \)
- \( O(n^2m^2) \) algorithm with \( n = |F_1| \) and \( m = |F_2| \)

- Zhang and Sascha: fewer subproblems are needed
- *relevant* problems: prefix of \( F^\circ \), where \( F \) is a root having degree \( > 2 \).
Same in our DP Notation

- forests $F$ and $G$: all regions $[i..i']$ and $[j..j']$.
  $\Rightarrow O(n^4)$ space and $O(n^6)$ time

- $\delta(F, G)$ then corresponds to $D(i, i', j, j')$ (alignment of subsequences)
- **But:** not all entries are considered

- **Hence:** $O(n^2)$-matrices $M_{a_2}^{a_1}(i, j)$ for all pairs of arcs $a_1, a_2$. 
Dynamic Programming for Sequence/Structure Alignment

- matrices $M_{a_2}^{a_1}(i, j)$: subsequences/substructures under arcs $a_1$ and $a_2$

- recursion:
Dynamic Programming for Sequence/Structure Alignment

- matrices $M_{a_1 a_2}^a(i, j)$: subsequences/substructures under arcs $a_1$ and $a_2$

![Diagram of subsequences/substructures]

- recursion:
  - base (mis-)match, indel $\Rightarrow M_{a_2}^{a_1}(i - 1, j)$ ... existing arcs broken in this case

![Diagram of recursion]

[Backofen&Will: JCB 04]
Dynamic Programming for Sequence/Structure Alignment

- matrices $M_{a_2}^{a_1}(i,j)$: subsequences/substructures under arcs $a_1$ and $a_2$

recursion:
- base (mis-)match, indel $\Rightarrow M_{a_2}^{a_1}(i - 1,j)$  
  existing arcs broken in this case
- arc match $\Rightarrow M_{a_2}^{a_1}(i' - 1,j' - 1) + M_{a_2}^{a_1'}(i,j)$

[Backofen&Will: JCBC 04]
Comparative RNA Analysis

Plan B

A:
B:

simultaneously
ALIGN and FOLD

[Sankoff 85]

A:
B:

consensus:
consensus structure:
Sankoff-like approaches

- Sankoff is the gold standard **BUT requires extreme amount of space and time** [Gardner & Giegerich 2004]
  
  (time: $O(n^6)$, space $O(n^4)$)

- hence: Sankoff-like approaches are restricted versions

Problem: Suboptimal Structures

- example: two hammerhead ribozymes
Problem: Suboptimal Structures

- example: two hammerhead ribozymes

- corresponding dotplots
Problem: Suboptimal Structures

- example: two hammerhead ribozymes

LocARNA
- alignment of dotplots
- efficient version of Sankoff

Probabilistic Consistency Transformation

- **remaining problem:** progressive alignment
  
  *use probabilistic consistency transformation ala ProbCons*

- **Idea:**
  
  - Given set of sequences $S$
  - for all pairs $x, y \in S$ of sequences calculated:
    
    *match probabilities $P(x_i \sim y_j | x, y)$*

Then:

[Diagram showing the relationship between $x_i$, $y_j$, $y_j'$, and a question mark.]
Probabilistic Consistency Transformation

- **remaining problem:** progressive alignment

  use probabilistic consistency transformation ala ProbCons

- Idea:
  - Given set of sequences $S$
  - for all pairs $x, y \in S$ of sequences calculated:
    
    $\text{match probabilities } P(x_i \sim y_j | x, y)$

Then:

\[ P(z_k y_j \sim ) y, z P(x_i z_k \sim ) z, x P(x_i y_j \sim ) y, x P(z_k y_j \sim ) y, z P(x_i z_k \sim ) z, x \]

do this for all intermediate sequences $z \in S$
remaining problem: progressive alignment

use probabilistic consistency transformation ala ProbCons

Idea:

- Given set of sequences $S$
- for all pairs $x, y \in S$ of sequences calculated:
  
  * match probabilities $P(x_i \sim y_j | x, y)$

Then:

$$P(z_k \sim y_j | z, y) \cdot P(x_i \sim z_k | x, z)$$

$$P(x_i \sim z_k | x, z) * P(z_k \sim y_j | z, y)$$
**remaining problem:** progressive alignment

*use probabilistic consistency transformation ala ProbCons*

**Idea:**
- Given set of sequences $S$
- for all pairs $x, y \in S$ of sequences calculated:
  - *match probabilities* $P(x_i \sim y_j | x, y)$

Then:
- do this for all intermediate sequences $z \in S$
Reliability Profiles

- genomic cluster with known ncRNAs
- align corresponding regions in 10/5 vertebrates
- show reliability profile for human DNA

cluster of 6 micro RNAs, length \(\approx 900\)

cluster of 10 CD-Box snoRNAs 'GAS5', length \(\approx 4000\)

Boundary Prediction

- median
- use LocaRNA/ boundary prediction / reliability as post-processing filter
Overview

- Motivation
- sequence motif with secondary structure properties.
- finding the structure: RNA sequence/structure alignment
- RNA classification: clustering RNA into structural classes
- barrier trees and HP-lattice models
Classification of putative ncRNAs

- RNAz: detects functional RNA secondary structures in **multiple sequence alignments**

- results of human RNAz-scan
Classification of putative ncRNAs

- **RNAz**: detects functional RNA secondary structures in multiple sequence alignments
- **Results of human RNAz-scan** [Washietl et al. Nature Biotech. 2005]

![Diagram showing classification process of ncRNAs]

- Known gene
- Intron of coding region
- `< 10 kb from nearest gene`
- `> 10 kb from nearest gene`
- `3’−UTR (exon or intron)`
- `5’−UTR (exon or intron)`
LocARNA: Clustering of RNAz ncRNA Predictions

Clustering of 3332 putative ncRNAs in Ciona intestinalis
LocARNA: Clustering of RNAz ncRNA Predictions

Clustering of 3332 putative ncRNAs in Ciona intestinalis

Classification of ncRNA: MicroRNA Example

- problem: how to classify ncRNAs from properties of RNA 2D structure
  ⇒ learn graph properties

A sequence and predicted hairpin secondary structure: only stem portions (shadow regions) of the hairpin are computed.

```
CUUUCUACACAGGUUGGGAUCGGGUUGCAUGCUUGUGUUCUGUAUGGGAUUGCAUGUUGCCGACUGUUGAGGUUUGG
```

Triplet element: continuous three structures with middle nucleotide. Taking "(" as the same as ")".

32 triplet element features --- 32-dimension vector:
```
(U{(}, U{., U{., U{., U{., U{., U{., U{., U{., U{., G{{}, G{{., ...}
```

Counting the appearances of the triplet elements:
```
(12, 4, 3, 1, 2, 0, 0, 0, 10, 1, ...)
```

Normalizing the triplet element count vector:
```
(0.1846, 0.0615, 0.0462, 0.0154, 0.0308, 0, 0, 0, 0.1538, 0.0154, ...)
```
Clustering: How to cluster RNAz predictions?

- Problem: still too much data for LocaRNA
  
  16,000 Drosophila or 36,000 human RNAz hits

- Solution: modified cluster pipeline (Fabrizio Costa)
  - Built graphs (using RNASHapes) from RNA sequences
  - Convert them to high dimensional sparse vectors (graph kernel)
  - Use LSH to efficiently retrieve neighbors and density
  - Return highly dense clusters
  - Refine RNA family models in clusters by LocARNA

![Diagram showing the workflow of the modified cluster pipeline](chart.png)
Input graphs with RNAshapes

search context (C)
region of influence (R)
target mRNA
miRNA
folding window size (W)

RNAshapes Bioinformatics, 22(4), 2006
Features: all pairs of near small subgraphs

Interpretation: consider the occurrence of each subgraph in the context provided by the other subgraphs
Overview

- Motivation
- sequence motif with secondary structure properties.
- finding the structure: RNA sequence/structure alignment
- RNA classification: clustering RNA into structural classes
- barrier trees and HP-lattice models
Kinetic versus Thermodynamic Folding

- so far: consideration of thermodynamic stable folding
  *minimum free energy*

- however: folding is a kinetic process
  *suboptimal structure favourable*

- example: co-transcriptional folding of RNA

- technique: investigation of energy landscapes
Landscape Schemes

- smooth
- energetic trap
- folding intermediate
- unfolding trap
- entropic trap
- superfunnel like
**HP Lattice Model of Proteins**

- only backbone structure positions $\equiv$ lattice positions

- simplified energy function: e.g. only **hydrophobic force**
  native $=$ maximal number of HH-contacts

![cubic FCC](image)
General Approach To Lattice Folding

- Algorithm consist of three steps:
- Step 1 and 2 are precomputation steps

Step 1:
- compute lower energy bounds
- estimate contacts (within layers, between layers)

Step 2:
- construct hydrophobic cores
- use bounds from last step, precomputed

Step 3:
- thread sequence to hydrophobic cores of size $n$
- using constraint propagation

⇒ Step 1 ⇒ Step 2 ⇒ Step 3

[Backofen&Will: CP2001]
General Approach To Lattice Folding

- Algorithm consist of three steps:
  - Step 1 and 2 are precomputation steps
    - **Step 1:** compute lower energy bounds
      - estimate contacts (within layers, between layers)
    - **Step 2:** construct hydrophobic cores
      - use bounds from last step, precomputed
    - **Step 3:** thread sequence to hydrophobic cores of size $n$
      - using constraint propagation

$\Rightarrow$ Step 1

[Backofen&Will: CP2001]
Algorithm consists of three steps:

Step 1 and 2 are precomputation steps

Step 1: compute lower energy bounds
estimate contacts (within layers, between layers)

Step 2: construct hydrophobic cores
use bounds from last step, precomputed

Step 3: thread sequence to hydrophobic cores of size $n$
using constraint propagation

$\Rightarrow$ Step 1 $\Rightarrow$ Step 2 $\Rightarrow$ Step 3
General Approach To Lattice Folding

- Algorithm consists of three steps:
  - **Step 1 and 2 are precomputation steps**
  - **Step 1**: compute lower energy bounds
    - estimate contacts (within layers, between layers)
  - **Step 2**: construct hydrophobic cores
    - use bounds from last step, precomputed
  - **Step 3**: thread sequence to hydrophobic cores of size $n$
    - using constraint propagation

⇒ Step 1 ⇒ Step 2 ⇒ Step 3
General Approach To Lattice Folding

- Algorithm consists of three steps:
  - Step 1 and 2 are precomputation steps
  - **Step 1:** compute lower energy bounds
    estimate contacts (within layers, between layers)
  - **Step 2:** construct hydrophobic cores
    use bounds from last step, precomputed

Step 1: $n_1=2 \quad n_2=6 \quad n_3=8 \quad n_4=4$

Step 2:

⇒ Step 1
⇒ Step 2
General Approach To Lattice Folding

- Algorithm consist of three steps:
  - Step 1 and 2 are precomputation steps
  - Step 1: compute lower energy bounds
    estimate contacts (within layers, between layers)
  - Step 2: construct hydrophobic cores
    use bounds from last step, precomputed
  - Step 3: thread sequence to hydrophobic cores of size $n$
    using constraint propagation

$n_1=2$  $n_2=6$  $n_3=8$  $n_4=4$
## Comparison of Results

### Small Selection of Previous Approaches:

<table>
<thead>
<tr>
<th>Authors</th>
<th>Model</th>
<th>Dim.</th>
<th>Maxlen</th>
<th>Algorithm</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yue &amp; Dill PhysRevE93</td>
<td>Cubic HP</td>
<td>3</td>
<td>36</td>
<td>Branch-and-bound</td>
<td>Optimality proven</td>
</tr>
<tr>
<td>Yue &amp; Dill PNAS95</td>
<td>Cubic HP</td>
<td>3</td>
<td>88</td>
<td>Branch-and-bound</td>
<td>Optimality proven</td>
</tr>
<tr>
<td>Sazhin et al. 01</td>
<td>Cubic HP, FCC</td>
<td>3</td>
<td>34</td>
<td>Branch-and-bound</td>
<td>Not always optimal</td>
</tr>
<tr>
<td>Cui et al. PNAS02</td>
<td>Square HP</td>
<td>2</td>
<td>18</td>
<td>Comple. enum</td>
<td></td>
</tr>
<tr>
<td>Hart &amp; Istrail JCB97</td>
<td>FCC side chain</td>
<td>3</td>
<td>—</td>
<td>Approximation</td>
<td>86% of optimum</td>
</tr>
<tr>
<td>Agarwala et al. JMB97</td>
<td>FCC HP</td>
<td>3</td>
<td>—</td>
<td>Approximation</td>
<td>3/5 of optimum</td>
</tr>
</tbody>
</table>

### Our Results:

- Native conformation up to length **300**
- Proof of optimality
- Number of conformations of length $n$: $\approx 4.5^n$
  \[\Rightarrow \text{search space handled } \approx 4.5^{190} \text{ bigger}\]
- Only existing non-heuristic algorithm for **FCC**

<table>
<thead>
<tr>
<th>Seq.</th>
<th>Length</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>135</td>
<td>9 s</td>
</tr>
<tr>
<td>S2</td>
<td>151</td>
<td>15 s</td>
</tr>
<tr>
<td>S3</td>
<td>161</td>
<td>18 s</td>
</tr>
<tr>
<td>S4</td>
<td>164</td>
<td>11 s</td>
</tr>
</tbody>
</table>
Goal: quantification of complexity in self-organising biomolecular systems

- determination of ensemble of low energy conformations
- calculation of barrier-trees
- determination of kinetic parameters
Investigation of Landscape

- **Goal:** quantification of complexity in self-organising biomolecular systems
  - determination of ensemble of low energy conformations
  - calculation of *barrier-trees*
  - determination of kinetic parameters

- application to lattice proteins

[Wolfinger et al.: Europhysics Letters 2006]
Application: Design of protein-like Sequences

- find sequences with *exactly* one optimal structure
- stochastic local search

**node**: accepted sequences
**edges**: simulation step/mutation

Degeneracy

![Degeneracy Graph](image-url)
Take-Home Messages

3 major problems in RNA

- sequence/structure motifs $\Rightarrow$ RNA-binding proteins
  Memeris: sequence motifs with structural properties.

- RNA comparative structure prediction
  LocaRNA: currently most efficient Sankoff-like approach

- RNA classification: clustering
  graph-kernel based approach

lattice model: NP-hard problems are solvable

- NP-hard doesn’t mean that you cannot do it
- here: folding HP-models to optimality
- message: don’t be afraid, ask your local computer scientist
Acknowledgements

Our group
- Anke Busch
- Fabrizio Costa
- Steffen Heyne
- Michael Hiller
- Sita Lange
- Rileen Sinha
- Robert Kleinkauf
- Kousik Kundu
- Dominic Rose
- Andreas Richter
- Martin Mann
- Daniel Maticzka
- Mathias Möhl
- Sebastian Will

Wien/Leipzig
- Peter Stadler
- Ivo Hofacker
- Kristin Reiche

Erlangen/U. Kentucky
- Stefan Stamm

MIT
- Bonnie Berger

SFU Vancouver
- S. Cenk Sahinalp
- Hamidreza Chitsaz
- Raheleh Salari

DFG-Excellence Cluster “BIOSS”

BMBF “FRISYS”

DFG-SPP “regulatory bacterial RNAs”

DFG-SPP “InKomBio”
Thank You for Your Attention
LocARNA: Test on RFAM seed alignments

- ROC curve the global comparison of clustering and RFAM families
Clustering of bacterial ncRNA predictions

- tRNAs
- 16S rRNA substructures
- 16S rRNA: 223771-225312
- 16S rRNA: 2727638-2729179
- 16S rRNA: 4206170-4207711
- 16S rRNA: 3425243-3426784
- 16S rRNA: 3939831-3941372
- 16S rRNA: 4033554-4035095
- 16S rRNA: 4164682-4166223
- ec_96: 4164536-4165136
- ec_76: 3425441-3425800
- ec_3: 223849-224225
- ec_46: 2727636-2727796
- ec_48: 2728201-2728321
- ec_47: 2727836-272819
- ec_75: 3425241-3425401
- ec_77: 3425806-3425926
- ec_78: 3426385-3426705
- ec_82: 3940362-3940602
- ec_81: 3939685-3940285
- ec_89: 4033408-4034008
- ec_90: 4034085-4034325
- ec_4: 224302-224542
- ec_97: 4165213-4165453
- ec_110: 4206701-4206941
- ec_109: 4206024-4206624
Clustering of bacterial ncRNA predictions
Mapping graphs into vector spaces

Given a feature (a pair of near small subgraphs) compute an integer encoding via a hashing technique

Complexity dominated by edge sorting or all-pairwise-distance computation in small subgraphs $\rightarrow$ efficient (linear) in practice