SETTER - RNA SEcondary sTructure-based TERtiary Structure Similarity Algorithm

David Hoksza, Daniel Svozil

SIRET Research Group
Department of Software Engineering, FMP, Charles University in Prague

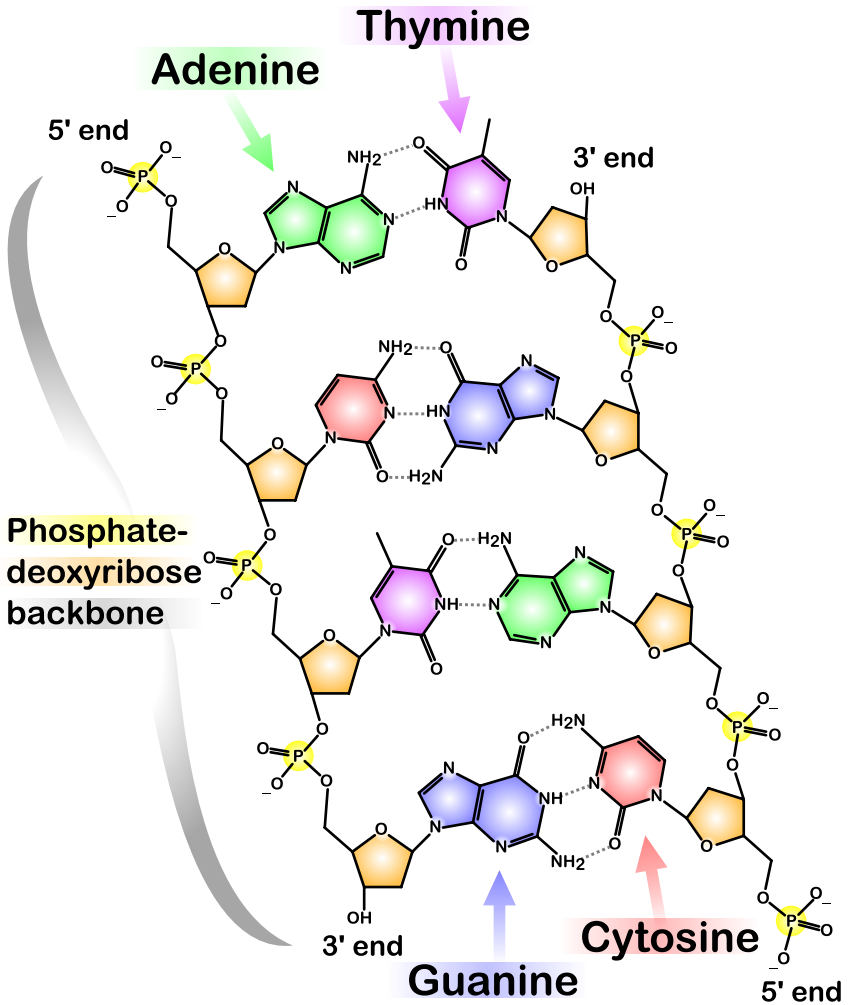
May 21, 2011

Outline

Genetic Information (DNA)

DNA (deoxyribonucleic acid)

- DNA strand (chain) of nucleotides
  - sugar-phosphate
  - base
- bases
  - adenine (A)
  - cytosine (C)
  - guanine (G)
  - thymine (T)
- doublestranded (Watson-Crick hydrogen base pairs)
  - A + T
  - G + C



DNA Expression

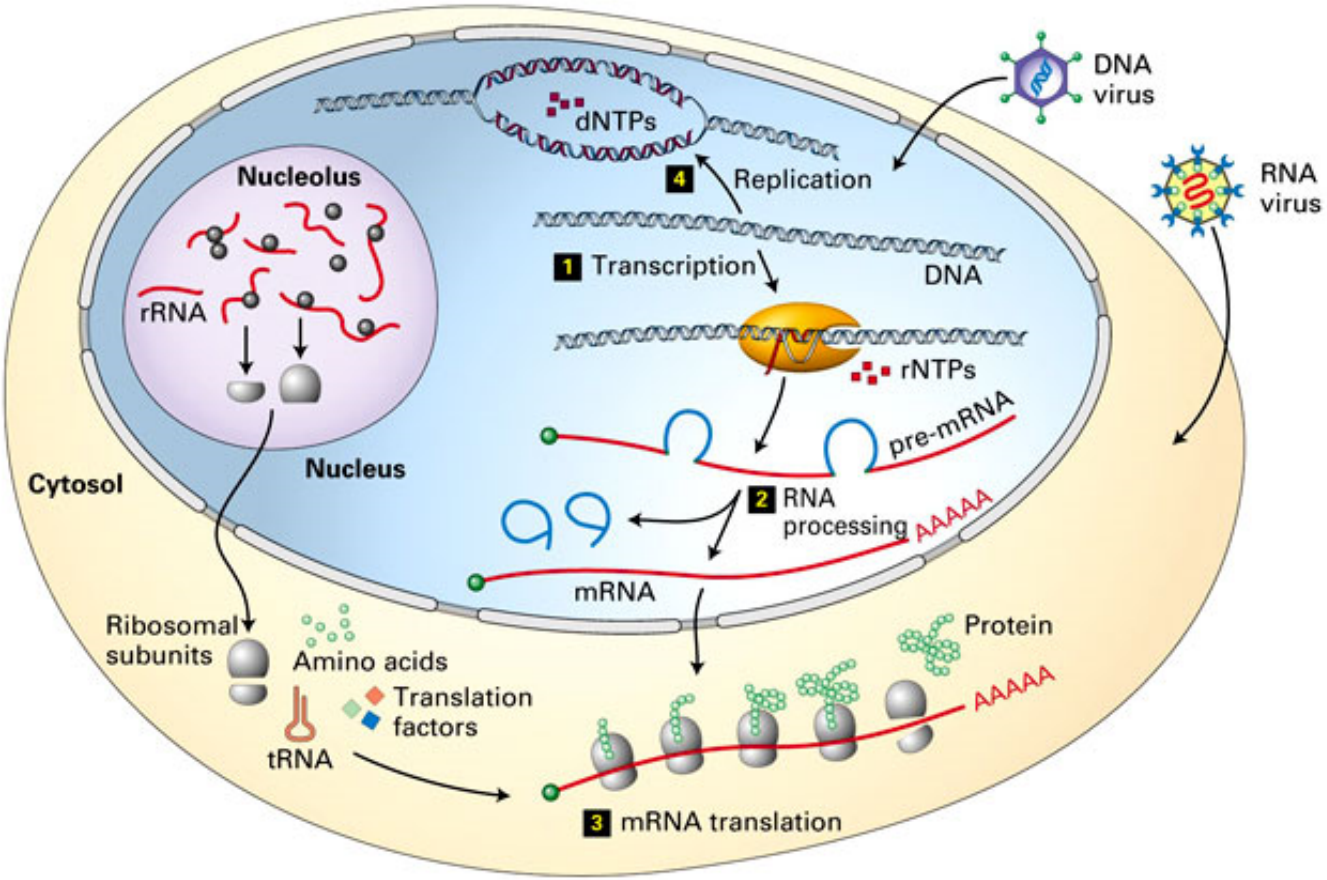
Central Dogma of Molecular Biology

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein}$$

DNA Expression

Central Dogma of Molecular Biology

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein}$$



RNA Function

- messenger RNA
  - DNA carrier
- transfer RNA
  - amino acid carrier
- ribosomal RNA
  - ribosome building blocks
- gene expression regulation
- catalytic function
- ...

RNA Function

- messenger RNA
  - DNA carrier
- transfer RNA
  - amino acid carrier
- ribosomal RNA
  - ribosome building blocks
- gene expression regulation
- catalytic function
- ...

- **function determined by structure**

RNA structure (PDB ID 1P6V)

Primary Structure

Secondary Structure

Tertiary Structure

RNA structure (PDB ID 1P6V)

Primary Structure

G G G G G C G G
A A A G G A U U C
G A C G G G G A
C U U C G G U C C
U C G G A C G C G
G G U U C G A U
U C C C G C C G C
C U C C A C C A

Secondary Structure

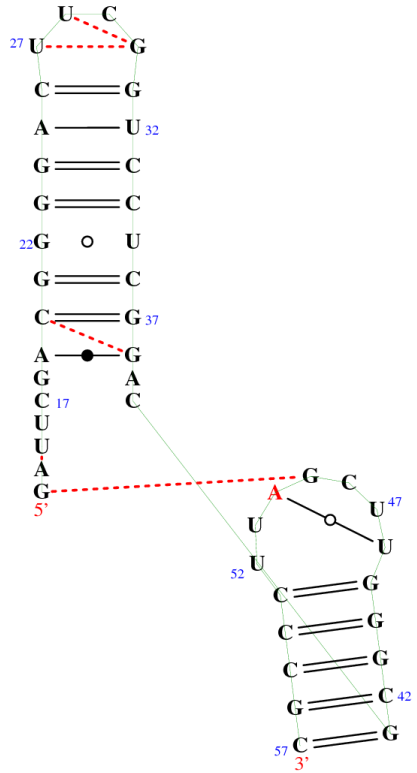
Tertiary Structure

RNA structure (PDB ID 1P6V)

Primary Structure

G G G G G C G G
A A A G G A U U C
G A C G G G G A
C U U C G G U C C
U C G G A C G C G
G G U U C G A U
U C C C G C C G C
C U C C A C C A

Secondary Structure



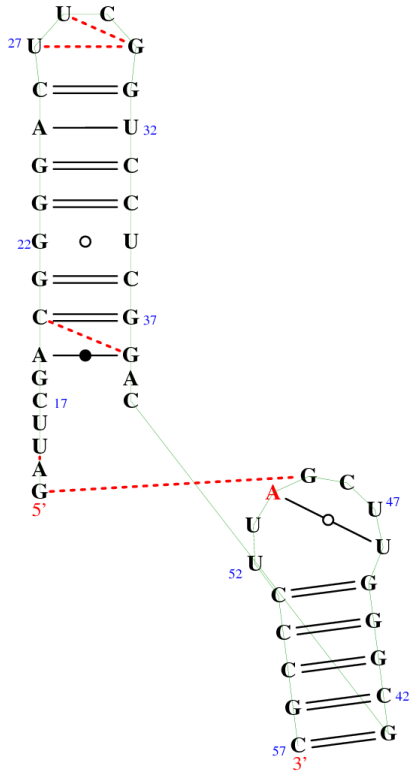
Tertiary Structure

RNA structure (PDB ID 1P6V)

Primary Structure

G G G G G C G G
A A A G G A U U C
G A C G G G G A
C U U C G G U C C
U C G G A C G C G
G G U U C G A U
U C C C G C C G C
C U C C A C C A

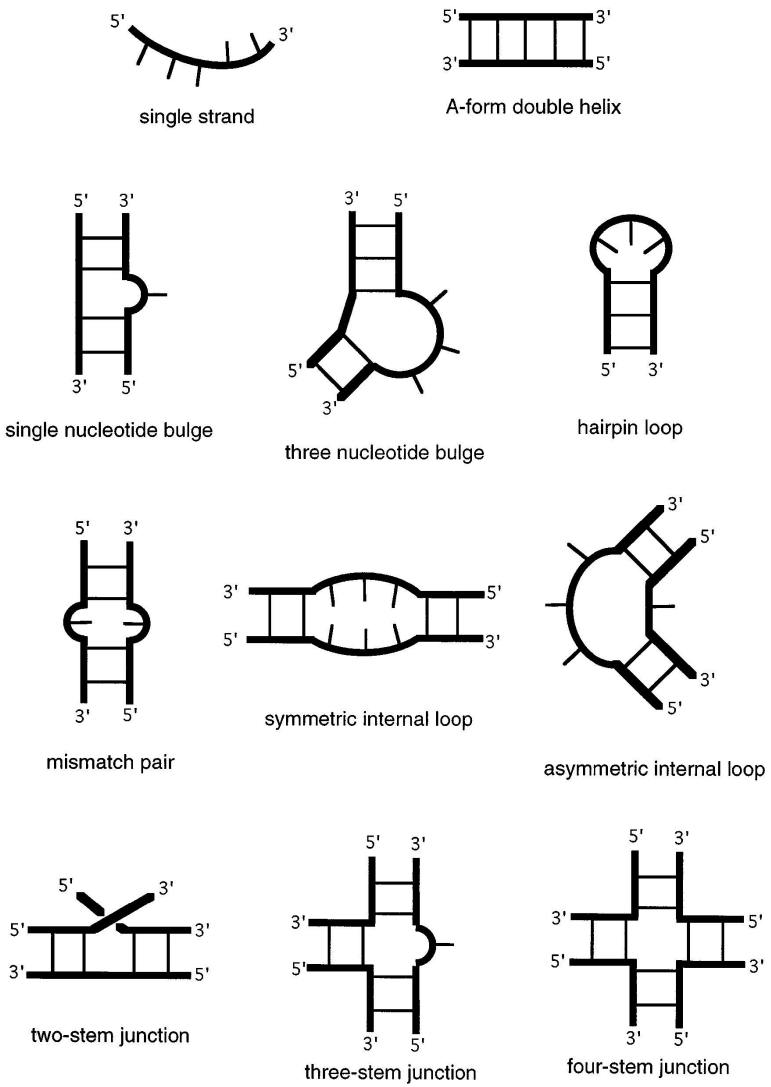
Secondary Structure



Tertiary Structure

Secondary Structure Motifs

- double helices combined with various types of loop structures
  - hairpin loop
  - internal loop
  - bulge loop
  - junction loop



RNA Databases

- Sequence Databases
  - Genbank
- Structure Databases
  - NDB — Nucleic Acid Database (5220 structures)
  - PDB — Protein Data Bank (2000 structures)
- Function Databases / Classification
  - SCOR — Structural Classification of RNA

RNA similarity

- pairwise sequence and structure similarity utilization:
  - RNA structure prediction
  - RNA function discovery
  - RNA design
  - RNA modeling

Outline

SETTER's Outline and Motivation

Outline:

- Two RNA structures
- Each structure is divided into *generalized secondary structure units* (GSSUs)
- Each pair of GSSUs (one from each of the structures) is superposed
- Optimal pair P of the GSSUs is identified
- Based on P, whole RNA structures are superposed and their distance is computed

SETTER's Outline and Motivation

Outline:

- Two RNA structures
- Each structure is divided into *generalized secondary structure units* (GSSUs)
- Each pair of GSSUs (one from each of the structures) is superposed
- Optimal pair P of the GSSUs is identified
- Based on P, whole RNA structures are superposed and their distance is computed

Motivation:

- Secondary structure represents clue to superposition
- Basic SSEs are too small and not enough robust

Generalized Secondary Structure Unit — GSSU

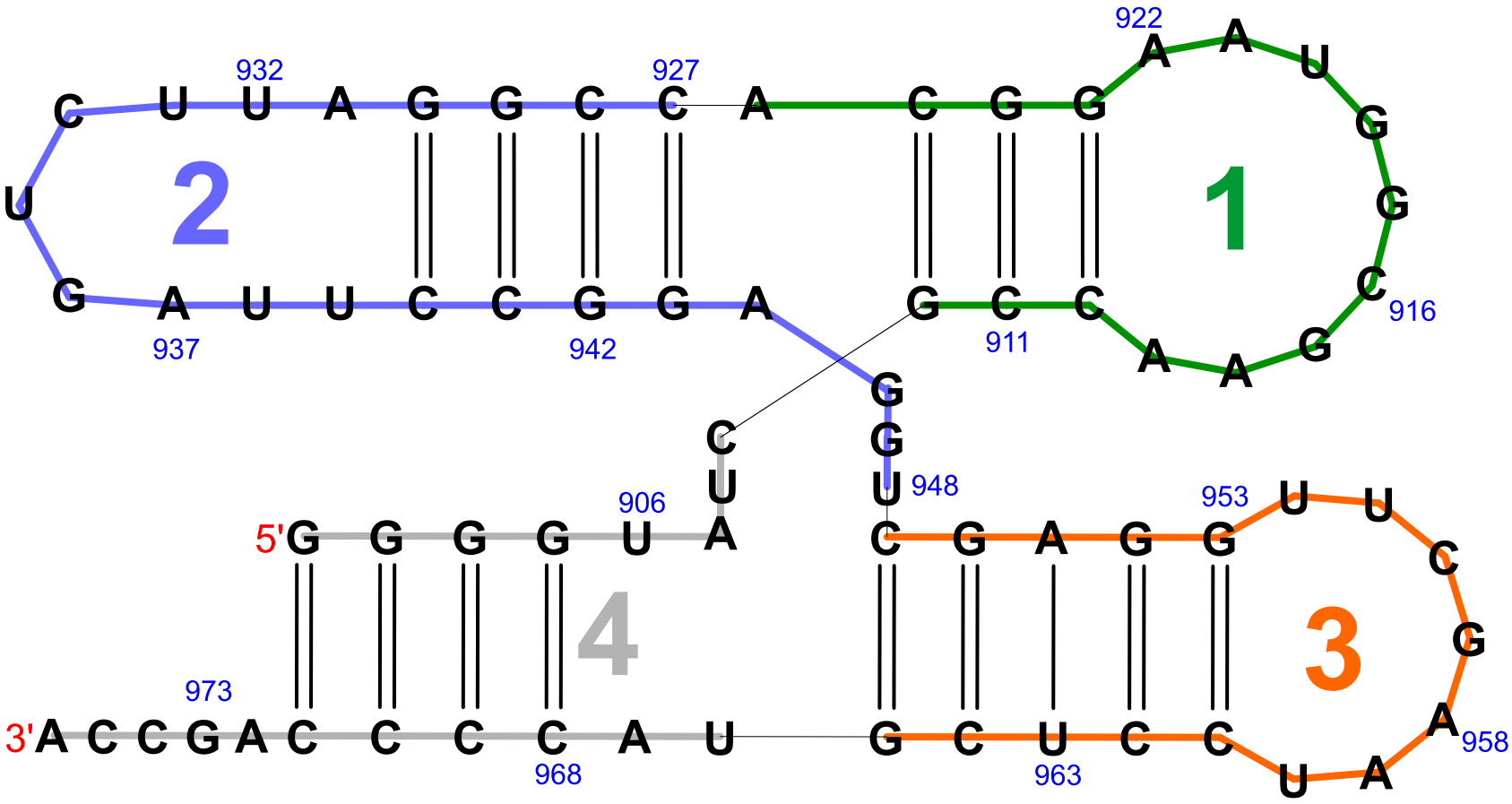
Definition

Let $\mathcal{R}$ be an RNA structure with nucleotide sequence $\{n_i\}_{i=1}^{n}$ and let $\mathcal{WC} \subset \mathcal{R}$ denote set of n_i participating in a Watson-Crick base pair. By **generalized secondary structure unit (GSSU)** $\mathcal{G}$, we understand a pair of substrings of $\mathcal{R}$, $\{n_i\}_{i=i_1}^{i_2}$ and $\{n_i\}_{i=j_1}^{j_2}$ of maximum lengths such that each nucleotide n_x:

- $i_1 \leq x \leq i_2 : n_x \notin \mathcal{WC}$ or n_x is paired with n_y where $j_1 \leq y \leq j_2$
- $j_1 \leq x \leq j_2 : n_x \notin \mathcal{WC}$ or n_x is paired with n_y where $i_1 \leq y \leq i_2$

Let i_{max} and j_{min} be highest indices of the Watson-Crick paired bases. We define **loop** as $\mathcal{L} = \{n_i\}_{i=i_{max}+1}^{j_{min}-1} \subset \mathcal{R}$ and **stem** as $\mathcal{R} \setminus \mathcal{L}$ and **neck** as the pair $\{n_{i_{max}}, n_{j_{min}}\}$.

Generalized Secondary Structure Unit — GSSU

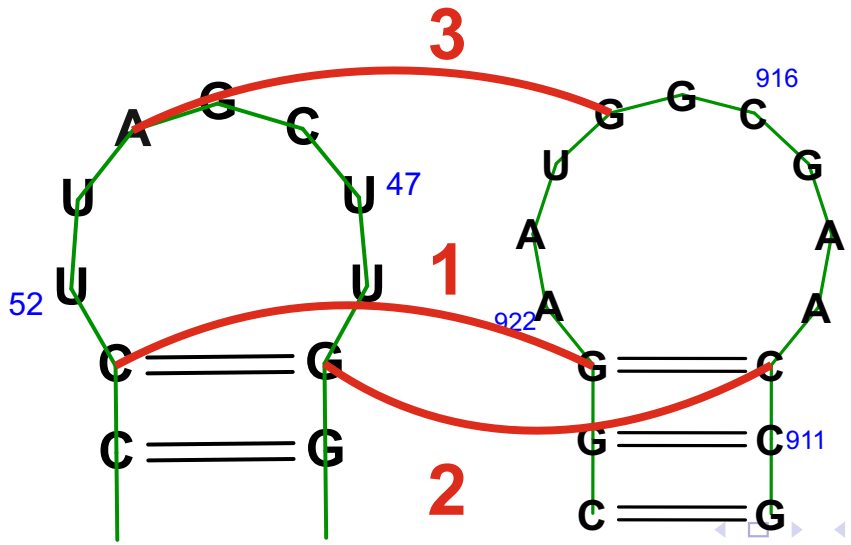


General Structure Similarity by Superposition

- given two sets of 3D coordinates / structures superpose them to optimize given distance/similarity measure
  - pairing
  - superposition (translation + rotation)
  - distance computation
    - root mean square deviation (RMSD)
    - TM-score
    - **Ad-hoc measure**

Single GSSU RNA Comparison

1. Identify candidate set of alignments of triplet pairs (two nucleotides from neck, one from loop).
2. Compute RMSD superpositions (i.e. set of rotation matrices and translations vectors) for each of the alignments.
3. For each rotation matrix and translation vector superpose the structures.
4. For each superposition identify nearest neighbors, sum the distances to get δ and normalize it to obtain the final distance Δ.



Single GSSU RNA Comparison

$$NN_{\zeta}(x, \mathcal{G}) = \begin{cases} \min_{1 \leq i \leq |\mathcal{G}|}\{d_{nt}(x, \mathcal{G}_i)\} \times \zeta & \text{if } x = \mathcal{G}_i \\ \min_{1 \leq i \leq |\mathcal{G}|}\{d_{nt}(x, \mathcal{G}_i)\} & \text{otherwise} \end{cases}$$

$$\gamma(\mathcal{G}^{\mathcal{A}}, \mathcal{G}^{\mathcal{B}}) = \sum_{i=1}^{|\mathcal{G}^{\mathcal{A}}|} \begin{cases} 1 & \text{if } NN_1(\mathcal{G}^{\mathcal{A}}{}_i, \mathcal{G}^{\mathcal{B}}) \leq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

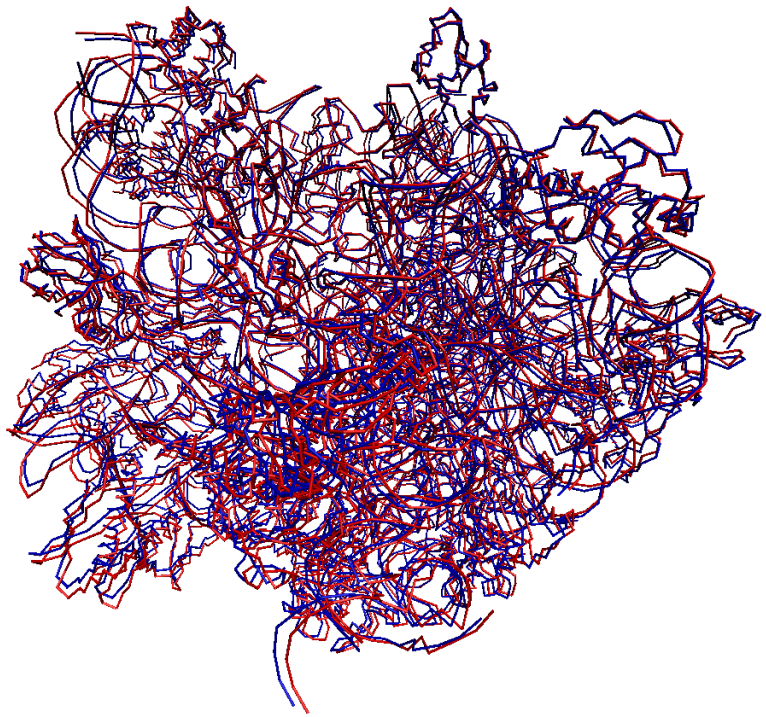
$$\delta(\mathcal{G}^{\mathcal{A}}, \mathcal{G}^{\mathcal{B}}) = \min_{t \in T} \left\{ \sum_{i=1}^{|\mathcal{G}^{\mathcal{A}}|} NN_{\alpha}(\mathcal{G}^{\mathcal{A}}{}_i, \tau(\mathcal{G}^{\mathcal{B}}, t)) \} \right\}$$

$$\Delta(\mathcal{R}^{\mathcal{A}}, \mathcal{R}^{\mathcal{B}}) = \Delta(\mathcal{G}^{\mathcal{A}}, \mathcal{G}^{\mathcal{B}}) = \frac{\frac{\delta(\mathcal{G}^{\mathcal{A}}, \mathcal{G}^{\mathcal{B}})}{\min\{|\mathcal{G}^{\mathcal{A}}|, |\mathcal{G}^{\mathcal{B}}|\}} \times \left(1 + \frac{||\mathcal{G}^{\mathcal{A}}| - |\mathcal{G}^{\mathcal{A}}||}{\min\{|\mathcal{G}^{\mathcal{A}}|, |\mathcal{G}^{\mathcal{B}}|\}}\right)}{\gamma(\mathcal{G}^{\mathcal{A}}, \tau(\mathcal{G}^{\mathcal{B}}, t_{opt}))} \quad (1)$$

Multiple GSSU Matching

$$\Delta(\mathcal{R}^{\mathcal{A}}, \mathcal{R}^{\mathcal{B}}) = \min_{\substack{1 \leq i \leq n_A \\ 1 \leq j \leq n_B}} \left\{ \mathcal{G}_i^{\mathcal{A}}, \mathcal{G}_j^{\mathcal{B}} \right\} \times (1 + |n_{\mathcal{A}} - n_{\mathcal{B}}|) \times \beta \qquad (2)$$

In the experimental dataset, 44% of structures contain more than one GSSU.



Speed Optimization

- Motivation
  - NN search has $O(n^2)$ time complexity with respect to GSSU length
  - NN search has to be done for each candidate triplet alignment
- Speed-up
  - decrease number of NN searches
  - skip NN search for triplets whose Δ is higher then best distance found so far
  - $\mathcal{T}^{\mathcal{A}} \subset \mathcal{G}^{\mathcal{A}}$, $\mathcal{T}^{\mathcal{B}} \subset \mathcal{G}^{\mathcal{B}}$ and $\Delta(\mathcal{G}^{\mathcal{A}}, \mathcal{G}^{\mathcal{B}}) = \chi \implies$ if $\Delta(\mathcal{T}^{\mathcal{A}}, \mathcal{T}^{\mathcal{B}}) \times \lambda > \chi$ then NN search can be skipped

Outline

Experimental Results — Classification

Datasets

- FSCOR — 420 structures from SCOR with unique classification
- R-FSCOR — structurally dissimilar subset of the FSCOR
- T-FSCOR — FSCOR \ R-FSCOR

Testing

- leave-one-out on FSCOR
- T-FSCOR vs R-FSCOR
- exact classification
- approximate classification

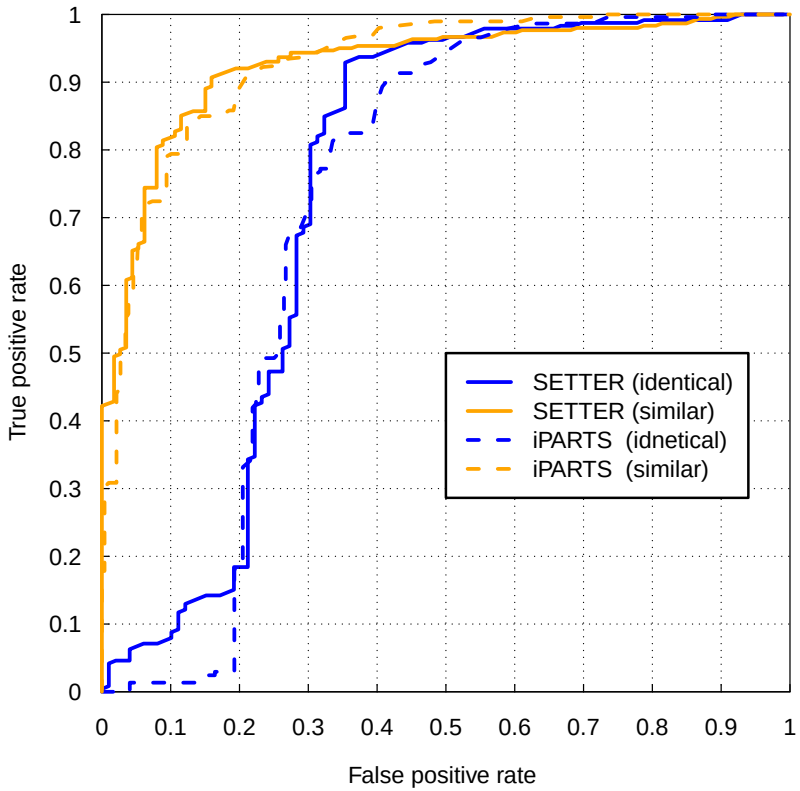
Qualitative measures

- ROC curves
- AUC (area under the curve)

Experimental Results — Classification

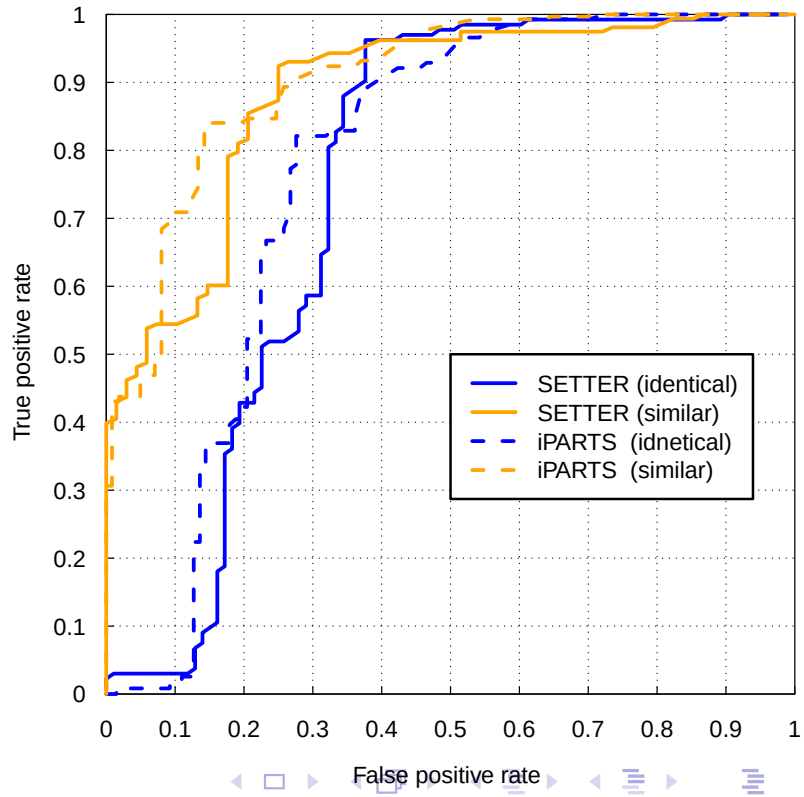
FSCOR

	SETTER	iPARTS	SARA
AUC(i)	**0.74**	0.72	0.61
AUC(s)	**0.93**	0.92	0.83



T-FSCOR vs R-FSCOR

	SETTER	iPARTS	SARA
AUC(i)	0.70	**0.77**	0.58
AUC(s)	0.88	**0.90**	0.85



Experimental Results — Runtime

Runtime comparison of iPARTS, SARA and SETTER. The *tRNA* set contains structures 1EHZ:A, 1H3E:B, 1I9V:A, 2TRA:A and 1YFG:A structures (average length 76 nucleotides), *Ribozyme P4-P6 domain* contains 1GID:A, 1HR2:A and 1L8V:A (average length 157 nucleotides), *Domain V of 23S rRNA* contains 1FFZ:A and 1FG0:A (average length 496 nucleotides) and *16S rRNA* contains 1J5E:A and 2AVY:A (average length 1522 nucleotides).

data set	iPARTS	SARA	SETTER
tRNA	1.1 s	1.7 s	**0.1** s
Ribozyme P4-P6 domain	2.6 s	9.2 s	**1.8** s
Domain V of 23S rRNA	17.0 s	?	**2.1** s
16S rRNA	2.8 min	?	**8.1** s

Conclusion and Future Work

- Conclusion
  - Proposal of a new RNA structure similarity method
  - GSSU introduction
  - Single GSSU and multiple GSSU structure comparison
  - Speedup by early termination
  - Promising experimental results
- Future Work
  - Introduction of significance computation into the classification process
  - More sophisticated multiple GSSU comparison
  - Web server

- Thank you. Questions?