Extracting Sense from Structure: an application to functional non-coding RNA polymers

F. Costa

Bioinformatics Group
Department of Computer Science
Albert-Ludwigs-University Freiburg, Germany

13-17 February 2012

What is this talk about?

Visualization of folding hypothesis landscape for a ncRNA family.
Semi-automatic construction of a vocabulary for structures.

- Allows finer grain view than clustering
- Useful to get an idea on the plasticity of a ncRNA family

Map of the talk

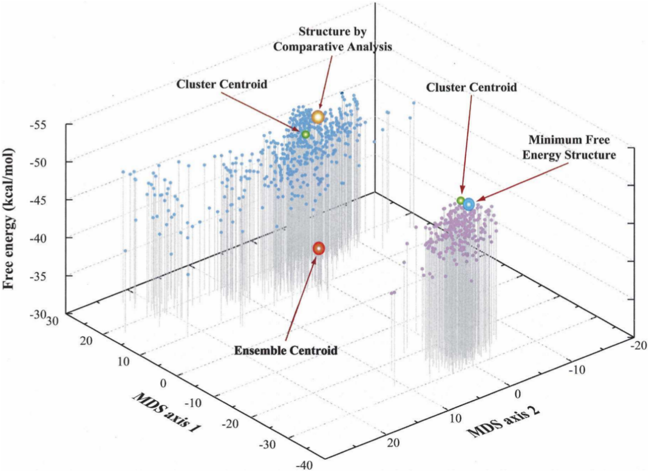
1. **Introduction:** Clustering induces an implicit prototype (cluster center) with which to measure the typicality of its members. Consensus structures allow to visualize the average agreement on different parts.
2. **Question:** Can we decompose the character of a RNA family into meaningful traits?
3. **Answer:** Represent the principal directions of change identify the parts characteristic for different directions.

- The identification of a single folding structure to characterize a functional ncRNA family is a difficult and ill-posed problem
- **Idea:** characterize the entire set of probable structures

Issues

How to best represent:

- multiple sub-groups
- continuous structural variation



MDS for all folding configurations for a single sequence[1].

[1] Source: Ding, Y. & Lawrence, C. E., *A statistical sampling algorithm for RNA secondary structure prediction*, Nucl. Acids Res. 2003

Issues with RNA structural representations

- The MFE yields a single folding hypothesis that can be (at times) non representative
- Partition function based dot plots represents only statistics on all folding structures
- Accessibility information marginalizes base-pairedness in an aggregate with loss of structural information
- Suboptimal sampling is expensive and requires an additional (heuristic) clustering step

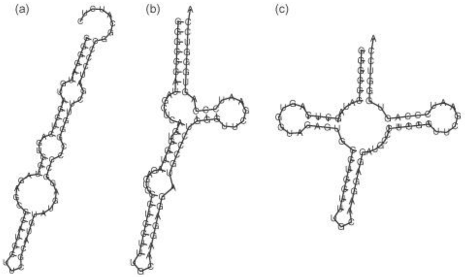
Proposal

- Use shape approach to derive a set of representative folding structures
- Represent each structure fully (i.e. as a labeled graph)
- Process set with graph kernels or explicit subgraph fingerprint techniques

Sampling representative structures

- Sample all folding hypothesis
- ...which exhibit significantly different structure
- ...and are in a small energy range above the minimum free energy ↦ representative structures: *shapes*[a]

[a] Giegerich, B, Voß and M. Rehmsmeier, *Abstract shapes of RNA*, NAR 2004



(d)

```
Shape     GGGCCCAUAGCUCAGUGGUAGAGUGCCUCCUUUGCAAGGAGGAUGCCCUGGGUUCGAAUCCCAGUGGGUCCA
[]        (((((((((((((((.((((.....(((((((...))))))).))))))))))).........)))))))).  -35.9 kcal/mol
[[] []]   ((((((((.....((.((((.....(((((((...))))))).))))))(((.......))).)))))))).  -32.2 kcal/mol
[[] [] []] ((((((...((((.......)))).(((((((...))))))).....(((((.......))))).)))))).  -31.7 kcal/mol
```

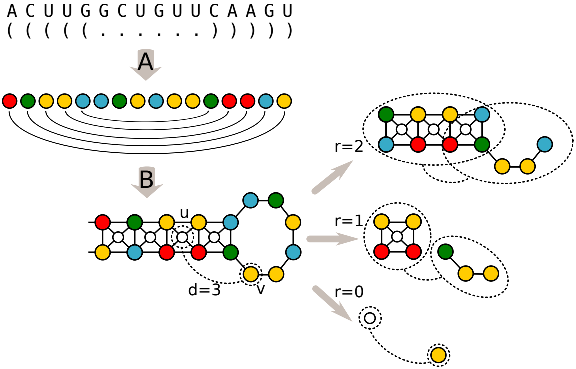
Representing RNA structure as graphs

Neighborhood Subgraph Pairwise Distance Kernel (NSPDK)[a]

Features as all pairs of near small neighborhood subgraphs

≈ a generalization of *k-mers* with gaps

[a] F. Costa, K. De Grave, *Fast Neighborhood Subgraph Pairwise Distance Kernel*. ICML 2010



From Linear Model to Importance Signal

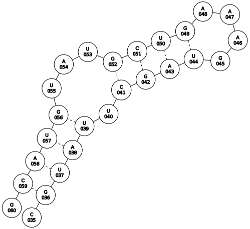
Given a binary classification task, induce linear models:

- **performance:** good generalization guarantees
- **fast and scalable:** linear in practice; can manage $> 10^5$ instances
- **interpretable:** model ↦ set of feature-weight pairs

Interpret the weight as importance score for each feature

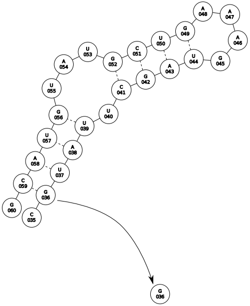
From Importance Signal to Important Parts

1. Compute the importance for each vertex v_i = cumulative importance of all subgraphs that involve v_i
2. Visualize regions with high vertex importance...



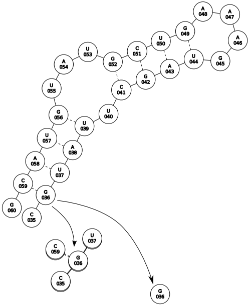
decomposition
of graph
in feaures

FIGURE: Cumulative vertex importance



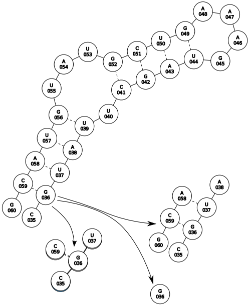
decomposition
of graph
in feaures

Figure: Cumulative vertex importance



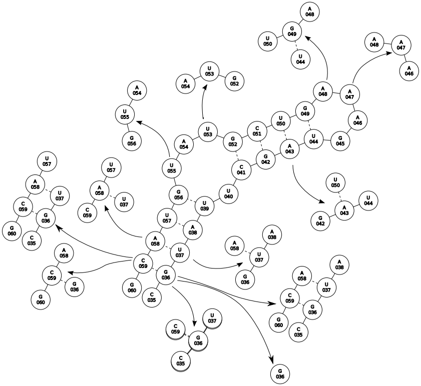
decomposition
of graph
in feaures

Figure: Cumulative vertex importance



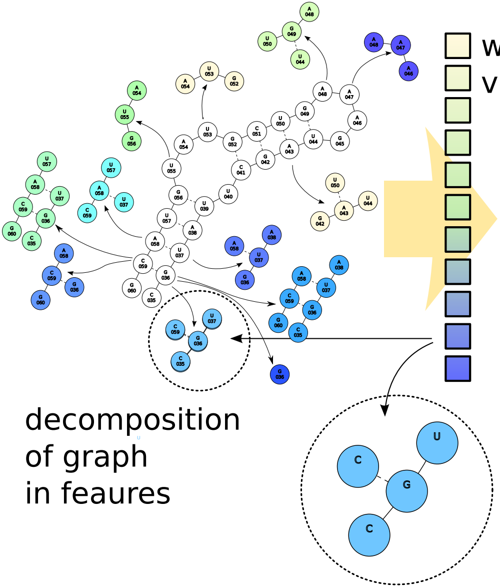
decomposition
of graph
in feaures

FIGURE: Cumulative vertex importance



decomposition
of graph
in feaures

Figure: Cumulative vertex importance

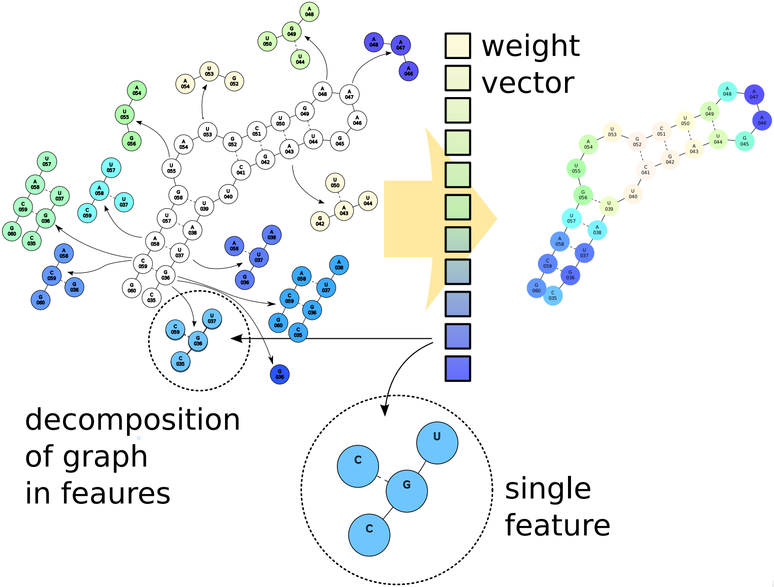


weight vector

decomposition of graph in feaures

single feature

Figure: Cumulative vertex importance



weight vector

decomposition of graph in feaures

single feature

FIGURE: Cumulative vertex importance

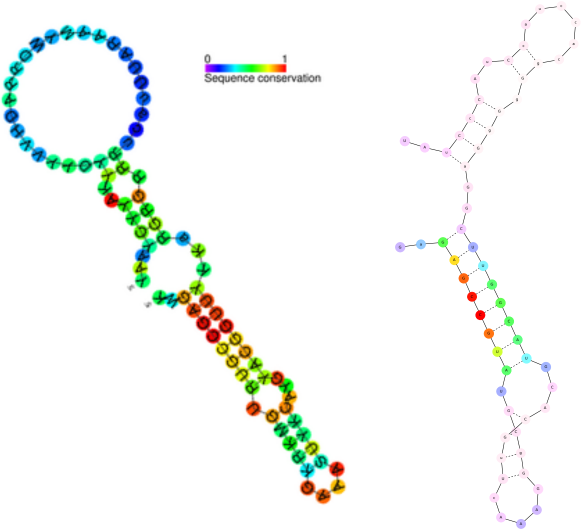


Figure: Consensus RF00029 Intron gpII (Ribozyme)

Applications of importance signal (ongoing work)

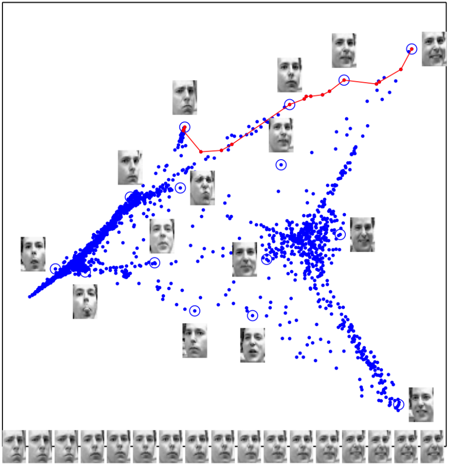
1. Important parts (=connected components with importance > threshold) are <u>structural motifs</u> that can be clustered for characterization and insights
2. Importance score can complement energetic score in <u>folding</u> algorithms
3. Important parts can be constrained to match in <u>alignment</u> procedures even when dissimilar at sequence level

Using NSPDK we can represent graphs in a very high dimensional vector space.

But how to map graphs onto a plane for visual inspection?

Dimensionality Reduction Techniques

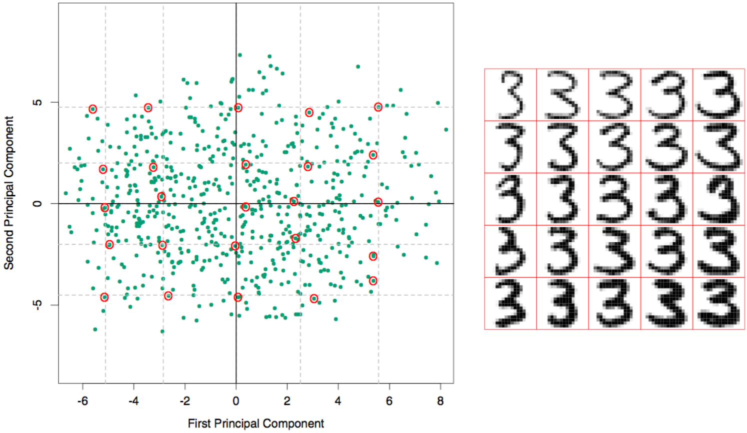
1. Multi Dimensional Scaling (MDS)
   Determin 2D coordinates so to maximally preserve the pairwise distances that instances originally had.
   - - Non/trivial identification of directions of change
   - - Non-convex optimization problem → locally optima
   - + Non-linear embedding
2. Singular Value Decomposition (SVD)
   or equivalently Principal Component Analysis (PCA)
   Rank ortogonal directions that induce the best reconstruction of the original vectors.
   - + Trivial identification of directions of change
   - + Convex optimization problem → global optimal solution
   - - Linear embedding



MDS for face image set[2].

Along the red line the expression moves from sad to happy.

[2] Source: T.Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. 09



SVD for digit image set[3].

X ↦ length of lower trait; Y ↦ thickness.

[3] Source: T.Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. 09

Proposal in a nutshell

Given a set of RNA sequences belonging to a functional class:

1. materialize *m* folding hypothesis for each sequence

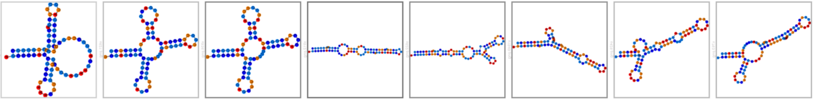
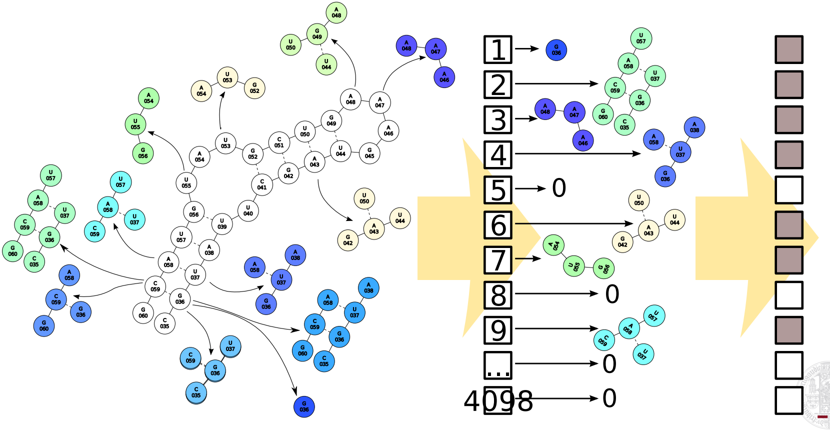


Figure: Color-code: G-C=blu-cyan A-U=red-orange

Proposal in a nutshell

Given a set of RNA sequences belonging to a functional class:

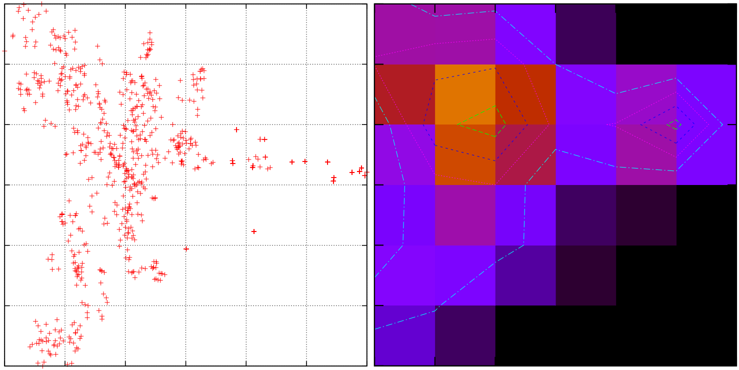
1. materialize *m* folding hypothesis for each sequence
2. structural graph ↦ vector representation



Proposal in a nutshell

Given a set of RNA sequences belonging to a functional class:

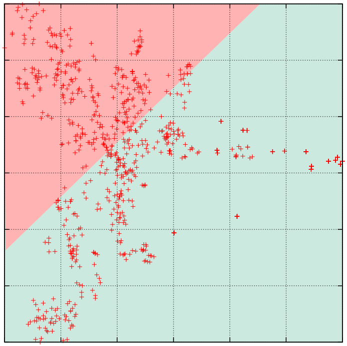
1. materialize *m* folding hypothesis for each sequence
2. structural graph ↦ vector representation
3. SVD ↦ compute 2 main components and embed



Proposal in a nutshell

Given a set of RNA sequences belonging to a functional class:

1. materialize *m* folding hypothesis for each sequence
2. structural graph ↦ vector representation
3. SVD ↦ compute 2 main components and embed
4. induce discriminative model on binary classification task: *instances in half space vs. instances in the other half*



Proposal in a nutshell

Given a set of RNA sequences belonging to a functional class:

1. materialize *m* folding hypothesis for each sequence
2. structural graph ↦ vector representation
3. SVD ↦ compute 2 main components and embed
4. induce discriminative model on binary classification task: *instances in half space vs. instances in the other half*
5. partition instances into $k \times k$ tiles in 2D plane

1. **Plot 1:** plot only one representative shape per tile (choose highest frequency shape)
2. **Plot 2:** plot importance signal on each vertex
3. **Plot 3:** plot consensus structures

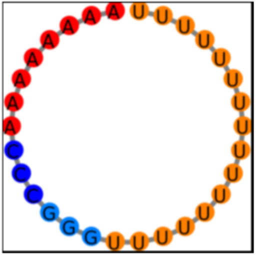
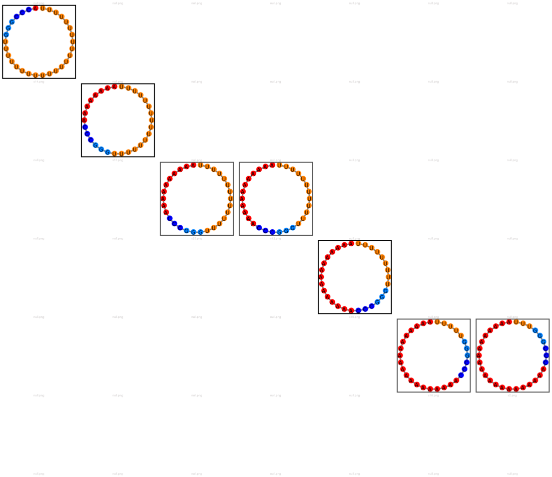
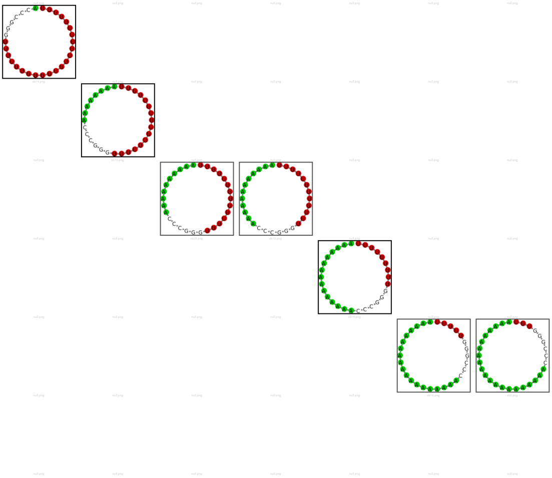


FIGURE: Artificial example: sequences with pattern $[U]^m GGGCCC[A]^n$





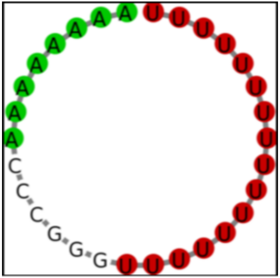


FIGURE: Artificial example: the part in common to all sequences that cannot be used to discriminate is white.

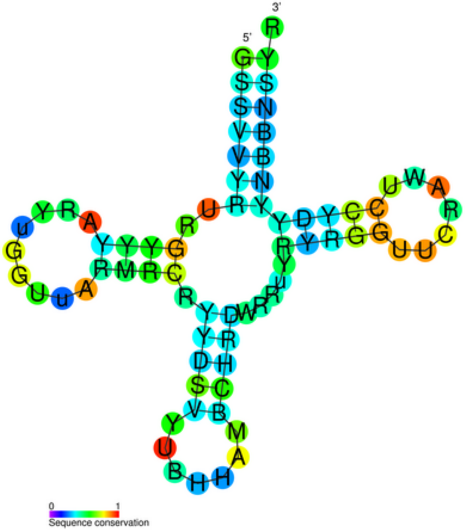
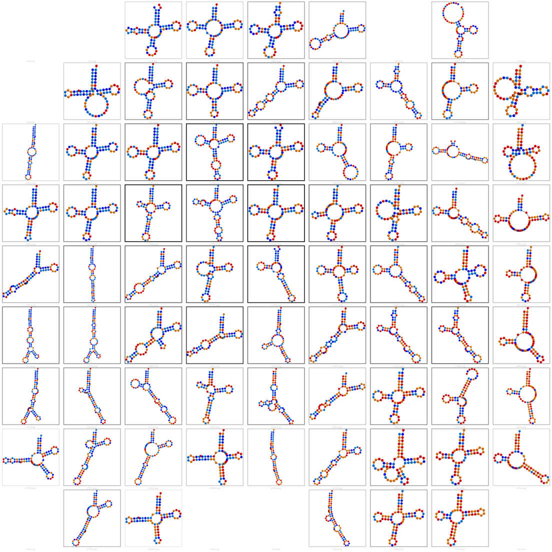
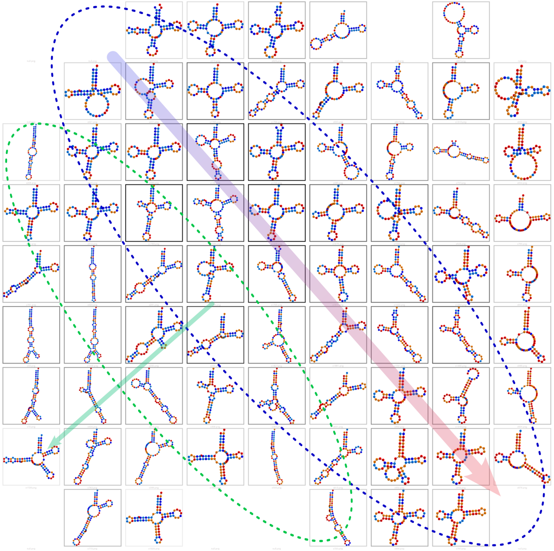
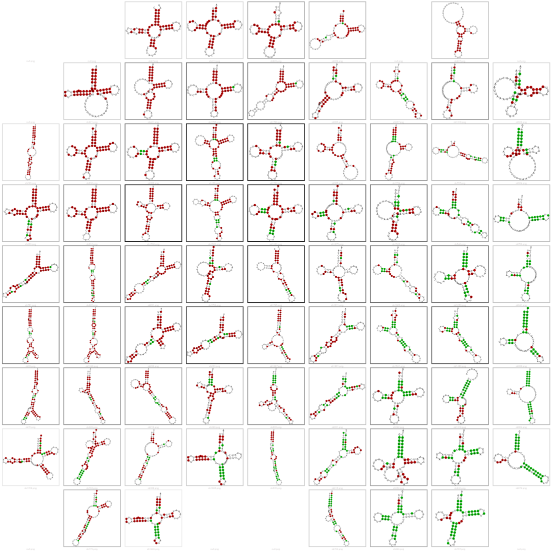
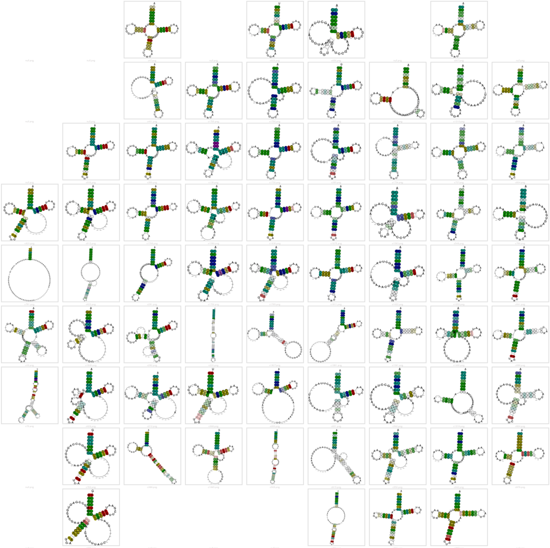


FIGURE: RF00005: tRNA









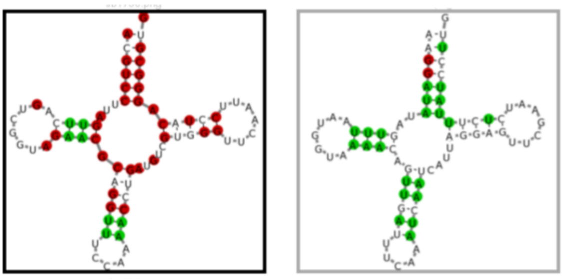


FIGURE: The poles represent GC vs AU content. Loop parts are white ↦ more interesting.

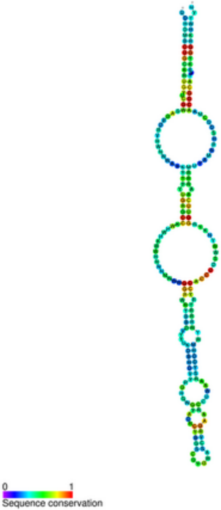
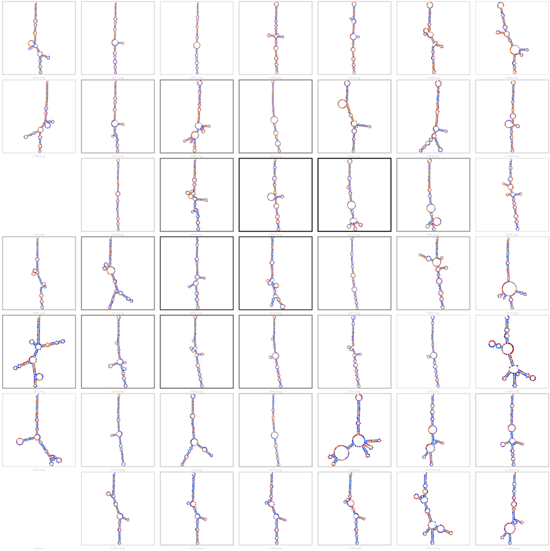
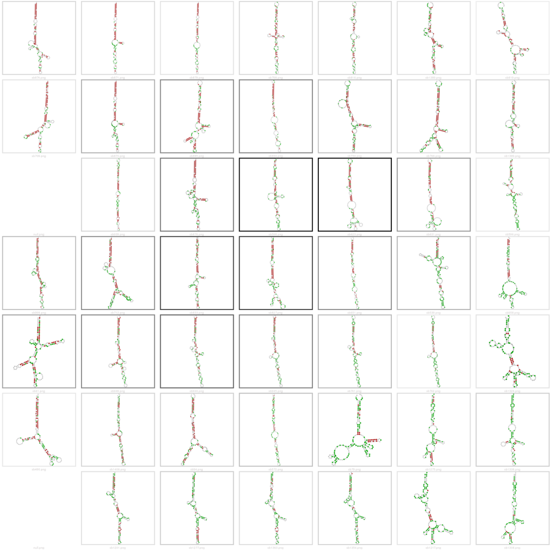
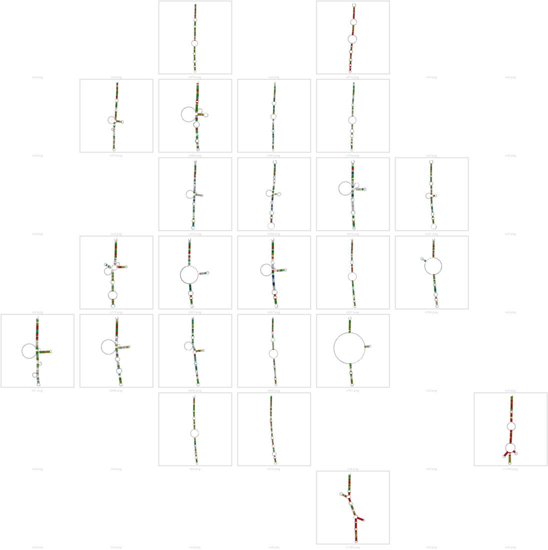


Figure: RF00013: 6S




sb874.png
sb871.png
sb879.png
sb798.png
sb816.png
sb1389.png
sb819.png
sb709.png
sb870.png
sb885.png
sb664.png
sb734.png
sb766.png
sb1399.png
null.png
sb939.png
sb876.png
sb598.png
sb623.png
sb631.png
sb596.png
sb698.png
sb712.png
sb833.png
sb827.png
sb651.png
sb539.png
sb858.png
sb61.png
sb949.png
sb946.png
sb945.png
sb741.png
sb742.png
sb1326.png
sb490.png
sb1406.png
sb94.png
sb572.png
sb79.png
sb1275.png
sb1338.png
null.png
sb1201.png
sb1277.png
sb1363.png
sb1354.png
sb1217.png
sb1308.png


null.png
null.png
c879.png
null.png
c816.png
null.png
null.png
null.png
c870.png
c885.png
c664.png
c734.png
null.png
null.png
null.png
null.png
c876.png
c598.png
c623.png
c631.png
null.png
null.png
c712.png
c833.png
c827.png
c651.png
c539.png
null.png
c61.png
c949.png
c948.png
c945.png
c741.png
null.png
null.png
null.png
null.png
c94.png
c572.png
null.png
null.png
c1338.png
null.png
null.png
null.png
null.png
c1354.png
null.png
null.png

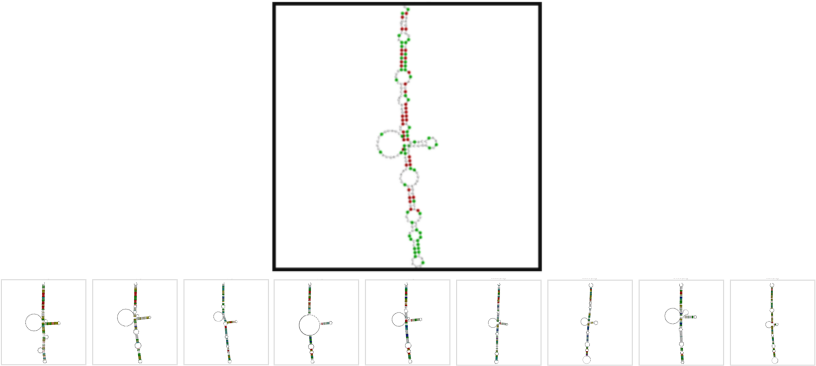


FIGURE: The presence of a (until recently unknown) functional hairpin is white ↦ important, and present in many consensus alignments.

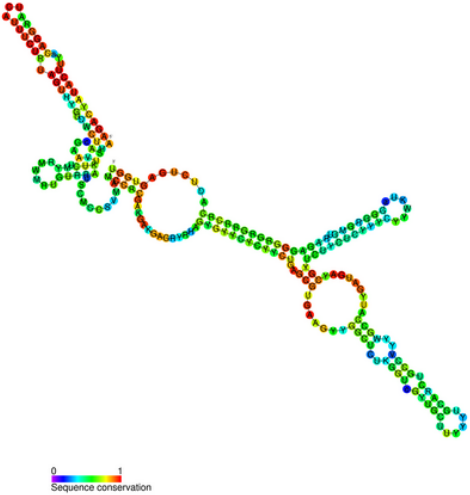
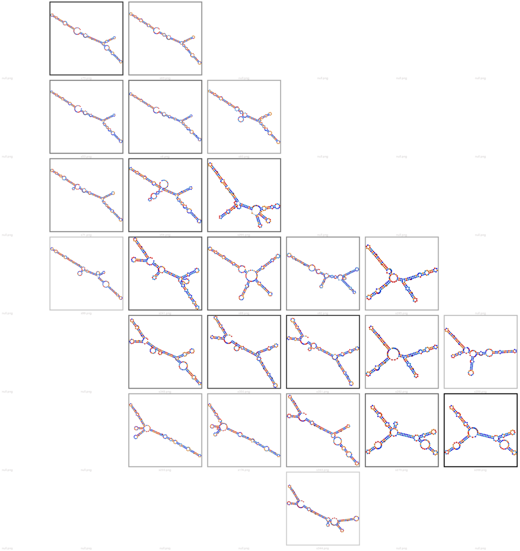
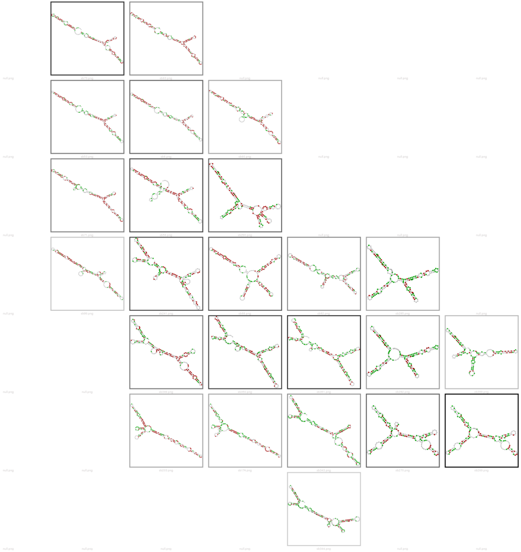
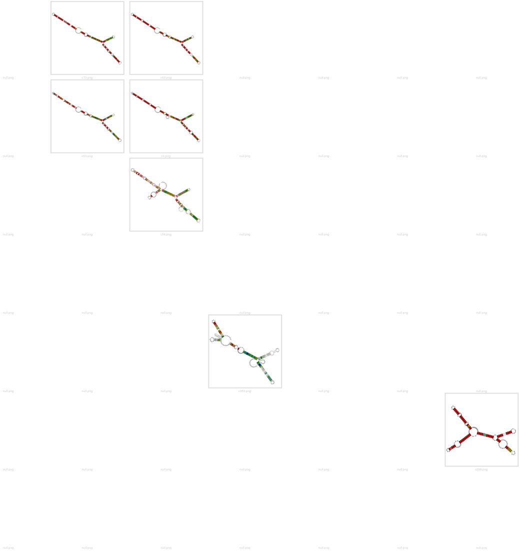


Figure: RF00012: U3


null.png
s73.png
s63.png
null.png
null.png
null.png
null.png
null.png
s53.png
s6.png
s60.png
null.png
null.png
null.png
null.png
s71.png
s56.png
s290.png
null.png
null.png
null.png
null.png
s99.png
s241.png
s88.png
s82.png
s285.png
null.png
null.png
null.png
s349.png
s350.png
s351.png
s282.png
s299.png
null.png
null.png
s203.png
s174.png
s343.png
s270.png
s269.png
null.png
null.png
null.png
null.png
s344.png
null.png
null.png


null.png
sb73.png
sb63.png
null.png
null.png
null.png
null.png
null.png
sb53.png
sb6.png
sb60.png
null.png
null.png
null.png
null.png
sb71.png
sb56.png
sb290.png
null.png
null.png
null.png
null.png
sb99.png
sb241.png
sb88.png
sb82.png
sb285.png
null.png
null.png
null.png
sb349.png
sb350.png
sb351.png
sb282.png
sb299.png
null.png
null.png
sb203.png
sb174.png
sb343.png
sb270.png
sb269.png
null.png
null.png
null.png
null.png
sb344.png
null.png
null.png


null.png
c73.png
c63.png
null.png
null.png
null.png
null.png
null.png
c53.png
c6.png
null.png
null.png
null.png
null.png
null.png
null.png
c56.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
c350.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png
c269.png
null.png
null.png
null.png
null.png
null.png
null.png
null.png

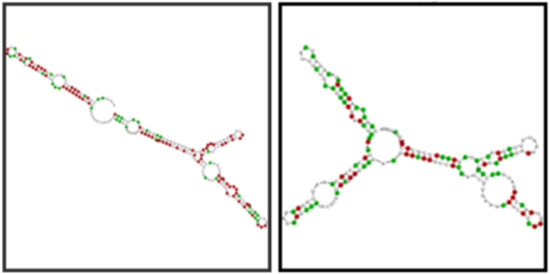


FIGURE: The opposite poles represent the variant vertebrate vs. non-vertebrate, characterized by one stem vs. two stems.

Conclusions

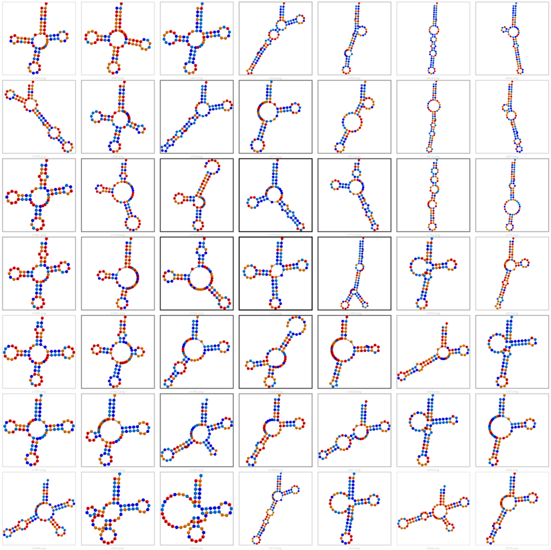
- We propose to:
  - visualize a set of folding structures
  - embed them in a plane...
  - ...whose coordinate system is aligned to the directions of major sequence-structural changes
- automatically learn how to discriminate between extreme cases
- ... to identify common regions

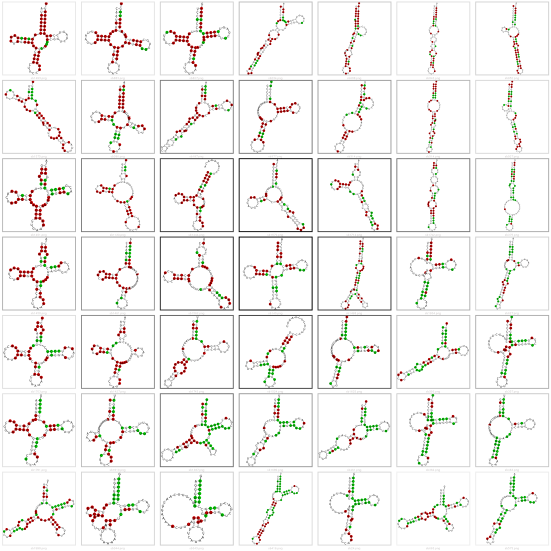
How can all this be used?

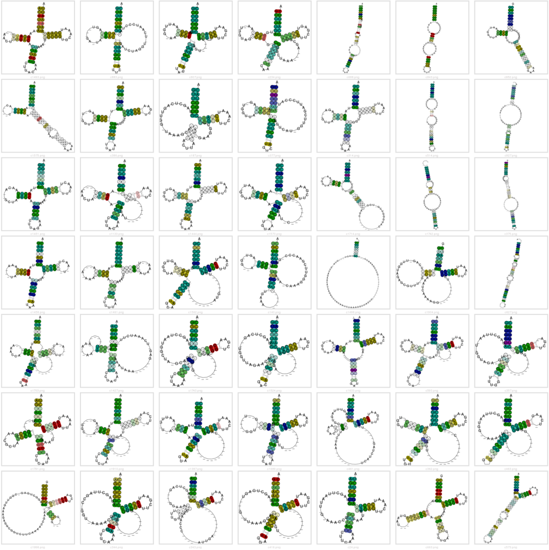
- Give biologists a new way to look into a ncRNA family
- Help them identify in a semi-automatic way interesting parts or sub-families
- Help them characterize and give a name to the structural *(and possibly functional)* traits
- Give a way to use biological knowledge to select a subset of meaningful sequences-structures to make better models

Thanks

Acknowledgments:
Steffen Heyne
Sita Lange
Robert Kleinkauf
Manja Marz
Rolf Backofen







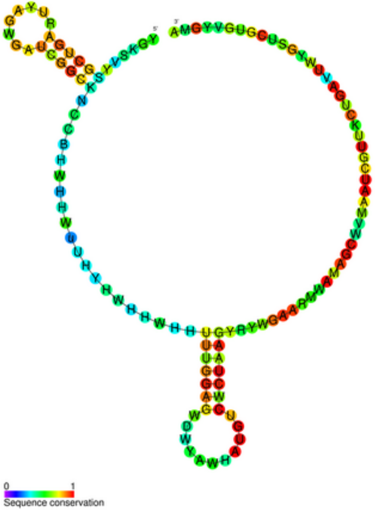
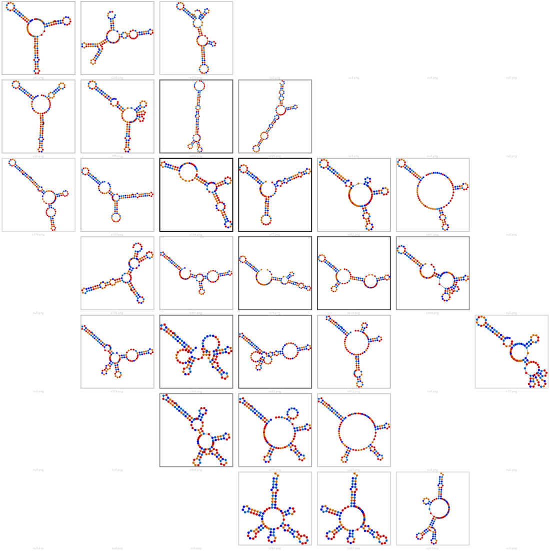
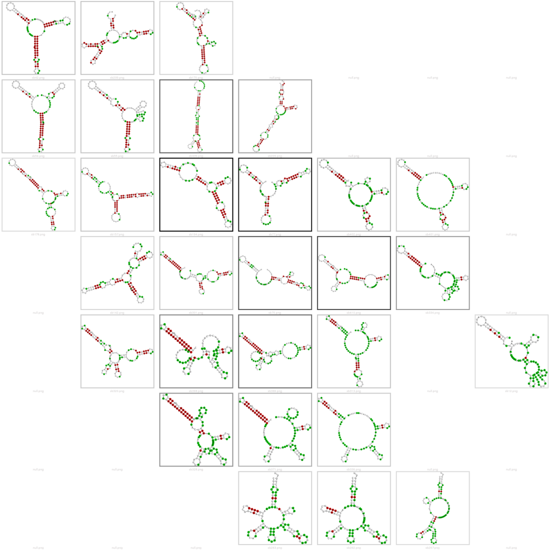


FIGURE: RF00114: Ribosomal S15 Leader




sb42.png
sb209.png
sb170.png
null.png
null.png
null.png
null.png
sb56.png
sb55.png
sb409.png
sb226.png
null.png
null.png
null.png
sb178.png
sb157.png
sb194.png
sb77.png
sb402.png
sb401.png
null.png
null.png
sb142.png
sb301.png
sb70.png
sb410.png
sb334.png
null.png
null.png
sb320.png
sb368.png
sb369.png
sb313.png
null.png
sb12.png
null.png
null.png
sb328.png
sb371.png
sb336.png
null.png
null.png
null.png
null.png
null.png
sb263.png
sb262.png
sb267.png
null.png

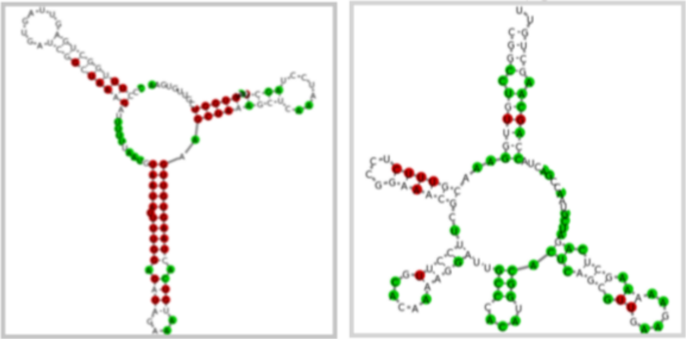


FIGURE: While one hairpin is common to the whole family (top-left white), the second hairpin (bottom green) seems to represent only one of the extreme cases.

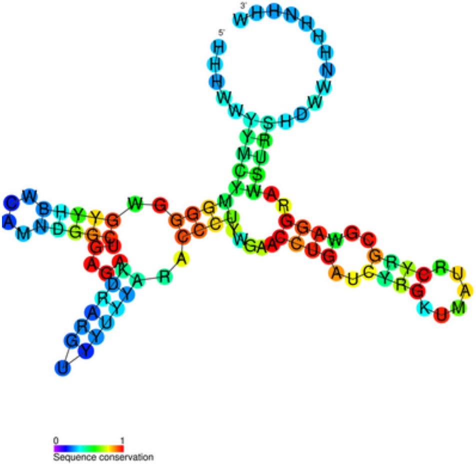
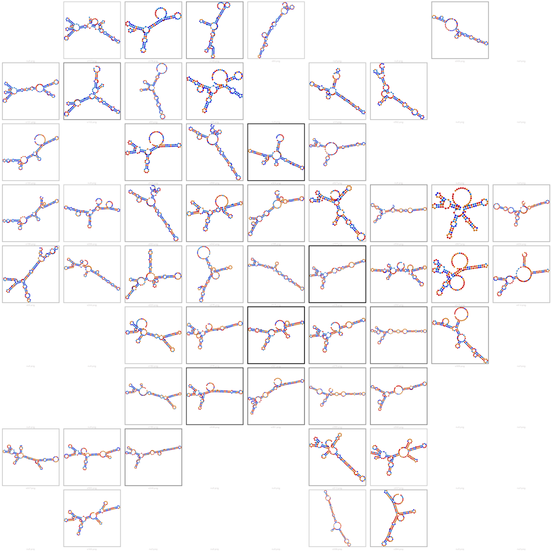
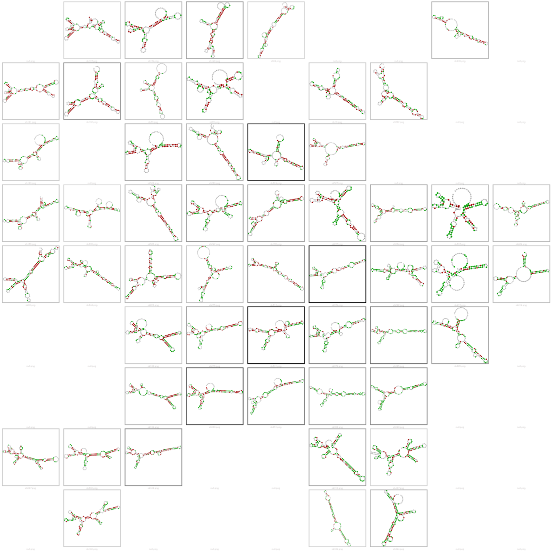


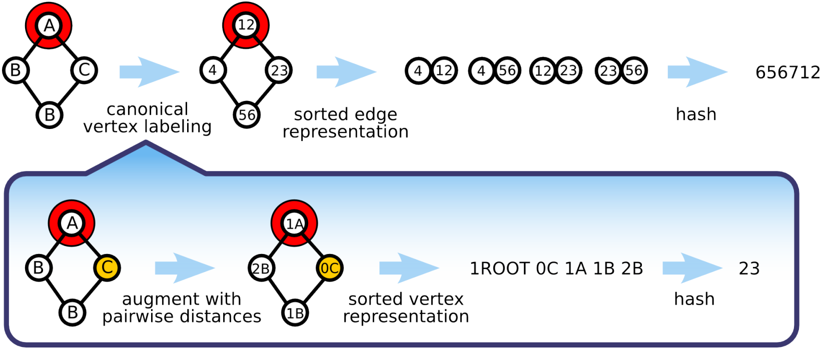
Figure: RF00059: TPP Ryboswitch





Gaining scalability via working in the primal: Computing the explicit mapping ϕ

Given graph as a (multi)set of pairs of near small subgraphs compute the explicit sparse representation via hashing techniques



A
B
C
B
canonical vertex labeling
12
4
23
56
sorted edge representation
4 12 4 56 12 23 23 56
hash
656712
A
B
C
B
augment with pairwise distances
1A
2B
0C
1B
sorted vertex representation
1ROOT 0C 1A 1B 2B
hash
23

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