Beyond Base Pair Probabilities and Single Nucleotide Accessibility

The Way of Graph Kernels

F. Costa

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

26th TBI Winterseminar in Bled
13-20 February 2011

How to process multiple RNA structure information?

- Commonly, all folding structures are evaluated but ...
- in the end individual structure information is lost/marginalized in an aggregate
- Only probability for single nucleotide of being paired/unpaired or probability for base pairing is retained

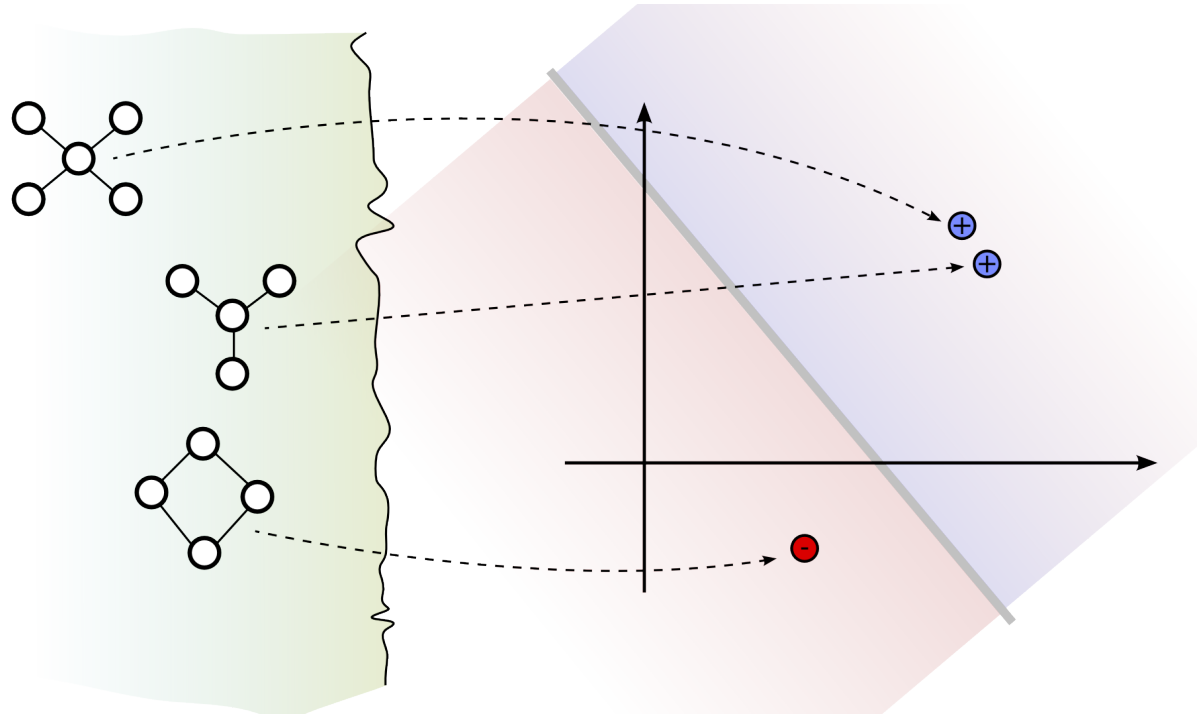
Proposal

- Increase information extracted from each structure
- **How:** consider occurrences of large(r) subgraphs
  Use graph kernels or explicit subgraph fingerprint techniques

Graph Kernels

How do we build robust/stable algorithms for learning relations in graph domains?

1. Embed graphs in vector space
2. Define dot product (aka the kernel function) and hence angles, lengths, distances
3. Use robust geometric algorithm for linear relations



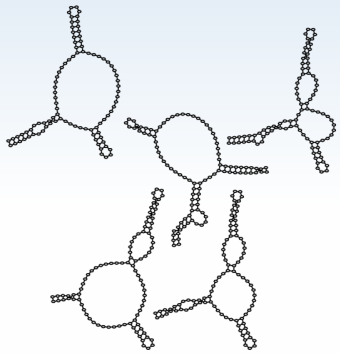
1

ACCCGUACUGGAACCACCCGUACUGGAACCACCCGUACUGGAACC

ACCCGUACUG ACCCGUACUG ACCCGUACUG
UACUGGAACC UACUGGAACC UACUGGAACC
GAACCACCCGU GAACCACCCG

2

GAACCACCCGU 



3

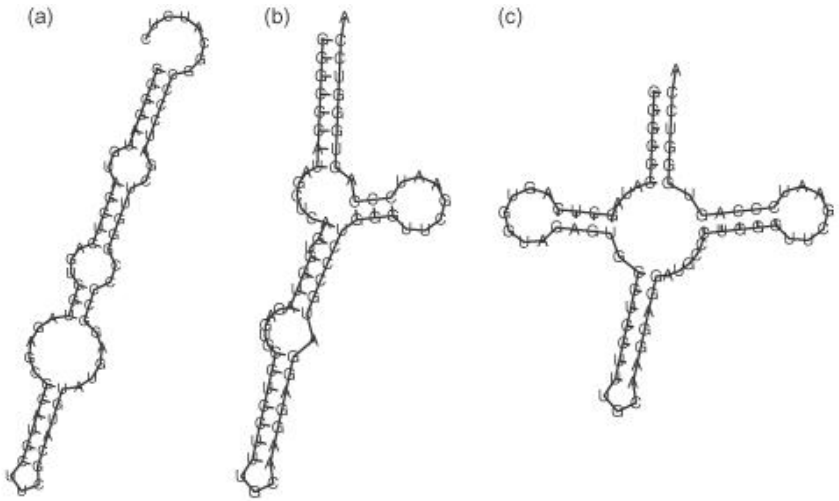
Method Overview

1. RNA sequence $\mapsto$ set of overlapping subsequences
2. subsequence $\mapsto$ set of representative structures
3. Structure $\mapsto$ set of features $\mapsto$ vector encoding

Sampling a representative set of structures

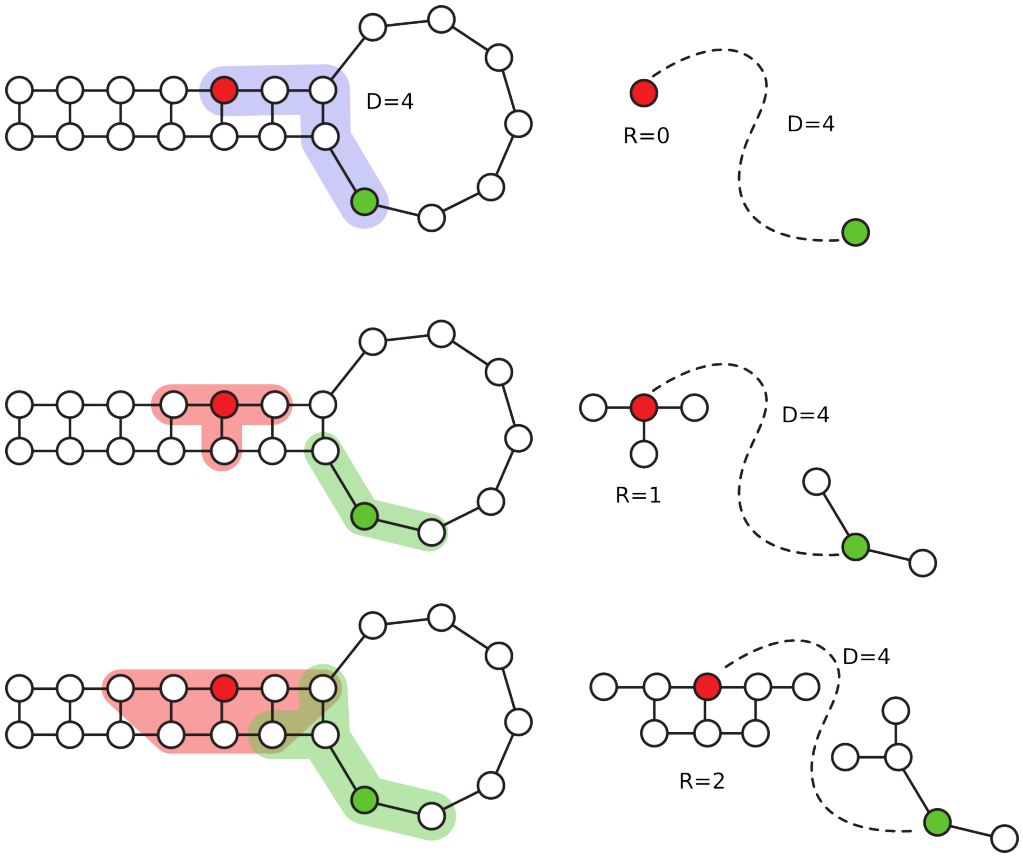
- We consider subset of representatives structures for significantly different shapes (shreps) in a small energy range above the minimum free energy

Robert Giegerich, Björn Voß and Marc Rehmsmeier, "Abstract shapes of RNA", Nucl. Acids Res. (2004) 32 (16): 4843-4851



Given a structure which features to extract? 1/3

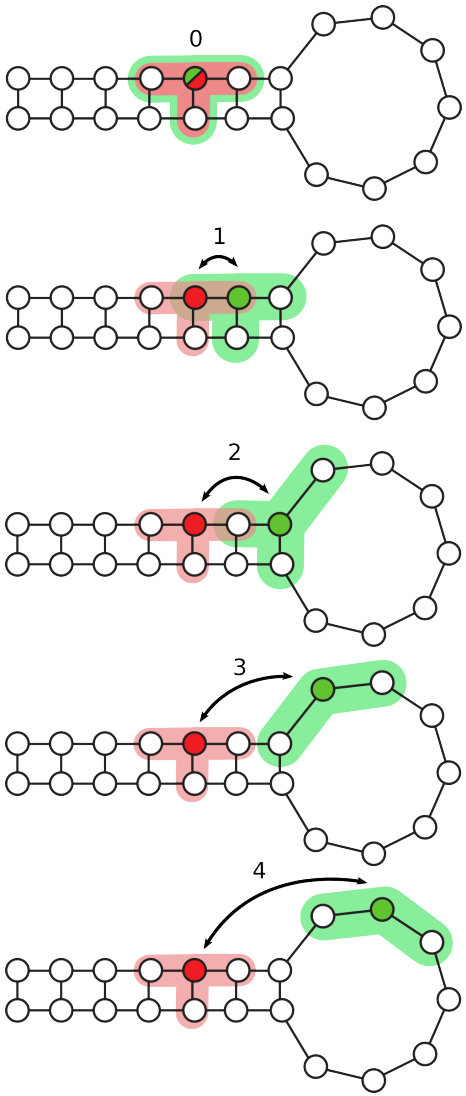
Features: all pairs of near small subgraphs



Given a radius R and a distance D, for each vertex v consider all pairs of neighborhood subgraphs rooted in v with radius ranging from 0 to R ...

Given a structure which features to extract? 2/3

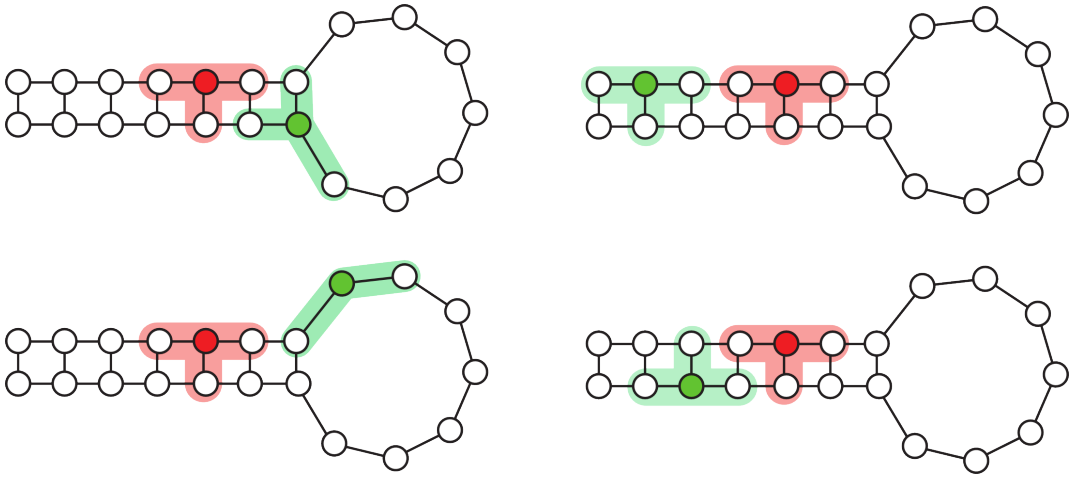
Features: all pairs of near small subgraphs



Given a radius R and a distance D, for each vertex v consider all pairs of neighborhood subgraphs rooted in v with radius ranging from 0 to R ... with distance between roots ranging from 0 to D

Given a structure which features to extract? 3/3

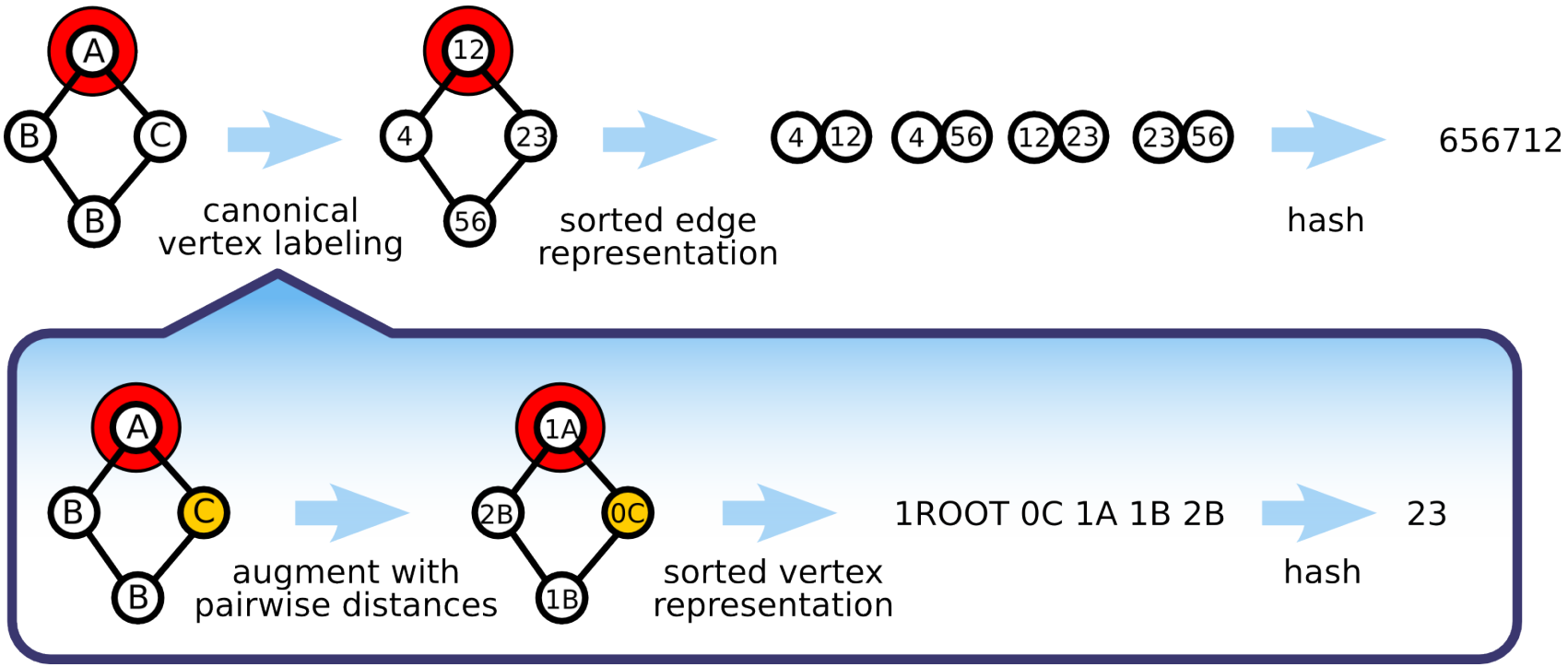
Features: all pairs of near small subgraphs



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

Explicitly mapping graphs into vector spaces

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



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

How to determine structures sample size?

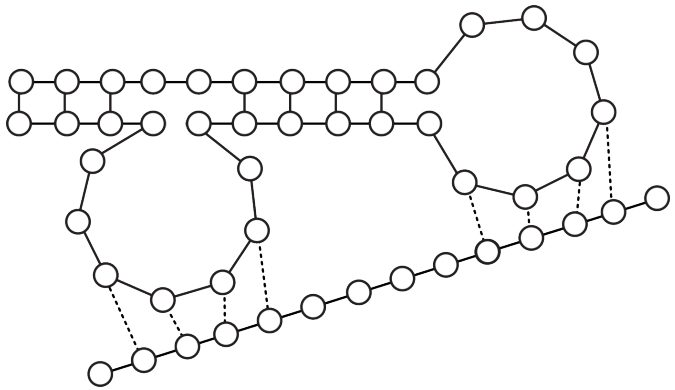
- Kernelized algorithms see data through the Gram (correlation) matrix
- **Criterion:** increase sample size until Gram matrix does not change significantly
  (i.e. until all pairwise similarities/distances do not change)

How to deal with different window sizes and different shreps?

- Represent each RNA as a set of graphs (one for each window size and shrep)
- Treat the set of graphs as a single graph with disconnected components
- ...local subgraphs (structural motifs) that appear often in many windows and many shreps have higher weight $\mapsto$ more important in similarity notion

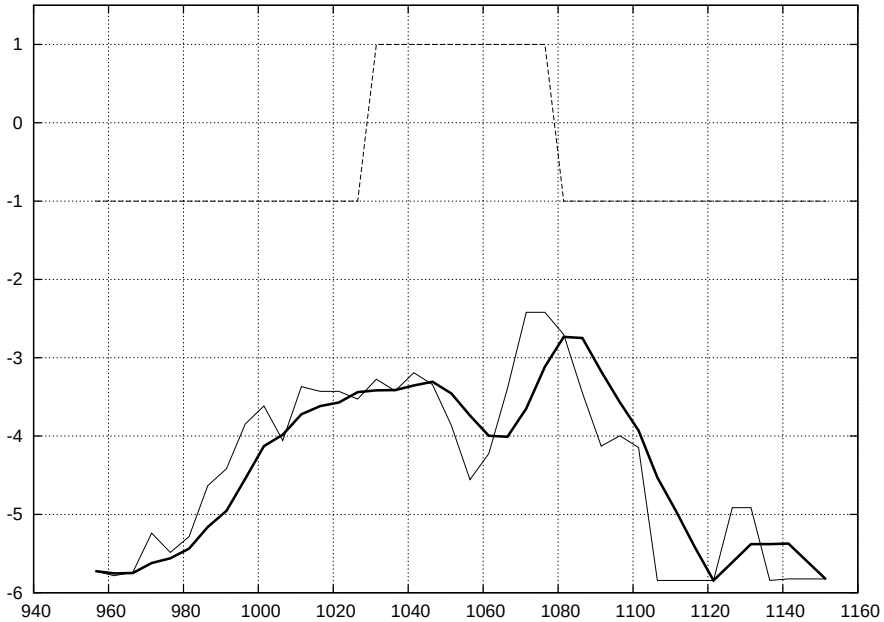
Advantages

- Flexible approach: encode domain knowledge via node/edge relabeling and/or node/edge insertion (ex. RNA-RNA interaction)
- Different types of tasks are natural:
  - classification $\mapsto$ signal localization
  - regression $\mapsto$ binding affinity prediction
  - similarity $\mapsto$ RNA family clustering
- Efficient: 10-100K RNA graphs per hour on desktop machine [graph encoding + train/test time]

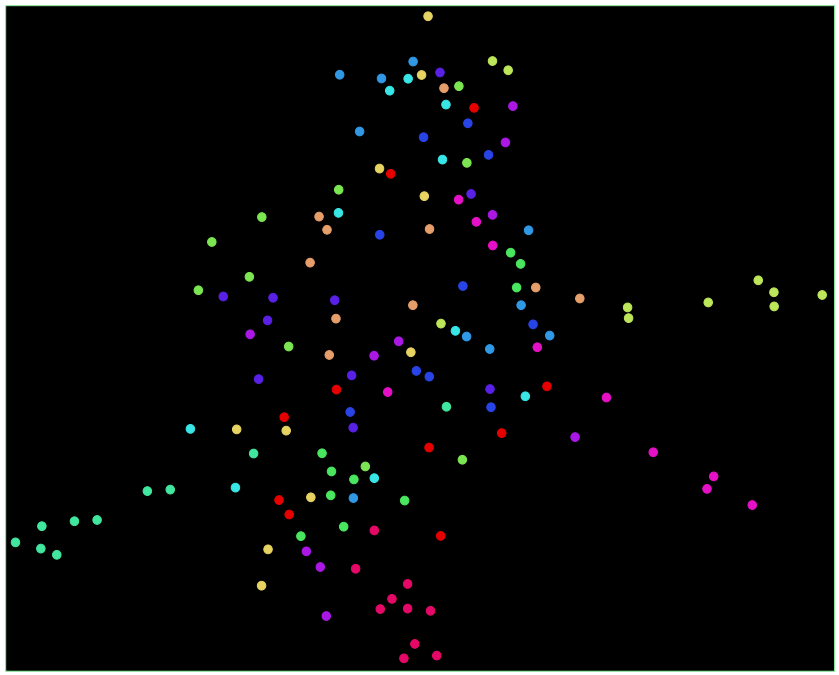


Modeling RNA-RNA interaction with edge insertion

Current Applications
(work in progress)



μRNA-mRNA interaction



RNA family clustering

Conclusions

It is possible to make use of more structure information in an efficient way

Future Work

Vector representation $\rightsquigarrow$ similarity notion $\rightsquigarrow$ density estimate $\rightsquigarrow$ robust p-value notion for structures