Making sense of non-coding RNA at genomic scale

The quest for efficient graph clustering

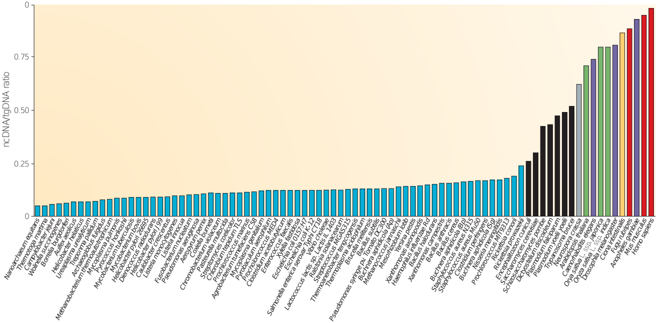
F. Costa

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

Spring meeting on Mining and Learning in Prüm
29-31 March 2011

Why is RNA important?

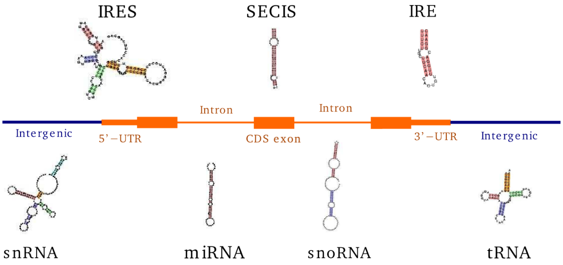
- While it is true that gene≡protein in prokaryotes
- ...in more complex organisms the quantity of non protein coding DNA ranges from 50% in plants to 98.5% in humans
- ncRNA has a regulatory function of paramount importance to allow organism complexity



Ratio of non-coding to coding DNA in increasingly complex organisms

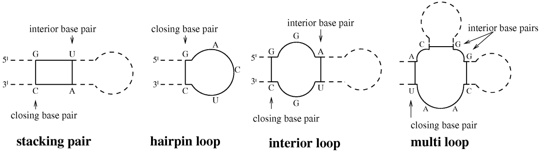
RNA function

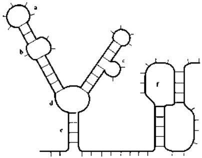
- RNA (single strand of 4 nucleotides: A,U,C,G) has many functions (translation, modification, catalytic, splicing, transport, silencing, regulatory, ...)
- Function is determined by sequence **and** structure
- Next generation sequencing technologies allow high-throughput data collection of sequence information
- ...but structure determination is (still) done algorithmically



How to compute RNA structure

- The minimum free energy structure of a given nucleotide sequence can be computed via dynamic programming in $O(n^3)$
- The best alignment of 2 RNA (considering simultaneously sequence and structure) in $O(n^4)$ $\rightsquigarrow$ good similarity notion





```
RNAfold < trna.fa
>AF041468
GGGGGUAUAGCUCAGUUGGUAGAGCGCUGCCUUUGCACGGCAGAUGUCAGGGGUUCGAGUCCCCUUACCUCCA
(((((((..((((........)))).(((((.......))))).....(((((.......)))))))))))). (-31.10)
```

What

- Given all ncRNA sequences in one *or several* organisms
- ...group together ncRNA either by function or structure
- ⇒ graph clustering
- **Goal:** discover novel groups/functions/structures ↦ families

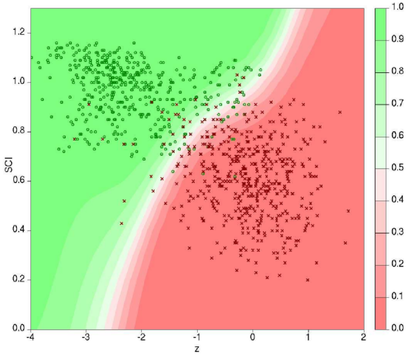
Issues

- Given a known family of ncRNA one can efficiently scan entire genomes to identify members
  ...but how to approach novel family discovery is open question
- Pairwise alignment is state-of-the-art technique to induce reliable similarity notion for RNA
  ...but it is very expensive (feasible only up to 2-3K sequences)

The proposal in a nutshell

Given one or more genomes ($n\times$ 1G nt):

1. Extract candidate ncRNA fragments (10K seq of 100 nt)
2. Cluster fragments according to sequence **and** structure
   Contribution:
   use graph kernel and locality sensitive hashing
   $\mapsto$ *linear efficiency*
3. Refine clusters/families C_i (via structural alignment)
4. Make models for C_i, scan genome to collect and remove all members of C_i
5. Iterate and find additional clusters/families



Features for extraction of candidate ncRNA

1. Minimum free energy (MFE)

   Has a natural occurring RNA sequence a lower MFE than random sequences of the same size and base composition?

2. Structure Conservation Index (SCI)

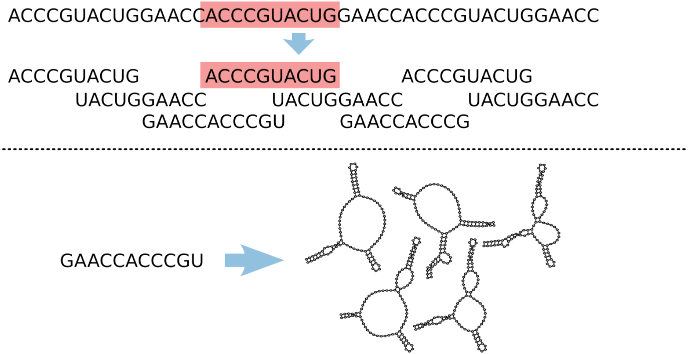
   Are there many sequences that are structurally conserved across related organisms?

RNAz: SVM on 2 features can reliably and efficiently identify RNA sequences that are likely to have a biological function (given pre-computed genome alignments)

Washietl, Hofacker & Stadler, Proc. Natl. Acad. Sci. USA (2005)

Representing ncRNA as graphs

- Given a ncRNA sequence consider all substrings obtained as windows of size $W_1, W_2, \ldots, W_p$ at intervals $I_1, I_2, \ldots, I_m$
- Consider a set of k most representative structures for each subsequence
- $\Rightarrow$ graph with disconnected components



ACCCGUACUGGAACCACCCGUACUGGAACCACCCGUACUGGAACC

ACCCGUACUG ACCCGUACUG ACCCGUACUG

UACUGGAACC UACUGGAACC UACUGGAACC

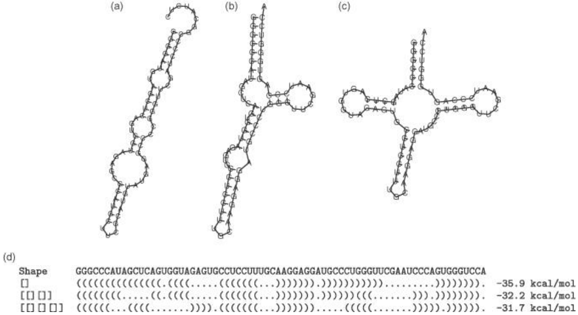
GAACCACCCGU GAACCACCCG

GAACCACCCGU

Representative structures

- Sample set of folding structures...
- which exhibit significantly different shapes (*abstraction levels*)
- in a small energy range above the minimum free energy
  ↦ representative structures

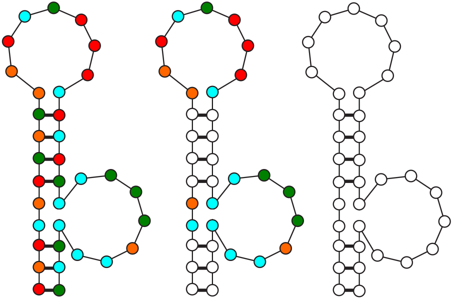
R. Giegerich, B, Voß and M. Rehmsmeier, “Abstract shapes of RNA”, NAR (2004)



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

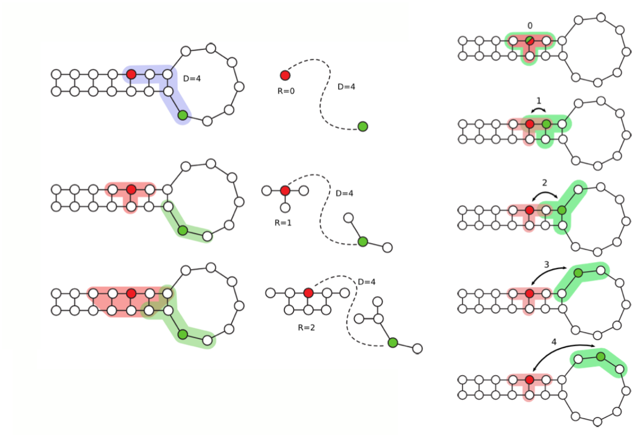
Encoding RNA domain specific knowledge

- The binding of nucleotides stabilizes and defines a structure
  - A pair of nucleotides can mutate provided that they still bind *(compensatory mutations)*
  - ⇒ exact sequence identity in these regions (stems) is at times not required to preserve functionality
- We encode this knowledge via structure replication and label equivalence enforcement



An efficient graph kernel: NSPDK

Enumerate all pairs of near small neighborhood subgraphs



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

Fast Graph Clustering

- Graph kernel ↦ efficient computation of pairwise similarity $\Rightarrow$ direct use in clustering
  ...but it is still $O(n^2)$ (breaks down at 10-100K instances)
- Locality Sensitive Hashing techniques allow fast ($O(1)$) approximate neighbor retrieval
- **Key idea:** use hash collision as surrogate for similarity

MinHash

- Jaccard set similarity $s(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$
- Signature $H(C_i) =$ smallest index of non-zero component of C_i after random permutation of components
- Surprising property: $P(H(C_i) = H(C_j)) = s(C_i, C_j)$
- Set of signatures ↦ similarity as fraction of common signatures (better approximation)
- Replace random permutation with re-hashing for efficiency

k-Neighbors search for a graph G

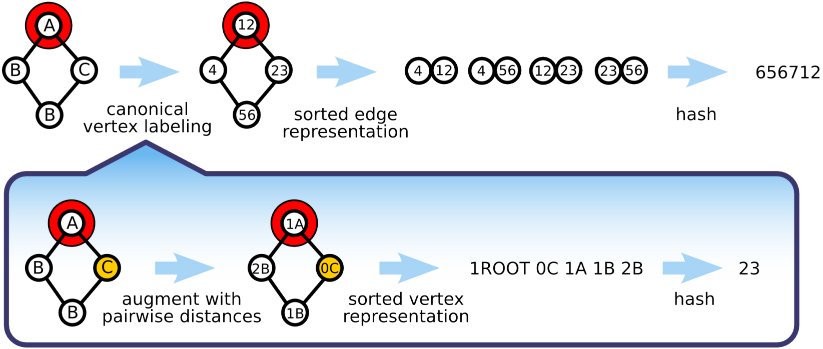
- Find all the signatures of G, $H^s(G)$
- Retrieve all G_i for each signature (efficient step $O(1)$)
- Retrieve the $m-$most frequent G_i in M
- Output the k-nearest neighbors between G and $G_i \in M$ (compute exact similarity/distance on few instances)

Clustering

- Define density using neighborhood
- Clusters as neighbors of graphs sampled from dense regions

Explicit sparse graph encoding 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 $\mapsto$ efficient (linear) in practice

Preliminary experimental results

Small scale

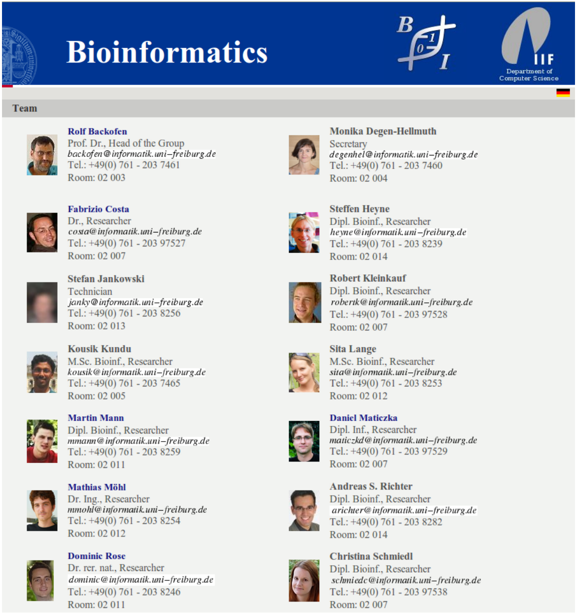
- RFam dataset: 23 ncRNA families, 6-20 sequences each (400 total) with 100 graphs of 50 nodes per sequence
- Clustering time: minutes
- Identification of 21 families ($>$ 0.8 F-measure) in hours
- Complete pairwise alignment $\approx$ days

Large scale

- Drosophila genome: extracted 16K ncRNA sequences (90-300 nt in length)
- Unknown number of correct families (ongoing analysis)
- Clustering time: hours
- Overall run-time of days vs. practical infeasibility (current limit $\approx$ 2-3K sequences) using pure alignment techniques

Conclusions

1. Manipulating graph representation is very flexible way to inject domain knowledge
2. Developing hash techniques for graphs allows to tackle interesting tasks on complex objects at large scales (i.e. clustering ncRNA structures at genomic scale)



Bioinformatics

Team

Rolf Backofen
Prof. Dr., Head of the Group
backofen@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 7461
Room: 02 003

Fabrizio Costa
Dr., Researcher
costa@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 97527
Room: 02 007

Stefan Jankowski
Technician
janky@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 8256
Room: 02 013

Kousik Kundu
M.Sc. Bioinf., Researcher
kousik@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 7465
Room: 02 005

Martin Mann
Dipl. Bioinf., Researcher
mmann@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 8259
Room: 02 011

Mathias Möhl
Dr. Ing., Researcher
mmohl@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 8254
Room: 02 012

Dominic Rose
Dr. rer. nat., Researcher
dominic@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 8246
Room: 02 011

Monika Degen-Hellmuth
Secretary
degenhel@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 7460
Room: 02 004

Steffen Heyne
Dipl. Bioinf., Researcher
heyne@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 8239
Room: 02 014

Robert Kleinkauf
Dipl. Bioinf., Researcher
robertk@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 97528
Room: 02 007

Sita Lange
M.Sc. Bioinf., Researcher
sita@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 8253
Room: 02 012

Daniel Maticzka
Dipl. Inf., Researcher
maticzkd@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 97529
Room: 02 007

Andreas S. Richter
Dipl. Bioinf., Researcher
arichter@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 8282
Room: 02 014

Christina Schmiedl
Dipl. Bioinf., Researcher
schmiedc@informatik.uni–freiburg.de
Tel.: +49(0) 761 - 203 97538
Room: 02 007

Thanks to the Bioinformatics Group in Freiburg