Structural Alignment of two RNA Sequences with Lagrangian Relaxation

Mandy Fuchs

16.7.2007

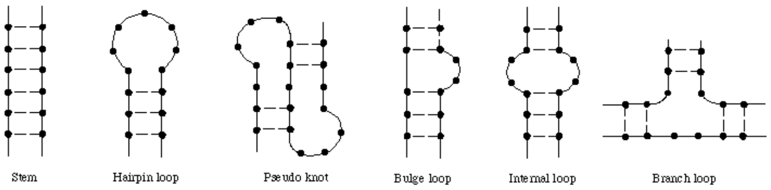
Overview

- Introduction
- Graph theoretical model
- ILP
- Relaxed Problem
- Lagrange Relaxation
    - Lagrange Function
- Results
- Summary

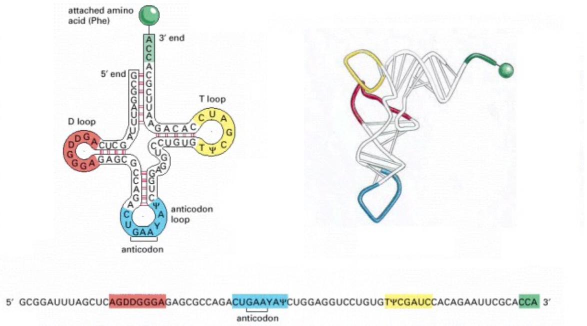
Introduction

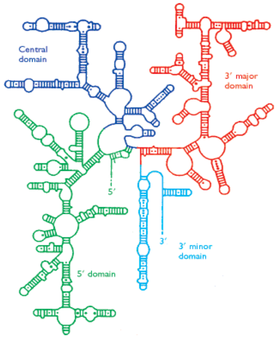
RNA

- single strand: A, C, G, U
- secondary structures are formed by hydrogen bonds: G-C (3), A-U (2)



- RNAs have different structure and function (functional motifs)


attached amino acid (Phe)
3′ end
5′ end
T loop
D loop
anticodon loop
anticodon
5′ GCGGAUUUAGCUCAGDDGGGAGAGCGCCAGACUGAAYAΨCUGGAGGUCCUGUGTΨCGAUCCACAGAAUUCGCACCA 3′
anticodon

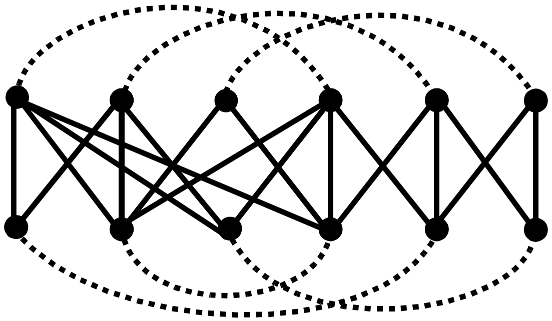


⇒ related functional RNAs often have low sequence but high structural similarity

Graph theoretical model

Given: Two annotated sequences $(S_{1,} P_1)$ and $(S_{2,} P_2)$

Find: Structural alignment of maximal weight

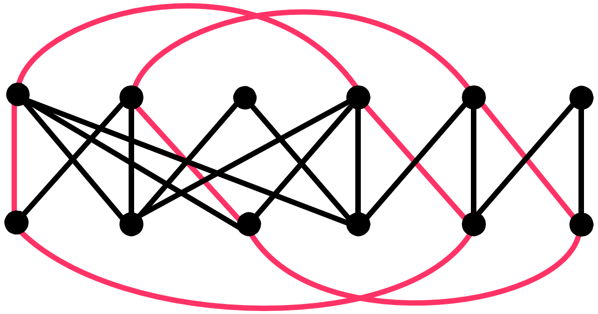


Alignment graph $G = (V, A \cup I, w)$:

- Vertices: characters in S_1 and S_2
- Alignment edges
  - are in conflict if they cross or touch each other
- Interaction edges
  - two interaction edges are in conflict if they share one base
  - two interaction edges are realized by alignment edges
- Each alignment edge and interaction match is assigned a positive weight

Integer linear programming (ILP)

Objective function: $$\max \sum_{m \in A} \sum_{l \in A} w_{lm} y_{lm} + \sum_{m \in A} w_m x_m \qquad (1)$$



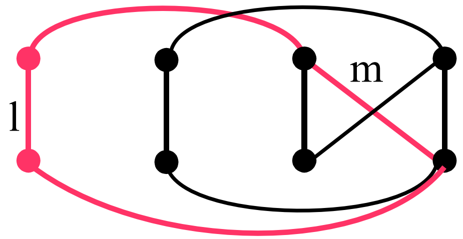
Variables: $x \in \{0,1\}^{A}, \; y \in \{0,1\}^{A \times A}$

$x_m = 1$, if the alignment edge m is part of the alignment

$y_{ml} = 1$, if the alignment edges l and m realize the interaction match $(1, m)$

Integer linear programming (ILP)

Constraints:



- No alignment edges are in conflict:

$$\sum_{l \in I} x_l \leqslant 1 \qquad \forall I \in \mathit{I} \tag{2}$$

- Interaction matches are realized by alignment edges:

$$y_{lm} = y_{ml} \qquad \forall l, m \in A, l < m \tag{3}$$

- Every vertex is incident to at most one interaction edge:

$$\sum_{l \in A} y_{lm} \leqslant x_m \qquad \forall m \in A \tag{4}$$

Lagrange Function

optimization problem: $\inf\left\{ f_0(x) \mid f_1(x) \leqslant 0 \right\}$ (1)

modified problem: $\inf\left\{ f_0(x) + y f_1(x) \mid x \in \mathbb{R} \right\}$ with $y \geq 0, y \in \mathbb{R}$ (2)

$x^*(y)$: optimal solution for a given y

y: describes a weight for (not) satisfied constraints

- $y = 0$

  optimal value $x^*(0)$ will violate the constraint $f_1(x) \leqslant 0$

- great value for y

  minimize f_1, but $x^*(y)$ is not optimal for eq. (1)

=> find $\bar{y} > 0$ such that $f_1(x^*(\bar{y})) = 0$, i.e. $x^*(\bar{y})$ is an optimal solution of eq. (1)

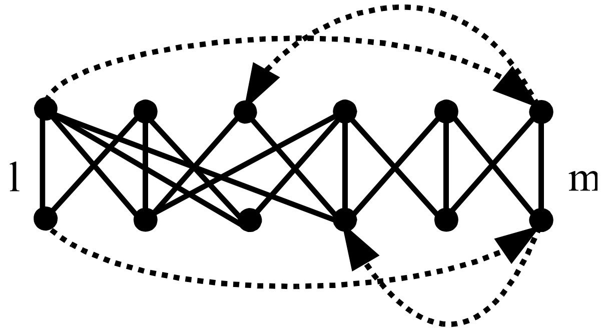
Lagrangian Relaxation

Relaxed problem:

- compute max. profit for each m and a conventional alignment
- time $O(|A|^2)$ but solution is not optimal

Integrate bad constraints in the objective function and penalize its violation:

$$\max \sum_{m \in A} \sum_{l \in A} w_{lm} y_{lm} + \sum_{m \in A} w_m x_m + \sum_{l \in A} \sum_{m \in A, l < m} \lambda_{lm} (y_{lm} - y_{ml})$$



$\Rightarrow$ every alignment edge choose the interaction that maximizes its overall score

Lagrangian multipliers

Find: Lagrangian multipliers that provide the best bound to the original problem

Iterative subgradient optimization:

$$\lambda_{lm}^{0} = 0 \text{ for all } m, l \in A$$

$$\lambda_{lm}^{i+1} = \begin{cases} \lambda_{lm}^{i} & \text{if } s_{lm}^{i} = 0 \\ \max(\lambda_{lm}^{i} - \gamma_i, -w_{lm}) & \text{if } s_{lm}^{i} = 1 \\ \min(\lambda_{lm}^{i} + \gamma_i, w_{lm}) & \text{if } s_{lm}^{i} = -1 \end{cases}$$

subgradient:

$$s_{lm}^{i} = \bar{y}_{lm} - \bar{y}_{ml} \text{ for all } l, m \in A, l < m$$

step size:

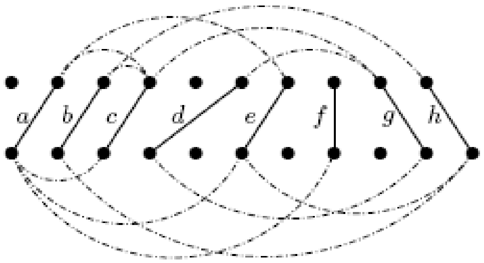
$$\gamma_i = \mu \frac{\text{UB} - \text{LB}}{\sum_{m,l \in A} (s_{lm}^{i})^2}$$

Computing an lower bound

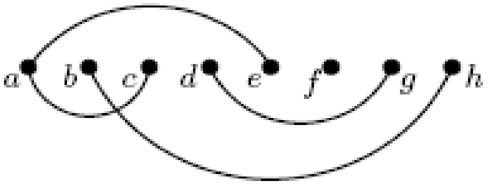
Given:

- traditional sequence alignment (alignment edges)
- set of interaction edges

Find: best possible set of interaction edges such that no pair is in conflict



Alignment graph



Interaction matching graph

⇒ Matching of maximum weight in the interaction graph

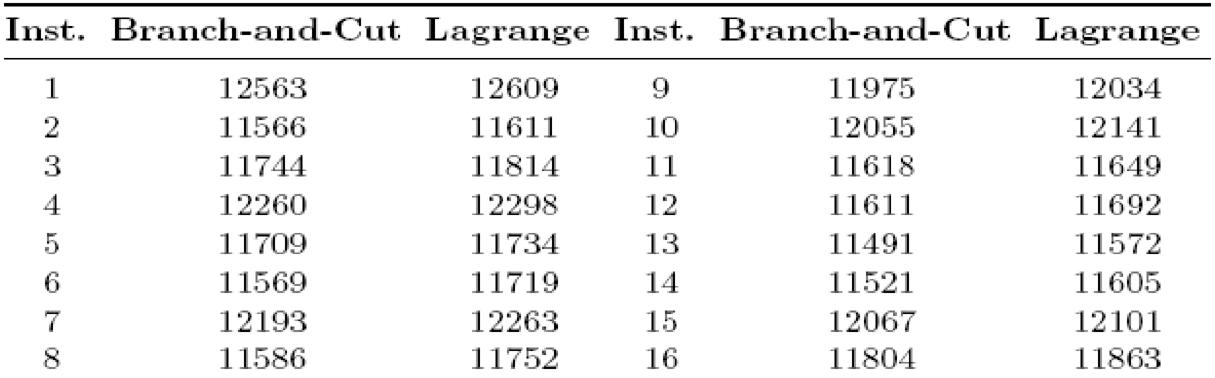
Results

Data: 23S ribosomal RNA

Alignment edges:

- conventional sequence alignment with affine gap cost (6, 2)
- insert all alignment edges realized by suboptimal alignments
- scoring schema: match 4, mismatch 1, interaction match 8

⇒ compute instances with 16000 – 21000 alignment edges in only 10 min



Inst.	Branch-and-Cut	Lagrange	Inst.	Branch-and-Cut	Lagrange
1	12563	12609	9	11975	12034
2	11566	11611	10	12055	12141
3	11744	11814	11	11618	11649
4	12260	12298	12	11611	11692
5	11709	11734	13	11491	11572
6	11569	11719	14	11521	11605
7	12193	12263	15	12067	12101
8	11586	11752	16	11804	11863

Results

Pseudoknots

- Data
  - 18 rRNA from Drosophila melongaster and human (1870 and 1995 bases)
  - three parts within the sequence form pseudoknots
- Traditional sequence alignment
  - structural score: 12031
  - realizing 366 interaction matches
- Lagrange method
  - structural score: 12662
  - realizing 409 interaction matches

Summary

- ILP formulation for RNA structural alignment
- Lagrangian Relaxation in time $O\left(|A|^2\right)$
- structural alignments with higher score
- method is faster than previous algorithm
- detect conserved structures containing pseudoknots
- can be extended to multiple sequence alignments

References

- Structural Alignment of Two RNA Sequences with Lagrangian Relaxation, Markus Bauer, Gunnar W. Klau, Vienna University of Technology
- Lagrangian Relaxation: An Overview, Discrete Mathematics, K. Reinert
- Optimierung; Jarre, Stoer, Springer-Verlag, 2004