Structural Alignment of RNA Sequences with Lagrangian Relaxation

Markus Bauer [a,c], Gunnar W. Klau [b], Knut Reinert [a]

a Institute of Computer Science, Free University of Berlin, Germany, b Institute of Mathematics, Free University of Berlin, Germany, c International Max Planck Research School on Computational Biology and Scientific Computing

Abstract. RNA is generally a single-stranded molecule where the bases form hydrogen bonds within the same molecule leading to structure formation. In comparing different homologous RNA molecules it is usually not sufficient to consider only the primary sequence, but it is important to consider both the sequence and the structure of the molecules. Traditional alignment algorithms can only account for the sequence of bases, but not for the base pairings. Considering the structure leads to significant computational problems because of the dependencies introduced by the base pairings. In this paper we address the problem of optimally aligning given RNA sequences either with or without known structure. We phrase the problem as an integer linear program and then show how to solve it using Lagrangian relaxation.

1 Introduction

Similarity searches based on primary sequence or the detection of structural features using multiple alignments are usually the first steps in the analysis of biomolecules. Unfortunately, many functional classes of RNA show little sequence conservation, but rather a conserved secondary structure which is formed by folding onto itself and forming hydrogen bonds between its bases. Among such RNAs are tRNA, rRNA, and SRP RNA [5].

Hence, algorithms to compute (multiple) alignments ought to take not only the sequence, but also the secondary structure into account. In the study of [7] the authors back up this consideration by showing that sequence based alignments are significantly worse than sequence-structure based alignments if their pairwise sequence identity sinks below $\approx 60\%$. Thus, the problem of producing RNA alignments that find a common structure has become the bottleneck in the the computational study of functional RNAs.

In this paper we deal with the computation of a *multiple* RNA sequence-structure alignment, given a number of RNA sequences together with their secondary structure.

2 Approach

We first describe the graph-theoretical model we use which is based on the description in [2] and [6]. Then, we present an integer linear programming for-

mulation for this model and devise a solution approach based on Lagrangian relaxation.

2.1 Graph-Theoretical Model for Structural RNA Alignment

Let S be a sequence $s_1, \ldots, s_n$ of length n over the alphabet $\Sigma = \{A, C, G, U, -\}$. A paired base (i, j) is called an *interaction*, if $s_i \neq -$ and $s_j \neq -$ and if (i, j) forms a Watson-Crick-pair. The set P of interactions is called the *annotation* of sequence S. Two interactions are said to be in *conflict*, if they share one base. A pair (S, P) is called an *annotated sequence*. Note that a structure where no pair of interactions is in conflict with each other forms a valid secondary structure of an RNA sequence.

We are given a set of k annotated sequences $\{(S_1, P_1), \ldots, (S_k, P_k)\}$ and model the input as a mixed graph $G = (V, L \cup I \cup A)$. The set V denotes the vertices of the graph, in this case the bases of the sequences, and we write v_j^i for the jth base of the ith sequence. The set L contains undirected *alignment edges* between vertices of two different input sequences (for sake of better distinction called *lines*) whereas the set I codes the annotation of the sequence by means of *interaction edges* between vertices of the same sequence. In addition to the undirected edges the graph has directed arcs A representing *consecutivity* of characters within the same string that run from each vertex to its "right" neighbor, i.e., $A = \{(v_j^i, v_{j+1}^i) : 1 \leq i \leq k, 1 \leq j < |S_i|\}$. A *path* in a mixed graph is an alternating sequence $v_1, e_1, v_2, e_2, \ldots$ of vertices v_i and lines or edges $e_i \in L \cup A$. It is a *mixed path* if it contains at least one arc in A and one line in L. A mixed path is called a *mixed cycle* if the start and end vertex are the same. A mixed cycle represents an ordering conflict of the letters in the sequences. In the two-sequence case a mixed cycle represents lines crossing each other. A subset $\mathcal{L} \subset L$ corresponds to an *alignment* of the input sequences $S_1, \ldots S_k$ if $\mathcal{L} \cup A$ does not contain a mixed cycle. In this case, we use the term alignment also for $\mathcal{L}$.

Two interaction edges $(i_1, i_2) \in P_i$ and $(j_1, j_2) \in P_j$ are said to be *realized* by an alignment $\mathcal{L}$ if and only if $\mathcal{L}$ contains the alignment edges $l = (i_1, j_1)$ and $m = (i_2, j_2)$. The pair (l, m) is called an *interaction match*. Note that (l, m) is an ordered tuple, that is, (l, m) is distinct from (m, l). Figure 1 illustrates the above definitions by means of an example.

Each line l and each interaction match (i, j) is assigned a positive weight w_l and w_{ij}, respectively, representing the benefit of realizing this edge or the match.

Approaches for traditional sequence alignment aim at maximizing the score of edges in an alignment $\mathcal{L}$. A structural alignment, however, must take the structural information encoded within the interaction edges into account as well. A structural alignment of the annotated sequences $\{(S_1, P_1), \ldots, (S_k, P_k)\}$ calls for an alignment such that the weight of the lines plus the weight of the realized interaction matches is maximal.

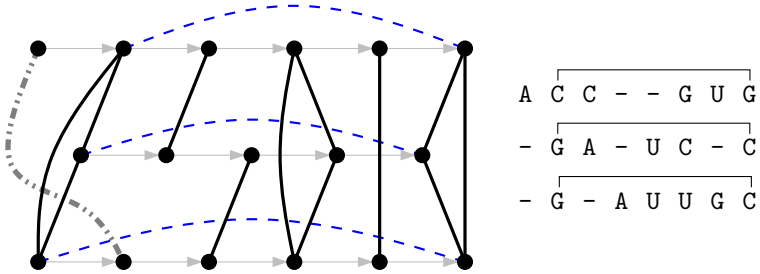

Fig. 1. Graph-theoretic concept of alignment. The right side shows a structural alignment of three annotated sequences, the left side the corresponding graph G. Thicker lines represent alignment edges in $\mathcal{L}$, adding the grey line (v_1^1, v_2^3) creates a mixed cycle. Lines $\mathcal{L}$ realize three interaction matches, namely $((v_2^1, v_1^2), (v_6^1, v_5^2))$, $((v_2^1, v_1^3), (v_2^1, v_1^2))$ and $((v_1^2, v_1^3), (v_5^2, v_6^3))$.

2.2 Integer Linear Programming Formulation

Modeling our problem as described above lets us very conveniently write it as the following integer linear program (ILP):

$$\max \quad \sum_{l \in L} w_l x_l + \sum_{l \in L} \sum_{m \in L} w_{lm} y_{lm} \tag{1}$$

$$\text{s.t.} \quad \sum_{l \in C} x_l \leq |C \cap L| - 1 \qquad \forall \text{ mixed cycles } C \tag{2}$$

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

$$\sum_{m \in A} y_{lm} \leq x_l \qquad \forall l \in L \tag{4}$$

$$0 \leq x \leq 1, \quad 0 \leq y \leq 1 \qquad \text{integer} \tag{5}$$

The variable x_l equals one, if line l is part of the alignment, whereas $y_{lm} = 1$ holds, if lines l and m realize the interaction match (l, m). One can easily verify that all properties for a multiple structural alignment are satisfied: (3) and (4) guarantee that interaction matches are realized by lines and that every vertex is incident to at most one interaction edge, whereas (2) ensures that the selection of lines forms a multiple alignment. The order $l < m$ within the equality constraints (3) denotes an arbitrary order defined on the elements of A (to avoid the same constraints to appear twice in the ILP).

This ILP formulation is similar to the one given in [6] where the authors present a branch-and-cut approach for case of structurally aligning two RNA sequences. Previous work on contact map alignment in the area of proteomics by [3] and for the two-sequence case of our problem by [2] indicates, however, that Lagrangian relaxation is better suited to obtain good solutions to this ILP than a direct branch-and-cut approach in terms of running time.

2.3 Lagrangian Relaxation Approach

Following the Lagrangian optimization method, we drop the constraints that complicate the original problem – in this case the equality constraints (3) – and incorporate them into the objective function with a penalty term for their violation.

Lemma 1. *The relaxed problem is equivalent to the general multiple sequence alignment problem.*

Proof. We distinguish two cases, depending on whether a line l is part of an alignment or not. First, assume $x_l = 0$. In this case, due to (4), all y_{lm} must be zero as well, and the contribution of line l to the objective function is zero. If, however, a line is part of an alignment, its maximal contribution to the score is given by solving

$$p_l := \max \; w_l + \sum_{m \in L} w_{lm} y_{lm} \tag{6}$$

$$\text{s.t.} \sum_{m \in L} y_{lm} \leq 1 \tag{7}$$

$$\sum_{m \in L} y_{lm} \leq 0 \qquad \forall\, m \in C : l \in C \tag{8}$$

$$0 \leq x \leq 1, \quad 0 \leq y \leq 1 \qquad \text{integer} \tag{9}$$

where C is a mixed cycle in G. Inequality (7) states that only one interaction match can be chosen. According to the objective function (6) it is clear that this will be the one with the largest weight w_{lm}. Inequality (8) constrains this choice by excluding interaction matches with lines m that are in conflict to l. This ILP is easily solvable by just selecting the most profitable interaction match $(l, \hat{m})$ such that l and $\hat{m}$ do not cross each other, which can be done in constant time. Thus, the profit p_l a line l can realize is given by its own weight w_l plus the weight $w_{l\hat{m}}$ of such an interaction match.

In the second step, we compute the optimal overall profit by solving the multiple sequence alignment problem

$$\max \; \sum_{l \in L} p_l x_l$$

$$\text{s.t.} \sum_{l \in C} x_l \leq |C \cap L| - 1 \qquad \forall \text{ mixed cycles } C$$

$$0 \leq x \leq 1 \qquad \text{integer}$$

Let x^* be the solution of this problem. We claim that an optimal solution of the relaxed problem is given by (x^*, y^*) with $y^*_{lm} = x^*_m y_{l\hat{m}}$. (proof omitted)

□

Having demonstrated how to formulate the relaxed problem as a pure sequence-based multiple alignment problem we now describe the Lagrangian method. Formally, we introduce appropriate Lagrangian multipliers λ^i with $\lambda^i_{ml} = -\lambda^i_{lm}$ for $l < m$ and with $\lambda^i_{ll} = 0$ and define the Lagrangian problem as

$$\begin{aligned}
\max \quad & \sum_{l \in L} w_l x_l + \sum_{l \in L} \sum_{m \in L} (\lambda^i_{lm} + w_{lm}) y_{lm} & \\
\text{s.t.} \quad & \sum_{l \in C} x_l \leq |C \cap L| - 1 & \forall \text{ mixed cycles } C \\
& \sum_{m \in L} y_{lm} \leq x_l & \forall\, l \in L \\
& 0 \leq x \leq 1, \quad 0 \leq y \leq 1 & \text{integer}
\end{aligned}$$

Note that, according to Lemma 1, we can solve instances of the Lagrangian problem by solving a multiple sequence alignment problem where the profits of the interaction matches are coded in the weights of the lines.

In the end, the task is to find Lagrangian multipliers that provide the best bound to the original problem. We do this by employing iterative subgradient optimization as proposed by Held and Karp in 1971. A detailed description is beyond the scope of this paper, for details the reader referred to [4].

References

1. E. Althaus, A. Caprara, H.-P. Lenhof, and R. K. Multiple sequence alignment with arbitrary gap costs: Computing an optimal solution using polyhedral combinatorics. *Bioinformatics*, 18(90002):S4–S16, 2002.
2. M. Bauer and G. W. Klau. Structural Alignment of Two RNA Sequences with Lagrangian Relaxation. In *Proceedings of the 15th Annual International Symposium on Algorithms and Computation (ISAAC)*, number 3341 in LNCS, pages 113–123. Springer, 2004.
3. A. Caprara and G. Lancia. Structural Alignment of Large-Size Proteins via Lagrangian Relaxation. In *Proceedings of the Sixth Annual International Conference on Computational Biology*, pages 100–108. ACM Press, 2002.
4. M. Held and R. Karp. The traveling-salesman problem and minimum spanning trees: Part II. *Mathematical Programming*, 1:6–25, 1971.
5. I. L. Hofacker, S. H. F. Bernhart, and P. F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20:2222–2227, 2004.
6. H.-P. Lenhof, K. Reinert, and M. Vingron. A Polyhedral Approach to RNA Sequence Structure Alignment. *Journal of Comp. Biology*, 5(3):517–530, 1998.
7. S. Washietl and I. L. Hofacker. Consensus folding of aligned sequences as a new measure for the detection of functional rnas by comparative genomics. *Journal of Molecular Biology*, 2004.