A Polyhedral Approach to Sequence Alignment Problems

Tallman Zacharia Nkgau
School of Computer Science, McGill University
3480 University Street, Montreal, PQ, Canada H3A 2A7
nkgautz@cs.mcgill.ca

Based on a Paper by
Kececioglu et. al. Discrete Applied Mathematics (2000)

1 Introduction

It is well known that living things are made of proteins and nucleic acids (DNA — DeoxyriboNucleic Acid and RNA — RiboNucleic Acid). A DNA sequence is made of nucleotides **A, T, C** and **G**, standing for **Adenine, Thymine, Cytosine** and **Guanine**, respectively. In RNA **U – Uracil** replaces **T**. In as so far as our work is concerned, we visualize DNA and RNA as strings over the alphabet $\sum = \{$ A, C, G, T, U$\}$.

Similar sequences (DNA/RNA strings) often have similar origin or function

Figure 1. Fact of Life # 1

The above fact of life motivates our study of DNA/RNA sequences, specifically, of molecular sequence comparisons. The objectives of the study being:

- To determine the origins of life
- To determine the functions of genes

The primary idea underlying sequence comparison is sequence **similarity** which we compute by **aligning** sequences [2, 1].

1.1 Sequence Alignment

Alignment An alignment of sequences S, T is a pair of sequences S', T' (with spaces) such that:

- $\mid S' \mid = \mid T' \mid$
- Removing spaces from S' and T' leaves S and T, respectively
- Space character "–" is not matched wit itself.

Ex. An alignment between $S = ACBCDB$ **and** $T = CADBD$

A	C	–	–	B	C	D	B
–	C	A	D	B	–	D	–

1.2 Alignment Scoring

The score of aligning x, y is $\sigma(x, y)$. The value of an alignment, hence the similarity between the two sequences, is

$$\sum_{i=1}^{|S'|} \sigma(S'[i], T'[i])$$

An **optimal alignment** is one of maximum value. A dynamic programming algorithm gives an exact solution in $\mathcal{O}\ (2^2 n^2)$ time and $\mathcal{O}\ (n^2)$ space, where $\mid S \mid = \mid T \mid = n$ and we also compute the alignment. For two sequences, we call this a **pair-wise alignment**, and is useful if one of the sequences has known properties.

1.3 Multiple Sequence Alignment (MSA)

Evolutionarily and functionally related molecular sequences can differ significantly and yet preserve the same three dimensional structures and the same active sites

Figure 2. Fact of Life # 2

The above fact of life gives us the motivation to study multiple sequence alignment. In this case we are given $k > 2$ sequences which we need to align. The final alignment gives us common features of the sequences, which might not be evident had we performed only pair-wise alignments. MSA is usually used for shape analysis and the sequences are known to be related. Dynamic programming (a straightforward extension of the pair-wise algorithm) computes the exact solution in $\mathcal{O}$ $(2^k n^k)$ time and $\mathcal{O}$ (n^k) space. The dynamic programming algorithm has some limitations:

- It is inefficient, performs well for $k \leq 10$ and $n \approx 300$ — this is not a surprise, the general problem is NP-complete.
- No "optimal" scoring function. Most people used the **sum-of-pair (SP)** scoring scheme [1].

1.4 Maximum Trace Problem (MT)

Let $S = \{S_1, S_2, \ldots S_k\}$ be $ksequences$, and let $S' = \{S'_1, S'_2, \ldots S'_k\}$ be an alignment of S. The characters x, y are said to be **aligned** if they are placed in the same column in an alignment. We construct the following graph $G = (V, E)$:

- $V =$ character positions in the input strings.
- $E =$ pairs of characters we like to be aligned in an alignment of the input strings
- Edge $e = (x, y) \in E$ is said to be **realized** by an alignment if x, y are aligned.

A subset E' of E is called **trace** of S', denoted by trace(S') if i is realized by S'. The MT problem [4] is then

Input: An alignment graph, $G = (V, E)$ with positive edge weights
Output: An alignment S' with trace(S') maximum.

Figure 3. The Maximum Trace Problem.

The problem MT is NP-complete and a **branch and bound** algorithm based on dynamic programming takes $\mathcal{O}$ $(k^3 2^k n^k)$ time and $\mathcal{O}$ (kn^k) space citeJK93.

2 Generalized Maximum Trace Problem (GMT)

In Kececioglu et al. [3], MT is generalized as follows

- multiple edges are allowed
- different scoring functions can be used

To do this E is partitioned into D **blocks** where a block is a trace of a pair of sequences and it is an indivisible unit — it either wholly realized on it is not realized at all. Each block is assigned a weight w_d. The value of an alignment is the sum of the blocks realized. The GMT problem [3] is then

Input: An alignment graph, $G = (V, E)$ with positive edge weights and the set of blocks D
Output: An alignment S' of maximum value (max $\sum w_d$ where w_d is realized).

Figure 4. The Generalized Maximum Trace Problem.

Since the GMT has the MT as a special case, it follows that GMT is as hard as MT. Therefore we do not expect an efficient algorithm for it.

2.1 Characterization as ILP

A **mixed graph** is a tuple $G = (V, E, A)$ where A is a set of arcs. A **path** p is an alternating sequence $v_1, e_1, \ldots v_k$ of edges such that either $e_i = (v_i, v_{i+1}) \in E$ or $e_i = (v_i, v_{i+1}) \in A$. A **mixed path** is a path that contains at least one edge in E and one arc in A. A **mixed cycle** is a mixed path where the last and first vertices are the same. The length of a mixed path P (cycle C) is the number of arcs and edges in it. The size of a mixed path P (cycle C) is the number of edges in it. An **extended alignment graph (EAG)** $G = (V, E, H)$ is an alignment graph with $H = \{(s_{ij}, s_{ij+1} : 1 \leq i \leq k, 1 \leq j <| S_i |)\}$ where s_{ij} is the vertex that corresponds to letter j of sequence i. Let S_i denote the vertices of sequence i. Define for an edge $e \in E$ between S_i and S_j with $i < j$, start(e) to denote the index of the letter of S_i where the edge starts and by end(e) the index of the letter of S_j where the edge ends. For two alignment edges e and f ($e \neq f$) define the relation $\prec$ as

- $e \prec f$ if and only if they both start in S_i and end in S_j for some $i < j$ and if (start(e) $>$ start(f) and end(e) $\leq$ end(f)) or (start(e) = start(f) and end(e) $<$ end(f)).

Two alignment edges e and f are in conflict if either $e \prec f$ or $f \prec e$. A mixed cycle R in G is critical if for all i all vertices in $R \cap S_i$ occur consecutively in R.

Theorem. Let $G = (V, E, H)$ be an EAG, let T be a subset of E and let $G' = (V, T, H)$ be the EAG induced by T. Then T is a trace if and only if there is no mixed cycle in G'.

As you are now aware, the above formulation was to make sure that what we get at the end is an really alignment. The GMT problem can then be cast as follows

Input: An EAG, $G = (V, E, H)$ and a partition D of blocks
Output: Compute a set M (subset of D) of maximum value such that edges in M do not induce a critical mixed cycle on G.

Figure 5. The Generalized Maximum Trace Problem.

Let T be the set of all feasible solutions, that is, subsets of D that do not induce a critical mixed cycle on G, and let P(T) be convex hull of the polytope of their incidence vectors. The GMT in disguise is

max $\sum w_d \cdot x_d$ for all d in D subject to x in P(T)

where x_d is 1 if w_d is realized and 0 otherwise.

3 Solution

The method of solution is via branch and cut [5] using the properties of the polytope defined by the constraints. It is too detailed to be covered here.

4 Results

The method was tried on 15 protein sequences and managed to compute the optimal alignment. Note that 15 sequences is beyond the scope of current dynamic programming techniques.

5 Conclusion

So far, the method can tackle problems that are more difficult to deal with using dynamic programming. However, the Kececioglu et al [3], do not discuss the time complexity of the algorithm, nor whether 15 sequences is the maximum. The size and structure of the extended alignment graph will play a major role in the complexity of the algorithms, so it is worth investigating its role in all of this, perhaps yielding faster approaches.

References

[1] D. Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, 1997.

[2] M. Hallett. *Course Notes for CS 760A*. M. Hallett, 2000.

[3] J. Kececioglu. A Polyhedral Approach to Sequence Alignment Problems. *Discrete Applied Mathematics*, (104):143–186, 2000.

[4] J. Kececioglu and D. Sankoff. The Maximum Weight Trace Problem in Multiple Sequence Alignment. In *LNCS 684, Combinatorial Pattern Matching*. Springer-Verlag New York Inc., 1993.

[5] L. Wolsey. *Integer Programming*. John Wiley & Sons, 1998.