BioGeometry NEWS
December 2004
biogeometry.cs.duke.edu

Research

Faster Multiple Sequence Alignment Algorithms Based on Pairwise Segmentation

by Pankaj K. Agarwal, Yonatan Bilu, Rachel Kolodny

Multiple Sequence Alignment (MSA) is a central problem in computational molecular biology --- it identifies and quantifies similarities among several protein or DNA sequences.The well-known dynamic programming (DP) algorithms align k sequences (each of length n) by constructing a k-dimensional grid graph of size O(n^k), with each of the sequences enumerating one of the dimensions of the grid. The optimal MSA is an optimal path from (n,…,n) to (0,…,0). Unfortunately, the exponential running time makes this approach prohibitive even for modest values of n and k. There is little hope for improving the worst-case efficiency of an algorithm for this problem since the MSA problem is NP-Hard [1]. However, the sequences constructed in the lower-bound constructions are not representative of protein and DNA sequences abundant in nature, and the alignments are not reminiscent of ones studied in practice. This led to the question whether faster algorithms could be developed for protein and DNA sequences.

Many MSA heuristics rely on the on the intuition that some of the input sequences are evolutionary related while others are not. Furthermore, all sequence pairs have been aligned as a pre-processing step. Lee et. al. [2] proposed an innovative heuristics: their POA (partial order alignment) program assumes that the only information in MSA is the aligned pair-wise sub-sequences and their relative positions. Building on Lee et. al.'s intuition, we use the geometric intuition that restricting the paths to the alignments found in pair-wise comparisons corresponds to restricting them to hyperplanes in the hypercube (see Figure1) and develop a formal framework that exploits this observation. We refer to this restricted version of MSA as Multiple Sequence Alignment from Segments (MSAS). Here, the input also includes a segmentation of the sequences, and a set of matching segment pairs. As in the original problem, we seek an MSA that optimizes the objective score. However, only positions that are in matching segments may be aligned. These are biologically reasonable properties of protein and DNA sequences that allow a speedup of the MSA algorithms.

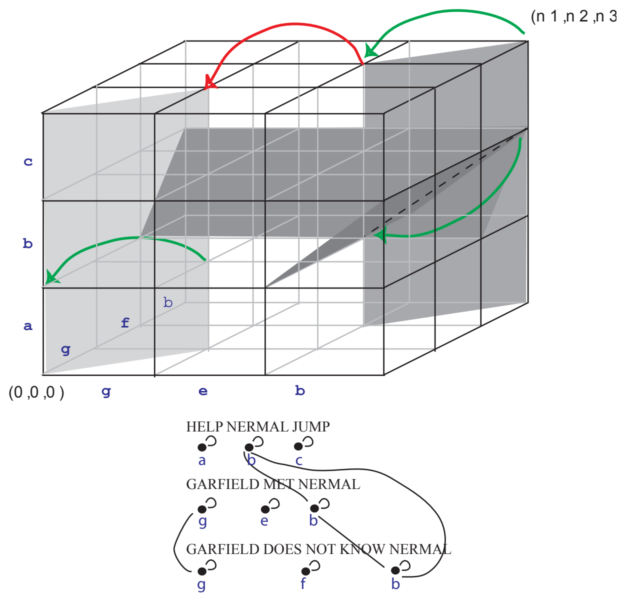


Figure 1: An example of aligning three sequences and their segmentation (as well as marking the aligned parts with edges). The DP matrix and the planes corresponding to the pairwise alignments are also shown. Only moves on these planes are allowed (shown in green), while ones outside of them are not (shown in red)

To find an optimal solution to the MSAS problem it is sufficient to match segments, rather than individual letters of the sequence. Therefore the complexity of a DP algorithm for MSAS depends on the number of segments in each sequence, rather than the length of the sequences. The problem is NP-Complete. Our experimental results show that this reduces the running time, measured in terms of the number of updates in the DP table, by several orders of magnitude. For example, aligning five human proteins (denoted by their Swiss-Prot identifiers) GBAS, GBI1, GBT1, GB11, and GB12 requires 4.3x10^8 rather than 6.6x10^{12} table updates.

We make the algorithm even faster, while still guaranteeing the optimal solution, by further decoupling the subproblems computation and thereby avoiding the computation of some of the cells in the k-dimensional grid. For example, in the above example it reduces the number of table updates to 1.5x 10^5

We consider two more biologically reasonable assumptions for align-ing DNA sequences. First, since a match in DNA indicates (near) identity, we assume that the segment matches have a transitive structure, i.e., if segment A matches segment B and B matches C, then A necessarily matches C. Second, we define an optimal alignment to be one of minimal width, rather than optimal under an arbitrary scoring function. We prove various structural properties of an optimal alignment under these two assumptions and obtain a faster algorithm.

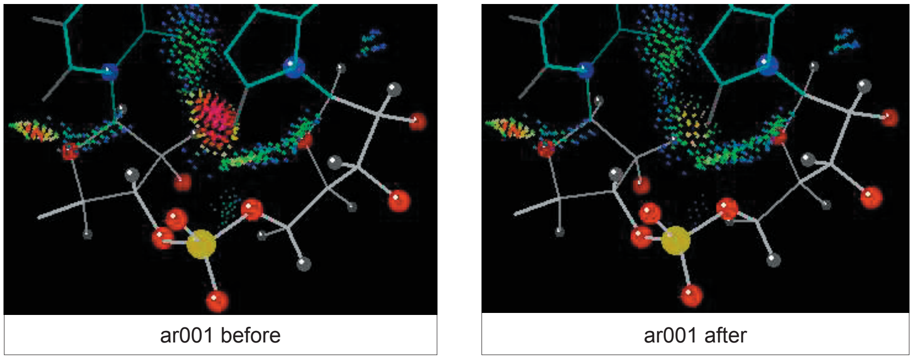
[1] T. Jiang and L. Wang. On the complexity of multiple sequence alignment. *J. Comp. Biol*. 1(4):337-48, 1994.

[2] C. Lee, C. Grasso, and M. F. Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics* 18(3):452-464, 2002.

Student Profile: Xueyi Wang

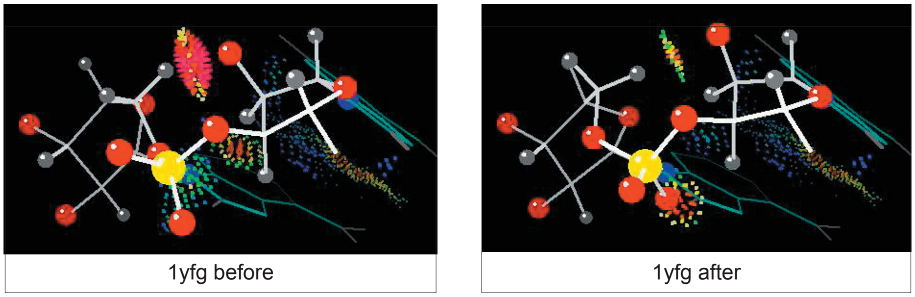
The story of Xueyi Wang, a second year PhD student at UNC Chapel Hill, shows that you never know where you will end up. After he finished his undergraduate study in Refrigeration and Cryogenics from Zhejiang University in 1996, Xueyi worked for 4 years as refrigeration engineer and software engineer in a company. He then returned to Zhejiang University to do an MSc in Computer Science, which he completed in 2003 in the area of Computer-Supported Collaborative Work. Xueyi joined UNC Chapel Hill in the fall of 2003, expecting to do more of the same.

It happened that the BioGeometry project had a research opening that could use his software engineering skills: Some of Prof. Jack Snoeyink's students had ideas to improve Reduce, a program for adding hydrogen atoms to a Protein DataBank (PDB) file, which was developed by Dr. Michael Word of GlaxoWellcome when he was in Richardsons' lab at Duke University. Xueyi replaced the original data structures with STL (Standard Template Library) routines, which were more efficient and easier to extend. He then worked with Andrew Leaver-Fay to replace a brute-force search by dynamic programming. In the process, he caught and repaired bugs in the evaluation of interatomic potentials, but seems to have himself caught the bug of computation in biology applications. He presented this work at the first BioGeometry Workshop after SoCG'04. (view the presentation file at http://biogeometry.cs.duke.edu/meetings/ITR/04jun12/presentations/wang-x.ppt)


ar001 before

ar001 after

Some clashes (indicated by red spikes at left) can be avoided (at right) by small backbones, as in ar001


1yfg before

1yfg after

Others need more significant conformational change, as in 1yfg, but this can still be done without disturbing the rest of the RNA structure.

Although it's a little hard to change research area, Xueyi has taken the Richardsons' course in protein structure and the prerequisites for the bioinformatics program at UNC. This past summer, Xueyi began to work on RNA structure in cooperation with the Richardsons' lab. The aim is to apply inverse kinematics and related geometrical techniques to solve the high-dimensional problem of fitting RNA backbone to experimental crystallographic data – much as Itay Lotan (Stanford) did for protein loops. In electron density maps for large RNA structures, only the phosphate, base and sugar pucker can be relatively clearly located. Since there are 6 dihedrals on the backbone for each RNA residue, it is harder to determine accurate backbones than for proteins; adding hydrogens reveals many errors in existing structures. The "before" and "after" figures show bad clashes (red spikes) that Xueyi's program has reduced or eliminated by adjusting the RNA backbone.

In protein structures the Ramachandran plot shows which phi and psi angle combinations are feasible in the backbone. In RNA the corresponding plot would have 6 dimensions, so another one of Xueyi's objectives is to derive a simpler Ramachandran-like plot for RNA backbones and investigate the collision boundary for allowable and forbidden conformations.

- Profile by Jack Snoeyink

The BioGeometry project is funded by the National Science Foundation under grant CCR-00-86013. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

BioGeometry News is the project's monthly newsletter. For more information, please visit http://biogeometry.cs.duke.edu/newsletter