An Investigation on the Efficiency of Multiple Sequence Alignment Tools

Justin Chao - juchao
Chase Meyer - cmeyer3
Bria Lacour - lacour

October 28, 2016

Proposal

We propose the development of an alignment program that uses both global and local alignment algorithms to create a Multiple Sequence Alignment (MSA) tool that is both accurate and efficient.

Our efforts will involve the usage of random walk algorithms for generating a BLOck SUbstitution Matrix (BLOSUM). The optimal sequence alignment will then be calculated using Needleman-Wunsch and Smith-Waterman algorithms, and a comparison of efficiency will be conducted between the two algorithms. A comparison of run-times will then be conducted between our developed MSA tool, BLAST, and MAFFT on a series of test sequences.

Background

A MSA tool can be used by computational biologists to track evolutionary relationships, predict structures, track mutations, etc. MSA tools come in a variety of forms and are tailored to address specific biological problems. These sequences are usually protein, DNA, or RNA, which can affect the methods of analysis in different ways.

BLAST, the most widely used MSA algorithm, works by generating a matrix of probability values called a BLOSUM matrix. It calculates the probability of certain regions in one sequence aligning with a number of other reference sequences, and generating matrices to store the probability values. This is referred to as scoring and is done using amino-acid similarity in the sequences. The final output from BLAST is the sequence alignment corresponding to the highest probability values determined from these calculated matrices.

The efficiency of these tools can affect scientific analyses, such as identifying significant patterns in protein families or assessing conservation of different genetic characteristics across species. BLAST, and other commonly used alignment tools, involve matrix calculations using the Needleman-Wunsch algorithm or the Smith-Waterman algorithm. While the Needleman-Wunsch algorithm is slightly more accurate, the Smith-Waterman algorithm is used more often for its greater efficiency, a necessity when performing calculations on large matrix sizes.

Tools

- BLAS (Matrix data storage, manipulation, and calculations)
- Probability and Statistical Methods
- Timers and Profilers

Algorithms

Needleman-Wunsch Algorithm

$$\begin{aligned} M(0,j) &= j \times p && \text{for first row, where p is the gap penaly} \\ M(i,0) &= i \times p && \text{for first column} \end{aligned}$$

$$M(0,j) = max \begin{cases} M(i-1,j) + p & \text{top} \\ M(i,j-1) + p & \text{left} \\ M(i-1,j-1) + s(a_j, b_i) & \text{diagonal} \end{cases}$$

Where $s(a_j, b_i)$ = match/mismatch score for sites j and i in sequences a and b.

Smith-Waterman Algorithm

$$\begin{aligned} M(0,j) &= j \times p && \text{for first row} \\ M(i,0) &= i \times p && \text{for first column} \end{aligned}$$

$$M(0,j) = max \begin{cases} 0 & \\ M(i-1,j) + p & \text{top} \\ M(i,j-1) + p & \text{left} \\ M(i-1,j-1) + s(a_j, b_i) & \text{diagonal} \end{cases}$$

Where $s(a_j, b_i)$ = match/mismatch score for sites j and i in sequences a and b.

Research Significance

The goal of BLAST is to find the best alignment, given a scoring system that varies in different multiple sequence alignment software. What makes this project interesting is the fact that there are so many ways to score these alignments and that a potentially significant alternative could be proposed. We would also like to compare the findings of our software to BLAST and other multiple sequence alignment methods. The goal is to develop and optimize or even propose a new way to align sequences that involves algorithms different than the standard to score the sequence alignments and generate substitution matrices. We would even like to attempt to implement parallelism, if time permits, which would add an interesting aspect in comparing efficiency and optimization.

References

- Multiple Sequence Alignment, https://goo.gl/2ulXvM
- Fast Fourier Transform Analysis of DNA Sequences, https://goo.gl/bCNHMC
- The Wilke Lab, https://goo.gl/QshFXK
- Department of Computer Science, Columbia University, https://goo.gl/tli4St