Comparison of Sequence Alignment Algorithms

Tejas Gandhi
Minnesota State University, Mankato

Comparison of sequence alignment algorithms

Tejas Gandhi

The fact that biological sequences can be represented as strings belonging to a finite alphabet (A, C, G, and T for DNA) plays an important role in connecting biology to computer science. String representation allows researchers to apply various string comparison techniques available in computer science. As a result, various applications have been developed that facilitate the task of sequence alignment. The problem of finding sequence alignments consists of finding the best match between two biological sequences. A best match can infer an evolutionary relationship and functional similarity. However, there is a lack of research on how reliable and efficient these applications are especially when it comes to comparing two sequences that might not be highly similar (but could have common patterns that are small yet biologically significant). This study compares two biological sequence comparison packages, namely WuBlast2 and Fasta3. which implement Blast and FastA algorithms, respectively. In order to do so, a framework was developed to facilitate the task of data collection and create meaningful reports. Amino acid sequences corresponding to related proteins, as well as the DNA sequences encoding these proteins, were analyzed with matching parameters for each application. Observations showed a trend of increasing variations between the matches produced by the two applications with decreasing sequence similarity.

Availability: The data and tools are available from the original author.

Contact: tejas.gandhi@mnsu.edu

Background

This research is concerned with presenting an analysis of how seemingly unlikely relationships can be determined by the use of different computational algorithms, which are step-by-step instructions that can perform DNA and protein sequence alignments. The problem of finding sequence alignments consists of finding the best match between two sequences. A best match that displays high sequence similarity potentially hints at an evolutionary relationship and functional similarity [1].

The DNA structure consists of two strands forming the shape of a double helix. Each strand is composed of four basic molecules called nucleotides, which are identical except that each contains a different nitrogen base. A DNA molecule is usually represented by a string of these four bases, namely A (adenine), T (thymine), C (cytosine), and G (guanine). The two strands are held together by hydrogen bonds between the bases so that each base bonds readily to only one other: A to T and C to G. DNA structure can be thought of as a zipper where two strands can be unzipped starting at one end and the unwinding of the DNA will expose single bases on each strand.

Since the pairing requirements imposed by DNA structure are strict (A-T and C-G), each base will bond only with its complementary base. When separated, this allows each strand to act like a template which can copy the other in a process known as replication. Faulty replication process can create a change in the sequence of nucleotides resulting in a mutation. Mutations, often a cause of genetic diseases, are also the force behind the evolution of the genetic makeup of organisms and creation of new species. The most common form of mutations is referenced as point mutation which includes insertion, deletion, and substitution of nucleotides. Given time, a sequence might mutate into two or more different sequences.

However, sequence mutation does not necessary translate into different structure. This is evident by the fact that there are far more protein sequences than protein structures. Therefore the ability to compare two sequences for their relatedness allows inferences to be made regarding their biological function and structure. Such comparisons can be made by aligning two sequences and then deciding if the alignment is by relation or chance. For example, sequence alignment would show a relationship between the eyeless gene in a fruit fly (where absence results in a lack of eyes) and the aniridia gene in humans (where absence of gene is associated to eye problems in humans). Researchers have found that when the aniridia gene is inserted into a fruit fly missing

the eyeless gene, it causes the production of normal fruit fly eyes [2]. Following is a small portion of the protein sequence of the eyeless gene (A) and the aniridia gene (B):

```
A:  IERLPSLEDMAHKGHSGVNQLGGVFV

B:  IPRPPARASMQNSHSGVNQLGGVFV
```

Alignment of the above sequences consists of finding two new sequences, A' and B', of equal lengths with no gaps at the same position. One possible alignment would be:

```
A':  IERLPSLEDMAHKGHSGVNQLGGVFV

     |  |  |      |     ||||||||||||

B':  IPRPPARASMQNS-HSGVNQLGGVFV
```

The gaps/dashes in the above alignment signify insertion/deletion of nucleotides in the original sequence whereas mismatches can be thought of as the substitution of nucleotides. Differentiation of a good alignment from a poor alignment requires quantification using some kind of scoring system. A simple scoring system might set a character match=1, mismatch=0, gap/dash=-1. In the above alignment we have 16 matches, one instance of a gap/dash, and 9 mismatches for a total score of fifteen: *16(1) + (-1) +9(0) = 15*. The above alignment shows that there is at least a partial match between the two genes. This allows us to discern that the human gene, about which not much is known, is potentially related to the fruit fly gene which has been thoroughly researched.

If n equals the length of the longest sequence, there are approximately 2^{2n} possible alignments between two sequences. Of these possible alignments, the best match or optimal alignment(s) can be considered as the one with the highest score. For a long DNA or protein sequence (large n), it would be time consuming to try all the possible alignments (brute force/naive algorithm). Also, sequence comparison can involve scanning not just a pair but hundreds of sequences which cannot be done manually. As of January, 2003, there were 22,318,883 sequence records in the online public database GenBank [3]. To find an optimal alignment for a pair of sequences based on a predetermined scoring system in a more timely manner is key. Hence, it is desirable to use efficient computational methods that can process such large amounts of data.

The fact that biological sequences can be represented as strings belonging to a finite alphabet (A, C, G, and T for DNA) plays an important role in connecting biology to computer science. Such a representation allows us to use a wide variety of available string algorithmic techniques in the computer science field to analyze and compare biological data. In this research, we have looked at two such techniques: FastA and Blast algorithms [4, 5].

What is FastA?

FastA is a dynamic programming algorithm that compares two sequences to find the best alignment. It finds regions of exact local matches between two sequences and then tries to connect them to get a global alignment.

What is Blast?

Blast stands for "basic local alignment search tool." Blast searches for common words or k-tuples in the selected sequence and each database sequence and then tries to extend them beyond a selected threshold. Both Fasta and Blast are heuristics of Smith-Waterman algorithm.

Significance

There has been an explosion of software tools that allow sequence comparison in a timely fashion based on the above mentioned algorithms. Today, a biologist can submit a new sequence to an online

implementation and, with the click of a few buttons, compare the new sequence with sequences whose functionalities and structures are already known. However, there is consensus among many researchers that there is a lack of data that compares the efficiency and reliability of these implementations [6]. The goal of this research is to compare the implementations of the above mentioned algorithms. Fasta3 (http://www.ebi.ac.uk/fasta33/index.html) and WuBlast2 (http://www.ebi.ac.uk/blast2/index.html) supported by European Bioinformatics Institute were the implementations chosen for this purpose.

Experimental Procedures

In order to facilitate the task of data collection and comparison, a framework was developed that utilized a software tool to perform much of the tedious work. The first step consisted of searching for the alignments of a pre-selected sequence using one of the two alignment programs. In this step, the program compared the selected sequence with sequences in a database and compiled a list of the top alignments that were formed from each comparison. The results were then sent back via e-mail. In the second step, the information in the e-mail was parsed by the software tool into a MS Access database. This procedure was repeated for each selected sequence using FastA and the Blast programs. The objective was to compare which alignments were reported by one program and not the other. For instance, Table 1 has the list of alignments that were missed by one of the programs for a *Bacillus Amyloliquefaciens* protein. The rows with value 0 for 'AppID' column signify that it was missed by WuBlast2 whereas the rows with value of 2 in that column signify that the alignment in that row was missed by Fasta3.

Table 1. *Bacillus Amyloliquefaciens* alignments that were missed by Fasta3 or WuBlast2.

MatchID01	MatchID02	SID	AppID	E	Name	Score
EM_PAT:AX655393	AX655393.1	51	0	50	Sequence 5263 from	314
EM_STS:G66143	G66143.1	51	0	50	sY1154 Miscellaneous Y (	678
EM_PAT:AX463254	AX463254.1	51	2	100	Sequence 11 from Patent WO0250...	241
EM_PAT:AX463246	AX463246.1	51	2	100	Sequence 3 from Patent WO0250257.	241
EM_PAT:AX463252	AX463252.1	51	2	100	Sequence 9 from Patent WO0250257.	241
EM_PRO:SMFTF	M18954.1	51	2	100	S.mutans fructosyltransferase gene,...	248
EM_PRO:AY191311	AY191311.1	51	2	100	Leuconostoc citreum strain CW2...	241
EM_PRO:SSFTFB	L08445.1	51	2	100	Streptococcus salivarius fructosyl...	401

Table 2 lists parameters that were kept as default while using the alignment programs. The Uniprot sequence database was used to compare the selected sequences and Blosum62 scoring matrix was used to score the alignments.

Table 2. Parameters used for comparison.

Parameter	Value
Database	Uniprot
Matrix	Blosum62
Scores	100
Alignments	100
Expected Threshold (WuBlast2)	100
Expected Threshold (Fasta3)	50
Sensitivity	Normal

Results

Table 3 contains the average number of misses that were reported for protein sequences belonging to a specific family. The 'Sequence Size' column relates to the size of the sequences that were used from the family

(corresponding to the family column) and the 'Misses' column relates to the average of total alignments that were missed by either Fasta3 or WuBlast2. It indicates that there is a relationship between the consistency of the programs and the protein family being used. For instance, sequences belonging to the Pepsin family have almost half the number of missed alignments compared to the sequences belonging to the Mycobacterium family.

Table 3. Average number of misses reported for sequences belonging to a specific protein family.

Family	Sequence Size	Misses (Avg. #)
Hemoglobin	142-146	1
H+ transporting ATP Synthase	386	2
Snake Neurotoxin	60-76	4
Pepsin	381-388	12
Mycobacterium FAP Ag85	325-341	26
Nucleoprotein	1760	107

Graph 1 shows the percent identity of the alignments that were missed for the sequences belonging to the *Mycobacterium FAP* family. Most of the missed alignments were in the so-called "grey zone" (25-30% Smith-Waterman identity). According to researchers, anything over 25% Smith-Waterman identity is significant for protein sequences [7]. Also, Fasta implementation missed approximately twice as many alignments as the Blast implementation. It is important to note that these results do not take into account the duplicate entries found in the sequence databases. As a result, the numbers are slightly inflated. Therefore, the numbers were consistent for all the protein families from Table 3.

Graph 1. A look at the homology of the sequences that were missed from *Mycobacterium FAP* family.
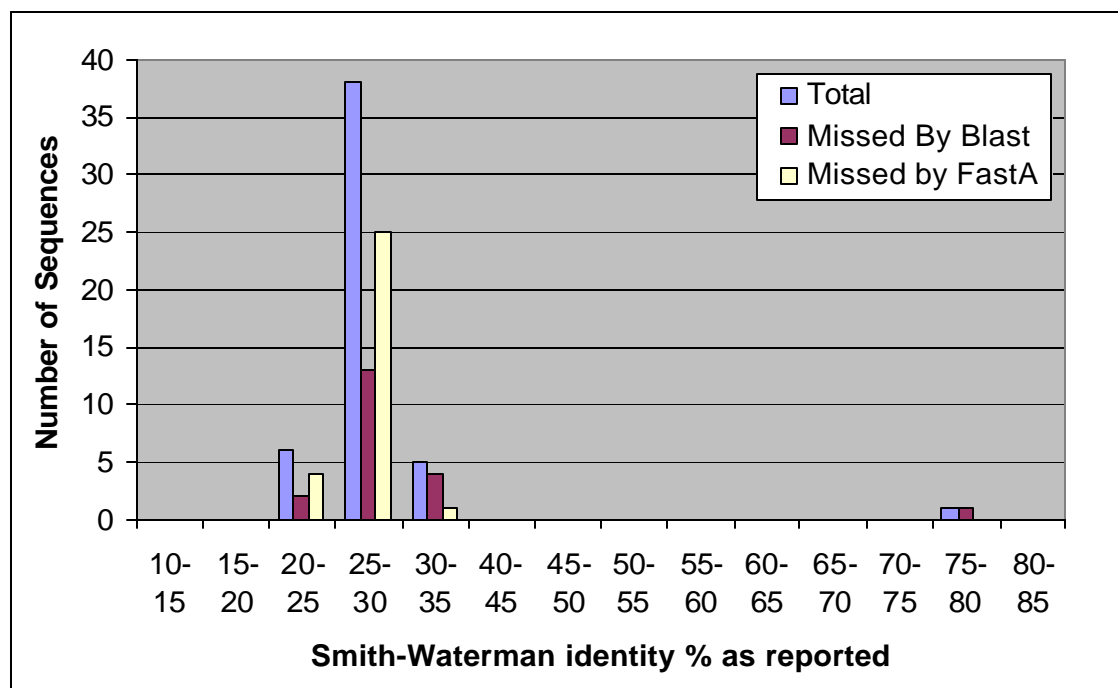
Graph 2 shows the percent identity of the alignments that were missed for the *sacB* DNA sequences. Since, DNA sequences contain four characters; they are bound to be more similar than protein sequences. This explains the higher identity misses. Unlike protein sequences, there is no hard and fast rule related to the significance of the misses based on Smith-Waterman identity. However, the graph does reveal inconsistencies in the results reported by the two programs.

In summary, both programs were highly consistent in reporting top scoring alignments. However, inconsistencies were reported at low scoring alignments that might still be biologically significant.

Graph 2. A look at the homology of the sequences that were missed from *sacB* DNA family
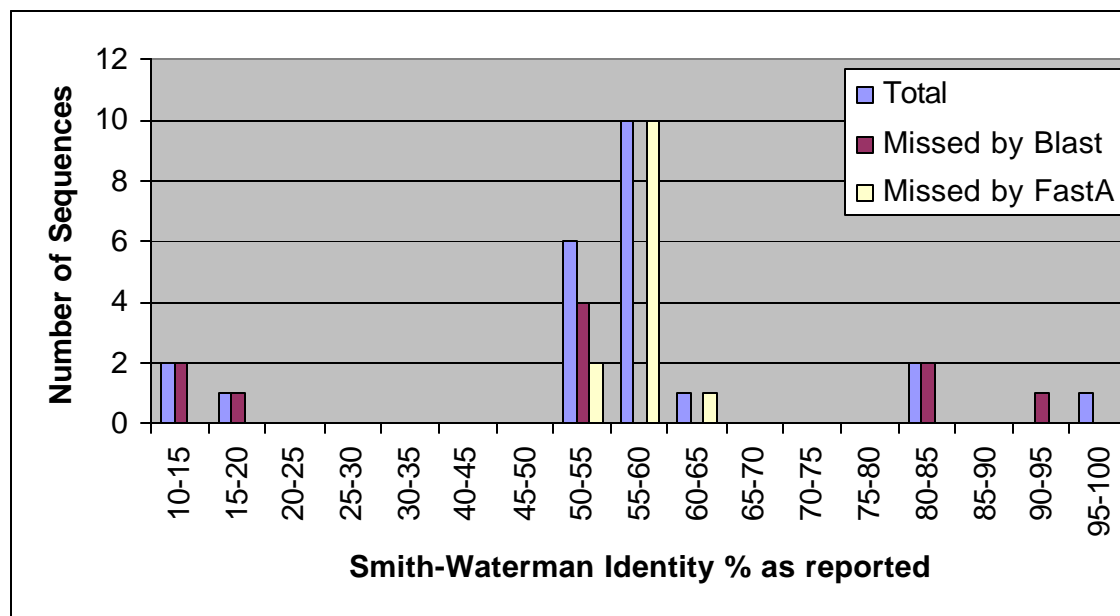
Future Direction

The next step for this research project is to expand the dataset with more sequences belonging to wider range of protein families. This would help discern any patterns that might exist that are not easily identified with a small data set. It would also be interesting to collect data by changing the scoring matrix being used since a scoring matrix is responsible for quantifying the alignments and has its own set of built-in assumptions regarding the sequences.

References

[1] D. Gusfield, <u>Algorithms on Strings, Trees, & Sequences</u>, New York, USA: Cambridge University Press, 1997.

[2] C. Gibas and P. Jambeckm, <u>Developing Bioinformatics Computer Skills</u>, O'Reilly & Associates, 2001.

[3] National Center for Biotechnology Information, "What is GenBank", [Online document], Rev. July 8, 2003,
[cited 2003 Sep. 28], Available: http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html

[4] D. Lipman and W. Pearson, <u>Improved tools for biological sequence comparison</u>, 85 ed. , 2444-2448: Proc. Natl. Academy science, 1988.

[5] W. Pearson, <u>Comparison of methods for searching protein sequence databases</u>, 4 ed. , 1145-1160: Protein Sci., 1995.

[6] J. Thompson, F Plewniak., O. Poch, <u>A comprehensive comparison of multiple sequence alignment programs</u>, Vol. 27, No. 13, p 2683: Nucleic Acids Research, 1999.

[7] B. Rost, <u>Twilight zone of protein sequences</u>, 2 ed., p 85-94: Protein Eng., 1999 Feb.

Author Biography:

Tejas Gandhi is a computer science senior at Minnesota State University, Mankato.

Faculty Mentors Biographies:

Dr. Christophe Veltsos is an assistant professor in the department of Computer and Information Sciences at Minnesota State University, Mankato. He completed his PhD at the University of Southwestern Louisiana (now the University of Louisiana at Lafayette). His research interests are software engineering methodologies and computer science education.

Dr. Timothy Secott is an Assistant Professor of Microbiology at Minnesota State University, Mankato. His research addresses the mechanisms used by pathogenic bacteria to attach to and invade host tissues. He received his Ph.D. from Purdue University. His doctoral thesis was an investigation of fibronectin-mediated attachment and invasion of intestinal epithelial cells by *Mycobacterium avium* subsp. *paratuberculosis.*