The 3rd Saudi International Conference
(SIC-2009)
UNITED KINGDOM- GUILDFORD
5th & 6th of JUNE 2009

Exact String Matching Algorithms for Searching Biological Sequence Databases

Ahmad Fadel Klaib and Hugh Osborne
Informatics Department, University of Huddersfield

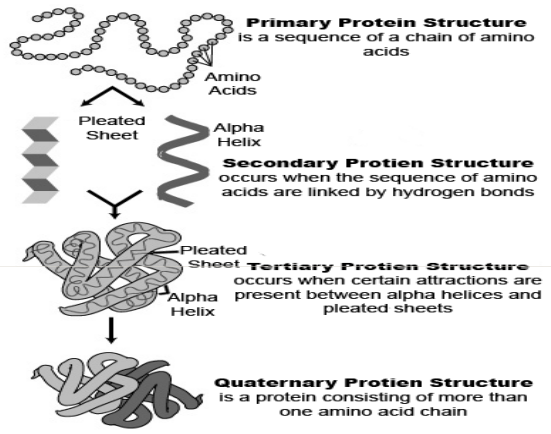
Abstract

Huge amounts of data are stored in linear files. This is also the case for biological sequence data. The quantities of data in these fields tend to increase year on year. For this reason efficient string-matching algorithms should be used which use minimal computer storage and which minimize the searching response time. In this research, new three algorithms which they are BRBMH, BRQS and OE algorithms have been developed, tested and compared with well known string matching algorithms. The experimental results show that the new algorithms are faster and perform fewer numbers of comparisons than other compared algorithms for any length of alphabets and patterns. So they are applicable for searching protein sequence databases as well as in any other string searching applications.

Biological Sequence Databases

There are a lot of distributed public databases over the world with different aims and contents which are designed to integrate many different types of data. There are a lot of biological databases contain the biological data. The SWISS-PROT database considers as one of the main sources for Protein sequences. The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) is Europe's primary nucleotide sequence resource. It is considered to be one of the main sources for DNA and RNA sequences. The Protein Data Bank (PDB) database is a repository for 3-D structural data of proteins and nucleic acids. The RNAdb is a comprehensive mammalian noncoding RNA database (RNAdb) containing sequences and annotations for tens of thousands of noncoding RNAs.

Proteins



String Matching Algorithms

String matching algorithms play a key role in many computer science problems, challenges and in implementation of computer software such as text processing, image and signal processing, information retrieval, speech recognition and analysis, and computational biology and chemistry such as Proteins, DNA and RNA searching. This problem has received a great deal of attention due to various applications in computational biology.

String-matching algorithms aim to find all the occurrences of a given pattern $P = p_1p_2...p_m$ in a text $T = t_1t_2...t_n$. They work as follows: they first align the left ends of the pattern and the text, then compare text characters with pattern characters and after a mismatch between the pattern and the text or a whole match between them they shift the pattern to the right. This procedure is repeated until the right end of the pattern reaches the right end of the text.

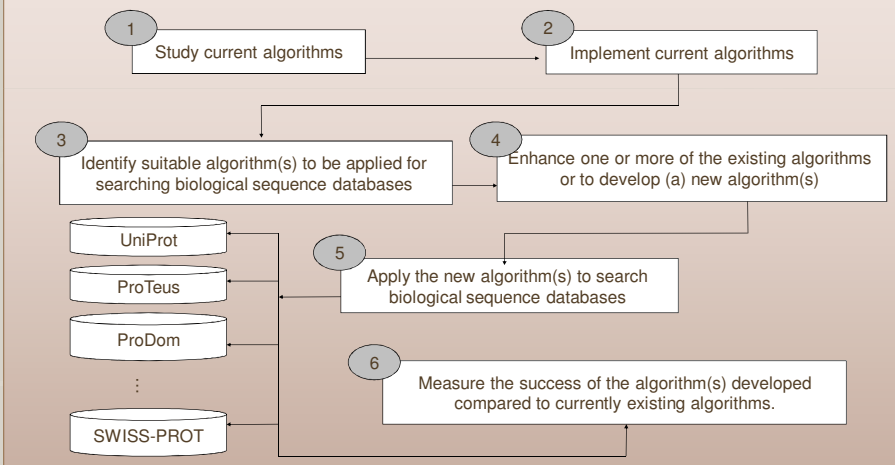
Objectives

In this section, we explain the main parameters and factors that we intend to include in our study. Thus our study includes the following objectives:

- To study existing string-matching algorithms in order to develop a taxonomy of such algorithms.
- To apply insights gained in the previous phase to enhance one or more of the existing algorithms, or to develop (a) new algorithm (s) to search biological sequence databases.
- To measure the success of the algorithm (s) developed compared to currently existing algorithms.

Research Methodology

The research methodology is divided into six phases as denoted by numbers 1, 2, 3, 4, 5 and 6 as shown in the following figure.

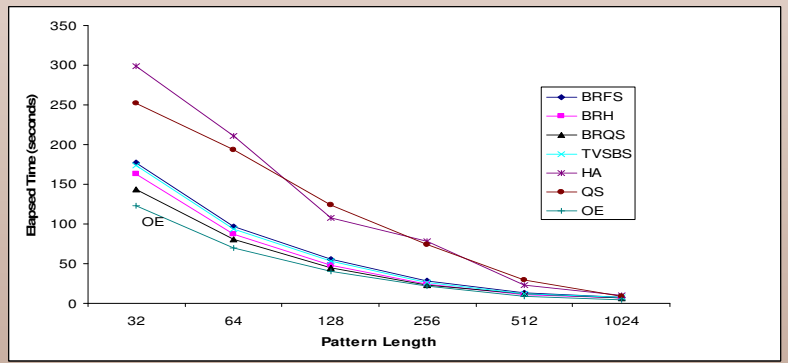


Implementation and Results

A sample file has been taken from the Swiss-Prot database which consists of 8740 proteins to test the efficiency of our new three algorithms which they are BRBMH, BRQS and OE algorithms compared to other well known existing algorithms. The performance of new algorithms has been evaluated using the number of comparison between the pattern and the text and the elapsed time of searching.

The following table shows the number of comparison while the following figure shows the average elapsed time (s.) for searching different length of patterns in the protein sample file.

Pattern Length	OE	BRQS	BRMH	BRFS	TVSBS	QS	BMH
32	**95384**	95498	95595	96356	95682	172936	161089
64	**50973**	51171	51202	52101	51258	133723	113597
128	**26985**	27099	27180	27388	27214	87426	59229
256	**10012**	10040	10058	11925	10075	45394	38005
512	**2950**	2978	2987	3186	2997	16120	8502
1024	**1233**	1235	1239	1282	1243	2647	2186



Conclusion

According to our research methodology, new three string matching algorithms have been developed and the experimental results show that they perform the search with less number of comparisons and faster elapsed searching time between the pattern and the text. Therefore the new algorithms are suitable for searching the proteins in Swiss-Prot database as well as in any other string searching applications.