# SEARCHING MOTIF PATTERN IN A PROTEIN SEQUENCE USING FUZZY PATTERN SEARCHING ALGORITHM

Khine Aye Mon Bo
Computer University (Mandalay)
khineayemonbo@gmail.com

**ABSTRACT**

*The analysis of protein and DNA sequence data has been one of the most active research areas in the field of computational biology. In protein sequence, often two sequences that share similar substring have similar functional properties. Learning the properties and characteristics of an unknown protein is much easier if its function can be predicted by finding the substring already known from other protein sequences. In real world biological applications, most sequences are similar instead of exactly the same, therefore, similar sequences searching process is required. This paper presents the pattern searching algorithm which searches for match text between a pattern and a sequence by using fuzzy logic and calculates the degree of similarity. Any motif patterns can be searched in a sequence and patterns for three protein families used in this system can be searched for extended interval values. This system is implemented by using visual C#.*

## 1. INTRODUCTION

Nowadays, the computer is not only viewed as an efficient machinery that performs numerical computations but also as potentially capable of storing knowledge in a human-like way, and displaying human-like intelligent behavior in reasoning and decision tasks such as information processing, information retrieving, and information exploitation. Fuzzy set methods offer useful tools for handling these tasks, due to their ability to provide a qualitative interface with data and to model graded notions such as uncertainty, preference and similarity, which play a key role in reasoning and decision [3].

Mining of sequence data has many world applications. Biological DNA and protein sequence data are sequence data where data mining technique can be applied. In contrast to ordinary data set, sequence data are dynamic and order dependent [6].

There are two fundamentally different tasks related to identifying new pattern in biological sequences. The first one is called pattern searching. This finds new occurrences of a known pattern. Many consensus sequences are known in biology and it is important to have tools that will allow one to find occurrences of known patterns in new sequences. The second is pattern discovery and it is to find a new pattern in a set of several sequences that are related in some way and expected to contain the pattern [9].

For symbolic sequential data, pattern searching can be considered as either exact searching or similar searching. Searching for a pattern in a text file is a very common operation in many applications ranging from text editors and databases to applications. Most text editors and searching programs do not support searching with errors because of the complexity involved in implementing it. The approximate pattern searching problem described in this paper was to find all substrings in a large text that are similar to a string under some measures of similarity using fuzzy logic. The three most significant feature of algorithm were

(1) searching for a approximate patterns,
(2) searching for records rather than just lines, and
(3) searching for multiple patterns with AND (or OR) logic queries.

The system was also competitive with other tools for exact string matching; it included many

options that made searching more powerful and convenient.

## 2. RELATED WORKS

Jamieson [10] focused on the development of pattern matching routines using recursion. They developed their routines using a methodical process based in analysis of the grammar of the items they are manipulating. Pattern machining was a process by which they searched for words include their length, character content and character order. A simple way of representing a pattern was by using a sequence of letters and special 'wild card' characters. When a letter appears in a pattern, it indicated that very same letter must appear in the same position in matching words. When a wild card appear in a pattern, it indicated that a number of different characters or sequences of characters were acceptable at that position.

Lovis [2] considered the problem of exact string pattern matching using algorithms that did not require any preprocessing.

In Knuth [4], the main idea was to shift the search string by more than one character to the right, whenever a mismatch was predictable. The method needed some preprocessing of the search string, to detect recurring sequences of letters.

The idea of Boyers and Moore [8] was to perform character comparisons from right to left ; if a mismatch occurs, the search string may be shifted up to m position to the right.

The idea Wu and Manber [12] was to scan the database one character at a time, keeping track of the currently matched characters in a clever bit-encoding.

## 3. PATTERN SEARCHING IN PROTEIN SEQUENCE

The Biologists often perform pattern searches due to the fact that sequences that have similar pattern usually have similar functional properties. When an unknown protein is sequenced, the scientist usually tries to get the functional properties of the unknown protein by doing pattern searching. The functionally properties of a protein can be scientifically determined using biological tests, however, the testing time period would be quite long. By doing the pattern search in protein sequence and if some subsequences that are similar to the pattern exist, the scientist would often test for the possibility of similar functional properties for the unknown sequences and hence, fast track the research [5].

Generally, there are two approaches of determining the functional properties of proteins using computer program: sequence alignment and motif pattern searching. Sequence alignment methods have two variations: local alignment and global alignment methods. Local alignment method aims to align two sequences so that the similarity between the regions of the two sequences is maximized. Global alignment methods aim to align two sequences that similarity between two sequences is maximized [5].

Motif pattern searching techniques identify the existence of motifs within an unknown protein sequence. A protein motif pattern is a portion of a sequence, called subsequences, may occur repeatedly in sequences of a protein family. PROSITE is one of the protein motif databases [7]. The protein motif pattern for TGF-Beta family, Leucine Zipper Family, and Integrins Beta Chain Cysteine-rich domain family are used in this system and they are

(1) [LIVM]-x(2)-P-x(2)-[FY]-x(4)-C-x(1)-G-x(1)-C, (2) L-x(6)-L-x(6)-L-x(6)-L and

(3) C-x(1)-[GNQ]-x(1,3)-G-x(1)-C-x(1)-C-x(2)-C-x(1)-C, respectively.

These patterns are obtained from PROSITE database [11]. From a statistically point of view, it is a repeated occurrence of sequential data. There are sequence and structure patterns that can be used to characterize proteins.

Protein sequence described in the language of twenty amino acids [9]. A protein conformation is often described in terms of three structural levels: (1) the primary structure, which is linear sequence of linked amino acids, (2) the secondary structure, which is regularly repeating local structures stabilized by hydrogen bonds, and (3) the tertiary structure, which is the overall shape of a single protein molecule. A protein's sequence which is also known as the primary structure of protein can often be used to predict its secondary structure.

However, the prediction of a protein's tertiary structure is still difficult [1].

In real world biological application, exact pattern do not exist due to the large number of possible combination. It is therefore useful to search subsequences that are similar to the pattern using fuzzy logic. This way, a scientist is able to identify an unknown sequence's functional properties using the past experience and expertise.

## 4. FUZZY PATTERN SEARCHING ALGORITHM

In exact pattern searching problems, which aim to find a substring in text T that is exactly the same as the searching pattern P. in biological sequence data applications, exact patterns are rare, but sequences belonging to the same functional family usually have "similar" substrings within each of the sequences. Hence, there is a requirement of a similar pattern searching system for biological data analysis.

The system aims to find a substring, P', within a text T, that is most similar to a searching pattern, P. A pattern can be interpreted as a series of events, $E_i$, separated by their event intervals, $I_i$. A pattern can be described as:

$$E_1 - I_1 - E_2 - I_2 ... - E_{n-1} - I_{n-1} - E_n$$

The concept of event interval is important when the searching pattern, P, contains wild cards. A wild card, usually represent by letter "x" in molecular biology, can match to any other symbols. The symbol x(i,j) represents the existence of i to j number of wild cards, whereas x(i) means that there are i number of wild cards. For example, the pattern A-x(2)-C-x(1,3)-G has three events A, C and G. The event interval between A and C is 2, and the event interval between C and G is 1 to 3 (i.e. $I_1 = 2; I_2 = 1,2,3;$).

There are four main steps in the system. Firstly, a searching pattern, P is decomposed to obtain events and event intervals. Then, the obtained events and event intervals are fuzzified in the pattern fuzzification step. A pattern inference step follows to determine the patterns, P', that are similar to the searching pattern P. Finally, a pattern search is conducted to determine the similarity between a text T and the pattern P..

### 4.1. Pattern decomposition

In this step, the searching pattern P is decomposed to obtain events and event intervals. An event can consist of one or more symbols/characters. As an example, a pattern A-[LF]-G can match 3-characters subsequences starting with A, ending with G and having either L or F in the middle and [LF] is called an ambiguous character.

### 4.2. Pattern fuzzification

The searching pattern P is fuzzified by applying fuzzification technique to the events and event intervals obtained from the previous step. In the fuzzification step, fuzzy membership function of events and event intervals are generated. The assignment of the fuzzy membership functions is depending on the requirement of the specific task or expert knowledge. An event can be fuzzified based on its content, or character(s) presented. The event contents for single character are not fuzzified in this system. For the ambiguous character, all characters inside the square bracket are given a membership degree of 1. The event intervals can be fuzzified according to the number of wild cards contain in the searching pattern.

### 4.3. Pattern inference

This step generates an array of sequences P' that are similar to the searching pattern P. The degree of similarity is determined by the fuzzy rule:

$I_i$: IF event $E_1$ occurs and event $E_2$ occurs AND event interval between $E_1$ and $E_2$ is $I_1$... AND event $E_{n-1}$ occurs AND event $E_n$ occurs and event interval between $E_{n-1}$ AND $E_n$ is $I_{n-1}$, THEN Pattern $P'_i$ is similar to $P$ with Degree $Y_i$.

Where

$Y_i$ =T-norm$(\mu(E_1), \mu(E_2), ..., \mu(E_n), \mu(I_1), \mu(I_2), ..., \mu(I_{n-1}))$

The T-norm used for multiplication in this system and the choice of functions will depend on the need of specific applications.

### 4.4. Pattern searching

The array of similar sequences $P'$ obtained from the previous step is then used for the determination of similarity between a sequence text $T$ and a searching pattern $P$. Each $P'_i$ is compared with sequence $T$ as an exact searching problem and if $P'_i$ exists in $T$, then the similarity between $P$ and $T$ is $Y_i$. Since the sequence $T$ can match to many of the sequences in $P'$, the similarity between $T$ and searching pattern $P$ is determined as:

$$Y = F(Y_i), \text{ for } P'_i \text{ exist in } T,$$

and the function $F$ is Maximum.

## 5. SYSTEM DESIGN

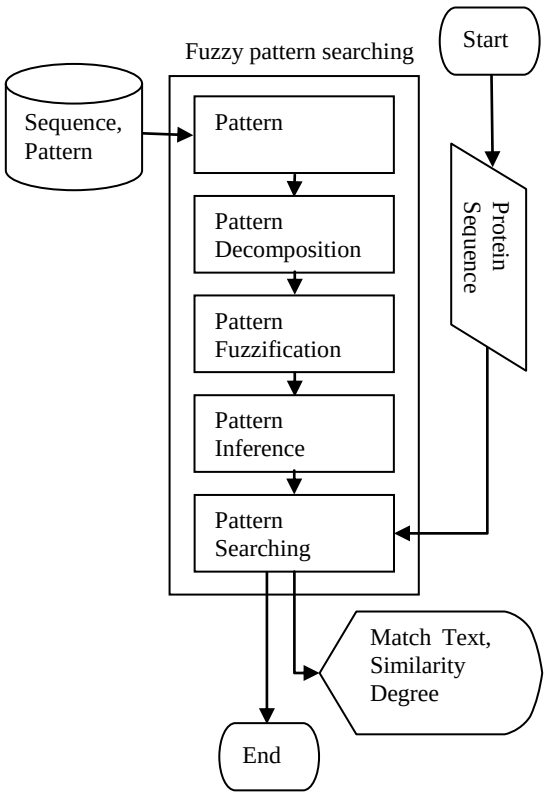


**Figure 1.** System flow diagram

The system starts when the user input protein sequence and click search button. The pattern is decomposed to obtain events and event intervals. In the fuzzification step, fuzzy membership functions of events and event intervals are generated. In the pattern inference step, the system generates an array of sequences that are similar to the searching motif pattern. The array of similar sequences is used in pattern searching process. Then, the system produces the degree of similarity between the motif patterns and protein sequence, and a portion of input sequence that are similar to the pattern.

From the similarity degree, the user can predict the function of input protein sequence. The system can search any motif patterns that are formatted to the motif form. Furthermore, the system can increase or decrease the interval values of three motif patterns used in this system with 1 and search these motif patterns with extended interval values.

## 6. IMPLEMENTATION

In Protein sequences data set, user can see protein sequences data set with Protein Sequence, Organism in which the protein sequence exist, and family name of protein sequence.

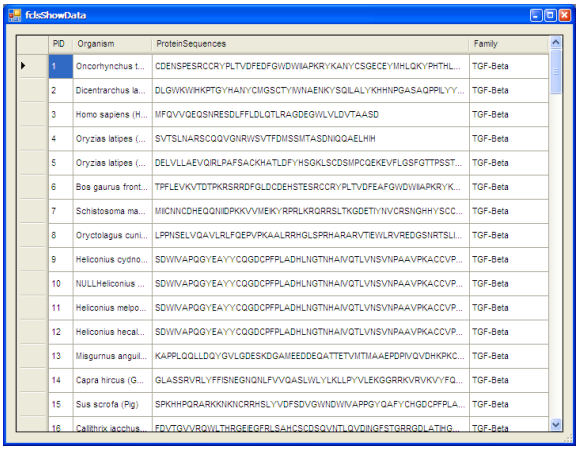


**Figure 2.** Protein sequences data set

The user can search one of three motif patterns with extended interval values. The system show match text and similarity degree and these form is illustrated in figure3.

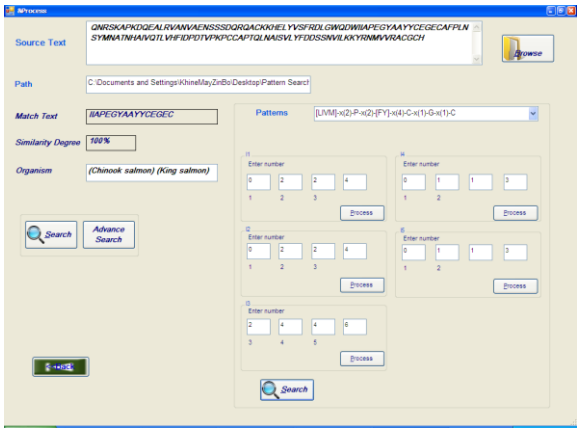


**Figure 3.** Searching motif pattern process with extended interval values

## 7. CONCLUSION

The system applies the fuzzy pattern searching algorithm to search the three known motif patterns in a protein sequence: TGF-Beta family, Integrins Beta Chain Cysteine-rich Domain family and Leucine Zipper family. The system can search pattern with variable length wild cards and the system can also search the motif pattern with extended interval values. The system calculates the similarity degree between one of three motif patterns and protein sequence, and displays the match motif pattern and a portion of match sequence. By searching motif patterns, the user can know the family and can predict the function of a new protein.

## REFERENCES

[1] Bill C.H. Chang and Saman K. Halgamuge, "Approximate Symbolic Pattern Matching for Protein Sequence Data", University of Melbourne, Australia, 2001.

[2] Lovis, MD and R.H. Boud, "Fast Exact String Pattern-matching Algorithm Adapted to the Characteristics of the Medical Language", J Am Med Inform Assoc, pp. 378-391, 2000.

[3] Didier Dubois and Henri Prade, "Perspectives of Fuzzy Systems", Institude of Research and Information Technology, 2000

[4] D.E. Knuth, J.H. Morris and V.R. Pratt, "Fast Pattern Matching in String", SIAMJ.Comput, Vol. 6, pp. 323-350, June 1977.

[5] D. Gusfield, "Algorithm on Strings, Trees, and Sequences", Cambridge University Press, 1997.

[6] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Elsevier, 2006.

[7] K. Hoffman, P. Bucher, L. Falquet and A. Bairoch, "The PROSITE Database, its status in 1999", Nucleric Acids Research, 1999.

[8] R.S. Boyer and J.S. Moore, "A first string searching algorithm", CACM, Vol. 20, pp. 762-772, October 1977.

[9] Stephen A. Krawetz and David D. Womble, "Introduction to Bioinformatics: A Theoretical and Practical Approach", Human Press, 2003.

[10] S. Jameison, "Pattern Matching Part I", March 3, 2003.

[11] http://au.expasy.org/sprot/, Swiss-Port Protein Database.

[12] S. Wu and U. Manber, "Agrep- A Fast Approximates Pattern-matching Tool", in Proceedings of the winter 1992, USENIX Conference: January, 1992, pp-153.