International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 26 Issue 5 – AUGUST 2019.

Rule Based Motif Pattern Searching In A Protein Sequence

Thi Thi Soe[#1], Zarni Sann[*2], Htwe Htwe Pyone[#3], Hnin Yu Yu Win[#4]

[#]*Faculty of Computer Science*
[*]*Faculty of Computer Systems and Technologies*
[#1, *2]*University of Computer Studies (Mandalay), Myanmar*
[#3]*University of Computer Studies (Myitkyina), Myanmar*
[#4]*University of Computer Studies (Taungoo), Myanmar*

***Abstract*— In the field of computational biology, protein and DNA sequence data analysis is one of the most active research areas. When two protein sequences that have similar substring, their functional properties are also similar. Therefore, similar sequences searching process is required in biological applications. This paper presents the pattern searching algorithm which searches by matching a pattern and a sequence using the fuzzy logic based generated rules. In this way, a motif pattern can be searched in a sequence. As a case study, three protein families are used in this system, the families are Leucine Zipper family, TGF-Beta family, and Integrins Beta Chain Cysteine-rich Domain family. This motif pattern searching system is implemented by using Visual C#.**

***Index Terms*—Computational biology, Motif pattern, Protein and DNA sequence,**

I. Introduction

Biologists often perform pattern searches due to the fact that sequences, which have similar pattern usually, have similar functional properties. There are two approaches related to motif pattern in proteins: motif pattern searching and motif pattern discovery. Motif pattern searching technique identifies the existence of motifs within an unknown protein sequence. A protein motif pattern is a short sequence that is found within sequences of a same protein family [1]. It is therefore useful to search sequences using fuzzy set theory. Discrete fuzzy sets for fuzzy expert systems may be numeric or non-numeric. Members of a numeric discrete fuzzy set always describe a numeric quantity. Such discrete fuzzy sets are called linguistic variables, with member linguistic terms [2]. In this study, the motif pattern of three families; Leucine Zipper family, TGF-Beta family, and Integrins Beta Chain Cysteine-rich Domain family are used to search the pattern. The motif patterns for these families are obtained from PROSITE [3] protein motif database. This work intends to realize the basic understanding and knowledge about bioinformatics and to know the family of an unknown protein sequence. The remainder of the paper is organized as follows. Section II explores how to search the motif pattern using fuzzy rule based approach. In section III, we test the motif pattern searching system by applying the data of three families as a case study. Finally, we conclude the paper in section IV and further extension is pointed out.

II. Rule Based Motif Pattern Searching

Protein motif patterns can contain ambiguous characters, wildcards and flexible gaps. Three protein families used in this system and corresponding motif patterns are described in the following section.

A. *Leucine Zipper Family*

The leucine zipper consists of a periodic repetition of leucine residues at every seventh position over a distance covering eight helical turns. The localization of the leucines are critical for the DNA binding to the proteins. Leucine zippers are present in both eukaryotic and prokaryotic regulatory proteins, but are mainly a feature of eukaryotes. They can also be annotated simply as ZIPs, and ZIP-like motifs have been found in proteins other than transcription factors and are thought to be one of the general protein modules for protein–protein interactions [4]. The motif pattern for leucine zipper family is L – x(6) – L – x(6) – L – x(6) – L in [5].

B. *TGF-Beta Family*

Transforming growth factor beta (TGF-β) is a protein that controls proliferation, cellular differentiation, embryonal development, hormone secretion and other functions in most cells. It plays a role in immunity, cancer, heart disease, diabetes, and Marfan syndrome. TGF-β acts as an antiproliferative factor in normal epithelial cells and at early stages of oncogenesis [6]. The motif pattern for TGF-beta family is [LIVM]-x(2)-P-x(2)-[FY]-x(4)-C-x-G-x-C in [7].

C. *Integrins Beta Chain Cysteine-Rich Domain Family*

Integrins are receptors that mediate attachment between a cell and the tissues surrounding it, which may be other cells or the extracellular matrix (ECM) [8]. The motif pattern for Integrins beta chain cysteine-rich domain signature is C-x-[GNQ]-x(1,3)-G-x-C-x-C- x(2)-C-x-C in [9].

D. *Fuzzy Pattern Searching*

In motif pattern searching, we follow the idea of fuzzy pattern searching algorithm [10].The algorithm aims to find a substring, P', within a text, T, that is “most similar” to a searching pattern, P. A sequence data can be interpreted as a series of events, E_i, separated by their event intervals, I_{ij}.

- Separation of the pattern

The searching pattern P is separated into events and event intervals.

Pattern decomposition for P = C - x - [GNQ] - x (1,3) - G - x- C -x- C - x(2) - C - x- C

Events:

$E_1 = C; E_2 = \text{[ENQ]}; E_3 = G; E_4 = C; E_5 = C; E_6 = C; E_7 = C$

Event Intervals:

$I_1 = 1;\ I_{2,} = 1{,}2{,}3;\ \ I_3 = 1;\ \ I_4 = 1;\ I_5 = 2;\ \ I_6 = 1$

Pattern decomposition for

P = [LIVM]-x(2)-P-x(2)-[FY]-x(4)-C-x(1)-G-x(1)-C

Events:

$E_1 = \text{[LIVM]};\ \ E_2 = P;\ \ E_3 = [FY];\ \ E_4 = C;\ E_5 = G; E_6 = C$

Event Intervals:

$I_1 = 2;\ I_2 = 2;\ \ I_3 = 4;\ \ I_4 = 1$

Pattern decomposition for P = L - x(6) - L - x(6) - L - x(6) – L

Events: $E_1 = L;\ \ E_2 = L;\ \ E_3 = L;\ \ E_4 = L$

Event Intervals: $I_1 = 6;\ I_2 = 6;\ \ I_3 = 6$

- Fuzzification of the Pattern

The searching pattern P is fuzzified by applying fuzzification techniques to the events and event intervals. The events for single character symbols are not fuzzified. Following figure represents the fuzzy membership functions for Integrins Beta Chain Cysteine-rich Domain signature, TBF-Beta protein family and Leucine Zipper family, respectively.

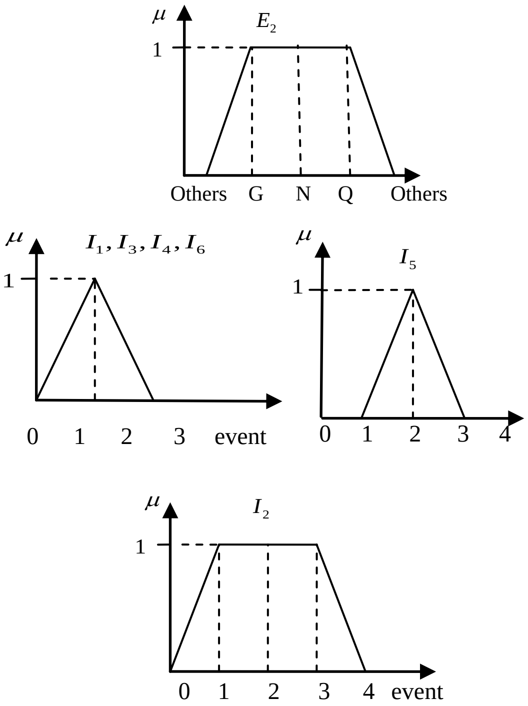
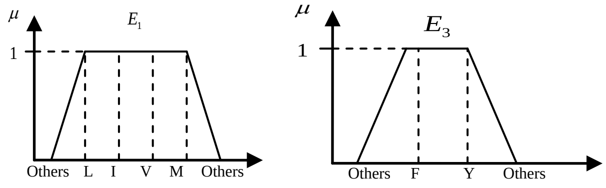


Fig 1. Membership Functions for Integrins Beta Chain Cysteine-Rich



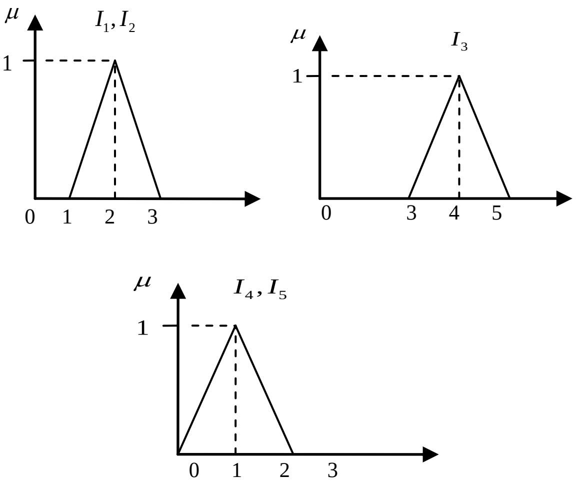


Fig 2. Membership Functions for TBF-Beta Family

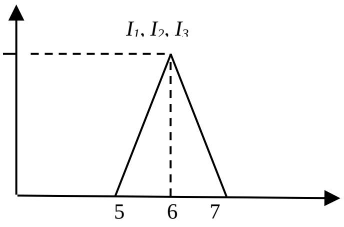


Fig 3. Membership Functions for Leucine Zipper Family

- Inference Rule

In the sequence inference step, the most common way to represent human knowledge is IF-THEN rule form expressed as [10]:

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 input pattern P with degree Y_i, where, $Y_i = \mu(E_1) * \mu(E_2) * \ldots * \mu(E_n) * \mu(I_1) * \mu(I_2) * \ldots * \mu(I_{n-1})$.

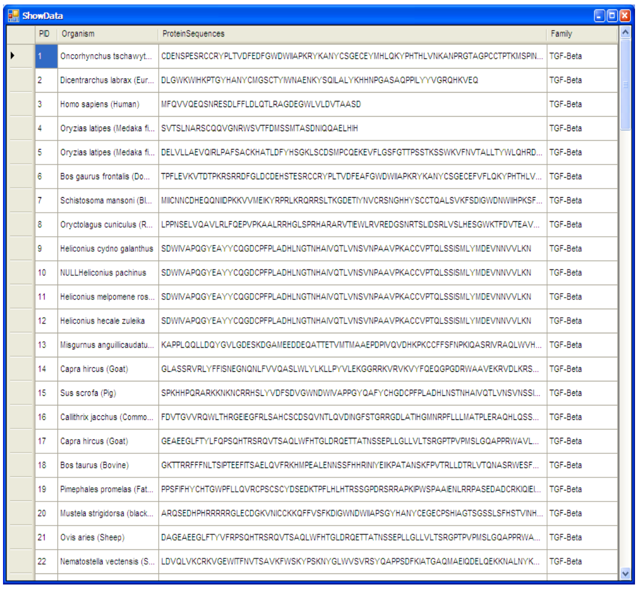
The process of obtaining the overall consequent (conclusion) from the individual consequents contributed by each rule in the rule-based is known as aggregation of rules. For determining an aggregation strategy, two simple cases are Max-Min Inference method and Max-Product Inference method.

- Searching the Pattern

The array of similar sequences $P_i^{'}$ obtained from the previous step is used for the determination of similarity between a protein sequence and a search pattern P. Each $P_i^{'}$ is compared with protein sequence as an exact matching. If $P_i^{'}$exists in protein sequence, then similarity between P and protein sequence isY_i. A protein sequence can match to many of the sequences in $P_i^{'}$. The similarity between protein sequence and searching pattern P is determined as: $Y = max(Y_i)$ for $P_i^{'}$ exists in protein sequence

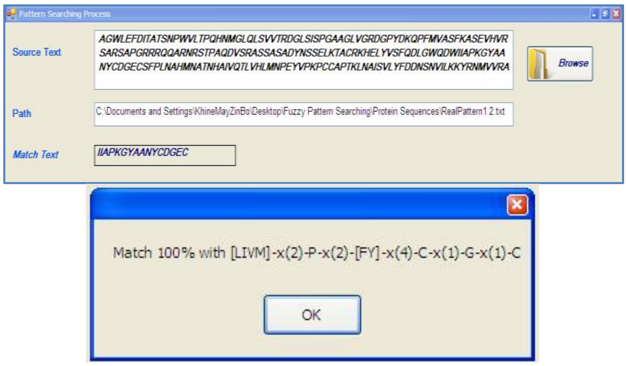
III. Initial Testing

The data set for known protein sequences of three protein families are used in this system. Example data is shown in fig 4. The system searches the motif patterns in an input protein sequence. And, the system shows the match percentage and a portion of match sequence if input sequence is matched with one of three motif patterns. If there is no match motif pattern, the system shows the message box that there is no match pattern. Searched result for searching motif pattern with exact match is illustrated in fig 5.



PID	Organism	ProteinSequences	Family
1	Oncorhynchus tschawyt...	CDENSPESRCCRYPLTVDFEDFGWDWIIAPKRYKANYCSGECEYMHLQKYPHTHLVNKANPRGTAGPCCTPTKMSPIN...	TGF-Beta
2	Dicentrarchus labrax (Eur...	DLGWKWIHKPTGYHANYCMGSCTYIWNAENKYSQILALYKHHNPGASAQPPILYYVGRQHKVEQ	TGF-Beta
3	Homo sapiens (Human)	MFQVVQEQSNRESDLFFLDLQTLRAGDEGWLVLDVTAASD	TGF-Beta
4	Oryzias latipes (Medaka fi...	SVTSLNARSCQQVGNRWSVTFDMSSMTASDNIQQAELHIH	TGF-Beta
5	Oryzias latipes (Medaka fi...	DELVLLAEVQIRLPAFSACKHATLDFYHSGKLSCDSMPCQEKEVFLGSFGTTPSSTKSSWKVFNVTALLTYWLQHRD...	TGF-Beta
6	Bos gaurus frontalis (Do...	TPFLEVKVTDTPKRSRRDFGLDCDEHSTESRCCRYPLTVDFEAFGWDWIIAPKRYKANYCSGECEFVFLQKYPHTHLV...	TGF-Beta
7	Schistosoma mansoni (Bl...	MIICNNCDHEQQNIIDPKKVVMEIKYRPRLKRQRRSLTKGDETIYNVCRSNGHHYSCCTQALSVKFSDIGWDNWIIHPKSF...	TGF-Beta
8	Oryctolagus cuniculus (R...	LPPNSELVQAVLRLFQEPVPKAALRRHGLSPRHARARVTIEWLRVREDGSNRTSLIDSRLVSLHESGWKTFDVTEAV...	TGF-Beta
9	Heliconius cydno galanthus	SDWIVAPQGYEAYYCQGDCPFPLADHLNGTNHAIVQTLVNSVNPAAVPKACCVPTQLSSISMLYMDEVNNVVLKN	TGF-Beta
10	NULLHeliconius pachinus	SDWIVAPQGYEAYYCQGDCPFPLADHLNGTNHAIVQTLVNSVNPAAVPKACCVPTQLSSISMLYMDEVNNVVLKN	TGF-Beta
11	Heliconius melpomene ros...	SDWIVAPQGYEAYYCQGDCPFPLADHLNGTNHAIVQTLVNSVNPAAVPKACCVPTQLSSISMLYMDEVNNVVLKN	TGF-Beta
12	Heliconius hecale zuleika	SDWIVAPQGYEAYYCQGDCPFPLADHLNGTNHAIVQTLVNSVNPAAVPKACCVPTQLSSISMLYMDEVNNVVLKN	TGF-Beta
13	Misgurnus anguillicaudatu...	KAPPLQQLLDQYGVLGDESKDGAMEEDDEQATTETVMTMAAEPDPIVQVDHKPKCCFFSFNPKIQASRIVRAQLWVH...	TGF-Beta
14	Capra hircus (Goat)	GLASSRVRLYFFISNEGNQNLFVVQASLWLYLKLLPYVLEKGGRRKVRVKVYFQEQGPGDRWAAVEKRVDLKRS...	TGF-Beta
15	Sus scrofa (Pig)	SPKHHPQRARKKNKNCRRHSLYVDFSDVGWNDWIVAPPGYQAFYCHGDCPFPLADHLNSTNHAIVQTLVNSVNSSI...	TGF-Beta
16	Callithrix jacchus (Commo...	FDVTGVVRQWLTHRGEIEGFRLSAHCSCDSQVNTLQVDINGFSTGRRGDLATIHGMNRPFLLLMATPLERAQHLQSS...	TGF-Beta
17	Capra hircus (Goat)	GEAEEGLFTYLFQPSQHTRSRQVTSAQLWFHTGLDRQETTATNSSEPLLGLLVLTSRGPTPVPMSLGQAPPRWAVL...	TGF-Beta
18	Bos taurus (Bovine)	GKTTRRFFFNLTSIPTEEFITSAELQVFRKHMPEALENNSSFHHRINIYEIIKPATANSKFPVTRLLDTRLVTQNASRWESF...	TGF-Beta
19	Pimephales promelas (Fat...	PPSFIFHYCHTGWPFLLQVRCPSCSCYDSEDKTPFLHLHTRSSGPDRSRRAPKIPWSPAAIENLRRPASEDADCRKIQIEI...	TGF-Beta
20	Mustela strigidorsa (black...	ARQSEDHPHRRRRRGLECDGKVNICCKKQFFVSFKDIGWNDWIIAPSGYHANYCEGECPSHIAGTSGSSLSFHSTVINH...	TGF-Beta
21	Ovis aries (Sheep)	DAGEAEEGLFTYVFRPSQHTRSRQVTSAQLWFHTGLDRQETTATNSSEPLLGLLVLTSRGPTPVPMSLGQAPPRWA...	TGF-Beta
22	Nematostella vectensis (S...	LDVQLVKCRKVGEWITFNVTSAVKFWSKYPSKNYGLWVSVRSYQAPPSDFKIATGAQMAEIQDELQEKKNALNYK...	TGF-Beta

Fig 4. Data for TGF-Beta Protein Family



Source Text: AGWLEFDITATSNPWVLTPQHNMGLQLSVVTRDGLSISPGAAGLVGRDGPYDKQPFMVASFKASEVHVR SARSAPGRRRQQARNRSTPAQDVSRASSASADYNSSELKTACRKHELYVSFQDLGWQDWIIAPKGYAA NYCDGECSFPLNAHMNATNHAIVQTLVHLMNPEYVPKPCCAPTKLNAISVLYFDDNSNVILKKYRNMVVRA

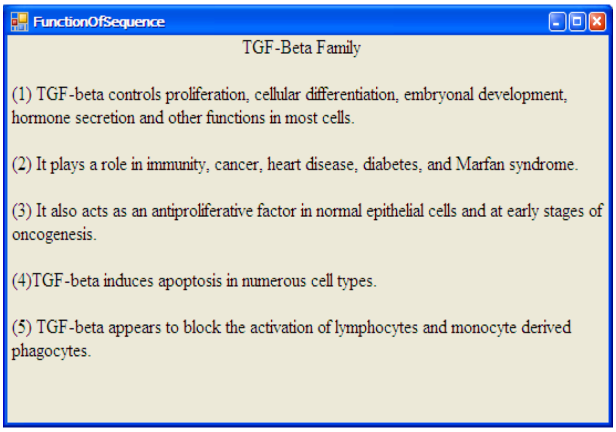
Path: C:\Documents and Settings\KhineMayZinBo\Desktop\Fuzzy Pattern Searching\Protein Sequences\RealPattern1.2.txt

Match Text: IIAPKGYAANYCDGEC

Match 100% with [LIVM]-x(2)-P-x(2)-[FY]-x(4)-C-x(1)-G-x(1)-C

Fig 5. Searching Motif Pattern with Exact Search

After finding out the pattern in the input sequence, the user can get the information about the function of the input protein sequence. The protein sequence function for TGF-beta family is shown in fig 6.



TGF-Beta Family

(1) TGF-beta controls proliferation, cellular differentiation, embryonal development, hormone secretion and other functions in most cells.

(2) It plays a role in immunity, cancer, heart disease, diabetes, and Marfan syndrome.

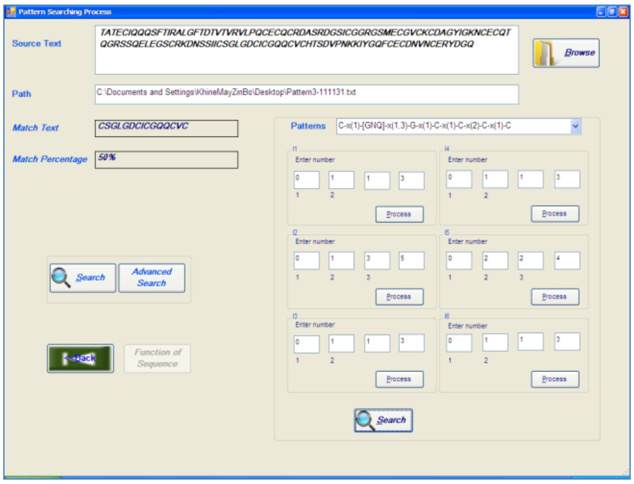
(3) It also acts as an antiproliferative factor in normal epithelial cells and at early stages of oncogenesis.

(4)TGF-beta induces apoptosis in numerous cell types.

(5) TGF-beta appears to block the activation of lymphocytes and monocyte derived phagocytes.

Fig 6. Function for TGF-Beta Protein Family

At advanced search, the user can search one of three motif patterns with the extensible interval values. For TGF-beta pattern, the user can extend the five interval values. For Leucine Zipper pattern, the user can extend the three interval values. For Integrins Beta Chain Cysteine-Rich domain pattern, the user can extend the six interval values. The system shows the match percentage and a portion of match sequence if input sequence is matched with the pattern of extended interval values. If the match percentage is 50%, the sequence may have varied in one interval value with exact motif pattern. The above search is shown as an example in fig 7.



Source Text: TATECIQQQSFTIRALGFTDTVTVRVLPQCECQCRDASRDGSICGGRGSMECGVCKCDAGYIGKNCECQT QGRSSQELEGSCRKDNSSIICSGLGDCICGQQCVCHTSDVPNKKIYGQFCECDNVNCERYDGQ

Path: C:\Documents and Settings\KhineMayZinBo\Desktop\Pattern3-111131.txt

Match Text: CSGLGDCICGQQCVC

Match Percentage: 50%

Patterns: C-x(1)-[GNQ]-x(1,3)-G-x(1)-C-x(1)-C-x(2)-C-x(1)-C

I1 Enter number: 0, 1, 1, 3

I4 Enter number: 0, 1, 1, 3

I2 Enter number: 0, 1, 3, 5

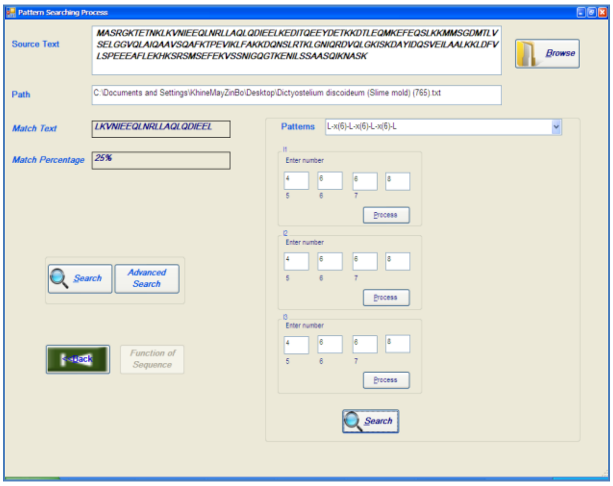
I5 Enter number: 0, 2, 2, 4

I3 Enter number: 0, 1, 1, 3

I6 Enter number: 0, 1, 1, 3

Fig 7. Searching Motif Pattern with Advanced Search

If the match percentage is 25% at advanced search, the sequence may have changes in two intervals value with exact pattern. The above course can be seen as an example in fig 8. But the match percentage is below 25%, the sequence is impossible to perform the protein function of the pattern. By advanced search, the user can detect the sequences that already experimentally identified as the proteins of one family but not detected by exact motif pattern.



Source Text: MASRGKTETNKLKVNIEEQLNRLLAQLQDIEELKEDITQEEYDETKKDTLEQMKEFEQSLKKMMSGDMTLV SELGGVQLAIQAAVSQAFKTPEVIKLFAKKDQNSLRTKLGNIQRDVQLGKISKDAYIDQSVEILAALKKLDFV LSPEEEAFLEKHKSRSMSEFEKVSSNIGQGTKENILSSAASQIKNASK

Path: C:\Documents and Settings\KhineMayZinBo\Desktop\Dictyostelium discoideum (Slime mold) (765).txt

Match Text: LKVNIEEQLNRLLAQLQDIEEL

Match Percentage: 25%

Patterns: L-x(6)-L-x(6)-L-x(6)-L

I1 Enter number: 4, 6, 6, 8

I2 Enter number: 4, 6, 6, 8

I3 Enter number: 4, 6, 6, 8

Fig 8. Advanced Search Match with 25%

IV. CONCLUSION

The system applied the rule based pattern searching algorithm to search the three motif patterns in a protein sequence: TGF-Beta family, Integrins Beta Chain Cysteine-Rich Domain family and Leucine Zipper family. The system not only search the motif pattern with variable length wild cards but also search the motif pattern with extended interval values. The searching system derived the similarity degree for three motif patterns. By searching motif patterns, the user can know the family and can predict the function of a new protein. An extension of the system is to search for motif patterns in a protein sequence of another family.

REFERENCES

[1] P. Bork, E. Koonin, Protein sequence motifs, Curr. Opin. Struct. Biol. 6 (1996) 366–376.

[2] William Siler and James J. Buckley “Fuzzy Expert Systems and Fuzzy Reasoning”, ISBN 0-471-38859-9, 2005.

[3] https//prosite.expasy.org/cgi-bin/prosite/prosite-list.pl

[4] Hakoshima, T. (2005). “Leucine Zippers". Encyclopedia of Life Sciences”. doi:10.1038/npg.els.0005049. ISBN 0470016175.

[5] https//prosite.expasy.org/PDOC00029

[6] Herpin A, Lelong C, Favrel P (May 2004). “Transforming growth factor-beta-related proteins: an ancestral and widespread superfamily of cytokines in metazoans”. Dev. Comp. Immunol. 28 (5): 461–85. doi:10.1016/j.dci.2003.09.007. PMID 15062644

[7] https//prosite.expasy.org/PDOC00223

[8] Giancotti FG, Ruoslahti E (August 1999). "Integrin signaling". Science. 285 (5430): 1028–32. doi:10.1126/science.285.5430.1028. PMID 10446041

[9] https//prosite.expasy.org/ PDOC00216

[10] Bill C.H. Chang and Saman K. Halgamuge, “Approximate symbolic pattern matching for protein sequence data”, International Journal of Approximate Reasoning 32(2-3): 171-186, February 2003 DOI:10.1016/S0888-613X(02)00082-8