A Fangled Greedy Approach For Protein Pattern Matching

Ramya.V* VidyaPriya.V*
* *Assistant Professor, Department of Computer Science, Meenakshi College for women, Chennai-24, India.*
**Assistant Professor,Department of ComputerScience, Quaid-e-Millath College for women, Chennai-2, India.*
E-mail: * ramyars89@gmail.com
*vdy_priya@yahoo.co.in

ABSTRACT*—Molecular biologists often search for the important information from the databases in different directions of different uses Pattern Matching an automated data analysis technique, usually performed on a computer, by which a group of characteristic properties of an unknown object is compared with the comparable groups of characteristics of a set of known objects, to discover the identity or proper classification of the unknown object. The sequence pattern match algorithm proposed in this paper searches for matches between patterns. With the increasing need for instant information, pattern matching will continue to grow and change as needed from time to time. To extract pattern from a large sequence it takes more time, in order to reduce searching time an approach is proposed that reduces the search time with accurate retrieval of the matched pattern from the given sequence of any size, the greedy method is used here to perform the pattern matching.*

KEYWORDS**:** *Pattern Matching, DNA and Protein Sequences, comparison per character*

I. INTRODUCTION

Proteins are a chain of amino acids which are the main parameter for pattern matching. There are about 20 common amino acids. They share a common structure except for one chemical group (R, side chain) attached to the central carbon atom [1]. The amino acids are taken to form a sequence called the protein sequence. The proposed algorithm takes the sequence and the pattern as input, matches the pattern with the sequence. Greedy method is used for performing matching of the given pattern.

II.PATTERN MATCHING

Pattern matching is the problem of finding a subsequence with some property in a sequence of symbols, the simplest case being finding a given string inside the sequence. This is one of the oldest and most pervasive problems in computer science. Applications requiring some form of string matching can be found virtually everywhere. However, recent years have witnessed a dramatic increase in interest in string matching problems, particularly from the rapidly growing communities of information retrieval and computational biology. Not only are these communities facing a drastic increase in the text sizes they have to manage, but they are demanding more and more sophisticated searches. The types of patterns of interest are not just simple strings but also include wild cards, gaps and regular expressions. One of the major problems in genomic field is to perform pattern comparison on protein sequences.

III.A FANGLED GREEDY APPROACH FOR PROTEIN PATTERN MATCHING (FGAPM)

A greedy algorithm is that follows the problem solving heuristic of making the locally optimal choice at each stage with the hope of finding a global optimum [6]. The most common approach is to improve efficiency is the idea of using the greedy method for matching of the pattern to the sequence. The method gets the sequence and the input from the user and stores each in different arrays and calculates the length of the pattern and stores it in a variable .The sequence, the pattern, pattern length and the starting value are inputs for processing of the retrieval of protein pattern. Now the method extracts the protein data length as the control has to drive through the entire sequence for finding the matching pattern, its position and the number of occurrence .Then once the start position and all necessary data are given the greedy method starts the matching process, it gets the first set of characters and stores in a variable likewise the process done for the entire sequence, for every time the identified pattern is returned and its position in the pattern along with the no of times it occurs in the pattern are retrieved.

ALGORITHM:

Step 1: Get Input String As Pattern (IP)

Step 2: Get Protein Data Source As (PD)

Step 3: Get IPlength as Input Pattern Length

Step 4: Store: Step 3 Result->IPLENGTH

Step 5: Greedy (IP,PD,IPLENGTH,STARTVAL){

Step 6: Extract Data from PD[IPLENGTH]

Step 7: Store :Step 6 Result -> PARTDATA

Step 8: Compute Length[PD]

Step 9: Store Step 8 Result-> PDLENGTH

Step 10: Loop (PDLENGTH !=0) do

Step 11: if'(Compare(IP,PARTDATA,IPLENGTH)) then

Step 12: FIND(START,END)-

>STORE :STARTPOS,ENDPOS

Step 13: Store :Output As

IPATTERN[IP,STARTPOS,ENDPOS]

Step 14 :Greedy(IP,PD,IPLENGTH,STARTVAL+ENDPOS)

Step 15 : End Loop}

Step 16 :Return IPATTERN

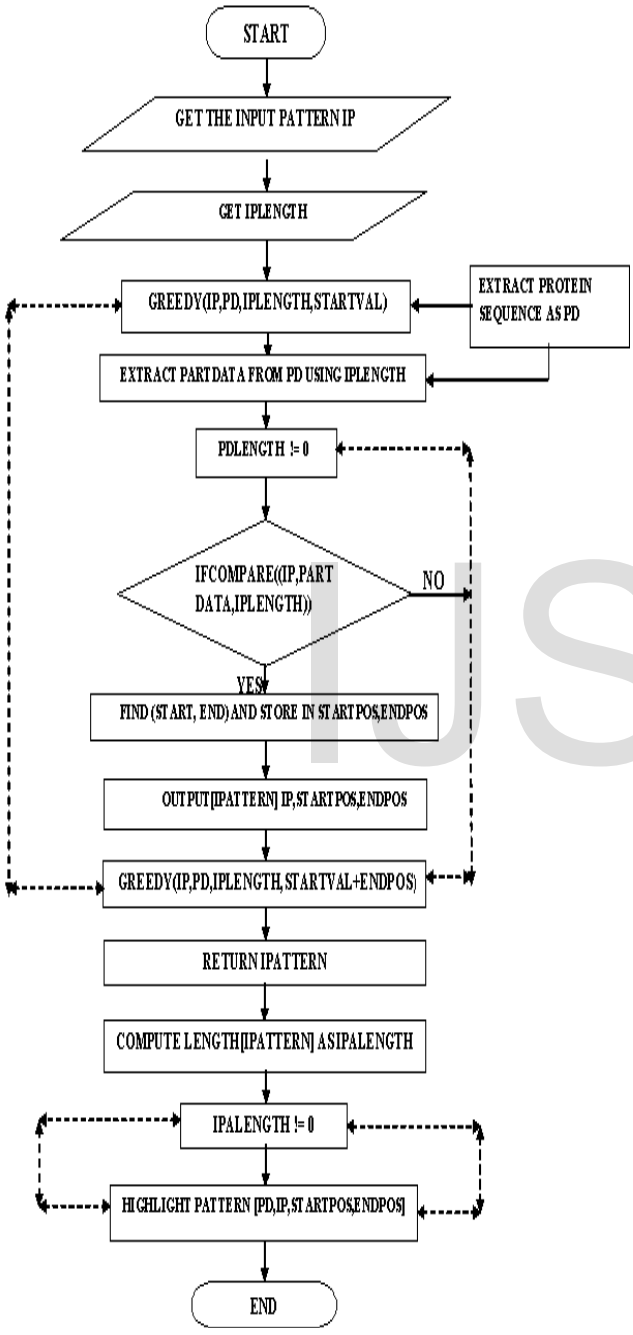
Step 17:Compute Length[IPATTERN] as IPALENGTH

Step 18:Loop(IPALENGTH!=0)

Step 19:Highlight Pattern[PD,IP,STARTPOS,ENDPOS]

Step 20: END Loop

Step 21:END



SAMPLE DATASET:

PD=JKLTVSHIBTVSGYRTVSPOTTVSJFETVSLOPTVSU
NVTVSKIZTVSRF
IP=TVS
IPLENGTH=3
GREEDY(TVS,PD,S,1)
PD[3] THE FIRST SET OF RESIDUES FROM PD =JKL
PARTDATA=JKL

Now length of the sequence is calculated ,till the end checking for the pattern occurrence is performed by,
Compare(TVS,JKL,3)
The start and end value of the pattern if occurred is stored and displayed with a multidimensional
array IPATTERN ,as TVS4 6
TVS1012,TVS1618,TVS2224,……

CONCLUSION:

. The FGAPM is for matching of the pattern with the sequence given, the future work can be enhanced in many ways like the comparison can be carried out for genomes or for different medical data and also the use of numerous medical language tools like UMLS.

REFERENCES:

1. Raju Bhukya, DVLN Somayajulu **"** Multiple Pattern Matching Algorithm using Pair-count" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 2, July 2011 ISSN (Online): 1694-0814 .

2. S. S. Sheik, Sumit K. Aggarwal, Anindya Poddar, N. Balakrishnan, and K. Sekar, "A FAST Pattern Matching Algorithm**"** J. Chem. Inf. Comput. Sci. 2004**,** *44*, 1251-1256.

3. Shann-Ching Chen, and Ivet Bahar, "Mining frequent patterns in protein structures:a study of protease families, bioinformatics.oxfordjournals.org.

4. ROBERT B.MOYER , " A FAST STRING SEARCH ALGORITHM", bioinformatics.oxfordjournals.org.

5. ALLOPROTEIN, http://www.springerreference.com/docs/html/chapterdbid/90965.html.

6. Alain Frisch, and Luca Cardelli, " Greedy Regular Expression Matching studies the problem of matching sequences against regular expressions in order to produce structured values"

7. Joan Hérisson, Guillaume Payen and Rachid Gherbi, "A 3D pattern matching algorithm for DNA sequences",

8. Shann-Ching Chen, and Ivet Bahar, "Mining frequent patterns in protein structures:a study of protease families, bioinformatics.oxfordjournals.org.

9. Joan Hérisson, Guillaume Payen and Rachid Gherbi, "A 3D pattern matching algorithm for DNA sequences", http://oxfordindex.oup.com/view/10.1093/bioinformatics/btl669.

10. http://www.answers.com/topic/bioinformatics.