Protein databank Filtering and Amino Acid Frequency Calculator (PFA_2FC): A Tool

Ranjana Dhuppar
M.Tech. Research Scholar,DCSE, M. M. University,
Mullana-133203,Haryana,India

Gurpreet Singh Bhamra
DCSE, M. M. University,
Mullana-133203,Haryana,India

ABSTRACT

Data Mining is the process of automatic extraction of useful patterns in the form of knowledge from the huge databases. Bioinformatics or computational molecular biology deals with the design and use of computer software to solve the complex biological problem. Proteins are important constituents of cellular machinery of any living organism and the functioning of proteins heavily depends upon its amino acids. A tool called Protein databank Filtering and Amino Acid Frequency Calculator (PFA_2FC) has been designed using Java language to mine the Protein databank and find the frequencies of each amino acid within a protein.

General Terms:

Bioinformatics,Data Mining

Keywords:

Bioinformatics Tools, Protein Sequence Analysis,Frequent Itemsets

1. INTRODUCTION

Data Mining (DM) provides the means for analysis and interpretation of large data for the extraction of interesting knowledge that could help in decision making. Frequent patterns are the patterns (such as itemsets, subsequences) that appear in a dataset frequently and finding such patterns plays an important role in mining associations, correlations, and many other interesting relationships among data[Han2006],[Fayyad1996]. In recent years, rapid developments in genomics and proteomics have resulted into the generation of large amount of biological data. There is a growing need of sophisticated computational analyses techniques for drawing conclusions from such kind of high throughput data. Bioinformatics, a computational discipline in molecular biology, deals with the management and automated analysis of high-throughput biological data to model and simulate the biological systems and processes. Today in-silico analysis is a fundamental component of biomedical research. Bioinformatics has now encompasses a wide range of subject areas from structural biology, genomics to gene expression studies. DM techniques constitute an active area of research in bioinformatics to solve biological problems. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc.[1, 2, 3, 6, 7].

The proteins sequences are made up of 20 types of Amino Acids (AA). Each AA is represented by a single letter code (see Table 1). Unique 3-dimensional structure of each protein is decided completely by its amino-acid sequence. A slight change in the sequence might completely change the functioning of the protein. The main aim of this research is to analyze the AA sequence to find the frequency of occurrence of each AA present in a protein.

Table 1. Single Letter Codes for Amino Acids.

Sr. No.	Amino Acid Name	Single Letter Code	Three Letter Code
1	Alanine	A	Ala
2	Cysteine	C	Cys
3	Aspartic Acid	D	Asp
4	Glutamic Acid	E	Glu
5	Phenylalanine	F	Phe
6	Glycine	G	Gly
7	Histidine	H	His
8	Isoleucine	I	Ile
9	Lysine	K	Lys
10	Leucine	L	Leu
11	Methionine	M	Met
12	Asparagine	N	Asn
13	Proline	P	Pro
14	Glutamine	Q	Gln
15	Arginine	R	Arg
16	Serine	S	Ser
17	Threonine	T	Thr
18	Valine	V	Val
19	Tryptophan	W	Trp
20	Tyrosine	Y	Tyr

2. PFA_2FC TOOL

PFA_2FC tool consists of two modules as shown in Figure 1. Both these components are developed in Java language. The first component is Protein Database Filtering(PDBF) that takes real Protein Data Bank(PDB) as input and filters it to generate another intermediate databank called Filtered Protein Data Bank(FPDB) that contains only those protein records in which the protein sequence length is in the range ≥ 50 and ≤ 400 amino acids. FPDB databank is further processed by another component known as Amino

Acid Frequency Calculator (AAFC) which generates a data bank of amino acids frequency called Amino Acid Frequency Data Bank (AAFDB) for each protein record in FPDB. Real PDB is taken from the Astral SCOP [4, 5],version 1.75 (http://scop.berkeley.edu). This PDB is further modified with the inclusion of a string tokenizer character '#' to separate protein description headers and protein sequences in each protein record. There are total of 10569 Protein records in this data set which are further filtered by PDBF module to generate FPDB. A total of only 9637 such filtered protein records are considered for the next stage of frequency mining. Working of PDBF module is shown in Algorithm 1 and Algorithm 2 describes the working of AAFC module. The screenshots of various datasets involved in the working of the tool are depicted in Figure 2.

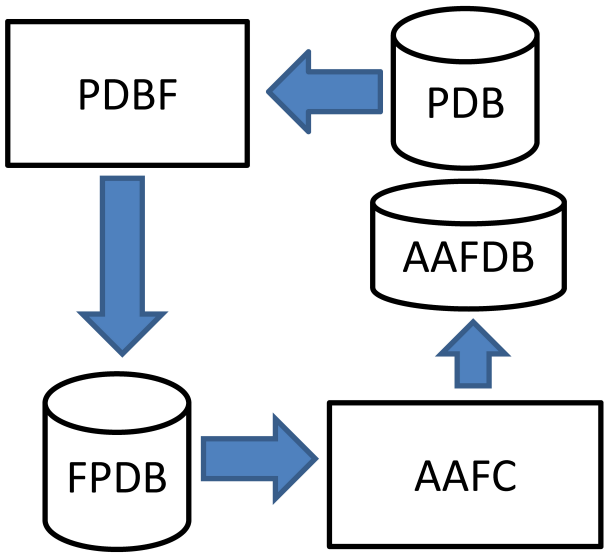

Fig. 1. Block Diagram of the Tool

Algorithm 1 PDBF

Input: $PDB, Real\ Protein\ Databank$
Output: $FPDB, Filtered\ Protein\ Databank$

procedure PDBF(PDB)
  $\alpha \leftarrow$ *open an input stream with* PDB
  $\beta \leftarrow$ *read a protein record from* α
  while $\beta \neq NULL$ **do**
    $\gamma \leftarrow$ *extract protein descriptor token from* β
    $\delta \leftarrow$ *extract protein sequence token from* β
    if $\delta.length \geq 50\ AND\ \delta.length \leq 400$ **then**
      add γ *into vector* PD
      add δ *into vector* PS
    end if
    $\beta \leftarrow$ *read a protein record from* α
  end while
  Add PD *and* PS *vectors into vector* $FPDB$
  Save $FPDB$ *object in the local file system*
  return $FPDB$
end procedure

Algorithm 2 AAFC

Input: $FPDB, Filtered\ Protein\ Databank$
Output: $AAFDB, Amino\ Acids\ Frequency\ Databank$

procedure AAFC($FPDB$)
  $\alpha \leftarrow$ *open an input stream with* $FPDB$
  $\beta \leftarrow$ *read* $FPDB$ *vector object from* α
  $\gamma \leftarrow$ *extract protein descriptor vector from* β
  $\delta \leftarrow$ *extract protein sequence vector from* β
  $AAF[\delta.size][20]$ ▷ amino acids frequencies
  $AACodes \leftarrow$ "$acdefghiklmnpqrstvwy$" ▷ AA codes
  for $ps \leftarrow 0, \delta.size$ **do**
    $PS \leftarrow \delta.get(ps)$
    for $aac \leftarrow 0, AACodes.length$ **do** ▷ for each AA
      $AA \leftarrow AACodes.charAt(aac)$
      $freq \leftarrow 0$
      for $psc \leftarrow 0, PS.length$ **do**
        $PSAA \leftarrow PS.charAt(psc)$
        if AA = PSAA **then**
          $freq \leftarrow freq + 1$
        end if
      end for
      $AAF[ps][aac] \leftarrow freq$
    end for
  end for
  Add γ *vector into vector* $AAFDB$
  Add $AAF[\delta.size][20]$ *array into vector* $AAFDB$
  Save $AAFDB$ *object in the local file system*
  return $AAFDB$
end procedure

3. CONCLUSION

Amalgamation of DM techniques and bioinformatics forms an active area of research. A software tool has been designed to mine the protein databank and finding the frequencies of each amino acids present in a protein. This tool would be used in the ongoing research to mine the frequent amino acids in a biological dataset to study the structure of proteins.

4. REFERENCES

[1] T. Attwood and D. Parry-Smith. *Introduction to Bioinformatics*. Prentice Hall, First edition, March 1999.

[2] J.M. Claverie. From bioinformatics to computational biology. *Genome Research*, 10(9):1277–1279, 2000.

[3] Arthur M. Lesk. *Introduction to Bioinformatics*. Oxford University Press, Fourth edition, February 2014.

[4] Loredana Lo Conte, Steven E. Brenner, Tim J.P. Hubbord, Cyrus Ghothia, and Alexey G. Murzin. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Research*, 30(1):264–267, 2002.

[5] Alexey G. Murzin, Steven E. Brenner, Tim Hubbord, and Cyrus Ghothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(1):536–540, 1995.

[6] Khalid Raza. Application of Data Mining in Bioinformatics. *Indian Journal of Computer Science and Engineering*, 1(2):114–118, 2010.

[7] J. C. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. PWS Publishing Company, 1997.

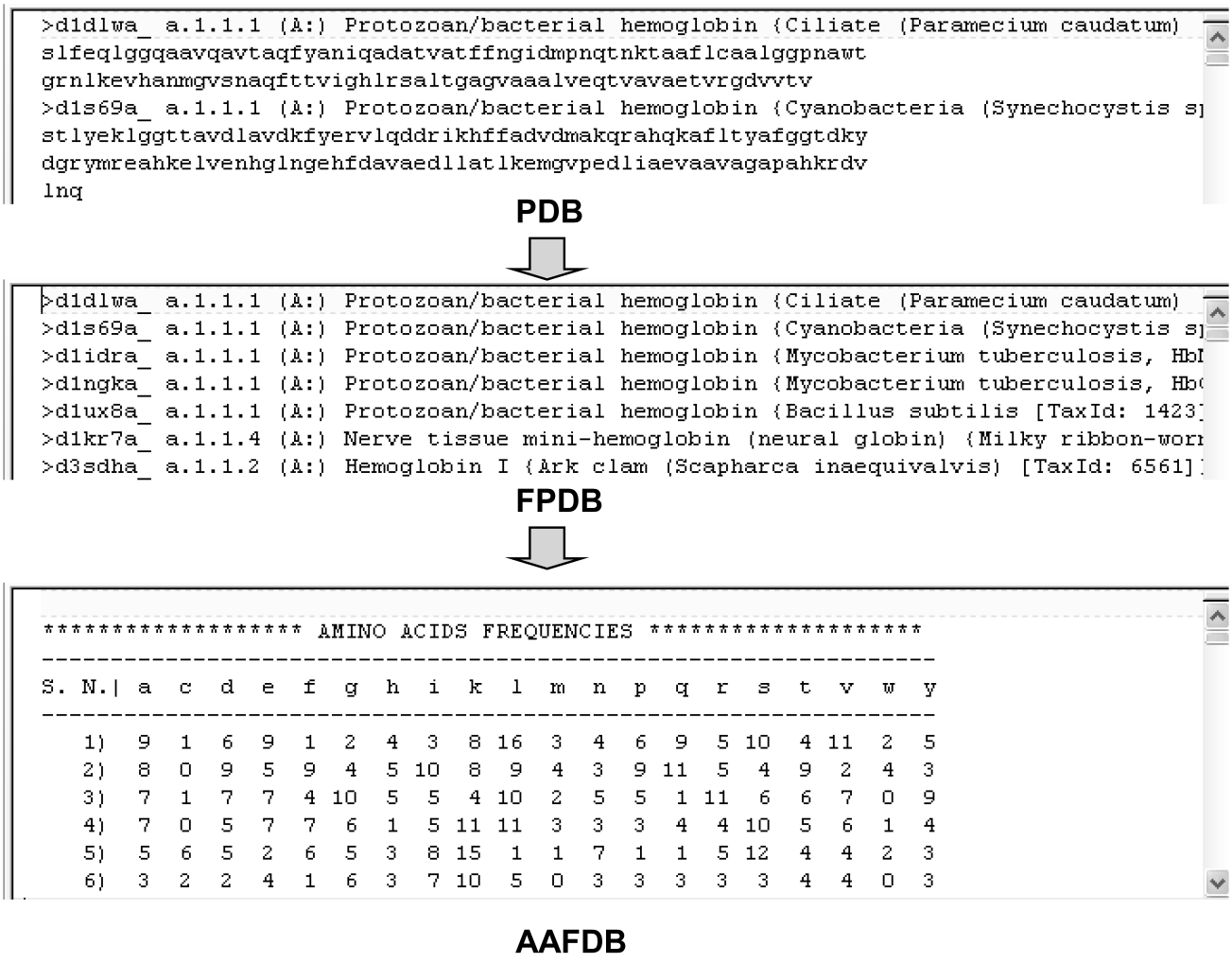

Fig. 2. Screenshots of PDB,FPDB and AAFDB