Available online at www.ijpab.com

ISSN: 2320 – 7051
Int. J. Pure App. Biosci. **2 (1):** 35-39 (2014)

Research Article

International Journal of Pure & Applied Bioscience

Extracting Database Properties for Sequence Alignment and Secondary Structure Prediction

Maulika S Patel[1]* and Himanshu S Mazumdar[2]

[1]Head, Department of Computer Engineering, G H Patel College of Engineering & Technology, Vallabh Vidyanagar, Gujarat, India, http://www.gcet.ac.in

[2]Head, Research & Development Center, Dharmsinh Desai University, Nadiad, India, http://www.ddu.ac.in

*Corresponding Author E-mail: Maulika.sandip@gmail.com, hsmazumdar@hotmail.com

ABSTRACT

A plethora of continuously increasing data exists in genomic and proteomic domains. Computational tools are of vital importance for research in these areas. Biologists, who are involved in identifying new sequences or genes would like to compare their findings with the existing data sets locally. In this paper, we present a set of utilities that can help the researchers to conveniently extract the fields of interest from the public protein databases. UniRef100 is a large comprehensive set of unique, non-redundant protein sequences. The utilities described are used to index, sort, access the records randomly, and extract the properties of UniRef100 database. The properties derived are used for creating a synthetic bio-random database for further research in sequence analysis and secondary structure prediction.

Keywords—*amino acid pair; amino acid trio; protein database; protein secondary structure prediction.*

INTRODUCTION

A genome is an organism's complete set of DNA, including all of its genes. Each genome contains all of the information needed to build and maintain that organism. The work of the Human Genome Project[1] has allowed researchers to begin to understand the blueprint for building a person. As researchers learn more about the functions of genes and proteins, this knowledge will have a major impact in the fields of medicine, biotechnology, and the life sciences. A large amount of genomic and proteomic information is available. This has attracted many researchers towards Computational biology[2] tasks such as multiple sequence alignment[3], sequence similarity, motif finding, and structure prediction[4].

The availability of free and massive data provides ample opportunities to computational biologists to experiment and test their softwares and tools for the tasks of biological relevance. This also ensures that machine learning tools that heavily depend on the training data may be used effectively. As more and more whole genomes are sequenced, the need for a central, publicly available and easily accessible archive for deposition, searching and analysis of sequence data continues to grow[5]. With the exponential increase in the size of the genomic and proteomic databases, there exists a need to hierarchically maintain these databases for fast and relevant retrieval. It is very common to compare two DNA sequences to identify the common genes. Sequence similarity in protein databases is also important for classifying the existing proteins and also categorizing the newly invented proteins.

Alignment based methods such as those based on Dynamic programming[6] and BLAST[1] have been developed for identifying sequence similarity. BLAST has been widely used by biologists for sequence analysis. Multiple sequence alignment[3] is a problem of aligning more than two sequences to identify the conserved regions in many sequences. Tools like ClustalW give an optimal alignment on the given input sequences. Multiple sequence alignment becomes computationally costly with the increase in the number of sequences and also with the length of the sequences. Alignment free sequence analysis approaches

have used techniques such as finding identical short substrings called words, using frequencies of all fixed length words[7] and Hidden Markov Models[8].

Sequence analysis reduces the scope of the protein structure prediction algorithms[4] and increases the accuracy of prediction. Protein structure prediction is a bioinformatics task of much relevance as it helps in finding the function of a protein. Proteins fold, and the fold determine the function of a protein. Understanding how proteins fold, and the functions of proteins can help the biologists in prediction of a state of a disease. Several machine learning tools have been used to predict the protein structure. Neural networks[9, 10], support vector machines[11], hidden markov models and statistical methods[12, 13], have been widely used for protein secondary structure prediction. These methods have achieved the accuracy of slightly greater than 75%, leaving a lot of space for researchers in this area. Release 2012_01 of 25-Jan-2012 of UniProtKB/TrEMBL contains 19434245 sequence entries[5], while the number of discovered protein structures as on 3rd Aug 2012 is 77,195 [14]. This clearly indicates that the gap between the discovered protein sequences and discovered protein functions is increasing. Any set of tools that reduces this gap is desirable.

It is obvious that there are researchers who are interested in specific and hence limited databases in terms of size and species. A biologist who is working on a family of proteins might not be interested in a generalized database. There are other researchers who need to work with less specific and large databases so as to generalize their findings. Searching for sequence similarity is one such task. Researchers use variety of data sets, of which some are very limited in size. This particularly saves time and energy but might not be representing the complete set. A set of utilities to customize and preprocess the entire data downloaded from the primary databases is developed and demonstrated.

The paper is organized in 4 sections. Description of how the databases can be customized and prepared for the necessary tasks is given in Section II. A detailed discussion on signature extraction based on the amino acid composition in various species is given in section III. Conclusion is presented in section IV.

AN INSIGHT INTO THE UTILITIES DEVELOPED

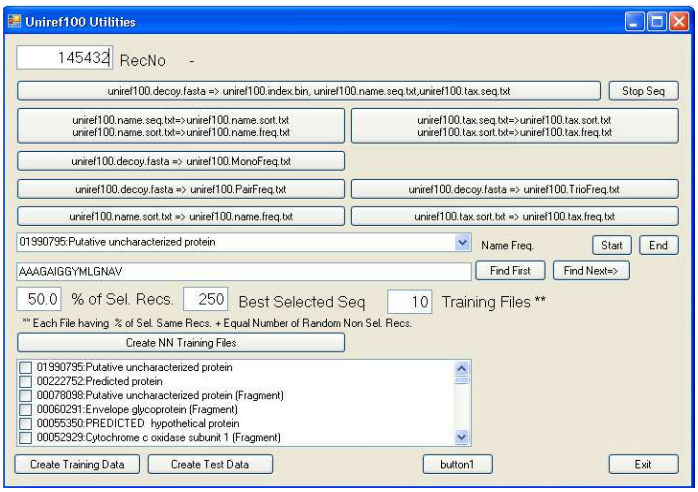
Authors chose to use the Uniref100.fasta database, available from www.uniprot.org[15,16]. The size of the database is 4.21 GB. It contains 9757328 protein records containing information such as name, id, tax, and amino acid sequence as on 31st Jan 2010. It is necessary to develop a memory resident program that facilitates to access and view the large database. In this view, a set of utilities were developed[17] so that the database can be handled with convenience and necessary features can be extracted from the database.

A. Customizing the database

The database in fasta format needs to be processed for various purposes. An integrated tool is developed in C# .NET that facilitates indexing of large database as well as viewing the database at run time. Figure 1 shows the snapshot of an integrated tool for following tasks.

1) Indexing the UniRef100.decoy.fasta database for random access

Fig. 1 Integrated tool for creating binary index file to retrieve records, calculating frequency of mono, pair and trio of amino acids, creating test and training data sets

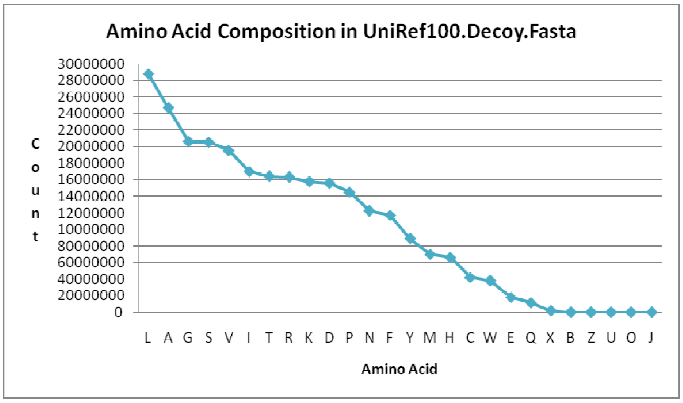


2) Generating separate files for storing amino acid sequences and storing related fields like name and TAX
3) Calculating the frequencies of substrings of amino acids of different lengths
4) Generating the data set for training and testing the neural network
5) Searching for a given word in the selected file

B. Protein sequence - secondary structure pair database

For addressing the problem of protein secondary structure prediction, a number of databases are available. Some of them act as benchmarks for testing the PSSP softwares. Among these are RS126, CB396 and CB513. Also, it has been found that researchers have their own databases for protein secondary structure prediction[18]. In this case, a set of 14292 protein records having secondary structure information was downloaded from Uniprot. These records had lot of information of biological relevance, but were filtered out for preparing the input-output pairs for using with neural network. This operation provides the sequence and structure pair database. The database so generated has been used by the authors in their ongoing work on protein secondary structure prediction using Backpropogation neural networks.

Fig. 2 Amino acid composition in UniRef100.decoy.fasta



C. Programmatically mutated bio-random database

A programmatically mutated bio-random database can be created using the tool as mentioned in[17]. Although databases are freely available, but it is observed that researchers develop their own customized databases so as to suitably test their application. To validate the similarity search algorithms, an artificially mutated known dataset may be used. If one is using a standard database, he might not have the information of the number of similar sequences present in the database for a given query. The authors feel this issue can be addressed by using a database with known properties. The parameters used are:

1) % mutation required
2) % insertions allowed
3) % deletions allowed
4) Group size

Based on the selected mutation rate, insertions and deletions will be performed at random locations in all sequences. For example, if mutation rate = 5%, then sequences having 95 % similarity in terms of the number of amino acids matching, will be generated. This allows forming groups of sequences. The number of sequences in each group is controlled by the group size parameter. This information is stored and is used when testing the similarity search algorithm. The amino acid statistics extracted from natural sequences have been used to preserve the proportion of the amino acids in the bio-random database.

SIGNATURE EXTRACTION USING FREQUENCY OF AMINO ACID RESIDUES

An algorithm to calculate the frequencies of potential 20 mono, 400 pairs, 8000 trios (sequence of three amino acids), and 160000 quads (sequence of four amino acids) of amino acids is developed and discussed by[17]. The method is used for finding the composition of single amino acids, amino acid pairs and amino acid trios in the uniref100.decoy.fasta database. The letters B, J, O, U, X and Z do not correspond to any amino acid. However, their occurrences are found in the uniref100.decoy.fasta database for unknown reasons.

TABLE I. SIGNATURE EXTRACTED IN ORDER OF FREQUENCY OF 20 AMINO ACID AND TOP TEN AMINO ACID PAIRS FOR VARIOUS SPECIES

S. No.	**Species**	**Signature (20 Amino Acid)**	**Signature (Top ten amino acid Pairs)**
1	Homosapiens1	LSAGEVPRTKIDQFNYMHCW	LLAALRALTLAVLALESSLS
2	Mouse	LSKETVGRAQIDPFNHCYMW	SSLSKSSLEELLEKLKGKCG
3	Mus_musculus	LSAEGVPKRTDQIFNYHMCW	LLLSSLALLAVLSSLVLGLE
4	Oryza-sativa	AGSLVPERDKTFINQYHMCW	AAGGLLVAAVLASSASVVAL
5	Pan-troglodytes	LSAVPGETRDKIQNFYCHMW	LLSSLSRRALAASLLPLAGL
6	Rattus	LSEAGVPKTDRQINFYHMCW	LLLSSLALLAVLLVLGSSAA
7	Rattus-norvegicus	LSEAGVKPTIRDQNFYMHCW	LLSLLSLAALVLLVLGGLSS
8	Solanumlycopersicum	LSKIREVGADTNFPQYHMCW	LLSSLSLAVLLVAASLLGAL
9	Zebrafish	LSEKTGVRAQIDPFNHCYMW	SSLSKSSLEELLEKLKGKCG

TABLE II. SIGNATURE CONSISTENCY OBSERVED ON DIFFERENT DATA SETS OF SAME SPECIES

Data set	**Species**	**Signature**	**Data set**	**Species**	**Signature**
1	Mouse	LSKETGVRQIDANFPCHYMW	1	Mus_musculus	LSAVEGTPKIRDFQNYHMCW
2	Mouse	SLEKTRGAVQIDPFNHCYMW	2	Mus_musculus	LSGAVEPRKTIDQFNYHCMW
3	Mouse	LSKETVGRAQIDPFNHCYMW	3	Mus_musculus	LSAEGVPKRTDQIFNYHMCW
4	Mouse	SLVTEGKAIDRPNQFYHCMW	4	Mus_musculus	LSAEGVPRKTDQIFNYHMCW
5	Mouse	LSEVATRKDGIQPNFHYCMW	5	Mus_musculus	LSAEGVPKRTDQINFHYMCW
6	Mouse	LSEGKAVTPRDQINFYCHMW	6	Mus_musculus	LSAEGPVRKTQDINFYHMCW
7	Mouse	SLEAVGKTRPDIQNFYHCMW	7	Mus_musculus	LSAGEVPRKTDQIFNCYHMW

Figure 2 shows the amino acid frequency in UniRef100.decoy.fasta. Please note that letters B, J, O, U, X, and Z don't correspond to any amino acid and their occurrence is almost zero. The frequencies have been sorted and suggest that amino acids L and A have the highest occurrence in Uniref100.decoy.fasta. It is found that amino acid pair LL is the highest occurring pair followed by pair LA, and the trio, LLL is the most occurring trio followed by AAA. We have analysed 100Kb of protein sequence data of 9 species. The amino acid residues in the descending order of frequency of occurrence can be used as a signature. The top 10 amino acid pairs form the signature as shown in table 1. Table 2 shows the consistency of the signatures of two species in different datasets.

We have extended the idea of frequency of amino acid residue words upto the length of 15. The same has been used for similarity searching in large databases[19,20]. Currently, we are working on secondary structure prediction using the frequency of amino acid residue words and their corresponding structure.

CONCLUSION

Useful database properties from the large primary databases can be extracted using the specifically designed tools. The properties of the databases can be known and further used to generate a bio-random database. We have demonstrated the tools and used the same for extracting useful information from Uniref 100 database and database of 9 species. The tools can be used to provide significant clues to the researchers to test their sequence analysis and secondary structure prediction algorithms on various data sets.

REFERENCES

1. S Altschul et al. Basic Local Alignment Search Tool. J. Mol. Biol., **215**: 403–410 (1990)
2. C. Setubal and J. Meidanis. Introduction to Computational Molecular Biology, Cengage Learning, (1997)
3. H. Carrillo, & D. Lipman. The multiple sequence alignment problem in biology. *SIAMJ Appl. Math*, **48(5)**: 1073-1082, (1981)
4. J. Cheng, A. Tegge & P. Baldi. Machine learning methods for protein structure prediction. IEEE Reviews in Biomedical Engineering, **1(49)**: (2008)
5. Europe's leading laboratory for basic research in molecular biology accessed from the World Wide Web: http://www.ebi.ac.uk/embl.
6. W. Pearson & D. Lipman. Improved tools for biological sequence comparison. Biochemistry, **85(8)**: 2444-2448, (1988)
7. M. Kantorovitz, G. Robinson, & S. Sinha. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics, **23(13):** i249–i255, (2007)
8. T. Pham and J. Zuegg. A probabilistic measure for alignment-free sequence comparison. Bioinformatics, **20(18):** (2004)
9. R. Kakumanim, V. Devabhaktuni, and M. Ahmad. A two stage neural network based technique for protein secondary structure prediction. 30th Annual International IEEE EMBS Conference, pages 1355-1358, IEEE, (2008)
10. J Chen and N. Chaudhari. Cascaded Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, **4(4)**: (2007)
11. B. Naul. A Review of Support Vector Machines in Computational Biology, (2009)
12. S. Vinga and J. Almeida. Alignment-free sequence comparison—a review. Bioinformatics, **19(4)**: (2003)
13. G. Reinert, D. Chew, F. Sun, and M. Waterman. Alignment-Free Sequence Comparison (I): Statistics and Power. J Comput Biol. **16(12)**: 1615–1634, (2009)
14. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, & P. Bourne. The Protein Data Bank. Nucleic Acids Research, **28(1)**: 235-242.
15. Database in fasta format- Uniref100.fasta, downloaded on 31st Jan 2010 from the World Wide Web: http://www.uniprot.org.
16. B. Suzek, H. Huang, P. Mcgarvey, R. Mazumder, & C. Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. Bioinformatics, **23(10)**: 1282-1288, (2007)
17. M. Patel and H. Mazumdar. Utilities for preprocessing large biological databases. Proceedings of WCECS 2010, **2**: IAENG, (2010)
18. S. Ray, S. Bandyopadhyay, P. Mitra and S. Pal. Bioinformatics in neurocomputing framework. IEE Proc.-Circuits Devices Syst., **152(5)**: (2005)
19. M. Patel and H. Mazumdar. Similarity Search Using Pre-Search In Uniref100 Database. International Journal of Hybrid Information Technology, **4(3)**: (2011)
20. H. Mazumdar and M. Patel. Protein sequence similarity search technique suitable for parallel implementation. International Journal of Computer Applications, **50(22)**: (2012)