Application of Genetic Algorithm in Various Bioinformatics Problems

Subhendu Bhusan Rout[1], Sasmita Mishra[2], Dillip Kumar Swain[3]
Department of CSEA,IGIT Sarang, Odisha, India

Abstract- **Genetic Algorithm is one of most popular soft computing technique in the field of Bioinformatics. There are so many soft computing techniques like Artificial Neural Network, Fuzzy Logic, Swarm Optimization, Genetic Algorithm etc. These techniques are very much use full in the format of processing of huge amount of data. These soft computing techniques are having several; applications in the field of drug design and medicine research. Though there are a large number of hard computing techniques in this regard still in these days as these bioinformatics data or genomic data are folded in to several times so soft computing techniques are very popular. There are many task of bioinformatics like protein structure prediction, gene mapping, DNA-RNA alignment etc. These problems can be easily solved by using soft computing techniques for which soft computing techniques are becoming very popular in the recent days. Genetic Algorithm which is one of the most popular techniques of bioinformatics has several applications for these bioinformatics problems. In this paper we have discussed various implementation and research works using genetic algorithm. We have also proposed a technique for protein structure prediction using genetic algorithm.**

Index Terms- **Bioinformatics, Swarm Optimization, ANN, Genetic Algorithm, Protein Structure Prediction, Gene Mapping**

I. INTRODUCTION

Bioinformatics is the application of computer technologies in the field of medical science. In the last few decades as these types of genomic data are folded into many times so it needs a good and upgraded computer technology for processing as well to find out fruitful result upon these medical data processing. There are many soft computing technologies like Artificial Neural Network, Fuzzy Logic, and Genetic Algorithm which has great applications upon the bioinformatics problems. There are also many hard computing techniques that are already implemented in this purpose to process or simulate these genomic data. There are several task of bioinformatics like Protein structure prediction, DNA RNA alignment, Gene Mapping which can be easily solvable by using soft computing techniques.

Artificial Neural Network has a huge application upon these various bioinformatics problems. There are many research works that are already carried out for these problems using Artificial Neural Network. Artificial Neural Network is a combination of artificial neurons which generally works just like a human brain. As our human brain controls all the part of human body and send signals through neurons to human brain similar type of work done through ANN. Artificial Neural Network has a great application upon Protein structure prediction, gene mapping, DNA-RNA alignment and comparison as well as other bioinformatics problems. Fuzzy Logic is one of the important soft computing techniques for the bioinformatics problems like DNA RNA alignment, Protein Structure Prediction, Gene Mapping etc. Though there are many implementations and research work that are already done for these problems still this field of research attracts more and more researcher to find out more and more new techniques and methods for the processing of genomic data.

No doubt Genetic Algorithm and other Hybrid soft computing techniques like swarm optimization, Ant Colony are very much useful for these problems. Though there are a less number of research are being carried out in this field still there are some specific and optimized research that are being carried out in this field. Genetic Algorithm has huge application for the bioinformatics problems. In this paper we have point out several research works that are being carried out in the field of bioinformatics. Our

literature review provides the brief research works that are being developed in this area.

This Paper is organized as follows. The Section II provides a brief literature review about Genetic Algorithm and its applications. In section-III we have discussed some applications of Genetic algorithm upon various bioinformatics problem. In this section in sub section A,B,C we have discussed about the application of Genetic Algorithm for the problems like PSP, Gene Mapping and DNA, RNA alignment etc. In section-IV we have provides a brief comparison of research articles that are associated with genetic algorithm and bioinformatics problem. The section-V provides a proposed idea and experimental result of a database 3FS4. Finally the paper concludes in section-VI with conclusion and future work.

II. GENETIC ALGORITHM AND ITS APPLICATIONS

Genetic algorithms are the heuristic search engines and methods, which are always, operate on many pieces of information like nature with genes during evolution. It deals with three levels like selection, crossover and mutation. For a biological research the Selection level in the Genetic Algorithm search is very very important. For example It may dictates which chromosomes are selected for breeding in the next crossover operator. The crossover operator is the heart of the Genetic Algorithm (GA). Mutation is always used to stop the GA from getting trapped in local minima and also in stalling. There are a lot of numerous ways of doing this, but one of the wise techniques is to use rotational moves to mutate a protein's conformation.

Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. The Genetic Algorithm is a model of machine learning which derives its behaviour in the form of a metaphor of the processes. GAs represents an intelligent exploitation of a random search which is used to solve various optimization problems. It is always better to use Genetics Algorithm instead of other conventional AI techniques because GAs doesn't break easily even if the inputs changed deviates slightly or in the presence of reasonable noise. The Genetic Algorithm Simulations deals with the process of natural evolution and mimicking the processes. These processes include Selection, crosses over, mutation and accepting. If we consider an example and if we go for a comparison we may observe that the caterpillars and butterflies may have identical genomes, but the difference in expression of the genes in DNAs lead to an obvious physical differentiation. These types of events motivate the modelling of the gene expression networks. The Rebuilding of these genetic networks from the gene expression profiles allows the inventions of various functions targeting over diverse domains including biochemistry, molecular biology, bioengineering, and pharmaceutics. [1]

The Genetic Algorithm is the natural evolution which provides a metaphor of the natural process. So by using Genetic Algorithm like an evolutionary manner in protein structure prediction problem one can easily predict the territory and quaternary changes in those amino acid sequences. As these type of genomic data gather in a huge amount so it need a reliable and suitable platform for integrating, processing and interpreting these data. The basic shape of these amino acid sequences generally changes in the secondary and its shape changes in the territory and also quaternary positions with the effect of drugs or external agents. The research in the changes of behaviour and structures in the primary, secondary, territory and quaternary is so much useful in the medicine or drug research. For providing the optimization technique in the evolutionary manner, Genetic Algorithm is always a good technique for these prime researches. [2]

III. APPLICATION OF GENETIC ALGORITHM UPON VARIOUS BIOINFORMATICS PROBLEMS

Genetic algorithm is one of the basic natural optimization techniques for many biological problems. As this technique works on an evolutionary process with proving output in creating the metaphor of the process, so it is very popular among all soft computing techniques. The Genetic algorithm is having several applications upon bioinformatics problems like protein structure prediction problem, DNA-RNA alignment and Gene mapping. We have taken several research works that

are already carried out in this area for these type of bioinformatics problems.

A. Applications in PSP Problems

Protein Structure prediction is the prediction of the three dimensional structure from the amino acid sequence. Generally the proteins are the large biological molecules of a living body which contain large number of amino acid sequence. The amino acid sequence generally creates a three dimensional structure. During the passing of time and with effect of external agents these amino acid sequence change their shape in the secondary, territory and quaternary stages. Though this process is a biological process but the Genetic Algorithm creates the mimicry of the process and creates the metaphor of the process to show the entire evolution.

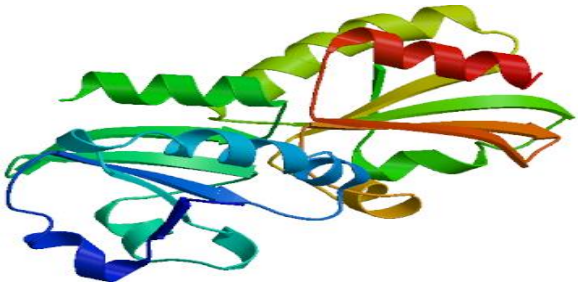

Fig.1. 3D structure of a protein

The Fig.1 shows the three dimensional structure of the protein. These shapes changes from time to time. If there is some models or research outputs are already in our hand than it is always easier for Genetic Algorithm to provide the metaphor of the process by studying and analyzing the previous models.

B. Applications in Gene Mapping

Gene mapping is also called as genome mapping is the creation of the genetic map assigning the DNA fragments towards chromosomes. Basically If we investigate a genome then the map is nonexistent. These maps improve with scientific applications, progress and become truthful when the genomic DNA sequencing of the species is being completed. When we have to map between two genes first of all we have to distinguish between these genes in the chromosomes. Throughout this process, and for the proper investigation of the differences in strain, the total fragments are marked by small tags. This can be called as genetic markers or unique sequence dependent pattern of the DNA-cutting enzymes. This ordering is being taken from genetic observations for the markers. The Gene Mapping is generally used in two different and related contexts. The Two different ways of these mapping are fully distinguished from each other[3].

Genetic mapping always use genetic techniques to calculate the sequence features within the genome. In this format using modern molecular biology techniques for same purpose is generally referred as physical mapping. In this physical mapping, the DNA is cut off by a restriction enzyme. Once it cuts up, the total DNA fragments are being separated by electrophoresis. After this the resulting pattern of DNA migration is used to identify what stretch of DNA is in the clone. By analyzing the fingerprints, contigs are assembled into overlapping DNA stretches manually or automatically. Macro restriction is also one type of physical mapping where in the high molecular weight the DNA is digested with a restriction enzyme with a low number of restriction sites. Once the map is determined, the clones may be used as a resource for efficiently contain large stretches of the genome. These types of mapping are also more accurate than that of genetic maps. Genes may be mapped prior to the complete sequencing with independent approaches like in situ hybridization.

C. Applications in DNA RNA Alignment

DNA-RNA alignment and comparison is one of the major tasks of bioinformatics. As these type of data are folded into many times in the recent years so this topics catch the eyes of man researchers to get several new idea and techniques in order to process these types of data. DNA comparison is the only concrete idea to differentiate between two persons and two generations. In order to determine the similarity in sequences and to get valuable evolutionary information it is very much essential for biologist to align multiple sequences of DNA, RNA or amino acids. The main intention of Gene prediction is to identify regions of genomic DNA that will encode into proteins. Computational methods also required to keep up with the annotation of the rapidly increasing sequencing of genomes[4].

Sequence alignment is one of the most essential parts of bioinformatics. The secondary, territory and quaternary sequence alignment is most essential in predict homology between new sequences. In [4] they have provides a new technique as well as

algorithm "SAGA" for the alignment of sequence using genetic Algorithm. This SAGA technique uses an automatic scheduling scheme to control the uses of different operators for the combination and mutation between generations. According to the authors this SAGA technique is able to find globally optimal multiple alignments in reasonable time, starting from unaligned sequences. In this method they have provides a cost function to align scores between two aligned sequences. An important thing about SAGA population structure that to neglect the duplicates. In the same generation all the alignments should be different. So this technique helps to keep a high level diversity in a population of small size. In order to maintain this level a child is checked to ensure whether he or she is identical to any of the children already generated. If it is not than the new born should be kept to new generation. This process is carried on until enough children have been successfully inserted in the new generation. The Evaluation/Breeding process will be carried on until the decision is made to stop the search[4].

IV. COMARISON WITH OTHER SOFT COMPUTING TECHNIQUES

Fuzzy logic has been proved to be a very useful tool for the representation of knowledge by mathematical expression. The optimized result of fuzzy logic is adopted as very powerful in the fuzzy expert system. Generally the fuzzy logic is having two components that is the discrete one that is the rules and the second one is the fuzzy sets. In the other hand Genetic algorithm are the optimization methods which are based on the mechanism of natural evaluation such as selection, mutation and reproduction. These techniques are introduced decades ago but used as a solution to many problems in the field of soft computing in the current years. In the current years both the techniques fuzzy logic as well as Genetic algorithm merged together to solve many problems in the recent years. Though both the soft computing and Genetic Algorithm are very promising techniques in the field of soft computing still Genetic Algorithm provides better and upgraded result and more accurate result than that of fuzzy logic. In general language Fuzzy systems are rule based systems which are capable of dealing with imprecise information. The most advantage is that everything inside a fuzzy system is interpretable for humans. Though prototyping can be done very easily and quickly but it takes more time to tune all the involved parameters which is a drawback.

There are also several experiments having predictions of protein structure using Artificial Neural Network. The Fig.2 provides a clean idea about the designing of Neural Network for various Bioinformatics problems. Though Neural Network can be used in several applications in day to day life still it has a huge application in Bioinformatics. In [8] they have provided an experiment of protein structure prediction using Artificial Neural Network. Their research work is upon "paired couple amino acid composition". They have used a self-consistency test with an independent data set. In this paper they have shown the error in self-consistency test and an independent data set test. The average over all error rate for self-consistency test is 0.022 and in the other hand the average overall error rate in independent data set is 0.025 which is little bit higher than that of previous one. From the total number of 52 proteins none of them is having more than 35% of homology with the others during the training dataset. The result is quite similar to that of Liu et. al. in [10]. Though Artificial neural network provides the results with the protein data set still Genetic Algorithm may provide better and low error rate with higher accuracy prediction rate[8].

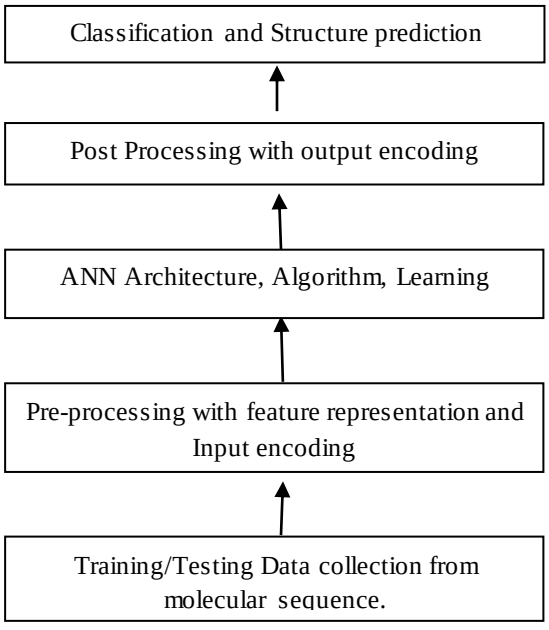

Fig. 2. Design of Neural Networks for bioinformatics

Support vector machine (SVM) is one type of learning machine based on statistical learning theory. Support Vector Machine (SVM) is also one more technique for the purpose of protein structure prediction and other bioinformatics problems. In [11] they have provided a technique for prediction of protein structural classes using SVM. The support vector machine is applied based on the data base derived from Structural classification of Protein (SCOP) data base which generally creates a three dimensional structure. In this paper both self constancy and jack knife tests are performed. According to the others support vector machine can be taken as a powerful tool and technology for the protein structure prediction problems. For the experimental purpose they have taken 204 protein chain among which 52 are all α proteins, 61 all-β proteins, 45 α/β proteins and 46 α+β proteins. finally they have got the correct prediction rate of 64%, 90%, 64% and 64%respectively. According to the authors the results of SVM both in case of self consistency test and jack knife test gives better performance than that of Neural Network [11]. But still the Genetic Algorithm is also popular soft computing technique for this purpose.

V. APPLICATION OF GENETIC ALGORITHM IN 3FS4 PROTEIN DATABASE

Genetic algorithm is having a huge application in the field of bioinformatics. It basically depend s upon three stages like selection, crossover and mutation. The Genetic Algorithm is basically creates a metaphor of the system by studying and realizing the behavior of the system. Each leaving body contains the proteins. Proteins are the large biological molecule in a human body and any leaving body. As the time rotates the number of these types of genomic data like DNA-RNA sequence, Genes, Protein Data bases are folded into many times. In many areas DNA-RNA sequence matching, sequence separation is required. In these fields the Genetic Algorithm as well as other soft computing techniques is very much useful. There are several website and organizations available in the worldwide for the purpose of sharing these types of biological information and research outputs. 'www.ncbi.org' is one of the popular website for this purpose to collect protein data base. In this experiment we have taken a ostrich PDB like "3FS4".

The 3FS4 is an ostrich pdb freely available with a small size like 500KB. We have applied the Genetic Algorithm technique for this pdb to calculate the secondary, territory and quaternary structures. In Fig.3 it provides the structure of the pdb 3FS4. It shows the protein data base with the form of ribbon with coloring each group with different color.

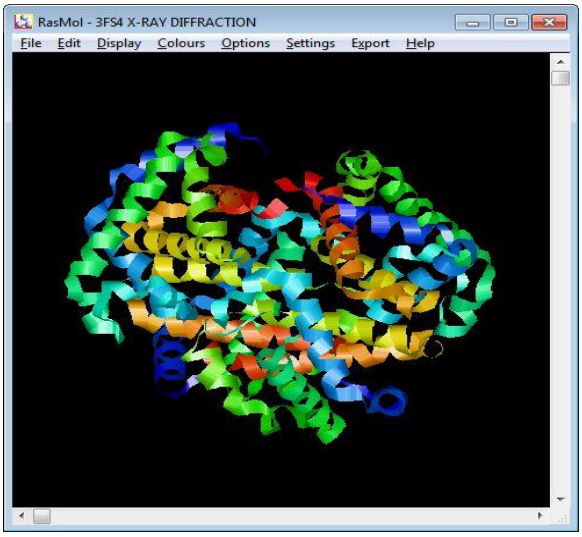

Fig.3. Screen shot of Three dimensional structure of 3FS4 with ribbon view

Two genomic bodies are different from each by their evolution, DNA-RNA sequencing, difference in their protein structure etc. The proteins are the large biological molecules which contains large number of amino acid sequenced in different format. These amino acid sequences generally creates a three dimensional structure which generally changes in the affect of external agents. In this experiment we have taken the 3FS4 pdb which is generally collected from ostrich. It is basically a protein data base collected from ostrich in a genomic format and stored in soft format. The alpha-helix, beta-sheet, and loop format are displayed in fig.3 in a graphical format. The genetic algorithm generally creates a metaphor of the system when there is the availability of the predefined system. In comparison to other soft computing techniques Genetic Algorithm provides a secure and better environment for the protein structure prediction.

Rather than constructing more complicated classifier or protein structure prediction model is always wise to design a simple model for the testing and

experiment purpose. In [12] they have a presented a simple SVM fusion network for predicting the protein secondary structure. In the first layer three classifiers are constructed with various statistical inputs. Then after the computational values combined it enters into a second layer. The result from two data sets represented that the fusion network can improve prediction accuracy. Though some small data set are taken here but this can be applied to large and huge data set for big experiment purpose. This experiment contains a very small data set with a predefined experimental platform. But this model will be helpful for the simulation of the big data sets as well as large protein data bases.

There are several environments to design and experiment of the protein data base. Swiss model is one of the premier and reliable tool to draw and research for the protein data base. Pymol is one of the best molecular designing tools for this purpose. The application of external agents and drugs changes in the shape of the protein structure. The change in shape and size of the protein structure makes the drug researcher to create a better drug or to create a better research platform to design various medicines. For an experimental purpose we have taken a small database as well as a small experiment to design and to display the protein structure with the application of Genetic Algorithm techniques. Generally these types of protein data comes in a huge amount as these types of data are folded into many time. So this type simulating platform will provide an environment for the simulation of big databases.

VI. CONCLUSION & FUTURE WORK

There are several bioinformatics problems for which soft computing techniques play a vital role in order of solving these problems. So soft computing techniques like ANN, Fuzzy logic, Genetic Algorithms are having several applications in the field of PSP problems, Gene Mapping, DNA-RNA alignment etc. In comparison to other hard computing techniques these soft computing techniques provide better result in the form of simulating the result as well as for the real time developments. Basically Genetic Algorithm provides a better result for these problems as well developments of a good research platform for the Genes, DNA-RNA, or PSP problem. This paper though provides a neat and clean idea about the researches that are already carried out in this topic but our upcoming research will focus upon the real time experiment of these bioinformatics problems using Genetic Algorithms.

Compliance with Ethical Standards:

(1) There is no source of funding for this research and there is no conflict of Interest among the authors.

(2) This article does not contain any studies with human participants or animals performed by any of the authors.

(3) Informed consent was obtained from all individual participants included in the study.

REFERENCES

[1] S. B. Rout, S.N. Dehury, B.S.P. Mishra, " Protein Structure Prediction using Genetic Algorithm" International Journal of Computer Science and Mobile Computing, Vol. 2(6), 2013.

[2] S.B. Rout, S. Kisan, S. Mishra, "Protein Secondary Structure Prediction of PDB 4HU7 using Genetic Algorithm (GA)", Proceedings of International Conference on Computer Communication and Informatics, IEEE, 2017.

[3] S. B. Rout, S.N. Mishra, B.S.P. Mishra, "Mapping of Genes using Cloud Technologies", IJRET, Vol.2(2), 2013.

[4] P. Agarwal, R. Gupta, T. Maheswari, P. Agarwal, S. Yadav, V. Bali "A Genetic Algorithm for Alignment of Multiple DNA Sequences" 2012.

[5] C. Notredame, D. G. Higgins, "Sequence Alignment by Genetic Algorithm" Journal of Nucleic Acids Residue, Vol. 24(8), 1996.

[6] Ulrich Bodenhofer "Tuning of Fuzzy Systems Using Genetic Algorithms" Institute for Mathematics, Austria, 1996.

[7] Amie Judith Radenbaugh "Applications of genetic algorithms in bioinformatics", San Jose State University, 2008.

[8] Y. D. Cai, X. jun liu, X. biao Xu, K. Chen Chou, "Artificial neural network method for predicting protein secondary structure content", Journal of Computers & Chemistry, Vol. 26(4), 2002.

[9] S. Agrawal, S. Silakari, "A Review on Application of Particle Swarm Optimization in Bioinformatics" Journal of Current Bioinformatics, Vol. 10(4), 2015.

[10] W.M. Liu, K.C. Chou, "Prediction of Protein Structure" Protein Engineering, pp. 1041–1050, Vol. 12, 1999.

[11] Y. Dong Cai, X. Jun Liu, X. biao Xu, K. Chen Chou "Prediction of protein structural classes by support vector machines", Computers and Chemistry, pp. 293–296, Vol. 26, 2002.

[12] C. Chen, X. Zhou, Y. Tian, X. Zou, P. Cai "Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network", Analytical Biochemistry, pp. 116–121, Vol. 357, 2006.

[13] Yu-Dong Caia, Guo-Ping Zhou, "Prediction of protein structural classes by neural network", Journal of Elsevier, pp. 783-785, Vol. 82, 2000.

[14] Kuo-Chen Chou, Yu-Dong Cai, "Predicting protein structural class by functional domain composition", Biochemical and Biophysical Research Communications, pp. 1007-1009, Vol. 321, 2004.

[15] J. Yang, T. Yang Zhu, X. Chi Dong, "An Application Software For Protein Secondary Structure Prediction Based On Peptide Triplet Units And Artificial Neural Networks", International Congress on Image and Signal Processing, IEEE, 2010.