International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 01, January 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

A COMPARATIVE ANALYSIS OF SOFT COMPUTING TECHNIQUES FOR PREDICTING PROTEIN 3D STRUCTURE

Manish Kumar[1], Hari Om[2]

[1,2]*Department of CSE, IIT(ISM), (India)*

ABSTRACT

It is broadly recognised that the prediction and classification of protein sequences has become one of the most important research topic in present scenario. A variety of soft computing techniques including gravitational search algorithm, particle swarm optimization, act colony optimization, genetic algorithms and fuzzy logic have been implemented in order to enhance the efficiency and accuracy in various aspects of protein structure prediction. In this paper, we present extensive review on few soft computing approaches and based on this study we made a comparative analysis of all the techniques over standard structural datasets. The main motive of our study is to predict the nature and behaviour of different soft computing approaches when subjected to standard dataset for solving real time complex biological problems. The results were summarized for analytical analysis between different soft computing techniques.

Keywords: Bioinformatics, Classification, Soft Computing, Structural Prediction

I INTRODUCTION

In the beginning of early 90`s, Dr. Zadeh introduced the concept of Soft Computing (SC). Soft Computing is the combination of different methodologies that are designed to solve problems related to real world. Soft Computing generally deals with the problems which are not modelled or too difficult to model, mathematically. By using soft computing methodology we can get robustness, achieve tractability and can get low cost solution. In order to have low-cost solutions, soft computing techniques exploit the tolerance for imprecision and uncertainty. The guiding principle behind soft computing techniques is to develop methods that provide low cost solutions for imprecisely or precisely formulated problem [1].

Soft computing is quite different from hard (conventional) computing. Unlike hard computing, it is tolerant of approximation, uncertainty, imprecision and partial truth. Many researchers have concluded that the role model for soft computing is the human brain. Conclusively we can say that, soft computing is an optimization technique to find solution of though problems which are very hard to answer [2].

ISSN (O) 2454-9665
ISSN (P) 2454-0663

Soft Computing (SC) is the combination of different modern methodologies that are designed in accordance with the problems that we face in modern era. As compared to mathematical techniques, SC tends to provide strong and good solution with minimum cost. As explained in previous paragraphs, soft computing generally used and applied for problem such as prediction, modelling, optimization and data mining [3]. Soft computing is an extensive method that have widely been implemented for finding similarity with biological reasoning and problem solving.

In this paper, we have presented a review of different soft computing techniques with their advantages and disadvantages. Based on these feature, we have made a comparative analysis of the performances of different soft computing techniques for predicting and classifying the protein tertiary structure. The performances were evaluated over standard structural datasets. All the soft computing approaches were made to run over the same parameters which are commonly used for experimenting biological sequences.

II DIFFERENT SOFT COMPUTING APPROACHES

In the following sub sections, we will details about the different soft computing approaches.

2.1 Optimization

Optimization is the method of framing inputs variables in order to find the minimum or maximum output. It is a process by which we can find the minimum or maximum value of a given function. Now days, optimization technique is used in all fields ranging from engineering science to economics. In scientific field, optimization processes help to maximize or minimize the given problem. With the advancement of computer in day to day life, it has emerged as an important tool for optimization problems.

2.2 Combinatorial Optimization Techniques

The entire testing problem belongs to some classes of combinatorial optimization problems. The techniques generally required to handle combinatorial optimization problems are the approximate methods and exact methods .In general, exact method solution are not considered to be good for real life problems because it requires huge computation time due to its complex nature. In comparison to exact methods, approximate methods solutions requires less time in its processing and are quite suitable for real time problems. Another method which can be applied for real life complex problems is the meta-heuristics.

2.3 Genetic Algorithm

Genetic algorithm is a type of iterative algorithms which allows an efficient and robust search. In the search process, a genetic algorithm starts with an initial state (population) in the solution space and in every search step, it produces a new and usually a better set of solutions. At each stage, GA moves forward towards producing a better solution

ISSN (O) 2454-9665
ISSN (P) 2454-0663

which may lead to minimize the change of getting trapped into a local extrema [4]. Genetic algorithms are capable of handling large and complex scale problems [5]. Some applications of genetic algorithms for solving bioinformatics related problems can be found in.The references cited above, explain the GA approach and its ability to produce optimal solution for solving bioinformatics problem. With addition to above, there are various merits of genetic algorithms which can be utilized for prediction, alignment and classification of protein, DNA and RNA sequences and their structural and behavioral study

The major elements of genetic algorithm consists of representing a solution space, a fitness function, reproduction, crossover and mutation. In every step of GA operation, the genetic operators were applied to the solution space in order to produce new and better individuals for coming generations. A search may terminate when no further improvement is observed in the coming generation as compared to its previous one or when a predefined condition is met.

2.4 Particle Swarm Optimization (PSO)

Although GAs provides very good solution quality but they don`t have the information about the optimum solution in the whole community. Therefore, the need to have information's about both the good and bad solutions is required. This requirement is fulfilled by the introduction of particle swarm optimization (PSO) [6]**.** This technique, provides an advantage that along with the local best solution, we also have the knowledge of global best solution. This is required to avoid the solution being tapped in local optima.

PSO developed by Kennedy and Eberhart .In this algorithm each particle represents a possible solution for the optimized problem. It works on the principle of maintaining several possible solutions in the search space. In each run of the algorithm, the candidate solutions are being evaluated over fitness function. The algorithm maintains a population potential where each particle represents a potential solution for the optimization problem. The PSO algorithm works by simultaneously maintaining several candidate solutions in the search space. Each candidate solution in PSO is seen as a particle "flying" through the fitness landscape in order to have maximum or minimum output from the objective function [7].

2.5 Ant Colony Optimization (ACO)

The colony optimization technique is inspired from the ant colonies [8]. Ant Colony Optimization (ACO) is a search technique, which is inspired by the pheromone of the moving ant colonies. As we all know that, the ant's moves in all possible directions in search of food and leaves behind pheromones while moving. Now the path or the directions having the large number of pheromones is considered to be the path which will be used by most numbers of ant colonies. ACO was first applied for travelling salesman problem [9]. Although, ACO suffered from the problem of premature convergence but it's being used widely around the world.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

2.6 Biogeography Based Optimization (BBO)

The study of species distribution over a biological ecosystem is known as Biogeography [10]. The species or the organism in a biological community is often depends and vary over the isolation, habitat area, latitude and elevation. The branch of biogeography that studies the distribution of plants is known as Phytogeography and those which deals about animals is known as Zoogeography .

2.6.1 Migration

Habitat or the island is defined as the places which suit the species or the solution based on some features of the habitat or the island. Feature of the habitat may vary from problem to problem. Depend on these features the immigration and the emigration of species takes place. The immigration and emigration rate is the rate by which a species can leave or join a habitat based on certain features. This feature may depends upon Habitat Suitability Index (HSI) or the Suitability Index Variable (SIV).

2.6.2 Mutation

A mutation is known to be a problem dependent mechanism and commonly are user defined. In the approach of BBO technique, mutation is used for bringing the diversity in the populations of different habitats. If a habitat having a higher set of good solutions then applying mutation technique can help the habitat to improve even more from its present's conditions. Furthermore, for a habitat having little good quality solution or less number of good solutions, a mutation can lead to improve the habitat by removing or replacing bad quality solution with the good once. Here, good or bad quality solution is judge by the scores so obtained by the fitness functions. In this thesis, we present the application of BBO for protein alignment and structure prediction. In our approach, we have applied improved mutation technique to improve the column score or to replace the residues in between the protein sequences so to get good quality solution in terms of score. Generally from the literature studies and the experimental analysis it can be concluded that, the mutation operator is more or less same as in genetic algorithm.

2.7 Gravatational Search Algorithm (GSA)

Gravitational Search Algorithm (GSA) was introduced in 2009 and is considered as one of the newest and finest technique for optimization. This population based optimization technique is based on the law of mass and gravity. In this approaches, the particle interact with each other with the help of gravity force. The agents or the particles are considered as objects and their performance is measured by their respective masses. The gravity force mentioned above, creates a global movement among all objects and all objects move towards other objects having higher or heavier masses. The slow movement of heavier masses gives the conclusion about the exploitation step of the GSA and corresponds to good solutions. The GSA are generally governed with the two equation mentioned hereunder.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

$F=G(M_1M_2/R^2)$…….(1)

a= F/M ……(2)

F in equation 1 tells about the magnitude of the gravitational force, whereas G is gravitational constant, M1and M2 represents the mass of the first and second objects respectively. R is the distance between the two objects.

While Equation 2 gives us the knowledge that when a force, F, is applied to an object, its acceleration (a) depends on the force and its mass, M.

In GSA, the agent has four different parameters which are follows: passive gravitational mass, active gravitational mass, position and inertial mass. The current position of the mass tells us about the solution of the problem, where the fitness functions is determined by the gravitational and inertial masses. The algorithm is preceded by positing the inertia and gravitational masses.

III RESULTS

In this section, we present the comparative results that are obtained using various soft computing techniques over different standard datasets.

In TABLE 1, we have presented the results over ASTRAL datasets. This is the standard structural datasets which are commonly used for predicting and classifying protein structure [10]. The two other most important datasets which stands with ASTRAL is the 1189 and 640 datasets. These three datasets are most commonly used around the world for structure prediction. The results were compared to calculate the overall accuracy in terms of percentages. The results analysis is based on calculating the overall accuracy achieved by using different methods.

In TABLE 1 we can see that the, five different soft computing approaches were implemented and results obtained by these methods were presented. TABLE 1, indicates that in overall comparison GSA came out to be the optimal method foe structural prediction. But, the performance of GSA was not optimal for all the classes mentioned in table 1. GSA is only best for classes I and II. For the classes III and IV, PSO and ACO performed well respectively. Fig. 1,2 and 3 represents bar graph representation of result comparison between different soft computing methods.

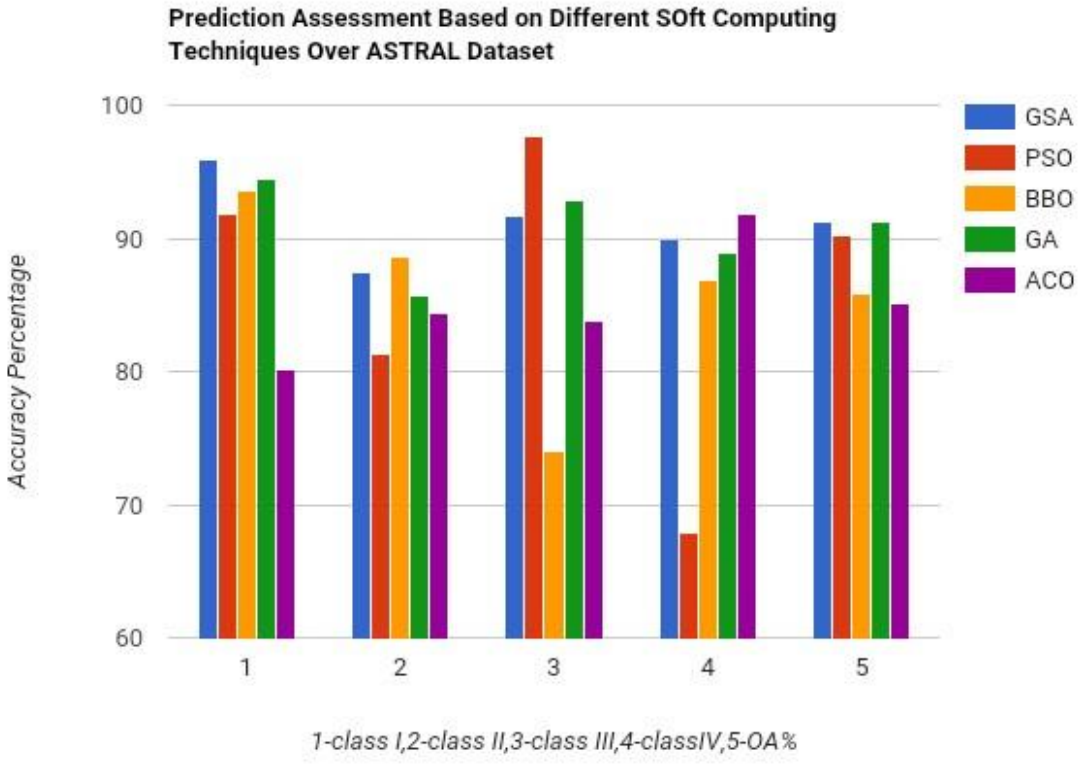
Similarly in TABLE 2, again GSA performed well for class I and II. BBO and PSO outperformed all other methods for class III and IV respectively. But, in overall comparison again GSA performed well and came out to be the best optimizing technique we encountered so far.

For dataset 640, the case is bit different than the ASTRAL and 1189 datasets. For dataset 640, GA performed well and came out to be a superior method as compared to other. As we can see that for 640 datasets, GSA only performed well for class IV and its overall accuracy also reduces. The bold faced data`s indicate the superiority of the results/methods.

International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 01, January 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

TABLE I: Prediction Assessment Based on Different Soft Computing Techniques over ASTRAL Dataset

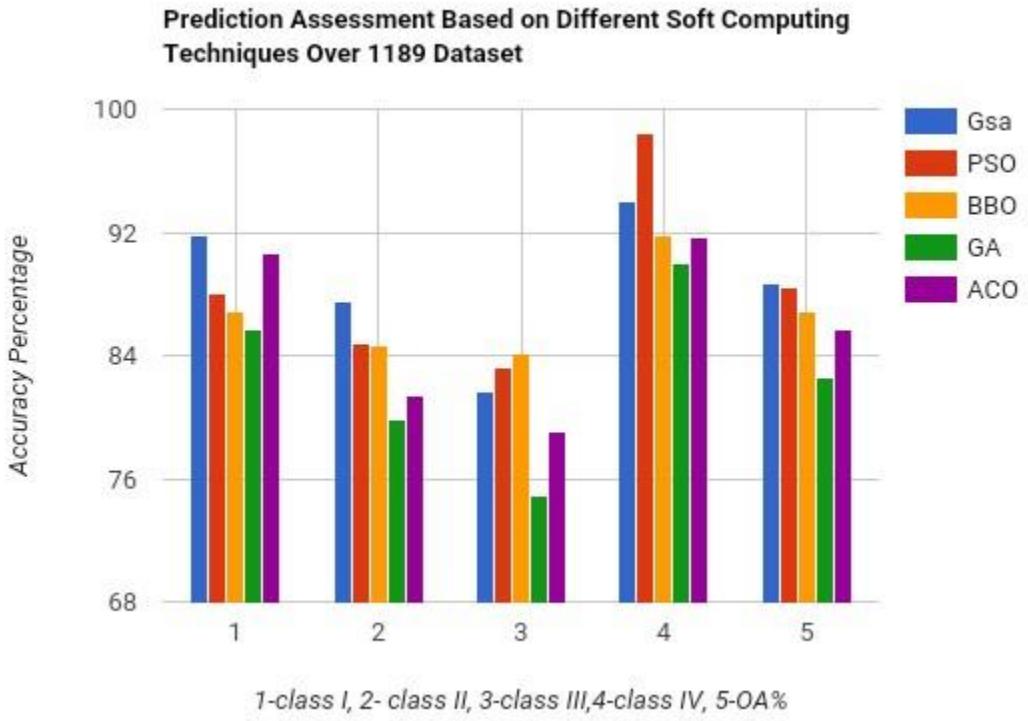
SOFT COMPUTING TECHNIQUES	Class I (all-α)	Class II (all-β)	Class III (α+β)	Class IV (α/β)	Over All Accuracy {O A (%)}
GSA	**95.87**	**87.45**	91.67	89.90	**91.22**
PSO	91.87	81.34	**97.67**	67.88	90.19
BBO	93.64	88.56	73.98	86.91	85.77
GA	94.45	85.67	92.78	88.89	91.19
ACO	80.21	84.32	83.78	**91.87**	85.04



International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 01, January 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

TABLE II: Prediction Assessment Based on Different Soft Computing Techniques over 1189 Dataset

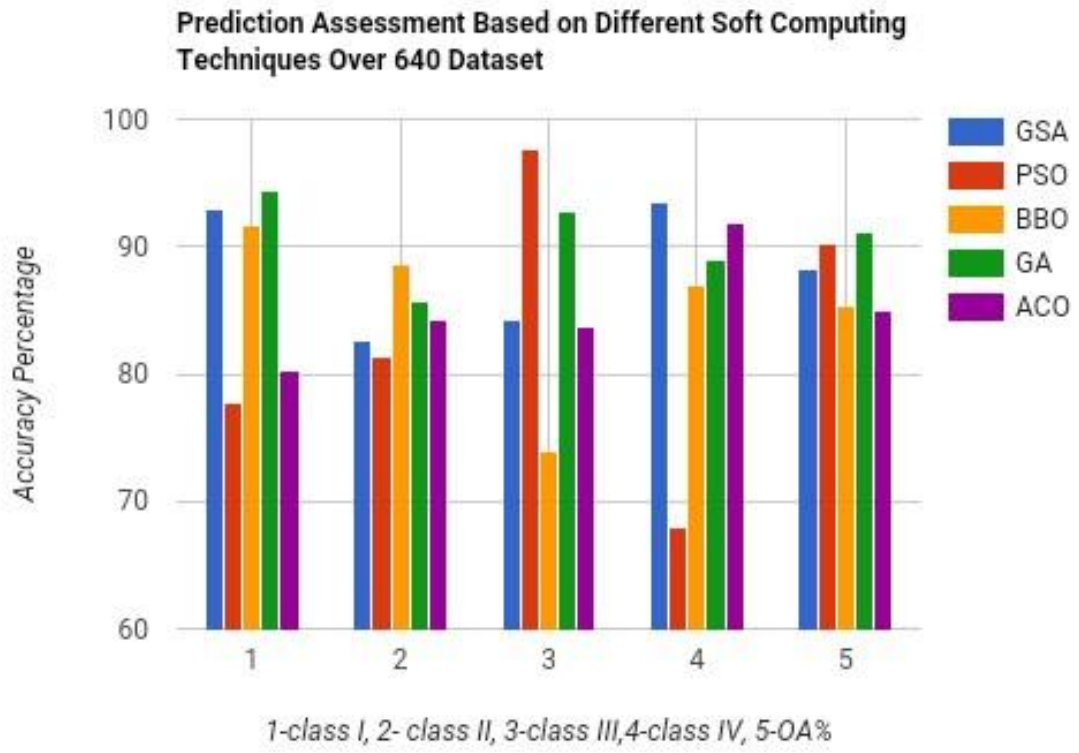
SOFT COMPUTING TECHNIQUES	Class I (all-α)	Class II (all-β)	Class III (α+β)	Class IV (α/β)	Over All Accuracy {O A (%)}
GSA	**91.77**	**87.45**	81.65	93.98	**88.71**
PSO	87.98	84.77	83.23	**98.45**	88.60
BBO	86.88	84.65	**84.12**	91.77	86.85
GA	85.70	79.87	74.89	90.01	82.61
ACO	90.65	81.39	79.05	91.72	85.70



International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 01, January 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

TABLE III: Prediction Assessment Based on Different Soft Computing Techniques over 640 Dataset

SOFT COMPUTING TECHNIQUES	Class I (all-α)	Class II (all-β)	Class III (α+β)	Class IV (α/β)	Over All Accuracy {O A (%)}
GSA	92.93	82.56	84.32	**93.44**	88.31
PSO	77.81	81.34	**97.67**	67.88	90.19
BBO	91.64	**88.56**	73.98	86.91	85.27
GA	**94.45**	85.67	92.78	88.89	**91.19**
ACO	80.21	84.32	83.78	91.87	85.04



IV CONCLUSION

We have presented the role of different soft computing approaches in solving structure prediction problem of protein sequences. By visualizing the literature review and results, we can easily conclude that each approach has its own way of reacting and solving a problem. Due to the different approach adopted by the soft computing techniques,

ISSN (O) 2454-9665
ISSN (P) 2454-0663

their outcome often varies. In this paper, GSA approach has performed better in achieving highest accuracy as compared to other methods in most of the cases. However, methods like PSO, ACO, GA and BBO have also performed well for some classes. At the last, it can be concluded that the soft computing approaches are quite efficient and can easily be modelled according to the need of the problem.

REFERENCES

[1].T. Ito, "Soft computing approaches towards design and decision support applications," 2007 IEEE International Fuzzy Systems Conference, London, 2007, 1-6.

[2].H. C. Huang, "Fusion of Modified Bat Algorithm Soft Computing and Dynamic Model Hard Computing to Online Self-Adaptive Fuzzy Control of Autonomous Mobile Robots," in *IEEE Transactions on Industrial Informatics*, 12 (3), 2016, 972-979.

[3].T. Anand, R. Pal and S. K. Dubey, "Data mining in healthcare informatics: Techniques and applications," 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, 4023-4029.

[4].A. Sonak, R. Patankar and N. Pise, "A new approach for handling imbalanced dataset using ANN and genetic algorithm," *2016* International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, Tamilnadu, India, 2016, 1987-1990.

[5].H. Aliee, S. Vitzethum, M. Gla?, J. Teich and E. Borgonovo, "Guiding Genetic Algorithms using importance measures for reliable design of embedded systems," 2016 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), Storrs, CT, 2016, 53-56.

[6].A. A. A. Saed and W. M. N. W. Kadir, "Applying particle swarm optimization to software performance prediction an introduction to the approach," 2011 Malaysian Conference in Software Engineering, Johor Bahru, 2011, 207-212.

[7].A. Andalib and S. M. Babamir, "A new approach for test case generation by discrete particle swarm optimization algorithm," 2014 22nd Iranian Conference on Electrical Engineering (ICEE), Tehran, 2014,1180-1185.

[8].P. Janacik, D. Orfanus and A. Wilke, "A Survey of Ant Colony Optimization-Based Approaches to Routing in Computer Networks," 2013 4th International Conference on Intelligent Systems, Modelling and Simulation, Bangkok, 2013, 427-432

[9].B. Jin, L. Zhang and S. Qian, "An Improved MMAS Algorithm for Path Optimization in Emergency Rescue," 2010 International Conference on Machine Vision and Human-machine Interface, Kaifeng, China, 2010, 612-616.

[10]. R. Srinivasan and G. D. Rose, "Protein structure prediction — An Ab initio approach," 2003 European Control Conference (ECC), Cambridge, UK, 2003, 3378-3385.