GENOME STRUCTURE PREDICTION A REVIEW OVER SOFT COMPUTING TECHNIQUES

Amandeep Sharma[1], Amanpal Singh[2]

[1]*Department of Computer Science & Engg .RIEIT Ropar India*
amansharma10408@gmail.com
[2]*Department of Computer Science & Engg. RIEIT Ropar India*
amanpalrayat@gmail.com

Abstract

There are some techniques like spectrometry or crystallography for the determination of DNA, RNA or protein structures. These processes provide very accurate results for the structure estimation. But these conventional techniques are very slow and could be applied over a few special cases only. Soft computing techniques guarantee a near appropriate results in much smaller time and have very large applicability. These techniques are much easier to apply. Different approaches have been used in soft computing including nature inspired computing for estimation of genome structures with a considerable accuracy of results. This paper provides a review over different soft computing techniques been applied along with application method for the determination of genome structure.

Keywords—*DNA, RNA, proteins, structure, soft computing, techniques.*

----------***----------

1. INTRODUCTION

Soft computing deals with having a near optimal results for decreased cost, effort and time. A much lesser time requirements and computing power may be required using the soft computing techniques. Some of the soft computing techniques are like genetic algorithm (GA), evolutionary algorithm, artificial neural networks (ANN), fuzzy logic (FL), ant colony optimization (ACO), artificial bee colony (ABC), cuckoo search (KS), tabu search (TS), particle swarm optimization (PSO). The techniques like artificial neural networks or fuzzy logic take into account the direct application of heuristics in their applications. While the approaches like GA, evolutionary algorithm, ACO, ABC, KS, TS and PSO first make some assumptions in the form of heuristics and then may be refining those heuristics to achieve better results. These generally make random searches by some instances of their agents. The instances move in the search space and have some mechanism to compare their results. After some time or passes of running the application a final result which is a near optimal one is generated from these agents. A lower intelligence of agents gives the result of higher level of intelligence. This behavior of achieving intelligence by simulating the techniques in nature is known as the nature inspired computing. These nature inspired techniques have a major role in the soft computing.

A huge amount of biological data is available. This include the sequences of some DNA, RNA and proteins. The primary, secondary and tertiary structures of these genome components could be estimated from these gene sequences. While, the ideas from these could be used for further estimating the sequences of other genome components. There are always some uncertainties for the estimation of these sequences or structures from techniques of estimation like mass spectrometry. The soft computing techniques play a very robust role in overcoming these uncertainties. Final results achieved using the soft computing are nearly accurate, even after having a number of uncertainties in the techniques which have been used earlier (like crystallization) for various estimates in the genomes.

A number of techniques have been used for the detection of structures and sequences of genome. There are conventional techniques like x-ray crystallography [1], nuclear magnetic resonance spectroscopy (NMR)[2]. These techniques further have strategies like small angle x-ray solution scattering [3], in-line probing [4], etc. The conventional techniques are very time consuming and are also very costly. Also not all the RNA and proteins could be crystallized. These are the major limitations in the conventional methods, although very accurate results may be achieved using these techniques. These accurate data from the conventional techniques could be used in the soft computing for achieving very accurate results. The soft computing techniques have very wide applicability and the results obtained from the application of these techniques are also very efficient. Soft computing could also handle the uncertainties in the outputs from conventional methods. Various soft computing techniques like K-mean clustering [5], ANN [6], FL [7], GA [8], Simulated Annealing (SA) [9], PSO [10], Accelerated Particle swarm optimization (APSO) [11], ACO [12], TS [13] have been applied for structure prediction.

1.1 Techniques For Estimating Genome Structures

There are a number of soft computing techniques been used for the prediction of genome secondary or tertiary structures. The soft computing techniques allow the calculation of

structures upto more than 90% efficiency in certain cases of secondary structure estimates. The accuracy drops as the sequence length for the given genome sequence increases. The genome sequences are predicted based on the minimum free energy estimates or these may predict based on the basis of predicted locations for folding of given genome. Some of the techniques used in soft computing for the structure estimation are as follows:

1.2 A For Genome Structure Prediction

In GA [8] the principle of natural selection i.e. survival of the fittest is followed. The algorithm is adaptive in nature and is based on natural selection genetics. These have a large amount of parallelism and start with some of the user defined components. The components consists of alternate close corresponding solutions to the problem. The GA could be applied for the optimization of some parameters, hence could be applied for large number of real world problems. GA are used effectively when :

1. There is very large search space and also it is very complex.
2. The conventional methods could not provide the optimal results in reasonable time.

The GA has been used for the RNA secondary prediction [14] and also for the estimation of protein structures. In GA each of the substructures are provided an integer number in the sequence. Some conflicting sequences may be generated using the GA algorithm which could be removed using some other optimization. Van Batenburg et al. [14] have developed GA for the prediction of RNA secondary structure. The method was based on free energy minimization and possible RNA folding. Several possible solutions to the problem are taken in the form of array with sequence number for each substructure. Each of the possible sequences for structure are given value 1 and the substructures which are not possible at all are given the value 0. The sequences are then crossed with other components in such a way that the structures with the value 0 are not carried any further but those with the value 1 are crossed further making new combinations. Initially this is applied to small part of RNA and then it is increased for the other iteration. The final result after the whole process results in the formation of a number of possible structures in the arrays. The required number of possible solutions could be taken from these children. If larger number of children are there then the structures which are matching in all the possible solutions are taken while the non matching are randomly dropped from the set of solutions. The given technique could be efficiently applied in a number of RNA classes.

1.3 Ann For Genome Structure Prediction

It is composed of a system containing number of operating elements known as nodes. These nodes function in parallel and emulate biological neural network. ANNs are used for function approximation, prediction, classification, feature extraction, and clustering. The artificial neural networks could be categorized as supervised neural networks and unsupervised neural networks. In the supervised neural networks the networks are trained with data which could provide useful features to the neural network. This helps the neural network to detect the higher order correlations in the data. Biological systems with non-linear characteristics are best suited to apply the supervised artificial neural network technique. Unsupervised neural networks are good for the feature extraction and clustering. These do not require a previous knowledge about the class of data. These allow an unsupervised learning and are easier to apply. Some of the features in ANN are :

1. These allow adaptation with new patterns of data.
2. Have tolerance to distorted data
3. If data at some node goes wrong, it does not have impact over performance
4. High speeds could be achieved because of parallelization
5. Greater examples lead to error minimization

S. Le et al. [15] have defined a tree system for the prediction of RNA structure. The stems in RNA been represented as edges, loops and bulges in RNA as vertices of degree two and junctions of vertices more than two degree. The structure could also be applied to the estimation of protein structures. Back propagation of the achieved results after passing through weighted nodes is used in training until the error is near to 0.

1.4 Genome Structure Prediction Using Fl

In FL we take multiple values for logic. The logic values in the case of FL are approximate. The membership variables in FL can have any values between 0 and 1. The non numeric variables, like high, low, medium etc. may also be used. The statements of if-then-else may be used for the derivation of results from the variable values. Some of the features of FL are:

1. These allow approximate results even in the case of ambiguity and low data availability
2. A high level of complexity could be handled

In an approach by D. Song et al [16] the dynamic programming is used along with fuzzy logic for the prediction of the RNA structures. In this approach at first all the possible base pairs are kept in triangular matrix. Sixteen such matrices would be generated for the base pairs. A probability value is provided to each of the pairs based on the occurrence of such base pairs e.g. the fuzzy value for AU pair would be high and for the AA pair would be low. For a particular position, a position specific calculation is made and the given pair is assigned to that location. Then positions are iteratively updated for the position matrices and the base pair structure is updated to give the final optimal structure. The given algorithm would also allow the addition of the base pairs into the structure to arrive at more sequences in the case of DNA sequences. The structure around the given sequence may be added later on in the given results.

1.5 PSO For Genome Structure Estimation

PSO technique is a nature inspired soft computing technique based over the bird flocking or the fish swarm spooling. In PSO a number of swarm particles or agents move randomly in the search space. At each of their locations the particles check for some condition of optimality. If the condition at the current location of swarm particles is lower than the condition of optimality at some of its previous locations, then the previous maximum is kept by the particle. The particle makes a comparison for the optimal value of its neighboring particles (the values may be having different weights for the comparison) and accordingly adjusts its velocity in the search space. At the end of certain time of flocking through the search space, the swarm particles may be oscillating at some optimal values in the search space.

PSO have not been implemented in its generic form for the structure prediction but SetPSO [17] and fuzzy logic based PSO[18] have been implemented for the structure prediction. In the fuzzy based PSO the structure is represented as combination of stems and then the free energies are minimized to predict the final structure. The globally best velocity vector along with the best unchanged fitness particle are used for the input into the fuzzy system. The fuzzy system then decides the learning parameter, particle velocity and the different weightage to be given to various neighbors in PSO.

A modified PSO, APSO have also been used for the structure prediction . In APSO the global maxima for the particles is taken into the consideration as compared with the generic PSO. Care for prevention of early termination is also taken. The learning parameter, particle velocity and the weightage to the various particles is decided globally and the particle velocity is adjusted accordingly. The APSO provides equivalent results as compared to the fuzzy PSO with much lower computational overheads as compared to the fuzzy PSO.

1.6 ACO For Genome Structure Prediction

ACO is another nature inspired computing technique been used for the genome structure prediction. This technique is based over the ants being able to search for the food resource in nature. When a searching ant finds a food source, it comes to its colony and picks a drone with itself for the location of food. The drone then brings the particle along with releasing pheromone trails to the food location. More ants go for the food locations by accessing the pheromone trails. If there are multiple paths to the food location then the pheromone over the shortest paths would fade the slowest and hence over time more and more ants are travelling over the shortest path to food location. At the end only the shortest path to the food location remains as the pheromone trails from the other paths fade away. Also if the food depletes, then the ants leave no pheromone on return, making ACO good for dynamic estimates.

N. McMilan have devised a technique based over ACO [19]. At first all the stems (straight sequences) are identified in given genome using a brute force method. Then new stems are added to the given stems to form the probable secondary structure by ants. The probability with which an ant would be adding a stem to the previous stems is based over the pheromone trail and the type of previous stem in the structure of genome. The process is repeated for a number of ants. The ant having lowest deviation in the free energy is given priority and is chosen as the best possible structure for the given case.

1.7 Other Techniques For Structure Prediction

Some of the other techniques which have been used for the genome structure estimation are as follows:

1. TS: In TS the formations in the structures are classified in the form of bases. At first a structure with longest linear sequence is taken. Then more structures are added to this structure based over the intensified search, which form a neighboring solution to the problem. A tabu list is maintained to stop repeating the recently added solutions in the problem. When all the neighboring solutions for the given stem have been obtained, these are arranged in ascending order of free energy. The initial structure is then modified using the values from the minimum free energy.
2. K-nearest neighbor classifier: It is based over the K-nearest neighbor voting in the feature space. The voting starts with the multiple sequences been generated for the structure. The majority of voting is selected as the first result. Then a consensus probability matrix is generated. The next consensus would use the results from the previous consensus and probabilistic values for the addition of new base pairs. Finally, the structure corresponding to best consensus score is selected as the best solution.

1.8 Calculating The Effectiveness Of Algorithms

The following parameters are used for prediction of algorithm effectiveness for the case of soft computing algorithms:

[1]. No. of accurately predicted base pairs: The DNA, RNA or proteins occur in the form of pairs. The usual pairs are A with T and G with C for DNA, but there is also possibility of other types of pairs. The algorithm which efficiently predicts the pairs weather usual or unusual inside the given sequence of genome is assumed to be better. The greater is the number of accurately detected base pairs the better the algorithm is.

[2]. Minimum Gibbs free energy: There is free energy associated with molecules which make those molecules stable. When an energy equivalent or greater is available the molecules react further. There is a minimum Gibbs free energy associated with a molecule in equilibrium state. The algorithm which is able to predict the minimum stable free energy for the given molecule is assumed to be more accurate for given conditions.

[3]. True positive number of base pairs: The base pairs which have accurate locations in the given sequence of genome with respect to the central location are known as true positive base pairs.

Sensitivity: The sensitivity of algorithm is calculated as:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

The above parameters could be compared with the conventional technique results for calculation of effectiveness.

2. FUTURE POSSIBILITIES

Structure prediction is important for drug development; it also helps in better classification and tracking mutations in the genome. The techniques like cuckoo search and ABC have not been effectively used for the prediction of genome structure. These techniques have the potential to be effective for the genome structure prediction. A work by J. Agrawal et al.[11] on APSO shows using some global procedure for the structure prediction could also be effective as compared to other algorithms. The global detection mechanism contained in the ABC algorithm, so becomes a good contender for the genome structure prediction. Also the ABC have a good solution pointing mechanism as all the agents are recording their observations. The ABC also overcomes the premature convergence as in the case of APSO. The cuckoo search have some mechanisms similar to genetic algorithm which makes the algorithm also a good contender for the structure prediction. There also has not been much focus on the hybrids of the nature inspired computing techniques for the genome structure prediction. Some structures in genome for e.g. RNA could be better estimated using some algorithms, other with some other algorithms. There is a scope for achieving good prediction results from hybrids.

3. CONCLUSION

The soft computing techniques are very fast as compared to the conventional techniques. A number of methods have been devised for adapting the soft computing techniques like GA, PSO, APSO, etc. to make these algorithms capable of calculating the genome structures. The strategies are like assigning tree or graph forms to the various structure formations, using integer values, using matrices or multi-dimensional arrays for the locations of various folds in the structures. These also include the probability of occurrence of bonds between the various pairs. The usual pairs having a high probability of occurrence, while the unusual pairs having the minimal probability of occurrence. A few techniques start with having prediction about small portions of molecules to be identified, while some techniques like GA use a large predicted genome sequence at beginning and then go by modifying the portions of the genome. These soft computing techniques could be tested for effectiveness by using the some parameters. These parameters include predicting the correct base pairs, having the exact location of base pairs as in comparison with the center of genomes, also could be based over the minimum free energy in the genome bonds for a complete genome sequence, could also be defined in the form of algorithm sensitivity. The algorithm as GA provides very good results as far as detection of true pairs of genes is concerned. The centralized detection algorithms like APSO have good results for relative location detection of genome structures. While some other soft computing techniques which are based over the conventional methods would also be very important for estimating the structure of genomes.

The soft computing techniques like ABC and cuckoo search could also be very important for the estimation of genome structures. ABC is good for pointing a particular solution's location in the search space. It is an algorithm for the global estimation of the feasible solutions. While, the cuckoo search have some of the good features of GA. This makes the cuckoo search also a good contender for the genome estimation. A new implementation of techniques could be proposed for the genome structure estimation; this could be one of the techniques of ABC or cuckoo search, as these could have a better efficiency as compared to the previous techniques (based on their features). A hybrid of techniques like GA or ACO or PSO could also be effective for the structure estimation.

REFERENCES

[1]. S.H. Kim, G. Quigley, F.L. Suddath, and A. Rich, "High-Resolution X-Ray Diffraction Patterns of Crystalline Transfer RNA that Show Helical Regions," Proc. Nat'l Academy of Sciences USA, vol. 68, pp. 841-845, 1971

[2]. A.E. Ferentz and G. Wagner, "NMR Spectroscopy: A Multifaceted Approach to Macromolecular Structure," Quarterly Rev. of Biophysics, vol. 33, pp. 29-65, 2000

[3]. R.P. Rambo and J.A. Tainer, "Improving Small-Angle X-Ray Scattering Data for Structural Analyses of the RNA World," RNA,vol. 16, pp. 638-46, 2010.

[4]. E.E. Regulski and R.R. Breaker, "In-Line Probing Analysis of Riboswitches," Methods Molecular Biology, vol. 419, pp. 53-67, 2008.

[5]. I.L. Hofacker, "Vienna RNA Secondary Structure Server," Nucleic Acids Research, vol. 31, pp. 3429-3431, 2003

[6]. G.P. Zhang, "Neural Networks for Classification: A Survey," IEEE Trans. Systems, Man and Cybernetics, Part C, vol. 30, no. 4, pp. 451-462, Nov. 2000.

[7]. L.A. Zadeh, "Fuzzy Sets," Information and Control, vol. 8, pp. 338-353, 1965.

[8]. Deb, K, Pratap, A. Agarwal, S. Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Transactions on, 2002.

[9]. S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by Simulated Annealing," Science, vol. 220, no. 4598, pp. 671-80, 1983.

[10]. Kennedy, J.; Eberhart, R.. "Particle Swarm Optimization". Proceedings of IEEE International Conference on Neural Networks 1995.

[11]. Acceleration based Particle Swarm Optimization (APSO) for RNA Secondary Structure Prediction, J. Agrawal, S Agrawal - Progress in Systems Engineering, 2015 – Springer

[12]. M. Dorigo, Optimization, Learning and Natural

Algorithms, PhD thesis, Politecnico di Milano, Italy, 1992.

[13]. Y. Liu, J. Hao, and J. Peng, "Predicting RNA Secondary Structure with Tabu Search," Proc. IEEE Int'l Conf. Cognitive Informatics,pp. 409-414, 2010.

[14]. F.H. Van Batenburg, A.P. Gultyaev, and C.W. Pleij, "An APLProgrammed Genetic Algorithm for the Prediction of RNA Secondary Structure," J. Theoretical Biology, vol. 174, no. 3,pp. 269-280, 1995.

[15]. D.R. Koessler, D.J. Knisley, J. Knisley, and T. Haynes, "A Predictive Model for Secondary RNA Structure Using Graph Theory and a Neural Network," BMC Bioinformatics, vol. 11,pp. S6-S21, 2010.

[16]. D. Song and Z. Deng, "A Fuzzy Dynamic Programming Approach to Predict RNA Secondary Structure," Proc. Sixth Int'l Conf. Algorithms in Bioinformatics, pp. 242-251, 2006.

[17]. M. Neethling and A.P. Engelbrecht, "Determining RNA Secondary Structure Using Set-Based Particle Swarm Optimization," Proc. IEEE Congress Evolutionary Computation, pp. 6134-6141, 2006.

[18]. C. Xing, G. Wang, Y. Wang, Y. Zhou, K. Wang, and L. Fan, "Psofold: A Metaheuristic for RNA Folding," J. Computational Information Systems, vol. 8, pp. 915-923, 2012

[19]. N. McMillan, "Rna Secondary Structure Prediction Using Ant Colony Optimisation," master's thesis, School of Informatics,Univ. of Edinburgh, pp. 1-63, 2006.