IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-5, Issue-7 E-ISSN: 2347-2693

A Survey of Metaheuristics Approaches for Application in Genomic data

Manu Phogat[1*], Dharmender Kumar[2]

[1*]Dept. of CSE, Guru Jambheshwar University of Science & Technology, Hisar, India
[2]Dept. of CSE, Guru Jambheshwar University of Science & Technology, Hisar, India

[*]*Corresponding Author: kunjean4181@gmail.com, Tel.: +91-8285557267*

Available online at: www.ijcseonline.org

Received: 04/Jun/2017, Revised: 14/Jun/2017, Accepted: 20/Jul/2017, Published: 30/Jul/2017

***Abstract*—** the present era is the revolutionary time in genomic applications. In recent years, genomes of various species have been sequenced; genes and proteins have been mapped and learned. Structures of genes and proteins have been implied and their behavior is being understood. Over the past two decades, there is a viable interest in to analysis of gene sequence and microarray data with the help of metaheuristics techniques. Therefore this survey intended to give some nature inspired methods to analyze genomic data such as sequence analysis of various genes, microarray analysis and multiple sequence alignment. The survey later on is followed by the types of main nature inspired algorithms both population and single solution based methods. These are followed by their different application in genomic data and their merits to address specific task.

***Keywords*—** Metaheuristics, Microarray, genome, genetic algorithm

I. INTRODUCTION

The term genome specify to the complete implementation of DNA for a given species such as human, mice, rice etc. Basically the human genome consists of 23 pairs of chromosomes [1], and roughly contains 20000 protein coding genes [2], and 25000 noncoding genes [3]. The human genome successfully sequenced in 2003 from then large range of biological data sources have emerged, resulting in inconsistency of data formats [4]. The exponential rise in the volume of such data has required the use of computational techniques for information gathering and retrieval [5]. Most of the genomic data related task is defined as hard combinatorial problems. So the requirement of metaheuristics and other approximate techniques is mandatory. In short a metaheuristics [6, 7] can be determined as a top-level general approach which guides other heuristics to search for good solutions. Metaheuristics algorithm is usually said to be non-deterministic.

The word metaheuristic derived from two Greek words, Meta means "beyond, in an upper level" and heuriskein means "to find". Thus metaheuristics methods are combine the high level concepts in heuristics [8]. In the 1970s, metaheuristics have been developed to merge basic heuristic techniques in superior level frameworks to explore a search space in an efficient and an effective way. There are two classes in metaheuristics, Trajectory based methods also called as single solution based and population based methods. The common Trajectory based methods are iterated local search (ILS), variable neighborhood search (VNS), Simulated annealing (SA), Tabu search (TS) etc. While the population based methods commonly consists of genetic algorithms (GAS), particle swarm optimization (PSO), scatter search (SS), ant colony optimization (ACO), Estimated Distribution Algorithms (EDAs) etc. The important feature of designing any metaheuristics methods is its capability of performing wide diversification and deep accretion. The term diversification generally tends to the exploration of search space, whereas the term accretion tends to the exploitation of the accumulated search experience.

SINGLE-SOLUTION BASED METAHEURISTICS METHODS

The single solution based methods are also called as trajectory methods. They apply the generation of the neighbourhood solutions iteratively from the present single solution. The process iterates until a given stopping criteria. The common examples of single solution based methods are guided local search (GLS), variable neighbourhood search (VNS), tabu search (TS), simulated annealing (SA), iterated local search (ILS).

TABU SEARCH

The basic concept of Tabu Search (TS) is established on ideas proposed by Fred Glover (1977, 1986) [9]. Tabu Search is relay on the presumption that problem solving, to certify as intelligent, must incorporate adaptive memory and sensible exploration. The important feature of TS method is the use of memory, which document the information related of the search process. TS generate a neighborhood solution from the present solution and take the optimum solution even if is not improving the current solution.

Simulated annealing

The concept of Simulated annealing (SA) proposed by Kirkpatrick [10], SA is probably the most widely used meta-heuristic in combinatorial optimization problem. The terminology and idea come from a technique called as annealing in metallurgy which involves heating and controlled cooling of a material that enhance the size of crystals and minimize their defects. In SA at each step, the current solution is exchange by another one that improves the objective function, randomly chosen from the neighborhood. It is a technique to proximate the global optimum of a given function which based on probability. Specifically, the technique is a nature inspired approach to proximate global optimization in an enormous search space. The technique is used when the search space is discrete. The main objective of SA method is to breakout from local optima and so to delay the convergence.

Variable Neighborhood Search

The concept of Variable neighborhood search (VNS) method proposed by Hansen and Mladenovic [11].It explores distant neighborhoods of the present necessary solution, and advance from there to a new solution if and only if a progress was made. The local search method is practiced again and again to get from solutions in the neighborhood to local optima. VNS was made for proximate solutions of discrete and continuous optimization problems

POPULATION-BASED META-HEURISTICS TECHNIQUES

These techniques start from an initial population of solutions. The common difference between population based and single solution based metaheuristics methods is that the population base starts with population of solutions but other start from a single solution. When the initial population is originated, the replacement phase is started by selecting a new population from the previous population. Many of the population based methods are nature inspired methods. The most popular are particle swarm optimization (PSO), ant colony (AC), evolutionary algorithms (EAs), deferential evolutionary (DE), group search optimizer (GSO).

Swarm Intelligence

Swarm behavior commonly seen in flocking of birds, insects, as well as in fish schools. The global nature of a swarm (group) of social organisms therefore emerges in a nonlinear manner from the behavior of the individuals in that swarm. Thus, there exists a strong bonding between individual behavior and the behavior of the entire group. There are many algorithms belong to SI such as), ABC (Artificial Bee Colony), PSO (particle swarm optimization), ACO (ant colony optimization), artificial immune systems, Bee colony optimization, BAT algorithm etc. In the following subsections we outline two of these algorithms, PSO, GSO and ABC algorithms**.**

Particle swarm optimization

PSO is an intelligent optimization algorithm belongs to population based nature inspired algorithms. PSO was initially proposed by Eberhart and Kennedy in 1995 [12]. PSO basically based on paradigm of swarm intelligence and inspired by social behavior of animals like fish and birds. PSO contains a population of candidate solutions called a swarm. A swarm consists of no of particles and every particle is a candidate solution to the problem. The particles fly over the search space with each having certain velocity and position. The main idea is to find out best solution among all the possible solutions.

The main steps in PSO algorithm

1. Initialize population in hyperspace.
2. Evaluate fitness of individual particles.
3. Change the velocities based on previous best and global (or neighborhood) best.
4. Terminate on some condition.
5. Go to step 2.

Artificial Bee Colony

The ABC is an optimization algorithm based on the intelligent foraging behavior of honey bee swarm. ABC proposed by Karaboga in 2005. ABC algorithm basically work in five phases: initialization phase, employed bee phase, probabilistic selection phase, onlookers bee phase and scout bee phase [13]. In ABC the colony consists of three groups of bees; employ bee, onlookers and scouts. Number of employed bees and colonies equals to the no of food source around the hive. The searching of new food source carries out by scout bees [14]. ABC is population based algorithm. The position of a food source represents a possible solution to the optimization problem. The nectar amount of the food source determines its quality.

Group search optimizer

A new swarm intelligence algorithm called as Group search optimizer (GSO) proposed by Saunders [15].group search optimizer (GSO) motivated by behavior of animals, especially their seeking (foraging) behavior. Their searching behavior may be expressed as an active movement by which an animal seek or pursuits to find resources such as food, mates or nesting sites. The group stands for population in the algorithm and each one in the group is called a member. Mainly three kinds of members in a population of GSO algorithm first are scroungers, producer, and rangers (dispersed) members. Only a single producer at each search iteration and other members are scroungers and rangers members.

Evolutionary Algorithms

The Evolutionary algorithms (EAs) are problematic (Population-metaheuristics) that have been profitably applied to many complex and real problems. EAs are based on the notion of competition. They are based on the evolution of a population of individuals this population is usually developed randomly. Each one of the member in the population is evaluated by using an objective function (fitness function). The most common EAs are genetic algorithm and differential evolution.

Genetic algorithm

The concept of GA is developed by Holland in the 1970s to figure out the adaptive processes of natural systems [16]. Genetic algorithm (GA) is nature inspired algorithm rely upon the concept of natural selection that dwells to the higher class of evolutionary algorithms (EA). The common use of GA is to generate high-aspect solutions to optimization and search problems which depend on biological motivated operators such as mutation, crossover and selection. After developing the initial population randomly, the algorithm expands through three operators, which are selection which relates to survival of the fittest, crossover which represents mating among individuals and mutation which introduces random modifications.

II. METAHEURISTICS AS A TOOL FOR GENOMIC DATA APPLICATIONS

The coming sections tell how metaheuristics methods described above can apply with following genomic data applications.

Selecting Genes from Gene Expression Data for various diseases Classification.

Gene selection is important aspect for gene expression in various diseases especially in cancer (tumor) classification. A large datasets of Hundreds of genes are produced by microarray experiments with interpreted values in order to be useful to predict cancer. In microarray most of the may be irrelevant genes or noisy genes which make these genes difficult to analysis. Many efficient metaheuristic methods such as evolution algorithms (EAs), GAs, simulated annealing (SA), Tabu search (TS), and particle swarm optimization (PSO). In [17] a tabu search and hybrid PSO method is proposed for selecting genes for cancer classification, the method is called (HPSOTS). The PSO based methods are intractable to efficiently produce a subset of very few descriptive genes for high classification accuracy and Tabu search has the ability to avoid convergence to local minima and it enhance the exploitation process of the algorithm. A novel approach based purely on a GA to predict genes is described in [18]. Here the fitness function calculated using site and the content statistics are based upon positional weight matrix and in-frame hexamer frequency. The results of experiment show that the system obtained moderately good results at the nucleotide level. A new unguided filter based gene selection technique called as microarray gene selection based on ant colony optimization (MGSACO) for microarray data classification [19]. The method is an iterative improvement process where a population of agents selects a subset of genes at each iteration and the performance of the found subsets of genes is evaluated using a new proposed fitness function without using any learning model.

Multiple Sequence Alignment and sequencing

The MSA is a task of comparing sequences of nucleic or amino acids and finds the similarity in the structure between genes and protein. It also predicts the 3D structure of protein. GA is used [20] for MSA, and an alignment is the combination of two (multiple alignment) or more (pairwise alignment) sequences of 'residues' (amino acids or nucleotides) that raised the equality between them. Algorithmically, the problem consists of extending and opening loops in the sequences to enhance an objective function (measurement of similarity). A new PSO based training method [21] for Hidden Markov models (HMMs) for solving the MSA problem. The most critical problem of computational and molecular biology is DNA sequencing. Its goal is to specify a nucleotides sequence an analyzed DNA sequence consists of. In a basic DNA sequencing one has to find out whether a given nucleotide is part of the DNA sequence or not. A model proposed by [22] using tabu search to determine if a given oligonucleotide appears once, twice or at least three times in the target sequence. The tabu search algorithms outperform the previously used greedy algorithm.

The RNA Secondary Structure Prediction

The RNA (ribonucleic acid) is a polymer of nucleic acids like DNA consisting monomers of nucleotides. The nucleotides of RNA contain rings of ribose and uracil unlike DNA which consist of deoxyribose and thymine. The secondary structure of RNA is formed when an RNA molecule folds to form secondary structures owing to hydrogen bonding between complementary bases on the same strand. The Secondary structure of RNA molecules can be foreseen computationally by evaluating the MFE (minimum free energy) structure for hydrogen bonds of all different combinations and domains. The Modified PSO [23] is used to optimize the structure of RNA molecules; the author proposed the set PSO which operate on mathematical sets in order to solve set-based combinatorial optimization problems in discrete search spaces.

Fragment Assembly Problem

The FAP deals with sequencing of DNA. In sequencing the multiple exact copies of original DNA are made, each copy is then cut into short fragments at random positions. The FAP is then to reassemble the original molecule's sequence from smaller fragment sequences. Mainly the FAP is permutation problem and also NP complete in nature. The ACO technique is designed to solve fragment reordering problem. The other metaheuristics like genetic algorithm (GA) scatter search (SS) [24] and simulated annealing (SA) [25] are also used to solve FAP.

Table. 1

Common issues In Genomic Data	Metaheuristics algorithms
Sequence comparison and alignment	EA, MA, ACO, PSO
DNA fragment assembly	EA, SA, ACO
Gene Finding and Identification	EA
Gene Expression Profiling	EA,MA, PSO
Structure Prediction	EA, MA, SA, EDA
Phylogenetic Trees	EA, SS, MA

EA- Evolutionary Algorithm, MA- Memetic Algorithm, ACO- Ant Colony Optimization, PSO- Particle Swarm Optimization, SS- Scatter Search , EDA- Estimated Distribution Algorithms

III. CONCLUSION

This paper discuss that how metaheuristics methods are work individually or combined together to produce good results when applied to genomic data applications. Metaheuristics methods are generally classified into two classes, single solution based and population based methods. The population based methods initialize the searching with a group of solution called population whereas single solution based methods or the trajectory methods initialize the searching with a single solution. Seven different methods are presented in this work, three of them are single solution based methods such as, simulated annealing (SA), variable neighborhood search (VNS) and tabu search (TS) the other four methods are population based methods such as, Artificial Bee Colony (ABC),genetic algorithm (GA), particle swarm optimization (PSO), and group search optimizer (GSO). These methods are applied to various different applications of genomic data. The first application is to select a gene from gene expression data for tumor classification by applying a hybrid particle swarm optimization method with tabu search methods. The second application is the MSA using genetic algorithm (GA) and Tabu search is used for DNA sequencing. The other applications such as RNA secondary structure prediction and FAP are also computed by algorithms like ACO, GA and SS. As these examples describe the advantage of using the metaheuristic method for different genomic data applications.

REFERENCES

[1] Michael K. K. Leung, Andrew Delong, Babak Alipanahi, and Brendan J. Frey,"*Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets*", Proceedings of the IEEE, Vol.104, No.1, January 2016.

[2] E. de Klerk and P. A. C.'t Hoen, ''*Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing*,'' Trends Gen., vol. 31, no. 3, pp. 128–139, 2015.

[3] J. Harrow et al., ''*GENCODE: The reference human genome annotation for the ENCODE project*,'' Genome Research., vol. 22, no. 9, pp. 1760–1774, 2012.

[4] V. Marx, ''*Biology: The big challenges of big data,'' Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

[5] K. Y. Yip, C. Cheng, and M. Gerstein, ''*Machine learning and genome annotation: A match meant to be?*'', Genome Biology, vol. 14, no. 5, pp. 205, 2013.

[6] Fred W. Glover and Gary A. Kochenberger, "*Handbook of Metaheuristics (International Series in Operations Research & Management Science)*", Springer, January 2003.

[7] Blum C, Roli A, "*Metaheuristics in combinatorial optimization: Overview and conceptual comparison*", ACM Computing Survey, vol. 35, no.3, pp.268-308, September 2003.

[8] Glover, F "*Future paths for integer programming and links to artificial intelligence*", Computer Operation Research, Vol. 13, pp.533–549, 1986.

[9] Kirkpatrick, S., Gelatt, C., Vecchi, M., "*Optimization by simulated annealing",* Science, New Series, vol. 220, No. 4598, pp.671–680, May 1983.

[10] Mladenovic.M, Hansen.P, "*Variable neighborhood search*", Computer Operation Research,. Vol.24, pp.1097–1100, 1997.

[11] James Kennedy, Russell Eberhart, "*Particle Swarm Optimization*", IEEE International Conference on Neural Networks, Vol. 4, pp. 1942-1948, December 1995.

[12] He, S., Wu, Q.H., Saunders, J.R, "*Group search optimizer–an optimization algorithm inspired by animal searching behaviour*", IEEE Transactions on Evolutionary Computer, vol. 13, no.5, pp.973–990, October 2009.

[13] Ajit Kumar, Dharmender Kumar and S.K. Jarial, "*A novel hybrid K-means and artificial bee colony algorithm approach for data clustering*", Decision Science Letters, vol. 7, pp. 65-76, April 2017.

[14] Ajit Kumar, Dharmender Kumar and S.K. Jarial, "*A Comparative Analysis of Selection Schemes in the Artificial Bee Colony Algorithm"*, Computación y Sistemas, vol.20, No.1, pp. 55-66, 2016.

[15] Holland.J.H, "*Adaptation in Natural and Artificial Systems*", University of Michigan Press, Ann Arbor, 1975.

[16] Shen, Q.,Wei-Min, S.,Wei, K, "*Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data*", Computational Biology and Chemistry, vol. 32, pp. 53–60, 2008.

[17] Neelam Goel, Shailendra Singh and Trilok Chand Aseri, "*A comparative analysis of soft computing techniques for gene prediction*", Analytical Biochemistry, vol.438, pp.14-21, March 2013.

[18] Sina Tabakhi, Ali Najafi, Reza Ranjbar and Parham Moradi," *Gene selection for microarray data classification using a novel ant colony optimization*", Neurocomputing, Vol. 168, Issue.C, pp. 1024-1036, May 2015.

[19] C. Gondro, B.P. Kinghorn," *A simple genetic algorithm for multiple sequence alignment*", Genetic and Molecular research. Vol.6, no.5, pp. 964-982, October 2007.
[20] Rasmussen TK and Krink T, "*Improved hidden Markov model training for multiple sequence alignment by a particle swarm optimization- evolutionary algorithm hybrid*", Bio systems, vol. 72, pp. 5-17, 2003.
[21] Kamil Kwarciak, Piotr Formanowicz, "*Tabu search algorithm for DNA sequencing by hybridization with multiplicity information available*", Elsevier, Computers & Operation Research, Vol.47, pp. 1-10, January2014.
[22] Neethling M and Engelbrecht AP, "*Determining RNA Secondary Structure using Set-based Particle Swarm Optimization*", Proc. Of congress on Evolutionary Computation (CEC), IEEE press, USA, 2006.
[23] J.H. Holland," *Adaptation in Natural and Artificial Systems*", the University of Michigan Press, Ann Arbor, Michigan, 1975.
[24] F. Glover, M. Laguna, and R. Marti," *Fundamentals of scatter search and path relinking*", Control and Cybernetics, Vol.39, no.5, pp. 653–684, 2000.
[25] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "*Optimization by simulated annealing. Science*", Science, Vol. 220, no.4598, pp.671–680, 1983.

Authors Profile

Mr. Manu Phogat pursed Bachelor of Engineering from MDU Rohtak, India in year 2007 and Master of Technology from DCRUST Murthal, India in year 2011. He is currently pursuing Ph.D. in Department of CSE, GJUS&T, Hisar, India since 2015. His main research work focuses on Data Mining, Machine Learning and Computational Bioinformatics.

Dr Dharmender Kumar pursed Bachelor of Technology from GJUS&T, Hisar, India in 1996 and Master of Technolgy from Kurukshetra University in year 2001. He is pursed Ph.D. in Computer Science from GJUS&T in year 2009 and currently working as Associate Professor in Department of Computer Science, GJUS&T, Hisar, India. He is a life member of the compuer Society of india, International Association of Engineers and Computer Science Teachers Association. He has published more than 40 research papers in reputed international journals including Scopus, Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Data Mining, Machine Learning, Big Data Analytics,. He has 15 years of teaching experience .

