NATURE INSPIRED ALGORITHMS TO SOLVE DNA FRAGMENT ASSEMBLY PROBLEM: A SURVEY

Indumathy R [1] and Uma Maheswari S [2]
[1]Research Scholar, [2]Associate Professor
Department of Electronics and Communication Engineering, Coimbatore Institute of Technology, Tamilnadu, India
[1]indumathy.rajagopal@gmail.com [2]umamaheswari.cit@gmail.com

ABSTRACT

Evolutionary search algorithms are becoming an essential advantage in the algorithmic toolbox for solving multi-dimensional optimization problems in a wide range of bioinformatics problems such as Genome fragment assembly which is a NP-hard problem. In computational biology, one of the most challenging problems is to reconstruct the original DNA sequence from a huge number of fragments, each one being numerous hundred base-pairs (bps) long. Recently it has become common for the researchers to apply the heuristic search algorithms to solve these kinds of complex problems with the help of combinatorial optimization. A genetic algorithm, the most well-known and representative evolutionary search technique has been the subject of the major part of such applications. Despite the fact that there have been different methods available in the literature, researchers are still facing difficulties in choosing the best method that could solve these problems efficiently and effectively. This paper aims to analyses the existing algorithms, evaluate them, point out the faults of these works and specify the emerging trend.

KEYWORDS

Computational Biology, Human Genome Project, Optimization Problem, DNA Fragment Assembly, Genetic Algorithms

1. INTRODUCTION

The Human Genome Project is developing a complete representation of the information underlying the genetic make-up of human as well as the genetic make-up of several other species that are either important for comparative scientific understanding, relied on, for progress in human medicine, or of direct industrial or agricultural applications [1,2]. This information, encoded linearly along the genomic DNA polymer, will be used as a platform to rapidly expand the ability to characterize and understand human disease and infection, design and develop preventative and therapeutic medicines, develop improved nutritional sources, and employ microbial tools in environmental and other applications. The Human Genome Fragment Assembly Project aims to identify the exact sequence of nucleotide base pairs for the entire human genome. The Human genome contains about three billion nucleotide base pairs; however, current technologies usually sequence DNA fragments shorter than 1000 bases [3]. Thus, the bioinformatics industry needs efficient algorithms for the precise assembly of long DNA sequences.

2. DNA OVERVIEW

Deoxyribonucleic acid (DNA) is a house hold word today. The chemical structure of DNA was discovered 59 years ago in 1953 by James Watson and Francis Crick. They were awarded the Nobel Prize in 1962. DNA is an organic polymer that is found within every cell of an organism. The polymer is composed of three specific parts: i) the phosphate backbone ii) the deoxyribose

sugar and iii) the nitrogenous base. The first two components remain constant across all individuals, while the third is what distinguishes each constitute of the polymer, and thus aids in distinguishing between individuals. There are four different bases comprising two purines [Adenine (A) and Guanine (G)] and two pyrimidines [Cytosine (C) and Thymine (T)]. It is the combination of these four bases that determines the precise function and coding capacity of the DNA. It is important to note that DNA is double stranded, meaning that two complementary strands of DNA are present in every cell. The double strand nature is based on complimentary of the bases, where G pairs with C and A pairs with T. This complementary base is critical for various DNA testing techniques and the basic principles of DNA chemistry. When put together in a unique way, the string of A, C, T or G can serve as a template for messenger Ribonucleic Acid (mRNA) which in turn codes for proteins. These proteins finally form the structure and function for each and every process inside the cell and inside an organism. As a result, sequences of DNA coding for proteins enable the process of constructing and maintaining the cell, which is finally responsible for all genetic processes.

3. THE BASICS OF DNA FRAGMENT ASSEMBLY

DNA sequence assembly is a method that aims to reconstruct the original DNA sequence from a huge number of fragments, each several hundred base-pairs long. DNA sequence assembly is a very complex problem in computational biology. The heuristic techniques are needed to solve the DNA fragment assembly problem because current technology, such as gel electrophoresis, cannot directly and precisely sequence DNA molecules longer than a few thousands bases. However, most of the eukaryotic genomes are much longer. For example, a human DNA is about 3.2 billion bases in length and cannot be read at once. The following technique has been developed to deal with this limitation. First the DNA molecule is cut at random locations to obtain fragments that can be sequenced directly. To get the original DNA molecule the overlapping fragments are assembled. This strategy is called shotgun sequencing. Many literatures provide solutions for DNA sequence assembly problem. Christian Burks [4] suggested that DNA sequencing throughput has to be increased by orders of magnitude to complete the task in the time frame of 15 years that was laid out for the Human Genome Project, and that such dramatic increases will rely in large part on automating the several experimental and interpretive steps involved in DNA sequencing.

3.1. Genetic Algorithms Applied to the DNA Fragment Assembly Problem

Many heuristic approaches are applied in DNA Sequence Assembly which can improve the process of DNA Sequence Assembly one of them is Genetic Algorithm. Parsons, R.J. and Forrest, S. and Burks, C. [5] analyzed various genetic algorithm operators for one permutation problem related with the Human Genome Project - the DNA sequence fragments are assembled from a parent clone whose sequence is not known into a consensus sequence equivalent to the parent sequence. The sorted order representation and permutation representation can be used and these representations do not require specialized operators. Parsons, R.J. and Johnson, M.E. [6] discussed the modifications to the previous genetic algorithm used, the experimental design process by which new results were obtained and have also made preliminary attempts to explain the results and answer the questionnaire. Kim, K. and Mohan, CK [7] presents a fragment assembler using a new parallel hierarchical adaptive variation of evolutionary algorithms. The innovative features include a new measure for evaluating sequence assembly quality and the development of a hybrid algorithm. Reinitialization of the population of individuals helps prevent premature convergence, and control the explosion in computational resources and in solving most bioinformatics problems. Fang, S.C. and Wang, Y. and Zhong, J. [8] approach maximizes the similarity (overlaps) between given fragments and a candidate sequence. It considers both whole fragments and the single base pair similarities in the sequence. Special genetic operators are designed to speed up the searching process.

Luque, G. and Alba, E. [9] present several methods, a canonical genetic algorithm, a Cross generational elitist selection method, a scatter search algorithm, and a simulated annealing, to solve the problem accurately instances that are 77K base pairs long. Kikuchi, S. and Chakraborty, G. [10] added two heuristic ideas with GA to make it more efficient. The first step is chromosome reduction (CRed) step which condense the length of the chromosomes, participating in genetic search, to improve the efficiency. The second step is chromosome refinement (CRef) step which is a greedy heuristics, rearranging the bits using domain knowledge, to locally improve the fitness of chromosomes. A.J. Nebro [11] proposed a grid based genetic algorithm in solving the DNA fragment assembly problem. It is a steady- state GA which uses a panmitic population, and it is based on computing parallel function evaluations in an asynchronous way. The advantage of a grid system is to enhance the scalability. Enrique Alba and Gabriel Luque [12] solve the DNA fragment assembly problem by using a new hybrid genetic algorithm. This hybrid method combines a promising heuristic, PALS, with a well-know metaheuristic, a genetic algorithm, to obtain a result that is a very efficient assembler that allows to find optimal solutions for large instances of this problem.

3.2. DNA Fragment Assembly Using Fragment Assembly Packages

A number of fragment assembly packages have been developed which are used to sequence different organisms. The widely used popular package PHRAP [13] is a program for assembling shotgun DNA sequence data. PHERD quality scores are initially developed by the program PHERD to help in the computerization of DNA sequencing in the Human Genome Project. PHERD quality scores are assigned to each base call in automated sequencer traces. PHERD quality scores have become broadly accepted to describe the quality of DNA sequences and can be used to evaluate the efficacy of various sequencing methods. Perhaps the most important use of PHERD quality scores is the automatic determination of accurate, quality-based consensus sequences. The TIGR assembler [14] overcomes several major obstacles to assembling projects such as: the large number of pair wise comparisons required for the presence of repeat regions, chimeras introduced in the cloning process and sequencing errors. STROLL [15] implemented a reliable technique to sequence DNA using primer walking approach and a fragment assembly program is designed for large-scale (mega base level) genome sequencing. In an attempt to sequence the one-megabase genome of Borrelia burgdorferi, the bacterium which causes Lyme disease, the strategy proposed: after a thin coverage shotgun sequencing phase, the gaps are closed in the primer walking phase. CAP3 [16] program includes a number of improvements and new features to improve DNA sequence assembly. The program has a capability to clip 5' to 3' low-quality regions of reads. It uses base quality values in computation of overlaps between reads, construction of multiple sequence alignments of reads, and generation of consensus sequences. The program also uses forward–reverse constraints to correct assembly errors and link contigs. Celera assembler [17] developed at Celera for publication of the first draft human genome sequence in 2001, Attacks repeats by screening high copy repeats, finding repeat boundaries and utilizing mate pair information. EULER [18] is a novel approach to fragment assembly that abandons the classical "overlap - layout - consensus" pattern is used in all currently available assembly tools. This pattern is useful in assembling clones; it faces difficulties in genomic shotgun assembly: the existing algorithms make assembly shortcomings and are often unable to resolve repeats even in prokaryotic genomes. Biologists are well-conscious of these errors and are forced to carry out additional experiments to verify the assembled contigs.

3.3. The Problem of DNA Fragment Assembly Using Ant Colony Optimization

Various kinds of methods and strategies for DNA fragment assembly have been proposed. One such approach is an ant colony optimization (ACO) Meksangsouy, P. and Chaiyaratana; N. [19] proposed an asymmetric ordering representation where a path, co-operatively generated by all ants in the colony represents the search solution. The optimality of the fragment layout obtained

is then determined from the sum of overlap scores calculated for each pair of consecutive fragments in the layout. Two types of assembly problem are investigated: single-contig and multiple-contig problems. The simulation results indicate that in single-contig problems, the performance of the ant colony system algorithm is approximately the same as that of a nearest neighbor heuristic algorithm. On the other hand, the ant colony system algorithm outperforms the nearest neighbor heuristic algorithm when multiple-contig problems are considered. Zhao Y. and et al. [20] improved sequence alignment method based on the ant colony algorithm. This new method could avoid a local optimum and remove especially the paths scores of great difference by regulating the initial and final positions of ants and by modifying pheromones in different times. Zuwairie Ibrahim and Tri Basuki Kurniawan, [21] in their approach model the DNA sequence design as a path-finding problem, which consists of four nodes, to enable the implementation of the ACS and compared their results with other methods such as the genetic algorithm.

3.4. DNA Fragment Assembly Based on Particle Swarm Optimization

There are few literatures available which represent solution for DNA Sequence Assembly problem using metaheuristic and nature inspired algorithms. PSO algorithm comes under nature inspired algorithm and it has been proven as an effective optimization technique to solve any optimization problems for optimum result. PSO algorithm can also be used to solve computational biology problem to give better result than the conventional methods. Ravi Vikas and Sanjay [22] proposed a solution for DNA sequence assembly problem using Particle Swarm Optimization (PSO) with Shortest Position Value (SPV) rule. DNA sequence assembly problem is a discrete optimization problem, so there is need of discrete optimization algorithm to solve it. It is a continuous version of PSO with SPV rule to solve the DNA sequence assembly problem. SPV rule transforms continuous version of PSO to discrete version.

3.5. DNA Fragment Assembly Using the Greedy Algorithm

One of the optimization algorithms for the fragment assembly problem is the Greedy Algorithm which is based on the Best Set of Maximum Weight Contigs Approach [23]. The algorithm considers of unknown orientation and missing fragments. The initial step of the algorithm is to build the Best Set of Maximum Weight Contigs (BSC). The complexity of this step is $O(n^2 l^2)$, where n is the number of fragments and l is the average length of fragments. The next step of the algorithm is to order the Maximum Weight Contigs (MWC) of BSC based on contig overlaps order. The complexity of this step is $O(m^2 l^2)$, where m is the number of MWCs. The Greedy Algorithm calculates the contig overlaps rather than the fragment overlaps. The advantages are twofold is it enables us to take only the true overlaps into account and it gives a better guarantee for finding the orientation of the fragments.

3.6. DNA Fragment Assembly Using Structured Pattern Matching

The Structured Pattern Matching Algorithm is based on a method called hybridization fingerprinting that is usually used by biologists to figure out the overlap information among DNA clones from biological probes. DNA clones are exact copies of a particular part of a genome and are much longer than fragments [24]. To tackle the DNA fragment assembly problem, the algorithm divides the task into three phases. The first phase is called probe matching. Instead of using biological probes, short probes (e.g. 12 bps) from each fragment are randomly selected. The second phase is called overlap map construction. This phase constructs a detailed map to show how fragments are ordered and how they align. The third phase is called sequence determination. It is relatively straightforward since all the information we need is available from the second phase. The time complexity of the Structured Pattern Matching

algorithm is approximately linear in the length of the target sequence. The efficiency is due to the compact encoding representation of fragments.

4. CONCLUSION & FUTURE WORK

This survey on different nature inspired algorithms has given a wide range of complexity involved in DNA fragment assembly and the efficient outcomes from each of the strategies. A group of interesting papers demonstrates the efficiency and the competitive accuracy of this NP - hard problem. The Genetic Algorithms use different genetic operators and give sorted order representation and permutation order representation. These GA's are improved with the chromosome reduction and chromosome refinement methods for an accurate read of DNA fragments. Parallel hierarchical adaptive variation of evolutionary algorithm is included to maximize the similarity score. The two types of assembly problem single contig and multi contig are well handled using Ant Colony Optimization. Particle Swarm Optimization is used with SPV rule is an effective technique to solve this optimization problem. An interesting opportunity for future research is to hybridize the algorithms and use different variants of those algorithms to obtain a most optimal long reads of DNA sequence. Parallel version of these algorithms can be used. Special techniques to handle repeats can be implemented for a stringent assembly. Multi objectives can be defined to optimize algorithms with high run time efficiency and to get a more reliable long DNA consensus sequence.

REFERENCES

[1] Collins F, and Galas D (1993), **"**A new 5-year plan for the United-States Human Genome Project", Science Direct, No. 262, pp.43-46.

[2] Cooper NG (1994), "The Human Genome Project: Deciphering the Blueprint of Herediry", University Science Books, Mill Valley, CA.

[3] Tammi, M. T. (2003), "The Principles of Shotgun Sequencing and Automated Fragment Assembly", Center for Genomics and Bioinformatics, Karolinska Institute, Stockholm, Sweden.

[4] Burks C (1994), "DNA sequence assembly, Engineering in Medicine and Biology Magazine", IEEE, Vol. 13, pp. 771- 773.

[5] Parsons, R.J., Forrest, S. and Burks C (1995), "Genetic algorithms, operators, and DNA fragment assembly", Machine Learning, Vol. 21, pp. 11- 33.

[6] Parsons R.J. and Johnson M.E (1995), "DNA sequence assembly and genetic algorithms- new results and puzzling insights", Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology (ISMB-95), pp. 277- 284.

[7] Kim, K. and Mohan, CK (2003), "Parallel hierarchical adaptive genetic algorithm for fragment assembly", Evolutionary Computation, CEC'03, Vol. 1, pp. 600-607.

[8] Fang, S.C., Wang, Y. and Zhong J (2005), "A Genetic Algorithm Approach to Solving DNA Fragment Assembly Problem", Journal of Computational and Theoretical Nanoscience, Vol. 2, pp. 499- 505.

[9] Luque, G. and Alba E (2005), "Metaheuristics for the DNA fragment assembly problem", International Journal of Computational Intelligence Research, Vol. 1, pp. 98- 108.

[10] Kikuchi, S. and Chakraborty G (2006), "Heuristically tuned GA to solve genome fragment assembly problem", Evolutionary Computation, IEEE pp. 1491-1498.

[11] Nebro A.J, Luque G, Luna F and Alba E (2008), "*DNA Fragment assembly using a grid-based algorithm*", Computers & Operations Research, Vol. 35, pp. 2776-2790.

[12] Enrique Alba and Gabriel Luque (2008) , " A Hybrid Genetic Algorithm for the DNA Fragment Assembly Problem", Recent advances in Evolutionary Computation for combinatorial optimization Studies in computational Intelligence, Vol. 153, pp. 101-112, Springer.

[13] P. Green. Phrap. http://www.phrap.org/.

[14] G.G. Sutton, O. White, M.D. Adams, and A.R. Kerlavage (1995), "TIGR Assembler: A new tool for assembling large shotgun sequencing projects", Genome Science & Tech., Vol. 1, pp. 9- 19.

[15] T. Chen and S. Skiena (1998), "Trie-based data structures for sequence assembly", Combinatorial Pattern Matching, pp. 206–223.

[16] X. Huang and A. Madan (1999), "CAP3: A DNA sequence assembly program", Genome Research, Vol. 9, pp. 868– 877.
[17] E.W. Myers (2000), "Towards simplifying and accurately formulating fragment assembly", Journal of Computational Biology, Vol. 2, pp. 275–290.
[18] P.A. Pevzner (2000) "Computational molecular biology: An algorithmic approach", The MIT Press, Vol. 1.
[19] Meksangsouy, P. and Chaiyaratana, N (2003), "DNA fragment assembly using an ant colony system algorithm", Evolutionary Computation, Vol. 3, pp. 1756- 1763.
[20] Zhao, Y and et al (2008), "An Improved Ant Colony Algorithm for DNA Sequence Alignment", International Symposium on Information Science and Engineering, pp. 683—688.
[21] Zuwairie Ibrahim and Tri Basuki Kurniawan (2009), "Implementation of an ant colony system for DNA sequence optimization", Journal of Artif Life Robotics, pp. 293-296.
[22] Ravi, Vikas and Sanjay (2011), "DNA Sequence Assembly using Particle Swarm Optimization", International Journal of Computer Applications Vol.28- No.10, pp. 33-38.
[23] Elloumi M. and Kaabi S (1999), "Exact and approximation algorithms for the DNA sequence assembly problem", SCI in Biology and Medicine, Vol. 8.
[24] Kim, S. and Segre, A. M (1999), "AMASS: A structured pattern matching approach to shotgun sequence assembly", Journal of Computational Biology, 6(2), pp. 163-186.

Authors

R. Indumathy holds a Bachelors degree in Computer Science in the year 2003 from Bharathiar University, Coimbatore, Masters Degree in Computer Applications in the year 2006 from Anna University, Chennai and Master of Philosophy Degree in Computer Science in the year 2008 from Bharathiar University, Coimbatore. She is currently pursuing her Doctoral degree in Anna University, Coimbatore and is a research scholar in Coimbatore Institute of Technology, Coimbatore affiliated to Anna University, Coimbatore. She has published five research papers in the National and International Journals / Conferences. Her research interests are Soft Computing, Bioinformatics and Genetic Algorithms

S. Uma Maheswari, BE (ECE), ME (Applied Electronics), PhD, is presently working as an Associate Professor in the Department of Electronics and Communication Engineering in Coimbatore Institute of Technology, Coimbatore. She has 25 years of teaching experience. She is a life member in Indian Society for Technical Education and Institution of Engineers (India). Digital Image Processing, Digital Signal Processing and VLSI are her fields of specialization.