Cibtech Journal of Zoology ISSN: 2319–3883 (Online)
An Open Access, Online International Journal Available at http://www.cibtech.org/cjz.htm
2015 Vol. 4 (2) May-August, pp.5-9/Kumar and Banka
Research Article

A NOVEL AND EFFICIENT APPROACH FOR ALIGNMENT OF PROTEIN BIOMOLECULES THROUGH RESERVE SELECTION SCHEME

[*]Manish Kumar and Haider Banka
Department of CSE, Indian School of Mines, Dhanbad-826004, India
[*]Author for Correspondence

ABSTRACT
This article provides a general solution for the problem of multiple sequence alignment (MSA) of biological sequences by using Genetic Algorithms (GA) with Reserve Selection (GARS). As in genetic algorithm, when the number and length of the sequence increases, the problem of early convergence is being encountered. To eradicate this problem, the concept of reserve selection combined with genetic algorithms is proposed. The main objective of this research work is to maximize the similarity between the sequences by adding and shuffling gaps and then calculating the score of each column of the alignment so as to avoid early convergence in GA and get reliable Multiple Sequence Alignment. Empirical studies has been performed on the results of GA and GARS by using benchmark datasets (BALiBASE). To test the performance of the proposed algorithm, it has been compared with existing well-known methods, such as CLUSTALX, ML-PIMA, SB-PIMA, HMMT and PILEUP8 by solving a number of benchmark datasets from BAliBase 2.0. Experiments on a wide range of data`s have shown the effectiveness of the proposed approach and its ability to achieve good quality solutions.

Keywords: *Alignment, Genetic Algorithms, Early Convergence, Reserve Selection*

INTRODUCTION
Multiple sequence alignment (MSA) (Hamidi *et al.*, 2013) has special significance in identify biologic sequence model of correlated sequences. MSA refers to the problem of optimally aligning three or more sequences of symbols with or without inserting gaps between the symbols. The objective is to maximize the number of matching symbols between the sequences and also use only minimum gap insertion, if gaps are permitted. Multiple comparison or alignment of protein sequences has become a fundamental tool in many different domains in modern molecular biology, from evolutionary studies to prediction of 2D/3D structure, molecular function and intermolecular interactions etc. By placing the sequence in the framework of the overall family, multiple alignments can be used to identify conserved features and to highlight differences or specificities.

MSA problems are solved using several different methods, such as classical, progressive (Kupis and Mandziuk, 2007) and iterative algorithms (Bayati *et al.*, 2005). These algorithms follow either global or local alignment (Chun *et al.*, 2013) strategies. In global alignments, sequences are aligned over their whole length. By contrast, local alignments identify regions of similarity within a sub sequence. Local alignments are often preferable, but can be more difficult because of the additional challenge of identifying the regions of similarity. A general global alignment technique is the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970), which is based on dynamic programming. The Smith–Waterman algorithm is a general local alignment method which is also based on dynamic programming. The dynamic programming (DP) approach (Zhimin and Zhong, 2013) is good at finding the optimal alignment for two sequences. However, the complexity of this method grows significantly for three or more sequences. Note that MSA is a combinatorial problem (NP-hard) where the computational effort becomes prohibitive with a large number of sequences. The progressive alignment algorithm (tree-base algorithm), proposed by Feng and Doolittle (Feng and Dolittle, 1987), iteratively utilizes the method of Needleman and Wunsch (Thompson *et al.*, 1994) in order to obtain an MSA and to construct an evolutionary tree to depict the relationship between sequences. The progressive alignment algorithms align sequences according to the branching order of a guide tree. The difficulty with these methods is that

they usually converge to local optima. To overcome such a limitation, it is recommended to use an iterative or stochastic procedure.

It has been stated in the literature (Devereux *et al.*, 1984) (Thompson *et al.*, 1997) that none of the existing algorithms were capable of providing accurate alignments for all the test cases .As a consequence, iterative algorithms were developed to construct more reliable multiple alignments, using for example iterative refinement strategies, Hidden Markov Models (Eddy, 1998) or Genetic Algorithms (Yong *et al.*, 2011). These methods were shown to be more successful at aligning the most conserved regions for a wide variety of test cases, although some accuracy was lost for distantly related sequences, in the 'twilight zone' of evolutionary relatedness.

In this study, genetic algorithm (GA) has been considered for experimental analysis .The main advantage of using genetic algorithms for MSA problem is that there is no need to provide a particular algorithm to solve a given problem. It only needs a fitness function to evaluate the quality of different solutions. Also since it is an implicitly parallel technique, it can be implemented very effectively on powerful parallel computers to solve exceptionally demanding large-scale problems.

In order to solve the MSA problem efficiently, genetic algorithms (GAs) were applied. But, GA also suffers from a problem know as early convergence. As the number and length of sequences increase, the GA usually suffered from a problem of early convergence in which they are easily trapped into local optima. Early convergence occurs when the population in a genetic algorithm is trapped in such a suboptimal state that most of the genetic operators can no longer produce offspring that outperforms their parents. To avoid early convergence problem the concept of reserve selection has been introduced in this paper, which advocate the principle of "survival of the fittest". The reserve selection is a modified selection scheme that derives from the idea that chromosomes of poor solution quality should also be inherited since they may contain potential building blocks contributing to future evolution.

Many studies and research has been performed around the world to solve the problem of early convergence in GA. The Social Disasters Technique (SDT) was introduced by Kureichick and colleagues in order to avoid the premature convergence to local optima. The general idea is to diagnose the situations of loss of genetic diversity of the population and in such a case to apply a catastrophic operator to it. These operators were defined with the purpose to return the population to an acceptable degree of genetic diversity, by replacing a number of selected individuals, by others, generated at random.

Some authors have also suggested the idea of Random Off spring Generation (ROG) . The idea behind the Random Off spring Generation (ROG) is to test the individual's genetic material, before the crossover operation and if a situation as the one just referred is detected, the operation is not performed. Instead, one offspring, or even two, are randomly generated; i.e., their genotype will code a random solution on the problem's domain.

For preventing the premature convergence to local optima a concept of adaptive mutation rate was introduced, in which mutation operator aims to introduce a random component into the search process, with the exploitation of new chunks of the solution space, thus promoting the increase of the genetic diversity of the population; i.e., it is not surprising to find out that one of the first steps to take in order to maintain the genetic diversity in a population is the MR's increase. However, a high value to this parameter introduces a certain degree of noise into the system, thus creating serious obstacles to the convergence process.

As the multiple alignments of protein sequences is an important application for the foreseeable future, therefore this research works focuses on alignment of protein sequences using genetic algorithm, after removing the early convergence problem of genetic algorithm. The number of newly available protein sequences still far outpaces the number of determined protein three-dimensional structures and therefore sequence homology remains the main method to infer protein structure, function, active sites and evolutionary history. In recent years, protein MSA tools (Gelly *et al.*, 2011) have improved rapidly in both scalability and accuracy. The future improvements in protein sequences are likely to come by combining sequence alignment with other information, such as known structures of some of the proteins being aligned or homology to a larger pool of proteins. Finally, better utilization of phylogenetic relationships

and incorporation of models of protein sequence evolution (Cai *et al*., 2000) also hold promise for improved alignment performance.

Literature studies say that there are still a number of challenges in aligning protein sequences. First, the locally conserved regions, that reflect functional specificities or that modulates a protein's function in a given cellular context, are less well aligned. Second, motifs in natively disordered regions are often misaligned. Third, the badly predicted or fragmentary protein sequences, which make up a large proportion of today's databases, lead to a significant number of alignment errors.

Based on the literature studies and in order to test the feasibility of the proposed approach a comparison study were made between the proposed method and some of the existing methods such as the CLUSTALX, ML-PIMA ,SB-PIMA, HMMT and PILEUP8 by calculating the corresponding Baliscore with the help of proposed fitness function. By the experimental analysis, it can be concluded that the proposed method outperforms all above mention methods for most of the test cases (datasets).

The rest of the paper is organised as follow. The next section introduces the concepts underlying the research work with detailed discussion on the proposed approach. Followed by section which explains about the detailed results over different datasets, along with the experiments setup required in order to validate and observe the results.

MATERIALS AND METHODS

Here, we assume that the readers are familiar with the basics of GA, and hence, we will not go deeper inside it. We will only discuss the changes made in producing the new generation for genetic algorithm.

GA is a highly parallel, random and self-adaptive algorithm which has many merits over traditional methods such as global optimization .However in practice, GA is often criticized for the lack of a solid theoretical foundation. Actually, a theoretical foundation is desired in order to gain deep understanding of the strength and weakness of GA. GA usually has the drawbacks such as early convergence and slow convergent speed.

When early convergence occurs, it is difficult for GA to get rid of a local optimum and reach a global optimum. Early convergence is the main obstacle to a genetic algorithms practical application. To avoid this, a reserve selection scheme has been introduced which will jointly work with genetic algorithms.

In this scheme, some less fit individuals has been selected with regard to their uniqueness for the next generation. These individuals will surely be considered for the coming generation ensuring some better solution. The less fit individuals buried inside the reserve area will help to maintain the diversity in population with diversified search which will help to overcome the problem of early convergence in GA.

In the experiment, the reserve area are taken as 20%, 30% and 40% of the total population size. It means that for the first experiment, the population is sorted in ascending order according to the fitness value of the individuals. Now, if the reserve area is 20% it means 20% area of current population will be filled up by some less fit individuals from previous population. It can also be understood in a way, that every new generation will inherits some less fit individuals from its previous generation which will fit to its 20% of reverse area. This process will be repeated in all coming generations and stopped when desired fitness values are found or maximum number of generation reached. This same process will be repeated for reserve area of 30% and 40% respectively.

Description of Algorithm

Step 1: Generate random population of *n* chromosomes (suitable solutions for the problem).

Step 2: Evaluate the fitness *f(x)* of each chromosome *x* in the population.

Step 3: With a crossover probability (0.6%) crossover the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.

Step 4: With a mutation probability (0.1%) mutate new offspring at each locus (position in chromosome).

Step 5: Sorting is being performed with the population, with reserve area of 20%, 30% and 40% respectively.

Step 6: Go for desired number of generation.

Step 7: Stop when desired number of fitness value found or maximum number of generation reached.

RESULTS AND DISCUSSION

In this section, the experimental methodology and experiments setup required in this work is detailed. Moreover, results obtained with the proposed method are presented and discussed.

The computational time required for finding good multiple sequence alignments is dependent on the number of sequences, sequence length and the similarities of the sequences. In addition, the selection of algorithmic parameters also plays an important role.

The main objective of the research work is to use the concept of reserve selection to avoid early convergence in GA for solving MSA problem of protein sequences. In this study, multiple alignment of protein sequences has been considered, as multiple alignment of protein sequences is an important application for the foreseeable future.

In order to evaluate the proposed approach, the experiment is carried out with different datasets (ref. 3) of different lengths from the BAliBASE database. Availability of literature about performance of other related algorithms on these data sets prompted the authors to select them for the study. For each of the experiment, alignments were performed both with the proposed method as well as with the other methods described in the literature. Performance, in terms of both efficiency and apparent alignment quality, are summarized for several of the experimental runs. Table 1shows the comparison results, by which one can conclude that the proposed algorithm has better results among these approaches. The bold faced data represents the best scores among the methods.

Table 1: Experimental results with Reference 3 Datasets of BAliBase 2.0

Name of Datasets		CLUSTAL X	SB−PIMA	HMMT	ML−PIMA	PILEUP8	Proposed Method 20% Reserve Area	Proposed Method 30% Reserve Area	Proposed Method 40% Reserve Area
Ref. 3	1idy	0.273	0.000	0.227	0.000	0.000	0.281	0.297	0.325
	1wit	0.565	0.645	0.323	0.323	0.210	0.425	0.568	0.678
	1pamA	0.678	0.546	0.169	0.590	0.754	0.571	0.742	0.879
	2myr	0.538	0.278	0.101	0.494	0.310	0.645	0.452	0.612
	4enl	0.547	0.393	0.050	0.438	0.498	0.325	0.483	0.602
	1ped	0.627	0.450	0.172	0.507	0.722	0.547	0.954	0.857
	1r69	0.524	0.000	0.000	0.905	0.000	0.542	0.854	0.697
	kinase	0.720	0.541	0.478	0.682	0.599	0.845	0.864	0.961
	1ajsA	0.163	0.000	0.006	0.000	0.110	0.352	0.214	0.012
	1uky	0.130	0.083	0.037	0.148	0.083	0.213	0.012	0.233
	1ubi	0.146	0.000	0.366	0.000	0.268	0.120	0.389	0.315
Average Score		0.446	0.267	0.175	0.372	0.323	0.442	0.529	0.561

The experiments were performed using C programming on an Intel Core 2 Duo processor with T9400 chipset, 2.53 GHz CPU and 2 GB RAM running on the Linux platform.

Furthermore, the population size was established to 1000 individuals and the maximum number of generations was 10 with a crossover probability of 0.6, mutation rate of 0.1and the tournament size is 2 for the experiment. The scoring matrix and the space score used for the experiment are PAM 250 and -10, respectively, for each Protein sequences.

For evolution of the proposed approach, the algorithm were executed for ten independent run for 11 datasets and as the fitness score depends upon the level of similarity among the residue in the sequences so that the score can be either positive or negative. Here, on point is to be noted that if the residues among the comparable sequences are similar, then small numbers of gaps (“-”) are needed to make the sequences aligned properly. On the other hand, if the majority of the residues are dissimilar then a large number of gaps are need for necessary sequence alignment.

Performance of the Proposed Method with ref. 3

In this experimental study, eleven test cases were considered from references 3 and out of 11 test cases the proposed method shows better solution for 10 test cases. Only, ML−PIMA for 1r69 dataset have shows better performance than the proposed method (Table 1). The average score comparison between the proposed and other method gives an idea that in overall average comparison, our proposed method is better. As we can see from table 1, that the increase in reserve area has yield a better alignment score for most of the datasets. We can also conclude that, if the reserve area is more or if we increase the reserve area in the experimental study than the outcome will be the optimal one in terms of score.

REFERENCES

Cai L, Juedes D and Liakhovitch E (2000). Evolutionary computation techniques for multiple sequence alignment, in Process 829–835.

Devereux J, Haeberli P and Smithies O (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12** 387–395.

Eddy S (1998). Profile hidden Markov models. *Bioinformatics* **14** 755–763.

Feng DF and Dolittle RF (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* **25**(4) 351–360.

Gelly JC, Joseph AP, Srinivasan N and de Brever AG (2011). iPBA: a tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Research* **39**(2) 18–23.

Hamidi S, Naghibzadeh M and Sadri J (2013). Protein multiple sequence alignment based on secondary structure similarity. *International Conference on Advances in Computing, Communications and Informatics* 1224-1229.

Kupis P and Mandziuk J (2007). Evolutionary-Progressive Method for Multiple Sequence Alignment. *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology* 291-297.

Mohsen B, Balaji P, Devavrat S and Mayank S (2007). Iterative Scheduling Algorithms. *IEEE INFOCOM Proceedings*.

Needleman SB and Wunsch CD (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**(3) 443–453.

Nguyen K and Pan Y (2011). An improved scoring method for protein residue conservation and multiple sequence alignment. *IEEE Transactions on NanoBioscience* **10**(4) 275-285.

Peng Y, Dong C and Zheng H (2011). Research on Genetic Algorithm Based on Pyramid Model. *2nd International Symposium on Intelligence Information Processing and Trusted Computing* 83-86.

Thompson JD, Higgins DG and Gibson TJ (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **224** 673–4680.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG (1997). The CLUSTAL−X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**(24) 4876–4882.

Wei CC, Yu JC, Chien CC and Der TL Jan MH (2013) .Optimizing a Map Reduce module of pre processing high-throughput DNA sequencing data. *IEEE International Conference on Big Data* 6-9.

Yang C, Jinglu H and Songnian Y (2008). Multiple Sequence Alignment Based on Genetic Algorithms with Reserve Selection. *IEEE International Conference on Networking, Sensing and Control* 1511-1516.

Zhimin ZH and Zhong WC (2013). Dynamic Programming For Protein Sequence Alignment. *International Journal of BioScience and Bio Technology* **5**(2).