AN ENHANCED FLOWER POLLINATION ALGORITHM FOR MULTIPLE SEQUENCE ALIGNMENT

AHMAD MOH'D AZIZ ISSA HUSSEIN

UNIVERSITI SAINS MALAYSIA

2020

AN ENHANCED FLOWER POLLINATION ALGORITHM FOR MULTIPLE SEQUENCE ALIGNMENT

by

AHMAD MOH'D AZIZ ISSA HUSSEIN

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

February 2020

ACKNOWLEDGEMENT

IN THE NAME OF ALLAH THE ALL-COMPASSIONATE, ALL-MERCIFIFUL

First and foremost, I thank Allah S.W.T for giving me the strength to complete this research study. I would also like to express my deepest gratitude towards my thesis advisor, Professor Dr Rosni Abdullah, and co-supervisor, Assoc. Prof. Dr. Nur’Aini Abdul Rashid, for their mentorship, wise counselling and feedback throughout this research. Indeed, this thesis would not have seen the light of the day without their support and guidance. I would also like to extend my gratitude to the School of Computer Sciences, specially the Parallel and Distributed Computing Center and the Institute of Postgraduate Studies of Universiti Sains Malaysia.

My special thanks go to my beloved parents for their unconditional love, patience, encouragement, and continuous support. To Aisha - my muse, my better half, my beloved wife; who extended a helping hand and supported me heart and soul all through this tiring endeavor while at the same time taking care of our while at the same time taking care of our precious birdlings Omar and Reem, all I can say is I'm just lucky to have you in my life. Last but not least, I thank all those who supported me during my research for their help, patience and advice, especially my brothers, sisters, and friends as well as many others whose I may not recall.

TABLE OF CONTENTS

CHAPTER 3 - RESEARCH METHODOLOGY

CHAPTER 5 - A HYBRID MODIFIED FPAPROFILE WITH GENETIC ALGORITHM FOR MSA

CHAPTER 6 - AN IMPROVED FPAPROFILE BY MODIFYING LOCALSEARCH FOR MFPAPROFILE

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ABC	Artificial Bee Colony
ACO	Ant Colony Optimisation
BAliBASE	Benchmark Alignment DataBASE
B-HC	Beta-hill Climbing
BlockMSA	global block-based algorithm
BLOSUM	Blocks Substitution Matrix
CFPAProfile-MSA	Crossover Flower Pollination Algorithm with Profile for Multiple Sequence Alignment
CPSO	Chaotic Particle Swarm Optimisation
CS	Column Score
CSM	Class Sample Matrix
CSO	Cat Swarm Optimisation
DAG	Directed Acyclic Graph
DCA	Divide-and-Conquer Approach
DMQPSO	Diversity-Maintained Quantum-Behaved Particle Swarm Optimisation
DNA	Deoxyribonucleic acid
DP	Dynamic Programming Technique
EA	Evolutionary Algorithm
EC	Evolutionary Computation
EP	Evolutionary Programming
FFT	Fast Fourier Transform
FIC	Fractal Image Compression
FM	Memory Location
FP	Flower Pollination
FPA	Flower Pollination Algorithm
FPAProfile-MSA	Flower Pollination Algorithm with Profile for Multiple Sequence Alignment CFPAProfile-MSA
FPPSO	Hybrid Flower Pollination Algorithm with Particle Swarm Optimisation
g_extension	Gap extension value

g_open	Gap open value
GA	Genetic Algorithm
GA-ACO	Genetic Algorithm and Ant Colony Optimisation
gap_Penalty	gap_Penalty value
GAPAM	Genetic Algorithm Progressive Alignment Approach
GP	Genetic Programming
H4MSA	Metaheuristics for Multiple Sequence Alignment
HMMs	Hidden Markov Models
HMOABC	Hybrid Multiobjective Artificial Bee Colony Algorithm
HSA	Harmony Search Algorithm
IFPAProfile-MSA	Improved Flower Pollination Algorithm with Profile for Multiple Sequence Alignment CFPAProfile-MSA
LS	Local Search
MA	Memetic Algorithm
M-BPSO	Mutation-based Binary Particle Swarm Optimisation
MC	Musical composition
MFPAProfile-MSA	Smart-mutation Flower Pollination Algorithm with Profile For Multiple Sequence Alignment
MGA	Multiple Genome Aligner
MSA	Multiple Sequence Alignment
MSASA	Multiple Sequence Alignment Simulated Annealing
NP	Non Polynomial
OF	Objective Function
PAM	Point Accepted Mutation
PC	Permutation Coding
PCMA	Profile Consistency Multiple Sequence Alignment
PCR	Polymerase Chain Reaction
PID	Percentage Identity
PLS	Pareto Local Search
POA	Partial Order Alignment
PSAFPA	Pairwise Alignment Flower Pollination Algorithm
PSO	Particle Swarm Optimisation
QPSO	Quantum-Behaved Particle Swarm Optimisation
RBT-GA	Rubber Band Technique and Genetic Algorithm

RNA	RiboNucleic Acid
SA	Simulated Annealing
SC	Solution Coding
Scen	Scenario
SOGA	Self-Organizing Genetic Algorithm
SPS	Sum-of-Pairs Score
TC	Total Column
T-Coffee	Tree-based Consistency Objective Function for Alignment Evaluation
TS	Tabu Search
TS-FPA	Hybrid of Tabu Search And Flower Pollination Algorithm
WSN	Wireless Sensor Network
WSP	Weighted Sum-of-Pairs Score

ALGORITMA POLINASI BUNGA YANG TELAH DITINGKATKAN UNTUK PELARASAN PELBAGAI SUSUNAN

ABSTRAK

Penyelarasan pelbagai aturan atau (MSA) adalah satu penyelarasan tiga aturan atau lebih. Kajian menunjukkan bahawa MSA mempamerkan satu cabaran, iaitu bagaimana menemui MSA yang boleh memaksimakan skor Mutu atau Kualiti (Q) dan skor jumlah ruang (TC). Masalah ini adalah masalah *NP-lengkap*. Oleh itu, banyak kajian telah berusaha mengatasi masalah MSA, namun masih terdapat kakurangan dari segi. Banyak algoritma meta-heuristik telah disarankan untuk mengatasi masalah MSA ini. Algoritma Polinasi Bunga (FPA) adalah satu metod meta-heuristik baru berdasarkan populasi. Kelebihan utama algoritma ini adalah berdasarkan prestasinya dalam menyatukan bahagian penting metod berdasarkan populasi dan metod meta-heuristik lain seperti Algoritma Genetik (GA) dalam satu model optima yang lebih kurang sama. Oleh itu, kajian ini menggunakan FPA sebagai satu penanda-aras untuk memperbaiki ketepatan MSA. Adaptasi ini termasuklah mengubahsuai pengoperasi-pengoperasi FPA. Khususnya, tiga versi yang telah diubahsuai dicadangkan: pertama, untuk menggunakan FPA dan menghibrid teknik profil (FPAProfile) untuk mengoptima skor Q dan TC pada MSA. Kedua, untuk memperbaiki dan mengkaji keberkesanan FPA yang diubahsuai dengan Profil (FPAProfile) dengan menggabungkan pengoperasi lintasan dan mutasi pintar dalam kalangan kaedah populasi. Ketiga, untuk memperbaiki dan mengkaji keberkesanan FPA yang diubahsuai dengan cara memperbaiki polinasi tempatan (carian). Seterusnya, keputusan dibandingkan dengan 16 kaedah lain menggunakan dua set data penanda-aras berpiawai. Metod IFPAProfile mencapai keputusan yang lebih baik untuk kumpulan RV12 dari set data Balibase 3.0 dari aspek skor Q dan TC dengan 94.89 dan 87.51, masing-masing. Dalam set data OXbench, metod IFPAProfile mencapai keputusan yang terbaik dari sudut skor Q dengan 90.11.

AN ENHANCED FLOWER POLLINATION ALGORITHM FOR MULTIPLE SEQUENCE ALIGNMENT

ABSTRACT

Multiple sequence alignment (MSA) is an alignment of three or more sequences. Studies show that MSA exhibits a challenge, that is, how to find the MSA that maximises the Quality (Q) score and Total Column (TC) score. This problem is an NP-complete problem. Hence, numerous researchers have devoted their efforts to tackle the MSA problem, yet, there is still a shortcoming in the accuracy. Many meta-heuristic algorithms have been proposed to tackle the MSA problem. Flower Pollination Algorithm (FPA) is a new meta-heuristic method based on population. The main advantage of this algorithm is based on its performance in merging the important compounds of population-based methods and other known meta-heuristic methods such Genetic Algorithm (GA) within a similar optimisation model. Therefore, this research adopts FPA as a yardstick for improving the accuracy of MSA. The adaptation includes modifying the FPA operators. In particular, three modified versions are proposed: First, to adopt the FPA and then hybrid profile technique (FPAProfile) to optimise the *Q* and *TC* score of MSA. Second, to improve and investigate the effectiveness of the modified FPA with Profile (FPAProfile) by incorporating crossover and smart-mutation operators among the population method. Third, to improve and investigate the effectiveness of the modified FPA by improving the local pollination (search). Thereafter, the results were compared against 16 other methods using two standard benchmark datasets. The IFPAProfile method achieved a better result for the RV12 group from Balibase 3.0 dataset in terms of Q and TC scores with 94.89 and 87.51 respectively. In the OXbench dataset, the IFPAProfile method achieved the best result in terms of Q score with 90.11.

CHAPTER 1

INTRODUCTION

1.1 Background

Bioinformatics combines biology and computer sciences in one field to manipulate biological data to solve the biological challenges. This field is an interdisciplinary research area. In reality, bioinformatics in early 1970 was used from Paulien Hogeweg Fangand Ben Hesper. After that, in 1978 (Dayhoff and Schwartz, 1978; Hogeweg, 1978) coined bioinformatics as *"the study of informatic processes in biotic systems"* (Paulien, 2011; Nebel, 2014).

Biological data has increased in size in recent years. This has increased the importance of bioinformatics research to handle biological problems better. The growing protein, RiboNucleic Acid (RNA), and DeoxyriboNucleic Acid (DNA) databases require more efficient and faster methods to solve problems (Rigden et al., 2016). The examples of the most popular protein databases are UniProt, EXPASY, and RCSB (UniProt, 2018; RCSB, 2019; EXPASY, 2019).

On January 1995, The Brookhaven Protein Databank PDB, (Abola et al., 1987) was known to contain about 3091 sequence entries, which have increased to about 100 per month (Murzin et al., 1995). The size of the UniProt database grew from 550116 to 559228 sequence entries in 2015 and 2019 respectively (Uniprot., 2016; Uniprot., 2019). As mentioned earlier, the rapidly growing protein database requires fast sequence comparison methods to control and manage large data. The main role of bioinformatics is to classify, store, collect (Jacques, 2004; Guerfali et al., 2019), analyse, process, predict, and retrieve information (Hassanien et al., 2013). These roles present researchers with challenges to generate efficient and faster algorithms for organising and analysing enormous amount of data.

The problem with Multiple Sequence Alignment (MSA) may be considered as a problem of optimisation (Taheri and Zomaya, 2009; Szalkowski, 2013; Zambrano-Vega et al., 2017b; Rani and Ramyachitra, 2018; Retzlaff and Stadler, 2018) whereas its objective involved the maximising, as well as the minimising of a scoring function (Subramanian et al., 2005; Xu and Chen, 2009; Retzlaff and Stadler, 2018).

This problem encourages researches to involve with heuristics (Zablocki, 2007) as well as being NP-complete (or Non-Polynomial) (Wang and Jiang, 1994; Bonizzoni and Della Vedova, 2001; Just, 2001; Sun et al., 2012; Mora-Gutiérrez et al., 2015; Issa and Hassanien, 2017; Rani and Ramyachitra, 2018; Retzlaff and Stadler, 2018). Several approaches have been considered in order to assess their effect on the alignment optimisation. The main goal of this optimisation is to determine the best score without exhausting all ideas (Van Walle et al., 2004; Zambrano-Vega et al., 2017c). Over the years, the two main methods used in approaching this problem are evolutionary and meta-heuristic algorithms. Thus, a majority of MSAs are backed up by methods involving heuristics in order to obtain accurate MSAs within the chosen parameters of the study. This includes fair computational time as well as near optimal alignment (Thompson et al., 1994). Initial alignments are achieved through the use of various MSA tools (Zhang and Kahveci, 2006).

Meta-heuristic has the featured of performing many different solutions concurrently possess the feature of consolidating an exploratory search based on the solution space and exploitation of the present solutions (Chengpeng Bi, 2008; Fister et al., 2013; Hamza et al., 2015; Ting et al., 2015; Jain et al., 2019). In this context, Chellapilla and Fogel (1999) designed an evolutionary programming (EP) method to address the MSA problem. At the same time, consolidating evolutionary and progressive-based approaches was produced by Kupis and Mandziuk (2007). There are no restrictions in evolutionary algorithm on the number of sequences or their length. Hence, an evolutionary algorithm is very flexible in optimisation with low complexity in short length of the sequences (Kupis and Mandziuk, 2007; Hussein et al., 2017). Although, there have been many attempts to apply the meta-heuristic in bioinformatics, available meta-heuristic methods are still not producing completely or professional solutions (Das et al.., 2008; Abu-Srhan Al Daoud, 2013). Flower Pollination Algorithm (FPA) is quite flexible in optimisation with a low complexity (Diab and El-Sharkawy, 2016).

Many characteristics in FPA help to construct the MSA over conventional optimisation techniques. FPA has two major forms to produce the flowers: global pollination and local pollination. Global pollination constructs the overall of the MSA by improving the best solution. However, the local pollination will construct the MSA by improving the neighbour solution over constancy and pollinator acts according to the subsequent rules: (a) Biotic and cross-pollination are considered a global pollination process with pollen-carrying pollinators performing Lévy flights. (b) Abiotic and self-pollination are considered as local pollination. (c) Flower constancy considers that the reproduction probability is proportional to the similarity of two flowers involved. (d) Local pollination and global pollination are controlled by switch probability $p \in [0, 1]$ (Yang, 2012). FPA's hybridisation has improvements in various disciplines, such as operations research, computer science, and engineering (Diab and El-Sharkawy, 2016).

MSA is aligned by the use of profile technique. To align another MSA to a single sequence using dynamic programming (Simossis et al., 2002) (Chakrabarti et al., 2006), the profile technique of the two MSAs have relevant applications in computational biology. The single sequence versus the increase in the amount of information in the profile can result in more precise homology detection and precise alignments (Pei et al., 2003; Ohlson et al., 2004; Ohlson and Elofsson, 2005). A profile refers to the representation of related sequence groups usually on the basis of a multiple alignment of the sequences (Yona and Levitt, 2002). MSA is aligned to another MSA or to a single sequence by means of dynamic programming through the use of profile technique (Simossis et al., 2003).

Many stochastic methods, however, are regarded for enhancing the preciseness of alignment. Out of this methods, Genetic Algorithm is oftenly used by many researchers. The genetic algorithm (GA) is an evolutionary algorithm and it is presumed to be the most commonly used. It has become a conventional and classic method. It has fundamental genetic operators that have inspired many later algorithms, much has been introduced in detail by Yang and He, (2019). Moreover, the fitness function is obviously detached from the other components of the algorithm,and that is the benefit of GA. Hence, one change in the fitness function is sufficient to modify the GA for a new problem (Chowdhury & Garai, 2017). GA was used by Horng et al. and he recounted that efficiency and good performance was obtained in most of the data sets with highly comparable sequences and long lengths (Horng et al., 2005).

1.2 Motivations and Research Problems

By deciphering the MSA problem, scientists may be able to extract and identify sequences in human genomes (Wang, 2007; Franke, 2011). An arrangement of sequences of two or more residues in which the similarities are maximised between sequence alignments is significant in this kind of research (Rohit and Banka, 2015).

A lot of global MSA methods are rooted from tree-based approaches (Altschul et al., 1989) wherein multiple alignments are developed from pairwise alignments. The problem with this method is that an assumption that a tree exists in which the relationship is described between sequences is made. In addition, taking into consideration that until if the tree is perfect, it is not certain that the alignment along a tree may yield an optimised alignment (Altschul and Lipman, 1989; Layeb et al., 2009).

The primary objective of global multiple sequences is the improvement of accuracy of the methods for the alignment of the sequences in the multiple sequence alignment. This is important as the reconstruction of the phylogenetic trees, determines the function of previously unknown proteins through the alignment of its sequences with other known proteins (Wang, 2007; Franke, 2011).

This pursuit of finding ways to improve the overall accuracy of MSAs are becoming very challenging in the field of bioinformatics (Wang, 2007; Bucak and Uslan, 2011; Zhu et al., 2016). Contrary to this, the available techniques available for this problem are very minimal techniques in terms of accuracy (Zablocki, 2007). Most, if not all, heuristic approaches (Notredame and Higgins, 1996; Uren et al., 2007; Rubio-Largo et al., 2016) are researched and developed in order to attain the most optimised alignment, approximating an efficient solution within the experiment’s timeframe (Gotoh, 2014; Zambrano-Vega et al., 2017b).

In dealing with intricate optimisation problems, a meta-heuristic method, such as population-based and local search-based methods are proven to be most successful (Al-Betar and Khader, 2012; Hadwan et al., 2013; Fong et al., 2015; Zambrano-Vega et al., 2017b; Yousri et al., 2019). In exploring search space in its entirety, population-based methods have been proven to be efficient. However, a disadvantage is that they don’t excel at finding the exact local optimal solution in the algorithm converging regions. A contrast to this, are local search based methods which are proficient in the calibration of the search space region, as well as in the location of a precise local optimal solution (Fesanghary et al., 2008). Nonetheless, the path they use in the search space allows them to explore it without the process of performing a wider examination of the whole search space. An approach that could be used to handle MSA problems is to ensure the equilibrium between the global exploration and local exploitation. The former is strengthened by methods based in population, while the latter is improved with local-search based methods.

The Flower Pollination Algorithm (FPA) is an example of inspired meta-heuristic algorithm. This algorithm has been very popular over the recent years. This is due to the fact that compared to other algorithms, FPA has achieved highly comparable results, as well as its suitability to solve various real-world problems. (Diab and El-Sharkawy, 2016) such as the improvement of Fractal Image Compression (FIC) (Kaur et al., 2013), to optimise the Wireless Sensor Network (WSN) lifetime (Sharawi et al., 2014). Additionally, it is utilised to resolve the graph colouring issue (Bensouyad and Saidouni, 2015). Afterwards, such algorithms are utilised for solving the planar graph colouring problem based on the use of four colours in a more effective and precise manner (Bensouyad and Saidouni, 2015; Wang et al. 2015), and prove to be useful in the determination of protein essentiality (Leiet al., 2018). Whereas its local search intensification ability is comparatively weak (Nabil , 2016; Xu et al, 2018; Nasser et al, 2018; Fouad and Gao, 2018). Hence, the FPA needs enhancement to be suitable to manipulate the MSA.

In all the previous references for this study (Notredame, 2002; Yamada et al., 2006; Ortuño et al., 2013; Kaya et al., 2016; Leiet al., 2018), it can be seen that there is no clear and concise method available in solving the MSA problem. Even as new sets of algorithms begin to emerge, much improvement in terms of accuracy, computational complexity, as well as scalability must still be required. The motivation of this research is to know if the flower pollination algorithm can be utilised, hybridised and modified to improve and optimised the global MSAs' for high accuracy.

The most important criteria that may determine an effective MSA method design is the accuracy. It has been observed that the finding of an accurate alignment from primary sequences can be considered a computationally NP-hard problem (Just, 2001; Bonizzoni and Della Vedova, 2001; Wang et al., 2015; Kaya et al., 2016; Rubio-Largo et al., 2016; Issa and Hassanien, 2017; Rani and Ramyachitra, 2018). Though several methods have been tried and tested, problems such as low accuracy, large order computational complexity, huge memory requirements, and their inability to align multiple number sequences still arise (Wang et al., 2015). Note that more explanations of justifying to choose to solve the MSA are detailed in Appendix A.

Likewise, the creation of the optimal global alignment with a number of possible solutions means that it is growing in an exponential rate with an increasing number of sequences and lengths. An example of this would be the possible alignments of two sequences of size w and r which is seen in Table 1.1. It is obvious that the number of alignments is growing in an exponential rate as the length of the sequences increases, meaning they are directly proportional as illustrated in Equation 1.1 (Arenas-Díaz et al., 2009).

$$f(w,r) = f(w-1,r) + f(w-1,r-1) + f(w,r-1) \tag{1.1}$$

where the value of the $f(w,0) = f(0,r) = f(0,0) = 1$

Table 1.1 The Number of Possible Alignments for Two Sequences

w,r	No. of possible alignment
1,1	3
2,2	13
3,3	63
4,4	321
5,5	1683
6,6	8989
7,7	48639
8,8	265729
9,9	1462563
10,10	8097453

1.3 Research Questions

The main research question is how to enhance the accuracy of the MSA. This research intends to answer the following questions:

- Can the adaptive flower pollination algorithm and then hybrid profile technique address the MSA problem and improve the accuracy?
- Can a crossover and smart-mutation operators in the flower pollination provide better accuracy than the modified flower pollination algorithm that

described in the previous point in order to increase the diversity and the intensification of the search in the solution and increase the exploitation?

- Can the improved local search for FPA that is escaping from local optima provide higher accuracy better than the modified algorithm that described in the previous point?

1.4 Research Objectives

The aim of this proposed algorithm is to conduct an empirical analysis by using the flower pollination algorithm as a meta-heuristic approach in order to solve each encountered MSA problem. To achieve this aim, this study embarks on the following objectives:

1. To investigate and adopt the Flower Pollination Algorithm (FPA) and hybrid profile technique for MSA problem.
2. To enhance the accuracy of the modified FPA by adding a crossover and smart-mutation operators within the population for the MSA.
3. To further enhance the accuracy of the modified FPA by improving the local search of FPA for MSA.

1.5 Research Scope and Limitation

This thesis concentrates on improving the accuracy of the MSA by adapting the meta-heuristic flower pollination algorithm to manipulate the MSA problem by utilising the sequences of protein data.

The limitation of this research is the run-time, which comes from the search space that is extremely huge in the MSA problem when using the profile technique. The reason behind that refers to the construction of the MSA that adds a single sequence at each time until the end of the additional time. Additionally, the results in all the groups are unstable, where the accuracy of the produced method (i.e The. IFPAProfile-MSA) demonstrates that there is a variation within each group of the overall six groups.

1.6 Research Contributions

To the best of our knowledge, this algorithm is the first algorithm that efforts in adapting the FPA in addressing the MSA problem. Note that the three proposed methods were proposed sequentially, each to beat the shortcomings of the past one and hence acquire a high accuracy for MSA.

The expected contributions of the present study are summarised as follows:

a) A method that is based on the adapted meta-heuristic algorithm and then hybrid profile technique known as the flower pollination algorithm with profile technique (the basic FPA for MSA, FPAProfile-MSA) is proposed in order to solve the MSA problem.

b) A modified method (includes crossover and smart-mutation operators to the FPA for MSA, CFPAProfile-MSA and MFPAProfile-MSA) is proposed for MSA by increasing the diversity of the population to intensify the search in the solution and exploitation to increase the accuracy score.

c) A modified method (improves local search for IMFPAProfile-MSA and IFPAProfile-MSA) is proposed to improve the local optima for MFPAProfile-MSA to generate a new solution and improve the accuracy score using the effective features from the first and second best solutions.

1.7 Overview of Thesis

The remaining chapters in this thesis include of seven chapters organised as follows:

Chapter 2 reviews the MSA approaches and methods. A broad literature review is presented explaining and analysing the existing MSA methods and an overview of the flower pollination (FP) algorithm is given.

Chapter 3 the analogy between pollination and optimisation terms is provided. In addition, present the research methodology employed for five variants of proposed Flower Pollination (FP) methods: (FPAProfile-MSA, CFPAProfile-MSA, MFPAProfile-MSA, IMFPA-MSA, and IFPA-MSA). Thereafter, reviews the biological dataset and the benchmark that will be used in this thesis. Also, includes a comprehensive description of the methodology and the procedures that were carried out.

Chapter 4 presents an initial investigation of adapting the FPA for the MSA where a hybrid profile technique is proposed in order to improve the MSA solving process. The profile technique is added to the FPA in order to overcome the limitation of the initial alignment that is built based on the FPA. The proposed method is called the FPAProfile-MSA method.

Chapter 5 introduces the hybrid of the FPA with GA in order to enhance the quality of the multiple sequence alignment. As discussed in Chapter 4, the quality of the flower from FPAProfile-MSA method is successfully applied by using the profile technique in order to produce a good quality for the multiple sequence alignment when approaching a close optimal solution. Therefore, the FPA is integrated with the GA to improve the exploitation capabilities based on two operators, which are the crossover and mutation. Additionally, the experimental test and results of the proposed methods (CFPAProfile and MFPAProfile) for the MSA is described in this chapter where the performance of the CFPAProfile and MFPAProfile for the MSA is presented.

Chapter 6 involves the improved local search of the Flower Pollination algorithm is presented by using efficient features from the best and the second-best solutions in order to improve the quality of the multiple sequence alignment. Therefore, the improvement of local optima (pollination) may lead to encounter two cases: (1) effectively explore the search space, and (2) select and exploit the global best solution efficiently. Finally, a comparison of the proposed methods' accuracy with commonly used methods and summary is presented.

Chapter 7 provides and draws the conclusion remarks of the research findings. Further, the suggestions and recommendations for the future research are presented in this chapter.

CHAPTER 2

RELATED WORK

2.1 Introduction

This chapter includes the background, materials and the state of art pertaining to the thesis: Multiple Sequence Alignment (MSA). Understanding these perfectly is needed to know the dataset benchmarks, gap penalty, alignment score, scoring scheme, MSA approaches, existing methods. This chapter contains the representations of MSA alignment in Section 2.2. Section 2.3 explains the measurement scoring scheme. The alignment score functions are explained in Section 2.4. Then, Section 2.5 provides details of current MSA approaches. The flower pollination algorithm is explained in Section 2.6. Finally, Section 2.7 the chapter ends with a conclusion.

2.2 Representations of MSA Alignment

MSA alignment can be represented in various ways. Ordinarily, sequences one above the other can produce the final alignment. There are many existing methods used to represent the MSA alignments as illustrated below:

2.2.1 Sequence Logo

Schneider and Stephens (1990) used a graphical method to display the patterns in the protein sequence alignment. They represent the sequences by using a stack to put each sequence over of each other to create a logo from them. On another hand to present the mean of the sites, they used sequence logo graphs (Schneider 2002; Yachdav et al., 2016; Gao et al., 2017; Dey et al., 2018). Sequence logo generated from a combination of a protein aligned and move the information position (Anzald et al., 2012).

2.2.2 Multiple Number Strings

The number and position of gaps in multiple alignments are represented by the number strings (Conrad et al., 2004; Pérez-Serrano et el., 2018) and many authors used it (Horng et al., 2000; Horng et al., 2005; Chengpeng Bi, 2008). Any sequence can be displayed as a number of location/position $(x_{1,1}, x_{1,2}, \dots, x_{m-1,1}, x_{1,m})$. Every number is unique and corresponds to the position of space in an alignment, and all the number of spaces of the sequence in an alignment is the string length.

2.2.3 A Matrix of Gaps Position

Gaps representation as a "-" in the matrices by inserting spaces in the sequences or a gap means that one amino acid residue has been deleted (Yamuna and Elakkiya, 2015; Mokaddem Lai et al., 2018). Lai et al. (2009) discussed and recommended this method. In literature, used the gaps position in the biological field when the sequence is not aligned with another (Jesper, 2010; Kumar and Om, 2019). The sequence alignment is represented as two-dimensional arrays of residue/letters, which are rows, represent the sequences, and every column represent the position of residue in that row. The duo of integers (x, y), which is the x^{th} row and y^{th} column and representing the x^{th} sequence (Notredame and Higgins, 1996; Gondro and Kinghorn, 2007; Silva et al., 2008; Nguyen, 2008; Mohsen et al., 2018).

2.2.4 de Bruijn Graphs

Pevzner et al. (2004), Raphael et al. (2004) and Jones et al. (2006) used the De Bruijn graphs and recommended. This type of graph uses recurring representation. Whereas, the de Bruijn Graph repeats families in the sequences and represents multiple alignments as a weighted graph (Compeau et al., 2011; Fu et al., 2015).

2.2.5 Partial Order Graphs

Partial order alignment (POA) is used to present the alignment of multiple sequences and replace the row-column for the sequences to the linear by a directed acyclic graph (DAG). It uses the partial order graphs in MSA and represents the letters by nodes (Lee et al. 2002; Kavya et al. 2018).

2.2.6 Summary of Representations of MSA Alignment

As aforementioned, some of the representations represent the sequences as a graph or gaps position to represent the MSA alignment. For example, the sequence logo, de Bruijn graphs, and partial order graphs used a graph to represent the alignment. On the other hand, multiple number strings and a matrix of gaps position have used the position of gaps to represent the MSA alignment. Lai et al. (2009) and Chakrabarti et al. (2013) recommended using the position of gaps especially when the sequences are not aligned. The recent methods used a position of gaps (Kaya et al., 2014; Zhu et al., 2016; Mohsen et al., 2018).

The advantage of the position of gaps is representing the sequences as two-dimensional arrays (Silva et al., 2008; Nguyen, 2008). On the contrary, the disadvantage of using the graph to represent the sequences is needed to play in k-dimensions, where *k* is the number of sequences (Banerjee et al., 2013; Chakrabarti et al., 2013; Nguyen et al., 2016). It concluded that the position of gaps is a better representation for sequences because this does not consume time in the first processing.

2.3 Measurement Scheme

In order to find a good alignment from a set of sequences to know which sequence related to another should be estimated by the measure of distance or similarity by the probability methods. Matches, similarity mismatch, deletion, insertion and substitution are used as a scored. Researchers split the scoring function into gap penalties and substitution matrices (Al-Shatnawi et al. 2015; Wang, 2007). To evaluate the alignment accuracy requires consideration of the measurement scheme such as object function.

The scoring scheme is divided into distance scoring scheme and a similarity scoring scheme. The former is a function to measure the variance between the sequences and the second is a function to measure the similarity of sequences (Gondro and Kinghorn, 2007). There is an inverse relationship between similarity and distance as smaller distance means larger similarity and vice versa.

2.3.1 Similarity Scoring Scheme

Wang (2007) mentioned that alignment is a mixing set of residue-space pairs and two residues pairs. Whereas, the gap penalty score the residue-space pairs and the substitution matrix score the two residue pairs.

2.3.1(a) Similarity Substitution Matrix

In the case of pairwise sequence alignment for scores, the scores on substitution matrix are divided into two classes are based on similarity or percentage identity. A substitution matrix contains the number of rows and number of columns to make a matrix. The matrix will be *4* × *4* for nucleic acids and will rise to *20* × *20* for amino acids in protein, and this type of measurement reflects the probability of evolutionary events or conservation occurrence (Gondro and Kinghorn, 2007; XU, 2007). In this kind of measurement, it provides the probability of a conservation occurrence or substitution (Gondro and Kinghorn, 2007). Measuring the distance among the sequences relies on the probability of the specific residue mutation to another (Nguyen, 2011). This substitution/mutation is calculated by checking matches and mismatches or identical and non-identical depending on identifying matrix for amino acid alignment. The identify matrix is the simplest scoring scheme, where residues are classified as matches pairs present as a 1 score and mismatch as a 0 score (Durbin, 1998; Gondro and Kinghorn, 2007).

For protein alignment, increasing the residue from *4* to *20* comparing with DNA and scoring scheme of *0* for mismatch and *1* for a match is not enough. It makes the scoring scheme more complicated than DNA. In evolutionary time the probability of the two residue types will change through mutations. This status is described as substitution matrix. This is used to estimate how effective is the match between two given types of residues when they are aligned within a sequence alignment. Take into account the Log-odds are the most mostly used matrices which compare the probability of mutating the amino acid to another amino acid during the time (Durbin, 1998; Gondro and Kinghorn, 2007). The most widely used matrices for substitution matrix and protein alignment are Point Accepted Mutation (PAM) which is the first matrix generated substitution matrix (Dayhoff et al., 1978) and in 1992 Blocks Substitution Matrix (BLOSUM) was proposed for protein alignment, which is PAM matrices derived from symmetric sequences but PLOSUM matrices derived for protein alignment (Dayhoff et al., 1978; Henikoff and Henikoff, 1992; XIN and WEI-MOU, 2006).

2.3.1(b) Gaps Penalty

Gusfield (1997) introduced a standard definition for a gap as “a maximal, consecutive run of spaces in a single string of a given alignment”. It can illustrate is as the following: Figure 2.1 shows an example of alignment with eight spaces distributed into four gaps (Gusfield, 1997; Pérez-Serrano et el., 2018; Turjanski and Ferreiro, 2018).

```
T  C  T  T  C  A  -  C  -  -  -  A
T  C  -  -  -  -  A  C  C  T  A  A
```

Figure 2.1 Alignment Example

The position of the gaps is essential in the alignments, whereas the gaps between more related sequences are high (Thompson et al., 1994). In order To evaluate multiple alignments, it must use a formula to find the maximum value of proteins. The spaces distributed in an alignment are defined as a gap penalty. In the biological field, many substitutions happen by insertion or deletion residue/s to a sequence. Insertion or deletion causes large gaps in alignment. Gap penalty represents a negative score in the function (Richer et al., 2007; Wang, 2007). Richer et al. (2007) mentioned that two extremely used models of gaps are:

- Linear gap model - in this model, same penalty always is given to the alignment wherever it is placed. It is measured this model by proportional to the length of the gap and calculated by following Equation:

$$Gap_{penalty} = m * g\_open \qquad (2.1)$$

where $g_{open} < 0$ is the opening penalty of a gap and m the number of consecutive gaps.

- Affine gap model - in this model deferent penalty are given to the insertion of a new (open) gap and extension gap, but a new gap has greater penalty than extension gap and can be stated by the following Equation:

$$Gap_{penalty} = g_{open} + \ (m-1) * \ g_{extension} \qquad (2.2)$$

where $g_{open} < 0$ is the gap opening penalty and $g_{extension} < 0$ the gap extension penalty and are such that $|g_{extension}| < \ |g_{open}|$.

In conclusion, according to Chao and Zhang (2008) and Kamal et al. (2012), the affine gap model is considered to be more appropriate when aligning protein sequences.

2.3.2 Distance Scoring Scheme

While most widely used pairwise alignment programs present alignment scores in terms of similarity between residues, most multiple sequence alignment programs use a distance measure (sometimes referred to as a cost) to compute an alignment score. Distances have properties that allow sensible discussions of the simultaneous relationship between three or more objects, such as sequences, while similarity measures generally only permit the simultaneous consideration of relationships between pairs of objects (Nicholas et al., 2002). The notion of distance among alphabets, for example, is extended to multiple alphabets (Altschul, 1989). Distance scoring is used in many fields such as bioinformatics, computer science, mathematics, and graph theory. In bioinformatics, multiple sequence alignment uses distance scoring or distance matrix is a matrix that has a group of values of results from the distance between the sequences. In other words, the value is proportionate according to the similarity among them, which is the distance among pairwise alignment. These values are used for many purposes such as building a guide tree, phylogenetic trees (Blackshields et al., 2010) and constructing MSAs.

2.4 Objective Functions

Because the high multiple sequence alignment is composed of the alignment from three sequences and more, it is considered as an NP-complete problem (Wang and Li, 2004; Kaya et al., 2014; Chatzou et al., 2015; Rubio-Largo et al., 2016; Issa and Hassanien, 2017; Rani and Ramyachitra, 2018). MSAs are the most used computational applications for analysing sequences in the biological field, such as function prediction, structure prediction, Polymerase Chain Reaction (PCR) primer design, phylogenetic analysis (Notredame et al., 1998), motif finding, phylogenetic reconstruction (Chatzou et al., 2015), gene prediction, protein classification, and identification of conserved motifs and domains (XU, 2007).

Heuristic algorithms used to find the accurate for multiple sequence alignment by using various criteria by reaching the approximate optimal solution. Object function (scoring function) is a core of heuristic algorithm to estimate alignments and improve the optimal solution (Wang and Li, 2004).

Object function is a critical parameter in multiple sequence alignments, as it defines the designing accuracy of an MSA and its predictive ability. It is used in evolutionary rebuilding involving the highest weighted similarities (Chatzou et al., 2015).

In order to evaluate the final sequence alignments, a score is used to assess the score of the sequence alignment when compared between the potential alignments. Each similar pair sequence means these sequences shared the same evolutionary ancestral state. Consequently, the object function contains codes of the sequence alignment. Most of the effective algorithms are applied an objective function to reach the near-optimal solution or an optimal solution. Thus, the objective function is a mathematical function (Anisimova et al., 2010).

There are many models to assess the objective function for a given MSA. In the last thirty years, many authors used score methods and their amendments to compute the alignment scores (see subsections 2.4.1 - 2.4.4).

2.4.1 Sum-of-Pairs Score (SPS)

The sum-of-pairs (sometimes refers to it as *SP*) score is widely utilised to evaluate an alignment (Hassanien et al., 2013). It was proposed by Carrillo and Lipman (1988). *SP* is used as a standard method to evaluate the quality of MSA (Durbin, 1998; Trystram and Zola, 2007; Yao et al., 2015). MSA is an NP-complete problem for *SP* scoring (Wang and Jiang, 1994; Do and Katoh, 2008; Botta and Negro, 2009; Kaya et al., 2014; Issa and Hassanien, 2017; Zambrano-Vega et al., 2017c). *SP* is a total number of exactly aligned pairs of residues in the alignment divided by the sum of numbers of aligned pairs in the reference alignment (Wang, 2007; Wilm et al., 2008; Hassanien et al., 2013). *SP* defines the measures of a multiple alignments of N sequences as the total of the measures of the $\frac{N\,(N-1)}{2}$ pairwise alignments (Durbin, 1998; Do and Katoh, 2008).

To calculate the score of *SP* for the alignments of *N* sequences as mentioned earlier, suppose M is a number of columns in the test alignment, the i^{th} column is represented as $A_{i,1}$, $A_{i,2}$, A_{i3}, $A_{i,4}$,, …, $A_{i,N}$ in the test alignment. For each dual of residues A_{ij} and A_{ik}, the values of $p_{ijk} = 1$ when the residues A_{ij} and A_{ik} are aligned together in the reference alignment. Otherwise the value of the $p_{ijk} = 0$. To define the score S_i for the ith column as following:

$$s_i = \sum_{j=1, j\neq k}^{N} \sum_{k=1}^{N} P_{ij}k \tag{2.3}$$

The SP for the alignment is then:

$$SP = \sum_{i=1, j\neq k}^{M} = \ S_i \ / \sum_{i=1}^{M^r} S_{ri} \tag{2.4}$$

where M_r is the number of columns in the reference alignment and S_{ri} is the score S_{ri} for the i^{th} column in the reference alignment. Furthermore, finding an alignment that is biologically meaningful is not trivial since the *SP* score may not reflect the biological significances (Nguyen, 2008; Silva et al., 2010). The challenge in MSA is to maximise the *SP* score (Thompson and Poch, 2005).

2.4.2 Weighted Sum-of-Pairs Score (WSP)

The weighted Sum-of-Pairs Score was first introduced by Altschul et al. (1989), in which it was implemented in the MSA package (Lipman et al., 1989). The presented *WSP* score showed an iterative MSA method in which the *WSP* score was used to optimise the given sequences. Once this was done, the *SP* score was them extended as the *WSP* score in order to make the scoring of the pairwise alignment differently scored from the previous scores. This is used in order to realize the relationship between sequences as well as to yield a cost in the columns of alignment between the pairs of aligned residues to create gap costs for the gaps. Calculating the *WSP* as a mathematical function by computing the total score for all sequences can be expressed as the following:

$$WSP\ (A) = \sum_{i=0}^{m} \sum_{j=i+1}^{m-1} W_{ij} \sum_{k=0}^{l} s\left(a_{ijk}, b_{ijk}\right) \tag{2.5}$$

where m represents the number of sequences, l represents the length of the alignment A, the column position of the alignment is given a variable k, W_{ij} is a representation of the weight given to the sequence pare, and $s\left(a_{ijk}, b_{ijk}\right)$ is the similarity cost of the two sequences a_{ijk} as well as b_{ijk}. The gap-open and gap-extend penalties are representations of the cost function which includes opening and extending gaps (Naznin et al., 2012; Yadav and Banka, 2015; Zhu et al., 2016).

2.4.3 Column Score (CS)

Column score (sometimes referred to as total column (*TC*)) is used to measure the sequence alignment, the definition of the *CS* is the number of exactly aligned columns in the alignment divided by the whole number of aligned columns in the reference alignment (Thompson et al., 1999b; Liu et al., 2010). To consider a column as a matched column, there should be no differences between the two columns. *CS* and *SP* methods are used to compute the score of the test alignment (Thompson et al., 1999b; Wang and Jiang, 1994; Nguyen, 2008; Yao, et al., 2015; Zhu et al., 2016).

As described earlier, i^{th} column in the test alignment, in each column if all the residues aligned in the reference alignment, the score $C_i = 1$, otherwise, $C_i = 0$.

The TC for the alignment is then:

$$TC = \sum_{i=1,}^{M} = \ C_i\ /M \tag{2.6}$$

Furthermore, an effort to gain a superior level accuracy raises the score of aligned sequences and perhaps the proportional weight of the amount of singular information is enclosed in the sequence (Altschul and Lipman, 1989; Sibbald and Argos, 1990). The *WSP* tries to decrease the impact of redundant information from highly related sequences (Zablocki, 2007). The Equation of the weight represents the rate equal to a percentage identity (PID) which calculates both of the matched and mismatched pairs of aligned sequences over the alignment (Zablocki, 2007) as follows (excluding gaps):

$$PID = \frac{matches}{(matched+mismatched)} \tag{2.7}$$

2.4.4 Objective Functions Summary

From the earlier description, each object function complements the other to get the best score for the aligned sequence. For instance, the most commonly method used appraoches are sum-of-pairs (*SP*) and column score (*CS*) or total column (*TC*) as a measurement score because it is acceptable accuracy, presently used in MSAIndelFR and has been noted for its robustness and relative speed (Al-Shatnawi et al., 2015). The speed is a primary advantage in SP because it does not need a tree (Nicholas et al., 2002).

Other researchers used the Weighted Sum-of-Pairs (*WSP*) Score object function such as MSAProbs. It has been noted as efficient by Liu et al., 2010 (2010). Liu et al. (2010) presented two reasons for the efficiency of this object function: 1) using weight approach is to gain more accurate alignments than the non-weighted one. 2) perform it by using two iterations (the default value) presents a good tradeoff among accuracy and run time.