ANCHOR POINT APPROACH FOR INITIAL POPULATION OF BAT ALGORITHM FOR PROTEIN MULTIPLE SEQUENCE ALIGNMENT

AZIZ NASSER BORAIK ALI

UNIVERSITI SAINS MALAYSIA
2016

ANCHOR POINT APPROACH FOR INITIAL POPULATION OF BAT ALGORITHM FOR PROTEIN MULTIPLE SEQUENCE ALIGNMENT

By

AZIZ NASSER BORAIK ALI

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

September 2016

ACKNOWEDGMENT

I am thankful to almighty Allah, most Gracious, who in his infinite mercy has guided me to complete this PhD work. May Peace and Blessings of Allah be upon his prophet Muhammad (peace be upon him). Although I am liable for this research and its findings, I have to show gratitude to those magnificent people who help me throughout the duration of this research.

I am very much grateful to my academic supervisor, Professor Rosni Abdullah, and my co-supervisor, Dr. Ibrahim Venkat for providing me with constant guidance support, and feedback throughout this research. Their dedication, the time they spent on coordinating and supervising the whole thesis and their enlightening direction enable me to remain focused on the study.

I owe many thanks to the staff of School of Computer Sciences in Universiti Sains Malaysia (USM) for their help and support in many ways.

Finally and the most important, I would like to thank my parents, my sisters and brothers, my wife, and my children for their unconditional support, prayers and for all of the sacrifices that they have made throughout my life. I dedicate this research to them.

TABLE OF CONTENTS

ACKNOWLEDGEMENT ii

TABLE OF CONTENTS iii

LIST OF TABLES ix

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xv

ABSTRAK xvi

ABSTRACT xviii

CHAPTER 1 - INTRODUCTION 1

1.1 Overview 1

1.2 Background 2

1.3 Multiple Sequence Alignment 4

1.4 Motivation and Problem Statement 6

1.5 Research Objectives 13

1.6 Research Contributions 14

1.7 Scope and Limitations 14

1.8 Thesis Outline and Organization 15

CHAPTER 2 - LITERATURE REVIEW 17

2.1 Introduction 17

2.2 Formal Definitions for MSA: 17

2.3 Sequence Alignment Measure 19

2.3.1 Substitutions Matrix 20

2.3.2 Gap Penalty 21

2.4 Multiple Sequence Alignment Approaches 22

2.4.1 Progressive Approach 22

2.4.2 Consistency-Based Approach 26
2.4.3 Iterative Approach 30
2.4.3 (a) Non-Stochastic Iterative Methods 31
2.4.3 (b) Stochastic Iterative Methods (Heuristic Methods) 36
2.4.4 Segment/Fragment-Based Approach 41
2.4.5 Summary of MSA Methods 46
2.5 Benchmark Datasets 52
2.6 Bat Algorithm 53
2.7 Summary 58
CHAPTER 3 - RESEARCH METHODOLOGY 60
3.1 Introduction 60
3.2 Proposed Research Methodology 60
3.2.1 Problem Identification 61
3.3 Analysis of Current Techniques 62
3.3.1 The Proposed Methods 62
3.3.2 Automatic Detection of Anchor Points by Using SNN Clustering Algorithm 62
3.3.3 Bat Algorithm with Anchor Points Technique for MSA 64
3.3.4 Modified Bat Algorithm with Enhanced Initial Population for MSA 64
3.4 Experimental Design 65
3.4.1 Parameters Settings 65
3.4.2 Benchmark Datasets 66
3.4.3 Evaluation Measurement and Comparison 67
3.4.4 Statistical Analysis 69
3.4.5 Comparison with Commonly Used Methods 70
3.4.6 Test Platform 70

3.5 Summary 70

CHAPTER 4 - AUTOMATIC DETECTION OF ANCHOR POINTS USING SNN CLUSTERING ALGORITHM FOR MSA 72

4.1 Introduction 72

4.2 Detect Anchor Points by Using Basic Shared Near Neighbours 74

4.2.1 Anchor Points by Shared Near Neighbours Algorithm (AP-SNN) 74

4.2.2 Overlapping Anchor Points 78

4.2.3 Consistent Partial Alignment Columns 80

4.3 SNN with Weighted Similarity (SNN-WS) 82

4.4 Combining SNN-WS with Segment-Based Alignment Approach (SNN-SB) 86

4.5 Constructing Final Multiple Alignment from Partial Alignment Columns 87

4.6 Results and Discussion 88

4.6.1 Anchor Points by Shared Near Neighbours Algorithm (AP-SNN) 89

4.6.1 (a) Choosing an Appropriate Parameters for AP-SNN .. 89

4.6.1 (b) The Performance of AP-SNN 93

4.6.2 Performance of Consistent AP-SNN 96

4.6.3 SNN with Weighted Similarity (SNN-WS) 98

4.6.3 (a) Choosing an Appropriate Similarity Threshold 98

4.6.3 (b) Performance of SNN-WS 100

4.6.4 Combining SNN-WS with Segment-Based Alignment Approach (SNN-SB) 102

4.6.4 (a) Choosing an Appropriate Segment Length 102

4.6.4 (b) Performance of SNN-SB 104

4.6.5 Comparison of Proposed methods Accuracy with Commonly Used Methods 110

4.6.5 (a) Global alignment benchmark: BaliBase 3.0 110

4.6.5 (b) Local Alignment Benchmark: IRMBASE 2.0 115

4.7 Summary 120

CHAPTER 5 - BAT ALGORITHM WITH ANCHOR POINTS TECHNIQUE FOR MSA PROBLEM 121

5.1 Introduction 121

5.2 Bat Algorithm with Anchor Points Based Technique for MSA (BA-MSA) 122

5.2.1 BA-MSA Score Function 124

5.2.2 Initialize the Algorithm Parameters 126

5.2.3 Structure of the Individuals 126

5.2.4 Constructing Initial Alignment by Using Anchor Points Technique 128

5.2.5 Movement of Virtual Bats 131

5.2.6 Local Search 132

5.2.7 Update the Solutions by Flying Randomly 132

5.2.8 Update the Current Global Best Solution 133

5.2.9 Verifying the Stopping Condition 133

5.2.10 Mapping a Solution to MSA 133

5.3 Modified Local Search of Bat Algorithm for MSA 134

5.3.1 Modify the Local search 135

5.3.2 Profile alignment 136

5.3.2 (a) Profile Matrix 137

5.3.2 (b) Alignment Process 138

5.4 Results and Discussion 139

5.4.1 Bat Algorithm with Anchor Points Technique for MSA (BA-MSA) 139

5.4.1 (a) Tuning the MSA Parameters 139

5.4.1 (b) The Effect of Initial Population on Alignment Quality 141

5.4.1 (c) Tuning the BA Parameters 143

5.4.1 (d) The BA Convergence for Protein Sequence 146

5.4.2 Modified the Local Search in Bat Algorithm for MSA 149

5.4.2 (a) Tuning BA Parameters 149

5.4.2 (b) The BA Convergence for Protein Sequence 150

5.4.3 Comparison of Proposed methods Accuracy with Commonly Used Methods 153

5.5 Summary 157

CHAPTER 6 - MODIFIED BAT ALGORITHM WITH ENHANCED INITIAL POPULATION FOR MSA 158

6.1 Introduction 158

6.2 Improved Initial MSA 159

6.2.1 Aligning All Residues (Nodes) 160

6.2.2 Using Various of Scoring Matrices 163

6.2.3 Including Horizontal Information into Alignments 164

6.2.4 A New Selection of Central Nodes 165

6.3 Improved Bat Algorithm for MSA (IBA-MSA) 168

6.3.1 A New Operator 169

6.4 Results and Discussion 173

6.4.1 The Performance of Improved Initial MSA Method 173

6.4.1 (a) The Effect of Aligning All Residues in Initial alignment 174

6.4.1 (b) The Effect of Using Various Scoring Matrices 175

6.4.1 (c) The Effect of Including Horizontal Information into Alignments 176

6.4.1 (d) The Effect of Using a New Selection of Central Node 177

6.4.2 The Performance of Improved Bat Algorithm for MSA (IBA-MSA) 179

6.4.2 (a) Tuning IBA-MSA Parameters 179

6.4.2 (b) The Convergence of IBA-MSA 181

6.4.3 Comparison of Proposed methods Accuracy with Commonly Used Methods 184

6.5 Discussion 188

6.6 Summary 191

CHAPTER 7 - CONCLUSION AND FUTURE WORK 192

7.1 Research Contributions 192

7.2 Future Work 194

REFERENCES 196

LIST OF PUBLICATIONS .....

LIST OF TABLES

Page

Table 1.1	Amino Acids Abbreviations (Smith and Waterman, 1981)	3
Table 2.1	The Summary of MSA Methods	47
Table 4.1	Mean of Alignment Quality (Q) Scores for Range of Scoring of Parameters (*Eps* and *MinPts*) in AP-SNN Method on BaliBase 3.0	91
Table 4.2	Mean of Total Column (TC) Score for Range of Scoring of Parameters (*Eps* and *MinPts*) in AP-SNN Method on BaliBase 3.0	92
Table 4.3	Wilcoxon P-Vaues of Q and TC Scores for Range of Scoring of Parameters (*Eps* and *MinPts*) in AP-SNN Method on BaliBase 3.0	93
Table 4.4	Performance of AP-SNN Comparing to Dialign-TX on Reference Sets in BaliBase 3.0 based on an Q and TC Scores	95
Table 4.5	Performance of AP-SNN Comparing with Dialign-TX on Reference Sets in IRMBASE 2.0 based on an Q and TC Scores	96
Table 4.6	Performance of Consistent AP-SNN Compared with Dialign-TX and AP-SNN on BaliBase 3.0	97
Table 4.7	Performance of Consistent AP-SNN Compared with Dialign-TX and AP-SNN on IRMBASE 2.0	98
Table 4.8	Mean of Alignment Quality (Q) Scores for Range of Scoring of Similarity Threshold in SNN-WS Method on BaliBase 3.0	99
Table 4.9	Mean of Alignment TC Scores for Range of Scoring of Similarity Threshold in SNN-WS Method on BaliBase 3.0	100
Table 4.10	Performance of SNN-WS Compared with Dialign-TX, AP-SNN and Consistent AP-SNN on BaliBase 3.0	101
Table 4.11	Performance of SNN-WS Compared with Dialign-TX, AP-SNN and Consistent AP-SNN on IRMBASE 2.0	102
Table 4.12	Performance of SNN-SB Compared with Other Methods in Terms of Q Score on BaliBase 3.0	105
Table 4.13	Performance of SNN-SB Compared with Other Methods in Terms of TC Score on BaliBase 3.0	105
Table 4.14	Performance of SNN-SB Compared with Other Methods in Terms of Q Score on IRMBASE 2.0	108

Table 4.15	Performance of SNN-SB Compared with Other Methods in Terms of TC Score on IRMBASE 2.0	108
Table 4.16	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of Q Score on BaliBase 3.0	112
Table 4.17	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of TC Score on BaliBase 3.0	112
Table 4.18	Wilcoxon P-values of Q Score of SNN-SB Compared to Other Aligners on BaliBase 3.0	113
Table 4.19	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of Q score on IRMBASE 2.0	117
Table 4.20	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of TC score on IRMBASE 2.0	117
Table 4.21	Wilcoxon P-values of Q Score of SNN-SB Compared to Other aligners on IRMBASE 2.0	118
Table 5.1	MSA Parameter Settings of Gap-Open and Gap-Extend	140
Table 5.2	Alignment Quality (Q) Scores for Different Gaps Penalties Settings in BA-MSA	141
Table 5.3	Alignment Quality (Q) Scores for BaliBase using BA-MSA with Random Initial and with Anchor Points Initial Solution	143
Table 5.4	Alternative BA Parameters	144
Table 5.5	Alignment Score of BA-MSA after Tuning BA Parameters	145
Table 5.6	Selected Tests from BaliBase 3.0	146
Table 5.7	Alignment Score of ProfileBA-MSA with Different Settings	150
Table 5.8	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of Q Score on BaliBase 3.0	154
Table 5.9	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of TC Score on BaliBase 3.0	155
Table 6.1	The Proposed Modification to Initial Alignment in QOMA	174
Table 6.2	Alignment Quality (Q) Scores of Modified Initial Alignment	175
Table 6.3	Alignment Quality (Q) Scores of Modified Initial Alignment with Various Scoring Matrices Compared with Other Initial Alignments	176

Table 6.4	Alignment Q Score of Initial Alignment after Including Horizontal Information (Modification 3) with Different Settings	177
Table 6.5	Alignment Quality (Q) Scores for Different Number of Groups in New Selection of Central Node Method (Modification 4) Based on BaliBase 3.0	178
Table 6.6	Mean of Alignment Quality (Q) Scores of Initial MSA of QOMA and Other Modified Methods of Initial MSA	179
Table 6.7	Alternative IBA-MSA Parameters	180
Table 6.8	Alignment Score of IBA-MSA after Tuning BA Parameters	181
Table 6.9	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of Q Score on BaliBase 3.0	185
Table 6.10	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of TC Score on BaliBase 3.0	186
Table 6.11	Wilcoxon P-values of Q Score of IBA-MSA Compared to Other aligners on BaliBase 3.0	188
Table 6.12	The Score of Objective Function (Alignment Score) for Some References	191

LIST OF FIGURES

Page

Figure 1.1	Alignment of Two Sequences	5
Figure 1.2	Local and Global Alignment	6
Figure 2.1	Unaligned Sequences	18
Figure 2.2	An Example of Multiple Alignment Sequences	19
Figure 2.3	Progressive Alignment (Cortada, 2013)	24
Figure 3.1	The Proposed Research Methodology	61
Figure 3.2	An Example of Reference Alignment from BaliBase 3.0 (Subset BB11001)	67
Figure 4.1	The Organization of Anchor Points Methods	73
Figure 4.2	The Similarity Measurement of AP-SNN	76
Figure 4.3	An Example of Overlap between Some Points Across the Sequences	79
Figure 4.4	An Example of the Output of Anchor Points from Reference Set BBS11001	80
Figure 4.5	BLOSUM62 similarity matrix	84
Figure 4.6	An Example of Segment Alignment	87
Figure 4.7	Average Alignment Quality (Q) Scores for Range of Segment Size in SNN-SB Method on BaliBase 3.0	103
Figure 4.8	Average TC Scores for Range of Segment Size in SNN-SB Method on BaliBase 3.0	104
Figure 4.9	Results of Comparison of Q scores of SNN-SB with Other Methods in BaliBase 3.0	106
Figure 4.10	Results of Comparison of TC scores of SNN-SB with Other Methods in BaliBase 3.0	106
Figure 4.11	Results of Comparison of Q scores of SNN-SB with Other Methods in IRMBASE 2.0	109

Figure 4.12	Results of Comparison of TC scores of SNN-SB with Other Methods in IRMBASE 2.0	109
Figure 4.13	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of Q score on BaliBase 3.0	114
Figure 4.14	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of TC score on BaliBase 3.0	114
Figure 4.15	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of Q score on IRMBASE 2.0	119
Figure 4.16	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of TC score on IRMBASE 2.0	119
Figure 5.1	Example of the Bat Representation	127
Figure 5.2:	Construction the Initial MSA Based on Anchor Points Technique	130
Figure 5.3	An Example of Mapping Individual Bat into MSA	134
Figure 5.4	An Example of Profile Alignment	137
Figure 5.5	The Behaviour of BA-MSA Method for Six Subsets	148
Figure 5.6	The Behaviour of ProfileBA-MSA Method for Six Subsets	152
Figure 5.7	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of Q score on BaliBase 3.0	156
Figure 5.8	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of Q score on BaliBase 3.0	156
Figure 6.1	An Example of Initial Alignment in QOMA.	161
Figure 6.2	The Final Alignment after Taking into Account the Ignored Relation between the Nodes.	163
Figure 6.3	An Example of Aligning the Nodes Starting from Left to Right	167
Figure 6.4	An Example of Aligning the Nodes by Using a New Proposed Selection	167

Figure 6.5	The Process of New operator with Profile Alignment Technique	172
Figure 6.6	The Behaviour of IBA-MSA Method for Six Subsets	183
Figure 6.7	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of Q score on BaliBase 3.0	187
Figure 6.8	Comparison of Alignment Accuracy of Proposed Methods and Commonly Used Methods in Terms of TC score on BaliBase 3.0	187

LIST OF ABBREVIATIONS

MSA	Multiple Sequence Alignment
Q	Alignment Quality
TC	Total column Score
NP-hard	Non-deterministic Polynomial-time Hard
BA	Bat Algorithm
GA	Genetic Algorithm
SA	Simulated Annealing
TS	Tabu Search
BaliBase	Benchmark Protein Alignment Database
SNN	Shared Near Neighbours
AP-SNN	Anchor Points by using SNN Clustering Algorithm
SNN-WS	AP-SNN with Weighted Similarity
SNN-SB	Combining SNN-WS with Segment-Based Alignment Approach
BA-MSA	Bat Algorithm with Anchor Points Based Technique for MSA
ProfileBA-MSA	Modified Local Search of Bat Algorithm for MSA
IBA-MSA	Improved Bat Algorithm for MSA
BLOSUM	Block Substitution Mutation
PAM	Point Accepted Mutation
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid

PENDEKATAN TITIK SAUH UNTUK POPULASI AWAL ALGORITMA BAT UNTUK PENJAJARAN PELBAGAI TURUTAN PROTIN

ABSTRAK

Penjajaran pelbagai turutan atau *Multiple sequence alignment* (MSA) adalah satu langkah asas kepada banyak aplikasi bio-informatik seperti pembinaan pokok filogenetik, ramalan struktur sekunder dan pengenalpastian motif domain dan yang dipulihara. Kebolehpercayaan dan ketepatan aplikasi-aplikasi ini bergantung kepada kualiti MSA. Walaupun terdapat banyak kaedah yang ada untuk MSA termasuklah kaedah meta-heuristik, ketepatan MSA masih menjadi satu isu yang dibimbangkan. Tambahan pula, mencari satu penjajaran yang optimal adalah sukar di bawah apa sahaja fungsi objektif yang dikira wajar. Sebaliknya, algoritma bat (BA) adalah satu algoritma metaheuristik yang baru sahaja digunakan, yang mana ia berjaya menyelesaikan pelbagai masalah pengoptimaan seperti penskedulan pelbagai pemprosesan, isu padanan imej dan lipatan protin. Kajian ini bertujuan mengkaji kemampuan BA, metod berasaskan populasi dengan ciri-ciri berasaskan carian tempatan, untuk menangani masalah ketepatan penjajaran pelbagai turutan. Generasi populasi awal dalam algoritma pengoptimaan untuk masalah MSA ini ialah salah satu faktor penting yang boleh mempengaruhi kualiti penjajaran. Dengan terlebih dahulu menentukan kedudukan spesifik (titik sauh) untuk menjana penjajaran separa, ini terbukti bernilai untuk menentukan ketepatan MSA dalam beberapa kajian, di mana titik sauh digunakan sebagai panduan untuk membina MSA. Maka itu, kajian ini menyarankan satu metod untuk mengesan titik sauh menggunakan algoritma pengklusteran Jiran Kongsi Berdekatan untuk menjana penjajaran separa. Kemudian,

satu BA asas untuk MSA (BA-MSA) yang berkebolehan menerima titik sauh telah ditampilkan. Selepas itu, satu BA kepada MSA yang telah ditingkatkan (Profil BA-MSA) telah dibangunkan dengan mengubahsuai carian setempat dalam BA. Untuk penambahbaikan seterusnya, satu MSA awal yang telah dipertingkatkan telah dibentangkan dan satu pengoperasi baru juga dimasukkan ke dalam BA (IBA-MSA) dengan menggabungkan satu teknik penjajaran profil dengan pengoperasi lintasan. Metod-metod yang disarankan telah dinilai dan dianalisa dan dibandingkan dengan lain-lain metod MSA yang sering diaplikasi menggunakan penanda aras BaliBase 3.0. Penggunaan titik sauh telah memperbaiki ketepatan metod BA-MSA dan Profil BA-MSA.

ANCHOR POINT APPROACH FOR INITIAL POPULATION OF BAT ALGORITHM FOR PROTEIN MULTIPLE SEQUENCE ALIGNMENT

ABSTRACT

Multiple sequence alignment (MSA) is a fundamental step for many bioinformatics applications such as phylogenetic tree construction, prediction of the secondary structure and identification of domains and conserved motifs. The reliability and accuracy of these applications depend on the quality of MSA. Although there are many approaches available for MSA including meta-heuristic, the accuracy of MSA remains a challenge. In addition, finding an optimal alignment is NP-hard problem under any reasonable objective function. On the other hand, bat algorithm (BA) is a recently used meta-heuristic algorithm, which is efficient in solving various optimization problems such as multiprocessor scheduling, image-matching problem and protein folding. This research aims to investigate the capability of BA, a population-based method with local search-based characteristics, to tackle the accuracy problem of multiple sequence alignment. The generation of initial population in optimization algorithms for MSA problem is one of the important factors that can influence the alignment quality. Determining beforehand specific positions (anchor points) to generate partial alignment has proved valuable for the accuracy of MSA in some research, where the anchor points is used as a guide to build the MSA. Therefore, this research proposes a method to detect the anchor points by using Shared Near Neighbours clustering algorithm to generate partial alignment. Then, a basic BA for MSA (BA-MSA) which has the ability to accept anchor points is presented. Afterward, an enhanced BA for MSA (ProfileBA-MSA)

was developed by modifying the local search in BA. For further improvement, an enhanced initial MSA is presented as well as a new operator is included into BA (IBA-MSA) by combining a profile alignment technique with crossover operator. The proposed methods were evaluated and comparatively analyzed against other commonly applied MSA methods using BaliBase 3.0 benchmark. The inclusion of anchor points has improved the accuracy of the BA-MSA and ProfileBA-MSA methods.

CHAPTER 1
INTRODUCTION

1.1 Overview

Biology is a fundamental aspect of sciences based on its links to medicine and human diseases (Notredame, 2002). Advances in the identification of new molecular targets for drug discovery are derived from the fundamental progress in deciphering biological challenges. In recent times, increasing the amount of biological data amassed has raised the demand to employ computing techniques to handle such type of data. For instance, the advancement in genome research has provided access to a vast volume of biological data. Therefore, the studies carried out on bioinformatics have rapidly evolved by incorporating genomic data and developing several tools and resources to amass and analyse acquired biological data.

Multiple sequence alignment (MSA) is a fundamental technique of molecular biology. The consequences of obtaining this kind of analysis have been significant to help the biologist to infer some information such as phylogenetic tree estimation, protein structure prediction and identification of conserved motifs and domains (J D Thompson *et al.*, 1999; Kemena and Notredame, 2009).

Sequence alignment is a method for arranging the sequences one stacked over the other to show the mutual similarities between the sequences. The sequence of proteins is an order list of set of alphabet symbols S (twenty amino acids) (Zhang *et al.*, 2007; Naznin *et al.*, 2012). The primary sequences of biological data is generally denoted as strings, although implementing the techniques into biological research requires the knowledge of computing to comprehend the basic terms employed in

molecular biology research. In addition, the obvious disparities in handling sequences of biological data should be considered.

1.2 Background

There are several kinds of data in molecular biology. The most rudimentary types of biological data are primary sequences, which are Ribonucleic Acid (RNA), Deoxyribonucleic Acids (DNA), and protein (amino acid) sequences (Lu and Sze, 2009). Proteins are the main elements in all living organisms, thus they play a significant role in the activities of living cells. In the human body, there are thousands of various kinds of proteins. The human cells contain a number of proteins, the largest chemical component of a cell, which in turn contains oxygen, hydrogen, carbon, and nitrogen, and sulphur in certain cases. The proteins play several vital biological functions that include transmitting biological signals, infection attack and enzymatic activity (Thompson *et al.*, 1994).

Proteins are amassed from series of amino acids. There are twenty naturally occurring common of amino acids of different types as presented in Table 1.1, where each amino acid is symbolized by a solitary letter or three letters (Thompson *et al.*, 2001). The Proteins vary depending on the number of amino acids they contain (from a small number of amino acids to several thousands), and the sequence of their amino acids. The length of a protein molecule also varies, where the majority of the proteins sequences range between 30 and 500 residues (AbdulRashid, 2008).

Table 1.1 : Amino Acids Abbreviations (Smith and Waterman, 1981)

Amino Acid Name	Three Letter Code	One Letter Code
Isoleucine	Ile	I
Asparginine	Asn	N
Aspartric acids	Asp	D
Alanine	Ala	A
Valine	Val	V
Glycime	Gly	G
Proline	Pro	P
Serine	Ser	S
Leucine	Leu	L
Arginine	Arg	R
Histine	His	H
Lysine	Lys	K
Threonine	Thr	Y
Phenylalanine	Phe	F
Tyrosine	Tyr	Y
Glutamic acid	Glu	E
Methione	Met	M
Tryptophan	Trp	W
Glutamine	Gln	Q
Cysteine	Cys	C

The twenty amino acids can be categorized according to their physiochemical properties. In each group, the members possess relatively similar physiochemical properties, which include hydrophobic (A, P, G, I, L, C, V, W, M, F), acidic (E, D), hydroxylic (T, S), basic (K, H, R), charged (H, K, R, D, E) and aromatic (Y, H, W, F). It apparent that some amino acids have a dual nature, for instance, the amino acid 'W' possess both aromatic and hydrophobic physiochemical properties (Lipman *et al.*, 1989; Wang and Jiang, 1994).

In the study of biochemistry, the proteins have been classified into four different structural levels comprising primary protein structure, secondary structure, tertiary protein structure and quaternary structure. This study focuses only on primary

structure of the proteins, working directly with the primary sequences of amino acids. The primary structure comprises linear chain of amino acids linked by peptide bonds. Each amino acid holds two components: a backbone or main chain and a side chain. The backbones of all amino acids types are similar while the side chain is specific for each different type of amino acid, where it determines the chemical and physical properties of the amino acid. Therefore, the twenty different kinds of amino acids are attached with twenty different types of side chains (Feng and Doolittle, 1987).

1.3 Multiple Sequence Alignment

Sequence alignment is a method of comparatively analysing two or more sequences of protein with each other to establish a consistency (match) and disparity (mismatch) between their amino acid residues. The comparison is carried out by searching for pattern of characters or sequences of individual characters of similar order in the specified protein sequences.

Given two sequences $x = x_1, x_2, x_n$ and $y = y_1, y_2, \ldots y_m$, the sequence alignment can be generated by assigning matches (identical) or mismatches (similar characters) in a column. Spacing positions are introduced to increase the alignment score, where the maximum score can initiate enhanced alignment quality (D G Higgins and Sharp, 1988). One more reason for inserting spaces is that the produced aligned sequences have to be in similar in length (Thompson *et al.*, 1994). A dash character '-' is generally used in the sequences to specify the space position, referred to as a gap. There are three cases of mutations in sequence alignment: insertions, deletions and point mutations. Therefore, the gap in any position is observed as a deletion in one sequence and an insertion in another. In cases where the match and mismatch

columns make a prediction regarding point mutations, the columns that comprise only gaps have no meaning, therefore, they are excluded from appearing in the final MSA. The aligned sequences are ultimately stacked over each other as shown on Figure 1.1

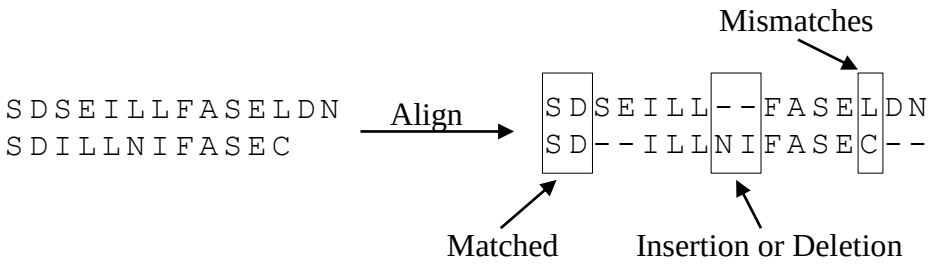

Figure 1.1 : Alignment of Two Sequences

Sequence alignment can be categorized into two types based on the number of input sequences, and they include pairwise sequence alignment and multiple sequence alignment (MSA). When the input sequences are equal to two, it refers to a pairwise sequence alignment. The pairwise alignment methods are most frequently applied in database search program such as Fasta (Pearson and Lipman, 1988) and Blast (Altschul *et al.*, 1990). In multiple sequence alignment, the amount of input sequences exceeds two (Barton and Sternberg, 1987). There are several applications developed based on multiple sequence alignments such as function prediction (Sjolander, 2004), phylogenetic tree estimation (Taheri and Zomaya, 2010) and protein structure prediction (Salamov and Solovyev, 1997). The precision reliability and consistency of these methods depends on the quality of the multiple alignments.

There are two concepts that govern alignment methods; local alignment and global alignment. In local alignment method, only the most similar parts and conserved blocks of the sequences are aligned. Conversely, in global sequence alignment method, all sequences that maintain a correspondence over their entire

length are aligned to detect the best alignment, i.e. an attempt is made to align every single amino acid in every sequence (Saitou and Nei, 1987). An example of local alignment and global alignment is shown in Figure 1.2.

Figure 1.2 : Local and Global Alignment

1.4 Motivation and Problem Statement

Multiple sequence alignment (MSA) is a fundamental step for virtually all facets of biological sequence analysis (Notredame *et al.*, 2000). There are several applications in bioinformatics that rely on MSA tools, for example, phylogenetic tree construction (Edgar *et al.*, 2004). A phylogenetic tree is a technique that classifies the relationships between homologous genes represented in the genomes (Katoh *et al.*, 2002). Identification of domains and conserved motifs is one more application that depends on MSA (Do *et al.*, 2005). Furthermore, MSA methods play an important part in structure prediction of the secondary and tertiary structure of proteins (Do *et al.*, 2005; Pei and Grishin, 2006). MSA plays a vital role in drug design such as the development of flu vaccines (Kim and Ma, 2014).

The quality of the results from MSA is a key factor in analyses of biological applications based on MSA. Given that the amount of biological data actually collected by molecular biologists tends to be extremely enormous, it is impractical to handle the sequence alignment manually to acquire an accurate MSA. Therefore,

automated methods are required to compute the alignments within a logical time frame.

In fact, finding an accurate alignment from primary protein sequences is considered to be computationally an NP-hard problem (Do *et al.*, 2005; Zhang *et al.*, 2007). Even though there are several approaches that have been proposed and developed to resolve issues associated with multiple sequence alignment (MSA), there still exist the major problems of low accuracy (Lalwani *et al.*, 2015), which is the key issue of sequence alignment and the high time complexity, where the time increases exponentially by increasing the number of input sequences and length of each sequence. For instance, in the case of two input sequences with equal lengths, the number of possible alignments that can be calculated is equal to 1683 for sequence length of 5, while the number of probable alignments that can be calculated for sequence length of 10 is 8097453 (Arenas-Díaz *et al.*, 2009). It is evident that the number of possible sequence alignments grows exponentially with increase in size of the sequences.

In addition, the majority of frequently applied multiple sequence alignment methods are based on a progressive approach (Abu-hashem *et al.*, 2015). Nonetheless, the major drawback that affects these methods is that the order profiles selected during the alignment process considerably affect the quality of the final alignment. Furthermore, the guide tree base in progressive approaches does not essentially explain the relationships between the input sequences, possibly resulting in the loss of very vital information regarding the other input sequences and their relationship, that can be very valuable in the whole alignment process (Henikoff and Henikoff, 1992; Modzelewski and Dojer, 2013). The final alignment acquired via the

progressive approach may differ when all possible guide trees are considered (Zhang *et al.*, 2007).

On the other hand, a consistency-based approach has been used to overcome the limitation of progressive approach. In consistency-based approach, the idea is that the final multiple alignment is the one that agrees the most with all the possible optimal pairwise alignments (Kemena and Notredame, 2009). However, the consistency-based methods still exhibit difficulty acquiring high quality alignments when sequences are highly variable, where conserved regions in the sequences may occur within long unaligned regions (Wang *et al.*, 2007). Several variations of methods have been developed for MSA. From those, MSAProbs (Liu *et al.*, 2010) which has been demonstrated to be one of the most accurate MSA methods in some recent literature (Modzelewski and Dojer, 2013; Abu-hashem *et al.*, 2015). MSAProbs is a tree-based progressive alignment tool based on the pair-Hidden Markov model.

It has been shown that MSAProbs could not obtain high-quality alignments for distantly related sequences (Zhan *et al.*, 2015), where neither progressive nor consistency-based approaches build optimal alignments when sequences are distantly related (Valenzuela *et al.*, 2013). The related studies reported in literature (Notredame, 2002; Duc, 2012; Valenzuela *et al.*, 2013; Daugelaite et al., 2013; Pramanik and Setua, 2014) indicate there is no generic approach available for solving MSA problem perfectly. Thus, there remains the need to develop an approach that improves the accuracy, and scalability. However, MSA can be formulated as an optimization problem with a logical objective function (Zhang *et al.*, 2007), thus the tree base problem in progressive approach can be disregarded. In contrast, the

alignment quality by using optimization approach for MSA depends on many factors such as the features of the optimization approach, the ability of optimization algorithm to balance between intensification and diversification, the quality of initial population, and the objective function (Mohsen, 2014).

Several optimization approaches have been applied to solve difficult problems such as local search-based and population-based approaches. In local search-based approach, the basic idea is to start from an initial solution and to search for successive improvements by examining neighbouring solutions (Qi, 2011). Thus, this approach is efficient in finding a precise local optimal solution (Mohsen, 2014). In contrast, population-based approach has a good ability to explore entire search space (Boussaïd *et al.*, 2013). However, this approach is not efficient to find a precise local optimal solutions in the regions to which the method converges (Mohsen, 2014). The best way to tackle the MSA problem is take the advantage of population-based and local search-based approaches by balancing the global exploration and local exploitation (Mohsen, 2014).

Recently, a bat algorithm (BA) as a new swarm intelligence algorithm is proposed (Yang, 2010). BA is based on the echolocation features of bats. It considered being a population-based method with local search-based characteristics (Fister, 2013). In BA, a frequency-tuning technique is used in order to increase the diversity of the solutions, while, it uses the automatic zooming to balance between global exploration and local exploitation during the search process by simulating the variations of pulse emission rates and loudness of bats when searching for prey (Yang and He, 2013). Furthermore, BA uses parameter control as the iterations proceed. This provides a way to automatically switch from exploration to

exploitation when the optimal solution is approaching (Fister, 2013). Additionally, the structure of the BA is relatively easier as it has few parameters.

BA has shown promising efficiency to solve numerous optimization problems with different applications such as; multilevel image thresholding (Alihodzic and Tuba, 2014a), classifications (Mishra *et al.*, 2012), image-matching problem (Zhang and Wang, 2012), and multiprocessor scheduling (Eliseo *et al.*, 2015). Furthermore, BA has been adapted and being successfully applied to optimization problems in bioinformatics field such as protein folding (Xingjuan *et al.*, 2014) and predict the protein-protein interaction (Chowdhury *et al.*, 2014). Thus, BA might possess the potential to be used for MSA problem. The multiple sequence alignment $P=(S, f)$ can be formulated as an optimization problem that contains finite set of sequence alignments S and objective function $f(x)$. The objective function (also called score function) assigns a cost value to MSA. Therefore, the target is to obtain the maximum/minimum score value in a reasonable amount of time (Naznin et al., 2011).

The way of generating the initial population in optimization algorithm for MSA problem is one of the important factors that can influence the alignment quality, where good quality initial population can be effectively converged faster (Zhang et al., 2007; Mohsen, 2014). The initial population for MSA problem can be generated using three strategies; (1) Generate the initial MSA randomly (Lee et al., 2008). (2) By using an existing MSA tool such as MUSCLE (Edgar, 2004). (3) Constructing the initial MSA by using the information from pairwise alignments (primary protein sequences), or any other resources such as secondary structure (Zhang et al., 2007; Deng and Cheng, 2011).

In the first strategy, specific numbers of gaps are inserted at random positions into given sequences. However, it has been reported in (Mohsen, 2014) that the random initial of MSA does not guarantee to produce a solution with reasonable quality, where the candidate solutions are far from either optimal or near-optimal solution. Furthermore, the search space of MSA is huge (Rodriguez et al., 2007). Thus, the search space increases exponentially as the number of sequences and the length of each sequence increase (Shyu et al., 2004). In the second strategy, the shortcoming that the initial MSA totally depends on other tools, which may be order-dependent such as ClustalW method (Thompson et al., 1994). Moreover, the initial solutions which are produced by this strategy can be similar to each other. In the third strategy, the main advantage that the produced alignment can be fairly accurate compared to the alignment that produced by random generation strategy. In this strategy, the information can be acquired from experts (Morgenstern *et al.*, 2006), secondary structure predictions (Subramanian *et al.*, 2010; Deng and Cheng, 2011) or from primary protein sequences (Zhang *et al.*, 2007). In this research, the information from the primary protein sequences only can be used based on the scope of this research.

From the primary protein sequences, determining beforehand some specific positions to be aligned has proved valuable for the accuracy MSA (Pitschi *et al.*, 2010; Fostier *et al.*, 2011). These specific positions (anchor points) can be detected automatically from the input sequences using some developed algorithm. The detected anchor points can appear as partial alignment, and then this partial alignment is used as a guide to build the MSA (Pitschi *et al.*, 2010). Min-cut algorithm have been used to construct anchor points (Corel *et al.*, 2010), however this approach is limited by the fact that incidence graphs become too large by

increasing the number of sequences. Moreover, Pitschi *et al.* (2010) have proposed an automatic detection of anchor points method. The anchor points in this approach are obtained directly from fragments (segments) for all input sequences. According to Rausch *et al.* (2008), segment based approach is able to identify homologies shared by two of the aligned of portion sequences to improve the alignment accuracy. However, in the segment approach methods, the search for consistent segment may become extremely time consuming (Subramanian, 2009). Moreover, there is no guarantee that the best possible set of consistent segment can be constructed. To avoid this limitation, the anchor points can be detected first from the input sequences based on the residues, and then the segments can be discovered based on the detected anchor points. To achieve that, clustering algorithm such as The Shared Near Neighbours (SNN) method (Jarvis and Patrick, 1973; Ert *et al.*, 2003) can be used.

The Shared Near Neighbours (SNN) method a common density-based clustering methods, has been used for several applications such as mobile computing (Kuenning and Popek, 1997), linguistic purposes (Ertöz *et al.*, 2004) and identification of a diverse (Santos and Silva, 2012). Its versatility is based on the fact it can identify the core points of varying density and thus, can handle data that hold clusters of different densities. In addition, the number of clusters is not requisite beforehand for this approach unlike other clustering algorithms such as K-means (Taheri and Zomaya, 2010). Besides, the SNN is not sensitive to the order of the data such as BIRCH algorithm (Zhang, 1997). The most important feature of SNN is that, the similarity measure between any two points is based on the number of neighbors that two points share. This concept can be useful to find strong similarity amino acids from the input sequences to find the anchor points. Hence, the partial alignment can

be intuitively explored by adapting SNN algorithm to detect anchor points in order to construct partial alignments.

Therefore, this study determines whether the partial alignment which produced by detecting of anchor points using SNN algorithm can be useful to improve the MSA. In this stage, the generated anchor points will be included to DIALIGN-TX method to produce a complete MSA from the detected anchor points, where this method has the ability to accept anchor points from the other tools, and it can be forced to align a list of amino acids that selected in advance, where only DIALIGN has such an explicit option. Moreover, this study determines whether the detected anchor points can be useful to generate the initial MSA for bat algorithm to solve the MSA problem. Furthermore, this research the will study whether the bat algorithm can be further improved by including a new operator by combining profile alignment technique with crossover as a refinement step.

1.5 Research Objectives

The objectives of this thesis are summarized as the following:

- To propose an automatic anchor point detection method for MSA based on Shared Near Neighbours (SNN) algorithm with segment-based alignment approach.
- To propose a bat algorithm with anchor point technique to resolve MSA problem.
- To enhance the bat algorithm by including a new operator which combines profile alignment technique with crossover.

1.6 Research Contributions

The research contributions can be summarized as the following:

- An automatic detection method of anchor points is proposed by adapting Shared Near Neighbours clustering algorithm (AP-SNN) to create partial alignment for MSA.
- A modified detection anchor point method (SNN-SB) is proposed to improve the accuracy of MSA by modifying the similarity measurement and by combining the SNN with segment-based alignment approach.
- Adapting the bat algorithm to solve multiple sequence alignment problem by including anchor points to improve the quality MSA.
- The quality of MSA is improved by using enhanced initial population for bat algorithm and by including a new operator as well.

1.7 Scope and Limitations

This research is restricted to developing an improved alignment by incorporating anchor points to resolve MSA problem and by enhancing initial alignment and adapting the bat algorithm. The MSA problem in this research is analyzed as an optimization problem based on a fitness function. Thus, the studying of prediction of structure and protein folding extends beyond the scope of this research. In addition, DNA and RNA are not considered in this thesis given that only the primary structure of protein sequences is examined.

1.8 Thesis Outline and Organization

This thesis organized into five chapters. This chapter introduces an overview of the research field, followed by background of biological data, motivation and problem statement, objectives, contributions, and scope. The remainder of this thesis is organized as the following:

Chapter 2 provides a broad review and analysis of existing MSA methods. After that, present a brief review of the available benchmark datasets for biological sequences and the benchmark datasets for MSA.

Chapter 3 provides information on the research methodology. It contains the general proposed framework of this study and the experimental design.

Chapter 4 presents the automatic detection of anchor points based on SNN algorithm for MSA problem, namely (AP-SNN). After that, the improvements of AP-SNN is discussed (they are SNN-WS and SNN-SB). Then, the performances of all proposed methods on benchmark datasets are discussed and statistically analyzed. Finally, the proposed methods are compared with other methods.

Chapter 5 presents the initial population is built based on anchor points technique which has been constructed by the methods that proposed in Chapter 4. Two proposed methods are presented in this chapter to solve the MSA using BA based on anchor point technique. The result and discussion of the proposed methods are described in this chapter.

Chapter 6 presents enhanced initial alignment. Thereafter, an improved bat algorithm for MSA is presented by included new operator by combing profile

alignment with crossover operator. The performance of proposed methods compared to the other common methods is also discussed in this chapter.

Chapter 7 concludes remarks with suggestions of possible future studies directions.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

This chapter describes various types of currently applied multiple sequence alignment methods. MSA is a broad subject, thus several methods have been proposed to obtain optimal alignment. To understand the MSA methods, basic definitions of sequence analysis, measurement scheme, substitution matrix, gap penalty, database benchmarks and existing approaches of MSA are reviewed in this chapter.

2.2 Formal Definitions for MSA:

To improve the understanding of the process of sequence alignment, there are specific formal definitions considered in this thesis as the following:

Definition 1: An alphabet *A*: a set of finite of symbols. In this work, the alphabet corresponds to the set of amino acids because the protein sequences only considered. The twenty different amino acids in protein are as follows: $A_{Protein}$ ={*A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}*. Thus, by using the twenty different alphabets, the sequences are constructed in which a sequence and sequence alignments are defined as follows:

Definition 2: a group of protein sequences, $S = (s_1, s_2, \ldots s_i)$, each sequence s_i is a series of finite ordered characters from a specific alphabet, that is, $s_i \in A$. The length of the sequence *n*, is denoted by $|s_i|$. For multiple sequence alignment, the number of sequences is greater than 2, i.e. $i > 2$.

Definition 3: given a protein sequence s_i, a sub-sequence, s[p, q], from the protein sequence s_i is a sequence of finite ordered characters from the p^{th} position to the q^{th} position.

Definition 4: a multiple sequence alignment (MSA) of $S = (s_1, s_2, \ldots s_i)$ is a rectangular matrix consisting of i rows of symbols of $A \cup \{-\}$.

There are certain properties that must be satisfied for alignment of sequences to be considered a valid alignment. In a case where all aligned sequences s_i must be of equal length, any column comprising only gap symbols must be removed from the sequences in the final alignment (Wang, 2007).

An example of MSA: Given a family $S=(S_1,\ldots,S_n)$ of $N=4$ unaligned sequences (input) as shown in Figure 2.1:

```
S1: HARFLEGVTHELASCFATMMG
S2: HARFLEGVTHEFASCMMG
S3: HARFLEGVTHEVERYFASCMMG
S4: THEFACMMG
```

Figure 2.1: Unaligned Sequences

The MSA is a process of determining residues in homologous sequences. Thus, the MSA of S is a new set of sequences $S'=S'(S'_1,\ldots,S'_N)$ such that all the strings is S' are equal length and each S'_i is generated from S_i by inserted gaps (denoted by "-"). The output of MSA (aligned Sequences) is shown in Figure 2.2:

```
S'1: HARFLEGVTHE----LASCFATMMG
S′2: HARFLEGVTHE----FASC---MMG
S′3: HARFLEGVTHEVERYFASC---MMG
S′4: --------THE----FA-C---MMG
```

Figure 2.2: An Example of Multiple Alignment Sequences

It can be denoted that, any column with all gaps can be removed from the aligned sequences without losing biological relevance. The mismatches that shared between any two sequences can be interpreted as substitutions (point mutations). In the Figure 2.2, the amino acid "F" has been substituted by "L" in S'_1. In MSA, when one or more amino acids are removed the mutation procedure is known as deletion. In the example, the subsequence "HARFLEGV" is removed from S'_4, this known as deletion. Finally, an insertion takes place when one or more amino acids are inserted. In the Figure 2.2 the subsequence "FAT" is inserted in S'_3.

2.3 Sequence Alignment Measure

Sequence alignment can be defined as an approach to find maximal sets of similar parts of many sequences to show the match and mismatch between their amino acid residues. Therefore, to determine how the sequences are alike, certain kinds of scoring measures need to be applied, where the maximum/minimum score can be applied to obtain better alignment quality (Pramanik and Setua, 2014). Therefore, specific measure of the score can be fixed for each probability of events, where the events in alignment sequences can be matched, mismatches, insertions and deletions. In practical terms, the score method consists of two parts: substitution matrices and gap penalties (Gondro and Kinghorn, 2007). The substitutions matrices provide a numerical value for each pair of residues, which can be whether matches, or mismatches. On the other hand, the gap penalties provide a numerical

quantification of each set of residue-space in aligning sequences, which indicate deletions and insertions.

2.3.1 Substitutions Matrix

A substitution matrix is a table of numbers with dimensions 20 x 20 for proteins and 4 x 4 for RNA and DNA (David, 2001). The substitution matrices provide a score of the probability of occurrence of conservation or substitutions (Gondro and Kinghorn, 2007). There are two types of scoring matrices conventionally used to quantify the sequence similarity: identity matrix and alphabet-weight matrix. The identity matrix is a sample matrix with fixed scores of matches and mismatches that is typically used to score DNA/RNA sequences. In identity matrix, the score between any pair of matches can be 1 while the score between any pair of mismatches can be 0. The alphabet-weight matrix is more complex scheme, where the effect of structure and functions of proteins are taken into consideration. Thus, the alphabet-weight matrices or amino acid substitution matrices are used to reflect the substitution that amasses between any two amino acid characters in the sequence alignment. The scoring between any pair of residues is symmetrical in nature, where the score of residue A substituting residue B is the same as residue B substituting residue A. For protein sequences, there are three common types of matrices, Blocks Substitution Matrix (BLOSUM) (Henikoff and Henikoff, 1992), Point Accepted Mutation (PAM) (Dayhoff *et al.*, 1978) and Gonnet matrix (Gonnet *et al.*, 1992).

In BLOSUM matrix, the scores are calculated based on frequencies of amino acid substitution which are deduced from a large number of related proteins. This approach is considered to be more appropriate for similarity searches in databases (AbdulRashid, 2008). There are several types of BLOSUM matrices with different

percentage of repetitions, ranging from 1 to 100. As an example, BLOSUM62 has been extracted from protein blocks for two sequences exceeding 62% identical matrices. PAM Substitution Matrix was developed by Margaret Dayhoff (1978). The scores in PAM are calculated according to a mutation model that predicts the types of amino acid changes over a long period. The Gonnet matrix has been developed via an extensive pairwise alignment on protein sequence databases.

2.3.2 Gap Penalty

Gap is defined as a consecutive number of spaces inserted into a single protein sequence alignment (Gusfield, 1997). The gap has a considerable influence over the distribution of spaces in an alignment to produce the best possible alignment, thus it is very important to define a gap penalty function. In biology, there are numerous mutation events that accrue by deletion or insertion of an entire substring in a protein sequence. This deletion or insertion causes gaps in an alignment (Wang, 2007).

A penalty or native score is assigned to a set of gaps to reflect the number of amino acids being inserted or deleted. There are three common types of gap models; a linear gap or constant gap penalty model; affine gap penalty model, and convex gap penalty model. In the linear gap or constant gap model, the penalty of gap is always fixed wherever it is placed in the string, thus there is no disparity between the opening penalty and extending penalty of the gap. The penalty in this model can be calculated using the equation: $gap = g_0 \times n$ where g_0 is the gap-opening penalty and n is the number of consecutive gaps. The affine gap penalty model is the most extensively used (Wang, 2007). In this model, the penalty assigned for opening gap is higher than the extension for the addition gap. It motivates the addition of the extension of gaps instead of the opening of new gaps. In this model, the gap can be

calculated using the equation: $gap = g_0 + (n - 1) \times g_e$ where g_0 is the gap opening penalty and g_e is the extension penalty of gap where $|g_0| > |g_e|$. The concept of the convex gap penalty model is that each additional gap will contribute less to the gap score than the previous gap. The affine gap penalty model is more advantageous compared to the constant model because of its relatively higher efficiency (Roy, 2014).

2.4 Multiple Sequence Alignment Approaches

Several methods have been developed to resolve the MSA problem. Comprehensive reviews of these approaches have been conducted in a number of studies (Notredame, 2002; Edgar and Batzoglou, 2006; Kemena and Notredame, 2009; Daugelaite *et al.*, 2013). MSA methods can be divided into different categories (Notredame, 2002); progressive, iterative algorithm, consistency based and segment based. However, numerous MSA methods include characteristics of several categories. In the underlying sub-sections, the characteristics of some MSA methods are concisely explained.

2.4.1 Progressive Approach

Most of the existing MSA methods are developed around the progressive alignment approach (Kemena and Notredame, 2009). Progressive methods are effective in resolving linear complexity, and they require only a small amount of memory (Taylor, 1988). The progressive approach was introduced by Hogeweg and Hesper (1984). Afterward, it was modified by DF Feng and Doolittle (1987) and Taylor (1988a). The general idea governing these conventional progressive alignments is that they begin by constructing the initial alignment by aligning the two

most closely related sequences, and then a new sequence is incorporated in order of increasing distance. The order for selecting the new sequences to be aligned firstly is determined using a previously obtained guide tree. The guide tree is conventionally constructed using a dynamic programming, such as the Needleman-Wunsch method (Needleman *et al.*, 1970), which has been applied by Feng and Doolittle (1987).

Several methods have been proposed based on the progressive approach. ClustalW method (Thompson *et al.*, 1994) is the most common progressive method, which is considered as the standard method for the this approach (Notredame, 2002). The ClustalW alignment process consists of three main stages to produce the final MSA (see Figure 2.3). The first stage is to construct a distance or similarity matrix. The distance matrix is a *n x n* matrix, where *n* is the number of the sequences to be aligned. It represents the pairwise distances between the input sequences in sequence space. The pairwise distances between the input sequences are calculated by doing all-against-all pairwise alignments between all input sequences and each input sequence is assigned a weight during the alignment process. In addition, different amino acid substitution matrices are used according to the divergence of the input sequences to be aligned. The second stage is to build a guide tree. The guide tree is constructed based on the relationships between the sequences defined by the distance matrix is produced. Neighbour-joining method (Saitou and Nei, 1987) is used to build the guide tree using the distance matrix. The last stage entails sequentially aligning input sequences to the growth of MSA based on the sequence order which is defined by the guide tree.

The advantages of the ClustalW alignment process include the ability to select the most appropriate substitution matrix and its effective performance. The ClustalX

alignment developed as a replacement of the command line program in ClustalW (Thompson, 1997) provides an easier graphical user interface.

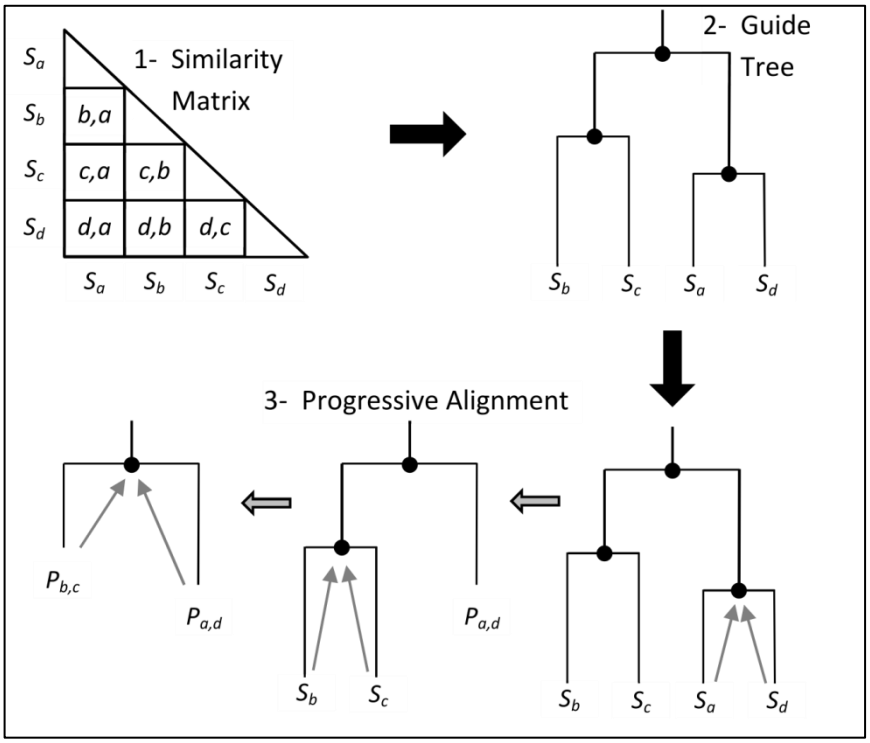

Figure 2.3 : Progressive Alignment (Cortada, 2013)

Other alignment algorithms based on the progressive alignment method have been proposed, such as MULTAL (Taylor, 1987), MULTALIGN (Corpet, 1988), PILEUP (Devereux *et al.*, 1984), Kalign (Lassmann and Sonnhammer, 2005) MUSCLE (Edgar, 2004). MULTAL uses a sequential branching algorithm to align the two closest sequences, and subsequent next closest sequences, and so on. PILEUP and MULTALIGN develop the final alignments from a guide tree, which is constructed using the Unweighted Pair Group Method Using Arithmetic Averages (UPGMA) (Sneath, Peter HA, 1973). Kalign method follows the progressive approach, but it uses the Wu-Manber string matching method (Wu and Manber, 1992) to find the initial scores of distance, which is faster than pairwise alignment. In