Republic of Iraq

Ministry of Higher Education and Scientific Research

University of Diyala

College of Science

Firefly Optimization Versus Genetic Algorithm For Inferring Well Supported Phylogenetic Trees

A Thesis

Submitted to the Computer Science Department \College of Science \University of Diyala
In a Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer

By
Omer Emad Ensaif

Supervised By

Prof.Dr.Taha Mohammed Hassan

Assist.Prof.Dr.Bashar Talib Al-Nuaimi

2021 A.D. 1443 A.H.

يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنْكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ وَاللَّهُ بِمَا تَعْمَلُونَ خَبِيرٌ (11)

صدق الله العظيم

سورة المجادلة (11)

ACKNOWLEDGMENTS

*First of all, Praise is to God, Lord of the worlds, for a blessing that helped me in achieving this research until the end of it, I would like to express my thanks and gratitude to my supervisor "Prof.***Dr.TahaMohammed Hassan andAsist.Prof. Dr.Bashar Talib Al-Nuaimi***" for supervising this research and for the bounty, patience and continued guidance throughout the work.*

My thanks to all academics and administrative staff at the Department of computer science.

Dedication.

I would like to dedicate this work to:

My father, my mother and all my family

Supervisor's Certification

I certify that this thesis entitled" **Firefly Optimization Versus Genetic Algorithm For Inferring Well Supported Phylogenetic Trees"** was prepared by **"Omer Emad Ensaif"** at the University of Diyala / College of Science /Department of Computer Science, has been evaluated scientifically; therefore, it is suitable for debate by the examining committee.

Signature:

Name: **prof Dr.Taha M. Hassan**

Date: / /2021

Signature:

Name: **Assist prof Dr. Bashar T. Al-Nuaimi**

Date: / /2021

Approved by the University of Diyala Faculty of Science Department of Computer Science.

Signature:

Name: **Assist Prof Dr. Bashar T. Al-Nuaimi**

Date: / /2021

(Head of Computer Science Department)

Scientific Certification

I certify that this thesis entitled "**Firefly Optimization Versus Genetic Algorithm For Inferring Well Supported Phylogenetic Trees"** was prepared by **"Omer Emad Ensaif"** at the University of Diyala / College of Science /Department of Computer Science, has been evaluated scientifically; therefore, it is suitable for debate by the examining committee.

Signature:

Name : **Assist Prof Dr .Mohammed S. Hamoud**

Date : / / 2021

Scientific Certification

I certify that this thesis entitled "**Firefly Optimization Versus Genetic Algorithm For Inferring Well Supported Phylogenetic Trees"** was prepared by **"Omer Emad Ensaif"** at the University of Diyala / College of Science /Department of Computer Science, has been evaluated scientifically; therefore, it is suitable for debate by the examining committee.

Signature:

Name : **Assists. Prof.Dr. Huda Abdulaali Abdulbaqi**

Date : / / 2021

Linguistic Certification

I certify that this thesis entitled ***"Firefly* Optimization Versus Genetic Algorithm For Inferring Well Supported Phylogenetic Trees"** was prepared by ***"Omer* Emad Ensaif"** at the University of Diyala / College of Science /Department of Computer Science, is reviewed linguistically. Its language was amended to meet the style of the English language

Signature:

Name : **Assists. Prof.Dr.Fatima M. Aboud**

Date : / / 2021

Examination Committee Certification

We certify that we have read the thesis entitled "**Firefly Optimization Versus Genetic Algorithm For Inferring Well Supported Phylogenetic Trees**" and an examination committee, examined the student "**Omer Emad Ensaif**" in the thesis content and that in our opinion, it is adequate as fulfill the requirement for the Degree of Master in Computer Science at the Computer Science Department, University of Diyala.

(Chairman)
Signature:
Name: **Prof.Dr. Dhahir A. Abdulah Salman**
Date: / / **2021**

Signature:
Name: **Assists. Prof.Dr. Ahmed Sabah Ahmed** **(Member)**
Date: / / **2021**

Signature:
Name: **Assists. Prof.Dr. Hassan Hadi Saleh** **(Member)**
Date: / / **2021**

Signature:
Name: **Prof Dr.Taha M. Hassan** **(Supervisor)**
Date: / / 2021

Signature:
Name: **Assist Prof Dr. Bashar T. Al-Nuaimi** **(Supervisor)**
Date: / / **2021**

Approved by the **Dean** of College of Science, University of Diyala

(The Dean)
Signature:
Name: **Prof. Dr. Tahseen H. Mubarak**

Date: / / **2021**

Abstract

The purpose of phylogenetic analysis research is to study the modifications that occur in various species throughout the process of evolution finding the connections between sequences of genomic DNA and discovering the ancestors' sequences and descendants' sequences. In disciplines like bioinformatics, phylogenetic trees are important for a systematic and comparative phylogenetic analysis.

Molecular data like DNA and protein sequences are used to construct phylogenetic trees. Nodes, or taxonomic units, in a phylogenetic tree represent genomic sequences. Phylogenetic tree construction is a complex yet important problem in the field of bioinformatics. Once constructed, a phylogenetic or evolutionary tree can lend insight into the evolution of different species. The issue is that for a large number of species the problem grows to a computational complexity that is not easily solved.

This thesis proposed a system to building phylogenetic trees faster and robustness to finding optimal solutions based on firefly optimization algorithm. Evaluation performance using bootstrap measurements and finally make the comparison between the genetic algorithm and the firefly optimization algorithm.

The implementation and results of each stage in the proposed system demonstrated that the proposed system has the ability to fast build an Inferring Well Phylogenetic Tree using firefly optimization algorithm and evaluate the performance of the proposed system based on the bootstrap scale which measures the extent of match between family trees and that the best value obtain in the proposed system =0.919013.

List of contents

CHAPTER One *1*

1.1 Introduction *1*

1.2 Related works *4*

1.3 Problem Statement *5*

1.4 Aims of Thesis *6*

1.5 Contribution *7*

1.6 Layout of thesis *7*

CHAPTER Two 8

2.1 Introduction *8*

2.2 Chromosomes and Genomes *8*

2.3 Sequence Alignment *9*

2.3.1 BLAST 10

2.3.2 LOCAL SEQUENCE ALI GNMENT : SMITH–WATERMAN ALGORITHM 11

2.3.3 GLOBAL SEQUENCE ALIGNMENT: THE NEEDLEMAN WUNSCH EXAMPLE 12

2.3.4 MULTIPLE SEQUENCE ALIGNMENT (MSA) 13

2.4 Boyer Moor Algorithm *14*

2.5 Firefly Optimization Algorithm *15*

2.6 *Genetic Algorithm (GA)* *18*

2.6.1 *Selection (Encoding of a Chromosome)* 19

2.6.2 *Crossover* 20

2.6.3 *Mutation* 20

2.7 Phylogenetic Tree Constructions *23*

2.7.1 Distance-based methods *24*

2.7.2 *Character-based methods* *24*

2.7.3 *Neighbor joining algorithm* 25

2.7.4 4 Unweighted Pair Group Method (UPGM) Tree 27
2.8 BootStrap *29*
2.8.1 *Non-Parametric Bootstrapping* 29
2.8.2 Parametric Bootstrapping 31
CHAPTER Three 32
3.1 Introduction *32*
3.2 The Objective of the Proposed System *32*
3.3 Architecture of the Proposed System *32*
3.3.1 *Input Data Set Genomes Stage* *34*
3.3.2 *Evaluation Data set Genomes Stag* *35*
3.3.3 *Select Genome Using Firefly Optimization Algorithm* *41*
3.3.4 Building Phylogenetic Tree stage *42*
4.3.2.1 Neighbor joining (NJ) Tree *42*
4.3.2.1 Unweighted Pair Group Method (UPGM) Tree *49*
CHAPTER Four 54
4.1 Introduction *54*
4.2 Initialization *54*
4.3 Implementation of the proposed system *55*
4.4 Results of The Proposed System *58*
4.4.1 *Load DNA Dataset* *58*
4.4.2 *Results evaluation Genome and Select Stages* *61*
4.4.3 *Results of Building Phylogenetic Tree* *73*
4.4.4 Results of Boot Strap Measure 78
4.5 Comparison between GA and FA *80*

CHAPTER Five.....*83*
5.1 Introduction.....*83*
5.2 Conclusions.....*83*
5.3 Future works.....*84*
References.....*85*

List of Tables

Table 3.1	A plant Genomes DNA dataset	34
Table 3.2	Case One.	38
Table 3.3	Case Two	39
Table 3.4	Case Three	39
Table 3.5	Strong Suffix Rule	40
Table 4.1	Description of Plant Genome of the Dataset.	59
Table 4.2	Results of Firefly algorithm with iteration =100.	61
Table 4.3	Fitness Function	63
Table 4.4	The Best Location.	68
Table 4.5	Performance Evaluation of the NJ and UPGM based on Bootstrap Measure.	78

List of Figures

Figure 2.1	Figure 2.1 : Global alignments and local alignment	10
Figure 2.2	Multiple sequence alignment of various sequences of Apiales order **.**	13
Figure 2.3	General Block Diagram of the Firefly Algorithm	18
Figure 2.4	Chromosomes	19
Figure 2.5	Crossover of Chromosome	20
Figure 2.6	Mutation of Offspring **.**	21
Figure 2.7	Flow chart of genetic algorithm	22
Figure 2.8	**(a) unrooted tree, (b) rooted tree**	24
Figure 3.1	General Block Diagram of the Proposed System.	33
Figure 3.2	distance matrix	43
Figure 3.3	Modify Distance Matrix.	44
Figure 3.4	Tree Topology.	45
Figure 3.5	Tree Topology After Change Value.	45
Figure 3.6	Tree Topology with U1 value	47
Figure 3.7	distance matrix based on calculate net divergence	47
Figure 3.8	Create a matrix MIj.	48
Figure 3.9	NJ final Tree.	49
Figure 3.10	Tree Consisting of 6 OTUs	50
Figure 3.11	Distance Matrix	50
Figure 3.12	Construct a Subtree .	51
Figure 3.13	Results of second Cycle of building UPGM Tree	51
Figure 3.14	Results of Third Cycle of building UPGM Tree	52
Figure 3.15	Results of Fourth Cycle of building UPGM Tree	52
Figure 3.16	UPGMA Tree .	53
Figure 4.1	The interface of load DNA plant dataset.	55
Figure 4.2	The Interface of Implementation of the Firefly Optimization Algorithm to Select Genome for Each Plant Family.	56
Figure 4.3	Interface of implementation of Distance Matrix Technique.	57
Figure 4.4	The Implementation Interfaces for (NJ) Tree.	57
Figure 4.5	The Implementation Interfaces for (UPGM) Tree.	58

Figure 4.6	Distance Matrix Results of Three Plan Families.	74
Figure 4.7	Results of NJ for Three Plan Families	76
Figure 4.8	Results of UPGM for Three Plan Families.	77
Figure 4.9	Comparison between GA and FA to building Phylogenetic Tree.	81

List of algorithms

Algorithm (3.1)	Evaluation all Genomes Boyer Moore Algorithm Good Suffix heuristic	35
Algorithm (3.2)	Evaluation all Genomes	37
Algorithm (3.3)	Select Genome Using Firefly Optimization Algorithm	41

Abbreviations

BLAST	Basic Local Alignment Search Tool
BM	Boyer Moore
DPSO	Discrete Particle Swarm Optimization
EST	Expressed sequence tag
FA	Firefly Optimization Algorithm
GA	Genetic algorithm
MSA	Multiple sequence alignment
NCBI	The National Center for Biotechnology Information
NJ	Neighbor joining
SW	Smith–Waterman algorithm
UPGM	Unweight Pair Group Method
DNA	deoxyribonucleic acid
RNA	Ribonucleic acid
OTU	Operational Taxonomic Unit

CHAPTTER ONE

General Introduction

Chapter One

General Introduction

1.1 Introduction

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyze and interpret biological data. Bioinformatics has been used in silicon analyses of biological queries using mathematical and statistical techniques [1].

Bioinformatics, a hybrid science that links biological data with techniques for information storage, distribution, and analysis to support multiple areas of scientific research, including biomedicine. Bioinformatics is fed by high-throughput data-generating experiments, including genomic sequence determinations and measurements of gene expression patterns [2].

Chloroplasts are one of the main organelles in plant cells. They are considered to have originated from cyanobacteria through endosymbiosis when a eukaryotic cell engulfed photosynthesizing cyanobacteria, which remained and became a permanent resident in the cell. Chloroplast can convert water, light energy, and carbon dioxide into chemical energy by using carbon-fixation cycle [3].

In bioinformatics, sequence analysis is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution. Methodologies used include sequence alignment, searches against biological databases, and others [4].

Genomics studies are a typical case of very large volume data investigations: they are of very high resolution (until nucleotide level) and very reliable (many

genome sequences contain less than one error every 10,000 bases), very abundant (100 sequenced eukaryotic genomes for instance, and more than 1,000 prokaryotes) and centralized in public databases. It informs about the likelihood of certain metabolic or developmental pathways that may exist in an organism and the importance of functions specific to each species. Genomes thus represent the foundation on which many advances can be achieved, and accessing such information in an ancestral genome provides a broad spectrum of these properties [5].

Phylogenetic is the science that studies the evolutionary relationship between species. TO make predictions about these relationships, phylogenetic trees are constructed which link the species. The Phylogenetic tree is a binary tree representation of the resulting relationship. It is construction methods are widely accepted to fall into one of two categories: distance-based and character-based. These two categories both offer a vast variety of options when constructing trees in two different directions [6]. Phylogenetic tree construction is a complex yet important problem in the field of bioinformatics. Once constructed, a phylogenetic or evolutionary tree can lend insight into the evolution of different species.

The issue is that for a large number of species the problem grows to a computational complexity that is not easily solved [7].

Robustness aspects of the produced trees can be evaluated too, for instance through bootstrap analyses. In other words, given a set of close plant species, their core genome (the set of genes in common) is as large and accurately detected as possible, to hope to be able to finally obtain a well-supported phylogenetic tree. However, all genes of the core genome are not necessarily constrained similarly, some genes having a larger ability to evolve than other ones due to their lower importance: such minority genes tell their own story instead of the species one,

blurring so the phylogenetic information. The link between the robustness and accuracy of the phylogenetic tree, and the amount of data used for this reconstruction, is not yet completely understood [2]. Genetic Algorithm (GA) using to solve the problem of finding the largest subset of core genes producing a phylogenetic tree as supported as possible. However, in some situations, this algorithm fails to solve the optimization problem due to a low convergence rate [8].

This thesis, proposed a system to building a phylogenetic tree in a faster and accurate manner to find optimal solution based on firefly optimization algorithm. In addition, the proposed system applies on a data set of the plants genome, this dataset contains a set of genome probabilities, and these genes are divided into primary and secondary genes. In the proposed system identifies the ancestral strains of the existing families and extracted the most identical hosts with each other based on the basic genes because they are the ones that carry the genes, noting that in Database there are many varieties of plants in the proposed system consisting of several main steps. First, the alignment process is intended to arrange the genome sequence so that will be known what the differences between each series. Second, used the Boyer Moore algorithm method that specifies a value for the series. Third, apply The firefly algorithm to generate the 1000 of the tree and then determine the percentage of match between the trees that were generated until choose the best.

1.2 Related Works

Several researchers have created many works about Genomics studies the following are some studies and discussions which associated with the proposed work in this thesis.

- **Alkindy, B., et al. (2015) [2]:** proposed Hybrid Genetic Algorithm for Inferring Well Supported Phylogenetic Trees. In this work focuses on how to extract the largest subset of sequences in order to obtain the most

supported species tree. Due to computational complexity, a distributed Binary Particle Swarm Optimization (BPSO) is proposed in sequential and distributed fashions.

- **Noutahi, E, et al. (2016) [9]:** proposed a new gene tree correction method, called Profile NJ, which can be directly used as a fast integrative method, without local search. It is a deterministic approach with a guaranteed time complexity
- **Alsrraj, R., et al.(2017) [1]:** A discrete particle swarm optimization algorithm has been proposed in this article, which focuses on the problem to extract the largest subset of core sequences with a view to obtain the most supported phylogenetic tree. The proposal of this research work is thus the application of a Discrete Particle Swarm Optimization (DPSO) that aims at finding the largest subset of core genes producing a phylogenetic tree as supported as possible. A new algorithm has been proposed and applied, in a distributive manner, to investigate the phylogeny of Rosales order.
- **Noutahi, E., and El-Mabrouk, N. (2018) [10]:** Several methods have been developed for the accurate reconstruction of gene trees. Some of them use reconciliation with a species tree to correct, a posterior, errors in gene trees inferred from multiple sequence alignments. The method is based on a genetic algorithm acting on a population of trees at each step. It substantially increases the efficiency of the phylogeny space exploration ,reducing the risk of falling in to local minima ,at a reason able computational time.
- **Zhao, T.,et al. (2020) [11]:** present a novel approach for phylogenetic tree reconstruction based on genome-wide synteny network data. The

proposed system results highlight that phylogenies based on genome structure and organization are complementary to sequence-based phylogenies and provide alternative hypotheses of angiosperm relationships to be further tested.

1.3 Problem Statement

The amount of completely sequenced chloroplast genomes increases rapidly every day, leading to the possibility to build large-scale phylogenetic trees of plant species. Considering a subset of close plant species defined according to their chloroplasts, the phylogenetic tree that can be inferred by their core genes is not necessarily well supported, due to the possible occurrence of "problematic" genes (i.e., homoplasy, incomplete lineage sorting, horizontal gene transfers, etc.) which may blur the phylogenetic signal.

However, a trustworthy phylogenetic tree can still be obtained provided such a number of blurring genes is reduced. The problem is thus to determine the largest subset of core genes that produces the best supported tree. To discard problematic genes and due to the overwhelming number of possible combinations.

1.4 Aim of Thesis

Each plant family contains a set of genome DNA and this genome DNA has a relationship with the other genome family of plants, in order to build a phylogenetic tree to find the relationship of the genomes of all plant families and to discover the relationship between the different family of plants, and therefore the objectives of this thesis are a proposed system to build A faster and more powerful phylogenetic tree to find optimal solutions based on the firefly optimization algorithm.

Also this work aims to conducting a comparative study between firefly optimization algorithm and Genetic algorithm depending on the performance of each of them in building the phylogenetic tree effectively, quickly and with high accuracy.

1.5 Contribution

The contribution of this thesis is using firefly optimization algorithm rather than genetic algorithm to provide best solutions in faster and more accuracy manner.

1.6 Layout of thesis

The rest of the thesis chapters are clarified as follow:

Chapter two: "theoretical background"

In this chapter explains about the algorithm of fireflies and algorithm used in constructing the phylogenetic tree

Chapter Three: **"**proposed System"

This chapter illustrates the design stages and the implementation requirements of the proposed system

Chapter Four: "Results and Discussion"

This chapter presents the implementation of the proposed algorithm

Chapter Five: "the conclusions and future work"

In this chapter explain conclusions about the implementation and results