# An Accurate Method for Phylogeny Tree Reconstruction Based on a Modified Wild Dog Algorithm

Essam Al Daoud

***Abstract*****—**This study solves a phylogeny problem by using modified wild dog pack optimization. The least squares error is considered as a cost function that needs to be minimized. Therefore, in each iteration, new distance matrices based on the constructed trees are calculated and used to select the alpha dog. To test the suggested algorithm, ten homologous genes are selected and collected from National Center for Biotechnology Information (NCBI) databanks (i.e., 16S, 18S, 28S, Cox 1, ITS1, ITS2, ETS, ATPB, Hsp90, and STN). The data are divided into three categories: 50 taxa, 100 taxa and 500 taxa. The empirical results show that the proposed algorithm is more reliable and accurate than other implemented methods.

***Keywords*****—**Least squares, neighbor joining, phylogenetic tree, wild dogpack.

## I. Introduction

A phylogenetic tree is a graph comprising nodes and branches, in which only one branch connects any two adjacent nodes. The units (nodes of the tree) can represent genes, species or populations—though not all at once, obviously. A gene tree represents the evolutionary history of a single gene (e.g., the evolution of the globin family, with its numerous gene duplication events). The first step in molecular phylogenetics is to select a suitable molecule that is homologous in all the taxa to be included in the phylogeny [1]. Homologous genes like ribosomal, 16s, 18s, 28s, Cox 1, and ITS-1. Inferring (reconstructing) a phylogeny consists of creating or selecting one tree out of perhaps millions of possible ones. Reconstructing the evolutionary history of molecular sequences through phylogenetic analysis is at the heart of many biological research areas such as comparative studies, epidemics, drug design, and forensic; it used both to explain and to predict. Therefore, several web sites provide phylogenetic tree constructions, and include PhyloBuilder and PhyloBlast both of which use a distance or a parsimony method. Phylemon 2.0 provides experts with a suite of online programs and a Java interface for building phylogeny trees [2]. Phylogeny.fr is the first web server designed for both non-specialists and experts, and it provides a complete automated phylogenetic analysis from a FASTA file to tree image [3].

The basic principle behind any tree construction method is to find the correct relationship among taxa from their DNA or RNA sequences, by using genes, proteins, species, populations, or higher taxonomic units [4]. The construction methods can be classified into two major categories: Methods based on distances and methods based on characters. Unfortunately, this task is very difficult from a computational perspective; the phylogeny problem is NP-hard. Therefore, several heuristic techniques have been used, ranging from simple constructive heuristics such as greedy methods and branch and bound techniques, to complex metaheuristics that use genetic algorithms, ant colonies and simulated annealing [5].

E. Al-Daoud is with faculty of Information Technology, Computer Science Department, Zarka University, Jordan (phone: +96279668000, e-mail: essamdz@zu.edu.jo).

## II. Related Work

Given a set of taxa $X$ and a distance matrix $D$, the unweighted pair group method using arithmetic averages (UPGMA) operates by clustering the given taxa at each stage, such that the average distance is:

$$d(i,j) = \frac{1}{|C_i||C_j|}\sum_{x\in C_i, y\in C_j} d(x,y) \quad (1)$$

where $C_i$ and $C_j$ are two disjointed sets of taxa and $d(x, y)$ is the distance between two taxa in matrix $D$. If $C_k$ is the union of two clusters, then

$$d(k,l) = \frac{d(i,l)|C_i| + d(j,l)|C_j|}{|C_i| + |C_j|} \quad (2)$$

At each step UPGMA determines two clusters $C_i$ and $C_j$ of which $d(i, j)$ is minimal, merging the clusters and at the same time creating a new node in the tree. Fig. 1 shows a sample output of UPGMA [6].

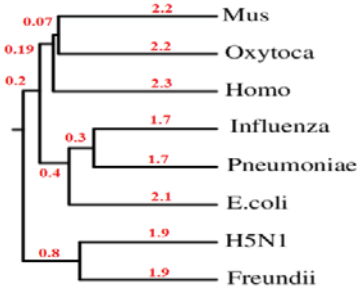


Fig. 1 Sample output of UPGMA

Neighbor Joining (NJ) is another quick distance-based method for approximating unrooted phylogenetic trees. NJ

starts with a completely unresolved tree and merges a new node for the smallest average distance of the pair *I* and *j*; then the distance is updated to

$$D_{k,(ij)} = (D_{ik} + D_{jk} - D_{ij})/2 \quad (3)$$

Under some conditions, the NJ method will yield a biased tree. Minimum least squares (MLS) has been used to enhance the distance base methods. Maximum parsimony (MP) is a form of character-based reconstruction. The criterion of the MP method is that the simplest explanation for the data is preferred, as it requires the fewest conjectures. The Fitch algorithm can be used to determine the minimum score for a fixed topology, such that the new set at each node is [7]:

$$R_i = \begin{cases} if\ R_j \cap R_k \neq \phi \rightarrow R_j \cap R_k \\ otherwise \longrightarrow R_j \cup R_k \end{cases} \quad (4)$$

The new character at node $i$ can be determined using the top-down phase.

$$s_i = \begin{cases} if\ s_j \in R_i \longrightarrow s_j \\ otherwise \longrightarrow arbitrary\ stat \in R_i \end{cases} \quad (5)$$

## III. A Modified Wild Dog Algorithm

The modified wild dog pack optimization (MWDPO) algorithm is suggested by Al Daoud [8]. The main advantage of MWDPO is its suitable balancing of exploitation and exploration strategies. MWDPO combines the advantages of harmony search algorithms and the wild dog algorithm; once a local minimum is detected, the algorithm will switch automatically to a new version of the harmony search with a suitable parameter to efficiently escape the local minimum [9]. Algorithm 1 summarizes the MWDPO**.**

**Algorithm1**: MWDPO

- Generate *n* dogs randomly and choose the best as alpha

**while**( *t* < *iter* )
if *flag*= 0
  Evaluate dogs Locations
if iteration % *q* =0
  Update the Parameters using self competition
  Select the new dogs Locations
  Evaluate Dogs
if no improvement for *v* iterations
*flag*=1
else
if (rand <$tol_1$)
  Update a dog randomly
else if (rand <$tol_2$)
adjust the alpha dog by *seq*
else
exclude a dog randomly and initialize a new one
after *k* sub-iterations switch *flag* to 0

In an MWDPO algorithm, the set *seq* is a set of numbers, such that the next number is equal to the previous number * 10. For each iteration the alpha dog is adjusted randomly, according to the set *seq*.

## IV. Least Squares Method

The MLS approach has been used to enhance distance-based methods. MLS is efficient with a fixed tree topology [10]. Let $D_{ij}$ be the pairwise distance between two species, and let $T_{ij}$ be the distance between $i$ and $j$ in a tree; the sum of the length of all the branches on the unique path between $i$ and $j$—for example, the tree distance between $a$ and $c$ in Fig. 2 is $T_{ac}=x+y+z$.

MLS can be used to find a tree minimizing (6).

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}(D_{ij} - T_{ij})^2 \quad (6)$$

The exact solution for a fixed tree can be found in polynomial time by solving:

$$\frac{dQ}{dx} = -2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} \left(D_{ij} - \sum_{x \in T_{ij}} x\right) = 0 \quad (7)$$

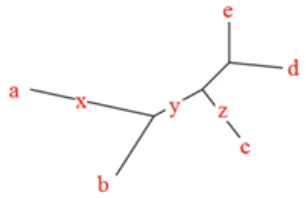


Fig. 2 Tree distance between two nodes

## V. A New Reconstruction Algorithm

Using MLS to optimize the general phylogenetics tree problem is an NP-hardness problem [11]; therefore, it is computationally infeasible. Algorithm 2 introduces the main steps by which to reduce the errors in MLS by using MWDPO.

**Algorithm 2**: MLS-MWDPO

- **Input**: The matrix of the pairwise distance of the taxa (*D*), the taxa size (*n*)
- **Output**: Near-optimum phylogenetic tree

Steps:

1- Compute the approximate phylogenetic tree using NJ or UPGMA
2- Calculate the approximated distances between the tree leaves and store them in *T*
3- Generate *m* vector randomly; each one consists of the $(n^2-n)/2$ element
4- Store the vectors in *m* lower matrices and call them $M_i$, where$i=1..m$.
5- Calculate new distances $C_i=A+M_i$.
6- Construct a tree $T_i$ for each of $C_i$
7- Find MLS for each of $C_i$, and choose the best $C_b$ and call $M_b$ alpha
8- Call MWDPO by considering $M_i$ as the dogs; the evaluation of each dog is achieved by using MLS as done in steps 5, 6, and 7.
9- Iterate this process until no further improvement can be obtained.

Fig. 3 shows the gene "Cox 1" from different taxa in FASTA format; Fig. 4 describes a part of the multiple sequences that are aligned by using ClustalW2. Fig. 5 shows

the distance matrix *D* for the aligned sequences; Fig. 6 illustrates the NJ tree, using the previous distance matrix; Fig. 7 shows a representation of the tree in the Newick format; and Fig. 8 is the corresponding distance matrix T, which is extracted from the tree in Fig. 6. The least squares error of *D* and *T* is *43.46/64=0.68*. However, this error can be quickly reduced by using Algorithm 2; for instance, this error becomes 0.02 after a few iterations. The random matrices $M_i$ (the dogs) are generated within the range *±max(T-D)*.

```
>Influenza B
MANNNMTTTQIEVGPGATNATINFEAGILECYERLSWQRALDYPGQDRLNRLKRKLESRIKTHNKSEPES
KRMSLEERKAIGVKMMKVLLFMNPSAGIEGFEPYCMNSSSNSNCTKYNWTDYPSTPERCLDDIEEEPEDV
DGPTEIVLRDMNNKDARQKIKEEVNTQKGKFRLTIKRDMRNVLSLKVLVNGTFLKHPNGYKSLSTLHRL
NAYDQSGRLVAKLVATDDLTVEDEEDGHRILNSLFERLNEGHSKPIRAAETAVGVLSQFGQEHRLSPEEG
DN

>Pneumoniae
MSVIGRIHSFESCGTVDGPGIRFITFFQGCLMRCLYCHNRDTWDTHGGKEITVEELMKEVVTYRHFMNAS
GGGVTASGGEAILQAEFVRDWFRACKKEGIHTCLDTNGFVRRYDPVIDELLEVTDLVMLDLKQMNDEIHQ
NLVGVSNHRTLEFAQYLAKKNINVWIRYVVVPGWSDDDDSAHRLGEFTRDMGNVEKIELLPYHELGKHKW
VAMGEEYKLDGVHPPKKETMERVKGILEQYGHKVMY

>Homo sapiens
MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPPSWYKVFDPSCGLPYYWNADT
DLVSWLSPHDPNSVVTKSAKKLRSSNADAEEKLDRSHDKSDRGHDKSDRSHEKLDRGHDKSDRGHDKSDR
DRERGYDKVDRERERDRERDRDRGYDKADREEGKERRHHRREELAPYPKSKKAVSRKDEELDPMDPSSYS
DAPRGTWSTGLPKRNEAKTGADTTAAGPLFQQRPYPSPGAVLRANAEASRTKQQD

>H5N1
MSLLTEVETYVLSIIPSGPLKAEIAQKLEDVFAGKNTDLEALMEWLKTRPILSPLTKGILGFVFTLTVPS
ERGLQRRRFVQNALNGNGDPNNMDRAVKLYKKLKREITFHGAKEVALSYSTGALASCMGLIYNRMGTVTT
EVAFGLVCATCEQIADSQHRSHRQMATITNPLIRHENRMVLASTTAKAMEQMAGSSEQAAEAMEIANQAR
QMVQAMRTIGTHPNSSAGLRDNLLENLQAYQKRMGVQMQRFK
```

Fig. 3 The gene "Cox 1" from different taxa, in FASTA format

```
Influenza       --MANNNMTTTQIEVGPGAT-NATINFEAGILECYERLSWQRALDYPGQDRLN-------
Pneumoniae      MSVIGRIHSFESCGTVDGPG-IRFITFFQGCLMRCLYCHNRDTWDTHGGKEIT-------
Homo_sapie      ---MPLPVALQTRLAKRGIL-KHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPPS-----
H5N1            MSLLTEVETYVLSIIPSGPL-KAEIAQKLEDVFAGKNTDLEALMEWLKTRPILS------
E.coli          --MKQPAPVYQRIAGHQWRH-IWLSGDIHGCLEQLRRKLWHCRFDPWR------------
Oxytoca         MAFVTTKDGVNIYFKDWGPKEAQPIVFHHGWPLSADDWDNQMLFFLAEGFRVIAIDRRGH
Freundii        --MLADIKYWENDAQNKHYAIAHFNVWNAEMLMGVIDAAEESKSPVIISFGTGFVGNTSF
Mus_muscul      -------------MIPNGYLMFEDENFIESSVAKLNALRKSGQFCDVR------------


Influenza       -RLKRKLESRIKTHNKSEPESKRMSLEERKAIGVKMMKVLLFMNPSAGIEGFEPYCMNSS
Pneumoniae      -VEELMKEVVTYRHFMNASGGGVTASGGEAILQAEFVRDWFRACKKEGIH----TCLDTN
Homo_sapie      -WYKVFDPSCGLPYYWNADTDLVSWLSPHDPNSVVTKSAKKLRSSNADAEEKLDRSHDKS
H5N1            ---PLTKGILGFVFTLTVPSERGLQRRRFVQNALNGNGDPNNMDRAVKLYKKLKREITFH
E.coli          ---DLLISVGDVIDRGPQSLRCLQLLEQHWVCAVRGNHEQMAMD----------------
Oxytoca         GRSDQVSDGHDMDHYAADASAVAEHLDLHNAVHVGHSTGGGQVARYVAKYGQPQGRVAKA
Freundii        EDFSHMMVSMAKKASVPVITHWDHGRSMEIIHNAWAHGMNSLMRDASAFDFEENIRLTKE
Mus_muscul      ----LQVCGHEMLAHRAVLACCSPYLFEIFNSDSDPHGVSHVKLDDLN------------
```

Fig. 4 Part of a ClustalW2 output

| Species | Influenza | H5N1 | E.coli | Homo | Oxytoca | Pneumoniae | Freundii | Mus |
|---|---|---|---|---|---|---|---|---|
| Influenza | 0.0 | | | | | | | |
| H5N1 | 6.6 | 0.0 | | | | | | |
| E.coli | 4.6 | 4.7 | 0.0 | | | | | |
| Homo | 5.7 | 7.2 | 5.3 | 0.0 | | | | |
| Oxytoca | 4.7 | 4.8 | 5.9 | 4.7 | 0.0 | | | |
| Pneumoniae | 3.5 | 4.7 | 3.8 | 4.8 | 4.8 | 0.0 | | |
| Freundii | 4.8 | 3.9 | 7.8 | 5.8 | 4.9 | 4.3 | 0.0 | |
| Mus | 4.6 | 4.7 | 4.8 | 4.6 | 4.5 | 4.7 | 5.8 | 0.0 |

Fig. 5 Distance matrix D

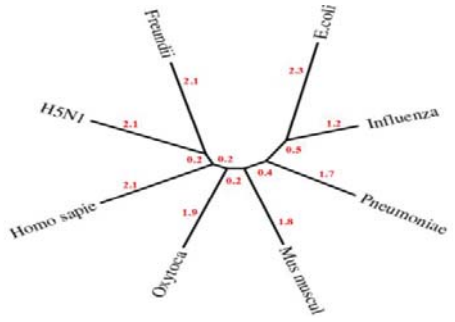


Fig. 6 Neighbor joining tree using the distance matrix D

```
(Oxytoca:1.974182,(Homo_sapiens:2.138110,(H5N1:2.126471,Freundii:
2.189509):0.270520):0.262500,( ( ( E.coli: 2.305189, Influenza_B:
1.277130):0.595765, Pneumoniae: 1.727292): 0.417546,Mus_musculus:
1.875921):0.298965);
```

Fig. 7 Newick format

| Species | Influenza | H5N1 | E.coli | Homo | Oxytoca | Pneumoniae | Freundii | Mus |
|---|---|---|---|---|---|---|---|---|
| Influenza | 0.0 | | | | | | | |
| H5N1 | 5.2 | 0.0 | | | | | | |
| E.coli | 3.6 | 6.3 | 0.0 | | | | | |
| Homo | 5.0 | 4.5 | 6.0 | 0.0 | | | | |
| Oxytoca | 4.6 | 4.6 | 5.6 | 4.4 | 0.0 | | | |
| Pneumoniae | 3.6 | 5.1 | 4.6 | 4.8 | 4.4 | 0.0 | | |
| Freundii | 5.3 | 4.3 | 6.3 | 4.6 | 4.7 | 5.2 | 0.0 | |
| Mus | 4.2 | 4.8 | 5.2 | 4.6 | 4.1 | 4.0 | 4.9 | 0.0 |

Fig. 8 Distance matrix T

## VI. Data and Experimental Results

In this study, ten homologous genes are selected and collected from the National Center for Biotechnology Information (NCBI) site. The selected genes are in RNA format and are taken from several species, organisms, and taxa. The species collected include animals, plants, fungi, bacteria, archaea, and viruses. Hundreds of taxa are used such asproteobacteria, firmicutes, actinobacteria, cyanobacteria, chlamydiae, euryarchaeotes and crenarchaeota. Table I describes the total number of genes, species, and taxa available through the NCBI site.

Five algorithms are implemented, including UPGMA, NJ, MLS with a harmony search [12], and the suggested algorithm (i.e., MLS–MWDPO). Tables II-IV show the least squares error using 50, 100, and 500 taxa, respectively. The suggested

algorithm outperforms the other methods, although it requires more implementation time.

TABLE I
TOTAL NUMBER OF GENES, SPECIES AND TAXA

| Gene | #Species | #Taxa |
|---|---|---|
| 16S | 7 | 41800 |
| 18S | 7 | 1786 |
| 28S | 7 | 6725 |
| Cox 1 | 6 | 7428 |
| ITS1 | 5 | 160 |
| ITS2 | 5 | 85 |
| ETS | 7 | 6348 |
| ATPB | 6 | 1112 |
| Hsp90 | 7 | 32532 |
| STN | 7 | 689 |

TABLE II
LEAST SQUARES ERROR USING 50 TAXA

| Gene | UPGMA | NJ | MLS-HS | MLS-MWDPO |
|---|---|---|---|---|
| 16S | 0.91 | 0.89 | 0.31 | 0.02 |
| 18S | 1.84 | 1.64 | 0.27 | 0.03 |
| 28S | 1.66 | 1.73 | 0.23 | 0.03 |
| Cox 1 | 2.12 | 2.02 | 0.10 | 0.01 |
| ITS1 | 1.38 | 1.50 | 0.09 | 0.02 |
| ITS2 | 1.24 | 1.33 | 0.08 | 0.02 |
| ETS | 2.22 | 1.62 | 0.12 | 0.01 |
| ATPB | 1.61 | 2.77 | 0.23 | 0.03 |
| Hsp90 | 3.40 | 3.24 | 0.41 | 0.11 |
| STN | 2.31 | 1.43 | 0.08 | 0.02 |

TABLE III
LEAST SQUARES ERROR USING 100 TAXA

| Gene | UPGMA | NJ | MLS-HS | MLS-MWDPO |
|---|---|---|---|---|
| 16S | 2.73 | 2.34 | 0.61 | 0.22 |
| 18S | 3.55 | 2.87 | 0.59 | 0.18 |
| 28S | 3.64 | 3.12 | 0.48 | 0.24 |
| Cox 1 | 4.27 | 3.94 | 0.32 | 0.21 |
| ITS1 | 2.02 | 1.86 | 0.33 | 0.12 |
| ETS | 3.11 | 2.70 | 0.28 | 0.18 |
| ATPB | 2.98 | 2.54 | 0.71 | 0.21 |
| Hsp90 | 5.62 | 4.03 | 1.02 | 0.29 |
| STN | 3.38 | 3.15 | 0.12 | 0.13 |

TABLE IV
LEAST SQUARES ERROR USING 500 TAXA

| Gene | UPGMA | NJ | MLS-HS | MLS-MWDPO |
|---|---|---|---|---|
| 16S | 15.28 | 12.67 | 3.74 | 1.08 |
| 18S | 18.56 | 13.45 | 2.78 | 1.15 |
| 28S | 17.15 | 14.72 | 3.90 | 0.82 |
| Cox 1 | 20.36 | 13.35 | 2.96 | 1.32 |
| ETS | 16.75 | 10.01 | 1.92 | 0.82 |
| ATPB | 13.22 | 9.18 | 2.56 | 1.32 |
| Hsp90 | 26.19 | 16.99 | 5.24 | 2.03 |
| STN | 21.37 | 10.62 | 3.46 | 1.26 |

## VII. CONCLUSION

Distance-based phylogenetic methods are widely used in biomedical research. The general thinking in using distance methods is to calculate the distance between each pair of species, and then find a tree that predicts the observed set of distances as closely as possible. One major problem with the phylogeny problem is that in order to find the optimal solution, all possible tree topologies need to be evaluated. However, for even a small number of species, the size of the tree space makes such an approach infeasible. In this study, modified wild dog pack optimization was used to reduce the least squares error. ten homologous genes were selected and collected from NCBI databanks, and the data were divided into three categories. The results indicate that the suggested algorithm is superior to the other implemented methods.

## REFERENCES

[1] J. Gatesy, M. S. Springer, "Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/ concatalescence conundrum. Molecular Phylogenetics and Evolution vol. 80, no. 11, pp231–266, 2014.

[2] M. Riester, C. S. Attolini, R. J. Downey, S. Singer, F. Michor, "A Differentiation-Based Phylogeny of Cancer Subtypes," *Plos Comp. Biol*., vol. 6, no. e1000777, pp1-14, 2010.

[3] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.-F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J.-M. Claverie and O. Gascuel, "Phylogeny.fr: robust phylogenetic analysis for the non-specialist," *Nucleic Acids Research*, vol. 36, pp465–469, 2008.

[4] S. Kumar, M., Nei, J., Dudley, K., Tamura, "MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences," *BriefBioinform*., vol 9, pp. 299–306, 2008.

[5] S. Roch, "Toward extracting all phylogenetic information from matrices of evolutionary distances," *Science,* vol 327, pp1376-1379, 2010.

[6] P. J. Waddell, H. Kishino, R. Ota, "Phylogenetic Methodology for Detecting Protein Complexes," *Mol. Biol. Evol.*, vol 24, pp650-659, 2007.

[7] O. Gascuel, M. Steel, "Neighbor-Joining Revealed," *Molecular Biology and Evolution*, vol23, no 11, pp1997-2000, 2006.

[8] E. Al Daoud, "A Modified Optimization Algorithm Inspired by Wild Dog Packs," *International Journal of Science and Advanced Technology*, vol. 4, no. 9, pp: 25-28, 2014.

[9] E. Al Daoud, R. Alshorman, F. Hanandeh, "A New Efficient Meta-Heuristic Optimization Algorithm Inspired by Wild Dog Packs," *International Journal of Hybrid Information Technology*, vol. 7, no. 6, pp: 83-100, 2014.

[10] R. Mihaescu, L. Pachter, "Combinatorics of least-squares trees," *Proc. Nat. Acad. Sci.*, vol. 105, pp13206-13207, 2008.

[11] C. C. Ribeiro, D.S. Vianna, "A hybrid genetic algorithm for the phylogeny problem using path-relinking as a progressive crossover strategy," *International Transactions in Operational Research*, vol. 16, no. 5, pp641–657, 2009.

[12] E. Al Daoud, "An Efficient Algorithm for Finding a Fuzzy Rough Set Reduct Using an Improved Harmony Search," *I.J. Modern Education and Computer Science*, vol. 7, no. 2, pp16-23, 2015.