July 2010

Sequence Alignment and Phylogenetic Tree Construction of Malarial Parasites

Sk. Mujaffor

M.Tech (2nd Yr.), ITER, S'O'A University, mujaffor09@yahoo.in

Tripti Swarnkar

Deptt. Of CA, SOA University Bhubaneswar, tripti_sarap@yahoo.com

Raktima Bandyopadhyay

Dept. of Bioinformatics, Vidyasagar University, raktima.bioinformatics@gmail.com

Follow this and additional works at: <https://www.interscience.in/ijcct>

Recommended Citation

Mujaffor, Sk.; Swarnkar, Tripti; and Bandyopadhyay, Raktima (2010) "Sequence Alignment and Phylogenetic Tree Construction of Malarial Parasites," *International Journal of Computer and Communication Technology*: Vol. 1 : Iss. 3 , Article 6.

DOL: 10.47893/IJCCT.2010.1042

Available at: <https://www.interscience.in/ijcct/vol1/iss3/6>

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer and Communication Technology by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

Sequence Alignment and Phylogenetic Tree Construction of Malarial Parasites

Sk. Mujaffor¹, Tripti Swarnkar², Raktima Bandyopadhyay³
M.Tech (2nd Yr.), ITER, S'O'A University
mujaffor09@yahoo.in

Department of Computer Applications
Institute of Technical Education & Research, S'O'A University, Bhubaneswar
tripti_sarap@yahoo.com

Dept. of Bioinformatics, Vidyasagar University
raktima.bioinformatics@gmail.com

Abstract-Sequence alignment is one of the basic problems in computational biology that has helped researchers analyze biological sequences. The analysis has helped biologists to detect pathogens; to develop drugs, and to predict the secondary and tertiary structure of a protein and identify common genes. The objective of the Phylogenetic tree is to determine the branch length and to figure out how the evolutionary tree has been generated. One way to tackle MSA is to use Hidden Markov Models (HMMs), which are known to be very powerful in the related problem domain of speech recognition. The fully trained model is applied to draw a valid conclusion about the evaluation of malarial parasites.

Keywords- Sequence alignment; Phylogenetic tree; HMM; MSA; ClustalW; Merozoite surface protein; BioEdit

I. INTRODUCTION

Multiple sequence alignment (MSA) [5] of nucleotides (or amino acids) is one of the basic problems in computational biology. Good alignments allow sequence comparison, which can be used for a variety of purposes, such as to determine the phylogenetic relatedness of organisms, to identify conserved motifs and to assist secondary and tertiary structure prediction. Through the sequence alignment it can be resolved about the transmission of disease by parasites. Zoonosis is a term that means transmission of a disease from subhuman vertebrate to human body. For the evolution of parasite and the evolution of parasitic disease, the study of Zoonosis is very important in respect to the epidemiology of the disease. India is endemic for malaria and it's a global problem also. Human malaria is basically caused by four

parasites – *Plasmodium vivax*, *Plasmodium falciparum*, *Plasmodium ovale* and *Plasmodium*

malariae. *Plasmodium cynomolgi* is a malarial parasite of monkey and *Plasmodium berghei* is the rodent parasite. Our objective is to find out the Zoonosis of malarial parasites.

A. Sequences in the realm of a biologist

A sequence for a biologist is either a RNA, DNA or protein string made of their respective alphabet set shown below :

DNA = { A, C, G, T }

RNA = { A, C, G, U }

Protein = { A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V }

B. Sequence Alignment

A Sequence alignment [1] means lining up the characters of strings, allowing mismatches as well as matches and allowing characters of one string to be placed opposite spaces made in opposing strings. Our objective is to find the regions of similarity which may provide additional information on the functional, structural, evolutionary and other interests between the sequences.

C. Phylogenetic Tree

The similarity of molecular mechanisms of the organisms that have been studied strongly suggests that all organisms on Earth had a common ancestor. Thus any set of species is related, and this relationship is called a phylogeny. Usually the relationship can be represented by a phylogenetic tree [4]. The task

of phylogenetics is to infer this tree from observations upon the existing organisms.

D. *Hidden Markov Model*

A hidden Markov model (HMM) [5] is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a *hidden* Markov model, the state is not directly visible, but output dependent on the state is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; Even if the model parameters are known exactly, the model is still 'hidden'. There are three canonical problems associated with HMM:

- Given the parameters of the model, compute the probability of a particular output sequence. This requires summation over all possible state sequences, but can be done efficiently using the [forward algorithm](#), which is a form of [dynamic programming](#).
- Given the parameters of the model and a particular output sequence, find the state sequence that is most likely to have generated that output sequence. This requires finding a maximum over all possible state sequences, but can similarly be solved efficiently by the [Viterbi algorithm](#).
- Given an output sequence or a set of such sequences, find the most likely set of state transition and output probabilities. In other words, derive the [maximum likelihood](#) estimate of the parameters of the HMM given a dataset of output sequences. No tractable algorithm is known for solving this problem exactly, but a local maximum likelihood can be derived efficiently using the [Baum-Welch algorithm](#) or the [Baldi-Chauvin algorithm](#). The [Baum-Welch algorithm](#) is also

known as the [forward-backward algorithm](#), and is a special case of the [Expectation-maximization algorithm](#).

E. *Multiple Sequence Alignment*

Multiple Sequence Alignment (MSA), is an extension of two-sequence/pairwise sequence alignment. Nowadays, multiple sequence alignment is an important tool in molecular biology and it provides key information for sequence analysis. There are several uses of MSA; finding sequence to determine patterns that characterize protein/gene families; detecting homology between new sequences and known protein/gene family sequences; predicting secondary and tertiary structures of new protein sequences; predicting function of new sequences and molecular evolutionary analysis.

F. *ClustalW*

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins. It is also based on HMM. It produces biologically meaningful multiple sequence alignment of divergent sequences[3]. It calculates the best match for the selected sequences and lining them up so that the identities, similarities and differences can be seen. Evolutionary relationship can be seen via viewing cladograms or phylograms.

G. *Merozoite surface protein*

A merozoite surface protein is a protein molecule taken from the surface of a merozoite. Merozoite surface proteins are used in researching malaria, caused by protozoans.

H. *BioEdit*

BioEdit is a biological sequence editor that runs in Windows 95/ 98/ 2000 and is intended to provide basic functions for protein and nucleic sequence editing, alignment, manipulation and analysis. It offers a graphical interface for users to run external analysis programs

II. MATERIALS OF METHOD

The sequences of merozoite surface protein of the malarial parasites i.e. *Plasmodium vivax*, *Plasmodium falciparum*, *Plasmodium berghei*, *Plasmodium cynomolgi* were downloaded from National Center for Biotechnology Information (NCBI). The sequences were FASTA [2] formatted and multiple sequence alignment was done by using ClustalW. It was also determined about the amino acid composition of the merozoite surface protein of all the parasites by BioEdit. Phylogenetic tree was constructed. The sequences of malaria parasites are

A. *Plasmodium berghei*

MKVIGLLFSFVFFAIKCKSETIEVYNDIIQKL
EKLESLSVEGLELFQKSQVIINASPPSETINP
FSDNTFAPKLQGFITP.....

B. *Plasmodium cynomolgi*

MKALLFLFSFIFFVTKCQCETEDYKQLLVK
LDKLEALVVDGYELFQKKKLEVKDIPVDN
NANENNVNSLAYKIR.....

C. *Plasmodium falciparum*

FINNAYNMSIRRSMAESKTPTGAGG
SGSAGGSGSAGGSGSAGGSGSAGST
TTTNDAEASTSTSSNPNNHNAET...

D. *Plasmodium vivax*

EIYDLAQEIRKNENKLIVENKDFSGVVELQ
VQKVLIIKKIEALKNVQNLLKNAKVKDDL
YVPKVYKTGEKPEPYLMVLKREIDKLKD
.....

III. RESULT DISCUSSION

From the sequence alignment and phylogenetic tree construction it has been observed that there is a very close relationship between *Plasmodium cynomolgi* and *Plasmodium vivax* (Max score, Total score, Query coverage and E-value). It has shown below :

Accession	Description	M ax sc ore	To tal sc ore	Quer y cove rage	E val ue
BAF74064.1	merozoite surface protein-1 [<i>Plasmodium cynomolgi</i>]	3645	3645	100%	0.0
BAF74054.1	merozoite surface protein-1 [<i>Plasmodium cynomolgi</i>]	2887	2887	100%	0.0
BAF74055.1	merozoite surface protein-1 [<i>Plasmodium cynomolgi</i>]	2866	2866	100%	0.0
BAF74060.1	merozoite surface protein-1 [<i>Plasmodium cynomolgi</i>]	2862	2862	100%	0.0
BAI82251.1	merozoite surface protein 1 [<i>Plasmodium cynomolgi</i>]	2856	2856	98%	0.0
AAN86210.1	merozoite surface protein 1 [<i>Plasmodium vivax</i>] >gb AAN86225.1 A F435612_1	2287	2287	98%	0.0
AAN86235.1	merozoite surface protein 1 [<i>Plasmodium vivax</i>] >gb AAN86239.1 A F435629_1	2274	2274	98%	0.0
AAN86238.1	merozoite surface protein 1 [<i>Plasmodium vivax</i>] >gb AAN86241.1 A F435631_1	2269	2269	98%	0.0
AAN86215.1	merozoite surface protein 1 [<i>Plasmodium vivax</i>] >gb AAN86216.1 A F435603_1	2266	2266	98%	0.0

Accession	Description	M To Quer	E
		ax tal y	val
		sc sco cove	ue
		ore re rage	

protein 1
[Plasmodium vivax]

A. Alignments

[dbj|BAF74064.1|](#) merozoite surface protein-1
[Plasmodium cynomolgi]

Length=1786, Score = 3645 bits (9453),
Expect = 0.0, Method: Compositional matrix
adjust. Identities = 1786/1786 (100%), Positives
= 1786/1786 (100%), Gaps = 0/1786 (0%)

Query 1

MKALLFLFSFIFFVTKCQCETEDYKQLLVK
LDKLEALVVDGYELFQKKKLEVKDIPVDN
N 60
MKALLFLFSFIFFVTKCQCETEDYKQLLVK
LDKLEALVVDGYELFQKKKLEVKDIPVDN
N

Sbjct 1

MKALLFLFSFIFFVTKCQCETEDYKQLLVK
LDKLEALVVDGYELFQKKKLEVKDIPVDN
N 60

.....

[dbj|BAD08401.1|](#) merozoite surface protein 1
[Plasmodium falciparum]

Length=1688 Score = 1084 bits (2804), Expect
= 0.0, Method: Compositional matrix adjust.
Identities = 707/1888 (37%), Positives =
1037/1888 (54%), Gaps = 311/1888 (16%)

Query 1

MKALLFLFSFIFFVTKCQCET-
EDYKQLLVKLDKLEALVVDGYELFQKKKL
EVKD----- 54

MK + FL SF+FF+ QC T E Y++L+ KL+ LE
V+ GY LFQK+K+ +KD

Sbjct 1

MKIIFFLCSFLFFIINTQCVTHESYQELVKKL
EALEDAVLGTGYSLFQKEKMLKDGANTQ
60

.....

[gb|AAN86210.1|AF435596_1](#) merozoite
surface protein 1 [Plasmodium vivax]
[gb|AAN86225.1|AF435612_1](#) merozoite
surface protein 1 [Plasmodium vivax]

Length=1728, Score = 2287 bits (5927), Expect
= 0.0, Method: Compositional matrix adjust.
Identities = 1241/1773 (69%), Positives =
1391/1773 (78%), Gaps = 82/1773 (4%)

Query 1

MKALLFLFSFIFFVTKCQCETEDYKQLLVK
LDKLEALVVDGYELFQKKKLEVKDIPVDN
N 60
MKALLFLFSFIFFVTKCQCETE YKQL+
KLDKLEALVVDGYELF KKKL DI V+ N

Sbjct 1

MKALLFLFSFIFFVTKCQCETESYKQLVAK
LDKLEALVVDGYELFHKKKLGENDIKVEA
N 60

.....

B. Phylogenetic Tree

The phylogenetic trees made by Neighbour
Joining method, Maximum parsimony method,
Unweighted pair group method with arithmetic
mean (UPGMA method), Minimum
Evolutionary distance method (ME method) are
shown by figure no. 2, 3, 4 and 5 respectively.

Fig.2

Fig.3

Fig.4

Fig.5

C. Amino acid composition - BioEdit

The amino acid composition of four malarial parasites are Plasmodium berghei, Plasmodium cynomolgi, Plasmodium falciparum and Plasmodium vivax shown by figure no. 6,7,8 and 9 respectively.

Fig.6

Fig.7

Fig.8

Fig.9

Protein: Plasmodium berghei

Length = 1787 amino acids

Molecular Weight = 198146.17 Daltons

Amino Acid	Number	Mol%
Ala A	139	7.78
Cys C	19	1.06
Asp D	74	4.14
Glu E	160	8.95
Phe F	52	2.91
Gly G	80	4.48
His H	12	0.67
Ile I	118	6.60
Lys K	176	9.85
Leu L	152	8.51
Met M	21	1.18
Asn N	128	7.16

Pro P	90	5.04
Gln Q	66	3.69
Arg R	39	2.18
Ser S	134	7.50
Thr T	169	9.46
Val V	79	4.42
Trp W	1	0.06
Tyr Y	78	4.36

Protein: *Plasmodium cynomolgi*

Length = 1786 amino acids

Molecular Weight = 198841.67 Daltons

Amino Acid	Number	Mol%
Ala A	153	8.57
Cys C	19	1.06
Asp D	94	5.26
Glu E	170	9.52
Phe F	46	2.58
Gly G	90	5.04
His H	24	1.34
Ile I	97	5.43
Lys K	206	11.53
Leu L	171	9.57
Met M	34	1.90
Asn N	112	6.27
Pro P	89	4.98
Gln Q	71	3.98
Arg R	28	1.57
Ser S	96	5.38
Thr T	121	6.77
Val V	88	4.93
Trp W	3	0.17
Tyr Y	74	4.14

Protein: *Plasmodium falciparum*

Length = 196 amino acids

Molecular Weight = 19665.25 Daltons

Amino Acid	Number	Mol%
Ala A	22	11.22
Cys C	0	0.00
Asp D	6	3.06
Glu E	13	6.63
Phe F	1	0.51
Gly G	24	12.24
His H	5	2.55
Ile I	2	1.02
Lys K	8	4.08
Leu L	1	0.51
Met M	3	1.53
Asn N	23	11.73
Pro P	14	7.14
Gln Q	15	7.65
Arg R	4	2.04
Ser S	28	14.29
Thr T	23	11.73
Val V	3	1.53
Trp W	0	0.00

Protein: *Plasmodium vivax*

Length = 338 amino acids

Molecular Weight = 37344.74 Daltons

Amino Acid	Number	Mol%
Ala A	35	10.36
Cys C	2	0.59
Asp D	15	4.44
Glu E	25	7.40
Phe F	7	2.07
Gly G	13	3.85
His H	4	1.18
Ile I	18	5.33
Lys K	36	10.65
Leu L	28	8.28
Met M	7	2.07
Asn N	21	6.21
Pro P	19	5.62
Gln Q	26	7.69
Arg R	3	0.89
Ser S	16	4.73
7.69	23	6.80
0.00	26	
Trp W	0	
Tyr Y	14	4.14

D. The Pairwise evolutionary distance
are shown below:

Title: para

Description

No. of Taxa : 4

Data File : para

Data Title : para

Data Type : Amino acid

Analysis : Disparity Index Analysis

Calculate : Conduct ID-Test (1000 reps;
seed=86348)

Include Sites

->Gaps/Missing Data : Complete Deletion

No. of Sites : 193

Prob (black) : Probability computed (must be
<0.05 for hypothesis rejection at 5% level
[yellow background])

Stat (blue) : Disparity Index.

[1] #*Plasmodium_berghei*[2] #*Plasmodium_cynomolgi*[3] #*Plasmodium_falciparum*[4] #*Plasmodium_vivax*

[1 2 3 4]

[1] [1.119][3.150][1.415]

[2] 0.001 [2.539][0.337]

[3] 0.000 0.000 [2.430]

[4] 0.000 0.018 0.000

.IV. CONCLUSION

Among the four human malarial parasites only *Plasmodium vivax* was found to be very close to monkey parasite i.e, *Plasmodium cynomolgi*. So it may be predicted that malaria was transmitted from monkey to man. As a case of Zoonosis, the *Plasmodium cynomolgi* might be mutated and modified in such a way so that it could adapt to the human body and ultimately established a human parasite.

V. REFERENCES

- [1] A. L. Delcher, et al., "Alignment of whole genomes," *Nucl. Acids Research*, vol. 27, pp. 2369-2376, 1999.
- [2] D. Gusfield, *Algorithms on Strings, Trees and Sequences: Computer science and Computational Biology*. Cambridge University Press, 1997.
- [3] M. Tompa, "Lecture notes on Biological Sequence Analysis," University of Washington, Seattle, Technical report, 2000.
- [4] Neil C. Jones and Pavel A. Pevzner, 2004 – "An Introduction to Bioinformatics Algorithms". [5] Richard Durbin, Eddy, Mitchison, 1998- "Biological Sequence Analysis".