Journal of Intellectuals (JoI)

Original article

Phylogenetic study of eukaryotic and prokaryotic haemoglobin using PHYLIP

Rajdeep Das [1] and Swarupa Bhattacharjee *[2]

[1] Department of Zoology, Dudhnoi College, Dudhnoi, Assam, India

[2] Department of Zoology, Bahona College, Jorhat, Assam, India

*Corresponding author email: bhattacharjeeswarupa@gmail.com

Citation: Das, R.; Bhattacharjee, S.; (2024). Phylogenetic study of eukaryotic and prokaryotic haemoglobin using PHYLIP. *Journal of Intellectuals*, 4(1), 44–52. Retrieved from https://journals.bahonacollege.edu.in/index.php/joi/article/view/joi2024-4-1-6

Received: 18 September, 2024
Revised: 09 November, 2024

Accepted: 12 December, 2024
Published: 25 December, 2024

Publisher's Note: JOI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: Haemoglobins are the most popular and essential protein groups found in wide range of organisms. The gene encoding haemoglobin is thus very old, going back to the ancestor common to essentially all life on this earth. Their phylogenetic analysis seems to be very crucial from the perspective of evolutionary knowledge and preparation of phylogenetic tree. Clustal W is a popularly used multiple sequence alignment algorithm. This study intends to draw a phylogenetic tree with amino acid sequences of the protein haemoglobin of selected species and thereafter study their phylogenetic relationship. Haemoglobin molecule found ubiquitously in plants, animals and microbes were analysed using PHYLIP software to estimate the relationships among the taxa and their hypothetical common ancestor. The analysis involved 22 amino acid sequences of haemoglobin from two bacteria, two fungus, four animals and 16 plants. This is a comparatively simpler method for the construction of phylogenetic tree. The results are encouraging and more promising in larger sets. Further scopes for the development of various algorithm in this line of study for future are also possible.

Keywords: Haemoglobin, PHYLIP, Clustal W, Multiple sequence alignment, Phylogenetics

1. Introduction

Haemoglobins are amongst the popular ubiquitous proteins in a wide variety of organisms which includes bacterias, protozoans, fungi, plants, and animals (Dordas et al 2003 & Vinogradov et al 1993). The most commonly known involvement of Haemoglobins in each of the mentioned variety of organisms confine functions such as the reversible binding of gaseous ligands along with the ability to bind other cellular molecules. The co-evolution of various regulated and contemporary functions within a very old gene family gives us a scope to examine these types of biologically relevant sequence data that play a

very crucial role for a large variety of tissues. These intricate genetic relationships can be depicted with the help of representative evolutionary tree called the tree of life. The tree of life represents the phylogeny of all organisms, which is actually the representation of the journey of organism's lineage and changes that occurred with time. Under the able umbrella and sponsorship of the National Science Foundation (NSF) Assembling the Tree of Life (ATOL) initiative many prolific projects are taking place. Over the time, organisms have evolved and acquired various changes from ancestors. These acquired new forms also preserved many of their ancestral features which are beneficial for the organisms in longer run for adjustment, adaptability and survival. In such situation phylogenetic studies can be very useful to analyse the similarities and differences among species (Tree of Life Web Project, 2008, http://tolweb.org). Various techniques are known for the preparation of phylogenetic tree and most of them use the aligned genetic sequence for this work. ClustalW can be considered as the most popular genetic sequence alignment. Largely speaking, ClustalW has gained a wide popularity and good success instead of being highly sensitive to very divergent sequences in its domain. Thus, the core of the project remains the modification of the ClustalW sequence alignment algorithm in order to achieve higher accuracy for constructing the trees where highly divergent sequences are present.

From the discovery and study of haemoglobins in almost all kingdoms of organisms it is apparent that the ancestral gene for haemoglobin is ancient, (Antoine & Niessing, 1984) and that haemoglobins besides performing the very essential function of transporting oxygen between tissues can also be seen engaged in other important activities which include intracellular oxygen transport and catalysis of redox reactions (Andersson et al, 1996). Haemoglobin is an ancient class of oxygen- carrying molecules that is found in the red blood cells and uniting almost all forms of life on this planet. The molecule haemoglobin derives its name from the globular structure in association with the prosthetic group haeme which binds oxygen. Like all other proteins, haemoglobins are also constituted from the building blocks of amino acids. Thus, it is quite surprising that non-blooded plants also encode haemoglobin (Gupta et al., 2011) and even more of an astonishment remains the fact that they are ubiquitous in nature and in responsibility of a number of functions in the animal world. Bacteria, fungi, protozoans and plants all use haemoglobin in addition to animals. This discovery of the presence of haemoglobin in all the kingdoms further ascertain the fact that the ancestral gene for haemoglobin is ancient. Also, the haemoglobins can perform additional functions besides oxygen transport as do fungi, protists and bacteria. ranging from intracellular oxygen transport to catalysis of redox reactions (Hardison, 1998). Apart from oxygen, nitric oxide (NO), carbon monoxide (CO), hydrogen sulphide(H2S) are among the other ligand that the haemoglobin binds. It also binds certain organic molecules such as several membrane lipids (Gupta et al., 2011). The haemoglobin concentration in the RBCs is in the order of ten milli molars while it is in the sub milli molar range when its about the skeletal muscle's myoglobin. The haemoglobin in plants was studied within the symbiotic nitrogen-fixing root nodules. There the leghaemoglobin (legHb) and play a crucial role transporting free oxygen away from the oxygen-sensitive-nitrogenase enzyme. In the nodules, the range of concentration of legHb may reach up to 0.7mM which is the main reason behind their characteristic red appearance. For the plants producing haemoglobins having a non-symbiotic role, the concentrations are usually in the range of 5 20 μM upon induction, which is very low for showing red colour in case of plants. (Gupta et al., 2011). Haemoglobins were first identified in the nitrogen fixing plant species. Further studies tell us that yet another class of haemoglobins by name non-symbiotic haemoglobins were present throughout the plant kingdom. This class of haemoglobins were expressed themselves differentially during the development of plants. From the limited available data, it can be ascertained that, the non-symbiotic haemoglobins were involved in hypoxic stress and oversupply of nutrients. From the recent studies, the non-symbiotic haemoglobin genes were identified from both nitrogen and non-nitrogen fixing dicot and from monocot species (Reddy, 2007). Thus, it can be said that among the plant kingdom, two different types of haemoglobin have been discovered,

a type which is widely distributed and perhaps ubiquitous among species known as the non-symbiotic type and a type that is induced upon nodulation, known as symbiotic type. (Hardison, 1998).

Phylogenetics

Phylogenetics is a very popular domain of research which deals with organisms and their genetic relationship. Earlier, phylogenetics used morphological features such as size, colour, fur, or other physical characteristics for determining the relations between various organisms. But, the modern phylogenetics is mostly dependent on the informations extracted from genetic materials - DNA, RNA or protein sequences (Shamir, 2001). Multiple Sequence Alignment is a way of arranging the sequences of DNA, RNA, or proteins so as to distinguish regions of similarity.

A sequence alignment of biological sequences, such as those of proteins, DNA or RNA, is called a multiple sequence alignment (MSA). The set of sequences is usually assumed to have an evolutionary link, meaning that they are all descended from a common ancestor. These areas can represent structural, functional or evolutionary connections between the sequences. A degree of evolutionary change between sequences descended from a common ancestor can be reflected in alignments. Sequence alignments and phylogenies are related (Felsenstein, 2004). If one employs a parsimony method and a certain scoring scheme, a dynamic programming technique can be employed to identify the globally optimal alignment. There is no universally accepted scoring scheme yet. Since it is a NP complete problem, this method is computationally costly and unfeasible (Bonizzoni &Gianluca 2001). Instead, the multiple sequence alignment is carried out using heuristics. This study concentrated on the heuristic technique known as progressive alignment. ClustalW is a popular program that uses a progressive alignment technique.

Clustal W

Clustal W is a widely used program for the multiple sequence alignment and for the preparation of phylogenetic trees. The reason for its widespread users is its portability amongst various computing platforms. Due to its wide popularity, user friendliness and the availability of source code, ClustalW was chosen for the project. For the purpose of performing multiple sequence alignment Clustal W uses progressive alignment algorithm which can be broken down into three major steps. Firstly, before calculating a distance matrix that indicates each pair of sequences divergence, each pair of sequences is first aligned independently. Two gap penalties – one for opening a gap and another for extending a gap – are used to calculate a full programming alignment. The distance matrix score is calculated by dividing the number of residues compared, excluding gap sites, by the number of identities in the optimal alignment. To get a value between 0 and 1.0, that number is then multiplied by 100 and subtracted from 1. Next step includes the calculation of a guide tree. It will be used to direct the final multiple alignment procedure. A Neighbour-Joining clustering algorithm and the distance matrix from the first stage are used to calculate this tree. Each sequence is also given a weight based on how far apart they are. Lastly, all the sequences are aligned progressively in accordance to the branching order that is found in the guide tree. To align many larger groups of sequences a series of pairwise alignment is used. From the tips of the rooted tree, it is proceeded towards the root. At each alignment a full dynamic programming algorithm issued with penalties for opening and extending gaps. Each step aligns two existing alignments or sequences. Gaps that are present in the older alignments stay in place. After having considered all the sequences, a final alignment is produced. The final alignment so produced can be then used for the purpose of constructing a phylogenetic tree for those species (Thompson et al 1994).

A major disadvantage of this approach of progressive alignment is that, once an alignment has been performed involving some of the species, this alignment is never reconsidered despite what other decisions are made for the remaining species. This can result into inaccuracies in the final alignment (Felsenstein, 2004) with the root of the tree. Whereas in the original Clustal W progressive alignment algorithm, all sequences were equally weighted.

2. Methodology

Materials

The dataset considered for the purpose of this experiment consisted of various protein sequences. Proteins are one of the main constituent materials in organism which is found in eukaryotic as well as prokaryotic organisms.

Protein sequences were given the preference based on their size. Generally, they contain only 150-500 amino acids whereas DNA sequence contains 16,000 to 20,000 base pairs and when it comes to the full genome of human it has approximately around 3 billion base pairs [13]. Since aligning the entire genome sequences and DNA sequences are practically not that feasible, therefore the protein sequences taken into account were taken from a database of sequences hosted by the http://www.ncbi.nlm.nih.gov/protein of NCBI protein data base. Thereafter all these sequences were converted into the FASTA format. This format can be used as input in Clustal W (Notredame et al, 2000).

Table 1. Table showing the species that were used in this study

	Animal Haemoglobin sequences
1	>gi|4504345|ref|NP_000508.1| haemoglobin subunit alpha [*Homo sapiens*]
2	>gi|183830|gb|AAA52634.1| beta haemoglobin [*Homo sapiens*]
3	>gi|388452798|ref|NP_001253705.1| haemoglobin subunit alpha [*Macaca mulatta*]
4	>gi|242878266|gb|ACS94046.1| beta haemoglobin [*Macaca mulatta*]
	Plant: Symbiotic haemoglobin sequences
5	>gi|2921626|gb|AAC04853.1| leghaemoglobin [*Lupinus luteus*]
6	>gi|169351|gb|AAA33767.1| leghaemoglobin [*Phaseolus vulgaris*]
7	>gi|169882|gb|AAA03002.1| leghaemoglobin, partial [*Sesbania rostrata*]
8	>gi|3980177|emb|CAA38024.1| leghaemoglobin [*Medicago sativa*]
9	>gi|77994689|gb|ABB13622.1| leghaemoglobin [*Astragalussinicus*]
1o	>gi|166402|gb|AAB41521.1| leghaemoglobin [*Medicago sativa*]
11	>gi|414378|gb|AAC60563.1| leghaemoglobin [*Psophocarpus tetragonolobus*]
12	>gi|2842550|dbj|BAA24685.1| leghaemoglobin [Pisum sativum]

	Plant: Non Symbiotic haemoglobin sequences
13	>gi|377643998|gb|AFB70892.1| non-symbiotic haemoglobin [*Vigna radiata*]
14	>gi|11095158|gb|AAG29748.1|AF172172_1 non-symbiotic haemoglobin [*Medicago sativa*]
15	>gi|657388444|gb|KEH30379.1| Non-symbiotic haemoglobin [*Medicago truncatula*]
16	>gi|474151151|gb|EMS56981.1| Non-symbiotic haemoglobin [*Triticum urartu*]
17	>gi|971516029|dbj|GAQ82196.1|Non-symbiotic haemoglobin [*Klebsormidium flaccidum*]
18	>gi|162462053|ref|NP_001104966.1| haemoglobin [*Zea mays*]
	Haemoglobin sequence of Bacteria, fungi
19	>gi|380471629|emb|CCF47182.1| haemoglobin [*Colletotrichum higginsianum*]
20	>gi|130377768|emb|CAA48729.2| haemoglobin [*Pichia norvegensis*]
21	>gi|441429904|gb|ELR67355.1| haemoglobin [*Photobacterium marinum*]
22	>gi|393182089|gb|EJC82128.1| haemoglobin [*Rhizobium leguminosarum bv. trifolii* WSM2297]

Methods

Step1: Acquiring the protein sequence

For the present study 22 protein sequence were chosen from the link http://www.ncbi.nlm.nih.gov/protein of NCBI protein data base

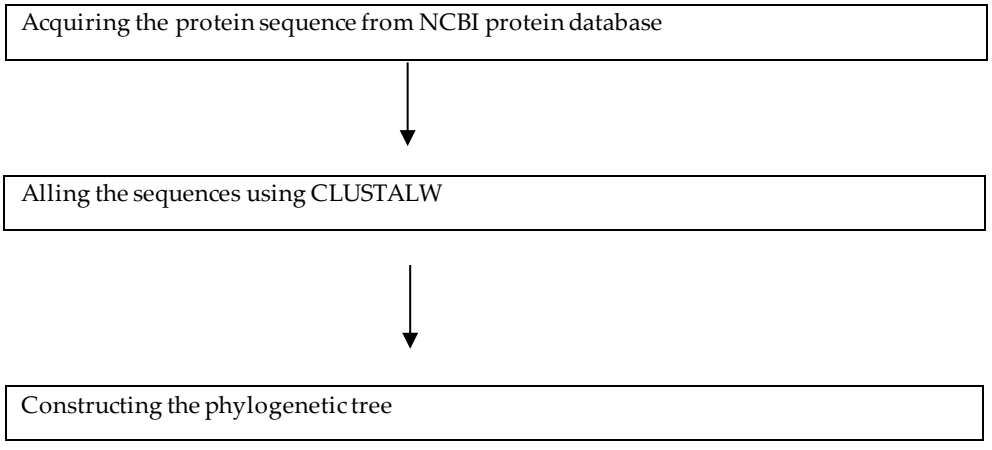
Step2:

The sequence are aligned using the multi-alignment feature of CLUSTALW with maximum fixed-gap and gap penalties to reduce the influence of gaps in the subsequent analysis

Step3:

The phylogenetic tree was constructed using the parsimony method (prot pars).

Flow Chart for Phylogenetic Tree Construction



3. Result and Discussion

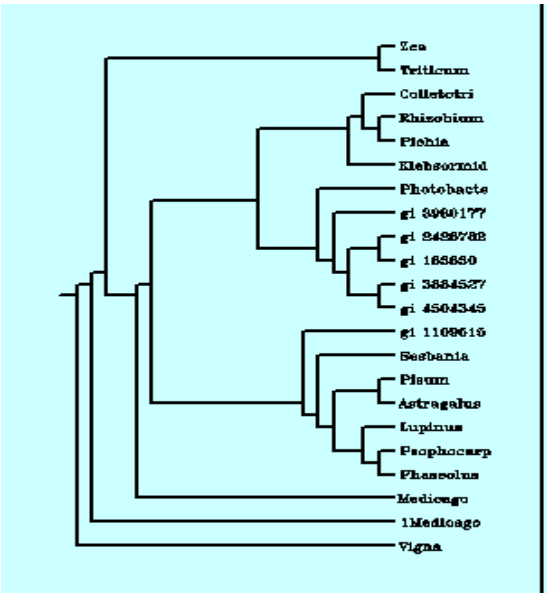


Fig 1: Molecular Phylogenetic analysis of prokaryotes and eukaryotes using Phylip

The list of the sequences depicted in Table 1 were taken from the Protein Database of NCBI. Then with the help of CLUSTAL W method of PHYLIP version the Multiple sequence alignment was performed. Using the Parsimony method, the molecular phylogenetic analyses of alignments of haemoglobins (Symbiotic and non-symbiotic) of selected plant with representative microbial and animal haemoglobins were done.

From the phylogenetic tree in the Fig. 1 the following inferences can be made. A separated clade was formed by the haemoglobin of plants (symbiotic and non-symbiotic), which in turn gave rise to two major Nested clades. One nested clade of plants were constituted by all the plants containing class I haemoglobin. On the other hand, Class II plant haemoglobin or the Leghaemoglobin of all 8 plants along with Class II haemoglobin of 2 non-symbiotic plants i.e symbiotic haemoglobins formed another nested clade. Although inside a common clade of the plant haemoglobin, the Class I and Class II haemoglobin of plants formed two nested clade. As expected, Class I and Class II haemoglobins of plants formed two nested clades. Such results can be for the differences in their functions (Gupta et al., 2011). The main function of Class I haemoglobins is the scavenging of nitric oxide (NO) at very low oxygen levels. In addition to that they have an extremely high affinity to oxygen. When it comes to the Class II haemoglobins, they have a lower affinity to oxygen then Class I and they are involved in the facilitation of the oxygen supply to developing tissues. Symbiotic haemoglobins in nodules have mostly evolved from class II haemoglobins. Thus, they can be considered as a subclass of Class II haemoglobins of plants. The haemoglobin subunits of animal altogether formed a separate clade

with alpha and beta globins forming the nested clades. (Goodman et al. 1987). He stated that the sequences of amino acids of the alpha and beta globins were nearly 50% identical, irrespective of its source of origin from whichever vertebrates. This indicates that these two genes would have come from a common ancestor about 450 million years ago, in the ancestral jawed vertebrate. According to the tree, haemoglobin sequences of Microbes form the outgroup.

The importance of this phylogenetic analysis lays in the fact that it proves the common ancestry of Haemoglobins of all organisms. This result in a different way further establishes the taxonomic classification. A clear distinction of Kingdoms is visible and they together constitute the main clades. The subsequent nested clades formed were according to the evolutionary order and functional differences.

Clustal W already employed a few features to address the divergent sequences. Firstly, it delays the alignment of divergent sequences until the more similar sequences are aligned. This may be crucial in placing within the alignment gaps aptly. The second one is that of sequence weights, which are calculated directly from the guide tree. Closely related sequences will receive low weights and highly divergent sequences will receive high weights. During the final alignment steps, these weights are used for scoring. The underlying objective of the step is to minimise the scoring bias from very similar sequences (Reddy, 2007). A major issue with this approach is the dependence of these weights on the guide tree. So, if the clustering algorithm provides bad results, then the guide tree could calculate incorrect weights. The results presented here cannot be considered as definitive as a larger test set is required but the improving and developing trend is undeniable and paves a path for the further investigations.

Apart from the Clustal W other progressive alignment programs are also available. Further works can be carried out to compare and analyse the result obtained from Clustal W with other programs such as TCoffeeto to look at and address the types of differences that exist (Notredame et al 2000). This approach can further be crucial in potentially determining whether a program is better suited for a specific type of dataset. Apart from the progressive alignment method, there are available other approaches in order to solve the multiple sequence alignment problem. Genetic algorithms, hidden Markov models and iterative methods are among the few currently employed methods used to try and find better alignments. Research in the coming days may include a thorough study of all the methods and underline their pros and cons with programs like ClustalW.

Another essential factor includes the software that is employed to infer the phylogenetic trees. Various methods are now-a-days available to construct the tree based on the multiple sequence alignment. Further refinement and required modifications in these methods could give us more promising results as well. But as these trees are based on genetic data there are important limitations to consider since there is still a lot that remains to be known about genetic sequences. As we advance through the better

understandings of the genetics behind different sequences, we can further apply them to the field of phylogenetics for obtaining better results. (Shamir 2001).

4. Conclusion

The results are progressive and encourages for testing it on larger test sets. But, like the other presently operative methods for the construction of evolutionary trees this method also does not guarantees that the correct phylogenetic tree will be produced. For better results user should develop more expertise on this domain so that they can interpret the results and adjust parameters within the program to get the best phylogenetic tree. Also, further development in the programs for alignment would open new doors in the near future.

References

1.Dordas, C., Rivoal, J. and Hill, R.D. 2003 Plant haemoglobins, nitric oxide and hypoxic stress. Ann. Bot.91, 173178.

2 Vinogradov SN, Walz DA, Pohajdak B, Moens L, Kapp OH,Suzuki T, Trotman CN (1993) Adventitious variability? The amino acid sequences of nonvertebrate globins. CompBiochemPhysiol (b) 106: 1–26.

3. Tree of Life Project. "What is Phylogeny?" Tree of Life Web Project. 6 May 2008 <http://tolweb.org>.

4. Andersson, C. R., Jensen, E. O., Llewellyn, D. J., Dennis, E. S., Peacock, W. J. (1996). A new hemoglobin gene from soybean:A role for hemoglobin in all plants. Proc. natn.Acad. Sci. U.S.A.93, 5682–5687.

5.Antoine, M. & Niessing, J. (1984). Intron-less globin genes in the insect Chironomus thummi. Nature 310, 795–798.

6.Gupta, K. J., Hebelstrup, K. H., Mur, L. A. and Igamberdiev, A.U. 2011b. Plant hemoglobins: important players at the crossroads between oxygen and nitric oxide. FEBS Letters 585:3843-3849.

7.Hardison, R. C. 1998. Hemoglobins from bacteria to man: evolution of different patterns of gene expression J. Exp. Biol., 201, pp. 1099 -1117.

8.Reddy, D. M. R. 2007. Evolutionary trace analysis of plant haemoglobins: implications for site-directed mutagenesis. Bioinformation 1(9):370-375.

9. Shamir, Ron. "Algorithms in Molecular Biology" Tel Aviv University School of Computer Science. Fall 2001.Tel Aviv University.

10. Felsenstein, Joseph. Inferring Phylogenies. Sunderland, Massachusetts: Sinauer Associates, Inc., 2004. 496 -520.

11. Bonizzoni, Paola, and Gianluca Della Vedova. "The Complexity of Multiple Sequence Alignment with SP-Score That is a Metric." Theoretical Computer Science. 259 (2001): 63-79.

12. Thompson, Julie D., Desmond G. Higgins, and Toby J. Gibson. "CLUSTALW: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position Specific Gap Penalties and Weight Matrix Choice." Nucleic Acids Research 22 (1994): 4673-4680.

13. International Human Genome Sequencing Consortium. "Finishing the Euchromatic Sequence of the Human Genome." Nature 431 (2004): 931-945.

14. Notredame, Cedric, Desmond G. Higgins, and Jaap Heringa. "T-Coffee: A novel method for fast and accurate multiple sequence alignment." J Mol Biol302 (2000): 205-217.

15. Goodman, M., Czelusniak, J., Koop, B. F., Tagle, D. A., & Slightom, J. L. (1987, January). Globins: a case study in molecular phylogeny. In Cold Spring Harbor symposia on quantitative biology (Vol. 52, pp. 875-890). Cold Spring Harbor Laboratory Press.