International Journal of Engineering & Technology, 7 (2.31) (2018) 13-15

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Organizing hematologic gene sequence data using neighbor joining phylogeny method

B.J. Bipin Nair[1*]

[1]Dept. of Computer Science, Amrita Vishwa Vidyapeetham, Mysuru Campus, Karnataka.
**Corresponding author E-mail:bipin.bj.nair@gmail.com*

Abstract

There are a number of methods to spot orthologous genes from homologous genes. Since identifying orthologous genes are main problem and play a major role in the hematologic genetic disorders. In this paper, we propose different approaches to discover orthologs of homologous hereditary diseases of the hematological system and find the evolutionary relationship between hereditary genetic diseases of the hematological system by adjoining the joining method of contiguous merging trees.

Keywords: *Paralogous gene, orhtologous gene, neighbor joining tree, neighbor joining method.*

1. Introduction

Genes are hereditary materials that are transferred from one generation to other. Genes are formed of DNA, and are found in proteins. Homology forms the basis of biology, Homologous are also called as homolog. Homolog is used to refer to both the genes and the proteins. Homologous genes are genes that share mutual ancestors. Homologous genes are further divided into two as paralogous genes and orthologous genes. Orthologous genes are homologous genes which are emerged during speciation event. Paralogs genes are homologous genes which are emerged during gene duplication. DNA is often subjected to mutations. These mutations can lead to missing or malfunction of proteins and then that can lead to genetic disorders. Mutations can occur due to many reasons. Mutations can be inherited from the parents and are called as germ-line mutations. Mutations may a occur at least once during one's life time. Some mutations are very bad. Some are beneficial too. Genetic mutations create genetic diversity and cause genetic disorder. Blood is very important factor of one's life. Anything that affects the blood is called hematologic disorder. There are many blood disorders which are inherited. These hematologic disorders are recessive. We found that orthologs are the cause of blood diseases.

Various works have been done in the orthologous genes. But no one has attempted in relating the orthologous genes and hematologic genetic syndromes. Here we introduce a novel approach to find the evolutionary relationship between different hematologic genetic disorders, and organizing hematologic genetic disorders based on their relationship. Using Neighbor joining tree we provide a better visualization of hematologic blood disorders.

2. Overview

The research work is focusing mainly on two fields of bioinformatics. They are orthologous genes and hematological genetic disorders. There is a great relationship between these two fields. We found that orthologs are a major contributor to hereditary genetic diseases. Here we propose a new technique to organize the hematologic genetic disorders based on the evolutionary relationship through Neighbor joining tree using Neighbor joining method.

3. Problem statement

DNA is often subjected to mutation, which unintentionally changes the sequence. These mutations can lead to diseases. They can be genetically mutated and get transferred from one generation to the other. There are many hematologic disorders which are genetically inherited, and cause hematologic genetic disorders. Orthologous genes plays very important role in hematologic genetic disorders. Till date no work is attempted to find the orthologous genes from the hematologic blood protein, find the evolutionary relationship between different hematologic genetic disorders and how one blood protein sequence differ from other.

4. Problem formulation

The proposed work is focused on finding the evolutionary relationship between different hematologic disorders through Neighbor joining tree using Neighbor joining method

The Proposed Work Performs,

- Gathering of different hematological blood protein sequences.
- Finding the orthologous gene from the hematological blood protein.
- Classify each disorder.
- Perform Neighbor Joining method.
- Construct Neighbor joining tree.

Neighbor joining method

Adjacency is a clustering method used to reconstruct the phylogenetic tree. Computer simulation studies have revealed its computational speed and inferred high accuracy of phylogeny and are therefore widely used. Most technical studies have quantified the overall performance of neighbors joining methods as far as the share of branches that are reasonably inferred or the percentage of replications that the correct tree is recovered from. The input here is "n" taxa. The output appears as a rooted tree with branch length.

Neighbor Joining Method Pseudo Code

Step 1**:** Construct Neighbor joining matrix D* form D
Step 2: Find minimum element $D^*_{x,y}$ of D*
Step 3: Compute
$\Delta_{x,y}$ = Total Distance $_{D}{}^{(x)}$ – Total Distance $_{D}{}^{(y)}$ / (n- 2)
Step 4: Remove x^{th} and y^{th} row / column from D and add an z^{th} row / column such that for any k,
$D_{k,z=}$ ($D_{k,x}$ + $D_{k,y}$ – $D_{x,y/}$) / 2
Given distance matrix D as the matrix D* whose x, y^{th} entry is given by,
$D^*_{x,\ y=}$ (n-2)* $D_{x,y}$ –Total Distance $_{D}{}^{(x)}$ –Total Distance $_{D}{}^{(y)}$
Where Total Distance$_{D}{}^{(x)}$ is the sum of distance from x to all other leaves.
Formula to calculate the distance of new node
d (g, k)=½[d(a, k) + d(b, k) – d(a, b)]

Neighbor joining tree construction

Neighbor joining tree is a phylogenetic tree that shows the evolutionary relationship among entities that share a common ancestor. Here the construction of tree is based on the sequence distance between each blood disorders. A method known as Neighbor joining method is proposed for the reconstruction of Neighbor joining tree. Here the tree shows the evolutionary relationship of blood disorders.

Dataset

Initially the dataset is Hematologic blood protein sequence. So through Neighbor Joining method we will get the evolutionary relationship among the blood proteins. Through that we can construct the neighbor joining tree. The protein sequence will change according to the disorders. Below Fig 1 we can see some sample dataset for hematologic blood protein sequence

```
Myl
Myl,MYSNVIGTVTSGKRKVYLLSLLLIGFWDCVTCHGSPVDICTAKPRDIPMNPMCIYRSPEKK
ATEDEGSEQKIPEATNRRVWELSKANSRFATTFYQHLADSKNDNDNIFLSPLSISTAFAMTKLGA
CNDTLQQLMEVFKFDTISEKTSDQIHFFFAKLNCRLYRKANKSSKLVSANRLFGDKSLTFNETYQ
DISELVYGAKLQPLDFKENAEQSRAAINKWVSNKTEGRIDDVIPSEAINELTVLVLVNTIYFKGL
WKSKFSPENTRKELFYKADGESCSASMMYQEGKFRYRRVAEGTQVLELPFKGDDITMVLILPKPE
KSLAKVEKELTPEVLQEWLDELEEMMLVVHMPRFRIEDGFSLKEQLQDMGLVDLFSPEKSKLPGI
VAEGRDDLYVSDAFHKAFLEVNEEGSEAAASTAVVIAGRSLNPNRVTFKANRPFLVFIREVPLNT
IIFMGRVANPCVK
Hemo
Hemo,MRLPQFVWTLGLLLPLLSTSQAFKDICNAKPKDVPLEPRCIYRSPEDEAPTGDAIPEKVP
ENTNPRVWELSKANSRFALSLFKQLAQGKPSESNIFMSPISISSAFAMTKLGACNNTLKQIMNVF
EFDTIKEKTSDQVHFFFAKLNCRLYRKDKTPELISANRLFGEKSLAFNEIYQNISELVYGAKLM
PLNFKEKFELSRVTINFWIANKTENRIQNTLPKDSLNSNTVLVLVNTIYFKGQWKSKFPKNNVFK
ADFYVSKSKTCPVSMMYQETKFHYGRFTEDKVQVLELPYRGDDITMVLILPLKDTPLSEVEENLD
LKKLIGWLHNMRETSVSVHLPRFRIEDSFSLKEKLQAMGLEDLFSPKDASLPGILEDEANYLYIS
DAFHKAFLEVNEEGSEAAAATAVMAVGRSINSNREMFVANKPFLLLIRESTINTMVFPGRVADPC
DP
Lym
Lym,RDIPVNPICIYRNPEKKPQERRGAGAGEGQDPGVHKPPVWELSRANSRFAVVFYKHLADSK
DNEENIFLSPLSISTAFAMTKLGACGDTLQQLMEVFQFDTISEKTSDQVHFFFAKLNCRLYKKAN
KSSELISANRLFGEKSLVFNETYQNISEIVYGAKLWPLNFKEKPELSRKIINEWVANKTERRITE
VIPEKGIDDLTVLVLVNTIYFKGHWKSQFPAPNTRLDLFHKANGETCNVPIMYQESRFPYAFIQE
DKVQVLELPYKGDDITMVLVLPKAGTPLVVEVERDLTDDKLQNWISRMEVRLVSFFRPTVEDSF
SVKEKLRKMGLEDLFSPENAKLPGIVAGDRTDLYVSEAFHKAFLEVNEEGSEASAATAVVISGRS
FPMNRIIFEANRPFLLFIREATLNTIIFMGRISDPCS

Ane,MYSPGAGSGAAGERKLCLLSLLLIGALGCAICHGNPVDDICIAKPRDIPVNPLCIYRSPGK
KATEEDGSEQKVPEATNRRVWELSKANSRFATNFYQHLADSKNDNDNIFLSPLSISTAFAMTKLG
ACNDTLKQLMEVFKFDTISEKTSDQIHFFFAKLNCRLYRKANKSSDLVSANRLFGDKSLTFNESY
```

Fig. 1

5. Result

D =

	Leukemia	Anemia	Lymphoma	Hemophilia	Myeloid
Leukemia	0	13	21	22	10
Anemia	13	0	12	13	12
Lymphoma	21	12	0	13	14
Hemophilia	22	13	13	0	15
Myeloid	10	12	14	15	0

Fig. 2

- This is the initial matrix D. It contains the sequence distance among different Hematologic disorders.

D* =

	Leukemia	Anemia	Lymphoma	Hemophilia	Myeloid
Leukemia	0	-77	-63	-63	-87
Anemia	-77	0	-74	-74	-65
Lymphoma	-63	-74	0	-84	-69
Hemophilia	-63	-74	-84	0	-69
Myeloid	-87	-65	-69	-69	0

Fig. 3

- Construct Neighbor joining matrix D* from D
- Find the minimum element from D*
- Combine the row and column of the minimum element and construct the new matrix.

D` =

	Leuk-My	Anemia	Lymphoma	Hemophilia
Leuk-My	0	7.5	12.5	13.5
Anemia	7.5	0	12	13
Lymphoma	12.5	12	0	13
Hemophilia	13.5	13	13	0

Fig. 4

- The new Matrix D`

D`* =

	Leuk-My	Anemia	Lymphoma	Hemophilia
Leuk-My	0	-51	-46	-46
Anemia	-51	0	-46	-46
Lymphoma	-46	-46	0	-51
Hemophilia	-46	-46	-51	0

Fig. 5

- Construct Neighbor joining matrix D`* from D`
- Find the minimum element from D`*
- Combine the row and column of the minimum element and construct the new matrix.

D`` =

	Leuk-My-An	Lymphoma	Hemophilia
Leuk-My-An	0	8.5	9.5
Lymphoma	8.5	0	13
Hemophilia	9.5	13	0

Fig. 6

- The new Matrix D``

D``* =

	Leuk-My-An	Lymphoma	Hemophilia
Leuk-My-An	0	-31	-31
Lymphoma	-31	0	-31
Hemophilia	-31	-31	0

Fig. 7

- Construct Neighbor joining matrix D``* from D``
- Find the minimum element from D``*
- Combine the row and column of the minimum element and construct the new matrix.

D``` =

	Leuk-My-An	Lym-Hemo
Leuk-My-An	0	2.5
Lym-Hemo	2.5	0

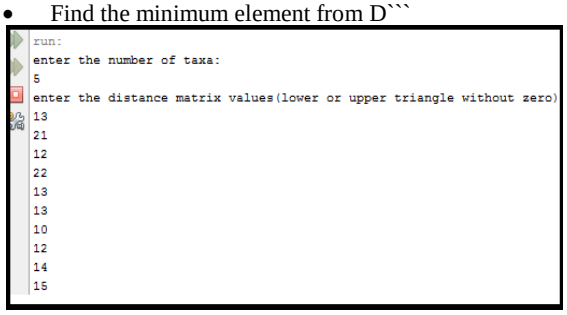
Fig. 8

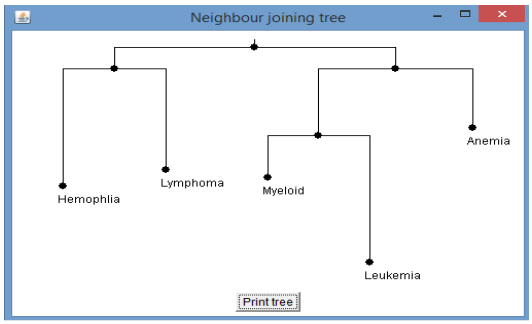
- The new matrix D```

	Leuk-My-An	Lym-Hemo
Leuk-My-An	0	-5
Lym-Hemo	-5	0

Fig. 9

- Construct Neighbor joining matrix D```* from D```
- Find the minimum element from D```





6. Conclusion

Here we show the evolutionary relationship among various hematologic genetic disorders. Here initially we have the sequence similarity distance among various hematologic genetic disorders like Myeloid, Hemophilia, Lymphoma, Anemia and Leukemia. So after the Neighbor joining method we can generate the Neighbor joining tree showing the evolutionary relationship between those disorders.

References

[1] Wong KC & Zhang Z, "SNPdryad: predicting deleterious non-synonymous human SNPs using only orthologous protein sequences", *Bioinformatics*, (2014).

[2] Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F & Perrière, G, "Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases", *Bioinformatics*, Vol.21, No.11,(2005), pp.2596-2603.

[3] Dessimoz C, Gabaldón T, Roos DS, Sonnhammer EL & Herrero J, "Toward community standards in the quest for orthologs", *Bioinformatics*, Vol.28, No.6,(2012), pp.900-904.

[4] Chen R & Jeong SS, "Functional prediction: identification of protein orthologs and paralogs", *Protein Science*, Vol.9, No.12, (2000), pp.2344-2353.

[5] Cannon SB & Young ND, "OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies", *BMC bioinformatics*, Vol.4, No.1,(2003).

[6] Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM & DeSalle R "OrthologID: automation of genome-scale ortholog identification within a parsimony framework", *Bioinformatics*, Vol.22, No.6,(2006), pp.699-707.

[7] Sonnhammer EL, Gabaldón T, da Silva AWS, Martin M, Robinson-Rechavi M, Boeckmann B & Dessimoz C, "Big data and other challenges in the quest for orthologs", *Bioinformatics* , (2014).

[8] Li YI & Copley RR, "Scaffolding low quality genomes using orthologous protein sequences", *Bioinformatics*, Vol.29, No.2,(2013), pp.160-165.

[9] Kim K, Kim W & Kim S, "ReMark: an automatic program for clustering orthologs flexibly combining a Recursive and a Markov clustering algorithms", *Bioinformatics*, Vol.27, No.12,(2011), pp.1731-1733.

[10] Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC & Sali, A, "Evolutionary constraints on structural similarity in orthologs and paralogs", *Protein Science*, Vol.18, No.6,(2009), pp.1306-1315.

[11] Mahmood K, Konagurthu AS, Song J, Buckle AM, Webb GI & Whisstock JC, "EGM: encapsulated gene-by-gene matching to identify gene orthologs and homologous segments in genomes", *Bioinformatics*, Vol.26, No.17,(2010), pp.2076-2084.

[12] Forslund K & Sonnhammer EL, "Benchmarking homology detection procedures with low complexity filters", *Bioinformatics*, Vol.25, No.19,(2009), pp.2500-2505.

[13] Yosef N, Sharan R & Noble WS, "Improved network-based identification of protein orthologs", *Bioinformatics*, Vol.24, No.16, (2008), pp.i200-i206.

[14] Storm CE & Sonnhammer EL, "Automated Ortholog inference from phylogenetic trees and Calculatio of orthologyreliability", *Bioinformatics*, Vol.18, No.1, (2002), pp.92-99.

[15] Alexeyenko A, Tamas I, Liu G & Sonnhammer EL, "Automatic clustering of orthologs and Inparalo shared by multiple proteomes", *Bioinformatics*, Vol.22, No.14,(2006), pp.e9-e15.

[16] DeLuca TF, Wu IH, Pu J, Monaghan T, Peshkin L, Singh S & Wall DP, "Roundup: a multi-genome repository of orthologs and evolutionary distances", *Bioinformatics*, Vol.22, No.16, (2006), pp.2044-2046.

[17] Moreno-Hagelsieb G & Latimer K, "Choosing BLAST options for better detection of orthologs as reciprocal best hits", *Bioinformatics*, Vol.24, No.3,(2008), pp.319-324.

[18] Arvestad L, Berglund AC, Lagergren J & Sennblad B, "Bayesian gene/species tree reconciliation and orthology analysis using MCMC", *Bioinformatics*, Vol.19, (2003), pp.i7-i15.

[19] Bipin Nair BJ, Sujith M & Alphonsa MV, "Self-regulating Exploration for Orthologous in Homologous Hematologic Gene Sequence Data Using UPGMA Method", *IJET*, Vol.8, No.1, (2016), pp.287-292.