Unique and Universal Proteins in Human Genome

Essam Al-Daoud

Computer Science Department, Faculty of Information Technology, Zarqa University, Zarqa, Jordan,

Abstract: *One of the major troubles with a comparative analysis between human and other species is that only similar amino acid sequences are selected for analysis. To find the relationship between the species and discover the unique, the common and the universal proteins, the whole genome of 40 species are compared with the human genome which is used as reference genome. More than 11 billion pairwise alignments are performed using blastp. Several findings are introduced in this study, for example, we found 330 unique proteins in human genome and have insignificant hits in all tested genomes, the number of universal proteins in human genome and conserved in all tested species is 82, and there are 180 proteins common in vertebrates genomes, but have insignificant hits in the other tested species. In contrary to the previous studies which use selected set of the genes and do not consider the whole genomes, this study proves that the similarity between human and chimpanzee is only 94.8.*

Keywords: *Genome, Species, blastp, unique protein, universal protein.*

1. Introduction

The previous two decades have seen a blast of the hereditary information. Countless DNA sequences and genotypes have been produced, and they have prompted noteworthy biomedical advances and provided new insights into biology [1]. In addition, these information have significantly expanded our comprehension of patterns of hereditary variety among individuals and populations [2]. Interpreting of a given genomic sequence is one of the focal difficulties of science today. Maybe the most encouraging way to deal with this problem is based on the pairwise alignment and multiple sequences alignment methods. For example, protein-coding subsequences tend to be conserved between species. Subsequently, a straightforward strategy for recognizing an functional exon is to look for its homologue from related species using the whole genome alignment. Hence, enthusiasm for quicker, estimated, or heuristic (instead of ideal) alignment algorithms has increased [3-5]. Two of the most well known heuristic alignment procedures are implemented in the FASTA and BLAST packages. Comparisons of full genome sequences empower scientists to make inquiries that were unthinkable with small subsequences. Large-scale comparisons can uncover the genetic basis of speciation and variation, increase our understanding of the biological processes in living cells, recognize shared biochemical functions, expand our knowledge in human diseases and offer important information about evolutionary histories of extinct and living kinds [6,7]. The whole genome is used in several studies such as utilizing data from one genome to understand another, identifying potential orthologs, comparison of genome content genome alignment and genome signature analysis based on di-nucleotide abundance among others [8-11].

Alignment of genomes implies identify differences that generated from mutational changes. In considering genome modifications, one differentiates between three important evolutionary operations: DNA mutations, genome rearrangements, and content alterations [12,13]. DNA mutations impact on one or few nucleotides, while genome rearrangements work on bigger genomic subsequences and lead to change the orientation and the order of genes. Lastly, content alterations are an outcome of gene losses and duplications. Genome duplication has clearly permitted the development of more complex life forms; it equips an organism with a cornucopia of extra gene copies, which are allowed to change to fill unique needs. While one copy evolved for use in the brain, say, another evolved for use in the liver or adjusted for a novel reason. Therefore, the duplicated genes allow for increased sophistication and complexity [14, 15].

In this study, we used 40 full genomes from 11 organisms to find the relationship between the species and discover the unique, the special, the common and the universal proteins. To trace the genes using top down approach, the human genome is used as reference genome.

2. Data Collection and Preprocessing

To find the distinguished genes and quantify sequence similarities, the full genome of 40 species from 11 organisms are downloaded from KEGG site (Kyoto Encyclopedia of Genes and Genomes http://www.genome.jp/kegg/catalog).

Table 1. The proteins details of bacteria Genomes

ID	Species	# protiens	#AA
1	Cronobact.	3842	1244298
2	Salmonella	4770	1385186
3	Shigella	6409	1293263
4	Enterobact.	4289	1375730
5	Chlamydia	1013	356049
6	Crono_sak	4442	1342730
7	Ecoli	4843	1508759

Table 2. The proteins details of protists, fung and archaea Genomes

ID	Species	# protiens	#AA
8	Entamoeba	8811	3563877
9	Babesia	3706	1856394
10	Plasmodiu	7353	3382406
11	Laccaria	18215	6700944
12	Aspergillus	9541	5067689
13	Neurospora	10813	5632539
14	Pyrococcus	1784	539209
15	Archaeoglo	1823	478828
16	Methanotorr	1772	506747

The species are selected to represent various branches of the phylogenetic tree of life and provide adequate coverage of main kinds within the evolutionary tree, including, seven bacteria, three protists, three fungi, three archaea, seven mammals, three birds, three fishes, five insects, a tick, a mollusk and four plants. Tables 1-6 summarize the name of the selected species, the number of proteins and the average length (number of the amino acid) of each one.

Table 3. The proteins details of mammals Genomes

ID	Species	# protiens	#AA
17	Human	109052	73449745
18	Chimp.	79947	55635610
19	Mouse	76217	52262429
20	Cow	28901	18146954
21	Camel	26729	15276008
22	Elephant	29784	17488002
23	Whale	34821	21600601

Table 4. The proteins details of birds and fishes Genomes

ID	Species	# protiens	#AA
24	Chicken	46346	32575322
25	Falcon	21235	12955188
26	Pigeon	18582	11198213
27	Zebrafish	52829	38449214
28	Platyfish	23478	13384899
29	Coelacant.	34251	20280708

Table 5. The proteins details of insects Genomes

ID	Species	# protiens	#AA
30	Fly	21304	13686004
31	Mosquito	14099	7371687
32	Bee	22451	15287002
33	Ant	10657	6082041
34	Butterfly	15232	6424480

Table 6. The proteins details of a tick, a mollusk and plants Genomes

ID	Species	# protiens	#AA
35	Octopus	23994	13806582
36	BlackTick	20467	5810072
37	ThaleCres	48350	20856276
38	Rice	28555	10301721
29	Wheat	33849	13570085
40	Chl_Rein	14489	6573428

3. Genomes comparisons and mining

Before comparing the human genome with other genomes, the similar proteins in the human genome with hit $<10^{-10}$ is removed. Thus the total number of human proteins is reduced to 16614. To align two proteins, *blastp* is downloaded and called using Matlab as follows:

```
system(['blastp -query human.fa -db sp1 -out
        results.out – evalue .01 –num_alignments
        5']);
```

where *human.fa* is a query that is formatted as fasta file which will be compared with the genome *sp1*. The results are saved as NCBI file for each pair has expectation value < 0.01, and then the results are interpreted and saved as a matrix:

```
M=ParseNCBI('results.out');
```

four important values are extracted for each pair of the compared sequences, the values are the score, the expectation, the percentage of identities and the match:

```
M.Hits(0).HSPs(1).Score;
M.Hits(0).HSPs(1).Expect;
M.Hits(0).HSPs(1).Identities.Percent;
M.Hits(0).HSPs(1).Identities.Match;
```

Algorithm 1 is used to find all universal genes with expectation value less than 10^{-33}:

Algorithm 1: *Universal genes*

```
For each protein in humanGenome j
   For each species i
      If exect(i, j)<1e-33
            count=count+1
   If count = num
      Print j
```

where num is equal to 40 for universal genes, more than 38 for near-universal genes and less than 3 for special and unique genes. Algorithm 2 is used to find the common proteins in one organism but not in the other organisms

Algorithm 2: *Common genes*

```
For each protein in humanGenome j
Flag=1
   For each species i in the target organism
      If exect(i, j)>Expet_value
            Flag=0
   For each species k not in the target organism
      If exect(i, j)< Expet_value
            Flag=0
   If  flag = 1
      Print j
```

Algorithm 3 is used to find the maximum identical protein in a given species

Algorithm 3: *Maximum identical protein*

```
For each protein in humanGenome j
   Max=0
   For each species i
      If Ident(i, j)> max
         Max=iden(i,j)
```

4. Unique and Universal Proteins

Five algorithms are implemented using *Matlab* and the package *Blastp*, where the human genome is used as reference genome, the implemented algorithms are to compare the proteins, interpret the results, find the common, the universal and maximum identical proteins. Human genome contains 16614 proteins, while the Chimpanzee genome contains 79947 proteins. Hence, to compare the both genomes, we have to implement 16614 × 79947 pairwise alignments, which took 25.6 hours using 2.3 GHz dual-core CPU. To mine all the selected genomes, more than 11 billion pairwise alignments are implemented and took about 36 days. Fig. 1 shows the score of first 100 proteins of human genome after aligning it to Chimpanzee and wheat genome, which illustrates the relationship between the both species and human proteins

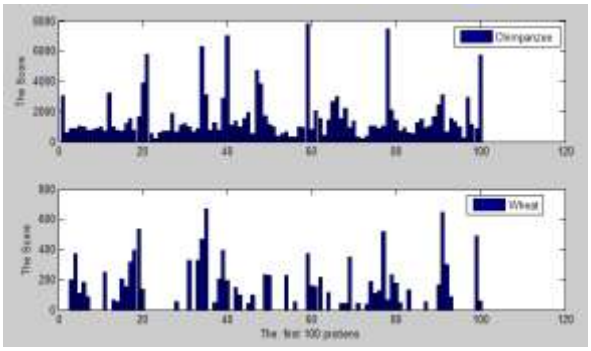

Fig. 1. The score of first 100 proteins for Chimpanzee (top) and wheat genome.

The following are some important findings:

- 330 unique proteins are found in human genome and have insignificant hits in all tested genomes, such as protein ID 99032 and 107876.

- Number of significant proteins with p-value $<10^{-10}$ and conserved in all tested species is 82 (universal proteins) such as protein ID 25020. While Number of significant proteins with p-value $<10^{-50}$ and conserved in all tested species is 3, namely protein ID: 7833, 10309 and 25020. The corresponding proteins name according to NCBI site are: signal recognition particle, beta-enolase isoform 2, and tRNA ligase. These proteins seem to be the core biological functions in all living cells. Fig. 2 shows the number of matched amino acid for each species when aligned to protein ID10309, Around 98% from this protein is the same in all the mammals. The following is the amino acid sequence of the protein ID 10309 in FASTA format:

>NP_001180432.1 beta-enolase isoform 2 [Homo sapiens]
MAMQKIFAREILDSRGNPTVEVDLHTAK
GRFRAAVPSGASTGIYEALELRDGDKGR
YLGKAKFGANAILGVSLAVCKAGAAEK
GVPLYRHIADLAGNPDLILPVPAFNVING
GSHAGNKLAMQEFMILPVGASSFKEAM
RIGAEVYHHLKGVIKAKYGKDATNVGD
EGGFAPNILENNEALELLKTAIQAAGYPD
KVVIGMDVAASEFYRNGKYDLDFKSPD
DPARHITGEKLGELYKSFIKNYPVVSIED
PFDQDDWATWTSFLSGVNIQIVGDDLTV
TNPKRIAQAVEKKACNCLLLKVNQIGSV
TESIQACKLAQSNGWGVMVSHRSGETE
DTFIADLVVGLCTGQIKTGAPCRSERLAK
YNQLMRIEEALGDKAIFAGRKFRNPKAK

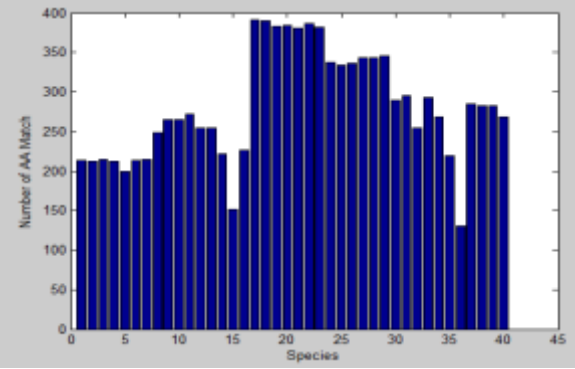

Fig. 2. Number of matched AA in protein ID10309 for each species

- There are 239 proteins common in human and chimpanzee genomes, but have insignificant hits in the other tested species.
- There are 3 proteins common in human and mouse genomes, but have insignificant hits in the other tested species.
- There are 78 proteins common in mammals genomes, but have insignificant hits in the other tested species, such as protein ID 540, 108393 and 52999. Coelacanth is seem to be the closest species to the mammals among non-mammals species, where there are 10 proteins common with the mammals, but not exist in the other non-mammals species. However, the next section illustrates another perspective.
- There are 180 proteins common in vertebrates genomes, but have insignificant hits in the other tested species, such as protein ID 99123 , 91265 and 36. Octopus is seem to be the closest species to the vertebrates among invertebrates species , where there are 19 proteins common with the Vertebrates. However, more studies should be accomplished and more genome should be included to decide what is the closest species to the vertebrates or to mammals. Figure 3 compares the score of a universal protein (ID 25020), a vertebrate protein (ID 36) and a Mammal protein (ID 540).

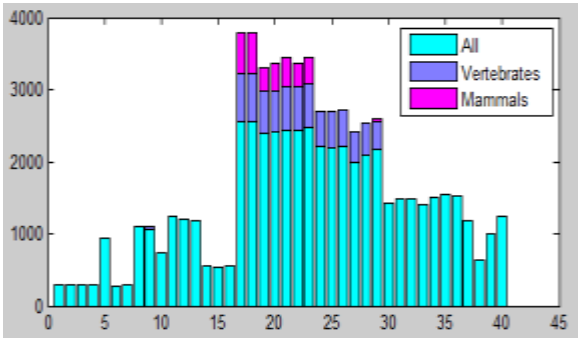

Fig. 3. Stacked scoring of protein ID 25020, 36 and 540

The conserved proteins in the mammals are compared with other organisms, the following results are obtained with expectation value $< 10^{-50}$:

- Three proteins are common in mammals and birds, and not exist in other tested species.

- Four proteins are common in mammals and fishes.
- 127 proteins are common in all the tested species except plant genome.
- Two proteins are common in mammals, birds, fishes, insects and plants.

Fig. 4 shows the number of matched amino acid in first 50 proteins when aligned to chimpanzee, coelacanh, octopus and wheat.

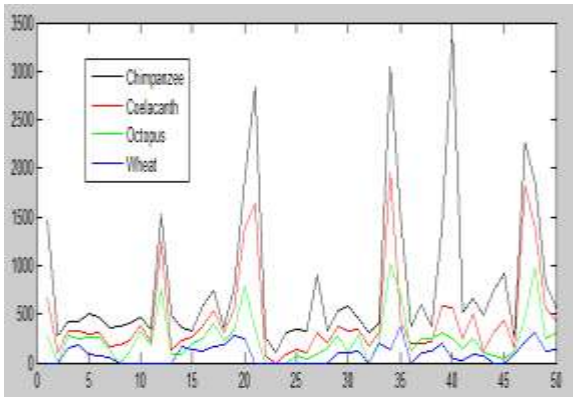

Fig. 4. the number of matched amino acid in first 50 proteins when aligned to different species

5. The Species Similarity

Fig. 5 shows the number of accepted proteins in human genome when aligned to each species and each category (the proteins is accepted if expectation value of the alignment is less than 10^{-50}). The histogram suggests that the bactria genomes (ID: 1-7), protists genomes (ID: 8-10) and archaea genomes (ID: 14-16) have the lowest homology, and the mammals genomes (ID: 18-23) have the highest homology with human genome and contains the most conserved proteins.

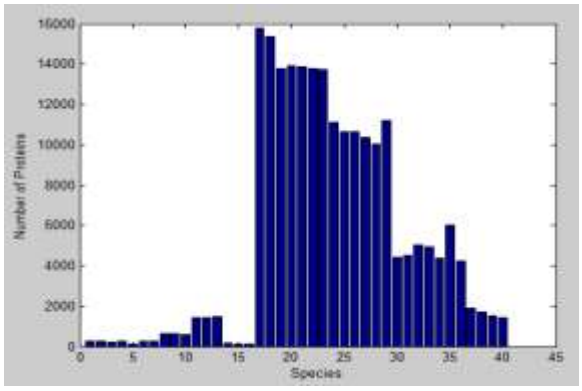

Fig. 5. The number of accepted proteins in each species.

Fig. 6 can be used to sort the families of species according to its distance from human genome, where the closer families (sorted ascendingly) are the mammals, fishes, birds, mollusks and then the insects. The farther families are plants, fungs, protists, bactria , and the farthest is the archaea genomes. Thus, the appearance time of these species will be similar to this order.

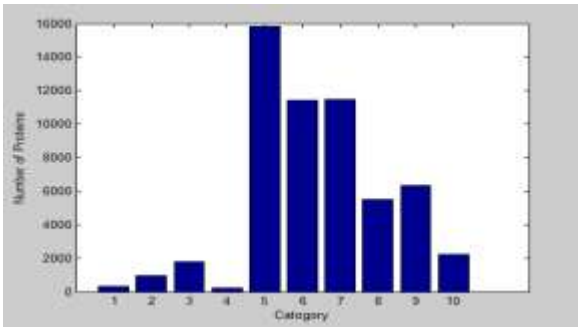

Fig. 6. Number of accepted proteins found in each category

Fig. 7 shows the match percentage of the amino acid for each species according to all proteins in the genomes. In contrary to the previous studies which use selected set of the genes and do not consider the whole genomes [16], this study proves that the similarity between human and chimpanzee is only 94.8. If Fig. 7 is compared with the previous two figures, we can conclude that the Octopus (ID 35) is closest species to the vertebrates among invertebrates species. The three figures have the same order of the species categories , but disagree whether the birds or the fishes are the closest to the mammals. Moreover, it is not clear whether the coelacanth fish (ID 29) is the closest species to the mammals (as it is given in Fig. 5) or the chicken (ID 24) is the closest to the mammals (as it is given in Fig. 7). Thus, we have two perspectives, the first based on the number of accepted proteins in the whole genome, and the second based on the similarity of the proteins content in the whole genome.

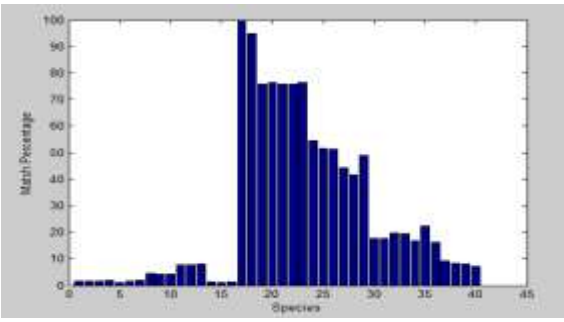

Fig. 7. The match percentage of the amino acid for each species

To find a relative relation between all the tested species and build a phylogenetic tree using human genome as reference, a distance matrix is constructed as following:

$$Distance_{ij} = \sum_{k=1}^{16614} (Score_k^i - Score_k^i)^2$$

where $Distance_{ij}$ is the distance between the species *i* and the species *j*. Therefore, its size is 40×40. The value $Score_k$ is the highest score of human proteins when aligned to the species *i*. The length of the vector *Score* is 16614. Fig. 8 shows phylogenetic tree based on the scoring of universal proteins (82 proteins). While Fig. 9 based on the scoring of all proteins. Neighbour-joining method is used in the both trees. All the tree branches are consistent with the previous figures except the birds and the fishes again. The first tree shows that the fishes and in particular coelacanth fish is closer to the mammals. While the second tree shows that the birds are closer to the mammals. This contradiction can be understood when we consider that the fishes seem to be closer to mammals from the common ancestry perspective, but the birds seem to be closer to the mammals from phenotype perspective.

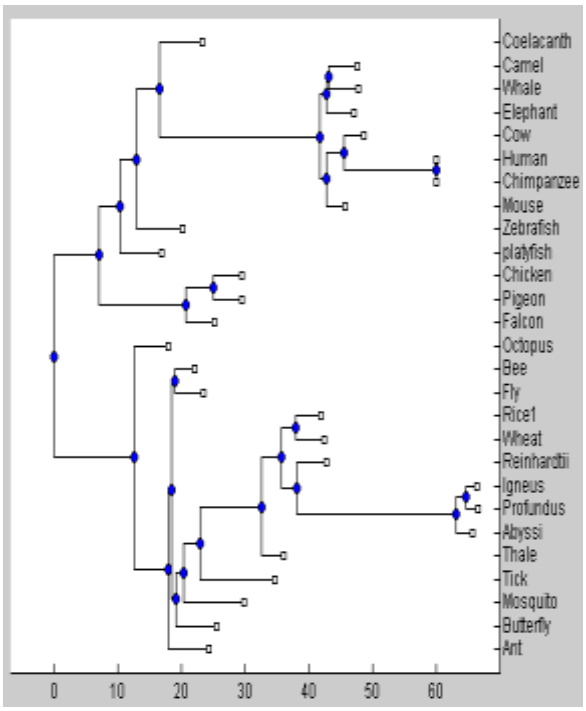

Fig. 8. Phylogentic tree based on universal proteins only

6. Conclusion

The aim of whole genomes alignment is to utilize an ensemble of related genomes to better see every individual genome in the set and to discover the core biological functions.

Comparison of proteins encoded in fourty complete genomes from ten major phylogenetic lineages allowed to identify the unique and the universal proteins in the human genome. This study found 330 unique proteins in human genome, no species besides humans have these proteins. The uniquely human proteins drive uniquely human traits which play an essential role in human dexterity, brain function, reasoning, language, speech, sensory perception and other strong cognitive components. on the other hand, 82 universal proteins are found, a significant number of them have unknown function, but they are likely to play key roles in cellular processes. Hence, there is a need for more intensive studies for these proteins.

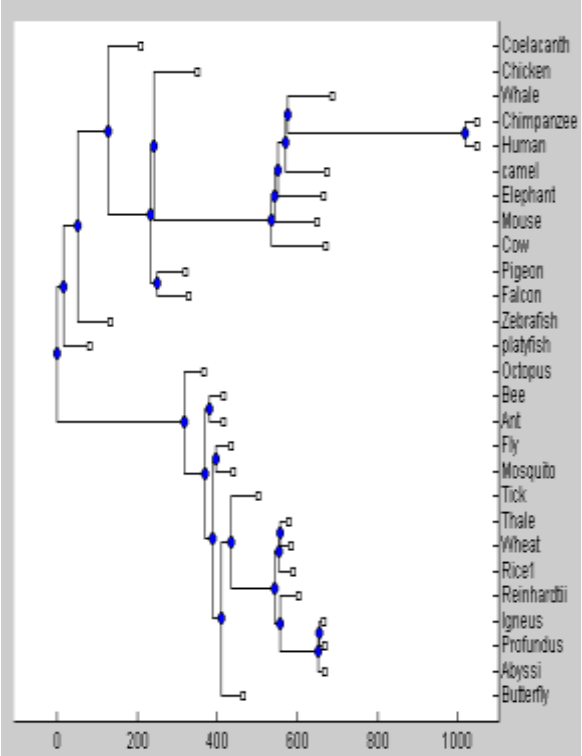

Fig. 9. Phylogentic tree based on all proteins only

Acknowledgement

This research is funded by the Deanship of Scientific Research in Zarqa University /Jordan.

References

1. Sian, L. Brain evolution: Genetic layering. Nature reviews Neuroscience 2017;18: 324-324.
2. Blanco P, Sargent CA, Affara NA. A comparative analysis of the pig, mouse, and human *PCDHX* genes. Mamm. Genome 2004*;* 15: 296–306.
3. Varkil A, Altheide KT. Comparing the human and chimpanzee genomes: Searching for needles in a haystack Genome Res 2005*;* 15: 1746-1758.
4. He Z, Han D, Efimova O, Guijarro P, Yu Q, Oleksiak A, et al. Comprehensive transcriptome analysis of neocortical layers in humans, chimpanzees and macaques. Nat. Neurosci 2017; 20: 886–895.
5. Dixon JR, Gorkin DU, Ren B. Chromatin Domains: The Unit of Chromosome Organization. *Mol Cell* 2016; 62: 668–680.
6. Bae BI, Jayaraman D, Walsh CA. Genetic changes shaping the human brain. *Dev Cell* 2015; 32: 423–434.
7. Lake BB. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* 2016; 352: 1586–1590.
8. Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, et al. EAGER: efficient ancient genome reconstruction. *Genome Biol. 2016;* 17: 60-74.
9. Pozzi L, Bergey CM, Burrell AS. The use (and misuse) of phylogenetic trees in comparative behavioral analyses. International Journal of Primatology 2014; 35: 32–54.
10. Al Daoud, E. Fast Protein Classification by Using the Most Significant Pairs. EXCLI Experimental and Clinical Sciences Journal *2010;* 9: 133-140.
11. McMahon MA, Rahdar M, Porteus M. Gene editing: not just for translation anymore. *Nat Methods* 2012; 9: 28–31.
12. Al Daoud, E. Identifying DNA Splice Sites using Patterns Statistical Properties and Fuzzy Neural Networks. EXCLI Experimental and Clinical Sciences Journal 2009; 8:195-202.
13. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 2015; 12: 59–60.
14. Fiz-Palacios DH, Fu CJ, Fehling J, Tsai CC, Baldauf SL. An alternative root for the eukaryote tree of life. *Curr Biol* 2014*;* 24: 465-470.
15. Patro K, Kumar P. Optimized Feature Selection with Mutual Information for ECG based Bio-Metric Recognition system using Genetic Algorithm. MAGNT Research Report 2018; 5: 222-231.
16. Tatsuya A, Takashi S, Natsuki K, Kazuyo Y, Sakae K, Atsuko S, et al. Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. Proceedings of the National Academy of Sciences 2003; 100: 7708-7713.