International Journal of Engineering & Technology, 7 (1.9) (2018) 162-166

International Journal of Engineering & Technology

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

Research Paper

Evolutionary clustering annotation of ortho-paralogous gene in a multi species using Venn diagram visualization

Bipin Nair B J [1] *, Sarath M S [1]

[1] *Department of computer science Amrita School of Arts and Science,Mysuru Campus Amrita Vishwa Vidyapeetham, India*
**Corresponding author E-mail: bipin.bj.nair@gmail.com*

Abstract

The evolutionary analysis of the genome of the immediate cluster is an important part of comparative genomics research. Identifying the overlap between immediate homologous clusters allows us to elucidate the function and evolution of proteins between species. Here, we report a network platform called Ortho-paralogous Venn-diagram representation that can be used to compare and visualize a wide range of ortho-paralogous clustering of genomes. In our work Ortho-paralogous Venn-diagram results show a functional summary of interactive Venn diagrams, summary counts, and interspecies shared cluster separations and intersections. Ortho-paralogous Venn-diagram also uses a variety of sequence analysis tools to gain an in-depth understanding of the cluster. In addition, Ortho-paralogous Venn identifies direct homologous clusters of single copy genes and allows custom search of specific gene clusters. It enables us in wide analysis of the genes and protein by comparing the genes using Venn diagram .Here the user can upload our own gene sequences into the application ,using three clustering approach to check the best clustering approches like SOM,K-means and advanced clustering after that we are using the Venn diagram repersentator to evolutionary cluster the genes having similar functionality and structural similarity from the uploaded data.Here we are using a venn diagram representation as an application which used to cluster the orthologous and paralogous gene on basics of their evolution and functional aspects.it enables us in wide analysis of the genes and protein by comparing the genes using venn diagram representation.here the user can upload our own gene sequences into the application where the venn diagram representator clusters.the genes having similar functionality and structural similarity from the uploaded data.

Keywords: *Venn Diagram Representator; Orthologous; Paralogous; UPGMA; SOM.*

1. Introduction

In a multi species segregation is a tedious process so clustering approach will give a clear idea of separation of various species for that we can use multiple clustering algorithm. Combining.
All three algorithms to check the efficiency.Orthologous and paralogous genes are the clusters of genes from different species where the first shows the similar character or functionality achieved from the common ancestry whereas the latter shows the similar character due to the constrained evolution. The orthologous gene shows the same functional similarity whereas the paralogous shows the same structural similarity. Thus, the comparative study of these two genes helps us to gain the information about their gene evolution and their structure. Such information from the comparative study helps us the phylogenetic of organisms.
In order for the study of the orthologous and paralogous genes we considered the Ensemble database for the pairwise comparison of the gene sequences. To provide the better understandability of overlapping of the genes, clusters are represented in different colors. Using Venn diagram which are all the genes orthologous and paralogous in nature using intersection, that is commonality, and all together common genes using union feature. It not only provides the visualization of the genes but also the enables the comparison of both the clusters.
In order to make the comparison we take the gene sequences from the Ensemble dataset which is a collection of large number of gene comprising of orthologous and paralogous genes.so from the large collection of gene sequences we need to identify and cluster the genes which are orthologous or paralogous and compare them and separate them using Venn diagram representation, if we apply clustering and Venn diagram representation we can overcome the limitation of improper visualization.

2. Litrature survey

In present literature they are explain about clustering of genes and visualizing of similar gene using various clustering algorithm, those literature are
Servant et.al [1] the paper propose ProDom graphical interface. The fundamental window gives general data on the area family (upper right): the section promotion number, the photo used to speak to the family on the graphical yield, a few measurements, and helpful connections. Wei et.al.
A novel diAnoA novel separation work called the Block Similarity measure, is proposed for evaluating the closeness between adjusted arrangements and for choosing which groupings ought to be bunch level separation work called the Block Similarity measure. Thomas et.al [3] Qualities were bunched and converged as portrayed in the content and in materials and strategies. An assortment of assets, particularly the UCSC Family Browser, were utilized to determine an enlightening family identifier and the protein family (PF) as portrayed on the Pfam site. The position is the area of the centroid of the quality group (see materials and techniques) in nucleotide directions of the WS120 informational collection at WormBas. Shannon et.al [4] Developmental Analysis To explore the phylogenetic connections existing amongst human and mouse

qualities, both nucleotide and anticipated amino corrosive groupings of the ZNF areas from the 31 related loci were adjusted utilizing CLUSTAL_X1.8 (Thompson et al. 1997) and looked at utilizing PAUP4.0b10 (Swofford 2002). A progression of phylogenetic trees was developed utilizing distinctive calculation. Bipin Nair et.al [5] this paper manages the recognizable proof of the change in the nucleotide arrangement of the genome of a life form, infection, or additional chromosomal hereditary component position. Changes in DNA arrangement comes about because of unrepaired harm to DNA grouping or to RNA genome. Sujith et.al [6] Gathering of different haematological blood protein sequences. Finding the orthologous gene from the haematological blood protein, Classify each disorder, Perform UPGMA method, Construct phylogenetic tree, Visualization of the Phylogenetic tree different haematological blood protein. Wang et.al [7] OrthoVenn provides detailed comparison details the orthologous gene of multiple species with detailed visualization using Venn diagrams. Peterson et.al
Orthologous and paralogous was studied based on their both functional and structural similarity rather than concentrating only on their functionality similarity. Fouts et.al
All four clustering Methods that were adopted agreed on approx. 70% of the clusters and approx86% of the proteins.
The unwanted data's that were present was removed manually.
Lechner et.al [10] High accuracy and applicable for large dataillustration of orthology connections is a critical advance both in quality capacity expectation and in addition towards understanding examples of arrangement development. Berglund sing et.al [11] Identifying orthology relationship between different species in INPARANOID database using Phylogenetic tree. Orthology relationship is calculating through mathematical equations. Less accuracy. Singh.et.al
Conceivably intervening articulation dissimilarity in paralogous. Examining the administrative changes possibly intervening articulation disparity in paralogous Gene families under particular conditions. Frias-Lopez.et.al [13] they report here worldwide examination of communicated qualities in a normally happening microbial group. We initially adjusted RNA enhancement innovations to create a lot of cDNA from little amounts of aggregate microbial group RNA. Fouts et. al.
PanOCT is an apparatus for dish genomic investigation of firmly related prokaryotic species or strains. PanOCT utilizes conservedgene neighborhood data to isolate as of late veered paralogous into orthologous bunches where homology-just grouping techniques can't. Muller et.al [15] the new system contains twice the number of proteins and species. It provide the three domain of life with several levels of resolutions. Alexeyenko et.al [16] MultiParanoid is an openly accessible independent program that empowers productive orthology examination much required in the post genomic period. An electronic administration giving access to the first datasets, the subsequent gatherings of orthologous. Liang et.al [17] MultiParanoid is an openly accessible independent program that empowers productive orthology examination much required in the post genomic period. An electronic administration giving access to the first datasets, the subsequent gatherings of orthologous. lee et.al [18] EST sequencing remains the primary method of genomic sequencing analysis .the TIGR gene indices which represents the most comprehensive publically available analysis of EST sequences. These processes provide a set of unique high fidelity virtual transcripts or tentative consensus (TC) sequences. Chen et.al [19] the emergence of genomics has changed the way by which genes are discovered and isolated. Previously, a peptide arrangement or phenotype has prompt the seclusion of another quality. Presently an examiner starts with the succession of a key quality and scans for homologous qualities in a creature of moment. Genomics, as a powerful approach, will certainly have tremendous impact on both fundamental and applied research. Uchiyama. Uchiyama et.al
We uses UPGMA algorithm for clustering orthologous genes and to identify the domain fusion or domain fission as the part of the clustering.

In our work we are taking the data set of various species an grouping the ortho para genes using clustering, clustered genes is representing in Venn-visualizer and applying the features like intersection and union and finally comparing the efficiency. But in present literature we are not comparing efficiency and not using the Venn-features and they are not grouped frequently the paralogous genes.

3. Problem formulation

In order to make the comparison in species grouping effectively is a tedious process. The present grouping strategy's are less efficient and time consuming, finding the proper evolutionary relationship .proper visualization of ortho-para representation is a difficult process in existing literature.

4. Problem definition

In order to make the comparison we take the gene sequences from the Ensemble dataset which is a collection of large number of gene comprising of orthologous and paralogous genes.so from the large collection of gene sequences we need to identify and cluster the genes which are orthologous or paralogous and compare them using ven diagram visualization.

5. Methodology

In our proposed methodology we are taking three species data set s like birds, fishes, mammals and applying three clustering algorithm finding the efficient one .from the grouped data set using Ven diagram separating them as ortho , para category, finally finding evolutionary relationship through phylogenetic tree representation.

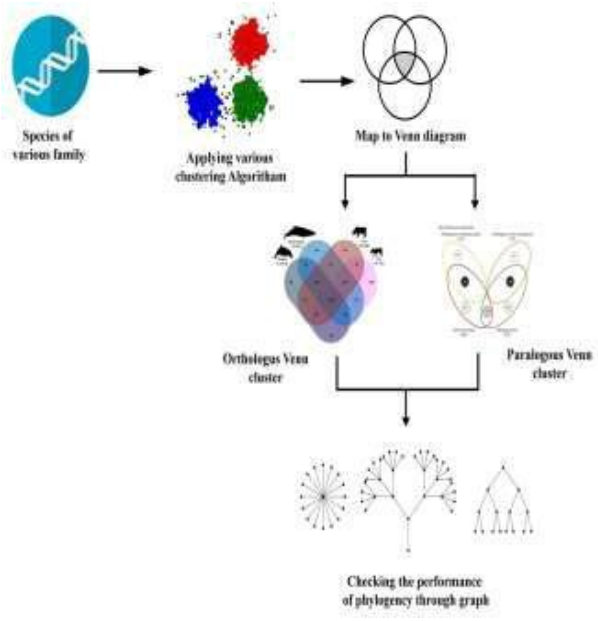

Fig: 1: Flow Diagram.

6. Algorithm

The algorithm which is used to cluster the various species is self-organizing map, k-means and advanced clustering, here we are showing the pseudo representation of effective clustering like SOM and phylogenetic tree construction using UPGMA algorithm is. It is a genome – scale calculation for gathering orthologous protein succession. It gives not just gatherings shared by at least two species yet in addition bunches speaking to species particular quality development families. The calculation begins with best corresponding hits over any two genomes as potential orthologous

sets. It is mainly used for the identifying the orthologous and paralogous genes. It helps us in constructing orthologous and paralogous genes across multiple eukaryotic taxa.
SOM algorithm
Step1: set the no of input layers
Step2: Iterations (100)
Step3: error approach
Step4: partition the group
UPGMA algorithm
Step1: input the range values for each species
Step2: construct the matrix
Step3: iterations
Step4: construct the tree

7. Orthologous and paralogous cluster

Orthologous cluster is a group of genes having similar character due to common ancestry and paralogous cluster is a group of genes having similar character due to constrained evolution. In Orthologous cluster contain group of genes having similar functionality and paralogous cluster contain group of genes having structural similarity. IN our work we are using ven diagram to segregate ortho paralogous gene from multiple species.in our work four category orthologous in nature, three category paralogo in nature, two category ortho-para nature

8. Related work

The existing work is used to cluster the genes sequences based on orthologous and paralogous with simple visualization. our work first phase we are passing our data set with three clustering algorithm then finding the efficient one ,later stage finding the evolutionary relationship. Ven diagram will provide the high detail information about their similarity and dissimilarity.
Data set

	A	B	C	D	E	F
1	taxon	common.r	class	order	family	genus
2	lake fishe	american	actinopte	anguillifo	anguillida	anguilla
3	river fishe	blacktail r	actinopte	cyprinifor	catostomi	moxoston
4	river fishe	central st	actinopte	cyprinifor	cyprinidae	camposto
5	river fishe	rosyside d	actinopte	cyprinifor	cyprinidae	clinostom
6	river fishe	longnose	actinopte	cyprinifor	cyprinidae	rhinichthy
7	river fishe	muskellur	actinopte	esociform	esocidae	esox
8	marine fis	pollack	actinopte	gadiforme	gadidae	pollachius
9	marine fis	saithe	actinopte	gadiforme	gadidae	pollachius
10	marine fis	lined surg	actinopte	perciform	acanthuri	acanthuru
11	marine fis	orangespi	actinopte	perciform	acanthuri	naso
12	marine fis	bluespine	actinopte	perciform	acanthuri	naso
13	marine fis	redlip ble	actinopte	perciform	blennidae	ophiobler
14	marine fis	giant trev	actinopte	perciform	carangida	caranx
15	lake fishe	rock bass	actinopte	perciform	centrarchi	ambloplit

Fig. 2: Sample Data Set.

9. Experimental result

Clustering is performed by uploading the species like birds, mammals, and fishes as a .csv file into our interface. The main advantage of the work is that we are comparing with three algorithms and finding the efficient one according to the error rate parameter. Then clustered data is representing in Ven diagram as well as phylogenetic tree for finding evolutionary relationship.
In k-means according to the various property of cluster it generate cluster, it is a basic clustering approach get the cluster we are first convert our data set in numeric value according to the property which we are selecting from a species. Based on one arbitrary centroid value it will iterate and finally form the exact fitting cluster.

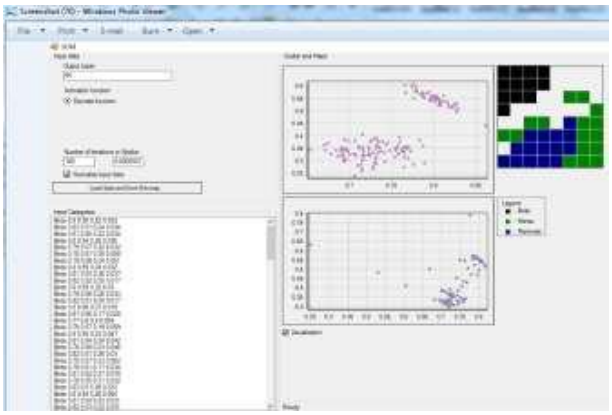

Fig. 3: SOM Cluster

In this self-organizing map we are using neural network approach, we are setting the input layer then it iterate 100 times, based on the error rate approach we are fining the exact fitting cluster, while passing the data set on each node we calculate the error rate based on the less error rate we are getting the effective cluster in the form of category, here we have used automatic generation in 3D visualization for clustering, finally we can say according to map generated fish and mammals show the similarity.

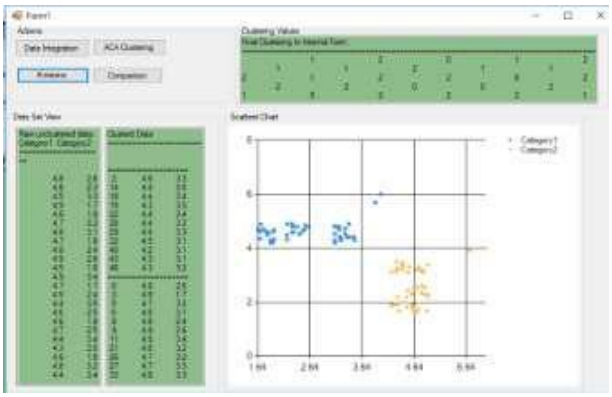

Fig. 4: K Means Cluster.

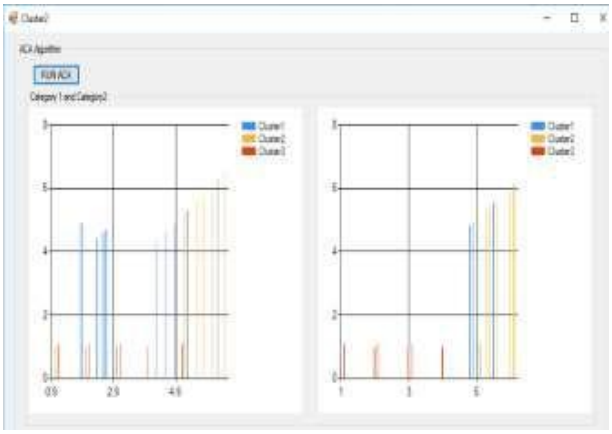

Fig. 5: Advanced Clustering.

In advanced clustering it is clustering according to frequency values that means height it is taking from the data set according to the property.it is not a stable algorithm it varies according to the property which we select from the species, based on the frequency it will form a cluster

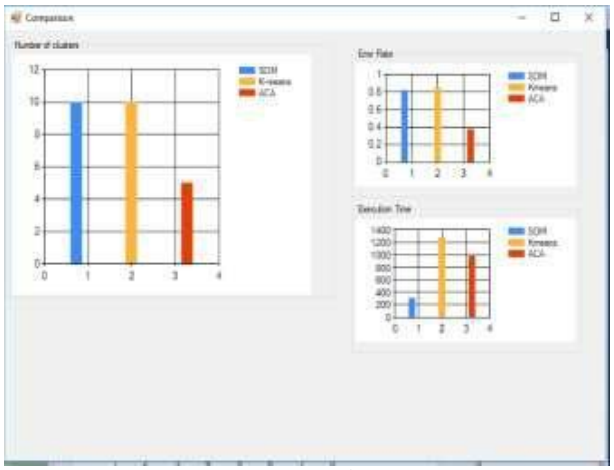

Fig. 6: Comparison of Clustering.

In the above graph shows the comparison of three clustering algorithms like SOM, K-means and advanced clustering algorithm. Based on error rate method SOM is and efficient algorithm, it will cluster efficiently as well as for each pass of data set it will show 3D visualization.

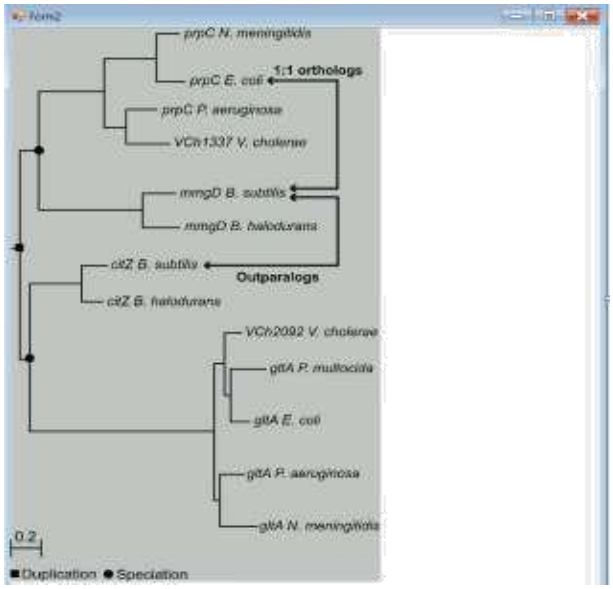

Fig. 7: Phylogenetic Tree.

Using UPGMA method we are constructing the phylogenetic tree, here using a mathematical model a range values are deciding from various species to construct the matix, and then using mathematical calculation of average method constructing the tree using various iteration method. Here we are using two process duplication and speciation. This three will give the evolutionary aspect of clustered data.

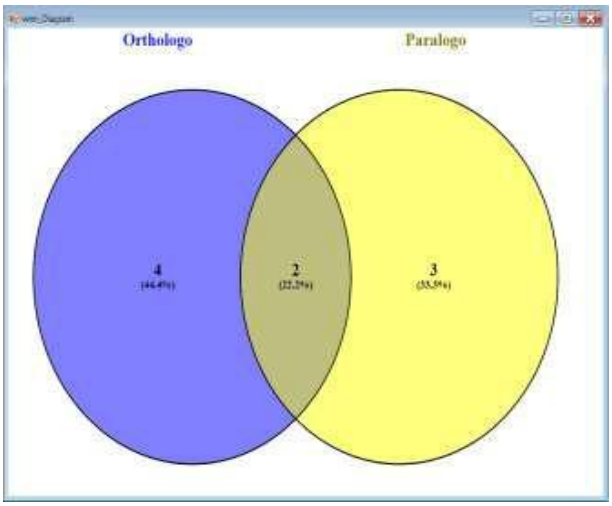

Fig. 8: Ven Diagram Representation.

From the clustered data we are representing orthologous, paralogous and ortho-para combination from various species using Ven diagram representation. Here in this Ven diagram we are using different set operations like difference for ortho only and para only, intersection for ortho-para combination.
This Ven diagram will give clear separation of functional similarity and difference of various taxa's.

10. Conclusion

The Venn diagram helps us to cluster the data with based on orthologous and paralogous gene with high visualization details and various clustering methods will give efficient grouping of various species also we are finding the evolutionary relationship through phylogenetic tree. The future scope of the Venn diagram could provide refined information on orthologous and paralogous gene for large volume of the data.

References

[2] Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D., & Kahn, D. (2002). ProDom: automated clustering of homologous domains.*Briefings in bioinformatics*, *3*(3), 246-251. https://doi.org/10.1093/bib/3.3.246.

[3] Wei, X., Kuhn, D. N., & Narasimhan, G. (2003, August). Degenerate primer design via clustering. In *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE* (pp. 75-83). IEEE. https://doi.org/10.1109/CSB.2003.1227306.

[4] Thomas, J. H. (2006). Analysis of homologous gene clusters in

[5] Caenorhabditis elegans reveals striking regional cluster Domains. *Genetics*, *172* (1), 127-143. https://doi.org/10.1534/genetics.104.040030.

[6] Shannon, M., Hamilton, A. T., Gordon, L., Branscomb, E., & Stubbs, L. (2003). Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. *Genome research*, *13*(6a), 1097-1110. https://doi.org/10.1101/gr.963903.

[7] Bipin Nair B J #1 # DNA Sequence Alignment Using Matching Algorithm to Identify the Rare Genetic Mutation in Various Proteins.

[8] Sujith, M., & Alphonsa, M. V. Self-regulating Exploration for Orthologous in Homologous Hematologic Gene Sequence Data Using UPGMA Method.

[9] Wang, Y., Coleman-Derr, D., Chen, G., & Gu, Y. Q. (2015). OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic acids research*, *43*(W1), W78-W84. https://doi.org/10.1093/nar/gkv487.

[10] Peterson, M. E., Chen, F., Saven, J. G., Roos, D. S., Babbitt, P. C., & Sali, A. (2009). Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Science*, *18*(6), 1306-1315. https://doi.org/10.1002/pro.143

[11] Fouts, D. E., Brinkac, L., Beck, E., Inman, J., & Sutton, G. (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and

closely related species.*Nucleic acids research*, *40*(22), e172-e172. https://doi.org/10.1093/nar/gks757.

[12] Lechner, M., Hernandez-Rosales, M., Doerr, D., Wieseke, N., Thévenin, A., Stoye, J., & Stadler, P. F. (2014). Orthology detection combining clustering and synteny for very large datasets. *PLoS One*, *9*(8), e105015. https://doi.org/10.1371/journal.pone.0105015.

[13] Berglund, A. C., Sjölund, E., Östlund, G., & Sonnhammer, E. L. (2007). InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic acids research*, *36*(suppl_1), D263-D266. https://doi.org/10.1093/nar/gkm1020.

[14] Singh, L. N., & Hannenhalli, S. (2009). Correlated changes between regulatory cis elements and condition-specific expression in paralogous gene families. *Nucleic acids research*, *38*(3), 738-749. https://doi.org/10.1093/nar/gkp989.

[15] Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., & DeLong, E. F. (2008). Microbial community gene expression in ocean surface waters. Proceedings of the National Academy of Sciences, 105(10), 3805-3810. https://doi.org/10.1073/pnas.0708897105.

[16] Fouts, D. E., Brinkac, L., Beck, E., Inman, J., & Sutton, G. (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species.*Nucleic acids research*, *40*(22), e172-e172. https://doi.org/10.1093/nar/gks757.

[17] Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., & Bork, P. (2009). eggNOG v2. 0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species andfunctional annotations. Nucleic acids research, 38(suppl_1), D190-D195.

[18] Alexeyenko, A., Tamas, I., Liu, G., & Sonnhammer, E. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, *22*(14), e9-e15. https://doi.org/10.1093/bioinformatics/btl213.

[19] Quackenbush, J., Liang, F., Holt, I., Pertea, G., & Upton, J. (2000). The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic acids research*, *28*(1), 141-145. https://doi.org/10.1093/nar/28.1.141.

[20] Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., & White, J. (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, *29*(1), 159-164. https://doi.org/10.1093/nar/29.1.159.

[21] Chen, R., & Jeong, S. S. (2000). Functional prediction: identification of protein orthologs and paralogs. *Protein Science*, *9*(12), 2344-2353. https://doi.org/10.1110/ps.9.12.2344.

[22] Uchiyama, I. (2006). Hierarchical clustering algorithm forcomprehensive orthologous-domain classification in multiple genomes. *Nucleic acids research*, *34*(2), 647-658. https://doi.org/10.1093/nar/gkj448.