Unraveling Evolutionary Relationship between Fungi Species

LINYUN LIU[1], TARANDEEP KAUR[2], AND SHIKA BORGE[3]

[1] *Trinity Western University*
[2] *22500 University Drive, Langley, BC, Canada, V2Y1Y1*
[3] *Linyun.Liu@mytwu.ca, Tarandeep.Kaur@mytwu.ca, Shika.Borge@mytwu.ca*

Compiled April 16, 2023

With the passage of time, organisms evolve. Organisms from all five kingdoms have undergone numerous genetic alterations, with some becoming extinct or losing/gaining genes. Fungal species are diverse in the eukaryotic kingdom, and their genes evolve throughout time as environmental circumstances change and through natural selection. We present a comparative analysis of 14 fungal species from the phylum Ascomycota's distinct class groups. We compared the 12 fungal species to two key species, Saccharomyces Cerevisiae and Schizosaccharomyces Pombe, to study gene duplication by extracting protein sequences and building phylogenetic trees. These studies have been compiled into a source code that serves as an online resource for comparative genomic research of the fungal species supplied.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

Genetic novelty in organisms is generated through gene duplication. Duplication is a mutation that involves the production of multiple copies of a DNA segment. Duplications can consist of as little as a couple of bases or as large as a chromosome. They play a crucial role in the evolution of humans as well as other organisms' genomes.[1]. It has been demonstrated that gene duplication has resulted in a wide variety of new gene functions that have contributed enormously to the development of different organisms as a result of unequal crossing over, retropositions, and chromosomal duplications[2]. The mechanisms of gene duplications and the dynamics between genes are vital for understanding evolution's influence on genome contents, evolutionary relationships, and interactions. [2]

Fungi are the largest group of organisms ranging from slime mold, to mushrooms to yeast. These eukaryotic organisms are classified as monophyletic Eumycota group. Their diversity ranges from 500 to 9.9 million while spanning over 1 billion years of evolutionary history. The diversity and abundance of these organisms can be attributed to their small size, cryptic lifestyle and symbiotic relationship with algae, fungi, biophytes, pteridophytes, in addition to higher plants and animals worldwide[3]. In addition to their diversity, they are also able to survive in various climates, and can be found dominating the world. Their ecological dominance in nature has allowed them to play a central role in human endeavor: mushrooms and truffles are consumed by humans while yeast is used to make bread and baked goods. Furthermore, fungi carry out nutrient cycling, produce antibiotics, enzymes, mycotoxins, alkaloid compounds, polyketides and mycotoxins[3]. The kingdom phyla are classified into several major phyla namely, " Ascomycota, Basidiomycota, Chytridiomycota, Monoblepharidomycota, Neocallimastigomycota, Blastocladiomycota, Glomeromycota, Entomophthoromycota, Stramenopiles and Micorsporidia and sub-phyla namely Kickxellomycotina, mucoromycotina and Zoopagomycotina" [3]. Fungi are valid from an evolutionary perspective because of their diverse ecological dominance. It is for these reasons that fungi should be subjected to intense phylogenetic ecological and molecular studies because of the diverse and ecological dominance of fungi throughout time[3]. The proteomic sequences of different fungal species from the Phylum Ascomycota collected from the NCBI Blast and the Protein Clusters are aligned using software Clustal Omega and generated by the Phylogenetic tree with Maximum Likelihood method using software PhyML-3.1. Maximum Likelihood method is used to find the branch lengths of the tree with the greatest probability of observing genomic and proteomic sequences [11]. Phylogenetic trees are an easy approach to visualize the common evolutionary history of several animals across millions or billions of years. The main focus of evolutionary trees is ancestors and descendants [7]. references.

2. OBJECTIVES AND METHODS:

The objective of this research is to address the under-utilization of tree-based clustering in bioinformatics applications for clustering homologous sequences based on their similarity [5]. Furthermore, Classification of proteins holds significant importance across diverse fields of biology [10]. Homologous molecular sequences across different species or within the same genome can show remarkable similarity due to their shared evolutionary history. This similarity can be due to the retention of ancestral sequence features or the convergence of sequences through natural selection. Clustering is a common technique used to group sequences with high similarity. In clustering, a set of sequences is partitioned into clusters, such that each cluster contains sequences that are more similar to each other than to sequences in other clusters [5]. In the context of protein clusters, classifi-

cation can help to identify groups of proteins that have similar functional properties or perform similar biological roles. By clustering proteins based on their sequence or structural similarities, we can identify protein families or superfamilies and gain insights into the functional properties and evolutionary relationships of these proteins [13]. The significance of this problem is the potential to provide more accurate results and reveal evolutionary relationships between sequences. The main technique used to address the problem is constructing a phylogenetic tree from protein clusters using maximum likelihood methods. The research project aims to compare the genomes of 12 distinct fungi from different phyla and classes to Saccharomyces cerevisiae and Schizosaccharomyces pombe to reveal their evolutionary relationships [2]. By using a tree-based approach, the researchers intend to provide a more accurate and informative understanding of the relationships between these fungi species.

The research project utilizes a robust theological framework as a guiding principle to construct a phylogenetic tree from protein clusters. Protein sequences were obtained from NCBI for the species listed in table 1, and a diverse range of software was used to analyze and interpret the data. The Google Cloud Compute Engine server was used for data processing and analysis. To generate protein clusters, the protein sequences of Saccharomyces cerevisiae were compared to Schizosaccharomyces pombe using ncbi-blast2.13.0+ software. Jalview was used for the visualization and analysis of multiple sequence alignments, including protein clusters, offering color-coding based on conservation to identify important regions in the proteins (figure 3). Biopython was used to convert protein sequence files into different formats and to construct phylogenetic trees based on evolutionary relationships between protein sequences. PhyML-3.1 software was used to generate tree data, and Clustal Omega was used to convert the original .fasta file into the required .phylip format[9]. Biopython and Python were used for file conversion and tree drawing, respectively, resulting in a robust and efficient workflow. This approach facilitated the accurate representation and analysis of the evolutionary relationships between the protein sequences.

A. Collect Protein sequences

We collect our GCF protein sequences from NCBI. By selecting the relevant genome and filtering the search for protein sequences specific to the species(fig 1),they were then downloaded into a FASTA format

B. Google Server

Using Google Cloud Compute Engine (Debian Linux Compute-optimized) Google Cloud Compute Engine is a cloud computing platform that allows you to create and manage virtual machines (VMs) on Google's infrastructure[11]. Debian Linux Compute-optimized is a machine type optimized for compute-intensive workloads, such as scientific computing, data analysis, and machine learning.

C. Blast Software

software Name: ncbi-blast-2.13.0+ the software can be found using the follwing link connect ftp to download the lastest blast software

D. Result Handling and Protein Cluster

The result is handled using Java, it invloves data structures and algorithm that sort the results find the best results and generate protein cluster fasta f

query acc.ver	subject acc.ver	% identity	alignment length	mismatches	gap opens	q. start	q. end	s. start	s. end	evalue	bit score
NP_001018029.1	XP_033764819.1	81.818	66	12	0	1	66	1	66	1.24E-35	112
NP_001018029.1	XP_033765465.1	21.212	33	26	0	34	66	2	34	3.7	23.1
NP_001018029.1	XP_033765560.1	42.105	19	11	0	1	19	142	160	4.9	23.1
NP_001018029.1	XP_033765162.1	46.429	28	14	1	18	45	141	167	6.6	22.3
NP_001018030.1	XP_033764893.1	94.167	360	21	0	1	360	1	360	0.0	709
NP_001018030.1	XP_033766890.1	54.601	326	147	1	5	329	4	329	7.4E-128	368
NP_001018030.1	XP_033765906.1	27.046	281	187	6	4	283	75	338	3.68E-16	77.4
NP_001018030.1	XP_033767421.1	24.606	317	203	9	6	317	21	306	6.94E-12	63.5
NP_001018030.1	XP_033766331.1	28.448	232	129	12	21	232	368	582	6.42E-07	48.9
NP_001018030.1	XP_033766492.1	21.244	193	145	4	22	212	38	225	3.76E-04	40.0
NP_001018030.1	XP_033766975.1	32.558	43	28	1	289	331	208	249	9.8	26.2
NP_001018030.1	XP_033765901.1	32.558	43	28	1	289	331	208	249	9.8	26.2
NP_001018031.2	XP_033764897.1	88.496	113	13	0	1	113	1	113	5.8E-67	195
NP_001018031.2	XP_033768692.1	31.373	51	30	1	54	99	1	51	0.19	28.5

Fig. 1. Example result of blasting the protein sequences of Saccharomyces cerevisiae against Schizosaccharomyces pombe using the ncbi-blast2.13.0+ software

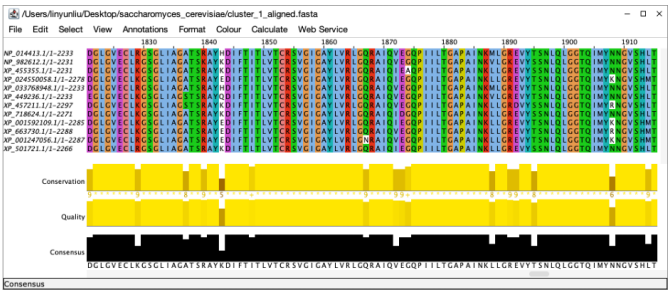

Fig. 2. Jalview is used to visualize the protein clusters in this case the protein cluster 1 of Saccharomyces cerevisiae is shown. Jalview conservation shading is a visualization tool that displays the level of sequence conservation. Darker colors indicate a higher level of conservation, and lighter colors indicate a lower level of conservation. [6] The level of conservation is determined by comparing the amino acid residues at each position in the alignment and calculating the degree to which they are conserved across all clusters. [6]

E. Biopython And Other

Biopython is a powerful open-source library for biological computation that provides tools for manipulating, analyzing, and visualizing biological data [12]. It includes modules for handling various types of biological data, such as DNA and protein sequences, structures, and alignments. On the other hand, Matplotlib is a widely used data visualization library in Python that allows users to create high-quality visualizations, including bar charts, line plots, scatter plots, and histograms. Together, Biopython and Matplotlib can be used to analyze and visualize biological data, making it easier for researchers to understand and interpret complex biological systems. By utilizing the strengths of both libraries, scientists can gain insights into biological processes and communicate their findings effectively [12].

F. Clustal Omega

Clustal Omega is a multiple sequence alignment program developed by the Computational Biology Research Group at University College Dublin. It is the successor to the widely used ClustalW program and is a powerful tool for aligning three or more biological sequences, such as DNA or protein sequences. Clustal Omega uses a sophisticated algorithm to produce high-quality alignments quickly and accurately. One of the key advantages of Clustal Omega over other alignment programs is its ability to handle large datasets quickly and efficiently. It uses a scalable parallel algorithm that can take advantage of multi-core

processors and distributed computing environments. Additionally, Clustal Omega can handle a wide variety of sequence types, including nucleotide and protein sequences, as well as RNA and DNA sequences with complex structures. In order to generate ML phylogenetic tree using other software (using cluster.fasta file), the fasta file need to be formatted The original cluster.fasta file contais protein sequences that have different length and they are not aligned Hence, the clustalo software is capable of formatting the cluster.fasta file

G. PhyML

PhyML is a popular software tool for constructing phylogenetic trees from molecular sequence data. It uses maximum likelihood methods to infer the evolutionary relationships among a set of sequences and provides several models of sequence evolution. PhyML offers a user-friendly interface with various customization options, making it a flexible and powerful tool for phylogenetic analysis. It also provides features for bootstrapping and statistical testing to assess the confidence of the inferred tree topology. PhyML is widely used by researchers in diverse fields such as evolutionary biology, genomics, and bioinformatics [9].

3. RESULTS

Our main aim was to study the evolutionary relationship among the 12 fungal species by comparing them to the 2 main species Saccharomyces Cerevisiae and Schinzosachharomyces Pombe. All the FASTA format for the protein sequences were collected in a form that could be easily assessed to demonstrate the evolutionary relatedness among the below listed fungal species in table 1. Table 1 shows an example table.

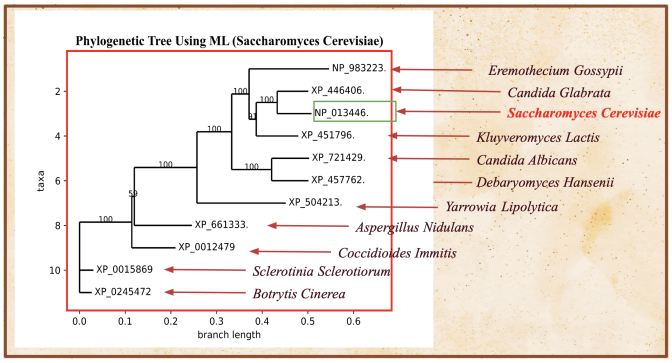

Fig. 3. Phylogenetic relationship between Saccharomyces Cerevisiae and 11 species using protein cluster.

The Maximum Likelihood method was used to generate the phylogenetic tree by aligning the fasta format of protein clusters using software Phylip. The figure 3 and 4 shows the evolutionary relationship between the main species Saccharomyces Cerevisiae and the Schizosaccharomyces Pombe with the other 11 species mentioned in table 1. In the Phylogeny tree, the X-axis shows the branch length and Y-axis shows the taxa of the species. The branches represent the time period when the species diverged from a common ancestor. The tip or leaf shows the species whose relationship is depicted using the tree diagram. Nodes are the point where the common ancestor splits into two different taxa. Overall, evolutionary relationships can be observed from the above given figures 3 and 4.

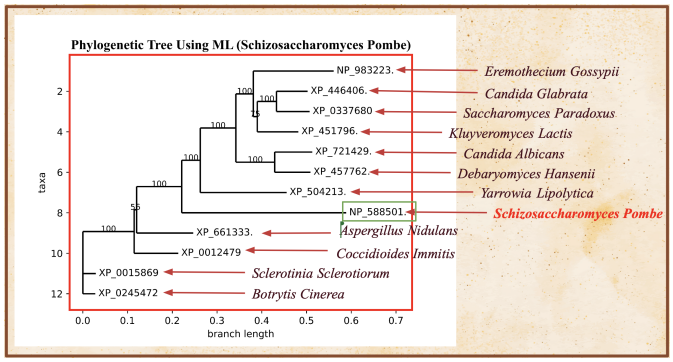

Fig. 4. Phylogenetic relationship between Schizosaccharomyces Pombe and 11 species using Protein Cluster

4. DISCUSSION

The FASTA protein sequences of all the 14 fungal species was used to generate the protein clusters on the BLASTp. The two species named Botrytis cinerea and Sclerotinia sclerotiorum that belongs to class Leotiomycetes has largest protein sequence size, i.e., 7.46 Mb and 6.19 Mb respectively among all the 14 fungal species. The Encephalitozoon cuniculi is an only fungal species belonging to Class Microsporidea has the protein sequence size of 876 kb. Many of the fungal species has protein size between 2.67 Mb to 3.39 Mb and all belongs to the class Saccharomycetes and falling in the same Phylum Ascomycota as shown in Table 1. Therefore, the largest protein sequence belongs to the Class Leotimycetes and the smallest protein sequence belongs to the Class Microsporidea and belongs to the different phyla named Acsomycota and Microsporidia respectively. Next, the FASTA protein sequences were aligned and Phylip software was used to generate the phylogenetic tree. The phylogenetic trees show the species belonging to the phylum Ascomycota comprising the different class divisions, Eurotiomycetes, Leotiomycetes, Sachharomycetes and Microsporiodia respectively. In the figure 4, the evolving relations between the Saccharomyces Cerevisiae and another 11 Species. The Phylogeny tree shows that the main species S. Cerevisiae and Candida Glabrata are diverged from the common ancestor which means a single population lineage splits into two descendant lineages [7]. Also both the species have their own specific properties like C.Glabrata is an infectious fungal and S. Cerevisiae is used in fermenting alcohols, baking etc. Moreover, comparing the species of Class Eurotiomycetes with Class Saccharomycetes, Aspergillus Nidulans and Coccidioides Immitis from Eurotiomycetes are evolved into Class Saccharomycetes with more genetic changes and more properties over time. From one common ancestor species C.Immitis is diverged and ancestor is again diverged into two descendant lineages , one showing species A. Nidulans and second descendant lineage diverged into the Class Saccharomycetes[7]. By observing the phylogenetic tree (fig 4), it can be seen that Class Leotiomycetes is diverged into Classes Eurotiomycetes and Saccharomycetes. Considering, Figure 5, evolutionary relationship between the Fungi Schizosachharomyces Pombe and 11 species from Class Sachharomycetes, Eurotiomycetes and Leotimycetes respectively. It can be observed that the fungal species Schizosaccharomyces Pombe and the other 7 species from Class Saccharomycetes diverged from a common ancestry. The Species Candida Albicans and Debaryomyces Hansenii diverged into separate descendant lineages

Table 1. Demonstrates the 14 different fungal species with their protein sequence based on the sizes in Mb

Fungal Species	Protein Sequence Size	CLASS	PHYLA
Aspergillus nidulans	5.57	Eurotiomycetes	Ascomycota
Botrytis cinerea	7.46	Leotiomycetes	Ascomycota
Candida glabrata	2.92	Saccharomycetes	Ascomycota
Candida albicans	3.27	Saccharomycetes	Ascomycota
Coccidioides immitis	4.87	Eurotiomycetes	Ascomycota
Debaryomyces hansenii	3.25	Saccharomycetes	Ascomycota
Encephalitozoon cuniculi	876kb	Microsporea	Microsporidia
Eremothecium gossypii	2.52	Saccharomycetes	Ascomycota
Kluyveromyces lactis	2.73	Saccharomycetes	Ascomycota
Saccharomyces paradoxus	2.96	Saccharomycetes	Ascomycota
Schizosaccharomyces pombe	2.67	Saccharomycetes	Ascomycota
Sclerotinia sclerotiorum	6.19	Leotiomycetes	Ascomycota
Yarrowia lipolytica	3.39	Saccharomycetes	Ascomycota
Saccharomyces cerevisiae	3.23	Saccharomycetes	Ascomycota

from common ancestor and Kluyveromyces Lactis and Candida Glabrata, S.Paradoxus from common ancestry but C. Glabrata and S. Paradoxus diverged into separate lineages. From the 7 species of Sacchromycetes on top in fig.5 the S. Pombe has more genetic changes over the time according to the branch length. Concluding, the Class Leotiomyctes evolved into the Class Eurotiomycetes and Class Euriomycetes further diverged into the Class Saccharomycetes of the Phylum Ascomycetes [7]. Apart from this, the molecular clock can estimate the valid time period for each species when it is evolved and tells about the age of the species. In this way, Phylogeny trees can be used to study the relationship between different species over time.

A. Expected Benefits

Gene duplication is a major force in evolution. Gene duplication can produce new genetic material for mutation, drift, and selection to operate on, leading to the emergence of specialized or novel gene functions. Without gene duplication, species might face difficulty to survive in the changing environmental conditions. Gene duplication helps to survive the thriving conditions of the environment like climate change [2]. The project helps to gain a deeper understanding of how fungi differ genetically and how they have evolved over time. It also helps to understand morphology and biological activity and their convergent evolution with fungal plant pathogens. By studying the genes, proteins, and essential functions in different fungal species can lead to new ways to use fungi in areas such as biotechnology, and its application. Fungi are lower eukaryotes and are classified as the fifth kingdom by modern biologists because of their absorptive mode of nutrition. In bioprocessing and manufacturing processes like baking and brewing, mostly fungal species from phylum Ascomycetes are used. In the medical field, fungi are used for producing organic acids, antibiotics,enzymes, and among other things. Here, the term "mycotechnology" is a reference to the significant influence that fungi have had on biotechnology [8].Thus, the possible identifications of unique fungal genes and proteins could lead to the new development of a new antifungal drug or biocatalyst for industrial processes. Apart from the mycotechnology, the study of genes of different fungal species can help to produce viable hybrids by fusing the closely related species. The fusing of two different species may give birth to a new species with the properties of two species or may develop a brand new species with more beneficial properties. Furthermore, through this approach, we can discover new fungi that produce similar or better medical compounds with reduced toxicity, which could lead to the development of safer and more effective therapeutic agents. To conclude, the study of protein clusters of fungal species provided in (Table 1) can also help to understand various biological functions and how protein structures are different or evolved over time in the provided fungal species.

B. Conclusion

In conclusion, the success of the is illustrated by a source code which aims to provide a framework for investigation of various fungus species from the Classes Saccharomycetes, Leotiomycetes, and Euriomycetes of the Phylum Ascomycota and may be employed for any future mycological studies. Understanding the evolutionary connection between the two important species, Saccharomyces Cerevisiae and Schizosaccharomyces Pombe, with all 12 species was made more accessible by using the proteome FASTA sequences to construct Phylogenetic trees. From the length of the phylogenetic tree, this study has shown that the two major species have higher genetic variety than all other species, and that the Class Leotiomycetes diverges into the Class Eurotiomycetes and the Class Eurotiomycetes into the Class Saccharomycetes of the Phylum Ascomycota.

5. REFERENCES

[1]National Human Genome Research Institute, “Duplication,” Genome.gov, 2019. https://www.genome.gov/genetics-

glossary/Duplication

[2]S. MAGADUM, U. BANERJEE, P. MURUGAN, D. GANGAPUR, and R. RAVIKESAVAN, "Gene duplication as a major force in evolution," Journal of Genetics, vol. 92, no. 1, pp. 155–161, Mar. 2013, doi: https://doi.org/10.1007/s12041-013-0212-8.

[3]T. K. Mohanta and H. Bae, "The diversity of fungal genome," Biological Procedures Online, vol. 17, no. 1, Apr. 2015, doi: https://doi.org/10.1186/s12575-015-0020-z. [4]M. J. Cornell et al., "Comparative genome analysis across a kingdom of eukaryotic organisms: Specialization and diversification in the Fungi," Genome Research, vol. 17, no. 12, pp. 1809–1822, Dec. 2007, doi: https://doi.org/10.1101/gr.6531807.

[5]M. Balaban, N. Moshiri, U. Mai, X. Jia, and S. Mirarab, "TreeCluster: Clustering biological sequences using phylogenetic trees," PLOS ONE, vol. 14, no. 8, p. e0221068, Aug. 2019, doi: https://doi.org/10.1371/journal.pone.0221068.

[6]D. Martin et al., "Jalview 2.11 Manual and Introductory Tutorial," 2019. Accessed: Apr. 16, 2023. [Online]. Available: https://www.jalview.org/tutorial/TheJalviewTutorial.pdf

[7]A. D. Scott and D. A. Baum, "Phylogenetic Tree," Encyclopedia of Evolutionary Biology, pp. 270–276, 2016, doi: https://doi.org/10.1016/b978-0-12-800049-6.00203-1.

[8]J. W. Bennett, "Mycotechnology: the role of fungi in biotechnology1Based on a lecture held at the symposium, 'Progress in US Biotechnology', at the 8th European Congress on Biotechnology (ECB8) in Budapest, Hungary, August 1997.1," Journal of Biotechnology, vol. 66, no. 2–3, pp. 101–107, Dec. 1998, doi: https://doi.org/10.1016/s0168-1656(98)00133-3.

[9]S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0," Systematic Biology, vol. 59, no. 3, pp. 307–321, Mar. 2010, doi: https://doi.org/10.1093/sysbio/syq010.

[10]B. Pelosi, "Developing a bioinformatics pipeline for comparative protein classification analysis," BMC Genomic Data, vol. 23, no. 1, Jun. 2022, doi: https://doi.org/10.1186/s12863-022-01045-x.

[11]D. Moraa, "Creating a Linux Virtual Machine Compute Engine in Google Cloud Platform," Medium, Mar. 22, 2022. https://diana-moraa.medium.com/creating-a-linux-virtual-machine-compute-engine-in-google-cloud-platform-f8059687feb8 (accessed Apr. 16, 2023).

[12]P. J. A. Cock et al., "Biopython: freely available Python tools for computational molecular biology and bioinformatics," Bioinformatics, vol. 25, no. 11, pp. 1422–1423, Mar. 2009, doi: https://doi.org/10.1093/bioinformatics/btp163.

[13]J. S. Bernardes, F. R. Vieira, L. M. Costa, and G. Zaverucha, "Evaluation and improvements of clustering algorithms for detecting remote homologous protein families," BMC Bioinformatics, vol. 16, no. 1, Feb. 2015, doi: https://doi.org/10.1186/s12859-014-0445-4.

6. APPENDIX (COMPUTER PROGRAMS SUMMARY)

A. Softwares

- Programming IDE: IntelliJ IDEA Ultimate, PyCharm Professional
- Blast: ncbi-blast-2.13.0+
- Sequence Alignment Visualization: Jalview
- Multi-Sequence Alignments: Clustal Omega
- Maximun Likelihood: PhyML-3.1

B. Blast Running Shell Scripts

Building Database:

```
makeblastdb -in sequence.fasta -parse_seqids
-dbtype prot -out /to/database/dir
```

Blast Sequences Against Customized Database:

```
blastp -db to/database/dir -query seq.fasta
-outfmt 10 -out result.csv
```

C. Results Handling Java Program

BestResults.java: using data structures and algorithm to extract best results (i.e. significant protein alignments).
Clusters.java: using data structures and algorithm to generate protein cluster information
SeperateCluster.java: parse cluster information and generate individual cluster by adding protein sequences.

D. Multi-sequence Alignments

Shell Script Example

```
./clustalo -i "input.fasta" -o output.fasta --auto -v
```

E. Maximun Likelihood

Using Biopython convert

```
.fasta file to .phylip file
SeqIO.convert(input_file, "fasta", output_file, "phylip")
```

Shell Script Example

```
./phyml -i cluster.phylip -d aa -m LG -b 100 -o n
```

The result is saved as newick file which can be interpreted using Biopython

F. Phylogenetic Tree

Python program example:

```
tree = Phylo.read("tree.txt", "newick")
plt.figure(figsize=(50, 50))
Phylo.draw(tree, do_show=False)
Save as PNG
plt.savefig('tree.png', dpi=1000)
Related Libraries
from Bio import Phylo
import matplotlib.pyplot as plt
```

G. Source Code

https://github.com/LinyunLiu/Bioinformatics