COM5216 Group Assignment report

Group project name: Serendipities Gene
Group number: 26
Group members:

Shengyuan Sun: 460257820

Kai Liu: 470369474

Yanlong Guan: 450481303

Background:

Nowadays, medical field became increasingly significant in our daily life. Some several problems such as mutation have already become more and more attention for biologists to research. In biology, mutation is a permanent changing on the nucleotide sequence of the genome of virus or organism and another kind of genetic elements. Mutations can result from errors in the DNA replication and other kinds of damage to the DNA, and some of the damage maybe undergo micro-homology-mediated or make an error on the other form of repair. Due to the mobile genetic elements, mutations can also result from deletion or insertion of DNA's segments. Mutation can change many different types in sequences. It can also change the gene product as well as prevent the gene from functioning properly. On one hand, mutation can include the duplication of DNA's large sections. This includes the genetic recombination and these duplications are a major source of material which can evolve new kind of genes. On the other hand, the mutation may change the gene on the virus and cause problems or diseases that can be throat to human beings. It can also change the DNA of virus and bacteria that can cause the resistance to drugs.A problem which is worthy to solve is that how to create an app that can search for the gene and make a comparison with gene sequence and gene combination.

1. What is the problem that my app will solve?

At the beginning of creating this kind of app, some problems surrounding us and we provide several solutions about how can users compare the gene sequence using the app on the mobile phone, as well as check the gene database at any time. That is to say, users can use our created app to search for the gene data at anywhere, anytime. In addition, we should achieve a goal which is that users' group can share their genetic discovery together as well as only for specific group members rather than other group's users. Finally, we should overcome a challenge that how to modify bacteria's resistance gene content. These kinds of questions are the priority challenges for us to overcome.

2. Why does this problem matter/background and requirement?

During using this app, users eager for the connection between the sequence of evolution and bacteria as well as a similar project for bacterial strain using the visualization. Nowadays, because of the swift update speed, it may lead to the browse the trail of information. Therefore, in order to increase the efficiency of work as well as liberating the researchers' lost productivity, we should add the updating data into the mobile phone which can solve this problem. Apart from this, safety, privacy, and management in data sharing also take a significant role in providing this service as well as extending to more kinds of platforms. Users should believe that our app can do with the problem and hidden danger including the privacy and safety problem. For example, each group has its own cloud folder and their own password to enter it. Thousands of group use this app and check their

documents separately and would not interrupt each other as well. Finally, we should avoid the wrong result be required by data which would interrupt the estimate. Thus, this app should be updated at the real-time in order to make sure the latest and up-to-date data.

3. According to these problems, how can this app solve?

In terms of comparing with gene sequence, we search for plenty kinds of measurements. After choosing and comparing, we finally use the phylogenetic tree as our measurement. In terms of phylogenetic tree, there are plenty kinds of measures that can build the tree, each method has its benefits and drawbacks.

a. Maximum parsimony：

Maximum parsimony is a standard which would be selected for the most wonderful choice. In this maximum-parsimony criterion, this optimal tree would try to minimize the quantity of homoplasy. This method is a simple and intuitive criterion. Therefore, it is popular for public to use. However, a major problem is that especially for paleontology, the maximum parsimony assumes the only two species which at the same position can be share the same nucleotide. Therefore, this method would be used under limitation.

b. Maximum likelihood estimation:

Maximum likelihood estimation is a way to deal with statistical model by estimating the parameters. It makes the observations which is given to the parameters and searches for the parameter values that is maximize the likelihood. This method is a useful way to estimate but this need a model as a statistical model and it does cost long time waiting for the results.

c. neighbor-joining method:

This kind of method usually used trees which is for DNA sequence protein data. It has three steps about first, second and final joining then make a conclusion about the additive distances.

This is a fast way to compare the least squares as the maximum parsimony method. If the input distance matrix in property is correct, the output tree will be also correct. It saves much time when using it and as for molecular clock hypothesis, this method does not assume the lineages in the similar rate.

Compared and considered with these four kinds of method which can build the evolutionary tree, we use homology analysis and use the unrooted tree to build evolutionary tree, combined with the neighbor-joining method.

As for checking data at any time as well as sharing data, we decide to use the Cloud service to solve this problem. We compare with many cloud platforms such as Alibaba Cloud, AWS and Azure Cloud.

As for Alibaba Cloud, it is the ancestor of China Cloud company. However, at first, the company decided to use the PaaS technology. After a while, they realized IaaS is much butter than PaaS, then they try to change it but they miss the good opportunity for Cloud. Apart from this, we try to use AWS as our major storage cloud. After the realization of this Cloud, we found that this cloud has many kinds of products in their application market and it can help IT users expand the traditional IT field. However, the cost of this Cloud is so high, compared with other cloud. In addition, some of its

architecture and implementation methods have not been able to adopt the latest technology, so the cloud's performance does not reach the high level.

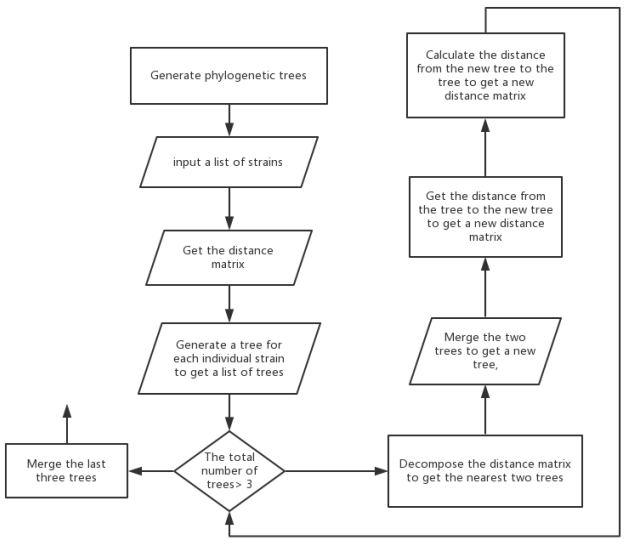
Finally, after selecting and analyzing, we choose the Azure Cloud for Microsoft company. This kind of cloud company has a great number of core engineers and it has abundant business experiences. Apart from this, it has been used in many major worldwide companies. By the way, the cost of this Cloud is cheaper than any other Cloud companies and the storage capacity is larger than others.

Apart from choosing cloud platform and phylogenetic tree, we should make sure that cloud data can save resistance gene information as well as real-time examination of new version updating.

4. How is our solution implemented?

A. Generate tree:

Function name	Build Phylogeny tree
Priority	extremely high
Business correspondence	comparison with the relationship between strains and get the phylogeny tree
Basic requirement	get phylogeny tree by NJ method
Constraint	input data is fasta
External interface	show the function of phylogeny tree



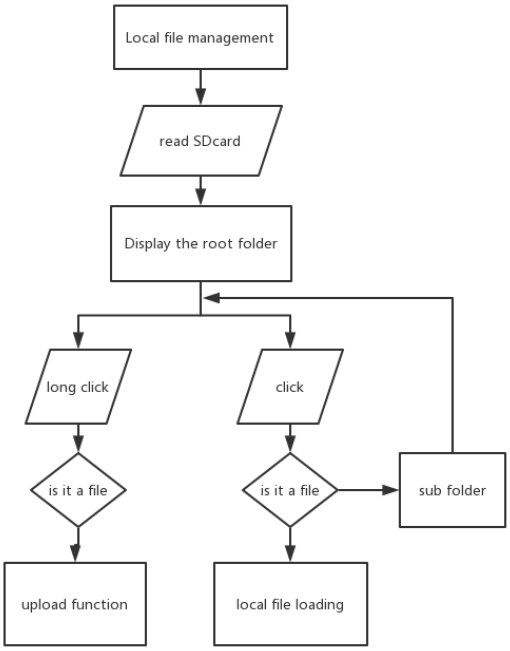
Name	Species	Method	Explanation

util	MatrixArr	removeTwoAddOne	Combine the nearest two trees in the distance matrix into a new tree and compute the new matrix
		findLowestValuePoint	Look for the nearest two trees in the matrix
		initQuadraticWithValue	Initializes a matrix that determines the size of each position in the matrix for a given value
sdCardutil	ReadFromSDCard	getSAAlist	Require all the gene sequences in a file
genetree	NJArr	calculateQMatrix	Get the two places of the minimum distance of the matrix
		updateClusters	upload phylogeny tree
		updateDistances	Update distance matrix
		generateFinalTree	When the last three trees are planted, the final tree will be merged
		distancesToNewNode	Calculate the distance from the tree to the new tree in the distance matrix
		distanceFromNewNode	Calculate the distance from the new tree to the other tree

B. local file management

Function name	Local file management
Priority	high
Business correspondence	Basic file management, loading files, analysis of documents and other services
Basic requirement	1. Browse the local folder structure 2, display the target file and general file 3, and extend the other functions of the service 4, database record
Constraint	1.Allow the operation of the fas file, and the file is not allowed to be modified and deleted.

External interface	Local database, fasta file analysis function

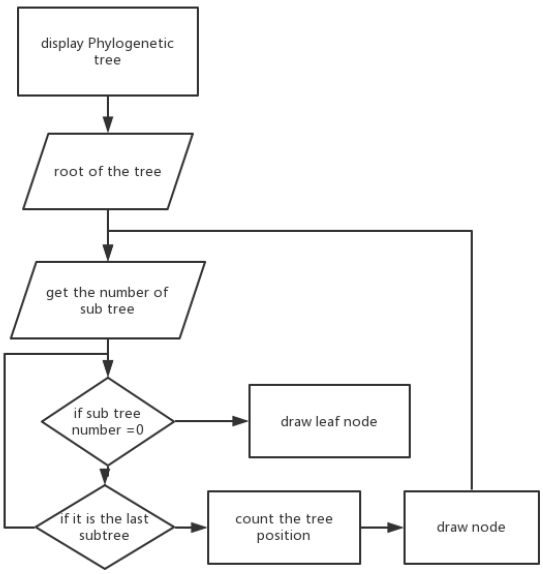


Name	Species	method	Explanation
sdCardutil	ReadFromSDCard	Checkfilebypath	Check the existent of the file
		FileFromSDCard	load file to become the input stream
		getSPNamelist	Gets the name of all the strains in a file
		getSAAlist	Get all the gene sequences in a file
databbaseutil	DatabaseHelper	AddOneFilePath	Record the name of the file and the location of the file
		AddOneSP	The name of the strain contained in the record file

C. view tree:

Function name	show the function of phylogenetic tree
Priority	Extremely high
Business correspondence	display of strain evolution tree by visualization

Basic requirement	Ensure that the displayed tree matches
Constraint	
External interface	file io

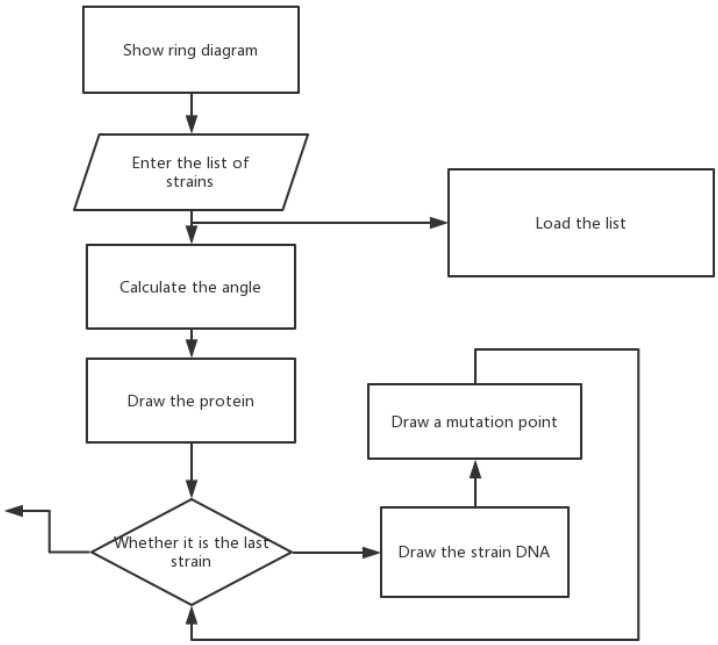


Name	Species	Method	Explanation
util	TreePrinter	renderNode	When traversing a tree, it is called when a node has a subtree and is used to continue to traverse the nodes of the next layer
		renderLeaf	
		indent	When traversing a tree, it is called when a node has no subtree
		DrawRoot	edit root node
		DrawNode	edit node
		DrawLeaf	edit leaf node

D. View circle:

Function name	Display annular and tabular
Priority	Extremely high
Business corripondence	Show strain information by visualization

Basic requirement	The list shows the information of all the resistance genes in the strain, and the annular image shows the gene and the location of the mutation
Constraint	
External interface	Database, file io

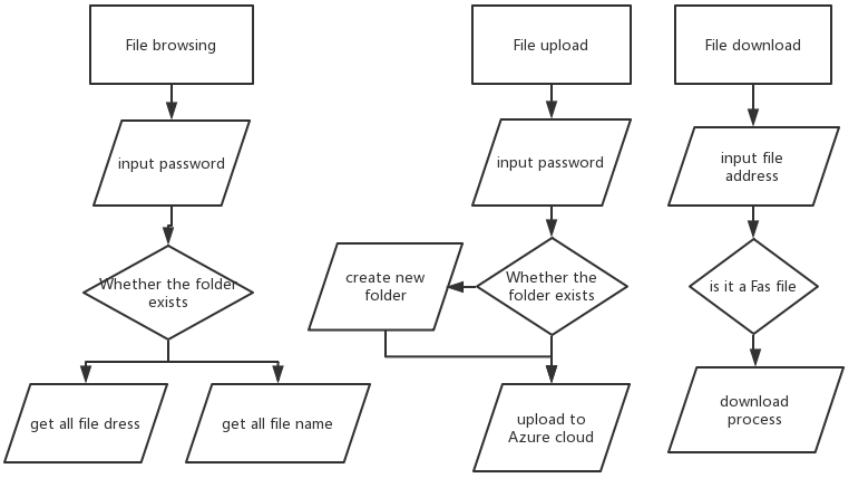


Name	species	Method	Explanation
views	VisualizationCircleView	initDrawCircle	Edit a protein ring
		DrawOneGene	Edit a single strain
		DrawCross	Mapping the variation points on the strain

E. cloud file management:

function name	Cloud documents management
priority	Extremely high
Business correspondence	safety and management during sharing the documents; check documents at any time
basic requirement	upload and download documents, folder password, documents in database

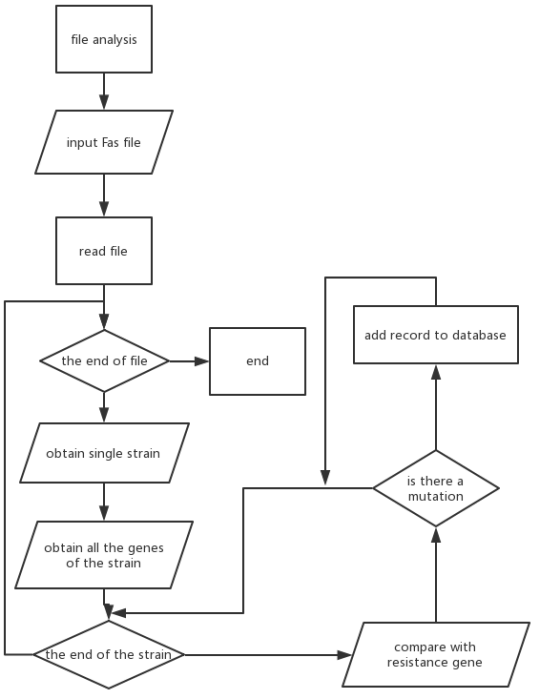
constraint	folder limited by password
External interface	data management and database



Name	species	method	explaination
azureUtil	AzureUtilNew	CloudBlobContainer	Encapsulates a method to connect to an Azure cloud container
		Uploadfasta	Upload documents to Cloud
		Listallfile	Displays all files in a specified container
		ListallfilePath	Displays all files in a specified container
	DownUtil	download	create a new thread to download a document for a specific file
		getCompleteRate	Require the process of downloading files in real time

F. file analysis

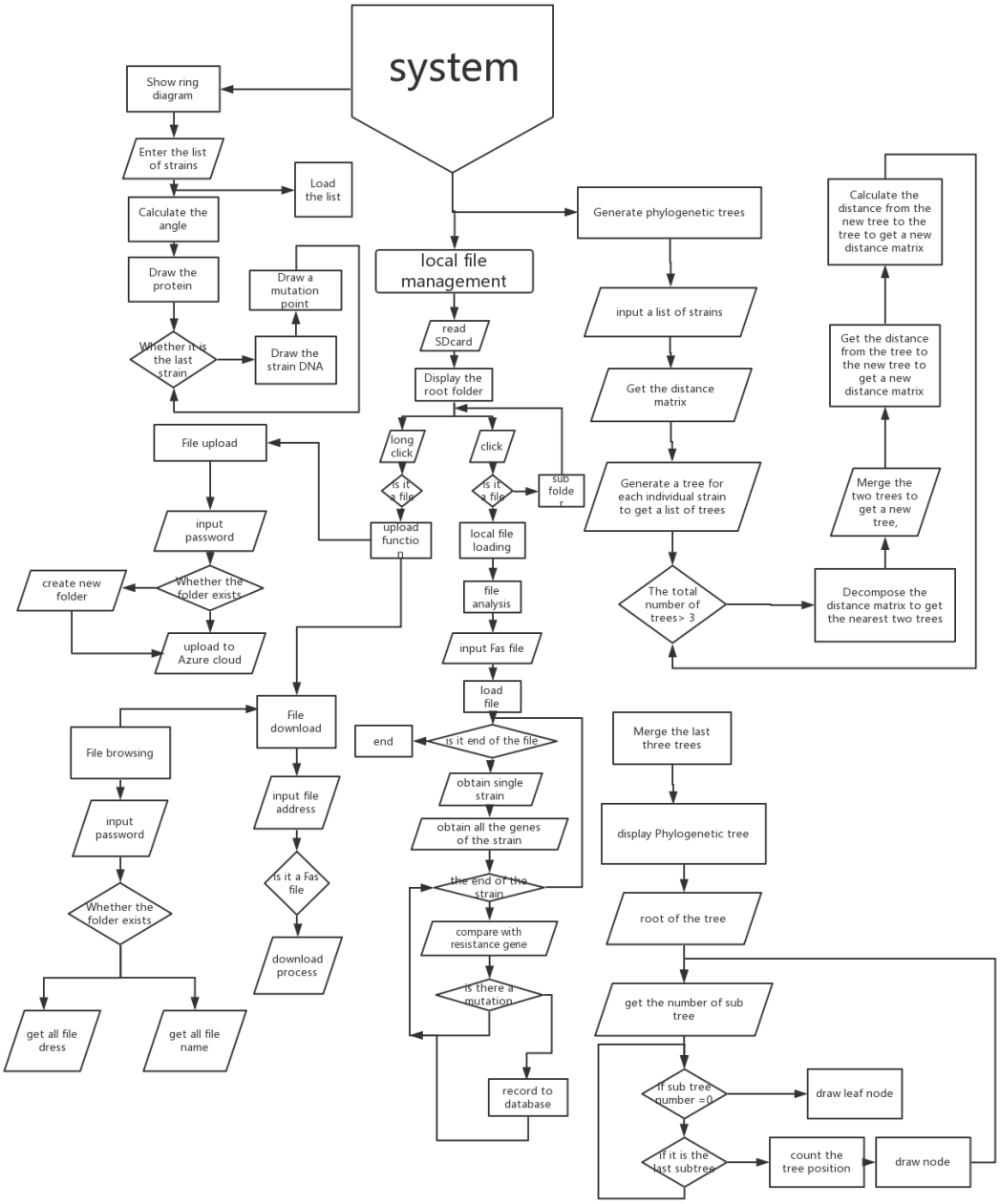
function name	Documents analysis
Priority	High
Business correspondence	comparison with gene sequence and present by visualization
Basic requirement	generate the genetic material, analyze gene resistance genes, record and save
Constraint	
External interface	Database, local file IO, cloud file IO



Name	Species	Method	Explaination
sdCardutil	ReadFromSDCard	FileFromSDCard	load the file
		getString	Load a portion of the fas file into a string
		getSAAlist	Get the DNA sequence of all the strains in a file
		getSPNamelist	Get the name of the entire strain in a file

databbaseutil	DatabaseHelper	AddOneSP	record a strain in a file
		AddOneMutationGene	Record a variation of information

5. How does our app work?



reflection :

We took about two weeks to finish this report. During writing this report, we search many documents from the Internet and books from the library. We require more knowledge about the algorithm, phylogenetic tree and cloud company and learn more about gene combination and gene sequences. In addition, we realize the advantages and drawbacks of each method to build the evolutionary tree and every Cloud company such as AWS, Alibaba, Azure Cloud and so on. During writing this report, we realize that we should make a brief introduction about the background information for gene mutation and phylogenetic tree. Then we should understand what do the hospital and research institution requirement for gene app. Next, we make a comparison with other apps that have been used in the society as well as comparing and selecting the appropriate cloud and method which can build the evolutionary tree. Therefore, this report includes our conception and measurement during creating this app as well as expansion about the algorithm.

After we finishing creating the new app for the gene, we find that it is amazing for the connection between medical domain to IT app. Scientists and researchers can use this app to achieve such requirement about gene sequence and gene combination. Of course, the code for operating the app is really complicated and cost us a period of time to overcome it.

In terms of presentation, we search some researchers and some videos on the Youtube about how can we explain our topic and system works in 3 minutes, as well as some presentation skills such as how to talk briefly and how to dress formally. We also look for powerpoint modules that can make our powerpoint tidy and awesome.

Obviously, there are also plenty of serious problems we faced. A big problem is that we can not find the correct method to use Azure Cloud. Then we search the Internet and check the documents for Official technology. Additionally, we take about one week to find the connection between medical treatment and information technology as well as a necessity for connecting them.