Order Tracking For Genome Assembly

Saira Tariq
dept.of SBE
UMT University
lahore, Pakistan
F2019313054@umt.edu.pk

M.Fahed Chaudhry
dept.of SBE
UMT University
lahore, Pakistan
F2019313046@umt.edu.pk

Hamza Arif
dept.of SBE
UMT University
khewra, Pakistan
F2019313011@umt.edu.pk

Farhat-ul-ain
dept.of SBE
UMT University
Multan, Pakistan
F2019313048@umt.edu.pk

I. Introduction

It is the foundation for the development of genetic resources such as gene annotation, high resolution maps of polymorphism, genomic structural variation, etc. These resources enable a wide range of applications involving genomic, transcriptomic, or epigenomic analysis in fields including biomedicine, agriculture, biotechnology, molecular ecology, and evolutionary biology1. Ideally, a fully sequenced genome, with long contiguous genomic segments anchored to full-length chromosomes should be produced, often by combining sequencing and mapping technologies. Such a genome project demands substantial funding, which is often reserved for medical or agricultural research. However, the rapid development of Next Generation Sequencing (NGS) over the last decade provided a relatively affordable and powerful tool fitting also for non-model organism research, where lower-quality draft genome assemblies are typically produced. Two main challenges affect de novo assembly of eukaryotic genomes: (1) repetitive sequences, including gene duplications, transposons, and short sequence repeats; and (2) polymorphism, including single-nucleotide polymorphisms (SNPs), insertions and deletions, and large genome rearrangement polymorphisms. Although sequencing technologies have advanced dramatically in the past decade, these issues still present a major hurdle, resulting in highly fragmented assemblies.

II. Methods

A. Data Details

we have one file with four sheets. There are four data-sets first is NBB4-Plasmid Second is Hamburgensis X14 third is Vibrio Cholerae and the last one is PAb1.First three data-set have 10 observations and 13 number of variables and PAb1 have 4 observations and 13 number of variables.

B. Normalization

Normalization is a pre-processing stage of any type problem statement.Normalization takes important role in the field of soft computing, cloud computing etc. for manipulation of data like scale down or scale up the range of data before it becomes used for further stage. There are so many normalization techniques are there namely Min-Max normalization, Z-score normalization and Decimal scaling normalization. As per Min-Max normalization technique,

$$A = (\frac{A - minvalueofA}{maxvalueofA - minvalueofA}) * (D - C) + C$$

Where A contains Min-Max Normalized data one. If predefined boundary is [C,D].If A is the range of original data B is the mapped one data.The technique which gives the normalized values or range of data from the original unstructured data using the concepts like mean and standard deviation then the Parameter is called as Z-score Normalization. So the unstructured data can be normalized using z-score parameter, as per given formulae:

$$vi = (\frac{vi - E}{std(E)})$$

Where,

$$std(E) = \sqrt{\frac{1}{n-1}\Sigma^n i \equiv 1(vi - E^2)}$$

The technique which provides the range between -1 and 1.

$$v^i = \frac{v}{10^j}$$

Where, vi is the scaled values
v is the range of values
j is the smallest integer Max($|vi|$) <1

C. Sampling

In statistics, quality assurance, and survey methodology, sampling is the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population. The whole data set is known as population, sometimes the population is too large that we cannot perform the analysis using the whole dataset so we usually take a small portion of the dataset known as to be sample. We take certain number of instances or observations in the sample, in such a way that the results of the analysis can be generalized to the whole population. In other words the sample should be a true representative of the whole population i.e. whole dataset. There are multiple ways to select a sample which will fulfill this condition. Statistical sampling is a large field of study, but in applied machine learning, there may be three types of sampling that you are likely to use: simple random sampling, systematic sampling, and stratified sampling. Simple Random Sampling: Samples are drawn with a uniform probability from the domain.However in our case the dataset is not that large, it will not be an issue for us to use the whole dataset as sample but we usually don't use the whole dataset. In machine learning we usually divide the whole dataset in two parts, one part is the training dataset on which we will train our algorithm and the other part is the testing dataset on which we will check the accuracy of our algorithm. The only question remaining is that in what quantity the data should be divided in each part. So in our project we are divid the dataset in 80 % and 20 % , 80 % data in training data and 20 % in testing dataset. For this purpose in python we use the following commands. In our project we are using "simple random sampling technique" and this is done by using the library sklearn. Splitting and simple random sampling both are done using this library.

D. Feature Extraction

In sampling we take care of the huge amount of data. In feature extraction/selection we do the same for but with the variables. In machine learning we are usually faced with huge datasets which contain a lot of variables or usually called as dimensions or features. Feature selection/extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process.Feature selection and feature extraction usually achieve the same purpose but they are different in their nature of the process. Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set. Feature selection is for filtering irrelevant or redundant features from your dataset. Again, feature selection keeps a subset of the original features while feature extraction creates new ones. In our project if we run the regression using all the variables the results are not useful because not all the variables help us in getting the result. That is because of the collinearity. In simple words there are multiple variables which negatively correlated to the predicted variable and some are highly positively correlated. So to find out which variables are not useful for us we can use multiple ways which can tell us the correlation values of all the variables with the predicted variable. We choose the correlation heat map. Correlation heat map is probably the easiest way to do this as you can easily distinguish between the useful and not useful variables though visuals shown by the heat map. It represents the high values using colors i.e. red color for highly negative values -1 and green color for the highly positive values +1. So we choose those variables which are closer to 1 because they are highly correlated and they are represented by green color.

E. Regression

A measure of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables. Regression takes a group of random variable and predict y. Equation is

$$y = ax + b$$

.and you tell the relationship throw state line. we use regression because we tell the relationship between independent and dependent variables relation explaining we use regression Analysis.In the regression analysis produce a regression equation the coefficient show the relationship of every independent variable with dependent variable.In our dataset used regression have different variables we perform analysis on these variables and check our order or right therefore we perform regression analysis on our dataset.How we apply a regression in dataset we use the libraries of python and applied a Multiple linear regression.We use multi linear regression we used because we have number of independent variables are more than one when you have more than one variable you select a feature and perform analysis and apply multiple linear regression and you can takes the results. There are different techniques to apply regression like ordinal regression , linear regression or multiple regression.In this part we can use multiple linear regression.

F. Accuracy

We check accuracy throw our predicting model how much our results are accurate. we check accuracy in two ways first one is library of python .score. In .Score we predict y and y-test the library compare the results of y-predict and y-test and tell how many percent your result is accurate .for example NBB4 plasmid data accuracy is 0.997 and other also checked like this.Pab1 we do not apply .score because the data set are small and test set have just one result score work on R^2 if you don't have R^2 you cannot use the library of .score and other three test data set result have more than one result therefore they can calculate the accuracy and can run the .score library. Its mean if you don't have R^2 you cannot use .score library. Second one is RMSE

$$RMSError = \sqrt{1 - r^2}SDy$$

which measure the every deviation of the actual data point from regression line.In this one we just import the one library and one more thing in RMSE if we have more than 1 more than 5 ,10 or 100 entries don't create any difference.RMSE subtracts the predicated value from actual value and take square that tell the how much training and testing have correlation.if you get the less RMSE you have good result. For example NBB4 plasmid have result 0.159 its really near to Zero and have less correlation therefore it's a very good model. Hambergenisis result is near to inaccurate the reason is multicorrlinearity is greater .

Genome Assembly	reg.score(X, y)	RMSE
NBB4 Plasmid	0.997	0.159
Hamburgenesis	0.851	0.577
VibroChlorea	0.946	0.464
PAb1	Nan	0.403

Table:1

III. Results And discussion

We are given four datasets in which there are different genome assemblies and we had to make an algorithm by which we can rank the assemblies from best to worst. We simply want to do a regression analysis by which we will analyze all the variables and choose the most useful variables for our analysis and by which we can successfully predict the dependent variable which in turn tells us which assembly is the best and which one is the worst. We started with the raw data and then normalized it so that it can be used properly in our analysis. Then we divided the data in the training and testing datasets. The regression algorithm was performed on the training dataset and then to confirm that whether it was giving the right results, we used the test dataset. The regression analysis on all the datasets were performed separately and in each dataset the assemblies were ranked from best to worst. We ran the regression algorithm to find that out. We took “Order” variable as the dependent variable and then by feature selection we chose the best variables to predict the dependent variable.In the dataset NBB4 plasmid, we had total 10 assemblies from which we were to rank best to worst. We used the correlation heat map and we got three variables “No. of Con. Tigs, $Contigs \geq N50$ and Sum of the Contig Lengths” which we used for the regression analysis. In any of the regression analysis the accuracy of the result can be measured by using the root mean square error. Lower the RMSE the better the result. In this analysis our RMSE is 0.159 which is considered to be very low. Then we ran the algorithm on the test dataset and the results were quite impressive. In one particular case the test value was 3 and the predicted value through our algorithm was 3.0069 which can be considered as 3 and the other test value was 9 and our predicted value was 9.25249 which can also be considered as 9. So we can say that we were quite successful in predicting the rank of the assemblies from best being 1 and worst being 10. In our dataset Mira2 is the best assembly and at second is MARAGAP and Maq is at third while Mira is the worst i.e. number 10. In the dataset Hamburgensis X14, we have similar number of assemblies i.e. 10 and we wanted to find which one was best and which one was worst. So we ran the regression analysis on this dataset as well, but first we found out that which variables give us the best results and which didn’t give us the best result we ignored them for our final regression analysis. We used correlation heat map and it gave us two variables which give us the best results which were No. of Cont.Tigs and $Contigs \geq N50$. By using these variables we successfully conducted regression analysis with a very low RMSE which was 0.577521. Then we ran this algorithm on test dataset to find out that to which extent our algorithm can predict the dependent variable successfully. The results were not as good as for the NBB4 plasmid but they were satisfactory as the test value at that particular time was 6 and our algorithm predicted 6.7 and the other value was 3 and we predicted 2.58. So according to the predictions we can rank the assemblies from best to worst as Mira2 is still number 1, second is MARAGAP and Maq is at third while Mira is the worst i.e. number 10 same as they were in NBB4 plasmid but there is some difference in middle ranks between these two datasets.In assembly of Vibrio Cholerae there are 9 assemblies. To rank these assemblies from best to worst we use similar regression analysis. The correlation heat map for this dataset gave us four variables which play crucial role in the prediction of the rank of the assemblies from best to worst. These four variables are “No. of Con. Tigs, $Contigs \geq N50$, Contigs$>$ 200 bp and Sum of the Contig Lengths”. By using these variables we successfully rank the assemblies from best to worst by using regression. Our regression analysis for this particular dataset gives us the RMSE 0.464 which is quite decent as well. So we continue with our analysis to predict the dependent variable as the RMSE value for our regression analysis algorithm is quite low. We applied this algorithm on the test dataset and we got the results quite satisfactory as the tested value was 5 and the predicted value is 5.67 and the other value was 9 and the predicted value was 8.97. So we consider these results satisfactory and thus the rank of the assemblies were MARAGAP as the best assembly, Mira2 at the number 2 and at number 3 it was IDBA and the SSAKE being the worst assembly. For the assembly of Pseudomonas Aeruginosa or PAb1 we had 4 assemblies and we were to find the rank of these assemblies from best to worst. We did this with the regression analysis as well and for that we had to select the variables best suited for this job. We used the correlation heat map for this purpose yet again and found out that there are three variables “No. of Con. Tigs, $Contigs \geq N50$ and Sum of the Contig Lengths” which can be used to find the dependent variable in this dataset. So we use these variables to run the regression analysis and fortunately we were able to get the satisfactory results as well. The RMSE for this regression analysis was also low at 0.403, so we could continue further with our analysis. We ran the algorithm on test dataset as usual and the result was 2.403 for the tested value of 2. In this way the rank of the assemblies were that MARAGAP at number one, IDBA at number two, QSRA at number three and VCAKE at number four.

IV. Conclusion

Genome assembly is a complex field; only to understand the vocabulary of the field it can take several hours or even days for a beginner. We were fortunate enough that we got the chance to work on a real life dataset. Regardless of the results whether we were successful in our effort or not, we got to learn many aspects of working with some real life problem of data science. The dataset provided was not perfect and we had to do different things to get it in proper form. It was necessary as well because without it we could not apply regression analysis on the dataset. Once the data was clean we had to divide the dataset for training and for the testing of our algorithm. Then by feature selection we chose the best suited variables or features for this job. By using the proper variables for the prediction of the dependent variables through regression we were able to figure out the rank of the assemblies in different datasets. Our results might not be perfect but we tried our best to come up with different solutions to tackle the problem at hand.

References

[1] Wajid, Bilal, and Erchin Serpedin. ”Do it yourself guide to genome assembly.” Briefings in functional genomics 15.1 (2016): 1-9.

[2] Wajid, B., Serpedin, E.,Nounou, M. and Nounou, H. (2015) ‘MARAGAP: a modular approach to reference assisted genome assembly pipeline’, Int. J. Computational Biology andDrug Design, Vol. 8, No. 3, pp.226–250.Biographical

[3] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al: The sequence of the human genome. Science. 2001, 291: 1304-1351.

[4] Huang X, Wang J, Aluru S, Yang SP, Hillier L: PCAP: A whole-genome assembly program. Genome Res. 2003, 13: 2164-2170. 10.1101/gr.1390403.