ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

In-Silico Modeling and Prediction from Metagenomic Metadata.

Abhisek Ranjan Bera[1]*, Vineet Vishal[1], Pankaj K Singh[2]^, Damayanti Chakravarty[3], Dipabarna Bhattacharya[4], Hirak Jyoti Chakraborty[5], and Sayak Ganguli[6].

[1]Bangabasi Evening College, Kolkata-700009, India.
[2]Computational Biology Division, The Biome Kolkata - 700064, India.
[3]University of Southern Mississippi, Hattiesburg, USA.
[4]Institute of Biotechnology, University of Helsinki, Finland.
[5]Central Inland Fisheries Research Institute, Barrackpore, India.
[6]TCB Division, AIIST Palta - 743122, India.

ABSTRACT

There has been an exponential rise in the number of environmental diversity data using next generation sequencing technologies. Community level analyses of various ecosystems studied so far has exhibited species richness and microbial identification of uncultivable microbes. With the increase in the number of submitted datasets it has become imperative to dig deep into the various facets of the datasets that are available and identify parameters which contribute towards overall quality assessment of the data. In this work we construct regression models using the few common parameters which are present in almost all metagenomic analyses pipeline - total reads, annotated reads, unclassified reads, average read length and GC content. These models and equations should enable future workers to assess the overall quality as well as predict facets of their datasets in comparison to existing datasets.

Keywords: *Metagenomics, Regression Analyses, GC content, Annotated Reads, Unclassified microbes*

Corresponding author

INTRODUCTION

Understanding the earth and its life processes has been the quest of science and with every successful endeavour we inch closer towards the systems level understanding of the various organismic processes that contribute towards life on earth. Anthropogenic influences and other factors that adversely affect the environment preliminarily attack the existing microbial community and influence the microcosm of a particular area by creating an imbalance amongst the microflora which in turn affects the primary consumers and in turn affect the entire food chain of the place.

Environmental Genomics approaches have enabled the rapid identification of the various microbial members that contribute towards the microflora of a particular geographical or environmental niche. Metagenomics has enabled us to get an insight into the segment or group of microorganisms which could not be cultured in vitro, thus opening up new vistas into the understanding of form and function of that community.

With the application of shotgun metagenomic approaches in metagenomics, metatranscriptomics have emerged to enable the quantification of the total number of transcripts of corresponding genes which are overexposed in a particular community. This approach has not only contributed towards the elucidation of functional classification of the microbial cohort, but also has paved way for correlation between the soil characteristics and microbial community assessment.

Despite the advances in the various technologies and computational analyses pipelines a quick survey of accumulated metagenomics data reveals that almost 50% of the total data that is generated through sequencing of microbial communities, remains unclassified due to the lack of proper information. A large proportion of this unclassified portion of the data is constituted of the microbes which are not cultureable. Further annotation of the available sequences also reveal the presence of a high proportion of hypothetical proteins which do not have any experimental evidence. To counter this issue several attempts have been made over the years to process data using various statistical techniques (Table 1) with average to good impact.

Table 1: Statistical Methods used to study Metagenomics Datasets

Serial	Name of Method	Purpose of Use	Reference
1	K Means Clustering	Unsupervised method for classifying observations in K groups	[1]
2	Cross Validation using Classification Tree	Applicable to small metagenomic datasets	[2]
3	Supervised Random Forest	for identification of standard variables which can differentiate between groups	[3]
4	Mean Decreasing Accuracy	For estimation of Gini index of a particular variable and its contribution to the tree, reducing the chances of misclassification	[4]
5	Multidimensional Scaling	Visualization procedure similar to principal component analyses	[3]
6	Linear Discriminant Analyses	For prediction of group membership of new data	[3]
7	Principal Component Analyses	For reduction in the dimensions of data	[1]
8	Canonical Discriminant analyses	Estimation of variance between classes	[1]
9	Multilevel Regularized Regression	Taxa identification and Network Construction	[5]

In this work we use simple linear regression to formulate the regression equations for comparing the major components of metagenomic data sets for creating a standard for the use and analyses of the generated data.

MATERIALS AND METHODS

Twenty four metagenomic datasets were selected randomly from the SRA archive and metadata from the samples were collected. Metadata was then classified based on optimal occurrence and was narrowed down to five important parameters: Total Reads, Average Length, Annotated Reads, Unclassified Reads and GC content. Simple Linear regression was performed to formulate the best fit regression line which shall enable the prediction of soil metagenomics studies. In simple linear regression, we predicted the scores on one variable using the scores of the second variable. The variable which is predicted is referred to as the criterion variable and is denoted as Y. The variable on which the assumptions are being based on is the predictor variable and is denoted as X. Since here for all the regression equations there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the predictions of Y when plotted as a function of X form a straight line. Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line which is calculated using the following equation:

$$Y' = bX + A,$$

The slope (b) can be calculated as follows:

$$b = r\, sY/sX$$

and the intercept (A) can be calculated as

$$A = MY - bMX.$$

All calculations were made using R statistical software and the graphs were plotted using default application of Mac - Numbers.

Once the data was obtained the prediction accuracy of the regression line was checked by first calculating the prediction accuracy and error percentage using the regression equation that was obtained. Finally the differences between the observed and expected values were checked using the Chi Square Test.

RESULTS

From the five variables under study the comparison of the following set of variables yielded reproducible regression equations.

a) Total Number of Reads: Total Number of Annotated Reads
b) Total Number of Reads: Total Number of Unclassified Reads
c) Average Length of Reads: GC Content.

The other combinations did not produce any reproducible results indicating that the two variables chosen for the combinations were not suitable to act as the criterion and predictor variable and vice versa.

The equations obtained (Fig 2 A, B and C) were then retested in the datasets used to train the model as well as on ten different soil metagenome datasets.

The prediction accuracy obtained were found to be 90%; which indicates that the regression equations generated in this present study holds good for testing fresh metagenomics datasets (Fig 1).

Table 2: Base Parameters used as criterion and predictor variables and their corresponding values

Sites	Total Reads	Average length	Annotated Reads	Unclassified Category	GC percentage	Reference
1	7,05,326	256	268157	302390	50	[6]
2	514784	255	196211	221259	50	[6]
3	1267409	566	511924	577278	50	[6]
4	1416928	563	569172	641832	51	[6]
5	854451	558	358555	404328	51	[6]
6	1045353	542	420389	474055	51	[6]
7	249993	235	80936	169057	55.75	[7]
8	231233	238	69953	161280	54.64	[7]
9	214921	248	69600	145321	56.36	[7]
10	217605	223	58575	159030	54.66	[7]
11	20857	349	4936	4742	57	[8]
12	25787	348	7331	6768	58	[8]
13	27348	342	4467	2531	55	[8]
14	23830	348	7663	5325	57	[8]
15	20179	349	4697	4888	61	[8]
16	334386	105	23	334363	49	[9]
17	388627	99	148	388479	44	[9]
18	351205	105	74	351131	46	[9]
19	209073	226	99310	109763	40	[9]
20	221744	239	117524	104220	39	[9]
21	782404	411	464748	317565	59	[9]
22	619288	310	363522	255766	62	[9]
23	217605	222	72898	144707	54	[9]
24	280753	190	24426	256327	52	[9]

The chi-square analyses that was performed indicated that the differences in the observed and expected samples were arising by chance and that the regression equation can be used for subsequent analyses

DISCUSSION AND CONCLUSION

Recently Liu et al. (2015) has used multilevel regularised regression for determining the network and selecting taxa from metagenomic count information. Their approach is aimed towards compelling of disease associated taxa and networks and is generally applicable after the sample has passed the quality control stage. The regression equations obtained in this study precisely enables the user to predict the possible number of outcomes following the generation of the QC passed read count.

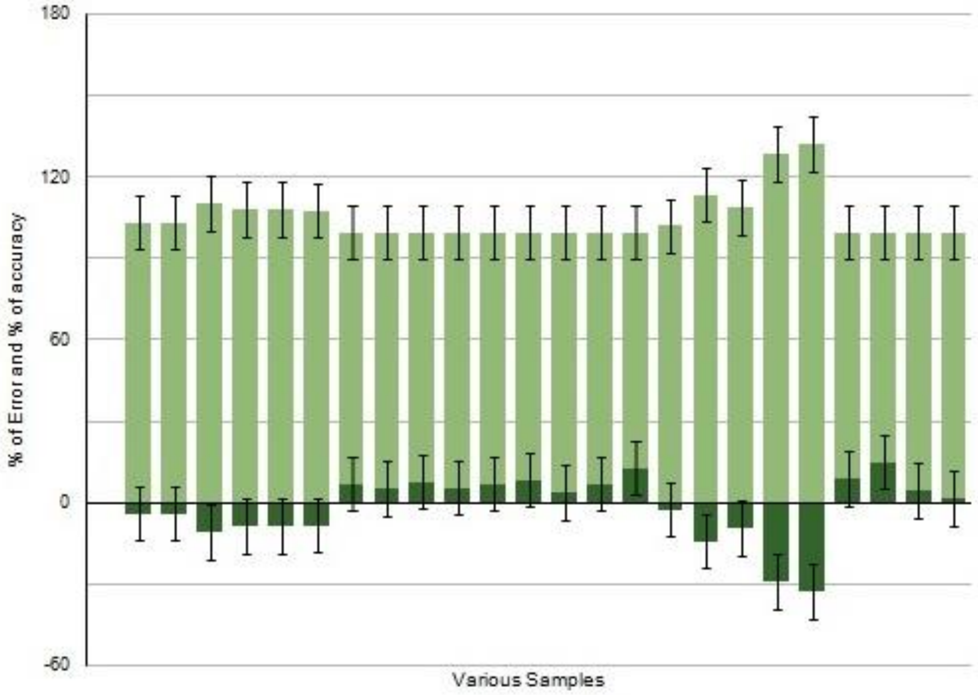

Figure 1: Prediction Accuracies and Error %

The high prediction accuracy of the equations also enable them to be used as a second level predictive quality control measure of metagenomics datasets. Goll et.al. (2010) in their report on METAREP, used regression analyses as a measure for elucidating differential abundance of taxa. Three independent regression equations were obtained in this study which displayed an average accuracy of prediction as high as 90%. Chi square analyses established that the differences in observation were arising by chance alone and thus we can safely conclude that these equations can be used further for the prediction of multi parametric metagenomic datasets.

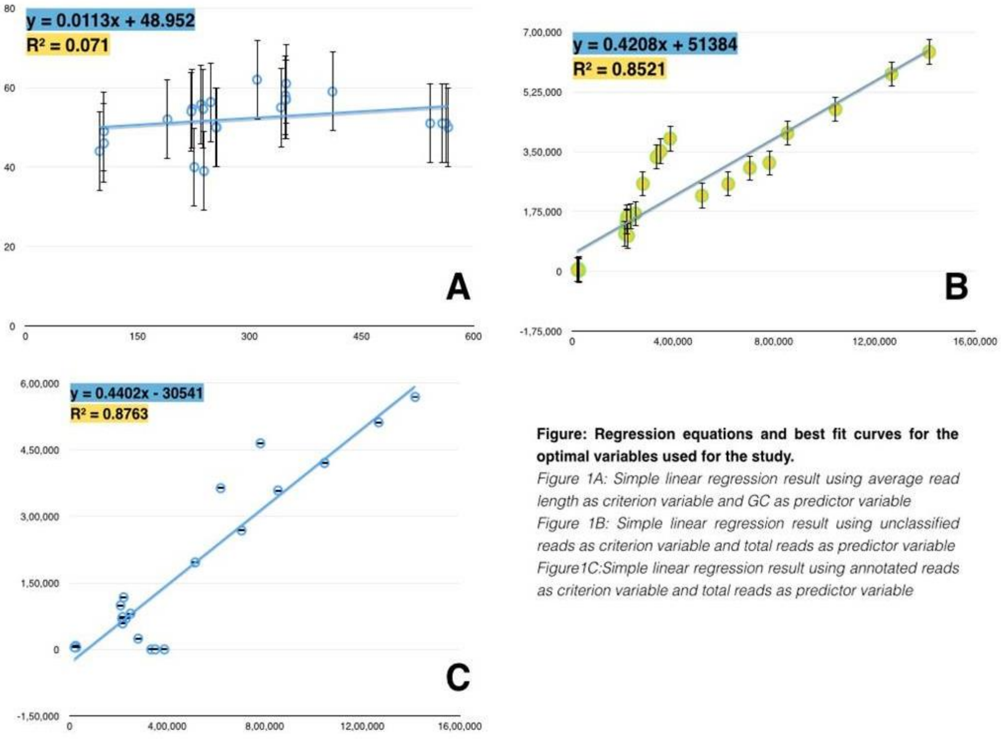

Figure: Regression equations and best fit curves for the optimal variables used for the study.

Figure 1A: Simple linear regression result using average read length as criterion variable and GC as predictor variable

Figure 1B: Simple linear regression result using unclassified reads as criterion variable and total reads as predictor variable

Figure1C:Simple linear regression result using annotated reads as criterion variable and total reads as predictor variable

Figure 2: Regression Analyses using optimal variables

Thus, the analyses clearly indicate that there are certain parameters in metagenomic metadata that can be used as metric for assessing the overall quality of the sequencing results. The regression equations developed in this study can be used for such assessments in the future.

ACKNOWLEDGEMENT

The authors are indebted to The Biome management for permitting to use their laboratory facilities for the entire duration of the work.

REFERENCES

[1] Marden, J. I. Multivariate Statistical Analysis http://istics.net/pdfs/multivariate.pdf accessed on 15th June 2017

[2] Shi, T., and Horvath, S. Unsupervised learning with random forest predictors. J. Comput. Graph Stat 2006;15: 118–138.

[3] Dinsdale Elizabeth A. et al. Multivariate analysis of functional metagenomes. Frontiers in Genetics 2013; 4:1-25.

[4] Brieiman, L. Friedman, J. M. Stone, C.J. and Olshen, R.A. Classification and Regression Trees 1984 Boco Raton: Chapman and Hall CRC.

[5] Liu et al. Multilevel regularized regression for simultaneous taxa selection and networkconstruction with metagenomic count data. Bioinformatics 2015; 31(7): 1067–1074.

[6] Alzubaidy H, Essack M, Malas TB, et al. Rhizosphere microbiome metagenomics of gray mangroves (Avicennia marina) in the Red Sea. Gene 2016; 576:626–636.

[7] Andreote FD, Jime´nez DJ, Chaves D, et al. The Microbiome of Brazilian Mangrove Sediments as Revealed by Metagenomics. Plos One 2012; 7 (6): e38600.

[8] Jing H, Xia X, Suzuki K, et al. Vertical Profiles of Bacteria in the Tropical and Subarctic Oceans Revealed by Pyrosequencing. Plos One 2013;8(11): e79423.

[9] Jimenez DJ, Dini-Andreote F, Elsas JDV. Metataxonomic profiling and prediction of functional behaviour of wheat straw degrading microbial consortia. Biomed Central 2014; 7:92.

[10] Goll J, Rusch DB, Tanenbaum DM, et al. METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. Bioinformatics 2010;26: 2631-2632