# An Efficient Comparative Machine Learning-based Metagenomics Binning Technique Via Using Random Forest

Helal Saghir, Dalila B. Megherbi

*Center for Human/Machine Intelligence, Networking and Distributed Systems, Department of Electrical and Computer Engineering, University of Massachusetts, Lowell, USA*

*Abstract*— **Metagenomics is the study of microorganisms collected directly from natural environments. Metagenomics studies use DNA fragments obtained directly from a natural environment using whole genome shotgun (WGS) sequencing. Sequencing random fragments obtained from whole genome shotgun into taxa-based groups is known as binning. Currently, there are two different methods of binning: sequence similarity methods and sequence composition methods. Sequence similarity methods are usually based on sequence alignment to known genome like BLAST, or MEGAN. As only a very small fraction of species is available in the current databases, similarity methods do not yield good results. As a given database of organisms grows, the complexity of the search will also grow. Sequence composition methods are based on compositional features of a given DNA sequence like K-mers, or other genomic signature(s). Most of these current methods for binning have two major issues: they do not work well with short sequences and closely related genomes. In this paper we propose new machine learning related predictive DNA sequence feature selection algorithms to solve binning problems in more accurate and efficient ways. In this work we use Oligonucleotide frequencies from 2-mers to 4-mers as features to differentiate between sequences. 2-mers produces 16 features, 3-mers produces 64 features and 4-mers produces 256 features. We did not use feature higher than 4-mers as the number of feature increases exponentially and for 5-mers the number of feature would be 1024 features. We found out that the 4-mers produces better results than 2-mers and 3-mers. The data used in this work has an average length of 250, 500, 1000, and 2000 base pairs. Experimental results of the proposed algorithms are presented to show the potential value of the proposed methods. The proposed algorithm accuracy is tested on a variety of data sets and the classification/prediction accuracy achieved is between 78% - 99% for various simulated data sets using Random forest classifier and 37% - 95% using Naïve Bayes classifier. Random forest Classifier did better in classification in all the dataset compared to Naïve Bayes.**

***Keywords: Bioinformatics; Metagenomics; Binning; Next generation Sequencing; Pattern Classification; Machine learning; Random forest; bagged decision tree; Computational intelligence; Reduction methods; forwaord sequential feature selection; ttest***

## I. INTRODUCTION

### *A. Introctuion and related research*

Metagenomics is the study of uncultured microorganisms collected directly from natural environments. Most of these collected organisms cannot be cultivated in a laboratory and hence cannot be sequenced as individual organisms. Metagenomics allows sequencing of genomes which cannot be cultured in a laboratory [1]. A vast majority of microbes, more than 99%, cannot be isolated and only survive in communities [2].

Metagenomics studies use DNA fragments obtained directly from an environment using whole genome shotgun (WGS) sequencing. Sequence data coming from Metagenomics are from heterogeneous microbial communities, and hence they are noisy and fragmented. The sequences obtained by Environmental Shotgun Sequence (ESS) are fragments coming from specific species. As there can be many different species in a sample, it is hard to know from which species did that DNA came from. When DNA sample is collected from a given environment, depending on the sample the DNA might only give a partial picture of the organisms in environment as genomic material from the more abundant species dominates the sample [3]. Many microorganisms live in symbiotic relationship and thus their host does not grow well in pure culture, thus making them ideal candidates for metagenomic analysis [4].

The length of fragments can be anywhere between 20 base pairs to 1000 base pairs depending upon the sequencing methods used. Due to the limitations in computational approaches short reads tend to miss genetically distant sequences and hence difficult to sequence. Longer read tends to produce better results [5].

To obtain complete genomes, usually ESS reads are assembled using the overlap-layout-consensus procedure [6]. Sequencing random DNA fragments obtained from whole genomes shotgun into taxa-based groups is known as binning [7].

Currently, there are two different methods of binning, sequence similarity methods and sequence composition methods. The sequence similarity methods are based on aligning sequence fragments to known genome based on homology like BLAST [8], or MEGAN [9] or USEARCH, UBLAST, and UCLUST[24] or CD-HIT[25, 26]. As only a very small fraction of species is available in the current databases, similarity methods do not yield good results and are very slow. They also do not work well for short reads. As a given database of organisms grows the complexity of the search will also grow. Genome sequence composition has shown to have organism-specific characteristics like k-mers, G+C content and codon usage [10, 11]. The sequence composition methods are

based on compositional features of a given DNA sequence like k-mers frequency, G+C content and codon usage or any other genomic signature(s) that are characteristics of evolutionary linkage. Most of the methods available today use di-, tri-or tetra- nucleotide frequencies. K-mers are any combination of the nucleotide A, C, T, G. So all possible combination of 2-mers would be AA, AC, AT, AG, CA, CC, CT, CG, TA, TC, TT, TG, GA, GC, GT, GG. So the total combination of 2-mers would be 16. Composition based methods can also be divided into supervised and unsupervised. Among the composition-based methods available are TETRA [12], Metacluster [13], Phylopythia [14], CompostBin [15], SOM [16], likelyBin [17], Maximal Quasi-clique Enumeration on Map-Reduce Clusters [22, 23]. TETRA [12] is an unsupervised method and uses z-scores of tetranucleotide by evaluating the over or underrepresenation of each tetranucleotide and then comparing sequences in pairs by evaluating the pearson's correlation coefficient of their z-scores for binning sequences. MetaClust [13] uses a combination of k-mers frequency to classify sequences in binning. Phylopythia [14] is a supervised method and uses a multiclass Support Vector Machine (SVM) classifier with the oligonucleotide composition to bin genome fragments. CompostBin [15] is a semi-supervised method and uses Principal Component Analysis (PCA) to reduce high- dimension into lower-dimensional space and then use normalized cut clustering algorithm to classify sequences into taxon-specific bins. Growing Self-organizing maps GSOM [18] and seeded GSOM or S-GSOM [19] is also used for binning method in Metagenomics. likelyBin [17] is an unsupervised binning method and uses markov chain Monte Carlo based on k-mers feature space and explicit likelihood model for low complexity communities between 2 to 10 species.

Most of the binning solutions are based on using the genomic signature or feature set of DNA fragments. Oligonucleotide (K-mers) has shown promising results and has been widely used for predicting genes and hence we in this paper we also used oligonucleotide (k-mers) of DNA fragment as a feature space to differentiate the different DNA segments for purpose of binning. Assuming there are N metagenomic sequences containing M different organism. Binning is defined as to divide the sequence reads into M bins that correspond to the species from which they are sequenced.

One of the problems of the sequence composition-based methods is using Oligonucleotide (k-mers) as feature sets. K-mers produce a high dimensional feature space. As one uses higher k-mers the number of feature rises exponentially. Using the higher dimension poses a significant computational challenge. Additionally computation of binning increases exponentially as the number of species increases. Other complications relate to the fact that in binning lateral gene transfer phenomenon exists [20]. Most of the current methods for binning have two major issues: they do not work well with short sequences because of the local variation of DNA composition characteristics and they do not work well with closely related genomes. A sequence composition method also usually degrades for shorter DNA fragments [21].

### *B. Contribution of this paper*

In this paper we use Oligonucleotide frequencies from 2-mers to 4-mers as features to differentiate between the sequences. By using forward sequential feature selection [27], we reduced feature set of k-mers to minimal set of features that has the same predictive power as the original features set. We then, based on the reduced feature set compared the classification accuracy between Random Forest classifier and Naïve Bayes classifier. We also analyze and compare the forward sequential feature selection to the popular ttest feature selection method, and analyze the effect of these two feature reduction and selection methods on the classification accuracy of the DNA sequences.

We also found, as sequence feature representation, that using the existence or non-existence of k-mers rather than using the actual frequency of K-mers in a particular reads, result in better and more accurate classification.

### *C. Orginazation of this paper*

The rest of the paper is organized as follows: The proposed methods for binning and sequence classification used are given in section II. Result and discussions are given in section III, and finally conclusions are given in section VI.

## II. PROPOSED METHODS

### *A. Generation of Data Sets*

Metagenomic being a new field, no standard data sets has yet been created for testing binning algorithms. So to test the accuracy of our algorithm, Metasim data set is being used. In our experimentation, Metasim was used to generate metagenomic data sets with an average length of 250bp, 500bp, 1000bp and 2000bp. Two metagenomic datasets D1 and D2 were created which varied in relative abundance and number of species. Each given dataset was tested with varied number of read length of 250bp, 500bp, 1000bp and 200bp. To demonstrate the scalability of the proposed methods, we did our testing from 600 reads to up to 5000 reads. We created 2 different datasets which are shown in TABLE I. The 1st column denotes dataset ID, the 2nd column denotes various species used in the dataset, and the 3rd column denotes the relative abundance of species in the dataset. Dataset D1 contains 5 different experimental sets (D1-1 to D1-5) containing 600 and 5000 shotgun fragments with different read length of 250-2000 bp contains two species and Dataset D2 contains 5 different experimental sets (D2-1 to D2-5) containing 600 and 5000 shotgun fragments with different read length of 250-2000 bp contains six species which are shown in Table II. We created training and testing data from the given dataset (ratio 80% training and 20% test data).

TABLE I. SPECIES DATASETS

| *ID* | *Species* | *Ratio* |
|---|---|---|
| D1 | `Acetohalobium arabaticum DSM 5501 chromosome'<br>`Acholeplasma laidlawii PG-8A'<br>(No of species = 2) | 1.5:1 |

| ID | Species | Ratio |
|---|---|---|
| D2 | `Acetohalobium arabaticum DSM 5501 chromosome'<br>`Acholeplasma laidlawii PG-8A chromosome'<br>`Aeropyrum pernix K1'<br>`Alcanivorax borkumensis SK2 chromosome'<br>`Arcobacter butzleri RM4018'<br>`Atopobium parvulum DSM 20469 chromosome'<br>(No of species = 6) | 1:1:1:2:1:1 |

TABLE II. EXPERIMENTAL DATASETS

| ID | Species | Read Size(bp) | Number of Reads |
|---|---|---|---|
| D1-1 | D1 | 1000 bp | 600 |
| D1-2 | D1 | 250 bp | 5000 |
| D1-3 | D1 | 500 bp | 5000 |
| D1-4 | D1 | 1000 bp | 5000 |
| D1-5 | D1 | 2000 bp | 5000 |
| D2-1 | D2 | 1000 bp | 600 |
| D2-2 | D2 | 250 bp | 5000 |
| D2-3 | D2 | 500 bp | 5000 |
| D2-4 | D2 | 1000 bp | 5000 |
| D2-5 | D2 | 2000 bp | 5000 |

### *B. Feature Selection Methods and Classification*

As mentioned in the introduction section, one of the problems of the sequence composition-based methods is using Oligonucleotide (k-mers) as feature sets. K-mers produce a high dimensional feature space as one uses higher k-mers the number of feature raises exponentially. To solve this issue some feature selection techniques have to be used. Feature selection techniques do not alter the original representation of the variables, but merely select a subset of them. The main functions of feature selection are to decrease dimensionality, to avoid over fitting and improve performance, to provide faster, more cost-effective models, and to model interpretability.

The advantage of feature selection techniques comes with a price, which is the search for a subset of relevant features comes with additional complexity in the modeling. In the context of classification, feature selection techniques can be of three different types, depending on how they combine the feature selection search with the construction of the classification model: filter methods e.g. t-test, correlation-based feature selection, wrapper methods i.e. sequential forward selection, and embedded methods i.e. embedded in the classifier such as decision tree(Random forest), or Naïve Bayes.

Here, we compare two feature selection methods, that is the t-test and Forward Sequential Feature Selection. We compare the binning accuracy obtained using both feature selection methods on Random Tree and Naïve Bayes classifiers. First, by using a forward sequential feature selection we identify a minimal set of features that has the same predictive power as the original model. Feature selection method reduces the dimensionality of data by selecting only a subset of features to create a model. Forward sequential feature selection works by sequentially adding features, until the addition of any more features does not significantly increase the accuracy of the original model. We start by testing each possible feature one at a time, identifying the single feature that gives a most accurate model and then add that to the model. Next, add each of the remaining features one at a time and identify the variable that improves the accuracy of the model the most and then test the two models whether they are statistical significant. The process is considered complete, if the new model is not significantly more accurate than the original model. If, however, the new model is statistically more significant, than the process to find the next feature continues. This process is repeated until no new feature can be identified that has a statistically significant impact on the model. For Dataset D1-1, 7 features were used from an initial given 256 features set. The numbers of features used for Dataset D2-1 are in shown in table III.

TABLE III. NUMBERS OF FEATURES USED IN DATASET D2-1 BY USING THE FORWARD SEQUENTIAL FEATURE SEELCTION METHOD

| K-mers | Total Number of features | Number of features used after using forward sequential selection |
|---|---|---|
| 2-mers | 16 | 11 |
| 3-mers | 64 | 12 |
| 4-mers | 256 | 10 |

The out-of-bag feature importance is derived using bootstrap aggregation (treebagger) which measures the importance for each predictor feature. The ranking of feature is based on the treebagger algorithm, which is based on randomly permuting each variable's values on the out-of-bag samples, and see how much the prediction error increases compared to the real predictions. We built the decision tree model for each of the 250 resampled datasets created. We compare the classification using both random forest classifier and Naïve Bayes classifier using 10-fold cross validation. The block diagram of Table-IV, and Table-V, shows the complete process of predicting sequence binning.

TABLE IV. PIPELINE FOR BINNING USING FORWARD FEATURE SELECTION AND RANDOM FOREST AND NAÏVE BAYES CLASSIFIER

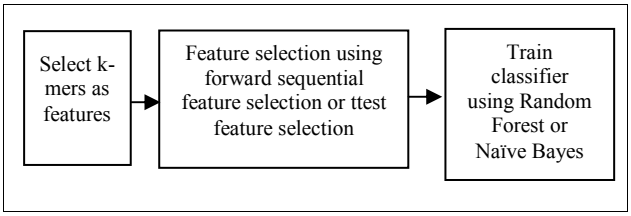
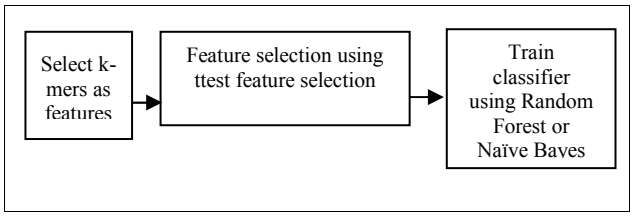


TABLE V. PIPELINE FOR BINNING USING TTEST USING RANDOM FOREST AND NAÏVE BAYES CLASSIFIER




The result for feature importance using forward sequential feature selection method is shown in figure 1 for dataset D1-1 and in figure 3 for dataset D2-1. Figure-2 illustrates the final 7 features, out-of 256 original features, which were sufficient to predict the 600 species in data set D2-1. Figure 4, Figure 5 and Figure 6 display similar results and effect of various variables on the prediction error for the various data sets used. The Computational Complexity for Random Forest is O(NlogN).

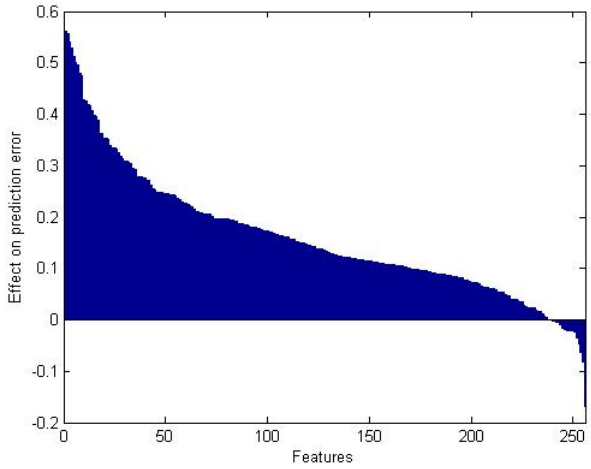



Figure 1. Displays the variables, sorted by the effect they have on prediction error for Dataset D1-1 using Forward sequenctial feature selection method.

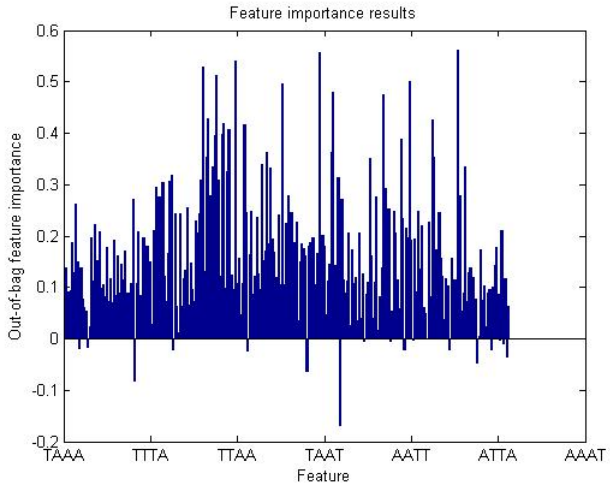



Figure 2. Forward sequential feature selection method was used to create feature set importance to identify species using 4 mers from dataset D1-1 containing 600 shotgun fragments with an average of 1000 bp size. 7 features were sufficient to predict the species correctly out of 256 features. Feature selection resulted in a good classification of the two-species genome sequences with result of accuracy up to 99% using random Forest.

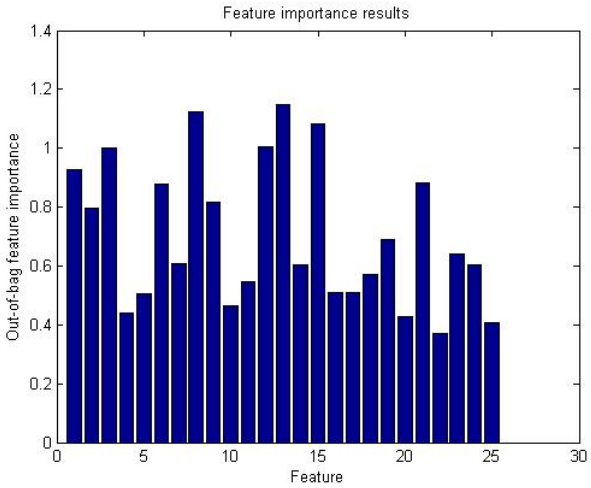



Figure 3. ttest feature selection method was used to create feature set importance to identify species using 4 mers from dataset D1-1 containing 600 shotgun fragments with an average of 1000 bp size. 7 features were sufficient to predict the species correctly out of 256 features. Feature selection resulted in a good classification of the two-species genome sequences with result of accuracy up to 98.33% using random Forest.

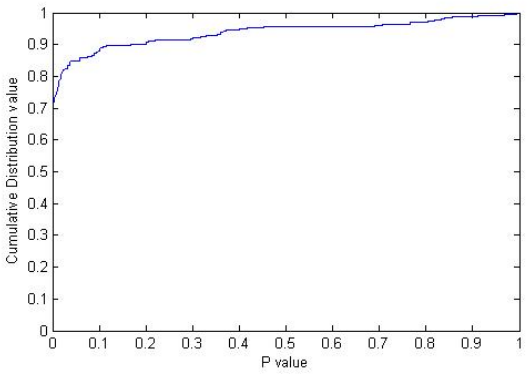



Figure 4. Displays the variables, sorted by the effect they have on the prediction error for dataset D1-1 using ttest feature selection. This graph shows that more than 70% of features have p-values close to zero and over 85% of features have p-values smaller than 0.05 meaning there are only about less that 15% of features that have a strong discrimination power that is less than 40 features out of 256 features.

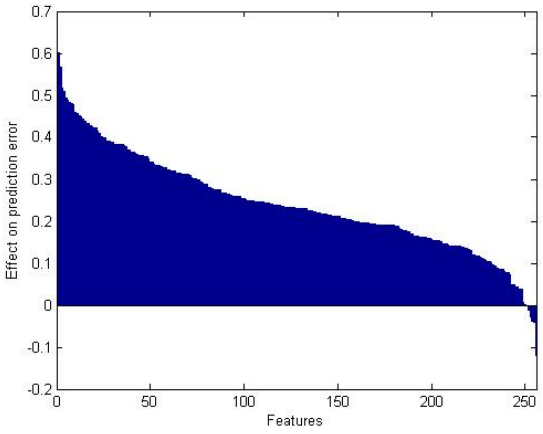



Figure 5. Displays the variables, sorted by the effect they have on prediction error for Dataset D2-1 using Forward sequenctial feature selection method.

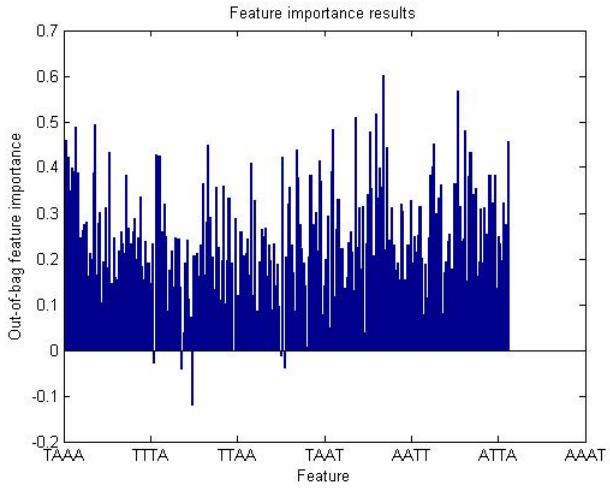



Figure 6. Forward sequential feature selection method was used to create feature set importance to identify species using 4 mers from dataset D2-1 containing 600 shotgun fragments with an average of 1000 bp size. 17 features were sufficient to predict the species correctly out of 256 features. Feature selection resulted in a good classification of the six-species genome sequences with result of accuracy up to 96.7% using random Forest.

## III BINNING/CLASSIFICATION PREDICTION : RESULT AND DISCUSSIONS

With the two metagenomic datasets considered, which vary in relative abundance and number of species and read length, accuracy of the results, as shown in table VI. The classification achieved an accuracy of more than 78% - 99% for various simulated data sets using Random forest classifier and 37% - 95% using Naïve Bayes classifier. We also observed that the random forest classifier produced better result in general than the Naïve Bayes classifier.

TABLE VI. RESULTS USING FORWARD SEQUENTIAL FEATURE SELETION USING NAÏVE BAYES AND RANDOM FOREST

| *Dataset* | *Accuracy using Naïve Bayes* | *Accuracy using Random Forest* |
|---|---|---|
| D1-1 | 95% | 99.17% |
| D1-2 | 56.6% | 78.4% |
| D1-3 | 63.3% | 86.9% |
| D1-4 | 56.5% | 91.1% |
| D1-5 | 37.9% | 86.3% |
| D2-1 | 93.13% | 96.7% |
| D2-2 | 65.3% | 95.5% |
| D2-3 | 65.3% | 95.5% |
| D2-4 | 62.5% | 95.7% |
| D2-5 | 62.9% | 96.8% |

We also compared the ttest to sequential forward feature selection for feature selection method and did not find much difference in accuracy. In ttest we were getting larger feature set for better performance. We used 25 top features for ttest whereas in forward feature selection we used 7 features to obtain the same level of accuracy. In the dataset D1-1, the accuracy using ttest as a feature selection, in naïve Bayes we obtained 95% whereas in random forest was 98.33%. Only drawback with the ttest feature selection is that if there are more than 2 species in a test, it cannot be used to find the feature selection.

The algorithm's accuracy was tested on a variety of simulated data sets. We also showed that 4-mer produce better results than 2-mers and 3-mers. Results are shown in Fig 7. The classification achieved an accuracy of more than 90% for various simulated data sets using Random forest classifier.

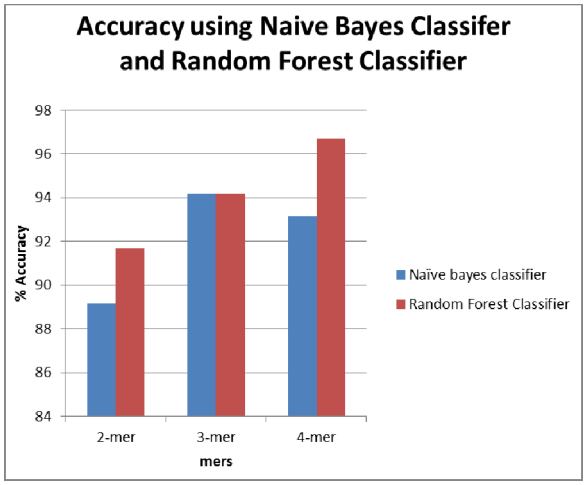



Figure 7. We compared the 2-mer, 3-mer, 4-mer and found that 4-mer produces a better result in general.

## III. CONCLUSIONS

Our findings showed that the 2-mers, 3-mers or 4-mers are sufficient for binning 1000 bp sequences. We also showed that 4-mer produce better results than 2-mers and 3-mers. We found out that the random forest classifier produces better results in general than Naïve Bayes classifier.

## REFERENCES


[1] Hugenholtz, P. 2002. Exploring prokaryotic diversity in the genomic era. Genome Biol. 3:REVIEWS0003.

[2] J. C. Wooley, Y. Ye, " Metagenomics: Facts and Artifacts, and computational Challenges" J Comput Sci Technol. 2009 January ; 25(1): 71–81. doi:10.1007/s11390-010-9306-4.

[3] Wooley JC, Godzik A, Friedberg I (2010) "A Primer on Metagenomics." PLoS Comput Biol 6(2): e1000667. doi:10.1371/journal. pcbi.1000667

[4] J. Handelsman, "Metagenomics: Application of Genomics to Uncultured Microorganisms," Microbiology and Molecular Biology Reviews, Vol. 68, No.4, pp. 669-685, 2004

[5] K. E. Wommack, J. Bhavsar, and J. Ravel, "Metagenomics: Read length matters," Applied Environmental Microbiology, vol. 74, no. 5, pp. 1453–1463, 2008.

[6] J.C. Venter, K.Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W.Nelson, D.E. Fouts, S. Levy, A.H. Knap, M.W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H.B. Tillson, C. Pfannkoch, Y.H. Rogers, H.O. Smith, "Environmental genome shotgun sequencing of the Sargasso Sea," Science, 304, 66-74, 2004.

[7] Victor Kunin, Alex Copeland, Alla Lapidus, Konstantinos Mavromatis, and Philip Hugenholtz "A Bioinformatician's Guide to Metagenomics" MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS, Dec. 2008, p. 557–578

[8] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. Journal of Molecular Biology 1990, 215:403-410.

[9] D. E. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "Megan analysis of metagenomic data," Genome Research, 2007.

[10] S. Karlin and C. Burge, "Dinucleotide relative abundance extremes: a genomic signature," Trends in Genetics, vol. 11, pp. 283–290, 1995.

[11] P. J. Deschavanne and et al., "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences," Molecular Biology and Evolution, vol. 16, pp. 1391–1399, 1999.

[12] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glockner, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," BMC Bioinformatics, vol. 5, no. March 30, 2009

[13] Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ, et al. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 2006;443:950–955. [PubMed: 16980956]

[14] A.C. McHardy, H.G. Martln, A. Tsirigos, P. Hugenholtz and I. Rigoutsos, "Accurate phylogenetic classification of variable-length DNA fragments," Nautre Methods, 4, (1), 63-72, 2007

[15] S. Chatterji and et al., "CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads," Springer Lecture Notes in Computer Science, 2008.

[16] T. Abe and et al., "Informatics for unveiling hidden genome signatures," Genome Research, vol. 13, pp. 693–702, 2003.

[17] Andrey Kislyuk , Srijak Bhatnagar, , Jonathan Dushoff and Joshua S Weitz1, Unsupervised statistical clustering of environmental shotgun sequences, BMC Bioinformatics 2009, 10:316

[18] C.-K. Chan and et al., "Using growing self-organising maps to improve the binning process in environmental wholegenome shotgun sequencing," Journal of Biomedicine and Biotechnology, 2008.

[19] Chan CKK, Hsu AL, Halgamuge SK, Tang SL (2008) Binning sequences using very sparse labels within a metagenome. BMC Bioinformatics 9: 215.

[20] J. A. Eisen, "Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes," PLoS Biology, vol. 5, no. 3, 2007.

[21] Arima, K. and Wooley, J, "Metagenomics," in Computational Methods for Understanding Archaeal and Bacterial Genomes, Edited by Xu, Y. and Gogarten, J.P., Imperial College Press, 2008.

[22] X. Yang, J. Zola, S. Aluru, "Large Scale Metagenomic Clustering on Map-Reduce Clusters", Journal of Bioinformatics and Computational Biology 11(1):1340001, 2013.

[23] X. Yang, J. Zola, S. Aluru, "Parallel Metagenomic Sequence Clustering Via Sketching and Maximal Quasi-clique Enumeration on Map-Reduce Clouds", In Proc. IEEE Int. Parallel and Distributed Processing Symposium (IPDPS), pp. 1223-1233, 2011.

[24] Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26(19), 2460-2461.doi: 10.1093/bioinformatics/btq461

[25] Beifang Niu, Limin Fu, Shulei Sun and Weizhong Li, Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics, (2010) 11:187

[26] Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22, 1658-1659.

[27] T.Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition," Springer Press.