CompostBin: A DNA Composition-Based Algorithm for Binning Environmental Shotgun Reads

Sourav Chatterji[1], Ichitaro Yamazaki[2], Zhaojun Bai[2],
and Jonathan A. Eisen[1,3]

[1] Genome Center, U C Davis, Davis CA 95616, USA
schatterji@ucdavis.edu,jaeisen@ucdavis.edu
[2] Computer Science Department, U C Davis, Davis CA 95616, USA
yamazaki@cs.ucdavis.edu,bai@cs.ucdavis.edu
[3] The Joint Genome Institute, Walnut Creek CA 94598, USA

Abstract. A major hindrance to studies of microbial diversity has been that the vast majority of microbes cannot be cultured in the laboratory and thus are not amenable to traditional methods of characterization. Environmental shotgun sequencing (ESS) overcomes this hurdle by sequencing the DNA from the organisms present in a microbial community. The interpretation of this metagenomic data can be greatly facilitated by associating every sequence read with its source organism. We report the development of CompostBin, a DNA composition-based algorithm for analyzing metagenomic sequence reads and distributing them into taxon-specific bins. Unlike previous methods that seek to bin assembled contigs and often require training on known reference genomes, CompostBin has the ability to accurately bin raw sequence reads without need for assembly or training. CompostBin uses a novel weighted PCA algorithm to project the high dimensional DNA composition data into an informative lower-dimensional space, and then uses the normalized cut clustering algorithm on this filtered data set to classify sequences into taxon-specific bins. We demonstrate the algorithm's accuracy on a variety of low to medium complexity data sets.

Keywords: Metagenomics, Binning, Feature Extraction, Normalized Cut, weighted PCA, DNA composition metrics, Genome Signatures.

1 Introduction

Microbes are ubiquitous organisms that play pivotal roles in the earth's biogeochemical cycles. Their most visible effects on human well-being arise through their roles as mutualistic symbionts and hazardous pathogens. The study of microbes is crucial to our understanding of the earth's life processes and human health. Most of our knowledge about microbes has been obtained through the study of organisms cultured in artificial media in the laboratory. Although this approach has provided profound biological insights, it is inadequate for studying

M. Vingron and L. Wong (Eds.): RECOMB 2008, LNBI 4955, pp. 17–28, 2008.

the structure and function of many microbial communities. One obstacle has been that the vast majority of microbes have not been cultured and may not be culturable [1]. Even though culture independent methods such as 16S rRNA surveys [2] have been developed, they are unable to simultaneously answer two fundamental questions: Who is out there? and What are they doing? The application of genome sequencing methods is revolutionizing this field by enabling us for the first time to address those two questions for unculturable microbial communities [3,4,5]. These techniques, called environmental genomics or metagenomics, study microbial communities by analyzing the pooled genomes of all the organisms present in the community.

In one specific metagenomic method, **e**_nvironmental_ **s**_hotgun_ **s**_equencing_ (ESS), DNA pooled from a microbial community is sampled randomly using whole genome shotgun sequencing. Thus, ESS data is made up of sequence reads from multiple species. This adds an additional layer of complexity compared to single-species genome sequencing, as it requires analysis of the metagenomic data in order to associate each sequence read with its source organism. Therefore, a critical first step in many metagenomic analyses is the distribution of reads into taxon-specific bins.

The difficulty of accurately binning ESS reads from whole genome data remains a significant hurdle in metagenomics. The taxonomic resolution achievable by the analysis depends on both the binning method and the complexity of the community. For instance, binning into species-specific bins can be achieved in low-complexity microbial communities (e.g., the dual-bacterial symbiosis of sharpshooters [6]). However, the problem becomes more difficult in high-complexity communities with hundreds of species, such as ocean microbial communities [7] and the human distal gut [5]. Because of these difficulties, many metagenomic studies (e.g., [8]) have resorted to analyzing at the level of the metagenome, essentially treating a microbial community as a bag of genes. This is not a satisfactory solution. Identifying and characterizing individual genomes can provide deeper insight into the structure of the community [6].

A variety of approaches have been developed for binning: assembly, phylogenetic analysis [9], database search [10], alignment with reference genome [7] and DNA composition metrics [11,12,13] Most current binning methods suffer from two major limitations: they require closely related reference genomes for training/alignment and they perform poorly on short sequences. To overcome the second difficulty, almost all current binning methods are applied to assembled contigs. However, most of the current generation assemblers can be confounded by metagenomic data since they implicitly assume that the shotgun data is from a single individual or clone. Therefore, we believe that assembly is risky when binning and that it is necessary to analyze raw sequence reads to get an unbiased look at the data.

To overcome the above-mentioned disadvantages of other binning methods, we have developed CompostBin, a binning algorithm based on DNA composition. CompostBin can bin raw sequence reads into taxon-specific bins with high accuracy and does not require training on currently available genomes.

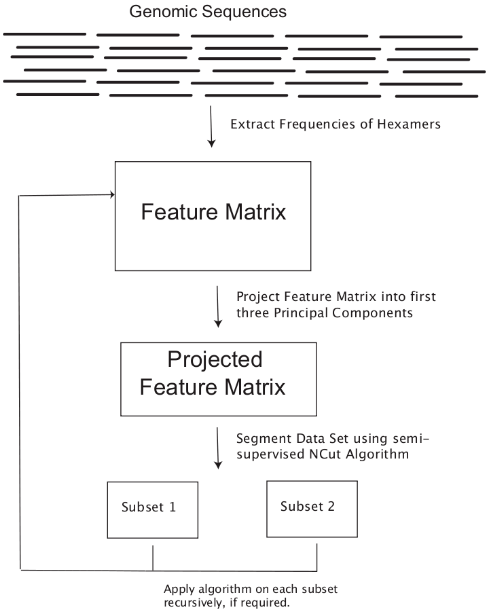

Genomic Sequences
Extract Frequencies of Hexamers
Feature Matrix
Project Feature Matrix into first three Principal Components
Projected Feature Matrix
Segment Data Set using semi-supervised NCut Algorithm
Subset 1
Subset 2
Apply algorithm on each subset recursively, if required.

Fig. 1. High-level overview of the CompostBin algorithm. Principal Component Analysis is used to project the data into a lower-dimensional space. A semi-supervised normalized cut algorithm is used to segment the data set into two subsets. The algorithm is applied iteratively on the subsets to obtain the desired number of bins.

Like other composition-based methods, it seeks to distinguish different genomes based on their characteristic DNA compositional patterns, termed ”signatures.” For example, one of the most commonly used metrics measure the frequency of occurrence of Kmers (oligonucleotides of length K) in a sequence. Biases in $Kmer$ frequencies were analyzed extensively by Karlin and colleagues (e.g. [14]).

These biases have been extensively used for binning metagenomic sequences. For instance, TETRA[11] uses z-scores from tetramer frequencies to classify metagenomic sequences. A related program, MetaClust uses a combination of $Kmer$ frequency metrics to score metagenomic sequences and was used to classify sequences from the endosymbionts of a gutless worm [15]. However, the final assignment of sequences to bins in both these programs involve a significant manual component. Another class of methods [12,13] train their classifier using existing whole genome sequences and these classifiers can be even used to classify sequences from closely related novel genomes. However, as we discuss later, a serious drawback of these methods is that the pool of available genomes is very small and biased. Finally, the interpolated Markov model of the genefinder *Glimmer* can be used for binning in specific cases and has been used to distinguish symbiont sequences from the host sequences [16]. Unfortunately, these composition-based binning algorithms do not perform well on short fragments. Poor performance in shorter fragments is caused by the noise associated with the high dimensionality of the feature space and the associated curse of dimensionality.

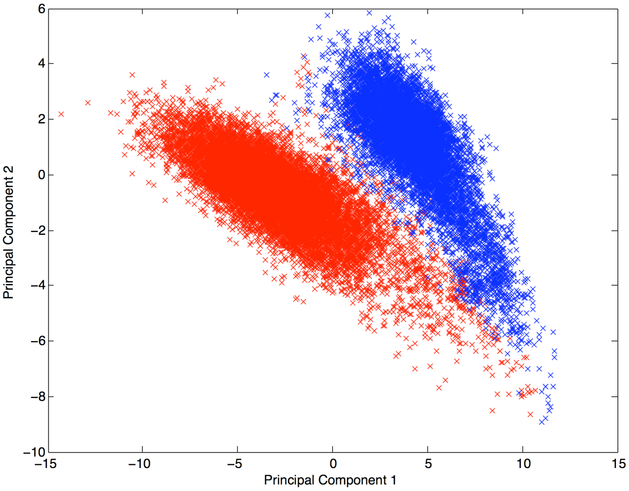


Fig. 2. Figure illustrating the separation of sequences according to species by using PCA. This data set contains sequences from two alphaproteobacteria, *Gluconobacter oxydans* (in red) and *Rhodospirillum rubrum* (in blue), which have GC content of 0.65 and 0.61, respectively. The data set is projected into the first two principal components.

When measuring the frequency of Kmers, the feature vector has 4^K dimensions (associated with measuring the frequencies of 4^K possible oligonucleotides of length K). Thus, for instance, if one looks at the frequency of hexamers in 2kb fragments, the dimensionality of the feature space is twice the length of the sequenced fragments.

CompostBin employs a new approach to deal with the noise arising from the high dimensionality of the feature vector (Figure 1). Instead of treating all components of the noisy feature space equally, we extract the most "important" directions and use these components for distinguishing between taxa. We use a weighted version of the standard Principal Component Analysis technique[17] to extract a “meaningful” lower dimensional sub-space. As shown in Figure 2, the algorithm can distinguish sequences from various species using just these first three principal components. The normalized cut clustering algorithm used to classify sequences into taxon-specific bins works on the lower dimensional sub-space and is guided by information from phylogenetic markers. We tested CompostBin on a wide variety of data sets and demonstrated that it is highly accurate in separating sequences into taxon-specific bins, even when processing raw reads of short sequences.

2 Methods

2.1 The CompostBin Algorithm

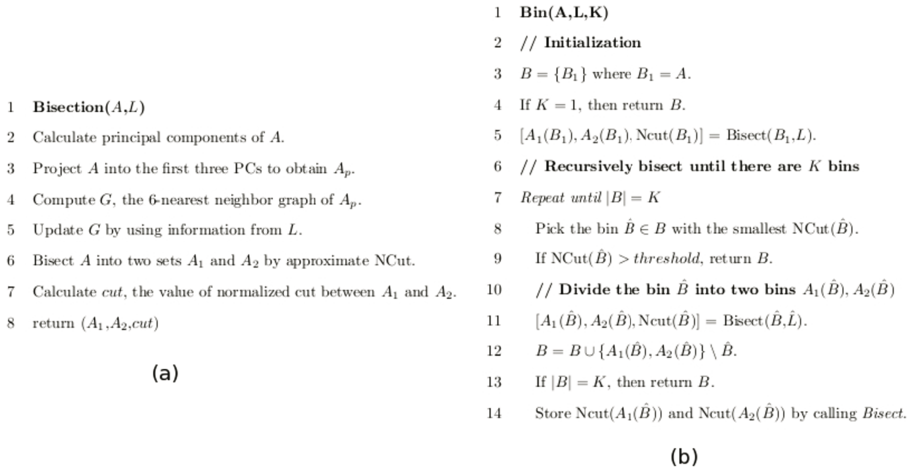
The input to CompostBin consists of raw sequence reads, along with mate pair information and the taxonomic assignment of reads containing phylogenetic markers. Either the number of abundant species or the number of taxonomic groups in the data set is provided to help the algorithm determine the number of bins in the output. This information can be obtained by analyzing the reads containing genes for ribosomal RNA or other marker genes [7]. In the simulation experiments, the number of bins is set to the number of species in the simulation. An overview of the algorithm is provided in Figure 1.

Feature Extraction by weighted PCA: Mate pairs are joined together and treated as a single sequence because they are highly likely to have originated from the same organism. Each sequence being analyzed is initially represented as a $4,096$-dimensional feature vector, with each component denoting the frequency of one of the $4,096$ hexamers. A weighted version of Principal Component Analysis (PCA) is then used to decrease the noise inherent in this high-dimensional data set by identifying the principal components of the feature matrix A.

In the standard form of Principal Component Analysis, the principal components are the orthogonal directions with highest variance and correspond to eigenvectors of the covariance matrix. If the *relative abundance* of various species is skewed, standard PCA might not be suitable for distinguishing the species. This is because the *within species variance* in the more abundant species might be overwhelming compared to *between species variance* and therefore the principal components cannot be used to distinguish between species. Therefore, we try to take the relative abundance into account in a weighting scheme that is used to normalize the effect of skewed relative abundance. We use a generalized variant of Principal Component Analysis that assigns a weight to each sequence and uses these weights to calculate the *weighted* covariance matrix of the data set. The principal components are the eigenvectors of this weighted covariance matrix. Further details about this generalization of PCA can be obtained from Chapter 14 of the book by Jolliffe [17].

We use a novel weighting scheme where the weight of each sequence is calculated by measuring its overlap with other reads in the data set. For each sequence, BLAT [18] was used to find overlaps with other sequences in the data set. This overlap information is then used to calculate the number of times a particular base in the sequence has been sequenced and thus estimate the coverage of the sequence (as defined in [19]). The weight of each sequence is set to the inverse of its *coverage*. The rationale behind this weighting strategy is that the sequences from the more abundant species will have higher coverage and thus will be weighed down. In fact, if there are sufficient number of sequences and the genome sizes of all species in the sample are equal, the average weight of the sequences from a particular species will be inversely proportional to the relative abundance of that species.

Determining the number of principal components required for analysis is crucial to the success of the algorithm. In our case, use of just the first three principal components is adequate to separate sequences from different species. For example, Figure 2 shows that for Data Set S5 which contains two alphaproteobacteria with similar GC content, almost complete separation is achieved by using only the first two principal components.



```
Bisection($A$,$L$)
Calculate principal components of $A$.
Project $A$ into the first three PCs to obtain $A_p$.
Compute $G$, the 6-nearest neighbor graph of $A_p$.
Update $G$ by using information from $L$.
Bisect $A$ into two sets $A_1$ and $A_2$ by approximate NCut.
Calculate $cut$, the value of normalized cut between $A_1$ and $A_2$.
return ($A_1$,$A_2$,$cut$)
```

(a)

```
Bin(A,L,K)
// Initialization
$B = \{B_1\}$ where $B_1 = A$.
If $K = 1$, then return $B$.
$[A_1(B_1), A_2(B_1), \text{Ncut}(B_1)] = \text{Bisect}(B_1,L)$.
// Recursively bisect until there are $K$ bins
Repeat until $|B| = K$
  Pick the bin $\hat{B} \in B$ with the smallest NCut($\hat{B}$).
  If NCut($\hat{B}$) > threshold, return $B$.
  // Divide the bin $\hat{B}$ into two bins $A_1(\hat{B}), A_2(\hat{B})$
  $[A_1(\hat{B}), A_2(\hat{B}), \text{Ncut}(\hat{B})] = \text{Bisect}(\hat{B},\hat{L})$.
  $B = B \cup \{A_1(\hat{B}), A_2(\hat{B})\} \setminus \hat{B}$.
  If $|B| = K$, then return $B$.
  Store Ncut($A_1(\hat{B})$) and Ncut($A_2(\hat{B})$) by calling Bisect.
```

(b)

Fig. 3. Pseudocode describing the bisection and binning algorithm. A is the $N \times 4,096$ feature matrix, with each $4,096$-length feature vector representing a sequence. L contains labeling information obtained from phylogenetic markers, and K is the the desired number of bins.

Bisection by Normalized Cuts: The projection of the data matrix A into the first three principal components produces an $N \times 3$ data matrix A_p. A clustering algorithm is then applied to A_p to separate the N points into taxon-specific bins. A bisection algorithm is used to bisect a data set into two bins as detailed below. If the data set is to be divided into more than two bins, this algorithm is used recursively. Figure 3(a) shows pseudocode for the bisection algorithm. Given the projected matrix and phylogenetic markers as inputs, the procedure first computes the weighted graph over the sequences where the edge weights measure the similarity between corresponding sequences. Then, the normalized cut clustering algorithm [22] is employed to bisect the graph such that sequences from the same taxonomic group stay together.

Computation of Similarity Measure: As described earlier, the $4,096$-dimensional feature vector is projected into the first three principal components, and each sequence is represented as a point in 3-dimensional space. The clustering algorithm initially creates a 6-nearest neighbor graph $G(V, E, W)$ to capture the structure of the data set. The vertices in V correspond to the sequences, and an edge $(v_1, v_2) \in E$ between two sequences v_1 and v_2 exists only if one of the sequences is a 6-nearest neighbor of the other in Euclidean space. The nearest-neighbor graph reveals the global relation of the data set through this easily-computable local metric [20]. Each edge between two neighboring sequences v_1 and v_2 is weighted by their similarity $w(v_1, v_2)$, which is defined as the exponential inverse of their normalized Euclidean distance:

$$w(v_1, v_2) = \begin{cases} e^{-\frac{d(v_1,v_2)}{\alpha}} & \text{if } (v_1, v_2) \in E, \\ 0 & \text{otherwise,} \end{cases}$$

where $d(v_1, v_2)$ is the Euclidean distance between v_1 and v_2, and $\alpha = \max_{(v,u) \in E} d(v, u)$.

Semi-supervision Using Phylogenetic Markers: Marker genes, such as the genes that code for ribosomal proteins, are one of the most reliable tools for phylogenetically assigning reads to bins. Since these marker genes appear in only a small fraction of the reads, we used taxonomic information from 31 phylogenetic markers [21] to improve the clustering algorithm. This taxonomic information is provided to the binning algorithm as a label for each sequence, with each label corresponding to a single taxonomic group. Sequences without a taxonomic assignment are assigned the label "unknown."

A semi-supervised approach is then employed to incorporate this information into the clustering algorithm. Two vertices v_1 and v_2 are connected with the maximum edge weight (i.e., $w(v_1, v_2) = 1$) if the corresponding sequences are from the same taxonomic group, and the edge between v_1 and v_2 is removed (i.e., $w(v_1, v_2) = 0$) if they are from different groups.

Normalized Cut and its approximation: Given a weighted graph $G(V, E, W)$, the association between two subsets X and Y of V $W(X, Y)$ is defined as the total weight of the edges connecting X and Y: $W(X, Y) \ = \sum_{x \in X, y \in Y} w(x, y)$. The normalized cut algorithm bisects V into two disjoint subsets U and $\bar{U}$ such that the association within each cluster is large while the association between clusters is small, i.e., the normalized cut value *NCut* is minimized, where

$$\text{NCut} = \frac{W(U, \bar{U})}{W(U, V)} + \frac{W(U, \bar{U})}{W(\bar{U}, V)}.$$

Since finding the exact solution to minimize *NCut* is an NP-hard problem, an approximate solution is computed using a spectral analysis of the Laplacian matrix of the graph [22].

Generalization to Multiple Bins: If the data set needs to be divided into more than two bins, an iterative algorithm is used, where the bins are bisected recursively until the required number of bins is obtained. Figure 3(b) shows the pseudocode describing the algorithm. A set of bins, B is kept, where each element of B is a set of data points belonging to the same bin. The set B is initialized to be the singleton set $\{A\}$, where A contains all points in the data set. At each subsequent step of the algorithm, the bin with the lowest normalized cut value is bisected. The bisection continues until either B has the required number of bins or we no longer have a good bisection as measured by the normalized cut value.

2.2 Generation of Test Sets

In our experiments, we simulated the sequencing of low- to medium-complexity communities in which the number of species ranged from two to six and their

Table 1. Test Data Sets and Binning Accuracy

ID	Species	Ratio	Taxonomic Differences	Error
S1	*Bacillus halodurans* [0.44] & *Bacillus subtilis* [0.44]	1:1	Species	6.48%
S2	*Gluconobacter oxydans* [0.61] & *Granulobacter bethesdensis* [0.59]	1:1	Genus	3.39%
S3	*Escherichia coli* [0.51] & *Yersinia pestis* [0.48]	1:1	Genus	10.0%
S4	*Rhodopirellula baltica* [0.55] & *Blastopirellula marina* [0.57]	1:1	Genus	2.05%
S5	*Bacillus anthracis* [0.35] & *Listeria monocytogenes* [0.38]	1:2	Family	5.49%
S6	*Methanocaldococcus jannaschii* [0.31] & *Methanococcus mariplaudis* [0.33]	1:1	Family	0.51%
S7	*Thermofilum pendens* [0.58] & *Pyrobaculum aerophilum[0.51]*	1:1	Family	0.28%
S8	*Gluconobacter oxydans* [0.61] & *Rhodospirillum rubrum* [0.65]	1:1	Order	0.98%
S9	*Gluconobacter oxydans* [0.61], *Granulobacter bethesdensis* [0.59], & *Nitrobacter hamburgensis* [0.62]	1:1:8	Family Order	7.7%
S10	*Escherichia coli* [0.51], *Pseudomonas putida* [0.62], & *Bacillus anthracis* [0.35]	1:1:8	Order Phylum	1.96%
S11	*Gluconobacter oxydans* [0.61], *Granulobacter bethesdensis* [0.59], *Nitrobacter hamburgensis* [0.62], & *Rhodospirillum rubrum* [0.65]	1:1:4:4	Family Order	4.44%
S12	*Escherichia coli* [0.51], *Pseudomonas putida* [0.62], *Thermofilum pendens* [0.58], *Pyrobaculum aerophilum* [0.51], *Bacillus anthracis* [0.35], & *Bacillus subtilis* [0.44]	1:1: 1:1: 2:14	Species, Order Family, Phylum Kingdom	4.52%
R1	Glassy-winged sharpshooter endosymbionts	-	-	9.04%

relative abundance ranged from 1:1 to 1:14. ReadSim [23] was used to simulate paired-end Sanger sequencing from isolate genomes with an average read length of $1,000$ bp. The reads from various isolates were then combined in ratios corresponding to their relative species abundance in the data set to yield a simulated metagenomic data set of known composition. The 12 simulated data sets are described in Table 1. The GC content of each species' genome is listed in squared-brackets and can be used for assessing the diversity of DNA composition. The taxonomic levels are obtained from IMG[24] and can be used for assessing the phylogenetic diversity.

In addition, we tested the algorithm on a metagenomic data set containing reads obtained from gut bacteriocytes of the glassy-winged sharpshooter. The original study [7] had used phylogenetic markers to classify the sequence reads into three bins: reads from *Baumannia cicadellinicola* in Bin 1, reads from *Sulcia muelleri* in Bin 2, and reads from the host and miscellaneous unclassified reads in Bin 3. Due to the heterogeneity of Bin 3, the accuracy of the algorithm was tested only on its ability to distinguish between reads from Bin 1 and Bin 2.

3 Results

CompostBin was coded in C and Matlab. It is publicly available for download from `http://bobcat.genomecenter.ucdavis.edu/souravc/compostbin/`. CompostBin was tested on a variety of low-to-medium complexity data-sets. Details of the test data sets and CompostBin's performance are provided in the next two sections.

3.1 Test Data Sets

Metagenomics being a relatively new field, very few standard data sets for testing binning algorithms have been developed [25]. One obstacle to their development has been that the ”true” solution is still unknown for the sequence data generated by most metagenomic studies. To test the accuracy of a binning algorithm, one can instead simulate the shotgun sequences that would be obtained from a combination of organisms of known genome sequences. Simulated sequence reads from multiple genomes were pooled to simulate the challenges of metagenomic sequencing. When designing our simulated data sets, we took into account several variables that affect the difficulty of binning: the number of species in the sample, their relative abundance, their phylogenetic diversity, and the differences in GC content between genomes.

We also tested CompostBin on a publicly available metagenomic data set whose solution is well accepted. This data Set (R1) contains sequence reads obtained from gut bacteriocytes of the glassy-winged sharpshooter, *Homalodisca coagulata*. The data sets used for testing CompostBin are described in Table 1, and experimental details are provided in Methods.

3.2 Performance

The most self-evident way of measuring error rates would be to report the percentage of reads misclassified by the algorithm. However, this method can artificially decrease the error-rates of data sets with skewed relative abundance of species. For example, consider a data set consisting of 90 sequences from species 1 and 10 sequences from species 2. If we classify 5 sequences of species 2 inaccurately, the error rate would be just 5%, even though 50% of the sequences have been misclassified. Therefore, we report a normalized error rate, where we compute the error rate for each bin and the error rate for the whole data set is the mean of these error rates.

CompostBin's accuracy in classifying reads from the test data sets is reported in Table 1. The normalized error rates is bounded by 10% in all the 13 data sets. The error rates are correlated mostly with the phylogenetic distances between the species and the relative abundance of species. For example, the highest error rates measured was 10% for Data Set S3 (sequences from *E. coli* and *Y. pestis*), where the phylogenetic distances between the genomes is small. Similarly, the error rates are comparatively high in S9 because there are very few sequences from the less abundant *Gluconobacter oxydans* and *Granulobacter bethesdensis*, which are also phylogenetically very close.

4 Discussion

In this paper, we report the development of CompopstBin, a new algorithm for the taxonomic binning problem associated with the analysis of metagenomic data. The principal novel aspect of our method is the observation that the high-dimensional Kmer frequency data for short sequences is noisy, and that one can deal with the noise by projecting the data into a carefully chosen lower-dimensional space. We illustrate that CompostBin can accurately classify sequences from low to medium complexity data sets into taxon-specific bins.

Unlike previous methods, CompostBin doesn't require training of the algorithm with data from sequenced genomes. This is critical for success when binning environmental shotgun data because more than 99.9% of microbes are currently unculturable and unlikely to be represented in the training data set. Even closely related organisms living in different environments may have divergent genome signatures. For example, Bacillus anthracis and Bacillus subtilis have widely differing GC content and genome signatures. One should also keep in mind that the currently available genomes are not a phylogenetically random sample, but rather are a highly biased collection of biomedically interesting genomes combined with an overabundance of strains of model organisms such as Escherichia coli.

We used the frequencies of hexamers (oligonucleotides of length 6) as the metric for our analysis of short sequences. The choice of hexamers was motivated by both computational and biological rationale. Since the length of the feature vector for analyzing Kmers is $O(4^K)$, both the memory and the CPU requirements of the algorithm become infeasible for large data sets when K is greater than six. Using hexamers is biologically advantageous in that, being the length of two codons, their frequencies can capture biases in codon usage. Similarly, hexamer frequencies can detect genomic biases resulting from the observed avoidance of specific palindromic words of lengths 4 and 6 from genomes due to the presence of restriction enzymes [26]. It should be noted that the frequencies of lower-length words are linear combinations of hexamer frequencies. For example: $f(AAAAA) = f(AAAAAA) + f(AAAAAC) + f(AAAAAG) + f(AAAAAT)$. Thus, our PCA-based method implicitly takes into account any biases in the frequencies of lower length words.

CompostBin is a work in progress, with several refinements of the algorithm planned for the future. Our method of analysis is based primarily on DNA composition metrics and, like all such methods, it cannot distinguish between organisms unless their DNA compositions are sufficiently divergent. Thus, our method would probably be unable to distinguish between strains of the same species. We believe that an ideal binning algorithm would also utilize additional types of information, such as assembly (depth of coverage and overlap information) and population genetics parameters. We have taken an initial step in this direction by using taxonomic information from phylogenetic markers to guide the clustering algorithm. We intend to develop other hybrid methods in the future that can tackle the very formidable problem of classifying sequences in complex metagenomic communities.

Acknowledgments

We thank Ambuj Tewari, Martin Wu, Lior Pachter, Jonathan Dushoff, Joshua Weitz, Dongying Wu, Amber Hartman, and Jenna Morgan for their helpful suggestions and comments. This work was supported in part by a Laboratory Directed Research and Development Program (Lawrence Berkeley Lab) Grant to Jonathan A. Eisen. S.C. and J.E. were partially supported by the Defense Advanced Research Projects Agency under grants HR0011-05-1-0057 and FA9550-06-1-0478. I.Y. and Z.B. were supported in part by the NSF under grant 0313390.

References

1. Rappe, M.S., Giovannoni, S.J.: The uncultured microbial majority. Annu Rev Microbiol 57, 369–394 (2003)
2. Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., Pace, N.R.: Rapid determination of 16s ribosomal rna sequences for phylogenetic analyses. Proc. Natl Acad. Sci. USA 82(20), 6955–6959 (1985)
3. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H., Smith, H.O.: Environmental genome shotgun sequencing of the sargasso sea. Science 304(5667), 66–74 (2004)
4. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., Banfield, J.F.: Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428(6978), 37–43 (2004)
5. Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., Nelson, K.E.: Metagenomic analysis of the human distal gut microbiome. Science 312(5778), 1355–1359 (2006)
6. Wu, D., Daugherty, S.C., Van Aken, S.E., Pai, G.H., Watkins, K.L., Khouri, H., Tallon, L.J., Zaborsky, J.M., Dunbar, H.E., Tran, P.L., Moran, N.A., Eisen, J.A.: Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol. 4(6), 188 (2006)
7. Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.H., Falcon, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M., Venter, J.C.: The sorcerer ii global ocean sampling expedition: Northwest atlantic through eastern tropical pacific. PLoS Biol. 5(3), e77 (2007)
8. Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., Mathur, E.J., Detter, J.C., Bork, P., Hugenholtz, P., Rubin, E.M.: Comparative metagenomics of microbial communities. Science 308(5721), 554–557 (2005)
9. von Mering, C., Hugenholtz, P., Raes, J., Tringe, S., Doerks, T., Jensen, L., Ward, N., Bork, P.: Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315(5815), 1126–1130 (2007)
10. Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C.: Megan analysis of metagenomic data. Genome Research (in press, 2007)

11. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., Glockner, F.O.: Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. BMC Bioinformatics 5(1471–2105 (Electronic)) (2004)
12. Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S., Ikemura, T.: Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. DNA Res 12(5), 281–290 (2005)
13. McHardy, A.C., Martin, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I.: Accurate phylogenetic classification of variable-length dna fragments. Nat Methods 4(1), 63–72 (2007)
14. Karlin, S., Burge, C.: Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11(7), 283–290 (1995)
15. Woyke, T., Teeling, H., Ivanova, N.N., Huntemann, M., Richter, M., Gloeckner, F.O., Boffelli, D., Anderson, I.J., Barry, K.W., Shapiro, H.J., Szeto, E., Kyrpides, N.C., Mussmann, M., Amann, R., Bergin, C., Ruehland, C., Rubin, E.M., Dubilier, N.: Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443(7114), 950–955 (2006)
16. Delcher, A.L., Bratke, K.A., Powers, E.C., Salzberg, S.L.: Identifying bacterial genes and endosymbiont dna with glimmer. Bioinformatics 23(6), 673–679 (2007)
17. Jolliffe, I.T.: Principal Component Analysis. Springer, Heidelberg (2002)
18. Kent, W.J.: Blat-the blast-like alignment tool. Genome Res 12(4), 656–664 (2002)
19. Lander, E.S., Waterman, M.S.: Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2(3), 231–239 (1988)
20. Tenebaum, J.B., Silva, V.D., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 190(5500), 2319–2323 (2000)
21. Wu, M., Eisen, J.: A simple, fast and accurate method for phylogenenomics inference approach (submitted, 2007)
22. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)
23. Schmid, R., Schuster, S.C., Steel, M.A., Huson, D.H.: Readsim- a simulator for sanger and 454 sequencing (in press, 2007)
24. Markowitz, V.M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, N., Kyrpides, N.C.: The integrated microbial genomes (img) system. Nucleic Acids Res. 34(Database issue), D344–348 (2006)
25. Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A.C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P., Kyrpides, N.C.: Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat. Methods 4(6), 495–500 (2007)
26. Gelfand, M.S., Koonin, E.V.: Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. Nucleic Acids Res. 25(12), 2430–2439 (1997)