CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads

Sourav Chatterji[1], Ichitaro Yamazaki[2], Zhaojun Bai[2], Jonathan A Eisen[*1,3,4]

[1]UC Davis Genome Center
[2]Department of Computer Science, UC Davis
[3]Section of Evolution and Ecology, UC Davis
[4]Department of Medical Microbiology and Immunology, UC Davis

Email: Sourav Chatterji - schatterji@ucdavis.edu; Ichitaro Yamazaki - yamazaki@cs.ucdavis.edu; Zhaojun Bai - bai@cs.ucdavis.edu; Jonathan A Eisen - jaeisen@ucdavis.edu;

[*]Corresponding author

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. It applies principal component analysis to project the 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 simulated data sets and on one metagenomic data set with known species assignments. CompostBin is a work in progress, with several refinements of the algorithm planned for the future.

Background

Microbes are ubiquitous organisms that play pivotal roles in the earth's bio-geochemical 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 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, 3] 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 [4–6]. These techniques, called environmental genomics or metagenomics, study unculturable communities by analyzing the pooled genomes of all the organisms present in the community. The genomic data obtained can be analyzed to make inferences about both who is out there and what they are doing (e.g., [7]).

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 [7]). However, the problem becomes more difficult in high-complexity communities with hundreds of species, such as the Sargasso Sea [4] and the human distal gut [6]. 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 [7].

A variety of approaches have been developed for binning: assembly, phylogenetic analysis [9], database

search [10], alignment with reference genome [11] and DNA composition metrics [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. 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 DNA metrics measure the frequency of occurrence of Kmers (oligonucleotides of length K)in a DNA sequence. Kmer frequencies have been used to distinguish between organisms since the 1960s [14]. With the explosion of available genomic data in the 1990s, Karlin and colleagues were able to establish that the relative abundances of various dinucleotide sequences (the dinucelotides odds ratio) is a genomic signature [15]. Subsequently, taxon-specific biases were also found in the frequencies of Kmers with lengths of four or more, leading to the use of a wide variety of methods exploiting this bias as a signature [12, 13, 16–19].

Unfortunately, many 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. 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. 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. The technique employed is Principal Component Analysis (PCA) [20], a multivariate analysis method previously applied in diverse biological areas ranging from ecology [21] to codon usage in genes [22] and even visualization of metagenomic binning results [23]. The normalized cut clustering algorithm used to cluster sequences into

taxon-specific bins is further 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.

Results

CompostBin was coded in C and Matlab on a 64bit Linux Machine. It is publicly available for download from the Eisen Lab website. As shown in the overview in Figure 1, CompostBin extracts the "principal components" of the DNA composition data and then uses PCA to project that data into a lower-dimensional space for further analysis. As shown in Figure 2, the algorithm can distinguish sequences from various species using just these first three principal components. Next, CompostBin uses the normalized cut clustering algorithm [24] to segment the data set into taxon-specific bins. Since the accuracy of phylogenetic assignment for reads containing phylogenetic marker genes is very high [9], we devised a semi-supervised approach which uses the phylogenetic information to guide the clustering algorithm. Simulated data sets were designed to evaluate the accuracy of CompostBin in binning metagenomic data sets of low and medium complexity. Additionally, we tested the data set on environmental shotgun reads from the gut of a glassy-winged sharpshooter [7]. Details of the test data sets and CompostBin's performance are provided in the next two sections.

Test Data Sets

Metagenomics being a relatively new field, standard data sets for testing binning algorithms have not yet been developed. 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. We used ReadSim [25], a publicly available program, to simulate Sanger sequences from known genomes. The 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.
Since environmental shotgun data is influenced by factors that may not be reflected in simulation experiments, we also tested CompostBin on a publicly available metagenomic data set whose solution is well accepted. 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.

Performance

CompostBin's accuracy in classifying reads from the test data sets is reported in Table 1. The percentage of misclassified reads is less than 6% in 11 of the 13 data sets. The highest error rates measured were 8.01% for Data Set S3 (sequences from *E. coli* and *Y. pestis*) and 7.24% for Data Set S5 (sequences from B. anthracis and L. monoytogenes). In both cases, the phylogenetic distance between genomes is comparatively small. However, the results from Data Set S1, which contains sequences from *Bacillus halodurans* and *Bacillus subtilis*, show that, in some instances, the algorithm can distinguish at the species level with high accuracy. The low error rate for the sharpshooter data set (R1) demonstrates the ability of the algorithm to handle the peculiarities of environmental shotgun data.

Discussion

In this paper, we report the development of a new approach to the taxonomic binning problem associated with the analysis of metagenomic data. Accurate binning is a crucial step in the application of environmental shotgun sequencing to the study of microbial communities. The problem of binning is intertwined with the fundamental questions of genomic signatures. Does the genome of every organism have a unique signature that distinguishes it from the genomes of all others? If so, what is the minimum length DNA sequence required to distinguish between two organisms? Even though it has been demonstrated that DNA-composition metrics such as the dinucleotide odds ratios are genome signatures [15], previous studies have typically worked with long sequences. In this study, we demonstrate that shrewdly analyzed Kmer frequency data from short sequences can also provide a signature. 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. This lower-dimensional space is determined by the principal components of the data. In a sense, it, too, is a "genome signature" that can be used to classify even short sequences into taxon-specific bins.

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.
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.
An ideal binning system would, like CompostBin, not 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.
CompostBin is a work in progress, with several refinements of the algorithm planned for the future.

- In the analyses reported here, we used PCA as the projection method for choosing the lower-dimensional space. Since PCA misses nonlinear structures of the underlying variables, we plan to look at alternative projection methods such as Projection Pursuit [27], ICA [28], and kernel PCA [29].

- CompostBin analyzes only the first three principal components in the data set. We plan to explore alternative approaches for choosing the optimal number of principal components (e.g., [30]).

- The clustering algorithm employed captures the global geometry of a data set using its k-nearest

neighbor graph. The highly accurate binning of the data sets reported in this paper was obtained using a fixed value of $k = 6$. However, the optimal value of k may depend on the characteristics of each individual data set. We plan to explore a technique which can automatically determine the optimal k through capture of the global geometry of the input data set.

- The similarity between two connected sequences in the nearest-neighbor graph was measured by the exponential inverse of their normalized Euclidean distance. We plan to explore alternative criteria for sequence similarity which have the potential to improve binning.

- The running time of our program can be improved by developing more efficient data-structures and by utilizing other numerical tools [31–33] to compute the principal components of the original data set and the eigenvectors of the similarity matrix.

- We observed that the separate clusters of rRNA genes can be outliers in many archaeal genomes and cause errors in the binning algorithm. Therefore, binning accuracy can be improved in future investigations by removing those genes prior to performing the DNA composition-based analyses.

- We plan to explore other potential applications of our algorithm to the study of genome structure and its variations within a single genome.

Methods

Generation of Test Sets

Genomic sequences of bacterial and archaeal isolate genomes were downloaded from the NCBI GenBank database [34]. ReadSim 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.

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 relative abundance ranged from 1:1 to 1:14. We included species of varying degrees of phylogenetic relatedness in order to test the ability of the program to discriminate between sequences at the species, genus, and family levels. The 12 simulated data sets created are described in Table 1.

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.

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 [11]. In the simulation experiments, the number of bins is set to the number of species in the simulation.

Feature Extraction by 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. As a result, all the sequences are initially represented as an $N \times 4,096$ feature matrix A, where N is the number of sequences being analyzed. 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.
The PCA algorithm [20] filters the noise and removes redundant variables to arrive at a new basis for expressing the data set. Furthermore, by using PCA, we may be able to find new underlying variables which reveal additional details about the mathematical structure of the system. Determining the number of principal components required for analysis is crucial to the success of the algorithm. Too few components and some important information may be lost. Too many components increases the noise in the data unnecessarily. When using PCA to bin sequences, use of just the first three principal components is adequate to separate sequences from different species. 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 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 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 [35]. 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 [36] 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 can then be employed [37, 38] to incorporate this information into the clustering algorithm.

Our binning algorithm uses the simplest approach to update the nearest neighbor graph. 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

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

The minimization of $NCut$ avoids the bias toward small segments, which results if the cut value is minimized without normalization [39]. 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 [24]. To generalize the algorithm for more than two bins, the binning algorithm uses PCA and the normalized cut algorithm iteratively, as described below.

Generalization to Multiple Bins

If the data set needs to be divided into more than two bins, an iterative algorithm is used and sequences in one of the bins are projected into their first principal components and bisected recursively until the required number of bins is obtained. Figure 4 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. If none of the bins in B have a small normalized cut value, the algorithm terminates. Both the principal components and the normalized cut of A can be computed using the Lanczos method [40] in $\mathcal{O}(N)$ space and $\mathcal{O}(Nm)$ time, where N is the number of sequences in A and m is a small constant representing the number of Lanczos iterations. By using kd-tree [41], a 6-nearest graph is computed in $\mathcal{O}(N)$ space and $\mathcal{O}(N \log(N))$ time. Computing and updating the similarity measures takes $\mathcal{O}(N)$ and $\mathcal{O}(l_{\max}^2)$ time, respectively, where $l_{\max}$ denotes the maximum number of phylogenetic markers for a particular species in A.

In order to separate A into K bins, the bisection algorithm needs to be called at most $2K-1$ times. Therefore, the running time of the whole algorithm is bound by $\mathcal{O}(NK(\log(N)))$.

Author contributions

S.C. and J.E. conceived the high level algorithm and designed the experiments to test the algorithm's performance. I.Y. and Z.B. were involved in the design and analysis of the clustering algorithm.

Acknowledgments

We thank Lior Pachter, Jonathan Dushoff, Joshua Weitz, Dongying Wu, Martin Wu, Amber Hartman, and Jenna Morgan for their helpful suggestions and comments. 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 MS, Giovannoni SJ: **The uncultured microbial majority.** *Annu Rev Microbiol* 2003, **57**:369–94.
2. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR: **Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses.** *Proc Natl Acad Sci U S A* 1985, **82**(20):6955–9.
3. Pace NR: **A molecular view of microbial diversity and the biosphere.** *Science* 1997, **276**(5313):734–40.
4. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304**(5667):66–74.
5. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428**(6978):37–43.
6. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome.** *Science* 2006, **312**(5778):1355–9.
7. Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL, Moran NA, Eisen JA: **Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters.** *PLoS Biol* 2006, **4**(6):e188.
8. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM: **Comparative metagenomics of microbial communities.** *Science* 2005, **308**(5721):554–7.
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* 2007, **315**(5815):1126–30.
10. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data**. *Genome Research* 2007, **in press**.
11. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcon LI, Souza

V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC: **The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.** *PLoS Biol* 2007, **5**(3):e77.

12. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO: **TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences.** *BMC Bioinformatics* 2004, **5**(1471-2105 (Electronic)):163.
13. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments.** *Nat Methods* 2007, **4**:63–72.
14. Swartz MN, Trautner TA, Kornberg A: **Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids.** *J Biol Chem* 1962, **237**:1961–7.
15. Karlin S, Burge C: **Dinucleotide relative abundance extremes: a genomic signature.** *Trends Genet* 1995, **11**(7):283–90.
16. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B: **Genomic signature: characterization and classification of species assessed by chaos game representation of sequences.** *Mol Biol Evol* 1999, **16**(10):1391–9.
17. Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ: **Evolutionary implications of microbial genome tetranucleotide frequency biases.** *Genome Res* 2003, **13**(2):145–58.
18. Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T: **Informatics for unveiling hidden genome signatures.** *Genome Res* 2003, **13**(4):693–702.
19. Delcher AL, Bratke KA, Powers EC, Salzberg SL: **Identifying bacterial genes and endosymbiont DNA with Glimmer**. *Bioinformatics* 2007, **23**(6):673–679.
20. Pearson K: **On Lines and Planes of Closest Fit to Systems of Points in Space**. *Philosophical Magazine* 1901, **2**(6):559–572.
21. Jackson DA: **Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches**. *Ecology* 1993, **74**(8):2204–2214.
22. Peden JF: **Analysis of Codon Usage**. *PhD thesis*, University of Nottingham 1999.
23. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ, Szeto E, Kyrpides NC, Mussmann M, Amann R, Bergin C, Ruehland C, Rubin EM, Dubilier N: **Symbiosis insights through metagenomic analysis of a microbial consortium.** *Nature* 2006, **443**(7114):950–5.
24. Shi J, Malik J: **Normalized Cuts and Image Segmentation**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2000, **22**(8):888–905.
25. Schmid R, Schuster SC, Steel MA, Huson DH: **ReadSim- A simulator for Sanger and 454 sequencing**. *in press* 2007.
26. Gelfand MS, Koonin EV: **Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes.** *Nucleic Acids Res* 1997, **25**(12):2430–2439.
27. Friedman JH, Tukey JW: **A Projection Pursuit Algorithm for Exploratory Data Analysis**. *IEEE Transactionon Computers* 1974, **C-23**(9):881– 890.
28. Comon P: **Independent component analysis, a new concept?** *Signal Process.* 1994, **36**(3):287–314.
29. Scholkopf B, Smola A, Miller KR: **Nonlinear component analysis as a kernel eigenvalue problem**. *Neural Comput.* 1998, **10**(5):1299–1319.
30. Cangelosi R, Goriely A: **Component retention in principal component analysis with application to cDNA microarray data**. *Biol Direct.* 2007, **2**(2).
31. Lehoucq RB, Sorensen DC, Yang C: *ARPACK User's Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.* SIAM, Philadelphia 1998.
32. Wu K, Simon H: **TRLAN User Guide**. Technical report LBNL-42953, Lawrence Berkeley National Lab 1999.

33. McCombs JR, Stathopoulos A: **PRIME: PReconditioned Iterative MultiMethod Eigensolver: Methods and software description.** Technical report WM-CS-2006-08, The College of William and Mary, Williamsburg, Va 2006.

34. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2007, **35**(Database issue):D21–5.

35. Tenebaum JB, Silva VD, Langford JC: **A global geometric framework for nonlinear dimensionality reduction.** *Science* 2000, **190**(5500):2319–2323.

36. Wu M, Eisen J: **A simple, fast and accurate method for phylogenenomics inference approach**. *submitted* 2007.

37. Kamvar SD, Klein D, Manning C: **Spectral learning.** *Proc. 17th Intl. Joint Conf. on Artificial Intelligence* 2003.

38. Yu S, Shi J: **Segmentation given partial grouping constraints.** *IEEE Trans. on Pattern Analysis and Machine Intelligence* 2004, **26**:173–183.

39. Wu Z, Leahy R: **An optimal graph theoretic approach to data clustering: theory and its application to image segmentation.** *PAMI* 1993, :1101–1113.

40. Golub GH, van Loan CF: *Matrix computations*, Johns Hopkins University Press, Baltimore 1989 chap. 9.

41. Mount DM, Arya S: **ANN: A library for approximate nearest neighbor searching.** *http://www.cs.umd.edu/∼mount/ANN/.*

42. Markowitz VM, 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 NC: **The integrated microbial genomes (IMG) system.** *Nucleic Acids Res* 2006, **34**(Database issue):D344–8.

Figures

Figure 1 - Overview of the Binning Algorithm

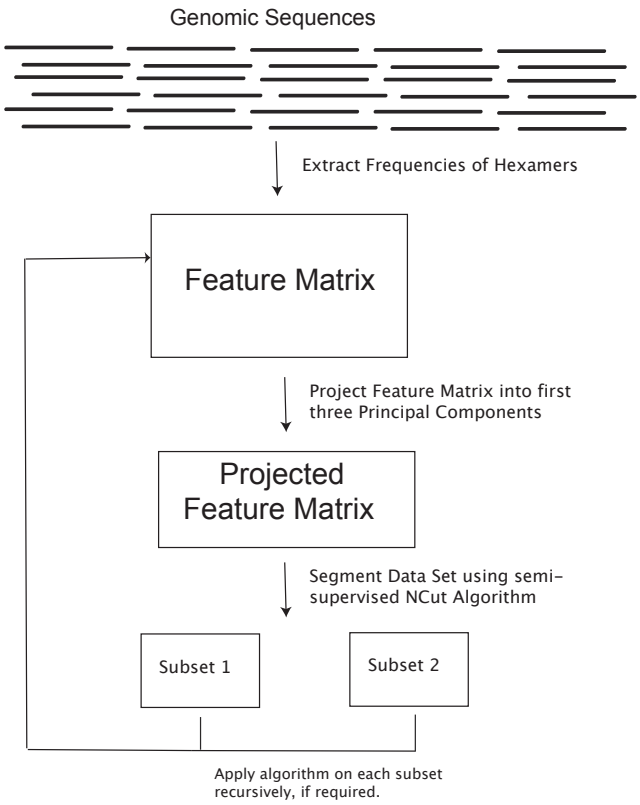

Figure 1: High-level overview of the CompostBin algorithm. Each sequence is represented by a $4,096$-length feature vector, where each component of the vector represents the frequency of one of $4,096$ hexamers. Thus, N sequences are initially represented as a $4,096 \times N$ feature matrix. Principal 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.

Figure 2 - Separation of sequences by PCA

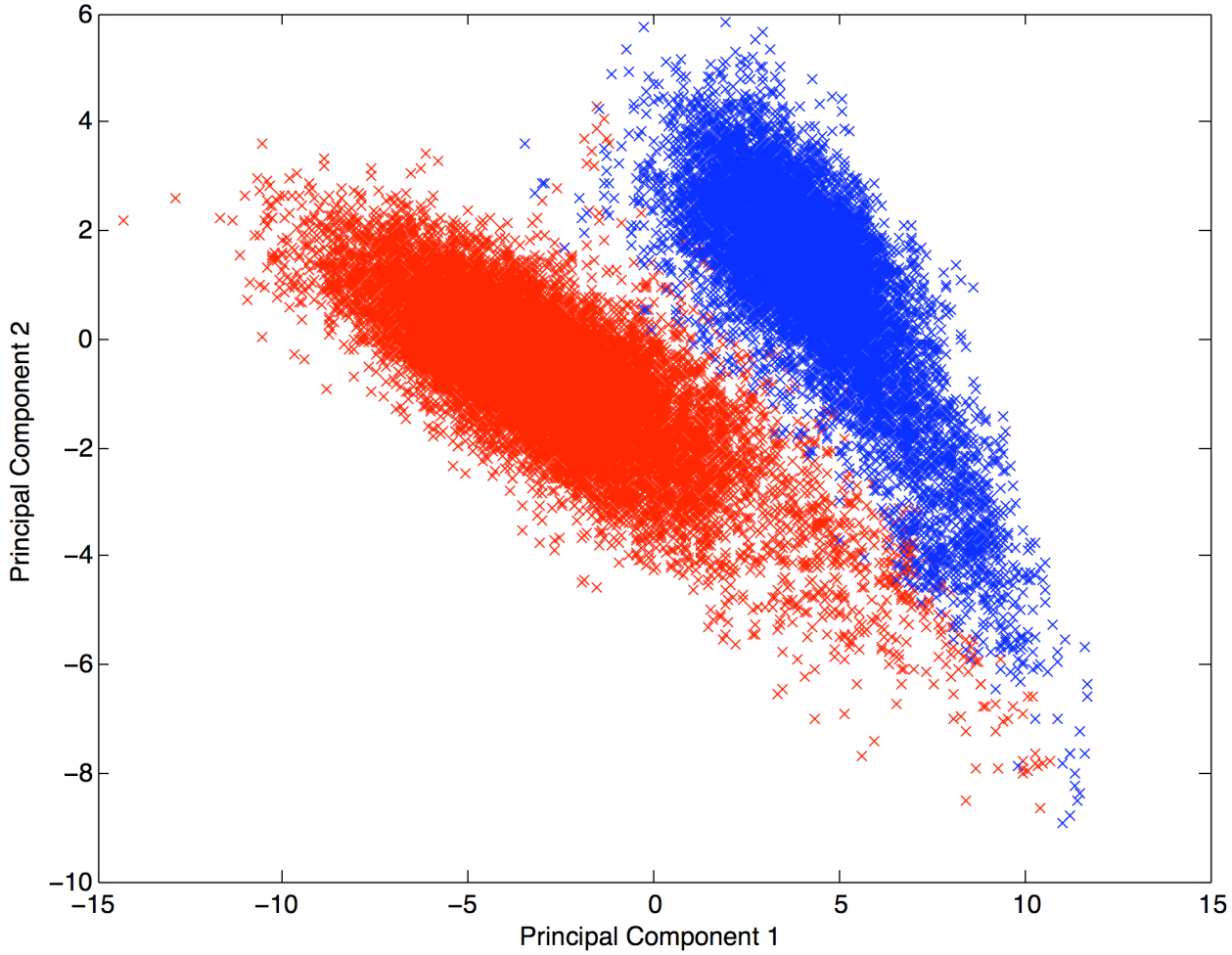

Figure 2: Figure illustrating the separation of sequences according to species by using just the first few principal components of the data. This data set contains sequences from two alphaproteobacteria, *Gluconobacter oxydans* and *Rhodospirillum rubrum*, which have GC content of 0.65 and 0.61, respectively.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

Figure 3 - The Bisection Algorithm

1 **Bisection**(A,L)

2 Calculate principal components of A.

3 Project A into the first three PCs to obtain A_p.

4 Compute G, the 6-nearest neighbor graph of A_p.

5 Update G by using information from L.

6 Bisect A into two sets A_1 and A_2 by approximate NCut.

7 Calculate cut, the value of normalized cut between A_1 and A_2.

8 return (A_1,A_2,cut)

Figure 3: Pseudocode describing the bisection algorithm used to bisect a data set into two taxon-specific subsets. A is the feature matrix and L contains the labeling information for A. This procedure is used iteratively by the binning algorithm described in 4.

Figure 4 - The Binning Algorithm

1 **Bin(A,L,K)**

2 **// Initialization**

3 $B = \{B_1\}$ where $B_1 = A$.

4 If $K = 1$, then return B.

5 Store the Ncut value for B_1 : $[A_1(B_1), A_2(B_1), \text{Ncut}(B_1)] = \text{Bisect}(B_1, L)$.

6 **// Recursively bisect until there are K bins**

7 *Repeat until* $|B| = K$

8 Pick the bin $\hat{B} \in B$ with the smallest $\text{NCut}(\hat{B})$.

9 If $\text{NCut}(\hat{B}) > threshold$, return B.

10 **// Divide the bin $\hat{B}$ into two bins $A_1(\hat{B}), A_2(\hat{B})$**

11 $[A_1(\hat{B}), A_2(\hat{B}), \text{Ncut}(\hat{B})] = \text{Bisect}(\hat{B}, \hat{L})$.

12 $B = B \cup \{A_1(\hat{B}), A_2(\hat{B})\} \setminus \hat{B}$.

13 If $|B| = K$, then return B.

14 Store the Ncut values for $A_1(\hat{B})$ and $A_2(\hat{B})$ by calling *Bisect*.

Figure 4: Pseudocode describing the iterative PCA and the normalized cut algorithm used for binning. 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. Lines in bold starting with "//" contain comments intended to help understand the code. Note that the calls to `Bisect` in Line 11 can be avoided at the cost of extra memory if one stores the optimal cut for each set in B during the calls to `Bisect` in Lines 5 and 14.

Tables

Table 1 - Test Data Sets and Binning Accuracy

Table describing the simulated and real data sets used to test the binning algorithm. Each data set is assigned a unique ID for reference. IDs of simulated data sets start with S and IDs of experimental data sets start with R. 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 [42] and can be used for assessing the phylogenetic diversity. The error rate of the binning algorithm on each test set is shown in the last column.

ID	Species	Ratio	Taxonomic Differences	Error
S1	*Bacillus halodurans* [0.44] & *Bacillus subtilis* [0.44]	1:1	Species	5.74%
S2	*Gluconobacter oxydans* [0.61] & *Granulobacter bethesdensis* [0.59]	1:1	Genus	3.69%
S3	*Escherichia coli* [0.51] & *Yersinia pestis* [0.48]	1:1	Genus	8.01%
S4	*Rhodopirellula baltica* [0.55] & *Blastopirellula marina* [0.57]	1:1	Genus	1.98%
S5	*Bacillus anthracis* [0.35] & *Listeria monocytogenes* [0.38]	1:2	Family	7.24%
S6	*Methanocaldococcus jannaschii* [0.31] & *Methanococcus mariplaudis* [0.33]	1:1	Family	0.56%
S7	*Thermofilum pendens* [0.58] & *Pyrobaculum aerophilum[0.51]*	1:1	Family	0.21%
S8	*Gluconobacter oxydans* [0.61] & *Rhodospirillum rubrum* [0.65]	1:1	Order	1.15%
S9	*Gluconobacter oxydans* [0.61], *Granulobacter bethesdensis* [0.59], & *Nitrobacter hamburgensis* [0.62]	1:1:8	Family Order	2.28%
S10	*Escherichia coli* [0.51], *Pseudomonas putida* [0.62], & *Bacillus anthracis* [0.35]	1:1:8	Order Phylum	1.73%
S11	*Gluconobacter oxydans* [0.61], *Granulobacter bethesdensis* [0.59], *Nitrobacter hamburgensis* [0.62], & *Rhodospirillum rubrum* [0.65]	1:1:4:4	Family Order	5.28%
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	3.35%
R1	Glassy-winged sharpshooter endosymbionts	-	-	5.9%

Additional Files

Additional file 1 — CompostBin Code

File 1, in tar gunzipped format contains the CompostBin source code in C/Matlab.

Additional file 2 — Test Data Sets

File 2, in tar gunzipped format contains the data sets that was used to test CompostBin's performance.