SIMCOMP: A Hybrid Soft Clustering of Metagenome Reads

Shruthi Prabhakara* and Raj Acharya

Department of Computer Science and Engineering
Pennsylvania State University, University Park, State College, PA 16801
{sap263,acharya}@cse.psu.edu

Abstract. A major challenge facing metagenomics is the development of tools for the characterization of functional and taxonomic content of vast amounts of short metagenome reads. In this paper, we present a two pass semi-supervised algorithm, SimComp, for soft clustering of short metagenome reads, that is a hybrid of comparative and composition based methods. In the first pass, a comparative analysis of the metagenome reads against BLASTx extracts the reference sequences from within the metagenome to form an initial set of seeded clusters. Those reads that have a significant match to the database are clustered by their phylogenetic provenance. In the second pass, the remaining fraction of reads are characterized by their species-specific composition based characteristics. SimComp groups the reads into overlapping clusters, each with its read leader. We make no assumptions about the taxonomic distribution of the dataset. The overlap between the clusters elegantly handles the challenges posed by the nature of the metagenomic data. The resulting cluster leaders can be used as an accurate estimate of the phylogenetic composition of the metagenomic dataset. Our method enriches the dataset into a small number of clusters, while accurately assigning fragments as small as 100 base pairs.

1 Introduction

Metagenomics is defined as the study of genomic content of microbial communities in their natural environments, bypassing the need for isolation and laboratory cultivation of individual species[1]. Its importance arises from the fact that over 99% of the species yet to be discovered are resistant to cultivation[2]. Metagenomics promises to enable scientists to study the full diversity of the microbial world, their functions and evolution, in their natural environments.

Metagenomics projects collect DNA from environments that are characterized by large disparity in sequence coverage and abundance of species distribution. Sequencing technologies are used to survey the metagenomic content. The recent ultra-high throughput sequencing technologies [3] produce relatively short reads, 25-400 base pairs(bp), and enormous datasets, thereby creating new computational challenges for metagenomics. It is critical that we develop fast and

* Corresponding author.

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 113–124, 2010.

accurate tools for assembling and characterizing the phylogenetic provenance and the relative abundance of different species in a metagenomic sample. Clustering of metagenome reads is one such tool that provides deeper insight into the structure of the community and hence, can be used to model the ecological and population parameters. This pre-processing step can lead to faster and more robust assembly by reducing the search space[14].

2 Related Work

Methods for clustering reads proposed so far in literature can be categorized into two main approaches; comparative(or similarity) and composition based. Comparative based methods align metagenomic sequences to close phylogenetic neighbors in existing databases and hence depend on the availability of closely related genomes in the database[7,6,11]. Such methods fail to find any homologs for new families. Composition based methods, on the other hand, distinguish between clades by using intrinsic features of reads such as oligomer frequencies[10,12,13], codon usage preferences[17] or GC content[16]. The strength of this approach is that no reference database is required. However, oligomer composition of reads shorter than 1 kbp carry insufficient signal to be able to differentiate between species. Composition based clustering of metagenome reads complements the comparative analysis[12].

The last decade has seen an explosion in the computational methods developed to analyze metagenomic data. A number of methods for classifying(as opposed to clustering) metagenome reads into taxon-specific bins have been proposed in literature. Phylopythia[10] is a supervised composition based classification method that trains a support vector machine to classify sequences of length greater than 1 kbp. Phymm uses interpolated markov models to characterize variable length DNA sequences into their phylogenetic groups[12]. Its accuracy of assignment drops drastically for short reads and reads from unknown species. Nasser et al.[14] demonstrated that a k-means based fuzzy classifier, trained using a maximal order markov chain, can separate 1kbp reads with a high accuracy at phylum level. All the above supervised methods depend on the availability of reference data for training. These methods assume the prior knowledge of the number of classes. A metagenomic dataset may contain reads from unexplored phyla which cannot be labeled into one of the existing classes.

Li et al. propose a composition based leader clustering algorithm that clusters highly homologous sequences in order to condense a large database[9]. More recently, Chan et al. developed a semi-supervised seeded growing self-organizing map to cluster metagenomic sequences[18]. It extracts 8-13 kbp of flanking sequences of highly conserved 16S rRNA from the metagenome and uses them as seeds to assign the remaining reads using composition based clustering. CompostBin uses weighted PCA to project the DNA composition data into lower-dimensional space, and then uses the normalized cut clustering to classify reads into taxon-specific bins[20]. Likely-Bin is an unsupervised method for binning short reads by taxonomy on the basis of their k-mer distributions[21].

MEGAN, a metagenome analysis software system [11], on the other hand, uses sequence homology to assign reads to common ancestors based on best match as given by BLAST(Basic Local Alignment Search Tool)[19]. As most of the extant databases are highly biased in their representation of true diversity, methods such as MEGAN fail to find any homologs for new families. Most metagenomic analysis methods until now have been relatively inaccurate in classifying reads as short as 100 base pairs.

Increased amounts of polymorphism and horizontal gene transfer in metagenome reads leads to conflicts in assembly and taxonomic analysis. Reads from closely related species will most likely have homologous sequences shared between clusters that fuzzify the cluster boundaries[18]. Another characteristic of these datasets is the incomplete and fragmentary nature of the metagenome reads that reduces the quality of annotation. However, clipping low quality reads such as chimeras can exclude potentially useful sequences. Hence, in light of the new data, we need to adapt the traditional approaches to metagenome analysis. Overlapping clusters generated by a soft clustering algorithm such as the one proposed in this paper elegantly handle the problems associated with the nature of metagenomic data while providing tolerance for the noise due to errors in sequencing and fragmentation. The soft boundaries between clusters provide the flexibility to capture the misplacements of reads due to polymorphism or over representation of conserved regions, thereby providing interesting insights into the data.

Our work is inspired by the works of Dalevi et al.[6] and Folino et al.[7]. In [6], the authors propose a method for clustering reads based on a set of proteins, called proxygenes. The protein hits are obtained by BLASTx (specialized nucleotide-protein BLAST) of the reads against a reference proteome database. Their work is extended in [7], where a method based on weighted proteins is used to cluster the reads, resulting in overlapping clusters, each represented by a proxygene. The underlying basis of the above methods is that a high sequence similarity between the read and the proxygene implies phylogenetic proximity of the organisms from which they originated [6]. Consequently, the taxonomic annotation of the proxygene can be used in assessing that of the reads in the cluster. Both the methods use the comparative approach and hence rely on the use of a reference database that contains closely related genomes. However, in a typical metagenome dataset, majority of the reads may exhibit no similarity to any known sequence in the database. In such a scenario, these methods will fail to assign these reads to any cluster.

In this paper, we propose a two pass semi-supervised algorithm for soft clustering of short metagenome reads. We call our method SimComp; a hybrid of similarity and composition based methods. The objective of our method is to enrich the dataset into a small number of clusters such that reads within a cluster are phylogenetically closer than reads from different clusters. Each cluster is defined by a core consisting of reads that definitely belong to the cluster and a fringe that has reads which may overlap with other clusters. We make

no assumptions about the taxonomic distribution of the metagenome dataset. SimComp makes use of a reference database, however is not dependent on it.

In the first pass, a comparative analysis of the metagenome reads against an existing database, using BLASTx, extracts reference sequences from within the dataset to form an initial set of seeded clusters. Reads that have a significant match to the database are clustered by their phylogenetic provenance. In the second pass, the global clade-specific characteristics(e.g. oligomer frequency) are used to cluster the remaining reads by a soft leader clustering algorithm described in [1]. Our algorithm groups the reads into overlapping clusters, each with its read leader. The fringes of the clusters accomodate the ambiguity associated with reads in the dataset. It automatically performs the selection of the number of clusters. Essentially, the comparative analysis of reads avails apriori biological knowledge in existing protein database to form initial set of seeded clusters.Then, the composition based characterization of remaining fraction of reads, thereby facilitating a means of exploring novel species.

3 An Overview of Methods and Algorithm

SimComp is based on the Adaptive Rough Fuzzy Leader Clustering presented by Asharaf et al.[8]. The authors use rough set theory to define the clusters. Each cluster has a core(lower bound) and a fringe(upper bound) and is represented by a read leader. The core contains all the reads that definitely belong to the cluster. Reads in the core are mutually exclusive between the clusters. There can be an overlap in the fringes of two or more clusters.

3.1 Comparative Clustering

In the comparative pass of the algorithm, as in [7,6], we associate a list of protein hits with each read, identified by BLASTx. Each hit consists of one protein, two score values called bits and identities which describe the significance of read-protein alignment, and a confidence value called E-value which describes the likelihood that the sequence will occur in the database by chance. We further use the measure defined in [7] (explained in the Appendix) for assigning weights to the each of the proteins, such that proteins that cover more reads are assigned smaller weights. Proteins that are below a predefined protein threshold form the proxygenes, the rest are discarded. The proxygenes are clustered with the corresponding best hit reads(as identified by BLASTx). For each cluster thus formed, the most representative read is chosen as the leader(seed of a cluster).

3.2 Composition Based Clustering

The reads remaining after the first pass are clustered using the soft leader clustering algorithm based on sequence composition. In this pass, each unclustered read is compared with the existing read leaders. The similarity between the read and the leaders along with the sequence thresholds determines whether the read gets added to the core of some cluster or fringes of one or more clusters, or the read itself gets added as a leader. The steps in SimComp are outlined below.

3.3 Definitions

Cluster. Each cluster consists of a read leader, representative of the set of reads in the cluster. A cluster is defined by the following parameters:

- Protein threshold (PT): Proteins with weight below the threshold form proxygenes. Each proxygene is representative of a cluster with the corresponding reads(as identified by BLASTx). Rest of the proteins are discarded. The weight assigned to a protein is measured by two score values, i.e. bits and identities, and a confidence value called E-value[7].
- User defined core and fringe sequence similarity threshold for clusters (RT_C and RT_F): If the similarity between the read and its nearest leader is greater than RT_C, the read is added to the core of a cluster. Otherwise, if the similarity between the read and the corresponding cluster leaders is greater than RT_F, the read is added to the fringes of one or more clusters.

Sequence similarity. Each sequence is represented by a vector of oligomer frequencies, $v = (f_1, f_2...f_q)$; where for each oligomer of length n, $O = (o_1, o_2...o_q)$ is the set of all possible oligomers, f_i is the frequency of oligomer pattern o_i in the read, q is the number of oligomer patterns of length n possible, i.e. 4^n. Each vector is normalized relative to the length of the sequence. $S(x, y)$ gives the similarity between read x and leader y. We define sequence similarity as the number of fixed length oligomers shared between x and y.

Fuzzy membership. U_{ik} is the fuzzy membership of the read r_i in a cluster represented by Leader L_k.

$$U_{ik} = \sum_{j=1}^{N} \frac{S(r_i, L_k)}{S(r_i, L_j)} \tag{1}$$

3.4 SIMCOMP : Outline of the Algorithm

The algorithm proceeds in two passes. Let $R = (r_1, r_2, ...r_n)$, be the set of all reads and N be the number of clusters at any point in the algorithm.

I. Comparative Clustering: In the first pass, metagenome reads are grouped into clusters based on similarity of the reads to the proteins in the reference database.

1. Extract all proteins that R has hits to(by BLASTx).
2. Assign weights to all the proteins based on equation described in [7] (see Appendix). Proteins with weight below PT form proxygenes.
3. Each proxygene, along with the corresponding best hit reads (identified by BLASTx) form a cluster.
4. For each of the clusters, find a read leader that is most representative of the reads in the cluster, i.e. one whose sum of sequence similarity from all the other reads in the cluster is maximum.

II. Composition Based Clustering: In the second pass, we use the similarity measure based on oligomer frequency(defined above) to cluster the remaining reads.

1. All the reads from the original dataset that have not yet been clustered form the remaining read set. For each read in the remaining read set, compare the read with the existing read leaders. Depending on the value of RT_C, RT_F and sequence similarity between the read and the leaders, one of the three cases can arise for assignment of the current read:
   (a) It gets added to the core of a cluster. The current read gets added to the core of a cluster represented by leader L_p, if $\max(S(r_i, L_k)/k = 1...N) = D_{ip}$ and $D_{ip} > RT_C$.
   (b) It gets added to the fringes of one or more clusters. r_i falls into the fringes of all the clusters L_p for which $S(r_i, L_p) > RT_F$ and $S(r_i, L_p) < RT_C$.
   (c) Otherwise, r_i gets added as leader since it is outside the region defined by any of the existing clusters.

4 Results

We implemented our algorithm in Matlab. All experiments were run on an IBM X3550 server with 8GB memory. We tested our method on the simulated metagenome datasets M1, M2 and M3, introduced in [6], each at a coverage of 0.1X. These datasets were sequenced at Joint Genome Institute using the 454 pyrosequencing platform that produces ∼100 bp reads. We present results from experiments on M1 dataset only due to constraints in space. The characterization of reads at the taxonomic level of an organism for M1 is as shown in Fig 1. We used the default parameters of BLASTx, and NR[15] (Non-Redundant) protein sequence database as our reference. We have conducted experiments for varying values of user-defined thresholds(RT_C, RT_F) and lengths of oligomers. Based on the evaluation of our method on M2 and M3, we observed that proteins with weight below the 1^{st} percentile cover all the taxonomies that reads belong to. Therefore, we selected the 1^{st} percentile of weight as our protein threshold. The most time consuming component of SimComp is generating the BLASTx output. Once this output has been generated, the algorithm performs a single pass over the BLASTx output and the dataset to cluster the reads and hence is very efficient.

4.1 Accuracy across Taxonomic Ranks

In this paper, we use two measures to evaluate the effectiveness of our method: Mode Cluster Purity and Leader Cluster Purity. Mode Cluster Purity is defined as the maximum fraction of reads in a cluster belonging to the same taxon[7]. We define Leader Cluster Purity as the fraction of elements in the cluster belonging to the same taxon as the read leader. This measure determines how well our algorithm models the problem of classifying reads from species that have never been seen before. Depending on the elements of the cluster that we evaluate on, cluster purity can be further divided into core cluster purity(all the reads in the core of the cluster) and total cluster purity(all the reads in the cluster). In evaluating both the measures, we take into account only the non-singleton clusters, as a singleton cluster has a cluster purity of 1.

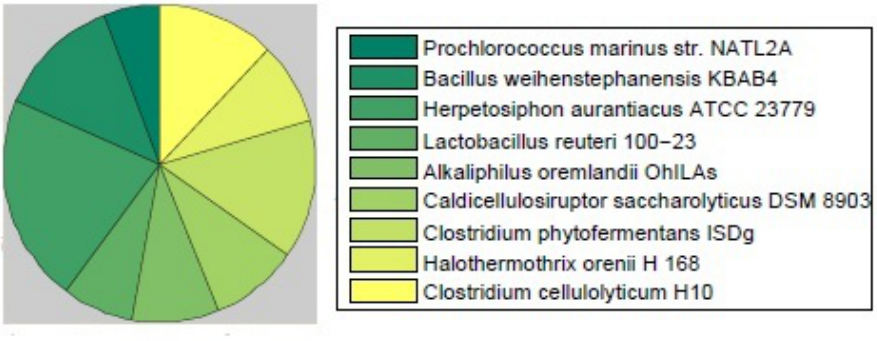

Fig. 1. Organism level characterization of M1 dataset

In Fig 2, we plot the taxonomic distribution of reads in M1 at phylum, class, order and family level($RT_C = 15$ and $RT_F = 12$ and length of oligomer = 6) as predicted by our algorithm. To measure the taxonomic distribution, all the reads in the cluster are assigned the same taxa as the read leader. Our method yields satisfactory results at all ranks. Hence, leaders of the clusters can be used as an accurate estimate of the phylogenetic composition of the metagenome. In [6,7], only those reads that have significant hits in the BLASTx output are selected for further clustering, the remaining reads are discarded. As opposed to this, in our method, we cluster all the reads in the dataset, even if no significant hits to the reference database are obtained. In Fig 3, we have plotted three measures for dataset M1 across all taxonomic ranks. By definition, mode cluster purity is greater than or equal to leader cluster purity. From the plot, we conclude that the cluster purity of the core is higher than that of the entire cluster at all ranks. This asserts our algorithms ability to filter out low quality reads into the fringe of a cluster.

4.2 Length of Oligomer

Oligomer frequency of genomes has been shown to reflect clade-specific characteristics and thus form a genome signature[4]. Teeling et al.[5] have shown that tetranucleotide frequency has a higher discriminatory power than GC content for phylogenetic grouping of reads. We have evaluated the accuracy of assignment of reads to clusters for a range of oligomers varying from trimers to hexamers. Fig 4 shows the plot of percentage of non-singleton clusters with purity values in the range [0.1,1] for varying lengths of oligomer. From our experiments, we conclude that hexamers have the best discriminatory power for clades at higher taxonomic ranks. With reads as small as 100 bp, not many reads cross that high a similarity threshold for hexamers. This explains the increase in number of singleton clusters with the increase in read threshold.

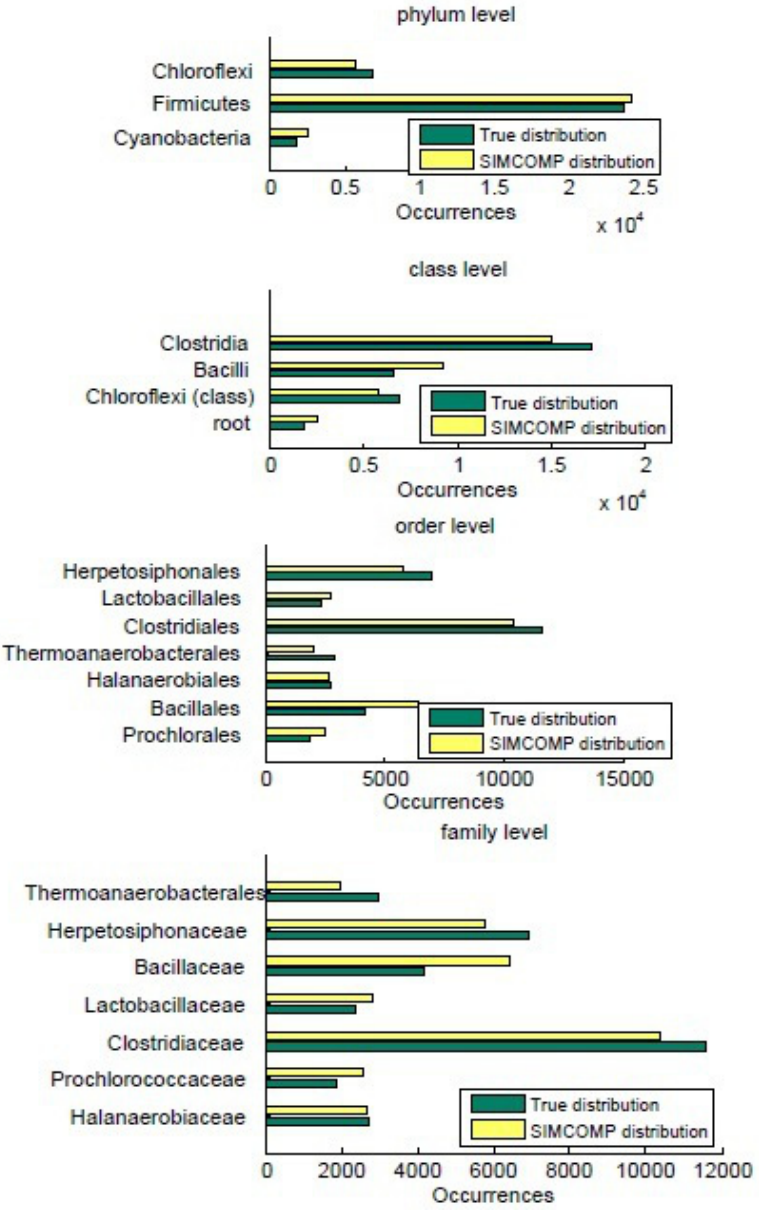

Fig. 2. Taxonomic Distribution Across Ranks (Phylum, Class, Order, Family

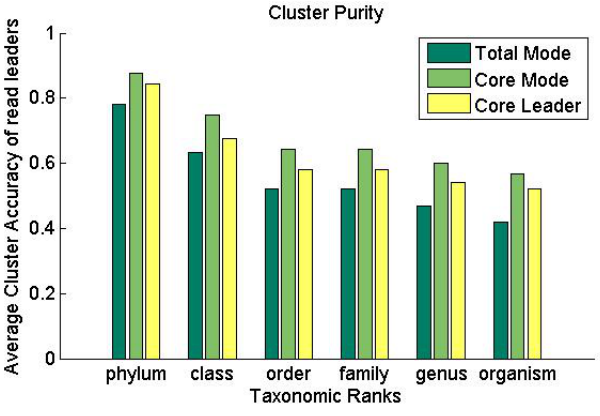

Fig. 3. Average cluster purity across taxonomic ranks for ($RT_C = 15$ and $RT_F = 12$ and length of oligomer = 6, Number of Clusters = 2430)

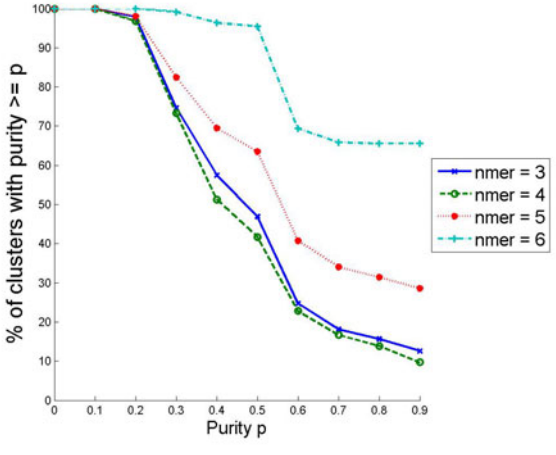

Fig. 4. Plot of percentage of non-singleton clusters for different values of purity with $RT_C = 25$ and $RT_F = 22$ and varying values of oligomers

4.3 Read Threshold

In our method, sequence similarity between two reads is measured as a function of number of fixed length oligomers shared between the two reads. A read is added to the core of an existing cluster only if the read similarity between the read and the cluster leader is above a certain threshold. Fig 5 plots the mode cluster purity for different values of read thresholds. The curve for $RT_C = 25$ clearly dominates the others. This is justified as clusters with large read thresholds are smaller in size and hence are likely to have a high purity. Table.1 summarizes the results for a fixed oligomer length of 6 and varying read thresholds. Cluster purity increases with the increase in read thresholds, for the reasons cited above.

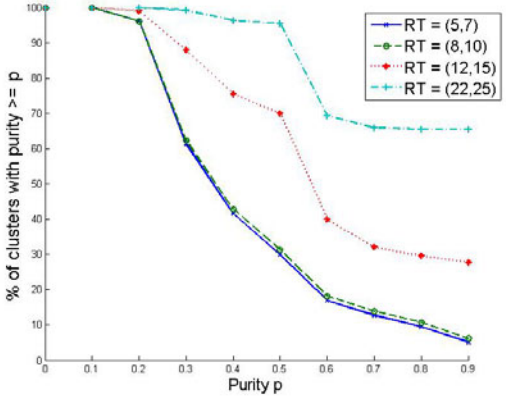

Fig. 5. Plot of percentage of non-singleton clusters for different values of purity with oligomer lenght = 6 and varying values of Read Threshold (Core, Fringe)

Table 1. Summary of the results of experiments for oligomer length = 6 and varying Read Thresholds

RT_C	10	15	20
RT_F	8	12	17
Number of Clusters	1482	2430	14250
Maximum size of clusters	320	415	288
Number of singleton clusters	6	67	5865
Reduction factor	0.042	0.068	0.4
Mode Cluster Purity at Phylum level	79.93	88.14	96.95
Mode Cluster Purity at Organism level	40.88	61.75	88.41

5 Conclusion

In this paper, we proposed SimComp, a soft clustering method that allows complete and accurate characterization of short metagenome reads that come from a spectrum of known and unknown species. We clustered a simulated dataset using a hybrid of comparative and composition based method. The overlap between the clusters accomodates the ambiguity associated with metegenomic data. It does not require assembled contigs or training on a reference set, nor does it make any assumptions on the number of species or the nature of the dataset.

The oligomer composition of reads as short as 100 bp does not provide sufficient signal to differentiate between species. For best results, we would like to test our algorithm on metagenome datasets with larger read length. Phenomena such as polymorphism and horizontal gene transfer can complicate phylogenetic clustering. As proposed in this paper, the soft boundary between clusters has the ability to capture such misplacements providing interesting insights into the data. We believe soft clustering has a promising role in classifying metagenome reads and we wish to investigate its scope in the future.

Acknowledgments. We would like to thank Mavrommatis Konstantinos for providing datasets in [6] and Elena Marchiori and Fabio Gori for providing useful information about their paper [7]. We would also like to thank Piotr Berman for suggesting useful changes to the paper.

References

1. Chen, K., Pachter, L.: Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comp. Biol., 1–24 (2005)
2. Rappe, M.S., Giovannoni, S.J.: The uncultured microbial majority. Annual Rev. Microbiol., 357–369 (2003)
3. Pop, M., Salzberg, S.L.: Bioinformatics challenges of new sequencing technology. Trends Genet. 24, 142–149 (2008)
4. Karlin, S., Ladunga, I., Blaisdell, B.E.: Heterogeneity of genomes: measures and values. Proc. Natl. Acad. Sci. USA 91, 12837–12841 (1994)
5. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., Glockner, F.: Application of tetranucleotide frequencies for the assignment of genomic fragments. Environmental Microbiology 6, 938–947 (2004)
6. Dalevi, D., Ivanova, N.N., Mavromatis, K., Hooper, S.D., Szeto, E., Hugenholtz, P., Kyrpides, N.C., Markowitz, V.M.: Annotation of metagenome short reads using proxygenes. Bioinformatics 24(16) (2008)
7. Folino, G., Gori, F., Jetten, M.S., Marchiori, E.: Clustering Metagenome Short Reads Using Weighted Proteins. In: EvoBIO '09: Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (2009)
8. Asharaf, S., Narasimha Murty, M.: An adaptive rough fuzzy single pass algorithm for clustering large data sets. Pattern Recognition 36(12) (2003)
9. Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22 (2006)
10. McHardy, A.C., Martin, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I.: Accurate phylogenetic classification of variable-length DNA fragments. Nature Methods 4, 63–72 (2007)
11. Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C.: MEGAN analysis of metagenomic data. Genome Res. 17, 377–386 (2007)
12. Brady, A., Salzberg, S.L.: Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nature Methods 1358 (2009)
13. 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, 163 (2004)
14. Nasser, S., Breland, A., Harris, F.C., Nicolescu, M.: A fuzzy classifier to taxonomically group DNA fragments within a metagenome. Annual Meeting of the North American Fuzzy Information Processing Society, 1–6 (2008)
15. Non-Redundant Proteome database, `ftp://ftp.ncbi.nlm.nih.gov/blast/db`
16. Bentley, S.D., Parkhill, J.: Comparative genomic structure of prokaryotes. Annual Review of Genetics 38, 771–792 (2004)
17. Bailly-Bechet, M., Danchin, A., Iqbal, M., Marsili, M., Vergassola, M.: Codon Usage Domains over Bacterial Chromosomes. PLoS Computational Biology 2(4), e37 (2006)

18. Chan, C., Hsu, A., Halgamuge, S., Tang, S.: Binning sequences using very sparse labels within a metagenome. BMC Bioinformatics 9, 215 (2008)
19. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990)
20. Chatterji, S., Yamazaki, I., Bai, Z., Eisen, J.: CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. In: Vingron, M., Wong, L. (eds.) RECOMB 2008. LNCS (LNBI), vol. 4955, pp. 17–28. Springer, Heidelberg (2008)
21. Kislyuk, A., Bhatnagar, S., Dushoff, J., Weitz, J.S.: Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinformatics 10, 316 (2009)

Appendix

As in [7], from each hit that BLASTx outputs for a given read r, we extract a 4-dimensional vector $h = (p; S_B; Id; E)$ where p is the matched protein, S_B the bit score, Id the identities score, and E the E-value of that match. For a read r let Hit_r be the sequence, sorted in increasing order of E-values, of its hits. Denote by $r_1, ..., r_m$ the set of reads r with non-empty Hit_r. Let $P = \{p_1, ..., p_n\}$ be the set of proteins occurring in $\cup_{i=1}^{m} Hit_i$ For each protein $p \in P$, the set H_p is defined as:

$$H_p = \{h \in \cup_{i=1}^{m} Hit_i | h(1) = p\} \tag{2}$$

where $h(1)$ denotes the first component of the hit vector h. Thus H_p consists of the selected hits containing p. We use the equation described in [7] to assign weights to the each of the protein hits that BLASTx outputs. Weight of protein p is defined as:

$$w_p = 1 + \lceil \frac{1}{|H_p|} \sum_{h \epsilon H_p} (100 \frac{max\_score - S_B(h)}{max\_score - min\_score} + 100 - Id(h)) \rceil \tag{3}$$

where H_p consists of hits containing p, $S_B(h)$ and $Id(h)$, the bit and identity score of hit h respectively. For further details, we refer the reader to [7].