Genomic signatures for metagenomic data analysis: exploiting the reverse complementarity of tetranucleotides

Fabio Gori*, Dimitrios Mavroedis*
*Radboud University Nijmegen
iCIS
Nijmegen, The Netherlands
Email: gori@cs.ru.nl, d.mavroedis@cs.ru.nl, elenam@cs.ru.nl

Mike S.M. Jetten†
†Radboud University Nijmegen
IWWR, Department of Microbiology
Nijmegen, The Netherlands
Email: m.jetten@science.ru.nl

Elena Marchiori*

***Abstract*—Metagenomics studies microbial communities by analyzing their genomic content directly sequenced from the environment. To this aim metagenomic datasets, consisting of many short DNA or RNA fragments, are computationally analyzed using statistical and machine learning methods with the general purpose of binning or taxonomic annotation. Many of these methods act on features derived from the data through a genomic signature, where a typical genomic signature of a fragment is a vector whose entries specify the frequency with which oligonucleotides appear in that fragment. In this article we analyze experimentally the ability of existing genomic signatures to facilitate the discrimination between fragments belonging to different genomes. We also propose new genomic signatures that take into account that fragments can have been sequenced from both strands of a genome; this is achieved by exploiting the reverse complementarity of oligonucleotides. We conduct extensive experiments on *in silico* sampled genomic fragments in order to assess comparatively the effectiveness of existing genomic signatures and those proposed in this article. Results of the experiments indicate that the direct use of the reverse complementarity of tetranucleotides in the definition of a genome signatures allows to have performances comparable to the best existing signatures using less features. Therefore the proposed genomic signatures provide an alternative set of features for analyzing metagenomic data. Online Supplementary material is available at http://www.cs.ru.nl/~gori/signature_metagenomics/.**

***Keywords*-genome signature; metagenome binning; taxonomic annotation; metagenomic data analysis; metagenomics**

I. Introduction

It has been observed that there is a relative intragenomic composition invariance in many prokaryotes [1]. This observation led to the idea of "genomic signature" [2], [3], a set of composition-based features to be derived from DNA sequences, that shows few differences between DNA fragments of the same genome. The biological underlying explanation of this property of genomic signatures is still unclear. It is conjectured to be the result of more contributing factors [4], like GC content and phylogeny [5] (although the correlation between signature and phylogenetic distance appear to be weak, mainly due to the absence of divergence of oligonucleotide composition in some phylogenetically distant species [6]).

In this article we focus on the use of genomic signatures for metagenomic data analysis [3]. Metagenomics is the study of genomic content of microbial communities [7]. Essentially, a metagenomic dataset is a set of DNA fragments sequenced from the genomes of an environmental sample. The complete genomes of the community members are usually not known, and some parts of them might not be represented in the metagenomic data. As a consequence, an important step of metagenomic data analysis is to detect to which (kind of) organism each fragment belongs to. This problem is tackled, for instance, by means of clustering methods (binning) [8], by prediction models constructed using available genomes (supervised taxonomic annotation) [9], [10], and by other evidence-based approaches that match fragments with sequences in a database of reference [11], [12]. Many of these methods act on genomic signatures: for instance, some binning and taxonomic annotation methods first compute signature value of all the fragments then apply either clustering [8], [13]–[18] or classification algorithms [19], [20] to the resulting signatures values.

Typically, oligonucleotide frequencies signatures are used: among these, tetranucleotide frequencies signature is the most used [13], [14], [17], [20]; sometimes frequencies of short oligonucleotides (length up to 6) are used together [16], [19]; a few tools adopted oligonucleotides longer than 6 nucleotides [8], [15], [21]. These signatures were not specifically designed to act on metagenomic fragments, and few of them were tested on fragments of 10kb [3]. Recently, signatures for metagenomic applications were introduced [18], [22]. For instance, the binning tool MetaCluster successfully tested and implemented a signature, based on tetranucleotide frequencies ranking [18] and showed their effectiveness for clustering metagenomic fragments.

However, to the best of our knowledge, a thorough comparative analysis of genome signatures for metagenomic fragments has not yet been provided. In this paper we conduct such a study, and propose new signatures suitable for metagenomic fragments that take into account the special

requirements of metagenomic applications. More precisely, a signature for metagenomic data needs to be effective on sequences generated by contemporary DNA sequencing technologies. These sequences have a length not higher than 1,000 bp [7]; existing genomic signatures, instead, were tested on fragments of 10 kb or more [3]. Moreover, signatures for metagenomic data should be effective even with sequences belonging to different strands of the genomes, because sequencing technologies might sample fragments from both strands. Finally, a signature cannot be based on information extracted from source genome, since composition of the sequenced community is often unknown, and new species might also be present in the metagenome. These observations motivated the development of new signatures that assume the same value for a sequence and its reverse complement; these signatures are based on tetranucleotide frequencies, and exploit the reverse complementarity of tetranucleotides.

In order to test the effectiveness of the proposed signatures on metagenomic data, simulated sequenced data are constructed. These consist of DNA fragments of 1,000 bp, like data studied in recent works on binning [18] and sequences generated by the new 454 GS FLX+ System[1]. As mentioned before, this sequence length is much lower than the one used in standard signature studies [3]. We assess the capability of the considered signatures to render fragments stemming from the same genome closer (with respect to a given similarity measure) to each other than to fragments belonging to other genomes; we also evaluate how well the signature distance of two fragments represent the taxonomic relationship between the source genomes. To this end, we employ various quality measures, such as statistical univariate tests (e.g., Mann-Whitney U test) and evaluation measures employed in machine learning (e.g., ROC curve and AUC).

Results of the experiments indicate that incorporating the above mentioned ad-hoc properties of metagenomic fragments into a definition of signature allows to use less features than the standard tetranucleotide frequency signature, obtaining comparable performances. New signatures provide a valid alternative to signatures used so far in metagenomic data analysis.

II. Material and Methods

A. Genomic Signatures

In this study we focus on signatures based on tetranucleotide (4-mer) frequencies, since previous works had demonstrated that these features carry a significative genetic signature signal [23].

Let W denote the set of all the 256 tetranucleotides, represented as words of length 4 in the alphabet {A,C,G,T}; given a tetranucleotide $w \in W$, we denote by w^C the reverse complement of w. Note that 16 of these 256 words are such that $w^C = w$, that is, they coincide with their reverse complement.

Given a metagenomic fragment s, we denote with f_i and f_i^C the frequency with which tetranucleotide w_i and w_i^C occur in s, respectively.

We view a signature as a function ρ^α mapping a fragment s to a vector $(a_1, \ldots, a_n)$ of real numbers, and consider the following signatures.

Frequencies Signature ρ^T: This signature is defined by setting the i-th component a_i of $\rho^T(s)$ to f_i, for $i = 1, \ldots, 256$. This signature was used in many tools for metagenomic sequences processing [13], [14], [17], [20]; among the analyzed signatures, it is the only one that does not exploit reverse complementarity of tetranucleotides.

Symmetrized Signature ρ^S: This signature is obtained by summing the frequencies of distinct reverse complementary 4-mers (see, e.g., [6]). It is defined as $\rho^S(s) := (a_1, \ldots, a_{136})$, with $a_i = f_i + f_i^C$ if $w_i \neq w_i^C$, and $a_i = f_i$, otherwise. Notice that the vector representation is of length 136, since 16 4-mers coincide with their reverse complement and 240 do not (i.e., $w_i \neq w_i^C$).

Symmetrized Rank Signature ρ^{Rank}: This signature is defined such that $\rho^{Rank}(s)$ is the ranking induced by sorting the elements of $\rho^S(s)$. For instance, if $\rho^S(s) = (0.7, 0, 0.3)$ then $\rho^{Rank}(s) = (1, 3, 2)$. This signature was used in recent works on metagenomics binning [2] [18].

Minimal and Maximal complementarity signatures ρ^{min} *and* ρ^{max}: These signatures are defined such that $\rho^{min}(s) := (a_1, \ldots, a_{120})$, with $a_i = \min(f_i, f_i^C)$ and $w_i \neq w_i^C$, and $\rho^{max}(s) := (a_1, \ldots, a_{120})$, with $a_i = \max(f_i, f_i^C)$ and $w_i \neq w_i^C$. Notice that in this signature we employ only the 4-mers that have a non-identical reverse complement.

Reverse Identity Signature ρ^I: This signature considers only the frequencies of the 16 4-mers that are equal to their reverse complement (i.e., $w_i = w_i^C$). That is $\rho^I(s) := (a_1, \ldots, a_{16})$ with $a_i = f_i$.

Ratio Signatures ρ^{Ratio1} *and* ρ^{Ratio2}: These signature are defined for the 4-mers that have different reverse complement as $\rho^{Ratio1}(s) = (a_1, \ldots, a_{120})$ and $\rho^{Ratio2}(s) = (b_1, \ldots, b_{120})$, where

$$a_i := \begin{cases} 1, & \text{if} f_i = f_i^C = 0, \\ \min(\frac{f_i}{f_i^C}, \frac{f_i^C}{f_i}), & \text{otherwise,} \end{cases},$$

$$b_i := \begin{cases} \frac{1}{2}, & \text{if} f_i = f_i^C = 0, \\ \frac{\min(f_i, f_i^C)}{f_i + f_i^C}, & \text{otherwise,} \end{cases}$$

for $w_i \neq w_i^C$.

JS Signature ρ^{JS}: This signature is based on Jensen-

[1] See http://454.com/products/gs-flx-system/index.asp.

[2] In that work, they compare the values assumed by signature ρ^S for different fragments through Spearman footrule distance, that is equivalent to compare the values assumed by ρ^{Rank} via L_1.

Shannon divergence [24], and is defined such that $\rho^{JS}(s) = (a_1, \ldots, a_{120})$ with

$$a_i := f_i \log \frac{f_i}{\frac{1}{2}(f_i + f_i^C)} + f_i^C \log \frac{f_i^C}{\frac{1}{2}(f_i + f_i^C)}.$$

This signature is defined only for the 4-mers that have different reverse complements (i. e. $w_i \neq w_i^C$).

Similarity between signatures was computed using L_1 distance (also known as Manhattan distance). The choice of this distance is motivated by its use in previous methods for taxonomic annotation of metagenomic sequences [17]. Moreover, the distances most often used in literature are based on L_1 multiplied by an averaging factor [6], [3] , [2]. Given a genomic signature ρ^a, the related signature distance between two nucleotide sequences s, z is defined by computing the L_1 distance between $\rho^a(s)$ and $\rho^a(z)$.

We also analyzed combinations of couples and triplets of signatures, like $(\rho^{min}(s), \rho^I(s))$ and $(\rho^{max}(s), \rho^{min}(s), \rho^I(s))$. Similarly, we studied combinations of normalized signatures, like $(\rho_N^{min}(s), \rho_N^I(s))$ and $(\rho_N^{max}(s), \rho_N^{min}(s), \rho_N^I(s))$, where $\rho_N^a(s)$ is defined as $\rho^a(s)$ divided by the maximum value that can be achieved by the related signature distance between two DNA fragments of 1,000bp. The list of maximum values of the signatures is provided in online Supplementary Material (Table 2). We also analyzed a few linear combinations of normalized signatures: $(\alpha\rho_N^a(s), \beta\rho_N^b(s))$ and $(\alpha\rho_N^a(s), \beta\rho_N^b(s), \gamma\rho_N^c(s))$. As for the individual signatures, distances between combination of signatures were evaluated with L_1 norm.

B. Data acquisition and preprocessing

Complete genomes of 1,284 prokaryotes were downloaded from the NCBI ftp server[3]. The complete list of the genomes is provided in online Supplementary Material (Table S1). From each genome, a set of sequences was randomly sampled. Each of these sets consists of 10,000 possibly overlapping sequences, each of them having length equal to 1,000bp. The NCBI taxonomy[4] [25] was used as reference taxonomy of the analyzed prokaryotes.

C. Comparing signature values

Following the methodology employed in related works [18], we evaluated the quality of a signature based on its property to assume similar values for fragments of the same genome, and different values for fragments of different ones. Therefore, as a first step, we generated sets of fragments and evaluated the dissimilarity of the signature values on pairs of these fragments. Specifically, we created 9 sets of fragment pairs, where each set corresponds to a different degree of diversity of the source genomes. Subsequently, signature distances between fragments for each pair of the sets were computed. From the resulting distance values, 9 distributions of mean distances were obtained for each signature. A first distribution was generated using the distances between fragments of a same genome (*intra-genomic* signature distances): for each genome, we computed all the pairwise signature distances between the 10,000 fragments of that genome, and the mean of the resulting set of distances was considered. The collection of these means across the considered genomes provided a distribution of *intra-genomic* (mean) distances for a given signature.

The other 8 distributions of distances were generated by computing distances between fragments from different genomes (*inter-genomic* signature distances), where each of the 8 distributions was obtained by considering a different level of taxonomic divergence of the compared genomes. Specifically, we created 7 sets of organisms' pairs, for each of the following taxonomic ranks: Species, Genus, Family, Order, Class, Phylum, Superkingdom. The set of organisms associated to rank r consisted of 1,000 different pairs of organisms randomly selected among those whose lowest common ancestor in the taxonomy tree is at rank r. For each pair of these organisms, we randomly selected 10,000 pairs of genomic fragments from the set of all fragments sampled from these genomes, and calculated the mean of the resulting distances. The collection of the mean distances over all these pairs of organisms provided a distribution of inter-genomic distances at rank r.

Furthermore, we also created a set of organisms' pairs where each element is made by a *bacterium* and an *archaeon*, that are the two superkingdom of the *Prokaryotes*. We computed a signature distance distribution for this set of organisms' pairs as described above. We refer to this distribution as the inter-genomic signature distances distribution at prokaryotes level.

D. Evaluating the effectiveness of signatures

We assessed the capability of a genomic signature to facilitate the discrimination between fragments sampled from the same genome and fragments sampled from different ones. More specifically, for each genomic signature, we tested if the related signature distance yielded small values for fragment pairs of the same genome, and greater values for fragments of different ones. To this aim, the signature distance was considered as a score for the fragment pair; the score quantifies the degree to which the two fragments of the pair belong to different genomes. Higher scores correspond to fragments that are more likely to belong to different genomes, according to the related signature.

Signature performances were tested and compared plotting the associated Receiver Operating Characteristic (ROC) curves [26]. A first class of fragment pairs, called "negatives", was made by the pairs sampled from the same genome; the remaining pairs, called "positives", formed a second class. Signature distance was considered to be effec-

[3]ftp://ftp.ncbi.nih.gov/genomes/Bacteria/, downloaded on March 2011.

[4]Available at ftp://ftp.ncbi.nih.gov/pub/taxonomy/

Table I
COMPARISON OF GENOMIC SIGNATURES AND THEIR COMBINATIONS, WITH RESPECT TO AUC, POSITIVE PAIRWISE MANN-WHITNEY U TEST, MEAN PAIRWISE AUC, NUMBER OF FEATURES

Signature[1]	AUC	U tests	Pairwise AUC	Feat.
$(\rho_N^S, \rho_N^{min}, \rho_N^{Ratio1})$	0.919	35	0.810	376
$(\rho_N^{max}, \rho_N^{min}, \rho_N^{Ratio1})$	0.913	35	0.808	256
(ρ_N^S, ρ_N^I)	0.913	35	0.807	152
(ρ^{max}, ρ^I)	0.909	35	0.807	136
$(\rho_N^{Rank}, 2\rho_N^{Ratio1})$	0.886	34	0.789	256
ρ^S	0.912	35	0.807	136
ρ^{max}	0.900	35	0.802	120
ρ^T	0.884	34	0.789	256
ρ^{min}	0.881	34	0.783	120
ρ^I	0.851	33	0.770	16
ρ^{Rank}	0.794	33	0.723	136
ρ^{Ratio1}	0.707	31	0.683	120
ρ^{Ratio2}	0.686	31	0.668	120
ρ^{JS}	0.573	26	0.562	120

[1] ρ_N^a denotes the normalized ρ^a signature (Section II-A).

tive if it yields small values for the negative pairs and high values for the positive pairs. In our data, the negatives were represented by the intra-genomic (mean) distance distribution; the union of all the 8 inter-genomic (mean) distance distributions represented the positives. Given a threshold signature distance, we considered as "true negatives" the negative pairs whose distance was below or equal to the threshold; similarly, "true positives" were made by the positives with distance above the threshold. Therefore, for each threshold we could compute the specificity and sensitivity and assign a point in the ROC space. Varying the threshold among the values of our distributions, we produced the ROC curve for the associated signature. We used the ROC curve because it clearly shows if a signature ρ^α is always better than signature ρ^β, namely if the curve of ρ^α is always above the one of ρ^β. Another index of discriminative power is the Area Under the ROC Curve (AUC) [26].

In order to evaluate whether the distribution related to the most taxonomically divergent genome pairs is significantly shifted toward higher values, we used the Mann-Whitney U test [27]. Specifically, we employed this test for evaluating whether inter-genomic distance distribution at rank r_2 was shifted toward higher values than the distribution at rank r_1, where r_2 is higher than r_1. The Mann-Whitney U test was applied to compare each of the 36 pairs of distance distributions (that is all possible pairs for the 8 inter-genomic and the intra-genomic class, i.e., $\binom{9}{2}$), for each signature. A significance level of 5% was chosen. We also checked if the genomic signatures were able to facilitate the discrimination between two pairs of fragments having different taxonomic dissimilarity (i.e., the distance measure should correlate with the taxonomic divergence of the genomes). This was done by computing a ROC curve as previously described, for each of the 36 pairs of distance distributions.

III. RESULTS

A. Intra-inter genomic discrimination

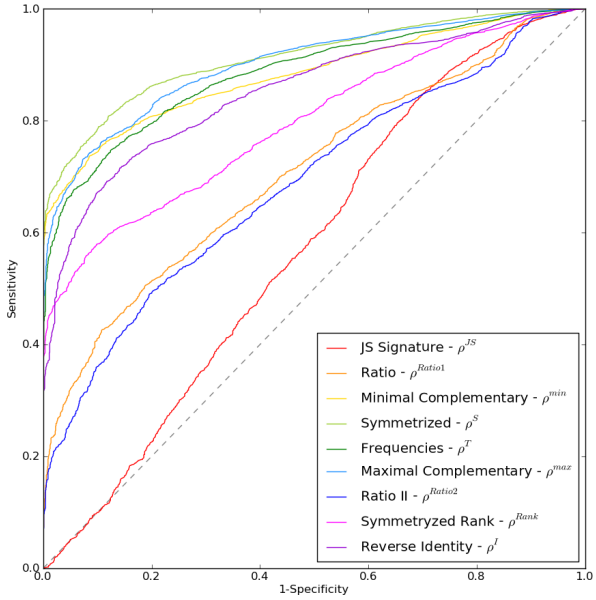
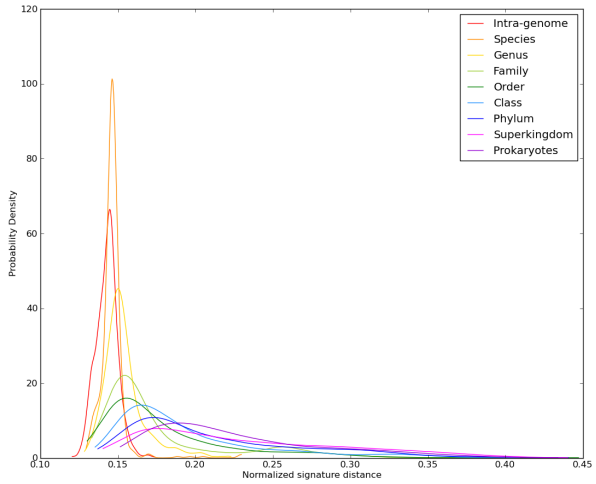


Figure 1. Receiver Operating Characteristic (ROC) curves for different genomic signatures. Sensitivity is the number of true positives divided by the number of positives; specificity is the number of true negatives divided by the number of negatives.

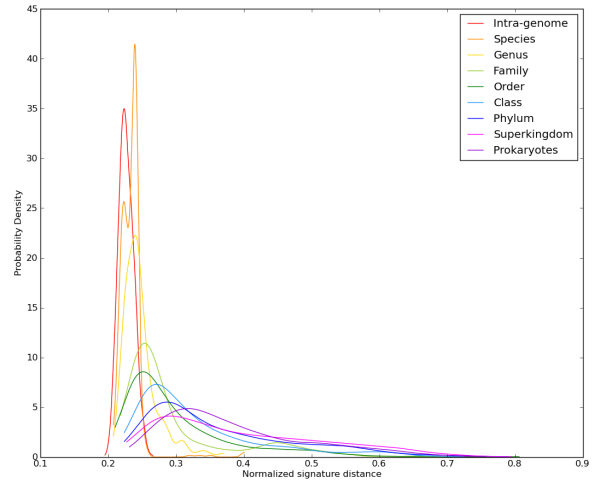
In this section we evaluate the ability of a signature to facilitate the discrimination between fragments belonging to different genomes. The best results were achieved by signatures ρ^S, ρ^{max}, ρ^T, ρ^{min} and ρ^I (Figure1). Indeed, the ROC curve of ρ^{min}, ρ^S, ρ^{max} and ρ^T were always above the remaining curves; their AUCs were all greater or equal than 0.88 (Table I). Signatures ρ^{max}, had AUC comparable to the symmetrized signature ρ^S, although it had 120 features instead of 136 (Table I). Similarly, ρ^{min} AUC was almost identical to the one of ρ^T, but these signatures had 120 and 256 features, respectively. Despite ρ^I had only 16 features, its ROC curve was higher than the remaining four curves except for low values of specificity; its AUC was 0.85.

B. Relation between signature distances and taxonomic diversity

In this section we evaluate whether the signature distances correlate with the taxonomic divergence of the genomes (i.e., higher distances are assigned to genomes that have higher taxonomic divergence). To this aim, we employed the Mann-Whitney U test for evaluating whether the difference between signature distance distributions sampled at different level of taxonomic divergence were statistically significant at 5% confidence level (see Section II-D). Results



(a) Maximal Complementarity Signature ρ^{max}



(b) Symmetrized Signature ρ^{S}

Figure 2. Probability density function of genomic signature distances, for different levels of taxonomic diversity. The functions were generated through the gaussian kernel density estimator included in SciPy library [28].

demonstrate that the signature distances tended to have a positive correlation with taxonomic diversity. This correlation is also shown by the estimated probability density functions (Figure2, Supplementary Figures 1-9). The third column of Table I shows the number of distance distribution pairs for which the Mann-Whitney U test was significant. For six signatures, the test was significant for at least 33 of the 36 distribution pairs (Table I). According to Mann-Whitney U test, no signature was able to distinguish the distance distributions at levels Superkingdom and Prokaryotes.

Nevertheless, all the signatures had difficulties in distinguishing between different levels of taxonomic diversity. This distinguishing capacity was evaluated through the mean of AUCs computed for each ROC curve of the 36 pairs of distance distributions (Table I, fourth column). For each signature, this mean AUC did not exceed 0.81. The best results were obtained for signatures whose features were tetranucleotide frequencies; their means AUC were above 0.77.

C. Combinations of signatures

The studied combinations of signatures could not achieve results significantly better than the individual signatures (Table I, Supplementary Tables 3-5). For example, the AUC of the best combination exceeded the AUC of the studied signatures, but not significantly (Table I). Similar remarks can be done for the other evaluation tests.

However, a few combinations of signatures had better performances than the genomic signatures that compose them. For instance, the combination of normalized ρ^{Rank} and normalized ρ^{Ratio1} had results significantly better than the two signatures taken separately; its results are comparable to best analyzed signatures (Table I).

IV. Conclusion

In this work, we compared the performances of new and already used genomic signatures for metagenomic data analysis. The signatures were tested with respect to their capacity to facilitate the discrimination between DNA fragments sampled from different prokaryotes genomes. The relation between signature distance and taxonomic diversity was also analyzed. Signature distances were evaluated for fragments of 1,000 bp randomly sampled from 1,284 genomes.

According to our experiments, the reverse complementarity can be successfully exploited for our scopes, as shown by the results for symmetrized signature and the newly introduced minimal and maximal reverse complementarity signatures. These two signatures, in particular, provide a valid alternative to symmetrized signature and frequencies signature, because they achieved comparable performances using less features. The appreciable results for reverse identity signature, that is made by just 16 tetranucleotide frequencies, shows that not all the tetranucleotides are necessary in defining a good signature for metagenomics; indeed, reduction of features space is exploited by some existing binning methods [16]. Results on combinations of signatures indicated that different signatures can carry complementary informations, like symmetrized rank signature and ratio signature, and can marginally improve the quality of results.

All the signature distances seemed to be related to taxonomic diversity; nevertheless this relation was not strong enough to estimate the taxonomic diversity from signature distance. This observation is consistent with previous studies on genomic signatures and phylogenetic distance [6], [29].

In future work we plan to compare the signatures performances on shorter sequences, resembling data produced by other contemporary sequencing technologies like Illumina and other 454 machines[5]. Results of preliminary experiments on sequences of 150 and 500 bp indicate that good performances are achieved by symmetrized and maximal reverse complementary signatures. We also aim to study signatures performances for taxonomic discrimination at different ranks and to compare signatures variations taxon-by-taxon.

REFERENCES

[1] S. Karlin and C. Burge, "Dinucleotide relative abundance extremes: a genomic signature." *Trends in Genetics*, vol. 11, no. 7, pp. 283–290, Jul 1995.

[2] S. Karlin *et al.*, "Comparative DNA analysis across diverse genomes," *Annual review of genetics*, vol. 32, no. 1, pp. 185–225, 1998.

[3] J. Bohlin, "Genomic signatures in microbes – properties and applications," *TheScientificWorldJournal*, vol. 11, pp. 715–725, 2011.

[4] M. W. van Passel *et al.*, "The reach of the genome signature in prokaryotes." *BMC Evolutionary Biology*, vol. 6, pp. 84+, Oct. 2006.

[5] J. Bohlin *et al.*, "Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering," *BMC Genomics*, vol. 10, no. 1, pp. 487+, 2009.

[6] J. Mrázek, "Phylogenetic signals in DNA composition: Limitations and prospects," *Molecular Biology and Evolution*, vol. 26, no. 5, pp. 1163–1169, 2009.

[7] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics." *PLoS Comput Biol*, vol. 6, no. 2, p. e1000667, Feb 2010.

[8] Y.-W. W. Wu *et al.*, "A novel abundance-based algorithm for binning metagenomic sequences using l-tuples," *Journal of computational biology : a journal of computational molecular cell biology*, vol. 18, no. 3, pp. 523–534, Mar. 2011.

[9] A. Brady *et al.*, "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models," *Nature Methods*, vol. 6, no. 9, pp. 673–676, August 2009.

[10] D. Kelley *et al.*, "Clustering metagenomic sequences with interpolated Markov models," *BMC Bioinformatics*, vol. 11, no. 1, pp. 544+, Nov. 2010.

[11] D. H. Huson *et al.*, "Megan analysis of metagenomic data," *Genome Research*, vol. 17, no. 3, pp. 377–386, March 2007.

[12] L. Krause *et al.*, "Phylogenetic classification of short environmental DNA fragments," *Nucleic Acids Research*, vol. 36, no. 7, pp. 2230–2239, 2008.

[13] H. Teeling *et al.*, "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environmental Microbiology*, vol. 6, no. 9, pp. 938–947, 2004.

[5]See http://www.illumina.com/ and http://454.com/.

[14] C.-K. K. K. Chan *et al.*, "Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing," *Journal of Biomedicine and Biotechnology*, vol. 2008, no. 1, 2008.

[15] S. Chatterji *et al.*, "CompostBin: A DNA Composition-Based Algorithm for Binning Environmental Shotgun Reads," in *RECOMB 2008*, ser. LNCS, 2008, vol. 4955, ch. 3, pp. 17–28.

[16] A. Kislyuk *et al.*, "Unsupervised statistical clustering of environmental shotgun sequences," *BMC Bioinformatics*, vol. 10, no. 1, pp. 316+, 2009.

[17] M. H. Mohammed *et al.*, "SPHINX–an algorithm for taxonomic binning of metagenomic sequences." *Bioinformatics*, vol. 27, no. 1, pp. 22–30, Jan 2011.

[18] B. Yang *et al.*, "Metacluster: unsupervised binning of environmental genomic fragments and taxonomic annotation," in *BCB*. ACM, 2010, pp. 170–179.

[19] K. R. Patil *et al.*, "Taxonomic metagenome sequence assignment with structured output models," *Nature Methods*, vol. 8, no. 3, pp. 191–192, Mar. 2011.

[20] N. N. Diaz *et al.*, "TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach," *BMC Bioinformatics*, vol. 10, no. 1, pp. 56+, 2009.

[21] O. Nalbantoglu *et al.*, "RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles," *BMC Bioinformatics*, vol. 12, no. 1, pp. 41+, 2011.

[22] I. Saeed *et al.*, "The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments." *BMC Genomics*, vol. 10 Suppl 3, p. S10, 2009.

[23] D. T. Pride *et al.*, "Evolutionary implications of microbial genome tetranucleotide frequency biases," *Genome Research*, vol. 13, no. 2, pp. 145–158, 2003.

[24] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[25] E. W. Sayers *et al.*, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D5–D15, 2009.

[26] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[27] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods, 2nd Edition*, 2nd ed. Wiley-Interscience, Jan. 1999.

[28] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–. [Online]. Available: http://www.scipy.org/

[29] S. C. Perry and R. G. Beiko, "Distinguishing microbial genome fragments based on their composition: evolutionary and comparative genomic perspectives." *Genome biology and evolution*, vol. 2, no. 0, pp. 117–131, Jan. 2010.