The importance of Chargaff's second parity rule for genomic signatures in metagenomics

Fabio Gori[1,*], Dimitrios Mavroeidis[1], Mike SM Jetten[2,3], Elena Marchiori[1]
1 Radboud University Nijmegen, Institute for Computing and Information Science, Nijmegen, The Netherlands
2 Radboud University Nijmegen, Institute for Water and Wetland Research, Nijmegen, The Netherlands
3 Delft University of Technology, Department Biotechnology, Delft, The Netherlands
∗ **E-mail: fabio.gori.phd@gmail.com**

Abstract

An important problem in metagenomic data analysis is to identify the source organism, or at least taxon, of each sequence. Most methods tackle this problem in two steps by using an alignment-free approach: first the DNA sequences are represented as points of a real n-dimensional space via a mapping function then either clustering or classification algorithms are applied. Those mapping functions require to be genomic signatures: the dissimilarity between the mapped points must reflect the degree of phylogenetic similarity of the source species. Designing good signatures for metagenomics can be challenging due to the special characteristics of metagenomic sequences; most of the existing signatures were not designed accordingly and they were tested only on error-free sequences sampled from a few dozens of species.

In this work we analyze comparatively the goodness of existing and novel signatures based on tetranucleotide frequencies via statistical models and computational experiments; we also study how they are affected by the generalized Chargaff's second parity rule (GCSPR), which states that in a given sequence longer than 50kbp, inverse oligonucleotides are approximately equally frequent. We analyze 38 million sequences of 150 bp-1,000 bp with 1% base-calling error, sampled from 1,284 microbes. Our models indicate that GCSPR reduces strand-dependence of signatures, that is, their values are less affected by the source strand; GCSPR is further exploited by some signatures to reduce the intra-species dispersion. Two novel signatures stand out both in the models and in the experiments: the combination signature and the operation signature. The former achieves strand-independence without grouping oligonucleotides; this could be valuable for alignment-free sequence comparison methods when distinguishing inverse oligonucleotides matters. Operation signature sums the frequencies of reverse, complement, and inverse tetranucleotides; having 72 features it reduces the computational intensity of the analysis.

1 Introduction

Metagenomics studies the genomic content of microbial communities, obtained through DNA sequencing technologies [1]. Essentially, a metagenomic dataset is a set of DNA sequences acquired from the genomes of an environmental sample. By bypassing the cultivation step, metagenomics is able to obtain microbial genomes unattainable through individual sequencing, since less than 1% of the microbes present in nature can be cultured [2]. Moreover, with metagenomics it is possible to infer the interactions occurring in a microbial community. Unfortunately, the potentialities given by metagenomics come with a price in terms of data analysis challenges: we do not know from which genome a sequence was sampled; in most of the cases, the full genomes of the community members are not available; even species number is unknown.

As a consequence, an important step of metagenomic data analysis is to detect to which organism, or at least to which taxon each sequence belongs to. This problem is tackled, for instance, by means of clustering methods (binning) [3], by prediction models constructed using available genomes (taxonomic assignment) [4,5], and by other similarity-based approaches that match sampled sequences with sequences in a database of reference [6, 7].

Many of the binning and taxonomic assignment methods tackle the problem in two steps: first the DNA sequences are represented as points of a real n-dimensional space via a *mapping function*, then either clustering [3, 8–13] or classification algorithms [14, 15] are applied to these points. Typically, the mapping functions adopted in literature represent a given sequence with the frequency counts of a set of oligonucleotides: among these, tetranucleotides are the most used [8, 9, 12, 15]; sometimes frequencies of short oligonucleotides (length up to 6) are used together [11, 14]; a few tools adopted oligonucleotides longer than 6 bases [3, 10, 16, 17]. More recent tools are based on tetranucleotides frequencies computed on contigs assembled from the reads of the metagenome [18, 19].

In order to be effective for binning and taxonomic assignment, these mapping functions have to be *genomic signatures* [20]. A genomic signature is a mapping function that has the following properties: sequences sampled from the same genome are mapped to relatively similar points; sequences sampled from different genomes are mapped to significantly different points, and this difference is related to the phylogenetic distance of the source genomes. Signatures are also used for alignment-free sequence comparison [21]. The biological underlying explanation of this property of existing genomic signatures is still unclear. It is conjectured to be the result of more contributing factors [22], like GC content and phylogeny [23]; however the correlation between signature and phylogenetic distance appears to be not very strong, mainly due to the absence of divergence of oligonucleotide composition in some phylogenetically distant species [24].

Despite the relevance of genomic signatures for metagenomic data analysis, existing signatures were not designed to take into account the special properties of metagenomic data. Signatures are usually tested on sequences of 10,000 base pairs (bp) or more [20], while the sequencing technologies used for metagenomics generate sequences of 50-1,000 bp. Signatures for metagenomic data cannot be based on information extracted from source genome of a sequence, like many existing signatures do, since composition of the sequenced community is often unknown, and new species might also be present in the metagenome. Signatures need also to be effective with sequences containing errors, that can be generated by sequencing machines. Moreover, signatures for metagenomic data have to be effective even with sequences belonging to different strands of the genomes, because sequencing technologies might sample sequences from both strands.

Furthermore, the development of binning and taxonomic assignment methods was more focused on the algorithmic part rather than on the adopted mapping function/genomic signature. Attempts to introduce signatures for metagenomic applications are recent and validated the signatures on error-free sequences, sampled from a few dozens of species, and longer than real metagenomic sequences. For instance, the binning tool MetaCluster successfully tested and implemented a signature based on tetranucleotide frequencies ranking [13, 25], and showed its effectiveness for clustering metagenomic sequences. RAIphy used a signature computable on sets of sequences [16]; MetaProb computed normalized tetranucleotide frequencies on sequence sets [26]. Signature OFDEG was designed for error-free sequences of at least 8,000 bp [27]. Recently, ICOs signature was tested on 50,000 error-free sequences of 1,000 bp or more, sampled from 60 species [28].

In this study, we analyze and compare theoretically the statistical distributions of some existing and novel signatures; in particular, for each signature we study its dispersion among sequences of the same species and the dispersion of signature's expected values among different species. A low within-species dispersion shows that the signature assumes similar values for sequences sampled from the same species. A signature's expected value for a species gives us an indication of its average values for that species; high dispersion of species expected values indicates that the given signature maps sequences of different species to significantly different points.

Furthermore, we investigate the biological and statistical rationales that make signatures effective in dealing with metagenomics data. In particular, we take into account the effects of the generalized Chargaff's second parity rule (GCSPR) on genomic signatures. This rule states that, in a given sequence of at least 50 kbp, an oligonucleotide and its reverse complement are approximately equally frequent [29,30];

the rule implies that the two oligonucleotides have approximately the same total count in a genome [31]. GCSPR rule applies to all the double stranded DNAs, organelles excepted [32].

We also provide a thorough comparative experimental analysis of those signatures on data more similar to the metagenomic ones than the ones used in signature studies [20, 27, 28], especially in terms of number of species, sequence length, and presence of errors. Our experimental setup is a significant extension of the one adopted in our previous work [33]: sequences have base-calling errors (1% rate); experiments are performed for multiple sequence lengths (150 bp, 500 bp, 1000 bp); we directly compare distributions of signatures' distances instead of their mean values; the Area Under Precision-Recall curve (AUPR) replaces the Area Under ROC curve as an evaluation measure, to take into account that within-species sequence pairs are much fewer than between-species ones; the number of sequences pairs on which within-species signatures' distances are computed has been increased hundred-fold.

The proposed signatures take into account the special properties of metagenomic data mentioned before. In particular, they are designed to be effective with sequences of the same genome sampled from different strands: these *strand-independent* signatures assume the same value for a sequence and its reverse complement, so that the source strand of the sequence become irrelevant for the signature. All the signatures we study are derived from the tetranucleotide frequencies signature; we will refer to this as the *standard signature*.

The most important novel signatures introduced in this study are the *combination signature* and the *operation signature*. Combination signature is a reordering of the features of the standard tetranucleotide frequencies signature; the adopted reordering makes combination signature strand-independent. Strand-independent signatures currently used in metagenomics [18, 26] are derived form the *symmetrized signature*, that is obtained by summing the frequencies of reverse complementary tetranucleotides [24]. Therefore, combination signature proves that summing frequencies of inverse oligonucleotides, and thus loosing information, is not necessary to achieve a strand-independent signature.

The 72-feature Operation signature is instead obtained by summing the frequency of a tetranucleotide with the ones of its complement, reverse, and inverse. This was developed to study if exploiting GCSPR and the reverse, complementarity, and inverse relations between tetranucleotides [34] can reduce feature space dimension and increase error tolerance. In data analysis, in general, reduction of feature space dimension can be beneficial in many aspects: performances are less dependent from the data (see bias-variance trade-off [35]), the computational cost of the analysis is reduced, and the data can become more interpretable [36]. Dependency from data is a very important issue for metagenomic data analysis, because metagenomes are likely to contain novel microbes and communities: hence, an analysis method based on fewer features can be more accurate when facing these types of data. The reduction of computational cost is also relevant, because the size of metagenomes is growing so much that analysing them is already becoming expensive in terms of computational power and data storage [37].

We also define and test new signatures capturing the divergence from GCSPR. Since metagenomic sequences are much shorter than 50 kbp [29, 30] the hypotheses behind GCSPR are not completely satisfied and hence the frequencies of reverse complementary oligonucleotides may differ. These differences could change according to the taxonomic classification of the source genome, making them exploitable for our purpose: indeed, previous research showed that purine-pyrimidine asymmetry in mammalian mitochondrial DNA carries phylogenetic information [38].

2 Materials and Methods

2.1 Genomic Signatures and their Rationale

In this study we focus on signatures based on tetranucleotide (4-mer) frequencies, since previous works had demonstrated that these features carry a significant phylogenetic signal [39]. Let w denote one of the 256 tetranucleotides, represented as words of length 4 in the alphabet {A,C,G,T}; we denote by

w^{RC} the reverse complement (also called *inverse*) of w. Note that 16 of these 256 tetranucleotides are *palindromic*, i.e, they coincide with their respective reverse complements ($w^{RC} = w$). A metagenomic sequence s is also represented as a word in the alphabet {A,C,G,T} but with no length limit; s^{RC} still denotes the inverse of s. For a given sequence s, we denote with f_i and f_i^{RC} the frequencies with which tetranucleotides w_i and w_i^{RC} occur in s, respectively. We also denote with w^C and w^R the complement and reverse of w, respectively. Note that w^R and w^C are reverse complement of each other. Therefore, for the 8 distinct tetranucleotides coinciding with their reverse ($w = w^R$), the relation $w^{RC} = w^C$ holds; if w is palindromic, then w^C is another distinct palindromic ($w^C = w^R$). The symbols f_i^C and f_i^R denote the frequencies of w_i^C and w_i^R in a given sequence s, respectively.

We view a signature as a function ρ^α mapping a sequence s to a vector $\rho^\alpha(s) = (a_1, \ldots, a_n)$ of real numbers. A signature ρ^α is called *strand-independent* if it assumes the same value for each possible genomic sequence s and the sequence sampled from exactly the same genomic region but on the complementary strand, that is the inverse of s: that is, if the relation $\rho^\alpha(s) = \rho^\alpha(s^{RC})$ holds for all possible sequences s.

First, we introduce the signature usually adopted in metagenomic data analysis:
Standard Frequencies Signature ρ^{T}: This signature is defined by setting the i-th component a_i of $\rho^{\mathrm{T}}(s)$ to f_i, for $i = 1, \ldots, 256$. This signature has been used in many tools for metagenomic data analysis [8, 9, 12, 15]; among the analyzed signatures, it is the only one that is affected by the source strand of the sequence and does not exploit reverse complementarity of tetranucleotides. It is also affected by the deviation from GCSPR.

All the other signatures under examination in this work are strand-independent. The first group of signatures we examined are novel signatures that can be derived from ρ^{T} by simply reordering its features or selecting only some of them:
Minimal and Maximal complementarity signatures $\rho^{\min}$ *and* $\rho^{\max}$: These signatures are defined such that $\rho^{\min}(s) := (a_1, \ldots, a_{120})$, with $a_i = \min(f_i, f_i^{RC})$ and $w_i \neq w_i^{RC}$, and $\rho^{\max}(s) := (a_1, \ldots, a_{120})$, with $a_i = \max(f_i, f_i^{RC})$ and $w_i \neq w_i^{RC}$. These signatures are affected by the deviation from GCSPR. Notice that in these signatures we employ only the 240 non-palindromic tetranucleotides.
Palindromic Signature ρ^{P}: This signature considers only the frequencies of the 16 palindromic tetranucleotides (i.e., $w_i = w_i^{RC}$). That is $\rho^{\mathrm{P}}(s) := (a_1, \ldots, a_{16})$ with $a_i = f_i$. The introduction of this signature is motivated by a study where the frequency distribution of palindromic tetranucleotides was shown to exhibit highest inter-species but low intra-species variance on 10,000 bp sequences [40].
Combination Signature $(\rho^{\max}, \rho^{\min}, \rho^{\mathrm{P}})$: As a combination of $\rho^{\max}$, $\rho^{\min}$ and ρ^{P}, it maps a sequence s to a vector $(a_1, \ldots, a_{120}, b_1, \ldots \ \ldots, b_{120}, c_1, \ldots, c_{16})$, where the features $a_i = \max(f_i, f_i^{RC})$ and $b_i = \min(f_i, f_i^{RC})$ are derived from the non-palindromic 4-mers ($w_i \neq w_i^{RC}$), while $c_i = f_i$ is computed for the 16 palindromic 4-mers. This signatures is actually a reordering of the 256 tetranucleotide frequencies composing ρ^{T}, in a way that makes the signature strand-independent and is still affected by the deviation from GCSPR. As a matter of fact, this combination maps a given sequence s to a permutation of the 256 components of $\rho^{\mathrm{T}}(s)$. Indeed, following the notation previously introduced, it can be proved that each feature of ρ^{T} corresponds to the frequency of a certain tetranucleotide w. If $w = w^{RC}$ holds, than its frequency will be a feature of ρ^{P}; if the equality does not hold, frequency of w will be a feature either of $\rho^{\max}$ or $\rho^{\min}$.

2.1.1 Signatures exploiting generalized Chargaff's second parity rule

Other signatures we examined reduce the number of features and increase error tolerance by exploiting GCSPR and other genomic symmetries; those signatures include a novel one called operation signature:
Symmetrized Signature ρ^{S}: This signature is obtained by summing the frequencies of distinct inverse 4-mers (see, e.g., [24],). It is defined as $\rho^{\mathrm{S}}(s) := (a_1, \ldots, a_{136})$, with $a_i = f_i + f_i^{RC}$ if $w_i \neq w_i^{RC}$, and $a_i = f_i$, otherwise. Notice that the vector $\rho^{\mathrm{S}}(s)$ has 136 features, since 16 tetranucleotides are palin-

dromic, i.e, they coincide with their inverse, and 240 are not (i.e., $w_i \neq w_i^{RC}$).
The symmetrized signature ρ^{S} can be seen as a simplification of the combination signature $(\rho^{\max}, \rho^{\min}, \rho^{\mathrm{P}})$, and hence ρ^{T}, because ρ_i^{S} is equal to $\rho_i^{\max} + \rho_i^{\min}$ if $w_i \neq w_i^{RC}$ holds, otherwise $\rho_i^{\mathrm{S}} = \rho_i^{\mathrm{P}}$. Indeed, a way to reduce the dimension of combination signature is to substitute some of its features with new features corresponding to their sum; the side effect of this approach is that it removes the distinction between the frequencies of the chosen tetranucleotides.
This signature exploits GCSPR to reduce feature space dimension. GCSPR states that, on sequences of 50 kbp or longer, w_i and w_i^{RC} have approximately the same frequency, i.e. the relation $f_i \approx f_i^{RC}$ holds. Therefore, if sequences were sampled all from the same strand, we would just need to choose one of the two frequencies as features to build effective signatures. Since this is not possible, it is sensible to replace those two frequencies with their sum, that corresponds also to the sum of $\max(f_i, f_i^{RC})$ and $\min(f_i, f_i^{RC})$. Indeed, thanks to the previous relation, we have that $\max(f_i, f_i^{RC}) + \min(f_i, f_i^{RC}) = f_i + f_i^{RC} \approx 2f_i \approx 2f_i^{RC}$ also hold; hence it is sensible to replace each of these 120 feature pairs with their sum $\max(f_i, f_i^{RC}) + \min(f_i, f_i^{RC}) = f_i + f_i^{RC}$, thus reducing the signature to 136 features. As a result, we obtain ρ^{S}.

Symmetrized Rank Signature ρ^{Rank}: This signature is defined such that $\rho^{\mathrm{Rank}}(s)$ is the ranking induced by sorting the elements of $\rho^{\mathrm{S}}(s)$ in descending order. This signature was used in recent works on metagenomic binning [1] [13,25]; however, it was not specified how $\rho^{\mathrm{S}}(s)$ ranking is performed when some $\rho^{\mathrm{S}}(s)$ elements have the same value. We decided to perform a second ranking between features having the same values according to the alphabetical order of the respective tetranucleotides. For example, if the frequency of the palindromic 4-mer 'ACGT' is equal to the sum of the frequencies of the reverse complementary pair 'AAAA' and 'TTTT', then the ρ^{Rank} value corresponding to the pair will be lower than the one of the single sequence, because 'AAAA' precedes 'ACGT' in the alphabetic order. Our choice of this second ranking was motivated by the simplicity of its implementation and computation.
Operation Signature ρ^{O}: This signature is obtained by summing the frequency of a tetranucleotide with the ones of its complement, reverse, and inverse. It is inspired by a publication where the set of oligonucleotides is partitioned in equivalence classes with respect to complement and reverse operations [34]. It is defined as $\rho^{\mathrm{O}}(s) := (a_1, \ldots, a_{72})$, with $a_i = f_i + f_i^C$ if $w_i = w_i^{RC}$ or $w_i = w_i^R$, and $a_i = f_i + f_i^C + f_i^R + f_i^{RC}$ otherwise. Notice that the vector $\rho^{\mathrm{O}}(s)$ has 72 features: 8 features are given by the 8 sets $\{w_i, w_i^C\}$ for which $w_i = w_i^{RC}$ (and $w_i^C = w_i^R$); other 8 features are associated ot sets of the same form where $w_i = w_i^R$ (and $w_i^C = w_i^{RC}$); the remaining 224 tetranucleotides are all distinct, from their reverse, complement, and inverse, thus leading to 56 features.

As ρ^{S} is a reduction of ρ^{T}, signature ρ^{O} can be seen as an additional simplification of ρ^{S}, derived by exploiting complement and reverse relation between the tetranucleotides to further reduce feature space dimension. Indeed, ρ^{S} can be additionally reduced by substituting some of its features with new features corresponding to their sum, as we did before. Given a tetranucleotide w, we can observe that its complement w^C and the reverse w^R are one the reverse complement of the other. Therefore we can replace the 56 ρ^{S} feature pairs $(f_i + f_i^{RC}, f_i^C + f_i^R)$ with their sum when the four tetranucleotides are distinct. 8 features $f_i + f_i^C$ are obtained from the 16 palindromic tetranucleotides $w = w^{RC}$ and $w^C = w^R$ (e.g. w ='ACGT', w^C ='TGCA'). The remaining 8 features of ρ^{S} corresponding to the w coinciding with their reverse (and hence $w^{RC} = w^C$) are not changed (e.g. w ='ACCA', w^{RC} ='TGGT').

2.1.2 Signatures capturing deviation from generalized Chargaff's second parity rule

The following novel signatures were designed to capture solely the deviation from GCSPR in the given sequence. The deviation is computed with respect to tetranucleotide frequencies; their features are derived from the non-palindromic tetranucleotides:

[1] In those works, distance between sequences was measured through Spearman footrule distance, that is equivalent to compare the values assumed by ρ^{Rank} via L_1.

Asymmetry Signature ρ^{A}: This signature is defined as $\rho^{\mathrm{A}}(s) := (a_1, \ldots, a_{120})$, where $a_i = |f_i - f_i^{RC}|$. This is the only signature that measures the deviation with respect to sequence length, because of the way f_i is defined.

Skew Signature ρ^{Skew}: This signature is based on the standard relative skew index usually adopted in literature, such as [41]. It is defined as $\rho^{\mathrm{Skew}}(s) := (a_1, \ldots, a_{120})$, where

$$a_i = \begin{cases} 0, & \text{if } f_i = f_i^{RC} = 0, \\ \frac{|f_i - f_i^{RC}|}{f_i + f_i^{RC}}, & \text{otherwise.} \end{cases}$$

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

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

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

for $w_i \neq w_i^{RC}$.

JS Signature ρ^{JS}: This signature is based on Jensen-Shannon divergence [42], and is defined as $\rho^{\mathrm{JS}}(s) = (a_1, \ldots, a_{120})$ with

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

This signature is based only on the non-palindromic tetranucleotides.

Similarity between signatures was computed by using L_1 distance (also known as Manhattan distance). The choice of this distance is motivated by its use in previous methods for taxonomic assignment of metagenomic sequences [12, 19]. Moreover, the distances most often used in literature on genomic signatures are based on L_1 multiplied by an averaging factor [20, 24, 43]. 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 a few combinations of pairs of signatures, such as $(\rho^{\mathrm{S}}(s), \rho^{\mathrm{A}}(s))$, $(\rho^{\min}(s), \rho^{\max}(s))$, and the remaining combinations of $\rho^{\min}$, $\rho^{\max}$ and ρ^{P}. Similarly, we studied the combinations $(\rho_N^{\mathrm{S}}(s), \rho_N^{\mathrm{Rank}}(s))$ and $(\rho_N^{\mathrm{S}}(s), \rho_N^{\mathrm{Skew}}(s))$, where $\rho_N^a(s)$ is the normalized version of a given signature $\rho^a(s)$; $\rho_N^a(s)$ is defined as $\rho^a(s)$ divided by the maximum value that can be achieved by the related signature distance. This maximum distance depends on the sequence length. The maximum values of the signatures are provided in Supplementary Material (Table 1). Performances of the sum of minimal and maximal complementary signatures, namely $\rho^{\min} + \rho^{\max}$, were also analyzed.

2.2 Data acquisition and preprocessing

Complete genomes of 1,284 prokaryotes were downloaded from the NCBI ftp server[2]. The list of the genomes is provided online[3].

From a given genome, three sets of sequences were randomly sampled from both strands, simulating a sequencing error of 1%. Each of these sets consists of 10,000 possibly overlapping sequences with same length. Three sequence lengths where considered: 150 bp, 500 bp, and 1,000 bp. A sequence of length l is sampled by copying a random sub-sequence of l consecutive bases from a strand of a genomic

[2] Available at `ftp://ftp.ncbi.nih.gov/genomes/Bacteria/`

[3] Available at `http://cs.ru.nl/~gori/download/Table_S1_list_genomes.txt`

sequence; only sequences made exclusively of the four bases {A,C,G,T} were considered. Sequences were sampled with a random base-calling error of 1%. The sequencing error was simulated with the method adopted in [13]: each base had 1% of probability of been wrongly sequenced. The probability was uniformly distributed among the other three bases (e.g A has 99% probability of being correctly sequenced as A; A has a probability of 0.33% of being sequenced as C,G, or T, separately). This simple error model was chosen to make results not affected by biases of specific sequencing technologies. The NCBI taxonomy[4] [44] was used as reference taxonomy of the analyzed prokaryotes. Sequence sampling, analysis of the results and plotting were carried out using the following Python packages: Biopython [45], SciPy [46], IPython [47], and Matplotlib [48].

2.3 Computing signatures values

Revising the methodology employed in related works [13], we generated sets of sequences and evaluated the dissimilarity of the signature values on pairs of these sequences. Specifically, we evaluated the quality of a signature based on its property of assuming similar values for sequences of the same genome, and different values for sequences of different ones. We also evaluated the signatures' performance at taxonomic levels, by considering the *taxonomic distance* of two sequences as the taxonomic rank of the lowest common ancestor of their source genomes in the taxonomy tree.

Specifically, for each of the three sequence lengths (150 bp, 500 bp, 1,000 bp) we created 9 sets of sequence pairs, where each set corresponds to a different degree of diversity of the source genomes. Subsequently, signature distances between sequences for each pair of the sets were computed. From the resulting distance values, 9 distributions of distances were obtained for each signature. A first distribution was generated using the distances between sequences of a same genome (*intra-genome* signature distances): for each genome, we computed all the pairwise signature distances between the 10,000 sequences of that genome. These $\sim 6.42 \cdot 10^{10}$ distances ($\binom{10,000}{2}$ sequence pairs for 1,284 genomes) provided a distribution of intra-genome distances for the given signature. Each distribution was stored as a histogram of distance frequencies.

The other 8 distributions of distances were generated by computing distances between sequences from different genomes (*inter-genomic* signature distances), where each of the 8 distributions was obtained by considering a different level of taxonomic distance of the compared genomes. Specifically, we created 7 sets of organism pairs, one for each of the following taxonomic ranks: Species, Genus, Family, Order, Class, Phylum, Superkingdom. The set of pairs associated to a rank r consisted of 1,000 different pairs of organisms randomly selected among those whose lowest common ancestor in the taxonomy tree was at rank r. For each pair of these organisms, we randomly selected 1,000,000 pairs of genomic sequences from the set of all the sequences sampled from these genomes, and calculated the resulting distances. These 10^9 distances (1,000,000 sequence pairs for 1,000 genome pairs) provided a distribution of inter-genomic distances at rank r. Each distribution was stored as a histogram of distance frequencies, whose bins were the same used for intra-genome distances histogram. Furthermore, we also created a set of organism pairs where each element is made by a *bacterium* and an *archaeon*, the two superkingdoms of the *Prokaryotes*. We computed and stored a signature distance distribution for this set of organism pairs using the same methodology applied for the other inter-genomic distances. We refer to this distribution as the inter-genomic signature distance distribution at prokaryotes level.

2.4 Evaluating the effectiveness of signatures experimentally

We assessed the capability of a genomic signature to preserve the taxonomic relations between the source genomes of pairs of sequences. Specifically, for each genomic signature, we tested if the related signature distance yielded small values for sequence pairs of taxonomically closely related source microbes, and

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

greater values for sequences of distantly related microbes. To this aim, the signature distance was considered as a score for the sequence pair; the score quantifies the degree of relation between the source genomes, according to the given signature. Higher scores correspond to sequences that are more likely to belong to taxonomically distant genomes, according to the related signature.

Performances of different signatures were compared through Precision Recall (PR) curves [49]. The performance of each signature was evaluated at different *representation levels*: Intra-genome, Species, ..., Superkingdom. For a given representation level, sequence pairs were partitioned in two sets: the pairs having taxonomic distances up to the associated level, called "positives", and the remaining pairs, the "negatives". Specifically, for intra-genome representation level, the set of positive sequence pairs was made by the pairs sampled from the same genome; the remaining pairs formed the set of negatives. For representation level corresponding to taxonomic rank r, instead, we considered as positives all the sequence pairs such that the lowest common ancestor of the taxa of their source genomes was at rank r or lower. The remaining pairs were the negatives. Having defined the set of positives and negatives, we could compute the set of "true positives" and "false positives" for a given signature distance threshold and derive the PR curve. Given a signature distance threshold, we considered as "true positives" the positive pairs whose distance was below or equal to the threshold; similarly, "false positives" were made by the negatives with distance below or equal to the threshold. Therefore, for each distance threshold t we could compute the Precision and Recall, defined as follows:

$$\text{Precision}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FP}(t)}, \quad \text{Recall}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)},$$

where FN(t), TP(t), and FP(t) indicate the number of false negatives, true positives and false positives for t, respectively. Plotting Recall on the x-axis and Precision on the y-axis, a point in the PR space is derived for a given t. Varying t among the values of our distance distributions, we produced the PR curve for the associated signature. We used the PR curve because it can clearly show if a signature ρ^α is always better than signature ρ^β, namely if the PR curve of ρ^α is always above the one of ρ^β. PR curve was preferred to Receiver Operating Characteristic curve because it is more informative when data are highly skewed with respect to negatives/positives abundances [49]. This property is relevant in our case, since in a generic metagenomic dataset the sets of sequence pairs corresponding to different levels of taxonomic diversity can have different sizes; for instance, it might happen that pairs belonging to related genomes, i.e. the positives, would be much fewer than pairs of distantly related ones. As an index of signature quality, we used the Area Under the PR Curve (AUPR) [49].

PR curves were derived from the histograms of distance frequencies previously obtained, simulating the signature performances on three different community structures. The community structures are given by the topology of the taxonomic tree of the community members. The first community structure, called *complex*, is a binary taxonomic tree: the taxa of each rank have two descent taxa, and each species has two distinct strains in the community. As shown in Supplementary Material (Section 2.2), with this structure the number of sequence pairs with taxonomic distance at a given rank increases exponentially with rank highness. We decided to study such a complex structure because it is known that binning methods have problems with communities made by many species. The second community structure, called *medium*, is made by a total of 11 strains distributed among 7 species (Supplementary Figure 1); 6 of these species belong to the same phylum. In this structure, there are no sequence pairs with taxonomic distance at rank Class and Superkingdom, because no species pair has lowest common ancestor at these ranks. The third community structure, called *simple*, is made by a species with 3 strains and by other 3 species with one strain each (Supplementary Figure 2). These 4 species belong to 3 phyla of the same superkingdom. In this structure, the sequence pairs are present only for taxonomic distance at intra-genome level and at ranks Class and Superkindom. The detailed description of these structures is available in the Supplementary Material. To analyze the effectiveness of a signature on a given community structure, the histograms of distance frequencies for the different ranks were rescaled: this was done to

take into account that, for a given structure, the numbers of sequence pairs corresponding to different levels of taxonomic diversity of the source genomes respect a certain distribution. Details about the rescaling are provided in the Supplementary Material (Section 2.1).

For a given representation level and community structure, two histograms of distance frequencies were derived from the rescaled histograms; these two histograms represented the distances for the positive and the negative sequence pairs, respectively. For intra-genome representation level, the rescaled histogram related to intra-genome distances gave us the distance frequencies for the positive sequence pairs, the ones sampled from the same genome. The remaining rescaled histograms were added bin by bin, giving us the distance frequencies for the negative sequence pairs. For representation level corresponding to rank r, the rescaled histograms related to taxonomic distance at rank r or lower were added bin by bin, giving us the distance frequencies for the positive sequence pairs, i.e., the ones such that the lowest common ancestor of the taxa of their source genomes was at rank r or lower. The distance frequencies for the negatives were obtained in an analogous way, using the remaining rescaled histograms.

For each signature, the PR curve was derived from the histograms of positive and negative signature distances, respectively; the histograms shared the same bins. The PR curve was produced varying the threshold t among the edges of the histograms. Given a histogram, the number of sequence pairs whose distances were lower or equal than t was computed adding the histogram values for bins whose edges were lower or equal than t. Similarly, the number of sequence pairs whose distances were higher than t was computed using the histogram bins higher than t. Therefore, from the histograms we could compute the number of positives, true positives and false positives and hence the PR curve.

3 Results and Discussion

3.1 Theoretical analysis on the effect of GCSPR on signatures

We analyze and compare theoretically the statistical distributions of seven signatures; in particular, we study the signature dispersion among sequences of the same species (that should be as low as possible) and the dispersion of its expected values among different species (that should be as high as possible). The seven signatures under examination in this section are: four strand-independent signatures, namely the standard signature ρ^{T}, maximal $\rho^{\max}$ and minimal $\rho^{\min}$ complementary signatures, combination signature $(\rho^{\max}, \rho^{\min}, \rho^{\mathrm{P}})$; two GCSPR-based signatures, namely symmetrized signature ρ^{S} and operation signature ρ^{O}; and asymmetry signature ρ^{A}, capturing the deviation from GCSPR.

We model signatures' features of sequences as random variables. By analyzing the statistical dispersion of random variables corresponding to signature feature, we can assess theoretically the intra-species dispersion of signatures: a lower dispersion would correspond to better results, because it means that the signature assumes similar values for sequences sampled from the same species. We also analyze the inter-species discrimination capacity of signatures by looking at their distributions of per-species expected values: in this case, a higher dispersion indicates a better performance of the signature, because overall the signature values for different species will be more distant between each other.

Let V_i be the random variable corresponding to the total occurrence of k-mer w_i in a sequence randomly sampled from a given strand of a given organism, for $i = 1, \ldots, 4^k$. Consistently with literature [26], we assume that $V_1, \ldots, V_i, \ldots, V_{4^k}$ follow a multinomial distribution with success probabilities $g_1, \ldots, g_i, \ldots, g_{4^k}$, respectively. It is sensible to take g_i as the total occurrence of w_i in the given strand divided by the genome size of the given organism; in particular, the sum of the success probabilities must be equal to one:

$$\sum_{i=1}^{4^k} g_i = 1 \tag{1}$$

In particular, each V_i follows the binomial distribution $\mathrm{B}(g_i, n)$, where $n := l - k + 1$ is the number of subsequences of length k in a sequence of length l. For small values of k (e.g. $k = 4$) it is sensible to assume that all the g_i are strictly positive ($g_i > 0$) because each k-mer will occur multiple times in the genome.

We now focus our attention on the random variable $X_i := V_i/n$, that is the relative frequency of k-mer w_i in sequences sampled from a given strand of a given organism. It is plain to see that X_i corresponds to the i-th feature of standard signature $\rho_i^{\mathrm{T}}(s)$. By analyzing the statistical dispersion of X_i, we can assess whether the feature of ρ^{T} corresponding to w_i tends to assume the same values for all the sequences sampled from the same strand of the same organism. We denote by X_i^{RC} the variable corresponding to the relative frequency of the reverse complement w_i^{RC}.

GCSPR indirectly explains why standard signature ρ^{T} is effective on metagenomics data despite being strand-dependent; it also allows us to study ρ^{T} for sequences sampled from different strands. As recently stressed, GCSPR implies that inverse k-mers have approximately the same total count in a genome [31] and therefore we can assume that the global frequencies of w_i and w_i^{RC} are equal:

$$g_i^{RC} = g_i. \tag{2}$$

Let s and z be two sequences sampled from the opposite strands of the same genome. To determine the similarity between the two sequences, $\rho_i^{\mathrm{T}}(s)$ actually should not be compared to $\rho_i^{\mathrm{T}}(z)$ because the latter contains the relative frequency in z of the reverse complement of w_i in the strand of s, namely w_i^{RC}. However, equation (2) implies that X_i and X_i^{RC} are identically distributed, and hence the relative frequencies of w_i and w_i^{RC} in sequences of the same genome follow the same probability distribution, irrespectively of the source strand (see Supplementary Material Section 1). Therefore it is sensible to compare $\rho_i^{\mathrm{T}}(s)$ and $\rho_i^{\mathrm{T}}(z)$; in particular the two paired features have identical mean and variance:

$$E[X_i] = E[X_i^{RC}] = g_i, \tag{3}$$

$$Var[X_i] = Var[X_i^{RC}] = \frac{g_i(1 - g_i)}{n}. \tag{4}$$

By assuming that word frequency g_i is approximately the same in all the organisms of the same species, we can extend our theoretical analysis of ρ^{T} to all the sequences sampled from the same species. For standard signature, as for most of the signatures we analyze, an increase in sequence length l leads to lower variance; this is sensible for X_i, because the longer the sequence, the narrower will be the difference between local and global frequencies of k-mers. For the rest of the manuscript, we assume that the indices of non-palindromic w_i are ordered such that $X_i^{RC} = X_{120+i}$, $i = 1, \ldots, 120$; the last 16 random variables $X_{241}, \ldots, X_{256}$ correspond to palindromic k-mers. Thanks to this ordering and to equation (2), the constraint (1) can be rewritten as:

$$2 \sum_{\substack{i=1 \\ w_i \neq w_i^{RC}}}^{4^k} g_i + \sum_{\substack{i=1 \\ w_i = w_i^{RC}}}^{4^k} g_i = 2 \sum_{i=1}^{120} g_i + \sum_{i=241}^{256} g_i = 1. \tag{5}$$

The expected value of random variable X_i associated to non-palindromic k-mers therefore cannot be higher than $1/2$.

As measures of intra-species dispersion of the i-th feature, we use the coefficient of variation (CV), defined as the ratio between the standard deviation and the mean of that feature for sequences sampled from the same species. For standard signature ρ^{T} the coefficient of variation is as follows:

$$CV[X_i] = CV[X_i^{RC}] = \sqrt{\frac{1 - g_i}{n g_i}} = \sqrt{\frac{1}{n}\left(\frac{1}{g_i} - 1\right)}. \tag{6}$$

Signature with features having lower intra-species coefficient of variations have lower dispersion, and therefore are better because they tend to assume similar values for sequences of the same species. To compare the inter-species discrimination power of signatures, first of all we rescale the distribution of the expected signature value to the same range $[0, 1]$: each feature is rescaled such that $\rho_i(0) = 0$ and $\rho_i(1/2) = 1$. If two signatures have the same distribution of species expected values for the i-th feature, then they have the same inter-species distinction power. Otherwise we look at the coefficient of variations of the random variable corresponding to the expected value of the i-th feature for a given set of organisms; in this case, the higher the coefficient of variation, the higher is the discrimination power.

3.1.1 Symmetrized signature has lower intra-species dispersion than standard signature

Symmetrized ρ^{S} and operation signature ρ^{O}, as all the signatures based on the sum of features of the standard signature ρ^{T}, have smaller intra-organism coefficient of variation than ρ^{T} (Supplementary Information Sections 1.1 and 1.4). That result is obtained by modelling the sum of features as a random variable $\grave{U}_j$ in the following way:

$$\grave{U}_j := \sum_{i \in I_j} X_i = \sum_{i \in I_j} V_i/n, \tag{7}$$

where $I_1, \ldots, I_h$ are a partition of the index set $\{1, \ldots, 4^k\}$; for example, for ρ^{S} and ρ^{O} the set I_j corresponds to the indices of words $\{w_j, w_j^{RC}\}$ and $\{w_j, w_j^{RC}, w_j^{C}, w_j^{R}\}$, respectively. It is worth to mention that the use of GCSPR was not required for this observation.

Nevertheless, summing features could reduce the inter-species discrimination efficacy of a signature: sequences s_1 and s_2 sampled from different organisms could have similar values for $\grave{U}_j$ despite having distinct g_i's. For example, let us suppose that a signature has a feature given by the sum of the first two features of ρ^{T}, and that $X(s_1) = (1/128, 1/64, \ldots)$ and $X(s_2) = (5/256, 1/256, \ldots)$. We have that the inequality $(X_1(s_1), X_2(s_1)) \neq (X_1(s_2), X_2(s_2))$ holds but $X_1(s_1) + X_2(s_1) = X_1(s_2) + X_2(s_2)$.

However, GCSPR gives symmetrized signature ρ^{S} the same inter-species discrimination power of the standard signature ρ^{T}. Our analysis therefore indicates that signature outperforms standard signature because it has a lower intra-species dispersion but the same inter-species discrimination power. To illustrate that, by following equation (7) we model $\rho_i^{\mathrm{S}}(s)$ as a random variable as follows:

$$\rho_i^{\mathrm{S}}(s) = Z_i := X_i + X_i^{RC};$$

as shown at the beginning of Section 3.1, GCSPR implies that X_i and X_i^{RC} are identically distributed and have mean and variance as in equations (3) and (4). Therefore Z_i has the following mean, variance, and coefficient of variation (see Supplementary Material Section 1.1):

$$E[Z_i] = 2g_i, \qquad Var[Z_i] = \frac{2g_i(1 - 2g_i)}{n},$$
$$CV[Z_i] = \sqrt{\frac{1}{n}\left(\frac{1}{2g_i} - 1\right)}. \tag{8}$$

We compare the inter-species discrimination capacities of symmetrized and standard signatures by looking at their distributions of per-species expected values. For the 240 non-palindromic k-mers, symmetrized signature has 120 features because it combines the values of X_i and X_i^{RC}. Therefore, we compare the distributions of the expected values of Z_i with the distribution of the average expected values of X_i and X_i^{RC}. By the definition of Z_i, we have that the expectation of Z_i is twice the average expectation of X_i and X_i^{RC}: $2(E[X_i] + E[X_i^{RC}])/2 = E[X_i + X_i^{RC}] = E[Z_i]$. Since Z_i takes values between zero and one, $X_i + X_i^{RC}$ is the rescaling of $(X_i + X_i^{RC})/2$ on that range and has the same expectation of Z_i. Since the

expected values are identical for the same species, the distributions of their expected values per-species are also identical.

Inter-species distinction reduction could actually occur for operation signature, although it is partially compensated by GCSPR. Let $U_i := X_i + X_i^{RC} + X_i^C + X_i^R$ be the random variable corresponding to ρ_i^{O}; as observed at the beginning of Section 2.1, w_i^R is the inverse of w_i^C and therefore the GCSPR implies that U_i follows a distribution with mean $2g_i + 2g_i^C$ and variance $2(g_i + g_i^C)[1 - 2(g_i + g_i^C)]/n$ (Supplementary Material Sections 1.1). This means that the random variable corresponding to the sum of four k-mers actually depends on only two of them; however, it is likely that summing the frequencies of w_i and w_i^R will have negative effects on the performance of the signature, because there is no symmetry rule for those kind of pairs. Specifically, operation signature cannot distinguish organisms having different values for k-mer frequencies of reverse words g_i and g_i^{RC} but the same value for the sum $g_i + g_i^C$.

3.1.2 Combination signature performs in between standard signature and symmetrized signature

To simplify the analysis of combination signature, in this section we consider the normal approximation of X_i. Specifically, we replace each binomial variable V_i with its normal approximation $\mathcal{N}(ng_i, ng_i(1-g_i))$. Consequently, even the variable X_i follows a normal distribution because it is now the product of a constant term $1/n$ and a Gaussian variable; its mean is g_i and its variance is $g_i(1-g_i)/n$:

$$\rho_i^{\mathrm{T}}(s) = X_i \sim \mathcal{N}(g_i, g_i(1-g_i)/n).$$

Let $X_i^{RC} \sim \mathcal{N}(g_i^{RC}, g_i^{RC}(1-g_i^{RC})/n)$ be the frequency of w_i^{RC} in sequence s. Given the normal distribution of ρ^{T} features, it follows that maximal $\rho_i^{\max}$ and minimal $\rho_i^{\min}$ features follow the distribution of the maximum and the minimum of the pair of Gaussian variables $\{X_i, X_i^{RC}\}$, respectively. We can therefore define the following random variables corresponding to the signature values:

$$\rho_i^{\max}(s) = \hat{Y}_i := \max(X_i, X_i^{RC}),$$
$$\rho_i^{\min}(s) = \tilde{Y}_i := \min(X_i, X_i^{RC}).$$

Formulas for first and second moment of maximum and minimum of pairs of Gaussian variable are known [50] and from those we can express mean and variance of $\hat{Y}_i$ and $\tilde{Y}_i$ (see Supplementary Material 1.2). Since those maximal and minimal complementarity signatures are strand-independent, random variables $\hat{Y}$ and $\tilde{Y}$ and their moments are also strand-independent.

The GCSPR allows us to show that maximal and minimal signature features, and thus the features of combination signature, have lower variance than the standard signatures. Equation (2) indeed drastically simplifies the first and second moment of $\hat{Y}$ and $\tilde{Y}$ and hence $\rho^{\max}$ and $\rho^{\min}$. By replacing g_i^{RC} with g_i we obtain (see Supplementary Material 1.2):

$$E[\hat{Y}_i] = g_i + \sqrt{\frac{g_i}{n\pi}}, \qquad E[\tilde{Y}_i] = g_i - \sqrt{\frac{g_i}{n\pi}}, \tag{9}$$

$$Var[\hat{Y}_i] = Var[\tilde{Y}_i] = Var[X_i] - \frac{g_i}{n\pi} = \frac{g_i}{n}\left(1 - \frac{1}{\pi} - g_i\right) \tag{10}$$

$$CV[\hat{Y}_i] = \frac{\sqrt{\frac{g_i}{n}\left(1 - \frac{1}{\pi} - g_i\right)}}{g_i + \sqrt{\frac{g_i}{n\pi}}} = \frac{\sqrt{\pi - \pi g_i - 1}}{\sqrt{n g_i \pi} + 1} \tag{11}$$

$$CV[\tilde{Y}_i] = \frac{\sqrt{\pi - \pi g_i - 1}}{\sqrt{n g_i \pi} - 1}. \tag{12}$$

Since $g_i/n\pi$ is strictly positive, it is plain to see from (10) that maximal and minimal signature features have lower variance than the standard signature features for all the non-palindromic k-mers - for the palindromic there is no difference.

The combination signature has lower intra-species dispersion than the standard signature but higher than the symmetryzed one (Figure 1). For the 16 palindromic k-mers, standard, symmetrized, and combination signatures are identical. Since combination signature for the 240 non-palindromic k-mers is a combination of maximal and minimal complementarity signature, the coefficients of variation of its features do not have the same formula. Therefore, we study the average value of $CV[X_i]/CV[\hat{Y}_i]$ and $CV[X_i]/CV[\tilde{Y}_i]$. As shown in Section 1.4 of Supplementary Material, this is higher than one, and therefore combination signature has a lower dispersion than standard signature. On the other hand, in the same section we show that combination signature has higher dispersion than symmetrized signature for the non-palindromic k-mers, as shown in Figure 1:

$$1 < \frac{1}{2}\left(\frac{CV[X_i]}{CV[\hat{Y}_i]} + \frac{CV[X_i]}{CV[\tilde{Y}_i]}\right) < \frac{CV[X_i]}{CV[Z_i]}.$$

Combination signature has the same inter-species discrimination power of the standard signature. For the 240 non-palindromic k-mers, combination signature has 120 pairs of features, namely $\hat{Y}_i$ and $\tilde{Y}_i$, and each of these pairs is a function of the values of X_i and X_i^{RC}. Therefore, we compare the distributions of the average expected values of $\hat{Y}_i$ and $\tilde{Y}_i$ with the distribution of the average expected values of X_i and X_i^{RC}. Thanks to equations (9) we have the following equality: $E[\hat{Y}_i] + E[\tilde{Y}_i] = 2g_i = E[X_i] + E[X_i^{RC}]$. Hence the distributions of $(E[\hat{Y}_i]+E[\tilde{Y}_i])/2$ and $(E[X_i]+E[X_i^{RC}])/2$ are identical, and the two signatures have the same inter-species discrimination power. Since the expected values are identical, the distributions of their expected values are also identical.

3.1.3 Divergence from GCSPR carries a phylogenetic signal

The expected values of asymmetry signature ρ^{A} are functions of the global word frequencies, and therefore ρ^{A} carries a phylogenetic signal. Indeed the i-feature of asymmetry signature is defined as the absolute difference between the frequencies of w_i and w_i^{RC}; we show that the random variable A_i corresponding to the value of ρ_i^{A} among sequences sampled from the given organism can be rewritten as a function of $\hat{Y}$ and $\tilde{Y}$:

$$A_i := |X_i - X_i^{RC}| = \max(X_i, X_i^{RC}) - \min(X_i, X_i^{RC}) = \hat{Y}_i - \tilde{Y}_i.$$

The mean, the variance, and the coefficient of variation for this signatures are (see Supplementary Material section 1.3):

$$\begin{aligned} E[A_i] &= 2\frac{g_i}{n\pi}, \\ Var[A_i] &= \frac{2g_i}{n}\left(1 - \frac{2}{\pi}\right), \\ CV[A_i] &= \sqrt{\frac{\pi}{2} - 1}. \end{aligned} \tag{13}$$

in contrast with all the other signatures studied, its intra-species dispersion does not decrease with read length and is actually constant (13).

The intra-species dispersion seems mostly worse than the symmetrized signature, and probably also of the plain signature. Indeed, the intra-species coefficient of variation strictly decreases with k-mer frequencies; it was lower than standard and symmetrized signatures for frequencies above 0.00351 and 0.00175, respectively (Figure 1).

We conduct an analysis of the intra-species coefficient of variation. Let F_i be the random variable corresponding to the expected values of k-mer w_i frequency for an organism belonging to a given set of organisms. The coefficient of dispersion of F_i measures the intra-species dispersion of the standard signature: the higher this coefficient, the more spread are the expected value for feature i of the standard

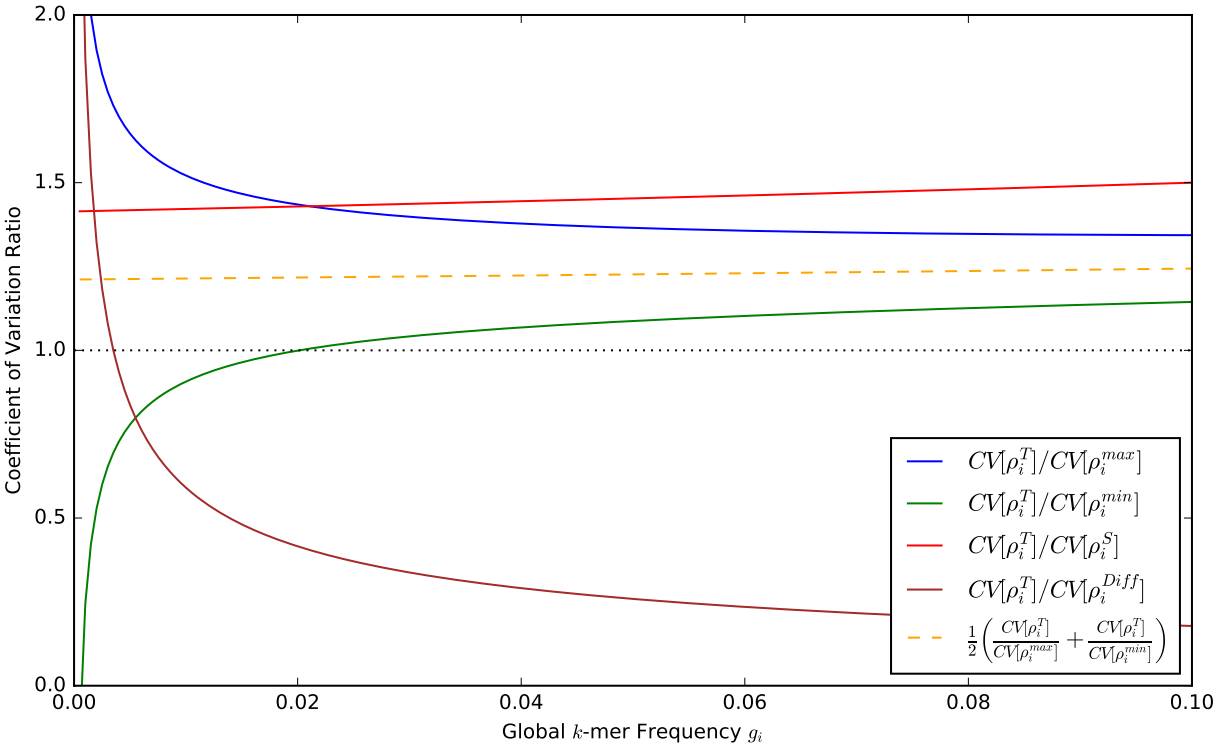

Figure 1. Ratio between coefficient of variations of some studied signatures and the coefficient of variation of the standard signature $\rho^{\mathbf{T}}$ for sequences of 500bp. Relevant ratio are displayed for symmetrized signature, combination signature, and asymmetry signature

signature. Therefore, a higher coefficient leads to better discrimination between the given organisms. Our analysis shows that asymmetry signature can outperform the standard signature intra-species discrimination power if and only if $CV[\sqrt{F_i}] > \sqrt{3}$ (see Supplementary Section 1.5).

3.1.4 GCSPR-based signatures have better error-tolerance

An advantage of symmetrized signature over combination signature is in its tolerance to local deviation from GCSPR. In particular, symmetric deviations from k-mers frequence parity will have no effect on symmetrized signature. Indeed, it is known that Chargaff's second parity rule, stating that complementary nucleotides have the same frequency along a strand, may not hold for sequences shorter than a species-dependent "critical fragment length"; this length is comprised between 6 kbp and 50 kbp [51]. The parity rule seems the result of alternating regions with different signs of deviation from parity. Consequently, it is reasonable to assume that even the version of the rule generalized for k-mers (the GCSPR) will not hold for some of the reads, whose length is for sure lower than the GCSPR-equivalent of the critical fragment length. For sequences of those regions, X_i will still follow a binomial distribution but with a success rate $\acute{g}_i = g_i + \delta_i$, with $\delta_i \in (-g_i, 1/2 - g_i)$ dependent on the sampled region and such that $\sum_i \delta_i = 0$. Let us focus on a special case of this deviation from parity here called *symmetric deviation*, where complementary k-mers frequency deviate from parity but with opposite deviation of the same size, i.e. $\delta_i^{RC} = -\delta_i$. Symmetrized and operation signature will be uneffected by this deviation because the sum of the success rates of complementary k-mers will be the same: $\acute{g}_i + \acute{g}_i^{RC} = g_i + \delta_i + g_i - \delta_i = 2g_i$

By summing frequencies of different tetranucleotides, the GCSPR-based signatures are more tolerant than the standard signature with respect to base-calling errors. It is plain to see that symmetrized signature ρ^{S} has some degree of tolerance to base-calling errors: if base-calling errors replace a tetranucleotide w with its inverse w^{RC}, then the sum of their frequencies will be preserved. Moreover, if a base-calling error replace tetranucleotide w_i with $w_j \neq w_i^{RC}$, then the features of ρ^{S} corresponding to them could still be preserved not only if the opposite error occurs (w_j replaced by w_i) but also if any of w_j and w_j^{RC} is replaced either by w_i or w_i^{RC}. Those two mechanisms of error-tolerance of ρ^{S} are strengthened in ρ^{O}. Indeed, feature associated with tetranucleotides w, w^C, w^R, w^{RC} will be preserved not only if a base-calling error replace one of those with its inverse, but also if any of the them is replaced by one of the other three. Moreover, if a base-calling error replace tetranucleotide w_i with $w_j \notin \{w_i^C, w_i^R, w_i^{RC}\}$, then the features of ρ^{O} corresponding to them could still be preserved if any of w_j, w_j^C, w_j^R, and w_j^{RC} is replaced by any of w_i, w_i^C, w_i^R, or w_i^{RC} (hence there are 16 possible compensating errors instead of 4 and 1 of ρ^{S} and ρ^{T}, respectively).

3.2 Experimental results

Signatures' experimental performances were assessed through Precision Recall curves, using the Area Under the Precision Recall curves (AUPRs). The AUPRs obtained by a signature for different levels of representation were compared with the ones of an artificial signature that cannot distinguish between the different taxonomic ranks, because it has the same distance distribution for each rank. We show in the Supplementary Material that the AUPR of this artificial signature is equal to the ratio of positives in the data. If the AUPR of a signature is higher than the one of this artificial signature, then it can be considerate efficacious; the higher the AUPR, the better the signature.

Experimental performances of symmetrized, combination, and standard signatures were consistent with the theoretical analysis. Indeed, symmetrized signature ρ^{S} outperformed combination signature $(\rho^{\max}, \rho^{\min}, \rho^{\mathrm{P}})$, which in its turn outperformed standard signature ρ^{T}; these relations held for each representation level and community structure. More in general, the best experimental results were achieved by strand-independent signatures summing or reordering most of the tetranucleotides frequencies, like ρ^{S}, $(\rho^{\max}, \rho^{\min}, \rho^{\mathrm{P}})$, $\rho^{\min} + \rho^{\max}$, and $(\rho^{\max}, \rho^{\min})$ (Figures 2, 3, 4, and Supplementary Figure 4).

The novel operation signature ρ^{O} was the signature with the lowest number of features among the ones superior or comparable to the standard signature ρ^{T} (ρ^{T} and ρ^{O} had 256 and 72 features, respectively). Theoretical analysis of ρ^{O} proved that it has the lowest intra-species dispersion among the signatures we study, but its inter-species discrimination power could be lower than the standard signature. Experiments indicate that its intra-species dispersion compensates its reduced inter-species discrimination power; this excellent intra-species dispersion allows operation signature to perform better than the standard signature. In agreement with theoretical analysis, experiments showed that operation signature gives the best performances when inter-species discrimination is easier. Adding frequencies of non-inverse k-mers probably leads to some information loss; however, the negative effect of this loss will be reduced when the microbial community is composed by few species. In this case, the likelihood of having organisms having different values for k-mer frequencies of reverse words g_i and g_i^{RC} but the same value for the sum $g_i + g_i^C$ is reduced. Indeed, ρ^{O} excelled ρ^{T} for community structures of simple and medium complexity (Figure 3 and Supplementary Figure 4), while it had similar results for complex community structure (Figures 2 and 4). It is likely that ρ^{O} achieved good performances because it was inspired by a genomic rule [34], but we cannot exclude that similar results could be obtained by comparable reduction of ρ^{S} (i.e. by summing frequencies of other sets of tetranucleotide frequencies that still lead to 72 features).

In accordance with our theoretical analysis, signatures capturing the deviation from GCSPR, like asymmetry signature ρ^{A}, carried a phylogenetic signal; however, this signal was too weak to lead to appreciable performances. Figure 5 shows that the distance associated to signature ρ^{A} tends to assume slightly higher values for sequences coming from distantly related species; indeed, the signature distance

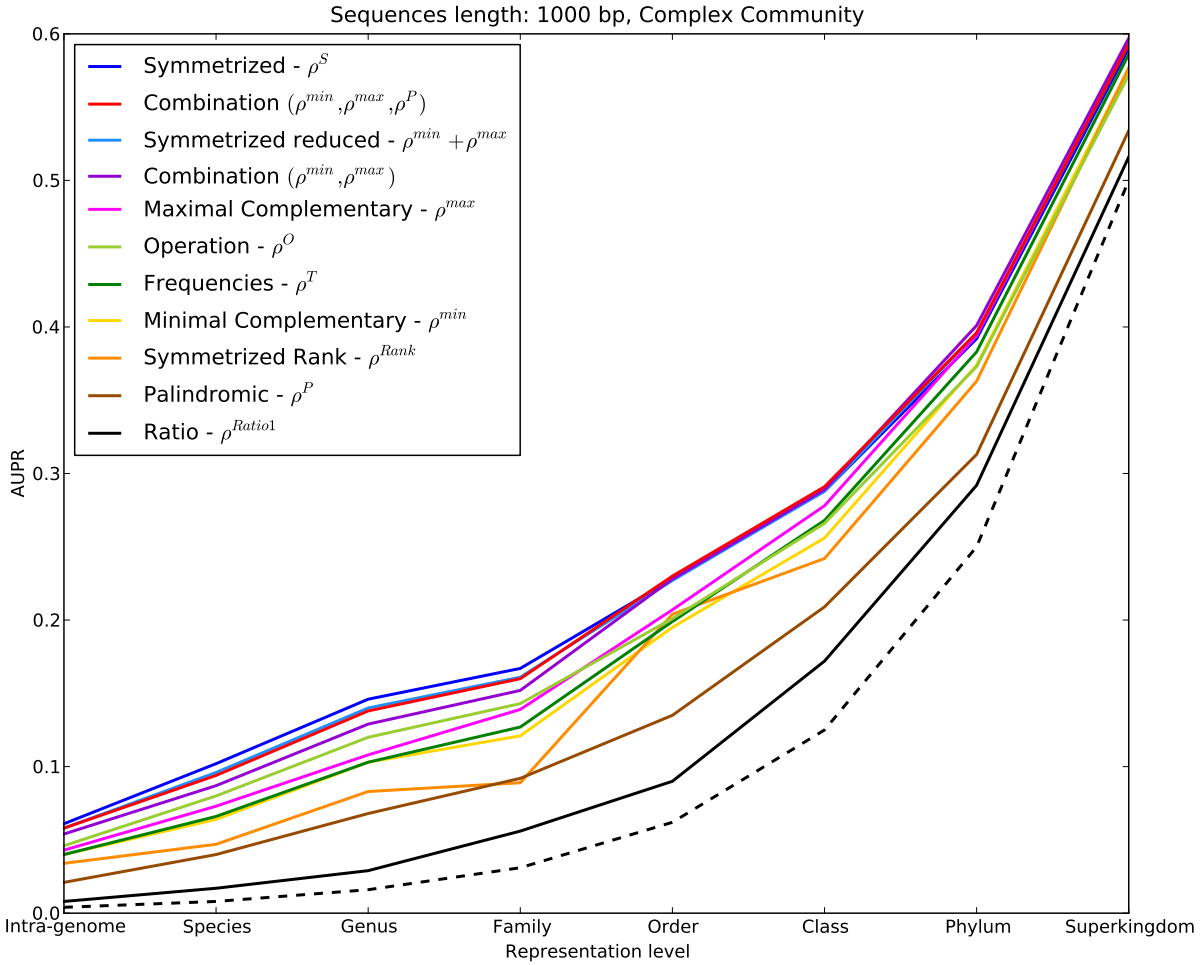

Figure 2. AUPRs obtained by the best signatures on complex community structure, for 1000 bp sequences. Dashed line is made by the AUPRs of a signature whose distance distributions are identical for each rank; the AUPRs of this signature do not depend on distribution shape. Signatures' names in the legend are sorted with respect to the sum of their AUPRs for the different levels of taxonomic distance.

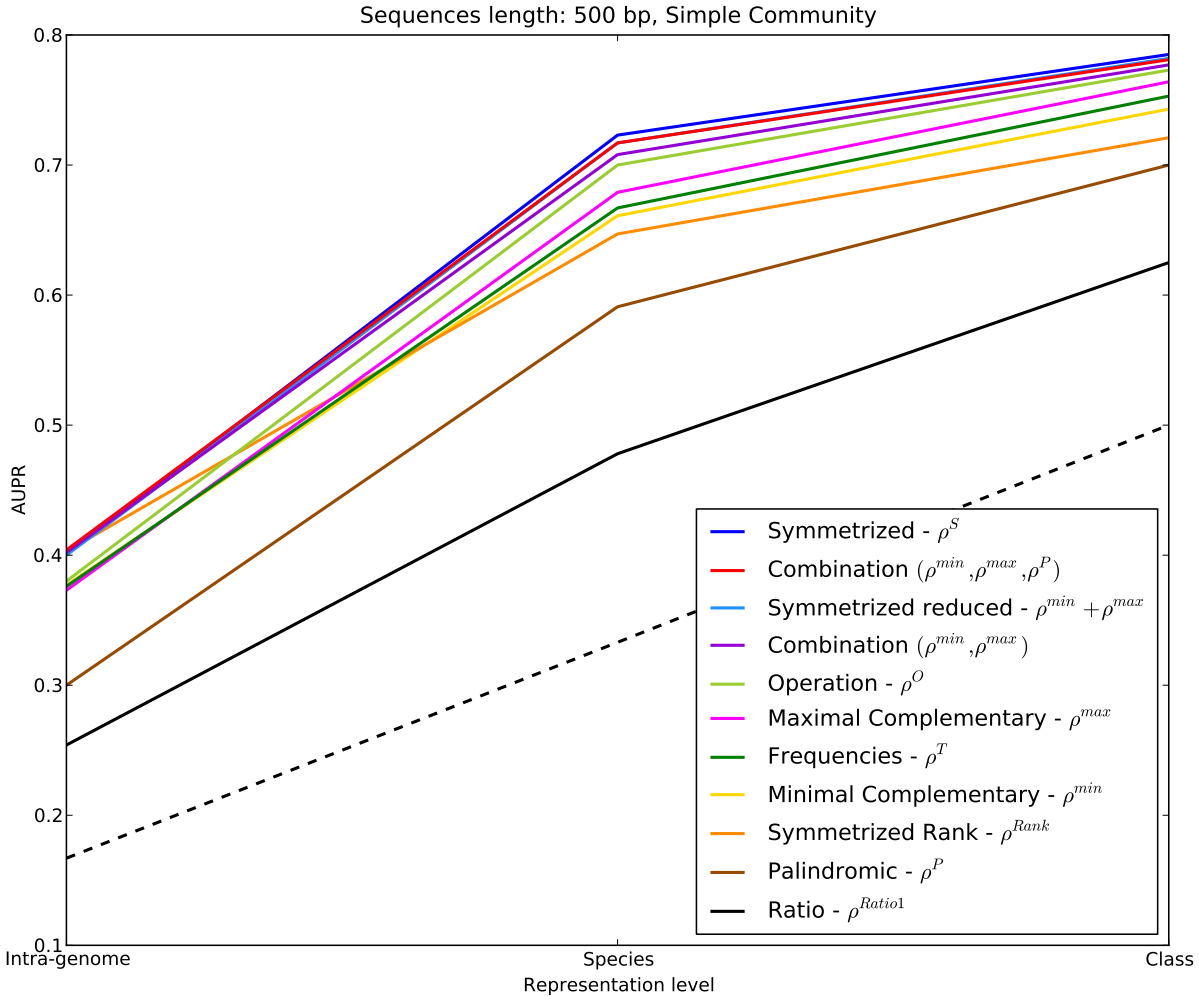

Figure 3. AUPRs obtained by the best signatures on simple-complexity community structure, for 500 bp sequences. Dashed line is made by the AUPRs of a signature whose distance distributions are identical for each rank; the AUPRs of this signature do not depend on distribution shape. Signatures' names in the legend are sorted with respect to the sum of their AUPRs for the different levels of taxonomic distance.

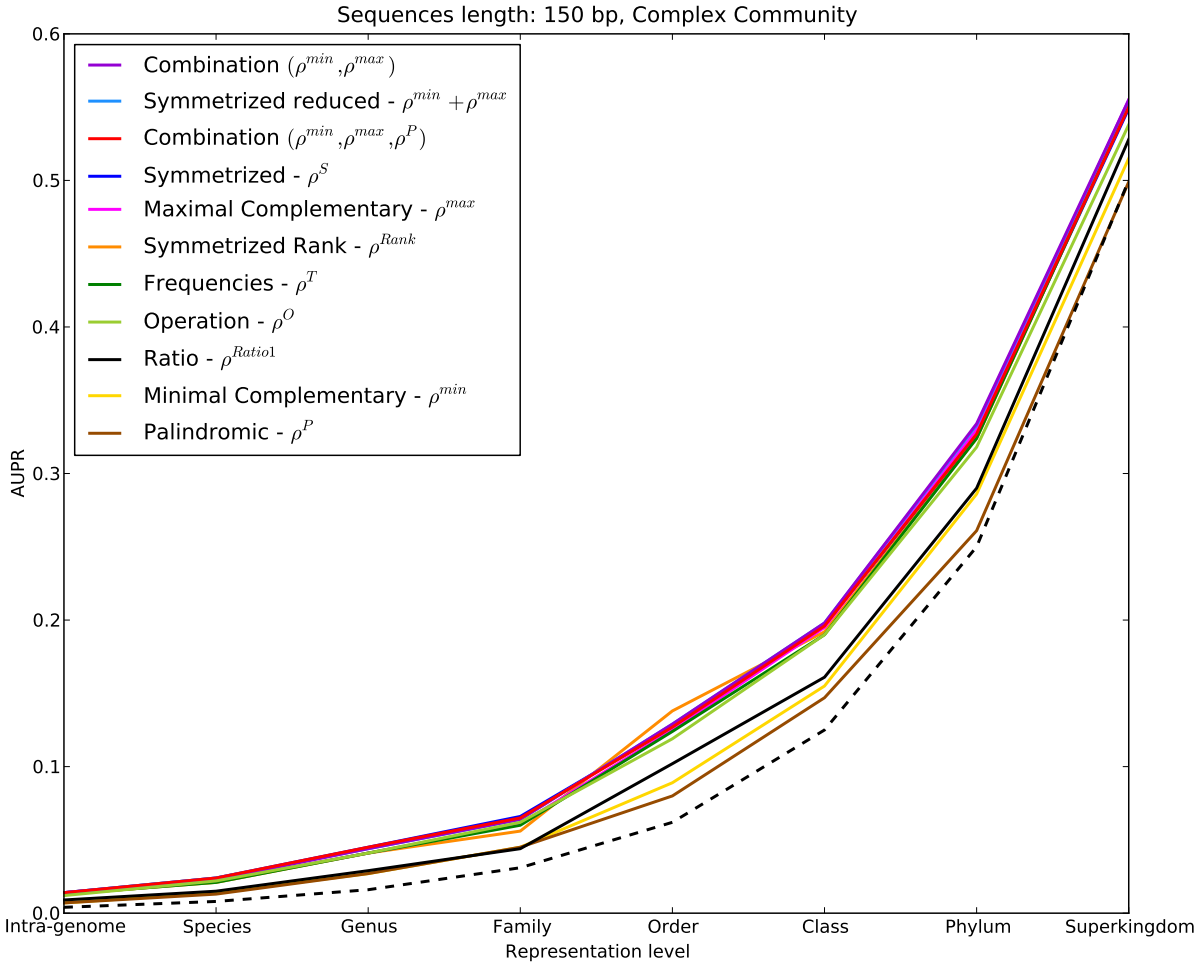

Figure 4. AUPRs obtained by the best signatures on complex community structure, for 150 bp sequences. Dashed line is made by the AUPRs of a signature whose distance distributions are identical for each rank; the AUPRs of this signature do not depend on distribution shape. Signatures' names in the legend are sorted with respect to the sum of their AUPRs for the different levels of taxonomic distance.

distributions associated to higher levels of taxonomic distance are a bit shifted to higher values. Despite carrying a phylogenetic signal, signatures capturing the deviation from GCSPR were the worst in almost any experiment; ratio signature, which was the best deviation-capturing signature, had actually the worst performance among the signatures displayed in Figures 2 and 3. Our theoretical analysis of ρ^{A} suggest that this is due to high intra-species signature dispersion.

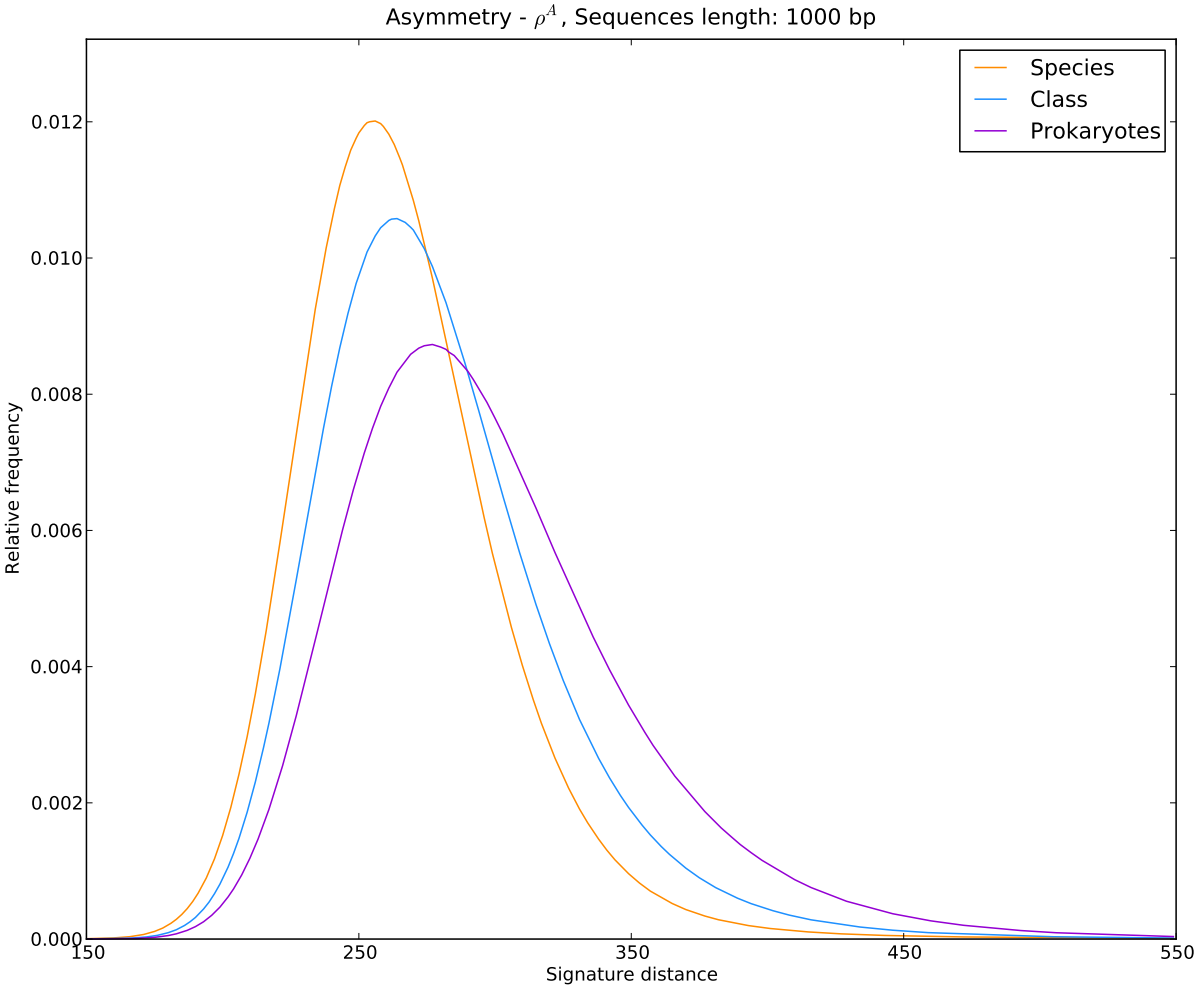

Figure 5. Normalized empirical distributions of Asymmetry Signature distances, for different levels of taxonomic diversity.

Signature performances increased with read length, as predicted by the theory. For instance, Figure 6 shows that longer sequences led to higher AUPRs for ρ^{S}; the same happened for the other signatures. Theoretical analysis of symmetrized and standard signatures indicates that this is due to the reduction of intra-species dispersion, that tends to zero as read length l (and thus $n = l - 4 + 1$) increases – see Equations (8) and (6). The same observation holds for compositional, maximal complementary and minimal complementary signatures (11)-(12). This phenomenon is sensible also because the longer a sequences is, the more information it contains; hence, the compositional properties that characterize the source genomes are more recognizable. The trend was weaker for signatures designed to capture the

deviation from GCSPR. In particular, their performances for complex community structure were not affected by sequence length (for ρ^{Ratio1}, see Supplementary Figure 3). This could be due to the fact that their intra-species dispersion could be independent from read length; this is the case of the asymmetry signature (13) as our theoretical analysis showed.

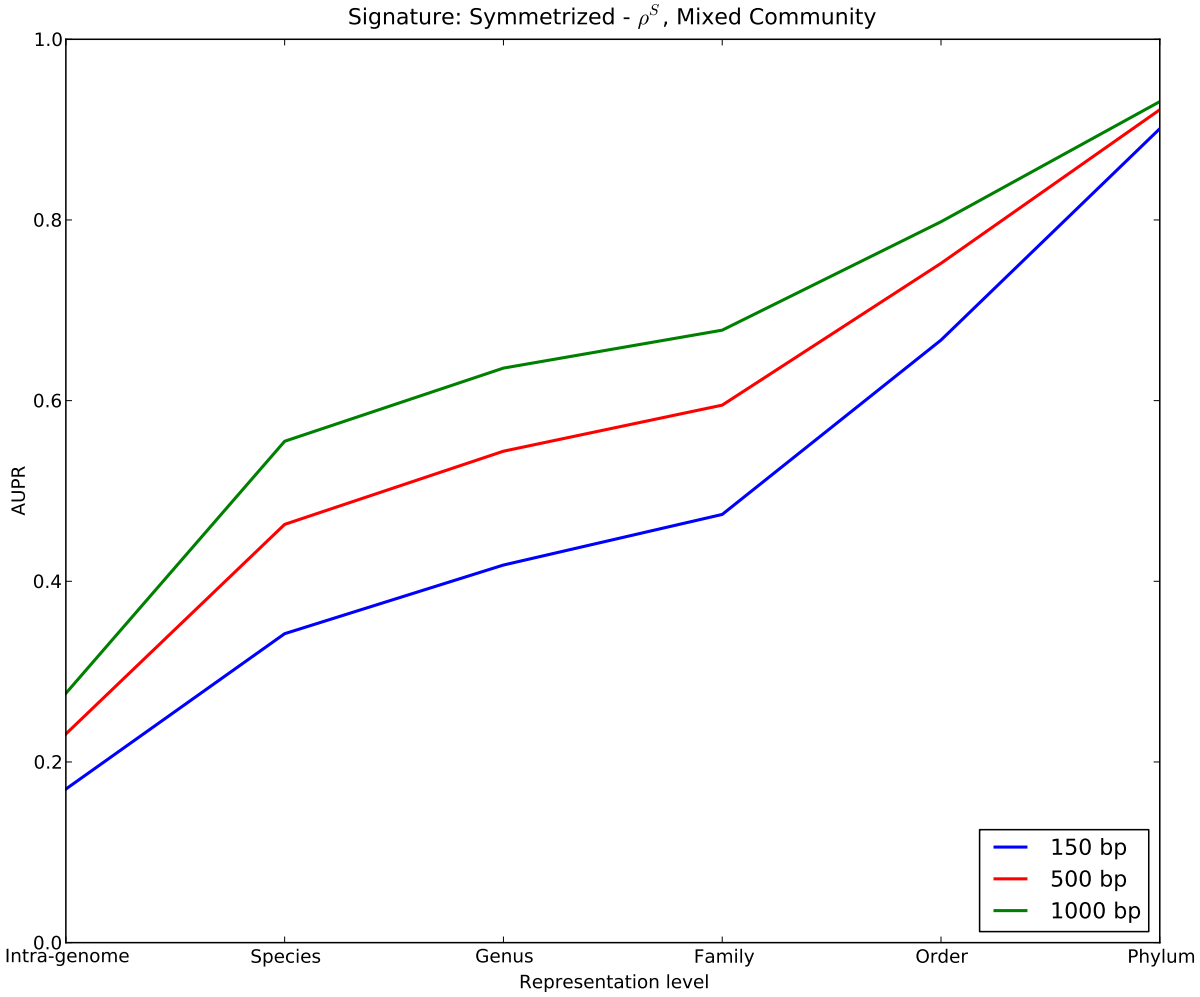

Figure 6. AUPRs obtained by Symmetrized Signature for all the sequence lengths. AUPRs were computed on medium-complexity community structure.

Reducing symmetrized and combination signatures to the features corresponding to the non-palindromic tetranucleotides worsened their performances, but they still outperformed the standard signature (Figure 7). Indeed, as mentioned before, signatures $\rho^{\min} + \rho^{\max}$ and $(\rho^{\max}, \rho^{\min})$ outperformed the standard signature; moreover, they coincide with the features of ρ^{S} and $(\rho^{\max}, \rho^{\min}, \rho^{\mathrm{P}})$ corresponding to the non-palindromic tetranucleotides, respectively. Our theoretical analysis indicates that symmetrized and combination signatures outperform the standard signature thanks to their reduced intra-species dispersion; experimental results thus indicate that this factor is so strong that those two signatures outperform the standard signature even if they loose those 16 features of ρ^{S}. Nevertheless, it might be possible that

good performances could still be obtained by removing features corresponding to a different and perhaps larger set of k-mers.

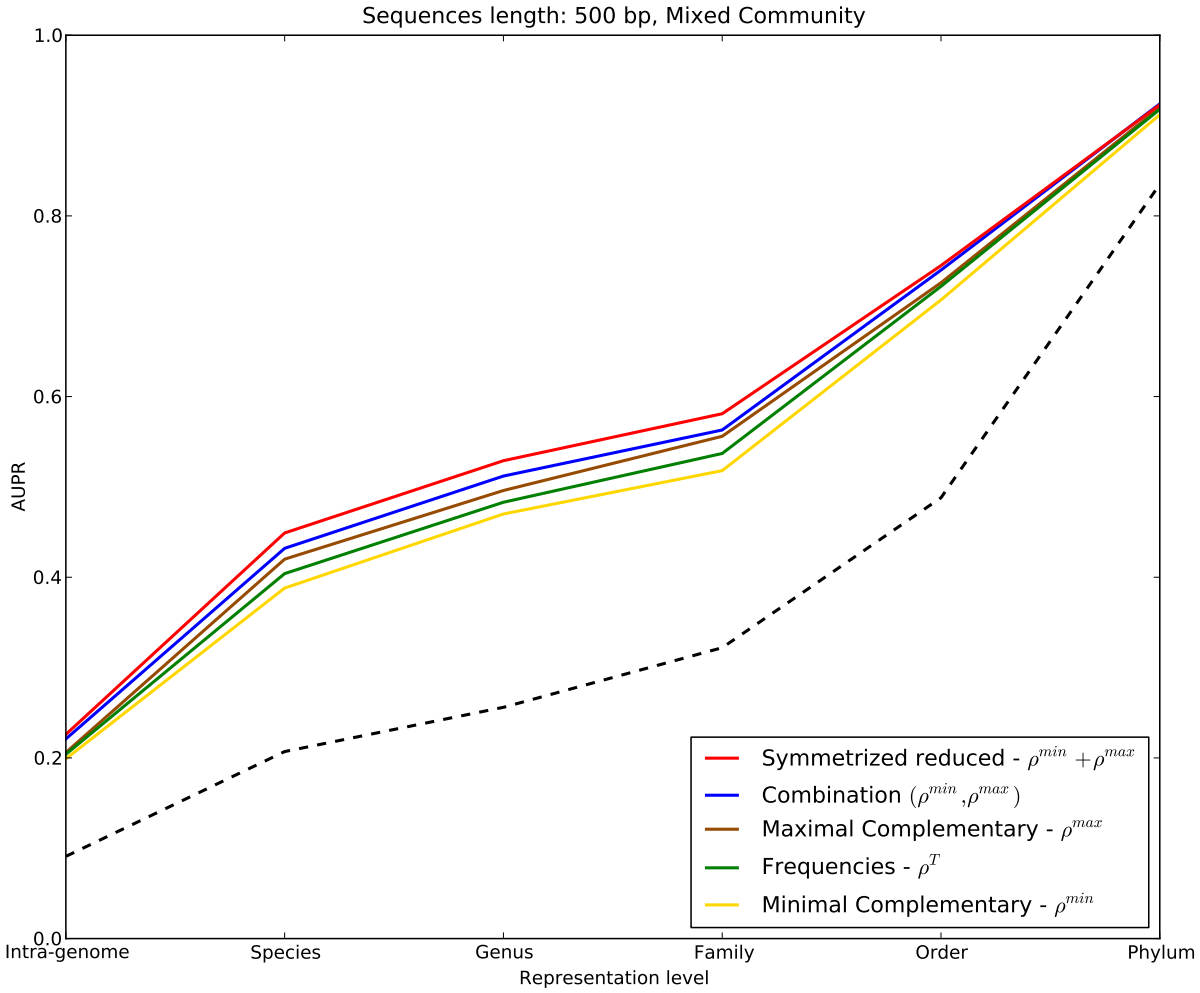

Figure 7. AUPRs obtained by a few asymmetry-related signatures on medium-complexity community structure, for 500 bp sequences. Dashed line is made by the AUPRs of a signature whose distance distributions are identical for each rank; the AUPRs of this signature do not depend on distribution shape. Signatures' names in the legend are sorted with respect to the sum of their AUPRs for the different levels of taxonomic distance.

4 Conclusions

In this work, we conducted a theoretical analysis of new and existing genomic signatures for metagenomes; we analyzed their intra-species dispersion and their inter-species discrimination power. Furthermore, signatures' performances were evaluated experimentally; the signatures were tested with respect to their capability to preserve the taxonomic relations of the source organisms of pairs of sequences. Signature

distances were evaluated for sequences of 1,000 bp, 500 bp, and 150 bp randomly sampled from 1,284 prokaryotic genomes.

In metagenomic literature, symmetrized signature is erroneously believed to outperform the standard tetranucleotide signature simply because it is strand-independent. Our theoretical analysis proved that even the standard signature is actually not affected by the sampling strand; this is due to the GCSPR, which implies that frequencies in equally long reads of inverse k-mers follow the same probability distribution.

Moreover, our theoretical analysis showed that the success of symmetrized signature is due to its direct exploitation of GCSPR. By summing frequencies of inverse k-mers, the inter-species discrimination power is unmodified: there is no information loss because the frequencies of the combined k-mers follow the same probability distribution. Moreover, this summation reduces the intra-species dispersion, thus leading to performance improvement. Experimental results confirmed that symmetrized signature had the best performance among all the tested signatures; the dimension of its feature space is about half of the one of the standard signature, making the data also more tractable.

On the contrary to what is believed, summing frequencies of inverse k-mers is not the only effective way to produce an effective strand-independent signature from the standard signature. The novel combination signature achieves strand-independence through a biologically-sensible reordering of standard signature's features. As predicted by the theory, experimental performances of combination signature were a bit below symmetrized signature but still higher than standard signature. Combination signature thus seems the ideal choice if oligonucleotide frequencies must be kept separated; in general, it could be used by alignment-free sequence comparison methods when source genomes do not respect GCSPR or distinguishing inverse oligonucleotides matters. In particular, it can be advantageous for tasks on metagenomic data that are not strictly related to taxonomy, like the identification of *cis*-regulatory modules [21]. It can still be helpful for taxonomic analysis of metagenomes when they contain sequences sampled from genomes not respecting GCSPR, like viral genomes. Experimental results also indicate that some features of symmetrized and combination signatures can be removed without too much decrease of performance.

The performances of the novel operation signature indicates that other genomic symmetries than GCSPR can be successfully exploited for designing low-dimensional signatures. Indeed, despite having lower performance than the symmetrized signature, it has about half of its features and was superior to the standard signature for communities with not-very-complex taxonomic structures (and at least comparable for very complex ones). Theoretical analysis proved that it has the lowest intra-species dispersion among the studied signatures; however, its inter-species discrimination power could be lower of the one of the standard signature. Therefore, operation signature can be particularly beneficial for the analysis of large metagenomes sampled from microbial communities whose taxonomic structure is not extremely complex; indeed, dealing with less features reduces the computational cost of data analysis [36].

Deviation from GCSPR seems related to the taxonomic classification of the species, but not strongly: signatures that are exclusively based on these asymmetries are not good enough to achieve the best performances. Theoretical analysis of asymmetry signature suggests that performances are low due to a high intra-species signature dispersion.

As predicted by the theoretical analysis and confirmed by the experiments, signature performances increase with sequence length. It is likely that no signature can be effective on very short sequences, due to the little information contained. Some works [16, 25, 26] dealt successfully with this issue by adopting the following approach: short sequences were grouped, and then the signature was computed on this set as if it were one long sequence. However, this procedure might be risky for metagenomes with high diversity or low species coverage [52]. An alternative method is to compute the signatures on reads assembled into contigs [18] but unfortunately the assembly process of a metagenome is computationally very intensive.

Our approach provides a unified framework to compare the performances of signatures by means of binomially and normally distributed random variables. This framework can be used to guide the development of novel signatures.

Acknowledgments

FG participated in conceiving the study, performing the data analyses, and interpreting the results; he developed the theoretical statistical analysis, generated the data, and wrote the manuscript. DM participated in conceiving the study, analyzing the data, interpreting the results, and supported the writing of the manuscript. MJ participated in conceiving the study and in interpreting the results. EM participated in conceiving and coordinating the study, analyzing the data, and helped writing the manuscript. All authors read and approved the final manuscript.

References

1. Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Computational Biology 6: e1000667.

2. Amann R, Ludwig W, Schleifer K (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews 59: 143–169.

3. Wu YWW, Ye Y (2011) A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. Journal of Computational Biology 18: 523–534.

4. Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nature Methods 6: 673–676.

5. Kelley D, Salzberg S (2010) Clustering metagenomic sequences with interpolated Markov models. BMC Bioinformatics 11: 544.

6. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Research 17: 377–386.

7. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, et al. (2008) Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Research 36: 2230–2239.

8. Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. Environmental Microbiology 6: 938–947.

9. Chan C, Hsu A, Tang S, Halgamuge S (2008) Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. Journal of Biomedicine and Biotechnology 2008.

10. Chatterji S, Yamazaki I, Bai Z, Eisen J (2008) CompostBin: a DNA composition-based algorithm for binning environmental shotgun reads. In: Research in Computational Molecular Biology, Berlin, Heidelberg: Springer Berlin / Heidelberg, volume 4955 of *Lecture Notes in Computer Science*, chapter 3. pp. 17–28. doi:10.1007/978-3-540-78839-3_3. URL `http://dx.doi.org/10.1007/978-3-540-78839-3\_3`.

11. Kislyuk A, Bhatnagar S, Dushoff J, Weitz JS (2009) Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinformatics 10: 316.

12. Mohammed MH, Ghosh TS, Singh NK, Mande SS (2011) SPHINX–an algorithm for taxonomic binning of metagenomic sequences. Bioinformatics 27: 22–30.

13. Yang B, Peng Y, Leung HCM, Yiu SM, Qin J, et al. (2010) MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation. In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology. New York, NY, USA: ACM, BCB '10, pp. 170–179. doi:10.1145/1854776.1854803. URL `http://doi.acm.org/10.1145/1854776.1854803`.

14. Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, et al. (2011) Taxonomic metagenome sequence assignment with structured output models. Nature Methods 8: 191–192.

15. Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics 10: 56.

16. Nalbantoglu O, Way S, Hinrichs S, Sayood K (2011) RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC Bioinformatics 12: 41.

17. Silva GGZ, Cuevas DA, Dutilh BE, Edwards RA (2014) FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. PeerJ 2: e425.

18. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, et al. (2014) Binning metagenomic contigs by coverage and composition. Nat Meth 11: 1144–1146.

19. Lu YY, Chen T, Fuhrman JA, Sun F (2016) COCACOLA: binning metagenomic contigs using sequence composition, read coverage, co-alignment, and paired-end read linkage. Bioinformatics .

20. Bohlin J (2011) Genomic signatures in microbes – properties and applications. TheScientificWorldJournal 11: 715–725.

21. Song K, Ren J, Reinert G, Deng M, Waterman MS, et al. (2014) New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Briefings in Bioinformatics 15: 343-353.

22. van Passel MW, Kuramae EE, Luyf AC, Bart A, Boekhout T (2006) The reach of the genome signature in prokaryotes. BMC Evolutionary Biology 6: 84.

23. Bohlin J, Skjerve E, Ussery D (2009) Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering. BMC Genomics 10: 487.

24. Mrázek J (2009) Phylogenetic signals in DNA composition: limitations and prospects. Molecular Biology and Evolution 26: 1163–1169.

25. Wang Y, Leung HCM, Yiu SM, Chin FYL (2012) MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. Journal of Computational Biology 19: 241–249.

26. Girotto S, Pizzi C, Comin M (2016) MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures. Bioinformatics 32: i567-i575.

27. Saeed I, Halgamuge SK (2009) The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments. BMC Genomics 10: S10.

28. Ding X, Cao CC, Sun X (2014) Intrinsic correlation of oligonucleotides: A novel genomic signature for metagenome analysis. Journal of Theoretical Biology 353: 9–18.

29. Prabhu VV (1993) Symmetry observations in long nucleotide sequences. Nucleic Acids Research 21: 2797–2800.

30. Rudner R, Karkas JD, Chargaff E (1968) Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proceedings of the National Academy of Sciences of the United States of America 60: 921–922.

31. Shporer S, Chor B, Rosset S, Horn D (2016) Inversion symmetry of DNA k-mer counts: validity and deviations. BMC Genomics 17: 1–13.

32. Mitchell D, Bridge R (2006) A test of Chargaff's second rule. Biochemical and Biophysical Research Communications 340: 90–94.

33. Gori F, Mavroeidis D, Jetten MSM, Marchiori E (2011) Genomic signatures for metagenomic data analysis: exploiting the reverse complementarity of tetranucleotides. In: Proceedings of the 5th IEEE International Conference on Systems Biology (ISB). pp. 149–154. doi:10.1109/ISB.2011.6033147. URL `http://www.cs.ru.nl/~gori/papers/Gori_signature_metagenomics_ISB_2011.pdf`.

34. Beleza Yamagishi ME, Herai RH (2011) Chargaff's "Grammar of Biology": new fractal-like rules. ArXiv e-prints .

35. Manning CD, Raghavan P, Schtze H (2008) Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press.

36. Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, et al. (2000) Feature selection for SVMs. In: Advances in Neural Information Processing Systems 13. MIT Press, pp. 668–674.

37. Sboner A, Mu X, Greenbaum D, Auerbach R, Gerstein M (2011) The real cost of sequencing: higher than you think! Genome Biology 12: 125.

38. Barral PJ, Cantini L, Hasmy A, Jiménez J, Marcano A (2005) Correlation between strand asymmetry and phylogeny in mitochondrial DNA. Journal of Theoretical Biology 236: 422–426.

39. Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Research 13: 145–158.

40. Lamprea-Burgunder E, Ludin P, Mäser P (2011) Species-specific typing of DNA based on palindrome frequency patterns. DNA Research 18: 117–124.

41. Xia X (2012) DNA replication and strand asymmetry in prokaryotic and mitochondrial genomes. Current Genomics 13: 16-27.

42. Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 37: 145–151.

43. Karlin S, Campbell AM, Mrzek J (1998) Comparative DNA analysis across diverse genomes. Annual Review of Genetics 32: 185–225.

44. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2009) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 37: D5–D15.

45. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423.

46. Jones E, Oliphant T, Peterson P, et al. (2001–). SciPy: Open source scientific tools for Python. URL `http://www.scipy.org/`.

47. Pérez F, Granger BE (2007) IPython: a System for Interactive Scientific Computing. Comput Sci Eng 9: 21-29.

48. Hunter JD (2007) Matplotlib: A 2D graphics environment. Computing In Science & Engineering 9: 90–95.

49. Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. New York, NY, USA: ACM, ICML '06, pp. 233–240. doi:10.1145/1143844.1143874. URL `http://dx.doi.org/10.1145/1143844.1143874`.

50. Nadarajah S, Kotz S (2008) Exact distribution of the max/min of two Gaussian random variables. IEEE Transactions on very large scale integration (VLSI) systems 16: 210–212.

51. Rapoport AE, Trifonov EN (2013) Compensatory nature of Chargaff's second parity rule. Journal of Biomolecular Structure and Dynamics 31: 1324-1336.

52. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT (2012) Individual genome assembly from complex community short-read metagenomic datasets. The ISME Journal 6: 898–901.