Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 513701, 10 pages
doi:10.1155/2008/513701

Methodology Report

Using Growing Self-Organising Maps to Improve the Binning Process in Environmental Whole-Genome Shotgun Sequencing

Chon-Kit Kenneth Chan,[1] Arthur L. Hsu,[1] Sen-Lin Tang,[2] and Saman K. Halgamuge[1]

[1] *Dynamic Systems & Control Group, Department of Mechanical Engineering, University of Melbourne, VIC 3010, Australia*
[2] *Research Center for Biodiversity, Academia Sinica, Taipei 115, Taiwan*

Correspondence should be addressed to Chon-Kit Kenneth Chan, c.chan22@pgrad.unimelb.edu.au

Received 31 August 2007; Accepted 18 November 2007

Recommended by Daniel Howard

Metagenomic projects using whole-genome shotgun (WGS) sequencing produces many unassembled DNA sequences and small contigs. The step of clustering these sequences, based on biological and molecular features, is called binning. A reported strategy for binning that combines oligonucleotide frequency and self-organising maps (SOM) shows high potential. We improve this strategy by identifying suitable training features, implementing a better clustering algorithm, and defining quantitative measures for assessing results. We investigated the suitability of each of di-, tri-, tetra-, and pentanucleotide frequencies. The results show that dinucleotide frequency is not a sufficiently strong signature for binning 10 kb long DNA sequences, compared to the other three. Furthermore, we observed that increased order of oligonucleotide frequency may deteriorate the assignment result in some cases, which indicates the possible existence of optimal species-specific oligonucleotide frequency. We replaced SOM with growing self-organising map (GSOM) where comparable results are obtained while gaining 7%–15% speed improvement.

1. INTRODUCTION

Metagenomics is an emerging area of genome research that allows culture-independent, functional, and sequence-based studies of microbial communities in environmental samples. Whole-genome shotgun (WGS) sequencing has been applied to most of the metagenomic projects [1–6]. These projects have unveiled remarkable information on microbial genomics and also brought an unprecedentedly comprehensive and clearer picture of microbial communities. In the WGS sequencing approach, random sampling of DNA fragments of all microbes that form a community in an environmental sample is performed. The individual DNA fragments are sequenced and then assembled into genomes by using computing techniques. However, a fundamental limit of WGS sequencing is that only the genomes of high-abundance species can be completely or near-completely assembled [7] due to the requirement of multiple overlapping fragments for a confident assembly. In one of the prominent metagenomic studies conducted by Venter et al. [2], about 1 Gb of DNA sequences has been successfully sequenced from Sargasso Sea samples. This study has clearly indicated the existence of far more diverse microbial communities than previously thought. Most of the environmental genomes sequenced to date contain only few high-abundance species but many low-abundance species in the communities that account for a large portion of the total genome size of an environmental sample. The presence of large amount of DNA fragments from the low-abundance species poses a problem for assembling genomes. In order to infer the biological functions of a microbial community from sequences, a process named "binning" is used to group these unassembled DNA sequence fragments and small contigs into biologically meaningful "bins," such as phylogenetic groups [8].

There are a number of tools currently available for the binning process. These include Chisel System [9, 10], MetaClust [4, 11], TETRA [12, 13], PhyloPythia [14], and the combination of oligonucleotide frequency and SOM [15]. The Chisel System helps binning the sequences according to the identification, characterisation, and comparative analysis

of taxonomic and evolutionary variations of enzymes. The remaining above-listed tools use the method of analysing nucleotide composition of sequences that is considered to have the potential of working well for the binning process in WGS sequencing [8]. MetaClust computes different DNA signatures followed by the use of a clustering algorithm to assign sequences into bins. TETRA bins the species-specific sequences by the use of tetranucleotide-derived z-score correlations. PhyloPythia uses a supervised-learning approach where it trains a multiclass support vector machine (SVM) classifier using all the known genome sequences in the existing database then assigns the unknown environmental sequences to the closest clade in the selected taxonomic level. This method has been demonstrated to be able to classify most DNA sequence fragments with high accuracy. However, considering the current amount of known genomes which is far less than 1% of the entire microbial genomes [7], it is reasonable to assume that the currently available training data is insufficient to represent all the extremely diverse microbial genomes for supervised-learning methods. Unsupervised-learning may provide the answer to this problem. The combination of oligonucleotide frequency and the well-known unsupervised learning method self-organising map (SOM) was used by Abe et al. [15] to explore genome signatures. They used the di-, tri-, and tetranucleotide frequencies as the training features of SOM to cluster the 1 kb and 10 kb DNA sequence fragments derived from 65 bacteria and 6 eukaryotes. Clear species-specific separations of sequences were obtained in the 10 kb fragment tests. Their results showed that the combination of oligonucleotide frequency and SOM can be used as a powerful tool to cluster or bin the DNA sequence fragments after WGS sequencing.

In order to successfully bin the DNA sequence fragments, using an appropriate genome signature as the training feature is important. In recent years, researchers have found that, due to oligonucleotide frequency bias in various prokaryotic genomes, the oligonucleotide frequency can be used as a possible genome signature. The di-, tri-, and tetranucleotide frequencies, which are the frequencies using two, three, and four nucleotides respectively, have been well studied. Karlin et al. [16, 17] has shown the compositional bias of the di- and tetranucleotide contents of 15 prokaryotic genomes. Weinel et al. [18] found that 80% of *Pseudomonas putida* KT2440 genome have a similar bias in GC contents and di- and tetranucleotide contents. Teeling et al. [12] showed that the tetranucleotide frequency has a higher discriminatory power than GC content and used it for the assignment of genomic fragments to the taxonomic group. In addition, Sandberg et al. [19] employed a Bayesian approach to classify the short sequences and found that the classification accuracy increases with a higher-order oligonucleotide frequency. Above-mentioned papers provide evidences that there is a trend of using oligonucleotide frequency as prokaryotic genome signature, rather than the GC content. Thus, high-order oligonucleotide frequency may also be an appropriate training feature for binning DNA sequence fragments by unsupervised clustering methods.

Since the combination of oligonucleotide frequency and SOM appears as a promising binning strategy that can be further explored, we focus in this paper on improving the training features and the clustering algorithm. In Abe et al.'s work [15], there was no systematic way of comparing the quality of the SOM results. We tested the traditional clustering evaluation measures (recall, precision, and F-measure) and discovered the inadequacy of using them for examining the similarity of phylogenetic levels. Therefore, we introduce a method to quantitatively measure and assess the results of clustering DNA sequence fragments from a collection of species. In the investigation of evaluating suitable training features, we attempt to compare results for the di-, tri-, and tetranucleotide frequencies as well as the pentanucleotide frequency (the frequency usage of five nucleotides) to test if higher-order oligonucleotide frequency yields better binning of DNA sequence fragments. We also study the effectiveness and efficiency of the combination of oligonucleotide frequency and SOM by employing alternative clustering algorithms. We compare SOM with a variant of it called growing self-organising map (GSOM), which has been successfully applied in several different applications [20–25] including microarray clustering [26]. These comparisons allow us to suggest a better compositional binning strategy for WGS sequencing using the method of combining oligonucleotide frequency and SOM-based clustering algorithm.

This paper is organized as follows: Section 2.1 gives a brief introduction to the SOM and GSOM clustering algorithms; Section 2.2 proposes a method of measuring inseparable species when DNA sequence fragments are clustered; Section 2.3 describes the procedures of preparing the three datasets used in this paper, and the data preprocessing step for preparing the input vectors; and Section 2.4 shows the details of the algorithm settings and the experiment set up for repeatability of the experiments. Section 3 presents the results of comparing the four orders of oligonucleotide frequencies and the comparison between SOM and GSOM; Finally, Section 4 gives the discussion, conclusion, and future work.

2. METHODS

2.1. Growing self-organising map

Growing self-organising map (GSOM) [27, 28] is an extension of self-organising map (SOM) [29]. GSOM is a dynamic SOM which overcomes the weakness of a static map structure of SOM. Both SOM and GSOM are used for clustering high-dimensional data. This is achieved by projecting the high-dimensional data onto a two- or three-dimensional feature map with lattice structure where every point of interest in the lattice represents a neuron or a node in the map. The mapping preserves the data topology, so that similar samples can be found close to each other on the 2D/3D feature map.

The SOM training consists of three phases: initialisation phase, ordering phase, and fine-tuning phase. The initialisation is crucial to achieve a quality-clustering result. The following parameters are determined in this phase:

(i) the map topology (either rectangular or hexagonal);
(ii) the number of nodes which is the resolution of the map;

(iii) the weight vector initialization of nodes;
(iv) the width/height (or aspect) ratio of the map.

The user determines the first two parameters and generally principle components analysis (PCA) is used for setting the last two parameters. The weight vectors are initialised by the first two principle vectors of the inputs and the aspect ratio of the map is determined based on the ratio of magnitudes of the first two principle components. In the ordering and fine-tuning phases, each input is presented to the map and the best matching unit or "winner," which has the smallest Euclidean distance to the presented input, is identified. The weight vector of the winner and its neighbouring nodes are updated by

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \alpha \times h \times [\mathbf{x}(k) - \mathbf{w}(t)], \quad (1)$$

where $\mathbf{w}(t)$ is the weight vector of the node at time t, $\mathbf{x}(k)$ is the kth input vector ($\mathbf{w}, \mathbf{x} \in \Re^D$ where D is the dimensionality of data), α is the learning rate and h is the neighbourhood kernel function.

GSOM employs the same weight adaptation and neighbourhood kernel learning as SOM, but has a global parameter of growth named Growth Threshold (GT) that controls the resolution of the map. The Growth Threshold is defined as

$$\mathrm{GT} = -D \times \ln(SF), \quad (2)$$

where $SF \in [0, 1]$ is the user defined spread factor with 0 representing minimum growth (coarsest resolution) and 1 representing maximum growth (finest resolution).

There are three phases in the GSOM training: initialisation phase, growing phase, and smoothing phase. In the initialisation phase, the GSOM is initialised with a minimum single "lattice grid" depending on whether the rectangular or hexagonal topology is chosen. Due to the small number of nodes in the beginning of training, the weight vector initialisation has less effect on the clustering quality and these weights will be corrected quickly in the growing phase. During the growing phase, every node has an accumulated error counter and the counter of the winner (E_{winner}) is updated by

$$E_{\text{winner}}(t+1) = E_{\text{winner}}(t) + ||\mathbf{x}(k) - \mathbf{w}_{\text{winner}}(t)||. \quad (3)$$

If the winner is at the boundary of the current map and E_{winner} exceeds GT, new nodes will be added to the surrounding vacant slots of the winner. In the case when E_{winner} exceeds GT and the winner is not a boundary node, E_{winner} is evenly distributed outwards to the winner's neighbouring nodes. The smoothing phase is for fine-tuning the weights of nodes and no new node will be added to the map.

The major advantages of GSOM over SOM are summarised as follows.

(i) The shape of GSOM represents the hidden data structure better than SOM that leads to better identifiable clusters.
(ii) New nodes are added to the necessary regions while keeping the order of nodes. Therefore, neither PCA nor ordering phase is required in the training.
(iii) Fewer nodes at the beginning of the training leads to the speed improvement.

2.2. Quality measurement of the clustering performance in the mixing region

In our preliminary test, the well-known F-measure [30], which computes both recall and precision into a single index from a contingency table, was used to evaluate the clustering results. However, after examining the cluster contents, it is apparent that, for binning applications, F-measure does not provide sufficient insight and description of ambiguities in terms of phylogenetic relationships (refer to Section 3). More specifically, one would expect phylogenetically-close groups as highly likely to be ambiguous, but F-measure does not account for such likelihoods. Therefore, we propose an alternative clustering evaluation measure specifically for this application.

When an SOM or GSOM is used to group species fragments into clusters on a 2D/3D map, it is often inevitable that regions with overlapping clusters (mixing regions where a neuron represents DNA sequence fragments from more than one species) will exist. To evaluate a clustering algorithm's ability to group DNA sequence fragments into species-specific or "pure" clusters, we define two criteria that measure the clustering quality in the mixing region: intensity of mix (IoM) and level of mix (LoM), where the former measures the percentage of mixing and the later indicates the taxonomic level of ambiguity for a given pair of clusters.

The IoM is evaluated based on the concept of mixed pair described below. Let A and B be sets of vectors belonging to species A and B, respectively, and $n(X)$ is the number of elements in set X. If A and B is a mixed pair, then the percentage of A in the mixing region of the two classes is $n(A \cap B \mid A)/n(A)$ and the percentage of B is $n(A \cap B \mid B)/n(B)$. As illustrated in Figure 1, 11.6% of A sequences is mixed with B sequences in the B cluster and the complementary mix indicates that 10.6% of B sequences is mixed with A sequences in the A cluster. The same concept of mixed pair applies for B and C. Therefore, there are two mixed pairs in Figure 1, one is A and B and the other is B and C. For k number of species, there can be up to $k(k-1)/2$ mixed pairs. Additionally, a pair of clusters is only considered to be truly mixed when both clusters are heavily overlapped. Thus, as in Figure 1, when $n((B \cap C \mid B)/n(B)) > \mathrm{THRESHOLD\ (TH)}$ but $n(B \cap C \mid C)/n(C) < \mathrm{TH}$, it indicates that only a small number of outliers of one species (C) is mixed with the other species (B). Therefore, this mixed pair is not considered as truly mixed. We use TH = 5% for the threshold of being truly mixed meaning that, statistically, we have a nonmixing confidence of 95%. The IoM measures the amount of mixing sequences and it is nonlinearly categorised into five levels: low (L) 5%–10%, medium low (ML) 10%–20%, medium (M) 20%–40%, medium high (MH) 40%–60%, and high (H) 60%–100%. For example, the IoM is ML for the truly mixed pair A and B in Figure 1.

To evaluate clustering results of species, we use LoM to describe the taxonomic level of the mixed species. For example, as in Figure 1, *Bacillus subtilis* is classified in Kingdom *Bacteria* and Phylum *Firmicutes. Acinetobacter* is classified in Kingdom *Bacteria* but Phylum *Proteobacteria.* Then the two species are mixed at the Phylum level. Because of the

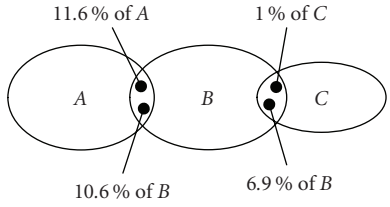


Figure 1: Concept of mixed pair: the mixed pair between A and B is truly mixed (IoM = ML and LoM = Phylum). The mixed pair between B and C is not truly mixed because $n(B \cap C \mid C)/n(C) < 5\%$.

evolution of organisms, nucleotide composition of genomes belonging to the same lower taxonomic levels can be very similar. Clustering organisms at higher level of taxonomy should be easier than at lower level of taxonomy. Therefore, if truly mixed pair occurs, lower LoM (e.g., Species) is more acceptable and more desirable than higher LoM (e.g., Kingdom).

In summary, the proposed two measures are defined as

$$\begin{aligned} &\text{IoM} \in \{\text{L}, \text{ML}, \text{M}, \text{MH}, \text{H}\}, \\ &\text{LoM} \in \{\text{Species}, \text{Genus}, \text{Family}, \text{Order}, \text{Class}, \\ &\qquad \text{Phylum}, \text{Kingdom}\}. \end{aligned} \tag{4}$$

The two proposed measures, IoM and LoM, are only defined for truly mixed pairs to evaluate the clustering quality in the mixing regions of a map by the following steps.

(i) Find truly mixed pairs for all pairs of species where if $n(X \cap Y \mid Y)/n(Y) \geq \text{TH}$ and $n(X \cap Y \mid X)/n(X) \geq \text{TH}$, then X and Y is a truly mixed pair.
(ii) If X and Y are truly mixed, determine IoM according to $\min\{n(X \cap Y \mid Y)/n(Y), n(X \cap Y \mid X)/n(X)\}$.
(iii) Identify LoM of X and Y.

Clustering results can now be assessed based on three criteria: number of truly mixed pairs, IoM, and LoM. However, which criterion should have higher priority may vary between applications. Therefore, in our assessment, one result is better than another only when it is superior on at least two of the three measures.

2.3. Dataset preparation and data preprocessing

The NCBI database (http://www.ncbi.nlm.nih.gov) contains 370 completed microbial genomes in early 2006, which includes 28 Archaea and 342 Bacteria. As the investigation seeks to cluster sequences of species, genomes are filtered so that there is no duplicating species. One strain was arbitrarily chosen if the same species genome contains more than one strain. After this process, there were 283 genomes remaining. Considering the available computing resources and the algorithm comparison focus of this paper, two artificial sets of prokaryotic DNA sequences (each of 10 different species out of the 283 species) were randomly sampled from the NCBI database.

In addition, three simulated metagenomic datasets were created by Mavromatis et al. [31] to facilitate benchmarking of metagenomic data processing methods, which include, but not limited to, binning methods. The three datasets vary in relative abundance and number of species that represent different complexity levels of real-world microbial communities. The sequence fragments in the simulated datasets were assembled using three commonly used sequence assembling programs: JAZZ, Arachne, and Phrap at U.S. Department of Energy (Wash, USA), DOE Joint Genome Institute (Calif, USA). In this paper, we tested one of the three simulated datasets named simMC and was assembled by Phrap (http://www.phrap.com). For simplicity, this dataset will be represented as simMC_Phrap throughout the paper.

The taxonomic distributions of the three sets of species are displayed graphically in Figure 3. Each letter represents a single species and species within a single rectangle have the same taxonomy at the specific level. The names of the species can be found in Section 1 of the supplementary material which consists of 6 sections showing the species names in Section 1, clustering evaluation methods in Sections 2 and 3, and the labelled cluster maps in Section 4 to 6 (http://www.mame.mu.oz.au/~ckkc/Binning). The numbers below the taxonomic levels in Figure 3 indicate the maximum possible number of mixed pairs at that taxonomic level. For example, in Figure 3(a), the maximum number of mixed pairs at taxonomic level of *Class* is 12, which consists a mixed pair each from (a,j) and (c,e) and 5 pairs each from (c,{b,d,g,h,i}) and (e,{b,d,g,h,i}).

In the experiments of Abe et al. [15] that attempts to separate 1 kb and 10 kb DNA sequence fragments of 65 bacteria genomes containing 54 different species, it is visually shown that 1 kb DNA sequence fragments do not carry enough discriminatory information and hence could not completely separate the fragments into species-specific groups. Therefore, a sequence fragment length of 10 kb is used for the analysis in the two artificial datasets to ensure appropriate separation of species-specific groups. Sequences used in the two artificial datasets are produced from complete genome sequences to simulate the environment of WGS sequencing. Such a complete genome sequence is segmented into 10 kb nonoverlapping fragments. A sliding window with the size n is used for counting the oligonucleotide frequency for each of the fragments in which n is the nucleotide length. For example, the dinucleotide frequency ($n = 2$) for a short sequence "AATACTTT" is shown graphically in Figure 2. The oligonucleotide frequency count for each of the fragments yields a single input vector for clustering. The input vectors to the clustering algorithm will have 4^n dimensions.

Whereas, the simMC_Phrap was preprocessed by extracting all sequences with contig length $\geq$ 8 kb. The oligonucleotide frequency count is applied to these sequences to generate the input vectors for clustering. Finally, each input vector is normalised by the sequence length.

2.4. Algorithms parameters and experiment details

In order to avoid the algorithm implementation bias, an in-house clustering program was developed consisting a

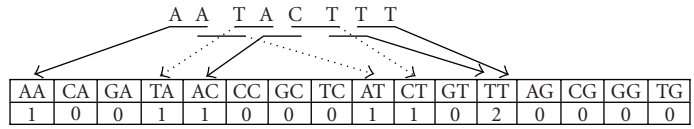


Figure 2: Dinucleotide frequency counting for the short sequence "AATACTTT."

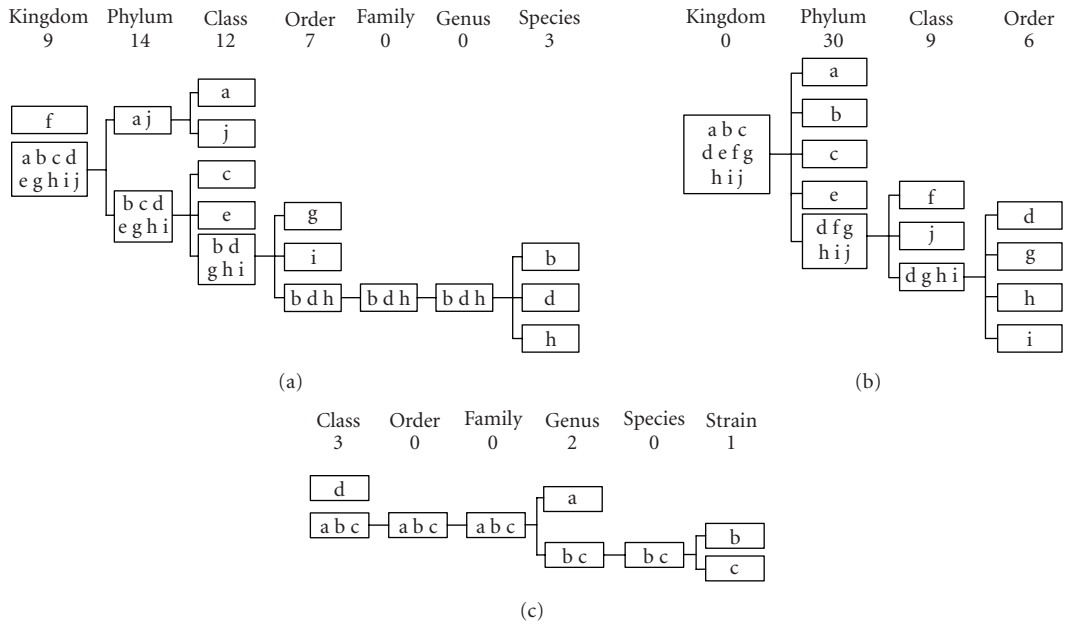


Figure 3: The taxonomy distribution of the 10 species in (a) Set 1, (b) Set 2, and the 4 species in (c) simMC_Phrap. Each letter represents a single species. The numbers below the taxonomic levels indicate the maximum number of mixed pairs at that taxonomic level. For example, in (a), the maximum number of mixed pairs at taxonomic level of Class is 12, which consists (a,j), (c,e), (c,{b,d,g,h,i}), and (e,{b,d,g,h,i}) mixed pairs.

Table 1: Training parameters used for the SOM and GSOM training.

Training parameter	Phase 2	Phase 3
Learning length	15 epochs	70 epochs
Learning rate	0.1	0.05
Neighbourhood size	3	1

data preprocessor, both SOM and GSOM algorithms, and a graphical neuron-map output. The program is written in C# with a Windows form interface. This program can be requested from the author for academic use. Since the hexagonal topology has better topology preservation [26, 29], it is used in both SOM and GSOM. In addition, the tests were conducted on a Pentium 4 3.2 GHz desktop PC running Windows XP and the same parameter settings are used in both algorithms for a fair computational speed comparison (as listed in Table 1).

To compare results from the 4 orders of oligonucleotide frequencies, we obtain similar map resolution (number of nodes) for both algorithms and for all nucleotide frequencies. Since the GSOM algorithm automatically determines the number of nodes, it can be used to determine the total number of nodes of SOM. This can be achieved by training GSOM with a specified resolution (we used SF = 0.4) for the dinucleotide frequency then the final number of nodes in GSOM is used to set the number of nodes in SOM, as well as determine the SF for other nucleotide frequencies. Using this scheme, we set SF = 0.4 for dinucleotide frequency (for a higher-resolution map) and experimentally determined that SF = 0.6 for trinucleotide frequency, SF = 0.8 for tetranucleotide frequency and SF = 0.9 for pentanucleotide frequency will result in similar map resolution. The SOM also requires setting the aspect ratio of the map and initializing the weight vectors. These two parameters are set by using PCA. The schematic diagram of this approach is shown in Figure 4.

3. RESULTS

The proposed binning method is tested on two artificial datasets and a simulated metagenomic dataset (simMC_Phrap) which was created and published for benchmarking the metagenomic data processing methods. The two artificial datasets of prokaryotic DNA sequences (each of 10 different species out of the 283 species) were randomly sampled from the NCBI database. Each set of the genome

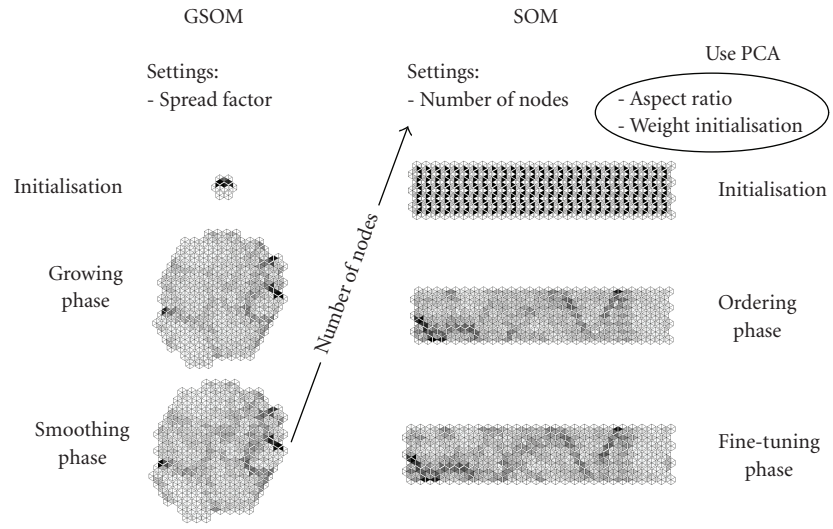


Figure 4: Illustration of method used to compare SOM and GSOM.

Table 2: The evaluation of clustering results using F-measure.

	Set 1		Set 2		simMC_Phrap	
	SOM	GSOM	SOM	GSOM	SOM	GSOM
Di	0.95	0.95	0.94	0.94	0.92	0.91
Tri	0.97	0.97	0.98	0.98	0.90	0.90
Tetra	0.97	0.97	0.98	0.98	0.92	0.90
Penta	0.97	0.97	0.99	0.99	0.89	0.89

sequences was preprocessed to obtain the 4 orders of oligonucleotide frequencies (di-, tri-, tetra-, and pentanucleotide frequencies) to form 4 datasets. This data preprocessing involves segmenting each genome sequence in the set into 10 kb lengths then produce input vectors by calculating the specific oligonucleotide frequency. After preprocessing, each of the 4 datasets from species Set 1 contains 4,145 input vectors. Whereas, each of the 4 datasets from Set 2 contains 2,398 input vectors. The simMC_Phrap was preprocessed by extracting all sequences with contig length $\geq$ 8 kb then obtaining the 4 orders of oligonucleotide frequencies to form 4 datasets. The produced input vectors were normalised by the sequence length. After preprocessing, each of the 4 datasets contains 401 input vectors. The details of preparing these three datasets can be found in Section 2.

To evaluate the clustering performance, we used well-known clustering evaluation measure F-measure. The results for the three datasets are shown in Table 2. A summary of F-measure calculation can be found in Section 2 of the supplementary material.

From these results, we can observe that F-measure does not distinguish the clustering quality clearly enough for this species separation application. For example, in the results of Set 1 using GSOM, F-measure equals 0.97 for the tri-, tetra-, and pentanucleotide frequencies. However, by using the proposed evaluation, more details of the ambiguities can be seen. It shows that there are only two mixed pairs with low taxonomic level in the mixing region when using the pentanucleotide frequency. However, there are four mixed pairs with two of them having higher taxonomic levels when using the tri- and tetranucleotide frequencies (as shown in Table 3). This suggests that the pentanucleotide frequency provides a higher level of phylogenetic resolution, which cannot be detected via F-measure because the numbers of incorrectly assigned sequence fragments are similar for these three nucleotide frequencies.

We use two approaches to evaluate the performance of clustering DNA sequence fragments of species. The first approach is to observe the cluster formation of species sequences to verify the cluster formation similar to the method used by Abe et al. [15]. The second approach is to compare the LoM and IoM in the mixing region. A simple example is given in Section 3 of the supplementary material. It highlights the difference between the calculation of F-measure and IoM.

After the training is completed, for display purpose, we use the label information of the input data to display the labelled cluster map. The labelled cluster maps from the training of SOM and GSOM for the pentanucleotide frequency of species Set 1 are shown in Figure 5. The following points can be observed from this labelled cluster maps.

(i) All species are clearly clustered and are marked in the figures.

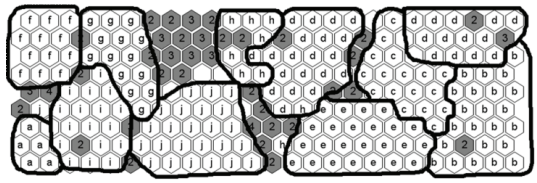
(ii) The nodes that contain more than 2 species (which are coloured in grey) are mostly located at the border of the clusters.

(iii) Species "d" is clustered as one group in GSOM, but separated by species "c" into two groups in SOM.

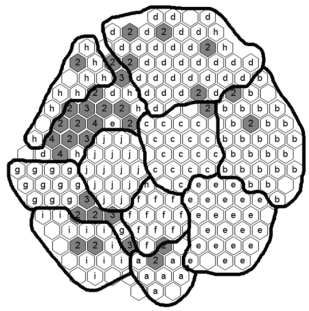
These observations show that the GSOM have better cluster formation in terms of cluster identification than the SOM due to the flexibility of feature map shape. The labelled cluster maps for other datasets also show a clear cluster

Table 3: Training results in the mixing regions for species Set 1.

Algorithm	SOM				GSOM			
Nucleotide Freq.	Di	Tri	Tetra	Penta	Di	Tri	Tetra	Penta
Kingdom	—	—	—	—	—	—	—	—
Phylum	—	—	—	—	—	—	—	—
Class	ML, ML, L	—	—	—	ML, ML, L	—	—	—
Order	ML, ML	ML, L	ML	L	M, L	ML, L	L, L	—
Family	—	—	—	—	—	—	—	—
Genus	—	—	—	—	—	—	—	—
Species	M, L	M, L	ML	ML	M, L	M, L	ML, L	ML, L



(a)



(b)

Figure 5: The labelled cluster maps for clustering species Set 1 by (a) SOM, (b) GSOM with the pentanucleotide frequency. Each hexagon represents a single node. If a node contains input samples from only a single species, it is displayed with a letter that uniquely identifies the species. Grey colour nodes correspond to two or more species in the node and the number of species is displayed on the node. A node without label means that there is no input sample "hits."

formation and can be found in Sections 4, 5, and 6 of the supplementary material.

We also interpret the clustering results of the species mixing regions by summarising the IoM and LoM in Tables 3, 4, and 5 for Set 1, Set 2, and simMC_Phrap, respectively.

By comparing the results of the four orders of oligonucleotide frequencies for Set 1, the number of truly mixed pairs for dataset that uses the dinucleotide frequency is almost twice the number of truly mixed pairs of the higher-order oligonucleotide frequencies as shown in Table 3. In addition, the LoM is also high for the dinucleotide frequency. Similarly in Table 4, there are three to four truly mixed pairs when using the dinucleotide frequency, but no more than one truly mixed pair when higher-order oligonucleotide frequencies are used. All four truly mixed pairs are of a very high LoM at the Phylum level. Since there are only very few species and number of sequence fragments in simMC_Phrap, the difference of using different nucleotide frequencies is not obvious. Nevertheless, the results of our two artificial sets indicate that dinucleotide frequency is not a strong signature for clustering the 10 kb fragments of species.

Furthermore, from Set 1, we can see that IoM and LoM tend to decrease as the order of oligonucleotide frequency increases. One would naturally suspect that higher-order oligonucleotide frequencies may carry more information so they can be used to achieve better clustering results. However, it is not the case for Set 2 and simMC_Phrap. In Set 2, there is a truly mixed pair in the tetranucleotide frequency but no mixed pair in the tri- and pentanucleotide frequencies when using SOM. For the GSOM algorithm, a truly mixed pair appears in the pentanucleotide frequency but not in the tri- and tetranucleotide frequencies. A detailed examination shows that they are the same pair, *Acinetobacter* sp.ADP1 and *Bacillus subtilis* subsp. *subtilis* str. 168, in both cases. Both of the IoM is low indicating a much better result than the dinucleotide frequency. As in Table 5 for simMC_Phrap, GSOM performs well for all four orders of nucleotide frequencies, whereas SOM shows inconsistent clustering quality for different nucleotide frequencies. There are two mixed pairs in the di- and pentanucleotide frequencies but only one mixed pair in the other two nucleotide frequencies. The mixed pair in the tetranucleotide frequency has the lowest IoM (IoM = M). These results also do not support the hypothesis that higher-order oligonucleotide frequencies are better clustering features. Therefore, we can only conclude that higher-order oligonucleotide frequencies are better features for clustering the species than using dinucleotide frequency. However, the optimal oligonucleotide frequency may vary in different species.

In terms of clustering quality, both SOM and GSOM have similar results. However, besides the mixing quality comparison, we also compare the training speed of them. The speed comparisons for the first two training phases and the overall training time of both algorithms for all 12 datasets are shown in Tables 6 and 7, respectively.

Comparing the time taken for SOM and GSOM to finish the first two training phases (as in Table 6), GSOM has more than 37% speed improvement than SOM. This speed improvement can be explained by considering the initial formation of the map structure. As discussed in Section 2, PCA

Table 4: Training results in the mixing regions for species Set 2.

Algorithm	SOM				GSOM			
Nucleotide Freq.	Di	Tri	Tetra	Penta	Di	Tri	Tetra	Penta
Kingdom	—	—	—	—	—	—	—	—
Phylum	MH, ML, ML	—	L	—	H, ML, L, L	—	—	L
Class	—	—	—	—	—	—	—	—
Order	—	—	—	—	—	—	—	—
Family	—	—	—	—	—	—	—	—
Genus	—	—	—	—	—	—	—	—
Species	—	—	—	—	—	—	—	—

Table 5: Training results in the mixing regions for the contigs ≥ 8 kb from simMC_Phrap.

Algorithm	SOM				GSOM			
Nucleotide Freq.	Di	Tri	Tetra	Penta	Di	Tri	Tetra	Penta
Kingdom	—	—	—	—	—	—	—	—
Phylum	—	—	—	—	—	—	—	—
Class	—	—	—	—	—	—	—	—
Order	—	—	—	—	—	—	—	—
Family	—	—	—	—	—	—	—	—
Genus	MH	MH	M	MH	MH	MH	MH	MH
Species	—	—	—	—	—	—	—	—
Strain	L	—	—	L	—	—	—	—

is used to initialise SOM. However, it increases the computational cost exponentially when the data dimension and size increases. Therefore, even though time consumed for PCA initialisation in this experiment is negligible, it is expected to be significant in large-scale metagenomic analysis. In addition, SOM starts with all nodes fully covering the whole input space then adjusts the weights of all nodes to represent the input data better. On the other hand, GSOM starts with the minimum number of nodes and grows more nodes in the required direction to get an abstract representation of the input data while still correcting the weights of existing nodes. At the end of the second phase, both algorithms will be roughly representing the abstraction of the data. However, GSOM saves time by avoiding PCA calculation and operates on fewer numbers of nodes in the first two phases. Although the last training phase is basically identical for both algorithms and the learning length of the last training phase is much longer than the first two phases, GSOM still has 7%–15% of speed improvement for the overall training. It was reported that the use of this strategy involving SOM was used for analysing a large amount of eukaryote genomes and one of the highest performance supercomputers in the world was required [32]. For the large-scale analysis, which can take weeks to complete, 7%–15% of speed improvement means 1-2 days of time saving for a two-week computation.

Besides using Tables 6 and 7 to compare the training speed of SOM and GSOM, the tables also show that the training time grows exponentially from one order of oligonucleotide frequency to the higher-order oligonucleotide frequency. It is because of the rapid increment of dimensions when the order of oligonucleotide frequency increases.

4. DISCUSSION, CONCLUSION, AND FUTURE WORK

We have investigated four orders of oligonucleotide frequencies: di-, tri-, tetra-, and pentanucleotide frequencies on two artificial sets of species and a published simulated metagenomic dataset. Each of the two artificial sets contains 10 randomly selected species from NCBI database. We noticed that the F-measure can not distinguish the clustering quality for this application. Therefore, two methods have been defined for evaluating the performance of clustering DNA sequence fragments of species. This is done by observing the cluster formation with the labelled cluster map and then qualitatively and quantitatively comparing the LoM and IoM in the mixing region.

The results have shown that dinucleotide frequency is not a sufficiently strong signature for the tested 10 kb DNA sequences on the SOM-based algorithm. Similar to other reports, we also found that higher-order oligonucleotide frequencies, such as tri-, tetra-, and pentanucleotide frequencies, are carrying reasonably adequate genomic information to group intraspecies sequences and separate interspecies sequences [12, 19]; but the required computational power increases exponentially for each increased order of oligonucleotide frequency. Additionally, we noticed that increase of the order of oligonucleotide frequency may deteriorate the assignment of DNA sequence fragments to classes in some cases, which indicates the possible existence of optimal species-specific oligonucleotide frequency. For example, the trinucleotide frequency has a better discrimination power for *Acinetobacter* sp. ADP1 and *Bacillus subtilis* subsp. *subtilis* str. 168 then the tetra- and pentanucleotide frequencies. Therefore, analysts are recommended to start with trinucleotide frequency in large-scale projects and higher-order oligonucleotide frequencies may not always be better.

We also compare the SOM and GSOM algorithms for clustering the DNA sequence fragments of species. Both SOM and GSOM have shown similar results. However, in term of speed comparison, GSOM has more than 37% speed improvement over SOM in the first two training phases and

Table 6: Speed comparisons for the first two training phases of SOM and GSOM, in which the improvement columns represent the percentage of speed improvement for GSOM comparing to SOM.

	Species Set 1			Species Set 2			simMC_Phrap		
	SOM (sec)	GSOM (sec)	Improvement	SOM (sec)	GSOM (sec)	Improvement	SOM (sec)	GSOM (sec)	Improvement
Di	54	34	37%	24	15	38%	2	1	50%
Tri	188	115	39%	74	45	39%	7	4	43%
Tetra	779	475	39%	236	147	38%	31	18	42%
Penta	3031	1847	39%	878	518	41%	144	80	44%

Table 7: Speed comparisons for the overall training time of SOM and GSOM, in which the improvement columns represent the percentage of speed improvement for GSOM comparing to SOM.

	Species Set 1			Species Set 2			simMC_Phrap		
	SOM (sec)	GSOM (sec)	Improvement	SOM (sec)	GSOM (sec)	Improvement	SOM (sec)	GSOM (sec)	Improvement
Di	313	274	12%	133	121	9%	11	10	9%
Tri	1048	942	10%	427	387	9%	39	36	8%
Tetra	4639	3932	15%	1297	1203	7%	173	158	9%
Penta	16839	15709	7%	4702	4387	7%	720	662	8%

7%–15% speed improvement in the overall training. Therefore, GSOM is potentially a better alternative clustering tool. As a result of this study, we would suggest to use GSOM and a higher-order oligonucleotide frequency (at least trinucleotide frequency) to improve the strategy proposed by Abe et al. [15] for the binning process after WGS sequencing.

The method of combining oligonucleotide frequency and the SOM-based algorithm has provided a promising way of binning after WGS sequencing. However, there are limitations with this method. Since SOM-based algorithms are essentially data visualisation techniques, it is difficult to identify the exact cluster boundaries when clusters severely overlap with each other. The overlapping cluster can often be misinterpreted as a single cluster when no label is available. Therefore, a further development to overcome this cluster-overlapping problem is necessary for such SOM-based binning method to be fully practical. Additionally, at the current state, due to the high diversity of microbial communities and the nature of WGS sequencing, most of the unassembled sequences are less than 10 kb. In order to maximize the use of this binning strategy, more investigation on the optimal sequence length will need to be performed in the future work. On the other hand, the rapidly advancing sequencing technology and techniques that are capable of faster sequencing, higher coverage, and longer contig length are continuously being developed [33, 34]. The length of the unassembled fragment is expected to increase in the near future. Therefore, this binning strategy is useful for the analysis after WGS sequencing. Alternatively, when one is attempting to identify a specific species in the metagenome, which has already been sequenced, supervised learning methods can be applied. While PhyloPythia employs SVM as its supervised learning classifier, one can opt for other well-known supervised learning methods that has been used in other various applications [35, 36].

REFERENCES

[1] G. W. Tyson, J. Chapman, P. Hugenholtz, et al., "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature*, vol. 428, no. 6978, pp. 37–43, 2004.

[2] J. C. Venter, K. Remington, J. F. Heidelberg, et al., "Environmental genome shotgun sequencing of the sargasso sea," *Science*, vol. 304, no. 5667, pp. 66–74, 2004.

[3] S. G. Tringe, C. Von Mering, A. Kobayashi, et al., "Comparative metagenomics of microbial communities," *Science*, vol. 308, no. 5721, pp. 554–557, 2005.

[4] T. Woyke, H. Teeling, N. N. Ivanova, et al., "Symbiosis insights through metagenomic analysis of a microbial consortium," *Nature*, vol. 443, no. 7114, pp. 950–955, 2006.

[5] D. B. Rusch, A. L. Halpern, G. Sutton, et al., "The sorcerer *II* global ocean sampling expedition: northwest atlantic through eastern tropical pacific," *PLoS Biology*, vol. 5, no. 3, p. e77, 2007.

[6] S. Yooseph, G. Sutton, D. B. Rusch, et al., "The sorcerer *II* global ocean sampling expedition: expanding the universe of protein families," *PLoS Biology*, vol. 5, no. 3, p. e16, 2007.

[7] K. Chen and L. B. Pachter, "Bioinformatics for whole-genome shotgun sequencing of microbial communities," *PLoS Computational Biology*, vol. 1, no. 2, p. e24, 2005.

[8] J. A. Eisen, "Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes," *PLoS Biology*, vol. 5, no. 3, p. e82, 2007.

[9] A. Rodriguez, Y. Zhang, N. Maltsev, and E. Marland, "Chisel—a framework for identification and characterization of taxonomic and phenotypic versions of enzymes," in *Proceedings of the Institute of Structural Molecular Biology (ISMB '06)*, Fortaleza, Brazil, 2006.

[10] N. Maltsev, M. Syed, A. Rodriguez, B. Gopalan, and F. Brockman, "A novel binning approach and its application to a metagenome from a multiple extreme environment," in *Proceedings of the Joint Genomics: GTL Awardee Workshop V and*

Metabolic Engineering and USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Workshop, North Bethesda, Md, USA, 2007.

[11] M. Huntemann, "MetaClust—entwicklung eines modularen Programms zum Clustern von Metagenomfragmenten anhand verschiedener intrinsischer DNA-Signaturen," Diploma thesis, University of Bremen, Germany, 2006.

[12] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glöckner, "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environmental Microbiology*, vol. 6, no. 9, pp. 938–947, 2004.

[13] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics*, vol. 5, no. 163, 2004.

[14] A. C. McHardy, H. G. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos, "Accurate phylogenetic classification of variable-length DNA fragments," *Nature Methods*, vol. 4, no. 1, pp. 63–72, 2007.

[15] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, "Informatics for unveiling hidden genome signatures," *Genome Research*, vol. 13, no. 4, pp. 693–702, 2003.

[16] S. Karlin, J. Mrázek, and A. M. Campbell, "Compositional biases of bacterial genomes and evolutionary implications," *Journal of Bacteriology*, vol. 179, no. 12, pp. 3899–3913, 1997.

[17] S. Karlin, "Global dinucleotide signatures and analysis of genomic heterogeneity," *Current Opinion in Microbiology*, vol. 1, no. 5, pp. 598–610, 1998.

[18] C. Weinel, K. E. Nelson, and B. Tümmler, "Global features of the Pseudomonas putida KT2440 genome sequence," *Environmental Microbiology*, vol. 4, no. 12, pp. 809–818, 2002.

[19] R. Sandberg, G. Winberg, C.-I. Bränden, A. Kaske, I. Ernberg, and J. Cöster, "Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier," *Genome Research*, vol. 11, no. 8, pp. 1404–1409, 2001.

[20] Y. Z. Zhai, A. Hsu, and S. K. Halgamuge, "Scalable dynamic self-organising maps for mining massive textual data," in *Proceedings of the Lecture Notes in Computer Science (including subseries: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4234, LNCS III, pp. 260–267, Springer, Berlin, Germany, 2006.

[21] R. Amarasiri and D. Alahakoon, "Applying dynamic self organizing maps for identifying changes in data sequences," in *Design and Application of Hybrid Intelligent Systems*, pp. 682–691, IOS Press, Amsterdam, The Netherlands, 2003.

[22] S. Chen, D. Alahakoon, and M. Indrawan, "Background knowledge driven ontology discovery," in *Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service, (EEE '05)*, pp. 202–207, 2005.

[23] H. Wang, F. Azuaje, and N. Black, "Improving biomolecular pattern discovery and visualization with hybrid self-adaptive networks," *IEEE Transactions on Nanobioscience*, vol. 1, no. 4, pp. 146–166, 2002.

[24] H. Wang, F. Azuaje, and N. Black, "Interactive GSOM-Based approaches for improving biomedical pattern discovery and visualization," in *Computational and Information Science*, vol. 3314 of *Lecture Notes in Computer Science*, pp. 556–561, Springer, Berlin, Germany, 2004.

[25] M. A. Karim, S. Halgamuge, A. J. R. Smith, and A. L. Hsu, "Manufacturing yield improvement by clustering," in *Neural Information Processing*, vol. 4234 of *Lecture Notes in Computer Science*, pp. 526–534, Springer, Berlin, Germany, 2006.

[26] A. L. Hsu, S.-L. Tang, and S. K. Halgamuge, "An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data," *Bioinformatics*, vol. 19, no. 16, pp. 2131–2140, 2003.

[27] A. L. Hsu and S. K. Halgamuge, "Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualisation," *International Journal of Approximate Reasoning*, vol. 32, no. 2-3, pp. 259–279, 2003.

[28] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, "Dynamic self-organizing maps with controlled growth for knowledge discovery," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 601–614, 2000.

[29] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, Germany, 2nd edition, 1997.

[30] C. J. van Rijsbergen, *Information Retrieval*, Butterworths, London, UK, 2nd edition, 1979.

[31] K. Mavromatis, N. Ivanova, K. Barry, et al., "Use of simulated data sets to evaluate the fidelity of metagenomic processing methods," *Nature Methods*, vol. 4, no. 6, pp. 495–500, 2007.

[32] T. Abe, H. Sugawara, S. Kanaya, M. Kinouchi, and T. Ikemura, "Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes," *Gene*, vol. 365, no. 1-2, pp. 27–34, 2006.

[33] T. Jarvie, L. Du, and J. Knight, "Shotgun sequencing and assembly of microbial genomes: comparing 454 and Sanger methods," *Biochemica*, pp. 11–14, 2005.

[34] L. Bonetta, "Genome sequencing in the fast lane," *Nature Methods*, vol. 3, no. 2, pp. 141–146, 2006.

[35] S. K. Halgamuge and M. Glesner, "Fuzzy neural networks: between functional equivalence and applicability," *International Journal of Neural Systems*, vol. 6, no. 2, pp. 185–196, 1995.

[36] S. K. Halgamuge, "Self-evolving neural networks for rule-based data processing," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2766–2773, 1997.