Local Binary Patterns as a Feature Descriptor in Alignment-Free Visualisation of Metagenomic Data

Samaneh Kouchaki*, Santosh Tirunagari†, Avraam Tapinos *, David L Robertson*
*Evolutionary and Genomic Sciences, Faculty of Biology, Medicine and Health, The University of Manchester
Email: samaneh.kouchaki@manchester.ac.uk
†Department of Computer Science, University of Surrey

***Abstract*—Shotgun sequencing has facilitated the analysis of complex microbial communities. However, clustering and visualising these communities without prior taxonomic information is a major challenge. Feature descriptor methods can be utilised to extract these taxonomic relations from the data. Here, we present a novel approach consisting of local binary patterns (LBP) coupled with randomised singular value decomposition (RSVD) and Barnes-Hut t-stochastic neighbor embedding (BH-tSNE) to highlight the underlying taxonomic structure of the metagenomic data. The effectiveness of our approach is demonstrated using several simulated and a real metagenomic datasets.**

.

I. Introduction

Metagenomic methods aim to study the genetic information of environmental samples. They have the potential for increasing our knowledge in a variety of fields including medicine, agriculture, and ecology. The shotgun sequencing approach enables sensitive profiling of the taxonomic composition of microbial communities. Recently, utilising this approach, several methods have been suggested to reconstruct the genomic fragments of single species within a community [1], [2]. However sequencing errors, sequence repetition, insufficient coverage, and genetic diversity can give rise to fragmented assemblies. Therefore, clustering the reconstructed genomic fragments into species-level groups is an important challenge in the analysis of the metagenomes. Clustering plays an important role in metagenomic analysis as it groups related reads or contigs. Therefore it helps in analysing any particular microbial community by identifying the underlying genomic compositions. Unsupervised clustering and visualisation of the metagenomic data is especially helpful when there is no related reference genomes or any other prior information about the taxonomic structure of the data.

A number of clustering techniques employ genomic signatures for clustering purposes. Studies have shown that species-specific genomic *signatures* extracted as features, can be used for clustering the microbial communities [3], [4]. One such signatures include, calculating the normalised frequency of k-mers (all possible subsegments of length k) of a specific size, e.g., k = 4. K-mer frequency of each sequence represents a feature vector in a high dimensional space. Across-sample coverage-profiles or a hybrid approach, with genomic signatures, are also common to describe genomic fragments [5], [6]. Emergent self organising maps (ESOM) based clustering is one such example that uses contour boundaries to visualise the clusters [6]. Unfortunately, ESOM plots are computationally very expensive. On the other hand, methods that consider coverage across multiple samples include CONCOCT [5] and MetaBAT [7], require a high number of samples to perform well, e.g., 50. One approach, VizBin [8], is a reference-independent visualisation approach that considers a single sample, however, it needs manual selection of the centroids for clustering.

The k-mer frequency feature has been extensvely used for metagenomic data analysis. However, here our aim is to use signal processing approaches that have addressed similar and highly relevant problems in various applications. For instance, feature descriptors capturing local texture changes can help segment an image into several meaningful partitions [9], [10]. Similar methods are also common in voice activation detection and recorded audio signal segmentation [11]. However, most of the signal processing methods require numerical data as an input. Thus, in order to use the existing signal processing tools for metagenomic data analysis, genomic sequences need to be mapped into one or several numerical representations [12]. A group of such representation methods are based on biochemical or biophysical properties of DNA molecules. Later, using signal processing tools features can be extracted from these numerical representations, that can be readily given to any clustering algorithm.

In this study, we numerically represent the genomic fragments. For extracting the features from the nucleotide for numerical mapping, we use a feature descriptor called local binary patterns (LBP) in one-dimension. We assume that each genomic fragment has a texture pattern that can be extracted using LBP. For visualisation, the feature vectors need to be projected from a higher dimensional to lower dimensional space (e.g., two-dimensions). The common feature reduction techniques are: (i) linear based such as, singular value decomposition (SVD) [13], [14] and (2) non-linear based such as, ESOMs [15] and Barnes-Hut t-Distributed stochastic neighbor embedding (BH-tSNE) [16]. Although the non-linear techniques preserve the underlying structure of the data, they are computationally expensive. Therefore, we have used randomised SVD (RSVD) [17] to first reduce the higher dimension to obtain "eigengenome" information [18] in a shorter time. Finally, there eigengenome features are given as an input to BH-tSNE for visualising the metagenomic dataset.

Our contributions can thus be summarised as follows:

- **Novel use of LBP for extracting species specific genomic signatures.** Although LBP has been used extensively as a feature descriptor in the fields of image, speech, and signal processing, its application to analyse metagenomic data is novel.
- **Novel use of nucleotide mapping.** Since various representations of nucleotide representations carry different properties of a genomic sequence, a combination of properties can improve the discriminating properties between sequences. Therefore, we have designed a nucleotide map that uses a combination of Electron-ion interaction potentials (EIIP), atomic, and paired nucleotide representations.
- **Proposal of LBP+RSVD+BH-tSNE pipeline.** LBP captures the descriptive information from the genomic sequences, while RSVD captures the eigengenome information in fewer dimensions, that can be readily given to BH-tSNE algorithm for visualisation.

II. METHODOLOGY

In this section we present our methodological pipeline (Fig. 1). We numerically represent the genomic fragments using three nucleotide mapping. After that LBP is used to extract features from these numerical representations. RSVD is then used to reduce the dimensions of the LBP feature vectors by capturing the eigengenome information. BH-tSNE is then used to map RSVD features on to a two-dimensional space for visualisation. For quantitatively evaluating the visualisation performance, we cluster the BH-tSNE projected data using k-means++ algorithm and calculate the rand index between the k-means++ assigned labels and the original labels.

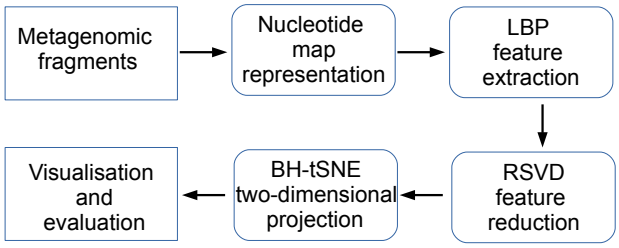


Fig. 1: Schematic overview of the proposed visualisation of the species relationship among metagenomic fragments.

A. The Nucleotide Mapping

The common methods to numerically represent the genomic reads can be categorised into two groups: (1) assigning an arbitrary value to each letter A, C, G, and T of the nucleotide sequence. Voss representation [19] and two and four bit binary representations [20], [21] can be considered as examples of this group. (2) defining numerical representations that correspond to certain biochemical or biophysical properties of the DNA molecules. EIIP [22], paired nucleotide representations [23], and atomic representations [24] are examples of this group.

Since various representations carry different properties (texture patterns) of each sequence, a combination can improve comparisons. Therefore, we designed a nucleotide mapping that uses a combination of EIIP, atomic, and paired nucleotide representations. Table I shows the value assigned to each nucleotide in each of the representations. Fig. 2 shows an example of mapping a nucleotide sequence to three numerical vectors.

TABLE I: EIIP, atomic, and paired nucleotide representations.

Letter	EIIP	Atomic	Paired
A	0.1260	70	0
C	0.1340	58	1
G	0.0806	66	1
T	0.1335	78	0

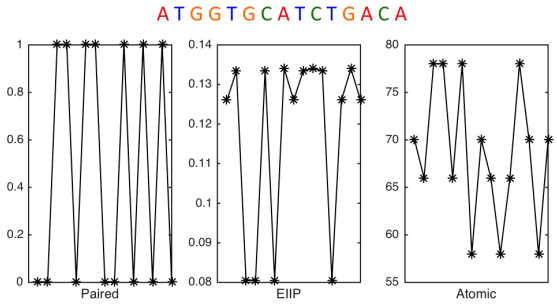


Fig. 2: A nucleotide sequence (top) and its EIIP, atomic, and paired representations. Each letters A, C, G, and T of nucleotide sequence assigns to a value depending on the representation.

B. Local Binary Patterns

LBP has gained significant popularity in the field of image, speech, and signal processing [25]. Using LBP, each two-dimensional window is mapped to a binary number with a fixed length. LBP codes illustrate the data patterns (e.g., for textural changes in images and frequency changes in speech) while the histogram distribution shows how often each pattern appears. These histograms are considered as the feature vectors which essentially extract the species specific genomic signatures. Here, we apply LBP to one-dimensional linear sequences.

LBP assigns a binary code to each sample by examining its neighbouring points. By considering $x(t)$ as the t^{th} sample of the numerical representation of a genomic segment, LBP is defined as

$$\begin{aligned}\text{LBP}(x(t)) = \sum_{i=0}^{p/2-1} \{&\text{Sign}(x(t+i-p/2)-x(t))2^i + \\ &\text{Sign}(x(t+i+1)-x(t))2^{i+p/2}\},\end{aligned} \tag{1}$$

where p is the number of neighbouring points and Sign indicates the sign function

$$\text{Sign}(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}. \tag{2}$$

Sign assigns a binary number by thresholding the difference between each neighbouring point and the centre point t. Consequently, it assigns a p-bit binary number to each window

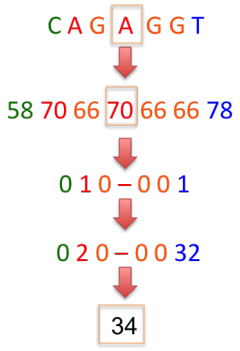


Fig. 3: Calculating the LBP code. A threshold of the atomic numerical representation of the sequence is determined by comparing the centre point and its neighbours. The LBP code is then obtained by using binomial weights.

of length $p+1$. Each binary number is converted to a LBP code using a binomial weight. An example of the LBP operator can be seen in Fig. 3 where $p = 6$. The value of the centred point (squared in Fig. 3) is compared with the six neighbouring points to produce the LBP code. This code describes the data changes locally all in a compressed format. Finally, by considering all the obtained codes, the distribution of the LBP codes can be defined as

$$\mathbf{h}_k = \sum_{p/2 \leq i \leq N-p/2} \delta(\text{LBP}_p(x(i), k), \tag{3}$$

where $k = 1, 2, ..., 2^p$ and N is the genomic fragment length.

C. Randomised Singular Value Decomposition

A metagenomic community can be considered as a linear combination of genomic variables. The sequence of LBP codes for each genomic fragment captures the changes in the pattern (the "texture") of each distinct fragment. By representing a vector of LBP codes for each fragment, low-rank matrix approximations can be used for efficient analysis of the metagenomic data.

SVD decomposition of a matrix $\mathbf{X}$ is defined as

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T, \tag{4}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors, Σ contains singular values, and $(\cdot)^T$ denotes the transpose operator.

SVD can be time consuming when dealing with large scale problems such as metagenomic data analysis. Therefore, RSVD is used as an accurate and robust solution to estimate a number of dominant eigen components quicker as shown in [26].

RSVD calculates the first ith eigen components of the data by using QR decomposition and mapping $\mathbf{X}$ to a smaller matrix as

$$\begin{aligned} \mathbf{\Omega} &= \text{randn}(N, i), \\ \mathbf{Y} &= \mathbf{X}\mathbf{\Omega}, \mathbf{Y} = \mathbf{Q}\mathbf{R} \\ \mathbf{B} &= \mathbf{Q}^T\mathbf{X}, \end{aligned} \tag{5}$$

where randn generates a random matrix of size of its inputs and N is the number of fragments. After decomposing $\mathbf{B}$ using SVD, the final factors are obtained using $\mathbf{Q}$ and the eigen factors of $\mathbf{B}$.

D. Barnes-Hut t-Distributed Stochastic Neighbor Embedding

BH-tSNE has become a common technique for high dimensional data visualisation in several applications [16]. It is based on the divergence minimisation of two distributions: pairwise similarities of the input objects and the corresponding low-dimensional points. As a result, the data in the final lower dimension keeps the original local data structure.

The ordinary similarity measure of data points is defined based on normalised Gaussian kernel values that scales quadratically to the number of data points. The main objective function also has been approximated by defining the similarity function based on a number of neighbouring points [16]. In addition, a vantage-point tree is employed for rapidly finding the neighbouring points. BH-tSNE is then a more efficient ($O(N\log N)$) data reduction approach and used in this paper for data visualisation.

E. Performance Evaluation

Finally, in order to check the performance, k-means++ [27] has been used to cluster the final results. The rand index [28] is calculated between the k-means++ assigned labels and the original labels to determine the performance as a measure of a clusters "purity".

III. Datasets

To validate the effectiveness of our methodology we consider several simulated datasets as well as a real dataset. The simulated datasets used in this study have been generated using the Grinder metagenomic simulator software [29]. Our simulations mainly focused on generating datasets with two main properties of microbial communities, i.e., mixtures of different species with (i) unevenly distributed taxa and (ii) closely related taxa. In the case of unevenly distributed taxa, we have simulated five datasets consisting of seven bacterial species. Their %GC and genome size are illustrated in Table II. Each dataset has a number of genomic fragments (Fig. 4).

TABLE II: Properties of the genomes used in IV *A*, *B*, and *C*.

Bacterial species	%GC	Genome size (nt)
Candidatus Carsonella ruddii PV	16.6	159662
Rickettsia prowazekii str. Dachau	29.0	1109051
Haemophilus influenzae PittGG	38.0	1887192
Bacillus amyloliquefaciens FZB42	46.5	3918589
Escherichia coli UTI89	50.6	5065741
Leifsonia xyli subsp. xyli str. CTCB07	67.7	2584158
Geodermatophilus obscurus DSM 43160	74.0	5322497

The five simulated datasets are as follows: each fragment with a length of $1000nt$ (set1), $1000nt$ with an error rate of 1% (set2), $1000nt$ with an error rate of 3% (set3), each fragment with a length of $800nt$ (set4) , and each fragment with a length of $400nt$ (set5). For closely related taxa, we have considered six bacterial species; their %GC and genome size are shown in Table III. The fragment length considered is $1000nt$ and the number of genomic fragments is demonstrated in Fig. 5 (set6).

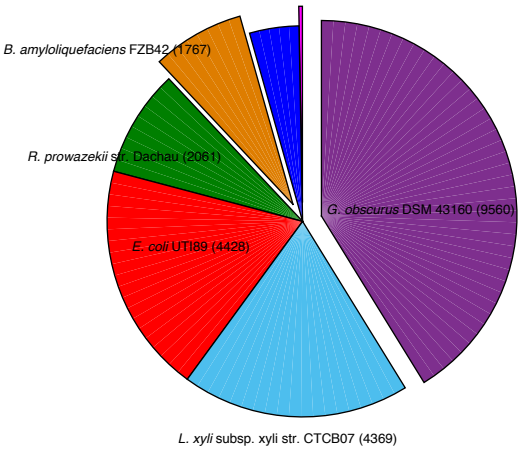


Fig. 4: Visualisation of the unevenly distributed metagenomic community. Abundance of genomic fragments (the number of fragments per species) is indicated in parentheses.

TABLE III: Properties of the genomes used in IV *D*.

Bacterial species	%GC	Genome size (nt)
Streptococcus thermophilus CNRZ 1066	39.1	1796226
Streptococcus suis A7	41.2	2038409
Lactobacillus brevis ATCC 367	46.2	2291220
Lactobacillus casei ATCC 334	46.6	2895264
Lactobacillus delbrueckii ATCC 11842	49.7	1864998
Shewanella amazonensis SB2B	53.6	4306142

Finally, an infant human gut dataset has been considered here that has been analysed in [6] for microbial genome reconstruction. They assembled the data into 2,329 contigs and assembly and binning information (carrol.scaffolds_to_bin.tsv) is provided in http://ggkbase.berkeley.edu/carrol/. Corresponding reads can be downloaded from the NCBI, SRA052203. It consists of 18 Illumina runs (SRR492065-66 and SRR492182-97).

IV. Experiments and Results

In this section we present our experiments and results:

1) **Effect of LBP parameter tuning.** For the LBP, we have to determine the optimal window length of the LBP method. Therefore, in this experiment we would like to investigate the optimal window length of the LBP that could obtain a maximum performance.
2) **Effect of RSVD.** The computational complexity of BH-tSNE increases as the dimensions of features increase. We conjecture that giving RSVD eigengenome features as an input to BH-tSNE would achieve similar results as that of BH-tSNE but in less time. Therefore, we test to find out the optimal number of RSVD components required to obtain a closer result to that of BH-tSNE.
3) **Performance of LBP+RSVD+BH-tSNE pipeline.** We conduct the experiments on all the datasets using the optimal parameters achieved through the above experiments on set1. We discuss the performance of our proposed pipeline for the aforementioned simulated and real data and discuss the results in detail in the following subsections.

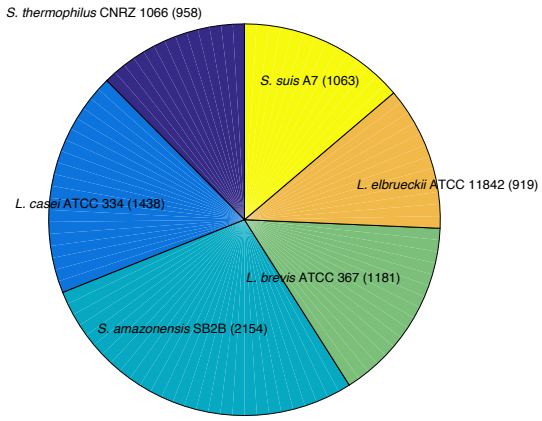


Fig. 5: Visualisation of metagenomic community with closely related taxa. Abundance of genomic fragments is indicated in parentheses.

A. Effect of LBP Parameter Tuning

The effect of considering various LBP window lengths on performance is show in Table IV for set1. We have selected length of 8 in this work to have good performance and less time complexity.

TABLE IV: RI Score (%) for the simulated data and various LBP window lengths ($p+1$).

LBP window size	5	7	9	11
Feature length	48	192	768	3072
RI	82.8190	83.9159	**83.9659**	83.4550

B. Effect with and without RSVD

Table V demonstrates the effect of different numbers of RSVD eigen components on the final performance. It can be seen that keeping only 50 eigen components leads to very close results compared to not using RSVD, however, the time complexity improves a lot as BH-tSNE has taken around 50% less time using RSVD.

TABLE V: RI score (%) for set1 without applying RSVD and with keeping various number of eigen components.

RSVD Dimensions					**no RSVD**
10	20	30	40	50	
82.0582	81.7622	82.5971	83.8189	**83.9634**	84.0261

C. Results on an Unevenly Distributed Community

For set 1, where each fragment has a length of the 1000nt, the final representation of metagenomic fragments clusters the genomic fragments of the same species together (Fig. 6). Our results demonstrate that visualisation of the data can help evaluate the underlying data structure. Moreover, RI is performed to demonstrate the clustering performance.

Furthermore, on the simulated metagenomic data with various error rates has been visualised in Fig. 7. It shows, even

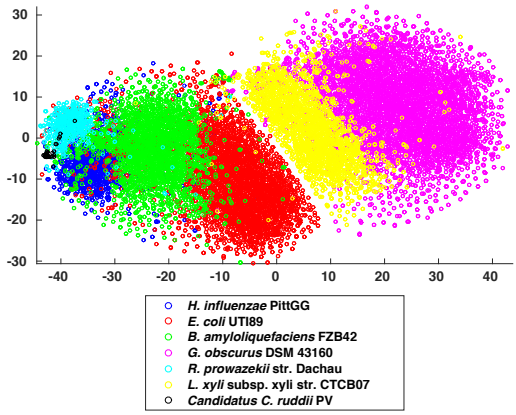
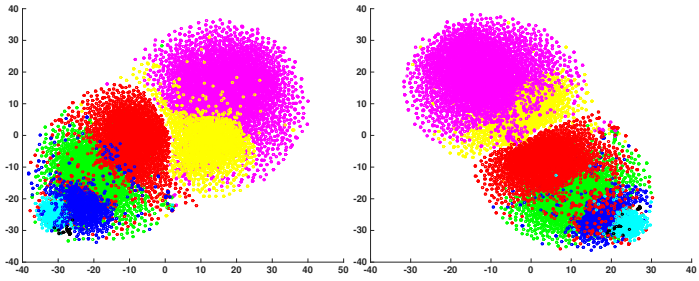


Fig. 6: Visualisation of the unevenly distributed metagenomic community. The result of applying the proposed method to the simulated mixture of unevenly distributed taxa. Each colour represents a different species (see key).

TABLE VI: RI score (%) for all the simulated data.

set1	set2	set3	set4	set5	set6
83.9561	83.1567	80.3416	80.4867	76.98450	80.4867

in the presence of some error, the proposed visualisation procedure returns similar results (Table VI set2 and set3).



(a) (b)

Fig. 7: Visualisation of the metagenomic community with genomic length of 1000nt and two error rates (a) 1% and (b) 3%. See Fig. 6 for species key.

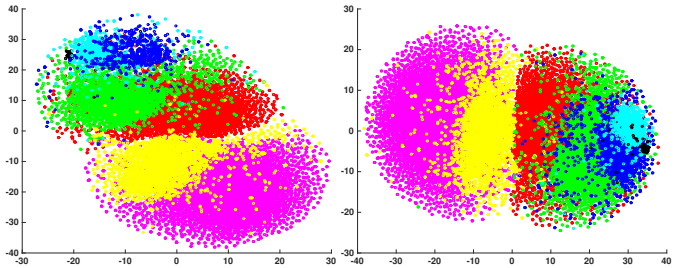
Moreover, in order to check the effect of segment length on the visualisation, two metagenomic data have been simulated with genomic fragments of length 800 and 400nt. The results as illustrated in Fig. 8 still resemble the underlying data structure (Table VI set4 and set5). However, shorter fragments can result in similar feature vector and lower performance.

D. Results on Data with Closely Related Taxa

The representation of metagenomic fragments clusters the genomic fragments of the same or closely related species together (Fig. 9 and Table VI set6). *S. amazonensis* SB2B and *S. suis* A7 genomic contigs overlapped in two-dimensional space. However, most of the remaining genomic fragments form separate clusters.

E. Results on Real Infant Gut

11 clusters is shaped by applying the proposed method to the real human gut (Fig. 10). *S. Lugdunensis* and *S. Aureus contigs*



(a) (b)

Fig. 8: Visualisation of the metagenomic community with genomic length of (a) 800 and (b) 400nt. See Fig. 6 for species key.

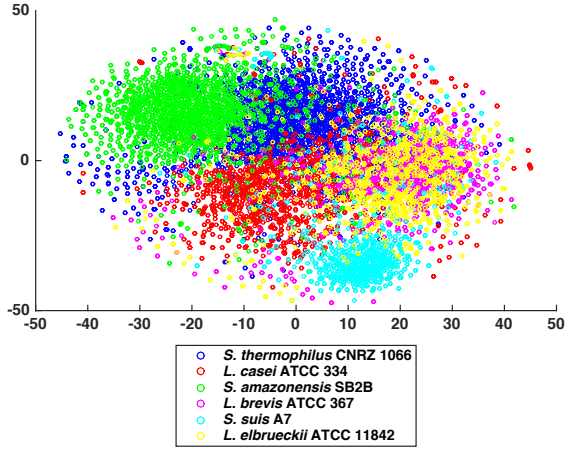


Fig. 9: Visualisation of metagenomic data with closely related taxa. Each colour represents a different species (see key).

and also *S. Hominis* and *S. Epidermidis* contigs has been grouped together. However, the remaining of species forms separate cluster that confirm the application of the proposed method to real data.

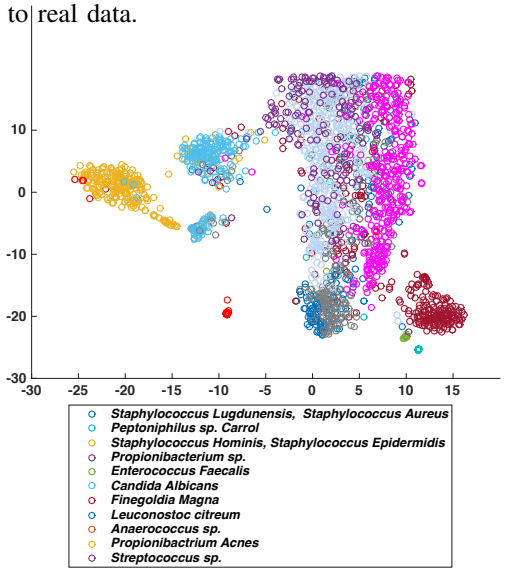


Fig. 10: Visualisation of human gut community. Each colour represents a different species (see key). The species clustered together has been shown with similar colours.

V. Conclusions

A metagenomic visualisation approach has been introduced by representing the nucleotide genomic fragments numerically. LBP has been employed to describe the genomic signature changes followed by a dimension reduction step to visualise the data in a lower dimension. Our results on simulated genomic fragments show the underlying taxonomic structure of the metagenomic data and verify the advantage of using signal processing approaches for metagenomic data analysis. Consequently, it shows the potential of the proposed method to analyse complex communities.

Several metagenomic communities were considered with various error levels and fragment lengths. As illustrated in Section IV, the proposed method can be used for visualisation and clustering of such data at genus or species level. In addition, only a limited number of contigs overlap with the clusters of other species. To have a better clustering, longer fragments are desirable. However, in this study we have considered only a limited number of genomic fragments of one sample, where considering their longitudinal aspects may result in better clustering of shorter fragments.

Acknowledgement

'SK' receives funding from the VIROGENESIS project. The VIROGENESIS project receives funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 634650. 'AT' receives funding from a BBSRC project grant, BB/M001121/1. 'ST' receives funding from the Medical Research Council funded project "Modelling the Progression of Chronic Kidney Disease under the grant number R/M023281/1, the details of the project can be found at www.modellingCKD.org. We would like to thank Bede Constantinides for his useful comments.

References

[1] J. Handelsman, "Metagenomics: application of genomics to uncultured microorganisms," *Microbiology and molecular biology reviews*, vol. 68, no. 4, pp. 669–685, 2004.

[2] H. B. Nielsen, M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. Le Chatelier *et al.*, "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes," *Nature biotechnology*, vol. 32, no. 8, pp. 822–828, 2014.

[3] G. J. Dick, A. F. Andersson, B. J. Baker, S. L. Simmons, B. C. Thomas, A. P. Yelton, and J. F. Banfield, "Community-wide analysis of microbial genome sequence signatures," *Genome biology*, vol. 10, no. 8, p. 1, 2009.

[4] K. C. Wrighton, B. C. Thomas, I. Sharon, C. S. Miller, C. J. Castelle, N. C. VerBerkmoes, M. J. Wilkins, R. L. Hettich, M. S. Lipton, K. H. Williams *et al.*, "Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla," *Science*, vol. 337, no. 6102, pp. 1661–1665, 2012.

[5] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince, "Binning metagenomic contigs by coverage and composition," *Nature methods*, vol. 11, no. 11, pp. 1144–1146, 2014.

[6] I. Sharon, M. J. Morowitz, B. C. Thomas, E. K. Costello, D. A. Relman, and J. F. Banfield, "Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization," *Genome research*, vol. 23, no. 1, pp. 111–120, 2013.

[7] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities," *PeerJ*, vol. 3, p. e1165, 2015.

[8] C. C. Laczny, T. Sternal, V. Plugaru, P. Gawron, A. Atashpendar, H. H. Margossian, S. Coronado, L. Van der Maaten, N. Vlassis, and P. Wilmes, "Vizbin-an application for reference-independent visualization and human-augmented binning of metagenomic data," *Microbiome*, vol. 3, no. 1, p. 1, 2015.

[9] M. Pietikäinen and T. Ojala, "Texture analysis in industrial applications," in *Image Technology*. Springer, 1996, pp. 337–359.

[10] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. Ho, "Detection of face spoofing using visual dynamics," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 762–777, 2015.

[11] Q. Zhu, N. Chatlani, and J. J. Soraghan, "1-D local binary patterns based VAD used in HMM-based improved speech recognition," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 1633–1637.

[12] A. Tapinos, B. Constantinides, D. B. Kelland, and D. L. Robertson, "Alignment by the numbers: sequence assembly using reduced dimensionality numerical representations," *bioRxiv*, p. 011940, 2014.

[13] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische mathematik*, vol. 14, no. 5, pp. 403–420, 1970.

[14] M.-S. Paukkeri, I. Kivimäki, S. Tirunagari, E. Oja, and T. Honkela, "Effect of dimensionality reduction on different distance measures in document clustering," in *International Conference on Neural Information Processing*. Springer, 2011, pp. 167–176.

[15] D. Deng and N. Kasabov, "ESOM: An algorithm to evolve self-organizing maps from on-line data streams," in *Proc. of IJCNN*, vol. 6, 2000, pp. 3–8.

[16] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms." *Journal of machine learning research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[17] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions," 2009.

[18] B. Cleary, I. L. Brito, K. Huang, D. Gevers, T. Shea, S. Young, and E. J. Alm, "Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning," *Nature biotechnology*, vol. 33, no. 10, pp. 1053–1060, 2015.

[19] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Physical review letters*, vol. 68, no. 25, p. 3805, 1992.

[20] R. Ranawana and V. Palade, "A neural network based multi-classifier system for gene identification in DNA sequences," *Neural Computing & Applications*, vol. 14, no. 2, pp. 122–131, 2005.

[21] B. Demeler and G. Zhou, "Neural network optimization for e. coli promoter prediction," *Nucleic acids research*, vol. 19, no. 7, pp. 1593–1599, 1991.

[22] A. S. Nair and S. P. Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)," *Bioinformation*, vol. 1, no. 6, pp. 197–202, 2006.

[23] P. Bernaola-Galván, P. Carpena, R. Román-Roldán, and J. Oliver, "Study of statistical correlations in DNA sequences," *Gene*, vol. 300, no. 1, pp. 105–115, 2002.

[24] T. Holden, R. Subramaniam, R. Sullivan, E. Cheung, C. Schneider, G. Tremberger Jr, A. Flamholz, D. Lieberman, and T. Cheung, "ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes," in *Optical Engineering+ Applications*. International Society for Optics and Photonics, 2007, pp. 669 417–669 417.

[25] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[26] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

[27] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[28] K. Y. Yeung and W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[29] F. E. Angly, D. Willner, F. Rohwer, P. Hugenholtz, and G. W. Tyson, "Grinder: a versatile amplicon and shotgun sequence simulator," *Nucleic acids research*, p. 251, 2012.