Self-organizing maps: A tool to ascertain taxonomic relatedness based on features derived from 16S rDNA sequence

D V Raje[1], H J Purohit[1,*], Y P Badhe[2], S S Tambe[2] and B D Kulkarni[2]

[1]*Environmental Genomics Unit, National Environmental Engineering Research Institute (NEERI), Nagpur 440 020, India*

[2]*Chemical Engineering and Process Development Division, National Chemical Laboratory (NCL), Dr Homi Bhabha Road, Pune 411 008, India*

**Corresponding author (Fax, +91 712 2249 883; Email, hemantdrd@hotmail.com)*

Exploitation of microbial wealth, of which almost 95% or more is still unexplored, is a growing need. The taxonomic placements of a new isolate based on phenotypic characteristics are now being supported by information preserved in the *16S rRNA* gene. However, the analysis of 16S rDNA sequences retrieved from metagenome, by the available bioinformatics tools, is subject to limitations. In this study, the occurrences of nucleotide features in 16S rDNA sequences have been used to ascertain the taxonomic placement of organisms. The tetra- and penta-nucleotide features were extracted from the training data set of the 16S rDNA sequence, and was subjected to an artificial neural network (ANN) based tool known as self-organizing map (SOM), which helped in visualization of unsupervised classification. For selection of significant features, principal component analysis (PCA) or curvilinear component analysis (CCA) was applied. The SOM along with these techniques could discriminate the sample sequences with more than 90% accuracy, highlighting the relevance of features. To ascertain the confidence level in the developed classification approach, the test data set was specifically evaluated for *Thiobacillus,* with *Acidiphilium, Paracocus* and *Starkeya*, which are taxonomically reassigned. The evaluation proved the excellent generalization capability of the developed tool. The topology of genera in SOM supported the conventional chemo-biochemical classification reported in the Bergey manual.

[Raje D V, Purohit H J, Badhe Y P, Tambe S S and Kulkarni B D 2010 Self-organizing maps: A tool to ascertain taxonomic relatedness based on features derived from 16S rDNA sequence; *J. Biosci.* **35** 617–627] **DOI** 10.1007/s12038-010-0070-y

1. Introduction

Molecular phylogenetic methods have revolutionized the classifying and identification of organisms that occur in microbial communities (Hugenholtz *et al.*1998). Prior to this development, the chemo-biochemical characteristics of strains were used to derive the coefficient of similarity (or percentage similarity) between the strains, leading to what is known as *numerical taxonomy* (Garrity *et al.* 2001). However, with the increase in the number of bacterial isolates, it became apparent that many such phenotypic criteria have limitations. For a new isolate based only on biochemical properties, quite often it is difficult to predict the phylogeny or additional associated characteristics of the isolate. Anticipating this, Woese (1987) suggested the use of nucleotide sequence differences in a single gene to investigate the evolutionary relationships. They pioneered the use of rRNA for phylogenetic analysis, which subsequently led to redrawing the universal tree of life and opened a new era of *molecular taxonomy.*

The *16S rRNA* gene is widely used to investigate the evolutionary relationships of prokaryotes. Over the years,

Keywords. Curvilinear component analysis; self-organizing maps; principal component analysis

Abbreviations used: ANN, artificial neural network; CCA, curvilinear component analysis; ID, intrinsic dimensionality; MDS, multi-dimensional scaling; PCA, principal component analysis; PE, processing element; SOM, self-organizing map; VQ, vector quantization the 16S rDNA database has grown tremendously. In taxonomy, the highly conserved regions group bacteria into higher taxonomic orders, whereas the variable regions allow classification at lower taxonomic levels, such as the genus or species level (Amann *et al*. 1995). The method consist of aligning the sequences using ClustalW/ClustalX and then obtaining the pair-wise distance matrix based on the alignment in order to provide taxonomic relatedness (Durbin *et al.* 1998). The analysis also provides an estimate of the evolutionary distance between sequences.

The early classification of life based on the *rRNA* gene by Carl Woese showed that bacteria could be divided into two different groups, and Archaea has been made as an additional group. The study led to the generation of the ribosomal RNA data base that contained the sequence data originating from either RNA or DNA versions of the 16S rRNA molecule (Maidak *et al*. 1994). The majority of the sequences in the database are generated from amplified PCR products of the *16S rRNA* gene, where the templates were derived from either genomic DNA of isolates or environmental DNA. However, use of 16S rRNA data does have some problems, which are mostly due to techniques involved in developing the sequence data. In cases in which cDNA is synthesized from the extracted total RNA from bacteria, the process is carried out by reverse transcriptase, which has its limitations. In cases in which rDNA is amplified from the metagenome and cloned, sometimes, hybrids are formed between two *16S rRNA* genes derived from different organisms present in the same environment and they generate chimeras that are difficult to classify (Ward *et al*. 1990; Amann *et al*. 1995). Yet, 16S rDNA is the usual data used for classifying microorganisms, and in this study as well, we have used 16S rDNA sequences in the analysis.

In this study, we have made an attempt to classify organisms based on abstract information from sequences, such short nucleotide patterns of length four or five base pairs. In other words, the interest in this study was to know whether the patterns and their occurrences in 16S rDNA sequences hold information that can be used to distinguish a set of genera from each other. This assertion rests on our earlier study wherein we showed that the five most dominant and closely related bacterial genera could be discriminated using the di-nucleotide compositions of *16S rRNA* gene (Raje *et al.* 2002).

In recent years, artificial neural network (ANN)-based tools have received wide acceptance in different application domains, especially when the variable space is large. One such well-known tool is the self-organizing maps (SOM). This tool undergoes unsupervised learning and is particularly useful in projecting/visualizing high-dimensional data (Kohonen 1990). The important application of SOM is classification (clustering), which has been exemplified in a number of recent genomic studies. To cite a few, SOM along with PCA was used for sub-cellular locations of bacterial proteins, resulting in a clear separation of cytoplasmic, periplasmic and extra-cellular proteins (Schneider 1999). In this study, the global sequence features based on amino acid composition were used for localization. Later, SOM was also used for the analysis of codon usage patterns of bacterial genomes wherein the clustering of average codon usage of main gene categories of genomes showed mixing of gene classes (Wang *et al.* 2001). Further, SOM was employed to analyse di- and tri-nucleotide frequencies in nine eukaryote genomes and were able to recognize the species-specific characteristics as a signature representation of each genome (Abe *et al.* 2003). Subsequently, they used SOM in a wide variety of prokaryotic and eukaryotic genomes wherein the analysis of 1- and 10-kb genomic sequences from 65 prokaryotes and 6 from eukaryotes provided a clear species-specific separation of major portions of the sequences using SOM. Also, Kasturi *et al.* (2003) studied the relative dissimilarity between the gene expression profiles in conjunction with an unsupervised SOM algorithm.

In this exercise, SOM has been used exclusively and also in conjunction with linear and non-linear feature extraction approaches such as PCA and curvilinear component analysis (CCA). For training purpose, 800 16S rDNA sequences belonging to 40 different genera were considered and each sequence was represented by the frequency of occurrence of various tetra- and penta-nucleotides. These tetra- and penta-nucleotide combinations were treated as features in this exercise. SOM along with the feature extraction techniques could discriminate the sample sequences with more than 90% accuracy. Moreover, the map also outlined the territory for different taxonomic classes for the selected genera, thereby highlighting the strength of the tool and the relevance of features.

2. Methods

2.1 *Feature extraction*

Feature extraction is a process of mapping a set of measurements (data) into fewer features, which preserve the main information of the original data structure. One of the important components of feature extraction is data projection, wherein the data in the original high-dimensional space is projected/transformed onto a lower-dimensional space (usually 2D or 3D) so that it can be visualized easily and the relationships and structure in the data can be clearly identified. While projecting the data onto a reduced dimensional space, it is necessary to perform dimensionality reduction of the original data. The methods performing dimensionality reduction search for a smaller set of variables (known as "features") for describing a large set of observed dimensions. Data projection onto a lower-dimensional feature space enables better understanding of the data structure, exploration of the intrinsic data dimensionality and analysis of the clusters present in the original high-dimensional data.

2.2 *Dimensionality reduction*

In principle, feature extraction and data projection can be formulated as a mapping Ψ from an n-dimensional input space to a p-dimensional mapping space (Ψ: $\Re^n \rightarrow \Re^p$, $p \leq n$) such that some criterion, E, is optimized. For data visualization purposes, the value of p is usually set to 2 or 3. A large number of approaches for feature extraction and data projection are available in the literature on pattern recognition. These approaches differ from each other in the characteristics of the mapping function, Ψ, and how Ψ is learned. The widely used feature extraction and dimensionality reduction method is PCA (Wold *et al.*1987). However, this technique captures only linear relationships among the multiple variables of a data set. Invariably, when data variables are correlated non-linearly, the linear PCA fails to capture these correlations. To overcome this drawback of linear PCA, a number of non-linear feature extraction and dimensionality reduction methods have been proposed. The significant ones among these are the multi-dimensional scaling (MDS)-based Kruskal mapping (Kruskal 1964), the Sammon mapping (Sammon 1967), isometric feature mapping IsoMap (Tenebaum *et al.* 2000) and locally linear embedding (LLE) (Roweis and Saul 2000). A significant drawback of these methods is that they do not possess the generalization capability of reducing/projecting new data without re-mapping the combined set and hence comprising the original and new data. The recently proposed CCA (Demartines and Herault 1997), which is an ANN-based non-linear feature extraction and dimensionality reduction paradigm, has overcome the above-stated (i.e. inability to generalize) drawback. Accordingly, the CCA formalism has been used in this study for conducting non-linear feature extraction of tetra- and penta-nucleotide frequency data. Specifically, the dimensionality of these two data sets has been reduced using the CCA formalism to afford subsequent classification and visualization using the SOM neural network (figures 1 and 2).

2.2.1 *Curvilinear component analysis:* The primary objective of the CCA is to generate a revealing representation of the original data in a lower-dimensional feature space so as to prepare a foundation for the further clustering of the input data. The CCA operates on the principle of preserving distances in its input and output spaces. However, in case of non-linearly correlated input data, it may not be possible to preserve distances of large magnitudes because the task necessitates unfolding of the manifold to effect dimensionality reduction in the projected space. For achieving the preservation of local distances, the CCA employs a neighbourhood function, which fulfills the condition of preservation of smaller distances while relaxing the condition for larger distances (Buchala *et al.* 2004a).

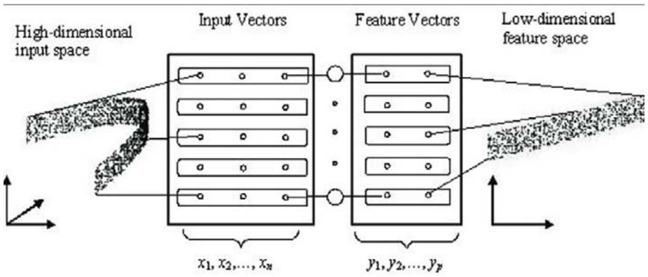

Figure 1. A schematic of the CCA network.

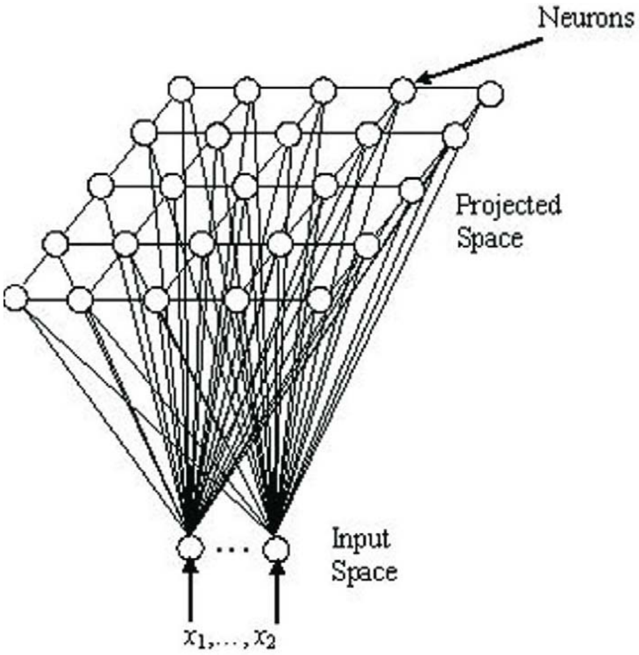

Figure 2. A schematic of self-organizing map.

The CCA can be considered as a self-organizing neural network (figure 2) that performs two tasks: (i) vector quantization (VQ) of the submanifold in the data set (input space) and (ii) non-linearly projecting the quantized vectors onto the output space. A vector quantizer maps n-dimensional vectors in the vector space, $\Re^n$, into a finite set of vectors, where $p < n$. That is, while dimensionality reduction methods reduce the dimension of the data, vector quantization reduces the number of data points that are termed “prototypes”. The main purpose of vector quantization in CCA is to reduce computational cost. If the data set is relatively small (a few hundred data points), then it may not even be essential to perform vector quantization.

In such a case, only the projection part of the CCA needs to be conducted (Buchala *et al.* 2004b). After training, the CCA network has the ability of generalization owing to which it can continuously map any new point in the forward or backward direction. The CCA as shown in figure 2 performs the VQ and non-linear projection tasks separately using two layers of connections. The first network layer performs VQ on the data set and the second layer, known as the "projection layer", conducts topographic mapping of the quantized vectors. The projection layer is a free space, which takes the shape of the sub-manifold of the data.

The training algorithm for the CCA network was proposed as an improvement to the Kohonen SOM, wherein the output is not a fixed lattice but a continuous space capable of taking the shape of the manifolds of the input data. In what follows, the procedural details of the CCA training are described.

Let $\{\mathbf{x}_i\}$, $i = 1, 2, \ldots, N$, be the set of data vectors ($\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^{\mathrm{T}}$) in an n-dimensional input space, and $\{\mathbf{y}_i\}$ be the corresponding lower-dimensional vectors ($\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{ip}]^{\mathrm{T}}$) in the p-dimensional ($p<n$) feature space. Accordingly, each of the n neurons (processing elements) in the CCA network has two weight vectors ($\mathbf{x}_i$ and $\mathbf{y}_i$) associated with it. During training of the network, the processing elements (PEs) in the first layer force the input vectors to become the prototypes of the distribution by using any standard VQ method. The output-layers PEs are required to construct a non-linear mapping of the input vectors. This objective is fulfilled by minimizing the structure differences between the quantized and output spaces. The structure differences can be described in terms of the Euclidean distances and the corresponding quadratic cost function (E) to be minimized for reducing the data dimensionality from n to p is given as

$$E = \frac{1}{2}\sum_{i}\sum_{j \neq i}\left[X_{ij} - Y_{ij}\right]^2 F\,(Y_{ij}, \lambda_y), \tag{1}$$

where $X_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ describes the Euclidean distance between n-dimensional vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, and $Y_{ij} = d(\mathbf{y}_i, \mathbf{y}_j)$ refers to the corresponding distance between p-dimensional vectors $\mathbf{y}_i$ and $\mathbf{y}_j$ in the CCA network's output space. The objective of minimizing E is to force Y_{ij} to match X_{ij} for each possible vector pair (i, j). As a perfect match between X_{ij} and Y_{ij} is not possible while mapping to a lower p-dimensional space, a weighting function $F(X_{ij}, \lambda_y)$ is used to favour the conservation of the local topology. For preserving this topology, a bounded and monotonically decreasing weighting function such as the decreasing exponential or sigmoid function is commonly chosen. The weighting function assigns greater weighting to points lying closer in the output space.

In the beginning, the set of p-dimensional output vectors $\{\mathbf{y}_i\}$ are initialized randomly to small magnitudes. The minimization of E with respect to the vectors $\{\mathbf{y}_l\}$ is performed by following the procedure outlined in Roweis and Saul (2000). This procedure temporarily fixes one of the $\mathbf{y}_i$ vectors and moves all the other $\mathbf{y}_j$ vectors around it without taking into consideration the interactions among the $\mathbf{y}_i$ vectors. The updating rule for $\mathbf{y}_j$ vector to effect minimization of E is given as:

$$\begin{aligned}\Delta\mathbf{y}_j\,(i) &= \alpha\,(t)\nabla_i E_{ij} = -\alpha\,(t)\nabla_j E_{ij}\\ &= \alpha\,(t) F\,(Y_{ij}, \lambda_y)(X_{ij} - Y_{ij})\frac{\mathbf{y}_j - \mathbf{y}_i}{Y_{ij}}, \forall\ \ j \neq i,\end{aligned} \tag{2}$$

where, i refers to the index of a randomly chosen vector; $\nabla_i\, E_{ij}$ represents the gradient of E with respect to $\mathbf{y}_i$, and $\alpha\,(t) = \frac{\alpha_0}{(1+t)}$ denotes the learning rate that decreases with time. The optimized $\mathbf{y}_j$-updation rule in Eq. (2) is numerically efficient, and its implementation results in the output vectors eventually converging to L number of prototypes (y_{i*}, $i = 1, 2, \ldots, L$) in a certain number (<100) of training iterations. The CCA training algorithm can now be briefly summarized as follows:

Step 1: Perform (if needed) VQ to reduce the size of the data set.
Step 2: Compute all pair-wise Euclidean distances $d(\mathbf{x}_i, \mathbf{x}_j)$ in the n-dimensional input space.
Step 3: Initialize the p-dimensional co-ordinates of all points $\{\mathbf{y}_i\}$ either randomly or on the hyperplane spanned by the first principal components (obtained using PCA).
Step 4: Initialize the iteration index t (to 1).
Step 5 Specify learning rate $\alpha(t)$ and neighbourhood width λ_y (width to be decreased with increasing magnitude of the iteration index, t).
Step 6 Compute Euclidean distances $d(\mathbf{y}_i, \mathbf{y}_j)$ in the p-dimensional output (projected) space.
Step 7 Update all the projected vectors according to equation 2.
Step 8 Increase t by 1.
Step 9 Repeat steps 5–8 until the change, $\Delta\mathbf{y}_{\mathrm{j}}(i)$, in the projected space is less than a pre-specified threshold or the maximum number of iterations ($t_{\max}$) is reached.

An important issue in the CCA implementation is to choose the dimension of the output space (p). Ideally, p should be chosen equal to the intrinsic dimensionality (ID) of the input data. The ID is defined as the smallest number (p) of variables that are needed to describe the set of data without any significant loss of information. Usually, the intrinsic dimensionality of a given data set is not known *a priori*. To overcome this difficulty, the above-stated step-wise CCA procedure is repeated a number of times by systematically varying the magnitude of p, and the magnitude resulting in the overall least converged value for the cost function

(E) is taken as a reasonable approximation of the intrinsic dimensionality, p.

The CCA is an efficient non-linear dimensionality reduction technique although other formalisms such as the SOM are necessary for classification and projection if the dimensionality of the projected space is very high ($p > 3$). Accordingly, in the present study, we have used purohit8655SOM for classifying and projecting the dimensionality reduced tetra- and penta-nucleotide frequency data onto a 2D lattice.

In the present study, all CCA-based non-linear dimensionality reduction simulations were performed using following parameter values: (i) initial step-size in the neighbourhood function (α_0) = 0.5, (ii) maximum number of iterations ($t_{\max}$) = 500, (iii) initial radius of influence (λ_0) = 3 times the maximum of the standard deviation of the input data and (iv)

$$F\left(d\left(\mathbf{y}_i,\mathbf{y}_j\right)\right)=\exp\left(\frac{-d\left(\mathbf{y}_i,\mathbf{y}_j\right)}{\lambda_0}\right),$$

where $d(y_i, y_j)$ denotes distances in the output space.

2.2.2 *Self-organizing map:* The SOM network architecture, as shown in figure 3, consists of a 2D array of units each of which is connected to all the p input nodes. It is also possible to use a grid of higher dimensions although such a grid is difficult to visualize conveniently. The SOM neural network architecture and its training method possess the following properties: (i) an array of neurons, which as a function of its input of arbitrary dimensionality, calculates the outputs using a simple output function, (ii) a criterion to determine the "winner" neuron possessing the largest output and (iii) an adaptive rule for updating the weights of the chosen neuron and its neighbours.

2.2.3 *SOM training algorithm:* Let $\mathbf{x}_i$, $i = 1,2,\ldots, N$ be the p-dimensional patterns and $\boldsymbol{w}_{ij}$ be the p-dimensional weight vector associated with the processing element at location (i, j) of the 2D array (figure 2). The step-wise procedure for training the SOM network is as given below.

Step 1 (Initialization): Choose small random values for the initial weights, $\boldsymbol{w}_{ij}(0)$, and fix the initial learning rate ($\hat{\alpha}_0$) and the neighbourhood.

Step 2 (Determining the winner): Select a sample pattern, $\mathbf{x}$, from the data set and determine the winner neuron (C_i, C_j) at time t, using the minimum-distance Euclidean criterion:

$$\left\|\mathbf{x}-\boldsymbol{w}_{C_iC_j}\right\|=\min_{i,j}\left\|\mathbf{x}-\boldsymbol{w}_{i,j}\right\|; \qquad (3)$$
$$i = 1, 2,\&., L; \quad j = 1, 2,\&, L,$$

where ‖.‖ refers to the Euclidean norm and L denotes the number of rows (as also columns) in the square 2D array.

Step 3 (Weight update): Update all the weights according to the kernel-based learning rule:

$$\begin{aligned}\boldsymbol{w}_{ij}(t+1) &= \boldsymbol{w}_{ij}(t) + \hat{\alpha}(t)\,\|\,\mathbf{x}(t) - \boldsymbol{w}_{ij}(t)\,\| \text{ if } (i, j) \in N_{C_iC_j}(t)\\ &= \boldsymbol{w}_{ij}(t), \qquad \text{otherwise}\end{aligned} \qquad (4)$$

where t denotes training iteration index; $N_{C_iC_j}(t)$ is the neighbourhood of the winner unit (C_i, C_j) at iteration t, and $\hat{\alpha}(t) = \hat{\alpha}_0/(1+t)$ is the learning rate.

Step 4 Decrease the value of the learning rate, $\hat{\alpha}(t)$, by incrementing the iteration index t, by unity and shrink the neighbourhood, $N_{C_iC_j}(t)$.

Step 5 Repeat steps 2–4 until the change in the weight values is less than the specified threshold, or the maximum number of iterations ($\hat{t}_{\max}$) is reached.

It should be emphasized that the success of SOM training depends critically on the judicious selection of the main algorithm-specific parameters (i.e., $\hat{\alpha}(t)$ and $N_{C_iC_j}(t)$), initial values of the weight vectors and the number of pre-specified training iterations, $\hat{t}_{\max}$. These are commonly optimised using a heuristic procedure. Further, the lattice of the 2D grid could be either rectangular or hexagonal. In this study, hexagonal lattice has been used since it is visually appealing. The square grid had a 45×45 neuron architecture, and the initial learning rate ($\hat{\alpha}_0$) was fixed to be 0.5. For the implementation of SOM as also CCA, the Matlab-based SOM Toolbox 2.0 was used (Vasanto *et al.* 1999). Neighbouring neurons on the map need not suggest a small distance in the data space. Although the SOM automatically groups together similar data items, for practical purposes it is still desirable to demarcate similar data items into clearly visible distinct clusters. This can be achieved by using the U-matrix method for demarcating boundaries between different clusters. The U-matrix determines the distance between neurons and in the present study different shades of grey are used to portray the distances. The dark-shaded borders between two neurons represent large distances between the data mapped into the respective neurons, whereas a light-shaded border indicates similarities between the data items. Additionally, the data points in the projected space, representing different organisms, are depicted using a colour coding to visualize the clusters formed by the data.

3. Results

The question we raised in this study was: Could 16S rDNA sequences be grouped based on their abstract information? The conceptual framework of the study has been presented in figure 1. From the selected 40 genera (table 1), 800 complete 16S rDNA sequences (20 from each genus) representing different species were retrieved from GenBank.

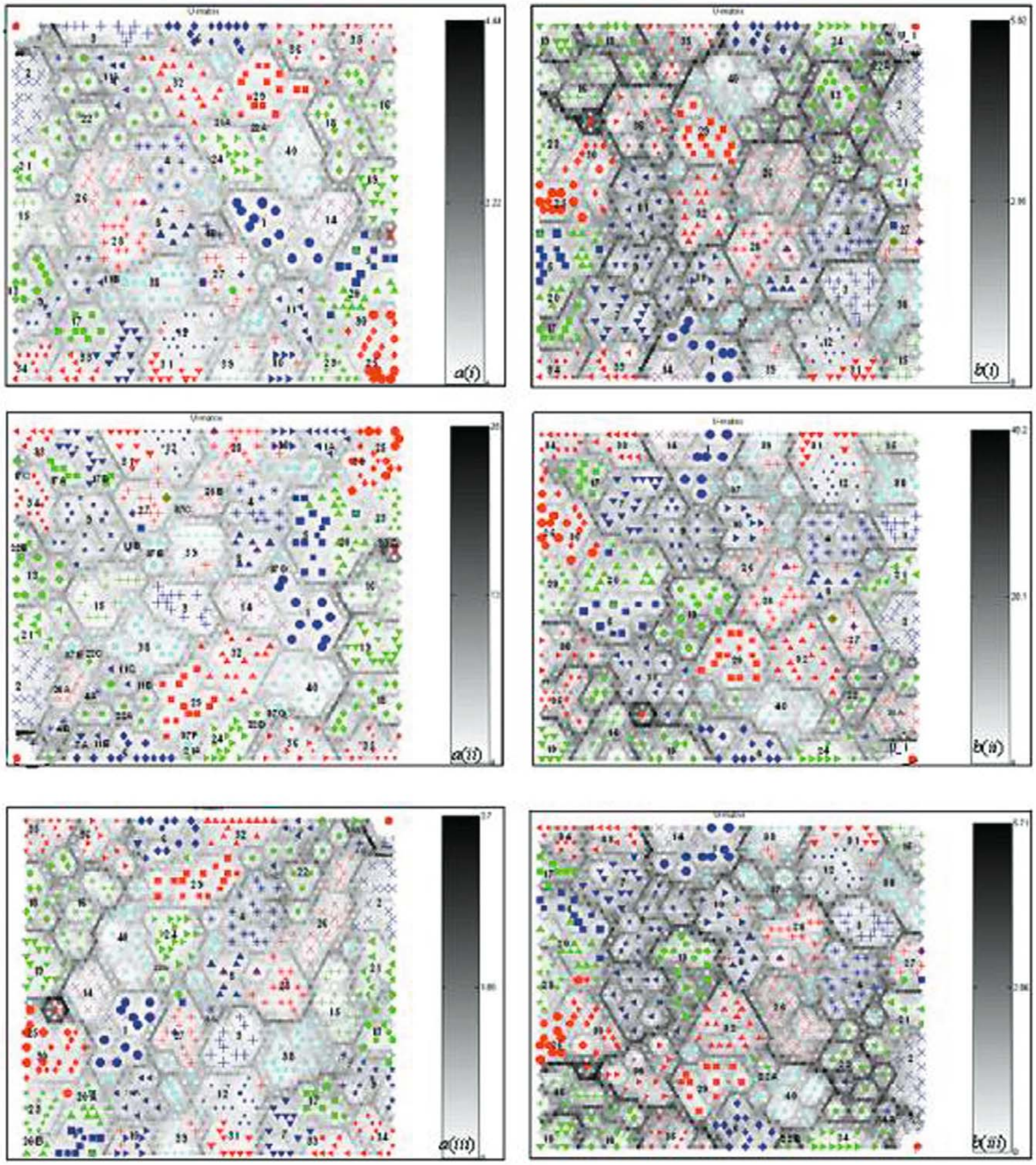

Figure 3. U-matrix visualization of a self-organizing maps (SOMs) showing the relatedness of the organisms and the groups of organisms using tetra-nucleotide (a(i)–a(iii)) and penta-nucleotide data (b(i)–b(iii)) using three methods SOM, PCA–SOM and CCA–SOM respectively. The numbers indicate the bacterial genera as referred in table 1. The polygons represent the territory for different bacterial genera. The organisms from the same genus but placed in different polygons are indicated by letters suffixing the genus identification number. The gray colour scale represents the magnitude of distance between the two adjacent groups.

The data on the number of occurrences (frequencies) of various tetra- and penta-nucleotides for each sequence were generated for the analysis. It may be noted that there are 256 (=4^4) and 1024 (=4^5) possible combinations for tetra- and penta-nucleotides, respectively. Thus, the tetra- and penta-nucleotide frequency data for each of the 800 sequences resulted into two data matrices of dimensions 800×256 (set I) and 800×1024 (set II), respectively. The frequency

Table 1. Bacteria selected from the effluent treatment plant

S. No.	Bacteria	S. No.	Bacteria	S. No.	Bacteria	S. No.	Bacteria
1.	*Acetobacter*	11.	*Desulfovibrio*	21.	*Moraxella*	31.	*Salmonella*
2.	*Acinetobacter*	12.	*Enterobacter*	22.	*Methylococcus*	32.	*Sphingomonas*
3.	*Aeromonas*	13.	*Flavobacterium*	23.	*Mycobacterium*	33.	*Staphylococcus*
4.	*Alcaligenes*	14.	*Gluconobacter*	24.	*Nitrobacter*	34.	*Streptococcus*
5.	*Arthobacter*	15.	*Haemophilus*	25.	*Nocardia*	35.	*Sulpholobus*
6.	*Azospirillum*	16.	*Halobacterium*	26.	*Nitrosomonas*	36.	*Thermus*
7.	*Bacillus*	17.	*Lactobacillus*	27.	*Pseudomonas*	37.	*Thiobacillus*
8.	*Burkholderia*	18.	*Methanococcus*	28.	*Ralstonia*	38.	*Vibrio*
9.	*Clostridium*	19.	*Methanosarcia*	29.	*Rhizobium*	39.	*Xanthobacter*
10.	*Commamonas*	20.	*Micrococcus*	30.	*Rhodococcus*	40.	*Xanthomonas*

values in each column of set I and II were separately normalized between 0 and 1, and the normalized data sets were used for SOM-based classification and projection. The architecture of SOM has been depicted in figure 2 and described in Methods. The classification analysis was done separately for tetra- and penta-nucleotide frequency data. In this study, three different approaches have been used for the classification and visualization of 16S sequence data: (i) SOM was used directly for the classification, (ii) dimensionality of the data sets was reduced using the linear PCA and the dimensionally-reduced data was used in SOM and (iii) dimensionality of the frequency data was reduced using the CCA and SOM was used subsequently for the classification. The CCA is a self-organizing neural network that performs two tasks, i.e. VQ and non-linearly projecting the quantized vectors onto the output space.

For tetra-nucleotides-based classification, the normalized matrix of dimension 800×256 was considered in SOM analysis. The grouping of organisms and the position of different groups obtained are as shown in figure 3a(i), wherein each point in the map represents an organism. Next, the linear relationships among the frequencies of various tetra-nucleotides were extracted using the standard PCA. It was observed that the first 100 latent variables (principal components) could capture nearly 97% variability of the original tetra-nucleotide frequency data. This suggests that the PCA could reduce the data dimensionality from 256 variables to 100. PCA-reduced data of dimensions 800×100 was then used in SOM-based classification, and the results obtained thereby are shown in figure 3a(ii). Further, for extracting non-linear relationships among the variables, CCA was performed on the same data matrix (800×256). It was observed that 80 non-linear principal components could capture approximately 99% variance of the original data. This indicated the existence of non-linearity among variables, and as a result CCA could impart more reduction in the dimensionality (i.e. reduction from 256 to 80 variables) as compared to the linear PCA (from 256 to 100 variables). Thereafter, SOM-based classification was performed using the CCA-reduced data set and the results were obtained are shown in figure 3a(iii).

Similar to the tetra-nucleotide frequency data, the analysis was performed using the penta-nucleotide frequency data of dimensions 800×1024. The classification results obtained using the above-stated three methodologies (i.e. SOM, PCA–SOM and CCA–SOM) have been shown in figures 3b(i)–b(iii), respectively. In the PCA–SOM method, PCA reduced the dimensionality of data from 1024 frequency variables to 275 principal components, accounting for 99% of variance of the original data; whereas in the CCA–SOM method, CCA reduced the dimensionality to 382 non-linear principal components, retaining nearly 99% of variation of the original data. Thus, the linear correlations were found predominant in the penta-nucleotide data, and therefore, the linear PCA could impart more reduction in dimensionality (1024 to 275) as compared with that of CCA (1024 to 382). It is pertinent to mention here that the *16S rRNA* gene is highly conserved across the species of particular genus, and for few species, the frequency of tetra- or penta-nucleotides might coincide, and hence, in some cases, there exists an obvious overlap of points within the polygon. The boundary of each polygon is indicated by gray pixels with varying intensities. It is also seen that the demarcation of polygons is much clearer for the penta-nucleotides as compared with the tetra-nucleotides, indicating that the organisms and their groups could be better distinguished on the basis of the penta-nucleotide frequencies. Moreover, the chances of misclassification of organisms are higher on the basis of tetra-nucleotide frequencies as compared with penta-nucleotide frequencies, because larger patterns consistently observed in sequences of particular genus provide better specificity than that of smaller patterns. Table 2 shows the percentage of correct classification of organisms obtained using the three approaches for the tetra- and penta-nucleotide data, respectively. These percentages have been deduced from 780 sequences belonging to 39 groups. For one group, i.e.

Thiobacillus, the points exhibited a high scatter irrespective of the method used. The reason for such scatter has been discussed later. Therefore, in the overall analysis we have omitted this genus while estimating the classification accuracy. Table 2 also reveals that the grouping of sequences is somewhat better for the penta-nucleotide frequencies as compared with tetra-nucleotides. Hence, the results obtained with penta-nucleotide frequencies were considered for further analysis.

The table also shows that when using penta-nucleotides, all the three approaches, viz. SOM, PCA–SOM and CCA–SOM, resulted into almost the same classification accuracy for the training data set, indicating the existence of significant linear relationships among the frequencies of penta-nucleotides. In order to validate the three classification methods and assess their generalization performance, a test set of 125 known 16S sequences belonging to the selected genera were retrieved from the NCBI GenBank.

Table 2. Classification accuracy of the training set using three different approaches

	SOM		PCA–SOM		CCA– SOM	
Method	Penta-mer	Tetra-mer	Penta-mer	Tetra-mer	Penta-mer	Tetra-mer
Sequences classified correctly out of 780 (%)	729 (93.46%)	712 (91.28%)	727 (93.20%)	721 (92.43%)	723 (92.69%)	714 (91.53%)

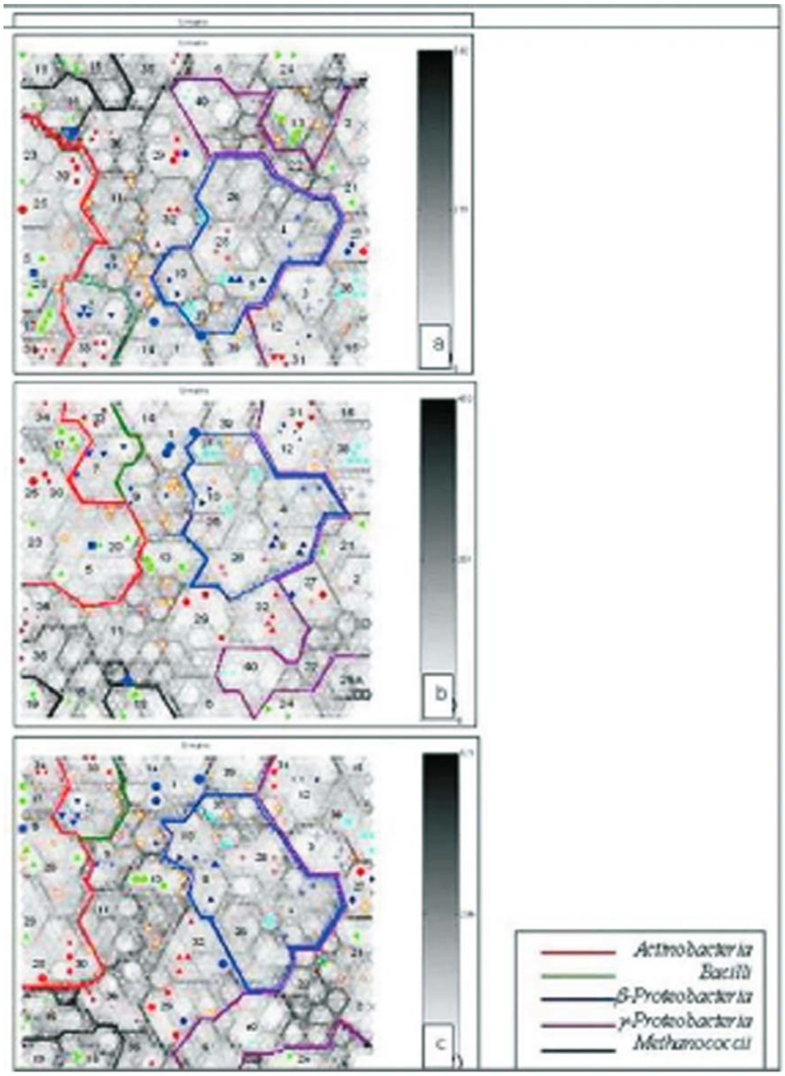

Figure 4. The grouping of test set of organisms based on the penta-nucleotide frequencies (a–c) using SOM, PCA–SOM and CCA–SOM, respectively. The marked areas indicate various classes of taxonomic hierarchy.

Interestingly, the classification accuracy for this data was found to be more than 95% (figures 4(i)–(iii)). This suggests the excellent generalization capability of SOM-based classification.

4. Discussion

We have earlier reported that on the basis of di- and tri-nucleotide frequencies, we could discriminate closely related sequences (Raje *et al*. 2002). However, when we applied the same strategy with sequences retrieved from different group of bacteria, the sequences could not be grouped even after using the tools developed in this study. This prompted us to use tetra- and penta-nucleotide features for classification to arrive at class-/group-specific features.

A closer inspection of the locations of different bacterial genera using the penta-nucleotide frequencies data led to some interesting findings. Figure 5 shows the groupings of 32 organisms taking into consideration the similarity of taxonomic hierarchy of organisms. For instance, *Methanococcus* and *Methanosarcia* (No. 18, 19 in table 1) have same hierarchy up to the "Class" level, but they split at the "order" level, while genera *Acetobacter* and *Gluconobacter* (No. 1, 14) share same hierarchy up to the "Family" level. The placing of these 32 genera was observed in the maps generated through each of the three approaches using penta-nucleotide data. The organisms within a group (figure 4) are mostly found adjacent to each other and this finding was consistent with all the three approaches. A closer inspection of the groups showing similarity up to the "Family" level (figure 4(i)–(iii)) reveals that the pixel intensities between the genera *Acetobacter* (1) and *Gluconobacter* (14), *Arthobacter* (5) and *Micrococcus* (20), *Enterobacter* (12) and *Salmonella* (31), *Nocardia* (25) and *Rhodococcus* (30), *Acinetobacter* (2) and *Moraxella* (21) were found to be much lower irrespective of the approach used. Such a close adjacency suggests that the penta-nucleotide distribution of the stated pairs of genera is necessarily similar but sufficiently distinct to avoid misclassification of the respective organisms. Similarly, the adjacency was observed for other groups of organisms showing taxonomic similarity up to the "Order" level or "Class" level; however, the pixel intensities are somewhat higher than in the earlier case. In this regard, it becomes interesting to ensure through some more case studies, using different organisms, whether the level of taxonomic relatedness between the organisms has any bearing on the pixel intensity of the boundary separating the organisms.

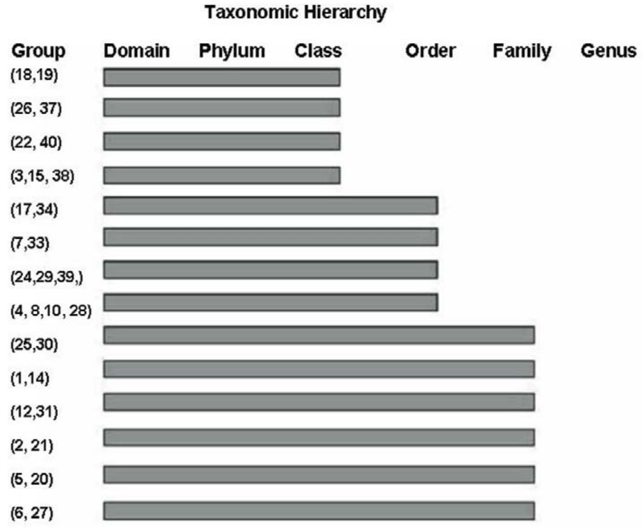

Figure 5. Taxonomic classification of the closely related bacteria based on the Bergey manual. The horizontal bars indicate the taxonomic level up to which the bacterial genera, indicated by numbers, show identity.

When the groups of organisms were formed at the "Class" level, there was an interesting finding that was consistent irrespective of the approach used. Table 3 shows the grouping of organisms based on "taxonomic class", and figure 4(i)–(iii) shows the territory for different classes, which includes the selected genera. A definite topological arrangement of taxonomic classes is seen in the panels of figure depicting the evolutionary trend of organisms. The left most area of the map is occupied by all *Actinobacteria* and *Methanococii*, which are ancient bacteria on the evolutionary scale. The *Bacilli* is adjacent to *Actinobacteria*, whereas the central area and the right most area is occupied by *Betaproteobacteria* and *Gammaproteobacteria* respectively, which imitate the evolutionary trend.

Another important observation was about genus *Thiobacillus* (No. 37). The sequences belonging to this genus exhibited a wide scatter; hence, the misclassified sequences were investigated for their non-memberships to genus *Thiobacillus*. Some of the sequences belonging to this

Table 3. Taxonomic class of the selected bacterial genera

S. No.	Taxonomic class	Bacterial genera*
1.	Actinobacteria	(5, 17, 20, 23, 25, 30, 34)
2.	*α*-Proteobacteria	(1, 14, 24, 29, 32, 39)
3.	*β*-Proteobacteria	(4, 8, 10, 26, 28, 37)
4.	*γ*-Proteobacteria	(2, 3, 6, 12, 15, 21, 22, 27, 31, 38, 40)
5.	Bacilli	(7, 33)
6.	Methanococci	(18, 19)
7.	Thermoprotei	(35)
8.	Halobacteria	(16)
9.	Flavobacteria	(13)
10.	Deinococci	(36)
11.	Clostridia δ-	(9)
12.	Proteobacteria	(11)

*The numbers in the parenthesis indicate the bacterial genera as mentioned in table 1.

genus showed nice grouping under the *Betaproteobacteria* class as evident in figure 4, while four sequences were placed at different locations falling in the *Alphaproteobacteria* region in the maps. *Thiobacillus acidophilus* (Acc. No. D86511), which is one of such sequences, exhibited good similarity with genus *Acidiphilium.* Another misclassified sequence (Acc. No. D32238) revealed that it belongs to genus *Paracoccus* and the sequence exactly matched with *Paracoccus alcaliphilu.* The sequence was earlier named as *Thiobacillus versutus* belonging to genus *Thiobacillus.* However, Katayama *et al.* (1995) noticed that the sequence resembles more to genus *Paracoccus*. The SOM-map also revealed that the species does not fall into *Thiobacillus* lattice. Further, similar result was obtained for sequence (Acc. No. D32241), which earlier was referred as *Thiobacillus versutus;* but now has been classified as belonging to genus *Paracoccus* and exactly matching with *Paracoccus kocurii.* Also the sequence with accession numbers D32247 belonged to genus *Starkeya.* All these sequences belonging to genus *Acidiphilium, Paracoccus* and *Starkeya* belonged to *Alphaproteobacteria*. The placing of these sequences in SOM-maps is also in the *Alphaproteobacteria* region, demonstrating the SOM capabilities based on penta-nucleotide sequence features.

Further, we tested seven unknown sequences (unculturable isolates) obtained from few effluent treatment plants. The 16S rDNA sequence clones of these isolates showed significant similarity with *unidentified bacterium* through BLAST. Equivalently, we obtained the classification of these isolates using the above three approaches using penta-nucleotide frequencies. Although these organisms are outside the set of selected 40 genera, we could predict their taxonomic classes using SOM analysis. The scatter of these isolates and their clustering in different taxonomic classes using the three approaches has been carried out. There was absolute consensus of methods for three isolates (No. 2 (DQ309369), 6 (DQ309372), 7 (DQ309373)) in regard to the taxonomic class, i.e. *Alphaproteobacteria*. For two other isolates (No. 3 (DQ309370), 4 (DQ309371)), the taxonomic class suggested by PCA–SOM agreed with CCA–SOM. For isolate No. 5 (DQ309379), the results of PCA and SOM matched with SOM, whereas for isolate No. 1 (DQ309368), there was no consensus amongst the methods. Since, the linear correlations are predominant in the penta-nucleotide frequency data, the classification suggested by PCA and SOM was relied upon in this study. So referring to this classification, it is possible to predict the belongingness of such unknown sequences at least at the "Class" level, which otherwise is difficult to know through the BLAST results.

Thus, a sample study with 40 genera revealed that sequence characterization, data reduction and data visualization approaches using abstract information on patterns could support the reported taxonomic knowledge. The relatedness of organisms could be displayed through elegant graphics, which is a novel representation. Another resultant of the analysis is the set of features (penta-nucleotides) that distinguish the selected organisms with higher accuracy. This opens further areas of research such as genus-specific features could be identified with due regard to their frequencies and could be used to generate regular expressions. The specificity of the regular expressions to genus could be tested using matching algorithms against the 16S rDNA database. The expressions yielding high specificity and sensitivity could act as signature to the genus. Also the features could be used to develop organism specific PCR primers for their rapid identification. Moreover, the variable regions between features could also be used as a seed to develop genus-specific signatures (Purohit *et al.* 2003; Liskiewicz *et al.* 2004).

5. Conclusions

In essence, this study describes a classification mechanism that provides a unique visual representation of the taxonomic relationships. Such a classification system could be developed spanning different taxonomic classes by using sample sequences from each of the classes. This could be an alternative approach to ascertain the membership of new sequence at least at the "Class" level in addition to the alignment-based approach. Today it is believed that nearly 95% of microbial diversity is still unknown (Torsvik *et al.* 2002). In the years to come, with the growing size of 16S rDNA database, the use of such tools in conjunction with the conventional classification tools could strengthen our understanding of relatedness of existing and the unexplored microbial wealth in nature.

Acknowledgements

This research was supported by the CSIR SIP16:4.1 Program, CSIR, New Delhi.

References

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T and Ikemura T 2003 Informatics for unveiling hidden genome signatures; *Genome Res.* **13** 693–702

Amann R, Ludwig W and Schleifer K 1995 Phylogenetic identification and in situ detection of individual microbial cells without cultivation; *Microbiol. Rev.* **59** 143–169

Buchala S, Davey N, Frank R J, Gale T M, Loomes M J, and Kanargard W 2004a *Gender classification of face images, 763–768. The role of global and feature-based information* (INCONIP)

Buchala S, Davey N, Frank R J and Gale T M 2004b Dimensionality reduction of face images for gender classification; *Intelligent*

Systems, 2004; *Proceedings. 2004 2nd International IEEE Conference*, Volume: 1, pp 88–93

Demartines P and Herault J 1997 Curvilinear component analysis: A self-organizing neural network for non-linear mapping of data sets; *IEEE Trans. Neural Network* **8** 148–154.

Durbin R, Eddy S, Krough A and Mitchinson G 1998 *Biological sequence analysis* (Cambridge: Cambridge University Press)

Garrity G M, Winters M and Searles D B 2001 *Taxonomic outline of prokaryotic genera-Bergey's Manual of systematic bacteriology, second edition* (New York: Springer-Verlag)

Hugenholtz P, Goebel B M and Pace N R 1998 Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity; *J. Bacteriol.* **180** 4765–4774

Katayama Y, Hiraishi A and Kuraishi H 1995 Paracoccus thiocyanatus sp. nov., a new species of thiocyanate-utilizing facultative chemolithotroph, and transfer of Thiobacillus versutus to the genus Paracoccus as Paracoccus versutus comb. nov. with emendation of the genus; *Microbiology* **141** 1469–1477

Kasturi J, Acharya R and Ramanathan M 2003 An information theoretic approach for analyzing temporal patterns of gene expression; *Bioinformatics* **19** 449–458

Kohonen T 1990 The self-organizing map; *Proc. IEEE* **78** 1464–1480

Kruskal J B 1964 Multidimensional scaling by optimizing goodness of a fit to a non metric hypothesis; *Phychometrica* **29** 1–27

Liskiewicz M, Purohit H J and Raje D V 2004 Relation of residues in the variable regions of 16S rDNA and their relevance to genus specificity; *Lect. Notes Comp. Sci.* **3240** 362–373

Maidak B L, Larsen N, McCaughey M J, Overbeek R, Olsen G J, Fogel K, Blandy J and Woese C R 1994 The ribosomal database project; *Nucleic Acids Res.* **17** 3485–3487

Purohit H J, Raje D V and Kapley A 2003 Identification of signature and primers specific to genus *Pseudomonas* using mismatched patterns of 16S rDNA sequences; *BMC Bioinformatics* **4** 19

Raje D V, Purohit H J and Singh R S 2002 Distinguishing features of 16S rDNA gene for five dominating bacterial genus observed in bioremediation; *J. Comp. Biol*. **9** 819–829

Roweis S and Saul L 2000 Nonlinear dimensionality reduction by locally linear embedding; *Science* **290** 2323–2326

Sammon J W 1969 A nonlinear mapping algorithm for data structure analysis; *IEEE Trans. Comput***. C-18** 401–409

Schneider G 1999 How many potentially secreted proteins are contained in bacterial genome; *Gene* **237** 113–121

Tenenbaum J, de Silva V and Langford J 2000 A global geometric framework for nonlinear dimensionality reduction; *Science* **290** 2319–2323

Torsvik V, Ovreas L and Thingstad T F 2002 Prokaryotic diversity--magnitude, dynamics, and controlling factors; *Science* **296** 1064–1066

Vasanto J, Alhoniemi E, Himberg J, Kiviluto K and Parvinainen J 1999 Self-organising map for data mining in Matlab: the SOM Toolbox; *Simulation News (Europe)* 25–54 *(http://www.cis.hut.fi/project/somtoolbox)*

Ward D M, Weller R and Bateson M M 1990 16S rRNA sequences, reveal numerous uncultured microorganisms in a natural community; *Nature (London)* **345** 63–65

Wang H C, Badger J, Kearney P and Li M 2001 Analysis of codon usage patterns of bacterial genomes using the self-organizing map; *Mol. Biol. Evol.* 18 792–800

Woese C R 1987 Bacterial evolution; *Microbiol. Rev*. **51** 221–271

Wold S, Esbensen K and Geladi P 1987 Principal component analysis; *Chemo. Intell. Lab. Syst*. **2** 37–52

MS received 14 October 2009; accepted 8 September 2010

*e*Publication: 29 October 2010

Corresponding editor: Vidyanand Nanjundiah