Sequence analysis

Hyperbolic SOM-based clustering of DNA fragment features for
taxonomic visualization and classification

Christian Martin¹,², *, Naryttza N. Diaz¹,², Jörg Ontrup¹ and Tim W. Nattkemper¹
¹AG Applied Neuroinformatics, Technical Faculty, Bielefeld University and ²Graduate School Bioinformatics and
Genome Research, CeBiTec, Bielefeld University, Bielefeld, Germany

Received on February 22, 2008; revised on May 16, 2008; accepted on June 3, 2008
Advance Access publication June 5, 2008
Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Modern high-throughput sequencing technologies
enable the simultaneous analysis of organisms in an environment.
The analysis of species diversity and the binning of DNA
fragments of non-sequenced species for assembly are two major
challenges in sequence analysis. To achieve reasonable binnings
and classifications, DNA fragment structure has to be represented
appropriately, so it can be processed by machine learning algorithms.

Results: Hierarchically growing hyperbolic Self-Organizing maps
(H²SOMs) are trained to cluster small variable-length DNA fragments
(0.2–50 kb) of 350 prokaryotic organisms at six taxonomic ranks
Superkingdom, Phylum, Class, Order, Genus and Species in the
Tree of Life. DNA fragments are mapped to three different types
of feature vectors based on the genomic signature: basic features,
features considering the importance of oligonucleotide patterns as
well as contrast enhanced features. The H²SOM classifier achieves
high classification rates while at the same time its visualization allows
further insights into the projected data and has the potential to
support binning of short sequence reads, because DNA fragments
can be grouped into phylogenetic groups.

Availability: An implementation of the H²SOM classifier in
Matlab is provided at www.techfak.uni-bielefeld.de/ags/ani/projects/
H²SOMSeqData

Contact: christian.martin@uni-bielefeld.de

Supplementary Information: Supplementary data are available at
Bioinformatics online.

1 INTRODUCTION

The emerging field of metagenomics allows, for the very first time,
to study the collective genomes (metagenomes) of microbes in
free-living microbial communities (Handelsmann et al., 2007). In
contrast to traditional research techniques that rely on laboratory
cultures and thus are only able to access around 1% of all microbes,
metagenomics enables to assess the complete species diversity and
genome variability in a natural environment. Besides a deeper
understanding of life, metagenomics offers great opportunities in
medicine as well as in biotechnology, where the exact profile of
organisms in a sample is of great interest. Two major developments
have cleared the way to metagenomics: first, novel methods
have been developed that bypass cultivation and thus enable to
sequence ≈ 99% of microbes that resisted all culturing efforts
so far. Second, the cost and time per sequenced base pair
decreases exponentially due to advances of the standard Sanger
sequencing technology (Sanger et al., 1997) as well as by the most
recently developed high-throughput 454 pyrosequencing technique
(Margulies et al., 2005).

The simultaneous study of a large number of species on a genomic
level demands for new algorithmic approaches to process the high
amount of sequenced DNA fragments. One major task is to assign
a DNA fragment to its originating species. Such a classification
would be a crucial step in the analysis of an environment on an
organismal level, leading to a more complex way of describing and
comparing environmental states. Another metagenomic challenge
results from the observation that many species have not yet been
sequenced, leaving us with a narrow sample as reference. To
obtain contigs from metagenomic data, a binning of the sequence
reads is helpful for assembly. We address this binning task by
classifying sequence reads to predefined categories at a specific
phylogenetic rank. The most challenging task is to classify them
at species rank, but a classification at a higher phylogenetic ranks
is also helpful. Genus, Order, Class, Phylum and Superkingdom
represent the most commonly used taxonomic ranks (Supplementary
Information 1).

In order to estimate phylogenetic differences between certain
species of interest, two approaches have been developed.

The traditional approach is based on sequence similarity
(homology) taking into account only part of the genome (i.e. a single
gene or a marker gene). The major drawback for the detection of
marker genes is the need for sequence alignment, a technique for
which many pitfalls can be encountered (Li, 1997), sampling of
representative sequences, lateral gene transfer or recombination are
among the most common. Other attempts to estimate phylogenetic
distance involve the use of ribosomal RNA molecules, gene
content, gene order, protein domain content, etc. Moreover, multiple
sequence alignments are computationally very expensive.

The second approach is alignment-free and is based on nucleotide
composition (Abe et al., 2006; McHardy et al., 2007; Teeling et al.,
2004). The information is directly computed from the nucleotide
sequence of genomes or non-assembled fragments disregarding prior
identification of functional regions which are taken into account
(Chapuis et al., 2005; Gupta, 1998; Qi et al., 2004; Stuart et al., 2002).
Such methods have the potential to classify even short genomic
sequences and help in the evaluation of trends in microbial diversity,
novel genes and genome evolution. A powerful characteristic that enables to capture evolutionary relationships between species and permits a classification of DNA fragments is the genomic signature which is defined as a set of short oligonucleotide patterns in a sequence (Karlin and Burge, 1995). Genomic signatures are species-specific and can be computed from complete genomes, but also from variable-length DNA fragments in any part of the genome (Deschavanne et al., 1999). A genomic signature can be regarded as a feature vector that can be processed by any data mining or machine learning algorithm. McHardy et al. (2007) applies the concept of genomic signature to train a support vector machine (SVM) on fragments of 100 kb in order to classify fragments of ≥1 kb of 340 species at the taxonomic ranks Domain, Phylum, Class, Order and Genus with high accuracy.

The classifier that performs the task of classifying sequence reads to support binning at a specific phylogenetic rank that should cover the following three aspects: first, it should be accurate in classification. Second, it should be scalable to process the large datasets that are present in metagenomics. Third, it should enable a data visualization to provide means for inspection and for feature space exploration, a feature that is usually not provided by sophisticated classifiers in the machine learning or data mining domain. Most black box classifiers such as SVMs do only address the first and sometimes the second of these requirements.

To cover all aspects of accurate classification, scalability and data visualization, we propose to process feature vectors representing genomic fragments by a special variant of Kohonen’s self-organizing map (SOM, Kohonen, 1990), the hierarchically growing hyperbolic SOM (H²SOM). In genomic studies, the standard SOM has already been applied very successfully to classify environmental DNA fragments with length of 10 kb or more (Abe et al., 2002, 2005, 2006, 2007). The standard SOM projects the data to a 2D flat Euclidean space, a space that might not correspond to the intrinsic structure of genomic data. It may be more likely that genomic sequences are structured hierarchically in the same way as their corresponding species are grouped into kinships relations as represented in the Tree of Life. Hierarchically organized data grows exponentially and thus requires a mapping into a geometric space with corresponding behavior. To this end, the hyperbolic SOM (HSOM) applies a tree-like grid in hyperbolic space. It has already been successfully applied in text mining (Ontrup and Ritter, 2006) and first successes have also been reported for the embedding of high-dimensional feature vectors representing genomic sequences (Martin et al., 2007). A special ultra-fast hierarchical training scheme is applied in the H²SOM, which enables to deal with the exponentially growing amount of data in metagenomic studies.

In order to further decrease the required minimum length of DNA fragments for a proper classification, three different representations (feature vectors) of genomic sequences and DNA fragments based on the genomic signature are evaluated: in a first experiment, frequencies of oligonucleotide patterns are directly taken as features, and are combined to a feature vector for each complete genome or DNA fragment. In a second approach, features are developed that take the general importance of each oligonucleotide pattern into account. These features are similar to the *term frequency – inverse document frequency* (tf-idf) features, successfully applied in text mining (Salton et al., 1975). In a third approach, features containing the enhanced contrast between over- and under-represented patterns are evaluated.

In our study we use genomic sequences of 350 prokaryotic organisms, which represents a vast majority of sequenced organisms from the two domains Archaea and Bacteria. DNA fragments of 0.2–50 kb are classified to the most commonly used taxonomic categories at the ranks Superkingdom, Phylum, Class, Order, Genus and Species. Thereby, we address the classification of DNA fragments sequenced by the Sanger technology or 454 pyrosequencing as well as assembled contigs. Feature vectors obtained from complete genomic sequences or DNA fragments are used to train a H²SOM. Subsequently the trained model is used as classifier to classify DNA fragments of shorter lengths. It is shown that a phylogenetic classification of DNA fragments of 0.2–50 kb is possible with a high accuracy in a H²SOM framework for all taxonomic ranks considered. At the same time, a visualization of the projected data permits further insights into the data and provides an intuitive browsing through DNA fragment clusters to support the binning process prior to assembly.

2 MATERIAL AND METHODS

2.1 Data

In this study we considered 350 different species (155 Genera, 66 Orders, 34 Classes, 18 Phyla) from either the Bacteria or Archaea superkingdom. This palette of species represents a vast majority of the microbial world sequenced up-to-date. The complete set of corresponding genomes, ranging from 432 kb to 9660 kb, were obtained from the SEED database1 (Overbeek et al., 2005). The complete taxonomic information from the set of species evaluated in this study were obtained from the taxonomy database located in the US National Center for Biotechnology Information (NCBI) (Wheeler et al., 2002).

The considered 350 species are categorized on the most commonly used taxonomic ranks Superkingdom, Phylum, Class, Order and Genus in the Tree of Life (Supplementary Information 1). On the first rank, the species are categorized in the two different superkingdoms Archaea and Bacteria. On the second rank, all species of the superkingdom Archaea are subdivided in the three phyla Crenarchaeota, Euryarchaeota and Nanoarchaeota. All species of the superkingdom Bacteria are subdivided in 15 different phyla (from Actinobacteria to Thermotogae). Finer subdivisions are obtained by the categories Class, Order and Genus on the ranks three to five. Each genus comprises at least one species.

2.2 Feature vectors

Each complete genomic sequence and each DNA fragment (genomic subsequence) is mapped to a feature vector. Its entries are computed from the numbers of oligonucleotide patterns of length k in a sliding window of step-size one in the forward and reverse DNA strand.

Let Σ be the alphabet of nucleotides Σ = {A, C, G, T}, and let k be an oligonucleotide pattern of length k ∈ Σ. This results in |Σ|km = 4k possible oligonucleotide patterns of length k, e.g. an oligonucleotide pattern of length k = 4, as used in our experiments, can be one of the following sequences: 4[1] = AAAA, 4[2] = AAAC, ..., 4[4] = TT TT. Let 4(j) be the genomic sequence of species j (with 1 ≤ j ≤ 350) of length |4[j]| each and 4[0] := Σ.

To find an appropriate representation for the genomic sequences and subsequences, three different types of feature vectors are computed: the *term frequency* (tf) feature vector directly consists of the frequencies of oligonucleotide patterns of a specified length in a sequence. Let 4(j)[k] be the number of the j-th oligonucleotide (with j = 1, ..., 4) in sequence 4(j).

The tf feature vector of the genomic sequence of species \( l \) is defined as
\[
\text{tf}_l \left( s^i \right) = \begin{bmatrix} t_f(1) & \cdots & t_f(l) \end{bmatrix}^T
\] (1)

The term frequency-term importance (tf-idf) feature vector is based on the tf feature vector, but considers the general importance of oligonucleotide patterns. This approach is similar to the tf-idf features successfully applied in the textmining domain (Salton et al., 1975). The traditional tfIdf measuring the fraction of documents in which a certain term is contained is not adequate to analyze genomic fragments, because oligonucleotide patterns up to length \( k = 6 \) are contained in almost every sequence at least once, resulting in an equal importance for each pattern. We propose to measure the importance of an oligonucleotide pattern by increasing the feature values of those patterns that

1. are rare among all species and
2. occur more frequently than other patterns in the considered species.

Let \( t_f = \sum_{s} t_f^i(1) \) be the number of the \( j \)-th oligonucleotide among all species. Let \( t_f^i = \sum_{s} t_f^i(1) \) be the number of all oligonucleotides in genomic fragment \( l \). The tf-idf feature vector of the genomic sequence of species \( l \) is defined as
\[
\text{idf}_l \left( s^i \right) = \begin{bmatrix} t_f^i(1) & \cdots & t_f^i(l) \end{bmatrix}^T
\] (2)

The oligo feature vector contains the enhanced contrast between over- and under-represented oligonucleotide patterns of length \( k \) given by the ratio of observed versus expected (Martin et al., 2007). For notation simplicity we consider only one sequence \( s \in \Sigma \) in the following.

For a sequence \( s \), the probability to observe a certain nucleotide \( \eta \in \Sigma \) can be computed by
\[
\rho(\eta) = \frac{1}{|s|} \sum_{t \in s} q(\eta, t)
\] (3)

with the indicator function
\[
q(\eta, t) = \begin{cases} 1 & \text{if } t = \eta \\ 0 & \text{else} \end{cases}
\] (4)

The expectation value for a certain oligonucleotide \( o \) in the sequence \( s \) can be estimated by
\[
E(o) = \frac{|s|}{|\Sigma|} \sum_{t \in s} \rho(t, o)
\] (5)

Let \( O \) be the number of observed oligonucleotides \( o \) in the same sequence \( s \). The contrast is performed by computation of the score
\[
g(o) = \begin{cases} 0 & \text{if } O(o) = 0 \\ \frac{O(o) - E(o)}{E(o)} & \text{if } O(o) > E(o) \\ \frac{O(o) - E(o)}{O(o)} & \text{if } O(o) < E(o) \end{cases}
\] (6)

The oligo feature vector of the sequence \( s \) contains the scores of all possible oligonucleotides:
\[
\text{oligo}_s = g(o(1)), g(o(2)), \ldots, g(o(|O|))
\] (7)

For further processing in learning algorithms, a normalization of the feature vectors is of decisive importance. Each type of feature vector works best with its specific normalization strategy (Supplementary Information 2): the tf features work best when the elements of each feature vector are scaled to unit variance, the tf-idf features when each component among all vectors is scaled to unit variance. No normalization is needed. The specific best normalization strategies are applied in our study.

In our experiments, a dataset either consists of 350 feature vectors, each encoding one complete genomic sequence, or of 2800 feature vectors, computed from eight\(^2\) disjoint genomic subsequences of either 0.2, 0.5, 1, 3, 5, 10, 15 or 50 kb of each of the 350 genomic sequences. Even though all 350 genomic sequences are used to generate feature vectors for the training and the testing dataset in each scenario, each computed feature vector is either contained in the training or the testing dataset, but never in both. A separation of the 350 genomic sequences into training and testing sequences might be an interesting approach at higher ranks, but it is infeasible at rank Species or Genus. A correct classification of a species or genus (which often only consists of one species) becomes impossible if the only existing species has been omitted during training.

### 2.3 The hyperbolic self-organizing map

Since the introduction of the SOM (Kohonen, 1990), it has become a widely used tool for exploratory data analysis, classification and visualization. Typically, the SOM projects the data to a 2D flat Euclidean space. However, this does not always correlate with the intrinsic structure of the considered data. Especially for hierarchically structured data, an exponentially growing display is more adequate, a property offered by hyperbolic space. Its uniform negative curvature results in a geometry such that the size of a neighborhood around any point increases exponentially with its radius \( R \). In a HSOM this exponential scaling property has already successfully been used to visualize high-dimensional text data (Ontrup and Ritter, 2001). The core idea of the HSOM is to employ a grid of nodes in the hyperbolic plane \( \mathbb{R}^2 \) which is then projected onto the \( \mathbb{R}^2 \) for inspection. The regular structure of formal neurons used by the HSOM is based on a tessellation of \( \mathbb{R}^2 \) with equilateral triangles (Ritter, 1999).

The HSOM is then formed in the standard self-organizing manner: each lattice node \( r \) is associated to a prototype vector \( w_r \in \mathbb{R}^2 \) from some \( D \)-dimensional feature space. During the learning phase, in each training step a best match node \( s \) is determined for a given input vector \( x \) by \( s = \arg \min |w_r - x| \).

The prototype vectors are then adjusted according to the familiar rule
\[
\Delta w_r(x, s) = \begin{cases} \eta(x, s(x \times w_r)) & \text{if } r(s) \neq r(s(x)) \\ 0 & \text{else} \end{cases}
\] (8)

with
\[
h(r, s(x)) = \exp \left( -\frac{d^2(r, s(x))}{2 \sigma^2} \right)
\] (9)

where \( h(r, s(x)) \) is a Gaussian shaped function centered at the winner node \( s(x) \) that decays with increasing node distance \( d^2(r, s(x)) \) on the hyperbolic lattice.

#### 2.3.1 Hierarchically growing HSOM

The H2SOM employs the same sort of regular lattice structure already used for the plain HSOM, but offers a hierarchically growing scheme: the HFSOM is initialized with the root node of the hierarchy placed at the origin of \( \mathbb{R}^2 \). Then the \( n_0 \) children nodes of the first sub-hierarchy are equidistantly placed around the center node as shown in Figure 1a. The radius of the first ring is chosen such that the hyperbolic distance of the first-level nodes to each other is the same as their distance to the center node. The ‘branching’ factor \( n_0 \) determines how many nodes are generated at each level and how ‘fast’ the network is reaching out into the hyperbolic space. \( n_0 \) is lower bounded by 7, but has no upper bound (Ontrup and Ritter, 2006). During a first phase, the top-level ring of nodes is trained in the standard self-organized fashion. After a fixed training interval, each node in the periphery is expanded as indicated in Figure 1b and their reference vectors become fixed. In a new learning phase adaptation ‘moves’ to the nodes of the new hierarchy level. This scheme is repeated until a desired hierarchical level is reached. Two advantages arise from this kind of training. First, the build-up hierarchy allows for a fast best match tree search permitting speed-ups of several orders of magnitude, as compared with a standard SOM or HSOM search. Second, the H2SOM forces the nodes in each ring to structure the data on different levels, i.e. hierarchies. In the first step the primary structure of the data is captured when the input data
σ

Best results are obtained for training steps for each ring and a linear decreasing learning rate \((\eta_1 = 0.9 \text{ to } \eta_{10000} = 0.1)\) and neighborhood size \((\sigma_1 = 10 \text{ to } \sigma_{10000} = 1)\). To address the issue of assigning DNA fragments to their originating species or to bin them for assembly, the DNA fragments are classified at rank Superkingdom to Species in the Tree of Life. The classification errors of the \(H^2\)SOM classifier and the \(k\)-nearest-neighbor (knn) classifier\(^4\) (Hastie et al., 2001) at rank Genus are displayed in Figure 2 for each training-classification scenario and each type of feature vector. The complete classification errors for all ranks (Superkingdom, Phylum, Class, Order, Genus and Species) can be found in Supplementary Information 4 and 5. Each scenario was repeated 10 times to account for variations in the stochastic learning process. The figures illustrate that the \(H^2\)SOM classifies DNA fragments of 0.2–50 kb with high accuracy to their correct category in the Tree of Life, but it is still outperformed by the knn classifier. The \(H^2\)SOM results prove to be relatively stable, only very few outliers were detected. The best results are obtained for the complete \(\rightarrow 50\) kb scenario.

Three major observations can be made: first, feature vectors of longer subsequences (50 kb, 15 kb, 10 kb) can be easily classified than those of shorter subsequences (0.2 kb, 0.5 kb, 1 kb). Especially for subsequences \(\leq 1\) kb the classification error increases considerably. Second, it seems to be of minor importance whether the \(H^2\)SOM has been trained on feature vectors from complete sequences, or on subsequences of smaller sizes. Finally, the introduced \(tf{-}ti\) feature proves to be the most powerful feature in this context, closely followed by the \(tf\) feature.

For visual inspection, a \(H^2\)SOM trained on the complete genomic data using \(tf{-}ti\) features is displayed using the Poincaré projection (Fig. 3). The background is painted applying to the U-matrix principle (Ultsch, 1993): blue areas indicate high node distances in the feature space whereas red ones indicate small node distances. This visualization feature permits to identify regions with high \((i)\) and low \((ii)\) variation among species. In Figure 3a the origin \((iii)\) of the HSOM is centered. At each node the most represented taxonomic order is displayed by an order-specific object, colored from red and orange (Archaea) to yellow, green, cyan and blue (Bacteria), with respect to the color code in Supplementary Information 4. It can be seen, that the species are not randomly mapped to the HSOM nodes, but that taxonomy related species are often mapped close to each other. Additional labels summarize the major content of each node. In the first ring \((iv)\), the node content is summarized at rank Phylum, i.e. it is counted how many Archaea and Bacteria are mapped to each node. In the second ring \((v)\), the node content is summarized at rank Class, Order and Genus in the third to fifth ring. The Moebius transformation permits to move any point of the hyperbolic plane to the center in the display while all other points are transformed accordingly (Fig. 3b and c). Such an interactive browsing and zooming into various regions of the trained HSOM allows to focus on regions of interest. In Figure 3b a region containing Archaea (red objects in \(vi\)) has been dragged towards the center, and in Figure 3c a region with Proteobacteria (cyan objects in \(vii\)) is focused. Training the \(H^2\)SOM with feature vectors from 350 species can be performed in less than 1 min on a standard PC with 2 GHz and

\(^{4}\)The knn classifier is often used as standard classifier in machine learning.
classification error
tf tf-ti oligo
tf
oligo
tf-ti
knn classifier
Fig. 2. Classification errors of the H2SOM and the knn classifier at rank Genus for each training-classification scenario and each type of feature vector, 10 runs each (box-and-whisker plots). It can be seen that the H2SOM classifies small DNA fragments with high accuracy to their correct category in the Tree of Life, but it is still outperformed by the knn classifier.

256 MB RAM, whereas the classification of the same number of species does not require more than a few seconds. The knn classifier requires about half a minute for classification.

4 DISCUSSION
H2SOMs were trained for the phylogenetic classification of DNA fragments of 0.2–50 kb at ranks Superkingdom, Phylum, Class, Order, Genus and Species. Three different types of feature vectors were applied to represent complete genomic sequences and DNA fragments of 350 prokaryotic organisms. The high classification accuracy at all ranks makes the H2SOM classifier a powerful tool to assign DNA fragments to their originating species or to bin fragments of unknown species for assembly.

Longer DNA fragments are easier to classify, as indicated by their smaller classification errors. The best results are obtained for classification of DNA fragments of 50 kb using the tf–ti features. Interestingly, the size of the DNA fragments is of minor importance for training, e.g. the classification result for subsequences of 500 bp is almost independent of the (sub-)sequences used for training (complete, 50 kb, 10 kb, 3 kb, or 1 kb).

SOMs have the advantage, that they can perform visualization, classification and clustering at the same time. The hierarchical class structure of the considered genomic data motivated the application of a hyperbolic SOM. The HSOM is organized on a tree grid, but also incorporates links between neighboring branches, which theoretically may permit to account for horizontal gene transfer observed in Bacteria. For visual inspection, the hyperbolic SOM can be visualized in Euclidean space using the Poincaré projection. Such a visualization allows an easy detection of homogeneous areas and permits to focus on areas of frequent misclassifications for further analysis. A hierarchical training scheme is applied in the H2SOM as used throughout our study, allowing considerable speed-ups of several orders of magnitudes compared to the SOM and HSOM. Regarding the classification accuracy, the H2SOM is still slightly outperformed by the knn classifier. However, the H2SOM classifier evolves its power with increasing number of sequenced species. The knn classifier requires $O(n)$ time to classify a novel genomic fragment, whereas the H2SOM classifier needs $O(\log m)$ for the same task when using the ultra-fast tree search, with $n$ being the number of training samples and $m$ being the number of nodes. In addition, the HSOM provides a visualization framework that allows interactive zooming into selected genomic feature space regions on the level of interaction or similar to the hyperbolic tree browser proposed by (Lamping et al., 1995). This makes the H2SOM well suited to deal with the increasing number of sequence reads and sequenced organisms in metagenomic studies and for the testing of different feature spaces or combinations of feature spaces.

In order to check if all features (derived from oligonucleotide patterns) in a feature vector are necessary for an accurate classification, we applied a feature selection algorithm to sort the oligonucleotide patterns according to their potential for classification. Experiments for the complete $\rightarrow$ 10 000 scenario and the tf–ti features revealed that 20% of all oligonucleotide patterns are already sufficient to produce similar classification rates (Supplementary Information 6). A slight improvement of the classification rates can even be obtained when half of the oligonucleotide patterns is used. However, for a proper use of computationally expensive feature selection methods, further
investigations about oligonucleotide pattern subsets and their stability need to be accomplished. To find an appropriate representation of both complete genomes and variable-length DNA fragments, three different feature spaces were analyzed in this study. Even though the $tf–ti$ features produces encouraging results, more sophisticated weighting schemes, combinations of feature spaces and more complex features based on additional knowledge about oligonucleotide patterns are imaginable. Similar to the research performed in the domain of textmining in the last decade, a new field opens for the development.
REFERENCES


