# Genome feature exploration using hyperbolic Self-Organising Maps

Christian Martin, Naryttza N. Diaz, Jörg Ontrup and Tim W. Nattkemper
Center for Biotechnology
Technical Faculty, AG Applied Neuroinformatics
Bielefeld University
email: {christian.martin, joerg.ontrup, tim.nattkemper}@uni-bielefeld.de
ndiaz@cebitec.uni-bielefeld.de

keywords: self organization, hyperbolic SOM, whole genome analysis, information visualization

***Abstract***— The advent of sequencing technologies allows to reassess the relationship between species in the hierarchically organized tree of life with respect to patterns in their genomic signatures. Self-Organizing Maps (SOM) in Euclidean and hyperbolic space are applied to genomic signatures of 350 different organisms of the two superkingdoms Bacteria and Archaea to link the sequence signature space to pre-defined taxonomic levels, i.e. the tree of life. In the hyperbolic space the SOMs are trained by either the standard algorithm (HSOM) or in a hierarchical manner (H$^2$SOM), which is naturally supported by a lattice structure in hyperbolic space. Genomic signatures containing estimated statistics of oligonucleotide of certain lengths are used as features for training. For evaluating the SOM performances, distances between organisms in the feature space, on the SOM grid and in the taxonomy tree are compared pair-wise using a correlation measure and Spearman's $\rho$. We show that the structure recovered using the different SOMs reflects the gold standard of current taxonomy. The distances between species are better preserved when using the HSOM or H$^2$SOM which makes the hyperbolic space better suited for embedding the high dimensional genomic signatures.

## 1 Introduction

All existing organisms have evolved from one common ancestor according to the theory of evolution proposed by Darwin in 1859 [1]. Studying the finches that inhabit the Galapagos archipelago, Darwin envisaged the fact that evolutionary forces can drive the bearing of new species from existing ones. Since then, the ultimate goal of many biologists is to obtain a hierarchical classification or taxonomy able to map the evolutionary relationships between species. Traditionally, evolutionary relationships were established using morphological characteristics (e.g. number of legs), still valid in the analysis of fossil record. However, with the advent of sequencing technologies yielding a vast amount of molecular data, it has become possible to reassess the relationship between species [2]. The evolutionary relationship between all existing species can be modeled and visualized by the "tree like structure" which is known as the *tree of life*. Superkingdom, Phylum, Class, Order and Genus represent the most commonly used taxonomic categories with Superkingdom being the most general class (Figure 1).

Using sequence alignment molecular biologists can estimate the differences between DNA sequences of certain species of interest, in order to estimate the degree of their relationship, namely sequence similarity (the closer their relationship the more similar they should be). Conventionally, only part of the genome (often a single gene) is used for this purpose such as ribosomal RNA molecules, gene content, gene order, protein domain content, etc. However, many pitfalls in sequence alignment can be encountered such as saturation of the underlying model of evolution (when far related species are compared), sampling of representative sequences, lateral gene transfer, or recombination. All these lead to very disparate results [3]. Moreover, multiple sequence alignments are computationally very expensive.

Nowadays, with the gargantuan amount of molecular information it is very valuable to count on methods that can make use of the information contained in the whole genome and do not depend on sequence alignment, but nevertheless can readily reconstruct the relationship between species and help in the evaluation of trends in genome evolution. New alignment-free approaches that take into account general characteristics of the genomes disregarding prior identification of functional regions have been developed [4, 5, 6, 7]. But this is still a challenge in computational biology. An intriguing characteristic that enables to capture evolutionary relationships between species is the genomic signature which is defined as the whole set of short sequences of oligonucleotide of certain length [8]. Genomic signatures are species-specific and can be measured in any part of the genome [9, 10, 11] allowing direct comparisons along the entire genome.

In this study we explore the suitability of several SOMs for reconstructing the hierarchical relation of whole genomic sequences. Hereby, compositional sequence prop-

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

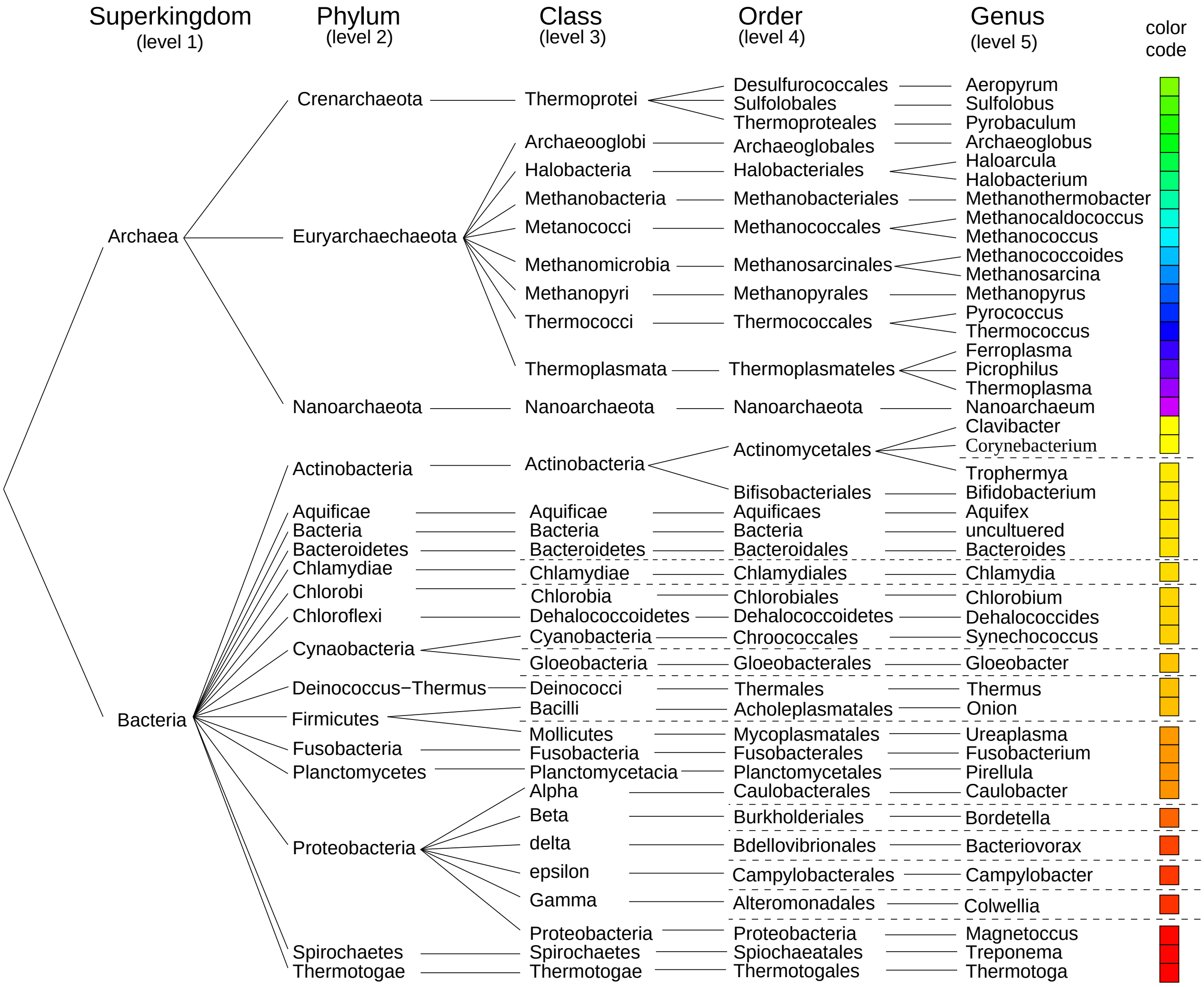

Figure 1: A subset of all 350 organisms of the superkingdoms Archaea and Bacteria is displayed. Due to space limitations, some categories are left out in the display which is indicated by dashed lines. The organisms are categorized on five different levels called Superkingdom, Phylum, Class, Order, and Genus. On the first level, the organisms are categorized in the two different superkingdoms Archaea and Bacteria. On the second level, all organisms of the superkingdom Archaea are subdivided in the three Phyla Crenarchaeota, Euryarchaeota and Nanoarchaeota. All organisms 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 levels three to five.

erties (genomic signature) are exploited. One way to find an answer is to apply dimension reduction techniques based on unsupervised learning like the SOM, to learn and project the structure of a large set of genomic signatures. Our data set of 350 organisms represents a vast majority of organisms sequenced up-to-date from the two domains of life. Genomic signatures are uniquely obtained from each complete sequenced genome without using any additional knowledge about the organisms. For each organism, the features are combined in a vector that is used to train a SOM in Euclidean and hyperbolic space. We evaluate if the structure recovered from the different SOMs reflects the gold standard of current taxonomy. Results are presented for the SOM, HSOM and H$^2$SOM. By comparing ranks of distances in the feature space, on the grid and in the tree of life, we show that the structure of the trained SOMs using only whole genome sequence data is biologically sound to the widely accepted *tree of life* based on RNA molecules. When the distances are directly compared, both the HSOM and H$^2$SOM perform better than the standard SOM, which makes them better suited for embedding of the high dimensional genomic signatures. Additionally, the H$^2$SOM allows considerable speed-ups of several orders of magnitudes which makes it well suited to deal with the increasing number of sequenced organisms and for the testing of different feature spaces.

## 2 Material and Methods

In this paper we consider 350 different organisms that are included either in the Bacteria or the Archaea superkingdom. This palette of organisms represents a vast majority of the microbial world sequenced up-to-date. For each genome, we measure the divergency from expectation of oligonucleotide patterns of length $k$, which is given as the ratio observed vs. expected. Each complete genomic sequence is encoded as a vector which entries are the divergencies on oligonucleotide patterns of length $k$. For each genome, feature vectors are computed as follows:

### 2.1 Feature vector computation from sequenced genome data

Let $\Sigma$ be the alphabet of nucleotides $\Sigma = \{A, C, G, T\}$. Let $\mathbf{o}$ be an oligonucleotide of length $k = |\mathbf{o}|$, with $o_i \in \Sigma$. Let $\mathbf{s}^{(l)}$ be the genomic sequence of organism $l$ (with $1 \leq l \leq 350$) of length $|s^{(l)}|$ each and $s_i^{(l)} \in \Sigma$. For notation simplicity we consider only one sequence $\mathbf{s} \equiv \mathbf{s}^{(l)}$ in the following.
For a sequence $\mathbf{s}$, the probability to observe a certain nucleotide $\eta \in \Sigma$ can be computed by

$$p(\eta) \quad = \quad \frac{1}{|\mathbf{s}|} \sum_{i=1}^{|\mathbf{s}|} h(s_i, \eta) \tag{1}$$

with the indicator function

$$h(s_i, \eta) \quad = \quad \begin{cases} 1 & \text{if } s_i = \eta \\ 0 & \text{else} \end{cases} \tag{2}$$

There are $|\Sigma|^{|\mathbf{o}|} = 4^k$ possible oligonucleotides of length $k$, e.g. an oligonucleotide of length $k = 3$ can be one of the following sequences: $\mathbf{o}^{(1)} = AAA$, $\mathbf{o}^{(2)} = AAG$, $\ldots$, $\mathbf{o}^{(4^k)} = TTT$. The sequence feature vectors are generated to encode the enhanced contrast between over- and underrepresented oligonucleotides in a sequence. The expectation value for a certain oligonucleotide $\mathbf{o}$ in the sequence s can be estimated by

$$E[\mathbf{o}] \quad \approx \quad |\mathbf{s}| \prod_{i=1}^{|\mathbf{o}|} p(o_i) \tag{3}$$

Let $O[\mathbf{o}]$ be the number of observed oligonucleotides $\mathbf{o}$ in the same sequence s. The contrast is performed by computation of the score

$$g(\mathbf{o}) \quad = \quad \begin{cases} 0 & \text{if } O[\mathbf{o}] = 0 \\ \frac{O[\mathbf{o}]}{E[\mathbf{o}]} & \text{if } O[\mathbf{o}] > E[\mathbf{o}] \\ -\frac{E[\mathbf{o}]}{O[\mathbf{o}]} & \text{if } O[\mathbf{o}] \leq E[\mathbf{o}] \end{cases} \tag{4}$$

The scores of all possible oligonucleotides for the sequence s are combined in one vector of dimension $4^k$.

$$g(\mathbf{s}) \quad = \quad \left(g(\mathbf{o}^{(1)}), g(\mathbf{o}^{(2)}), \ldots, g(\mathbf{o}^{(4^k)})\right)^T \tag{5}$$

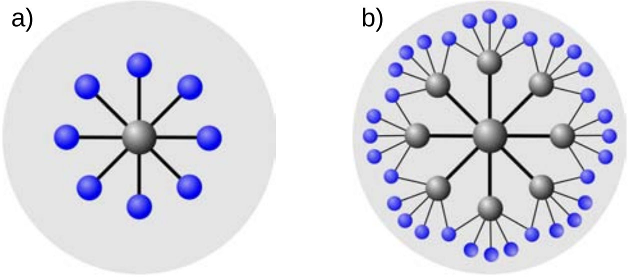

Figure 2: Construction of the H$^2$SOM: The H$^2$SOM is initialized with the root node of the hierarchy placed at the origin of $I\!H^2$. Then the $n_b$ children nodes of the first sub hierarchy are equidistantly placed around the center node (**a**)). During a first phase, the top level ring of nodes is trained in the standard self-organized fashion. After a fixed training interval, each node in the periphery is expanded as indicated in **b**).

The feature vector of the sequence $\mathbf{s}$ is defined as

$$f(\mathbf{s}) = \frac{g(\mathbf{s})}{\|g(\mathbf{s})\|} \tag{6}$$

### 2.2 Data

The complete set of 350 genomes included in the bacteria and archaea superkingdoms are obtained from the SEED database [1] [12]. The complete taxonomic information from the set of organisms evaluated in this survey are obtained from the taxonomy database located in the US National Center for Biotechnology Information (NCBI) [13].

The considered 350 organisms are categorized on different levels in the tree of life. We consider five such categorizations (Figure 1). On the first level, the organisms are categorized in two different superkingdoms archaea and bacteria. On the second level, all organisms of the superkingdom archaea are subdivided in the three phyla Crenarchaeota, Euryarchaeota and Nanoarchaeota. All organisms of the superkingdom bacteria are subdivided in 15 different phyla (from Actinobacteria to Thermotogae). Finer subdivisions are obtained by the categories order, class and genus on the levels three to five.

### 2.3 The Hyperbolic Self Organizing Map (HSOM)

In 1990, Kohonen introduced the Self-Organizing Map (SOM) [14]. Since then, it has become a widely used tool for exploratory data analysis. Typically, the SOM projects data to a two-dimensional 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 ad-

[1] http://www.nmpdr.org/FIG/index.cgi

equate, 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 hyperbolic SOM (HSOM) this exponential scaling property has already successfully been used to visualize high dimensional text data [15]. The core idea of the HSOM is to employ a grid of nodes in the hyperbolic plane $I\!H^2$ which is then projected onto the $I\!R^2$ for inspection. The regular structure of formal neurons used by the HSOM is based on a tessellation of $I\!H^2$ with equilateral triangles - for more details please refer to [16].

The HSOM is then formed in the standard self-organizing manner: Each lattice node $r$ carries a prototype vector $\mathbf{w}_r \in \mathcal{R}^D$ 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 $\mathbf{x}$ by $s = \text{argmin}_r \|\mathbf{w}_r - \mathbf{x}\|$. The prototype vectors are then adjusted according to the familiar rule

$$\Delta \mathbf{w}_r = \eta h(r,s)(\mathbf{x} - \mathbf{w}_r) \tag{7}$$

with

$$h(r,s) = \exp\left(-\frac{d^2(r,s)}{2\sigma^2}\right), \tag{8}$$

where $h(r,s)$ is a Gaussian shaped function centered at the winner node $s$ and decaying with increasing node distance $d(r,s)$ on the hyperbolic lattice.

### 2.3.1 Hierarchically growing HSOM (H$^2$SOM)

The H$^2$SOM employs the same sort of regular lattice structure already used for the plain HSOM, but offers a hierarchically growing scheme: The H$^2$SOM is initialized with the root node of the hierarchy placed at the origin of $I\!H^2$. Then the $n_b$ children nodes of the first sub hierarchy are equidistantly placed around the center node as shown in Figure 2a). 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 $n_b$ determines how many nodes are generated at each level and how "fast" the network is reaching out into the hyperbolic space. $n_b$ is lower bounded by 7, but has no upper bound [17]. 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 2b) 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 H$^2$SOM 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 is projected to the $n_b$ nodes of the first ring. A finer data categorization is obtained in the second step and so on. Thus it may have great potential to truthfully project data with an intrinsic hierarchical structure.

## 2.4 Evaluation measures

The correlation between distances, or Spearman's $\rho$ when applied to distances can measure to which extent the distances of point pairs in two different spaces $\mathcal{S}^1$ and $\mathcal{S}^2$ are correlated. In our work, distances between organisms can be computed in three different ways. First, the *feature space distance* $d^f_{ij} \in [0,1]$ can be obtained by computing the Euclidean distance $d\left(f\left(\mathbf{s}^{(i)}\right), f\left(\mathbf{s}^{(j)}\right)\right)$ between two organisms $i$ and $j$ in the feature space and normalizing it to [0,1]. Second, a *grid distance* $d^g_{ij} \in [0,1]$ can be obtained by computing the minimal distance on the SOM grid between the two nodes to which the organisms $i$ and $j$ have been mapped. The grid distances are normalized such that the maximal possible distance on the grid is one. Third, the *taxonomy distance* $d^t_{ij} \in [0,1]$ of two organisms $i$ and $j$ is defined as follows:

$$d^t_{ij} = \begin{cases} 0 & \text{if they have the genus in common} \\ 0.2 & \text{if they have the class in common} \\ 0.4 & \text{if they have the order in common} \\ 0.6 & \text{if they have the phylum in common} \\ 0.8 & \text{if they have the superkingdom in common} \\ 1 & \text{if they have nothing in common} \end{cases}$$

For notation simplicity, let the *distance vector* $\tilde{d}^*_l$, $l = 1, \ldots, n$ denote all distances $d^*_{ij}$, $i \neq j$, $n = \frac{N(N-1)}{2}$, $N = 350$.

### 2.4.1 Correlation

The correlation $c$ between the two distance vectors $\tilde{\mathbf{d}}^1$ and $\tilde{\mathbf{d}}^2$ is defined as

$$c = \frac{1}{n}\sum_{l=1}^{n} \tilde{d}^1_l \tilde{d}^2_l \tag{9}$$

$c$ is bounded by $[-1,1]$. $c = 1$ indicates a perfect correlation between the two distance vectors whereas $c = 0$ indicates no correlation.

### 2.4.2 Topology Preservation

In [18], Spearman's $\rho$ was applied to compute the quality of a *metric topology preserving* (MTP) transformation by computing the linear correlation coefficient of ranks of distances in the feature space and the projected space. In fact, Spearman's $\rho$ can be used for any two distance vectors $\tilde{\mathbf{d}}^1$ and $\tilde{\mathbf{d}}^2$ when defining it as the linear correlation coefficient of the ranks $R_l$ and $S_l$

$$\rho_{\text{Sp}} = \frac{\sum_l (R_l - \bar{R})(S_l - \bar{S})}{\sqrt{\sum_l \left(R_l - \bar{R}\right)^2}\sqrt{\sum_l \left(S_l - \bar{S}\right)^2}} \tag{10}$$

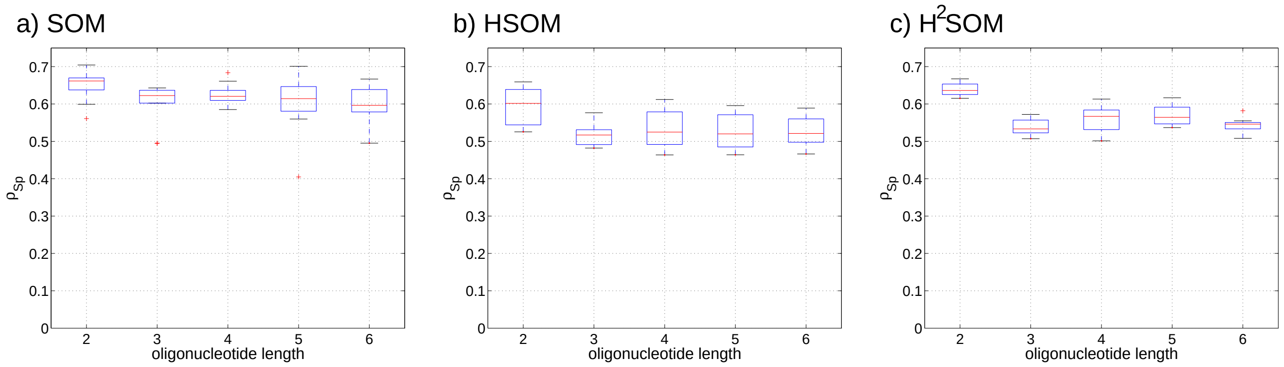

Figure 3: Spearman's $\rho$ between feature space distances and grid distances is displayed. A $\rho \approx 0.6$ indicates that the ranks of distances are very well preserved. This observation is rather independent of the oligonucleotide length and the SOM used.

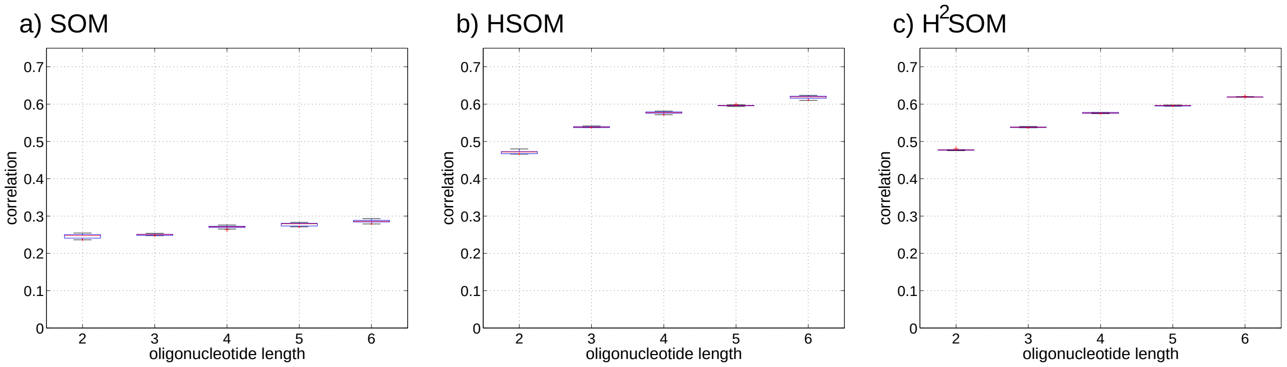

Figure 4: The correlation between feature space distances and grid distances is shown. It can be seen that SOMs in hyperbolic space better preserve the distances than the SOM in Euclidean space. The performance of the SOMs increases with the oligonucleotide length.

where $R_l$ and $S_l$ are the ranks of the considered distance vectors $\tilde{\mathbf{d}}^1$ and $\tilde{\mathbf{d}}^2$. $\rho_{\mathrm{Sp}}$ is a measure for the global metric preservation of a projection and is bounded by [-1,1]. $\rho_{\mathrm{Sp}} = 1$ indicates a complete metric preservation. As $\rho_{\mathrm{Sp}}$ decreases from one, the projection is becoming less MTP, and $\rho_{\mathrm{Sp}} = 0$ indicates a complete random projection in terms of distance preservation.

# 3 Results

A standard SOM, a HSOM, and a H$^2$SOM are trained on the data described in section 2.1. The features are defined by oligonucleotides of length $k$ ranging between 2 and 6 in five different datasets. All SOMs are trained with 10000 training steps and a linear decreasing learning rate ($\eta_1 = 0.9$ to $\eta_{10000} = 0.1$) and neighborhood size ($\sigma_1 = 10$ to $\sigma_{10000} = 1$). All hyperbolic SOMs consist of five rings with a branching factor $n_b = 8$, resulting in 2281 nodes. The standard SOM is initialized using the eigenvectors of the first and second largest eigenvalue. Its dimension is determined by the relation between the first and second largest eigenvalue such that the number of nodes is approximately the same as in the hyperbolic SOMs. For each dataset and for each training algorithm, the SOMs are trained 10 times. The following two issues are analyzed: To which extent does the structure of the grid correspond to *i)* the structure in the feature space, and *ii)* the taxonomy? To this end, we analyze the correlations between distances (section 2.4.1) and correlations between ranks of distances using Spearman's $\rho$ (section 2.4.2).

In Figure 3 Spearman's $\rho$ between feature space distances and grid distances is displayed. A $\rho \approx 0.6$ indicates that the ranks of distances are very well preserved. This observation is rather independent of the oligonucleotide length and the SOM used. When considering the direct correlation of feature space distances and grid distances (Figure 4), the SOMs in hyperbolic space better preserve the distances than the SOM in Euclidean space. The performance of the SOMs increases with the oligonucleotide length. The difference between Spearman's $\rho$ and the direct correlation can be explained by the different distributions of distances on the SOM grid. In Euclidean space the distribution of node distances favors smaller distances whereas in the hyperbolic case the exponential scaling behavior of $I\!H^2$ allows a larger proportion of the nodes to have a rather large

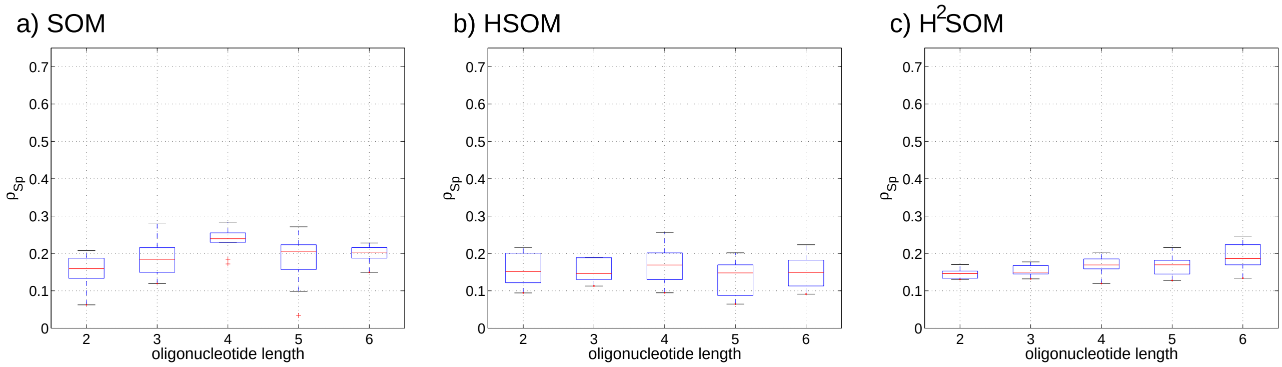

Figure 5: Spearman's $\rho$ between grid distances and taxonomy distances is shown. The slight positive $\rho_{\mathrm{Sp}}$ indicates that there is a link between the structure found by the SOMs and the taxonomy.

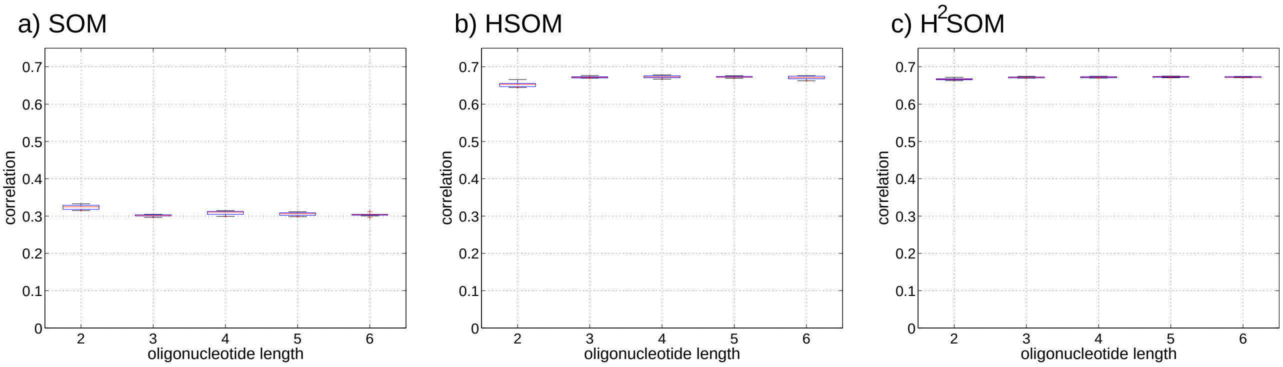

Figure 6: The correlation between grid distances and taxonomy distances is displayed. The SOMs in hyperbolic space preserve the distances better than the SOM in Euclidean space. This observation is independent of the oligonucleotide length.

distance to each other. This feature allows the hyperbolic SOM to better reflect the true distribution of distances in the original high dimensional oligonucleotide space.

In Figure 5 Spearman's $\rho$ between grid distances and taxonomy distances is shown. The slight positive $\rho_{\mathrm{Sp}}$ indicates that there is a link between the structure found by the SOMs and the taxonomy. When considering the direct correlation between grid distances and taxonomy distances (Figure 6), the SOMs in hyperbolic space preserve the distances better than the SOM in Euclidean space. This observation is independent of the oligonucleotide length. For visual inspection, a randomly chosen HSOM trained with oligonucleotides of length 4 and organisms colored according to their position in the taxonomy tree is displayed using the Poincaré projection (Figure 7). In Figure 7 a) the origin of the $I\!H^2$ is centered. In Figure 7 b) to e) different nodes of the HSOM (21, 39, 19, 1) are moved to the center of the display. All other points of the $I\!H^2$ are moved accordingly allowing us to inspect various regions in the hyperbolic space. The organisms are visualized as colored circles at the node to which they are mapped. Each color represents a genus as illustrated in Figure 1. The circle area is proportional to the number of species that are mapped to the node, but also decreases with the distance to the center node. It can be seen that organisms are not randomly mapped to the HSOM nodes, but that taxonomy related organisms are often mapped close to each other.

## 4 Discussion

We apply a standard SOM in Euclidean space, a hyperbolic SOM (HSOM) and a hierarchical hyperbolic SOM (H$^2$SOM) on genomic signatures of 350 different organisms of the two superkingdoms Bacteria and Archaea. The three different types of SOMs are evaluated by comparing Spearman's $\rho$ and correlations of feature space distances, grid distances and taxonomy distances. We find that the relatively simple features obtained from genomic signatures are sufficient to allow a reasonable SOM projection. By comparing Spearman's $\rho$ and correlations of distances in the feature space, on the grid and in the tree of life, we show that the structure of the trained SOMs using whole genome sequence data is biologically sound to the widely accepted tree of life based on RNA molecules. When the distances are directly compared, both the HSOM and H$^2$SOM perform better than the standard SOM, which makes the hy-

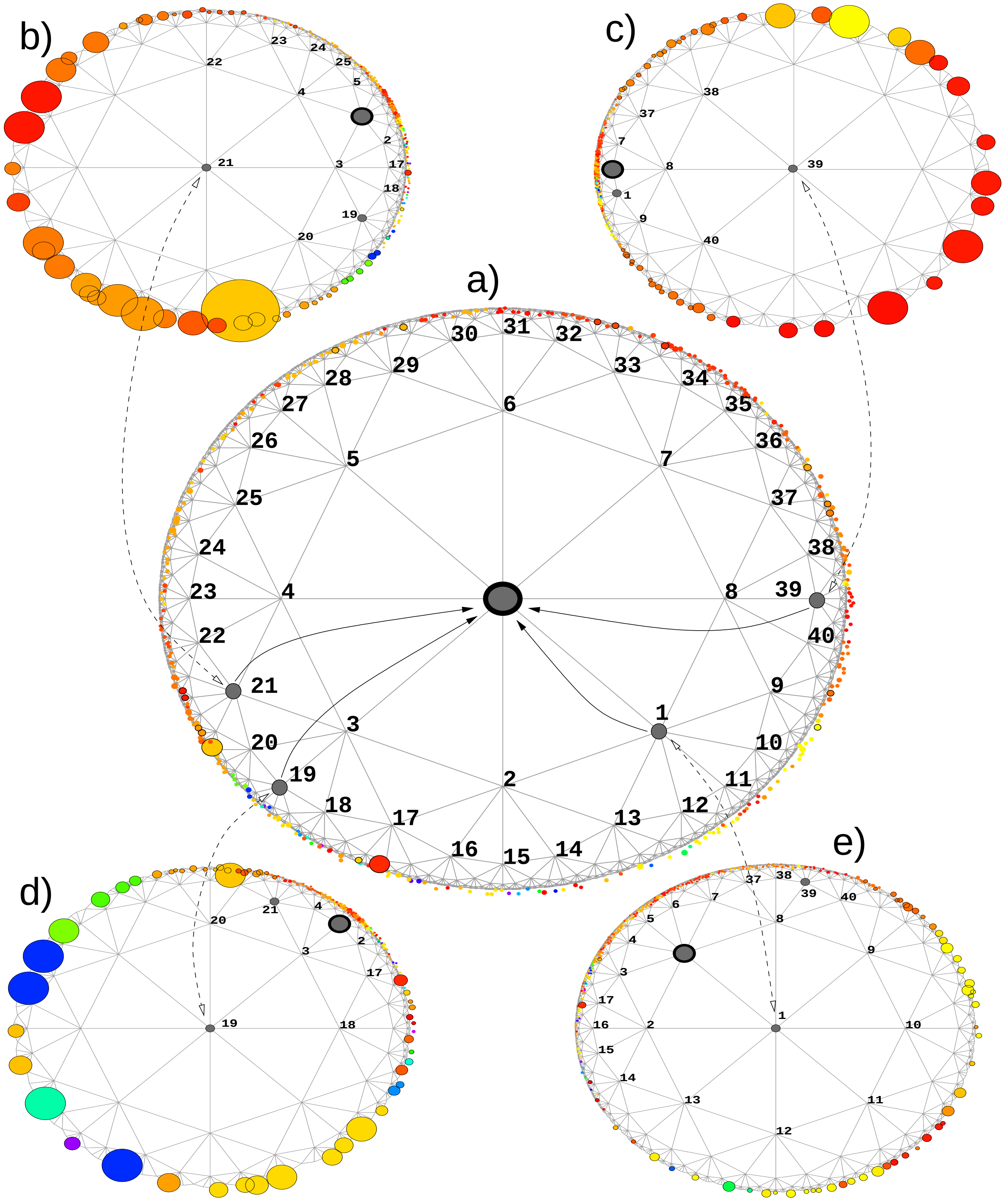

Figure 7: For visual inspection, a randomly chosen HSOM trained with oligonucleotides of length 4 and organisms colored according to their position in the taxonomy tree is displayed using the Poincaré projection. In **a)** the origin of the $\mathbb{H}^2$ is centered. In **b)** to **e)** different nodes of the HSOM (21, 39, 19, 1) are moved to the center of the display. All other points of the $\mathbb{H}^2$ are moved accordingly allowing us to inspect various regions in the hyperbolic space. The organisms are visualized as colored circles at the node to which they are mapped. Each color represents a genus as illustrated in Figure 1. The circle area is proportional to the number of species that are mapped to the node, but also decreases with the distance to the center node. It can be seen that organisms are not randomly mapped to the HSOM nodes, but taxonomy related organisms are often mapped close to each other.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

perbolic SOMs better suited for visualization issues. Additionally, the $H^2$SOM allows considerable speed-ups of several orders of magnitudes. This makes it well suited to deal with the increasing number of sequenced organisms and for the testing of different feature spaces or combinations of feature spaces.

## Acknowledgements

This research is supported by the International Graduate School for Bioinformatics and Genome Research at Bielefeld University, Germany as well as by the Technical Faculty of Bielefeld University, Germany. Naryttza N. Diaz is supported by the Deutscher Akademischer Austausch Dienst (DAAD).

## References

[1] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.

[2] E Zuckerkandl and L Pauling. Molecules as documents of evolutionary history. *J Theor Biol*, 8(2):357–366, Mar 1965.

[3] L Brocchieri. Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol*, 59(1):27–40, Feb 2001.

[4] R S Gupta. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev*, 62(4):1435–1491, Dec 1998.

[5] G W Stuart, K Moffett, and J J Leader. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol Biol Evol*, 19(4):554–562, Apr 2002.

[6] J Qi, B Wang, and B-I Hao. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J Mol Evol*, 58(1):1–11, Jan 2004. Comparative Study.

[7] C Chapus, C Dufraigne, S Edwards, A Giron, and P Fertil, B Deschavanne. Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol Biol*, 5:63, 2005.

[8] S Karlin and C Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, 11(7):283–290, Jul 1995.

[9] P J Deschavanne, A Giron, J Vilain, G Fagot, and B Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol*, 16(10):1391–1399, Oct 1999.

[10] T Abe, S Kanaya, M Kinouchi, Y Ichiba, T Kozuki, and T Ikemura. A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: self-organizing map of oligonucleotide frequency. *Genome Inform*, 13:12–20, 2002.

[11] T Abe, H Sugawara, M Kinouchi, S Kanaya, and T Ikemura. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res*, 12(5):281–290, 2005.

[12] R Overbeek, T Begley, R M Butler, J V Choudhuri, and et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33(17):5691–5702, 2005. Evaluation Studies.

[13] D L Wheeler, D M Church, A E Lash, D D Leipe, and et al. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res*, 30(1):13–16, Jan 2002.

[14] T Kohonen. The self-organizing map, 1990.

[15] J.Ontrup and H.Ritter. Hyperbolic Self-Organizing Maps for Semantic Navigation. In *NIPS*, 2005.

[16] H. Ritter. *Self-organizing maps in non-euclidian spaces*, pages 97–110. Amer Elsevier, 1999.

[17] J.Ontrup and H.Ritter. Large Scale Data Exploration with the Hierarchical Growing Hyberbolic SOM. *Neural Networks, Special Issue on New Developments in Self-Organizing Systems*, 19:751–761, 2006.

[18] J. Bezdek and N. Pal. An index of topological preservation for feature extraction. *Pattern Recognition*, 28(3):381–391, 1995.