Publication IV

Merja Oja, Göran Sperber, Jonas Blomberg, and Samuel Kaski. Grouping and visualizing human endogenous retroviruses by bootstrapping median self-organizing maps. In *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2004), 7-8 October, San Diego, USA*, pages 95–101, 2004.

IV

Grouping and Visualizing Human Endogenous Retroviruses by Bootstrapping Median Self-organizing Maps

Merja Oja, Göran Sperber, Jonas Blomberg, and Samuel Kaski

Abstract— **About eight percent of the human genome consists of human endogenous retrovirus sequences. Human endogenous retroviruses (HERV) are remains from ancient infections by retroviruses. The HERVs are mutated and deficient, but they still may give rise to transcripts or may affect the expression of human genes. The HERVs stem from several kinds of retroviruses. The possible current functioning of the HERV sequences may reflect the origin of the HERVs. Hence, the classification of the diverse HERV sequences is a natural starting point when investigating the effect of HERVs in humans. The current HERV taxonomy is incomplete: some sequences cannot be assigned to any class and the classification is ambiguous for others. A Median Self-Organizing Map (SOM), a SOM for data about pairwise distances between samples, can be used to group all the HERVs found in the human genome. It visualizes the collection of 3661 HERV sequences found by the RetroTector system, on a two-dimensional display that represents similarity relationships between individual sequences, as well as cluster structures and similarities of clusters. The SOM, as any dimensionality reduction method, necessarily has to make compromises when representing the data. In this work we extend the visualizations by bootstrap-based estimates on which parts of the visualization are reliable and which not, and use the SOM to find potentially new HERV groups.**

Index Terms— **Bootstrap, human endogenous retroviruses, self-organizing maps, visualization.**

I. Introduction

About eight per cent of human DNA consists of *human endogenous retroviruses (HERV)* [1]. Human retroviruses, such as HIV, are viruses capable of copying their genetic code into the DNA of humans, and they become endogenous once they have been copied to the germ-line. During the time the HERV sequences have inhabited the human genome they have become mutated and broken in crossovers or when transposons have moved to overlap them. Hence the sequences are noisy and incomplete, but it has been suggested that they may have functions in regulating the activity of human genes, and may produce proteins under some conditions [2], [3].

The HERVs stem from several kinds of retroviruses. Functions of HERV sequences in the human genome will probably correlate with their origin, and vary according to which kinds of functional parts are still present in the sequences. HERV categories formed according to sequence similarity could capture these relationships, and thus help in studying functions of HERVs.

A traditional way of classifying HERVs is to group them according to the similarity of a short region, the primer binding site (PBS), from which their transcription (activation) starts [4], [5]. In this grouping obviously a lot of information is lost, and recently the HERVs have been grouped according to phylogenetic analyses based on one of their genes: *pol* [6], [7] or *env* [8]. The phylogenetic trees are constructed together with representatives from exogenous retroviruses, to reflect the other widely used option; to classify HERVs according to their similarity to types of exogenous retroviruses, from which they presumably stem.

The taxonomy of HERVs is still far from complete. The groupings based on the PBS have been revised somewhat to present the groups with different origins (review of current groups in [3], [9], [10]). But as new instances of HERVs are detected from the human genome, it has become obvious that these groupings (classes) are not adequate. Some sequences can not be assigned unambiguously to any class. In addition, some current classes are mixed with sequences from other classes in phylogenetic trees constructed from large HERV collections. Furthermore, sequences from some classes appear in more than one branch. A new classification able to resolve these problems is needed. A better and clearer classification of the endogenous retroviruses will also help organize the overall retrovirus universe as most retroviruses are endogenous.

The phylogenetic methods are based on a multiple alignment of the sequences. Due to the exponential computational complexity of the alignment step, they can operate on only a limited number of sequences – normally of the order of hundreds at most even with heuristic alignment algorithms. To extend the phylogenetic methods to larger collections of related sequences, the sequences must first be clustered; only the cluster representatives are used in the multiple alignment and tree construction. This brings an extra step to the algorithm, which can introduce biases to the results.

The Self-Organizing Map (SOM) [11] is an algorithm capable of handling large amounts of data. The computational complexity of a large SOM is $O(n^2)$, with n sequences, and by reducing the resolution (size of the SOM) this can

Manuscript received June 18, 2004. This work was supported in part by the Academy of Finland under grant 201352 and in part by the Graduate School in Computational Biology, Bioinformatics, and Biometry.

M. Oja is with Neural Networks Research Centre, Helsinki University of Technology, Finland (phone: +358-9-451 4351, fax: +358-9-451 3277, email: merja.oja@hut.fi)

G. Sperber is with Unit of Physiology, Department of Neuroscience, Uppsala University, Sweden (email: Goran.Sperber@neuro.uu.se)

J. Blomberg is with Section of Virology, Department of Medical Sciences, Uppsala University, Sweden (email: Jonas.Blomberg@medsci.uu.se)

S. Kaski is with Department of Computer Science, University of Helsinki, Finland. Part of the work was carried out while he was with Neural Networks Research Centre, Helsinki University of Technology, Finland (email: samuel.kaski@hut.fi)

be reduced. The SOM operates in a data driven manner, producing a visualization of the cluster structures in the data set. The SOM can reveal groups of similar sequences, and visualize their relationships to other groups. The SOM displays the similarities in a two dimensional plane, and enables the visualization of many neighbors per sequence. In addition, by using SOM to group the HERV sequence data, we can get a visualization for all the data at the same time.

The Median Self-organizing Map [11], [12] is a variant of SOM capable of handling sequence data. It can be used on any nonvectorial data where pairwise distances can be defined between all input samples. Here we use pairwise distances between HERV protein sequences.

The reliability of the results is always a major issue in data analysis. The SOM is a dimensionality reduction method which represents a high-dimensional data set in a two-dimensional display. Any dimensionality reduction method will have to make compromises, and so will SOM. Some sequences are represented with lower precision in order to achieve a good overall projection of the data. For a comparative study between SOM and some alternatives see [13].

In this paper we complement our earlier work [14] on median SOMs of HERVs by assessing the reliability of the results. We will measure reliability in representing the similarities between sequences in each location on the SOM display. The reliability is estimated with the bootstrap method [15], [16], a statistical technique developed for estimating the sampling distribution of an interesting random variable, such as the mean of a distribution. The bootstrap has been used in the context of clustering [17]–[21] and phylogenetic trees [22], [23] to estimate the repeatability of the observed groupings. Here we will use the bootstrap method to estimate the sampling variability of the observed neighborhoods on the SOM display.

We will apply the combination of median SOM and bootstrap to grouping and visualizing a collection of 3661 HERV sequences found from the human genome. We extract new groups of sequences, and suggest that these could be new HERV classes.

II. Methods

A. Principle of the Median SOM

The Self-Organizing Map (SOM) [11], [12] is an algorithm used to visualize and interpret large high-dimensional data sets. We will outline the SOM algorithm here only briefly. An overview of the basic SOM algorithm can be found for example in [24] or in the book [11]. The Median SOM algorithm is explained in more detail in [12].

The SOM consists of a regular grid of units. A model, normally a vector representing the inputs, is associated with each unit. The map attempts to represent all the available input samples using the restricted set of models. At the same time the models become ordered on the grid so that similar models are close to each other and dissimilar models far from each other. The input samples are mapped onto the SOM grid to their best-matching models (the model closest to the input sample).

The SOM can be used to order nonvectorial data such as DNA sequences by a variant of the method in which each model on the map becomes the *generalized median* of the input samples mapped into the neighborhood of the model [12]. For this method it is sufficient that some similarity measure is definable between each sample and each model, as well as between all pairs of the data samples. This variation of the SOM, called the Median SOM, resembles the Batch Map method [11], [25].

In this work, the Median SOM has been applied to the production of similarity diagrams, and showing the clustering tendency of HERV sequences. The similarities between the sequences were computed by the FASTA method [26].

The generalized median is in practice often approximated by the set median. The generalized median is defined as the hypothetical data sample from which the sum of distances to the other elements in a data set is minimized. Similarly, the set median is the data sample from which the sum of distances to the other elements of the data set is minimized. The set median is an exact copy of one of the data samples in the data set.

The computation of the SOM using set medians as models is performed as the iteration of the following two steps. At the first step, the input (teaching) sequences are mapped to their best-matching models. At the second step, for each unit in the map, a new value for the model is determined as the set median of those input sequences that were mapped to the said unit or its neighboring units on the SOM grid. These two steps, namely, searching best-matching models for each input sequence, and computation of the new models as the set medians of sequences mapped into the neighborhood of each unit, are repeated, until the models can be regarded as stationary.

B. The SOM visualization

The SOM grid is visualized as a two-dimensional display. The visualization represents the similarities of the input samples. Samples located at proximate units are similar to each other whereas samples located far from each other are typically dissimilar.

To get insight into the cluster structure of the data, the distances between neighboring units are visualized with gray scale coloring of the unit boundaries on the SOM display. A cluster is an area of the map where the units are close to each other i.e. the unit boundaries inside a cluster have light coloring. Borders between clusters appear as dark edges or areas on the map where distances between neighboring units are considerably larger.

C. Reliability of the SOM visualization

The SOM algorithm aims at placing proximate points of the input space to SOM units that are neighbors (or even into the same unit). Here we want to measure the performance of the SOM in this respect. We ask the following question: If two sequences are observed as neighbors on the SOM, is this co-occurrence reliable? We will use the bootstrap method

[15], [16] to measure the sampling variability of the observed proximities.

The bootstrap method [15] is applicable to the following problem: Given a random sample $\mathbf{X} = (X_1, X_2, ..., X_n)$ from an unknown distribution F, estimate the sampling distribution of some prespecified random variable $R(\mathbf{X}, F)$, on the basis of observed data $\mathbf{x}$. The sampling distribution of $R(\mathbf{X}, F)$ is estimated by producing new samples $\mathbf{x}^*$ with replacement from $\mathbf{x}$ and computing $R(\mathbf{x}^*, F)$ for each sample $\mathbf{x}^*$. The histogram of $R(\mathbf{x}^*, F)$ values represents the sampling distribution. The sampling distribution can then be used to estimate e.g. the mean, variance and confidence intervals for $R(\mathbf{X}, F)$.

The bootstrap approach has been used in clustering [17]–[21] to estimate the stability of the discovered clusters. It is assumed that the cluster composition should not change radically between two samples of the same underlying data distribution. Therefore, if a clustering is robust to sampling variability, we can assume that it represents the real structure of the data. This reasoning can also be applied to SOMs: if the neighbors of a sequence are retained in SOMs constructed from different samples, we can assume that those are reliably neighbors.

The bootstrap approach has been previously applied to self-organizing maps in [18]. The article describes significance tests for the quantization error and for the stability of neighborhoods on the SOM. In this article we will not aim at a significance test but will look at the stability of the neighborhoods in each map unit separately.

We will estimate the confidence of the SOM visualization by counting how often a pair of sequences appear as neighbors in bootstrap repetitions of the SOM. Confidences for individual map units in the visualization are derived as averages over the sequences in that map unit. The next section describes our algorithm for bootstrapping SOMs.

D. Bootstrapping the SOM

The data set is sampled B times with replacement to produce B bootstrap data sets of the size of the original data set. Some samples will appear several times in a bootstrap data set, and some samples will be missing.

A self-organizing map is computed from each bootstrap data set to produce B bootstrap SOMs. The bootstrap data set used to construct a bootstrap map is then discarded and the original data set is projected to each of the bootstrap maps. Thus each data sample has a location on each of the bootstrap maps.

We estimate the stability of the neighborhood separately for each pair of sequences. Here we will consider the immediate neighborhood on the map (the same map unit and its bordering units); other choices of neighborhood size are possible as well. We count the frequency $f_{i,j}$ of samples i and j appearing as neighbors on the bootstrap maps:

$$f_{i,j} = \frac{\sum_{b=1}^{B} \mathrm{Neighbors}(i,j,b)}{B}, \qquad (1)$$

where $\mathrm{Neighbors}(i,j,b)$ is an indicator function that returns 1 if i and j are neighbors on bootstrap sample b, and otherwise zero. The frequency $f_{i,j}$ gets values between 0 and 1. A perfect stability would lead to $f_{i,j}$ being either 1 (always neighbored) or 0 (never neighbored).

The pairwise frequencies $f_{i,j}$ are collected into a matrix $\mathcal{F}$ of size $N \times N$, where N is the number of sequences in the data set. The matrix is symmetric and has ones on the diagonal (a sequence is always a neighbor to itself). This matrix can be used to compute summary statistics accounting for the stability of groups of sequences. Here we will use a simple average over the pairwise frequencies of the included sequences, but other options could be used as well.

We measure the reliability of each map unit by computing the average stability among the sequences in that unit:

$$s_k = \frac{1}{N_k(N_k - 1)} \sum_{i \neq j, i,j \in \text{ unit } k} f_{i,j}, \qquad (2)$$

where N_k is the number of sequences in the unit k.

A measure similar to (2) can be computed for larger groups of sequences as well, for instance for clusters of SOM units.

E. Collection of human endogenous retroviruses

The data set consists of 3661 HERV sequences automatically collected from the human genome by RetroTector© [27], [28]. The RetroTector© is a program developed for the detection of endogenous retroviruses and similar structures in genomes. It uses a combination of expert knowledge and machine learning to detect the retroviral-like parts in genomes. It locates known conserved features and strings them together into longer chains. This is combined with alignment (pairwise or to known sequences) through dynamic programming.

The current data set contains all the HERV sequences from the April 2003 (hg15) version of the human genome, from which the *pol* gene sequence can be found. The data contains DNA and translated *pol* protein ("putein"[1]) sequences for the *pol* area. In addition, the primer binding site is known for 1159 sequences. Finally, the RetroTector's estimate of the genus (alpha-, beta-, gamma-, delta- or epsilonretrovirus, spuma- or lentivirus) of the retrovirus is available as well.

The HERVs have traditionally been classified on two different grounds. The first classification stems from the tRNA used to prime DNA synthesis [4], [5]. The classes are named after the primer binding site (PBS); for instance the viruses that are primed by leucine (L) tRNA are called HERVL and those utilizing arginine (R) HERVR. The PBS based classification is, however, incomplete in such cases where HERVs of different origin are primed by the same tRNA, or when the PBS sequence is missing from the HERV.

The other widely used option is to classify HERVs to three classes according to their similarity to types of exogenous retroviruses, from which they presumably stem (see [2], [7], [8], [10]). Class I HERVs are related to gammaretroviruses such as Feline leukemia virus or Gibbon ape leukemia virus

[1] A "putein" is an estimated protein sequence for the ancient retroviral element. During evolution the retroviral element has gone through deletion and insertion mutations in addition to point mutations. In the construction of the "putein", the locations of deletion and insertion mutations are estimated and the translation of the DNA sequence is shifted accordingly to produce a full length protein sequence (with minimal amount of stop codons).

and include HERVH and HERVW, among many other subgroups. Class II HERVs are related to betaretroviruses (Mouse Mammary tumor virus) and alpharetroviruses (Rous sarcoma virus) and include several types of HERVK elements (the HML groups [29]). Class III HERVs are distantly related to spumaviruses (Human foamy virus) and include HERVL and HERVS.

For 2462 sequences in the data set a classification based on sequence similarity of translated *pol* protein sequences to a groups of previously characterized HERV sequences is given. The classification follows to some extent the primer binding site-based grouping, with extra classes for sequences with same PBS but different origins, and for groups with no identified PBS. This classification reflects the current state of the HERV classification, however only 67% of the data set could be rigorously classified in this manner. The classification is one of the following: ERV9, ERV3, HERVRb, HERVI, RHERVI, HERVE, HERVW, HERVH, HUERSP3, MER41, HERVT, MER66, HERV48, HERVFRD, HERV19, HERVFb, HERVFc, HERVADP, HERVS, HERVL, HERVL66, HML1, HML2, HML3, HML4, HML5, HML6, HML7, HML8, HML9, HML10. The nomenclature of the HERV classification is not always the same, mappings between different names are offered in [9], [10].

III. SOM OF THE HUMAN ENDOGENOUS RETROVIRUS COLLECTION

A. Computation of the SOM

The SOM was computed in two stages. In the first organization stage, the sequences were encoded into vectorial representations and the basic self-organizing map algorithm was used to spread the SOM models to cover the whole feature space. In the second stage the Median SOM algorithm was applied. First the model vectors were replaced by the local set medians of the data. The computation was then continued using FASTA-based [26] sequence similarities. This two-stage training scheme has proved to be useful in earlier studies [12], [14], [30]. The rough ordering attained in the first stage enables faster learning of the Median SOM.

In the first stage, we used 4-gram histogram representations of the DNA sequences of the HERV *pol* genes. The feature vectors were 256-dimensional and normalized to unit length. For the 3661-sequence data set, we selected a 20-by-30 units hexagonal SOM in order to achieve a resolution of approximately 6 samples per unit. The 256-dimensional model vectors were initialized randomly. The SOM was computed using the Batch Map algorithm for vectors with standard parameter values [11]. The SOM algorithm is robust to the exact choices of the parameters [11]. Here the width of the Gaussian neighborhood function decreased linearly from 15 to 4 during the 20 iterations of the organization phase and from 4 to 1 during the 20 iterations of the finetuning phase of the algorithm.

The SOM models were then converted to sequences by setting the model to the set median of the sequences in the unit in question and its neighboring map units.

Ten iterations of the Median SOM algorithm were then carried out. A Gaussian neighborhood function was used. Its effective width covered the nearest neighbors on the hexagonal map grid. The distance matrix used in the median SOM algorithm was based on the FASTA similarity scores [26] of the *pol* protein sequences. The FASTA scores were computed with default parameters: BLOSUM50 substitution matrix, penalty for opening a gap $= -10$, and penalty for continuing a gap $= -2$.

Since the lengths of the sequences varied greatly, we normalized the effect of sequence length in the FASTA scores by using the Tanimoto distance [31]. First, the FASTA scores were computed for each pair of sequences. These scores were converted to Tanimoto similarities,

$$s(i,j) = \frac{f(i,j)}{f(i,i) + f(j,j) - f(i,j)}, \tag{3}$$

where $f(i,j)$ denotes the FASTA similarity score between sequences i and j. The Tanimoto similarities are between 0 and 1. The similarities were converted to the Tanimoto distance by taking the negative logarithm of the Tanimoto similarity: $d(i,j) = -\log s(i,j)$.

The 20-by-30-unit Median SOM of HERV sequences is shown in Fig. 1. The shade of gray represents the distance between the models of adjacent map units.

Besides the map shown in Fig. 1, we also computed several other maps with different random vector initializations. Similar data clusterings were generally observed on different maps. The map in Fig. 1 gave the best quantization error.

B. Bootstrapping the SOM

The confidence of the SOM was estimated with the bootstrap procedure. We resampled the data set 100 times and counted the frequencies of each pair of two sequences appearing as neighbors on the bootstrap maps. Then we computed the reliability score (2) for each map unit. A visualization of the reliability scores for each map unit is presented in Fig. 2.

IV. RESULTS

The SOM reflects the division of HERVs into the standard classes I-III. The darkest borders in Fig. 1 divide the map into three major and a few small areas (see Fig. 3). Each major area contains sequences of mainly one genus. The spumaviruslike sequences (Class III HERVs: HERVL and HERVS) are separated to the lower left corner. Similarly, the betaretroviruslike sequences (Class II HERVs: the HML groups) form their own area on the upper left side of the map. The right side of the map is covered by gammaretroviruslike (Class I) elements like HERVH and HERVW.

The SOM display was visually compared to phylogenetic trees (not shown) constructed from the same data set. The main groupings were similar in both methods. Both method separated the three major groups as well as some smaller ones (like HERVE, HML5, HML6, HERVH, HERVF). The SOM had some interesting differences when compared to the phylogenetic trees. Some classes that were separate in the phylogenetic trees were mixed together on the SOM (for example the ERV9, HERVW and HUERSP3 area described later in the text). Furthermore the SOM found several groups

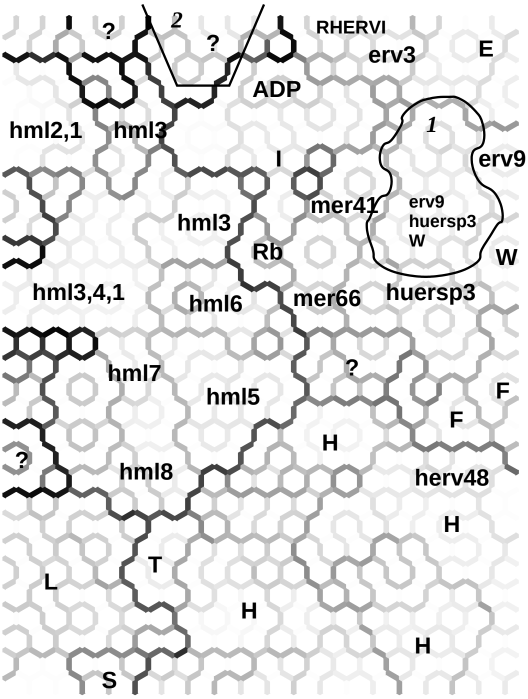

Fig. 1. The SOM of human endogenous retroviruses. The labels in the figure are manually assigned names for different areas of the map. The labels describe the class of the sequences in each area (class names like HERVADP, HERVH, HERVRb etc. have been abbreviated by dropping the "HERV" from the beginning). The question marks are used to mark areas where most of the sequences are unclassified. The gray scale coloring describes the distances between map units; black denotes large distance and white small. The darkest borders divide the three major groups (classes I-III; see Fig. 3) and lighter borders the different groups inside the major groups.

consisting of mainly unclassified sequences. These groups were not visible on the phylogenetic trees. In what follows we describe examples of interesting groups from the reliable areas of the SOM.

The map has an area where ERV9, HERVW and HUERSP3 sequences are mixed together (marked with "1" in the figures). A more detailed picture of this area, showing the mixing of the classes, is presented in Fig. 4. This group of sequences is also a stable structure according to the bootstrap (see Fig. 2). The mixing of the class labels in this group of sequences suggests that the old classifications of these sequences needs to be updated either to form a fourth independent class or to form one large class of all the sequence in classes (ERV9, HERVW and HUERSP3).

To verify that this finding is truly present in the data set and not merely an artifact caused by the visualization, we compared the classification accuracy within this found set with the expected classification accuracy (computed from other samples of the same classes). If the classification accuracy is significantly lower than expected accuracy, it supports the hypothesis that the classes are really mixed for the HERVs

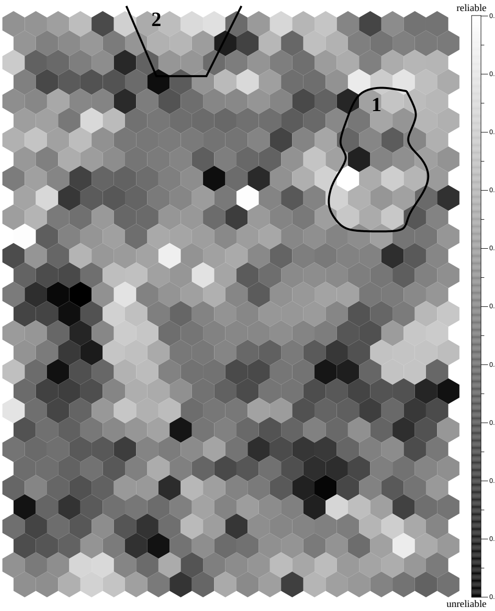

Fig. 2. Gray scale coded reliability of the map units. The reliability value (from 0=black to 1=white) tells the average stability of the neighborhoods of the sequences in each unit. In a white (or light gray) unit most of the sequences appear together on all of the bootstrap maps.

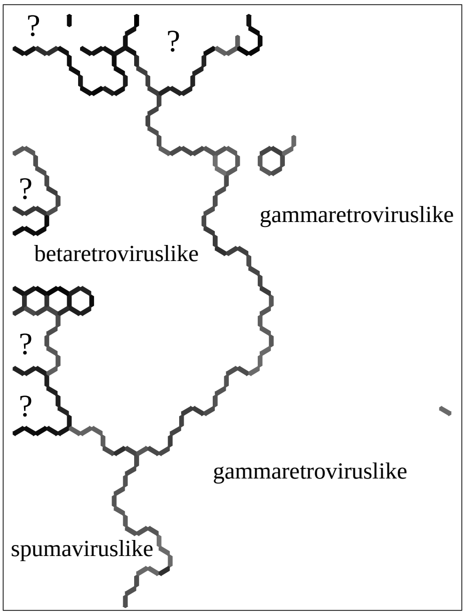

Fig. 3. The three major areas of the SOM of human endogenous retroviruses. The visualization shows only the darkest borders from Fig. 1 (with a suitable cutoff). A dark border represents a large distance between neighboring units, and a continuous dark borderline separates clusters of sequences from each other. Three major areas are visible. They are marked according to the genus of the sequences in each area. The smaller areas (marked with a question mark) contain mainly unclassified sequences of diverse genera.

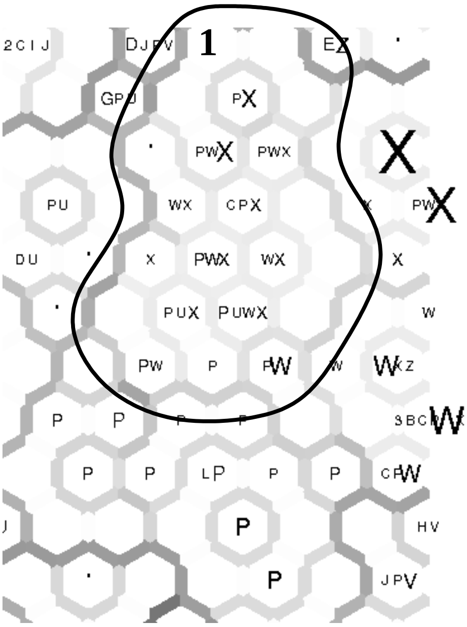

Fig. 4. A close-up of the area where classes HERVW, ERV9 and HUERSP3 are mixed together. The gray-scale is the same as in Fig. 1. The text inside each unit describes the classification given to the sequences mapped into the unit. The one letter coding used to represent the classes is: ERV9 (X), HERVW (W), HUERSP3 (P), ERV3 (Z), HERVRb (B), HERVE (E), MER41 (U), MER66 (G), HERVI (I), HERV48 (V), HERVFRD (D), HERVH (H), HERV19 (J), HERVFc (C), HERVL (L), HML2 (2) and HML3 (3). The size of the letters are proportional to the number of sequences with that label in the map unit e.g. largest X denotes 64 sequences, and the smallest X only 1 sequence – the scale is linear in between. The map units labeled with a dot contain only unclassified data.

within the region.

We compared the K-nearest-neighbor (KNN) classification errors in this selected group of sequences (group A, the selected area is marked with “1” in the figures) to the classification error of the other sequences in the classes ERV9, HERVW and HUERSP3 (group B). This comparison tells us whether the nearest neighbors of the sequences in group A truly are from other classes than the sequence itself, and if this variation differs from the common behavior for the sequences in these three classes. The comparison was done by computing the average KNN error rates (over $K = 1, 2, ..., 10$) for each sequence in each group (A and B). The distributions of the classification errors of the sequences in the two sets were compared with the Wilcoxson rank sum test. The distributions were found to be significantly different with P-value of $p < 10^{-11}$.

On the map there are also areas which do not have a clear interpretation based on either the earlier traditional classifications, the primer binding sites or the retrovirus genera. These areas are marked with a question mark in Fig. 1. For example, the area marked with “2” in the figures is very reliable based on the bootstrap analysis, but only 6 of the 49 sequences within that area have a classification. The reliability of this group of sequences is 0.72 (counted with (2) for the whole group of sequences). Fig. 5 presents the relevant part from the matrix $\mathcal{F}$. It can be seen that all the sequences in this group appear almost always together in the bootstrap repetitions of the map. This unclassified compact group of sequences might be a previously undiscovered HERV class.

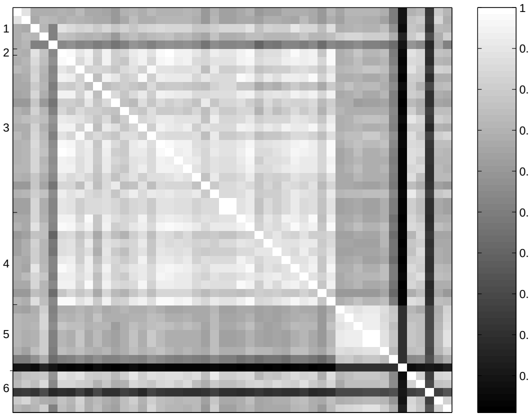

Fig. 5. A selection from the matrix $\mathcal{F}$ corresponding to the area marked with “2” in Fig. 1 and 2. The sequences in this unclassified area are ordered in the figure according to the (nonempty) SOM unit they belong to. The gray levels indicate the frequency for a pair of sequences to appear in the neighborhood of each other on the bootstrap repetitions of the map.

V. Conclusion

The Median Self-organizing Map is suitable for visualizing large collections of sequence data. The major cluster structures visible on the map are in accordance with the current knowledge about human endogenous retroviruses. In addition, the relationships of the HERV classes on the SOM are similar to the results obtained by phylogenetic trees constructed from HERV sequence collections. The phylogenetic trees and the SOM can complement each other when constructing a “final” grouping for all HERV sequences. The phylogenetic trees represent the evolutionary connections between groups of sequences. The SOM, on the other hand, is well suited for analyzing larger collections of sequences simultaneously and for visualizing them on a two-dimensional display. In this work we showed that the SOM was able to extract new knowledge from a HERV sequence collection previously analyzed with phylogenetic trees.

The SOM of human endogenous retrovirus sequences revealed two new groups of HERV sequences. In forthcoming articles we will analyze further these two groups of sequences to verify if they truly are new HERV classes and to characterize their properties.

Our results demonstrate that visualization of the reliability of the SOM is a valuable help in SOM data analysis. Here the bootstrap method was used to estimate the reliability of each map unit. The visualization revealed clusters of high confidence and areas where the visualized similarities are unreliable. The overall reliability of the visualization could be improved by removing the least reliable sequences. This approach will be discussed in future work.

Acknowledgment

The authors would like to thank Teuvo Kohonen and Panu Somervuo (Neural Networks Research Centre, Helsinki University of Technology) developers and implementators of the Median Self-Organizing Map algorithm. The authors would also like to thank Patrik Jern (Department of Medical Sciences, University of Uppsala, Sweden) for useful discussions concerning HERVs.

References

[1] E. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, February 2001.

[2] D. J. Griffiths, "Endogenous retroviruses in the human genome sequence," *Genome Biology*, vol. 2, no. 6, pp. reviews1017.1–1017.5, June 2001.

[3] P. N. Nelson, P. R. Carnegie, J. Martin, H. Davari Ejtehadi, P. Hooley, D. Roden, S. Rowland-Jones, P. Warren, J. Astley, and P. G. Murray, "Demystified ... human endogenous retroviruses," *Molecular Pathology*, vol. 56, pp. 11–18, 2003.

[4] E. Larsson, N. Kato, and M. Cohen, "Human endogenous proviruses," *Current Topics in microbiology and immunology*, vol. 148, pp. 115–132, 1989.

[5] M. Bock and J. P. Stoye, "Endogenous retroviruses and the human germline," *Current Opinion in Genetics & Development*, vol. 10, pp. 651–55, 2000.

[6] M. Lindeskog, "Transcription, splicing and genetic structure within the human endogenous retroviral HERV.H family," Ph.D. dissertation, Lund University, Lund, Sweden, 1999.

[7] M. Tristem, "Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database," *Journal of Virology*, vol. 74, no. 8, pp. 3715–30, April 2000.

[8] L. Bénit, P. Dessen, and T. Heidmann, "Identification, phylogeny, and evolution of retroviral elements based on their envelope genes," *Journal of Virology*, vol. 75, no. 23, pp. 11 709–19, Dec 2001.

[9] D. L. Mager and P. Medstrand, *Encyclopedia of the Human Genome*. Nature Publishing Group, 2004, ch. Retroviral repeat sequences.

[10] R. Gifford and M. Tristem, "The evolution, distribution and diversity of endogenous retroviruses," *Virus Genes*, vol. 26, no. 3, pp. 291–315, 2003.

[11] T. Kohonen, *Self-Organizing Maps*, ser. Springer Series in Information Sciences. Berlin, Heidelberg: Springer, 1995, vol. 30, (Third edition 2001).

[12] T. Kohonen and P. Somervuo, "How to make large self-organizing maps for nonvectorial data," *Neural Networks*, vol. 15, pp. 945–52, 2002.

[13] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen, and E. Castrén, "Trustworthiness and metrics in visualizing similarity of gene expression," *BMC Bioinformatics*, vol. 4, p. 48, 2003.

[14] M. Oja, P. Somervuo, S. Kaski, and T. Kohonen, "Clustering of human endogenous retrovirus sequences with median self-organizing map," in *WSOM'03 Workshop on Self-Organizing Maps, 9-14 Sep 2003, Hibikino, Japan*, 2003.

[15] B. Efron, "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, vol. 7, no. 1, Jan 1979.

[16] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[17] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, no. 1-2, pp. 91–118, July-August 2003.

[18] E. de Bodt and M. Cottrell, "Bootstrapping self-organising maps to assess the statistical significance of local proximity," in *8th European Symposium on Artificial Neural Networks. ESANN"2000. Proceedings. D-Facto, Brussels, Belgium*, 2000, pp. 245–54.

[19] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Pasific Symposium on Biocomputing*, vol. 7, 2002, pp. 6–17.

[20] E. Levine and E. Domany, "Resampling method for unsupervised estimation of cluster validity," *Neural Computation*, vol. 13, no. 11, pp. 2573–2593, 2001.

[21] A. K. Jain and J. V. Moreau, "Bootstrap technique in cluster analysis," *Pattern Recognition*, vol. 20, no. 5, pp. 547–68, 1987.

[22] J. Felsenstein, "Confidence limits on phylogenies: An approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–91, July 1985.

[23] S. Holmes, "Bootstrapping phylogenetic trees: Theory and methods," *Statistical Science*, vol. 12, no. 2, pp. 241–55, 2003.

[24] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.

[25] ——, "Self-organizing maps of symbol strings," Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, Tech. Rep. A42, 1996.

[26] W. Pearson and D. Lipman, "Improved tools for biological sequence comparision," *Proc. Natl. Acad. Sci. USA*, vol. 85, pp. 2444–8, 1988.

[27] G. O. Sperber and J. Blomberg, "RetroTector," unpublished.

[28] http://www.kvir.uu.se/RetroTector/RetroTectorProject.html.

[29] M.-L. Andersson, M. Lindeskog, P. Medstrand, B. Westeley, F. May, and J. Blomberg, "Diversity of human endogenous retrovirus class II-like sequences," *Journal of General Virology*, vol. 80, pp. 255–60, 1999.

[30] P. Somervuo and T. Kohonen, "Clustering and visualization of large protein sequence databases by means of an extension of the self-organizing map," in *Discovery Science. Third International Conference, DS 2000. Proceedings (Lecture Notes in Artificial Intelligence Vol.1967). Springer-Verlag, Berlin, Germany*, 2000, pp. 76–85.

[31] D. Rogers and T. Tanimoto, "A computer program for classifying plants," *Science*, vol. 132, no. 3434, pp. 1115–8, 1960.