Publication V

Electronic version of an article published as:
Merja Oja, Göran O. Sperber, Jonas Blomberg, and Samuel Kaski. Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups. *International Journal of Neural Systems*, 15(3):163–179, 2005. [http://dx.doi.org/10.1142/S0129065705000177]

V

SELF-ORGANIZING MAP-BASED DISCOVERY AND VISUALIZATION OF HUMAN ENDOGENOUS RETROVIRAL SEQUENCE GROUPS

MERJA OJA

Department of Computer Science, University of Helsinki
P.O. Box 68, FI-00014 University of Helsinki, Finland and
Neural Networks Research Centre, Helsinki University of Technology
P.O. Box 5400, FI-02015 HUT, Finland
E-mail: merja.oja@hut.fi

GÖRAN O. SPERBER

Unit of Physiology, Department of Neuroscience, Uppsala University
Box 572, 751 23 Uppsala, Sweden
E-mail: goran.sperber@neuro.uu.se

JONAS BLOMBERG

Section of Virology, Department of Medical Sciences, Uppsala University
Akademiska sjukhuset, 751 85 Uppsala, Sweden
E-mail: jonas.blomberg@medsci.uu.se

SAMUEL KASKI

Department of Computer Science, University of Helsinki and
Neural Networks Research Centre, Helsinki University of Technology
E-mail: samuel.kaski@cs.helsinki.fi

Received (to be inserted
Revised by Publisher)

About 8 per cent of the human genome consists of human endogenous retroviral sequences (HERVs), which are remains from ancient infections. The HERVs may give rise to transcripts or affect the expression of human genes. The first step in understanding HERV function is to classify HERVs into families. In this work we study the relationships of existing HERV families and detect potentially new HERV families. A Median Self-Organizing Map (SOM), a SOM for non-vectorial data, is used to group and visualize a collection of 3661 HERVs. The SOM-based analysis is complemented with estimates of the reliability of the results. A novel trustworthiness visualization method is used to estimate which parts of the SOM visualization are reliable and which not. The reliability of extracted interesting HERV groups is verified by a bootstrap procedure suitable for SOM visualization-based analysis. The SOM detects a group of epsilonretroviral sequences and a group of ERV9, HERVW, and HUERSP3 sequences which suggests that ERV9 and HERVW sequences may have a common origin.

1. Introduction

About eight per cent of human DNA consists of *human endogenous retroviral sequences (HERVs)*.[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 HERVs have inhabited the human genome they have become mutated and broken in crossovers or when transposons[2] have moved to overlap them. It has been suggested that they nevertheless may have functions in regulating the activity of human genes, and may produce proteins under some conditions.[3,4]

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.

The taxonomy of HERVs is still far from complete. Currently the sequences have been classified into ∼30 families,[4,5,6] but as new instances of HERVs are being detected in the human genome, it has become obvious that the classification is not adequate. Some sequences cannot be unambiguously assigned to any family. In addition, in phylogenetic trees constructed from large HERV collections, some of the families are mixed with sequences from other families. A new classification able to resolve these problems is needed. A better and clearer classification of the endogenous retroviral sequences will also help organize the overall retrovirus universe, as most retroviruses are endogenous.

In this work we visualize a massive HERV collection and use the visualization to refine the relationship among the current families, and to detect potentially new families. The traditional way to analyze HERVs is to use phylogenetic trees (PTs) that are based on a multiple alignment of the sequences. The exponential computational complexity of the multiple alignment step makes it difficult to use PTs for more than hundreds of sequences. Heuristic methods exist to overcome this limitation, but the results may be biased.

We will use the Self-Organizing Map (SOM)[7] to visualize the HERV collection. The SOM 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 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 on a two dimensional plane which enables visualizing more neighborhood relationships per sequence than a (phylogenetic) tree. Another advantage is that a separate clustering method is not needed; we can get a visualization for all the data at the same time. Furthermore, new sequences can easily be added to the visualization later on. In an earlier study[8] the SOM was found the be the most reliable alternative among several visualization and clustering methods used for visualizing relationships of input samples.

The Median Self-Organizing Map[7,9] is a variant of SOM capable of handling sequence data. It can be used on any non-vectorial data where pairwise distances can be defined between all input samples. Here we use pairwise distances between HERV protein sequences. This HERV analysis case is the first real application of the Median SOM algorithm.

The SOM-based analysis is complemented with estimation of the reliability of the results. First, we measure reliability of the representation of similarities between sequences in each location on the SOM display. The reliability is estimated with a trustworthiness measure,[8] which is extended here so that the trustworthiness of different areas on the display can be measured. An area in the visualization is trustworthy, if the local neighbors of a sequence can be trusted to be neighbors in the original space as well. The novelty in this approach is the visualization of the relative reliability in each location on the SOM display.

We propose two new measures for evaluating the reliability of groups of sequences extracted from a SOM visualization. These measures are based on the bootstrap method.[10,11] They differ from the bootstrap-based stability measures used in clustering[12,13,14] in that they take the orderedness of the SOM into account. The proposed measures are suitable for visualization-based SOM analysis, where the groups of samples are extracted manually from the SOM.

We apply the combination of median SOM, trustworthiness measure 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 families.

2. Methods

2.1. *Principle of the Median SOM*

The Self-Organizing Map (SOM)[7,9] is an algorithm used to cluster, 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 in several articles[16] or in Kohonen's book.[7]

The SOM can be used to order any kinds of samples, including non-vectorial items. It is sufficient that some distance measure is definable between the elements of a data space. The SOM variant used to order non-vectorial data is called the Median SOM.[9] It resembles the batch-learning version of the SOM.[7,17]

A SOM consists of a regular grid of units. Each unit is represented by a model $\mathbf{m}_i$, normally a vector in the high-dimensional data space. During the SOM teaching process the model vectors spread out in the data space in order to represent all the available input samples. At the same time they become ordered on the grid so that close-by units contain similar model vectors. The input samples are mapped onto the SOM grid to their best-matching models (see Eq. (2)). Each model is then used as a representative for the input samples mapped to it.

In the Median SOM, the model of each unit is defined as the *generalized median* of the input samples mapped into the neighborhood of the unit.[9] The generalized median $\mathbf{m}$ is defined as the hypothetical data sample from which the sum of distances to the other elements $\mathbf{x}(j)$ in a data set $\mathcal{D}$ is minimized, that is

$$\mathbf{m} = \arg\min_{\xi} \sum_{\mathbf{x}(j)\in\mathcal{D}} d(\mathbf{x}(j), \xi), \qquad (1)$$

where $d(\mathbf{x}(j), \xi)$ is some distance measure defined between the $\mathbf{x}(i)$ and ξ. In practice, the generalized median is often approximated by the *set median*, in the above equation ξ is then restricted to being an element of $\mathcal{D}$. The set median is an exact copy of one of the samples in the data set.[a]

The Set Median SOM is computed by iterating the following two steps. In the first step, the input (teaching) sequences $\mathbf{x}(j)$ are mapped to their best-matching models $\mathbf{m}_c$, where

$$c = c(\mathbf{x}(j)) = \arg\min_i d(\mathbf{x}(j), \mathbf{m}_i). \qquad (2)$$

Here $d(\mathbf{x}(j), \mathbf{m}_i)$ is some distance measure defined between $\mathbf{x}(j)$ and $\mathbf{m}_i$. In the second step, for each map unit a new value for the model $\mathbf{m}_i$ 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. This group of sequences is denoted by $N(i)$. In other words, the new $\mathbf{m}_i$ is given by Eq. (1), with $\mathcal{D}$ set to $N(i)$ and with the restriction $\xi \in N(i)$. These two steps are repeated until the models can be regarded as stationary.

In this work, the Median SOM is applied to producing similarity diagrams, and for showing the clustering tendency of HERVs. The similarities between the sequences are computed by the FASTA method (see section 4.1 for more details).[18]

2.2. *The SOM visualization*

The SOM grid is visualized as a two-dimensional display which reveals 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. However, the SOM can sometimes be folded in the input space, which may result in a group of similar samples being divided into two locations on the SOM.

To get insight into the cluster structure of the data, the distances between neighboring units are visualized with gray shading of the unit boundaries on the SOM display (the U-matrix visualization, example in Fig. 4).[19] A cluster is an area of the map where the models of neighboring units are close to each other, that is, the unit boundaries inside a cluster have light shading. Borders between clusters appear as dark edges: at the borders distances between neighboring units are considerably larger.

2.3. *Reliability of the SOM visualization*

[a] If several samples in the data set satisfy Eq. (1), one of them is chosen randomly.

The SOM algorithm searches for a low-dimensional presentation for the data collection. This task requires compromises, as no projection algorithm is capable of visualizing all the similarities among input sequences simultaneously. We accept the fact that some areas of the visualization and the locations of some sequences might be more distorted than others. We complement the SOM visualization with measures of the trustworthiness of the areas (this subsection) and reliability of extracted findings (next subsection). Based on these measures and their visualizations, the data analyst can be more cautious when interpreting the more distorted areas.

We will use a trustworthiness[8] criterion to measure the relative reliability of parts of the visualization. An area on a display is considered *trustworthy*[8] if all samples close to each other on the display can be trusted to have been proximate in the original space as well. The trustworthiness measure has been previously used to estimate the reliability of the whole visualization. The new contribution here is visualization of the trustworthiness value for each unit on the SOM display, computed by defining trustworthiness for individual sequences and averaging them.

By trustworthiness we mean whether we can trust that sequences visualized to be similar (i.e., being close-by on the display) really are similar. We compare the sequences included into the neighborhood of a sequence—on the SOM and in the original data space. Those samples that appear in the neighborhood of a sequence on the map but are not close-by neighbors in the original data space will lower the trustworthiness of the visualization. In other words we are measuring the number of false positives in the observed neighborhood of a sequence, or in information retrieval terminology, precision of the similarities.

When measuring the trustworthiness we must decide how to define the neighborhood of each sequence s. A practical choice would be to select the sequences from the same map unit and its neighboring units up to a pre-selected radius. The problem with this approach is that it disregards clusteredness in the data. Furthermore, the number of neighbors would vary and could lead to low quality trustworthiness estimates when the number is small. To get reliable estimates and to take into account the clustering visible on SOM displays, we select sequences from close-by map units in the order of their distance from the unit where s is. This is computed along the minimal path on the map grid where the distance of neighboring units is defined as their distance in the data space (the U-matrix distance which measures visual closeness). We collect sequences until their number equals or exceeds a preselected number k.[b] This number k should be close to the number of sequences a person looking at the SOM will assume similar.

More formally, let $r(s_i, s_j)$ be the rank of the sequence s_j in the ordering according to distance from s_i in the original data space. Denote by $U_k(s_i)$ the set of those sequences that are in the k-neighborhood of sequence s_i in the visualization display but not in the original data space. Our measure of untrustworthiness of the neighborhood of the sequence s_i is defined by

$$M_{seq}(k, i) = \sum_{s_j \in U_k(s_i)} (r(s_i, s_j) - k) \; . \qquad (3)$$

The untrustworthiness of the map display is estimated at each map unit u separately. Denote the set of sequences within u by I_u. A measure of the untrustworthiness of a map unit is computed as an average over the N_u sequences in the unit by

$$M_{unit}(k, u) = \frac{1}{N_u} \sum_{s_i \in I_u} M_{seq}(k, i) \; . \qquad (4)$$

Similarly, a trustworthiness measure for the whole SOM can be computed as the average over all sequence-wise untrustworthiness values:

$$T(k) = 1 - A(k) \sum_{i=1}^{N} M_{seq}(k, i) \; , \qquad (5)$$

where N is the number of sequences in the data collection and $A(k) = 2/(Nk(2N - 3k - 1))$ is a scaling factor used to scale the trustworthiness values between zero and one.[c] Note that Eqs. (3) and (4) are raw unscaled untrustworthiness values (small value is good) while Eq. (5) is a scaled overall trustworthiness measure (large value is good). The measure of Eq. (5) was used to compare visualization methods in our earlier work.[8]

In this work we will not consider the other aspect of reliability, i.e., the false negatives—neighbors

[b] If the number exceeds k, the average over all possible selections to fill up k should be computed. However, to save time we approximate by the average of the furthest and closest sequences (in terms of the true data space distance).
[c] $A(k)$ is the inverse of a sum of worst-case sequence-wise trustworthinesses.

in the original data space ending up far away on the SOM. There exists a method for measuring this other aspect of the visualization as well—the continuity measure.[8] It could be adapted, similarly as the trustworthiness here, to represent the conservation of neighborhoods of sequences in each map unit. Such a visualization might reveal the foldedness of the map (see section 2.2). Further discussion, and implementation of the unit-wise continuity measure, is left as future work.

2.4. *Measuring the reliability of groups extracted from the SOM*

The SOM visualization can be used to extract interesting groups of similar sequences. The U-matrix visualization shows with gray shades how close the models of neighboring map units are, and clusters can be defined as sets of close-by units. The exact borders of the cluster area on the visualization are selected subjectively. Such decisions are based on various labelings describing the contents of each SOM area, the U-matrix visualization, the trustworthiness visualization, and all background knowledge the analyst has about the data set. Before the group of sequences included in the cluster is analyzed further we want to be certain that the sequences really form a reliable cluster. We use the bootstrap method[10,11] to evaluate this.

The bootstrap method has been used in clustering[12,13,14] to estimate the stability of the discovered clusters. It is assumed that the cluster composition should not change radically between two sets of samples of the same underlying data distribution. Therefore, robustness of a clustering to sampling variability gives support to its validity. This reasoning can also be applied to SOMs: If the neighbors of a sequence are retained in SOMs constructed from different sampled data sets, we can assume that those are reliably neighbors. This assumption can also be extended to any group of sequences always appearing together; we can then assume that the sequences represent a true group (cluster) in the data. However, the measures of cluster stability[12,13,14,15] can not be directly applied here because they apply to a complete clustering, whereas we are primarily interested in single manually extracted clusters. They are formed of sets of close-by map units on the SOM display.

The bootstrap has been applied to Self-Organizing Maps previously.[20] The article describes significance tests for the quantization error and for the stability of neighborhoods on the SOM. In this article we will use bootstrap for a different purpose.

The data set is here 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 SOM is computed from each bootstrap data set to produce B bootstrap SOMs. The bootstrap data sets are 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.

The reliability of a cluster (a group of sequences located in close-by map units) can be measured by observing how the sequences in the cluster behave in the bootstrap repetitions. In an optimal case the whole group would appear together. In practice this does not hold, and the errors need to be quantified. We define two measures, the compactness and purity of the cluster, to estimate the deviations from a perfect clustering on the SOM display. Compactness measures how close together the group of sequences is on the SOM. Purity, on the other hand, measures how many foreign sequences are mapped to the same area as the interesting group. The purity of the cluster is analogous to precision in information recall, and false alarms in detection theory. To our knowledge these measures have not been previously used in this form.

The compactness and purity (C_b and P_b respectively) of the selected group of sequences are measured in each bootstrap repetition. Thus, we get a sampling distribution for each measure. We then estimate the mean and variance of this distribution and use these to analyze the clusteredness of the group of sequences.

Compactness and purity are measured as functions of the varied number of samples in the cluster, to take into account possible substructure in the clusters. First the measures are evaluated for the whole group, then the sequence which most worsens the measure is removed from the group and the measure is evaluated again. The removal of the worst sequence and re-evaluation of the measure is repeated until no sequences are left. The removal of sequences is carried out separately for the two measures. In the

case of purity this removal is optimal. Usually also the compactness improves steadily with this removal procedure which is a greedy approximation.

More formally, the compactness $C_b(k)$ after k removals is defined as

$$C_b(k) = 1 - \frac{\max_{i,j \in \mathcal{C}_k} d(u(i), u(j))}{D_{max}}, \qquad (6)$$

where $\mathcal{C}_k$ indexes the sequences still remaining in the set after k sequences have been removed, $u(i)$ is the location of the map unit containing the sequence i, d denotes the Euclidean distance on the SOM display (distance between the centers of bordering units is one), and D_{max} is the maximum distance between units on the SOM. The purity $P_b(k)$ after k removals is defined to be

$$P_b(k) = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \frac{\sum_{j \in \mathcal{C}_k} \text{InSameUnit}(i, j, b)}{\sum_{j=1}^{N} \text{InSameUnit}(i, j, b)}, \qquad (7)$$

where $\text{InSameUnit}(i, j, b)$ is the indicator function that returns 1 if i and j are in same unit on the SOM of the bootstrap sample b, and otherwise zero.

We can draw curves representing the measure as a function of the number of sequences removed. The curves are analogous to receiver operating characteristic (ROC) curves. A sharply rising curve is better that a nearly linear one. A sharp incline tells that the group of sequences is very homogeneous. Removing merely the few worst sequences brings the group's performance to the highest level.

2.5. *Comparing the trustworthiness of SOM and phylogenetic trees*

The ability of the Median SOM to visualize the HERV sequences is verified by comparing it to the traditional method of HERV sequence analysis, the phylogenetic trees. A phylogenetic tree (PT) can be constructed from a similarity matrix using the Neighbor-Joining Method.[21] The methods can be compared by applying both to the same similarity matrix. We need to use the pairwise similarity matrix, as a multiple alignment-based similarity matrix is too heavy to compute for thousands of long sequences.

We evaluate the results of SOM and the Neighbor-Joining Tree by measuring how well each *visualizes* the original neighborhood relations among the sequences. This is done using the trustworthiness measure described in section 2.3.

The definition of a neighborhood is not clear-cut in the phylogenetic tree. It can be defined along paths through the structure of the tree, taking into account the length of the branches, but in practice a user may mainly look at the linear order of the tree leaves. Defining fixed-size neighborhoods on the SOM has technical complications as well. Thus, we have included, for both methods, several ways of defining the neighborhood. For SOM, we have reported the best possible selection of k neighbors as well as the average over the possible selections (see section 2.3).

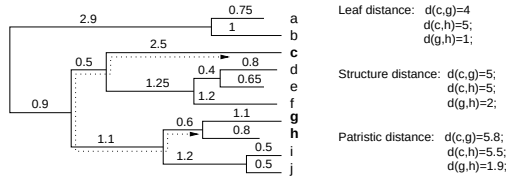


Fig. 1. Sample phylogenetic tree and distance measures between leaves of the tree.

For the phylogenetic tree the leaf order is not defined unambiguously since any two dividing branches can be flipped around in the tree, and the tree still remains the same. Bar-Joseph *et al.* have proposed an algorithm to produce an optimal leaf order for a tree structure.[22,23] We used their algorithm and then defined the neighborhood to be the k closest sequences along the linear leaf order.

For comparison we compute also neighborhoods through the structure of the tree. The patristic distance is computed by adding the branch lengths together along a path leading from one leaf to another. The structure distance is the patristic distance in a tree where all branches are set to have unit length. See Figure 1 for an example.

Note, however, that the data analyst looking at a tree visualization is unable to perceive the distance along the structure reliably for sequences not in the same branch. Furthermore, the exact branch lengths are even harder to perceive and they are often even left out from the visualization of a phylogenetic tree (see Fig. 10 for an example). In other words, the patristic and structure distances do not reliably describe the *visualized* distances between sequences. As we are interested in the visualization potential of the PT, we will not use the patristic distance to com-

pare SOM and PT. However, we can assume that the data analyst will be able to perceive some of the structure, and not only the linear order at the leaves. Thus we will use the structure distance, as an alternative to the leaf order, to compare the SOM and PT visualizations.

3. Collection of human endogenous retroviral sequences

The data set consists of 3661 HERVs automatically collected from the human genome by RetroTector©[24] RetroTector© is a program developed for the detection of endogenous retroviral sequences and other related 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 HERVs from the April 2003 (hg15) version of the human genome, from which RetroTector© finds the *pol* gene sequence. The data contains DNA and translated protein ("putein")[d] sequences for the *pol* area. The average length of the protein sequences is approximately 880 base pairs (bp). The shortest sequence is 114 bp and the longest 2463 bp long, and 90 % of the sequences have lengths between 610-1170 bp. The primer binding site is known for 1159 sequences. Finally, 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 into families on two different grounds. The first classification stems from the tRNA used to prime DNA synthesis.[25,26] The families 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 those cases where HERVs of different origin are primed by the same tRNA, or when the PBS sequence is missing from the HERV. Nevertheless, the names for most of the current families stem from this classification.

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.[3,27,28,6] These classes are broad and include the various families. Class I HERVs are related to gammaretroviruses such as Feline leukemia virus or Gibbon ape leukemia virus 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 the translated *pol* protein sequences to a group of previously characterized HERVs is given. The classification follows to some extent the primer binding site-based grouping, with extra groups for sequences with the same PBS but different origins, and for groups with no identified PBS. This classification reflects the current state of the HERV classification,[5,6]. However, only 67% of the data set could be rigorously classified in this manner. The current HERV families are: 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, and HML10. The nomenclature of the HERV classification is not always identical: mappings between different names are offered in.[5,6]

4. SOM of the human endogenous retrovirus collection

4.1. *Computation of the SOM*

The SOM was computed in two stages. In the first, initialization stage, the sequences were encoded into vectorial representations and the basic Self-

[d] 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 number of stop codons).

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. The rough ordering attained in the first stage enables faster learning of the Median SOM. This two-stage training scheme has proved to be useful in earlier studies.[9,30,31]

In the initialization stage, the DNA sequences of the HERV *pol* genes were transformed into vectorial representations. We used a 4-mer histogram representation, where each component in the vector measures how often a specific 4-mer, a contiguous subsequence of length 4, is observed in the DNA sequence. The DNA sequence was used instead of the protein sequence to limit the length of the vectors; the dimensionality of the DNA 4-mer vector is 4^4=256, where as a protein 2-mer would be 20^2=400-dimensional. The 256-dimensional feature vectors were normalized to unit length. For the 3661-sequence data set, we selected a 20-by-30-unit hexagonal SOM, fixing the resolution to approximately 6 samples per unit. The 256-dimensional model vectors were initialized randomly. The SOM was computed, with standard parameter values,[7] using batch-learning. The SOM algorithm is robust to the exact choices of the parameters.[7] 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 fine-tuning phase of the algorithm.

The SOM models were then converted to sequences by setting each model to the set median of the sequences in the map unit and its neighboring units using Eq. (1), with $d(x, y)$ being the Euclidean distance between the 4-mer feature vectors.

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[18] of the *pol* protein sequences. FASTA is a heuristic method that first searches for short identical stretches from the pair of sequences, and then applies the optimal Smith-Waterman algorithm[32] to a limited search space around the best initial alignment. Here 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.[33] First, the FASTA scores were computed for each pair of sequences. These scores were converted to Tanimoto similarities defined by

$$s(i,j) = \frac{f(i,j)}{f(i,i) + f(j,j) - f(i,j)}, \qquad (8)$$

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

The 20-by-30-unit Median SOM of HERVs is shown in Fig. 4. Besides the map shown in Fig. 4, we also computed several other maps with different random vector initializations. Similar data clusterings were generally observed on different maps. The map in Fig. 4 gave the best quantization error.[e]

The initialization SOM was compared to the final Median SOM to verify that the Median SOM phase was indeed necessary. The trustworthiness (see Eq. (5)) of the initialization SOM was worse for all k.

4.2. *Visualizing the reliability of the SOM*

The relative reliability of the different areas of the SOM visualization was estimated with the trustworthiness measure. First $k = 40$ nearest neighbors on the SOM for each sequence were collected using the procedure described in section 2.3. The parameter k was set so that it approximates the number of sequences in one map unit and its 6 immediate neighbors. The SOM has, on average, a little over 6 sequences per unit. Hence, a map unit and its neighboring units contain approximately 42 sequences.

The trustworthiness for each map unit was computed using Eq. (4); the untrustworthiness values are visualized on the SOM display in Fig. 2. In a white (or light gray) unit the average trustworthiness of

[e]The quantization error $E_q = \sum_{\mathbf{x}(i)\in\mathcal{D}} d(\mathbf{x}(i), \mathbf{m}_{c(\mathbf{x}(i))})$, where same notation is used as in Section 2.1, and d is the Tanimoto distance.

the sequences in the unit is very good. There are few sequences in the SOM neighborhood that are not in the neighborhood of the sequences in the original data space. The marked areas are examples that will be analyzed in the Results section. As can be seen from the image, while the whole SOM is reasonably reliable the marked areas are not particularly reliable and hence some caution is in place.

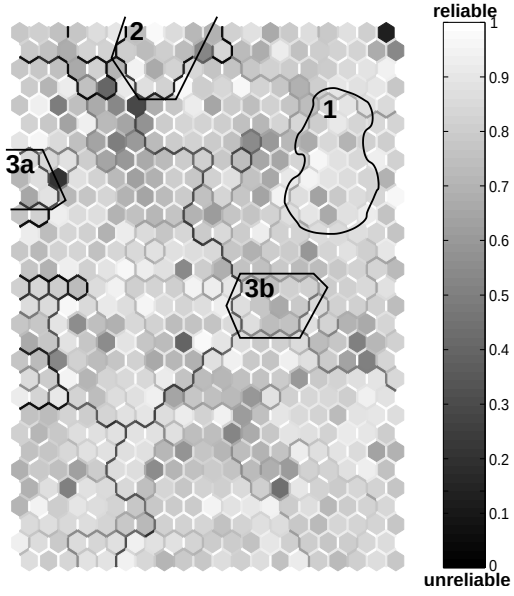


Fig. 2. Relative reliability of the map units visualized by gray shading. The reliability value tells the average stability of the neighborhoods of the sequences in each unit. The scale is from white="all neighbors preserved" to black="all neighbors could as well have been selected randomly". Empty map units have been colored with a gray value which is the average of its nonempty neighbors' trustworthiness values.

4.3. *The trustworthiness of SOM and phylogenetic tree*

A phylogenetic tree was constructed from the same pairwise similarity matrix as the SOM (the Tanimoto similarity), using the NEIGHBOR algorithm in the PHYLIP program package.[34]

Then the trustworthiness of each method was computed using the trustworthiness measure of Eq. (5). Fig. 3 shows the results for the neighborhood definitions described in section 2.5 (leaf order, structure and patristic distances for the phylogenetic tree (PT) and best possible and average trustworthinesses for the SOM).

The separation of the two curves for SOM is large due to the quantization of samples into map units. A large number of samples are within an equal distance from a selected sequence. This stems from the summarization ability of the SOM, which causes the difference between the best trustworthiness value achievable and the average trustworthiness to be dependent on the number of sequences in the map units. In a phylogenetic tree only two sequences can have an equal distance from a sequence in the linear leaf order.

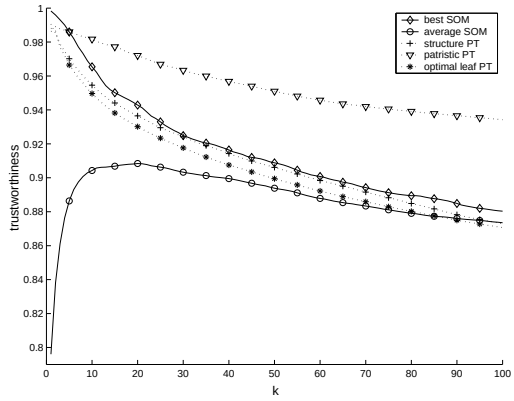


Fig. 3. Trustworthiness of SOM and Phylogenetic tree (PT) for varying number of neighboring sequences considered.

At its best the SOM outperforms the phylogenetic tree. The curves representing the trustworthiness of the SOM enclose the curves representing the leaf order distance and the structure distance-based trustworthiness measures for the phylogenetic tree. When the number of sequences in the neighborhood increases to a number used when analyzing the SOM in practice (40 and above), the difference between the SOM and PT curves is relatively small. Hence, the SOM is a reasonable although not clearly better choice for large scale sequence analysis.

The patristic distance is here included only for completeness. When comparing the visualizations of SOM and PT, the patristic distance is not realistic. A human observer looking at a patristic tree will not

be able to *visually* compare the exact branch lengths between several sequences at once. See section 2.5 for more details.

5. Biological Results

The Self-Organizing Map of human endogenous retroviruses is shown in Fig. 4. The SOM finds the division of HERVs into the standard classes I-III, providing it further support. Each class (i.e. genus) is localized to its own area on the display as shown in Fig. 5. Note how the classes are separated from each other by the darkest U-matrix distances in Fig. 4.

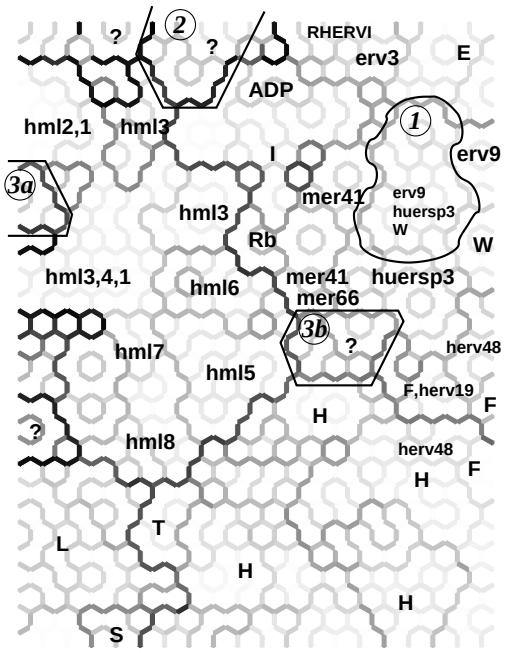


Fig. 4. The SOM of human endogenous retroviruses. The labels in the figure are manually assigned names for different areas of the map, describing the HERV families of the sequences in each area (group names like HERVADP, HERVH, HERVRb etc. have been abbreviated by dropping the "HERV" from the beginning). The question marks denote areas where most sequences are unclassified. The gray shading describes the distances between map units; black denotes large distance and white small.

The previously characterized HERV families can also be detected by the SOM. The class distribution of each family is focused on a group of nearby map units. Only few families spread out more, or mix with other families, reflecting the uncertainty in the current HERV classification. Some examples have been collected to Fig. 6.

The SOM display was visually compared to phylogenetic trees (not shown) constructed from representative subsets of the HERV sequence collection (500 sequences). The main groupings were similar

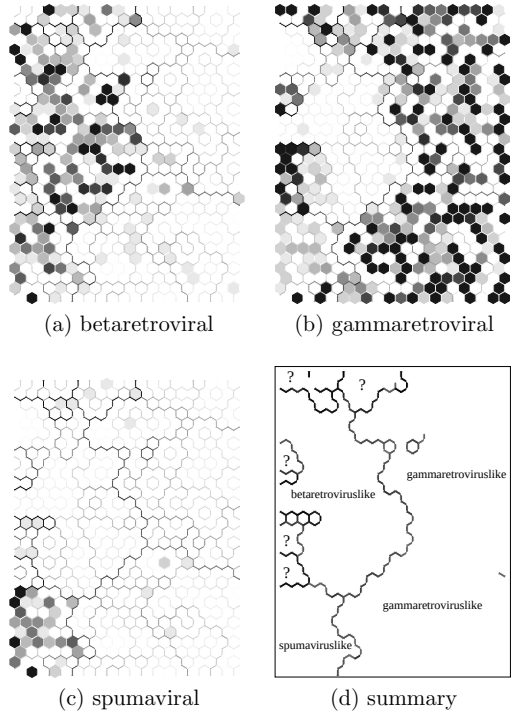
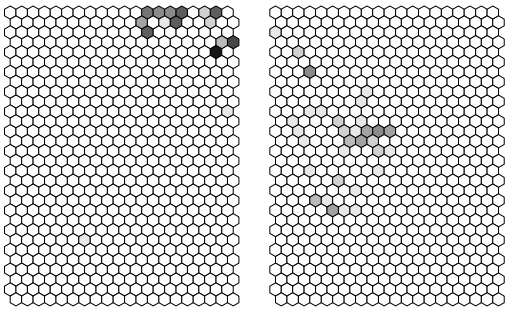
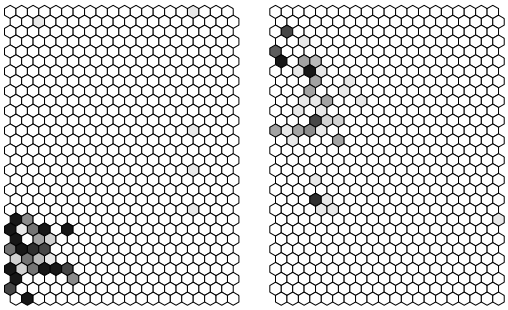


Fig. 5. Genera of the human endogenous retroviruses. The distribution of sequences of each genus is plotted on the displays (a)-(c), one genus per display. The scale is linear between 0 (white) and 10 or more (black). The genera of the sequences divide the SOM to three major areas, summarized in (d). This visualization shows only the darkest borders from Fig. 4 (with a suitable cutoff). The spumaviral sequences are the oldest type of HERVs, thus their genus is difficult to estimate and some have been additionally classified as betaretroviral or gammaretroviral (a sequence can have multiple genus labels).

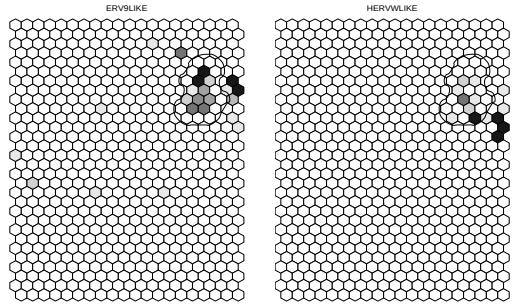


(a) ERV3 (b) HML6

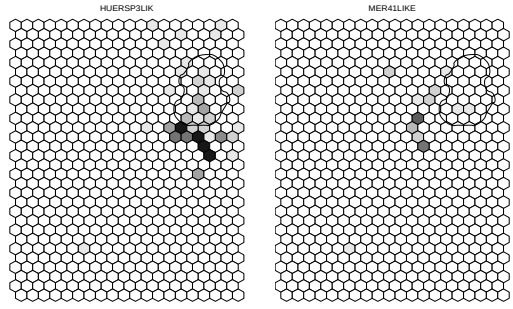


(c) HERVL (d) HML1

Fig. 6. Sample class distributions on the HERV SOM. Distributions of ERV3 and HML6 are more widespread than the labeling in Fig. 4 is able to represent. Most of the families remain in one cluster area on the SOM, such as HERVL and HML1. The scale in all the displays is the same: linear between 0 (white) and 10 or more (black).



(a) ERV9 (b) HERVW



(c) HUERSP3 (d) MER41

Fig. 7. Class distributions of ERV9, HERVW, HUERSP3 and MER41.

in both methods. However, SOM detected biologically interesting sequence groups that were not visible in the phylogenetic trees. One is a group where three HERV groups, previously thought to be separate mix together. Furthermore, SOM detects several groups of unclassified sequences (marked with '?' in the figures), one of which turns out to be a group of chimeric HERV elements and another to be a group of epsilonretroviral sequences. Epsilonretroviruses have not been previously detected in humans. These groups are described and analyzed in the following subsections.

5.1. *Area of ERV9, HERVW, and HUERSP3 sequences*

The map has an area where ERV9, HERVW, and HUERSP3 sequences are mixed together. The mixing suggests that the old classifications of the sequences in the area need to be updated, either to form a fourth independent family or to form one large family of all the ERV9, HERVW and HUERSP3 sequences. The interesting area is marked with "1" in the figures. The borderline of the area was set manually to follow continuous dark gray edges in U-matrix and a string of empty units. A group of 125 sequences was extracted from the area for further analysis. The class distributions of the families are visualized in Fig. 7.

In the following, we will analyze in more detail the set of sequences within Area 1. First, it is verified that mixing of the HERV families is truly happening in this group. Second, the group of sequences is found to be reliable by the bootstrap analysis. Third, biological analysis shows that that the HERV families HERVW and ERV9 form a single HERV family, and that other HERV families such as HUERSP3 are

very similar to this new HERV family.

First, we will want to verify that there really is mixing between the families ERV9, HERVW and HUERSP3. We formulate this as a technical hypothesis that the nearest neighbors for the sequences in Area 1 are not necessarily from the same family as the sequence itself. If the families weren't mixed this would, of course, not be true. To test our hypothesis we computed the K nearest neighbor classification errors for 1) the group of sequences from Area 1 and 2) all sequences from families ERV9, HERVW and HUERSP3. The distributions of the classification errors in the two sets were compared with the Wilcoxson rank sum test. The distributions were found to be significantly different ($p < 10^{-11}$), which supports the hypothesis that the classes are really mixed for the HERVs in Area 1.

Second, we will estimate the reliability of Area 1. First of all, the area is trustworthy (see Fig. 2), which is one of the reasons why we selected this area for analysis. We additionally wanted to verify that the group of sequences in Area 1 is a true group (cluster) in the data set. This was estimated using the bootstrap procedure explained in Section 2.4.

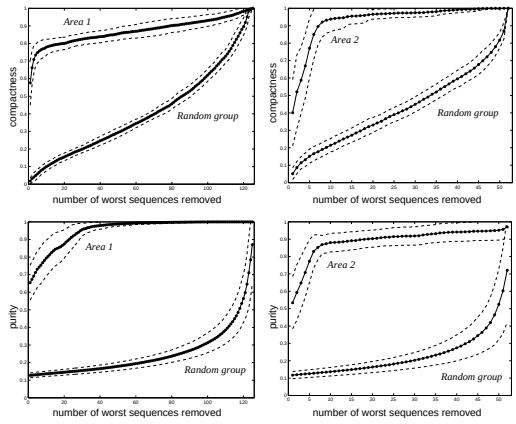


Fig. 8. Compactness (first row) and purity (second row) for the groups of sequences in Areas 1 (left) and 2 (right). The solid curves represent the average compactness (purity) over the 100 bootstrap maps, the dashed lines the mean ± standard deviation of the distribution. Note that the compactness value of one is not attainable in practice for large groups of sequences as they usually do not fit into a single map unit. For reference, the compactness of the group of sequences in Area 1 in Fig. 4 is 0.86.

The compactness and purity of this cluster are very good compared to an equal-sized randomly sampled control set (see Fig. 8). Both compactness and purity rise rather quickly to the reasonable level of 0.86 and 1, respectively; these figures are measured from the original SOM of Fig. 4 used for *defining* the group, and hence represent a kind of best possible reasonable values. This confirms that the cluster is not an artifact, but really exists in the HERV sequence collection. The biological analysis below agrees with this statement.

Third, we analyze the biological significance of this group. A phylogenetic tree was constructed from the sequences in Area 1 and areas surrounding it to analyze the evolutionary relationships of the families. The tree depicted in Fig. 10 consists of sequences from families HERVW, ERV9, HUERSP3, MER41, MER66, and HERVRb. It is evident that the sequences of Area 1 are mixed together in the tree and do not follow the traditional separation of the three families HERVW, ERV9, and HUERSP3. The sequences in Area 1 form a clear cluster in the tree, as predicted by the SOM and bootstrap reliability analysis. The tree suggests that ERV9 sequences may be ancestral to HERVW. The relationship between ERV9 and HERVW has been unclear previously.

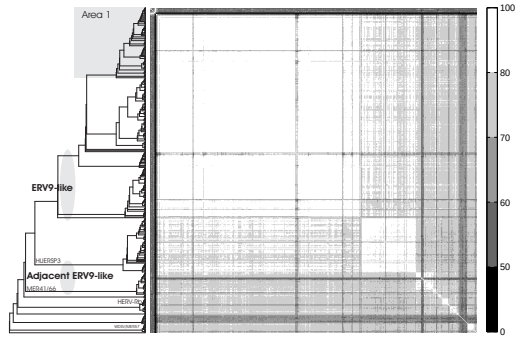


Fig. 9. Similarity matrix of HERV sequences from classes HERVW, ERV9, HUERSP3, MER41, MER66, and HERVRb. The distance matrix is in the same order as the leaves of the cladogram in Fig. 10. Similarities above 80% are visualized with white, and similarities below 50% with black. More detailed versions of this image are available at WWW supplemental information page.[35]

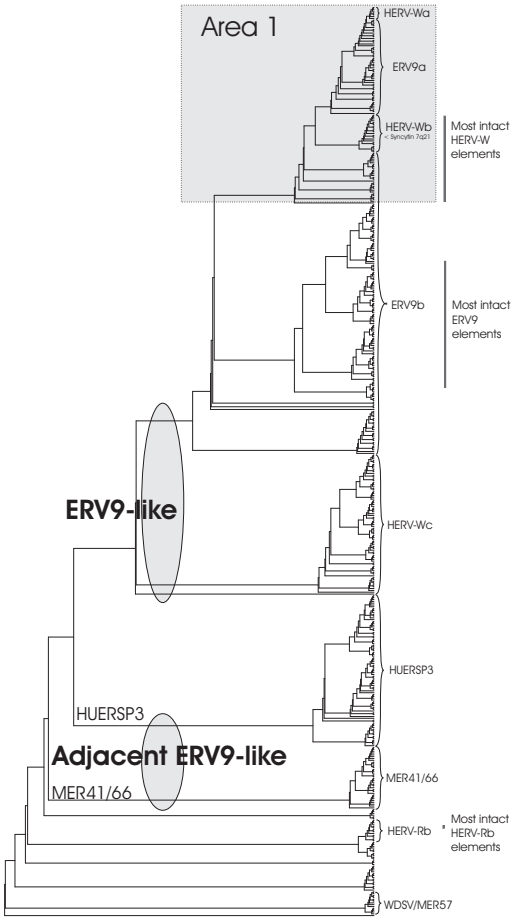


Fig. 10. Cladogram of a phylogenetic tree of HERVs from classes HERVW, ERV9, HUERSP3, MER41, MER66, and HERVRb. A more detailed version of this image, including a description for each sequence, is available at WWW supplemental information page.[35]

Fig. 9 shows the multiple alignment-based similarity matrix for the sequences included in the phylogenetic tree in Fig 10. Sequences marked as ERV9-like in the tree bear a similarity above 80% with each other. Similarly, within the branches HUERSP3 and MER41/66 the sequences have high mutual similarity. These sequences marked as adjacent ERV9-like are above 70% similar to the ERV9-like sequences.

To conclude, the ERV9 and HERVW sequences cannot be separated into two families but form one group of ERV9-like sequences. The HUERSP3 sequences and the MER41/66 sequences are rather similar to this group and will thus be called "adjacent ERV9-like" sequences.

5.2. *An unclassified area*

The SOM detected several areas containing mainly unclassified sequences which could not be unambiguously assigned to any HERV family. For example, Area 2 contains 46 unclassified sequences and only 6 classified ones. The manually set borders of Area 2 follow the black edges in the U-matrix on the left and lower right borders of the area. The top right corner could have included three more units, but these were left out because two of them are relatively untrustworthy.

The reliability of this area was analyzed and reasons for the observed lower reliability are discussed below. Biological analysis showed that the sequences are chimeric elements containing parts similar to HERVs and other parts similar to non-HERV repetitive elements.

First of all, we studied the reliability of the group of sequences using bootstrap. The compactness and purity measures were evaluated for this area similarly as for Area 1, and are presented in Fig. 8. The compactness reaches quickly the reference level (compactness of this group on the original SOM) of 0.89, which means that the sequences are always very close to each other on the bootstrap SOMs. On the other hand, the purity of this group of sequences in not so good suggesting that other sequences are repeatedly placed into this group on the bootstrap SOMs. The added sequences are probably outliers: they have large distance to most other data points and do not have a representative area on the SOM which naturally focuses mostly on representing the bulk of data. Some outliers may become projected more or less randomly, in some cases to the cluster in consideration.

A multiple alignment for these sequences revealed that they are similar to LINE elements (Long Interspersed Nuclear Elements). LINEs occur in over 100,000 copies in the human genome.[36] They transpose actively, and occasionally integrate into HERVs. The SOM detected these LINE-containing HERVs and grouped them separately from the pure HERVs.

The trustworthiness of this area is lower when compared to some other regions of the map. This

is probably because the sequences in Area 2 have high similarity also to sequences that do not contain LINE segments. Each Area 2 sequence will probably be similar to a different non-LINE HERV based on the retroviral part of its sequence.

5.3. *Epsilonretroviral group*

During our analyzes we found a group of epsilonretroviral sequences. This is the first study where epsilonretroviral sequences have been found from the human genome. We briefly describe the properties of the group and their location on the SOM.

Epsilonretroviruses are primarily known from fish and amphibians.[37,38,39,40] Although related to gammaretroviruses, epsilonretroviruses are a separate retroviral genus. It was therefore a surprise to find similar sequences also in the human genome.

The epsilonretroviral sequences have been included into the phylogenetic tree in Fig. 10, where they form a tight group of their own (marked with WDSV/MER57).[f] The epsilonretroviral branch of the tree has two subparts; the upper and lower branches.

The epsilonretroviral HERVs appear at two locations on the SOM (marked with "3a" and "3b" in the figures; the manually set borders follow dark gray U-matrix edges). The SOM areas 3a and 3b are relatively similar to each other, because the SOM is folded here (see section 2.2). Area 3a contains the nine sequences from the upper branch (in the PT) of epsilonretroviral sequences, and 12 unclassified others. Area 3b contains four sequences from the lower branch and 44 other unclassified sequences. Analyzing the remaining sequences in the Areas 3a and 3b will possibly bring forth more elements related to epsilonretroviral sequences.

6. Conclusion

The Median Self-Organizing Map is suitable for visualizing large collections of sequence data. In our endogenous retrovirus study, the major cluster structures visible on the map were in accordance with the current knowledge about human endogenous retroviruses. In addition, the relationships of the HERV families on the SOM are similar to the results obtained with phylogenetic trees constructed from HERV collections. The phylogenetic trees and the SOM can complement each other when constructing a "final" classification for all HERV sequences. The phylogenetic trees try to represent 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 addition, the SOM is able to shed light into non-hierarchical (polyphyletic) structures, such as the chimeric sequences in Area 2, and the mixture of families ERV9 and HERVW observed in Area 1.

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. The epsilonretroviral sequences are a completely new group, previously undetected in the human genome. The new "adjacent ERV9" group consists of HERVW and ERV9 -families, previously thought to be separate. The SOM was an invaluable part of the process of defining the relationships of ERV9 and HERV-W sequences.

The combination of Median Self-Organizing Maps and reliability estimation can be applied to any large non-metric data set, where a distance measure can be defined between the data samples. One application area is the analysis of biological sequence collections.

The proposed reliability estimates give strength to visualization-based data analysis. The bootstrap-based reliability estimates validate the clusteredness of groups of sequences selected manually based on the SOM display and all available background information. These estimates can be applied also in other cases where a SOM is used. The estimates are not limited to pairwise distance matrixes, but can be applied also to vectorial SOMs. Furthermore, the trustworthiness visualization can be applied to various visualization methods. For example, all individual points in a projection or all leaves and inner vertexes in a hierarchical clustering tree could be colored according to their trustworthiness.

In this work we selected the Median SOM for the visualization and clustering method based on an earlier study where a vectorial SOM performed

[f] Walleye dermal sarcoma virus (WDSV) is an exogenous epsilonretrovirus.

better in the sense of trustworthiness than hierarchical clustering and some alternative visualization methods. Later we have compared additionally with newer methods, with similar conclusions.[41] However, the results obtained from the comparison of phylogenetic tree and Median SOM in the present paper suggest that the performance of Median SOM is not necessarily better than some of the alternatives. A more extensive comparison is needed to clarify the differences between Median SOM and other visualization and clustering methods for pairwise data sets.

The current work used pairwise distances in the computation of the SOM for computational reasons. The algorithms for computing multiple sequence alignments (MSA) could be used instead, in principle at least. They have traditionally been formidably slow but are currently improving rapidly, and soon accurate alignment of thousands of long sequences will be possible.[42] At such a stage, a quantitative comparison between MSA-based phylogenetic trees and the approach used in this article will become possible. Still, the comparison of the actual results obtained by each method will remain qualitative, because of the subjective aspect in visualization-based data analysis. Nevertheless, a comprehensive comparison between MSA-based methods and pairwise distance-based methods is interesting also in the larger perspective. If the pairwise methods will be able to perform almost as well as the MSA methods, being faster they can be used instead of MSA methods in preliminary analyses.

7. Acknowledgment

The authors would like to thank Teuvo Kohonen and Panu Somervuo (Neural Networks Research Centre, Helsinki University of Technology), developers and implementors of the Median Self-Organizing Map algorithm; Jarkko Venna (Neural Networks Research Centre, Helsinki University of Technology) for the implementation of the trustworthiness measure evaluation; and Patric Jern (Department of Medical Sciences, University of Uppsala, Sweden) for useful discussions concerning HERVs.

8. References

1. Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome," *Nature*, **409**, 860–921, 2001.
2. H. H. Kazazian Jr. "Mobile elements: Drivers of genome evolution," *Science*, **303**, 1626–32, 2004.
3. D. J. Griffiths. "Endogenous retroviruses in the human genome sequence," *Genome Biology*, **2**, reviews1017.1–1017.5, 2001.
4. 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*, **56**, 11–18, 2003.
5. D. L. Mager and P. Medstrand. "Retroviral repeat sequences." *Encyclopedia of the Human Genome*. Nature Publishing Group, 2004.
6. R. Gifford and M. Tristem. "The evolution, distribution and diversity of endogenous retroviruses," *Virus Genes*, **26**, 291–315, 2003.
7. T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 1995. (Third edition 2001).
8. S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen, and E. Castrén. "Trustworthiness and metrics in visualizing similarity of gene expression," *BMC Bioinformatics*, **4**, 48, 2003.
9. T. Kohonen and P. Somervuo. "How to make large self-organizing maps for nonvectorial data," *Neural Networks*, **15**, 945–52, 2002.
10. B. Efron. "Bootstrap methods: Another look at the jackknife," *Ann. Stat.*, **7**, 1979.
11. B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall, 1993.
12. A. K. Jain and J. V. Moreau. "Bootstrap technique in cluster analysis," *Pattern Recogn.*, **20**, 547–68, 1987.
13. E. Levine and E. Domany. "Resampling method for unsupervised estimation of cluster validity," *Neural Comput.*, **13**, 2573–2593, 2001.
14. 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," *Mach. Learn.*, **52**, 91–118, 2003.
15. S. Dudoit and J. Fridlyand. "A prediction-based resampling method for estimating the number of clusters in a dataset," *Genome Biology*, **3**, research0036.1–0036.21, 2002.
16. T. Kohonen. "The self-organizing map," *Neurocomputing*, **21**, 1–6, 1998.
17. T. Kohonen. "Self-organizing maps of symbol strings." Technical Report A42, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.
18. W. Pearson and D. Lipman. "Improved tools for biological sequence comparision," *Proc. Natl. Acad. Sci.*, **85**, 2444–8, 1988.
19. A. Ultsch and H. P. Siemon. "Exploratory data analysis: Using Kohonen networks on transputers." Technical Report 329, Univ. of Dortmund, Dortmund, Germany, 1989.
20. E. de Bodt and M. Cottrell. "Bootstrapping self-

organising maps to assess the statistical significance of local proximity," in *Proc. ESANN'2000, 8th European Symposium on Artificial Neural Networks*. 245–54, 2000.
21. N. Saitou and M. Nei. "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Mol. Biol. Evol.*, **4**, 406–25, 1987.
22. Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola. "Fast optimal leaf ordering for hierarchical clustering," *Bioinformatics*, **17**, S22–S29, 2001.
23. Z. Bar-Joseph, E. D. Demaine, D. K. Gifford, N. Srebro, A. M. Hamel, and T. S. Jaakkola. "K-ary clustering with optimal leaf ordering for gene expression data," *Bioinformatics*, **19**, 1070–8, 2003.
24. G. O. Sperber and J. Blomberg. "RetroTector." In preparation.
25. E. Larsson, N. Kato, and M. Cohen. "Human endogenous proviruses," *Curr. Top. Microbiol.*, **148**, 115–132, 1989.
26. M. Bock and J. P. Stoye. "Endogenous retroviruses and the human germline," *Curr. Opin. Genet. Dev.*, **10**, 651–55, 2000.
27. M. Tristem. "Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database," *J. Virol.*, **74**, 3715–30, 2000.
28. L. Bénit, P. Dessen, and T. Heidmann. "Identification, phylogeny, and evolution of retroviral elements based on their envelope genes," *J. Virol.*, **75**, 11709–19, 2001.
29. M. Andersson, M. Lindeskog, P. Medstrand, B. Westley, F. May, and J. Blomberg. "Diversity of human endogenous retrovirus class II-like sequences," *J. Gen. Virol.*, **80**, 255–60, 1999.
30. M. Oja, P. Somervuo, S. Kaski, and T. Kohonen. "Clustering of human endogenous retrovirus sequences with median self-organizing map," in *Proc. WSOM'03, Workshop on Self-Organizing Maps*, 2003.
31. 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. Proceedings of the Third International Conference*, 76–85, 2000.
32. T. F. Smith and M. S. Waterman. "Identification of common molecular subsequences," *J. Mol. Biol.*, **147**, 195–7, 1981.
33. D. Rogers and T. Tanimoto. "A computer program for classifying plants," *Science*, **132**, 1115–8, 1960.
34. J. Felsenstein. "PHYLIP (phylogeny inference package) version 3.6." Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2004.
35. Supplemental information for the article is available at: `http://www.cis.hut.fi/projects/mi/data/ijns05/supplemental.html`.
36. A. V. Furano. "The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons," *Prog. Nucleic Acid Re.*, **64**,255–94, 2000.
37. D. L. Holzschu, D. Martineau, S. K. Fodor, V. M. Vogt, P. R. Bowser, and J. W. Casey. "Nucleotide sequence and protein analysis of a complex piscine retrovirus, walleye dermal sarcoma virus," *J. Virol.*, **69**, 5320–31, 1995.
38. R. Kambol, P. Kabat, and M. Tristem. "Complete nucleotide sequence of an endogenous retrovirus from the amphibian, xenopus laevis," *Virology*, **311**, 1–6, 2003.
39. L. A. LaPierre, D. L. Holzschu, P. R. Bowser, and J. W. Casey. "Sequence and transcriptional analyses of the fish retroviruses walleye epidermal hyperplasia virus types 1 and 2: Evidence for a gene duplication," *J. Virol.*, **73**, 9393–9403, 1999.
40. Z. Zhang, D. Du Tremblay, B. F. Lang, and D. Martineau. "Phylogenetic and epidemiologic analysis of the walleye dermal sarcoma virus," *Virology*, **225**, 406–12, 1996.
41. J. Venna and S. Kaski. "Visualized atlas of a gene expression databank," in *Proc. KRBIO'05, Symposium on Knowledge Representation in Bioinformatics*, 2005. Accepted for publishing.
42. R. C. Edgar. "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics*, **5**, 113, 2004.