Clustering of human endogenous retrovirus sequences with median self-organizing map

Merja Oja, Panu Somervuo, Samuel Kaski, and Teuvo Kohonen
Neural Networks Research Centre,
Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, FINLAND
Tel: +358-9-4518201, FAX: +358-9-7554892, merja.oja@hut.fi

Keywords: human endogenous retrovirus (HERV), median self-organizing map (SOM), retrovirus, taxonomy

***Abstract*—** Mutual relationships of human endogenous retroviruses (HERVs) and their similarities to other DNA elements are studied in this paper. We demonstrate that a completely data-driven grouping is able to reflect same kinds of relationships as more traditional biological classifications and phylogenetic taxonomies. The clusters and their visualization were computed with the Median Self-Organizing Map algorithm of pairwise FASTA-based distances. The whole-sequence distances are able to distinguish between the different known types of endogenous elements, and exogenous retroviruses. The HERVs become grouped meaningfully.

1 Introduction

Only about two percent of human DNA codes for proteins. The function of the rest is unknown, and it has been called "junk DNA." It is, however, far from random, and numerous studies (for a review see [10]) have already shown that it may serve for meaningful functions.

About 45 per cent of the DNA [8] is derived from *transposons*, parts of genome capable of moving or copying themselves in the genome. About eight per cent consists of specific kinds of transposons, called *human endogenous retroviruses (HERV)*. Human retroviruses such as HIV in general are viruses capable of copying their genetic code to the DNA of humans, and they become endogenous once they have been copied to the germ-line. Human endogenous retroviruses, in contrast to some other human transposons, are not capable of moving any longer 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].

The HERVs stem from several kinds of retroviruses. Functions of HERV sequences existing 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 hence help in studying functions of HERVs. The problem is that it is not known exactly which parts of HERV sequences are important. Furthermore, during the time the sequences have inhabited the human genome they have become mutated and broken in crossovers and when other transposons have moved to overlap them. Hence the sequences are noisy and incomplete.

A traditional way of classifying HERVs is to group them according to the similarity of the short region, the primer binding site, from which their transcription (activation) starts. 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 [9, 15]. Phylogenetic analysis is a form of hierarchical clustering that produces trees describing the descent, under the assumption that a sufficiently representative sample set is available.

Since the samples in practice are far from extensive, phylogenetic trees have little more justification than being a form of clustering. On the other hand, there exists evidence [11] for the Self-Organizing Map-based visualizations being in a sense more trustworthy than hierarchical clustering.

In this paper we investigate whether it is feasible to extract taxonomic relationships based on mutual similarities computed from the whole sequences of retroviruses. We will group the sequences and visualize their similarity relationships. If they correlate sufficiently well with the earlier findings that have focused on specific parts of the sequences, the result suggests that a completely data-driven analysis of retroviruses is feasible. Noisy and incomplete sequences not amenable to the focused analyses could be included in the more comprehensive analysis. Tolerance to noise and incompleteness will of course need to be studied further.

In this first feasibility study we group a set of known samples of retroviruses and related sequences to find out whether the known groups comply with similarities computed from the whole sequences. We will use the Median Self-Organizing Map [7] capable of organizing the sequences based on a priori computed pairwise mutual distances. Here the distances are computed by the FASTA [12] method.

2 Methods

2.1 Principle of the Median SOM

The Self-Organizing Map (SOM) can be used to order nonvectorial data such as DNA sequences by a variation of the method in which each model $\mathbf{m}_i$ on the map becomes the *generalized median* of the input items mapped into the neighborhood of $\mathbf{m}_i$ [5, 7]. For this method it will be sufficient that some similarity measure is definable between each input item $\mathbf{x}$ and each model $\mathbf{m}_i$, as well as between all pairs of the input items $\mathbf{x}$. This variation of the SOM resembles the Batch Map method [5, 6].

In the work in presentation, the above variation has been applied to the production of similarity diagrams, and showing the clustering tendency of DNA sequences. The similarities between the DNA sequences were computed by the FASTA method [12].

The generalized median, which is defined as the hypothetical data item from which the sum of distances to the other elements in a data set is minimized, can in practice often be approximated by the set median. The set median is an exact copy of one of the data items in the data set, namely, that one from which the sum of distances to the other elements of the data set is minimized. Usually the set median is a good approximation of the generalized median, but because it is quantized to the values of the set elements, neighboring models on the map often become identical. These duplicates then give rise to ties in the determination of the best-matching models, and special measures are necessary to break the ties (cf. [7]).

The computation of the SOM using set medians as models is carried out as the iteration of the following two steps. At the first step, copies of the input (teaching) sequences are listed under their best-matching models, taking into account the tie breaks in matching. At the second step, for each node in the map, a new value for each model is determined as the set median of those input sequences that lie in the neighborhood of the said node, i.e., in the union of the data lists existing in the neighborhood of that node. These two steps, namely, listing of copies of input sequences under the best-matching models, and computation of the new models as the set medians of sequences mapped into the neighborhood of each node, are repeated, until the models can be regarded as stationary.

The tendency of the map to create duplicate models at neighboring nodes depends on how densely populated the data space is. This can be observed from the smoothness of the distance matrix.

2.2 Data

The data set used in this work contains three types of sequences. We are mainly interested in the human endogenous retroviruses (HERVs), but long interspersed repeats (LINE) and exogenous retroviruses have been included for reference. Both HERVs and LINEs belong to the so-called transposable elements. They reside in the human genome and unless they are defective they are capable of transposing and copying them selfs to multiple locations in the genome. The different copies of the same sequence have diverged from each other during the tens of millions of years they have been in our genome. The HERVs and LINEs have been grouped into families based on their origin. The HERV and LINE sequences were derived from the RepBase database [3, 4]. Libraries retrovir.lib (90 sequences) and humlines.lib (103 sequences) contain consensus sequences for the known HERV and LINE families, respectively.

The exogenous retrovirus sequences were derived from the NCBI Taxonomy database [16] by searching out all *Retroviridae* genomes and then fetching the sequences from the GenBank®. The search performed in April 2003 resulted in 50 complete genome sequences.

The data set contains 243 sequences in total. The lengths of the sequences vary from about 500 to 10,000 base pairs (bp). The LINE elements are shortest with mean length of 1550 bp. The endogenous retroviruses are on the average 5850 bp long and the exogenous retroviruses 8430 bp long. The HERV sequences are shorter than exogenous retroviruses because their database entries contain only the internal sequence of the endogenous retrovirus. The exogenous retroviruses are represented with long terminal repeat sequences (LTR) at each end.

The HERVs have traditionally been classified on two different grounds. We will use these classifications in verifying the feasibility of our data-driven grouping. The first classification stems from the tRNA used to prime DNA synthesis [15]. The classes are named after the primer binding site (PBS); e.g. the viruses that are primed by leucine (L) tRNA are called HERV-L and those utilizing arginine (R), HERV-R. The PBS based classification is, however, incomplete in such cases where HERVs of different origin are primed by the same tRNA. There exist some evidence [15] that the traditional classification may be misleading; we will take this into account in the interpretation of the results.

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, 15]). Class I HERVs are related to gammaretroviruses such as Feline leukemia virus or Gibbon ape leukemia virus and include HERV-W and HERV-H, 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 HERV-K elements. Class III HERVs are distantly related to spumaviruses (Human foamy virus) and include HERV-L and HERV-S. Class I also includes the MER4 group in RepBase nomenclature.

2.3 Computation of SOMs

The SOM was computed in two stages. In the first organization stage, the sequences were encoded into vectorial representations. The computation was then continued using FASTA-based [12] sequence similarities. This two-stage training scheme has been found useful in earlier studies [7, 14]. The first stage ensures smooth spread of the SOM models to cover the feature space, since the vectorial representations facilitate smooth interpolation and averaging. When the SOM has attained a rough ordering after the first stage, the neighborhood function need not be wide any longer in the second stage, and the set median computation attains the final result much faster.

In the first stage, we used n-gram histogram representations of the data sequences. Each sequence was encoded into a histogram of 4-grams of the symbol alphabet consisting of the four symbols A, C, G, and T. Hence the feature vector was 256-dimensional. Besides A, C, G, and T there were also symbols B, D, H, K, L, M, N, R, S, V, W, X, and Y in the HERV and LINE sequences. However, their total number was less than one per cent from all nucleotide elements so they were ignored. The largest proportion of the non-ACGT symbols was in sequence PRIMAX-int, which contained five per cent of these. The feature vectors were normalized to unit length. For the 243-sequence data set, we decided to use a 9-by-10 unit hexagonal SOM. The 256-dimensional model vectors were initialized by random values between 0 and 1. The SOM was computed using the Batch Map algorithm for vectors [5]. The width of the Gaussian neighborhood function decreased linearly from 10 to 1 during the 20 iterations of the algorithm.

The model vectors were then replaced in the second stage by the indices of the local set medians of the data. The set median in each node was determined from the union of the data lists covering the neighboring map nodes.

Ten iterations of the Median SOM algorithm were then carried out. Details of the algorithm can be found in [7]. 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 [12]. The FASTA scores were computed with default parameters.

Since the lengths of the sequences varied greatly, we normalized the effect of sequence length in the FASTA scores by using the Tanimoto distance [13]. First, the FASTA scores were computed for each sequence pair. These scores were converted to Tanimoto similarities,

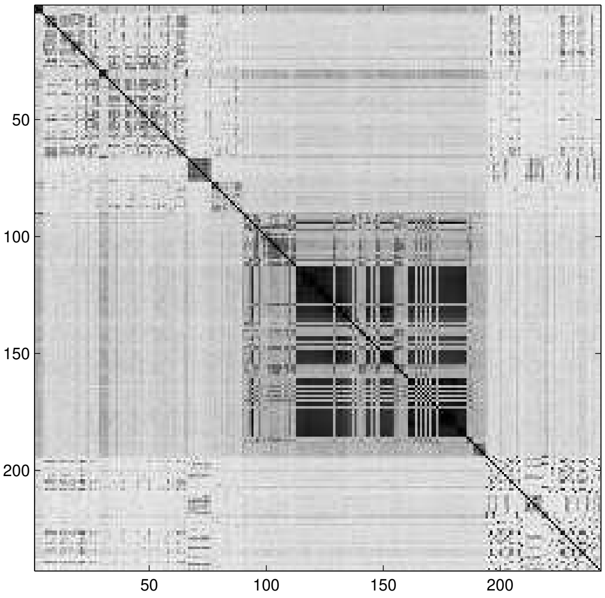

Figure 1: Pairwise Tanimoto distances between the 243 data sequences based on the FASTA score. The first 90 rows correspond to HERV sequences, the next 103 rows to LINEs and the last 50 rows to exogenous retroviruses. Black: zero distance; increasing lightness: Larger distance.

$$s(i,j) = \frac{f(i,j)}{f(i,i) + f(j,j) - f(i,j)}, \qquad (1)$$

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 finally converted to the Tanimoto distance

$$d(i,j) = -\log s(i,j). \qquad (2)$$

The distance matrix containing all pairwise sequence distances (2) is shown in Figure 1.

The 9-by-10-unit Median SOM of virus sequences is shown in Figure 2. The shade of gray represents the distance between the models of adjacent map nodes. Data sequences are listed at their best-matching units (BMUs).

There are some empty nodes in the map in Figure 2. This is because of the duplicates of the models, resulting from the discrete nature of the data. These do not cause any problems in the SOM training; in case of duplicates, BMUs can still be unambiguously determined by means of the models in the neighborhood of the BMU candidate as explained in [7].

Besides the map shown in Figure 2, we also computed several other maps with different random vector initializations. The sizes of the maps were also slightly

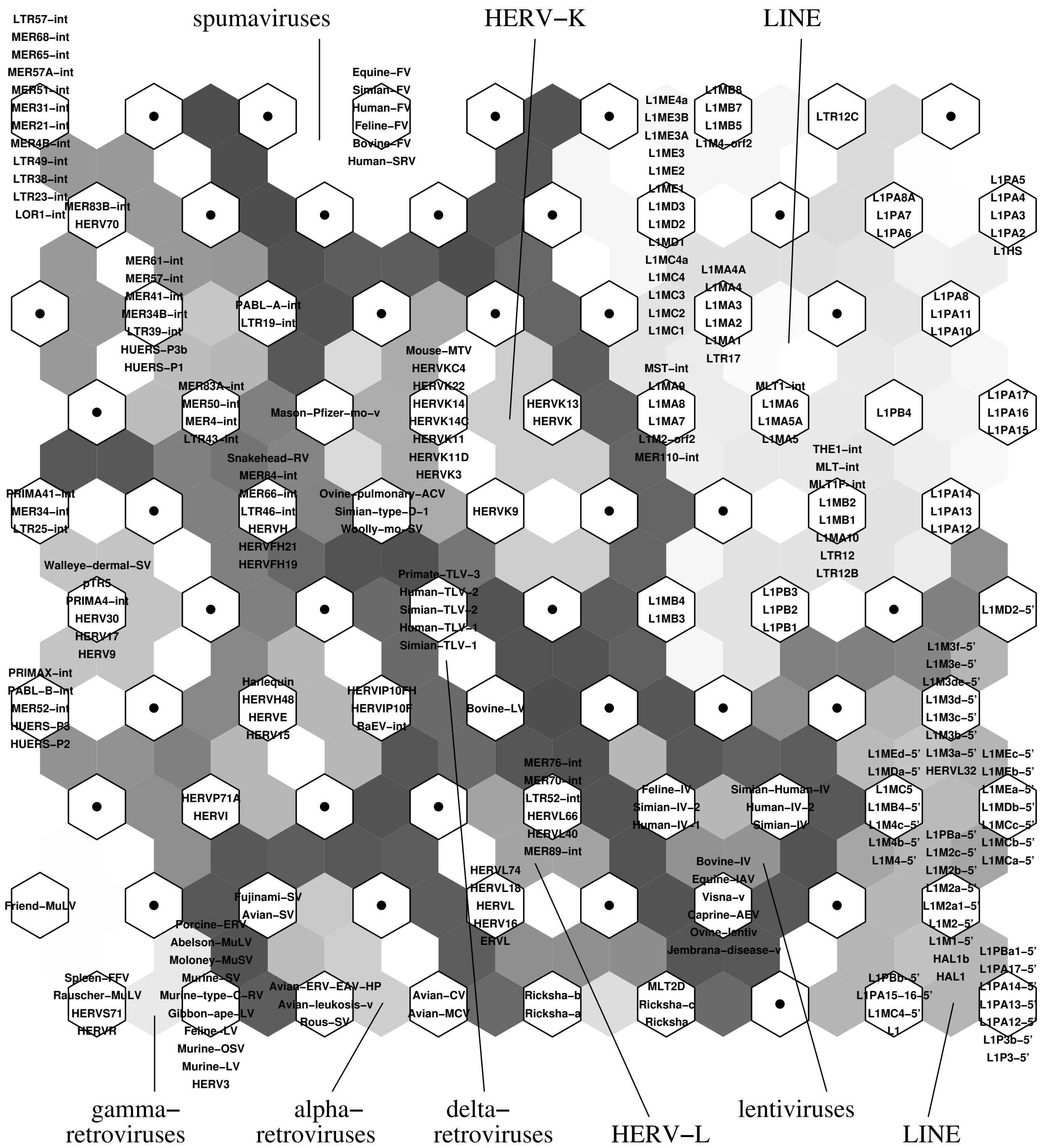

Figure 2: Median SOM of HERV, LINE, and exogenous retrovirus sequences. Every second (bordered, and dotted if not being a best match for any sequence) hexagon denotes a SOM unit, and the rest are U-matrix entries indicating distance between the units. The resulting light areas are clusters and black stripes borders between them. Symbols of the sequences have been inserted to the locations where the sequences have been mapped. Manually assigned names for the clusters are presented beside the map. SV=sarcoma virus, FV=foamy virus, LV=leukemia virus, TLV=T-lymhocytic virus, CV=carcinoma virus, MuLV =murine leukemia virus, MTV=mammary tumor virus, ASV=adenosarcoma virus, OSV=osteosarcoma virus, SRV=spumaretrovirus, AEV=arthitis-encephalitis virus, IAV=infectious anemia virus, MCV=myelocytomatosis virus, FFV=focus forming virus.

varied. In some cases different random initializations resulted in the exactly same Median SOMs. But even if final model sequences were not exactly same in different maps, similar data clusterings were generally still observed.

As for the conventional vector SOM, the quantization error of data can be computed also for the Median SOM. This requires only computing the distance of the data sequence to its best-matching unit and averaging this number over all data sequences. In our studies, for fixed map sizes, different initializations yielded very similar quantization errors. The map in Figure 2 gave the best quantization error among the 9-by-10-unit maps with five different random initializations.

The comparison of normal SOM to the Median SOM revealed that 4-grams are not adequate at presenting the information in the DNA sequences of the retroviruses. The normal SOM, used to initialize the Median SOM, organized the data samples differently than the Median SOM. The organization was not as meaningful when compared to the known classification of the retrovirus sequences. The LINE elements were separated from the retroviruses, but the different types of endogenous and exogenous retroviruses were mixed on the map.

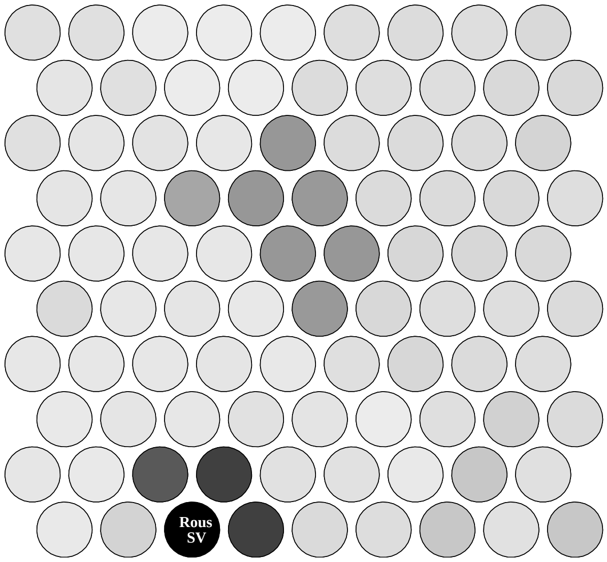

Figure 3: Tanimoto distance from the map node marked as Rous sarcoma virus to all the other map nodes (the circles). It can be seen that the alpharetrovirus cluster at the bottom of the map is not so far from the HERV-K cluster at the center of the map (compare to Fig. 2). The SOM has in effect folded to be able to visualize the many-sided similarities in 2D. Black: zero distance; increasing lightness: Larger distance.

3 Interpretation of the biological results

The different types of reference sequences have become grouped into different clusters (Fig. 2). The LINE-elements (L1 in the figure) form one large cluster with two subparts. Different types of exogenous retroviruses form compact and clear clusters: lentiviruses[1] (e.g. immunodeficiency viruses, IV), deltaretroviruses (T-lymphotropic viruses, TLV), alpharetroviruses (sarcoma viruses, SV), gammaretroviruses (leukemia viruses, LV), and spumaviruses (foamy viruses, FV).

The human endogenous retroviruses form clusters as well. The class II HERVs (HERV-K) are nicely clustered all together. In addition, the Mouse mammary tumor virus (Mouse MTV) and other betaretroviruses, are clustered with them, clearly in accordance with the traditional classification. The alpharetroviruses are further away on the map, but closer inspection revealed that the map node in which Rous sarcoma virus resides is actually rather close to the HERV-K cluster (see Fig. 3).

The human endogenous retroviruses HERV-3, HERV-R and HERV-S71 have been clustered together with gammaretroviruses which supports the classification of these HERVs to class I. The similarity of the other class I retroviruses (all sequences on the left side of the map which are not exogenous retroviruses) to gammaretroviruses is more vague.

Most of the class III HERVs (HERV-L) are clustered together on the map. The traditional view is that they bear similarity to spumaviruses, which however are at the other side of the map. The similarity between HERV-L and spumaviruses has been reported on the *pol* gene region, but in this study whole genome length samples have been used. The similarity of the other regions of HERV-L and spumaviruses needs further studying.

The class I HERVs form multiple neighboring clusters which span the whole left side of the map. In this clustering families primed by the same tRNA are not necessarily grouped together. This is in accordance with the current view that they are families with independent origin even though they share the PBS. See for example the sequences HERV-H and HERV-H48. The HERV-H48 seems to be more similar with HERV-E and HERV-15 (primed by isoleucine (I) tRNA) than HERV-H.

The class III HERVs (HERV-L) are more focused than class I HERVs, but a few sequences have diverged from the HERV-L cluster. The sequences LTR57-int and MER68-int (at the top left node) are HERV-L-type sequences but are clustered together with class I HERVs. In contrast, the sequence MER89-int is a class I HERV clustered together with HERV-Ls. These placings reflect the uncertainty of the classification of these

[1]See [16] and [17] for clarifications on the different types of viruses.

three sequences which is also stated at their entries in RepBase [3, 4].

The Rauscher murine leukemia virus is an unclassified retrovirus [16]. Its current position on the map suggests that it could belong to gammaretroviruses. This should of course be verified with thorough sequence alignments to protein and DNA sequences from other gammaretroviruses.

4 Concluding comments

We have explored mutual similarities of human endogenous retrovirus sequences by grouping them together with other endogenous DNA elements and related exogenous retroviruses. The grouping and visualization was done based on whole-sequence similarity with the new Median Self-Organizing Map algorithm, and resulted in findings consistent with earlier classifications.

A potential technical problem in using set medians as models arises if there are small data clusters which look like outliers for the rest of the data. In case there are no model sequences on the map for representing these small clusters, their best-matching units have to be chosen from among sequences which have no similarity between them. This results in determining their best-matching units based on noisy elements in the distance matrix, which may scatter the members of the small cluster to random places on the map. However, if the distance matrix is smooth and there are no small isolated data groups, this problem does not occur.

The method groups similar sequences together, and in a sense carries out unsupervised "class discovery." In this first work we grouped clean consensus sequences to verify that the FASTA-based distance computed from whole sequences gives meaningful groupings. The next question is whether the whole-sequence similarity is sufficiently tolerant to noise and incompleteness of sequences extracted automatically from the human genome. Methods for the detection are currently being intensively studied by e.g. Blomberg et al. [1].

Acknowledgements

This work was supported by the Microbes and Man Research Programme, The Academy of Finland (grant 200836). Study of this application area ensued from our contacts with Professor Jonas Blomberg of Uppsala University.

References

[1] Blomberg, J. and Sperber, G. (2003). The human genomes content of retroviral sequences. *The second workshop on retrotransposons and genome evolution*, Sochi, April 2003.

[2] Griffiths, D. J. (2001). Endogenous retroviruses in the human genome sequence. *Genome Biology*, 2:1017.1–1017.5.

[3] Jurka, J. (1998). Repeats in genomic DNA: mining and meaning. *Current Opinion in Structural Biology*, 8:333–7.

[4] Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends in genetics*, 16(9):418–420.

[5] Kohonen, T. (1995). *Self-Organizing Maps*. Springer, Third edition 2001.

[6] Kohonen, T. (1996). Self-Organizing Maps of Symbol Strings. *Technical Report A 42*, Helsinki University of Technology, Laboratory of Computer and Information Science.

[7] Kohonen T. and Somervuo P. (2002). How to make large self-organizing maps for nonvectorial data, *Neural Networks*, 15(8-9):945-952.

[8] Lander, E. et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

[9] Lindeskog, M. (1999). *Transcription, splicing and genetic structure within the human endogenous retroviral HERV.H family*. PhD thesis, Lund University, Lund, Sweden.

[10] Löwer, R. (1999). The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends in Microbiology*, 7(9):350–56.

[11] Nikkilä, J., Törönen, P., Kaski, S., Venna, J., Castrén, E., and Wong, G. (2002). Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks, Special issue on New Developments on Self-Organizing Maps*, 15:953–966.

[12] Pearson W. and Lipman D. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444-2448.

[13] Rogers D. and Tanimoto T. (1960) A Computer Program for Classifying Plants. *Science*, 132(3434):1115-1118.

[14] Somervuo P. and Kohonen T. (2000) Clustering and Visualization of Large Protein Sequence Databases by Means of an Extension of the Self-Organizing Map. *Proceedings of the Discovery Science (DS'2000), Lecture Notes in Artificial Intelligence* 1967:76-85. Springer.

[15] Tristem, M. (2000). Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *Journal of Virology*, 74(8):3715–30.

[16] Search for *Retroviridae* at http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html

[17] International Committee of Virus Taxonomy (ICTV), Index to Virus Classification and Nomenclature Taxonomic lists and Catalogue of viruses. http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/