CLASS DISCOVERY AND VISUALIZATION FOR HUMAN ENDOGENOUS RETROVIRUSES BY BOOTSTRAPPING MEDIAN SELF-ORGANIZING MAPS

Merja Oja[1], Jonas Blomberg[2] and Samuel Kaski[3,1]

[1] Neural Networks Research Centre, Helsinki University of Technology, Finland.
[2] Section of Virology, Department of Medical Sciences, University of Uppsala, Sweden.
[3] Department of Computer Science, University of Helsinki, Finland.

Introduction

- We have characterized a large collection of human endogenous retrovirus sequences by a Median Self-organizing Map.
- Human endogenous retroviruses (HERV) are sequences in non-coding human DNA, but they may have functions that affect human genes.
- Similar HERV sequences have probably a similar function – class discovery summarizes similarities.
- We want to refine and complement the current (inadequate) classification system for HERV sequences.
- We use a Median Self-Organizing Map (SOM) to group and visualize all the HERVs found in the human genome.
- We estimate the confidence of the SOM visualization with bootstrap.

Data

- We have a collection of 3661 human endogenous retrovirus sequences automatically picked from the human genome by the RetroTector© system.

Methods

- The Median Self-organizing Map (SOM) is well suited for analyzing thousands of samples simultaneously.
- The SOM visualizes the similarities between individual samples as well as between groups of samples.
- Confidence of the visualization is estimated by counting how often similar neighborhoods appear in bootstrap repetitions of the SOM. We assume that if two sequences are often neighbors on the bootstrap SOMs, they are truly similar.

Methods

The Median Self-organizing Map

The Median Self-Organizing Map (SOM) [2,3,4] is a SOM for non-metric data. It requires only that pairwise distances are definable between the input data samples.

Here we have used FASTA similarity scores between HERV protein sequences. The effect of the sequence length is normalized.

Bootstrapping the Self-organizing Map

We sampled the data set 100 times with replacement to produce 100 bootstrap data sets of the size of the original data set [5].

We computer a median Self-organizing map from each bootstrap data set and then the original data set was projected onto each bootstrap map.

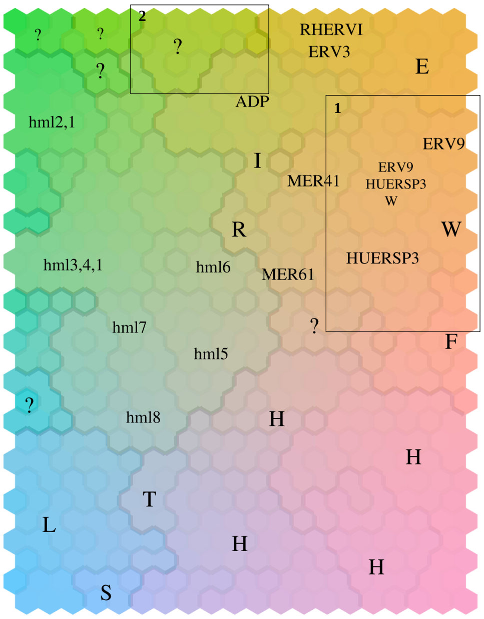
We counted the frequency $f_{i,j}$ of samples i and j appearing as neighbors on the bootstrap maps.

For each map node k we computed a confidence value:

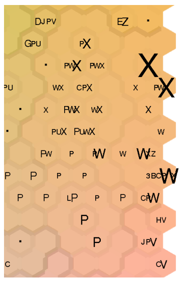
$$s_k = \frac{1}{N_k(N_k-1)} \sum_{i \neq j, i,j \in \text{node}\, k} f_{i,j},$$

where N_k is the number of sequences in node k.

Figure 1. The SOM of human endogenous retrovirus sequences – visualization of the cluster structure

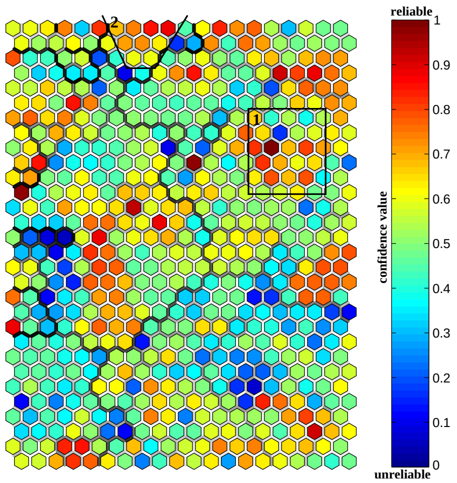


Top: The dark borders separate dissimilar areas from each other. The labels, referring to current HERV classes, are manually assigned descriptions for the map areas. The change in the shade of color between map nodes describes the distance between the nodes.



Right: A close up of area 1. The letters denote sequences with that label, the size of the letter is proportional to the amount of sequences. Note the mixing of the labels X, W and P. X=ERV9, P=HUERSP3.

Figure 2. Color coded confidence of the map nodes – red areas are reliable



Red areas of the map are reliable. The confidence value for a map node tells how often the node contents are next to each other in bootstrap repetitions of the SOM.

http://www.cis.hut.fi/projects/bioinf/

Results

- The main cluster structure of the map is in accordance with the current knowledge about human endogenous retroviruses.
- We found areas on the map that are both reliable (based on bootstrap) and unexplainable by the current classification:

  On area 1, current class labels are mixed together suggesting regrouping of ERV9, HERV-W and HUERSP3 sequences. We suggest that a new class is assigned for this mixed group or that the three mixed classes are combined to one class.

  Area 2 contains 49 sequences, from which only 6 are classified. This group can be a thus far undiscovered HERV class.

Conclusions

- The Median Self-organizing Map is able to extract new knowledge from the human endogenous retrovirus sequence collection.
- The bootstrap method can be used to estimate and visualize the confidence of areas on a Self-organizing Map.

What human endogenous retroviruses are?

- one type of transposable elements
- mutated and unable to transpose or to produce protein
- remains of ancient infections by retroviruses
- HERVs might have viral functions
- they might regulate human gene activity
- About 8% of the human genome is HERV sequence.

HERV	=	Human endogenous retrovirus
SOM	=	Self-organizing map

References

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

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

[3] Kohonen, T. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997, Third Extended edition 2001).

[4] Oja, M., Somervuo, P., Kaski, S., and Kohonen, T. (2003). 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.*

[5] Efron, B. and Tibshirani, R. J.(1993). *An introduction to the Bootstrap.* New York and London: Chapman and Hall.

[6] Kaski, S., Venna, J., and Kohonen, T. (2000) Coloring that reveals cluster structures in multivariate data. In *Australian Journal of Intelligent Information Processing Systems*, 6:82-88.