Bioinfromatics'02
April 2002, Bergen, Norway

Learning metrics for SOM-based clustering and visualization of yeast gene expression data

M. Oja[1], P. Törönen[2], J. Nikkilä[1], E. Castrén[2] and S. Kaski[1]

[1] Neural Networks Research Centre, Helsinki University of Technology, Espoo, Finland.
[2] A.I.Virtanen Institute, University of Kuopio, Kuopio, Finland

Introduction

- We have studied the application of a new *clustering and information visualization* methodology to functional genomics.
- The *learning metrics principle* is a new approach to finding important aspects of data. It is assumed that changes of gene expression are important only to the extent that they cause changes in certain auxiliary data, in this study the functional classes of the genes.
- We cluster the yeast genes based on their expression in a set of knock-out mutation experiments. (PD data set [1].)
- The SOM computed in the new metric is more accurate in modeling the functional classes of the genes. The visualizations are as intuitive as in the usual inner product (correlation) metric.

Methods

- The self-organizing map is a method well suited for analyzing high-dimensional gene expression data. (For more information see [4].)
- In a learning metric the SOM can concentrate on the essential properties of gene expression data and the effect of the arbitrarily chosen metric (Euclidean, inner product or correlation) diminishes. The learning metric is able to scale down or ignore uninteresting dimensions locally at each data sample.

Experiments

- The data analysed in this study is from [1]: measurements for all yeast (*Saccharomyces cerevisiae*) genes in 300 knock-out mutations.
- Each expression profile was normalized to unit length and the SOMs (learning metric and inner-product metric) were restricted to the same hypersphere as the data.

Methods

Self-organizing maps (SOMs)

SOM [3] is an neural network algorithm that maps high-dimensional data nonlinearly onto a low-dimensional lattice in a topology preserving manner.

SOM has features from both nonlinear projection methods and clustering methods.

Computation of a SOM is an iterative process where for each sample $\mathbf{x}(t)$ a winner unit $\mathbf{m}_{w(t)}$ is searched:

$$w(\mathbf{x}(t)) = \arg\min_i d^2(\mathbf{x}(t), \mathbf{m}_i(t)).$$

Then all model vectors are updated towards the data sample.

Learning metrics

If auxiliary information relevant to the goal of the data analysis is available it can be used to guide the data analysis method.

The dimensions of the data space will be scaled locally reflecting the changes in the auxiliary information. The Fisher matrix $\mathbf{J}(\mathbf{x})$ will represent the effect of the learning metric in each data point. The $\mathbf{J}(\mathbf{x})$ is computed from a PDF estimator.

The distances in learning metric can be calculated by

$$d_L^2(\mathbf{x}, \mathbf{x}+d\mathbf{x}) \equiv D(p(c|\mathbf{x})||p(c|\mathbf{x}+d\mathbf{x})) = d\mathbf{x}^T\mathbf{J}(\mathbf{x})d\mathbf{x},$$

SOM in learning metrics

The winner search is done using d_L as distance measure.

The update rule for learning metrics turns out to be the same as in the Euclidean metric:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{wi}(t)(\mathbf{x}(t)-\mathbf{m}_i(t)).$$

Figure 1. Goodness of the SOMs

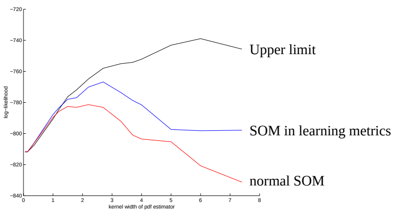
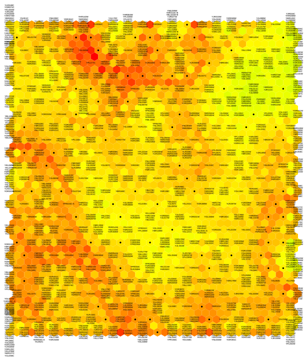
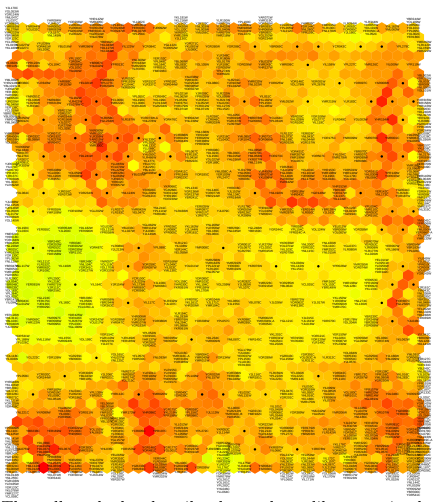


Figure 2. A SOM in inner-product metric



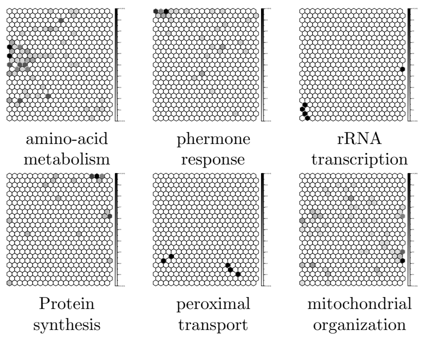
The yellow shades describe dense clusterlike areas in the data and red shades less dense areas. The names of genes mapped to each unit are shown on the map.

Figure 3. A SOM in learning metrics (SOM-LM)



The yellow shades describe dense clusterlike areas in the data and red shades less dense areas. The names of genes mapped to each unit are shown on the map.

Figure 4. Distributions of selected functional classes on the SOM in learning metrics



For more information see:
http://www.cis.hut.fi/projects/bioinf/

Results

- The SOM in learning metrics (Fig. 3) models the conditional class distribution of the functional classes better than the plain SOM (Fig. 2). For numerical results see Figure 1.
- The distributions of the functional classes were similar on both maps. Hence, even if the learning metric was able to model the functional classes better the difference isn't big. This suggests that the correlation between functional classes and this gene-expression data isn't high, i.e. the classes are very scattered in the expression space. Examples of class distributions on SOM in learning metrics are shown in Figure 4.
- Comparison of the SOMs in different metrics revealed an interesting small cluster of genes related to mitochondria on the SOM-LM. Three of the genes were associated with branched-chain amino acid biosynthesis and were grouped with two genes related to fermentation and carbohydrate utilization (YOL059W, YER073W) and genes involved in the threonine and lysine an leusine biosynthesis (YDR234W, YER086W,YGL009C). An unclassified gene YKL120W was clustered with these genes, its function is unknown but it has been suggested from the sequence similarity that it is related to mitochondrial proteines. Results with SOM in learning metrics support this assumption.
- Projection of pathway data on the map revealed that the genes involved in purine biosynthesis were localized on SOM-LM. The genes and other genes close by were classified to nucleotide metabolism. In addition gene YDL241W, whose function is unknown, was situated nearby on the map. This gene might also be related to nucleotide metabolism or purine biosynthesis pathway and should be examined closer.

Conclusions

- The learning metrics principle together with SOMs provides new, interesting insights to gene expression data with auxiliary data.
- In future work, the approach will be applied to other data sets and/or other kinds of auxiliary data.

References

[1] T. R. Hughes et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[2] S. Kaski, J. Sinkkonen, and J. Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.

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

[4] M. Oja, J. Nikkilä, P. Törönen, G. Wong, E. Castrén, and S. Kaski. Exploratory clustering of gene expression profiles of mutated yeast strains. In W. Zhang and I. Shmulevich, editors, *Computational And Statistical Approaches To Genomics*. Kluwer Academic Publishers, 2002. In press.

[5] J. Sinkkonen and S. Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–39, 2002.