Proceedings of STeP 2002, Oulu, Finland
STeP 2002 - Intelligence, The Art of Natural and Artificial

LEARNING METRICS FOR VISUALIZING GENE FUNCTIONAL SIMILARITIES

*Merja Oja**, *Janne Nikkilä**, *Petri Törönen*+, *Eero Castrén*+ *and Samuel Kaski**,

*Neural Networks Research Centre, Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland

+A.I. Virtanen-institute, University of Kuopio
P.O. Box 1627, FIN-70211 Kuopio, Finland

The usual first step in analyzing the large and high-dimensional data sets measured by microarrays is visual exploration. In this work self-organizing maps have been used to visualize similarity relationships of data samples. In all unsupervised data analysis methods the measure of similarity determines the result; we propose to use the learning metrics principle to derive a metric from interrelationships between data sets.

A metric is derived for a gene knock-out expression data set by considering those changes in the expression space that cause changes in the functional classes of the genes to be more important. The genes for which the new metric is the most different from the usual correlation metric are listed and visualized with a self-organizing map computed in the new metric.

1 INTRODUCTION

There are usually two successive phases in statistical data analysis: exploratory and confirmatory. In the first phase the data is inspected and explored to form hypotheses that are then verified in the second, confirmatory phase.

For data sets measured with microarrays the exploratory phase is particularly important for two reasons. First, if the number of plausible research hypotheses is very large it is reasonable to start with thorough exploration. Second, when collecting data to answer a particular research problem a large gene expression database is created as a side product, and it can be explored later for other purposes.

In this work we use self-organizing maps (SOM) to explore and visualize a high dimensional data set. The main goal is to visualize the pairwise similarities, but additionally clusters (groups of mutually similar data) may be revealed, and moreover the similarity relationships between clusters. Note, however, that it is a hard task to visualize very high-dimensional similarities on a low-dimensional display. In particular, it is impossible to project the samples to a low-dimensional display such that all similarity relationships would be preserved. The SOM will represent similarities so that samples close by on the map are also close in the original data space. The inverse, however, is not always true.

The SOM will be applied to functional genomics data measured by DNA microarrays in gene knock-out mutation experiments [2]. Visualized similarity of the genes is an indication of functional similarity revealed by the set of mutations.

A major question with visualization of similarity, and as a side note in clustering as well, is how to measure similarity. Gene expression measurements in a variety of treatments potentially include valuable information about the function and co-regulation of genes. The important variation is, however, hidden within all the biological and measurement noise in the very high-dimensional expression space.

For the knock-out mutation data the question is about which mutations to choose, and how to weight the mutations so that the functionally meaningful variation is emphasized and irrelevant variation is suppressed. Moreover, the weighting should be different for different genes, i.e., at different locations of the expression space.

The *learning metrics principle* [4, 5, 11] is a new approach to finding important aspects of data based on the interrelationships between two data sets. It is assumed that changes of gene expression are important only to the extent they cause changes in certain auxiliary data. The metric of the expression data space is changed to measure locally only the important changes. Note that it is crucial to choose proper auxiliary data since it determines the metric. Here, as we try to find genes that have similar function based of the expression data, we will incorporate prior biological knowledge into the metric by using the functional classes of the genes as the auxiliary data.

We will extract a set of genes for which the new metric is most different from the usual inner product (correlation) similarity measure; their earlier visualizations and clusterings have been the most badly misleading. Finally, the similarities of all genes in the new metric will be visualized with a self-organizing map computed in the new metric.

2 METHODS

2.1 Data

The data set was provided by Hughes *et al.* [2]. It consists of measurements of the expressions of all yeast (*Saccharomyces cerevisiae*) genes in 300 knock-out mutations. Hughes *et al.* had additionally derived error estimates based on replicated measurements, and tested whether the expression of the genes differed significantly from noise.

A subset of the data containing saliently expressed genes and mutations that induced expressions was selected. Only genes and mutations with at least two measurements that differed significantly from noise ($P < 0.01$) and were expressed over 2-fold when compared to the control were selected, resulting in a data set of the size of 1410 genes measured for 179 mutations.

We have earlier [9] compared different preprocessing methods (normalization of measurement error and/or standard deviation) by using the classification error of a k-nearest neighbor classifier as the performance measure. At the same time the effect of normalization of the expression vector length and/or mean was analyzed. For this data the following alternative showed the best performance: the data was preprocessed by dividing each measurement by its measurement error and then normalizing the standard deviation of each mutation. Finally all gene expression profiles were normalized to unit length.

The auxiliary data was chosen from the MIPS functional classification [8] for yeast. The classification consists of over two hundred classes at different levels of hierarchy. Many of the functional classes are known to correlate with gene expressions, although some classes undoubtedly are very heterogeneous at the level of gene expression. A set of 46 non-hierarchical classes, which cover all classifications and are as homogeneous as possible, were chosen from the functional classification. The complete list of chosen classes can be found in [10].

2.2 The self-organizing map (SOM)

The self-organizing map (SOM) [6] is an algorithm that maps high-dimensional data nonlinearly onto a low-dimensional lattice in a topology-preserving manner. The SOM can be used both as a nonlinear projection method and a clustering method. The SOM visualizes the density structures of the data and it is also possible to extract clusters from the computed SOM (see e.g. [13]).

The SOM is a discrete lattice of map nodes (marked by the dots and labels in Fig. 2). There is a *model vector* $\mathbf{m}_i$ attached to each map unit i. A data sample $\mathbf{x}$ is projected onto the SOM display at the location of the node having the closest model vector $\mathbf{m}_w$, defined by

$$w(\mathbf{x}) = \arg\min_i d^2(\mathbf{x}, \mathbf{m}_i) \ , \tag{1}$$

where d is the distance measure, here the inner product (correlation).

The SOM algorithm computes such values for the model vectors that (i) the projection becomes ordered: close-by samples on the SOM display are close-by in the data space as well, and (ii) the projection models the data distribution: each model vector becomes the centroid of all data samples mapped to it and to its neighborhood on the map display.

There are several variants of the SOM algorithm; here we describe the original sequential one which we will complement with new metrics.

During the iterative computation of the SOM, at step t a data sample $\mathbf{x}(t)$ is chosen randomly, and the model vectors are updated towards the data sample according to

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{w(\mathbf{x}(t)),i} \frac{\partial}{\partial \mathbf{m}_i} d^2(\mathbf{x}(t), \mathbf{m}_i(t)) \ . \tag{2}$$

If inner product is used as the distance measure, the model vectors should be normalized after the adaptation step. Here $h_{w(\mathbf{x}),i}$ is the *neighborhood function*, a decreasing function of the distance of the units w and i on the SOM lattice.

We computed SOMs of the size of 20x25 map units (about 3 genes in a unit on the average) in two phases, of lengths 50,000 and 750,000 iterations. In the first, organizing phase the standard deviation of the Gaussian neighborhood function decreased from 11 to 3, and its height (at the peak) decreased from 0.2 to 0.02. In the second, fine-tuning phase the standard deviation decreased from 3 to 1 and the height from 0.02 to zero. The best map of 4 randomly initialized SOMs was chosen according to the (local) cost function (cf. [6]).

2.3 Learning metrics

2.3.1 The learning metrics principle

We acknowledge that there is not enough prior knowledge to choose a proper metric for measuring similarity of the knock-out mutation profiles of genes. The relevant mutations would need to be chosen and scaled suitably for each gene. Our aim is to *learn* the proper metric from interrelationships between the expression data set and another data set which is known to be relevant to gene function: the functional classification of the genes.

The learning metrics [4, 5, 11] are constructed based on the assumption that changes in the primary data space are only important to the extent they cause changes in another (auxiliary) data space. Here the primary data are the knock-out expression profiles of the genes, and the auxiliary data are the functional classes. The new metric is defined by a *local* scaling of the original data space. Hence, it does not tear the space but emphasizes those directions that cause changes in the functional classes. In other words, the metric is supervised but the topology (loosely speaking) of the data space remains intact. With learning metrics it is possible to discover similarity and clustering relationships not included in the present functional classification.

Formally, denote the primary data by $\mathbf{x}$ and the functional class by c. During learning the data occurs in pairs $(\mathbf{x}, c)$. The squared distance measure of the data space is changed locally to measure the important differences, i.e. the differences between the distributions of the functional classes $p(c|\mathbf{x})$. When the differences are measured by the Kullback-Leibler divergence D_{KL}, the distances become locally

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) \equiv D_{KL}(p(c|\mathbf{x})||p(c|\mathbf{x} + d\mathbf{x})) = d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x}\,, \tag{3}$$

where $\mathbf{J}(\mathbf{x})$ is the Fisher information matrix with parameters $\mathbf{x}$. Since the conditional distribution $p(c|\mathbf{x})$ is not known we have estimated it with a standard estimator for the joint distribution, Mixture Discriminant Analysis (MDA2; [1]), and derived the conditional distribution by the Bayes rule. For more details see [4]. The metric can in principle be extended to non-local distances by computing path integrals, but for computational reasons we have resorted to local approximations. The equation (3), used also for non-local distances, has worked satisfactorily for nearest-neighbor searches in empirical tests, particularly when complemented with a kind of regularization; in practice the metric will often be singular for very high-dimensional spaces, and hence we will add a portion of Euclidean distance,

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) \equiv d\mathbf{x}^T \left[\lambda \mathbf{I} + (1 - \lambda)\mathbf{J}(\mathbf{x})\right] d\mathbf{x}\ , \tag{4}$$

where $\mathbf{I}$ is the identity matrix. The coefficient λ is chosen using a validation set. The regularization makes the local approximations more feasible for non-local distances as well.

A difference from the earlier similar works on learning metrics here is that the data lies on the surface of a hypersphere. For such data it is better to define the density estimators on the hypersphere as well. Technically, instead of using Gaussian kernels in the MDA2 we used the so-called von Mises-Fisher kernels that are analogs of Gaussians on the hypersphere [7]. (We used 30 kernels fitted to the data with the EM algorithm.) The local distances (4) are still Euclidean on the hypersphere, but in practice non-local distances need to be computed as well. When computing distances from $\mathbf{x}$, we have projected all vectors to the tangent plane of the hypersphere at $\mathbf{x}$, with the distance from $\mathbf{x}$ scaled to be equal to the arc length from $\mathbf{x}$, and used (4) to compute the distances on the tangent plane.

2.3.2 Visualizing the similarity relationships in the new metric

We used the SOM to visualize similarity relationships of the genes in the learning metrics.

In the first step of a SOM iteration, the best matching unit is sought in the new metric d_L (eqn 4; cf. also the discussion after the equation). The steepest descent update rule for learning metrics turns out to be the same as in the Euclidean metric [4]. Here the update is applied in the tangent plane, and the results are transformed back to the hypersphere. It can be shown that the resulting update rule moves $\mathbf{m}_i$ towards $\mathbf{x}$ along the shortest route on the hypersphere, such that their angle reduces by a fraction of $h_{wi}(t)$.

2.4 Detecting genes for which the inner product metric is suspect

It is to be expected that the new similarity relationships between all gene pairs do not differ notably from the old. The reason is the necessary regularization of the metric. Hence we propose to search for genes for which the set of closest neighbors has changed the most, and

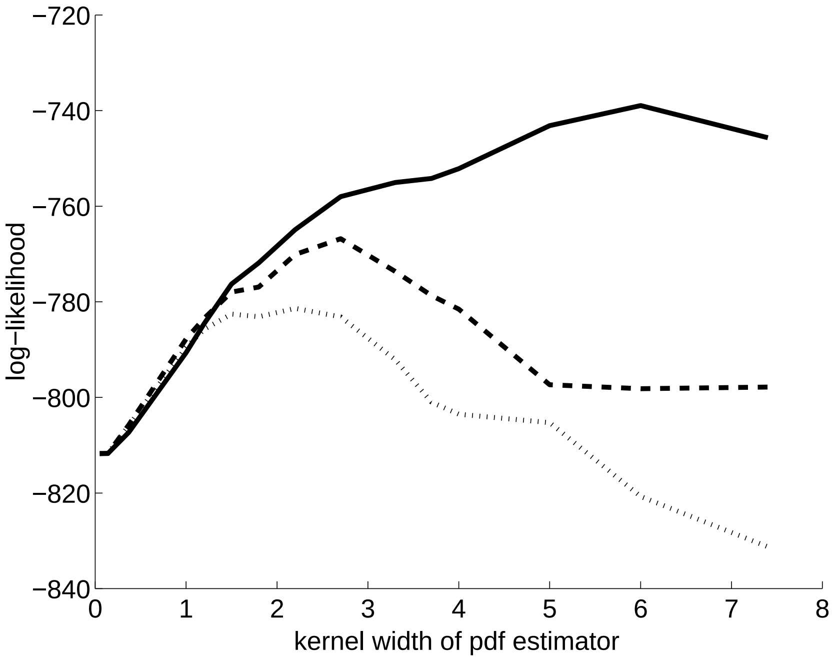

Figure 1: *Accuracy of SOMs in representing the functional classes. The goodness measure is the log-likelihood of the conditional density estimator at the closest SOM unit for each data point. The horizontal axis is the "smoothness" of the density estimator. Dashed line: SOM in learning metrics, dotted line: SOM in inner product metrics, solid line: the approximate upper limit, i.e. the estimate computed at the data point instead of at the closest model vector.*

focus the analysis on them. Assuming the auxiliary data is of good quality, these genes are the ones for which the old analyses using the inner product metric have given the most misleading results.

We sought 20 nearest neighbors (approx. equals the average of genes in a map unit and its direct neighbors) of each gene both in the old inner product and the learning metrics. The two sets were compared and the proportion of neighbors that had remained the same was computed. The 202 genes with the smallest proportion (at most 13 neighbors remained the same) were selected and visualized on the SOM (underlined in Figure 2).

3 RESULTS

First we measured quantitatively how well SOMs in learning metrics represented the functional classes, compared to SOMs in the inner product metric. The results shown in Figure 1 confirm that the SOMs in learning metrics are more accurate. The kernel width that produced the best results was chosen for the further experiments. The amount of regulation (λ in Eqn. 4) was validated at the same time as the kernel width. The selected parameters were 0.5 for regulation and 2.7 for kernel width. In Figure 1 only the curve for the selected λ is visualized (dashed line).

The similarity relationships of genes in the learning metrics are visualized in Figure 2. Genes close-by on the display are close-by in the original space as well. The gray shades in the figure indicate clusteredness [12]: in the light areas the genes are mutually more similar than on the dark areas; hence very light areas are clear clusters and dark stripes are gaps in between them. Technically, a hexagon is added in between each pair of SOM units and shaded according to the distance between their model vectors. The shade of the hexagons of the actual SOM units

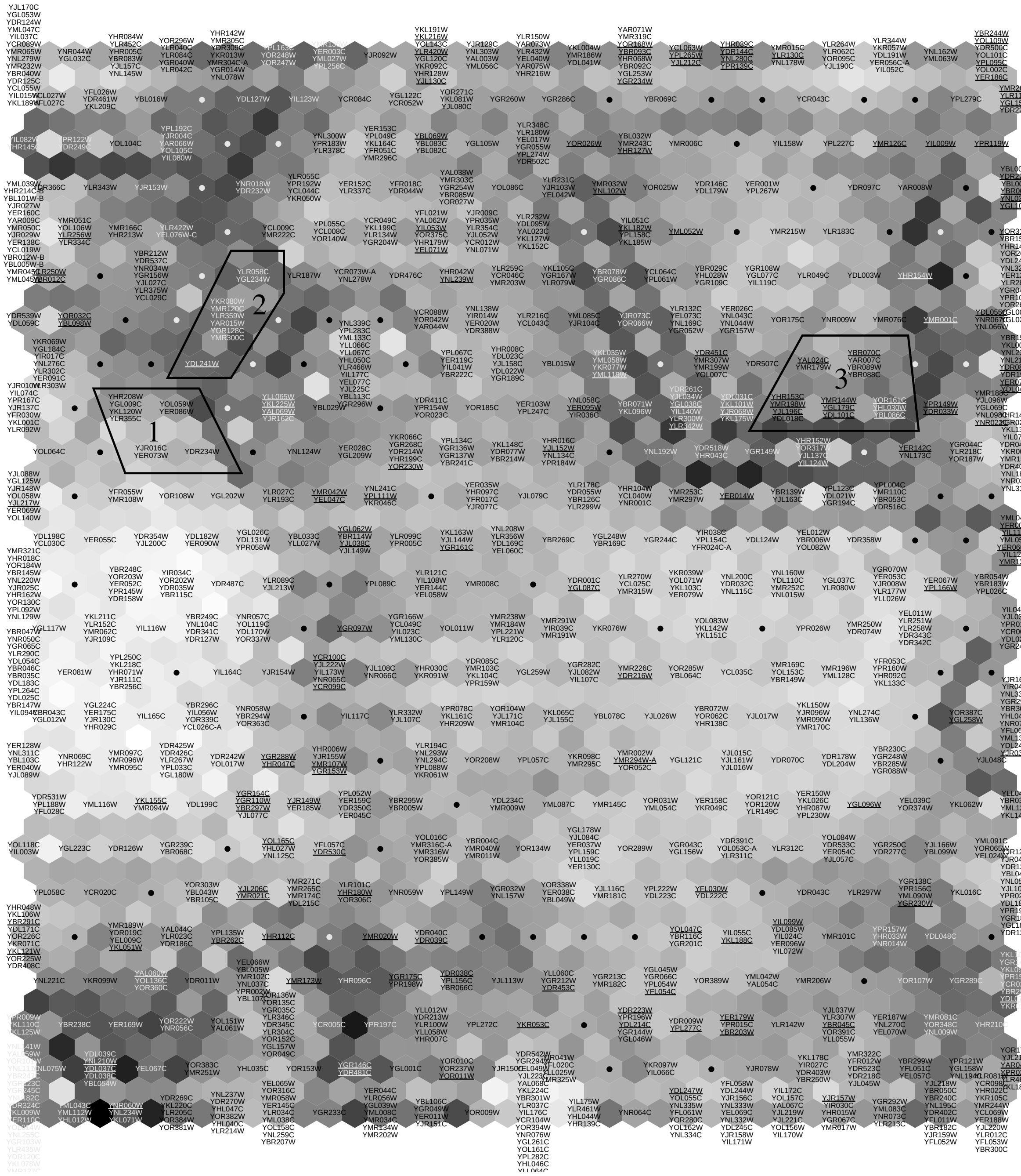

Figure 2: *Visualization of the similarity relationships of the yeast genes with a SOM in learning metrics. The names of the genes have been marked on the SOM units onto which they have been mapped, and the units with no genes have been marked with dots. The 202 genes for which the metric is most different from the usual inner product metric have been underlined. Gray shades indicate clusteredness (light: cluster; dark: sparse area). Three sample areas analyzed briefly in the text have been encircled and numbered.*

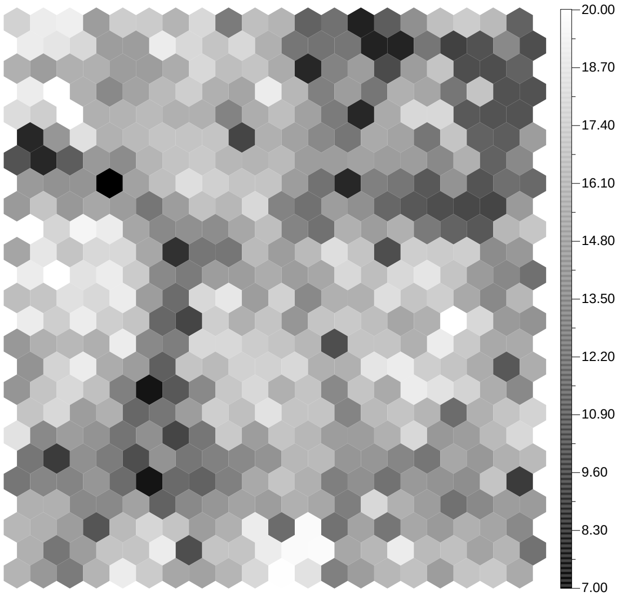

Figure 3: *Summary of where the metric has changed. The gray shade denotes, within the map unit, the average number of preserved neighbors in the neighborhood of 20 genes caused by changing the metric. The SOM is the same one shown in Figure 2; note that here each hexagon corresponds to one SOM unit, whereas in the U-matrix display of Figure 2 there is an extra hexagon in between each pair of neighboring SOM units.*

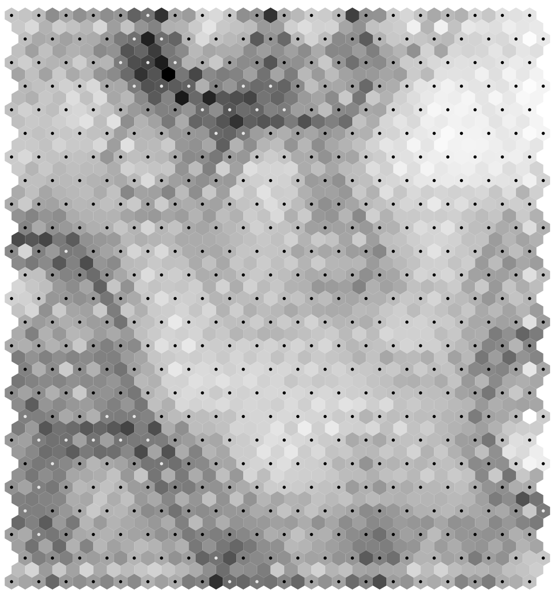

Figure 4: *Visualization of the similarity relationships of the yeast genes in inner product metric. Gray shades indicate clusteredness (light: cluster; dark: sparse area). Compare with the corresponding visualization in learning metrics in Figure 2.*

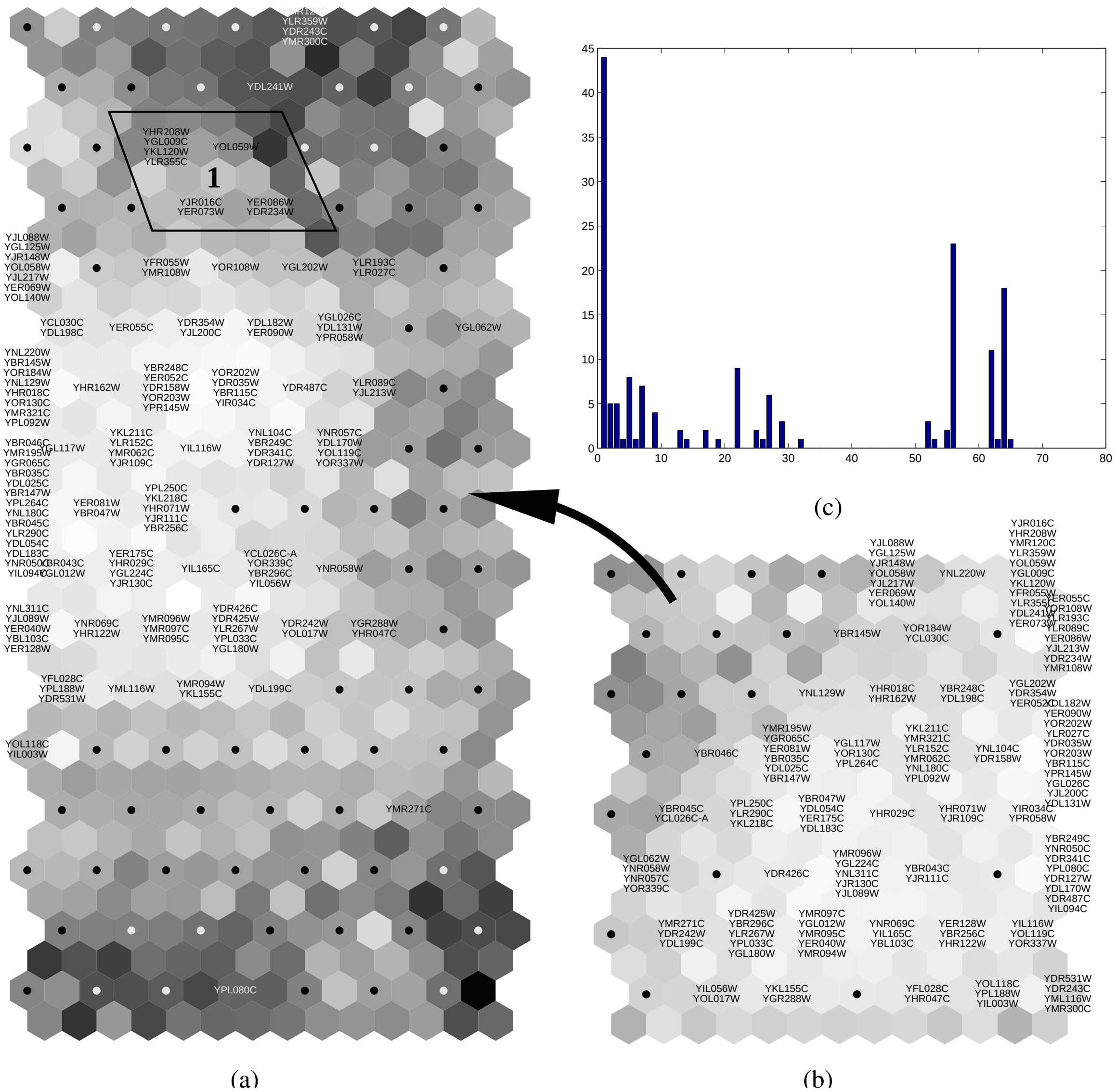

Figure 5: *Differences in cluster structure. The genes in the upper right hand corner cluster of the SOM in inner product metric (SOM-IP) (b) were projected to the SOM in learning metrics (SOM-LM) (a). In the SOM-LM the genes were found from the large cluster in the left side of the map and from its surroundigs. The genes from the SOM-IP cluster spread out to the SOM-LM because the functional classifications of the genes differ from each other. For example the group of genes marked with 1 in (a) consists of genes which belong to amino acid metabolism class and mitochodrial organization class when other genes in the SOM-IP cluster belong mainly to amino acid metabolism class. The class distribution of the projected genes is represented in (c), which shows clearly that the amino acid metabolism class (class no. 1) is predominent in the group.*

is the median of the neighboring hexagons.

A summary of the changes caused by changing the metric is shown in Figure 3. Comparing the display with Figure 2 reveals that the main cluster structures are similar in both metrics, and the most salient changes have occurred in the sparser areas of the expression space. The similarity of the main cluster structures can also be observed by comparing Figures 2 (SOM in learning metric) and 4 (SOM in inner product metric).

Differences in the clusterings of the different metrics was further analyzed by mapping the genes of a cluster in the SOM in inner product metric to the SOM in learning metrics (see Figure 5). Spreading of the genes on the SOM in learning metrics probably results from differences in the functional classifications. For example, the small group or subcluster marked with 1 in Figures 2 and 5(a) is related to mitochondrial organisation, in contrast to the other genes in the large cluster area in Figure 5(a). Four of the genes were also associated with branched-chain amino acid biosynthesis (YJR016C, YHR208W, YLR355C, YCL009C). These were grouped together with two genes related to fermentation and carbohydrate utilization (YOL059W, YER073W) and genes involved in threonine and lysine, as well as leusine biosynthesis (YDR234W, YER086W). All the genes with a known function in this cluster were located to mitochondria. The gene YKL120W was clustered with these genes; its function is unknown but it has been suggested based on sequence similarity to be related to mitochondrial proteins. Hence, results with the SOM in learning metrics support this assumption.

We then analyzed the genes for which the metrics had changed the most, based on the visualization in Figure 2 (such genes are underlined). The analysis of functional classifications and annotations of area number 3 revealed that 7 out of the 16 genes were associated with cell cycle and/or transcription. Besides these there were only two known genes, YBR070c and YJL196c, that were associated with stress response and fatty acid elongation.

Finally, we sought groups of genes belonging to known pathways from the visualization. Some of such groups occurred clustered on the SOM. The genes involved in purine biosynthesis occurred together on the area number 2 in Figure 2. These genes and other genes close-by were classified to nucleotide metabolism. In addition, the gene YDL241W, whose function is unknown, was located nearby. Hence, it might be worthwile to examine closer whether this gene is also related to nucleotide metabolism or the purine biosynthesis pathway.

4 DISCUSSION

The learning metrics principle was applied to derive a local metric based on functional classes of the genes. The resulting metric measures changes in gene expression but weights the changes according to how much they contribute to changes of the functional classes. The resulting metric is a step towards a more comprehensive picture of the functional similarity of the genes, incorporating prior biological knowledge to the measurements.

The new metric did not differ drastically from the old inner product (correlation) metric, which may either indicate that the inner product metric is a fairly good choice, or then there may be problems with the method or with the auxiliary data. The accuracy of the approximations of the new metric may be too low. An alternative explanation is that the functional classification may be on a too general level for the purpose of deriving metrics for expression data.

In future work, we will try other data sets and still improve the accuracy of the methods.

5 ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland, in part by the grants 50061 and 52123.

REFERENCES

1. Hastie, T., Tibshirani, R. & Buja, A. (1995). Flexible discriminant and mixture models. In Kay, J. & Titterington, D., (eds.) *Neural Networks and Statistics*. Oxford University Press.

2. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffrey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. & Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

3. Kaski, S., Nikkilä, J., Törönen, P., Castren, E. & Wong, G. (2001a). Analysis and visualization of gene expression data using self-organizing maps. In *Proc. NSIP-01, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*. (Proceedings on CD-ROM.)

4. Kaski, S., Sinkkonen, J. & Peltonen, J. (2001b). Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Trans. Neural Networks*, **12**, 936–947.

5. Kaski, S., & Sinkkonen, J.(2002). Principle of learning metrics for data analysis *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, special issue on Data Mining and Biomedical Applications of Neural Networks*. Accpted for publication.

6. Kohonen, T. (1995). Self-Organizing Maps. Springer, Berlin. (Third, extended edition 2001).

7. Mardia, K. V. (1975). Statistics of directional data. *JRSSB*, **37**, 349–393.

8. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. & Weil, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

9. Oja, M., Nikkilä, J., Törönen, P., Wong, G., Castrén, E. & Kaski, S. (2002). Exploratory clustering of gene expression profiles of mutated yeast strains. In Zhang, W. & Shmulevich, I., (eds.) *Computational And Statistical Approaches To Genomics*. Kluwer, pp. 64–78.

10. Oja, M. (2002). Masters thesis: Gene expression data mining using self-organizing maps and learning metrics Helsinki University of technology. (In Finnish).

11. Sinkkonen, J. & Kaski, S. (2002). Clustering based on conditional distributions in an auxiliary space. *Neural Comput.*, **14**, 217–239.

12. Ultsch, A. (1993). Self-organizing neural networks for visualization and classification. In Opitz, O., Lausen, B. & Klar, R., (eds.) *Information and Classification*. Springer-Verlag, Berlin, pp. 307–313.

13. Vesanto, J. & Alhoniemi, E. (2000). Clustering of self-organizing map. *IEEE Trans. Neural Networks*, **11**, 586–600.