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

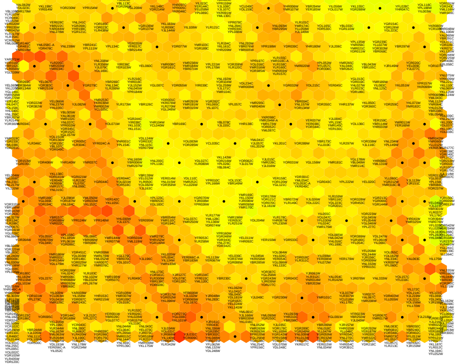
$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) \equiv D(p(c|\mathbf{x})||p(c|\mathbf{x} + d\mathbf{x})) = d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x},$$

SOM in learning metrics

The winner search is done using d_L as distance measure.

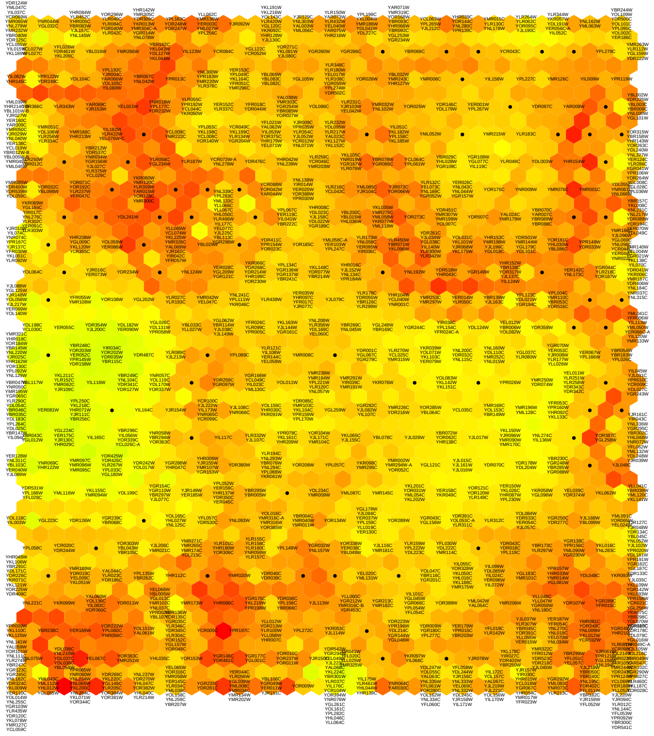
The update rule for learning metrics turns out to be the same as in the Euclidean metric:

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



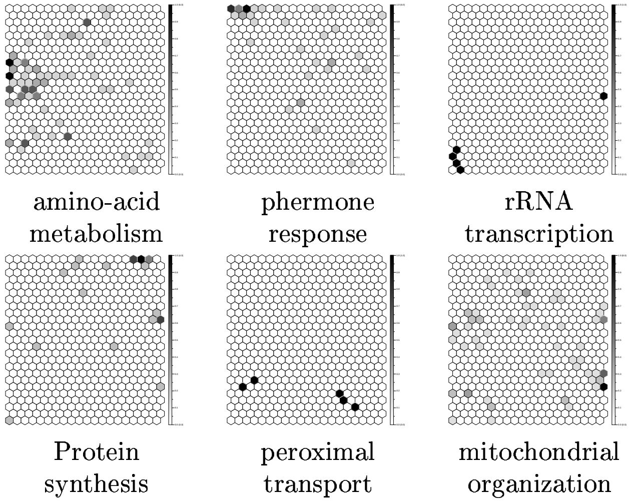
clusterlike areas in the data and red shades less dense areas. The names of genes mapped to each unit are shown on the map.

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



amino-acid metabolism

phermone response

rRNA transcription

Protein synthesis

peroximal transport

mitochondrial organization

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

- Comp
vealed
to m
genes
acid b
relate
tion (
in th
thesis
classi
genes
sugge
relate
SOM
- Proje
that t
locali
close
In ad
know
might
purin
ined c

Conclu

- The
SOM
expre
- In fu
other
data.

Referen

[1] T. R.
pendi
2000.

[2] S. Ka
analys
IEEE
2001.

[3] T. Ko
Sprin
Berlin
1997,

[4] M. Oj
and S
sion p
I. Shr
Appro
ers, 20

[5] J. Sin
dition
Comp