Method for exploring dependencies in the expression of orthologous man-mouse gene pairs

L. Lahti[1], J. Knuuttila[2], J. Nikkilä[1] and S. Kaski[3,1]
[1] Laboratory of Computer and Information Science, Helsinki University of Technology, Finland.
[2] Neuroscience Center, University of Helsinki, Finland.
[3] Department of Computer Science, University of Helsinki,Finland.

Introduction

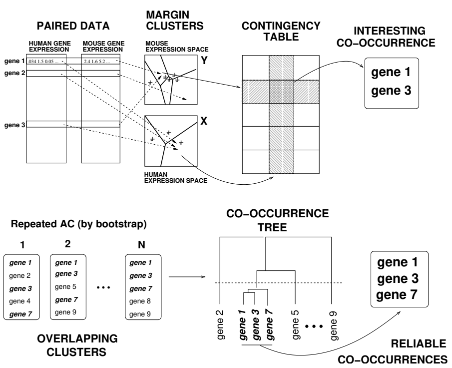
Functions of human genes are often studied indirectly, by model organisms such as the mouse. An underlying assumption is that genes with a common evolutionary origin i.e. orthologous genes, have similar functional roles in both species.

It is critical to know whether the orthologues really function in the same way. Orthologous gene groups with unexpectedly many similarly expressed gene pairs may highlight important physiological similarities between the species. Diverged gene function may refer to significant evolutionary changes.

Regularities and exceptions can be sought by analyzing expression data from two organisms. Here a dependency exploration tool called associative clustering (AC) [1] is used to identify orthologous man-mouse gene pairs with potentially interesting regularity in their expression (Figure 1).

AC has been designed for the first stage of analysis of large data sets. It explores dependencies in paired continuous-valued multivariate data, clustering the two sets of samples simultaneously. The results are represented as easily interpretable clusters.

Figure 1. The principle of associative clustering

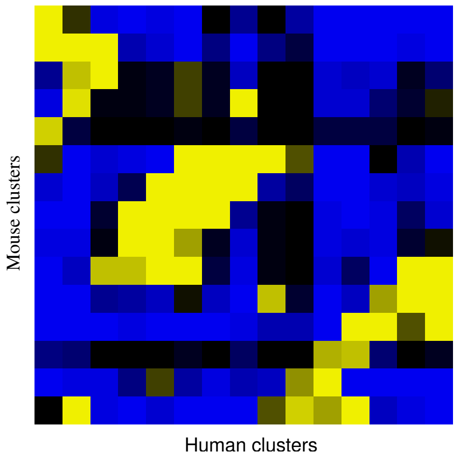


Principle of AC. Two sets of gene expression profiles are clustered simultaneously, and the results are represented on a contingency table where clusters of the two data sets correspond to the rows and columns of the contingency table, respectively. For this reason, they are called the margin clusters of the contingency table. Table cells are called cross clusters, and they contain orthologous genes from the two data sets. Cluster centroids are optimized to produce a maximally dependent contingency table. Potentially interesting co-occurrences are revealed by the contingency table cells with significant deviation from the hypothesis of independent margins. To enhance the reliability of the results, the clustering is repeated with slightly differing samples. Then reliable gene co-occurrences are identified from a co-occurrence tree with a specified threshold. Frequently co-occuring orthologues are selected for further analyzes.

Methods

- During optimization, the two sets of clusters are assigned on the rows and columns of the contingency table, respectively. Cluster centroids are optimized to produce a maximally dependent contingency table, in the sense of the Bayes factor of dependent vs. independent margins.

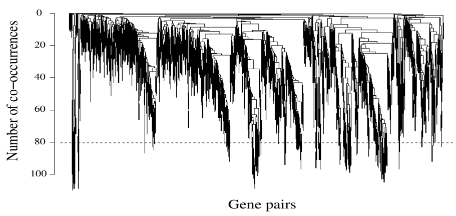
Figure 2. Contingency table of AC



Contingency table of the AC for the orthologous man-mouse gene expression data. Each human and mouse gene cluster corresponds to one of the contingency table columns and rows, respectively. The margin clusters imply the cross cluster for each orthologous man-mouse gene pair. Colors indicate the deviation of cross cluster size from the hypothesis of independent margins. Exceptionally small cross clusters are blue whereas exceptionally large cross cluster are yellow. The table is ordered using the cross-cluster p-values.

- Significantly deviating cross clusters reveal orthologues with potentially interesting regularity in their expression profiles.
- Reliability of the results is enhanced by repeating the clustering with slightly differing samples and inspecting the repeated co-occurrences. Here bootstrap and hierarchical clustering are used to identify such frequencies.

Figure 3. Hierarchical clustering of gene pair co-occurrences



Gene pair co-occurrence in the 100 AC runs with differing bootstrap samples is illustrated using hierarchical clustering. Gene pairs that frequently (> 80%) co-occur in dependent cross-clusters refer to regularity in the expression of orthologous genes.

Experiments

- 2818 paired expression profiles from orthologous man-mouse gene pairs were available [2] with 46 and 45 human and mouse tissues, respectively. For 21 tissues, expression values were available for both species.
- AC was performed with 15 margin clusters for both expression data sets, resulting in a 15 × 15 contingency table with 12 samples per cluster on average (Figure 2).

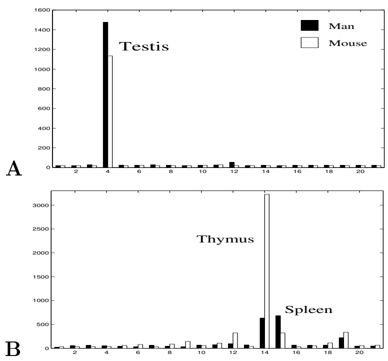
For more information see:
http://www.cis.hut.fi/projects/bioinf/

- AC was repeated 100 times with different bootstrap samples. Frequently co-occurring gene pairs were identified using hierarchical clustering (Figure 3).

Results and conclusions

- AC produced significantly more dependent clusters than standard K-means clustering. However, the cluster homogeneity was not significantly reduced by focusing on dependency modeling. The advantage of AC in comparison with other dependency modeling algorithms, e.g. information bottleneck [3] is its ability to produce more easily interpretable localized clusters.
- Two examples of frequent co-occurrences in unexpectedly large cross clusters demonstrate the capability of AC to capture biologically meaningful clusters (Figure 4).
- Orthologous genes often share functional similarity and this is also visible at the expression level. Thus, correlated expression is likely found in cross clusters. Additionally, AC is able to identify other types of functional frequencies.
- Clusters with potential implications in biomedical research are chosen for a further study based on the biological interpretation of the results.

Figure 4. Frequent orthologue expression trends



Averaged expression profiles of two reliable clusters from the hierarchical clustering of bootstrapped AC in Figure 3 with cluster size of 3 gene pairs in both cases. Only the common 21 tissues are shown here for clarity. **A** Function of these genes are related to reproductive system. **B** In this case the linking nominator within the gene cluster is T cell maturation and activation. There are notable differences in the expression profiles of the two species.

References

[1] J. Sinkkonen et al. Associative Clustering by Maximizing a Bayes Factor *Technical report A68*, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 2003

[2] A. I. Su et al. Large-scale analysis of the human and mouse transcriptomes *PNAS*, 99:4465–4470, 2002.

[3] N. Friedman et al. In Proc. 17th Conference on Uncertainty in Artificial Intelligence, pp. 152-161 Morgan Kaufmann Publishers, San Francisco, CA, 2001.