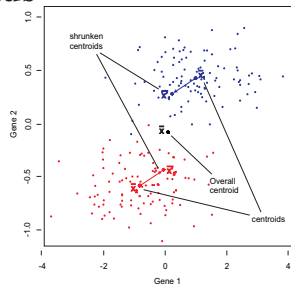


## Problem Statement:

- **Goal:** Medical diagnosis using gene expression patterns.
- **Data:** Many genes - few samples - no structure.
- **Approach:** Use functional annotation in addition to expression data.
- **Special feature:** Intuitive rationale for computational results.

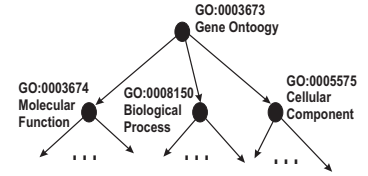
## Class Prediction - Nearest Shrunken Centroids [Tibshirani et al., 2002]

- **Centroids:** Characteristic expression levels for each class.
- **Shrinkage:** move centroids towards overall centroid.
- **Soft thresholding:** covariates progressively loose influence with increasing within-class variance.
- **Shrinkage parameter  $\Delta$ :** Controls the amount of shrinkage, is fine-tuned by cross-validation.
- **Classification:** New cases are attributed to the nearest shrunken centroid.

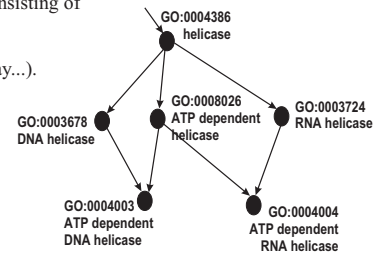


## Gene Ontology (GO) - Structuring Biological Knowledge [GO Consortium, 2000]

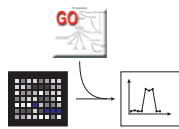
- Structure biological knowledge using a directed acyclic graph.
- Defined by an international consortium consisting of many academic and industrial members.



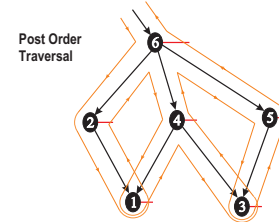
- **Nodes:** biological terms (function, pathway...).
- **Relations:** "member of" and "is a".
- Several thousands of nodes and relations are currently registered in the Gene Ontology.



## StAM - Structured Analysis of Microarray Data Combining Local to Global Predictors

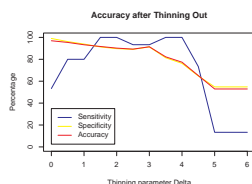
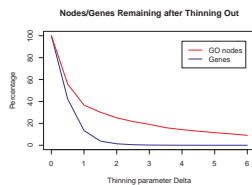


- Probe-sets are annotated to GO-nodes.
- There is a diagnostic predictor for each GO-node.
- It assigns samples according to expression similarity of genes annotated to this node or its descendants.
- Each node predictor combines two classifiers: one has the genes annotated to the node as inputs, the other has the output of its children as inputs.
- All classifiers are based on the nearest shrunken centroids method.
- Bottom-up information flow: training and classification start with leaf nodes and proceed towards the root (post-order traversal)



## Regularization - Thinning Out the Ontology

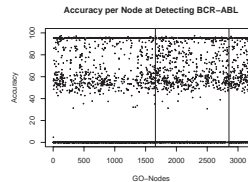
- Shrinkage leads to exclusion of nodes.
- With a parent node all its children disappear, thus the Gene Ontology is thinned out.
- This thinning is controlled by the shrinkage parameter  $\Delta$ .
- Leave-one-out cross validation is used to determine the appropriate amount of thinning.
- Thinning fights overfitting.
- Figures: separate BCR-ABL translocation from others



## Model Inspection - Pin-point Hot Spots

- Dataset: 327 cases of leukemia, various translocation types, hybridised on HG-U95Av2 (12625 probe-sets) [Yeoh et al., 2002].
- 2979 GO-nodes have 7115 probe-sets annotated.
- Evaluation per node reveals particularly reliable classifiers to detect a given class.

- Figure: nodewise Accuracy for Classification of BCR-ABL



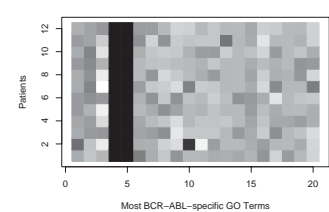
### Top-List

- GO:0003706: ligand-regulated transcription factor (97.2%)
- GO:0008652: amino acid biosynthesis (97.2%)
- GO:0005261: cation channel (96.9%)
- GO:0003674: molecular\_function (96.%)
- GO:0003673: Gene\_Ontology (96.9%)
- GO:0004519: endonuclease (96.6%)
- GO:0019932: second-messenger-mediated signaling (96.6%)
- GO:0005605: basement lamina (96.6%)

## Diagnosis - Which Class and Why?

- The BCR-ABL group has multiple molecular characteristics.
- The diagnosis of different patients is based on different characteristics.
- StAM allows for tracing back the evidence for diagnosis.

Color-coded BCR-ABL Probabilities



- Figure: GO Terms are sorted according to accuracy, best performing nodes to the left (see-Top List besides). Dark colors represent high probability for classification into BCR-ABL. All patients shown here are correctly classified into this group.

## References

- [GO Consortium, 2000] The Gene Ontology Consortium. Gene ontology: Tool for the unification of biology. Nature Genetics, 25:25-29, May 2000.
- [Tibshirani et al., 2002] R. Tibshirani, T. Hastie, B. Narasimhan, Chu G: Multi-class diagnosis of cancers using shrunken centroids of gene expression. Proc Natl Acad Sci USA 99: 6567-6572, 2002.
- [Yeoh et al., 2002] E. J. Yeoh, M. E. Ross et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell, 1:133-145, March 2002.

## Acknowledgements

We have recently started a collaboration with Christian Hagemeier, Wolf-Dieter Ludwig, Karl Seeger, Renate Kirschner Charité (Universitätsklinikum Medizinische Fakultät der Humboldt Universität zu Berlin) on ALL relapse expression profiling. StAM is designed for use in this study.