BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–6

Predicting Pathway Membership via Domain Signatures

Holger Fröhlich*, Mark Fellmann, Holger Sültmann, Annemarie Poustka, Tim Beißbarth

German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Functional characterization of genes is of great importance for the understanding of complex cellular processes. Valuable information for this purpose can be obtained from pathway databases, like KEGG. However, only a small fraction of genes is annotated with pathway information up to now. In contrast, information on contained protein domains can be obtained for a significantly higher number of genes, e.g. from the InterPro database.

Results: We present a classification model, which for a specific gene of interest can predict the mapping to a KEGG pathway, based on its domain signature. The classifier makes explicit use of the hierarchical organization of pathways in the KEGG database. Furthermore, we take into account that a specific gene can be mapped to different pathways at the same time. The classification method produces a scoring of all possible mapping positions of the gene in the KEGG hierarchy. Evaluations of our model, which is a combination of a SVM and ranking perceptron approach, show a high prediction performance. Moreover, for signaling pathways we reveal that it is even possible to forecast accurately the membership to individual pathway components.

Availability: The R package *gene2pathway* is a supplement to this paper.

Contact: {h.froehlich, t.beissbarth}@dkfz-heidelberg.de

1 INTRODUCTION

Biological characterization of genes is of fundamental importance for the understanding of complex cellular processes, like cancer. Valuable information can be obtained from databases, like the Gene Ontology (GO) (The Gene Ontology Consortium, 2004) or KEGG (Kanehisa et al., 2008). However, usually only a small fraction of genes have known functions. Most genes are annotated in GO, only few in KEGG. For example, the total number of human genes annotated in KEGG currently is about 4,000. This contrasts remarkably with the estimated number of putative protein-coding genes, which is 20,000 - 25,000 (Pennisi, 2007). It is therefore highly important to link other sources of information with these databases to improve the quality of biological characterization. Especially interesting for this purpose is the InterPro database (Mulder et al., 2008), which offers predicted protein domain annotation for $\sim$19,000 genes. Of the 4,000 genes in the KEGG database nearly all have at least one InterPro domain. Together, these comprise $\sim$3,000 distinct InterPro domains. Protein domains very often directly correspond to some core biological function, such as DNA binding, kinase or phosphorylation activity, or to cellular localization. Hence, predicted protein domains are often utilized for prediction annotations, such as in the GO database.

Hahne et al. (2008) introduced a method linking protein-domain signatures with assignments of genes to KEGG pathways. In this approach one looks for a protein domain signature being significantly enriched in a list of genes. This information is then used to find the most probable pathway these genes come from by comparing the enriched protein domain signature with all pathway domain signatures.

In contrast to Hahne et al., our aim is to make a prediction and thus a biological characterization for *individual* genes. This broadens the applicability of our method significantly. We explicitly take into account that a particular gene can be mapped to different pathways at the same time. Furthermore, our classifier makes use of the hierarchical organization of the KEGG database in 3 levels: At the top hierarchy there are the 4 branches "Metabolism", "Genetic Information Processing", "Environmental Information Processing" and "Cellular Processes" (we do not consider "Human Diseases" here). On the next hierarchy level each of these branches is divided further. For instance, "Environmental Information Processing" contains the branches ""Membrane Transport", "Signal Transduction" and "Signaling Molecules and Interaction". On the third hierarchy level we have the individual KEGG pathways. We expect that a good classifier should give especially precise predictions at the top levels of the KEGG hierarchy, while at the bottom levels misclassifications are more tolerable. That means it is worse to predict a MAPK pathway (branch "Signal Transduction" in "Environmental Information Processing") gene to be involved in "Olfactory transduction" (branch "Sensory System" in "Cellular Processes") than to predict it as a member of some other signal transduction pathway. This behavior, leading to a hierarchical classification scheme, is encoded into an appropriate loss function within our framework. Our classifier is also able to indicate the reliability of a pathway prediction. A 10×10-fold crossvalidation experiment with 2346 genes having both, a KEGG annotation and a unique protein domain signature, shows that our method yields good classification performance. We further demonstrate the usefulness of our method on a microarray dataset, where we obtain meaningful results.

Signaling pathways are of special importance for the functioning of biological systems. In an extension of our approach we demonstrate that it is not only possible to reliably predict a gene's membership to the different signaling pathways, but also to connected pathway components within individual signaling pathways. Again,

*to whom correspondence should be addressed

results on our microarray dataset show the biological relevance of our method.

2 METHODS

2.1 Hierarchical KEGG Pathway Classification

2.1.1 Classification Scheme We suppose that each gene product p is represented by a binary vector $\mathbf{x}$ with component $x_i = 1$, if the corresponding InterPro domain is contained in the protein and 0 otherwise. We hereby have to take into account that InterPro domains are organized in a hierarchical fashion. Hence, if domain i is contained in p, also all its parent domains are contained in p, and therefore all corresponding positions in $\mathbf{x}$ have to be 1 as well.

The mapping position(s) of a gene to the KEGG hierarchy can be encoded into a binary vector $\mathbf{C}$ as well. The dimension K of this vector equals the number of individual KEGG pathways + the number of branches at level 2 + the number of branches at top level. We set component $C_l = 1$, if the gene maps to the corresponding branch or any of its sub-branches. Note that any position code vector $\mathbf{C}$ can contain more than one 1, if the corresponding gene maps to more than one branch in the KEGG hierarchy.

Given a binary vector representation $\mathbf{x}$ for a gene product p, our classification scheme now consists of two basic steps, which are an adaption of an approach proposed by Melvin et al. (2007) for classifying proteins within the SCOP hierarchy:

1. On each hierarchy level we use Support Vector Machine (SVM) classifiers, trained to separate one specific branch from all others. Linear kernels are used, and all soft margin parameters $C = 1$. Each SVM classifier j will produce a decision value $f_j(\mathbf{x}) \in \mathbb{R}$. Please note that the decision value is not the same as the predicted class label, which is the sign of the decision value. For each gene product p represented by a binary vector $\mathbf{x}$ we summarize the decision values of all K SVMs into a input code vector $\vec{f}(\mathbf{x}) = (f_1(\mathbf{x}), ..., f_K(\mathbf{x}))$.
2. Each input code vector $\vec{f}(\mathbf{x})$ is mapped on the best matching position code vector(s)

$$\mathbf{C}^* = \mathbf{C}_{\hat{j}} \quad (1)$$

$$\hat{j} = \arg\max_j \langle \mathbf{C}_j, \vec{f}(\mathbf{x}) * \mathbf{w} \rangle \quad (2)$$

where $\{\mathbf{C}_1, ..., \mathbf{C}_m\}$ is a dictionary of possible position vectors, $\mathbf{w}$ is a weight vector and $*$ indicates component-wise multiplication. The dictionary of position vectors consists of all unique position vectors from a training set of gene products with both, KEGG and InterPro domain annotation. The weight vector $\mathbf{w}$ is chosen to minimize the mismatch between predicted and true KEGG hierarchy positions on the training data.

Please note that the maximum in Eq. (2) is not necessarily unique. In other words, it is possible to predict several positions vectors, which are all equally likely. Hence, we capture the often appearing situation that a gene maps to several positions in the KEGG hierarchy at the same time.

2.1.2 Training Procedure Similar to the classification scheme, the training procedure consists of two steps.

1. All K binary SVM classifiers are trained to obtain a position labeled data set $D = \{(\vec{f}_1(\mathbf{x}_1), \mathbf{C}_1), ..., (\vec{f}_n(\mathbf{x}_n), \mathbf{C}_n)\}$. For training the individual SVMs we only use genes belonging to the same super-branch. E.g. for training the SVM classifier detecting signal transduction, we only use genes mapping to other branches than signal transduction in "Environmental Information Processing" as negative examples. Each SVM classifier is thus trained to detect one specific branch in the KEGG hierarchy only.
2. Given the position labeled data set D, we employ the modified ranking perceptron algorithm presented in Melvin et al. (2007) to learn a weight vector $\mathbf{w}$ of the input code vectors $\vec{f}_i(\mathbf{x}_i)$. In the spirit of SVM classifiers the weight vector is optimized to maximize the margin between position code vectors $\mathbf{C}_i$, $\mathbf{C}_j$ with $\mathbf{C}_i \neq \mathbf{C}_j$ in input code vector space. The algorithm shown in Figure 1 involves updating $\mathbf{w}$ proportional to the loss we obtain by predicting a wrong position vector $\mathbf{C}_j$ instead of the true position vector $\mathbf{C}_i$. The choice of this loss function is the essential part of the algorithm, because it reflects our knowledge about the KEGG hierarchy. Making a wrong prediction at the higher levels of the hierarchy should be punished more than confusing individual KEGG pathways at the bottom level. We therefore set up the following loss function:

$$\ell(\mathbf{C}, \mathbf{C}') = \sum_{i=1}^{K} c_i \mathbf{1}\{C_i \neq C_i' \text{ and } ((C_j = C_j' \ \forall j \in Anc(j)) \text{ or } (Anc(j) = \emptyset))\} \quad (3)$$

where Anc denotes the set of all ancestors of branch j and $\mathbf{1}$ is the indicator function. By this loss function we punish the first mismatch on the path down the hierarchy to the final predicted position. The higher in the hierarchy the mismatch occurs, the higher the punishment c_i should be. We thus choose

$$c_i = \frac{|T(i)|}{|T(root)|} \quad (4)$$

where $|T(i)|$ denotes the size of the hierarchy down of branch i and $|T(root)|$ is the size of the complete KEGG hierarchy.

Algorithm 1 Pseudocode for the ranking perceptron algorithm to learn the input code vector weighting. The learning rate η was set to 0.1 here.

Input: Learning rate η, position labeled data set D
Output: weight vector $\mathbf{w}$
Define $F(\mathbf{x}, y) = \langle \mathbf{C}_y, \vec{f}(\mathbf{x}) * \mathbf{w} \rangle$
$\mathbf{w} = \mathbf{0}$
for $i = 1$ to n
  $foes(i) = \{1, ..., n\} - \{p | \ell(\mathbf{C}_i, \mathbf{C}_p) = 0\}$
  $l = \arg\max_{p \in foes(i)} F(\mathbf{x}_i, p)$
  if $F(\mathbf{x}_i, i) - F(\mathbf{x}_i, l) < 2$
    $\mathbf{w} \leftarrow \mathbf{w} + \eta \cdot \ell(\mathbf{C}_i, \mathbf{C}_l) \cdot (\vec{f}(\mathbf{x}_i) * \mathbf{C}_i - \vec{f}(\mathbf{x}_i) * \mathbf{C}_l)$
end for

2.2 Hierarchical Signaling Pathway Component Classification

Viewing all gene-gene interactions as an undirected graph, we calculated the connected components for each signaling pathway (Siek et al., 2002). Our hierarchy for signaling pathways thus consists of two levels: At the first level we have all individual signaling pathways and at the second level we have their corresponding connected components. The training and classification procedure is then the same as described above.

3 RESULTS

3.1 Estimating Prediction Performance

3.1.1 Hierarchical KEGG Pathway Classification We used all human genes annotated in both, KEGG and InterPro. KEGG annotation was retrieved via the R package KEGG 2.0.1 (released

August 2007). InterPro annotation was retrieved directly from the Ensembl database (Flicek et al., 2008) via the R package biomaRt 1.12.1 in March 2008. Hierarchy information for KEGG and InterPro was obtained from the corresponding homepages via FTP in March 2008. 3705 genes had both, KEGG and InterPro annotation. Since for estimating the classification accuracy we employed a 10-fold cross-validation procedure, we decided to remove genes with the same InterPro annotation, thus avoiding an overoptimistic prediction performance estimation by having one of the duplicates in the training and one in the test set. This way our set of genes was reduced to 2346, containing 2752 distinct InterPro domains in total.

As already noted by Hahne et al. (2008) it is unlikely to reliably separate metabolic pathways based on their InterPro domain signatures. We thus decided to prune the KEGG hierarchy in order to improve the prediction accuracy for branches of especially high importance. We cut the hierarchy for metabolic pathways at the top and for "Genetic Information Processing" pathways at the 2nd hierarchy level. At the same time we required to have more than 30 genes to be mapped to the corresponding hierarchy branch in order to consider it in the classification hierarchy. This way we ended up with a total of 53 hierarchy branches to distinguish (Table 1).

We ran a 10 times repeated 10-fold cross-validation procedure to assess the prediction performance of our hierarchical classification model. The classification performance was evaluated using four different measures:

1. the *accuracy*, measured as 1 - average classification loss (Eq. 3)
2. the *precision* (also known as positive predictive value), defined as $\frac{TP}{TP+FP}$, where TP and FP are the number of true positives and false positives summed over all hierarchy branches. I.e. we first calculated true and false positives for each component in the position code vector individually and then summed up.
3. the *recall* (also known as sensitivity), defined as $\frac{TP}{FP+FN}$, where FN are the number of false negatives summed over all hierarchy branches.
4. the $F1$ value, defined as $\frac{2 \cdot precision \cdot recall}{precision+recall}$

The results, depicted in Figure 1 (left) as boxplots showed a high median accuracy of >95% and a median $F1$ value of $\sim$60% with precision and recall being in the same range. It should be noted that only the accuracy measure takes into account the KEGG hierarchy via the loss function Eq. 3, whereas the other 3 measures weight all errors equally. Further analysis of the median $F1$ values for all top level and second level hierarchy branches approximately showed a uniform distribution, i.e. all branches could be predicted equally well within each hierarchy level.

To train our final hierarchical classification model, which we employed to give predictions on further unseen datasets, we used the complete set of 3705 genes without removing duplicates. The number of hierarchy branches to distinguish was 58 now (see Table 1). For further improvement of predictive power and in order to obtain confidence scores for predictions, our final model was bagged (Hastie et al., 2001). That means we drew 11 bootstrap training datasets with replacement and trained our classification model on each of them. To give a prediction, the majority vote among these 11 sub-models was used. This was done for each component in the position code-vector separately. A confidence score for the complete prediction can then be calculated as

$$score = \frac{1}{2}(\bar{v}_+ + 1 - \bar{v}_-) \quad (5)$$

where $\bar{v}_+$ is the average of all vote proportions >50% and $\bar{v}_-$ the average of all vote proportions $\leq 50\%$.

3.1.2 Hierarchical Signaling Pathway Component Classification A setup similar to the one described above was chosen. The number of human genes with a unique InterPro domain signature and a corresponding KEGG annotation was 515 and the total number of used InterPro domains 795. A minimum of 10 mapping genes per pathway component was required. Therefore, we ended up with 19 hierarchy branches to distinguish (see Table 1).

The result, depicted in Figure 1 (right) showed a high median accuracy of $\sim$100% and a median $F1$ value of $\sim$70% with precision and recall being in the same range. Again, the median $F1$ values for all top level and second level hierarchy branches approximately followed a uniform distribution, i.e. all branches could be predicted equally well within each hierarchy level.

To train our final hierarchical classification model, the same procedure was used as described above. The total number of genes used for training was 788, and the number of hierarchy branches to distinguish was 22 (see Table 1).

3.2 Application to Microarray Data

We applied our method to predict the KEGG pathway membership for a microarray dataset produced in our department: human MCF-7 breast cancer cells were treated with 100 nM tamoxifen for 48 hours. On mRNA level effects were measured with in-house developed cDNA two-color microarrays having 26,722 functioning probes (Barth et al., 2006). After VSN normalization (Huber et al., 2002) 2937 differentially expressed genes were found with limma (Smyth, 2004) using a Benjamini-Hochberg FDR cutoff of 5% (Benjamini and Hochberg, 1995). Further details on the experiment can be obtained from the authors upon request. The 26,722 probes correspond to 12,692 genes with an Entrez gene ID, of which for 10,057 InterPro annotation and for 2760 KEGG annotation was available. Comparison of our predicted and the original KEGG pathway annotations for the 2760 common genes indicated a very good median accuracy of $\sim$100% with a median $F1$ value $\sim$80% and precision and recall in the same range (Figure 2 left). There were a few outliers, as indicated in the boxplot. These genes are mostly linked to the KEGG category "Human Diseases", which we did not include in our model.

By our model we could predict pathway memberships for several genes with previously unknown KEGG annotation: E.g. *NR2C2* is a member of the nuclear hormone receptor family and acts as ligand-activated transcription factor (Yoshikawa et al., 1996). We predicted *NR2C2* to belong to the branch "Neuroactive ligand-receptor interaction" (confidence = 99.66%), which exactly fits this knowledge. As another example we predicted *TOMM34* to be a member of the branches "Folding, Sorting and Degradation" and "Cell Cycle" (confidence = 100%). Indeed, the protein encoded by *TOMM34* is involved in the import of precursor proteins into mitochondria. The encoded protein has a chaperone-like activity, binding the mature portion of unfolded proteins and aiding their import into mitochondria (Chewawiwat et al., 1999).

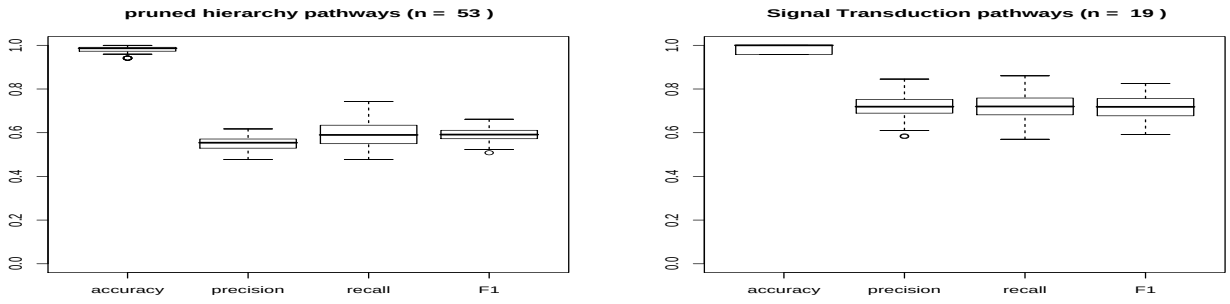

Figure 1. Prediction performance of our method (10×10-fold cross-validation). The accuracy measure uses the same loss-function, which was used to train the classifier, and which takes into account the KEGG hierarchy. **Left:** Pathway prediction within pruned KEGG hierarchy (53 branches). **Right:** Pathway component prediction for signaling pathways (19 branches)).

In a second step of our analysis we filtered those genes, which were either known to be involved in signal transduction by KEGG annotation (458 genes), or which were predicted by our model to map to the corresponding KEGG hierarchy branch with confidence >99% (164 genes). Comparison of our pathway component predictions for the 458 genes with the original KEGG information, revealed a very high median accuracy of ∼100% with a median $F1$ value >80% and precision and recall in the same range (Figure 2 right). As an example application of our model in Figure 3 we depict the predicted connected component for *PLCH2* (confidence = 100%) in the calcium signaling pathway, for which previously no KEGG annotation was available. The gene has an associated GO function "calcium ion binding" and GO process "intracellular signaling cascade" (The Gene Ontology Consortium, 2004).

In a final step we looked for those KEGG branches, which were statistically overrepresented in the set of differentially expressed genes compared to the rest. We used all predicted and all original KEGG annotation for this purpose. Fisher's exact test was employed to assess statistical significance, and a multiple testing correction using the method of Benjamini and Yekutieli (Benjamini, Y. and Yekutieli, D., 2001), which assumes statistical dependence of the individual tests, with a 10% cutoff was performed. The test shows an enrichment of metabolic, cell motility and cancer related pathways. None of this would have been found using KEGG annotation only.

4 CONCLUSION

We presented a novel hierarchical classification method, which can predict the KEGG annotations of individual genes based on their InterPro domain signatures. In an extension of our approach we showed that it is also possible to classify individual signaling pathway components via InterPro domain information. We think that linking KEGG with InterPro is an important step to generate new hypotheses about genetic pathways, which is finally of fundamental importance for a better understanding of human diseases like cancer. With our method it is not only possible to analyze lists of genes, as done in Hahne et al. (2008), but to give predictions for individual genes of interest. This way we can drop the unrealistic assumption that all genes in the list come from the same pathway. Moreover,

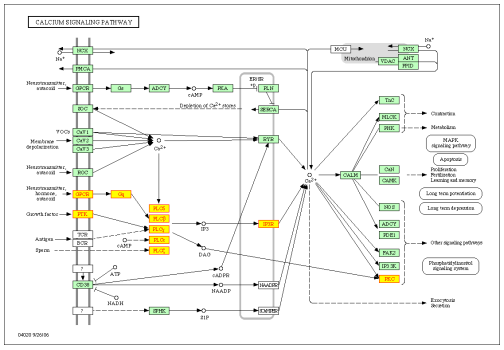

Figure 3. Predicted pathway component (marked yellow) for *PLCH2* in the Calcium signaling pathway.

our method is not restricted to microarray experiments any more, but can be used in a much broader spectrum of applications.

We have implemented our method in the R package *gene2pathway*, which is available as a supplement to this paper.

ACKNOWLEDGMENTS

TB and HF were supported by the National Genome Research Network (NGFN) of the German Federal Ministry for Education and Research (BMBF) - grants SMP Bioinformatics (01GR0450, subprojects PBF-S19T10, PBF-S02T11) and exploratory project EP-S19T03. MF and HS were supported by the NGFN SMP RNA (01GR0418). We thank Markus Ruschhaupt and Ruprecht Kuner for help and discussions, and Dirk Ledwinda for IT support.

REFERENCES

Barth, A. S., , Kuner, R., Buness, A., Ruschhaupt, M., Merk, S., Zwermann, L., Kääb, S., Kreuzer, E., Steinbeck, G., Mansmann, U., Poustka, A., Nabauer, M., and Sültmann, H. (2006). Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *Journal of the American College of Cardiology*, 48(8).

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. Royal Statist. Soc., Series B*, 57:289 – 300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165 – 1188.

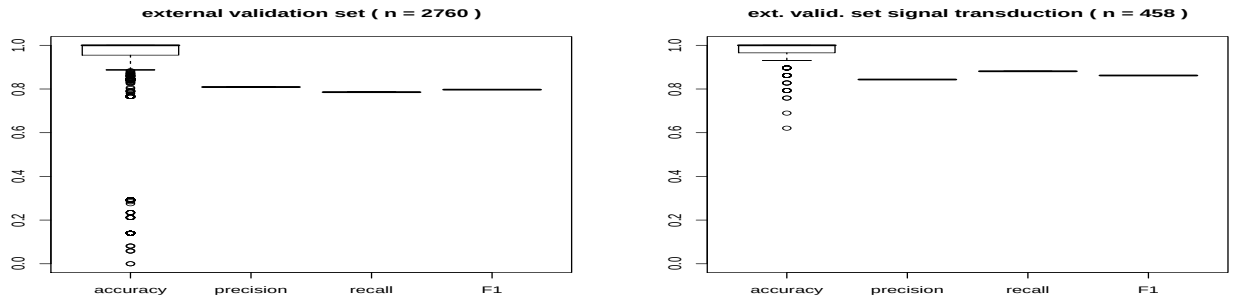

Figure 2. Prediction performance of the hierarchical classification model on an external validation set for the pruned KEGG hierarchy (left, 2760 genes) and for signaling pathway components (right, 458 genes).

Chewawiwat, N., Yano, M., Terada, K., Hoogenraad, N. J., and Mori, M. (1999). Characterization of the novel mitochondrial protein import component, tom34, in mammalian cells. *J Biochem*, 125(4):721–727.

Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kähäri, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., and Searle, S. (2008). Ensembl 2008. *Nucleic Acids Res*, 36(Database issue):D707–D714.

Hahne, F., Mehrle, A., Arlt, D., Poustka, A., Wiemann, S., and Beissbarth, T. (2008). Extending pathways based on gene lists using interpro domain signatures. *BMC Bioinformatics*, 9:3.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Huber, W., Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:S96 – S104.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). Kegg for linking genomes to life and the environment. *Nucleic Acids Res.*, 36:D480 – D484.

Melvin, I., Ie, E., Weston, J., Noble, W., and Leslie, C. (2007). Multi-class Protein Classification Using Adaptive Codes. *J. Machine Learning Research*, 8:1557 – 1581.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P. S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., nd Robert Petryszak, C. O., Selengut, J. D., Sigrist, C. J. A., Thomas, P. D., nd Derek Wilson, F. V., Wu, C. H., and Yeats, C. (2008). New developments in the InterPro database. *Nucleic Acids Res.*, 35:D224 – D228.

Pennisi, E. (2007). Genetics. working the (gene count) numbers: finally, a firm answer? *Science*, 316(5828):1113.

Siek, J. G., Lee, L.-Q., and Lumsdaine, A. (2002). *The Boost Graph Library: User Guide and Reference Manual*. Addison-Wesley, Pearson Education Inc.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

The Gene Ontology Consortium (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32:D258–D261.

Yoshikawa, T., DuPont, B. R., Leach, R. J., and Detera-Wadleigh, S. D. (1996). New variants of the human and rat nuclear hormone receptor, tr4: expression and chromosomal localization of the human gene. *Genomics*, 35(2):361–366.

Table 1: Pruned KEGG hierarchy used for our classification model. Hierarchy branches marked with '–' are left out in the cross-validation procedure, but are included in the final model. For signaling pathways the number in brackets indicate the number of connected pathway components. The first number refers to the number of connected pathway components used in the final model, and the second (italic) to the number used in the cross-validation procedure.

Level 1	**Level 2**	**Level 3**
Metabolism	—	—
Genetic Inf. Proc.	Transcription (–)	—
	Translation	—
	Folding, Sorting, Degradation	—
Env. Inf. Proc.	Membrane Transport (–)	—
	Signal Transduction	MAPK pathway ErbB pathway (2, *0*) Wnt pathway (2) Notch pathway Hedgehog pathway TGF-β pathway (3, *2*) VEGF pathway Jak-STAT pathway Calcium signaling (4) Phosphatidylinositol system mTOR signaling (–)
	Signaling Molecules and Interaction	Neuroactive ligand-receptor interaction Cytokine-cytokine receptor interaction ECM-receptor interaction Cell adhesion molecules
Cellular Processes	Cell Motility	—
	Cell Growth and Death	Cell Cycle Apoptosis p53 pathway
	Cell Communication	Focal adhesion Adherens junction Tight junction Gap junction
	Endocrine System	Insulin pathway Adipocytokine pathway PPAR pathway GnRH pathway Melanogenesis
	Immune System	Hematopoietic cell lineage Complement and coagulation cascades Toll-like receptor pathway Natural killer cell mediated cytotoxicity Antigen processing and presentation T cell receptor signaling B cell receptor signaling Fc-ϵ RI pathway Leukocyte transendothelial migration
	Nervous System	Long-term potentiation Long-term depression
	Sensory System	Olfactory transduction (–) Taste transduction (–)
	Development	—