Gene2Pathway
Predicting Pathway Membership via Domain Signatures

Holger Fröhlich*

October 2, 2012

Abstract

Functional characterization of genes is of great importance, e.g. in microarray studies. Valueable 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.

The R package *gene2pathway* implements 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. For signaling pathways it is even possible to forecast accurately the membership to individual pathway components.

1 Introduction

Microarray expression experiments have become a major high throughput analysis method during the last years. In a typical biological research setup people first rank all probes according to their differential expression, using tools like SAM or limma [6, 8]. In a second step a biological characterization and interpretation of differentially expressed genes is needed. For this purpose valueable information can be obtained from databases, like the Gene

*German Cancer Research Center, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. eMail: h.froehlich@dkfz-heidelberg.de

Ontology [7] or KEGG [2]. However, usually only a small fraction of differentially expressed genes is annotated within these databases. 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 encoding genes, which is more than 23,000 (counted as Entrez gene IDs in the IPI human database [3, 4]). 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 [5], which offers predicted protein domain annotation for 19,000 of all 23,000 genes in the IPI human database. Of the 4,000 genes in the KEGG database nearly all have at least one InterPro domain. Together, these comprise 3,000 distinct InterPro domains. Protein domains very often directly correspond to some core biological function, such as DNA binding, kinase or phophorylation activity, or to cellular localization. Hence, predicted protein domains are often utilized for prediction annotations, such as in the GO database.

Hahne et al. [1] introduced a first 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 compairing 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 specific 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, was encoded into an appropriate loss function within

our framework. Our classifier is also able to indicate the reliability of a pathway prediction via a bagging procedure.

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

More details on our hierarchical classification models can be found in the accompanying paper.

2 Example Usage

Usage of the R package *gene2pathway* mainly involves two functions: `gene2pathway` and `gene2pathway.signaltrans`. `gene2pathway` predicts the KEGG pathway membership for a given list of genes. The mapping of genes to InterPro domains can be done automatically via Ensembl, if Entrez gene IDs (or FlyBase IDs) are passed. Alternatively, the user can provide its own mapping in form of a list. In this case arbitrary gene identifiers can be used. Please refer to the manual pages for exact information.

By default a pruned KEGG hierarchy is used in order to improve the prediction quality. More specifically, metabolic pathways are not distinguished further, and the KEGG hierarchy for "Genetic Information Processing" is cut at the second level. That means we only distinguish between "Transcription", "Translation" and "Folding, Sorting and Degradation", but not between "RNA polymerase" and "Basal transcription factors". Please have a look at http://www.genome.jp/dbget-bin/get_htext?ko00001.keg+-f+F+C for a complete overview over the KEGG ontology. This behavior can also be changed, when the complete model is retrained, which is recommended to do regularly. For this purpose there exists the functions `retrain`. We refer to the manual pages for the exact usage here.

If for a given gene we suppose that it is related to signal transduction in some way, we can use the function `gene2pathway.signaltrans` in order to predict the exact signaling pathway or even the signaling pathway connected component. The latter is, however, only possible for those pathways, where there exists enough mapping genes. Again this behavior can be changed by retraining the model using the function `retrain.signaltrans`. The motivation for not including certain connected components into the hierarchy was that the number of mapping genes was below a certain cutoff (here: 10), which may spoil prediction performance.

Below we show an example analysis with *gene2pathway* for two genes: For

the first we predict the branch in the KEGG hierarchy, and for the second the connected component in a specific signaling pathway. The connected component may then be visualized using the function `color.pathway.by.elements`. This function uses the KEGG SOAP service and will return a URL with a gif-file.

```
> library(gene2pathway)
```

```
> gene2pathway("FBgn0030327", flyBase=TRUE, organism="dme")
```

```
> pred.comp = gene2pathway.signaltrans("43856", organism="dme") # prediction of the
```

```
Loading classification model ...
Using KEGG information from SOAP service ...
Mapping to signal transduction pathway components via KEGG database ...
xmlns: URI SOAP/KEGG is not absolute
xmlns: URI SOAP/KEGG is not absolute
xmlns: URI SOAP/KEGG is not absolute
xmlns: URI SOAP/KEGG is not absolute
xmlns: URI SOAP/KEGG is not absolute
xmlns: URI SOAP/KEGG is not absolute
---> Information found for  0 genes
done.
1  genes to predict
Retrieving information from InterPro database for organism ' dme ' via Ensembl ...
done: Information found for  1 out of  1  genes
Model prediction possible for  1  genes ..........done
Preparing output
finished
```

```
> pred.comp
```

```
$gene2Path
$gene2Path$`43856`
[1] "Hedgehog signaling pathway"


$scores
$scores$`43856`
[1] 0.9318182
```

```
$byKEGG
43856
FALSE

$elemIDs
$elemIDs$`43856`
list()


$votes
$votes$`43856`
[1] 0.9090909
```

It is important to mention that a separate prediction model for each organism is needed. Due to space restrictions of the R package we have only included a model for drosophila ("dme"). Other models can be created using the functions `retrain` and `retrain.signaltrans`, as mentioned above. In principle one could also train a model for one organism and apply it to another one. This can be achieved by setting "organism" according to the model, one would like to use.

It is possible to have a detailed look at each model in order to know, for example, which KEGG hierarchy levels can be predicted. In the following we first load the bagged model for drosophila and then explore it a little bit. It is important to know that here the bagged model consists of 11 individual models of class "model".

```
> data(classificationModel_dme)     # load the bagged model
> modelKEGG[[1]]$allpathways        # all employed KEGG hierarchy levels

 [1] "Metabolism"
 [2] "Genetic Information Processing"
 [3] "Environmental Information Processing"
 [4] "Cellular Processes"
 [5] "Organismal Systems"
 [6] "Translation"
 [7] "Folding, Sorting and Degradation"
 [8] "Replication and Repair"
 [9] "Signal Transduction"
[10] "Signaling Molecules and Interaction"
[11] "Endocrine System"
[12] "Neuroactive ligand-receptor interaction"
[13] "Lysosome"
```

```
> head(modelKEGG[[1]]$used_domains[[10]]) # Which InterPro domains are used by the
```

```
[1] "IPR005809" "IPR005810" "IPR003781" "IPR010097" "IPR001557" "IPR008209"
```

```
> head(modelKEGG[[2]]$W)                  # How are input code vectors weighted?
```

```
                          Metabolism       Genetic Information Processing
                           2.02196695                           0.11523449
Environmental Information Processing                  Cellular Processes
                           0.12492978                           1.80736892
                  Organismal Systems                          Translation
                           0.13770039                           0.09771773
```

References

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

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

[3] D. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler. The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985 – 1988, 2004.

[4] D. Maglott, J. Ostell, K. Pruitt, and T. Tatusova. Entrez: Gene-Centered Informaiton at NCBI. *Nucleic Acids Res.*, 35:D26 – D31, 2007.

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

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

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

[8] V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci.*, 98:5116–5121, 2001.

Session Information

The version number of R and packages loaded for generating the vignette were:

- R version 2.15.1 (2012-06-22), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=C`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: AnnotationDbi 1.20.0, Biobase 2.18.0, BiocGenerics 0.4.0, DBI 0.2-5, KEGGSOAP 1.32.0, RBGL 1.34.0, RSQLite 0.11.2, biomaRt 2.14.0, gene2pathway 2.12.0, graph 1.36.0, hgu95av2.db 2.8.0, keggorthology 2.10.0, kernlab 0.9-14, org.Dm.eg.db 2.8.0, org.Hs.eg.db 2.8.0
- Loaded via a namespace (and not attached): IRanges 1.16.2, RCurl 1.95-0.1.2, SSOAP 0.8-0, XML 3.95-0.1, XMLSchema 0.7-2, codetools 0.2-8, parallel 2.15.1, stats4 2.15.1, tools 2.15.1