Hunting for gene function: Using phenotype data mining as a large-scale discovery tool

In: Function prediction using combined methods

Philip Groth[1, 2] *, Bertram Weiss[1], Ulf Leser[2]
[1]Research Laboratories of Bayer Schering Pharma AG, Berlin, Germany
[2]Knowledge Management in Bioinformatics, Humboldt University, Berlin, Germany
*To whom correspondence should be addressed: groth@informatik.hu-berlin.de

1. INTRODUCTION

The importance of gene function information lies in understanding how specific genotype alterations may contribute to the development of certain diseases. In an experimental setting, gene function can be inferred, e.g., by looking at particular phenotypes, and in particular diseases derived from high-throughput methods, such as RNA interference or gene knock-out studies [1, 2]. The systematic use of these types of data has only recently given rise to the new field of research named 'Comparative Phenomics', i.e. analyzing phenotype data for gene function annotation.

In 2005, we have introduced PhenomicDB [3, 4] – a cross-species genotype-phenotype database which is one of the largest available collections of phenotype descriptions across species from a broad variety of sources, including the clinical descriptions of diseases, the characterization of naturally occurring mutants, as well as experimentally generated mutants from RNAi screens or gene knock-out experiments.

We present the results of a study where we analyzed these data in a Comparative Phenomics approach, using text-clustering to group together similar phenotypes and analyzing the induced gene clusters, therein creating a novel approach to cluster genes [5]. To validate the biological usefulness of the created 'phenoclusters' (clusters of genes with similar phenotypes), we examined the relatedness of the genes in a cluster using several independent measures. We found that 'phenoclusters' are highly enriched in terms of coherence of their functional annotation from the Gene Ontology [6]. We use this enrichment to predict new functions for genes. Using cross-validation, we found that this method yields a precision of 76.2% for predicting more than 150 GO-terms.

We conclude that the intrinsic nature of phenotypes to visibly reflect genetic activities underlines their usefulness for systematic analysis on a large scale, offering many new possibilities for inferring functional annotations for genes.

In our talk, we will present this study showing that text mining of phenotypes plays an important role in this process and will furthermore provide insights into newly developed even more powerful methods to make use of large-scale cross-species phenotype data for gene function prediction.

2. METHODOLOGY

We obtained textual description of phenotypes and a reference to their associated gene from the PhenomicDB database. For text mining purposes, the descriptions had to be properly adapted and prepared (stemming, etc.). We use the working term 'phenodoc' in the following to refer to this adjusted form of phenotype description and use 'phenocluster' to refer to a cluster of 'phenodocs'. We clustered the resulting 39,610 'phenodocs' associated to 15,426 genes from 7 different species into 1,000 clusters based on the cosine distance between 'phenodocs' using the k-means algorithm on a vectorized representation of the documents. We studied the resulting groups from a number of perspectives to assess whether or not the grouping itself is biologically reasonable. Finally, we predicted gene function within each cluster and evaluated this method using cross validation.

To estimate precision of our approach, we considered all clusters with at least three associated genes. We assume GO-terms to be descriptive for a cluster if common to at least 50% of its members and filtered clusters with no descriptive terms. We randomly partitioned each of the resulting clusters into a training set of 90% of its genes and a test set of at least one gene or 10% of genes, respectively. The descriptive terms of the training set were 'predicted' as new annotations to all genes in the test set of the same cluster. We then compared these predictions to the real annotation of the test genes to judge prediction correctness. This procedure was repeated 200 times (with different training / test sets) and averaged precision of the suggested terms was computed. As empirical threshold (p-value smaller than 0.05), we used randomly populated gene groups of equal size.

3. RESULTS

For making predictions from our 1,000 'phenoclusters', we defined a number of filters for selecting clusters, based on criteria such as the number of genes they contain, the number of available annotations, and their scores for in-group annotation coherence and in-group connectedness. We calculated precision and recall of function prediction in all clusters selected by different combinations of those filters.

The number of clusters was reduced to 856 by filtering all clusters containing less than 3 genes and reduced once more to 295 by filtering all clusters without any descriptive GO-terms (i.e. any Biological Process terms assigned to at least 50% of cluster members). We predicted 345 distinct GO-terms from the Biological Process subtree at a precision of 67.9%, averaged over all selected clusters. Relaxing the criteria for GO-term identity, now allowing for a single deviation towards the root (i.e., a predicted term is considered correct if it exactly matches a removed term or if it matches a parent of the removed term) resulted in an average 75.6% precision (191 unique terms for 2,686 genes in 279 groups). Allowing one more step towards the root, we predicted 151 unique terms with 76.2% precision.

4. CONCLUSIONS

We show that a great deal of the heterogeneous nature of phenotype data can be overcome by using text-mining. Within one framework, we systematically analyzed textual descriptions of clinical diseases, naturally occurring mutants, RNAi screens, gene knock-out experiments, and many others. Using clustering, a reasonable fraction of the associated genes can be grouped into biologically meaningful categories. Grouping genes based on certain properties is a powerful tool that has often been applied for function prediction before, using criteria such as participation in the same pathway [7, 8], participation in PPi cliques [9], or mentioning in the same Medline abstracts [10] – but not on phenotypes. We believe (and have shown) that this is an important new approach, as phenotypes, e.g. in contrast to interaction data, yield more information on the high diversity of biological meaning that is innate to any gene. It is, in fact, the intrinsic nature of phenotypes to visibly reflect genetic activity. Thus, phenotype data has the potential to be more useful for functional studies than most other types of data.

5. REFERENCES

1. Hannon GJ: **RNA interference**. *Nature* 2002, **418**(6894):244-251.
2. Shi Y: **Mammalian RNAi for the masses**. *Trends in Genetics* 2003, **19**(1):9-12.
3. Groth P, Pavlova N, Kalev I, Tonov S, Georgiev G, Pohlenz HD, Weiss B: **PhenomicDB: a new cross-species genotype/phenotype resource**. *Nucleic Acids Research* 2007, **35**(Database issue):D696-699.
4. Kahraman A, Avramov A, Nashev LG, Popov D, Ternes R, Pohlenz HD, Weiss B: **PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics**. *Bioinformatics* 2005, **21**(3):418-420.
5. Groth P, Weiss B, Pohlenz HD, Leser U: **Mining phenotypes for gene function prediction.** *BMC Bioinformatics* 2008, **9**(1):136.
6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25**(1):25-29.
7. Huynen MA, Snel B, von Mering C, Bork P: **Function prediction and protein networks**. *Current Opinion in Cell Biology* 2003, **15**(2):191-198.
8. Jaeger S, Leser U: **High-Precision Function Prediction using Conserved Interactions**. In: *German Conference on Bioinformatics (GCB).* Potsdam, Germany; 2007.
9. Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(21):12123-12128.
10. Raychaudhuri S, Chang JT, Sutphin PD, Altman RB: **Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature**. *Genome Research* 2002, **12**(1):203-214.