Ontologies improve cross-species phenotype analysis

Philip Groth[1,*], Bertram Weiss[1] and Ulf Leser[2]

[1] Research Laboratories of Bayer Schering Pharma AG, 13442 Berlin, Germany
[2] Knowledge Management in Bioinformatics, Humboldt-University of Berlin, 12489 Berlin, Germany

ABSTRACT

As phenotype data analysis has become an important component of functional genomics, many methods for analyzing these data have been published in the recent past. For example, RNA interference (RNAi) in mice has significantly improved our understanding of gene regulation, even for human disease. However, as phenotypes are obtained through species-specific experiments, they are usually described with unstructured text using very specific terminology. Such descriptions lack a common vocabulary, i.e. a universal phenotype ontology. This heterogeneity considerably hampers the analysis of phenotypes across species.

We have shown recently that clustering the free-text descriptions of phenotypes (phenoclustering) induces a clustering of genes that leads to gene function prediction with high precision. However, the method suffered from the fact that phenotypes of different species are rarely clustered together, impeding the power of comparative phenomics.

Here, we describe how matching terms from textual phenotype descriptions to biomedical ontologies like the Medical Subject Headings (MeSH), the Gene Ontology (GO) and the Mammalian Phenotype Ontology (MP) can significantly help to overcome heterogeneity in species-specific terminology. Using ontologies, the percentage of mixed-species clusters could be raised from 14.8% to 25.0%. Also, both precision and recall of gene function prediction could be improved. This shows that ontologies lead to a shift from a mere methodical (i.e. descriptive) towards a functional homogeneity of vocabulary.

1 INTRODUCTION

For over a century, phenotypes have been studied with regard to health and disease in order to reveal genotype-phenotype relationships. In the past few years, especially with the advent of RNA interference (Tuschl and Borkhardt, 2002), phenotype analysis has become an acknowledged and widely used tool for functional genomics. Mostly, these data are interpreted for a gene-by-gene functional annotation, since genotype-phenotype relationships are the most immediate results of such screens. However, this type of evaluation is also commonly applied using model organisms, e.g. mice, for generating hypotheses in humans with the goal to uncover the involvement of genes in diseases which may lead to novel therapeutic approaches.

The number of available methods to generate such data has grown significantly, leading to the availability of large amounts of phenotype-related data, scattered over a multitude of heterogeneous data sources that are mostly dedicated to single species or diseases (see the review by Groth and Weiss (Groth and Weiss, 2006) for an overview). Several ad-hoc integration methods were developed for pursuing meta-analyses of phenotypes across species or studies, thus gaining insights into the genetic origins of health and disease (Lussier and Li, 2004). Also, systematic approaches to phenotype integration have emerged, such as PhenomicDB (Groth, et al., 2007).

However, it is an evident obstacle of such cross-species phenotype analyses that many phenotype descriptions are highly heterogeneous (in concept and in content). Also, researchers often use domain-specific terminology. In our recent comparative phenomics study (Groth, et al., 2008), we could show that using the common denominator of most phenotypes, i.e. their textual descriptions, can overcome in part the diversity of these data, and that a clustering of phenotype descriptions (phenoclustering) can be used to predict gene function from the groups of associated genes with high precision. Still, in our cross-species setting, almost 90% of the phenotypes were grouped into single-species clusters due to the highly species-specific terminology used to describing phenotypes. One of our goals in cross-species clustering was to generate cross-species clusters. Such clusters should be homogeneous with regard to the biological function of the phenotypes (and the responsible genes) as compared to a clustering in which species-specific clusters merely indicate a common descriptive vocabulary due to terminology usage within a specific community. The use of such species-specific terminology can only be partly justified. Many observations in phenotypes are valid across species, e.g. in regard to survival, fertility, motility, growth, etc. and any discrepancies of similar terms will lead to artifacts, i.e. phenotypes being clustered separately which in fact describe a very similar observation but not in the same organism – or laboratory. Here, ontologies can help.

Recently, Washington et al. have shown that rigorous application of ontology-based phenotype descriptions to 11 gene-linked human diseases from the Online Mendelian

* To whom correspondence should be addressed.

Inheritance in Man (OMIM) (McKusick, 2007) can be used as a means to identify gene candidates of human diseases by utilizing similarities to phenotypes from animal models, showing that ontologies applied to cross-species phenotypes can significantly improve results in comparative phenotype studies (Washington, et al., 2009).

In this paper, we describe an improved approach to phenotype clustering built on the usage of ontologies. Therein, we first identify terms of phenotype-related ontologies, such as MeSH (Nelson, et al., 2004), GO (GeneOntologyConsortium, 2010) and MP (Smith, et al., 2005). Occurrences of such terms are overweighted when computing document similarity to increase the number of mixed-species clusters. We show that with the use of biomedical ontologies, the percentage of mixed-species clusters raised from 14.8% to 25.0%. Also, the precision for gene function prediction, i.e. the percentage of predicted terms that are correct, could be raised from 65.5% to 70.9%. At the same time, recall, i.e. the percentage of terms that should have been predicted, could be raised from 20.0% to 21.3%. We hypothesize that a universal cross-species ontology, such as it has been finally announced by Mungall et al. (Mungall, et al., 2010), could yield even better results.

2 METHODS

First, phenotype descriptions are extracted from the cross-species genotype/phenotype database PhenomicDB (Groth, et al., 2007). Herein, 347,689 phenotypes of species ranging from yeast to human have been assembled from their original repositories, e.g. OMIM (Hamosh, et al., 2005) for *H. sapiens*, the Mouse Genome Database (Bult, et al., 2008) for *M. musculus*, WormBase (Stein, et al., 2001) for *C. elegans*, FlyBase (Drysdale, 2008) for *D. melanogaster*, the Comprehensive Yeast Genome Database (Guldener, et al., 2005) for *S. cerevisiae*, the Zebrafish Information Network (Sprague, et al., 2008) for *D. rerio*, DictyBase (Chisholm, et al., 2006) for *D. discoideum* and the MIPS *Arabidopsis thaliana* database (Schoof, et al., 2004).

These descriptions are converted into feature vectors, in which each feature represents a token of a phenotype description. Features are weighted according to their importance within the description and within the entire set of phenotype descriptions, applying the term frequency-inverse document frequency (TFIDF) weighting. The multiplication of the term frequency by inverse document frequency (=TFIDF) 'discounts frequent words with little discriminating power' (Steinbach, et al., 2000). We then artificially over-weighted those features believed to have a high importance on the study goal, i.e., comparison of phenotypes. To this end, we extracted 224,387 ontology terms and 446,449 synonyms to these terms from GO, MP and MeSH. We searched occurrences of those in the set of 38,656 relevant phenotype descriptions by exact matching. Each ontology term with such a match in the phenotype description was ten-fold over-weighted with regard to its original score (TFx10).

The resulting feature vectors were clustered using CLUTO version 2.1.1 (Zhao and Karypis, 2005). The clusters were used for predicting gene functions using the method described in our study (Groth, et al., 2008). Evaluation of accuracy of predictions was carried out using cross-validation. We computed results with and without overweighting and compared them according to the percentage of cross-species clusters and the number and precision of predicted gene functions.

3 RESULTS

The percentages of terms from the different ontologies that could be matched to phenotype descriptions ('hits') varied between 44.66% and 0.55% (see 'ontology usage' in Table 1).

The most terms (in absolute numbers) that could be matched came from the MeSH Descriptors vocabulary. However, since this a very large dictionary, the coverage of matches was small (only 7.50%). The largest ontology usage was observed with the MeSH Qualifiers vocabulary, a small vocabulary (only 318 terms and synonyms) of rather broad concepts. However, as almost 45% of the phenotypes used here derived from either human or mouse, it was surprising that only slightly more than a third (37.66%) of all terms from MP could be found in at least one phenotype description. The reason for this may be that most annotators commonly use leaf terms for annotations, thus (possibly unintentionally) disregarding many terms in the upper levels of the ontology; as a consequence, those terms do not show up in our analysis.

Table 1. Result of matching terms from controlled vocabularies to phenotype descriptions.

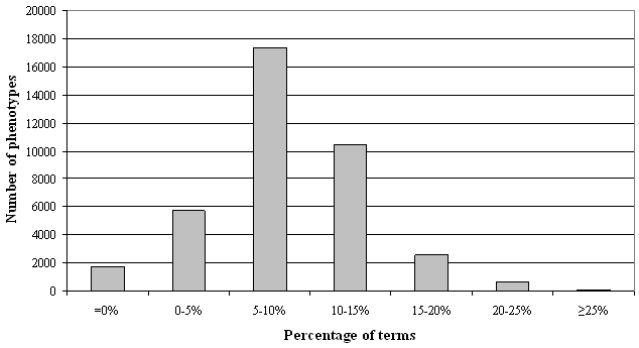
	MP	MeSH Supplementary Concept Records	MeSH Descriptors	MeSH Qualifiers	GO
Terms	5,606	170,663	24,357	83	23,678
Synonyms	1,980	268,792	152,166	235	23,276
Hits	2,857	2,406	13,242	142	1,792
Ontology usage	37.66%	0.55%	7.50%	44.66%	3.82%
Weighted terms	1,093	1,966	8,412	67	1,642
Mean term length	24.74	27.41	17.57	12.63	38.46

Terms: Number of terms in the vocabulary. **Synonyms:** Number of synonyms to terms in the vocabulary. **Hits:** Number of unique terms or synonyms found in at least one phenotype description. **Ontology usage:** Percentage of hits in all terms and synonyms. **Weighted terms:** Number of unique concept hits, where hits to a synonym are assigned as hits to their corresponding term.

It is noteworthy that the phenotypes can be generally regarded as fairly well 'annotated' with ontology terms (see Figure 1). In almost half of the phenotype descriptions

(17,351 of 38,656) 5-10% of the description consists of ontology terms. Only to 5% (1,727 of 38,656) of the phenotype descriptions, no ontology term could be matched at all.

Figure 1. The number of phenotypes and the number of ontology terms matching to words in the phenotype descriptions as percentage of the total number of words in the description.



After matching the ontology terms, phenotypes were processed as described above and features representing an ontology term were over-weighted 10-fold in comparison to all other features. In the next step, the feature vectors were clustered using k-means clustering with k=1,000 (other results not shown). The resulting groups of genes were evaluated in respect to the distribution of clusters in species and precision and recall of the derived functional predictions (see Table 2).

When looking at Table 2, the most obvious result is that the percentage of clusters with mixed species is much higher in the weighted scheme (25.0%) than in the unweighted scheme (14.8%). Thus, weighting helps to overcome to some extent the boundary set by the usage of species-specific vocabulary (e.g. screening methodology).

However, an increase from 4.6% to 6.7% can also be observed for human-specific clusters. This can most likely be explained by the fact that human disease descriptions from OMIM are by far the longest free-text descriptions. With these, an over-weighting of a few keywords will not compensate for the sheer number of vocabulary-specific mainly clinical descriptions.

Besides this, over-weighting has not only decreased the number of species-specific clusters (improving the clusters in regard to common functional terminology). Also, the precision for gene function prediction could be raised to 70.9% and recall could be raised to 21.3%. In consequence, the F-measure (the harmonic mean of precision and recall) from the over-weighted clustering (0.3271) is higher than that from the unweighted clustering (0.3063).

Table 2. Results of phenotype clustering using the TFIDF weighting. Here, the clustering of phenotypes with 10-fold over-weighted ontology terms is compared to clustering without over-weighting (k = 1,000).

TFIDF		**unweighted**	**weighted**
Distribution of clusters according to species	Mm	15.8%	14.3%
	Sc	1.6%	1.3%
	Dr	1.5%	1.3%
	Ce	11.2%	9.3%
	DM	46.8%	40.1%
	Hs	4.6%	6.7%
	Dd	3.7%	2.0%
	Mixed	14.8%	25.0%
Gene function prediction	# valid clusters	234	201
	# predicted terms	417	336
	# genes	3,857	3,537
	Recall	19.99%	21.26%
	Precision	65.48%	70.86%
	F-measure	0.3063	0.3271

Comparison of the results using normal and over-weighted TFIDF scores. Results are shown for the distribution of clusters in species, functional predictions of GO-terms (counting only unique and exactly matching GO-terms as correctly predicted). Ce = *Caenorhabditis elegans*; Dm = *Drosophila melanogaster*; Hs = *Homo sapiens*; Mm = *Mus musculus*; Sc = *Saccharomyces cerevisiae*; Dr = *Danio rerio*; Dd = *Dictyostelium discoideum*

These figures clearly show that term overweighting has a positive effect on the functional coherence of the clusters. Since the total number of predicted terms does not rise, the increase in precision and recall is due to an increase in the number of true positive predictions, implicitly reducing number of false negatives and false positives. Such an increase in true positive predictions can only be attributed to the fact that there is more congruent functional annotation for each member within a cluster. This shows directly the usefulness of applying ontologies in such a setting.

On another note, this method yields a prediction precision of almost 71%, which is an outstanding competitive result compared to other state-of-the art function prediction methods (see e.g. the survey by Pandey et al. (Pandey, et al., 2006) for precision values from other methods).

4 DISCUSSION

Cross-species phenotype clustering is a direct extension of other functional genomics methods with the goal to infer results for model organisms for hypothesis generation in human disease. However, the use of free text as a common denominator within phenotype descriptions needs to overcome the problems of species-specific vocabulary. Such vocabulary can be found in the terminology used to describe certain characteristics, but also by the descriptions of methodology of examination, which often is different in each of the species. We have shown that the problem can be addressed by applying a term-weighting that over-weights domain-specific terminology, e.g. using terms from

ontologies like MeSH or the MP, thus implicitly down-weighting highly species- or method-specific terms. With a ten-fold over-weighting, it was possible to push the portion of mixed-species clusters by over 40% to almost one third of total clusters. Since this overweight was chosen arbitrarily, we expect that choosing a slightly lower or any higher factor may somewhat alter the numeric outcome, but not the general observation.

On the other hand, applying over-weighting still leaves many clusters species-specific, especially for *Homo sapiens*. This tendency of phenotypes to accumulate into species-specific clusters shows that the terminology used to describe a phenotype, especially for humans, depends on the way we are used to defining the many complex processes and diseases, e.g. in differing fields of medicine like oncology, neurology or gynecology. However, such a separation of vocabulary is only partly justified, as many phenotypic effects are highly similar across species, and must therefore be considered an artifact.

These issues have finally been recognized by a larger community, and first efforts towards a more general phenotype ontology are underway (Mungall, et al., 2010). We strongly believe that such a unified ontology will open the door to more powerful ways of analyzing phenotypes, in the same manner as the establishment of GO has opened the door for many new approaches to analyzing biological knowledge in genes.

REFERENCES

Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E. and Blake, J.A. (2008) The Mouse Genome Database (MGD): mouse biology and model systems, *Nucleic Acids Res*, **36**, D724-728.

Chisholm, R.L., Gaudet, P., Just, E.M., Pilcher, K.E., Fey, P., Merchant, S.N. and Kibbe, W.A. (2006) dictyBase, the model organism database for Dictyostelium discoideum, *Nucleic Acids Res*, **34**, D423-427.

Drysdale, R. (2008) FlyBase : a database for the Drosophila research community, *Methods Mol Biol*, **420**, 45-59.

GeneOntologyConsortium (2010) The Gene Ontology in 2010: extensions and refinements, *Nucleic Acids Res*, **38**, D331-335.

Groth, P., Pavlova, N., Kalev, I., Tonov, S., Georgiev, G., Pohlenz, H.D. and Weiss, B. (2007) PhenomicDB: a new cross-species genotype/phenotype resource, *Nucleic Acids Res*, **35**, D696-699.

Groth, P. and Weiss, B. (2006) Phenotype Data: A Neglected Resource in Biomedical Research?, *Current Bioinformatics*, **1**, 347-358.

Groth, P., Weiss, B., Pohlenz, H.D. and Leser, U. (2008) Mining phenotypes for gene function prediction, *BMC Bioinformatics*, **9**, 136.

Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N., van Helden, J., Lemer, C., Richelles, J., Wodak, S.J., Garcia-Martinez, J., Perez-Ortin, J.E., Michael, H., Kaps, A., Talla, E., Dujon, B., Andre, B., Souciet, J.L., De Montigny, J., Bon, E., Gaillardin, C. and Mewes, H.W. (2005) CYGD: the Comprehensive Yeast Genome Database, *Nucleic Acids Res*, **33**, D364-368.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Res*, **33 Database Issue**, D514-517.

Lussier, Y.A. and Li, J. (2004) Terminological mapping for high throughput comparative biology of phenotypes, *Pac Symp Biocomput*, 202-213.

McKusick, V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM, *Am J Hum Genet*, **80**, 588-604.

Mungall, C.J., Gkoutos, G.V., Smith, C.L., Haendel, M.A., Lewis, S.E. and Ashburner, M. (2010) Integrating phenotype ontologies across multiple species, *Genome Biol*, **11**, R2.

Nelson, S.J., Schopen, M., Savage, A.G., Schulman, J.L. and Arluk, N. (2004) The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation. In Fieschi, M. (ed), *11th World Congress on Medical Informatics*. Amsterdam: IOS Press, San Francisco, CA, 67-69.

Pandey, G., Kumar, V. and Steinbach, M. (2006) Computational Approaches for Protein Function Prediction: A Survey. *Technical Report*. Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN.

Schoof, H., Ernst, R., Nazarov, V., Pfeifer, L., Mewes, H.W. and Mayer, K.F. (2004) MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics, *Nucleic Acids Res*, **32**, D373-376.

Smith, C.L., Goldsmith, C.A. and Eppig, J.T. (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information, *Genome Biol*, **6**, R7.

Sprague, J., Bayraktaroglu, L., Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Haendel, M., Howe, D.G., Knight, J., Mani, P., Moxon, S.A., Pich, C., Ramachandran, S., Schaper, K., Segerdell, E., Shao, X., Singer, A., Song, P., Sprunger, B., Van Slyke, C.E. and Westerfield, M. (2008) The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes, *Nucleic Acids Res*, **36**, D768-772.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J. and Spieth, J. (2001) WormBase: network access to the genome and biology of Caenorhabditis elegans, *Nucleic Acids Res*, **29**, 82-86.

Steinbach, M., Karypis, G. and Kumar, V. (2000) A Comparison of Document Clustering Techniques. *KDD Workshop on Text Mining*.

Tuschl, T. and Borkhardt, A. (2002) Small interfering RNAs: a revolutionary tool for the analysis of gene function and gene therapy, *Mol Interv*, **2**, 158-167.

Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M. and Lewis, S.E. (2009) Linking human diseases to animal models using ontology-based phenotype annotation, *PLoS Biol*, **7**, e1000247.

Zhao, Y. and Karypis, G. (2005) Data clustering in life sciences, *Mol Biotechnol*, **31**, 55-80.