A translational medicine approach to orphan diseases

Robert Hoehndorf[1] and Georgios V. Gkoutos[1,2]

[1]*University of Cambridge,* [2]*University of Aberystwyth*

Correspondence: gg295@cam.ac.uk, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK

Abstract

Phenotypes are investigated in model organisms to understand and reveal the molecular mechanisms underlying disease. Their computational analysis has been greatly facilitated by the introduction of phenotype ontologies developed to capture and compare phenotypes within the context of a single species. We have recently developed a method to transform phenotype ontologies into a formal representation, combine phenotype ontologies with anatomy ontologies, and apply a measure of semantic similarity to construct the cross-species phenotype network *PhenomeNet*. PhenomeNet relies on the descriptions of diseases and disorder with clinical signs to identifying causative genes for human diseases based on experimental data from animal organisms. We describe our integration of the Orphanet clinical signs for rare and orphan disease with PhenomeNet, and demonstrate that our approach can identify candidate genes through the systematic comparison of experimentally derived phenotypes in mice with clinical signs associated with Orphanet disorder (0.798 area under ROC curve).

Introduction

Animal models provide a valuable tool for the investigation of gene functions and the study of human disease. In phenotype experiments, genetically modified organisms are created and the observable characteristics resulting from the modification are recorded. Large scale knockout projects that aim to record the phenotypes resulting from deactivation of every protein-coding gene in an organism have been completed for yeast [1] and are currently underway for the mouse model organism [2]. The systematic generation of phenotype information associated with genetic modifications has the potential to lead into novel insights regarding the molecular mechanisms underlying orphan diseases based on the phenotypic similarity between an experimentally recorded phenotype in a model organism and the clinical phenotype of the patient. In order to systematically analyze and compare clinical phenotypes and phenotypes in animal models, it is necessary to integrate phenotype descriptions across species.

Within model organism communities, *ontologies* are being employed to record the outcome of phenotype experiments. For example, the Mammalian Phenotype Ontology (MP) [3] is used to record normal and abnormal phenotypes resulting from mouse phenotype experiments. In the clinical research context, the Human Phenotype Ontology (HPO) [4] has been developed and is currently being applied to describe the phenotypes associated with disorders in the OMIM database [5]. The PATO framework [6] has been applied to formally define the terms in species-specific phenotype ontologies [7] and enable their integration across species. We have developed the PhenomeBLAST software tool that integrated phenotype ontologies and employs automated reasoning to provide alignments of phenotype terms across multiple species. This integration enables the direct comparison of phenotypes in different species, including the comparison of phenotypes in animal models and phenotypes of human diseases. The PhenomeNET approach [8] performed systematic pairwise comparisons between phenotypes in five species and human diseases, demonstrating that it is possible to automatically reveal the molecular origins of orphan disease by analyzing the phenotypes associated with mutations in animal organisms.

Here we describe an extension to this approach by integrating clinical signs associated with disorders from Orphanet [9] – a database dedicated to information on rare diseases and drugs. Our goal is to employ this resource for identifying possible candidate genes for orphan diseases based on their phenotypic similarity with phenotypic manifestations observed in mice. In order to quantitatively evaluate this ap-

proach, we report the area under the ROC curve (i.e., a plot of the true positive rate as a function of the false positive rate) for identifying known human gene-disease associations.

Methods

We have created a phenotypic representation of the disorders in OrphaNet based on the phenotype ontologies for human (HPO) and mouse (MP). To generate this representation, we used a combination of lexical and structural approaches. First, we use the Needleman-Wunsch algorithm [10] to find the labels and synonyms of phenotype terms in the HPO and MP that are lexically most similar to the labels of clinical signs in OrphaNet and assign these MP or HPO classes as *equivalent* to the clinical sign in OrphaNet. Second, we use the taxonomic structure of clinical signs in OrphaNet and identify a *super-class* in HPO or MP for clinical signs. As a result, we can associate 2,507 disorders from OrphaNet with 52,002 phenotype terms from HPO and 11,674 phenotype terms from MP.

Results

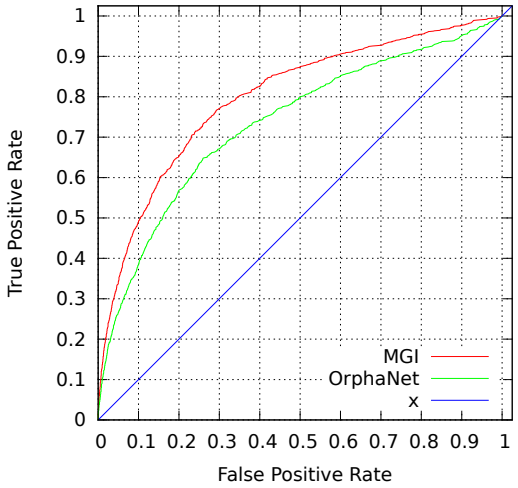

Figure 1:

We incorporate the OrphaNet phenotypes in PhenomeNET in order to compare phenotypes recorded in mice with human disease phenotypes and rank diseases based on their phenotypic similarity to the observed mouse phenotype. We then compare these ranking against known gene–disease associations taken from the Mouse Genome Informatics (MGI) database and against OrphaNet's gene-disease associations. To evaluate and quantify our approach, we use an analysis of the receiver operating characteristic (ROC) curve. A ROC curve can be used to visualize the performance of a classifier and plots the true positive rate of the classifier as a function of the false positive rate. The area under the ROC curve (AUC) is a quantitative measure of the classifier's performance. Using OrphaNet's gene–disease associations as positive instances and all remaining gene–disease pairs as negative, we achieve an AUC of 0.734, while we achieve an AUC of 0.798 using MGI's gene–disease associations as positive instances.

PhenomeNET and its associated tools and resources are freely available on `http://phenomeblast.googlecode.com`. The similarity between animal models and diseases can be explored using our PhenomeBrowser webserver at `http://phenomebrowser.net`.

Acknowledgements

Funding for RH was provided by the European Commission's 7th Framework Programme, RICORDO project, grant number 248502. Funding for GVG was provided by the NIH (grant number R01 HG004838-02).

References

[1] Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, et al. Systematic screen for human disease genes in yeast. Nature Genetics. 2002;31(4):400–404. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12134146.

[2] Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, et al. A conditional knockout resource for the genome-wide study of mouse gene function. Nature. 2011 Jun;474(7351):337–342. Available from: http://dx.doi.org/10.1038/nature10163.

[3] Smith CL, Goldsmith CAW, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biology. 2004;6(1):R7.

[4] Robinson PN, Koehler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. American journal of human genetics. 2008;83(5):610–615. Available from: http://dx.doi.org/10.1016/j.ajhg.2008.09.017.

[5] Amberger J, Bocchini C, Hamosh A. A new face and new challenges for online mendelian inheritance in man (OMIM). Hum Mutat. 2011;32:564–567.

[6] Gkoutos GV, Green EC, Mallon AMM, Hancock JM, Davidson D. Using ontologies to describe mouse phenotypes. Genome biology. 2005;6(1). Available from: http://dx.doi.org/10.1186/gb-2004-6-1-r8.

[7] Mungall C, Gkoutos G, Smith C, Haendel M, Lewis S, Ashburner M. Integrating phenotype ontologies across multiple species. Genome Biology. 2010;11(1):R2+. Available from: http://dx.doi.org/10.1186/gb-2010-11-1-r2.

[8] Hoehndorf R, Schofield PN, Gkoutos GV. PhenomeNET: a whole-phenome approach to disease gene discovery. Nucleic Acids Research. 2011;39(18):e119. Available from: http://nar.oxfordjournals.org/content/39/18/e119.

[9] Weinreich SS, Mangon R, Sikkens JJ, Teeuw MC M E amd Cornel. Orphanet: a European database for rare diseases. Ned Tijdschr Geneeskd. 2008 Mar;9(152):518–9.

[10] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970;48(3):443 – 453. Available from: http://www.sciencedirect.com/science/article/B6WK7-4DN8W3K-7X/2/0d99b8007b44cca2d08a031a445276e1.