PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics

Abdullah Kahraman1, Andrey Avramov², Lyubomir G. Nashev², Dimitar Popov², Rainer Ternes³, Hans-Dieter Pohlenz³ and Bertram Weiss³,*

1Department of Bioinformatics, University of Applied Science Giessen, 35596 Giessen, Germany, ²Metalife AG, Im Metapark 1, 79297 Winden, Germany and ³Research Laboratories, Schering AG, 13342 Berlin, Germany

Received on June 25, 2004; revised on August 11, 2004; accepted on August 30, 2004
Advance Access publication September 16, 2004

ABSTRACT

Summary: We have created PhenomicDB, a multi-species genotype/phenotype database by merging public genotype/phenotype data from a wide range of model organisms and Homo sapiens. Until now these data were available in distinct organism-specific databases (e.g. WormBase, OMIM, FlyBase and MGI). We compiled this wealth of data into a single integrated resource by coarse-grained semantic mapping of the phenotypic data fields, by including common gene indices (NCBI Gene), and by the use of associated orthology relationships. With its use-case-oriented user interface, PhenomicDB allows scientists to compare and browse known phenotypes for a given gene or a set of genes from different organisms simultaneously.

Availability: PhenomicDB has been implemented at Schering AG as described below. A PhenomicDB implementation differing in some technical details has been set up for the public at Metalife AG http://www.phenomicDB.de

Contact: bertram.weiss@schering.de

Supplementary information: database model, semantic mapping table.

MOTIVATION AND CONCEPT

More and more phenotypic data are being generated for both model and non-model organisms. New technologies such as RNAi now make genome-wide knock-down studies feasible and have already been applied in a high-throughput manner, for instance, to Homo sapiens (Berns et al., 2004; Fraser, 2004; Paddison et al., 2004). Valuable resources for phenotypic data are already available, but only for a given organism, for example, OMIM (Hamosh et al., 2002), WormBase (Harris et al., 2004), FlyBase (The FlyBase Consortium 2003) and MGD (Blake et al., 2003). Scientists have realized that there is an additional need to make phenotypic data from different organisms simultaneously searchable, visible and, most importantly, comparable (Lussier and Li, 2004).

Currently, research scientists looking for genes involved in a given disease have to search different phenotype databases. They need to figure out manually the orthology relationships among all genes concerned in order to understand the different genotypic effects on the phenotype of a certain gene in different organisms. These species-specific databases are scattered over the Internet and tailored to different objectives, and they store phenotypic data in different formats. Tedious handwork is therefore necessary to compare the phenotype of a gene in different organisms. A simple meta-search engine for these databases alone does not resolve this kind of problem, and this is exactly the functionality we were aiming to develop.

Currently, the different source databases all use different gene loci description systems (i.e. gene indices) and the orthology relationships are not always obvious, so that many important phenotypic relationships may be difficult to discover. As others (Claustres et al., 2002; Lussier and Li, 2004) have already stated, a common data model combining the data with a common gene index is required. Orthology data must be available and an use-case-oriented user interface should facilitate access to phenotypic data. Most data are available, but to the best of our knowledge, an integrative system, as described here, is not yet available.

In order to remedy to this situation, we set out to gather phenotype and genotype data from the different public resources and to map the data semantically into a single data model. To allow for direct comparison of phenotypes of orthologous genes from yeast to humans, we also uploaded these mapped data together with a gene index-like database [NCBI Gene (Pruitt et al., 2001)] and the associated orthology data [HomoloGene database (Wheeler et al., 2004)].
PhenomicDB is thought as a first step towards comparative phenomics and will improve our understanding of gene function by combining the knowledge about phenotypes from several organisms. PhenomicDB has to compromise between data depth as available in the source databases and data compatibility. It is not intended to compete with the much more dedicated primary source databases but tries to compensate its partial loss of depth by linking back to the primary sources. The basic functional concept of PhenomicDB is an integrated meta-search engine for phenotypes.

Users should be aware that comparing genotypes or even phenotypes between organisms as different as yeast and humans may involve serious scientific hurdles. Nevertheless, finding, for instance, that the phenotype of a given mouse gene is described as 'similar to psoriasis' and at the same time that the human orthologue has been described as a gene linked to skin defects can lead to novel and interesting hypotheses. Similarly, a gene involved in cancer in mammalian organisms could show a proliferation phenotype in a lower organism such as yeast, and this knowledge may lead to further insights.

IMPLEMENTATION

We implemented scripts to download phenotype/genotype data from public databases for *Mus musculus* (MGD), *H.sapiens* (OMIM), *Drosophila melanogaster* (FlyBase), *Caenorhabditis elegans* and *Caenorhabditis briggsae* (WormBase), *Arabidopsis thaliana* (MATDB) (Schoof et al., 2004) and *Saccharomyces cerevisiae* (CYGD) (Mewes et al., 2004). In addition, NCBI Gene and HomoloGene were downloaded. Whenever possible, the given source genotypes (here meant as equivalent of gene loci) were mapped to a NCBI Gene entry.

We performed coarse-grained semantic field mapping to bring the very heterogeneous phenotype data from different organisms into a common data model. Fields in the different source databases with the same content-type (e.g. containing the phenotype description) were identified and, irrespective of their original name there, uploaded in the corresponding PhenomicDB data field (e.g. ‘phenotype description’). Details of all semantic mappings and how we connected the data are shown in the semantic mapping table and the database schemata both available as Supplementary Information.

PhenomicDB was designed as a normalized relational Oracle v. 8.1.7.4 database. The database scheme comprises three parts: common data, genotype data and phenotype data. The common part is used for information shared between genotypes as well as phenotypes, e.g. name, symbol, organism or literature references. The genotype-specific part contains specific genotype information (e.g. gene description, chromosomal location, gene ontology, etc.). Each genotype entry can relate to a NCBI Gene identifier in order to bin identical genes that are represented by, for example, different transcript identifiers in the source databases. The use of NCBI Gene identifiers is a prerequisite for making use of the orthology relationships uploaded from the HomoloGene database. They are captured as pairs of orthologous NCBI Gene identifiers (as determined by HomoloGene). The phenotype-specific part stores the free-text phenotype descriptions, and the data that describe the underlying experiments (e.g. mutagenesis, RNAi and k.o. mice). Associated phenotype keywords or catalogue terms are stored as well. The genotype and phenotype parts are treated separately in our database. The connection between the two parts is implemented by genotype and phenotype id-mapping. Owing to the nature of the Oracle RDBMS, all knowledge containing fields are searchable. For performance reasons, those search functions which are accessible through the user interface are supported by interMedia text indices (Oracle).

User is presented with a search interface that allows free text search with the ability to restrict the search to certain fields (e.g. gene name, organism, etc.) or to show only genotypes, if they have a phenotype associated or vice versa. The result is a list of genotypes with their associated phenotypes. From the result list, user can select genotypes/phenotypes of further interest and expand these with their orthologues and the phenotypes of their orthologues. This allows direct and simultaneous comparison of all available phenotypes for a certain gene in all available organisms. Each result is hyperlinked to a dedicated genotype/phenotype report and directly to the source database. Genotype/phenotype reports contain the associated data about the genotype (usually a gene) and the associated phenotype(s) including phenotype experiments.

DISCUSSION AND OUTLOOK

Data-integrative approaches over a range of organisms decrease conceptual accuracy or eliminate data: a genotype in our database can mean a gene, a mutated gene or a chromosomal region. Phenotype description ranges from the mere mention of ‘non-viability’ in yeast to the detailed characterization of a knockout mouse including all the experimental details. However, only these general concepts allow for data integration. The detailed and extensive description of the data or dedicated mining tools, e.g. PhenoBlast (Gunsalus et al., 2004) should stay within the realm of the organism-specific source databases. Others (Lussier and Li, 2004) have started with the integration of phenotypic notation and terminology over several species or have proposed common semantics for genome-wide phenotype databases (Claustres et al., 2002). They have also discussed in more detail the associated difficulties of integrating phenotypic terminology that differs significantly between each organism-specific research community. We adopted a practical approach clearly intended to allow for high-level data integration and easy integration of new upcoming data. The data content will be updated every 8 weeks. PhenomicDB now has to prove that the method...
of integration applied here can add value to the scientific exploitation of phenome data.

Most of the valuable phenotypic data reside in the public literature not captured in databases. Effective text mining is needed to gather these data as well. A prerequisite for text mining, however, is the availability of specified thesauri, catalogues and validated terms. Those are not yet available for phenotypic data (Gunsalus et al., 2004). First steps are underway (Lussier and Li, 2004) and PhenomicDB could be used as a resource to extract such phenotype-specific vocabulary. We have started compiling thesauri from PhenomicDB to use them for the extraction of phenotypic data from literature by text mining.

ACKNOWLEDGEMENTS

We thank Dr Bernard Haendler (Schering AG) for useful discussion of the manuscript, and Dr Stephan Brock (Metalife) and Prof. Michael Schoenemann (Metalife) for their continuous support.

REFERENCES


