33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

PhenOMIM: An OMIM-Based Secondary Database Purported for Phenotypic Comparison

Han J. W. van Triest (*Student Member, IEEE*), Danqi Chen, Xinglai Ji, Shouliang Qi, Jesse Li-Ling

***Abstract* — Phenotypic comparison may provide crucial information for obtaining insights into molecular interactions underlying various diseases. However, few attempts have been made to systematically analyze the phenotypes of hereditary disorders, mainly owing to the poor quality of text descriptions and lack of a unified system of descriptors. Here we present a secondary database, PHENOMIM, for translating the phenotypic data obtained from the Online Mendelian Inheritance in Man (OMIM) database into a structured form. Moreover, a web interface has also been developed for visualizing the data and related information from the OMIM and PhenOMIM databases. The data is freely available online for reviewing and commenting purposes and can be found at http://faculty.neu.edu.cn/bmie/han/PhenOMIM/.**

I. Introduction

THE completion of the Human Genome Project [1] combined with the development of high-throughput technologies for genetic screening have made it possible to conduct genome-wide genotyping and accelerated elucidation of genetic contributions to various human disorders. Several large-scale computational approaches have been taken for genetic as well as protein analyses, which have provided valuable insights into gene functions and protein interactions. The massive amount of biological data that has been obtained in this fashion however, has signified complexity rather than simplicity with regard to the correlation between genotypes and phenotypes. In this work we aim to introduce a new tool for analyzing the phenotypic data, and to correlate symptoms with their underlying genetic causes.

A. Phenotypic comparison

In recent years, phenotypic analysis has been recognized to have great importance for the understanding of the molecular basis of various diseases [2, 3]. Attempts have been made for human diseases [4] and model organisms such as yeast [5], mouse [6], rat [7] and arabidopsis [8]. Phenomic databases have been constructed in both independent and combinatorial fashions [9, 10]. Large-scale efforts are still being undertaken to collect phenotypic data, in particular phenotypic features of disorders. To better delineate the genetic components and their contribution to Mendelian and complex traits, Human Phenome Project (HPP) has been proposed to systematically collect phenotypic information and for developing new approaches for analyzing the resulting data [11].

In recent years, a rich variety of computational techniques have been applied to text-mining on research literatures and biological databases that contain phenotypic data, which featured various levels of granularity, in different formats, and/or with different aims. Lussier et al. [12] developed *PhenoGO* database by applying *Natural Language Processing* (NLP) on scientific literature combined with the *Gene Ontology* database (GO) [13]. The database is rich in information on genotype-phenotype associations and relationships between genes, GO concepts and phenotypes. Van Driel et al. [14] used text-mining to extract human phenotypes from the *Online Mendelian Inheritance in Man* database (OMIM [15]) and correlated them with MeSH terms. Bajdik et al. [16] have developed *CGMIM* to identify genetically-associated cancers and candidate genes from the OMIM by means of automated text-mining. Masseroli et al. [17] also developed *GFINDer* to define a hierarchical structure for describing relationships between genetic diseases. Unfortunately, although the above approaches were able to encode phenotypes and generate large networks of genotype-phenotype correlations with high throughput and efficiency, they generally lacked accuracy, robustness and fell short of medical significance.

B. The OMIM database

Generally regarded as the ’phenotypic companion to the Human Genome Project (HGP)’, the OMIM database is a continuously updated catalog for human genes and genetic disorders. Despite its comprehensiveness, however, the unstructured record-based textual descriptions of OMIM data have made it less suitable for bioinformatics analysis. Computationally, it is difficult to extract coded phenotypic data from the OMIM database, since it is originally designed to be read by humans, and not by computers. In the OMIM dtabase, Clinical Synopses (CS) of various diseases, i.e.

Manuscript received April 15, 2011. This work was supported by the National Natural Science Foundation of China (No. 60574040, 81072899).

J. L-L. (corresponding author) is jointly appointed by the Department of Medical Genetics, China Medical University, Shenyang 110001, and Sino-Dutch Biomedical and Information Engineering School, Northeastern University, Shenyang 110003, China (Tel/Fax: 86-24-83689205; e-mails: liling@bmie.neu.edu.cn, jliling2000@gmail.com).

H. J. W. van T. and S. Qi are at Sino-Dutch Biomedical and Information Engineering School, Northeastern University, Shenyang 110003, China. e-mails: han@bmie.neu.edu.cn, qisl@bmie.neu.edu.cn.

D. C. is at the Department of Medical Genetics, China Medical University, Shenyang 110001, China. e-mail: orient_nancy@yahoo.com.cn.

X. J. is at the Laboratory for Conservation and Utilization of Bio-Resources & Key Laboratory for Microbial Resources of the Ministry of Education, Yunnan University, Kunming 650091, China. e-mail: rich.gee@gmail.com.

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

structured descriptions of disorders in terms of keywords, have not been presented in a uniform manner. Other problems include a considerable amount of spelling errors, and the usage of synonyms and overlapping concepts. Furthermore, some features seemed to be too complex to be described neatly (for instance abnormally shaped organs). To exploit the value of the data in the OMIM database, it seems necessary to manually check and standardize the data in order to enhance its comparability.

Due to the lack of a unified system of phenotypic descriptors, so far few attempts have been made to systematically analyze human phenotypic data. Considering the complexity of knowledge and enormous effort that is required to standardize the large amount of phenotypic data from the OMIM database, progress has to rely primarily on robust taxonomies of phenotypes and accurate clinical descriptions, for which participation of both clinicians as well as engineers, and novel tools in computation and informatics are required [3].

To encode and present phenotypic data in a controlled ontology has become a pending task for automating phenotypic comparison. Here we present a secondary database, PhenOMIM, which is purported to translate the Clinical Synopses (CS) derived from the OMIM database. The embedded data was manually checked and classified according to their anatomical relations. The resulting controlled vocabulary has been organized in an ontology. A web interface and a stand-alone program have been developed for visualization of the controlled vocabulary and the translated clinical features of particular disorders.

Besides the standardization of the terms and symptoms, another major difference between the structures of the original OMIM and new PhenOMIM ontologies lies in the fact that the OMIM tree only has a depth of up to 3 levels, whereas PhenOMIM has a depth of up to 7 levels. The purpose of this is to enhance the comparability of the dataset, as symptoms that share more categories in their paths can be assumed to be more similar. By introducing multiple extra levels, it becomes possible to introduce various degrees of compatibilities between symptoms.

II. Material and Methods

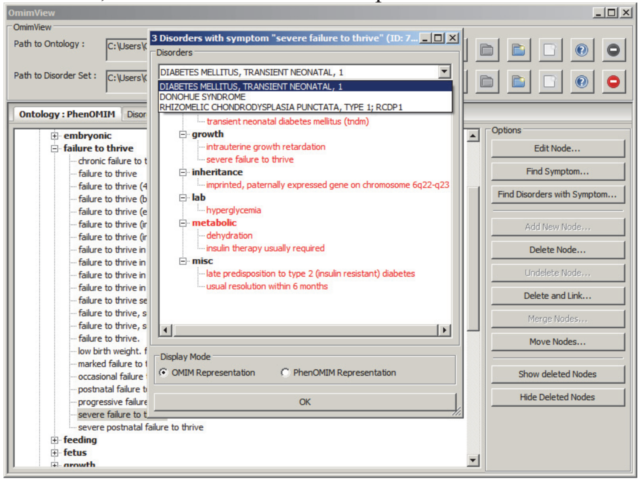
A. Materials

For creating the PhenOMIM database, we have used data retrieved from the OMIM database. All entries that contain a clinical synopsis were downloaded and stored in a plain text file. The latest dataset, downloaded on March 1, 2011, contained 4807 disorders, which have involved a total of 39,683 unique labels distributed over categories, sub-categories and symptoms.

B. Methods

In the original OMIM data, clinical features contained in the *Clinical Synopsis* (CS) fields are divided into categories and/or subcategories, e.g., "*organ system*", "*laboratory findings*", "*mode of inheritance*", in addition to detailed clinical features, i.e. symptoms.

A C++ program, named *OmimView* (see Fig. 1), has been created to parse these CS and to generate an ontology for describing the synopses. Through text parsing, a total of 70,983 clinical features were retrieved from the CS fields, of which 39,683 were found to be unique.



Path to Ontology : C:\Users\(

Path to Disorder Set : C:\Users\(

3 Disorders with symptom "severe failure to thrive" (ID: 7...

Disorders

DIABETES MELLITUS, TRANSIENT NEONATAL, 1

DIABETES MELLITUS, TRANSIENT NEONATAL, 1
DONOHUE SYNDROME
RHIZOMELIC CHONDRODYSPLASIA PUNCTATA, TYPE 1; RCDP1

transient neonatal diabetes mellitus (tndm)
growth
intrauterine growth retardation
severe failure to thrive
inheritance
imprinted, paternally expressed gene on chromosome 6q22-q23
lab
hyperglycemia
metabolic
dehydration
insulin therapy usually required
misc
late predisposition to type 2 (insulin resistant) diabetes
usual resolution within 6 months

Display Mode
● OMIM Representation ○ PhenOMIM Representation

embryonic
failure to thrive
chronic failure to t
failure to thrive
failure to thrive (4
failure to thrive (b
failure to thrive (e
failure to thrive (ir
failure to thrive (ir
failure to thrive in
failure to thrive in
failure to thrive in
failure to thrive in
failure to thrive se
failure to thrive, s
failure to thrive, s
failure to thrive.
low birth weight. f
marked failure to t
occasional failure
postnatal failure t
progressive failure
severe failure to t
severe postnatal failure to thrive
feeding
fetus
growth

Fig. 1. OmimView, a C++ interface for displaying and editing OMIM and PhenOMIM data

The data were then sorted according to a hierarchical ontological system designed by a group of doctors, based on anatomical connections between symptoms. Ambiguous descriptions were manually checked and standardized in order to clarify the semantics. To best preserve information, the original symptom descriptions were kept as much as possible. As such, each symptom node from the original synopses is linked with at least one term from the controlled vocabulary (PhenOMIM).

In contrast to the original OMIM dataset which can be represented as a tree structure with a depth of up to 3, the PhenOMIM ontology has a depth of up to 7 levels. The depth has been chosen to make the data more comparable, i.e. symptoms sharing more categories in their path are assumed to be more similar.

When a symptom has been found in the PhenOMIM ontology to replace a given symptom in the OMIM ontology, the OMIM node is marked as being deleted. However, the node is not actually deleted. Instead, a *reference link* is set to signify the replacement relationship. The purpose of this is that when a new dataset becomes available, it can be merged against the original OMIM ontology thus reducing the amount of work for rechecking previously classified symptoms so to keep the system maintainable. Moreover, it enables the user to view and compare the CS both in PhenOMIM as well as OMIM formats, with the information will remain to be stored in the ontologies.

III. Results and Discussion

A. Results

The *Clinical Synopses* of a total of 4,807 diseases were parsed into the OMIM Ontology. This has produced a total of 39,683 nodes, of which 39,491 are symptoms (see Table 1).

TABLE I
NUMBERS OF NODES IN OMIM AND PHENOMIM ONTOLOGIES

	OMIM	PhenOMIM [a]
Number of Nodes	39,683	40,173
Symptom nodes	39,491	34,889
Category nodes	192	5,284

During the reclassification process, many obvious errors and differences are resolved by mapping to new labels. For example, in the original data categories such as '*abd*' and '*abdomen*' both contained symptoms as '*inguinal hernia*' and '*umbilical hernia*'. Symptoms from both categories have now mapped to PhenOMIM symptoms at '*root\abdomen\abdominal wall\hernia\*' and '*..\umbilicus*'. Other examples include the categories in the original data named '*immune*', '*immunol*', '*immunologic*', '*immunology*', and '*imunology*' which descendants are mostly mapped to nodes in the category '*root\hematological\blood\immune*'.

Other solved problems include errors due to erroneous line breaks in the original files. Nodes created due to these dangling strings are marked deleted without linking to other symptoms in the PhenOMIM ontology. Original symptom names however are kept as much as possible, as often the symptom names contain references to indicate severity or extend of presence of the symptom. An example of this can be found in the category '*root\development\failure to thrive*', where there are many symptoms relating to *'failure to thrive*'.

The structured and controlled vocabulary has been stored in a SQL database for viewing through an AJAX based web-interface which can be found at http://faculty.neu.edu.cn/bmie/han/PhenOMIM/. The OMIM and the PhenOMIM ontology as well as the references between nodes are stored in separate tables. In the web-interface, users can browse both the OMIM as well as the PhenOMIM ontologies for comparison. A subset of the clinical synopses can be browsed both in OMIM as well as PhenOMIM representation. In the web-interface facilities are also added for users to give comments on the current structure and specific nodes, and users are invited to comment on the created ontology. In this fashion, the ontology can be continuously improved.

B. Discussion

We have presented the current state of a continuing effort to standardize the OMIM vocabulary. To this extent an off-line tool has been constructed to browse and edit the dataset as well as a web-based interface. The approach taken in this work is to keep as much as possible the original information for the symptoms in order to maintain modifiers to denote the extent of severity of the symptoms. This, on the other hand, has caused a problem that leafs (i.e., symptoms) are not very comparable, since there may be multiple labels describing the same symptom. We argue that instead of directly comparing the leaf nodes, comparing the category nodes may be a better measure for comparability of the symptoms, as categories in themselves do encode important information on the nature of the symptom. Moreover, by introducing extra levels (from 3 levels in the original OMIM data to up to 7 levels in the PhenOMIM data), comparisons can become more fuzzy of nature, and thus introducing extra information that is encoded in the categories and sub-categories.

Based on the above assumption, one may conclude that the structure of the ontology (i.e. the structure of categories and subcategories) is rather important for the results of any comparison algorithms. An ontology as such can thus be build up from various perspectives. One could take an anatomical approach where the fingers are connected to the hand, the hand connected to the wrist, and so on, but also other approaches such as based on the embryologic origins of the affected tissue.

IV. CONCLUSION

Although currently available systems like MeSH and UMLS have enlisted comprehensive semantic coverage, it is still difficult to directly apply such systems for the OMIM data. Several conceptual gaps seem to exist, which include 1) the MeSH and UMLS system have been established on the basis of anatomy and therefore may be unsuitable for definitions of birth defects; 2) conventional embryology has focused primarily on the origins of various tissues and details of development rather than pathogenesis of malformations; 3) for gross features such as '*coarse face*', '*short neck*' and '*joint contracture*', etc., no classification may be found from the existing semantic systems. Therefore, our vocabulary may serve as a special tool for analysis of the OMIM data.

The controlled vocabulary developed here can be used as a platform on which various phenomics approaches may be undertaken. For instance, text-mining procedures may be designed to automatically parse other phenotypic information based on this vocabulary. Along with other measures such as sequence and protein-protein interaction analyses, phenotype grouping may be performed for the prediction of gene functions. Furthermore, phenotype mapping may also be used for comparing different diseases to gain insights into their underlying molecular mechanisms.

Our work has featured a first attempt to develop a structured and controlled vocabulary based on the clinical features of hereditary human disorders. Since it has been built up manually, the vocabulary may provide more accurate genotype-phenotype relationships. Involvement of several doctors in our team has made this vocabulary rich in, or even biased towards medical knowledge. As the OMIM database is continuously updated, our controlled vocabulary will also need to be regularly updated. For the next step, we will continue to expand and improve the vocabulary to include other phenotypic information (for instance those of gene expression), and to make it compatible with other phenotype databases and/or controlled entities, in particular the Gene Ontology, the MeSH and/or the UMLS system. Furthermore, algorithms may be applied to compare disorders using the PhenOMIM vocabulary in order to derive potential interactions underlying diseases with similar phenotypes.

ACKNOWLEDGMENT

The authors are indebted to Drs. Jingshu Zhang, Jun Wang, Zhijun Sun and Hongkun Jiang for their assistance and criticisms on the database construction. Furthermore, the authors would like to thank He Huang for the help with the construction of the website.

REFERENCES

[1] International Human Genome Sequencing Consortium, "Initial Sequencing and analysis of the human genome", *Nature*, vol. 409, pp. 860-921, 2001.

[2] C. R. Scriver, "After the genome, the phenome?", J. Inherit Metab Dis. Vol. 27, no. 3, pp. 715-717, 2004.

[3] Y. A. Lussier and Y. Liu, "Computational approaches to phenotyping: High-throughput Phenomics", in Proc Am Thorac Soc, Vol. 4, no. 1, pp. 18-25, 2007

[4] K. M. Zbuk and C. Eng, "Cancer Phenomics: RET and PTEN as illustrative models", Nat Rev Cancer, vol. 7, no. 1, pp 35-45, 2007.

[5] L. Fernandez-Ricaud, J. Warringer, E. Ericson, K. Glaab, P. Davidsson, F. Nilsson, G. J. L. Kemp, O. Nerman and A. Blomberg, "PROPHECY – a yeast phenome database, update 2006", Nucleic Acids Res, vol.35, pp. 463-467, 2007.

[6] K. Paigen and J. T. Eppig, "A mouse phenome project", Mamm Genome, vol. 11, no. 9, pp. 715-717, 2000.

[7] N. de la Cruz, S. Bromberg, D. Pasko, M. Shimoyama, S. Twigger, J. Chen, C.-F. Chen, C. Fan, C. Foote, G. R. Gopinath, G. Harris, A. Hughes, Y. Ji, W. Jin, D. Li, J. Mathis, N. Nenasheva, J. NIe, R. Nigam, V. Petri, D. Reilly, W. Wang, W. Wu, A. Zuniga-Meyer, L. Zhao, A. Kwitek, P. Tonellato, and H. Jacob, "The Rat Genome Database (RGD): developments towards a phenome database", Nucleic Acids Res, vol. 33, pp. 485-491, 2005.

[8] T. Kurimori, T. Wada, A. Kamiya, M. Yuguchi, T. Yokouchi, Y. Imura, H. Takabe, T. Sakurai, K. Akiyama, T. Hirayama, K. Okada, and K. Shinozaki, "A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of Arabidopsis", Plant J. vol. 47, no. 4, pp. 640-651, 2006.

[9] A. Kahraman, A. Avramov, L. G. Nashev, D. Popov, R. Ternes, H.-D. Pholenz, and B. Weiss, "PhenomicDB: a multi-species genotype/ phenotype database for comparative phenomics", Bioinformatics, vol. 21 no. 3, pp. 418-420, 2005.

[10] M. A. Bogue, S. C. Grubb, T. P. Maddatu and C. J. Bult, "Mouse Phenome Database (MPD)", Nucleic Acids Res, vol. 35, pp. 643-649, 2007.

[11] N. Friemer and C. Sabatti, "The human phenome project", Nat Genet, vol. 34, pp. 15-21, 2003.

[12] Y. Lussier, T. Borlawsky, D. Rappaport, Y. Liu, and C. Friedman, "PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing", Pac Symp Biocomput, pp. 64-75, 2006.

[13] The Gene Ontology Consortium, "Gene Ontology: Tool for the unification of biology", Nat Genet, vol 25, pp. 25-29, 2000.

[14] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner and J. A. M. Leunissen, "A text-mining analysis of the human phenome", Eur. J. Hum Genet, vol. 14, no. 5, pp. 535-542, 2006.

[15] V. A. McKusick, "Mendelian Inehritence in Man, A Catalog of Human Genes and Genetic Disorders", 12th edition, John Hopkins University Press, Baltimore, 1998.

[16] C. D. Bajdik, B. Kuo, S. Rusa, S. Jones, and A. Brooks-Wilson, "CGMIM: Automated text-mining of Online Mendelian Inheritance in Man to identify genetically-associated cancers and candidate genes", BMC Bioinformatics, vol. 6, pp. 78, 2005.

[17] M. Masseroli, O. Galati and F. Pinciroli, "GFINDer: Genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists", Nucleic Acids Res, vol. 33, pp. w717-w723, 2005.