Mining Human Phenome to Investigate Modularity of Complex Disorders

Ranga C Gudivada[1, 4], Yun Fu[6], Anil G Jegga[2, 3, 4], Xiaoyan A. Qu [1, 4], Eric K. Neumann[5], Bruce J Aronow[1, 2, 3, 4]

Departments of Biomedical Engineering[1] and Pediatrics[2] University of Cincinnati, Center for Computational Medicine[3] and Division of Biomedical Informatics[4], Cincinnati Children's Hospital Medical Center, Cincinnati OH 45229, USA , Clinical Semantics Group [5], Lexington, MA 02420 and Beckman Institute[6], University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Abstract

A principal goal for biomedical research is to improve our understanding of factors that control clinical disease phenotypes. Among genetically-determined diseases, identical mutations may exhibit substantial phenotype variance by individual and background strain, suggesting both environmental and genetic mutant allele interactions. Moreover, different diseases can share phenotypic features extensively. To test the hypothesis that phenotypic similarities and differences among diseases and disease subvariants may represent differential activation of correlated feature "disease phenotype modules", we systematically parsed Online Mendelian Inheritance in Man (OMIM) and Syndrome DB databases using the UMLS to construct a disease – clinical phenotypic feature matrix suitable for various clustering algorithms. Using Cardiovascular Syndromes as a model, our results demonstrate a critical role for representing both phenotypic generalization and specificity relationships for the ability to retrieve non-trivial associations among disease entities such as shared protein domains and pathway and ontology functions of associated causal genes.

Introduction

Analyzing the overlaps and interrelationships of clinical manifestations of a series of related diseases may provide a window into biological modules that lead to pathophysiological processes to produce disease phenotype. This particular aspect of phenotype grouping reflects modularity in human disease genetics[1]. However, few computational methods[2-5] have been investigated to systematically cluster diseases and gain insights into the molecular processes underlying them. This is partly due to intrinsic difficulties in accessing controlled clinical descriptions and its availability in structured form suitable for computational genome – wise analysis. Our hypothesis is that the modular nature of complex disorders can be attributed to their clinical feature overlap associated with mutations in different genes that are part of same biological module. Genes that are part of a functional module are connected at various biological levels such as interacting partners, steps in a biochemical pathway, network of biological process or components of a multi protein complex[1-3, 5, 6]. Clustering diseases based on their shared clinical features provides an informational framework to analyze disease modularity and to explore underlying biological associations between various genotypic entities. Indeed, this approach of logical grouping of genes by their associated phenotype clusters is referred as phenomics[5]. We used Online Mendelian Inheritance in Man (OMIM)[7] and Online Congenital Multiple Anomaly/Mental Retardation Syndromes (SyndromeDB) (http://www.nlm.nih.gov/archive//20061212/mesh/jablonski/syndrome_toc/toc_c.html) as principal data sources for diseases and their corresponding clinical features. The phenotypic data presented in these data sources is not complete, unavailable in a well-ordered computable form and is not optimal to perform computations in forming accurate disease clusters. Therefore, this study only provides a proof of concept but certainly not a finished product. Our work has examined the possibility of using existing standard terminologies [8] and text mining tools[9, 10] to semantically normalize extracted clinical feature term variations and also dimensionality reduction methods to overcome the complexity in dealing with large number of clinical features.

Although there are limitations with this approach, our analysis revealed there is a detectable correlation between phenotype similarities to multiple levels of gene annotations. As, a pilot project we classified only Cardiovascular Syndromes (CVS) present in OMIM forming a cardio-phenome system. We considered an OMIM disease as a CVS if it has at least one cardiovascular symptom mentioned in the clinical synopsis (CS) section or occurrence of terms

such as “heart” or “cardiovascular” or “cardiac” in the free text section (TX) of OMIM.

Methods

Data Sources

Our data include both phenomic sources for forming disease clusters and genomic sources to validate the clusters by investigating phenotype – genotype correlations.

Genomic Data Sources

Human Gene Ontology - gene (GOA) annotations (ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz) and protein info (ftp.ncbi.nih.gov/reffseq/H_sapiens/mRNA_Prot/human.protein.gpff.gz)was downloaded from the NCBI FTP site. Protein domains at NCBI's Conserved Domain Database (CDD)[11] were not available for download and were parsed from the file ‘human.protein.gpff.gz’ using the Biojava software package (http://biojava.org) .

Phenomic Data Sources

A total of 977 records were downloaded in XML format from OMIM by searching for terms ‘cardiovascular’ or “heart” or “cardiac” occurring in CS or TX sections. A Java XML parser (http://xerces.apache.org/xerces-j/) was used to extract OMIM ID, disease name and the associated CS and TX sections from each OMIM record. OMIM ID and the corresponding gene associations were downloaded from NCBI Entrez Gene FTP site (ftp.ncbi.nlm.nih.gov/gene/DATA/mim2gene).Syndrome DB is not available for download and a Java HTML parser (http: //htmlparser. sourceforge.net/) was used to extract the relevant data directly from their website. Each Syndrome DB entry has a ‘major features (MF) section’ (e.g. mouth and oral structures, abdomen and skin) similar to the CS section of OMIM. A subset of 152 records having corresponding OMIM identifier and ‘cardiovascular system’ as one of the major features were extracted.

Matrix construction

The pheno-matrix is a binary matrix in which all the rows are OMIM CVS and columns are clinical features which comprise both clinical symptoms and affected anatomy. We assigned a value of ‘1’ for the presence of clinical feature associated with an OMIM CVS and ‘0’for absence of feature. From the total of 977 OMIM records, we took a subset, of 455 (46%) having atleast one causative gene and a clinical feature. These 455 CVS are associated with 585 genes.

Refining Clinical Features

We performed a three step process to refine and reduce the clinical feature dimensional space.

- Semantic Normalization
- Utilizing subsumption relations
- Principal Components Analysis (PCA)

Semantic Normalization

The TX and CS sections of OMIM and MF section of Syndrome DB are presented as loosely defined free textual descriptions. There is inconsistency in the use of clinical feature terms both semantically (e.g. increased sweating and diaphoresis) and syntactically (e.g. neonatal hypotonia and hypotonia, neonatal). In order to accomplish semantic normalization, based on several earlier approaches [3, 12], we have chosen to directly map these terms to Unified Medical Language System (UMLS) [8] concepts (CUIs), using MetaMap[9]. It’s a NLP (Natural Language Processing) tool which takes free text from biomedical domain and maps noun phrases to a potential list of matching concepts from UMLS metathesaurus. We used an online version of MetaMap program, available as part of Semantic Knowledge Representation project (SKR) (http://skr.nlm.nih.gov/), which aims to provide a framework for exploiting UMLS knowledge resources for NLP. The MetaMap output was first refined by restricting the mappings belonging only to UMLS Semantic Network semantic types under ‘disorders’ and ‘anatomy’ semantic groups. These sets are further refined between scores ranging from 570 to 1000 and after careful manual curation, incorrectly assigned concepts were eliminated.

OMIM TX section contains large sections of free text as opposed to small phrases in CS. SKR-MetaMap works well for short phrases but throws exceptions while handling the TX section of OMIM. As an alternative, we used GATE (General Architecture for Text Engineering [10], a text mining toolbox, for clinical feature entity recognition in the TX section of OMIM. GATE uses gazetteers, a component to hold a list of members of a particular category. Here, the input to gazetteers is a list of clinical feature keywords supplied from UMLS concept names and synonyms belonging to ‘disorders’ and ‘anatomy’ semantic groups. GATE

scans through each OMIM TX section and identifies the clinical features matching to the keywords present in its gazetteers, a post-processing step is performed to find the appropriate UMLS concepts for the extracted clinical features. Figure 1 illustrates the significant difference, a 50% reduction in total features by using UMLS concepts instead of raw clinical terms from unstructured text.

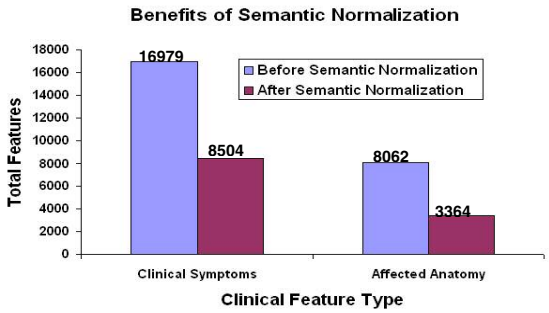


Figure 1: Reduction in Clinical features after semantic normalization. Count above each bar indicates the total number of features

Utilizing subsumption relations

UMLS Metathesaurus is a comprehensive database having terms from more than 100 various source vocabularies, which are mostly biomedical terminologies having hierarchical relations (parent/child) providing surrogate subsumption relations (is a, subclass of). MRHIER table from UMLS provides the required hierarchical relations between UMLS concepts. As all the clinical features in our data set are mapped to corresponding UMLS concepts, we devised a method to further reduce clinical feature space utilizing the subsumption relations present in MRHIER table. The entire clinical feature CUIs were scanned to find the most root parent concept for every specific subset of clinical features. OMIM CVS associated with that particular subset of clinical features (child CUIs) are now linked to the parent CUI instead of to the child CUI. By using subsumption relations and also ignoring clinical features associated with only one specific CVS, the final clinical feature set was further reduced to 1916 features.

Principal Components Analysis

Since the number of diseases (455 CVS) is much smaller than the number of the features (1916 clinical features from earlier step) and the features are all binary, the data contains a great deal of redundancy. We selected Principal Components Analysis in order to detect meaningful underlying dimensions from our high dimensional data set, as it efficiently reduces high dimensional data into low dimensional map. To select the minimum number of principal components required for our analysis, we gathered literature evidences of similar diseases associated with Marfan syndrome[13] and Coronary heart disease[14]. From the normalized cumulative sum (Figure 2), we see that only 68% spectrum energy for the top 82 principal components are good enough to reproduce the classifications. Selecting the top 82 principal components (from earlier 1916 clinical features obtained by using subsumption relations) clearly explicates an effective dimensionality reduction. Matlab 7.0 (http:// www.mathworks.com/), a popular Mathematical toolset was used to perform PCA and to generate the plot.

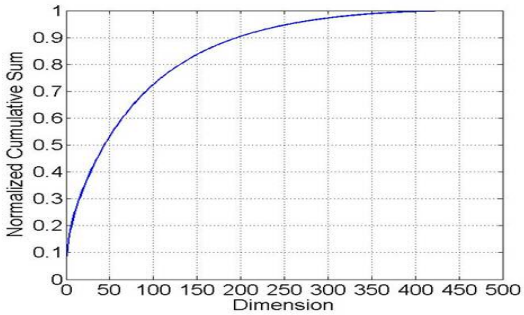


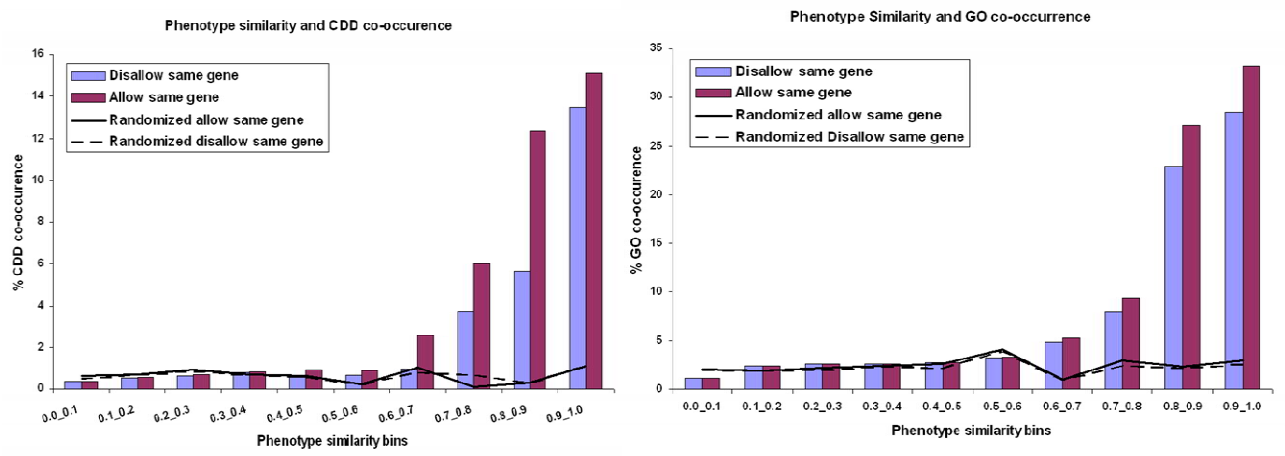
Figure 2: Plot of normalized cumulative sum versus dimension

Similarity measure and Clustering

The similarity between two CVS is calculated by measuring the cosine of angle between the associated clinical features vectors obtained after PCA. Hierarchical Clustering was performed on the resulting 455 x 455 phenomap (distance matrix obtained after applying cosine distance on the 455 x 82 pheno-matrix, where 82 are the top principal components). 'R' statistical software package (http://www.r-project.org/) was used for this analysis.

Results

The average of 10 randomized phenomaps was used as a control for background signal. Clinical feature vectors were randomly permutated using Fisher – Yates shuffling[15] before applying PCA and we



(a) **(b)**

Figure 3: Phenotype similarity versus gene annotation similarities (a) proteins associated with similar phenotypes and sharing at least one CDD domain (b) genes associated with similar phenotypes and sharing three or more GOA at the sixth or more detailed level. The average signal of 10 randomized phenomaps is at the lowest level. Disallow same gene analysis skips the disease pairs having same implicated gene.

considered the top principal components associated with 68% of spectral energy as like original matrix. We used similar validation approaches as [5], to test the initial hypothesis that similarities at the phenotypic level correlate to similarities in gene/protein function. Though we performed a thorough analysis correlating phenotype overlap to multiple levels of gene annotations, due to space limitations, here we present the results only for co-occurring domains and GO annotations.

Phenotype similarity – Domain co-occurrence

Proteins share functional domains and a mutation occurring in a shared domain might disrupt a specific biological process or a pathway leading to similar phenotypes [5]. Figure 3a shows the percentage of protein pairs that share a CDD domain as a function of the phenotypic similarity scores. The percentage of shared domains increases with increasing phenotype similarity score from 0.3% to 15%. For instance, 'Cardiomyopathy, Dilated, 1E' [OMIM: 601154] is caused by a mutation in SCN5A [NCBI GENE ID: 6331] and shares phenotypic characteristics with 'Jervell And Lange-Nielsen Syndrome' [OMIM: 220400] that is caused by a mutation in KCNQ1 [NCBI GENE ID: 3784]. These two proteins have a common 'sodium ion transporter' domain [CDD: 70001].

Phenotype similarity – Gene Ontology correlations

To explore possible functional relations between genes associated with overlapping CVS, we compared GOA. Similar to the earlier work[16] , we defined GO similarity by the sharing of atleast three GOA at the sixth or more detailed GO level. From Figure 3b the percentage of CVS pairs that share three or more GOA increased (from 1.15% to 33.33%) as a function of the phenotypic similarity. The signal we find is well above the average of 10 randomized matrices (~2%) over all bins.

Discussion and Conclusion

Though our work is closely related to [3, 5], we did deviate in several ways by using Syndrome DB in addition to OMIM and further reducing clinical feature dimension space by utilizing subsumption relations from UMLS and also implementing PCA were novel. Our work primarily emphasizes novelty in using combination of different existing methods in a sequential manner. We were not able to compare our results with the earlier work (analyzed 1653 OMIM phenotypes) as here we concentrated on subset of OMIM diseases (455 CVS). This approach substantially differs by work of others, such as Goh et al [17], who used OMIM information directly to link genes with diseases. Our approach has the advantage of using the breadth of symptomatic evidence to establish possible associations between different diseases, there by not relying solely on our current limited knowledge of the genetic basis. At present, we investigated how semantically rich information of symptoms can be used to relate their causative diseases to underlying genetic components. In previous work [18], we have shown how Semantic

Web (SW) standards can be used to aggregate broad sets of data, and prioritizing them for relative contribution using a page-ranking approach. We are currently working towards in combining these phenome networks obtained from clustering to integrated genome networks using SW standards and applying centrality analysis based ranking algorithms to discover the associated biological modules for definitive overlapping phenotypes. As more evidence and interpretations get compiled, we anticipate broad semantic analysis becoming more robust and yielding more insights. We believe the use of diverse sets of evidence and hypotheses will greatly advance the set of testable models for overlapping diseases mechanism and have a direct impact in the development of both new and re-directed therapeutic applications.

References

[1] H. G. Brunner and M. A. van Driel, "From syndrome families to functional genomics," *Nat Rev Genet,* vol. 5, pp. 545-51, Jul 2004.

[2] M. N. Cantor and Y. A. Lussier, "Mining OMIM for insight into complex diseases," *Medinfo,* vol. 11, pp. 753-7, 2004.

[3] K. Lage, E. O. Karlberg, Z. M. Storling, and P. I. Olason, et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nat Biotechnol,* vol. 25, pp. 309-16, Mar 2007.

[4] Y. Liu, J. Li, L. Sam, C. S. Goh, M. Gerstein, and Y. A. Lussier, "An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits," *PLoS Comput Biol,* vol. 2, p. e159, Nov 17 2006.

[5] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. Leunissen, "A text-mining analysis of the human phenome," *Eur J Hum Genet,* vol. 14, pp. 535-42, May 2006.

[6] M. Oti and H. G. Brunner, "The modular nature of genetic diseases," *Clin Genet,* vol. 71, pp. 1-11, Jan 2007.

[7] V. A. McKusick, "Mendelian Inheritance in Man and its online version, OMIM," *Am J Hum Genet,* vol. 80, pp. 588-604, Apr 2007.

[8] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Res,* vol. 32, pp. D267-70, Jan 1 2004.

[9] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," *Proc AMIA Symp,* pp. 17-21, 2001.

[10] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan., "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, July 2002.

[11] A. Marchler-Bauer, J. B. Anderson, and M. K. Derbyshire, et al., "CDD: a conserved domain database for interactive domain family analysis," *Nucleic Acids Res,* vol. 35, pp. D237-40, Jan 2007.

[12] A. J. Butte and I. S. Kohane, "Creation and implications of a phenome-genome network," *Nat Biotechnol,* vol. 24, pp. 55-62, Jan 2006.

[13] P. N. Robinson, E. Arteaga-Solis, and C. Baldock, et al., "The molecular genetics of Marfan syndrome and related disorders," *J Med Genet,* vol. 43, pp. 769-87, Oct 2006.

[14] S. A. Ritchie and J. M. Connell, "The link between abdominal obesity, metabolic syndrome and cardiovascular disease," *Nutr Metab Cardiovasc Dis,* vol. 17, pp. 319-26, May 2007.

[15] Y. F. Fisher RA, *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver and Boyd, 1938.

[16] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)," *Proc Natl Acad Sci U S A,* vol. 100, pp. 8348-53, Jul 8 2003.

[17] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabasi, "The human disease network," *Proc Natl Acad Sci U S A,* vol. 104, pp. 8685-90, May 22 2007.

[18] R. C. Gudivada, X. A. Q., A. G. Jegga, E. K. Neumann, and B. J. Aronow, "A Genome – Phenome Integrated Approach for Mining Disease-Causal Genes using Semantic Web," *Proceedings of the WWW2007 Workshop on Healthcare and Life Sciences,* 2007.