Biomechanisms of Comorbidity: Reviewing Integrative Analyses of Multi-omics Datasets and Electronic Health Records

N. Pouladi1-3*, I. Achour1-3*, H. Li1-3, J. Berghout1-3, C. Kenost1-3, M. L. Gonzalez-Garay1-3, Y. A. Lussier1-4

1 BIO5 Institute, The University of Arizona, Tucson, AZ, USA
2 Center for Biomedical Informatics and Biostatistics, The University of Arizona, Tucson, AZ, USA
3 Department of Medicine, The University of Arizona, Tucson, AZ, USA
4 University of Arizona Cancer Center, The University of Arizona, Tucson, AZ, USA

Summary

Objectives: Disease comorbidity is a pervasive phenomenon impacting patients’ health outcomes, disease management, and clinical decisions. This review presents past, current and future research directions leveraging both phenotypic and molecular information to uncover disease similarity underpinning the biology and etiology of disease comorbidity.

Methods: We retrieved ~130 publications and retained 59, ranging from 2006 to 2015, that comprise a minimum number of five diseases and at least one type of biomolecule. We surveyed their methods, disease similarity metrics, and calculation of comorbidities in the electronic health records, if present.

Results: Among the surveyed studies, 44% generated or validated disease similarity metrics in context of comorbidity, with 60% being published in the last two years. As inputs, 87% of studies utilized intragenic loci and proteins while 13% employed RNA (mRNA, lncRNA or miRNA). Network modeling was predominantly used (35%) followed by statistics (28%) to impute similarity between these biomolecules and diseases. Studies with large numbers of biomolecules and diseases used network models or naïve overlap of disease-molecule associations, while machine learning, statistics, and information retrieval were utilized in smaller and moderate sized studies. Multiscale computations comprising shared function, network, topology, and phenotypes were performed exclusively on proteins.

Conclusion: This review highlighted the growing methods for identifying the molecular mechanisms underpinning comorbidities that leverage multiscale molecular information and patterns from electronic health records. The survey unveiled that intragenic polymorphisms have been overlooked for similarity imputation compared to their intragenic counterparts, offering new opportunities to bridge the mechanistic and similarity gaps of comorbidity.

Keywords

Comorbidity, disease similarity, electronic health records, phenome, biomolecules, function, omics, modularity, integrative analysis

Yearb Med Inform 2016:194-206
http://dx.doi.org/10.15265/IY-2016-040
Published online November 10, 2016

1 Introduction

The last century has witnessed some of the greatest medical, scientific, and technological advances, substantially increasing life expectancy by 40% through better health care and prevention [1]. Global public health has been instrumental in reducing neonatal and transmissible disease mortality [1, 2]. However, in this common age of industrialization and urbanization, living longer increases the risk of developing non-com- municable diseases, often chronic, complex, and comorbid [3-5]. Disease comorbidity is becoming widely pervasive counting for 35% to 80% of case reports among 20 to 75 year-old patients [3, 6]. Although a growing body of research studies used associative analysis to identify risk factors of disease comorbidity, approaches leveraging both phenotypic and molecular patient information to understand the biological underpinnings of comorbidity remain elusive.

Disease comorbidity includes illnesses that mutually or gradually occur during one’s lifetime [3]. In the United States, with over 1.7 million sampled electronic health records, 42.2% of patients have been reported with at least one other morbidity [7]. Among individuals living with multiple diseases, there are those with two or more co-existing diseases (e.g., diabetes, Alzheimer’s, and cancer) and those with associated diseases developed as secondary conditions (e.g., diabetics with hypertension or retinopathy) [3, 8]. Further, disease comorbidity can be deleterious or protective [9-11], for example, immune-mediated inflammatory disorders increase the odds of infections and developing lymphoma, [12] while solid tumors in Down Syndrome patients lower the risk of developing leukemia and testicular cancer (inverse comorbidity) [11].

The sequence of events and the type of diseases developed during the course of time impact disease comorbidity management [3, 13]. Furthermore, disease comorbidity can limit the use of standard therapeutic options and medical interventions, e.g., a) the use of oral corticosteroids to treat patients with chronic obstructive pulmonary disease (COPD) patients and diabetes mellitus [14], b) the use of beta-blockers to treat asthmatic patients who are hypertensive [13], or c) resecting malignant lung tumors in patients with severe COPD [15]. Often, cancer patients with comorbidity found their survival outcomes compromised with a 5-year mortality hazard ratio between 1.1 and 5.8 years due to suboptimal or non-adequate chemotherapeutics [16, 17]. Different risk factors of comorbidity have been identified, such as frailty due to aging, drug side effects, and poor socio-economic status [3, 6, 18]. However, epidemiology alone has been

* These authors contributed equally
insufficient to understand fully the causes of developing multiple diseases and inform on adequate healthcare solutions.

In the last decade, the biological underpinning of disease comorbidity is coming into focus to complement epidemiology studies and bridge the mechanistic gap between phenotypes and genotypes. Increasingly, studies are looking at disease overlap and similarity based on disease-shared cellular and molecular mechanisms [6, 10, 19-23] such as shared disease genes, genetic variants, associated proteins [24, 25], or biological pathways, [26] etc. The probability of developing a secondary disease is about 3-fold and 1.8-fold if the primary disease shares the same genes or metabolic fluxes, respectively [24, 27]. Biologically informed disease similarity metrics have been developed using computational and statistical methods ranging from simple overlap and machine learning to network-based methods. Goh et al. and others constructed a disease network where diseases are connected given their associated molecular modules [19, 21, 28-30] including both at the topological (e.g. distance in protein-protein interactions (PPI) [31]) and functional levels (e.g. enzyme functions or apoptosis [19, 28]. The overlaps of disease modules can be significantly enriched with disease genetic variants [32, 33] supportive of shared pathogenesis between diseases and their comorbidity [19, 24, 27, 34].

In this review, we primarily focused on studies that incorporated both detailed phenotypes and biomolecular information to uncover disease similarity and understand the underpinning of comorbidity.

2  Methods

We retrieved publications using combinations of keywords queried in PubMed and Google database search engines such as comorbidity, phenotypic information, electronic health records (EHR), patient health records (PHR), diseasesome, genome, gene expression, data integration, computational methods, gene ontology, disease-disease similarity, disease network, shared mechanisms, molecular network, PPI, etc. We selected over 135 publications and studies including 59 that were surveyed based on the number of phenotypes, type of biomolecules, methods, and similarity metrics used and analyzed in the context of multiple diseases and/or comorbidity. The 59 studies, ranging from 2006 to 2015, were selected under the following criteria, (i) as an input, at least five diseases or traits and one type of biomolecule (single nucleotide polymorphism, (SNP) gene, RNA and/or proteins), (ii) at least used one of the five methods based on network, statistics, machine learning, information retrieval, and overlap, and (iii) at least produced one of the five disease similarity metrics based on shared loci, function, phenotype, both function and topology and/or jointly function, topology and phenotype (Fig 1 and Fig 2).

We also produced two plots reflective of the distribution of the methods used (Fig 3) as well as the similarity metrics generated (Fig 4) across the 59 surveyed publications and based on the type and number of biomolecules relatively to the number of phenotype used as an input. Abbreviations, key concepts and databases are outlined and defined in Table 1.

3  Results

To present and highlight previous, current, and future research directions underpinning the biology of comorbidity, we first sought to evaluate 59 studies and approaches that (i) integrated phenotypic and molecular information, (ii) used different computational and statistical methods, and (iii) uncovered disease similarity with and without a comorbidity context (Fig 1).

3.1  Survey

In the last decade, among these 59 studies that incorporated both phenotypic and molecular information, 44% generated or validated disease similarity metrics in a context of comorbidity. Interestingly, during the last two years, there has been a surge in studies assessing comorbidity, counting for 60% of our selected publications (Fig 2A). Looking across the spectrum of biomolecules, gene and protein information were predominantly used as input, counting respectively for 38% and 35% of surveyed publications, with at least a third being assessed in the context of disease comorbidity (Fig 2B). However, among the publications that analyzed SNP (14%) and RNA (13%), only 2% or none have evaluated disease comorbidity, respectively (Fig 2B). In the context of methods, studies mainly used network (34%) and statistics (28%) based methods followed by overlap (18%), machine learn-
Fig. 2 Publication survey based on the biomolecules, methods, and metrics used and generated without and with assessing comorbidity. A. Histogram illustrating number of publications analyzing disease similarity without and with assessing comorbidity. B. The studies are grouped based on the input biomolecules, mathematical framework, and the similarity metrics, which highlight the extensive lack of contribution of RNA species and the use of FTP metric for comorbidity analysis. Each pie chart presents the number of distinct studies for each captured element. The legend describes (clockwise starting at the top) the segments of the pie from lighter color to darker. In some cases, a subset of studies includes comorbidities (c), thus leading to two segments with the same color. For example, SNPs with and without comorbidities are relevant to 12% and 2% of the studied papers, respectively.

The results showed that the most commonly used biomolecules were SNPs (26%) and information retrieval (8%). With the exception of machine learning, a third or half of the publications using these methods analyzed shared molecular mechanisms between diseases including those that are comorbid (Fig 2B). Finally, among the 59 surveyed publications, the five disease similarity metrics were comparably represented (15-24%) with at least a third or more were generated to understand the biology of comorbidity with the exception of integrative disease similarity metrics based shared function, topology and phenotype (2% out of 17%). In light of these results, we focused our efforts on pioneer and emerging studies highlighting the importance and applicability of one-level molecular and phenotypic scales as well as integrative data based disease similarity metrics to decipher the biological underpinning of comorbidity.
### Table 1  Abbreviations and key concepts used in this review

#### A. Phenotypic database and ontology

<table>
<thead>
<tr>
<th>Abbreviations or concepts</th>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>Disease</td>
<td>Clinical phenotypes: diseases, traits, digital database of medical and/or personal health records of patients (Electronic Health Records, EHR) which include age, gender, race, clinical laboratory test results, response to therapy, medications, signs and symptoms, medical imaging, quantitative risk factors (endophenotypes), etc.</td>
</tr>
<tr>
<td>OIMM</td>
<td>Disease</td>
<td>Online Mendelian Inheritance in Man: collection of human genes and genetic phenotypes, which includes gene-phenotype interactions</td>
</tr>
<tr>
<td>PubMed</td>
<td>Literature</td>
<td>Collection of citations for biomedical literature</td>
</tr>
</tbody>
</table>

#### B. Molecular database, ontology and association study

<table>
<thead>
<tr>
<th>Abbreviations or concepts</th>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>Biomolecule</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>mRNA</td>
<td>Biomolecule</td>
<td>messenger RNA: RNA coding-protein</td>
</tr>
<tr>
<td>microRNA</td>
<td>Biomolecule</td>
<td>microRNA: non-coding RNA known to control mRNA expression.</td>
</tr>
<tr>
<td>IncRNA</td>
<td>Biomolecule</td>
<td>Long non-coding RNA: non-coding RNA</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
<td>Annotation of molecular function and biological process of genes and gene products</td>
</tr>
<tr>
<td>GWAS</td>
<td>Association study</td>
<td>Genome Wide Association Study: study assessing a correlation between SNPs and disease occurrence within a given population</td>
</tr>
<tr>
<td>PhenWAS</td>
<td>Association study</td>
<td>Phenome Wide Association Study: study assessing phenotypes associated to genotypes</td>
</tr>
</tbody>
</table>

#### C. Methods

<table>
<thead>
<tr>
<th>Abbreviations or concepts</th>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIN</td>
<td>Network</td>
<td>Molecular Interaction Network: representation of biomolecules (nodes) and their physical interactions or functional annotations (edges)</td>
</tr>
<tr>
<td>PPIN</td>
<td>Network</td>
<td>Protein-Protein Interaction (PPI) Network: representation of proteins (nodes) and their physical interactions (edges)</td>
</tr>
<tr>
<td>Disease network</td>
<td>Network</td>
<td>Disease Network: representation of diseases (nodes) and their similarity degrees (edges)</td>
</tr>
<tr>
<td>Topological feature</td>
<td>Network</td>
<td>Identification of structural and functional distances by which biomolecules or diseases are arranged and connected in the network, such as shortest distance between every two proteins in PPIN</td>
</tr>
<tr>
<td>Long range interaction</td>
<td>Statistics</td>
<td>Indirect interactions or relationships between two biomolecules in their associated networks</td>
</tr>
<tr>
<td>Genetic or Disease Profile</td>
<td>Machine Learning</td>
<td>Set of scores computed SNPs associated to diseases and vice versa</td>
</tr>
<tr>
<td>Clustering</td>
<td>Information Retrieval</td>
<td>Method for grouping elements into sets based on their relative similarities</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>Machine Learning</td>
<td>Term Frequency–Inverse Document Frequency: weighing method to adjust the frequency of occurrences of variables</td>
</tr>
</tbody>
</table>

#### D. Similarity Metrics

<table>
<thead>
<tr>
<th>Abbreviations and concepts</th>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shared loci</td>
<td>Luci</td>
<td>Number of genetic loci, SNPs, or genes and their relationships shared between diseases</td>
</tr>
<tr>
<td>Shared Function</td>
<td>Function</td>
<td>Biomolecules that are bioproducts of genetic loci such as RNAs, proteins, metabolites, etc., with assigned biological function and that are shared between diseases. The biological function is derived from gene ontology database, PPI, protein complex, and biological, molecular and metabolic pathways, etc.</td>
</tr>
<tr>
<td>Shared Phenotype</td>
<td>Phenotype</td>
<td>Any relationship that might be inferred between phenotypes, such as overlap, association, correlation, or co-occurrence</td>
</tr>
<tr>
<td>Shared Function and Topology (FT)</td>
<td>Function, Topology</td>
<td>Disease relationships based both on functional and topological features of biomolecules such as structure of feedback loop, indirect PPI, protein subnetworks, or the whole MINs or PPIN</td>
</tr>
<tr>
<td>Shared Function, Topology, and Phenotype (FTT)</td>
<td>Function, Topology, Phenotype</td>
<td>Disease relationships based both on functional and topological features of biomolecules in addition to phenotypic attributes</td>
</tr>
</tbody>
</table>
3.2 Methodology Used to Generate Disease Similarity Metrics

Several reviews have extensively described and evaluated the performance of varying methods for generating disease similarity metrics in the context of data integration strategies and modeling [35-41]. Here, we focused on five categories of methods based on network, statistics, machine learning, information retrieval and overlap (Fig 2-3, and Table 1). We evaluated the distribution of these methods based on the number and type of biomolecules relatively to the number of phenotypes analyzed (Fig 3). Distribution of the types of methods widely changes according to the number of biomolecules and phenotypic/disease inputs. We found that predominantly network based methods utilize proteins as inputs and very few methods employ RNA. We also found that big datasets with ~10^4 biomolecules and as many phenotypes largely rely on proteins information and to a lesser extent on Loci (SNP and gene). It is worth noting that the SNPs analyzed are intragenic (Fig 3).

3.3 Uncovering Disease Similarity and its Impact in the Biological Underpinning of Comorbidity

In this section we outline, describe, and highlight pioneering and emerging studies that utilize different disease similarity metrics to understand and reveal shared biology and etiology between multiple diseases including those that mutually or gradually occur in a patient’s lifetime (Fig 4).

3.3.1 Disease Similarity Metrics Based on Shared Phenotype

The use of very fine-grained phenotypes, such as disease-associated traits, phenotypic ontologies, clinical synopsis, electronic health records, laboratory tests, and billing information, increases the likelihood of uncovering disease similarity associated molecular mechanisms [19, 42-47]. Following the work of Swanson et al., [48] who highlighted the usefulness of Medline for knowledge discovery, several groups have utilized text-mining strategies to extract and integrate detailed phenotypes from medical datasets with underlined molecular mechanisms [48-51]. For instance, Butte et al. generated a network connecting 7,466 genes to 281 biomedical concepts via 64,003 relationships, highlighting a high connectivity of gene-phenotype-environment interactions [52]. Similarly, van Driel et al. extracted specific disease-associated phenotypes from the full text and clinical synopses of the OMIM database to build a human phenome network made of 5,000 human diseases in which the strengths of connection between disease pairs correlates with their phenotypic similarities [53].

Integrative analysis of the phenome (22) and PPI networks facilitated the discovery of novel disease-associated genes [54, 55]. Recently, Zhou et al. utilized the similarity between clinical manifestations of diseases queried from PubMed to construct a Human Symptoms Disease Network (HSDN) made of 322 symptoms, such as pain and diarrhea, and 4,219 diseases [56]. This integrative analysis of the HSDN with disease networks based on shared genes and PPIs uncovered a positive correlation of clinical presentations of diverse diseases with that of their underlying molecular interactions. Despite the discoveries that made use of phenotypic...
Biomechanisms of Comorbidity: Reviewing Integrative Analyses of Multi-omics Datasets and Electronic Health Records

As expected, proteins were not utilized in shared intragenic loci (top left) and minimally in the two other first row metrics; however, they are predominantly employed in the leftmost two metrics of the second row. Utilization of RNA is mostly found in “Shared Function” (top middle). No studies comprised intergenic loci for imputation of similarity thus none shown in the legend.

Electronic Health Records (EHR) are important sources of unbiased and detailed phenotypic data. Text-mining strategies [49, 57, 58] are often utilized to mine the electronic health records and extract rich phenotypes to stratify the patients and find comorbidities [59], analyze disease progression to uncover directional comorbidity [60], or to derive, in combination with genotypic data, pleiotropic genetic loci [61]. For instance, Razhetsky et al., analyzed the EHR by developing a statistical model in which the overlapping of the latent genetic susceptibilities across multifactorial diseases determines if they are positively or negatively correlated [62]. They showed that some diseases are inversely comorbid such as the negative correlations found between breast cancer and both schizophrenia and bipolar disorder. While, Hidalgo et al. adopted a network-based approach showing positive correlations between comorbidities and mortality rates dependent on age, gender, and disease progression, such as the higher prevalence of hypercholesterolemia in white males [63]. The landmark study of Blair et al. analyzed 110 million EHRs and uncovered striking comorbidities between 95 Mendelian and 150 complex diseases. They found that each of the complex diseases is attributable to a set of combinations of Mendelian-associated loci [64]. Simi-
larly, Melamed et al. found a comorbidity between Mendelian disorders and common cancers by evaluating shared genes and pathways [65]. They showed significant co-expression of genetically-altered genes associated with cancers with those of a set of Mendelian diseases. Altogether, the analysis of EHR uncovers novel phenotypic relationships among diseases stemming from common molecular and cellular mechanisms and that in part could explain their comorbidities [24, 66].

3.3.2 Disease Similarity Metrics Based on Shared Loci

During the last decade, many studies showed that genetic variants impact the prevalence of disease risk [67, 68]. Given the reported pleiotropic effects of such genetic variants, many studies assessed disease similarity based on their associated genetics (e.g. SNPs or genes) [69, 70]. Den- ny et al. analyzed the associations between 3,144 SNPs (single nucleotide variant) conferring risk of various diseases and traits with 1,358 phenotypes from 13,835 EHR in a combined phenotype-genome wide association study [71]. They uncovered 63 SNPs with pleiotropic effects that mechanistically link different phenotypes together such as actinic keratosis, seborrheic keratosis, and non-melanoma skin cancer. Other studies developed methods that make use of GWAS p-values of SNPs to build the association profiles for SNPs or diseases. Sirota et al. calculated, for each disease-SNP, a variant score in which its associated odds ratio and p-value account for its allelic directionality and the disease association strength, respectively [72]. They segregated six common autoimmune diseases into two classes based on their genetic profiles with variant scores of 573 SNPs. The allele directionality analysis prioritized SNPs conferring risk to one disease class while protecting against another one, underlying the inverse occurrence of each of the disease classes. For example, the minor allele of rs1752919 increases the risk of developing multiple sclerosis (MS) and autoimmune thyroid disease (ATD), while decreasing that of rheumatoid arthritis (RA) and ankylosing spondylitis (AS). In contrast to traditional meta-analysis, Cotsapas et al. developed a novel statistical method that evaluates concurrent associations of a SNP to several phenotypes [73]. They applied these statistics to GWAS data of seven autoimmune diseases and found 47 SNPs associated with multiple diseases, where clustering the SNP associated disease profiles identified four distinctive groups with each one being associated with one or multiple diseases, indicating shared genetic structure.

Recently, some studies developed other gene-based statistics to construct and compare the genetic profiles of diseases. Chang et al. presented a method in which they first construct a gene-based score using SNPs associated GWAS p-values, and measure how strongly each gene is associated with different diseases through randomization [74]. They applied a principal component analysis (PCA) to a gene disease interaction matrix of 31 distinctive diseases for which they identified shared pathogenic genetic profiles. On the other hand, Li et al. built a disease-trait network with 1,439 genes connecting 69 disease and 85 traits together [75]. Specifically, they considered a disease–trait pair similar if the cosine of their genetic profiles, based on their gene associated frequency-inverse document frequency (TF-IDF) weighted scores, are statistically significant. The time of occurrence between traits and their associated diseases, derived from the EHR, segregated the inferred associations into three different categories: risk factors, diagnostics, and complications. This is of value to the prognosis of potential comorbidity, such as the evaluation of decreased platelet counts before the diagnosis of alcohol dependence syndrome.

3.3.3 Disease Similarity Metrics Based on Shared Function

Multiple studies developed function-based disease similarity metrics beyond a simple assessment of shared genetics [76] using functional activity of biomolecules such as the expression of microRNAs [77, 78] and long non-coding RNA [79], interactions between proteins (e.g. PPI) [80], and gene ontology of biological pathways and processes [26, 81-83]. In this context, Sam et al. used enrichment analysis to identify significant disease similarities by taking into account the number of direct interactions between proteins belonging to different disease nodes [80]. Similarly, Yang et al. developed a metric based on differentially co-expressed molecular pathways [83]. On the other hand, Li and colleagues developed a novel disease similarity metric based on shared gene function derived from gene ontology (GO) [81] using information theoretic measure [84]. They showed that disease similarity based on GO metrics correlate with the shortest distances of their associated protein in the PPI network.

Multiple lines of evidence showed that comorbidity might stem from shared functional properties and domains such as from PPI or co-expressed genes [24]; however, simple metrics based on the number of the shared genes are not always sufficient to capture comorbid diseases with no shared genes [27, 76, 85]. For example, Lee et al. showed the observed comorbidity between metabolic diseases, not reflected from the disease network of shared genes but from shared metabolic fluxes [27]. Using enrichment analysis, Zhernakova et al. identified three major immune pathways (e.g. T-cell differentiation, immune-cell signaling and the innate immune response) that explain similar pathogenesis among 11 immune-related diseases with significant comorbidities [26]. Wang et al. utilized a similarity metric based on the number of shared protein complexes between diseases showing a high connectivity between diseases associated with different classes such as linking the glycolipid metabolic diseases with multiple types of cancers [25]. They also observed a two-fold increase in the odds of comorbidity between diseases sharing protein complexes.

3.3.4 Disease Similarity Based on Shared Function and Topology

Leveraging the topological features of biomolecules, in addition to their functional attributes, has the benefit to retain and uncover relevant structural properties of the molecular relationships and interactions underpinning disease similarity [86]. Using topological properties of protein interac-
tions and k-nearest neighbor’s classifier algorithm, Xu et al. showed that hereditary disease genes tend to cluster together in PPI networks curated from the literature rather than those predicted from high-throughput yeast two-hybrid mapping approach [87]. On the other hand, Kohler et al. utilized a random walk algorithm and PPI networks to reveal disease gene candidates and infer disease associated sub-networks [88]. They showed that similarity measures of long-range distance interactions outperform those of direct neighbors of disease genes. Recently, Hamenh et al. made use of an information flow algorithm for long-range interactions in the PPI networks to assign score relevance to all the proteins in the entire network for each disease [89]. These relevance score vectors were correlated to assess the degree of similarity between each disease. Other studies utilized long-range interaction based similarities in the context of long non-coding RNAs [90, 91]. For example, Yang et al. formulated the lncRNA-disease bi-partite graph with a resource-allocation process and applied a propagation algorithm to assess the relevance of each lncRNA and its associated protein-coding genes to various diseases [90]. Similarly, Chen X et al. showed, via machine learning methods, how disease similarity can be captured with the Gaussian kernel of interaction profiles of diseases with their associated lncRNAs [91].

Functionally related comorbidities can be identified through the analysis of similarity derived from the topological features of their associated PPI networks [66]. Duc-Hau et al. showed through Boolean simulation that diseases sharing shorter feedback loops in cell signaling network, particularly those positive, are more prone to be comorbid. This result suggests that amplified cell signals underpin the comorbidity [92]. Similarly, Paik et al. used network-based methods to study the shortest distance paths connecting the protein subnetworks and identified the overlapping degrees of subnetworks of the diseases and their associated traits [93]. They found disease overlap associated with PPI subnetworks reflective of their comorbidity with their associated traits. Importantly, Menche et al. assessed disease similarity using a novel network based distance measuring the overlapping degree between disease modules mapped to PPI networks [94]. With over 30 million EHR, they showed that the overlap of disease modules captured phenotypically different diseases with high comorbidity. Interestingly, this metric showed that diseases with high comorbidity and disease module overlap, such as lymphoma and myocardial infarction (RR=2.1), do not necessarily share similar genes. Following this study, Ghissi et al. developed an algorithm that utilized hypergeometric distribution and the connectivity patterns of disease associated proteins for detecting disease modules in the PPI [95].

3.3.5 Disease Similarity Based on Shared Function, Topology, and Phenotype

Modular nature of diseases reflects that sharing molecular mechanisms between diseases underlies their phenotypic similarities [53]. The natural extension of the use of long-range interactions is to incorporate the phenotypic similarity between diseases. By doing so, it is possible to infer the pathogenesis of diseases with unknown molecular mechanisms or to infer novel mechanisms for expanding known pathogenic disease basis [23]. There are several combined phenome-genome metrics that have been primarily devised to prioritize novel disease genes [96-100]. However, they make it possible to derive gene relevance profiles for each disease, using the propagation-based algorithm [96] or graph Laplacian [98], and optimizing the information flow in the genome-phenome network [97]. Some groups showed the adaptability of their gene prioritization metrics for assessing the disease similarity, for example, Li et al. used a coranking method to apply a random walk with restart algorithm to the bi-partite phenome-genome network [101]. They uncovered 122 associations between diseases sharing no obvious phenotype. On the other hand, Wu and colleagues [55] adapted a network alignment technique, which had been previously developed for finding conserved protein networks across different species [102]. They applied this approach to the genome-phenome networks and identified 39 bi-modules with diseases within each module belong to the same disease category. However, Hwang et al. addressed the genome-phenome integration problem with the use of regularized, non-negative matrix factorization and simultaneously identified the association between the dense clusters of similar phenotypes and those of genes [103]. While the diseases within the same class share similar phenotypic annotations, the ones from different classes also showed a similar pattern, suggesting similar molecular mechanisms could underpin different phenotypes.

Other groups also defined the disease gene similarity profiles. Wu et al. defined a score that includes topological distances between the genes on the PPI network in order to evaluate the correlation of the variation of the phenotypic profile of a disease relative to the other diseases with that of its associated genes’ profile relative to those of the other disease genes’ profiles [54]. The higher the score, the more similar the diseases are. Applying a bi-clustering method to the gene-disease similarity matrix identifies bi-modules with biological significance. Similarly, Zhao et al. developed a method based on gene closeness index, which makes an index by incorporating the relative closeness of a drug and target disease genes in the PPI network. Through the use of this index, the authors compute the gene closeness profile, which is subsequently partitioned to gene modules that importantly capture the diseases comorbidity [104].

4 Discussion

This literature review shows how the use of phenome enhances our knowledge about shared pathogenesis of diseases. Mining the phenotypic data from databases, such as OMIM and PubMed, helped uncover the correlation between phenotypic similarity with shared molecular mechanisms and extensive interrelationships between phenome, genome, and environment. In addition, we showed the joint analysis of the rich phenotypic information buried in EHR with molecular layers would give rise to discovering cellular mechanisms under-
pinning the interaction between diseases, such as inferring the pleiotropic loci [71], prioritizing diseases associated traits as risk factors [75], and also explaining the genetics of comorbidities [64]. It was through the use of EHR that Rzhetsly et al. [62] in their premier work linked the comorbidity with shared genetic susceptibility mechanisms, and later by Park et al. [24] and Lee et al. [27] for explaining comorbidity in disease modularity framework. The landmark work by Blaire et al. [64] highlights the immediate practicality of EHR for geneticists to systematically screen different combinations of rare or common phenotypes that might be genetically linked or vice versa, which would be otherwise discovered with lengthy and costly epidemiological studies. Importantly, several ongoing national efforts will facilitate studying the gene-phenotype-environment interactions, such as i2b2 [105], Vanderbilt BioVu [106], and the Electronic Medical Records and Genomics (eMERGE) consortium [107-109]. The future of genomic research depends on the effective use of EHR with sophisticated text mining methods to extract the informative phenotypic data from all of their different tiers, namely structured, codified and narrative [110-112].

There are unique limitations for each type of “omics” dataset that differentially impact data mining. Advanced technologies, such as microarrays or next generation sequencing (NGS), have enabled cost-effective, high-throughput measures of the whole genome biological activity. However, they present sequence assembly and accuracy challenges affecting each of the produced “omics” datasets. These limitations introduce noise and biases. For example, in the context of NGS, DNAseq and RNAseq introduce gene-specific biases [113] and can produce low mappability reads, causing base mis-calls and mis-alignments often being filtered out, which leads to missing data [114]. This particularly plagued pseudogenes and their coding paralogs or orthologs, as well as genomic duplication blocks, repetitive regions, and multi-gene clusters such as cytochromes [113, 115, 116]. Highly polymorphic regions, such as HLA or regions often subject to DNA rearrangements, also suffer from mappability issues [117], leading to a mismatch to the reference scaffold. In addition, computational tools [118], library preparation [119], and sequencing platforms [120] frequently introduce biases.

In contrast to RNAseq, batch effect issues are recurrent in gene expression microarrays and often require cross-batch and cross-sample normalization methods that don’t necessarily correct for technical variations, while the identification of differentially expressed genes is highly dependent on these analytical methods [122] and are notoriously platform (or even probe) specific [121]. Moreover, microarrays often present probe specificity issues [121]- due to hybridization steps and/or the lack of appropriate probe design taking into account isoforms. Nanostring technology can also be rate limiting and highly expensive to measure the entire transcriptome in addition to requiring distinct analyses from those used in expression arrays or RNAseq [118]. Lastly, protein-protein interaction datasets are incomplete and limited to more studied cases [123]. Data mining can be improved when explicitly modeled with the specific biases present in different datasets. For example, integrating expression data from high coverage RNAseq provides the opportunity to better evaluate a larger dynamic range of mRNA expression and their associated alternative splicing isoforms than expression microarrays.

This review unveiled a major opportunity to use more machine learning methods, as they have only been utilized in a limited number of manuscripts to date. Further, none of the reviewed methods addressed intergenic SNPs located far from the protein coding loci. Statistical modeling predominantly employed protein omics datasets as inputs, while straightforward overlap of nodes was the predominant method for locus assignment. Network modeling appears more versatile as it was applied equally to all types of omics datasets. Published network models always employed a larger number of biomolecules than statistical methods, suggesting a lack of predictive power of the former on smaller datasets. Here, we provide more details on similarity methods. We visualized the distribution of biologically informed similarity metrics according to multiple variables, such as the computational methods and the types and counts of biomolecules or diseases. These results demonstrate the efficacy of naïve as well as complex similarity metrics between biomolecular functions. The common theme among these metrics is the integration of various interaction types of data capturing the cellular variation, such as genetic polymorphisms, GO, protein complexes, and long-range interactions among biomolecules. Interestingly, devising biologically informed metrics or solely employing computational approaches could converge to similar findings. For example, Lee et al. highlighted the role of biologically inspired metrics thorough finding that metabolic diseases rarely share genes; however, they interact through metabolic fluxes and their similarities reflect their comorbidities [27]. Without making this assumption, Li and Patra reached the same conclusion by applying a random walk algorithm to the phenotype-genome network. We also showed how the different computational methods could integrate cellular and phenotypic data [101]. The performances of the computational methods have been primarily assessed in the context of gene prioritization. Relative to the algorithms strictly relying on the shortest topological distances, random walk and diffusion-based ones perform better, as they evaluate the global topology of the network while profiling the similarity of a disease and its known associated genes relative to other genes [124]. Additionally, some studies showed integration of the disease phenotypic similarities enhancing their performances [96, 125] since the primary premise implies that the genes associated with a disease also underlie the pathogenesis of those phenotypically similar. Different frameworks and methods and their relative performances for integrative data analysis have been covered in previous reviews [35-41]. Despite the significant contribution of a disease similarity framework to uncover the underpinning molecular mechanisms of comorbidity, its current formulation needs the integration of other types of data to explain how the interaction between different cellular domains underpins comorbidity. Omics data provide functional and physical interactions be-
between various molecular tiers inferred from biochemical measures rather than functional associations based on knowledge base and correlations studies. The integration of omics data with different degrees of specificity and sensitivity [126-128] enhances the performance of modeling [129-131] and optimizes accuracy and recovery of more comprehensive disease mechanism similarities. For example, the use of eQTL data (expression quantitative trait loci) [132, 133] linking mRNA expression to SNPs enabled the recovery of not only shared mRNAs among disease-associated SNPs but also their shared downstream biological pathways and their relative interactions [134]. Other studies showed that disease similarity stemmed from shared regulatory mechanisms through the mapping of disease-SNPs with transcription factor (TF) binding sites or genomic regulatory elements such as enhancers (e.g. ChIP-seq, DNaseq) along with RNA gene expression (RNAseq) and long-range chromatin interaction regions (Chla-PET) [135-138]. Cell type-specific interaction networks have also been used to uncover disease similarity [139, 140] [141-144]. Moreover, the use of other regulatory molecular players, such as microRNAs and lncRNAs, improves the formulation of the interaction between diseases for better mechanistic explanation of comorbidity.

This review is limited to recent scientific publications that met stringent inclusion criteria of a minimum of five analyzed diseases. There is an overabundance of manuscripts using rate-limiting biological or curation approaches to identify similarity of mechanisms between two diseases, but most of these would not scale up affordably to a large number of disease comorbidities. This review is therefore steered by design towards computational methods.

5 Conclusion

The premise of biomodularity of diseases implies that diseases should no longer be treated as simple traits due to their complex phenotypic presentation and genetic pleiotropy. This review highlights the significance of improving our understanding of disease’s underpinnings using similarity metrics that uncover shared molecular mechanisms. Future studies shall benefit from the rapid increase of large cohorts and personal omics data to infer biological activity and function, especially using previously overlooked intergenic loci. It is also important that future studies model genetic and biomolecular interactions as non-linear interdependencies may modulate the severity of comorbidities. Finally, the use of detailed phenotypes and availability of clinical databases provide new opportunities to develop more robust and comprehensive diseases similarity metrics that account for the spatio-temporal epidemiology of disease comorbidity.

Acknowledgments

Dr. Lussier is supported in part by The National Library of Medicine (K22 LM008308-04), The University of Arizona Cancer Center (P30CA023074), the Center for Biomedical Informatics and Biostatistics of the University of Arizona Health Sciences, and the BJO5 Institute.

References


Correspondence to:
Dr. Yves A. Lussier
The University of Arizona
Bio5 Building
1657 East Helen Street
Tucson, AZ 85721
USA
Fax: +1 520 626 4824
E-Mail: Yves@email.arizona.edu

For personal or educational use only. No other uses without permission. All rights reserved.