Protein interactions and disease: computational approaches to uncover the etiology of diseases

Maricel G. Kann

Submitted: 9th April 2007; Received (in revised form): 6th June 2007

Abstract

The genomic era has been characterised by vast amounts of data from diverse sources, creating a need for new tools to extract biologically meaningful information. Bioinformatics is, for the most part, responding to that need. The sparseness of the genomic data associated with diseases, however, creates a new challenge. Understanding the complex interplay between genes and proteins requires integration of data from a wide variety of sources, i.e., gene expression, genetic linkage, protein interaction, and protein structure among others. Thus, computational tools have become critical for the integration, representation and visualization of heterogeneous biomedical data. Furthermore, several bioinformatics methods have been developed to formulate predictions about the functional role of genes and proteins, including their role in diseases. After an introduction to the complex interplay between proteins and genetic diseases, this review explores recent approaches to the understanding of the mechanisms of disease at the molecular level. Finally, because most known mechanisms leading to disease involve some form of protein interaction, this review focuses on the recent methodologies for understanding diseases through their underlying protein interactions. Recent contributions from genetics, protein structure and protein interaction network analyses to the understanding of diseases are discussed here.

Keywords: computational biology; diseases; genes; phenotype; proteins; protein interactions

INTRODUCTION

‘Life is a relationship between molecules, not a property of any one molecule. So is therefore disease, which endangers life’, wrote Zuckerkandl and Pauling (1962) in their chapter on ‘Molecular disease, evolution and genic heterogeneity’ [1]. Over 40 years later, we are still far from unraveling the molecular mechanisms of most diseases and pondering about the role of molecular interactions on healthy and diseased organisms. Indeed, proteins do not function in isolation but rather within the cell, interacting mostly with other proteins but also with other molecules such as DNA, RNA and small molecules. Thus, studies of proteins and their interactions are essential to understand their role within the cell. Here, the term ‘protein interaction’ includes a great range of events, such as transient and stable complexes, as well as physical and functional interactions.

The focus of this review is protein interactions and their role in understanding diseases. The main topic is divided into three fields of research, addressed in three sections. The first section reviews the association of genes or proteins and their interacting partners with a particular disease. The second addresses the structural analysis of disease-related proteins, protein complexes and their mutants. The third section covers the analysis of the global properties of the protein interaction networks, i.e., those related to diseases. The methods...
and hypotheses presented here were formulated for general application to any kind of disease. When it aims to illustrate an idea or application, a particular disease is singled out and a brief description of the disease is provided. Readers are referred to the cited work for more details.

This review offers a computational perspective on a broad emerging field that considers the role of protein interactions in the etiology of diseases and the generation of new hypothesis derived from this knowledge.

**Proteins and genetic diseases**

This section provides a brief introduction to phenotype–genotype association studies and an overview of recent computational methodologies that prioritise disease-related genes. Gene-phenotype association and protein interaction studies are intimately related. Uncovering the mechanism by which genes (and their interactions) cause disease reveals information about the interplay between their corresponding protein products. Conversely, protein interaction studies play a major role in the prediction of new gene-phenotype relationships.

**From genes to phenotype**

Progress in genetic studies towards the association of phenotype with genotype has led to the identification of an increasing number of genes that contribute to human disease. Mendelian traits or diseases, named after Gregor Mendel, are those inherited and controlled by a single gene. This gene can be isolated based on its position in the chromosome by a process known as positional cloning [2]. Some examples of human disease-related genes which were identified by positional cloning are the genes associated with cystic fibrosis [3, 4], Huntington disease (HD) [5] and breast cancer susceptibility [6, 7]. The first step of positional cloning is linkage analysis, in which the gene is mapped using a group of DNA polymorphisms from families that segregate the disease phenotype [8]. Once the gene that predisposes a disease is identified, its protein products and mutations can be studied to clarify the nature of the disease process. Even in simple Mendelian diseases, however, the correlation between the mutations in the genome of the patient and the symptoms might not be clear [9]. Several reasons have been suggested for this apparent lack of correlation between genotype and phenotype, as illustrated in Figure 1 [10–12]. Among them are pleiotropy (the ability of some genes to produce multiple phenotypes), environmental factors and the influence of other genes. Genes could influence each other in several ways: they can interact synergistically, one could mask the phenotypic effect of the other (phenomenon known as epistasis), or a gene could modify another gene (having a small quantitative effect on the expression of the other gene). For instance, cystic fibrosis and Becker muscular dystrophy, previously considered classical examples of Mendelian pattern of inheritance,
are believed to be caused by a mutation of one gene modified by other genes [13, 14]. These observations lead to the evolving concept of oligogenic diseases, which require the interaction of a few genes, presenting inheritance patterns somewhere between monogenic and polygenic (reviewed in [15, 16]). This and other studies have demonstrated that even simple Mendelian diseases can lead to complex genotype–phenotype associations [12].

To add to the complexity, most common diseases such as cancer, metabolic, psychiatric and cardiovascular disorders (e.g. diabetes, schizophrenia and hypertension) are believed to be caused by several genes (multigenic) and affected by several factors including environmental ones (e.g. diet, infection by bacteria) [17]. Despite an increasing understanding of the multigenic inheritance, the study of these complex diseases remains challenging [18].

**From genes to protein complexes and back**

One of the main challenges scientists face today is deciphering the molecular details that lead to diseases. Even when the genetic basis of a disease is well understood, not much is known about the molecular mechanisms leading to the disorders. For oligogenic diseases, synergistic contribution of genes from several loci could explain disruptions in their products, in particular when these proteins are directly or indirectly interacting. Two models, namely the dosage [19] and the poison [20] model, have been used to explain the molecular mechanisms of the disruption (reviewed for oligogenic diseases in [16]).

The dosage model explains disruptions of two proteins within a complex. Mutations in one protein alone weaken the interaction but do not affect the phenotype. Only when the two proteins are mutated, the complex is not formed and the phenotype is affected. For instance, mutations that affect ligand-receptor interactions could be explained with such a model.

In the poison model, mutations in one of the proteins disrupt the complex but enough of the unchanged complexes are still available to maintain the function. Addition of another mutated subunit will further decrease the already reduced number of normal complexes, resulting in phenotype changes. The molecular models described earlier could be also used to explain indirect interactions between proteins (i.e. proteins that do not physically interact but participate in the same functional pathway).

The increasing knowledge about protein networks can be used towards identifying new genes and genetic mechanisms behind diseases. For instance, if the gene products (proteins) have any functional interaction, one could trace these proteins back to their respective genes and identify the genes responsible for the disease. Identifying genes associated with complex diseases from all possible candidates generated from genome-wide genetic linkage studies would involve searching through hundreds of genes. Several computational approaches to prioritise genes related to diseases have been developed to aid linkage analysis and association studies. Some of these methods rely on sequence and functional differences between disease-causing proteins and others not related to any disease. For example, sequence-based properties such as length, conservation across species and number of paralogs have been used to create disease classifiers [21–24].

Other methods tend to integrate several sources of data, like gene expression, gene ontology (GO) annotation, and disease phenotype annotation from the Medical Subject Headings (MeSH) and Online Mendelian Inheritance in Man (OMIM) [25–36]. The performance of seven of these computational methods [21, 23, 25, 30–32, 36] was recently reviewed by Tiffin et al. [37] and applied to the analysis of candidate genes for type 2 diabetes (T2D) and obesity genes. The authors identified a set of primary candidates with nine T2D genes and five obesity genes (selected by six out of seven methods). They also generated a secondary set of 94 and 116 gene candidates for T2D and obesity, respectively (found by five out of seven methods), of which 58 of these genes were common to both diseases. This study reviews seven independent computational methods and illustrates how these methodologies can be used to identify genes related to complex diseases, followed by an interesting discussion on the integration of results from different methodologies.

Several of these methods explicitly incorporate protein interaction data. Oti et al. [38] specifically analysed the effect of incorporating protein–protein interaction into the prediction of disease-causing gene candidates. Their results suggest that the inclusion of interaction data will result in approximately a 10-fold improvement on gene candidate identification. These methods are limited by the quality and sparseness of the experimental protein interaction data. Advances in the experimental
and computational approaches towards an accurate identification of the interactions will have an impact on the enhancement of these methods.

The problem of gene-phenotype association is complicated further by the fact that not only mutations in multiple genes could cause one disease but multiple syndromes could be caused by mutations in the same gene. These provide an explanation for the continuum of syndromes and their overlap (illustrated in Figure 2) [39]. For instance, the overlap of human malformation syndromes leads to the concept of ‘syndrome families’ [40]. It is possible that these diseases arise from the disruption of strongly related genes (i.e. in the same protein complex or pathway). For example, Fanconi anemia and Usher syndrome type 1 both represent diseases produced by mutations in several genes involved in the same complex [41, 42]. Sam et al. [43] have developed a method to compare diseases based on their shared network of protein interactions. Upon manual examination of significantly correlated disease pairs found by this method, the authors confirmed that, for several of them, the correlation was previously reported in literature. For instance, Cockayne syndrome and Xeroderma Pigmentosum are predicted to be correlated through their protein–protein interaction networks as expected from the literature [44].

A recently published method, developed by Lage et al. [45] illustrates the use of knowledge regarding protein interactions to prioritise genes within a linkage interval (these intervals were obtained from genetic linkage analysis data extracted from OMIM and GeneCards databases). The patient phenotype associated with this interval databases, is compared to the phenotype of all disease-related proteins interacting with each of the candidate proteins (gene products). A Bayesian predictor based on the pairwise score between phenotypes (obtained using text mining techniques) is used to rank all gene candidates extracted from OMIM’s 870 linkage intervals linked to disease. This method, which capitalises on the fact that interacting proteins are often responsible for similar phenotypes, produced several novel putative disease-causing genes. This approach, and related ones, are limited by the reliability of protein networks and the sparseness of protein-disease association data and would greatly benefit from a more accurate description of phenotypes (as described subsequently).

Another major challenge is the integration and organization of phenotypic databases. NIH recently acknowledged this need by launching the whole genome association studies. The NCBI’s database, dbGAP [46] provides open and controlled access to summary and individual data, respectively for several genotype association studies. A recent review by Lussier et al. [47] points to the challenges faced by the emerging field of high-throughput approaches for studying genotype-phenotype relationships, namely the field of phenomics. Advances in this area will depend on robust taxonomies for phenotypes and on the accuracy of their clinical description [48, 49].

**CONTRIBUTIONS OF STRUCTURAL ANALYSIS TO THE UNDERSTANDING OF DISEASES**

In many cases, a clear understanding of the malfunction that ultimately causes one or several diseases can only be achieved when the molecular level of the protein interactions are known. The three dimensional structure of the protein interaction complex, whether available or modeled,
can provide such detail. Furthermore, understanding the binding at such level is critical for the rational design of new therapeutic agents targeted to disrupt interactions that cause disorders. The following section examines the contribution of structural analysis to the understanding of diseases, and provides an overview of several disease-related resources.

### Protein structure, protein complexes and disease

To complement classical structural biology, the structural genomic (SG) projects aim to solve the X-ray or NMR structures in a high-throughput manner. The SG initiative’s goal is to provide three-dimensional structural models for all proteins encoded by complete genomes [50]. Most experimentally derived structures, however, might not be directly related to any human disease [51]. This requires computational homology studies to obtain models for the human or pathogens proteins relevant to diseases. For instance, from over 40,000 proteins with known structure deposited in the protein data bank (PDB) [52], only a few hundred are known to be related to diseases. Several computational approaches have been implemented to predict function from protein sequence and structure information (see reviews in [53, 54]). However, experimental techniques are still needed to validate the functions of these proteins.

Studies of inherited or somatic non-synonymous mutations constitute the main source for the analysis of the etiology of diseases at the molecular level. A distinction should be made among those rare mutations responsible for functional disruption that lead to disease and the large number of common variations in the human genome derived from high-throughput single nucleotide polymorphisms (SNPs) analysis experiments [55]. The majority of the non-synonymous SNPs (nsSNPs), in particular those that are present in a large number of individuals, are probably not associated with any disease. Rare variants (found in a very small percentage of the population), on the other hand, tend to occur on structurally and functionally relevant sites. This suggests that structural information can be valuable for understanding the effect of mutations and nsSNPs [56]. Several computational methods based on stability, evolutionary and structural information have been developed to predict the impact of a mutation on the protein function. Resources related to this methods are listed in Table 1 (see review by Mooney for more details [57]). The main drawback of these methods is their low accuracy, which has been shown to improve with the addition of structural information [58, 59]. Even if the disruption of the function is correctly predicted, none of these methods offer insight on how the mutation affects the function. Table 2 provides a list of several publicly available genomic databases containing disease information. This list is by no

### Table 1: Resources for SNP validation

<table>
<thead>
<tr>
<th>Resource</th>
<th>Website</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs3D [132]</td>
<td><a href="http://www.snps3d.org/">http://www.snps3d.org/</a></td>
<td>Based on structure and sequence analysis.</td>
</tr>
<tr>
<td>MutDB [133]</td>
<td><a href="http://mutdb.org/">http://mutdb.org/</a></td>
<td>Provides structural and functional annotation</td>
</tr>
<tr>
<td>PolyPhen [56]</td>
<td><a href="http://genetics.bwh.harvard.edu/pph/">http://genetics.bwh.harvard.edu/pph/</a></td>
<td>Uses straightforward physical and comparative considerations</td>
</tr>
<tr>
<td>SIFT [135]</td>
<td>blocks.fhcrc.org/sift/SIFT.html</td>
<td>Based on sequence homology and the physical properties of amino acids</td>
</tr>
</tbody>
</table>

### Table 2: Databases with disease annotation

<table>
<thead>
<tr>
<th>Database</th>
<th>Website</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swissprot</td>
<td><a href="http://www.ebi.ac.uk/unisprot">http://www.ebi.ac.uk/unisprot</a></td>
<td>A database of protein sequences with disease annotation</td>
</tr>
<tr>
<td>PhenoGo [61]</td>
<td><a href="http://www.phenogo.org">http://www.phenogo.org</a></td>
<td>Computed database that provides phenotypic context to existing associations between gene products and gene ontology (GO) for multiple organisms</td>
</tr>
<tr>
<td>PharmaGKB [64]</td>
<td><a href="http://www.pharmgkb.org">http://www.pharmgkb.org</a></td>
<td>Pharmacogenetics research database</td>
</tr>
</tbody>
</table>
means exhaustive. For instance, for well-studied diseases such as cancer, there are several disease-specific resources available that might or might not be encoded in the data sources listed subsequently (e.g. GeneCards, described subsequently, includes information from CGAP, the National Institute of Cancer’s Cancer Anatomy Project). The OMIM database [60], manually curated and updated daily, is one of the largest catalogs of human genes and disorders. As part of the NCBI Entrez database, OMIM is freely available and contains over 11 000 genes with known sequence and over 6 000 phenotypes. It should be noted that only a few hundred of the genes with known sequences currently annotated in OMIM have known phenotypes. Automatic approaches for linking genotype with phenotype information have the potential to overcome the data scarcity problem inherent in manual efforts. To this effect, several approaches (such as PhenoGo [61] that use natural language processing in combination with GO data) have been developed to create a collection of over 500 000 phenotype-GO associations, including approximately 33 000 genes from 10 species. Similarly, Gene2Disease automatically assigns priorities to genes related to a disease, and provides a list of candidates based on PubMed MeSH terms and GO. Another resource, Genecards [62], provides a suite of tools that integrate information from over 70 sources including OMIM, constituting a single location to retrieve available information for over 24 000 genes including relationships to diseases when available. The PhenomicDB [63] database uses associated orthology relations to provide multi-species genotype-phenotype mappings across human and several model organisms. The Orthodisease database provides a cluster of more than 3 000 disease genes comprising 26 Eukaryotic organisms. Swissprot is a database of protein sequences that includes disease annotations for about 2 600 of its 270 000 entries (16 600 are for human proteins). Finally, PharmaGKB [64] is a catalog of over 300 genes and 400 diseases (with genes involved in drug response), providing a single platform to study relationships between drugs, diseases and genes. Users will find that most of these databases are freely available (GeneCards is limited to non-profit institutions) and their interface varies in flexibility and convenience. Almost all of them can be easily searched using related words in the query (disease or gene). In addition, the use of standard vocabularies and ontologies within all these databases needs to expand beyond GO, so that descriptions of disease phenotypes, cytological changes, and molecular mechanisms can be well-defined and standardised for better discoverability, correlations and mining. In general, while these databases provide an excellent resource, only a small proportion of the genomic data known to be involved in an inherited disease have both known gene sequence and known phenotype. Despite the scarcity of the structural data related to disease, Moul and collaborators have shown that for a set of genes associated with monogenic diseases, the loss of protein stability is a major factor contributing to disease [65].

Structural data, when used in combination with information about mutations responsible for disease, could be essential in unveiling the molecular mechanisms leading to diseases. A study by Thornton and collaborators showed that the patterns of mutations of residues associated with human inherited diseases (from OMIM database) are different from that of the large number of nsSNPs (from dbSNP database) [66]. Their results showed that mutations that lead to major changes in hydrophobicity were more frequent in the OMIM data than in dbSNP. In addition, they found that disease-related mutations are more likely to be buried in the protein structure than what would be expected for the average protein residues. These results are consistent with several other studies, including Viktup et al. [67] and a more recent study by Ye et al. [68]. Conversely, common nsSNPs are less likely to be in protein cores than expected on average, a feature that could be useful when predicting the functional impact of an nsSNP (discussed in the previous subsection).

On the other hand, the protein core represents only a fraction of all residues. Consequently, many of the disease related mutations lie in solvent accessible sites, suggesting that the analyses of these mutations might also shed a light on the mechanisms of the disease. For instance, Thornton and collaborators estimated that more than half of the disease-related mutations analysed in their study occur at solvent accessible sites [66]. In their analysis, Ye et al. found that disease-related mutations located in the protein surface tend to be clustered, forming surface patches, while SNPs are uniformly distributed [68]. These could explain the role of mutations in disease since mutations in the binding site would likely disrupt the protein interaction and function. To reach
this conclusion, the authors compared the location and distribution of disease-related mutations with nsSNPs on a set of protein domains obtained through homology modeling of disease-related proteins. The authors verified that, for a smaller subset of experimentally determined structures, disease-related mutations are located mostly on the binding interfaces of proteins.

Protein structural analysis has helped to elucidate the molecular basis of several diseases. For example, a protein interaction disruption in Von Hippel-Lindau syndrome (VHL), a common mutation from Tyrosine to Histidine at residue 98 (represented as Tyr98His), which is part of the binding site, disrupts the binding of the VHL protein to a protein called the hypoxia-inducible factor (HIF). As a result, the VHL protein no longer degrades the HIF leading to the expression of angiogenic growth factors and local proliferation of blood vessels [69, 70]. Another example, given its central role in cancer, is the extensive study of mutations of the p53 tumor suppressor protein. Structural analysis of mutations in p53 could facilitate the dissection of their functional role, in particular their effect on DNA-binding that seems to be key in human cancers [71, 72]. For instance, a mutation in the DNA binding region (Arg273His) has been associated with Li-Fraumeni syndrome and low p53 DNA binding. On the other hand, a mutation of an Arginine in position 175 to Histidine is important in the stabilization of the protein which might regulate the binding to DNA.

In some cases, interactions between two proteins might involve order-disorder transitions in partially disordered regions of the interacting proteins during the binding process. These unstructured or disordered regions have been found to be involved in many disease mechanisms [73–75] (see [76] for a review on intrinsically disordered proteins). For instance, the cancer suppressor BRCA1 has been shown to contain intrinsically disordered regions through which it binds to several proteins [49]. Similarly, some bacteria pathogens’ surface proteins contain intrinsically disordered protein regions. The structural analysis of the host-pathogen protein interactions constitutes an excellent system for targeting by drug designers. The NMR determination of the structural complex of one of these surface proteins, namely the streptococcal fibronecting-binding protein (FnBP), bound to the human fibronectin provided the mechanistic details on how the bacterial target the host cell (see review [77]).

**DISEASES AND PROTEIN INTERACTION NETWORKS**

This section explores the study of protein networks, with a focus on protein–protein interactions, and their impact on understanding diseases. First, it provides an overview of the experimental and computational approaches that have been used to reconstruct the network of human protein interactions (or human *interactome*). It then lists the basic concepts that define the general properties of a network and introduces recent contributions to biology from this theoretical perspective. Finally, recent approaches to create disease-related protein interaction networks are discussed. Examples of experimental and computational methodologies for network reconstruction are provided.

**From one to thousands of interactions**

In the past, experimental techniques were limited to reveal a handful of protein–protein interactions at a time. For instance, genetic, biochemical and biophysical techniques mostly study individual interactions [78]. Recent high-throughput experimental analyses represent a dramatic change in the number of interaction data they generated, making possible the reconstruction of whole genome protein networks (see [79–84] and reviews in [85, 86]). These genome-wide analyses rely on the development of computational approaches to understand and visualise these data. Bioinformatics tools could also generate predictions of new functional roles of proteins from existing genomic data. Therefore, bioinformatics has a dual role in the context of protein interaction and diseases: prediction of putative protein interactions and of new gene-disease associations (see previous section), and development of a framework to integrate, represent, and visualise experimental data.

Computational techniques to predict protein interactions have been developed in parallel with experimental advances. These approaches rely on the fact that interacting proteins are more likely to be present in the same set of organisms [87, 88], to conserve the gene order [89, 90], or be fused in some organism [91, 92]. These methods have been successfully used to predict protein interactions but still have many limitations (see reviews [93–96]). The assumption that interacting proteins co-evolve to preserve their function has led to methods that rely on similarities between the evolutionary histories of interacting protein families to predict interacting
partners [97–104]. These methods are widely applicable and only require the protein sequence as input. However, the signal from functional co-evolution can sometimes be difficult to detect, resulting in low accuracy in the predictions. Addressing this problem, this technique was recently improved by subtracting the signal from speciation events of unrelated sequences [105, 106] and removing high entropy regions (i.e. regions poorly conserved across species) of the sequences [107].

**Proteins meet graph theory**

Protein–protein interaction data obtained from high-throughput experimental approaches can be represented as a graph [108, 109]. Proteins constitute the nodes of this graph and interactions between the proteins are represented as lines connecting the nodes. Biological networks have been found to be comparable to communication and social networks. Protein–protein interaction and communication networks share several commonalities, such as scale-free and small-world properties (see definitions in Table 3) [110]. Scale-free networks are fairly robust against random errors but are highly vulnerable to perturbations in highly connected nodes [111]. Certain properties of the protein network could be used to differentiate disease from non-disease proteins. Based on this approach, Xu et al. [112] devised a classifier based on several topological features of the protein interaction network to predict genes related to disease. The classifier was trained on a set of non-disease genes and one of disease genes (from OMIM) and applied to a set of over 5000 human genes. As a result, 970 disease genes were identified with 792 of them already listed in OMIM. Some of the 178 newly predicted disease gene candidates were validated by biological experiments.

Protein interaction networks could be used to improve functional annotation since the function of some proteins could be inferred from their role in pathways or protein complexes [113, 114]. Likewise, information about key nodes on disease-related networks could be used in drug discovery. Drug target identification constitutes a good example of the potential of integrating structural data with high-throughput data [115]. The structural details on the binding or allosteric sites could be used to design molecules that affect protein function. On the other hand, the reconstruction of the different protein networks in which the potential target is involved (signaling, metabolic, regulatory, etc), is needed to predict the overall impact of the disruption. If, for example, the target is highly connected (a hub), its inhibition may affect many activities that are essential for the proper function of the cell and is therefore unsuitable as a drug target. Less connected nodes affecting mainly the pathway that leads to disease, on the other hand, could constitute vulnerable points of the disease-related network, thus, they are better candidates for drug target. Ultimately, a more complex system biology approach that integrates and mathematically models the gene, protein and pathway responses would be needed to fully characterise the effects of the system disruptions caused by the drug.

Reconstructing the interaction networks of proteins and its mutants involved in a disease might be the key to understanding the differences between healthy and disease organisms. Recent work by Goehler et al. [116] on HD illustrates the potential of those approaches. HD is an autosomal dominant disease.

**Table 3: Glossary of terms**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single nucleotide polymorphism (SNP)</td>
<td>Sites in the human genome where individuals differ in their DNA sequence, often by a single base, usually with population frequencies greater than 1%.</td>
</tr>
<tr>
<td>Non-synonymous SNP (nsSNP)</td>
<td>SNP that produce an amino acid substitution.</td>
</tr>
<tr>
<td>Coding SNP (cSNP)</td>
<td>SNP located within the coding regions of genes.</td>
</tr>
<tr>
<td>Mendelian disease</td>
<td>Inherited disease with an inheritance pattern based on a single affected gene. The most common cause of Mendelian disease is caused by a cSNP.</td>
</tr>
<tr>
<td>Complex disease</td>
<td>Inherited disease caused by a number of factors and therefore do not follow simple Mendelian rules of inheritance.</td>
</tr>
<tr>
<td>Epistasis</td>
<td>When the action of one gene is modified by one or several genes.</td>
</tr>
<tr>
<td>Pleiotropic gene</td>
<td>A gene that affects more than one phenotype.</td>
</tr>
<tr>
<td>Moonlighting enzymes</td>
<td>A case of pleiotropic gene that produces metabolic enzymes with additional functional activities.</td>
</tr>
<tr>
<td>Network</td>
<td>Series of points or nodes interconnected by edges, edges can have direction or different weights. Ability of two nodes to interact with each other.</td>
</tr>
<tr>
<td>Diameter</td>
<td>Highly clustered networks with relatively short distance between any two nodes. The majority of the nodes have few links and a few of them (hubs) are highly connected.</td>
</tr>
<tr>
<td>Small-world networks</td>
<td></td>
</tr>
<tr>
<td>Scale-free networks</td>
<td></td>
</tr>
</tbody>
</table>
neurodegenerative disease. Currently, there is no pharmacological treatment to prevent the progression of this rare inherited disorder [117]. HD is caused by the repeat expansion of the trinucleotide CAG in the Huntingtin (Htt) gene and is one of several polyglutamine (or polyQ) diseases. This expansion causes aggregation of the mutant Htt in insoluble neuronal inclusion bodies which consequently leads to neuronal degeneration. Goehler et al. [116] reported an experimental strategy to generate the protein–protein interaction network of all proteins related to HD, revealing many new interactions and permitting the functional annotation of several uncharacterised proteins. Most importantly, they discovered an interaction of the Htt with GIT1, a GTPase-activating protein which seems to be required for the Htt aggregation. Upon further validation, the GTI1 could constitute an excellent target for therapeutical strategies [118, 119]. Towards a similar goal, Lim and collaborators [120] developed the network of the interactions among proteins related to ataxias and disorders of Purkinje cell degeneration. They found that most of the proteins related to ataxias interact directly or indirectly with each other. A more recent study corroborates these findings across all of the disease proteins from OMIM [121]. Thus, proteins related to a disease are more likely to interact with proteins already known to cause similar diseases. This motivates several of the gene prioritization computational studies presented in the previous section. Chen et al. [122] presented a computational approach to test and confirm this principle for the subnetwork of interacting proteins associated with Alzheimers disease (AD) (see [123] for more details about AD and other diseases associated with aggregates, namely amyloid fibrils, that result from protein misfolding). Chen et al. [122] devised a computational method to enrich the AD’s subnetwork based on a heuristic score. The score prioritises proteins with high specificity (favoring the addition of low promiscuously connected proteins) and with high confidence on their interaction data (this weighting addresses the problem of unreliability of the interaction data). In an attempt to derive common features among cancer proteins, Jonsson and Bates [124] performed a systematic computational study of a subset of proteins related to cancer. The authors found that the network topology of the cancer related proteins is quite different from those not involved in the disease, i.e. it was found that cancer proteins are highly connected with other cancer-related proteins. In addition, a study of the protein network of herpesvirus performed by Uetz et al. [125] indicates that viral networks differ significantly from cellular networks, which raises the hypothesis that other intracellular pathogens might also have distinguishing topologies. In a recent study, Goh et al. [126] explored the properties of the human disease network for all known phenotype disease gene associations. The authors found that genes that are essential in early development tend to encode highly connected proteins (hub proteins). Surprisingly, their results suggest that the vast majority of disease-related genes are non-essential and show no tendency to encode hub proteins.

CONCLUSIONS AND FINAL REMARKS
Protein interactions are involved in metabolic, signaling, immune and gene-regulatory networks. A better understanding of protein interactions, either with other proteins or with DNA, RNA, membrane or small molecules, could reveal the molecular mechanism of the processes leading to diseases. Mutations in the protein interaction interface (or related sites, e.g. active sites, allosteric binding sites) could evidently disrupt the protein interaction. For example, the etiology of several of the diseases mentioned in this review lies in the disruption of the protein–DNA interaction (e.g. p53 in cancer). For other diseases, the main cause is the disruption of the stability or protein-folding, thus destroying one or several protein–protein interactions (e.g. VHL), or creating new undesired ones (e.g. unfolded proteins tend to aggregate, as in AD and HD). Clearly pathogen-host protein interactions are central for bacteria or viruses hijacking the host immune system.

The study of the phenotypic commonalities of several disorders points to common modules that are responsible for inherited diseases. Evidence of these modules, namely protein complexes or functionally related proteins, emerges from several studies of protein and gene interactions reviewed here. However, several other factors could influence the final disease outcome. These confounders challenge the concept of modularity of genetic diseases (see [39]). Understanding the role of these factors, as illustrated in Figure 1, is a challenge central to the etiology of both Mendelian and complex diseases.
The processes leading to diseases are extremely complex, and so are the proteins and interactions involved in them. Most of the methodologies reviewed here used a simplistic ‘static’ view of the protein and their networks. In reality, proteins are continuously being synthesised from and degraded into amino acids. The kinetics of processes and network dynamics need to be considered to achieve a complete understanding of how the disruptions of proteins and their interactions lead to disease. Finally, it is important to also consider the context-specific (tissue, disease stage and response) functions of protein interactions.

It is clear, then, that a gap still exists between the identification of the disease-associated protein interaction network and the complete understanding of the disease mechanism. The gap is filled, unfortunately, with more questions than answers. The approaches reviewed here generated a considerable amount of valuable data, but also the need for further validation. To that purpose, a number of studies have used data from simpler organisms, such as worm (Caenorhabditis elegans), fly (Drosophila melanogaster) or yeast (Saccharomyces cerevisiae). There are limitations, however, to the transfer of interaction annotation across species, in particular those from distantly related organisms [127]. A study of the overlap between the interaction networks of fly, worm, yeast and human data showed that there are only a few conserved interactions among these organisms [121]. Despite these limitations, modeling different aspects of a disease in simpler organisms has proven to be extremely useful. For instance, several aspects of the polyQ diseases (see previous section) were modeled using worm, fly and yeast [128–131].

We are still far from the goal of understanding the etiology of most diseases, further advances on relevant experimental technology, i.e. genetic linkage, protein interaction, protein structure, gene expression, along with computational tools to organise, visualise and integrate these data will provide a step forward in that direction. In particular, the completion of the human protein interactome will provide data that could enhance several of the methodologies reviewed here. In addition, a systematic experimental genome-wide study of protein interactions between host and pathogen, which is not yet available in the literature, could provide insight into the bacteria, virus or parasite mechanisms of pathogenicity. In addition, valuable information about disease and protein interactions is buried within millions of biomedical records. Text mining approaches are therefore essential to recover such information. Indeed, several of the databases and methods to prioritise disease-related genes discussed in previous sections (e.g. Lage et al. [45]) have successfully incorporated text mining techniques.

Ideally, since network and structural approaches are complementary, the combination of network studies with a more detailed structural analysis has the potential to be an excellent framework for the study of disease mechanisms and rational design of drugs. In future, this strategy, and others discussed here, should be integrated into multidisciplinary disease-specific projects that provide a better understanding of a particular disease and help identify disease modules (if any) that are common to related disorders.

### Key Points
- Disruption of an existing protein interaction (by changing the stability of the protein and/or inhibiting the ability to bind to other molecules), production of new undesirable interactions (through mutations that result in misfolding of the protein and aggregation) and disruption of protein-DNA interactions (by affecting gene regulation) are the causes of many diseases.
- Mutations in one gene could affect other genes. If their protein products interact, the resulting disruption of this interaction could lead to a disease.
- Analysis of disease-related protein networks confirms that proteins involved in a disease tend to interact with other proteins involved in the same disease.
- A gene could have a role in several diseases, thus, many diseases could share interaction subnetworks.

### Funding
Support for this work was provided by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

### Acknowledgments
Thanks to all the scientists (included or not in this review) that contributed with their excellent work to the field of research reviewed here. Many thanks to Sonia Leach, Leonardo Marino, Eric Neumann, Anna Panchenko, Pedja Ravidojac, Willy Valdivia, Adam Godzik and anonymous reviewers for their helpful comments on the manuscript, to the NIH Fellows Editorial Board for their editorial work, and to Robert Yates for his help in the graphic design of the pictures.

### References


