Published in final edited form as:

Expert Rev Proteomics. 2009 December ; 6(6): 647. doi:10.1586/epr.09.86.

Role for protein–protein interaction databases in human genetics

Kristine A Pattin, PhD Candidate (BS Biology) and

Computational Genetics Laboratory and Departments of Genetics, Dartmouth Medical School, Lebanon, NH, USA

Dartmouth-Hitchcock Medical Center, Norris Cotton Cancer Center, 1 Medical Center Drive, Lebanon, NH 03756, USA Fax: +1 603 653 9900 kristine.a.pattin@dartmouth.edu

Jason H Moore, PhD[†]

Computational Genetics Laboratory, Departments of Genetics, Community and Family Medicine, and Norris-Cotton Cancer Center, Dartmouth Medical School, Lebanon, NH, USA

Department of Computer Science, University of New Hampshire, Durham, NH, USA

Department of Computer Science, University of Vermont, Burlington, VT, USA

Translational Genomics Research Institute, Phoenix, AZ, USA

706 Rubin Building, HB 7937, 1 Medical Center Drive, Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756, USA Tel.: +1 603 653 9939, Fax: +1 603 653 9900

jason.h.moore@dartmouth.edu www.epistasis.org

Abstract

Proteomics and the study of protein–protein interactions are becoming increasingly important in our effort to understand human diseases on a system-wide level. Thanks to the development and curation of protein–interaction databases, up-to-date information on these interaction networks is accessible and publicly available to the scientific community. As our knowledge of protein–protein interactions increases, it is important to give thought to the different ways that these resources can impact biomedical research. In this article, we highlight the importance of protein–protein interactions in human genetics and genetic epidemiology. Since protein–protein interactions demonstrate one of the strongest functional relationships between genes, combining genomic data with available proteomic data may provide us with a more in-depth understanding of common human diseases. In this review, we will discuss some of the fundamentals of protein interactions, the databases that are publicly available and how information from these databases can be used to facilitate genome-wide genetic studies.

Keywords

epistasis; expert knowledge; multifactor dimensionality reduction; protein–protein interaction; single nucleotide polymorphism

Sequencing complete genomes was a dream that has now become a reality. However, we are still striving to gain a complete understanding of the hundreds of thousands of proteins that are encoded by the fewer than 30,000 genes that comprise the human genome. The word ‘proteome’ was first introduced in 1995 and described as the ‘total protein complement of a

genome' [1], and it was then proposed that mapping the complete human proteome would help us to dissect the biochemical and physiological systems that influence complex human diseases at the molecular level. The concept of the proteome and the field of proteomics are rapidly developing, as new technology and high-throughput techniques are making the mapping of the entire human proteome seem less like a dream – as the complete sequencing of genomes once was – and more of a possible reality.

Similarly, the advent of proteomics did not put a halt to technological progress in the field of genomics. Measuring 1 million or more single nucleotide polymorphisms (SNPs) across the entire human genome is now technically and economically possible. As a result, the domain of human genetics is experiencing an information explosion; yet, at the same time, there is a lack of methods to analyze this information. That is, our ability to generate genetic data is far outpacing our ability to make sense of it. It is true that there are several important challenges that must be overcome before we can take full advantage of genome-wide data in order to decompose the genetic architecture of a complex trait into its interacting components [2].

The idea that the proteome is a complement of the genome is important and suggests that there is a strong relationship between the two. One criticism against genomics has been that sequence information provides us with a basic 'snapshot' of how a cell might utilize its genes, when, in reality, the cell is a dynamic entity that reacts and responds to its environment [3]. To follow with the idea of complement, we feel that the field of genomics could benefit greatly by being complemented by proteomic data, specifically by providing a means to deal with the widening gap in understanding that the field is experiencing. To our advantage, information on protein–protein interactions (PPIs) is right at our fingertips in a number of publicly available databases.

The challenges we face conducting genetic analyses are derived from the complexity of the genotype- and phenotype-mapping relationship that results from phenomena such as epistasis (gene–gene interactions) and plastic reaction norms (gene–environment interactions) [2]. First, there is the need for powerful statistical methods to model the relationship between combinations of SNPs and disease susceptibility because traditional methods, such as logistic regression, are not effective when genotype combinations increase exponentially with the number of SNPs being analyzed, and there are not enough subjects to represent each combination. Second, there is the challenge of selecting which SNPs should be included in the analysis, because analyzing all pair-wise and multiple-way SNP combinations in a genome-wide genetic study with thousands of SNPs is computationally infeasible. A third challenge is the interpretation of statistical gene–gene interaction models in the context of human biology, so that they may be used for the benefit of disease treatment and prevention.

The goal of this article is to explore how PPIs may play an important role in genetic research by providing a source of expert knowledge that can be used to guide genome-wide association studies (GWAS), in an effort to ease the computational burden of detecting and characterizing interactions while, at the same time, facilitating biological interpretation of the data. It is already known that the location of SNPs in the genome can have an effect on protein structure by changing amino acid sequences. Whether they are in a coding region of the genome or not, these SNPs may affect PPI, protein expression, alternative splicing, stability, folding, ligand binding or catalysis, thus inducing or influencing the disease state in individuals [4,5]. Several studies have led to the development of tools aimed at understanding the functional effect of SNPs by investigating their impact on protein structure or functional sites and even functional sites at the DNA level [6–8]. We will focus on the role of information extracted from PPI databases for improving the computational efficiency of genome-wide studies of epistasis.

Challenges of detecting epistasis

To begin to understand how protein interactions may play an important role in human genetics research, we need first to understand the concept of epistasis and the challenges encountered when trying to detect epistasis in large-scale studies. Recognized for many years, epistasis has been described from two different perspectives: biological and statistical [9,10]. Biological epistasis, as defined by Bateson who coined the term, results from physical interactions among biomolecules in gene-regulatory networks and biochemical pathways at the cellular level in an individual [11]. Statistical epistasis, as defined by Fisher, is deviation from additivity in a linear mathematical model that describes the relationship between multilocus genotypes and phenotype variation at the population level [12]. While determining the relationship between the two remains a challenge, it is an important endeavor if we wish to infer biological conclusions from statistical results.

Although epistasis has acquired different meanings over the years, it is a concept that exhibits a clear contrast to Mendelian, or single-gene, traits that have been characterized, such as multiple endocrine neoplasia type 1 (MEN1) or Marfan syndrome. MEN1 is a rare autosomal-dominant disorder that displays an autosomal-dominant pattern of inheritance; that is, each patient has a 50% probability of passing the gene defect to their progeny. MEN1 patients present with symptoms such as parathyroid adenoma, gastrinoma and prolactinoma, or other endocrine and nonendocrine tumors, which develop after the inactivation of both *MEN1* gene copies at chromosome 11q12–13 at tissue level [13]. Marfan syndrome is also an autosomal-dominant connective tissue disorder that affects the skeletal, ocular and cardiovascular systems. Marfan syndrome is primarily caused by a mutation in fibrillin-1 (*FBI*), which produces abnormal fibrillin proteins that normally act as extracellular matrix fibrillar components with functions in elastic and nonelastic tissues [14].

Epistasis, along with other phenomena, such as locus heterogeneity, phenocopy and gene–environment interactions, are major sources of complexity in the mapping relationship between genotype and phenotype in common human diseases. However, earlier studies of biological epistasis began in model organisms, such as *Caenorhabditis elegans*, that are still being used to order genetic pathways by means of crossing two genetic knockouts for a particular function and observing the resulting phenotypes [15,16]. More high-throughput studies began to take place approximately a decade ago that involved pairwise analysis of genes to determine combinations that conferred a phenotype, one of the first having utilized a synthetic genetic array analysis in yeast to examine the interactions of a series of different deletion mutants against more than 4000 deletion backgrounds [17]. Additional approaches have used RNAi libraries to systematically knockdown genes in *C. elegans*, given that deletion libraries for yeast are unlikely to be widely applicable to multicellular organisms [18]. While detecting epistasis on the molecular level with high-throughput approaches has led to a greater understanding of complex gene networks, detecting and characterizing epistasis on the statistical level to gain an understanding of the genetic susceptibility to these diseases is not a simple task.

Moore argues that epistasis is a ubiquitous component of common human disease based on the concept that epistasis has been prominent in literature for more than 100 years and many single locus results from linkage and association studies have demonstrated a further lack of replication [19–21]. The completion of the Human Genome Project in 2003 and the International HapMap Project in 2005 gave rise to an abundance of research tools, such as genome-wide genotyping, that allow researchers to conduct GWAS for detecting genetic variants which confer increased or decreased susceptibility to disease and have allowed for the detection of epistasis in a number of diseases, such as diabetes [22], bipolar disorder [23,24] and coronary artery disease [25]. Although the technical details of measuring a large

representative set of SNPs in an accurate and efficient manner are now well established [26], the analytical methods for determining which SNPs are important are in their infancy and are based on important assumptions, such as each SNP having a large and independent effect on disease risk [27]. Traditional methods of analysis, such as linear and logistic regression, have had limited success owing to the sparseness of data in high dimensions. For example, when interactions among multiple SNPs are considered, there are many multilocus genotype combinations that may have very few or no data points. This can lead to an increase in type I and type II errors owing to parameter estimates with very large standard errors. It is evident that we need research strategies that embrace, rather than ignore, the complexity of the relationship between genotype and phenotype [10,28–33].

There have been many methods proposed to identify SNP interactions and their association with diseases. These proposed methods include multifactor dimensionality reduction (MDR) [34], combinatorial partitioning method [35], symbolic discriminant analysis [36], Monte Carlo logistic regression [37], recursive partitioning method [38], focused interaction testing framework [39], backward genotype–trait association [40], Bayesian epistasis association mapping [41], a forest-based approach [42], penalized logistic regression [43], grammatical evolution neural networks [44] and MegaSNPHunter [45]. Each method has its advantages and disadvantages, and while some have shown improved computational efficiency over others, analyzing high-order interactions with any method is a computational burden that has yet to be overcome. To illustrate this, we next discuss MDR, which has been widely applied to detect gene–gene interactions in a number of common diseases.

Multifactor dimensionality reduction was developed as a nonparametric and model-free data mining method for detecting, characterizing and interpreting epistasis in the absence of significant independent effects in genetic and epidemiologic studies of complex traits such as disease susceptibility [33,34,46–50]. The goal of MDR is to change the representation of the data using a constructive induction algorithm to make nonadditive interactions easier to detect using any classification method, such as naive Bayes or logistic regression [49,50]. This is accomplished by first labeling each genotype combination as high or low risk using some function of a discrete end point, such as case–control status. A new MDR variable with two levels is constructed by pooling all high-risk genotype combinations into one group and all low-risk combinations into another group. Traditionally, MDR-constructed variables have been evaluated with a probabilistic naive Bayes classifier that is combined with tenfold cross validation to obtain an estimate of predictive accuracy or generalizability of epistasis models. While the MDR method has proven to be an effective way of detecting epistasis in a number of diseases, to analyze all of the combinations of SNP interactions in large datasets or genome-wide studies would be impractical, even with access to the largest and most powerful computers available.

To illustrate the scope of such an analysis, consider a recent report from the International HapMap Consortium that suggests that approximately 300,000 carefully selected SNPs may suffice to represent all of the relevant genetic variations across the human Caucasian genome [51]. If this is to be regarded as the lower limit of a GWAS, then approximately 4.5×10^{10} pairwise combinations (300,000 choose two) and 4.5×10^{15} three-way combinations (300,000 choose three) would need to be exhaustively analyzed to detect low-order epistasis using MDR. If 10^6 MDR evaluations can be computed each second, then evaluation of each individual SNP would require less than 1 s of computer time. However, computing all two-way and three-way MDR models would require more than 52,000 days of computer time. Access to a 1000-processor supercomputer might reduce this to 52 days, which is within the realm of possibility. However, extending this to all four-way combinations is not computationally feasible. This adds to the many challenges of detecting epistasis on a genome-wide scale [2,32].

As previously mentioned, an important and understandably difficult goal in human genetics is to determine which of the many thousands of SNPs are useful for predicting who is at risk for common diseases. It was close to a decade ago that Risch and Merikangas first seriously proposed the testing of all known SNPs in the human genome for disease association either directly or by linkage disequilibrium with other SNPs [52], and today, this ‘genome-wide’ approach is expected to revolutionize the genetic analysis of common human diseases [32, 53–57]. Currently, it is possible to measure more than 1 million SNPs with a genome-wide human SNP array available from Affymetrix, and Illumina® has released the Human1M DNA Analysis BeadChip, which is capable of profiling 1 million SNPs on a single array across the human genome.

Certainly, with these technologies now available to the scientific community at more affordable costs, it is clear to see how such large amounts of data are being rapidly produced. Unfortunately, owing to the lack of logical methods to summarize this quantity of information within a biological context, investigators are at a loss when they reach the analysis stage in their research. In fact, our ability to measure genetic information, and biological information in general, is far outpacing our ability to interpret it. It is probable that most of these large-scale studies harbor a wealth of information concerning susceptibility genes that can be used to improve the prevention, diagnosis and treatment of common diseases. However, to access this information to our full advantage, we need to address the specific technical challenges that confront researchers in the analysis process, such as the computational limitation of a large-scale genetic analysis with methods such as MDR and our ability to interpret it. We believe that we can overcome this limitation by utilizing expert knowledge from sources, such as PPI databases, that can be used to help guide our analysis as well as to help us to interpret our results in a biological context.

Expert knowledge

Expert knowledge can be defined as existing biological or statistical information regarding the problem at hand that can be incorporated into the analysis process to guide an algorithm in a more directed fashion. For example, when considering SNPs or genes in a genetic analysis, biological expert knowledge may be derived from what is known about the function of biochemical pathways, the gene ontology (GO) or expression information for that gene. Emily *et al.* used experimental knowledge concerning biological networks to narrow the search for two-locus epistases that confer susceptibility to Crohn’s disease, bipolar disorder, hypertension and rheumatoid arthritis. This group utilized the protein-interaction database, Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), which we discuss later, and expert knowledge from protein interactions to guide their search for epistasis. They were able to identify 71,000 high-confidence potential PPIs in the database and, from there, identified all of the SNPs that corresponded to the genes for the relevant proteins. Subsequently, this interaction information was used to prioritize SNPs in a large genomic dataset from the Wellcome Trust Case–Control Consortium that covered a number of diseases. They were able to identify four significant cases of epistasis between unlinked loci in all four diseases [58]. Another approach that similarly utilizes expert knowledge is a pathway-based approach or analysis, which employs computational methods to define sets of genes based on common biological attributes, such as GO or biological pathways. This information is used to define a measure of enrichment of each gene set among disease-associated markers [24]. Shriner *et al.* explained how, in linkage studies of complex traits, testing each candidate gene from every region is a computational challenge, as similarly described for genome-wide epistasis studies. They demonstrate this concept of pathway-based analysis with their commonality of functional-annotation method, which operates by testing individual GO terms for enrichment in candidate gene pools and ultimately ranks genes based on the number of quantitative trait loci regions where genes with such annotations are found. When this method was applied to

published linkage studies that examined the relationship between age of onset of Alzheimer's and BMI, new candidate genes, as well as previously published candidate genes, were identified [59].

In addition to biological information, it is also useful to use prior statistical knowledge to help guide an epistasis analysis. For example, logarithm (base 10) of odds scores from a prior linkage analysis could be used to weight SNPs in certain chromosomal regions higher during a combinatorial epistasis analysis. That is, SNPs from a certain pathway or chromosomal region would be evaluated for interactions with a higher probability than others in the dataset. Statistical knowledge could also come from filter algorithms that explicitly assess the quality of a SNP based on their relationship with the clinical end point. The Tuned ReliefF (TuRF) algorithm is an example of an algorithm that can assign high-quality scores to SNPs involved in complex interactions. The TuRF algorithm uses a nearest-neighbor approach to assess SNP quality and, thus, does not suffer from the computational limitations of an algorithm that explicitly considers combinations of SNPs. As such, it is very useful for preprocessing the data prior to analysis. Once computed, the TuRF scores can be used to select some reduced number of SNPs for combinatorial analysis or can be used to help guide a computational search algorithm [60]. While both biological and statistical expert knowledge may be useful for facilitating genome-wide studies, we wish to exploit the strong relationship between genes and proteins for these purposes. Statistical epistasis may be indicative of biological epistasis and *vice versa*, making this a valid application as well as a potential method to gain an understanding of the relationship between the two.

Protein–protein interactions

Proteomics and the study of PPIs are becoming increasingly important in our effort to understand human diseases on a system-wide level. While mass spectrometry has been a useful technology applied to the discovery of components in protein complexes [61], PPIs have traditionally been measured using a variety of assays, such as immunoprecipitation and yeast two-hybrid (Y2-H). To reconstruct the entire network of PPIs within cells remains a challenge; yet, this is becoming a more approachable problem. The field of proteomics is advancing, and the aforementioned techniques to detect PPIs have been scaled up to measure interactions on a genome-wide level. High-throughput techniques have also been developed to identify protein complexes using affinity pull-down followed by mass spectrometry [62], and systematically constructed double-knockout strains in yeast have proven to be useful for constructing a large-scale view of genetic-interaction networks [63].

To complement these experimental techniques, a number of computational methods have been developed that include algorithms that are capable of predicting interactions. Originally, the computational prediction of proteins was limited to proteins whose 3D structure had been determined, but now, on account of complete genome sequencing, prediction methods can extend to the genomic level, and many of the available systems for predicting PPIs are based on gene-coexpression data, given that two proteins must be spatially and temporally coordinated for an interaction to be meaningful [64], while others have been founded on the concept that if two proteins have co-evolved, then these proteins have a higher probability of interacting within a cell [60,65–67]. One such approach that encompasses the use of sequence information is based on the understanding of domain–domain interactions in proteins, and many computational methods have been developed in the past 8 years to infer domain–domain interactions from PPI databases that are then subsequently used to predict new PPIs. Other recent and common approaches are based on co-occurrence, which assumes a biological PPI based on their frequency of co-occurrence in text or rule-based approaches, which employ predefined phrase-pattern rules [68]. For example, one system, the Protein Interaction Information Extraction system, was developed as an online system for predicting PPIs from

text. Since it has been shown that PPI information exhibits a pattern in articles, machine-learning techniques have been used for many PPI prediction methods, given that machine learning has proven useful for discovering hidden patterns in data. This system uses both co-occurrence and rule-based approaches in a machine-learning framework [68]. Additional prediction techniques are reviewed by Ta *et al.*; these include the association method, as previously mentioned, which determines pairs of correlated domains that co-occur in the PPIs more frequently than by chance, the Maximum Likelihood Estimation technique, used to calculate the interaction probability for all possible domain pairs observed in a PPI dataset, and a parsimony-explanation approach, which uses a specific type of programming to derive a statistical score for domain–domain interactions [69]. Owing to the development and curation of protein-interaction databases, which will be elaborated upon later, up-to-date information on these interactions, both experimentally determined and predicted, is accessible and publicly available to the scientific community.

Some of these techniques have been actively employed in the study of the pathogenesis of Huntington's disease (HD), an autosomal neurodegenerative disorder that causes cognitive impairment, psychiatric problems and motor dysfunction. This inherited disease is caused by the expansion of a polyglutamine tract in the huntingtin (htt) protein. Although this protein was discovered more than a decade ago, constructing the protein-interaction network that it belongs to is still an ongoing process that is providing clues on the function of htt and its role in HD pathogenesis [70]. Many interaction partners for mutant and wild-type htt have been elucidated over the past decade by Y2-H, affinity chromatography and immunoprecipitation. To follow these studies, it was recently hypothesized that genetic modifiers of HD neurodegeneration should be enriched among htt protein interactors, and to test this, both high-throughput Y2-H screening and affinity pull-down followed by mass spectrometry were utilized [71]. This group was able to identify 104 htt interactions with Y2-H and 130 interactions with their pull-down method. To elucidate the biological relevance of these interactions, using a high-content validation assay, they also tested a set of 60 genes encoding interacting htt proteins for their ability to act as genetic modifiers of neurodegeneration in the HD *Drosophila* model. Results showed that 45% of these genes were high-confidence genetic modifiers (much higher than the 1–4% observed in unbiased genetic screens), and that these genes were similarly represented among proteins discovered with their Y2-H and pull-down and/or mass spectrometry methods. These results demonstrate that these methods are equally useful for identifying biologically relevant interactions [71].

Making use of information available in PPI databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Human Protein Reference Database (HPRD), along with PubMed publications, another group evaluated the commonality of molecular pathogenic mechanisms of neurodegenerative disorders, including HD, along with Alzheimer's disease, Parkinson's disease, dentatorubral–pallidolusian atrophy and prion disease, as well as amyotrophic lateral sclerosis. The investigators examined the PPI networks associated with causative proteins, such as htt, and found 19 proteins common to all diseases from literature, as well as 81 new common proteins from their network constructed using database information. Many of these identified proteins were previously characterized as being associated with the respective diseases. A relatively high correlation between all diseases for all of their analysis was seen, including commonality in characteristic protein domains. They concluded that the interactions found in this study *in silico* may serve to function in the common pathogenic mechanisms among neurodegenerative disorders [72].

As the effort continues to reconstruct the entire proteome, it would be to our advantage to exploit the breadth of knowledge contained in PPI databases. Not only will we gain a greater understanding of numerous biological processes but also, presumably, be able to apply this knowledge to advance other fields of research, such as drug discovery, disease prognosis and

the study of disease susceptibility. High-throughput molecular-profiling approaches, such as microarray technology, have already been successful in the advancement of these fields; yet, similarly, a rate-limiting step has been the ability to interpret the biological meaning of the data. Often, this problem has been approached from a pathway perspective that involves investigating which pathways are perturbed in a case-control population, which pathways determine a good or bad prognosis, or which pathways are activated or repressed in response to certain stimuli or compounds [57,62]. Such an approach can simplify the analysis and interpretation of genome-wide or large-scale datasets. Similarly, the abundance of information available in protein-interaction databases can be used for similar purposes in the field of human genetics, which we will discuss in greater depth later in this review.

Human protein-interaction databases

Currently, there exist numerous publicly available protein-interaction databases that contain information regarding human-specific interactions (Table 1). The majority of PPIs in these databases are from curation of the literature by biologists; however, some are incorporated by direct deposit prior to publication by the investigator [73]. In a majority of the PPI databases, the user will enter a protein of choice by either protein name or accession number, according to RefSeq, Genbank, Online Mendelian Inheritance in Man (OMIM), SwissProt or Entrez Gene, and, in return, receive a list of protein interactors, information pertaining to the experimental evidence for that interaction, as well as information concerning the protein itself. Another common feature of most databases is the ability to visualize the network of the queried protein and its interactors. We will touch on some of the key features of certain databases. For a more comprehensive review and additional information on these databases and others, please see [73–75].

One of the largest publicly available databases is the HPRD, which, to date, has more than 38,000 PPIs, more than 270,000 Pubmed links and access to curated pathways, as well as information on post-translational modifications (PTMs), domain architecture, protein functions, enzyme-substrate relationships, subcellular localization, tissue expression and disease association of genes. An interesting feature of this database is the Protein Distributed Annotation System, which enables researchers to annotate proteomic information in the context of HPRD data so that it is easily shared with the rest of the scientific community [76]. Another large and growing database that has similar components is the BioGrid, which currently houses approximately 42,800 human PPIs, but, altogether, contains more than 200,000 interactions from *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Mus musculus* and *Drosophila melanogaster*, in addition to *Homo Sapiens*.

Other available databases that are smaller than the HPRD and BioGrid yet offer additional unique features are The Biomolecular Interaction Network Database (BIND), which is a component of the Biomolecular Object Network Database, the Molecular Interaction database (MINT), the Database of Interacting Proteins (DIP) and Reactome. For example, BIND and MINT provide confidence scores for each interaction, specifically, for Y2-H experiments in BIND. In MINT, this score is based on the number of interactions, the number of citations and the type of experiment conducted to detect that interaction, while in BIND, the score is based on shared or related GO annotations, phenotypic profiling, homologous interactions, domain structure and the number of publications [73]. MINT also contains information pertaining to protein interactions with promoter regions and mRNA. Unique to DIP, the user can select to have certain PPIs evaluated based on paralogous interactions or common expression profiles of interactors or through domain-interaction preferences.

Reactome is not specifically a PPI database, but a curated resource for human-pathway data based on biologic reaction networks. Reactome reactions are described as taking place between

‘physical entities’, which include not only proteins but also nucleic acids, single small molecules, macromolecular complexes and even subatomic particles. All proteins, genes and reactions are cross referenced to a variety of widely used databases, such as Entrez Gene, Online Mendelian Inheritance in Man (OMIM) and KEGG, and each reaction is supported by evidence from biomedical literature, as well as documented with approved citations [77]. The user has the ability to search the database using a reaction name, gene name, protein name or any of several alternative identifiers. Reactions in the output are represented graphically, and the user has the option to click on ‘top-level’ pathways to delve deeper into the hierarchy with increasing detail at each level. Additionally, one can select for nonhuman species, and all accession numbers for all genes and proteins involved can be downloaded. For a more in-depth review of this database, please see [77].

Resources such as the STRING, Unified Human Interactome (UniHI) and GeneNetwork access a number of the reviewed databases to integrate protein-interaction information. The newest version of STRING, 8.0, covers approximately 2.5 million proteins from 630 different organisms and incorporates PPI information from a number of interaction databases, such as HPRD, BioGrid, MINT, BIND and DIP, and also imports known reactions from Reactome and KEGG pathways. Recent additions to this database incorporate interactions from IntACT, EcoCyc, NCI-Nature Pathway Interaction Database and GO. Automated text mining of PubMed abstracts, OMIM and information from other databases, such as the *Saccharomyces* Genome Database, Wormbase and the Interactive Fly, supplement this information [78,79]. For interactions in organisms that have not been confirmed experimentally, STRING is capable of running a set of prediction algorithms and transferring known interactions from model organisms to other species based on the prediction orthology for those proteins [79]. The user, however, has the option to select which organism the queried protein and its interactors will pertain to. Each interaction is given a numerical confidence score based on the experimental evidence and orthologous evidence behind that interaction, which allows the user to filter networks according to a desired confidence threshold.

Unified Human Interactome not only integrates PPIs from large Y2-H screens and curated databases, such as HPRD, DIP, BIND and Reactome (as well as others we have not discussed), but also predicts interactions based on orthology and computational text-mining approaches. This database also provides detailed information on each interaction, including statistical-interaction validation by gene coexpression data and validation by shared path length according to GO co-annotation hierarchy. The sources of the interactions are also documented and provided, along with links to access more information about that particular source of evidence. A useful feature of UniHI is that it allows for a highly targeted search, by which the user can exclude certain mapping approaches, such as Y2-H, display only proteins that are common interaction partners to multiple proteins in a query, display only interactions that occur in multiple maps, or display only direct interactions [80].

GeneNetwork is comprised of known interactions from BIND, HPRD, Reactome and KEGG. Similar to STRING and UniHI, GeneNetwork supplies predicted interactions based on biological process and molecular function annotation from the GO database. Additional experimental data are incorporated, such as coexpression data from approximately 450 microarrays from the Stanford Microarray Database and the National Center for Biotechnology Information Gene Expression Omnibus. Human Y-2H interactions and interactions based on orthologous high-throughput PPIs from lower eukaryotes are also included. After submitting a query for a given gene, the user is returned a list of interactors, each of which has an overall likelihood score, along with likelihood scores for that interaction based on microarray coexpression, human PPI prediction and orthologous PPI predictions. Positive evidence of known interactions from HPRD, BIND, KEGG and Reactome is indicated in additional columns. A recent study used this database to rank the best positional candidates in

susceptibility loci on the basis of their interactions using a method they developed known as the 'Prioritizer' [81].

A recent study examining PPI networks for human inherited neurodegenerative disorders characterized by ataxia (i.e., loss of balance or coordination) illustrates how these databases have been used to help us better understand pathogenic mechanisms underlying human diseases. Lim *et al.*, examined protein-interaction networks involved in cerebellar Purkinje cell degeneration, which is the primary cause of coordination and balance loss in inherited ataxias [82]. They developed a network for 54 proteins involved in 23 ataxias first by Y2-H screens and then expanded this network based on information from literature-curated and evolutionarily conserved interactions. Relevant direct PPIs were added from available interaction networks developed by Rual *et al.* and Stelzl *et al.* [83,84], and binary interactions were identified for the 54 ataxia-associated baits and 561 interacting prey proteins using literature-based information from BIND, HPRD, DIP, MINT and the mammalian PPI database. Furthermore, 1527 potential human interlogs (i.e., potentially evolutionarily conserved interactions) were identified from more than one species using the InParanoid database. Since 68 and 63% of literature-curated and interlog interactions are annotated to similar GO compartments, respectively, this group suggests that these identified interactions are of similar quality to interactions they identified in their Y2-H screens. The network demonstrated that several ataxia proteins interact and that there are shared pathways and mechanisms in this class of diseases. This study by Lim *et al.* will hopefully be able to provide additional knowledge about individual protein function and candidate genes for other diseases with similar phenotypes.

These are just a few of the more widely used publicly available databases that provide information on PPIs. Certainly, each one has unique features that allow researchers to gain access to vast amounts of useful biological information that can be broadly applied. In particular, we believe that this information would be extremely useful when applied to genome-wide studies that aim to detect epistatic or gene-gene interactions. There are many challenges when it comes to detecting epistasis, and we propose that we will be able to utilize this abundance of information to not only improve the computational efficiency of genome-wide studies of epistasis but also to facilitate the biological interpretation of the results.

Expert knowledge from PPI databases

We wish to exploit protein-interaction data to improve the genetic analysis of common human diseases. As we have illustrated, the information available to the scientific community in curated protein-interaction databases is abundant, and we certainly believe that proteomic data will be a useful complement to genetic data and that there is valuable relationship between the two. We can begin to dissect this relationship by simply examining whether epistatic interactions detected statistically, such as with MDR or other methods, are also found to exist at the protein level in the interaction databases previously mentioned.

To illustrate this, we used a number of databases to query the protein interactions represented by significant SNP interactions in three genetic-association studies. Coutinho *et al.* used MDR to analyze seven candidate genes in the serotonin metabolic and neurotransmission pathways mapping autism linkage regions and reported a significant interaction between polymorphisms in the 5-hydroxytryptamine (serotonin) receptor 5A (*HTR5A*), integrin- β 3 precursor (*ITGB3*), and sodium-dependent serotonin transporter (*SLC6A4*; $p = 0.001$) [85]. Evidence for physical interactions between *SLC6A4* and *HTR5A* was found for these genes in the STRING database when querying *HTR5A* and was based on evidence from text mining. When querying *SLC6A4* or *ITGB3* in STRING, evidence for interaction between these two genes was provided and was also based on information from text mining. Both interactions use this specific paper along

with others as sources of evidence. Asselbergs *et al.* analyzed interactions in polymorphisms influencing levels of tissue plasminogen activator and plasminogen activator inhibitor 1, which influence the risk of arterial thrombosis [86]. Using a twoway analysis-of-variance statistical test, the investigators found significant interactions between a polymorphism in the bradykinin B2 gene (*BDKRB2*) and the angiotensin-converting enzyme ($p = 0.003$) on tissue plasminogen activator in females, and between polymorphisms in bradykinin B2 and angiotensin II type 1 receptor (*AT1R/AGT1R*) on tissue plasminogen activator in males ($p = 0.006$). This latter interaction was also significant for plasminogen activator inhibitor 1 levels in both males and females. Strong evidence for interaction for all three of these genes is seen when querying STRING (Figure 1) and is supported by both experimental, based on *in vivo* assays, and text-mining evidence. One or more of these interactions are found in the databases that STRING integrates, such as HPRD, BIND, Reactome, MINT, BioGrid, DIP and KEGG annotated pathways, and DIP. Another more recent study examined SNPs in topoisomerase 3- α , RECQ-mediated genome instability 1 protein (*RMII*) and Bloom syndrome protein (*BLM*) and their association with cancer risk in acute myeloid leukemia/myelodysplastic syndromes, malignant melanoma, breast cancer and bladder cancer. Since mutations in *BLM* are known to be associated with elevated cancer risk, it was reasoned that genetic variants of *BLM* and the proteins that complex with it might play a role in influencing the risk for different cancers. It was determined that variant interactions in topoisomerase 3- α and *BLM* showed increased risk in all four cancers. While this study did not show statistical evidence of interaction with *RMII*, it was shown to confer increased risk of acute myeloid leukemia/myelodysplastic syndromes and malignant melanoma [87]. STRING showed that all three of these proteins interacted according to both text mining and experimental evidence based on co-immunoprecipitation, molecular sieving and fluorescence-imaging assays. While this provides stronger evidence that PPI databases can be useful as expert knowledge, we need to find a logical way to incorporate this information into the analysis process.

Similar to the approaches developed by Emily *et al.* (2009), one possible method would be to identify all of the genes associated with the SNPs in a dataset for whose protein products have evidence of direct interaction with each other and filter that dataset accordingly. Filtering based on the direct interactions may prove to be a simple solution, but doing so may ignore potentially important biological information. Interactions do not have to be direct, and it may be beneficial to include the SNPs and genes to a certain level according to their indirect interactions, in other words, by taking a more pathway-based approach, as did Shriner *et al.* (2008) and Askland *et al.* (2009). Another option would be to utilize or develop a confidence score for present interactions based on information available from a PPI database or even multiple databases. As mentioned, MINT, BIND, STRING, UniHI and GeneNetwork all provide a confidence score for interactions based on information such as the type of experiment conducted to detect that interaction and the supporting literature for that interaction. Specific metrics could be developed that would allow all SNPs or genes to be prioritized or weighted based on biological information on their interactions or allow investigators to filter SNPs or genes based on a determined interaction confidence threshold.

If one were to take any of these approaches, it would appear that the vital information to extract from these databases would be the direct and indirect interaction partners found in the dataset (to a certain level) as well as the evidence or confidence score to support those interactions. While this also seems to be a rather simple approach, one needs to consider that the number of databases available that could provide this information is abundant and that this information may not be consistent between databases. For example, Mathivanan *et al.* thoroughly reviewed the features of a number of databases, including MINT, BIND, HPRD, DIP and Reactome, and concluded that while there may be good overlap at the protein level between these databases, the level of overlap between PPIs is not as great [73]. They also find that for PPIs that do overlap between databases, there exists a difference in annotation, partly on account of

differences that arise according to how biologists interpret the experimental results. This presents an obstacle when attempting to apply this expert knowledge from multiple databases and may lead to the exclusion of important interactions or the inclusion of noninfluential interactions in a dataset. Considering this, it may be beneficial to use an integrated database, such as STRING, UniHI or GeneNetwork, which have, in their own respective ways, brought together the various information in a number of databases.

Expert commentary

We believe the methods we have discussed to bring expert knowledge into genome-wide studies to guide an analysis are one way to deal with the computational infeasibility of these large studies. Indeed, Bush *et al.* and Saccone *et al.* have shown that using biological knowledge to guide genetic association studies may provide more meaningful results [88,89]. Yu *et al.* provide a hypothesis-testing framework for combining multiple SNPs from the same gene or from multiple genes in a pathway-based manner [90]. As mentioned earlier, Askland *et al.* recently showed that patterns of SNPs in biological pathways are more likely to replicate than individual SNPs in GWAS [24]. Wilke *et al.* have suggested that we should not even begin to analyze a GWAS study until we have exhaustively studied each candidate gene and each pathway [91]. Only then will we have the appropriate knowledge base to make sense of GWAS results. As Moore noted, there is major shift in the field of genetic epidemiology away from the purely statistical approaches to these problems to a more bioinformatics approach that considers knowledge on gene function, gene networks and biochemical pathways [57]. This year may mark the turning point toward more of a systems approach that recognizes the role of epistasis and other complexities in the genetic architecture of common diseases.

While current work has demonstrated promise in the idea of combining many types of data as an analysis strategy, to narrow the evaluation to gene combinations that have been shown to interact experimentally provides a biologically concise reason why those two genes may be detected together statistically. However, it is important to acknowledge the potential limitations of using PPI databases for this purpose. We first must acknowledge the dynamics of PPIs and the fact that, as mentioned earlier, although a database may claim an existing interaction, for that interaction to be meaningful for the disease at hand, these two proteins must be both spatially and temporally coordinated [64]. PPIs are largely context dependent and require the appropriate cellular conditions in order for certain structural modifications that enable the interaction to occur. While mammalian two-hybrid systems that allow for assayed proteins to undergo these modifications in the appropriate cellular context have been developed to complement Y-2H systems, these tools are still under development and optimization. However, a large number of PPI networks available are representative of a static and not a dynamic network [92].

In addition to PPI dynamics, it is important to recognize that bias may exist across all databases, as well as the fact that genes and SNPs in the dataset may be unannotated or anonymous. How does one deal with anonymous SNPs or SNPs that are not in coding regions? It may be that a researcher wishes to consider this SNP as part of the gene that it is closest to or perhaps they may consider what annotated SNPs are in linkage disequilibrium with it. Furthermore, there are many proteins in the human proteome that have not been studied thoroughly or even studied at all and may be under-represented or nonexistent in these databases. To add to this, a bias of experimental methods for capturing certain interactions exists; for example, Y2-H experiments are not entirely adequate for detecting interactions with integral membrane proteins [73]. It is important to bear in mind the point that even if an interaction is not detected on the biological level, this does not mean that this interaction does not exist or will not be seen at the statistical level, and conversely, to remember that what is detected statistically may not have any biological relevance [10]. Therefore, we must be aware of, and concerned about, the amount

of important information we may potentially be missing owing to bias and lack of annotation, and what types of studies this expert knowledge is appropriate for.

Similarly, some may argue that using expert knowledge is biased in and of itself, despite bias in the databases. The ability to conduct a GWAS has been said to ‘relax’ the need for a strong prior hypothesis because the whole genome can be analyzed at once [93]. Individuals in support of ‘genomic agnosticism’ believe that when conducting a genome wide analysis, they will assume every SNP in the genome to be equally functional [94]. This brings us back to the issue of what information we may be missing by applying expert knowledge. While the benefits of conducting an unbiased GWAS study are valid, such as having no prior hypothesis, elimination of bias and inclusion of all information, we are still at a loss for computational power to conduct these studies and fully explore all interactions.

These issues demonstrate the need for a method to evaluate the metrics we develop from PPI expert knowledge and information we extract from these databases. Making use of data that are available and that have already been evaluated for interactions would aid this process. With access to data where the biological importance of the results and the interactions are known, we could determine if this same information is retained after applying expert knowledge. Therefore, we may be able to gauge the amount of information we gain or, on the other hand, which we may lose by applying expert knowledge. Since we are also concerned with the efficiency of evaluating genome-wide studies, we may want to explore simulated large-scale datasets that are already imbedded with known epistatic interactions, both physical and statistical. We would then be able to compare metrics we develop based on information from these databases. This will, hopefully, allow us to achieve the most meaningful results in the most efficient manner.

Five-year view

The field of proteomics is expanding with the availability of high-throughput methods to detect and characterize protein interactions, and there is continuing development of curated protein-interaction databases to provide the scientific community with access to this information. While in the field of human genetics, the availability of high-dimensional datasets from genome-wide studies is making it computationally expensive and impractical to carry out a genetic analysis study utilizing data-mining methods such as MDR. Since there is no indication that technological advancements in either field will come to a halt any time soon, the amount of valuable genetic and proteomic data produced will continue to grow.

We have proposed that expert knowledge extracted from protein-interaction databases may reduce the computational burden of large-scale and genome-wide studies, as well as facilitate the biological interpretation of the data. We foresee that, in the future, the methods that we propose to will not only be applicable to SNP studies but also to studies involving other forms of genetic variation (i.e., copy number variation and sequence repeats) that may run into similar problems with large-scale data analysis. The exploitation of expert knowledge will eventually not only ease our computation burdens but also aid in understanding the relationship between biological and statistical epistasis, as well as helping us to better understand the relationship between the proteome and the genome.

To fully exploit the knowledge in PPI databases, we need to develop a logical method to evaluate the information in these databases and the metrics developed from this information in order to incorporate this type of expert knowledge into our analysis. While we do not expect this to be a simple task, success in similar endeavors have assured us that it is an important and worthwhile task that needs to be explored [10,58,59]. Once we are able to successfully develop these methods, not only will we improve the ease with which we will be able to identify important epistatic interactions in genome-wide studies, but we will gain an understanding of

the physical biology that underlies these interactions and perhaps their role in a given disease. We expect these expert knowledge-based methods to enhance our comprehension of common human diseases and eventually lead to an improvement in the prevention, treatment and diagnosis of these diseases.

Key issues

- Owing to the feasibility of conducting genome-wide studies, the domain of human genetics is experiencing an information explosion; yet, at the same time, there is a lack of methods to fully explore the abundance of data.
- It is computationally infeasible to analyze all pair-wise and multiple-way SNP combinations in a genome-wide genetic study.
- Protein–protein interactions demonstrate one of the strongest functional relationships between genes; therefore, combining genomic data with available proteomic data may provide us with a more in-depth understanding of common human diseases.
- There exists a number of publicly available protein-interaction databases from which to extract biological expert knowledge.
- Protein–protein interaction information, such as confidence scores for interactions, could be used to prioritize or weight SNPs in a dataset in order to filter it accordingly or to guide an algorithm and improve the computational efficiency of genome-wide studies.
- We must be aware of the amount of important information we may miss owing to bias and lack of annotation in databases and which types of studies this expert knowledge is appropriate for.
- There are benefits of conducting unbiased genome-wide association studies, such as having no prior hypothesis, the elimination of bias and the inclusion of all information; however, we still lack the computational power to fully explore all interactions.
- There is a need for a method to evaluate the type of information we extract from these databases.
- We expect these expert knowledge-based methods to enhance our comprehension of common human diseases and lead to an improvement in the prevention, treatment and diagnosis of complex diseases.

Acknowledgments

Financial & competing interests disclosure

This work was supported by NIH grants LM009012, LM010098 and AI59694. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

- of interest
- of considerable interest

Expert Rev Proteomics. Author manuscript; available in PMC 2010 October 1.

1. Wasinger VC, Cordwell SJ, Cerpa-Poljak A, et al. Progress with gene-product mapping of the Mollicutes. *Mycoplasma genitalium*. *Electrophoresis* 1995;16:1090–1094.
2. Moore JH, Ritchie MD. The challenges of whole-genome approaches to common diseases. *JAMA* 2004;291:1642–1643. [PubMed: 15069055] • • Elaborates on the challenged of conducting whole-genome studies and potential ways to address these issues.
3. Persidis A. Proteomics. *Nat. Biotechnol* 1998;16:393–394. [PubMed: 9555734]
4. Wang Z, Moulton J. SNPs, protein structure, and disease. *Hum. Mutat* 2001;4:263–270. [PubMed: 11295823]
5. Cavallo A, Martin AC. Mapping SNPs to protein sequence and structure data. *Bioinformatics* 2005;21(8):1443–1450. [PubMed: 15613399] • Provides an understanding of how mutations such as SNPs can have an effect on protein structure.
6. Kim BC, Kim WY, Park D, et al. SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions. *BMC Bioinformatics* 2009;9:S2.
7. Bauer-Mehren A, Furlong LI, Raustschka M, et al. From SNPs to pathways: integration of functional effect of sequence variations on models of cell signaling pathways. *BMC Bioinformatics* 2009;10:S6.
8. Ryan M, Deikhans M, Lein S, et al. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics* 2008;25(11):1431–1432. [PubMed: 19369493]
9. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet* 2002;11:2463–2468. [PubMed: 12351582]
10. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* 2005;27:637–646. [PubMed: 15892116] • Provides a description of both biological and statistical epistasis (i.e., gene–gene interaction), its role in common complex diseases, and the challenges of detecting epistasis in genetic association studies.
11. Bateson, W. *Mendel's Principles of Heredity*. Cambridge, UK: Cambridge University Press; 1909.
12. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb* 1918;52:399–433.
13. Falchetti A, Marini F, Luzi E, Tonelli F, Brandi ML. Multiple endocrine neoplasms. *Best Pract. Res. Clin. Rheumatol* 2008;22(1):149–163. [PubMed: 18328987]
14. Gao LG, Luo F, Hui RT, Zhou XL. Recent molecular biological progress in Marfan syndrome and Marfan-associated disorders. *Ageing Res. Rev.* 2009 DOI:10.1016/j. arr.2009.09.001 (2009) (Epub ahead of print).
15. Avery L, Wasserman S. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet* 1992;8(9):312–316. [PubMed: 1365397]
16. Goodwin EB, Ellis RE. Turning clustering loops: sex determination in *Caenorhabditis elegans*. *Curr. Biol* 2002;12:R111–R120. [PubMed: 11839294]
17. Tong AH, Lesage G, Bader GD, et al. Global mapping of the yeast genetic interaction network. *Science* 2004;303(5659):808–813. [PubMed: 14764870]
18. Lehrer B, Crombie C, Tischler J, et al. Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat. Genet* 2006;38(8):896–903. [PubMed: 16845399]
19. Hirschhorn JN, Lohmueller K, Byrne E, et al. A comprehensive review of genetic association studies. *Genet. Med* 2002;4:45–61. [PubMed: 11882781]
20. Fink U. The future of genetic association studies in Alzheimer disease. *J. Neural Transm* 2003;110:253–266. [PubMed: 12658374]
21. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered* 2003;56:73–82. [PubMed: 14614241]
22. Wiltshire S, Bell JT, Groves C, et al. Type 2 diabetes susceptibility loci on chromosomes 1q21–25 and 10q23–26 in Northern Europeans. *Ann. Hum. Genet* 2006;70:726–737. [PubMed: 17044847]
23. Abou Jamra R, Fuerst R, Kaneva R, et al. The first genomewide interaction and locus–heterogeneity linkage scan in bipolar affective disorder: strong evidence of epistatic effects between loci on chromosomes 2q and 6q. *Am. J. Hum. Genet* 2007;81:974–986. [PubMed: 17924339]

24. Askland K, Read C, Moore JH. Pathway-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum. Genet* 2009;125:63–79. [PubMed: 19052778]
25. Tsai CT, Hwang JJ, Ritchie MD, et al. Renin–angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene–gene interaction. *Atherosclerosis* 2006;195(1):172–180. [PubMed: 17118372]
26. Spencer CC, Su Z, Donnelly P, et al. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009;5(5):E1000477. [PubMed: 19492015]
27. Clark AG, Boerwinkle E, Hixson J, et al. Determinants of the success of whole-genome association testing. *Genome Res* 2005;15:1463–1467. [PubMed: 16251455]
28. Templeton AR. Epistasis and complex traits. In: Wolf, J.; Brodie, B., 3rd; Wade, M., editors. *Epistasis and Evolutionary Process*. NY, USA: Oxford University Press; 2000.
29. Sing CF, Stengard JH, Kardia SL. Genes, environment, and cardiovascular disease. *Arterioscler. Thromb. Vasc. Biol* 2003;23:1190–1196. [PubMed: 12730090]
30. Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics, and human disease: analytical retooling for complexity. *Trends Genet* 2004;20:640–647. [PubMed: 15522460]
31. Rea TJ, Brown CM, Sing CF. Complex adaptive system models and the genetic analysis of plasma HDL-cholesterol concentration. *Perspect. Biol. Med* 2006;49(4):490–503. [PubMed: 17146134]
32. Moore JH, Williams SM. Epistasis and its implications for personal genetics. *Am. J. Hum. Genet* 2009;85(3):309–320. [PubMed: 19733727]
33. Ritchie MD, Hahn LW, Roodi N, et al. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet* 2001;69:138–147. [PubMed: 11404819]
34. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet. Epidemiol* 2003;24:150–157. [PubMed: 12548676]
35. Nelson MR, Kardia SL, Ferrell RE, et al. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 2001;11:458–470. [PubMed: 11230170]
36. Reif DM, White BC, Moore JH. Integrated analysis of genetic, genomic and proteomic data. *Expert Rev. Proteomics* 2004;1:1095–1104.
37. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. *Genet. Epidemiol* 2005;28:157–170. [PubMed: 15532037]
38. Young SS, Ge N. Recursive partitioning analysis of complex disease pharmacogenetic studies. I. Motivation and overview. *Pharmacogenomics* 2005;6(1):65–75. [PubMed: 15723607]
39. Millstein J, Conti DV, Gilliland FD, et al. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am. J. Hum. Genet* 2006;78:15–27. [PubMed: 16385446]
40. Zheng T, Wang H, Lo SH. backward genotype–trait association (BGTA) – based dissection of complex traits in case–control design. *Hum. Hered* 2006;62:196–212. [PubMed: 17114886]
41. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case–control studies. *Nat. Genet* 2007;39(9):1167–1173. [PubMed: 17721534]
42. Chen X, Liu CT, Zhang M, Zhang H. A forest-based approach to identifying gene and gene gene interactions. *PNAS* 2007;104(49):19199–19203. [PubMed: 18048322]
43. Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics* 2008;9(1):30–50. [PubMed: 17429103]
44. Motsinger-Reif AA, Dudek SM, Hahn LW, et al. Comparison of approaches for machine-learning optimization of neural networks for detecting gene–gene interactions in genetic epidemiology. *Genet. Epidemiol* 2008;32:325–340. [PubMed: 18265411]
45. Xiang W, Yang C, Yang Q, et al. MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association studies. *BMC Bioinformatics* 2009;10:13. [PubMed: 19134182]

46. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics* 2003;19:376–382. [PubMed: 12584123]
47. Hahn LW, Moore JH. Ideal discrimination of discrete clinical endpoints using multilocus genotypes. *In Silico Biol* 2004;4(2):183–194. [PubMed: 15107022]
48. Moore JH. Computational analysis of gene–gene interactions in common human diseases using multifactor dimensionality reduction. *Expert Rev. Mol. Diagn* 2004;4(6):795–803. [PubMed: 15525222]
49. Moore JH, Gilbert JC, Tsai CT, et al. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol* 2006;241:252–261. [PubMed: 16457852]
50. Moore, JH. Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. In: Zhu, X.; Davidson, I., editors. *Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data*. PA, USA: IGA Press; 2007. p. 17–30.
51. Altshuler D, Brooks LD, Chakravarti A, et al. A haplotype map of the human genome. *Nature* 2005;437:1299–1320. [PubMed: 16255080]
52. Risch NJ, Merikangas KR. The future of genetic studies of complex human disease. *Science* 1996;273:1516–1517. [PubMed: 8801636]
53. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet* 2005;6:95–108. [PubMed: 15716906]
54. Wang WY, Barratt BJ, Clayton DG, et al. Genome-wide association studies: Theoretical and practical concerns. *Nat. Rev. Genet* 2005;6:109–118. [PubMed: 15716907]
55. Kingsmore SF, Lindquist IE, Mudge J, et al. Genome-wide association studies: progress in identifying genetic biomarkers in common, complex diseases. *Biomarker Insights* 2007;2:283–292. [PubMed: 19662211] • Discusses the basics of genome-wide association studies, how they are useful for biomarker discover in complex diseases compared with other methods, as well as their limitations.
56. Greene CS, Moore JH. Ant colony optimization for genome-wide genetic analysis. *Lect. Notes Comp. Sci* 2008;5217:27–47.
57. Moore JH. From genotypes to genotypes: putting the genome back in genome-wide association studies. *Eur. J. Hum. Genet* 2009;17(10):1205–1206. [PubMed: 19277066]
58. Emily M, Mailund T, Hain J, et al. Using biological networks to search for interacting loci in genome-wide association studies. *Eur. J. Hum. Genet* 2009;17(10):1231–1240. [PubMed: 19277065]
59. Shriner D, Tesfaye BM, Padilla MA, et al. Commonality of functional annotation: a method for prioritization of candidate genes from genome-wide linkage studies. *Nucleic Acids Res* 2008;36(4):e26. [PubMed: 18263617]
60. Greene, CS.; Gilmore, JM.; Kiralis, J.; Andrews, PC.; Moore, JH. *Lecture Notes in Computer Science: Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Vol. Volume 5483/2009. Germany: Springer Berlin/Heidelberg; 2009. Optimal use of expert knowledge in ant colony optimization for the analysis of epistasis in human disease.
61. Yates JR. Mass spectrometry: from genomics to proteomics. *Trends Genet* 2000;16:5–8. [PubMed: 10637622]
62. Pellegrini M, Haynor D, Johnson JM. Protein interaction networks. *Expert Rev. Proteomics* 2004;1(2):239–249. [PubMed: 15966818] • Reviews several experimental approaches used to measure protein–protein, protein–DNA, and gene–gene interactions and how these approaches have been scaled up to produce genome-wide data used to depict interaction networks.
63. Tong AH, Evangelista M, Parsons AB, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 2001;294:2364–2368. [PubMed: 11743205]
64. Shrabaneek L, Saini HK, Bader GD, et al. Computational prediction of protein–protein interactions. *Mol. Biotechnol* 2007;38:1–17. [PubMed: 18095187]
65. Pellegrini M, Marcotte EM, Thompson MJ, et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 1999;96:4285–4288. [PubMed: 10200254]
66. Tan SH, Zhang Z, Ng SK. ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. *Nucleic Acids Res* 2004;32(Web Server issue):W69–W72. [PubMed: 15215353]

67. Makino T, Gojobori T. Evolution of protein–protein interaction network. *Genome Dyn* 2007;3:13–29. [PubMed: 18753782]
68. Kim S, Shin SY, Lee IH, Kim SJ, Sriram R, Zhang BT. PIE: an online prediction system for protein–protein interactions from text. *Nucleic Acids Res* 2008;36(Web Server Issue):W411–W415. [PubMed: 18508809]
69. Ta HX, Holm L. Evaluation of different domain based methods in protein interaction prediction. *Biochem. Biophys. Res. Commun.* 2009 DOI: 10.1016/j.bbrc.2009.09.130 (Epub ahead of print).
70. Li SH, Li XJ. Huntington-protein interactions and the pathogenesis of Huntington’s disease. *Trends Genet* 2004;20:146–154. [PubMed: 15036808]
71. Kaltenbach LS, Romero E, Becklin RR, et al. Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PloS Genet* 2007;3(5):e82. [PubMed: 17500595]
72. Limviphuvadh V, Tanaka S, Goto S, et al. The commonality of protein interaction networks determined in neurodegenerative disorders. *Bioinformatics* 2007;23:2129–2138. [PubMed: 17553855]
73. Mathivanan S, Periaswamy B, Gandhi T, et al. An evaluation of human protein–protein interaction data in the public domain. *BMC Bioinformatics* 2006;7:S19. [PubMed: 17254303] •• Gives a thorough review and evaluation of a number of protein–protein interaction databases comparing their overlap and a number of other statistics.
74. Chautard E, Thierry-Mieg N, Ricard-Blum S. Interaction networks: from protein function to drug discovery. *Pathol. Biol. (Paris)* 2009;57(4):324–333. [PubMed: 19070972]
75. Lehne B, Schlitt T. Protein–protein interaction databases: keeping up with growing interactomes. *Hum. Genomics* 2009;3(3):291–297. [PubMed: 19403463]
76. Prasad TSK, Goel R, Kandasamy K, et al. Human Protein Reference Database – 2009 update. *Nucleic Acids Res* 2009;37(Database issue):D767–D772. [PubMed: 18988627]
77. Matthews L, Gopinath G, Gillespie M, et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 2009;37(Database issue):D619–D622. [PubMed: 18981052]
78. Jensen LJ, Kuhn M, Stark M, et al. STRING 8 – a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 2009;37(Database issue):D412–D416. [PubMed: 18940858]
79. von Mering C, Jensen LJ, Kuhn M, et al. STRING 7: recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 2007;35(Database issue):D358–D362. [PubMed: 17098935]
80. Chaurasia G, Malhotra S, Russ J, et al. UniHI 4: new tools for query, analysis and visualization of the human protein–protein interactome. *Nucleic Acids Res* 2009;37(Database issue):D657–D660. [PubMed: 18984619]
81. Franke L, van-Bakel H, Fokkens L, et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet* 2006;78:1011–1025. [PubMed: 16685651]
82. Lim J, Hao T, Shaw C, et al. A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 2006;125(4):801–814. [PubMed: 16713569]
83. Raul TJ, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 2005;437:1173–1178. [PubMed: 16189514]
84. Stelzl U, Worm U, Lalowski M, et al. A human protein–protein interaction network: a resource for annotating the proteome. *Cell* 2005;122:857–968.
85. Coutinho AM, Sousa I, Martins M, et al. Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in the determination of platelet serotonin levels. *Hum. Genet* 2007;121:243–256. [PubMed: 17203304]
86. Asselbergs FW, Williams SM, Hebert PR, et al. Epistatic effects of polymorphisms in genes from the renin–angiotensin, bradykinin, and fibrinolytic systems on plasma t-PA and PAI-1 levels. *Genomics* 2007;89(3):362–369. [PubMed: 17207964]
87. Broberg K, Huynh E, Engstrom KS, et al. Association between polymorphisms in *RMII*, *TOP3A*, and *BLM* and risk of cancer: a case–control study. *BMC Cancer* 2009;9:140. [PubMed: 19432957]

88. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac. Symp. Biocomput* 2009;368–379. [PubMed: 19209715]
89. Saccone SF, Saccone NL, Swan GE, et al. Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics* 2008;24:1805–1811. [PubMed: 18565990]
90. Yu K, Li Q, Bergen AW, et al. Pathway analysis by adaptive combination of P-values. *Genet. Epidemiol.* 2009 DOI: 10.1002/gepi.20422 (Epub ahead of print).
91. Wilke RA, Mareedu RK, Moore JH. The pathway less traveled: moving from candidate genes to candidate pathways in the analysis of genome-wide data from large scale pharmacogenetic association studies. *Curr. Pharmacogenomics Personalized Med* 2008;6:150–159.
92. Lievens S, Lemmens I, Tavernier J. Mammalian two-hybrids come of age. *Trends Biochem. Sci.* 2009 DOI: 10.1016/j.tibs.2009.06.009 (Epub ahead of print).
93. Chanock SJ, Manolio T, Boehnke M, et al. Replicating genotype–phenotype associations. *Nature* 2007;447(7145):655–660. [PubMed: 17554299]
94. Carlson CS. Agnosticism and equity in genome-wide association studies. *Nat. Genet* 2006;38(6):605–606. [PubMed: 16736010] • Describes the standpoint of genomic agnosticism where each SNP is considered to be equally functional as applied to two studies, and how some believe this approach to have certain advantages.
95. Breitkreutz BJ, Stark C, Reguly T, et al. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* 2008;36(Database issue):D637–D640. [PubMed: 18000002]
96. Alfaro C, Andrade CE, Anthony K, et al. The Biomolecular Interaction Network Database and related tools: 2005 update. *Nucleic Acids Res* 2005;33(Database issue):D418–D424. [PubMed: 15608229]
97. Chatr-aryamontri A, Ceol A, Palazzi LM, et al. MINT: the Molecular INTERaction database. *Nucleic Acids Res* 2007;35(Database issue):D572–D574. [PubMed: 17135203]
98. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 2004;32(Database issue):D449–D451. [PubMed: 14681454]

Figure 1. Protein-interaction network as displayed in Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) when querying ACE

Interactions between angiotensin-converting enzyme, *BDKRB2* and *AGTR1*, are shown. Evidence for these interactions is derived from both experimental evidence (purple lines) and text-mining evidence (green lines).

ACE: Angiotensin-converting enzyme; *AGTR1*: Angiotensin II type I receptor; *BDKRB2*: Bradykinin B2 gene.

Table 1

Protein–protein-interaction databases.

PPI database	Website	Interactions	Additional features	Ref.
HPRD	www.hprd.org	>38,000	Additional information on phosphorylation motifs, signaling pathways, domain architecture, protein functions, enzyme–substrate relationships, subcellular localization, tissue expression and disease association of genes	[76]
BioGRID	www.thebiogrid.org	>200,000	Interaction directionality, phenotype, post-translational modification, domains and motifs are being added	[95]
BIND	http://bind.ca	>100,000	Includes information on genetic, biopolymer–biopolymer and protein–small molecule interactions and DNA, RNA and protein sequences	[96]
MINT	http://mint.bio.uniroma2.it	>100,000	Represents complexes, biomolecules, detailed experimental descriptions and protein structure information	[97]
DIP	http://dip.doe-mbi.ucla.edu	>57,000	Provides experimental-quality assessments to identify the most reliable interactions; represents complexes	[98]
Reactome	www.reactome.org	>2900 reactions	Extensive cover of human pathways in 46 domains of human biology; hypergeometric testing is used to display statistically over-represented events in the event hierarchy	[77]
STRING	http://string.embl.de	Incorporates a number of databases >2,500,000 proteins	Imports pathway information, provides confidence scores based on evidence from conserved genomic neighborhoods, gene-fusion events, co-occurrence events, coexpression data, experimental data, database information, text mining and homology	[78]
GeneNetwork	www.genenetwork.nl	Incorporates a number of interaction databases (unlisted)	Provides an overall likelihood score, along with likelihood scores for that interaction, based on microarray coexpression, human PPI prediction and orthologous PPI predictions	[81]
UniHI	www.mdc-berlin.de/unihi	Incorporates a number of interaction databases >250,000	Allows for the construction of tissue-specific networks, statistical-interaction validation by gene coexpression data,	[80]

PPI database	Website	Interactions	Additional features	Ref.
			and validation by shared path length according to gene ontology co-annotation hierarchy	

BIND: Biomolecular Interaction Network Database; BioGRID: Biological General Repository for Interaction Datasets; DIP: Database of Interacting Proteins; HPRD: Human Protein Reference Database; MINT: Molecular Interaction database; PPI: Protein–protein interaction; STRING: Search Tool for the Retrieval of Interacting Genes/Proteins; UniHI: Unified Human Interactome.