Chapter 4
Computational Approaches for Human Disease Gene Prediction and Ranking

Cheng Zhu, Chao Wu, Bruce J. Aronow, and Anil G. Jegga

Abstract While candidate gene association studies continue to be the most practical and frequently employed approach in disease gene investigation for complex disorders, selecting suitable genes to test is a challenge. There are several computational approaches available for selecting and prioritizing disease candidate genes. A majority of these tools are based on guilt-by-association principle where novel disease candidate genes are identified and prioritized based on either functional or topological similarity to known disease genes. In this chapter we review the prioritization criteria and the algorithms along with some use cases that demonstrate how these tools can be used for identifying and ranking human disease candidate genes.

4.1 Introduction

The majority of common diseases, common traits, and pharmacological drug response are genetically intricate, polygenic, multifactorial, and often result from an interaction of genetic, environmental, and physiological factors. Although

Cheng Zhu and Chao Wu contributed equally to this work.

C. Zhu • C. Wu
Department of Computer Science, College of Engineering and Applied Science,
University of Cincinnati, Cincinnati, OH, USA

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center,
3333 Burnet Ave. – MLC 7024, Cincinnati, OH 45229-3039, USA

B.J. Aronow • A.G. Jegga, D.V.M., M.S. (✉)
Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati,
OH 45229, USA

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center,
3333 Burnet Ave. – MLC 7024, Cincinnati, OH 45229-3039, USA
e-mail: Anil.Jegga@cchmc.org

N. Maltsev et al. (eds.), *Systems Analysis of Human Multigene Disorders*, Advances in Experimental Medicine and Biology 799, DOI 10.1007/978-1-4614-8778-4_4,

high-throughput, genome-wide studies like linkage analysis and gene expression profiling are useful for classification and characterization, they often fail to provide sufficient information to identify specific disease causal genes or drug targets. Both of these approaches typically result in the identification of hundreds of potential candidate genes and cannot effectively reduce the number of target genes to a manageable figure for further validation.

4.2 Bioinformatic Tools for Gene Prioritization

Several computational approaches (Table 4.1) have been developed for gene prioritization to overcome the limitations of high-throughput, genome-wide studies like linkage analysis and gene expression profiling, both of which typically result in the identification of hundreds of potential candidate genes [1–3, 8, 10, 16, 59, 61, 62, 65, 76]. See recent reviews [7, 29, 43, 46, 50, 60, 64, 76] for technical and algorithmic details of various gene prioritization tools. While a majority of these tools are based on the assumption that similar phenotypes are caused by genes with similar or related functions [9, 20, 27, 55, 65], they differ by the strategy adopted in calculating similarity and by the data sources utilized [63]. Further, no single source of data can be expected to capture all relevant relations. For example, using coexpression data alone will fail to detect many effects of posttranscriptional modifications, while relying on protein–protein interaction data alone will fail to capture transcriptional regulation. Since these different data types are complementary, they need to be merged not only to improve coverage but to infer stronger relationships through the accumulation of evidence [43]. While this is true, except for Endeavour [3, 63] and ToppGene [9, 10], most of the existing approaches mainly focus on the combination of only a few data sources.

4.2.1 Functional Annotation-Based Approaches

The functional annotation-based candidate disease gene prioritization approaches are usually based on the guilt-by-association principle which asserts that reliable predictions about the disease involvement (“guilt”) of a gene can generally be made if several of its partners (e.g., genes with correlated expression profiles or protein interaction partners or genes involved in same biological process or pathway) share a corresponding “guilty” status (“association”) [43]. Incorporating the prior information or knowledge about a disease is thus critical for this type of approach. One of the fundamental challenges for these approaches is the ability to gather, normalize, and integrate heterogeneous data from multiple sources and keeping them current. There are now several online tools available which make carrying out such analyses intuitively without the need for having programming knowledge or direct support of a bioinformatics expert (see [29, 46, 64] for a list of such Web-based

Table 4.1 List of current bioinformatics approaches and tools to rank human disease candidate genes

Approach	Online availability	Data types used	Training set (Input)
Approaches based on disease gene properties			
DGP [39]	http://cgg.ebi.ac.uk/services/dgp/	Sequence	Not applicable (N/A)
Prospectr [1]	http://www.genetics.med.ed.ac.uk/prospectr/	Sequence	N/A
Approaches using links between genes and phenotypes			
Genes2Diseases [48, 49]	http://www.ogic.ca/projects/g2d_2/	Sequence, gene ontology (GO), literature mining	Phenotype GO terms Known genes
BITOLA [23]	http://www.mf.uni-lj.si/bitola/	Literature mining	Concept
GeneSeeker [68, 69]	http://www.cmbi.ru.nl/GeneSeeker/	Expression, phenotype, literature mining	N/A
GFINDer [41, 42]	http://www.bioinformatics.polimi.it/GFINDer/	Expression, phenotype	N/A
TOM [52]	http://www-micrel.deis.unibo.it/~tom/	Expression, GO	Known genes and/or disease loci
Approaches using functional relatedness between candidate genes			
OMIM phenome map [67]	http://www.cmbi.ru.nl/MimMiner/	Phenotype, sequence, GO, protein interactions	N/A
Suspects [2]	http://www.genetics.med.ed.ac.uk/suspects/	Sequence, expression, GO	Known genes
Prioritizer [15]	http://www.prioritizer.nl/	Expression, GO, protein interactions	Disease loci
Endeavour [3]	http://www.esat.kuleuven.be/endeavour/	Sequence, expression, GO, pathways, literature mining	Known genes
ToppGene [9]	http://toppgene.cchmc.org	Mouse phenotype, expression, GO, pathways, literature mining	Known genes
ToppNet [8]	http://toppgene.cchmc.org	Protein interactions	Known genes

The first column has the source or the name of the tool. The second column shows the URL of the corresponding web application. The third column shows the list of genomic annotation types/features used by each of the methods for candidate gene ranking. The last column has details of the training or the input data, if used (*Note: modified from Kaimal et al.* [29]*, this list is extensive, but not exhaustive; references* [43, 50] *provide an additional list of tools*)

tools). While the usage of multiple heterogeneous data in the ranking makes the functional annotation-based approaches more thorough and less biased global assessment of candidate genes, they still suffer with a bias towards the training set and have some limitations. For instance, by using a training set, it is assumed that the disease genes yet to be discovered will be consistent with what is already known about a disease and/or its genetic basis. This assumption may not always be true. Additionally, since these approaches rely on known gene annotation, they tend to be biased towards selecting better annotated genes. For example, a "true" candidate gene can be missed if it lacks sufficient annotations. Thus, the effectiveness of this approach depends critically on how well the disease under investigation is defined both molecularly and physiologically. Second, it is important to note that the annotations and analyses provided, and the prioritization by these approaches, can only be as accurate as the underlying original sources from which the annotations are retrieved. For instance, only one fifth of the known human genes have pathway or phenotype annotations, and there are still more than 30 % genes whose functions are not well-defined. Third, using an appropriate or "true representative" training set is critical. For instance, in an earlier study, we observed that using larger training sets (>100 genes) decreases the sensitivity and specificity of the prioritization compared to smaller training sets (7–21 genes) [10]. Lastly, almost all of the current disease gene identification and prioritization approaches are coding-gene-centric, while it has been speculated that complex traits result more often from noncoding regulatory variants than from coding sequence variants [32, 35, 40].

4.2.2 Network-Based Approaches

A majority of the current computational disease candidate gene prioritization methods [1–3, 10, 16, 59, 61, 62, 65, 76] rely on functional annotations, gene expression data, or sequence-based features. The coverage of the gene functional annotations, however, is still a limiting factor. Currently, only a fraction of the genome is annotated with pathways and phenotypes [10]. While two thirds of all the genes are annotated by at least one functional annotation, the remaining one third has yet to be annotated. Interestingly, because biological networks have been found to be comparable to communication and social networks [28] through commonalities such as scale-freeness and small-world properties, the algorithms used for social and Web networks should be equally applicable to biological networks.

Recent biotechnological advances (e.g., high-throughput yeast two-hybrid screening) have facilitated generation of proteome-wide protein–protein interaction networks (PPINs) or "protein interactome" maps in model organisms and humans [53, 56]. Additionally, the shift in focus to systems biology in the post-genomic era has generated further interest in these networks and pathways. As a result, PPINs have been increasingly used not only to identify novel disease candidate genes [17, 30, 34, 73, 74] but also for candidate gene prioritization [8, 11, 34, 45, 73]. At the same time, network topology-based analyses hitherto used in social and Web network analyses have been successfully used in the identification and prioritization of disease candidate genes [8, 12, 19, 24, 34, 36, 54, 57, 70, 73]. Broadly, network topology-based candidate gene ranking approaches can be grouped into two categories: parameter-based and parameter-free methods. The parameter-based methods, such as PageRank with Priors (PRP [8]), Random Walk (RW [34]), and PRIoritizatioN and Complex Elucidation (PRINCE [70]), as the name indicates require additional auxiliary parameters that need to be trained by using available data sets. The PRP, for example, needs a parameter β to control the probability of jumping back to the initial node [8]. Similarly, the PRINCE algorithm uses a parameter to describe the relative importance of prior information [70]. However, selecting optimal parameters is often a challenge, and therefore the more "user-friendly" parameter-free approaches are preferred [24]. Further, most of the parameter-based approaches take into account the global information in the entire network, and thus they typically require extensive computation. For instance, in PRP, scores of all the vertices in the network need to be updated iteratively until they converge. This process tends to be slow and inefficient especially when the network size is large. The parameter-free methods (e.g., interconnectedness or ICN [24]), on the other hand, measure closeness of each candidate gene to known disease genes by taking into account direct link and the shared neighbors between two genes and therefore are relatively less intensive computationally. However, the performance of parameter-free methods was not comparable to those of parameter-based approaches. To address this, we recently developed a novel network-based parameter-free framework for discovering and prioritizing human rare disease candidate genes [75]. Our goals were to (a) enhance prioritizing performance compared to current parameter-free methods and (b) achieve a comparable performance to the parameter-based ones. Using several test cases, we compared the performance of our method (Vertex Similarity (VS)-based approach) to two approaches, one each from parameter-based (PRP) and parameter-free methods (ICN), and also used it to rank the immediate neighbors of known rare disease genes as potential novel candidate genes.

Network-based approaches using protein–protein interaction data while useful have some practical limitations [29]. First, high-throughput protein–protein interaction sets, especially yeast two-hybrid sets, are inherently noisy and may contain several interactions with no biological relevance [18, 26, 37, 66]. Surprisingly, only 5.8 % of the human, fly, and worm yeast two-hybrid interactions have been confirmed by the HPRD (Human Protein Reference Database), a manually curated compilation of protein interactions [47]. Second, the protein interactome tends to be biased towards well-studied proteins. Third, some of the human protein interactome data is derived by extrapolating high-throughput interactions from other species. Even though previous studies have shown that PPINs are conserved across species [25], there is a possibility for species-specific protein interactions. Fourth, two interacting proteins need not lead to similar disease phenotypes when mutated—for instance, they may have redundant or different but overlapping functions, or one may be more dispensable than the other [47]. Additionally, disease proteins may lie at different points in a molecular pathway and not necessarily interact directly. Fifth, disease mutations need not always involve proteins (e.g., telomerase RNA component in congenital autosomal dominant dyskeratosis) [47]. Lastly, most of the network topology-based algorithms were originally developed to identify "important" nodes in networks. Although extended versions of these algorithms are used to prioritize nodes to selected "seeds," they could still be biased towards hubs.

4.3 ToppGene Suite: A One-Stop Portal for Candidate Gene Prioritization Based on Functional Annotations and Protein Interactions Network

In this section, we describe the ToppGene Suite (http://toppgene.cchmc.org) [8–10], a unique, one-stop online assembly of computational software tools that enables biomedical researchers to perform candidate gene prioritization based on (a) functional annotation similarity between training and test set genes (ToppGene) [10], (b) protein interactions network analysis (ToppNet) [8], and (c) identify and rank candidate genes in the training set interactome based on both functional annotations and PPIN analysis (ToppGeNet) [8]. The ToppGene knowledgebase combines 17 gene features available from the public domain. It includes both disease-dependent and disease-independent information in the nature of known disease genes, previous linkage regions, association studies, human and mouse phenotypes, known drug genes, microarray expression results, gene regulatory regions (transcription factor target genes and microRNA targets), protein domains, protein interactions, pathways, biological processes, and literature co-citations.

4.3.1 ToppGene: Functional Annotations-Based Candidate Gene Prioritization

In the first step, ToppGene generates a representative profile of the training genes using as many as 17 features and identifies over-representative terms from the training genes. Each of the test set genes is then compared to this representative profile of the training set, and a similarity score for each of the 17 features is derived and summarized by the 17 similarity scores. Different methods are used for similarity measures of categorical (e.g., GO annotations) and numeric (i.e., gene expression) annotations. For categorical terms, a fuzzy-based similarity measure (see Popescu et al. [51] for additional details) is applied, while for numeric annotation, i.e., the microarray expression values, the similarity score is calculated as the Pearson correlation of the two expression vectors of the two genes. The 17 similarity scores are combined into an overall score using statistical meta-analysis, and a *p-value* of each annotation of a test gene G is derived by random sampling of the whole genome. The *p-value* of the similarity score S_i is defined as:

$$p(S_i) = \frac{\textit{count of genes having score higher than}\ \mathrm{G}\ \textit{in the random sample}}{\textit{count of genes in the random sample containing annotation}}.$$

To combine the *p-values* from multiple annotations into an overall *p-value*, Fisher's inverse chi-square method, which states that $-2\sum_{i=1}^{n}\log p_i \rightarrow \chi^2(2n)$ (assuming the p_i values come from independent tests) is used. The final similarity score of the test gene is then obtained by 1 minus the combined *p-value*. Additional details explaining the development of this method along with the validation process and comparison with other approaches have been previously published [9, 10].

4.3.2 ToppNet: Network Analysis-Based Candidate Gene Prioritization

ToppNet gene prioritization is based on the analysis of the protein–protein interaction network. Motivated by the observation that biological networks share many properties with social and Web networks [28], ToppNet uses extended versions of three algorithms from White and Smyth [72]: PageRank with Priors (PRP), HITS with Priors, and K-step Markov. The disease candidate genes (test set) are ranked by estimating their relative importance in the PPIN to known disease-related genes (training set). The PageRank with Priors, based on White and Smyth's PageRank algorithm [72], mimics the random surfer model wherein a random Internet surfer starts from one of a set of root nodes, R, and follows one of the links randomly in each step. In this process, the surfer jumps back to the root nodes at probability β, thus restarting the whole process. Intuitively, the PRP algorithm generates a score that is proportional to the probability of reaching any node in the Web surfing process. This score indicates or measures the relative "closeness" or importance to the root nodes. The second algorithm is HITS with Priors, an extension of HITS (Hyperlink-Induced Topic Search) developed by Jon Kleinberg to rank Web pages. It determines two values for a page: "hubness," representing the value of its links to other pages, and "authority," which estimates the value of the content of the page [33]. Here, too, the surfer starts from one of the root nodes. In the odd steps he/she can either follow a random "out-link" or jump back to a root node, and in the even steps he/she can instead follow an "in-link" or jump back to a root node. As in the case of PRP, HITS with Priors also estimates the relative probability of reaching a node in the network. The third algorithm is the K-Step Markov method which mimics a surfer who starts with one of the root nodes and then follows a random link in each step before returning to the root node (after K steps) and restarts surfing. For additional details readers are referred to our original published study [8].

4.3.3 ToppGeNet: Prioritization of Disease Gene Neighborhood in the Protein Interactome

ToppGeNet allows the user to rank the interacting partners (direct or indirect) of known disease genes for their likelihood of causing a disease. Here, given a training set of known disease genes, the test set is generated by mining the protein interactome and compiling the genes interacting either directly or indirectly (based on user input) with the training set genes. The test set genes can then be ranked using either ToppGene (functional annotation-based method) or ToppNet (PPIN-based method).

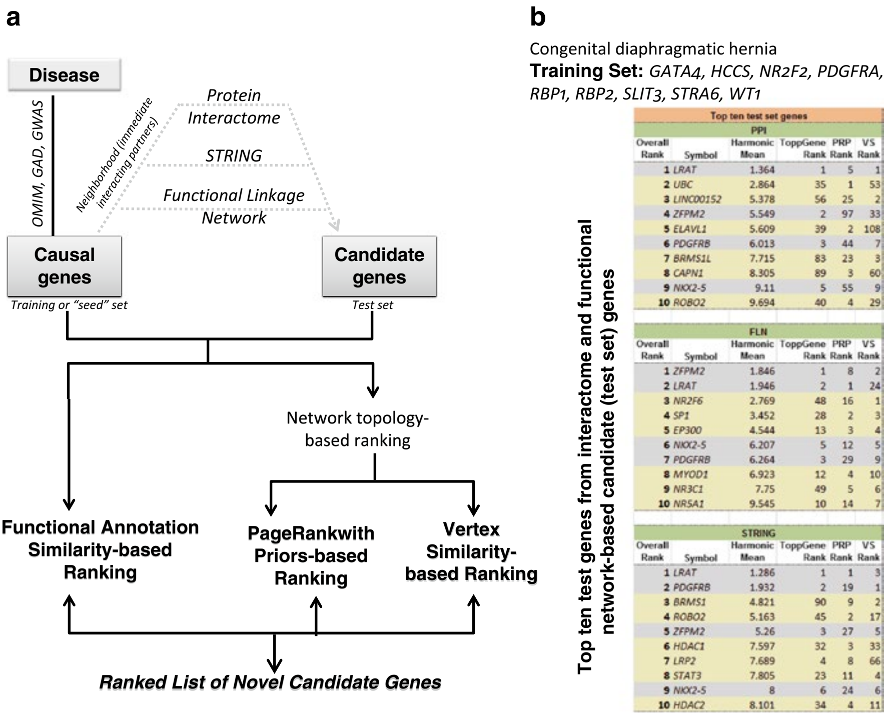
4.4 Case Studies to Demonstrate the Utility of Computational Approaches for Human Disease Gene Prediction and Ranking

In the following sections we present two sets of case studies to demonstrate the utility of computational approaches in discovering and ranking novel candidate genes for human diseases. In an earlier study, Tiffin et al. [61] used some of the computational approaches for disease gene identification and prioritization and concluded that using the methods in concert was more successful in prioritizing candidate genes for disease than when each was used alone. Hence, in the first case study, we select ten diseases and use both functional annotations-based and network-based approaches to identify and rank novel candidate genes for these diseases. We used ToppGene [9] for functional annotation-based ranking, and for network-based ranking we used both parameter [8]- and nonparameter [75]-based approaches (see next section for details). In the second case study, we present two recent examples that demonstrate the power of using bioinformatics techniques with the exome sequencing technologies in identifying novel candidate genes for rare disorders.

4.4.1 Case Study 1: Identifying and Ranking Novel Candidate Genes for Ten Human Diseases

The workflow (Fig. 4.1) described here is based on a simulation of a researcher's approach to selecting and ranking candidate disease genes. In this process, a variety of relevant database sources are mined for compiling both the training and test set genes. Known disease-associated genes for the ten selected diseases (from a recent review [43]) were obtained by combining gene lists from OMIM [21], the Genetic Association Database [4], GWAS [22], and diseases biomarkers from the Comparative Toxicogenomics Database [13] (see Table 4.2 for the list of selected ten diseases and their training sets or known causal genes). The test set or candidate genes to be ranked are compiled mining protein interactome and functional linkage networks. Briefly, for each of the training set genes (known disease causal gene), we extracted their interacting partners (both from the protein interactome and functional networks). The protein interactome data was downloaded from the NCBI (ftp://ftp.ncbi.nih.gov/gene/GeneRIF/interactions.gz), while for functional networks, we used two sources: (a) Functional Linkage Network (FLN) [38] and (b) STRING (score $\geq$ 700) [58]. Thus, for each disease, we compiled three test sets using the three databases.

The test sets were then ranked by three approaches: (a) functional annotations-based ranking (using ToppGene), (b) PageRank with Priors (parameter-dependent network topology-based approach), and (c) Vertex Similarity (parameter-free network topology-based approach). We used the harmonic mean of the individual ranks from the three approaches to obtain the final-ranked list. We repeated the


a
Disease
OMIM, GAD, GWAS
Neighborhood (immediate interacting partners)
Protein Interactome
STRING
Functional Linkage Network
Causal genes
Training or "seed" set
Candidate genes
Test set
Network topology-based ranking
Functional Annotation Similarity-based Ranking
PageRankwith Priors-based Ranking
Vertex Similarity-based Ranking
Ranked List of Novel Candidate Genes
b
Top ten test genes from interactome and functional network-based candidate (test set) genes

Congenital diaphragmatic hernia
Training Set: *GATA4, HCCS, NR2F2, PDGFRA, RBP1, RBP2, SLIT3, STRA6, WT1*

Top ten test set genes

PPI

Overall Rank	Symbol	Harmonic Mean	ToppGene Rank	PRP Rank	VS Rank
1	*LRAT*	1.364	1	5	1
2	*UBC*	2.864	35	1	53
3	*LINC00152*	5.378	56	25	2
4	*ZFPM2*	5.549	2	97	33
5	*ELAVL1*	5.609	39	2	108
6	*PDGFRB*	6.013	3	44	7
7	*BRMS1L*	7.715	83	23	3
8	*CAPN1*	8.305	89	3	60
9	*NKX2-5*	9.11	5	55	9
10	*ROBO2*	9.694	40	4	29

FLN

Overall Rank	Symbol	Harmonic Mean	ToppGene Rank	PRP Rank	VS Rank
1	*ZFPM2*	1.846	1	8	2
2	*LRAT*	1.946	2	1	24
3	*NR2F6*	2.769	48	16	1
4	*SP1*	3.452	28	2	3
5	*EP300*	4.544	13	3	4
6	*NKX2-5*	6.207	5	12	5
7	*PDGFRB*	6.264	3	29	9
8	*MYOD1*	6.923	12	4	10
9	*NR3C1*	7.75	49	5	6
10	*NR5A1*	9.545	10	14	7

STRING

Overall Rank	Symbol	Harmonic Mean	ToppGene Rank	PRP Rank	VS Rank
1	*LRAT*	1.286	1	1	3
2	*PDGFRB*	1.932	2	19	1
3	*BRMS1*	4.821	90	9	2
4	*ROBO2*	5.163	45	2	17
5	*ZFPM2*	5.26	3	27	5
6	*HDAC1*	7.597	32	3	33
7	*LRP2*	7.689	4	8	66
8	*STAT3*	7.805	23	11	4
9	*NKX2-5*	8	6	24	6
10	*HDAC2*	8.101	34	4	11

Fig. 4.1 Panel (**a**) shows schematic representation of the workflow for identifying and ranking novel disease candidate genes using functional annotation- and network-based approaches. Candidate genes are compiled using both protein interactions and functional associations (Functional Linkage Network and STRING). The candidate genes are ranked using both functional annotations (ToppGene) and network topology (PageRank with Priors and Vertex Similarity-based approaches). The final ranks are generated by taking the harmonic mean of the ranks of a gene from the three methods (ToppGene, PRP, and VS). Panel (**b**) shows the top-ranked genes for congenital diaphragmatic hernia using functional annotation- and network-based approaches. Highlighted genes (*LRAT*, *ZFPM2*, *NKX2-5*, and *PDGFRB*) represent those that have been ranked among top ten by different approaches

same process for two other test sets obtained from functional networks (FLN and STRING). In the final step, we intersected the top ten genes from the three networks (PPIN, FLN, and STRING) to see the intersection. The last column in Table 4.2 shows those genes that are ranked among the top ten in the three networks. For example, in congenital diaphragmatic hernia (CDH), four genes (*LRAT, ZFPM2, NKX2-5*, and *PDGFRB*) were ranked among top ten in all the three networks. Interestingly, the retinol status in newborns is associated with CDH, and genetic analyses in humans suggest a role for retinoid-related genes in the pathogenesis of CDH [6]. *LRAT* (lecithin retinol acyltransferase) ranked among the top mediates cellular uptake of retinol and plays an important regulatory role in cellular vitamin A homeostasis [31]. Similarly, Wat et al. [71] identified three unrelated patients

Table 4.2 Top-ranked novel candidate genes for ten select diseases

Disease name	Known disease-causing genes (training set)	Top-ranked novel candidate genes (using different approaches and data sets)
Congenital diaphragmatic hernia	*GATA4, HCCS, NR2F2, PDGFRA, RBP1, RBP2, SLIT3, STRA6, WT1*	*LRAT, NKX2-5, PDGFRB, ZFPM2*
Bipolar disorder	*ABCA13, BCR, BDNF, BRCA2, COMT, CUX2, DRD4, HTR4, PALB2, SLC6A3, SLC6A4, TRPM2, XBP1*	*ADRB2, BRCA1, DRD2, NTRK2*
Nasopharyngeal carcinoma	*CCND1, CDH13, COX7B2, CTLA-4, CYP2A6, CYP2E1, CYP2F1, ERCC1, FAS, GABBR1, GSTM1, HHATL, HLA-A, HLA-B, HLA-C, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1, HLA-E, HLA-F, HP, HSPA1B, IFNA17, IL10, IL12A, IL16, IL18, IL1B, IL8, ITGA9, LOC344967, MDM2, MECOM, MICA, MMP1, MMP2, N4BP2, NAT2, NFKB1, OGG1, PLUNC, PTGS2, RASSF1A, TAP1, TGFB1, TLR10, TLR3, TLR4, TNF, TNFRSF19, TP53, UBAP1, VEGFA, XPC, XRCC1*	*HLA-G, HLA-DPA1, HLA-DRA*
Testicular germ cell tumor	*ATF7IP, BAK1, DMRT1, FGFR3, KIT, KITLG, LTA, SPRY4, STK11, TGFB1, TNF*	*LTB, IFNG*
Crohn's disease	*ATG16L1, C11orf30, CCR6, CDKAL1, FUT2, ICOSLG, IL12B, IL23R, IRGM, ITLN1, JAK2, LRRK2, MST1, MUC19, NKX2-3, NOD2, ORMDL3, PTGER4, PTPN2, PTPN22, STAT3, TNFSF15, ZNF365*	*IL12RB1, IL23A, JAK1, STAT1, STAT5B*
Asthma	*ACE, ADAM33, ADRB2, CC16, CCL11, CCL5, CD14, CMA1, CSF1R, CTLA4, FLG, GPRA, GSTM1, GSTP1, GSTT1, HAVCR1, HLA-DPB1, HLA-DQB1, HLA-DRB1, IL10, IL13, IL18, IL4, IL4R, LTA, LTC4S, NAT2, NOS1, SPINK5, STAT6, TBXA2R, TGFB1, TNF*	*IL1B, HLA-DRA*
Metopic craniosynostosis	*FGFR1, FGFR2, FGFR3, GLI3, TWIST1*	*FGF9, FGF2*
Nonsyndromic cleft lip/palate	*BMP4, IRF6, MSX1, MTR, PVRL1, STOM, SUMO1, TP63*	*MSX2, PAX3*
Arthrogryposis	*MYH3, TNNI2, TNNT3, TPM2, UTRN*	*ACTA1, DMD, TNNC1, TNNC2, TNNT1, TPM1*
Bipolar schizoaffective disorder	*ABCA13, BCR, BDNF, COMT, CUX2, DRD4, GABRR1, HTR4, PALB2, SLC6A3, SLC6A4, TRPM2, XBP1*	*ADRB2, DRD2, ITPR3, SLC6A9*

with CDH who had a heterozygous deletion of chromosome 8q involving *ZFPM2*, which was ranked among the top five in the three networks. It is beyond the scope of this chapter to discuss about the top-ranked genes for all the ten diseases. The supplementary file (Supplementary File 1) shows the complete lists of training and ranked test set genes for the ten select diseases along with the details of rankings from each of the three approaches using three different networks (PPIN, FLN, and STRING).

4.4.2 Case Study 2: Exome Sequencing and Bioinformatics Applications to Identify Novel Rare Disease Causal Variants

In the following sections we present two examples from recently published studies [5, 14] where computational approaches for candidate gene ranking were used in concert with exome sequencing to identify novel disease causal variants.

The first example [14] illustrates the potential of combining genomic variant and gene level information to identify and rank novel causal variants of rare diseases. Combining computational gene prediction tools with traditional mapping approaches, Erlich et al. [14] demonstrated how rare disease candidate genes from exome resequencing experiment can be successfully prioritized. In this study, a familial case of hereditary spastic paraparesis (HSP) was analyzed through whole-exome sequencing, and the four largest homozygous regions (containing 44 genes) were identified as potential HSP loci. The authors then applied several filters to narrow down the list further. For instance, a gene was considered as potentially causative if it contains at least one variant that is either under purifying selection or not inherited from the parents or absent in dbSNP or the 1,000 Genomes Project data. Because majority of the known rare disease variants affect coding sequences, the authors also checked if the variant is non-synonymous. After this filtering step, 15 candidate genes were identified and this list was further prioritized using three computational methods (Endeavour [3], ToppGene [9], and Suspects [2]). As a training set, a list of 11 seed genes associated with a pure type of HSP was compiled through literature mining. Interestingly, the top-ranking gene from all the three bioinformatics approaches (each of which uses different types of data and algorithms for prioritization) was *KIF1A*. Subsequent confirmation of *KIF1A* as the causative variant was done using Sanger sequencing.

In the second example, Benitez et al. [5] used disease-network analysis approach as supporting in silico evidence of the role of the adult neuronal ceroid lipofuscinosis (NCL) candidate genes identified by exome sequencing. In this case, the authors used Endeavour [3] and ToppGene [9] to rank the NCL candidate variant genes identified by exome sequencing. Known causal genes of other NCLs along with genes that are associated with phenotypically close disorders were used as training set. Interestingly, the three variants identified by exome sequencing (*PDCD6IP, DNAJC5,* and *LIPJ*) were among the top five genes in the combined analysis using ToppGene and Endeavour, suggesting that they may be functionally or structurally related with NCL encoded genes and constituting true causative variants for adult NCL.

4.5 Final Remarks

The selection of "best" computational approach for identifying and ranking disease candidate genes is not an easy task and depends on several various factors. Since a majority of these approaches are based on guilt-by-association principle, having a "good" or representative training set is critical. The training set may not necessarily be always a set of known causal genes but can be an implicated pathway or biological process or even a list of symptoms (or phenotype). Additionally, prior knowledge can sometimes be also inferred from related or similar diseases. This similarity can be either similar manifestation or symptoms or similar molecular mechanisms of related or similar diseases. Second, selecting an appropriate approach is also important and frequently depends on the disease type and the molecular mechanism that causes it. For example, using protein–protein interaction data for identifying novel candidates may be useful when a disease is known to be caused by the disruption of a larger protein complex. On the other hand, using a protein interaction network may not be totally justified for a disease known to be caused by aberrant regulatory mechanisms. In such cases, either using gene regulatory networks and/or high-throughput gene expression data may be more apt [50]. Third, since several previous studies have shown that the computational approaches for disease gene ranking are largely complementary [5, 14, 44, 61], we recommend using a combination of at least two different approaches (e.g., functional annotation-based and network topology-based approaches).

References

1. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2005) Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics 6:55
2. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics 22(6):773–774
3. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24(5):537–544
4. Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nat Genet 36(5):431–432. doi:10.1038/ng0504-431, ng0504-431 [pii]
5. Benitez BA, Alvarado D, Cai Y, Mayo K, Chakraverty S, Norton J, Morris JC, Sands MS, Goate A, Cruchaga C (2011) Exome-sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis. PLoS One 6(11):e26741. doi:10.1371/journal.pone.0026741, PONE-D-11-16499 [pii]
6. Beurskens LW, Tibboel D, Lindemans J, Duvekot JJ, Cohen-Overbeek TE, Veenma DC, de Klein A, Greer JJ, Steegers-Theunissen RP (2010) Retinol status of newborn infants is associated with congenital diaphragmatic hernia. Pediatrics 126(4):712–720. doi:10.1542/peds.2010-0521, peds.2010-0521 [pii]
7. Bornigen D, Tranchevent LC, Bonachela-Capdevila F, Devriendt K, De Moor B, De Causmaecker P, Moreau Y (2012) An unbiased evaluation of gene prioritization tools. Bioinformatics 28(23):3081–3088. doi:10.1093/bioinformatics/bts581, bts581 [pii]
8. Chen J, Aronow BJ, Jegga AG (2009) Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics 10:73. doi:1471-2105-10-73, [pii] 10.1186/1471-2105-10-73
9. Chen J, Bardes EE, Aronow BJ, Jegga AG (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 37(Web Server issue):W305–W311. doi:gkp427, [pii] 10.1093/nar/gkp427
10. Chen J, Xu H, Aronow BJ, Jegga AG (2007) Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics 8(1):392

11. Chen JY, Shen C, Sivachenko AY (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. Pac Symp Biocomput 367–378
12. Chen X, Yan GY, Liao XP (2010) A novel candidate disease genes prioritization method based on module partition and rank fusion. OMICS 14(4):337–356. doi:10.1089/omi.2009.0143
13. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res 37(Database issue):D786–D792. doi:gkn580, [pii] 10.1093/nar/gkn580
14. Erlich Y, Edvardson S, Hodges E, Zenvirt S, Thekkat P, Shaag A, Dor T, Hannon GJ, Elpeleg O (2011) Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. Genome Res 21(5):658–664. doi:gr.117143.110, [pii] 10.1101/gr.117143.110
15. Franke L, Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78(6):1011–1025
16. Freudenberg J, Propping P (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 18(Suppl 2):S110–S115
17. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic Acids Res 34(19):e130
18. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM (2003) A protein interaction map of Drosophila melanogaster. Science (New York, NY) 302(5651):1727–1736. doi:10.1126/science.1090289, 1090289 [pii]
19. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, Herbst M, Suopanki J, Scherzinger E, Abraham C, Bauer B, Hasenbank R, Fritzsche A, Ludewig AH, Bussow K, Coleman SH, Gutekunst CA, Landwehrmeyer BG, Lehrach H, Wanker EE (2004) A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. Mol Cell 15(6):853–865. doi:10.1016/j.molcel.2004.09.016, S1097276504005453 [pii]
20. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. Proc Natl Acad Sci U S A 104(21):8685–8690. doi:0701361104, [pii] 10.1073/pnas.0701361104
21. Hamosh A, Scott A, Amberger J, Bocchini C, McKusick V (2005) Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33:D514–D517
22. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362–9367. doi:0903103106, [pii]10.1073/pnas.0903103106
23. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM (2005) Using literature-based discovery to identify disease candidate genes. Int J Med Inform 74(2–4):289–298
24. Hsu C, Huang Y, Hsu C, Yang U (2011) Prioritizing disease candidate genes by a gene interconnectedness-based approach. BMC Genomics 12(3):S25
25. Huynen MA, Snel B, van Noort V (2004) Comparative genomics for reliable protein-function prediction from genomic data. Trends Genet 20(8):340–344
26. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98(8):4569–4574. doi:10.1073/pnas.061034498, 061034498 [pii]
27. Jimenez-Sanchez G, Childs B, Valle D (2001) Human disease genes. Nature 409(6822): 853–855

28. Junker BH, Koschutzki D, Schreiber F (2006) Exploration of biological network centralities with CentiBiN. BMC Bioinformatics 7:219
29. Kaimal V, Sardana D, Bardes EE, Gudivada RC, Chen J, Jegga AG (2011) Integrative systems biology approaches to identify and prioritize disease and drug candidate genes. Methods Mol Biol 700:241–259. doi:10.1007/978-1-61737-954-3_16
30. Kann MG (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform 8(5):333–346
31. Kim YK, Wassef L, Hamberger L, Piantedosi R, Palczewski K, Blaner WS, Quadro L (2008) Retinyl ester formation by lecithin: retinol acyltransferase is a key regulator of retinoid homeostasis in mouse embryogenesis. J Biol Chem 283(9):5611–5621. doi:M708885200, [pii] 10.1074/jbc.M708885200
32. King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science (New York, NY) 188(4184):107–116
33. Kleinberg J (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
34. Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82(4):949–958. doi:S0002-9297(08)00172-9, [pii] 10.1016/j.ajhg.2008.02.013
35. Korstanje R, Paigen B (2002) From QTL to gene: the harvest begins. Nat Genet 31(3):235–236
36. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N et al (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25(3):309–316
37. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M (2004) A map of the interactome network of the metazoan C. elegans. Science (New York, NY) 303(5657):540–543. doi:10.1126/science.1091403, 1091403 [pii]
38. Linghu B, Snitkin ES, Hu Z, Xia Y, Delisi C (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol 10(9):R91. doi:10.1186/gb-2009-10-9-r91, gb-2009-10-9-r91 [pii]
39. Lopez-Bigas N, Ouzounis CA (2004) Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 32(10):3108–3114
40. Mackay TF (2001) Quantitative trait loci in Drosophila. Nat Rev 2(1):11–20
41. Masseroli M, Galati O, Pinciroli F (2005) GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. Nucleic Acids Res 33(Web Server issue):W717–W723
42. Masseroli M, Martucci D, Pinciroli F (2004) GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. Nucleic Acids Res 32(Web Server issue):W293–W300
43. Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev 13(8):523–536. doi:10.1038/nrg3253, nrg3253 [pii]
44. Navlakha S, Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. Bioinformatics 26(8):1057–1063. doi:10.1093/bioinformatics/btq076, btq076 [pii]
45. Ortutay C, Vihinen M (2009) Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies. Nucleic Acids Res 37(2):622–628. doi:gkn982, [pii]10.1093/nar/gkn982
46. Oti M, Ballouz S, Wouters MA (2011) Web tools for the prioritization of candidate disease genes. Methods Mol Biol 760:189–206. doi:10.1007/978-1-61779-176-5_12
47. Oti M, Snel B, Huynen MA, Brunner HG (2006) Predicting disease genes using protein-protein interactions. J Med Genet 43(8):691–698

48. Perez-Iratxeta C, Bork P, Andrade MA (2002) Association of genes to genetically inherited diseases using data mining. Nat Genet 31(3):316–319
49. Perez-Iratxeta C, Wjst M, Bork P, Andrade MA (2005) G2D: a tool for mining genes associated with disease. BMC Genet 6:45
50. Piro RM, Di Cunto F (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J 279(5):678–696. doi:10.1111/j.1742-4658.2012.08471.x
51. Popescu M, Keller JM, Mitchell JA (2006) Fuzzy measures on the gene ontology for gene product similarity. IEEE/ACM Trans Comput Biol Bioinform 3(3):263–274
52. Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, Benini L, Volinia S (2006) TOM: a web-based integrated approach for identification of candidate disease genes. Nucleic Acids Res 34(Web Server issue):W285–W292
53. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437(7062):1173–1178. doi:nature04209, 10.1038/nature04209
54. Sam L, Liu Y, Li J, Friedman C, Lussier YA (2007) Discovery of protein interaction networks shared by diseases. Pac Symp Biocomput 76–87
55. Smith NG, Eyre-Walker A (2003) Human disease genes: patterns and predictions. Gene 318:169–175
56. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122(6):957–968. doi:S0092-8674(05)00866-4, 10.1016/j.cell.2005.08.029
57. Sun PG, Gao L, Han S (2010) Prediction of human disease-related gene clusters by clustering analysis. Int J Biol Sci 7(1):61–73
58. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39(Database issue):D561–D568. doi:10.1093/nar/gkq973, gkq973 [pii]
59. Thornblad TA, Elliott KS, Jowett J, Visscher PM (2007) Prioritization of positional candidate genes using multiple web-based software tools. Twin Res Hum Genet 10(6):861–870
60. Tiffin N (2011) Conceptual thinking for in silico prioritization of candidate disease genes. Methods Mol Biol 760:175–187. doi:10.1007/978-1-61779-176-5_11
61. Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M, Lopez-Bigas N, Ouzounis C, Perez-Iratxeta C, Andrade-Navarro MA, Adeyemo A, Patti ME, Semple CA, Hide W (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Res 34(10):3067–3081
62. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 33(5):1544–1552
63. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res 36(Web Server issue):W377–W384
64. Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y (2011) A guide to web tools to prioritize candidate genes. Brief Bioinform 12(1):22–32. doi:10.1093/bib/bbq007, bbq007 [pii]
65. Turner FS, Clutterbuck DR, Semple CA (2003) POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol 4(11):R75
66. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T,

Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770): 623–627. doi:10.1038/35001009

67. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA (2006) A text-mining analysis of the human phenome. Eur J Hum Genet 14(5):535–542
68. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG (2003) A new web-based data mining tool for the identification of candidate genes for human genetic disorders. Eur J Hum Genet 11(1):57–63
69. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. Nucleic Acids Res 33(Web Server issue):W758–W761
70. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 6(1):e1000641. doi:10.1371/journal.pcbi.1000641
71. Wat MJ, Veenma D, Hogue J, Holder AM, Yu Z, Wat JJ, Hanchard N, Shchelochkov OA, Fernandes CJ, Johnson A, Lally KP, Slavotinek A, Danhaive O, Schaible T, Cheung SW, Rauen KA, Tonk VS, Tibboel D, de Klein A, Scott DA (2011) Genomic alterations that contribute to the development of isolated and non-isolated congenital diaphragmatic hernia. J Med Genet 48(5):299–307. doi:10.1136/jmg.2011.089680, 48/5/299 [pii]
72. White S, Smyth P (2003) Algorithms for estimating relative importance in networks. Paper presented at the KDD '03: proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining
73. Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. Mol Syst Biol 4:189. doi:msb200827, [pii] 10.1038/msb.2008.27
74. Xu J, Li Y (2006) Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics 22(22):2800–2805. doi:btl467, [pii] 10.1093/bioinformatics/btl467
75. Zhu C, Kushwaha A, Berman K, Jegga AG (2012) A vertex similarity-based framework to discover and rank orphan disease-related genes. BMC Syst Biol 6(Suppl 3):S8. doi:10.1186/1752-0509-6-S3-S8, 1752-0509-6-S3-S8 [pii]
76. Zhu M, Zhao S (2007) Candidate gene identification approach: progress and challenges. Int J Biol Sci 3(7):420–427