Improving disease gene prioritization using the semantic similarity of Gene Ontology terms

Andreas Schlicker†, Thomas Lengauer and Mario Albrecht∗
Max Planck Institute for Informatics, Department of Computational Biology and Applied Algorithmics, Campus E1.4, 66123 Saarbrücken, Germany

ABSTRACT

Motivation: Many hereditary human diseases are polygenic, resulting from sequence alterations in multiple genes. Genomic linkage and association studies are commonly performed for identifying disease-related genes. Such studies often yield lists of up to several hundred candidate genes, which have to be prioritized and validated further. Recent studies discovered that genes involved in phenotypically similar diseases are often functionally related on the molecular level.

Results: Here, we introduce MedSim, a novel approach for ranking candidate genes for a particular disease based on functional comparisons involving the Gene Ontology. MedSim uses functional annotations of known disease genes for assessing the similarity of diseases as well as the disease relevance of candidate genes. We benchmarked our approach with genes known to be involved in 99 diseases taken from the OMIM database. Using artificial quantitative trait loci, MedSim achieved excellent performance with an area under the ROC curve of up to 0.80 and a sensitivity of over 70% at 90% specificity when classifying gene products according to their disease relatedness. This performance is comparable or even superior to related methods in the field, albeit using less and thus more easily accessible information.

Availability: MedSim is offered as part of our FunSimMat web service (http://www.funsimmat.de).

Contact: mario.albrecht@mpi-inf.mpg.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

More than 1800 human hereditary disorders are known to be caused by mutations in a single gene (O’Connor and Crystal, 2006). However, most of these diseases are very rare. In contrast, many diseases of major importance to public health, like cancer, diabetes and cardiovascular disorders, are influenced by simultaneous alterations in several genes (Gibson, 2009). In order to identify genes involved in such multi-factorial diseases, genomic linkage and association studies are performed (Altshuler et al., 2008; Cordell and Clayton, 2005; Teare and Barrett, 2005). The genomic regions resulting from these studies may comprise as many as several hundreds of candidate disease genes, most of them unrelated to the disease of interest. Experimental testing of the complete list of candidate genes is generally impractical because of the time and cost involved in such an extensive procedure. Therefore, several studies examined the specific properties of genes and their products known to be associated with human genetic disorders and explored networks linking diseases based on the involved genes (Feldman et al., 2008; Goh et al., 2007; Jimenez-Sanchez et al., 2001; Lee et al., 2008; van Driel et al., 2006). In particular, the discovered relationships between properties of genes and gene products as well as their involvement in genetic disorders are exploited by a number of bioinformatics approaches for ranking and prioritizing disease gene candidates (Ala et al., 2008; Ideker and Sharan, 2008; Kann, 2007, 2010; Navlakha and Kingsford, 2010; Otii and Brunner, 2007; Tranchevent et al., 2010; Turner et al., 2003; van Driel and Brunner, 2006; van Driel et al., 2006; Yu et al., 2008).

Most computational approaches rely on the integration of several sources of heterogeneous data such as sequence features, gene expression data and protein–protein interactions (PPIs). For example, PROSPECTR is a sequence-based approach that uses decision trees trained on features such as the length of gene and protein sequences and the number of exons (Adie et al., 2005). The subsequent method SUSPECTS by the same authors combines sequence features with gene expression, protein domains and Gene Ontology (GO) term similarity of candidates and known disease genes (Adie et al., 2006). Endeavour is another method that relies on the integration of biological evidence resulting from many different kinds of data, for instance, PPIs, pathways, gene expression and sequence similarity (Aerts et al., 2006). The characteristics of known disease genes were extracted from each data source separately to rank candidate genes; the resultant ranking lists were then combined to a final overall ranking.

Recently, several methods have been published (Chen et al., 2009; Franke et al., 2006; Ortuñay and Vihinen, 2009; Ozgür et al., 2008; Shrirer et al., 2008) that build on both interaction networks and GO annotations (Ashburner et al., 2000). In particular, Chen et al. (2009) applied different algorithms originating from the analysis of social and web networks to disease gene prioritization. They concluded that methods using functional annotation are generally better than network-based methods, but that network data provide some valuable information. Ortuñay and Vihinen (2009) integrated GO annotation and protein interactions for finding genes involved in immunodeficiencies. To this end, three different network topology parameters were computed pertaining to an interaction network of genes known to be related to the immune system. For each of these parameters, a set of genes was selected from the gene network and then subjected to GO enrichment analysis. Genes received higher priority if they were annotated with enriched terms and achieved some significant network parameter value.

A number of methods for disease gene prioritization uses similarity measures for phenotypes, which leverage cross-references to structured vocabularies (Chen et al., 2007; Freudenberg and
In the following, we present MedSim, a novel approach to
disease gene prioritization that exploits the similarity between the
functional annotations of diseases and candidate genes (Fig. 1). This
methodological advance is in contrast to other methods that
consider only identical annotations or are based on GO enrichment
computations. In particular, we automatically derive functional
profiles consisting of GO terms for a certain disease phenotype based
on the genes and proteins that are already known to be related to the
phenotype.

Since the annotation of the human genome with GO terms is
rather incomplete, we introduce and test several new strategies for
automatically extending the available annotations of disease and
candidate genes or proteins. The resulting functional profiles are
compared with each other using GO and our sophisticated functional
similarity measures (Schlicker et al., 2007). Using different sets
of proteins encoded by known disease genes, we demonstrate
that our novel method allows for assigning known disease genes
specifically to the correct phenotype. Most importantly, we show that
MedSim is able to significantly outperform previous more
complex methods that rely on more diverse and voluminous,
and thus harder accessible data and we further explore the effect of
different semantic similarity measures on prediction performance.

MedSim also affords the distinction of disease phenotypes with
a common functional basis from unrelated phenotypes. Finally, we
implemented the best MedSim method in our FunSimMat web server
(http://www.funsimmat.de), making it easily usable by biological
and medical users (Schlicker and Albrecht, 2010).
Table 1. Summary of the different annotation strategies used to create functional profiles of diseases

| Annotation strategy | GO annotation source                      |
|---------------------|------------------------------------------|
| AS-base             | Known disease genes/proteins            |
| AS-ortho            | Known disease genes/proteins            |
| AS-inter            | Orthologs of known disease genes/proteins|
| AS-sem              | Interaction partners of known disease genes/proteins |
|                     | Semantically similar terms               |

The table lists sources of GO annotation used by the different annotation strategies. Term filtering can be applied to functional profiles created by any of these annotation strategies.

annotated disease genes and proteins may not cover all functions and processes involved in the respective disease. Therefore, we explored several possibilities to automatically extend the available annotation. The second annotation strategy (AS-ortho) adds GO terms from mouse orthologs of human disease proteins to the functional profile, and the third annotation strategy (AS-inter) augments the profile with GO terms from direct interaction partners of disease proteins (Table 1). Both strategies involve the removal of redundant GO terms after adding the new terms to the profile. A fourth strategy for expanding the functional profiles (AS-sem) is based solely on GO. The simRel measure (see Section 2.6 below) is used to identify terms that are highly related to at least one other term in the same profile. Two different simRel cut-offs, 0.90 and 0.95, are applied for selecting and adding related terms to a profile. Functional profiles of candidate disease genes and proteins are always generated by applying the same annotation strategy as used for the disease phenotype.

If a protein has many interaction partners with diverse functions or the dataset contains false positive interactions, the described automatic strategies might lead to a diffuse functional profile containing diverging GO annotations for BPs, MFs and CCs. Therefore, we implemented a term filtering step for removing unrelated terms from the functional profiles. In this step, terms are retained only if they have a simRel score above a predefined threshold with at least one other term in the profile. For example, if we consider a functional profile consisting of four GO terms and two of these terms are similar to each other and the other two terms are not related to any term in the profile, the latter two are removed from the profile. In contrast, if the latter two terms are similar to each other as well, all four terms are retained in the profile. We tested the two simRel thresholds 0.60 and 0.80. The term filtering step was applied to all functional profiles consisting of at least three GO terms. If the functional profile of a disease contained no GO term pair with simRel exceeding the threshold, the respective disease was not included into the benchmark.

2.3 Benchmark set 1

Several prioritization methods assess the probability of a gene or protein to be generally associated with some disease, but are unspecific for the disease. In order to test whether MedSim allows for specifically assigning known disease genes and proteins to the correct disease phenotype, we conducted leave-one-out cross-validation on a set of diseases and known disease-associated proteins. For this benchmark, we selected a preliminary set of 99 OMIM disease phenotypes, each of which is associated with at least three known disease genes (Supplementary Table S5). For each of these phenotypes, one disease protein was randomly selected and removed. Subsequently, the functional profiles of the 99 phenotypes were derived using annotation strategies AS-base, AS-ortho or AS-inter based on the remaining known disease genes. Disease phenotypes were discarded if either the phenotype or the randomly selected protein was not annotated with terms from all three GO ontologies. This led to benchmark set 1 consisting of 78 phenotypes with 78 randomly selected known disease genes. Five of these proteins are known to contribute to two diseases in the test set and were coincidentally chosen for both phenotypes. Supplementary Table S1 summarizes the number of GO terms annotated to phenotypes and randomly selected proteins in benchmark set 1.

2.4 Benchmark set 2

Genomic loci found to be associated with a disease may contain up to several hundred candidate genes. The second benchmark simulates such a genomic experiment, which results in a quantitative trait locus (QTL) and the corresponding list of candidate disease genes. For each of the 519 disease gene-encoded proteins associated with one of the 99 phenotypes in benchmark set 1, leave-one-out cross validation was performed for classifying the protein according to its disease relatedness. After a protein \( p \) was removed from the list of known proteins for some disease, the functional profile of this disease was derived using the remaining other proteins. An artificial QTL (\( a\text{-QTL} \)) of size 10 Mbp was centered at the genomic start position of the gene encoding \( p \), and all proteins translated from any gene in this \( a\text{-QTL} \) were added to the list of putative disease proteins. Benchmark set 2 contains 519 different \( a\text{-QTLs} \) for 99 phenotypes. All four annotation strategies were applied to annotate benchmark set 2. Additionally, term filtering with both thresholds 0.60 and 0.80 was applied together with AS-base and AS-inter, as well as term filtering using threshold 0.80 with AS-sem. As control, random PPIs were used for AS-inter (Section 2.1).

2.5 Benchmark set 3

Several approaches, for example, Endeavour (Aerts et al., 2006), had been benchmarked using random artificial QTLs (\( r\text{-QTLs} \)) that contain one known disease gene and 99 random genes. To facilitate a performance comparison between MedSim and these methods, we created a third benchmark set. This set was compiled using the same set of phenotypes as benchmark set 2 but differs in the methodological details of creating the \( r\text{-QTLs} \). Here, each disease protein annotated with terms from all three ontologies was complemented with 99 proteins randomly drawn from the set of all human proteins annotated with terms from all three ontologies. Benchmark set 3 consists of 287 distinct \( r\text{-QTLs} \) for 99 different phenotypes. To the phenotypes and \( r\text{-QTLs} \) in this benchmark set, we applied AS-base without and with term filtering (threshold 0.80) as well as AS-sem (cut-off 0.95) with term filtering (threshold 0.80).

2.6 Functional similarity measures

The similarity between functional profiles of diseases and candidate proteins was computed using the Functional Similarity Search Tool (FSST version 1.3.1) (Schlicker et al., 2007). The computed functional similarity scores are based on a semantic similarity measure for comparing two GO terms. The simRel score (Schlicker et al., 2007), which assesses the differences and commonalities between GO terms, was used to determine the semantic similarity of GO terms. This score is affected by the level of detail of the annotated terms. In order to find out whether the performance of MedSim depends on the choice of the semantic similarity measure, Lin’s (1998) measure was used as well. This similarity score measures the commonalities and differences between two GO terms and is not affected by the degree of specificity of some term as given by the GO hierarchy. To compare two functional profiles, several similarity scores are evaluated: BPscore for BP, CCscore for CC, MFscore for MF, rfunSim combining BPscore and MFscore and rfunSimAll combining BPscore, CCscore and MFscore. A detailed description of all semantic and functional similarity scores can be found in the Supplementary Data.

2.7 MedSim implementation

We implemented the MedSim approach in our FunSimMat database and web service (http://www.funsimmat.de). FunSimMat contains precomputed...
Additionally, we calculated the sensitivity and specificity of the predictions. Sensitivity is the percentage of correctly identified disease proteins ranked above a preset rank or score cut-off. Specificity is the percentage of proteins not involved in the disease ranked below this cut-off. When stating sensitivity values, we will always refer to a specificity threshold of 90%. The performance values presented in the remainder of the text constitute conservative estimates due to the following two reasons. Firstly, the ranking list of proteins may contain several proteins associated with a disorder, but solely the randomly left-out protein is considered a true positive. Second, proteins labeled as true negative might, in fact, be as yet unknown true positives.

A detailed discussion of the results for benchmark set 1 can be found in the Supplementary Data. Briefly, MedSim achieved an AUC of up to 0.81 on this set using strategy AS-ortho. This shows that MedSim effectively assigns top ranks to the correct protein in a list of known disease proteins.

Benchmark set 2 was designed for simulating the most common application scenario for disease gene prioritization methods. The task is to rank a list of candidate disease genes or proteins such that the most likely candidates are on top of the list (Fig. 1). Benchmark set 2 contained 519 aQTLs of size 10Mbp, which encompass 312 protein interaction data. We also applied AS-orto to benchmark set 2 using two different simRel cut-offs, 0.90 or 0.95, for adding terms. In both cases, the AUC and sensitivity values are similar to the scores obtained with AS-base (Supplementary Figs S15 and S17).

When inspecting the availability of GO annotation (Supplementary Table S3), it becomes evident that AS-ortho improves the coverage with functional annotation while preserving the performance. AS-inter increases coverage even more, but it negatively affects the prediction performance slightly. We carefully checked that this performance decrease is not due to an implementation error, and the application of AS-inter to a set of random PPIs yielded AUC values as expected for random prioritization (see Supplementary Data for details).

By increasing the coverage, AS-ortho and AS-inter potentially allow for ranking candidate disease genes and proteins that are not amenable to analysis using AS-base due to the lack of direct GO annotation. Thus, we studied the results with the rfunSim and rfunSimAll scores for aQTLs to which we could not apply the strategy AS-base. For these cases, the sensitivity of MedSim using AS-orto and AS-inter is 46% and 25%, respectively, with rfunSimAll. This indicates that both annotation strategies help ranking candidates if known human disease genes and proteins are not yet annotated with GO terms.

3.2 Improving prediction performance by filtering dissimilar terms

The findings above indicate that prediction performance is negatively influenced by semantically unrelated terms. Thus, we applied a semantic similarity term filter to the functional profiles of benchmark set 2 created by AS-base and AS-inter. The term filter removes all terms that do not have a simRel score greater than a specific threshold (here, 0.60 or 0.80) to any other term in the profile. With respect to AUC, the results are inconclusive for AS-base (Fig. 2 and Supplementary Figs S5 and S6). The AUC drops slightly for BP and MF using both thresholds, but the AUC of CC and of the combined scores are larger than without term filtering. The best AUC is achieved with AS-base using the rfunSim score (AUC 0.85) and term filtering with the threshold 0.80. If the functional profiles are complemented by PPIs in AS-inter, term filtering improves the AUC in most cases (Supplementary Figs S13 and S14). The rfunSimAll score
AUC of the BPscore and the MFscore, but increases the sensitivity. Therefore, term filtering has a much higher impact on the combined term filtering.

Such examples are inflammatory bowel disease (OMIM #266600) and familial hypertrophic cardiomyopathy (OMIM #192600). Using AS-base with term filtering (threshold 0.80) improves the sensitivity (57%) over the use of other methods that are based on GO annotations.

3.3 Performance on rQTLs increased over aQTLs

Benchmark set 3 was created in a fashion that is similar to previous publications for facilitating a comparison of the performance of MedSim and other prioritization methods. This benchmark set consists of 287 rQTLs, each containing 100 proteins annotated with BP, MF, and CC. Functional profiles for diseases in benchmark set 3 were derived using AS-base without and with term filtering (threshold 0.80), and AS-sem (cut-off 0.95) with term filtering (threshold 0.80). The ranking results for benchmark set 3 are listed in Supplementary Table S8. Using AS-base (Fig. 2 and Supplementary Fig. S19), the best performance is achieved with the combined scores, rfunSim and rfunSimAll (AUC 0.84). Using each combined score improves the sensitivity (57%) over the use of any other score (42–53%). Applying term filtering, deteriorates the AUC of the BPscore and the MFScore, but increases the sensitivity of the CCscore from 42% to 57% and of the MFScore from 47% to 51% (Supplementary Fig. S20). In case of the combined scores, both performance measures improve if AS-base is applied with term filtering (Fig. 2). The rfunSimAll score reaches a maximal AUC of 0.90 and a sensitivity of 73%. Virtually the same AUC and sensitivity are achieved when applying term filtering to AS-sem (Supplementary Fig. S21).

The impact on the coverage with GO annotation caused by the removal of unrelated GO terms from functional profiles was already described for benchmark set 2. For benchmark set 3, term filtering reduces the coverage to 36–59% in the cross-validations (Supplementary Table S4). To calculate the combined scores, the functional profiles have to contain either both BP and MF terms for rfunSim or terms from all three ontologies for rfunSimAll. Therefore, term filtering has a much higher impact on the combined scores, reducing the coverage to ~90% compared to ~59% without term filtering.

3.4 Results for exemplary diseases

Several inherited diseases involve cellular processes whose functional relationship on the molecular level is not clear yet. One such example is inflammatory bowel disease (OMIM #266600) (Schreiber et al., 2005). UniProtKB currently maps five proteins reported by genome-wide association studies to this disease (Cho, 2008): the nucleotide-binding oligomerization domain-containing protein 2 (NOD2, Q9HC29), the solute carrier family 22 members 4 and 5 (SLC22A4, Q9H015; SLC22A5, Q76082), interleukin 10 (IL10, P22301) and the interleukin 23-receptor (IL23R, Q5VWK5). In benchmark set 2, MedSim ranks all proteins except NOD2 in the top 22% when applying strategy AS-inter and the rfunSimAll score. Notably, SLC22A5 and SLC22A4 are ranked in the top 6% and top 11%, respectively. NOD2 is ranked in the top 11% using the rfunSim score and strategy AS-base. Further exemplary prioritization results for photosensitive trichothiodystrophy (OMIM #601675), susceptibility and resistance to human immunodeficiency virus type 1 (HIV-1) (Q909423), Parkinson disease (OMIM #168600), prostate cancer (OMIM #176807) and familial hypertrophic cardiomyopathy (OMIM #192600) are described in the Supplementary Data.

3.5 Comparison with other prioritization methods

First of all, it is important to note that several aspects hamper a fully objective comparison between different disease gene prioritization methods. Many methods are not readily available, making it impossible to apply them on exactly the same benchmark set. Furthermore, the biological contents of the datasets used by different methods influences the prediction results, which limits any detailed comparison. Nevertheless, it is possible and necessary to conduct a general performance comparison by utilizing large-scale benchmark sets that are created in a methodologically similar way. To this end, the procedure applied for creating benchmark set 3 is very similar to previous publications (Aerts et al., 2006; Chen et al., 2007).

Endeavour (Aerts et al., 2006) is a state-of-the-art method based on the integration of multiple data sources. It can be used to prioritize genes based on single data sources or a combination of different sources. The authors validated their approach with a benchmark set of rQTLs that were constructed with a strategy similar to benchmark set 3. With GO annotation as the only data source, Endeavour achieved a sensitivity of 73% when relying only on GO annotation. In case of prioritization using all data sources, Endeavour was reported to achieve an AUC value of 0.87 and a sensitivity of 74% (at 90% specificity), which is comparable to the performance of the less complex MedSim approach using only GO annotations as data source.

Recently, Chen et al. (2007) devised the ToppGene method that uses annotation with terms from the Mammalian Phenotype (MP) ontology (Smith et al., 2005) among other data sources, for instance, biomedical literature and protein interactions. For comparing their tool to Endeavour, the authors used a benchmark similar to benchmark set 3. The reported AUC values are 0.91 and 0.89 with and without using MP annotation, respectively, and a sensitivity of 74% with MP annotation. This means that MedSim performs comparatively, while using a much simpler prediction approach based on GO annotation alone. Further comparisons to other methods that are based on GO annotations and PPI data are provided in the Supplementary Data.

4 CONCLUSIONS

We presented the new approach MedSim for disease gene prioritization that introduces several novel strategies for
AUC and sensitivity for benchmark set 3 are generally higher. Therefore, it is important to take into account how a benchmark set was constructed when comparing the performance of different prioritization approaches. All benchmarks used for validating the MedSim approach were compiled in such a way that every candidate gene and proteins accurately if the latter do not already possess a suitable GO annotation. In particular, term filtering greatly increases coverage (up to 41%), but can have a negative impact on the overall performance. Nevertheless, our results provide evidence for the fact that the use of GO annotations from orthologous mouse orthologs to human proteins is particularly useful for increasing the coverage with GO annotation without lowering performance. Adding annotation from protein interaction partners greatly increases coverage (up to 41%), but can have a negative impact on the overall performance. Nevertheless, our results provide evidence for the fact that the use of GO annotations from orthologous mouse proteins or protein interaction partners aids in ranking candidate genes and proteins accurately if the latter do not already possess a suitable GO annotation. In particular, term filtering increases the performance and allows for finding a tradeoff between high coverage and high performance, especially when applied to functional profiles created with the help of protein interaction data.

In general, our comparison of the prediction results from different benchmarks demonstrated that the assessed performance of a method depends on the actual construction of the benchmark set. The AUC and sensitivity for benchmark set 3 are generally higher than for benchmark set 2 using the same annotation strategy for both sets. This effect was also observed in our exemplary study of susceptibility to HIV-1. The effect is most likely due to the fact that the rQTLs in benchmark set 3 are of smaller size on average and that the unrelated proteins are randomly drawn from the whole proteome.

Therefore, it is important to take into account how a benchmark set was constructed when comparing the performance of different prioritization approaches. All benchmarks used for validating the MedSim approach were compiled in such a way that every candidate list contains exactly one true positive. However, in real settings, it might happen that none of the candidates is related to the disease of interest. In such situations, the whole list might be rejected if no candidate scores significantly better than the rest of the candidates. If the functional similarity scores obtained for different disease are compared, it is important to normalize the absolute values because they are not directly comparable.

In addition, we presented strategies for automatically extending the existing GO annotation of human genes and proteins using orthologs from model organisms or interaction partners. Our approach is not restricted to GO as functional annotation source. Since the semantic and functional similarity measures used are applicable to any vocabulary that is organized as a tree or directed acyclic graph, MedSim could also leverage annotations from other vocabularies like the Human Phenotype Ontology (Robinson et al., 2008). The availability of functional annotations is generally expected to improve considerably in the near future because of comprehensive annotation efforts like the Reference Genome Annotation Project (Reference Genome Group of the Gene Ontology Consortium, 2009). The functional profile of a phenotype might also be used to predict functions for uncharacterized genes and proteins implicated in this phenotype. In particular, AS-ortho and AS-base are useful for transferring GO annotations from functionally annotated orthologs from model organisms or interaction partners, respectively.

It should be noted that the use of OMIM has some limitations. First, OMIM was initiated as database of Mendelian disorders and contains many entries describing single genes. These cannot be used for benchmarking methods that aim at the prioritization of candidates for polygenic diseases. Second, the information in OMIM is manually curated, which increases the quality but is labor-intensive. Therefore, OMIM does not contain all currently known genes affecting diseases as it became apparent in our exemplary study of susceptibility to HIV-1. Third, OMIM does not provide a hierarchical classification of phenotypes and contains free-text descriptions. This renders it difficult to automatically derive ontologies like the Human Phenotype Ontology and to use this information without further manual curation.

Finally, the most promising MedSim annotation strategy, AS-base with term filtering (threshold 0.80), is available via our FunSimMat online service (Schlicker and Albrecht, 2010). In particular, FunSimMat contains functional profiles for all OMIM disease entries and human proteins derived by annotation strategy AS-base with and without term filtering (threshold 0.80). The pre-computation of functional similarity scores affords the fast ranking of genes in QTLs or even of the whole genome with respect to the disease of interest. Moreover, the MedSim approach can be easily incorporated into other disease gene prioritization methods.

ACKNOWLEDGEMENTS

We are grateful to Fidel Ramirez for providing the protein interaction data used in this study and to Alejandro Piroini for providing data processing scripts.

Funding: German National Genome Research Network (NGFN, in part); German Research Foundation (DFG, contract number KFO 129/1-2, in part); European Commission (grant number LS9G-CT-2003-503265). The work was conducted in the context of the DFG-funded Cluster of Excellence for Multimodal Computing and Interaction and the BioSapiens Network of Excellence.

Conflict of Interest: none declared.

REFERENCES

Adie,E.A. et al. (2005) Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics 6, 55.
Adie,E.A. et al. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics, 22, 775–774.
Aerts,S. et al. (2006) Gene prioritization through genomic data fusion. Nat Biotechnol., 24, 537–544.
Ala,U. et al. (2008) Prediction of human disease genes by human-mouse conserved coexpression analysis. PLoS Comput Biol., 4, e1000483.
Alkholdar,D. et al. (2008) Genetic mapping in human disease. Sciento, 322, 881–888.
Amburne M. et al. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25, 25–29.
Brygand,A.C . et al. (2008) InParanoid 6: eukaryotic ortholog clusters with impurities. Nucleic Acids Res., 36, D263–D266.
Blake,J.A . et al. (2009) The Mouse Genome Database: gene symbols, transcripts, SNPs and CPG islands. Nucleic Acids Res., 37, D712–D719.
Chat-Aryamontri,A . et al. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res., 35, D572–D574.
Chen,J . et al. (2007) Improved human disease-causal gene prioritization using mouse phenotype. BMC Bioinformatics, 8, 392.
Chen,J . et al. (2009) Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics, 10, 73.
Chu,J.H . (2008) The genetics and immunopathogenesis of inflammatory bowel disease. Nat Rev. Immunol., 8, 456–466.
Cordell,H.J . and Clayton,D.G . (2009) Decanalization and the origin of complex disease. Proc. Natl Acad. Sci. USA, 106, 17390–17395.
Cordell,H.J . and Clayton,D.G . (2005) Genetic association studies. Nat. Genet., 37, S110–S115.
Gibson,G . (1998) An information-theoretic definition of similarity. In: Shavlik, J.W . (ed.) Data Mining and Knowledge Discovery Handbook (pp. 296–304). Kluwer, Dordrecht, The Netherlands, pp. 296–304.
Goh,K.I . et al. (2007) The human disease network. Proc. Natl Acad. Sci. USA, 104, 8665–8668.
Kann,M.G . (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief. Bioinform., 8, S33–S46.
Kanehisa,M . et al. (2008) The Human Phenome Interactome Network of Protein Complexes. Genome Res., 18, 1222–1230.
Kerrien,S . et al. (2007) IntAct-open source repository for molecular interaction data. Nucleic Acids Res., 35, D561–D563.
Kierlian,F . et al. (2007) A human phenotype-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol., 25, 300–316.
Lage,K . et al. (2007) A human phenotype-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol., 25, 300–316.
Lage,K . et al. (2007) The implications of human metabolic network topology for disease comorbidity. Proc. Natl Acad. Sci. USA, 104, 9845–9885.
Lin,D . (1998) An information-theoretic definition of similarity. In: Shavlik, J.W . (ed.) Data Mining and Knowledge Discovery Handbook (pp. 296–304). Kluwer, Dordrecht, The Netherlands, pp. 296–304.
Lowe,H . and Barrett,G . (1994) Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA, 271, 1103–1108.
Novotny,A . and Karpfmond,C . (2010) The power of protein interaction networks for associating genes with diseases. Bioinformatics, 26, 1057–1063.
O’Connor,T.P . and Crystal,R.G . (2006) Genetic medications: treatment strategies for hereditary disorders. Nat. Rev. Genet., 7, 261–276.
Oti,M . and Brunner,H.G . (2007) The modular nature of genetic diseases. Clin. Genet., 71, 1–11.
Ortagu,A . et al. (2008) Identifying gene-disease associations using centrality on a literature mind map: gene-interaction network. Bioinformatics, 24, 2277–2283.
Perez-Brato-C . et al. (2002) Association of genes to genetically inherited diseases using data mining. Nat. Genet., 31, 316–319.
Perez-Brato-C . et al. (2007) Update of the G2D tool for prioritization of gene candidates to inherited diseases. Nucleic Acids Res., 35, W212–W216.
Prasad,T.K . et al. (2009) Human Protein Reference Database-2009 update. Nucleic Acids Res., 37, D767–D772.
Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology’s Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput. Biol., 5, e1000431.
Robinson,P.N . et al. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet., 83, 610–615.
Rosetta,A . et al. (2008) CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res., 36, D666–D670.
Salwinski,L . et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res., 32, D449–D451.
Schlicker,A . et al. (2007) DITFar: investigating biological processes and biochemical activities along the taxonomic tree. Genome Biol., 8, R33.
Schröder,S . et al. (2005) Genetics of Crohn disease, an archetype inflammatory barrier disease. Nat. Rev. Genet., 6, 576–588.
Sharan,R . et al. (2003) Commonality of functional annotation: a method for prioritization of candidate genes from genome-wide linkage studies. Nucleic Acids Res., 31, c26.
Smith,C.L . et al. (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol., 6, R7.
Stein,M.D . and Barrett,J.H . (2005) Genetic linkage studies. Nat. Genet., 37, 1344–1352.
Tennissen,L.C . et al. (2010) A guide to web tools to prioritize candidate genes. Brief. Bioinform., in press.
Tumer,J.S . et al. (2003) POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol., 4, R75.
UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res., 37, D149–D154.
van den Ma, M.A . et al. (2006) A text-mining analysis of the human phenome. Eur. J. Hum. Genet., 14, 535–542.
vandrieling,M . et al. (2006) Bioinformatics methods for identifying disease gene candidates. Hum. Genomics, 2, 429–432.
Velmankar,S . et al. (2005) E-MeD: an integrated data resource for bioinformatics. Nucleic Acids Res., 33, D262–D265.
Wei,X . et al. (2008) Network-based global inference of human disease genes. Mol. Syst. Biol., 4, 189.
Yilmaz,S . et al. (2009) Gene-disease relationship discovery based on model-driven data integration and database view definition. Bioinformatics, 25, 230–236.
Yu,S . et al. (2008) Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. Bioinformatics, 24, i419–i425.