BIOINFORMATICS

Vol. 26 ECCB 2010, pages i561–i567
doi:10.1093/bioinformatics/btq384

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 0.90 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; Oti 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 proteins (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; Ortutay and Vihinen, 2009; Ozgür *et al*., 2008; Shriner *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. Ortutay 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

*To whom correspondence should be addressed.

†Present address: The Netherlands Cancer Institute, Division of Molecular Biology, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

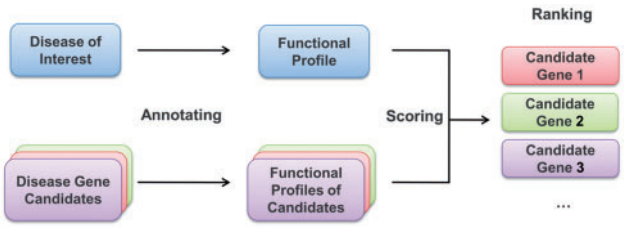

Fig. 1. Flow chart of the MedSim approach. First, the functional profiles of the disease of interest and the disease gene candidates are created using one of the annotation strategies. Afterwards, the functional profile of the disease is scored against each functional profile of a candidate, and the candidates are ranked according to this functional similarity score.

Propping, 2002; Lage *et al*., 2007; Oti and Brunner, 2007; Perez-Iratxeta *et al*., 2002, 2007; Robinson *et al*., 2008; Tiffin *et al*., 2005; van Driel and Brunner, 2006; van Driel *et al*., 2006; Wu *et al*., 2008; Yilmaz *et al*., 2009; Yu *et al*., 2008). Different controlled vocabularies such as MeSH (Lowe and Barnett, 1994) and eVOC (Kelso *et al*., 2003) have already been utilized, and the ACGR method by Yilmaz *et al*. (2009) is specifically based on manual annotation of diseases with GO terms. Candidate genes are selected based on the number of annotated GO terms shared with the disease in question. Subsequently, each candidate is assigned a similarity value based on the annotation similarity to the input disease. The authors did not perform a large-scale validation of their approach, but limited themselves to using three rare syndromes (AICARDI syndrome, CHARGE syndrome and focal dermal hypoplasia) as case studies. The required manual annotation of the disease with GO terms is a major hurdle for the large-scale application of the ACGR method.

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).

2 METHODS

2.1 Data sources

OMIM is a database of human genes and genetic disorders. OMIM entries describe either a single gene that is involved in some genetic disease or a phenotype with known or putative, but unknown, genetic basis. We extracted all phenotypes (entries starting with '#' or '%') from OMIM (downloaded on October 10, 2007). The mapping of proteins encoded by human disease genes to the OMIM phenotypes was obtained from UniProtKB release 12.3 (UniProt Consortium, 2009). Additionally, protein annotations with GO terms from all three ontologies, that is, biological process (BP), molecular function (MF) and cellular component (CC), were extracted from this UniProtKB release. We included annotations based on all GO evidence codes in our analysis; ∼62% of the human GO annotations in our dataset were derived by automatic methods (IEA).

Our set of human PPIs was compiled from the Human Protein Reference Database (HPRD, version 7) (Prasad *et al*., 2009), IntAct (downloaded on May 16, 2008) (Kerrien *et al*., 2007), the Molecular Interactions Database (MINT, downloaded on April 7, 2008) (Chatr-Aryamontri *et al*., 2007), the Database of Interacting Proteins (DIP, downloaded on February 14, 2008) (Salwinski *et al*., 2004), protein complexes extracted from SIFTS (downloaded on March 4, 2008) (Velankar *et al*., 2005) and the CORUM database (downloaded on May 19, 2008) (Ruepp *et al*., 2008). All protein and gene identifiers used by these sources were mapped to UniProtKB accession numbers. Since members of the same protein complex possibly affect the same diseases, the matrix model (all possible pairs of interacting proteins in the complex) was chosen for decomposing protein complexes into pairwise PPIs. A set of random PPIs was created by keeping one partner of an interaction fixed and randomly assigning a new partner from the interacting proteins.

Mouse orthologs for human proteins were obtained from Inparanoid version 6.1 (Berglund *et al*., 2008). Mouse and human proteins with an inparalog score of 1.0 were extracted as ortholog pairs from each Inparanoid cluster. An inparalog score of 1.0 indicates that the two proteins form the reciprocally best matching pair of orthologs. MGI (Blake *et al*., 2009) and Ensembl (Hubbard *et al*., 2009) accessions used by Inparanoid were converted to UniProtKB accessions using data from Ensembl BioMart (downloaded on May 14, 2008). Additionally, the chromosomal location of human genes and the cross-references to UniProtKB proteins were obtained via BioMart on October 21, 2008.

2.2 Functional profiles

Human diseases are usually described using natural language and are annotated with genes or proteins known to be involved in the respective diseases. However, they are not directly annotated with structured vocabularies like GO. GO consists of the three ontologies BP, MF and CC, which are organized as directed acyclic graphs (Ashburner *et al*., 2000). Biological concepts are represented as nodes in these graphs and relationships between concepts as edges. If a gene product is annotated with a GO term, the so-called 'true path rule' states that all of its parents are also valid annotations.

For functional comparisons, we developed several new strategies for automatically annotating OMIM disease entries with GO terms (Table 1). In the remainder of this article, we refer to the GO annotation of a disease phenotype or a candidate gene product as its functional profile. The first annotation strategy (AS-base) transfers all GO terms annotated to proteins encoded by known disease genes in UniProtKB to the corresponding OMIM entry. Genes and proteins are often annotated with terms from different levels of the GO hierarchy, which can lead to functional profiles that contain ancestral terms. Since annotation with a term implies annotation with all its predecessors, ancestral terms are redundant. Therefore, a term is removed if one of its descendants from the GO hierarchy is also contained in the functional profile.

In case of AS-base, OMIM entries cannot be annotated if the known disease genes and proteins lack any GO annotation. Furthermore, the

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
	Orthologs of known disease genes/proteins
AS-inter	Known disease genes/proteins
	Interaction partners of known disease genes/proteins
AS-sem	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 gene products 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 proteins (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 proteins. 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 proteins. 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 associated proteins. An artificial QTL (aQTL) 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 aQTL were added to the list of putative disease proteins. Benchmark set 2 contains 519 different aQTLs 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 QTL (rQTLs) 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 rQTLs. 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 rQTLs for 99 different phenotypes. To the phenotypes and rQTLs 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 apply the best-match average approach, which determines whether a function contained in one profile is also contained in the second profile. The 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, but 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

functional similarity values for proteins and protein families, accessible through a web front-end as well as XML-RPC and RESTlike interfaces. The functional profiles for all OMIM entries and human proteins in UniProtKB were derived using strategy AS-base without and with term filtering (threshold 0.80), and all functional scores are pre-calculated. The FunSimMat web page offers a simple HTML form for prioritizing a list of candidates, which requires the user's input of the OMIM accession of a specific disease and the UniProtKB accessions of the corresponding candidate disease proteins. The results table contains the candidates ranked by the functional similarity score. An alternative for providing a candidate list is the possibility of scoring all human proteins against the disease of interest. Additionally, programmatic access to the data is possible through the XML-RPC and RESTlike interfaces.

3 RESULTS

3.1 Performance and coverage using different annotation strategies

To measure the ability of MedSim to detect the correct protein for each disease, we applied receiver operating characteristic (ROC) analysis and determined the area under the ROC curve (AUC). 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. First, 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 contains 519 aQTLs of size 10 Mbp, which encompass 312 proteins on average, including one known disease protein. FSST was used to calculate functional similarity between diseases and proteins in the corresponding aQTLs. Supplementary Table S2 displays the number of annotated diseases and proteins in the aQTLs, and Supplementary Table S3 contains the mean and median number of annotated terms for benchmark set 2. The results for benchmark set 2 using the different annotation strategies are listed in Supplementary Table S7. Regarding strategy AS-base, the best prediction AUC of 0.81 is achieved using the BPscore and the rfunSim score with a sensitivity of 0.51 and 0.50, respectively (Fig. 2 and Supplementary Fig. S4). Adding ortholog annotation leads to virtually identical AUC (Supplementary Fig. S7) and sensitivity values. However, prediction performance using MFscore drops slightly, which also affects the results obtained with the rfunSim score. AS-inter performs worse, the best AUC being 0.71 for the rfunSimAll score (Supplementary Fig. S8). Sensitivity, however, is only slightly decreased by adding

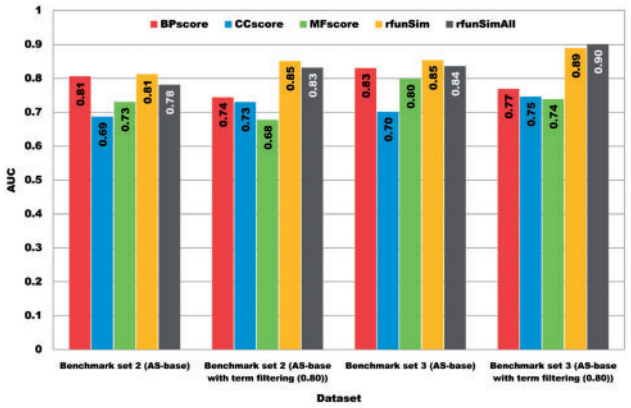

Fig. 2. AUC values of MedSim on benchmark sets 2 and 3 using AS-base with term filtering (0.80) and without.

protein interaction data. We also applied AS-sem 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-ortho 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 has an AUC of 0.82 using AS-inter and term filtering (threshold 0.80), which is even better than the best performance of AS-base without term filtering. The sensitivity values show the same trend, and the maximum is reached at 65% using AS-base with term filtering (threshold 0.80). The annotation coverage of proteins in aQTLs is lower after applying the term filtering procedure, but when using AS-inter, it is about as high as with AS-base without term filtering. In case of the combined scores, rfunSim and rfunSimAll, however, the coverage is significantly lower using term filtering. Applying term filtering (threshold 0.80) before adding terms based on high-semantic similarity does not improve the results compared to term filtering alone.

We have already shown above that adding terms from protein interaction partners helps ranking candidates in aQTLs that are not amenable to analysis with AS-base. When considering only cases in which the disease or the left-out protein could not be annotated using AS-base with term filtering (threshold 0.80), AS-inter with term filtering achieves a sensitivity of 31% with rfunSim and 36% with rfunSimAll. This further confirms that data from PPIs aids in identifying disease-related proteins if known human disease proteins are not annotated with GO terms.

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 (AUC 0.85) 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 ∼10% compared to ∼95% 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, O76082), 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) (#609423), 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 an AUC of slightly over 0.75. MedSim, on the other hand, reached an AUC value of up to 0.90 at 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

automatically annotating diseases with GO terms from known disease genes or proteins, and from their mouse orthologs or interacting human proteins. We also explored the possibility of increasing prediction performance by augmenting the functional profiles with semantically similar terms and filtering out dissimilar terms. The results obtained with several extensive benchmark experiments show that MedSim is able to specifically associate diseases with known proteins. Furthermore, despite its simplicity, MedSim achieves high AUC (up to 0.90) and sensitivity (up to 73%) values and is able to perform at least as well as more complex state-of-the-art methods like Endeavour (Aerts *et al*., 2006) and ToppGene (Chen *et al*., 2007). Moreover, we find that functional similarity can be used to distinguish diseases with a common functional basis from unrelated diseases, which enables further research on clustering diseases using functional criteria.

In detail, the functional similarity scores BPscore, rfunSim and rfunSimAll perform best for differing benchmark sets and annotation strategies. The transfer of GO annotations from 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-inter 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 Ramírez for providing the protein interaction data used in this study and to Alejandro Pironti 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 LSHG-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**, 773–774.

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**, e1000043.

Altshuler,D. *et al.* (2008) Genetic mapping in human disease. *Science*, **322**, 881–888.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Berglund,A.C. *et al.* (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.*, **36**, D263–D266.

Blake,J.A. *et al.* (2009) The Mouse Genome Database genotypes::phenotypes. *Nucleic Acids Res.*, **37**, D712–D719.

Chatr-Aryamontri,A. *et al.* (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.

Chen,J. *et al.* (2007) Improved human disease candidate 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.

Cho,J.H. (2008) The genetics and immunopathogenesis of inflammatory bowel disease. *Nat. Rev. Immunol.*, **8**, 458–466.

Cordell,H.J. and Clayton,D.G. (2005) Genetic association studies. *Lancet*, **366**, 1121–1131.

Feldman,I. *et al.* (2008) Network properties of genes harboring inherited disease mutations. *Proc. Natl Acad. Sci. USA*, **105**, 4323–4328.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.

Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, **18** (Suppl. 2), S110–S115.

Gibson,G. (2009) Decanalization and the origin of complex disease. *Nat. Rev. Genet.*, **10**, 134–140.

Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Hubbard,T.J.P. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.

Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res.*, **18**, 644–652.

Jimenez-Sanchez,G. *et al.* (2001) Human disease genes. *Nature*, **409**, 853–855.

Kann,M.G. (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief. Bioinform.*, **8**, 333–346.

Kann,M.G. (2010) Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief. Bioinform.*, **11**, 96–110.

Kelso,J. *et al.* (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.*, **13**, 1222–1230.

Kerrien,S. *et al.* (2007) IntAct-open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Lee,D.S. *et al.* (2008) The implications of human metabolic network topology for disease comorbidity. *Proc. Natl Acad. Sci. USA*, **105**, 9880–9885.

Lin,D. (1998) An information-theoretic definition of similarity. In: Shavlik, J.W. (ed.) *Proceedings of the 15th International Conference on Machine Learning (ICML-98). Madison, WI, USA*. Morgan Kaufmann Publishers, San Francisco, CA, USA, pp. 296–304.

Lowe,H.J. and Barnett,G.O. (1994) Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA*, **271**, 1103–1108.

Navlakha,S. and Kingsford,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 medicines: treatment strategies for hereditary disorders. *Nat. Rev. Genet.*, **7**, 261–276.

Ortutay,C. and 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**, 622–628.

Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, **71**, 1–11.

Ozgür,A. *et al.* (2008) Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, **24**, i277–i285.

Perez-Iratxeta,C. *et al.* (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.

Perez-Iratxeta,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.S.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.

Ruepp,A. *et al.* (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, **36**, D646–D650.

Salwinski,L. *et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Schlicker,A. *et al.* (2007) GOTax: investigating biological processes and biochemical activities along the taxonomic tree. *Genome Biol.*, **8**, R33.

Schlicker,A. and Albrecht,M. (2010) FunSimMat update: new features for exploring functional similarity. *Nucleic Acids Res.*, **38**, D244–D248.

Schreiber,S. *et al.* (2005) Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nat. Rev. Genet.*, **6**, 376–388.

Shriner,D. *et al.* (2008) Commonality of functional annotation: a method for prioritization of candidate genes from genome-wide linkage studies. *Nucleic Acids Res.*, **36**, e26.

Smith,C.L. *et al.* (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.*, **6**, R7.

Teare,M.D. and Barrett,J.H. (2005) Genetic linkage studies. *Lancet*, **366**, 1036–1044.

Tiffin,N. *et al.* (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.*, **33**, 1544–1552.

Tranchevent,L.C. *et al.* (2010) A guide to web tools to prioritize candidate genes. *Brief. Bioinform.*, in press.

Turner,F.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**, D169–D174.

van Driel,M.A. *et al.* (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, **14**, 535–542.

van Driel,M.A. and Brunner,H.G. (2006) Bioinformatics methods for identifying candidate disease genes. *Hum. Genomics*, **2**, 429–432.

Velankar,S. *et al.* (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.

Wu,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**, i119–i125.