Predicting Protein-Disease Relationships Using Sequence, Physicochemical Properties, and Molecular Function Information

Predrag Radivojac[1*], Kang Peng[1], Wyatt T. Clark[1], Brandon J. Peters[2],
Amrita Mohan[1], Sean M. Boyle[1], and Sean D. Mooney[2]
1) School of Informatics, Indiana University, Bloomington, Indiana 47408, USA
2) Center for Computational Biology and Bioinformatics, Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA
*To whom correspondence should be addressed: predrag@indiana.edu

1. INTRODUCTION

One of the most important tasks of modern bioinformatics is the development of computational tools that can be used to understand and treat human disease. To date, a variety of methods have been explored and algorithms for predicting whether a protein is involved in disease are gaining in their utility. Here, we describe an algorithm for detecting protein-disease associations based on the human protein-protein interaction network, known gene-disease associations, protein sequence, and protein functional information at the molecular level. Our method, PhenoPred (www.phenopred.org), is supervised: first, we map each protein onto the spaces of disease and functional terms based on distance to all annotated proteins in the protein interaction network. We also encode sequence, function, physicochemical, and predicted structural properties, such as secondary structure and flexibility. We then train support vector machines to detect a protein's disease function for a number of terms in Disease Ontology (DO). We provided evidence that, despite the noise/incompleteness of experimental data and unfinished ontology of diseases, identification of candidate genes and proteins can be successful even when a large number of candidate disease terms are predicted on simultaneously.

Predicting protein-disease associations has been previously considered by a group of techniques predominantly based on statistical principles (for all references see Ref. 1). For example, Freudenberg and Propping clustered a number of diseases from OMIM based on phenotypic data such as age at onset, tissue, inheritance, and then scored each gene-disease relationship proportional to the shared Gene Ontology (GO) annotation between a query gene and disease clusters associated with given disease. Another approach, POCUS, calculates the probability that different loci share the observed functional annotation by chance. TOM uses gene co-expression and GO annotation to find genes at particular loci that are likely to co-express or share functional annotation with the seed genes. Several other groups have analyzed protein-protein interaction (PPI) networks and proposed Bayesian approaches or various heuristics to gene prioritization. Prediction of disease associations has also been carried out in a broader context where various data sources are integrated together. In one of the earliest approaches, Perez-Iratxeta et al. calculated gene-disease associations by linking phenotype to protein function. RefSeq genes were first connected to GO terms and protein function was then connected to pathological condition through a Medline article search. Franke et al. developed Prioritizer, a Bayesian method, which utilizes functional annotation, microarray data, and predicted experimental protein-protein interactions. George et al. developed Gentrepid, a method based on PPI data and domain sharing, while Aerts et al. developed Endeavour, also based on statistical principles. Finally, Lussier et al. connect genomic and clinical data, whereas Butte and Kohane extend the concept of identifying disease associated genes from microarray data by considering a number of environmental and phenotypic factors. They use statistical principles to associate genes with Unified Medical Language System (UMLS) concepts, in effect creating a phenome-genome network.

Here, we present our novel approach to the prediction of protein-disease associations based on an experimental PPI network, known protein-disease associations, as well as protein sequence and functional annotation. We propose a method to associate genes or proteins to various levels of disease classification by considering DO information (http://diseaseontology.sourceforge.net) which organizes disease terms into a hierarchical structure expanding from the "disease" term to the most specific disease names in a top-down manner. Similarly to GO, DO is represented as a directed acyclic graph and is based on UMLS and International Classification of Diseases (ICD-9). The hierarchical organization of DO is beneficial for gene-disease prioritization algorithms in that it aggregates various levels of disease annotation into more general nodes thus enabling statistical inference with higher confidence.

2. METHODS

For each protein *p*, we constructed three sets of features for predicting disease associations: (i) PPI-DO features were constructed based on the distribution of shortest distances from *p* to other proteins in the PPI network known to be associated with specific disease terms; (ii) PPI-GO features were constructed in a similar way, but based on the shortest distances to other proteins known to be associated with specific GO terms; (iii) SPP-GO features encode various sequence, physicochemical, and other predicted properties of the protein as well as its GO terms.

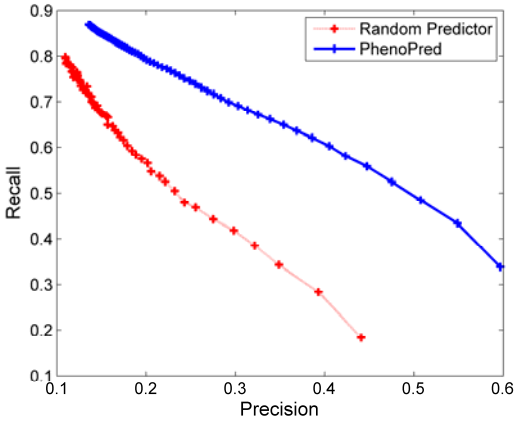

Figure 1. *Recall as a function of precision for PhenoPred (solid blue line) against the uniformly random predictor (dotted red line).*

To construct PPI-DO (and equivalently PPI-GO) features, we first computed the shortest distances between all pairs of proteins in the PPI network. For each combination of (*p*, *d*), where *d* is a disease term, we find the distribution of shortest distances from protein *p* to all proteins known to be associated with *d*, or simply the distribution of distances to disease *d*. In addition, we encoded fractions of proteins associated with disease *d* amongst *p*'s level-t neighbors (*t* = 1, 2…). Our assumption is that a protein *p* associated with disease *d* is more likely to share the distribution of distances to the DO terms with the proteins associated with *d* than the remaining proteins. The sequence-based and functional features (SPP-GO) were constructed based on (i) the real-valued vector data that is obtained for each physicochemical or predicted property and (ii) binary encoding of the known GO annotation and PROSITE matches. The real-valued data representation of a protein can be easily obtained by predicting its properties, e.g. secondary structure or intrinsic disorder, which effectively map an amino acid sequence into a real-valued vector (signal) of the same length. If we consider *s* to be such a property signal corresponding to protein *p*, then a set of features was generated based on the following: (i) the length of *s*; (ii) the mean and standard deviation of *s*; (iii) percentage of *s* that is above *n*-th percentile of the range of *s* (for various *n*) and (iv) the number of times each signal crosses these thresholds. We used the following properties: predictions of helix, sheet, coil, accessible surface area (ASA) and relative ASA as predicted by PHD, hydrophobic moment, flexibility predictions and predictions of intrinsically disordered protein regions. In addition, we calculated amino acid composition of each protein, as well as the number, orientation, and separation between predicted transmembrane helices by TMHMM. Physicochemical properties included aromatic content and charge. Finally, the GO and PROSITE information was encoded using a binary representation. The rationale for the use of property signals is that certain classes of disease-associated proteins have strong biases in their physicochemical properties. For example, it has been shown that cancer-related proteins and proteins involved in cardiovascular disease are significantly enriched in intrinsic disorder. All types of information are incorporated through a supervised framework using two layers of support vector machines. The first layer was built for each individual type of encoding (PPI-DO, PPI-DO, SPP-GO), while the second layer is simply a weighted average of the outputs of the first layer.

3. RESULTS

The predictor accuracy was evaluated using cross-validation on 422 DO terms for which 10 or more associated proteins were available. We evaluated all three individual models described above and a combined model that integrates the individual models (PhenoPred). Precision-recall curve for PhenoPred is shown in Figure 1, and it was shown that it achieved higher accuracy than any of the individual models, indicating usefulness of all lines of experimental evidence.[1] The average area under the ROC curve (AUC) over the 422 DO terms was 73.1%, however, it includes terms on which prediction results were nearly random (42 terms had AUC below 60%). In addition, the predictor was manually evaluated on several disease terms, of which we will discuss some candidate proteins that will be interesting future targets.

4. REFERENCES

1. Radivojac P. et al. An integrated approach to inferring gene-disease associations in humans. *Proteins* (2008) 72(3): 1030-1037. (http://www.ncbi.nlm.nih.gov/pubmed/18300252)