

## DISCOVERY OF PROTEIN INTERACTION NETWORKS SHARED BY DISEASES<sup>#</sup>

LEE SAM<sup>§,1</sup>, YANG LIU<sup>§,1</sup>, JIANRONG LI<sup>§,1</sup>, CAROL FRIEDMAN<sup>\*,2</sup>, YVES A. LUSSIER<sup>\*,1,2</sup>

*1- Center for Biomedical Informatics and Section of Genetic Medicine,  
Dept. of Medicine; The University of Chicago, IL, 60637*

*2- Department of Biomedical Informatics; Columbia University, NY, NY 10032*

The study of protein-protein interactions is essential to define the molecular networks that contribute to maintain homeostasis of an organism's body functions. Disruptions in protein interaction networks have been shown to result in diseases in both humans and animals. Monogenic diseases disrupting biochemical pathways such as hereditary coagulopathies (e.g. hemophilia), provided a deep insight in the biochemical pathways of acquired coagulopathies of complex diseases. Indeed, a variety of complex liver diseases can lead to decreased synthesis of the same set of coagulation factors as in hemophilia. Similarly, more complex diseases such as different cancers have been shown to result from malfunctions of common proteins pathways. In order to discover, in high throughput, the molecular underpinnings of poorly characterized diseases, we present a statistical method to identify shared protein interaction network(s) between diseases. Integrating (i) a protein interaction network with (ii) disease to protein relationships derived from mining Gene Ontology annotations and the biomedical literature with natural language understanding (PhenoGO), we identified protein-protein interactions that were associated with pairs of diseases and calculated the statistical significance of the occurrence of interactions in the protein interaction knowledgebase. Significant correlations between diseases and shared protein networks were identified and evaluated in this study, demonstrating the high precision of the approach and correct non-trivial predictions, signifying the potential for discovery. In conclusion, we demonstrate that the associations between diseases are directly correlated to their underlying protein-protein interaction networks, possibly providing insight into the underlying molecular mechanisms of phenotypes and biological processes disrupted in related diseases.

### 1. Introduction and Related Work

Currently, common diseases are mainly defined by their clinical appearance, with little reference to their molecular mechanism. For example, syndromes are defined in medicine as a set of phenotypes which, occurring together, serve to define a trait or disease. These phenotypes overlap in the case of many syndromes. This overlap brought about the concept of 'syndrome families' though consideration of the commonality of features shared between diseases [1]. Conceptually, what we have learned about 2000 human single gene diseases with a defined genetic phenotype is that each monogenic disease has a specified collection of specific phenotypic features. For example, hemophilias with

---

\* Corresponding authors

§ These authors have contributed equally to the work

# This study was supported in part by NIH/NLM grants 1K22 LM008308-01, R01 LM007659, R01 LM008635, and the National Center for the Multi-scale Analysis of Genomic and Cellular Networks (U54CA121852-01A1)

deficiencies in coagulation factors, otherwise called hereditary coagulopathies, are single gene diseases with clear Mendelian inheritance that have provided significant insight in the biochemical pathways of acquired coagulopathies. Indeed, a variety of complex liver diseases can lead to decreased synthesis of the same set of coagulation factors as in hemophilia, leading to the same disease phenotype despite very different causes. In some cases, the clustering of syndromes into these families in combination with genetic insights has led to the discovery that what were often thought as two different disorders were really variable expressions of the same disorder [2-4]. Conversely, it has long been known that mutations at different loci can lead to the same genetic disease [5]. It has also been hypothesized that this genetic heterogeneity has its roots at the protein interaction level, suggesting that other genes associated with the phenotype also have some functional role [6]. Therefore, it is plausible to theorize that phenotypic overlap of diseases may reflect, at multiple biological scales, the relationships and functional properties of shared underlying molecular networks. As signal transduction pathways are less understood than biochemical pathways, protein-protein interactions networks provide unique opportunities for exploring the signaling pathways of diseases.

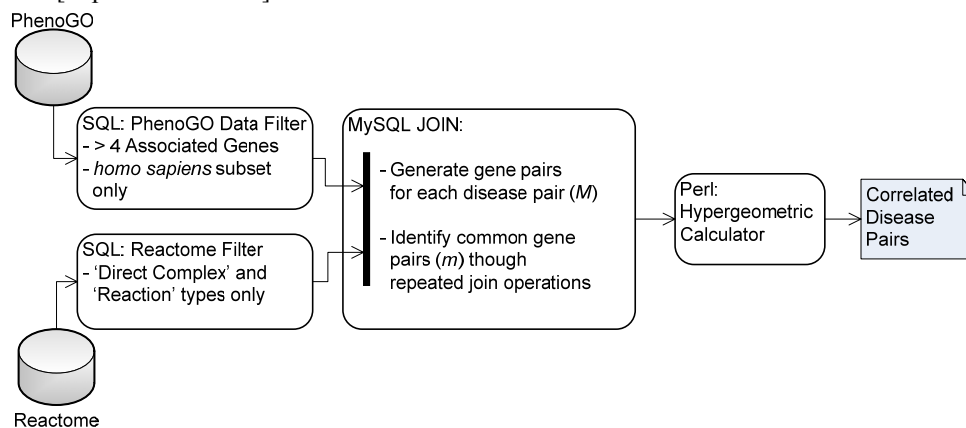
The shift in focus to systems biology has resulted in an increased interest in biological pathways and protein-protein interaction networks. As a result, large scale knowledge bases representing them are being rapidly developed [7-14]. These resources enable us to study complex biological problems using high throughput computational tools. While there is a wealth of protein-disease relationships in the published literature and a number of readily computable protein-protein interaction resources, there has been a paucity of work relating diseases using protein interactions from this kind of knowledge. Making use of these networks is a relatively new challenge in the field. Network-based analyses have been developed with a number of goals in mind [15], including protein function prediction [16], identification of functional modules [17], interaction prediction [18-21], and the study of network structure and evolution [22-26].

To explore the possibility of using protein-protein interaction networks to identify correlations between diseases, we hypothesize that protein-protein interactions shared by two diseases or more can be accurately identified in a protein interaction network by integrating knowledge from the literature and using statistical methods.

### *1.1 Related Work*

The method reported in this paper utilizes the PhenoGO database [25][[www.phenoGO.org](http://www.phenoGO.org)] that provides protein-GO-phenotype relations and the human-curated Reactome knowledgebase [8] that provides protein interactions to link protein-protein interactions with diseases. The recently developed PhenoGO database provides phenotypic context to protein-GO annotations, as an example, lymphoid tissue (a phenotype) is linked to interleukin 2 receptor (a protein) and interleukin 2 receptor

activity (a GO concept). It augments Gene Ontology (GO) annotations [27] by extracting protein-GO-phenotype relations from the literature using MeSH terms [28] and a natural language processing (NLP) system, BioMedLEE, combined with the PhenOS phenotype ontology organizer system. The phenotypic information, including diseases, tissues, and organs, is encoded into Unified Medical Language System (UMLS) codes as well as other ontological coding systems. PhenoGO was evaluated for anatomical and cellular context in mice, demonstrating a recall of 92% and a precision of 91% [29]. PhenoGO has since been extended to comprise over 523,000 unique entries associating disease phenotypes, ontological concepts, and proteins. In total, PhenoGO now contains data from 8,509 distinct PubMed articles, representing 7,016 distinct proteins classified under 3,214 distinct GO concepts in 3,102 distinct diseases. From a random sample of 120 Protein-disease-GO ternary annotations, precision was estimated at 85%, and recall at 76% [unpublished result].



**Figure 1. Method to correlate human diseases based on their underlying protein interactions.  $M$  and  $m$  refer to parameters of the hypergeometric calculations as described in equation 1.**

## 2. Methods

In order to identify associations between diseases by mapping their respective protein interaction networks with statistical significance values, we took the following steps. An overview of the process is pictured in **Figure 1**.

**Extraction of human protein-disease relationships** was achieved through Structured Query Language querying of the PhenoGO database. We extracted all UMLS-coded diseases classified under the “Disease” semantic type hierarchy along with their associated proteins. In this study, we chose to stay on a more conservative side, and only extracted diseases associated with more than 4 proteins to avoid errors stemming from mis-assignment in PhenoGO and to reduce spurious predictions in the next step from the

hypergeometric distribution because a single error contributes proportionally to a larger statistical impact on a smaller sample of protein in the statistical method that follows (**equation 1**). These UMLS-coded terms fall under the UMLS semantic types ‘Congenital Abnormality’, ‘Disease or Syndrome’, ‘Experimental Model of Disease’, ‘Anatomical Abnormality’, and ‘Neoplastic Process’. The resultant set consists of 154 diseases and their 1,931 associated proteins (<http://phenos.bsd.uchicago.edu/PSB2007/>).

**Integration and Discovery.** The second step is to correlate diseases with their underlying protein-protein interaction networks using a statistical approach. In this study, we used the Reactome protein interaction dataset [8] to define the underlying topological networks associated with these diseases. The common proteins between disease-associated proteins in PhenoGO and proteins in the Reactome were identified by using the identifiers in the UniProt [30]. The Reactome data set defines four distinct types of reactions: 1) neighboring reactions, which define interactions that occur consecutively; 2) indirect complexes, which define interactions which involve subcomplex interaction, but not direct binding/interaction; 3) direct complex, defining protein-protein complexes; and 4) reaction, representing situations where the two proteins participate in the same reaction [8]. The Reactome dataset was normalized to a set of paired Swiss-Prot accession numbers, and filtered to remove neighboring reactions and indirect complexes, leaving only entries for binary interactions and direct complexes. This data set contains 20,317 distinct interactions corresponding to 1,140 distinct proteins. From the 154 diseases, we generated combinations of pairs of diseases, and for each pair of diseases, proteins in both diseases were also paired for all potential combinations. These protein pairs were then cross-referenced with our filtered Reactome data set to determine if they participated in reactions or formed direct complexes with one another. There are two basic types of relationships used in calculations in our methods. These relationships correspond to the two scenarios we considered to determine whether two diseases share interaction networks: 1) an identity relationship where common proteins are shared by two diseases, and 2) direct interactions between protein A in one disease and protein B in the other disease. As related diseases can share both types of relations, and due to the requirements of the hypergeometric distribution, we consider both in the underlying protein-protein interaction network in diseases. Based on this, we calculated the correlations between all possible pairs of diseases by applying the hypergeometric distribution function to identify significantly correlated diseases (**equation 1**) and adjustments for multiple *a posteriori* comparisons (**equation 2**), as shown below:

$$p(i \geq m | N, M, n, m) = \sum_{i=m}^n \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}} \quad \text{(Equation 1)}$$

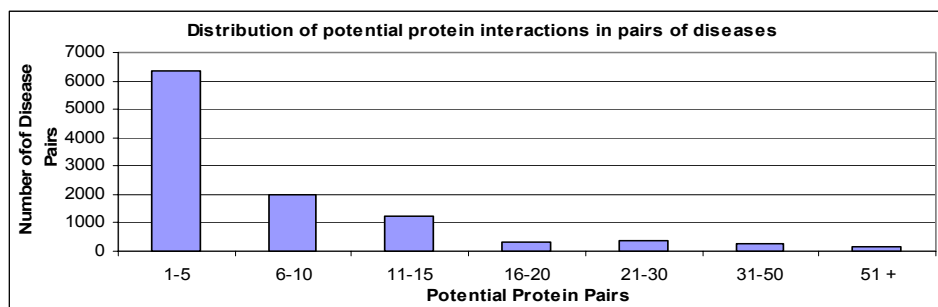
In **equation 1**, ‘ $N$ ’ represents the total number of all pair combinations between proteins of any two diseases in the experiment that includes the possibility of sharing the

same proteins (identical protein pair between two diseases), ‘ $M$ ’, as the sum of number of observed distinct pairs of interacting proteins that exist in the Reactome database for all the diseases in the experiment (direct interaction only), ‘ $n$ ’ as the putative total number of pairs of proteins that could exist in a pair of disease, and ‘ $m$ ’ as the sum of the observed number of common proteins shared between two specific diseases and the number of distinct pairs of interacting proteins observed in the Reactome database for these two specific diseases ( $M \cap n$ ). This measure gives a p-value which is then adjusted for multiple comparisons with the Dunn-Sidak method (a derivative of the Bonferroni method):

$$p' = 1 - (1 - p)^r \quad \text{(Equation 2)}$$

In **equation 2**,  $p'$  and  $p$  represent the corrected and uncorrected p-values, respectively, and  $r$  represents the number of independent comparisons, which is the number of disease pairs ( $r=11,703$ ) used in the study. These corrected p-values are then thresholded at  $p < 0.05$  to determine the final set of significantly correlated disease-disease relationships. Multiple diseases and genes sharing the same PubMed IDs can artificially boost the statistical significance of these disease pairs, therefore relationships mapping to more than 2 overlapping PubMed IDs were removed to reduce this artifact. A total of 11,703 disease pairs passed the filter out of 11,780 candidates. 77 combinations had more than two PMID overlaps and were filtered out as a result of this process. An example of values used for the calculation is described in the results section.

**Evaluations.** Two evaluations were conducted. The first one, a *quantitative evaluation*, was designed to control for the error rate in either assigning a protein disease relationship in PhenoGO or a protein-protein interaction in Reactome. It consisted of establishing the reliability of the predictions if we introduced noise in the integrated database network (10% more protein-protein interactions in the same set of diseases). The second one, a *qualitative evaluation*, consisted of carefully examining the discovered protein interactions shared by two diseases and identifying references in the scientific literature that validate the phenotypic overlap and potentially the protein interactions.



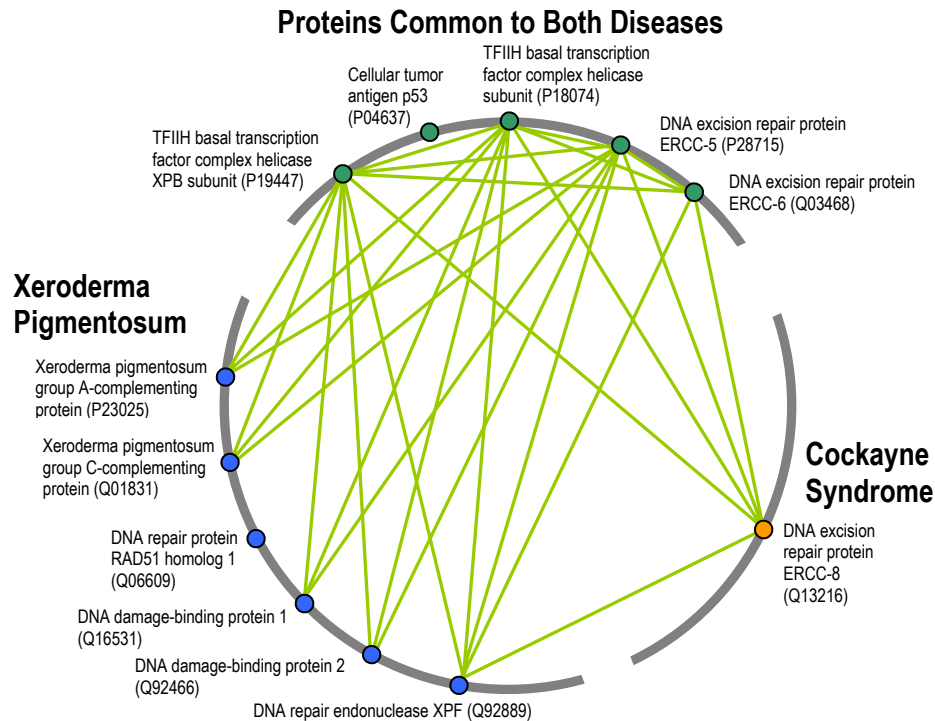
**Figure 2.** Distribution of the number of disease pairs from PhenoGO according to the number of possible protein interactions observed between their proteins in the Reactome.

### 3. Results

In this study, we examined a subset of PhenoGO pertaining to human diseases in order to identify relationships between these diseases according to criteria described in the methods. This filtering resulted in a set of 154 diseases and their 1,931 associated proteins. The intersection between the proteins of the Reactome and those of PhenoGO further reduced the set of proteins to 286. The number of candidate proteins per disease was greatly reduced by the need to be present in the Reactome dataset, and therefore the totals are smaller than observed in the PhenoGO database alone. We lose approximately 70% of the proteins in this process due to the limited content of the Reactome. In order to identify relationships between these diseases, we analyzed their underlying protein-protein interaction maps by applying a statistical method (details of equations in the Method Section). Of the 154 selected diseases, there are  $(285 \cdot 286 / 2 + 286) = 41,041$  distinct combinations of protein pairs and identical protein overlap (term  $N$ , **equation 1**) possible for all possible disease pairs, of which only 4,857 exist in the Reactome (term  $M$ , **equation 1**). **Figure 2** summarizes the distribution of protein-protein pairs per combination of diseases in our set. In ~60% of the 11,703 disease pairs under consideration, the number of potential protein-protein interactions is five or less (no significant predictions from this category), and about 40% of them have more than five interactions. We then proceeded in calculating the correlation between groups of pairs of interacting proteins associated with every pair of diseases according to **equations 1 and 2** (file available at <http://phenos.bsd.uchicago.edu/PSB2007/>). Based on the correlations of the shared protein interacting pairs between diseases, we identified 10 pairs of diseases that are significantly correlated due to their shared proteins and protein-protein interactions out of 11,703 disease pairs examined in this study (**Table 1**).

**Quantitative Evaluation.** We added 2031 “false positive” interactions between random nodes in the network to evaluate the robustness of the method to 10% noise in the network. We found that even with the introduced noise, none of the p-values in the top 10 entries changed. We also attempted adding 10% noise (46 “false” interactions) in just the 286 proteins under study, which changed the p-values of the top 10 entries, but left their rank order relatively intact (results available at <http://phenos.bsd.uchicago.edu/PSB2007/>).

**Qualitative evaluation.** The top ranked disease pairs are shown in **Table 1**, all of which have a significant adjusted pvalue less than 5%. The last column of **Table 1** provides strong scientific evidence in support of the predictions. We have manually examined all the significant disease pairs, and confirmed their correlations in the literature, demonstrating our method can successfully predict non trivial correlations between different diseases. Among these pairs of diseases, Cockayne Syndrome (**CS**) and Xeroderma Pigmentosum (**XP**) provide a very interesting example on how two diseases are correlated through their protein-protein interaction networks. Xeroderma Pigmentosum is a disorder conferring susceptibility of the skin to ultraviolet radiation,



**Figure 3: Protein interactions between Xeroderma Pigmentosum and Cockayne Syndrome.** Shared proteins between the two diseases are represented (top), proteins in the Cockayne syndrome (right), and those in the Xeroderma Pigmentosum (left). Protein interactions between the two diseases are linked by lines.

**Table 1. Top ranked significantly correlated diseases.**

UMLS ID	Disease 1	UMLS ID	Disease 2	P-PI (#)	pvalue	Corrected pvalue	Ref
C0009207	Cockayne Syndrome	C0043346	Xeroderma Pigmentosum	38	7.3e-22	8.5e-18	[31]
C0043346	Xeroderma Pigmentosum	C0085390	Li-Fraumeni Syndrome	24	6.7e-11	4.9e-06	[32]
C0007001	Carbohydrate Metabolism, Inborn Errors	C0002514	Amino Acid Metabolism, Inborn Errors	9	8.3e-10	6.2e-05	*
C0009404	Colorectal Neoplasms	C0950123	Genetic Diseases, Inborn	16	6.7e-10	5.0e-05	[33]
C0085390	Li-Fraumeni Syndrome	C0009207	Cockayne Syndrome	16	2.7e-09	1.9e-04	NA
C0009404	Colorectal Neoplasms	C0015625	Fanconi's Anemia	8	1.5e-05	6.7e-01	[9]
C0009404	Colorectal Neoplasms	C0085413	Polycystic Kidney, Autosomal Dominant	8	1.5e-05	6.7e-01	[21]
C0024141	Lupus Erythematosus, Systemic (LES)	C0004364	Autoimmune Diseases	4	9.3e-05	9.9e-01	*
C0024314	Lymphoproliferative Disorders (LD)	C0004364	Autoimmune Diseases	6	1.3e-04	9.9e-01	[34]
C0024314	LD	C0024141	LES	6	1.3e-04	9.9e-01	[35]

\* self-evident relations between disease pairs. Ref = references confirming the predictions, P-PI = Protein –Protein Interaction #

due to deficiencies in one of the XPA-XPG complementation group genes involved in nucleotide excision repair [36]. Similarly, Cockayne Syndrome involves deficiencies in transcription-coupled repair genes ERCC6 and ERCC8 leading to a number of conditions including abnormal sensitivity to sunlight. As shown in **Figure 3**, there are 27 direct protein-protein interactions and 5 common proteins (term  $m = 27+5$ , **equation 1**) that are shared by these two diseases. A total of 66 potential combinations of protein-protein interaction pairs (term  $n$ , **equation 1**) can be formed between the 11 proteins of XP and the 6 proteins of CS.

As shown in the **Figure 3** and described in **Table 2**, we find that most proteins in the common networks between the two diseases are related to DNA repair processes, which are Global Genomic Nucleotide Excision Repair (NER) and Transcription-coupled NER. The Global Genomic NER repairs lesions from non-transcribed regions of genome, a process independent to transcription, and the Transcription-coupled NER repairs UV-induced damage in the transcribed strands of active genes. Both Cockayne syndrome and Xeroderma Pigmentosum are associated with these processes, suggesting defects in the repair of DNA damage are the cause of the diseases, as indicated in the literature [36]. Our computational approach allows us to quickly identify the shared networks between these two diseases, demonstrating the method we used is able to identify the underlying molecular basis shared by these diseases.

In some cases, disruptions in any of the proteins or genes lying on a pathway can lead to a disease phenotype. This is the case with both Xeroderma Pigmentosum and Cockayne syndrome. At a higher classification level, these two previous diseases are a result of deficiencies in the DNA repair pathway, a class also shared with Li-Fraumeni Syndrome [37]. Though these three single gene diseases have a known initial molecular cause, how this cause is related to DNA repair pathways and whether the diseases share the same pathway or related disjoint pathways may be poorly understood.

In another example, Fanconi's Anemia (FA) is a hereditary DNA-repair deficiency characterized by hypersensitivity to DNA damaging agents. This disorder is caused by a mutation in any one of genes in the Fanconi's Anemia complementation group: FANCA, FANCB, FANCC, FANCD1, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, or FANCM [38-40]. Its phenotype is complex and includes anemia, several congenital malformations, and a *strong predisposition to cancers* [38, 39]. Kutler et al. (2003) analyzed clinical data from 754 FA patients from North America enrolled in the International Fanconi Anemia Registry, of whom 173 (23%) had a total of 199 neoplasms (28 distinct types of cancers) [9]. Among 14 potential protein interactions between Fanconi's Anemia and Colorectal Neoplasms, 8 were found to exist in the Reactome.

An evaluation of the relationship between the generality of a disease class (based on graph-theoretic distance from the "MeSH Descriptor" node in the UMLS) and the number of proteins annotated to it found no correlation (available at <http://phenos.bsd.uchicago.edu/PSB2007/>).

#### 4. Discussion

The protein-protein interaction network constructed by the Reactome dataset provides us a framework for structuring the knowledge of human diseases, which enables an objective approach to examine the molecular underpinnings of diseases in the context of their known molecular interactions on genomic scale. This method not only allows us to conduct high throughput computational analysis of the relations between diseases, but also reveals the underlying molecular relationships between diseases. Furthermore, new relationships between well-known diseases and new diseases could be revealed based on their overlapping molecular networks.

Although many diseases have been associated with their genetic and proteomic underpinnings, little research has been focused on bridging the gap between protein interactions and the relationships between diseases. Phenotype clustering methods achieve this to some extent. For example, Brunner and van Driel used a text mining approach based on MeSH terms as keywords over the OMIM database [6] to cluster similar disease phenotypes. Our implementation of the hypergeometric distribution significantly differs from its common use in bioinformatics. Other authors have used this distribution in large scale gene expression studies to identify “over-represented” gene classes (e.g. Gene Ontology classes) and find systemic patterns [41]. This classical implementation would be efficient in recognizing overlapping proteins or proteins sharing annotated pathways in GO, but would not recognize novel protein interaction based on newly discovered or predicted protein interactions. In contrast, we focused on protein interactions and thus counted the protein pairs rather than the genes’ assignments to categories. The proposed analytical approach could scale up in two ways. First, we could extend it to proteins interacting indirectly through a pathway rather than directly interacting in the Reactome (through additional join operations in the database in order to determine those interacting with one or more intermediate proteins in pathways). In doing this, the Bonferroni-type adjustment would have to be replaced with a data-derived control for multiple comparisons such as bootstrap or permutation resampling in order to interpret the results. A second, probably more useful way in which this analysis can scale up is its use with the rapidly expanding number of protein-interaction databases, many of which are not publicly available. The subset of the PhenoGO database used in this study can readily be reused in a similar manner over another protein interaction database containing more genes and provide other specialized predictions.

*Limitations.* One question about the use of this technique is its reliability when conditions change. Since we used well established statistics and one of the most severe multiple comparison criteria for controlling for false predictions, we believe this method is robust. As this technique relies on integrating accurate protein-protein interactions with accurate gene-disease associations, and both of these datasets likely contained at least 10% false positive relationships, we conducted an evaluation adding false relationships in

the network and confirmed that the identified disease pairs sharing protein networks were reliable in spite of the noise. Nonetheless, this approach remains limited by the quality of the underlying protein networks, and the accuracy of protein-disease mapping. Currently, the protein-protein interaction network is still at the early developmental stage. In this study, we extracted 1,931 proteins from 154 diseases, of which only 288 proteins exist in the Reactome dataset that contains 1,140 proteins. Therefore, the interaction network we used to correlate relationship between diseases is relatively small. Certainly, as bioinformatics databases become larger and more accurate this discovery method could become a valuable tool to identify relationships between diseases.

*Future studies.* We intend to explore a permutation based resampling in order to unveil additional valid relationships. A resampling-based approach would help determine the optimal relationship between quantity and quality in the dataset. We also plan to significantly extend the protein-disease associations by mining additional genetic datasets. Besides using the Reactome, we also demonstrated we could use DIP [7], although it is smaller than Reactome [results not shown]. Since the UMLS is used to encode the diseases, we plan to compare related diseases and their associated protein-protein interactions in order to establish the molecular basis of disease relationships in ontologies.

## 5. Conclusion

We developed and evaluated an automatic system to predict protein interactions shared by two or more diseases. It augments current protein interaction networks by integrating literature-based knowledge of protein-disease associations and systematically identifying the statistically significant Protein Interactions of Diseases (PID). Results demonstrated that the PID system provides accurate predictions and is scalable in a number of dimensions: (i) it enables high throughput predictions, and (ii) it scales across different protein-interaction datasets. Beyond direct protein-protein interactions, it also provides the theoretical framework to compare shared pathways between diseases. In the future, this framework could be applied to more complex diseases to determine if their shared phenotypes are a result of the shared molecular mechanism and pathways.

## References

1. Pinsky, L., *The polythetic (phenotypic community) system of classifying human malformation syndromes*. Birth Defects Orig Artic Ser, 1977. **13**(3A): p. 13-30.
2. Bertola, D.R., C.A. Kim, L.M. Albano, H. Scheffer, R. Meijer H. van Bokhoven, *Molecular evidence that AEC syndrome and Rapp-Hodgkin syndrome are variable expression of a single genetic disorder*. Clin Genet, 2004. **66**(1): p. 79-80.

3. Sorasio, L.G.B. Ferrero E. Garelli G. Brunello C. Martano A. Carando, *et al.*, *Eur J Med Genet*, 2006.
4. Zenteno, J.C., C. Venegas S. Kofman-Alfaro, *Evidence that AEC syndrome and Bowen--Armstrong syndrome are variable expressions of the same disease*. *Pediatr Dermatol*, 1999. **16**(2): p. 103-7.
5. Morton, N.E., *Am J Hum Genet*, 1956. **8**(2): p. 80-96.
6. Brunner, H.G.M.A. van Driel, *Nat Rev Genet*, 2004. **5**(7): p. 545-51.
7. Xenarios, I., D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte D. Eisenberg, *Nucleic Acids Res*, 2000. **28**(1): p. 289-91.
8. Joshi-Tope, G.M. Gillespie I. Vastrik P. D'Eustachio E. Schmidt B. de Bono, *et al.*, *Reactome: a knowledgebase of biological pathways*. *Nucleic Acids Res*, 2005. **33**(Database issue): p. D428-32.
9. Kutler, D.I.B. Singh J. Satagopan S.D. Batish M. Berwick P.F. Giampietro, *et al.*, *A 20-year perspective on the International Fanconi Anemia Registry (IFAR)*. *Blood*, 2003. **101**(4): p. 1249-56.
10. Ogata, H., S. Goto, W. Fujibuchi M. Kanehisa, *Computation with the KEGG pathway database*. *Biosystems*, 1998. **47**(1-2): p. 119-28.
11. Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono M. Kanehisa, *KEGG: Kyoto Encyclopedia of Genes and Genomes*. *Nucleic Acids Res*, 1999. **27**(1): p. 29-34.
12. Bader, G.D., I. Donaldson, C. Wolting, B.F. Ouellette, T. PawsonC.W. Hogue, *Nucleic Acids Res*, 2001. **29**(1): p. 242-5.
13. Zanzoni, A., L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-CitterichG. Cesareni, *FEBS Lett*, 2002. **513**(1): p. 135-40.
14. Breitkreutz, B.J., C. StarkM. Tyers, *Genome Biol*, 2002. **3**(12): p. PREPRINT0013.
15. Sharan, R.T. Ideker, *Modeling cellular machinery through biological network comparison*. *Nat Biotechnol*, 2006. **24**(4): p. 427-33.
16. Nabieva, E., K. Jim, A. Agarwal, B. ChazelleM. Singh, *Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps*. *Bioinformatics*, 2005. **21 Suppl 1**: p. i302-10.
17. Lubovac, Z., J. GamalielssonB. Olsson, *Combining functional and topological properties to identify core modules in protein interaction networks*. *Proteins*, 2006.
18. Rhodes, D.R.S.A. TomlinsS. VaramballyV. MahavisnoT. BarretteS. Kalyana-Sundaram, *et al.*, *Nat Biotechnol*, 2005. **23**(8): p. 951-9.
19. Jansen, R.H. YuD. GreenbaumY. KlugerN.J. KroganS. Chung, *et al.*, *A Bayesian networks approach for predicting protein-protein interactions from genomic data*. *Science*, 2003. **302**(5644): p. 449-53.
20. Lee, I., S.V. Date, A.T. AdaiE.M. Marcotte, *A probabilistic functional network of yeast genes*. *Science*, 2004. **306**(5701): p. 1555-8.
21. Wong, S.L.L.V. ZhangA.H. TongZ. LiD.S. GoldbergO.D. King, *et al.*, *Proc Natl Acad Sci U S A*, 2004. **101**(44): p. 15682-7.
22. Berg, J., M. LassigA. Wagner, *Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications*. *BMC Evol Biol*, 2004. **4**(1): p. 51.

23. Rzhetsky, A.S.M. Gomez, *Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome*. Bioinformatics, 2001. **17**(10): p. 988-96.
24. Barabasi, A.L.R. Albert, *Emergence of scaling in random networks*. Science, 1999. **286**(5439): p. 509-12.
25. Wagner, A.D.A. Fell, *The small world inside large metabolic networks*. Proc Biol Sci, 2001. **268**(1478): p. 1803-10.
26. Eisenberg, E.E.Y. Levanon, *Preferential attachment in the protein network evolution*. Phys Rev Lett, 2003. **91**(13): p. 138701.
27. Ashburner, M.C.A. BallJ.A. BlakeD. BotsteinH. ButlerJ.M. Cherry, *et al.*, *The Gene Ontology Consortium*. Nat Genet, 2000. **25**(1): p. 25-9.
28. National Library of Medicine. *Medical Subject Headings (MeSH®) Fact Sheet*. 27 May 2005 <http://www.nlm.nih.gov/pubs/factsheets/mesh.html>.
29. Lussier, Y., D. Rappaport, T. BorlawskyC. Friedman. *PhenoGO: a Multistrategy Language Processing System Assigning Phenotypic Context to Gene Ontology Annotations*. in *Pacific Symposium on Biocomputing*. 2006. Maui, HI, USA.
30. Bairoch, A.R. ApweilerC.H. WuW.C. BarkerB. BoeckmannS. Ferro, *et al.*, *Nucleic Acids Res*, 2005. **33**(Database issue): p. D154-9.
31. Online Mendelian Inheritance in Man, *MIM Number: Excision-Repair, Complementing Defective, In Chinese Hamster, 5; ERCC5 (#133530); last edited 11/29/2005: <http://www.ncbi.nlm.nih.gov/omim/>*
32. Hoogervorst, E.M.C.T. van OostromR.B. BeemsJ. van BenthemS. GielisJ.P. Vermeulen, *et al.*, *Cancer Res*, 2004. **64**(15): p. 5118-26.
33. Online Mendelian Inheritance in Man, *MIM Number: Colorectal Cancer (#114500); last edited 5/17/2006: <http://www.ncbi.nlm.nih.gov/omim/>*
34. Worth, A., A.J. Thrasher H.B. Gaspar, *Br J Haematol*, 2006. **133**(2): p. 124-40.
35. Blanco, R., B. McLaren, B. Davis, P. SteeleR. Smith, *Systemic lupus erythematosus-associated lymphoproliferative disorder: report of a case and discussion in light of the literature*. Hum Pathol, 1997. **28**(8): p. 980-5.
36. Spivak, G., *The many faces of Cockayne syndrome*. Proc Natl Acad Sci U S A, 2004. **101**(43): p. 15273-4.
37. Hanawalt, P.C., *Subpathways of nucleotide excision repair and their regulation*. Oncogene, 2002. **21**(58): p. 8949-56.
38. Cotran, R.S., V. KumarT. Collins, *Robbins Pathologic Basis of Diseases*. 1999: p. 169,296.
39. Online Mendelian Inheritance in Man, *MIM Number: Fanconi Anemia (#227650); last edited 3/15/2006: <http://www.ncbi.nlm.nih.gov/omim/>*
40. Joenje, H.K.J. Patel, *The emerging genetic and molecular basis of Fanconi anaemia*. Nat Rev Genet, 2001. **2**(6): p. 446-57.
41. Martin, D., C. Brun, E. Remy, P. Mouren, D. ThieffryB. Jacq, *GOToolBox: functional analysis of gene datasets based on Gene Ontology*. Genome Biol, 2004. **5**(12): p. R101.