Chapter 6

Predicting Disease-Related Proteins Using Human Disease Network

Based on

Hossein Rahmani, Hendrik Blockeel and Andreas Bender, "Predicting Disease-Related Proteins using Informative Human Disease Network" submitted to IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB).

6.1 abstract

Identification of novel proteins likely involved in diseases is an important issue in the area of computational biology. Protein-Protein Interaction (PPI) networks have been widely used for the task of predicting proteins involved in diseases. Previous methods assume to have a set of proteins which are previously known to be involved in disease (i.e., seed proteins) and then, they try to extend the seed proteins by predicting new disease-related proteins. While the initial seed proteins of each disease is incomplete and suffers from 'False Negative' cases, dependency of previous methods to the incomplete seed proteins is the main drawback of these methods. In this paper, we reduce the number of False Negative cases in the initial seed proteins of 20 analyzed diseases by proposing an informative Human Disease Network (HDN) considering both functional and structural information in the PPI network. After building a biologically meaningful HDN, we cluster the HDN nodes based on the connectivity and then, we augment the seed proteins of each disease based on the cluster it belongs to. Finally, we predict new disease-related proteins based on augmented seed proteins. Literature mining of newly predicted proteins proved the usefulness of the proposed HDN.

6.2 Introduction

In recent years, much effort has been invested in the construction of protein-protein interaction (PPI) networks [118]. Much can be learned from the analysis of such networks with respect to the metabolic and signalling processes present in an organism, and the knowledge gained can also be prospectively employed e.g., for the task of protein function prediction [78, 98, 18], identification of functional modules [71], interaction prediction [48, 129], identification of disease candidate genes [27, 109, 26, 58, 106, 37, 87, 130, 132] and drug targets [104, 81], according to an analysis of the resulting network [72].

Wu et al. [130] present an excellent overview of multiple methods for detecting proteins involved in disease or cancer. Among the different methods discussed in [130], "guilt-by-proximity" methods are well known. Methods classified in this category are based on the assumption that genes that directly interact, or, more generally, lie close to each other in the network, are more likely to be involved in the same diseases (as argued by, e.g., Gandhi et al. [31]). The methods vary based on how they define proximity: Some methods consider only direct neighbors to be in the proximity (e.g., [87, 3]), some quantify proximity of two proteins using the length of the shortest-path between them, some compute a "Global Distance Measure" that also takes into account how many paths there are between the two proteins, and how long these are; an example is the approach by Chen et al. [16], who use a PageRank based model for this.

The methods discussed by Wu et al. [130] mostly rely on notions of proximity (to genes known to be disease-related) from the area of graph analysis. An entirely

different type of approaches are those that rely on feature-based descriptions [132, 77, 29, 66]. There, each individual protein is described by means of a fixed set of features. Next, using machine learning methods, a model is learned that links some of these features to disease-relatedness.

In almost all the discussed methods, prediction accuracy depends directly on the initial disease-related proteins, which we refer to as seed proteins. While the initial seed proteins of each disease suffers from several 'False Negative' cases (i.e., disease-related proteins which are not annotated as being involved in disease), dependency of previous methods to the incomplete seed proteins is the main drawback of these methods. In this paper, first, we propose an informative Human Disease Network (HDN) considering both functional and structural information in the PPI network. Second, we cluster the HDN nodes based on connectivity. Third, we augment the seed proteins of each disease based on the cluster it belongs to. Fourth, we predict new disease-related proteins based on augmented seed proteins. Finally, we analyze the literature to prove the usefulness of the proposed HDN.

6.3 Methods

6.3.1 Formal Definition

We consider a PPI network as an undirected annotated graph $(P, E, \lambda_F, \lambda_D)$ where P is a set of proteins, $E \subseteq P \times P$ is a set of interactions between these proteins, and λ_F and λ_D are so-called annotation functions; for each p, λ_F and λ_D denote the additional information we have about p. In this work, we assume that $\lambda_F(p)$ simply lists all the GO functions that are associated with p; we call it the function set (or function vector) of p, and denote it $FS(p)$. $\lambda_D(p)$ lists all the diseases that protein p is involved in; we call it the disease list of p and denote it $dizList(p)$. If $D = \{diz_1, diz_2, \ldots, diz_m\}$ is the list of m analyzed diseases in our paper, then $dizList_i(p) = 1$ if p is involved in diz_i and 0 otherwise. We also define seed proteins $SP(diz_i)$ as the set of proteins involved in disease diz_i ($diz_i \in dizList(p) \Leftrightarrow p \in SP(diz_i)$).

6.3.2 Human Disease Network

We consider a Human Disease Network (HDN) as a directed graph $HDN(D, R)$ where D is a set of diseases and $R \subseteq D \times D$ is a set of directed relationships between these diseases. We build our proposed HDN as follows: For each disease $d_i \in D$:

1. We build $testSet$ as a union of the seed proteins of each disease $d_k \in D$ where $k \neq i$.
   $testSet = \bigcup_{d_k \in D \text{ and } k \neq i} SP(d_k)$

2. We consider the remaining proteins in P as $trainSet$. We assume the seed proteins $SP(d_i)$ as positive cases and the remaining proteins in the $trainSet$ as negative cases.

3. We choose a prediction method M, we train M with $trainSet$ and then, we use method M to calculate the prediction-value $PV(p)$ for each protein $p \in testSet$. M will return high PV values for more relevant disease-related proteins.
4. We repeat step 3 , 10 times and we calculate the average prediction-value of each protein $p \in testSet$ (APV(p)).
5. For each disease $d_j \in D(j \neq i)$, we add a directed edge $d_i \rightarrow d_j$ in HDN based on Formula 6.1.

$$weight(d_i \rightarrow d_j) = \frac{\sum_{\forall p \in SP(d_j)} APV(p)}{|SP(d_j)|} \tag{6.1}$$

In Formula 6.1, the $||$ operator returns the number of seed proteins of disease d_j.

The resulting HDN is the directed fully-connected network in which each node is a disease and each weighted edge shows a relationship between two diseases. In order to focus on the most important relationships in HDN, we prune the network by keeping only the highest-ranked edges.

Although our proposed approach for building HDN is very general and any prediction method M could be used in step 3 of building HDN, the quality of the resulted HDN still depends on the prediction method M. We next discuss some recommended prediction methods.

6.3.3 Recommended Prediction Methods

In this section, we discuss about three categories of methods used for predicting proteins involved in diseases. *Structural* methods predict proteins involved in diseases based on the topological location of the proteins in the PPI network while *functional* methods use the functional annotation of the proteins for the prediction. *Hybrid* methods take both structural and functional information into account.

Structural Category: Random Walk based Method (*ST-RW*)

Berger et al. [6] assume that disease-related proteins fall closer on average to the seed proteins than they do on average to the rest of the network. They calculate the score of each protein p_j in the network based on Formula 6.2 and then, select high-scoring proteins as disease-related proteins.

$$score_s(p_j) = \frac{\frac{\sum_{i \in C'} T_{ij}}{|C'|} - \frac{\sum_{i \in C} T_{ij}}{|C|}}{\frac{\sum_i T_{ij}}{|C|+|C'|}} \tag{6.2}$$

In Formula 6.2, T_{ij} is the average number of steps a random walker takes to walk from a specified node i to another specified node j, C is the set of seed proteins and C' is the set of all other proteins in the network. In the rest of this paper, we refer to this method as *ST-RW*.

Structural Category: ANOVA based Method (*ST-Anova*)

Rahmani et al. [98] proposed a relevance measure for proteins that is inspired by statistical ANOVA (analysis of variance), and showed that shortest-path distance to a relatively small number of proteins (selected according to the ANOVA-based measure) is informative for the task of function prediction in the PPI network. Since the ANOVA method works well for function prediction, it is natural to check whether it also gives good results for the task of predicting disease-related proteins. We therefore propose the use of similar features for predicting proteins involved in disease.

The ANOVA-inspired selection measure (briefly, ANOVA) is defined as follows. Let P^+ be the set of proteins labeled as being involved in disease *diz*, and P^- the set of proteins not labeled as such. For each protein q, we introduce a feature d_q; $d_q(p)$ denotes the shortest-path distance between p and q (viewed here as a feature of p). We consider for each q the mean and variance of $d_q(p)$, taken over all *diz*-related (m_q^+ and var_q^+ respectively) and non-*diz*-related p (m_q^- and var_q^- respectively).

$$m_q^+ = \frac{\sum_{p \in P^+} d_q(p)}{|P^+|} \tag{6.3}$$

$$m_q^- = \frac{\sum_{p \in P^-} d_q(p)}{|P^-|} \tag{6.4}$$

$$var_q^+ = \frac{\sum_{p \in P^+} (d_q(p) - m_q^+)^2}{|P^+| - 1} \tag{6.5}$$

$$var_q^- = \frac{\sum_{p \in P^-} (d_q(p) - m_q^-)^2}{|P^-| - 1} \tag{6.6}$$

Seeing P^+ and P^- as two groups of proteins, the following formula compares the variance between groups to the variance within groups (as it is used for relative ranking only, constant factors are dropped):

$$A_q = \frac{(m_q^+ - m_q^-)^2}{var_q^+ + var_q^-} \tag{6.7}$$

A high A_q means that d_q varies little within groups and/or much between groups, which indicates that d_q has high predictive power for the group. Features d_q can be ranked according to A_q, and the top-k features selected as actual features to be included in the description of all proteins. In the end, we apply the naive Bayes classifier to the proteins descriptions for predicting *diz*-related proteins. In the rest of this paper, we refer to this method as *ST-Anova*.

Functional Category: Individual based Method (*Func-Indiv*)

In this method, first, we use a χ^2-based feature selection method to select the most relevant individual functions. Let D and $\bar{D}$ be the set of proteins that are disease-related (D) or not ($\bar{D}$), and let, for each function f_i, P_i be the set of proteins annotated

Table 6.1: List of the 4 different hybrid methods considering structural and functional information in the network.

Structural Method	Functional Method	Hybrid Method
ST-RW	Func-Indiv	RW-Indiv
ST-RW	Func-Collab	RW-Collab
ST-Anova	Func-Indiv	Anova-Indiv
ST-Anova	Func-Collab	Anova-collab

with f_i and $\bar{P}_i$ the set of proteins not annotated with it. With $a = |D \cap P_i|$, $b = |D \cap \bar{P}_i|$, $c = |\bar{D} \cap P_i|$ and $d = |\bar{D} \cap \bar{P}_i|$, we have

$$\chi^2(f_i) = \frac{(ad - bc)^2 * (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \tag{6.8}$$

We calculate the chi-square of each individual function f_i in the network. Then, we describe each protein p_j in the network based on the high-scored individual functions. In the end, we apply the naive Bayes classifier for predicting disease-related proteins. In the rest of this paper, we refer to this method as *Func-Indiv*.

Functional Category: Collaboration based Method (*Func-Collab*)

Selecting individual discriminative functions based on $\chi^2(f_i)$ does not consider the network topology and the way different functions interact with each other in the network. Rahmani et al. [100] showed that for the task of predicting cancer-related proteins, it is possible that a function f_i does not correlate itself with cancer-involvement, but interaction of the same function with some function f_j does correlate with the former protein being involved in a cancer. Rahmani et al. [100] proposed a new way of calculating the χ^2 of the function pairs in the PPI network. They select high-ranked collaborative function pairs and then, they describe the proteins based on the high-ranked function pairs. In the end, they applied the naive Bayes classifier for predicting the proteins involved in cancer. In the rest of this paper, we refer to this method as *Func-Collab*.

Hybrid Category: Integrating Functional and Structural Information

Structural-based and functional-based methods can be combined into hybrid methods as shown in Table 6.1. The hybrid method is calculated as follows:

$$score_h(p) = norm(score_s(p)) + norm(score_f(p)) \tag{6.9}$$

In Formula 6.9, $score_s(p)$ and $score_f(p)$ show disease-relatedness score of p using *Structural* (*ST-RW* and *ST-Anova*) and *Functional* (*Func-Indiv* and *Func-Collab*) methods, respectively. In order to avoid a bias toward either of these categories, we use Formula 6.10 to normalize the disease-relatedness scores. In Formula 6.10,

$min(x)$ and $max(x)$ return minimum and maximum values taken over all values of x, respectively.

$$norm(x_i) = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{6.10}$$

6.4 Empirical Results

6.4.1 Dataset

We applied our method for building HDN to the PPI network used by Milenkovic et al. [77]. This dataset is the union of three human PPI datasets: HPRD [91], BIOGRID [116] and the dataset used by Radivojac et al. [97] and contains 47,303 physical interactions among 10,282 proteins. When we say "union", we mean that the new network contains all the nodes and edges (proteins and interactions) found in either of these networks. The aim of merging these three datasets was to obtain as complete a human PPI network as possible, i.e., a network that covers with its edges as many proteins in the human proteome as possible. Milenkovic et al. [77] provide details on the construction of the integrated network.

Table 6.2 shows the list of 20 different diseases analyzed in this paper in addition to the number of proteins involved in each disease (seed count).

6.4.2 Comparing Recommended Prediction Methods

In this section, we use the following leave-one-out cross validation to compare the different prediction methods discussed in section 6.3.3:

For each disease $d_i \in D$:

1. We select 99 proteins randomly from the PPI network ($randSet$).
2. For each seed proteins $p_i \in SP(d_i)$
   (a) We build the $trainSet$ by excluding the $\{p_i \cup randSet\}$.
   (b) We apply different prediction methods M to rank p_i relative to the 99 randomly selected proteins ($rank(p_i)$). M should return small rank values for more relevant disease-related proteins.
3. We repeat steps 1 to 2b, 10 times and we calculate the average rank of each $p_i \in SP(d_i)$ over different iterations ($avg(rank(p_i))$).

Figure 6.1 compares the discussed prediction methods for the 20 different diseases shown in Table 6.2 with respect to overall rank of seed proteins among 99 randomly selected proteins (Formula 6.11). *RW-Indiv* achieves the best overall performance, compared to the other methods, and is therefore a good candidate method for building HDN.

Table 6.2: List of the 20 different diseases analyzed in this paper.

Disease ID	Disease Name	Seed Count
D-1	Alzheimer	7
D-2	Amyotrophic	4
D-3	Anemia	36
D-4	Breast Cancer	21
D-5	Cataract	14
D-6	Charcot-marie-tooth	11
D-7	Colorectal-cancer	20
D-8	Deafness	28
D-9	Diabets	23
D-10	Dystonia	5
D-11	Ehlers-danlos	7
D-12	Emolytic-anemia	11
D-13	Epilepsy	11
D-14	Long QT Syndrome	13
D-15	Lymphoma	27
D-16	Mental-retardation	19
D-17	Parkinson	8
D-18	Usher-syndrome	5
D-19	Xeroderma	10
D-20	Zellweger	8

$$overallRank(d_i) = \frac{\sum_{p_i \in SP(d_i)} avg(rank(p_i))}{|SP(d_i)|} \tag{6.11}$$

For each discussed method M, Table 6.3 shows the set of diseases for which M produces the best result. It is clear that *Func-Indiv* and *RW-Indiv* are overall the best performing methods.

Figure 6.2 compares the three best methods, *ST-RW*, *Func-Indiv* and *RW-Indiv*, to each other for each disease. The figure shows that, for those diseases where *Func-Indiv* scores best (e.g., D-6, D-11 and D-19), it is only slightly better than the second-best method, whereas in those cases where it is not best, the difference with the best method can be large (e.g., D-3, D-7, D-15, D-16). *RW-Indiv*, on the other hand, never differs much with the best method, making it it the most stable method for predicting disease-related proteins in PPI networks.

Figure 6.3 compares all methods on six different diseases where neither *Func-Indiv* nor *RW-Indiv* achieves the best overall performance. Although *RW-Indiv* is not the best method for any of these, Figure 6.3 shows that on average it ranks second, with a very small difference compared to the best method.

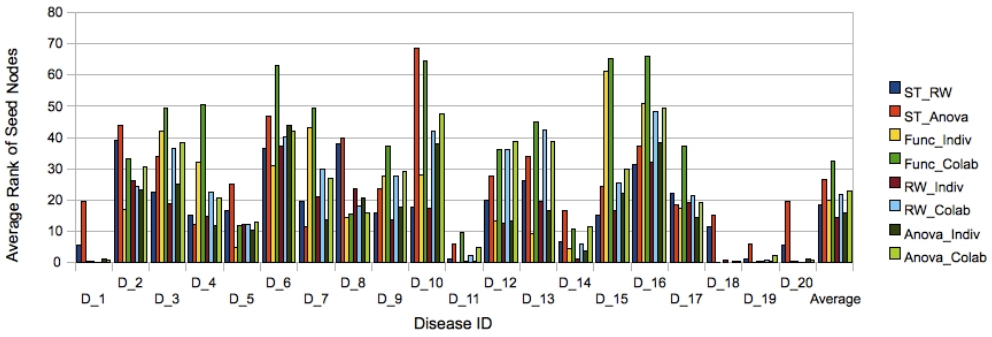

Figure 6.1: Average rank of seed proteins in 20 different diseases shown in Table 6.2. *RW-Indiv* achieves the best overall performance comparing to the other methods and is therefore a good candidate method for building HDN.

Table 6.3: Set of diseases in which each method produces the best result.

Method M	Set of diseases which M produces the best results	Count
ST-RW	D15, D16	2
ST-Anova	D7	1
Func-Indiv	D2, D5, D6, D8, D11, D13, D19	7
Func-Collab	D18	1
RW-Indiv	D1, D3, D9, D10, D12, D14, D20	7
RW-Collab	–	0
Anova-Indiv	D4, D17	2
Anova-Collab	–	0

6.4.3 Informative Human Disease Network

We choose the *RW-Indiv* prediction method to build our proposed HDN for 20 different diseases shown in Table 6.2. There are 380(20×19) possible edges in the original HDN. We prune HDN by sorting the edges based on their weight descendingly and then, keeping the 38 (10% of original HDN) highest-weighted edges. Figure 6.4 shows the pruned HDN. For each edge $(d_i) \xrightarrow{rank} (d_j)$, Figure 6.4 shows the rank of the relationship between two diseases d_i and d_j among all the 380 disease pairs. The highest-ranking found relationship is $(deafness) \xrightarrow{1} (usher\ syndrome)$. Analyzing the literature, we found biological evidence for most of the relationships shown in Figure 6.4.

Goh et al. [38] propose a simple method for building undirected Human Disease Network. They connect two diseases d_i and d_j in the network if there is at least one gene that implicated in both. We applied the Goh et al. [38] to our disease dataset and

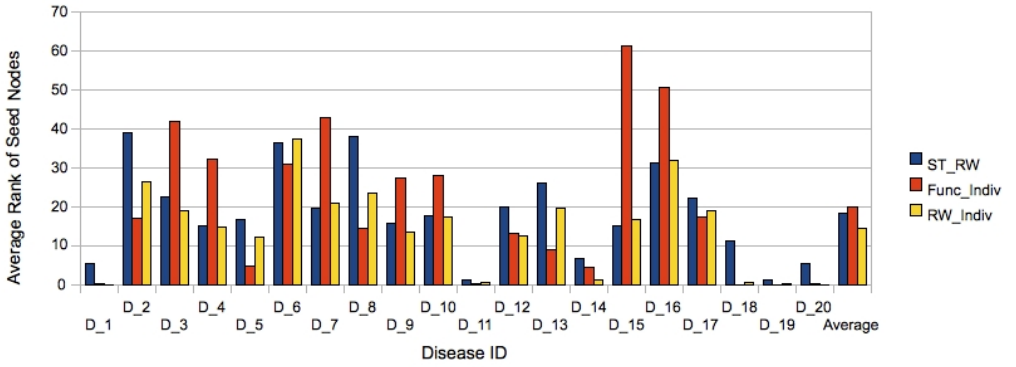

Figure 6.2: Comparing *ST-RW*, *Func-Indiv* and *RW-Indiv* methods with each other with respect to average rank of seed proteins in 20 different diseases shown in table 6.2.

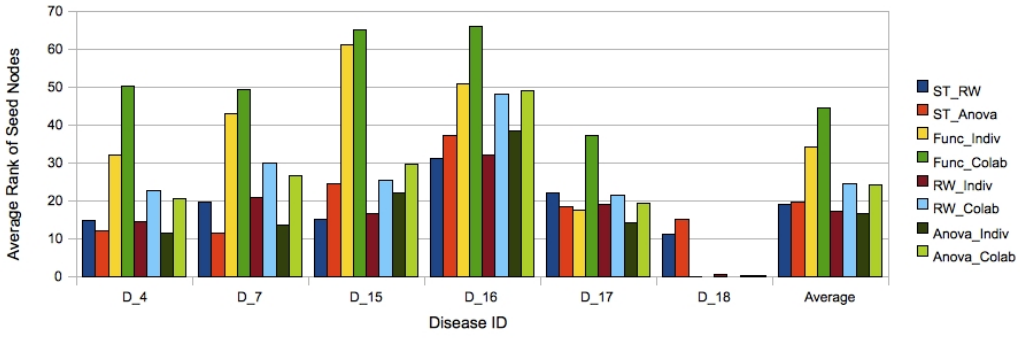

Figure 6.3: Comparing different prediction methods in 6 different diseases in which neither *Func-Indiv* nor *RW-Indiv* achieves the best overall performance. According to the average rank column , *RW-Indiv* is the second best method for these diseases with a very small difference with the best method.

the resulted HDN is shown on Figure 6.5. For each edge $(d_i) \leftrightarrow (d_j)$, Figure 6.5 shows the number of proteins involved in both diseases d_i and d_j ($|SP(d_i) \cap SP(d_j)|$). The best found relationship is $(anemia) \leftrightarrow (emolytic\ anemia)$. Comparing our proposed HDN (Figure 6.4) with the disease network discussed by Goh et al. [38] (Figure 6.5), we observe that our HDN is more informative than the network proposed by Goh et al. [38].

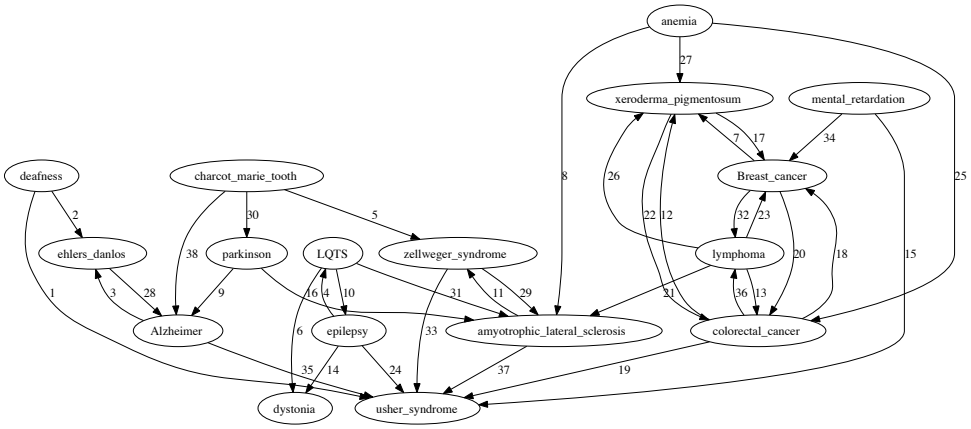

Figure 6.4: Pruned Human Disease Network by keeping only 38 (10% of original HDN) high-ranked relationships among different diseases. The best found relationship is $(deafness) \xrightarrow{1} (usher\ syndrome)$.

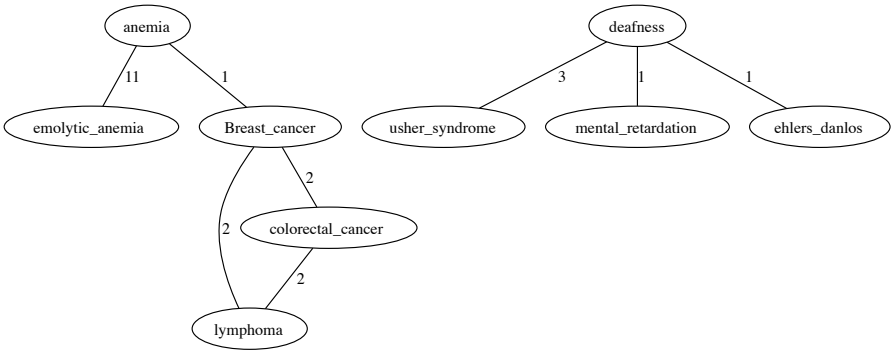

Figure 6.5: Human Disease Network based on the common proteins (Proposed by Goh et al. [38]). Edge's weight shows the number of common proteins between two related diseases.

6.4.4 Biological Interpretation of the Pruned HDN

We will now briefly discuss the biological significance of the observed findings. The highest ranked connection (1.) between deafness and Usher's Syndrome is appar-

ent, given the latter is an inherited form of deafness. The link between Deafness and Ehlers-Danlos syndrome however may be attributed also to misdiagnosis of joint laxity, given that the combination of this observation with deafness is more likely to be correctly classified as Stickler syndrome [65]. Epilepsy and Dystonia are both characterized by seizures, and given the proximity of both terms in the Figure also a mechanistic connection between both disorders can be elucidated. Interesting is the relationship of Long QT Syndrome (LQTS) to both Dystonia and Epilepsy, which hints at the importance of ion channels being important in all of those cases. On the other hand, Anyotrophic Lateral Sclerosis (ALS) is more related to Parkinson's disease (but neither epilepsy nor dystonia), hinting at fundamentally different mechanisms behind those, on the surface similar, disorders characterized by seizures. Apart from this seizure-cluster, also various cancer variations are found to be closely related, namely Xeroderma Pigmentosum (leading to sensitivity to UV light), breast cancer, lymphomas and colorectal cancer. What is interesting is the close link of ALS with the cluster of cancers, since indeed it is assumed that ALS, as a motor neuron disease, may represent a particular case of paraneoplastic encephalomyelitis [122].

6.4.5 Predicting Disease-Related Proteins using the Pruned HDN

In the context of involvement in disease, one main drawback of previous methods is their dependency on a list of seed proteins which is likely incomplete. In this section, we use our proposed HDN for augmenting the seed proteins of different diseases as follows: First, we cluster the pruned HDN into n clusters $C_1 \dots C_n$ based on the network connectivity. Second, we augment the seed proteins of each disease member d_i of cluster C_j by unioning the seed proteins of all the disease members of cluster C_j ($d_i \in C_j \Rightarrow Aug(SP(d_i)) = \cup_{d_k \in C_j} SP(d_k)$). $Aug(SP(d_i))$ is the augmented list of seed proteins of disease d_i. Third, we use a hybrid prediction method *RW-Indiv* for predicting new proteins involved in the disease. Table 6.4, Table 6.5, Table 6.6 and Table 6.7 show the four clusters extracted from the pruned HDN shown in Figure 6.4 in addition to the 10 highest-ranked proteins predicted for each cluster. The first cluster, covering Alzheimer and Ehler-Danlos syndrome, covers both known and potential novel protein targets to treat those diseases. In case of Alzheimer's, BACE2, HSD17B10 and TM2D1 have been implicated in literature before, while COL5A3, which encodes one of the fibrillar collagens, has been established to be involved in Ehler-Danlos syndrome. On the other hand, genes (and proteins) not explicitly associated with those diseases are TGBF2, THBS1 and SPON1, all of which are known to be involved in cell-to-cell interactions, cell-to-matrix interactions, and cell adhesion, respectively. In particular SPON1 can readily be understood to be of importance, given its involvement of attachment of neuron cells and neurite outgrowth.

Similar results covering both established and novel genes are observed for the second cluster, with LQTS, Epilepsy and Dystonia. Dopamine levels and epilepsy have been linked for a long time (DRD1, DRD3 and DRD4; [117] Dystonia) and they are of practical relevance for treatment. The KCNQ4 ion channel on the other hand

Table 6.4: 10 highest-ranked proteins predicted for cluster 1 = {Alzheimer, ehler-danlos}.

Index	Protein Symbol	Full Protein Name
1	COL5A3	Collagen, type V, alpha 3
2	THBS1	Thrombospondin 1
3	TGFB2	Transforming growth factor, beta 2
4	COL5A2	Collagen, type V, alpha 2
5	PDGFA	Platelet-derived growth factor alpha polypeptide
6	SPON1	Spondin 1, extracellular matrix protein
7	HSD17B10	Hydroxysteroid (17-beta) dehydrogenase 10
8	HADH2	Hydroxysteroid (17-beta) dehydrogenase 10
9	BACE2	Beta-site APP-cleaving enzyme 2
10	TM2D1	TM2 domain containing 1

has been previously linked with Long QT Syndrom (LQTS). What is interesting, with potential practial implications, is the importance of ALG10 in this cluster, which gates rat ether-a-go-go (the human homolog of the hERG channel involved in LQTS) and which might hence also play an important role in human. No explicit involvement of the EPM2AIP1 gene, encoding laforin, has been described in literature yet; however, our analysis makes a rather strong disease implication for the three diseases present in this cluster.

The third cluster of neoplastic diseases, covering Xeroderma pigmentosum, breast cancer, lymphoma and colorectal cancer gives relatively little surprises, with agreement on MSH3 and MSH6 which are both involved in DNA repair, on the APC tumor suppressor protein, and the RELA oncogene (which bings to the NF kappa b transcription factor with known involvement in cancerogenesis).

The fourth and final disease cluster, of Zellweger syndrome, ALS, and Usher's Syndrome, involves the myosins MYO6, MYO3A and MYO15A which are all known to be involved either in hearing loss or, in the latter case, the actin organization in the hair cells of the cochlea. What is apparent is the link of this set of disorders to the peroxisome, which has been established for this disease cluster before (the involvement of PEX7 and PEX12 which are involved in the assembly of peroxisomes is characteristic, but also ABCD1 is involved in fatty acid transport into the peroxisome, and PXMP3 is involved in its biogenesis). The potentially most surprising gene located in this disease cluster is SIRT3, which is known to be involved in epigenetic silencing and which has been characterized as a potential antineoplastic target - given its prominent role in this analysis, it might hence also play a role for drug treatments of this set of diseases in the future.

Table 6.5: 10 highest-ranked proteins predicted for cluster 2 = {LQTS, Epilepsy, Dystonia}.

Index	Protein Symbol	Full protein Name
1	DRD4	Dopamine receptor D4
2	DRD3	Dopamine receptor D3
3	DRD1	Dopamine receptor D1
4	ALG10B	Asparagine-linked glycosylation 10, alpha-1,2-glucosyltransferase homolog B (yeast)
5	KCR1	A membrane Protein That Facilitates Functional Expression of Non-inactivating K+ Currents Associates with Rat EAG Voltage-dependent K+Channels
6	EPM2AIP1	EPM2A (laforin) interacting protein 1
7	KCNQ4	Potassium voltage-gated channel, KQT-like subfamily, member 4
8	TOR1B	Torsin family 1, member B (torsin B)
9	HSPC163	–
10	GCHFR	GTP cyclohydrolase I feedback regulator

Table 6.6: 10 highest-ranked proteins predicted for cluster 3 = {xeroderma-pigmentosum, breast-cancer-leon, lymphoma, colorectal-cancer}.

Index	Protein Symbol	Protein Full Name
1	MSH6	MutS homolog 6 (E. coli)
2	MSH3	MutS homolog 3 (E. coli)
3	APC	Adenomatous polyposis coli
4	RELA	V-rel reticuloendotheliosis viral oncogene homolog A (avian)
5	TGFBR1	Ransforming growth factor, beta receptor 1
6	PTK2B	PTK2B protein tyrosine kinase 2 beta
7	HIPK2	Homeodomain interacting protein kinase 2
8	RPS6KB1	Ribosomal protein S6 kinase, 70kDa, polypeptide 1
9	TGFB1	Transforming growth factor, beta 1
10	ERBB2	V-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)

6.4.6 Case Study: Long QT Syndrome

In this section, we examine Long QT Syndrome (LQTS) in more details. According to [120], LQTS is a disorder of the heart's electrical activity which can cause sudden, uncontrollable, dangerous arrhythmias in response to exercise or stress. Table 6.8

Table 6.7: 10 highest-ranked proteins predicted for cluster 4 = {zellweger-syndrome, amyotrophic-lateral-slecrosis, usher-syndrome}.

Index	Protein Symbol	Protein Full Name
1	MYO15A	Myosin XVA
2	MYO3A	Myosin IIIA
3	MYO6	Myosin VI
4	DDO	D-aspartate oxidase
5	PEX12	Peroxisomal biogenesis factor 12
6	PEX7	Peroxisomal biogenesis factor 7
7	PXMP3	Peroxisomal membrane protein 3
8	SIRT3	Sirtuin 3
9	AGXT	Alanine-glyoxylate aminotransferase
10	ABCD1	ATP-bindende cassette, sub-familie D (ALD), lid 1

Table 6.8: Proteins associated with the Long QT Syndrome. The data is taken from Berger et. al.[6].

Index	Protein sybmbol	Full Protein name
1	KCNQ1	Potassium voltage-gated channel, KQT-like subfamily, member 1
2	KCNH2	Potassium voltage-gated channel, subfamily H (eag-related), member 2
3	SCN5A	Sodium channel, voltage-gated, type V, alpha subunit
4	ANK2	Ankyrin 2, neurona l
5	KCNE1	Potassium voltage-gated channel, Isk-related family, member 1
6	KCNE2	Potassium voltage-gated channel, Isk-related family, member 2
7	KCNJ2	Potassium inwardly-rectifying channel, subfamily J, member 2
8	CACNA1C	Calcium channel, voltage-dependent, L type, alpha 1C subunit
9	CAV3	Caveolin 3
10	SCN4B	Sodium channel, voltage-gated, type IV, beta
11	AKAP9	A kinase (PRKA) anchor protein (yotiao) 9
12	SNTA1	Syntrophin, alpha 1
13	ALG10	Asparagine-linked glycosylation 10 homolog (yeast, alpha-1,2-glucosyltransferase)

shows the set of proteins involved in LQTS.

Table 6.9: 10 most discriminative functions according to $\chi^2(f_i)$ (Formula 6.8).

Index	Function	Short Description
1	GO:0008016	Regulation of heart contraction
2	GO:0060307	Regulation of ventricular cardiomyocyte membrane repolarization
3	GO:0060299	Regulation of heart contraction
4	GO:0002095	Caveolar macromolecular signaling complex
5	GO:0014819	Regulation of skeletal muscle contraction
6	GO:0031579	Membrane raft organization
7	GO:0033292	T-tubule organization
8	GO:0005251	Delayed rectifier potassium channel activity
9	GO:0005244	Voltage-gated ion channel activity
10	GO:0008015	Blood circulation

Most Relevant Features for Long QT Syndrome

The number of different functions occurring in our human dataset is 9833; this is also the dimensionality of the *Func-Indiv* method if no dimensionality reduction is used. As we discussed in section 6.3.3, we can use a χ^2-based feature selection methods to reduce this number; at the same time, this techniques rank functions according to how relevant they are for prediction of disease relatedness.

Table 6.9 shows the ten most discriminant individual functions obtained. It can be seen that the top three GO annotations are explicitly related to cardiac action potential (regulation of heart contraction, regulation of ventricular cardiomyocyte membrane repolarization and negative regulation of sarcomere organization). Positions 4 and 5 are concerning caveolar signaling (which is also very prominent in the heart) and regulation of skeletal muscle contraction, alluding to the fact that muscle contraction in the skeleton and in the heart is goverened by related processes. Membrane rafts (as well as caveolae) are important for cardiac ion channel function as has been found before, [73] which is also correctly identified in Table 6.9. T-tubule organization, while not immediately apparent, has been linked to a 'new paradigm' for human arrhythmias recently [2]. It is interesting that explicit potassium and ion channel activity are appearing only low in this list, along with the broad term of blood circulation. Hence, overall it can be said that the most discriminative functions are overall meaningful, with specific functions appearing at the top, biologically derived functions (raft organization, T-tubule organization) in the middle, and general terms at the bottom of the terms derived from the analysis.

Our dataset contains 10,282 proteins. The Anova based method uses the ANOVA measure to select the most relevant among these. More detailed information could be obtained from an ANOVA analysis of the most relevant proteins among the full set of 10,282 proteins. Table 6.10 now shows the ten proteins with the highest ANOVA measure obtained using our analysis. Interestingly, no ion channel has been most

Table 6.10: 10 most discriminative proteins according to Anova (Formula 6.7).

Index	Protein	Short Description
1	NDUFS6	NADH dehydrogenase [ubiquinone] iron-sulfur protein 6, mitochondrial
2	KCNH1	Potassium voltage-gated channel subfamily H member 1
3	KCNH5	Potassium voltage-gated channel, subfamily H (eag-related), member 5
4	KCNF1	Potassium voltage-gated channel subfamily F member 1
5	AKAP6	A-kinase anchor protein 6
6	ALG10B	Asparagine-linked glycosylation 10, alpha-1,2-glucosyltransferase homolog B
7	KCR1	A membrane Protein That Facilitates Functional Expression of Non-inactivating K+ Currents Associates with Rat EAG Voltage-dependent K+Channels
8	KCNE1	Potassium voltage-gated channel subfamily E member 1
9	KCNH2	potassium voltage-gated channel, subfamily H (eag-related), member 2
10	KCNE2	Potassium voltage-gated channel subfamily E member 2

signficiant, but the NADH dehydrogenase NDUFS6. It has been found that HDUSF6 knockouts cause mitochondrial complex I deficieny [54], causing various cardiac problems such as reduced systolic function and cardiac output. On the one hand, this might relate to a functional relationship between diseases; on the other hand it might indicate imperfect diagnosis, hence confusing different underlying disease biology. The six Potassium channels listed can be understood to be involved in direct polarization and depolarization of the cardiac action potential; however the three remaining proteins, namely AKAP6, ALG10B and KCR1 deserve particular attention here. AKAP6 (also called mAKAP) anchors Protein Kinase A to RYR2 which is able to generate Ca2+ 'sparks' due to simultaneous activation within a certain neighborhood radius [124], and hence importance to the cardiac action potential and deviations thereof. ALG10B (also known as KCR1) is interestingly thought to be able to reduce KCNH2 sensitivity to proarrhythmic drug blockade which may be due to glycosylation of this potasssium channel [60, 92], hence our method was able to not only identify protein directly involved in causing LQTS, but also modifer proteins such as AKAP6 and ALG10B.

Predicting Disease-Related Proteins using Individual method *RW-Indiv*

The following steps were performed for predicting new LQTS-related proteins:

1. A new *trainSet* was built containing all the proteins annotated as being involved in LQTS (positive set) in addition to 100 randomly selected proteins (negative set).

2. A *testSet* was built containing all the remaining proteins in the network.

3. The *RW-indiv* method was used to rank the proteins in the *testset*.

Table 6.11 lists the highest ranked newly identified LQTS-related genes. In agreement with expections, many of the genes identified are (as hERG itself) voltage-gated Potassium channels; however also Sodium channels (SNC4A), Calcium channels (CACNB3 and CACNA1A) and solute carriers (SLC8A1) appear in the list. This is in agreement with the known proteins involved in the regulation of cardiac action potential, which are known to involve all three types of ions. KCNJ8 seems to be involved in cardiovascular sudden death at least in mouse models [51], indicating that while focusing on LQTS is of high practical relevance in today's drug development environment, one can in turn also assume that other ion channels involved in drug adverse reactions are currently not receiving sufficient attention. SLC8A1, as a sodium/calcium exchanger, is known to be involved in regulating action potential as well [1], though it is not easy to find a specific link to the QT interval prolongation in this case. SCN4A mutations have bene found to be insignificant under standard conditions, but become relevant in patients treated with LQ-inducing drugs [89]. This finding is interesting since it appears also synergistic adverse relations between genes and LQTS syndrome can be identified using our network approach. One of the potassium channels newly identified to be involved in cardiac action potential regulation (and, hence, with potential LQTS liability) is KCJN12 [49], which is indeed thought to be invovled in providing the cardiac inward rectifier current (IK1). A similar observation can be made regarding KCNA1, where it is thought that a brain-driven cardiac dysfunction can be made responsible for sudden death syndrome in epilespy patients [35]. Mutations in CACNA1 are classified as 'LQTS8' and, while rare, have been shown to be linked to LQTS [75]. Hence, overall we can find associations between the genes identified here and LQTS in many cases - and, interestingly, often they are dependent on the particular genetic or drug treatment conditions of the patients (such as in case of SCN4A and KCNA1).

6.5 Compare individual and network based prediction for LQTS

Considerable differences are apparent from the proteins included in the cluster including LQTS along with Epilespy and Dystonia (Table 6.5), and the prediction of LQTS-related proteins (Table 6.11). The receptors identified in Table 6.5 are on the one hand G-Protein Coupled Receptors (GPCRs) such as the Dopamine D1, D3 and D4 receptor subtypes identified with the highest rank in the disease cluster. The only ion channel selected is KCNQ4, which has been linked to deafness [21]; however, only related potassium channels appear to have been linked to LQTS until this stage. On the other hand, KCR1 (ALG10B), which is though to modulate sensititvity to drugs causing LQTS, also appears in this list (as well as in Table 6.10, in the list of most significant proteins according to ANOVA-based selection). On the other hand, Table

Table 6.11: Newly identified LQTS-related proteins by applying *RW-Indiv* method to the original seed proteins.

index	Gene-Name	Short Description
1	KCNH1	Potassium voltage-gated channel, subfamily H (eag-related), member 1
2	KCNH5	Potassium voltage-gated channel, subfamily H (eag-related), member 5
3	KCNJ8	Potassium inwardly-rectifying channel, subfamily J, member 8
4	SLC8A1	Solute carrier family 8 (sodium/calcium exchanger), member 1
5	SCN4A	Sodium channel, voltage-gated, type IV, alpha subunit
6	KCNJ4	Potassium inwardly-rectifying channel, subfamily J, member 4
7	CACNB3	Calcium channel, voltage-dependent, beta 3 subunit
8	KCNJ12	Potassium inwardly-rectifying channel, subfamily J, member 12
9	KCNA1	potassium voltage-gated channel, shaker-related subfamily, member 1 (episodic ataxia with myokymia)
10	CACNA1A	Calcium channel, voltage-dependent, P/Q type, alpha 1A subunit

6.11 is very much dominated by the different subtypes of voltage-gated potassium channels, which occupy 6 out of the 10 positions when RW-indiv is applied to the selection of novel proteins, with the remaining genes selected being ion channels or exchangers of sodium and/or calcium ions. Hence, it can be seen that both methods arrive at a very different selection of genes involved in the disease cluster, as well as the identification of novel disease genes using the RW-Indiv method. Combined with the fact that very disease relevant genes were identified in Table 6.11 (as discussed above), we believe that this illustrates the performance of the method implemented in this work.

6.6 Conclusions

Prediction accuracy of almost all the previous work on predicting disease-related proteins depends directly on the initial disease-related proteins (seed proteins). While the initial seed proteins of each disease suffers from several 'False Negative' cases, dependency of previous methods on the incomplete seed proteins is the main drawback of these methods.

In this article, we reduced the number of the False Negative cases in the initial seed proteins by proposing informative Human Disease Network (HDN). We analyzed different *Structural* and *Functional* prediction methods and we concluded that

a hybrid method which considers both Structural and Functional information in the PPI network is the best method for building the HDN. We built a HDN based on 20 diseases and we showed that resulting HDN is biologically meaningful. Then, we clustered HDN and we augmented the seed proteins of diseases based on the cluster they belong to. Finally, we predicted disease-related proteins based on the augmented version of seed proteins. Literature mining of the newly found disease-related proteins proved the usefulness of using our proposed HDN for predicting disease-related proteins.

6.7 Acknowledgment

This research is funded by the Dutch Science Foundation (NWO) through a VIDI grant.