International Journal of Innovative Computing, Information and Control
Volume **8**, Number **2**, February **2012**

ICIC International ©2012 ISSN 1349-4198
pp. 1327–1338

PRIORITIZING DISEASE GENES BY INTEGRATING DOMAIN INTERACTIONS AND DISEASE MUTATIONS IN A PROTEIN-PROTEIN INTERACTION NETWORK

Bongjun Song and Hyunju Lee*

Department of Information and Communications
Gwangju Institute of Science and Technology
1 Oryong-dong, Buk-gu, Gwangju 500-712, Republic of Korea
bjsong@gist.ac.kr; *Corresponding author: hyunjulee@gist.ac.kr

Received November 2010; revised March 2011

Abstract. *Complex diseases such as cancer are involved in inter-relationship among several genes, with protein-protein interaction networks being extensively studied in attempts to reveal the relationship between genes and diseases. Although these studies have shown promising results for identifying disease genes, it is not systemically studied that a protein functions differently depending on its interaction partners in the network since a protein can have multiple functions. In this study, domains are considered as functional units of proteins and we investigate how disease-related mutations in domains can be used to identify other disease genes in a domain-domain interaction network. We subsequently propose a computational method to predict disease genes based on the following two assumptions. The first assumption is that proteins closely interacting with known disease proteins in a protein interaction network are likely to be involved in the same disease. Second, although two proteins are in the same distance from known disease genes in a protein interaction network, the protein interacting with known disease genes through a domain with mutation is more likely to be related to the disease than other proteins that interact through domains with no mutation. As a result, when the proposed approach is applied to five diseases, it highly ranks disease-related genes compared to a model using only a protein interaction data set.*

Keywords: Disease gene prediction, Bioinformatics, Domain-domain interactions, Disease related mutations

1. **Introduction.** Complex diseases such as cancer and metabolic disorders are involved in inter-relationship among several genes. To date, revealing the underlying mechanisms of these diseases remains challenging due to the complexity of their interactions [1]. The first step in the study of complex diseases is to identify disease-causing genes; mutations of genes related to diseases are identified by experimental examinations. However, the complex underlying mechanisms of many diseases are not fully explained by the known disease related genes. Thus, identifying disease-related genes and their correlations still requires further study.

For this task, based on the growth of various biological data sources, extensive studies in developing computational methods for identifying disease genes have been made. For example, Aerts et al. [2] developed Endeavour, a method for integrating several genomic data sources that uses order statistics to prioritize disease genes. It first gathers information extracted from various data sources, such as literature, gene expressions, protein-domains and protein-protein interactions, to train genes. For each data source, test genes are then ranked based on their functional similarity to the training genes. Finally, rankings from the separate data sources are fused into a single ranking using order statistics. In this way, Endeavour improved the accuracy and coverage of predictions by

integrating several data sources. In another study, Wu et al. [3] proposed a network-based regression model (CIPHER) for predicting disease genes. This model uses three types of data sources, including manually curated protein-protein interactions from the HPRD database [4] for gene-gene networks, disease-gene associations from the OMIM database [5] for disease-gene networks, and similarities between diseases calculated via text mining [6] for disease-disease networks. By integrating the protein interaction data sets and disease similarities, CIPHER showed comparable performance to Endeavour.

Even though many studies have investigated disease genes based on protein-protein interaction networks, few have systemically incorporated interactions at a domain level. In one such case, Wang et al. [7] suggested that mutations can disrupt bindings between domains, and that this disruption can change pathways related to disease, thereby causing the disease. However, they did not develop a method for predicting disease genes based on their observations. As such, in this study, we extend the investigations of protein-protein interactions by using the functional units of their domains in order to develop a computational method for predicting disease genes, based on the following assumptions:

- First, proteins more closely interacting with the proteins of known disease genes in the protein-protein interaction networks are likely to be involved in the same disease.
- Second, if a protein P_1 has a mutation related to a given disease, and it interacts with a protein P_2 through a shared domain, then P_2 is more likely to be involved in the same disease than other proteins that interact with P_1 through domains with no mutation.

In Figure 1(a), when the protein P_1 contains a domain that has disease-related mutations, we assume that proteins P_2 and P_3, interacting with P_1 through the domain, are more closely related to the same disease as P_1 than P_4 and P_5. Based on these assumptions, we develop a method to measure the similarities between disease proteins and other proteins by considering both the protein-protein and domain-domain interaction networks. Using the domain-domain interactions, the proposed method allows us to distinguish direct and indirect disease-related interactions from among other complicated protein interactions. Then, when applied to five diseases, we observed in three diseases that the model incorporating interactions between domains with mutations in the protein interaction network helped to highly rank disease-related proteins, as compared to a model using only the protein interaction data set.

2. **Methods.**

2.1. **Data sources.**

In this study, we collect protein-protein interactions, protein-domains, domain-domain interactions, disease genes and disease-related mutation data sets for humans.

2.1.1. *Protein-protein interactions.*

We use a human protein-protein interaction data set obtained from the HPRD database (January 2008 version) [4], which is the same data set used in [3]. This set contains 34,364 manually curated interactions from among 8,919 human proteins.

2.1.2. *Protein domains.*

The Pfam database is a large collection of protein domain families [8]. We use this database (version 21.0) to map domains to proteins. Among two types of Pfam families (Pfam-A and Pfam-B), we use the manually curated Pfam-A. Proteins in Pfam and proteins in HPRD are then mapped using the accession number of the proteins in a SwissProt database [9].

2.1.3. *Domain-domain interactions.* The domain-domain interaction data set was obtained from iPfam (version 21.0) [10]. This database provides domain-domain interaction information between two interacting proteins. The database uses Pfam-A domains (version 21.0). Among all 7,265 domain-domain interactions with protein information in iPfam, we use 359 interactions between two different human proteins. Note that proteins in these interactions are also found in the HPRD human protein-protein interactions.

2.1.4. *Human diseases, disease genes and mutations.* OMIM is a continuously updated catalog of human genes and genetic disorders [5]. We collected human diseases, disease genes and disease-related protein mutations from the OMIM database (June 19, 2009 version). Disease genes were obtained using the following information:

- Genes are in the TEXT field of the disease phenotype in OMIM.
- Genes of the disease phenotype are in a Morbid Map list, an alphabetical list of diseases described in OMIM and their corresponding cytogenetic locations.
- Genes with mutation information are related to the disease phenotype.

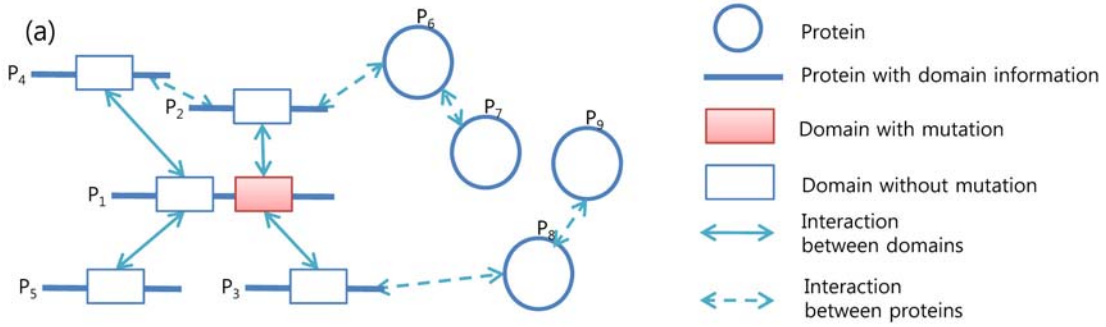
In addition, the allelic variants information of genes in OMIM are used for collecting the mutation information for disease genes. In the mutation data, a disease name consists of comma-separated multi-level names, and we use only the first level name of the disease name. For example, we consider all of the following disease names as breast cancer: 'breast cancer, somatic', 'breast cancer, sporadic', 'breast cancer famillial', 'breast cancer' and 'breast cancer, lobular, somatic.' Finally, we are able to categorize mutation data into 3,525 diseases, based on the location of the mutation in the protein sequence and the corresponding sequence domain. Here, protein identifiers in HPRD and OMIM are mapped using Entrez gene identifiers [11].

2.1.5. *Domain-domain interactions with disease-related mutations.* We subsequently integrate domain-domain interactions with disease-related mutations in a protein-protein interaction network, referred to as a domain-domain interaction in a disease-related protein with mutation (DDI-DRPM). In DDI-DRPM, at least one of two proteins has disease-related mutation; the protein with mutation is referred to as DDI-DRPM-P_m.

We classify DDI-DRPM into a particular disease type if at least one of two interacting proteins has a domain with disease-related mutations. Note that DDI-DRPM may be included into more than one disease if proteins have mutations related to more than one disease. Through this process, we obtain the DDI-DRPM list for 160 diseases. After filtering out diseases that have less than two DDI-DRPMs or have less than two proteins with disease-related mutation data, we obtain the DDI-DRPMs of 9 diseases. For these diseases, we map the diseases into disease phenotypes in OMIM, and then we add proteins with mutations into the disease gene list. With the manual inspection of the 9 diseases, we also filter out diseases that are too general or contain less than three disease genes. In addition, we divide 'cardiomyopathy' into the two diseases 'cardiomyopathy, dilated' and 'cardiomyopathy, hypertrophic'. Finally, we use the DDI-DRPM of five diseases, as listed in Table 1.

2.2. **Computational methods.**

2.2.1. *Proposed method for prioritizing disease genes by combining domain interactions and mutations in a protein interaction network.* We propose a computational method to prioritize disease genes based on the two assumptions discussed in the Introduction section. Our model prioritizes test proteins based on a protein-protein interaction network with DDI-DPRM information for a given disease. Here, proteins with mutation information (DDI-DRPM-P_m) in DDI-DRPM are used as training proteins. We then calculate



(b)

h	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
d_{P_1h}	0	1	1	1	1	2	3	2	3
d'_{P_1h} =min $\{d_{P_2h}, d_{P_3h}\}$	2 {2,-}	0 {0,-}	0 {3,0}	1 {1,-}	3 {3.-}	1 {1,-}	2 {2,-}	1 {4,1}	2 {5,2}

d_{P_1h}: the shortest distance between P_1 and h in PPI, d_{P_2h}: the shortest distance between P_2 and h in PPI – $\{(P_1, P_2)\}$,
d_{P_3h}: the shortest distance between P_3 and h in PPI – $\{(P_1, P_3)\}$, '-' : no path between two proteins

h	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
$cptn_{P_1h}$	-	4	3	3	2	2	1	2	1

$cptn_{P_1h}$: the number of common interaction partners between Extended_PTN(P_1) and PTN (h)

Extended_PTN(P_1) = PTN(P_1) ∪ PTN'(P_2) ∪ PTN'(P_3) = {P_1, P_2, P_3, P_4, P_5, P_6, P_8},
where PTN(P_1) : partners of P_1 in PPI = {P_1, P_2, P_3, P_4, P_5}
PTN'(P_2) : partners of P_2 in PPI' = {P_2, P_4, P_6}, where PPI' = PPI – {(P_1, P_2)}
PTN'(P_3) : partners of P_3 in PPI' = {P_3, P_8}, where PPI' = PPI – {(P_1, P_3)}

PTN(P_2) = {P_1, P_2, P_4, P_6}, PTN(P_3) = {P_1, P_3, P_8}, PTN(P_4) = {P_1, P_2, P_4}, PTN(P_5) = {P_1, P_5}, PTN(P_6) = {P_2, P_6 ,P_7},
PTN(P_7) = {P_6, P_7}, PTN(P_8) = {P_3, P_8 , P_9}, PTN(P_9) = {P_8 , P_9}

(c)

h	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
rsd_{P_1}(h)	1	1	1	1	5	7	5	7
rsd'_{P_1}(h)	1	1	3	8	3	6	3	6
$rcptn_{P_1}$(h)	1	2	2	4	4	7	4	7
rank (h)	1	2	3	6	4	7	4	7

FIGURE 1. A schematic of our approach. (a) In a protein-protein interaction network with domain-domain interaction and mutation information, P_1 is used as a training protein and other proteins are test proteins. P_2 and P_3 interact with P_1 through the domain with mutation. In contrast, P_4 and P_5 interact with P_1 through the domain without mutation. (b) d_{P_1h}, the shortest distance between P_1 and the other proteins, and d'_{P_1h}, the shortest distance between the DDI-DRPM-partner (P_1) = {P_2 and P_3} and the other proteins, are calculated. Then, $cptn_{gh}$, the number of common proteins between the Extended_PTN(g) and PTN(h), is presented. (c) After similarities of sd_{P_1h} and sd'_{P_1h} using d_{P_1h} and d'_{P_1h} are calculated, three respective rankings of test proteins using sd_{P_1h}, sd'_{P_1h} and $cptn_{gh}$ are determined. Then, the final ranking is shown in rank(h).

the similarity between the training and test proteins by using three similarity measures; the first measure is based on the protein-protein interaction network, and the second and the third are based on DDI-DRPM information.

Similarity between training proteins and test proteins: Three similarity measures between a training protein g and a test protein h are calculated. For the given training protein g, let DDI-DRPM-partner(g)= $\{g_p |\ g_p$ interacting with g in DDI-DRPM$\}$ be the

TABLE 1. List of DDI-DRPM for five diseases. For each disease, interacting proteins, in which interacting domains are known and at least one of proteins has disease related mutation information, are listed.

Disease	DDI-DRPM			
	P_m		P_m's partner	
	Protein 1	Domain 1	Protein 2	Domain 2
Cardiomyopathy, dilated	VCL	PF01044	TLN1	PF09141
	TNNT2	PF00992	TNNI3	PF00992
	TNNC1	PF00036	TNNI3	PF00992
Cardiomyopathy, hypertrophic	TNNI3	PF00992	TNNC1	PF00036
			TNNT2	PF00992
	TNNT2	PF00992	TNNI3	PF00992
Prostate cancer	AR	PF00104	NCOA2	PF08832
	MAD1L1	PF05557	MAD2L1	PF02301
Colorectal cancer	TP53	PF00870	TP53BP1	PF00533
			TP53BP2	PF00018 PF00023
	FGFR3	PF00047	FGF1	PF00167
Breast cancer	TP53	PF00870	TP53BP1	PF00533
			TP53BP2	PF00018 PF00023
	RAD51	PF08423	BRCA2	PF00634

interaction partners of the training protein g in DDI-DRPM. We represent the protein-protein interaction network as PPI, and the protein-protein interaction network excluding the interaction between g and g_p as PPI'.

- *Shortest path between g and h*: Let the shortest distance between g and h in PPI be d_{gh}. Their similarity is defined as $sd_{gh} = \exp(-d_{gh}^2)$.
- *Shortest path between DDI-DRPM-partner(g) and h*: By assuming that g_p, the protein in DDI-DRPM-partner(g), is closely involved in the same disease as g, we use the distance between g_p and h as a similarity measure between g and h. As g interacts in multiple DDI-DRPMs, let the shortest distance between them be $d'_{gh} = \min\{d_{g_{p_i}h}\}_{i=1}^{m}$, where m is the DDI-DRPM-partner number (g) and $d_{g_{p_i}h}$ is the shortest distance between g_{p_i} and h in PPI'. Then, the similarity between them is defined as $sd'_{gh} = \exp(-d'^2_{gh})$.
- *Common interaction partners between g and h*: Let the extended interaction partners of g be Extended_PTN(g) = $\{g\} \cup$ {all direct interaction partners of g in PPI} $\cup$ $\{g_p\} \cup$ {all direct interaction partners of g_p in PPI'}. Then, the extended interaction partners of g include both the directly interacting proteins of g as well as the directly interacting proteins of g_p. Let the interaction partners of h be PTN(h)=$\{h\} \cup$ {all direct interaction partners of h in PPI}, with their similarity defined as $cptn_{gh}$ = {the number of common proteins between Extended_PTN(g) and PTN(h)} .

Prioritizing of test proteins for disease: To prioritize test proteins for a given disease, we calculate the rankings of the test proteins based on their similarity to the training proteins. Let us assume that there are k test proteins of $\{h_i\}_{i=1}^{k}$. Then, for each training protein g, the three ranks of test protein h_i are first calculated as follows.

- $rsd_g(h_i)$: Rank of the test protein h_i in $\{sd_{gh_i}\}_{i=1}^{k}$, the shortest path similarities between a training protein g and k test proteins of $\{h_i\}_{i=1}^{k}$.

- $rsd'_g(h_i)$: Rank of the test protein h_i in $\{sd'_{gh_i}\}_{i=1}^k$, the shortest path similarities between DDI-DRPM-partner(g) and k test proteins of $\{h_i\}_{i=1}^k$.
- $rcptn_g(h_i)$: Rank of the test protein h_i in $\{cptn_{gh_i}\}_{i=1}^k$, common interaction partner similarities between g and k test proteins of $\{h_i\}_{i=1}^k$.

Next, when there are n training proteins of $\{g_j\}_{j=1}^n$ for a given disease, the overall ranks of h_i are calculated using order statistics.

$$rank(h_i) = \text{OrderStatistics } (rsd_{g_1}(h_i), rsd'_{g_1}(h_i), rcptn_{g_1}(h_i), \ldots, rsd_{g_n}(h_i), rsd'_{g_n}(h_i), rcptn_{g_n}(h_i)).$$

The order statistics are calculated using the formula defined in [2]. An example of the three similarity measures and the ranks of test proteins is shown in Figure 1. Figure 1(a) is a protein interaction network with domain interaction information; P_1 is a training protein, and P_2 and P_3 interact with P_1 through a domain with mutation. Figure 1(b) shows three types of similarities between P_1 and the other proteins. In the figure, $d_{P_1,h}$ is the shortest distance between P_1 and the protein h in the PPI network. In addition, $d'_{P_1,h}$ is the shortest distance between P_1 and protein h when the interaction between P_1 and P_2 and the interaction between P_1 and P_3 are removed from PPI network, where P_2 and P_3 interact with P_1 through a domain with mutation. And $cptn_{P_1,h}$ is the number of extended common interaction partners between P_1 and h. In Figure 1(c), the ranks from the three similarity measures are obtained and the overall ranking between P_1 and the other proteins are calculated. Note that proteins P_6 and P_8, which interact with P_2 and P_3, are ranked higher than protein P_5, which directly interacts with P_1 through a domain with no mutation.

2.2.2. *Method for prioritizing disease genes in a protein interaction network.* We then compare our method with a simple method using only a protein-protein interaction network. For a training protein g and a test protein h, the similarity between them is calculated using the distance d_{gh} of the shortest path. If there are n training proteins in the protein-protein interaction network, the overall similarity between n training proteins and the test protein h is $\sum_{j=1}^n \exp(-d_{gjh}^2)$, where g_j is the j-th training protein. This similarity is calculated for all test proteins, and their rankings are then determined. This model is referred to as a protein-protein interaction model (PIM) in the following sections.

3. **Results.** The proposed method was applied to five diseases; dilated cardiomyopathy, colorectal cancer, hypertrophic cardiomyopathy, breast cancer and prostate cancer have 25, 29, 25, 13 and 18 known disease related genes, respectively. Among them, 3, 2, 2, 2 and 2 genes were used as training genes in this study because they have DDI-DPRM information. Using these training genes, 8,919 genes that have interacting partners in a protein-protein interaction network, were then ranked using the proposed model. The higher the method ranks the disease-related genes, the better the method is perceived. We first show the performance of our method by comparing it with other methods, and then explain the five diseases in further detail. In the comparison, we focus on genes ranked in the top 300 since they might can be considered as candidate genes. These top 300 were previously used by [3].

3.1. **Performance comparison among PIM, CIPHER and proposed model.** We compared PIM, CIPHER [3], and the proposed method by counting the number of disease genes ranked in the top 300, as shown in Table 2. PIM predicts several genes in the same rank because many genes are the same distance from the training genes in the protein-protein interaction network; hence, when several genes are assigned into the same rank, the medium rank is assigned into these genes. For example, if 10 genes need to be ranked

in the same rank between 11 and 20, the rank of these genes is considered to be 15. Our model places disease genes at higher rankings compared to PIM, for three diseases: dilated cardiomyopathy, colorectal cancer and breast cancer.

CIPHER integrates the protein-protein interaction network with disease phenotype similarity to prioritize disease genes. Since our method uses different data sources than CIPHER, and the number of domain-domain interaction data in the protein interaction network is not sufficient, it is hard to directly compare the performances of the two methods. However, we used CIPHER for the comparison because CIPHER is one of best methods that use a protein-protein interaction network and shows comparable performance to Endeavour [2], which integrates several biological data sets. When the same training genes were used for both methods, our method had better performance than CIPHER for two diseases, similar performance for one disease, and worse performance for two diseases. This result indicates that our assumption of incorporating domain interaction and mutation data sets for predicting disease genes is promising.

In the following, highly ranked genes for five diseases are examined in further detail. Also, ranks of known disease genes from both the PIM model and our model are presented in order to show the effect of incorporating disease mutation information in the domain-domain interaction network.

TABLE 2. For five diseases, the number of disease genes ranked in the top 300 are shown for the proposed model, PIM and CIPHER

Disease	Proposed model	PIM	CIPHER
Dilated cardiomyopathy	8	6	14
Colorectal cancer	4	1	1
Hypertrophic cardiomyopathy	4	4	10
Breast cancer	13	10	10
Prostate cancer	2	3	2

3.2. **Dilated cardiomyopathy.** Cardiomyopathy is a heart muscle disease, and is classified into four types: dilated, hypertrophic, arrhythmogenic right ventricular and restrictive cardiomyopathy. In particular, the left ventricle of a patient with dilated cardiomyopathy becomes stretched, at which time the heart muscle becomes weak and thin, and is unable to pump blood efficiently [12].

Among 25 known dilated cardiomyopathy genes, three genes (VCL, TNNT2 and TNN C1) are used here as training genes because they have DDI-DPRM information. Predicted rankings of the remaining 22 disease genes calculated from PIM and the proposed model are shown in Table 3. Among 8,916 test genes, 8 out of 22 genes are ranked within the top 300 using our model (p-value < 0.0001, Fisher's exact test, one sided). Let us now examine these genes in detail. A part of the protein-protein interaction network, including 3 training genes and 8 genes ranked in the top 300, are illustrated in Figure 2 with domain-domain interactions. Note that TTN and DMD in our model are highly ranked compared to the PIM method; TTN and DMD share a common interaction partner (ACTA1) with TLN1, which is a DDI-DRPM-partner of the training gene VCL. It should also be noted that even though ACTA1 is not included as a dilated cardiomyopathy causing gene in the OMIM, multiple lines of evidence show that ACTA1 mutations result in congenital myopathies [13]. Also, ACTN2 shares a common interaction partner (PKD2) with TNNI3, a DDI-DRPM-partner of the training genes TNNT2 and TNNC1, which assigns ACTN2 to the top 300. ACTC and MYBPC3 are ranked in the top 300 using the proposed method, as both proteins share a common interaction partner (TNNI3K) with TNNI3,

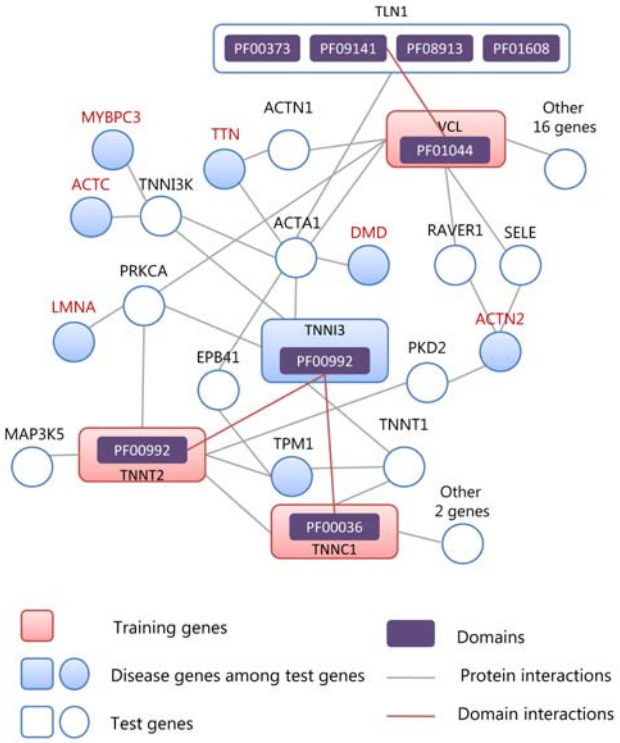


FIGURE 2. Protein-protein interaction network for dilated cardiomyopathy. Three training genes (VCL, TNNT2 and TNNC1), disease genes ranked in the top 300, and test genes closely connected with the training genes are shown in the PPI network. Training genes and their DDI-DPRM partners (TNNI3 and TLN1) are presented for domains, and their interactions are shown at a domain level. Training genes' domains have disease-related mutations. Among the disease genes, names of genes highly ranked using our method (as compared to PIM) are represented in red.

which is a DDI-DRPM-partner of the training genes TNNT2 and TNNC1. TNNI3 itself is a known dilated cardiomyopathy gene and it has recently been reported that TNNI3K, a TNNI3 interacting kinase, plays an important role in the progression of cardiomyopathy in the murine model [14]. This result indicates that our model successfully can incorporate indirect interactions with the training genes through TNNI3K.

On the other hand, 6 genes are ranked within top 300 using PIM. Three disease genes (TTN, DMD and ACTN2) are ranked within the top 288 because they are two units apart from one training gene in the PPI network. In addition, ACTC and MYBPC3 are ranked in the top 887 because they have the shortest distance to three training genes in the PPI network. Since many genes are two or three units apart from training genes in PPI networks, disease genes are not highly ranked using only PPI information.

3.3. **Colorectal cancer.** Among 29 known colorectal cancer genes, the two genes TP53 and FGFR3 are used for training. The prediction results of the other 27 disease genes are shown in Table 4. Using our proposed method, 4 out of 27 genes are ranked in the top 300 among the 8,917 test genes (p-value = 0.0120, Fisher's exact test, one sided). Using PIM, only 1 gene is ranked in the top 300. Although the number of colorectal cancer genes in top 300 is relatively small, Figure 3 illustrates that our proposed method can successfully distinguish disease-related genes from among other genes having the same topological distance from training disease genes in a protein-protein interaction

TABLE 3. Comparison of PIM and the proposed model for dilated cardiomyopathy genes. Three genes (VCL, TNNT2 and TNNC1) are used for training and 22 other dilated cardiomyopathy genes are ranked using PIM and the proposed model. Among 8,916 test genes, 8 out of 22 genes are ranked within the top 300 using our model, whereas 6 genes are ranked within the top 300 using PIM. The bold and italic fonts indicate the genes higher ranked in the proposed model than in PIM among the top 300 genes.

Genes	Rank		Genes	Rank	
	Proposed model	PIM		Proposed model	PIM
TNNI3	2	1	***TTN***	62	288
TPM1	81	7	***DMD***	97	288
LMNA	145	169	***ACTN2***	255	288
ACTC	280	887	***MYBPC3***	280	887
CSRP3	484	502	PSEN1	484	502
DES	715	670	SCN5A	929	1983
TMPO	2099	1983	PLN	2685	3535
TCAP	2685	3535	LDB3	3146	3535
TAZ	3438	1983	PSEN2	3533	3535
DSG2	4293	4501	SGCD	5847	6546
ABCC9	5847	6546	MYH7	7641	7515

TABLE 4. Comparison of PIM and the proposed model for colorectal cancer genes. Two genes (TP53 and FGFR3) are used for training and 27 other known colorectal cancer genes are included as the test genes. Using our proposed method, 4 out of 27 genes are ranked in top 300 among 8,917 test genes, while only 1 gene is ranked using PIM. The bold and italic fonts indicate the genes higher ranked in the proposed model than in PIM among the top 300 genes.

Genes	Rank		Genes	Rank	
	Proposed model	PIM		Proposed model	PIM
EP300	6	21	***AKT1***	91	362
CTNNB1	99	1290	***CCND1***	137	362
CHEK2	383	1290	SMAD7	451	1290
MSH2	489	1290	BRAF	551	1290
MSH6	551	1290	MLH1	705	1290
KRAS	957	1290	TGFBR2	1042	1290
APC	1042	1290	BUB1	1173	1290
DCC	1381	1290	MUTYH	2005	2524
AXIN2	2005	2524	BUB1B	2349	1290
PIK3CA	2579	3645	DLC1	3395	3645
PMS1	4450	5597	PMS2	4450	5597
MLH3	4450	5597	MCC	4450	5597
MYH11	5924	5597	PLA2G2A	6995	7471
MTCO1	8494	8705	–	–	–

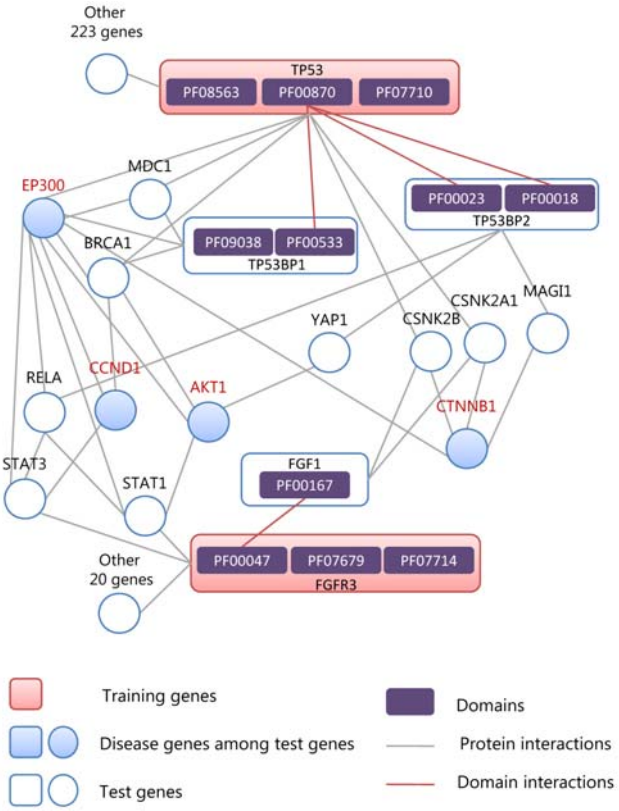


FIGURE 3. Protein-protein interaction network for colorectal cancer. Two training genes (TP53 and FGFR3), disease genes ranked in the top 300, and test genes closely connected to training genes are shown in the PPI network. Training genes and their DDI-DPRM partners (TP53BP1, TP53BP2 and FGF1) are presented with domains and their interactions are shown at a domain level. Training genes' domains have disease-related mutations. Among the disease genes, names of genes highly ranked using our method (as compared to PIM) are represented in red.

network, which allows us to predict candidate genes located a long distance from known disease genes. Here, EP300 is ranked in the top 6 because it directly interacts with the training gene TP53 and also interacts with TP53BP1, which is a DDI-DRPM-partner of the training gene TP53. In addition, CCND1 and AKT1 are distinguished among 271 genes with distance of two from the training gene TP53 and the training gene FGFR3; they share a common interaction partner (BRCA1) with TP54BP1, which is a DDI-DRPM-partner of the training gene TP53. And even though the relationship between the BRCA1 mutation and colorectal cancer is not specified in OMIM, several studies have indicated that BRCA1 plays a role in the development of colorectal cancer [15]. CTNNB1 is also highly ranked in our model, as it has common interaction partners with TP52BP2 and FGF1, both of which are DDI-DRPM-partners of the training genes TP53 and FGFR3.

3.4. **Breast cancer, hypertrophic cardiomyopathy and prostate cancer.** In breast cancer, among 25 known disease genes, RAD51 and TP53 are used as training genes. Using the proposed method, 13 out of these other 23 genes are ranked in the top 300 among 8,917 test genes (p-value < 0.0001, Fisher's exact test, one sided).

In hypertrophic cardiomyopathy, among 13 known disease genes, the two genes TNNI3 and TNNT2 are used for training. Using the proposed method, 4 out of the other 11

genes are ranked in the top 300 among 8,917 genes (p-value = 0.0003, Fisher's exact test, one sided). This disease also illustrates that our model can distinguish disease genes from among other genes at the same topological distance from the training disease genes in a protein-protein interaction network.

In prostate cancer, among 18 known disease genes, the two genes AR and MAD1L1 are used for training. Using our proposed method, 2 out of the remaining 16 genes are ranked in the top 300 among 8,917 test genes (p-value = 0.0993, Fisher's exact test, one sided). However, though the results of this test do not show any statistically significant prediction, we expect that this result might be improved if sufficient domain-domain interaction data with mutation information is made available in the future.

4. **Discussion and Conclusion.** The model proposed in this study demonstrates that domain interactions and disease-related mutations are helpful for prioritizing disease genes in protein-protein interaction networks. Here, the analysis of interaction networks in five diseases shows that test genes that interact with DDI-DPRM partners of training genes are more likely to be involved in the given disease. And although some genes do not contain disease-related mutations in their own sequences, their functions might be affected through their interaction with proteins containing domains with mutations. This observation suggests that pathways related to the disease can be revealed using domain-domain interactions.

The advantage of this study is that it successfully incorporated domain-domain interactions and mutation information at a domain level, resulting in the improvement in predicting disease genes compared to other methods that only use protein-protein interactions. However, this study is limited by a small number of overlaps between the disease mutation information and domain-domain interactions. Indeed, even though we collected all available disease-related mutation data sets and experimental domain interaction data sets, the proposed method was applied only to five diseases. Nevertheless, this study shows promising results in these diseases, so we expect that the proposed approach might help to identify disease genes with high accuracy for many other diseases as more experimental data sets become available. For instance, the shortage of domain-domain interaction data sets might be resolved by using computationally predicted domain-domain interactions; there have also been extensive efforts to develop methods for predicting domain-domain interactions by integrating biological data sets [16].

Recently, Wang et al. [7] proposed a computational method for predicting domain-domain interactions in two interacting proteins. When the reliability of these predicted data sets increases, it is expected that these computationally predicted domain interactions can be applied to our proposed method. Currently, one of reasons that the proposed method can only be applied to only five diseases is that it requires proteins have domains with disease mutations and domain-domain interactions. In our future work, we will improve our method in order to use for cases in which the disease mutation information and domain-domain interactions do not commonly occur in the same protein.

In addition to revealing disease related genes, the classification of patients from 'normal' people is also an important problem. Many studies have been performed in attempts to find a small set of genes that can be used to correctly identify patients [17, 18, 19]. The main issue in these studies is the selection informative genes; the approach used in our study to incorporate mutation information and domain-domain interaction might be helpful to find a small number of disease related genes and thereby improve classification accuracy. Thus, as a future work, it is expected that our approach of incorporating disease mutation information and domain-domain interactions can be extended to the disease classification problem.

Acknowledgment. This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (MEST) (2010-0003597).

REFERENCES

[1] M. Kann, Protein interactions and disease: Computational approaches to uncover the etiology of diseases, *Brief Bioinform*, vol.8, no.5, pp.333-346, 2007.

[2] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet and Y. Moreau, Gene prioritization through genomic data fusion, *Nat Biotechnol*, vol.24, no.5, pp.537-544, 2006.

[3] X. Wu, R. Jiang, M. Zhang and S. Li, Network-based global inference of human disease genes, *Mol. Syst. Biol.*, vol.4, no.189, 2008.

[4] S. Peri, J. Navarro, R. Amanchy, T. Kristiansen, C. Jonnalagadda, V. Surendranath et al., Development of human protein reference database as an initial platform for approaching systems biology in humans, *Genome Res.*, vol.13, no.10, pp.2363-2371, 2003.

[5] V. McKusick, Mendelian inheritance in man and its online version, OMIM, *Am. J. Hum. Genet.*, vol.80, no.4, pp.588-604, 2007.

[6] M. van Driel, J. Bruggeman, G. Vriend, H. Brunner and J. Leunissen, A text-mining analysis of the human phenome, *Eur. J. Hum. Genet.*, vol.14, no.5, pp.535-542, 2006.

[7] H. Wang, E. Segal, A. Ben-Hur, Q. Li, M. Vidal and D. Koller, InSite: A computational method for identifying protein-protein interaction binding sites on a proteome-wide scale, *Genome Biol.*, vol.8, no.9, 2007.

[8] A. Bateman, L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, D. Studholme, C. Yeats and S. Eddy, The Pfam protein families database, *Nucleic Acids Res.*, vol.32, pp.D138-D141, 2004.

[9] R. Apweiler, Functional information in SWISS-PROT: The basis for large-scale characterisation of protein sequences, *Brief Bioinform*, vol.2, no.1, pp.9-18, 2001.

[10] R. Finn, M. Marshall and A. Bateman, iPfam: Visualization of protein-protein interactions in PDB at domain and amino acid resolutions, *Bioinformatics*, vol.21, no.3, pp.410-412, 2005.

[11] D. Maglott, J. Ostell, K. Pruitt and T. Tatusova, Entrez gene: Gene-centered information at NCBI, *Nucleic Acids Res.*, vol.33, pp.D54-D58, 2005.

[12] J. N. Jameson, D. L. Kasper, T. R. Harrison, E. Braunwald, A. S. Fauci, S. L. Hauser and D. L. Longo, *Harrison's Principles of Internal Medicine*, 16th Edition, McGraw-Hill Medical Publishing Division, New York, 2005.

[13] J. Feng and S. Marston, Genotype-phenotype correlations in ACTA1 mutations that cause congenital myopathies, *Neuromuscul Disord*, vol.19, no.1, pp.6-16, 2009.

[14] F. Wheeler, H. Tang, O. Marks, T. Hadnott, P. Chu, L. Mao, H. Rockman and D. Marchuk, Tnni3k modifies disease progression in murine models of cardiomyopathy, *PLoS Genet.*, vol.5, no.9, pp.e1000647, 2009.

[15] H. Grabsch, M. Dattani, L. Barker, N. Maughan, K. Maude, O. Hansen, H. Gabbert, P. Quirke and W. Mueller, Expression of DNA double-strand break repair proteins ATM and BRCA1 predicts survival in colorectal cancer, *Clin. Cancer Res.*, vol.12, no.5, pp.1494-1500, 2006.

[16] H. Lee, M. Deng, F. Sun and T. Chen, An integrated approach to the prediction of domain-domain interactions, *BMC Bioinformatics*, vol.7, no.269, 2006.

[17] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, vol.7, no.6, pp.673-679, 2001.

[18] M. S. Mohamad, S. Omatu, M. Yoshioka and S. Deris, A cyclic hybrid method to select a smaller subset of informative genes for cancer classification, *International Journal of Innovative Computing, Information and Control*, vol.5, no.8, pp.2189-2202, 2009.

[19] M. S. Mohamad, S. Omatu, M. Yoshioka and S. Deris, A three-stage method to select informative genes for cancer classification, *International Journal of Innovative Computing, Information and Control*, vol.6, no.1, pp.117-125, 2010.