International Journal of Management, Economics and Social Sciences
Special Issue-International Conference on Medical and Health Informatics (ICMHI 2017)
2017, Vol. 6(S1), pp.274 – 292.
ISSN 2304 – 1366
http://www.ijmess.com

Unveiling Disease-Protein Associations by Navigating a Structural Alphabet-Encoded Protein Network

[*]Chi-Hua Tung[1]
Jih-Hsu Chang[2]
Jose C. Nacher[3]

[1]Dept. of Bioinformatics, Chung-Hua University, Taiwan
[2]Dept. of Bioinformatics, Chung-Hua University, Taiwan
[3]Dept. of Information Science, Toho University, Japan

The identification of genes associated with human disorders is a major goal in computational biology. Although the rapid emergence of cellular network-based approaches has been successful in many instances, all of these methodologies are partially limited by the incompleteness of the interactome. Here, we propose a novel method that may overcome the inherent problem of these incomplete molecular networks and assist already established network techniques. Instead of using protein-protein interaction networks, we encode the local three-dimensional structure of a protein into a series of letters, called the Structural Alphabet, and define a proteomic structural network in which each node represents a unit of the structural alphabet (USA) and each pair of USAs is linked based on their structural similarity. This novel structural network is the platform by which a diffusion-based algorithm determines the potential involvement of proteins in disease phenotypes. Computational experiments show that the combination of diffusion-based methods with the constructed structural alphabet network offers better predictive performance than the results obtained using interactome networks and provides a new avenue to assist in identifying disease-related proteins.

Keywords: Local structure similarity network, random walk with restart, protein modularity, structural alphabet

The identification of genes associated with human disorders is a major goal in computational biology, and it has direct medical, therapeutic and clinical implications. Various methodologies have been proposed to address this challenging problem, ranging from linkage mapping to genome-wide association (GWA) studies [1]. Recently, the rapid emergence of network-based methods has also expanded the technical tools available for predicting disease-associated genes [2, 3]. However, as recently highlighted by Menche, Sharma, Kitsak, Ghiassian, Vidal, Loscalzo and Barabasi [4], network-based approaches, although successful in many instances, are partially limited by the incompleteness of the interactome. Here, we

Manuscript received May 20, 2017; revised September 28, 2017; accepted October 28, 2017.
*Corresponding author: chihua.tung@chu.edu.tw

propose a novel method that may overcome the inherent problem of the largely incomplete protein interaction network. Instead of using interactome networks, we encode the local three-dimensional (3D) structure of a protein into a series of letters, called the Structural Alphabet (SA), and define a proteomic structural network in which each node represents a unit of the structural alphabet (USA) and each pair of USAs is linked based on their structural similarity. The USA and its similarity network are designed for local structure approximation to describe the significant core of protein and for investigating the structure-function relationship with a limited part of protein conformation. As shown below, this structural network-based framework combined with a diffusion-based algorithm improves predictions for detecting disease gene products compared with the results from interactome networks.

Let us briefly put the research into a wider perspective. In linkage methods, the first-degree neighbors of disease proteins are considered to be potentially associated with the same disease phenotype [2, 5-8]. Methods based on structural building blocks, such as the identification of disease modules, have also led to the prediction of disease genes. In this approach, all molecules that define functional or disease modules (i.e., highly connected regions of the network) are assumed to have a high probability of being associated with the same disorder. By combining local information derived from pairwise linkage methods and the entire network clusters of disease genes using the disease module assumption in an effort to increase predictive power, diffusion-based algorithms have been proposed and applied to identify complete routes and molecules that are associated with known disease genes. In this framework, dynamic random walkers, starting from the protein products of known disease genes, jump sequentially to neighboring proteins, navigating the entire interactome. The most often visited proteins in the protein-protein interaction network exhibit the highest likelihood of being involved in the same disorder [9, 10]. The biological significance of this apparently random navigation is rooted in the idea that causal molecular pathways tend to overlap with the shortest molecular routes between known disease-associated molecules. This fact is also known as the network parsimony principle [2]. Because the diffusion-based algorithm effectively embeds local and global information simultaneously, it has shown promising results for a variety of diseases, from prostate cancer to Alzheimer' s disease [9, 11], and it has the best predictive performance in comparative studies [9]. Other

studies have based their disease-protein correlation hypotheses on the interactions between protein sequences [12] and functional annotation [13] and use the known disease similarity and human protein interaction networks to perform disease gene prediction [14].

The rapid developments in prediction techniques that prioritize disease gene associations in molecular networks are, however, inevitably limited by the assembled interactome network in terms of data quality, completeness and the biological significance of the interactions among molecules. Here, we structurally scale down the entire interactome by using protein structural information and decomposing the proteins into smaller structural units. Several research programs have aimed at capturing the relationships between protein sequence and structure [15, 16]. Furthermore, these structural units can be encoded into a database, named the Unit of Structural Alphabet DataBase (USA-DB). This encoding can be performed using the protein structure database search tool 3D-BLAST [17-19]. Each USA will then become a node of the newly generated structural interactome network, and each pair of USAs is linked based on their structural similarity. Then, a random walk algorithm is used to prioritize candidate disease molecules, but instead of navigating a protein network as usual, it uses the structural alphabet similarity network in which the nodes are the elementary USAs. The process of walking through the new structural interactome is then used to prioritize the relationships between proteins and diseases, with the aim of identifying disease-causing proteins. The proposed methodology, which is the first to integrate structural protein information with diffusion-based disease gene prioritization algorithms, offers better predictive performance than the use of interactome networks. This finding suggests that the newly assembled structural interactome network containing information on protein structure represents a powerful tool that may aid our efforts to improve the current state of the art of methods for prioritizing candidate disease genes and may eventually elucidate the interplay of multiple molecular components that result in a disease phenotype.

MATERIALS AND METHOD

-Encoding Local Structure

In previous research, we encoded the three-dimensional (3D) structure of a protein into a series of letters, called the SA [17, 18]. The principle is to map the 5-mer local structural segment into corresponding alphabet

characters. The kappa and alpha angles can be calculated for each local structure. The former is the angle of the two connections between the first and third Cα -atoms and between the third and fifth Cα -atoms, and it is used to describe the bending degree of the local structure; the latter is the dihedral angle between the surface formed from the second, third and fourth Cα -atoms and that formed from the third, fourth and fifth C α -atoms. The alpha angle is used to distinguish the chirality of each structural fragment. Different kappa and alpha angles can be used to decide how the local structure can be abbreviated in the SA.

This innovative approach eliminates the need to perform a search for the appropriate residues between Euclidean distances; instead, the SA sequence character by itself can determine the similarity of the protein structure. In addition, we have developed a novel BLOSUM-like substitution matrix, called the structural alphabet substitution matrix, which is used to rapidly search the SA database. The hits are evaluated with the expectation value (E-value) to provide the statistical significance of the protein similarity search. Previous research has shown that the encoding of the local structure based on SA can not only be successfully used in the examination of the structural homology of query proteins [17, 18] but also provide evidence regarding protein function and the classification of well-known protein families [20].

Building the USA Complex Network

The research framework is shown in Figure 1. The steps, in order, are as follows. (1) The protein structures are converted into SA sequences, and the protein is divided into fragments containing USAs. The SA is 23 letters that represent the local structural fragment, which is five residues long and has specific kappa and alpha angles. Based on the SA, a 3D protein structure can be encoded as a 1D sequence [17, 18]. (2) A complex network is assembled in which nodes represent USAs and links indicate structural similarity based on the results of all-against-all USA alignment. (3) The OMIM database is used to compile the associations between human disease phenotypes and proteins [21]. (4) Human disease phenotypes are mapped onto the USA network. (5) The USA complex network is used in the calculation of RWR. The resulting score for each node in the network is used to measure the associations between protein structures and diseases (see also Figures 2 and 3).

3D-BLAST was used to define the characteristics of the 23 representative letters for local protein

segments, which are collectively known as the SA. Subsequent studies then proposed an extended SA fragment that could be applied in the field of network biology. This extended SA fragment was defined as the USA [19].

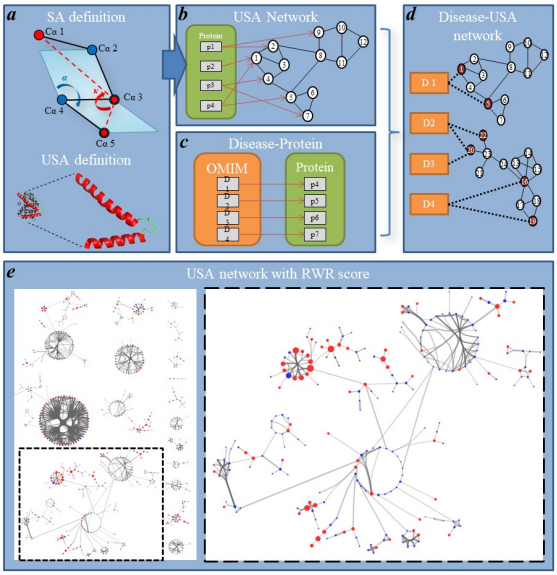

Figure 1. Research framework. (a) Encoding the Local Structure based on the SA definition. The USA is defined as extending the SA in a broader range and including the flexibility and stability of the secondary structure, (b) Assembling the USA complex network, in which nodes represent USAs and links represent structural similarity based on the results of all-against-all USA alignment, (c) Compiling the associations between diseases and proteins from the OMIM database, (d) Mapping both protein-USA and disease-protein relations, (e) The figure illustrates only a fragment of the USA complex network and the scores resulting from the RWR computation. Each node in the graph represents a USA, and the entire network is composed of 1511 nodes (USAs) and 3091 links (structural similarity relationships). The red and blue nodes represent disease-causing and non-disease-causing USAs, respectively. The size of each node represents its RWR score. A connection between two nodes means that the USAs have good structural similarity, and the width of the connection is related to the E-value of the USAs' structural similarity.

Each USA protein structure consists of a sequence of secondary protein structures connected first to a loop structure with no fixed shape and then to the next secondary protein structure. As a result, each USA not only has the structural variability and flexibility of a loop structure but also the stability of a secondary protein structure, as shown in Figure 2.

The USA-DB contains 5525 protein units which are separated from 1603 proteins [19]. In this database, 1603 human protein structures are collected from nr-PDB-50 [22, 23], which has a sequence homology of less than 50% of each proteins. After deriving local protein structure and translating into SA sequence, there is a total of 5525 USAs in this database.

3D-BLAST can then be used to quickly conduct all-to-all structural comparisons for each USA structure and to screen the post-comparison results based on *E*-values. An *E*-value less than 10^{-5} indicates that the USA structures are similar and that the two USA nodes can be linked together. A complex local structural similarity network was established based on all the USA structural similarities [19], as shown in Figure 1(e).

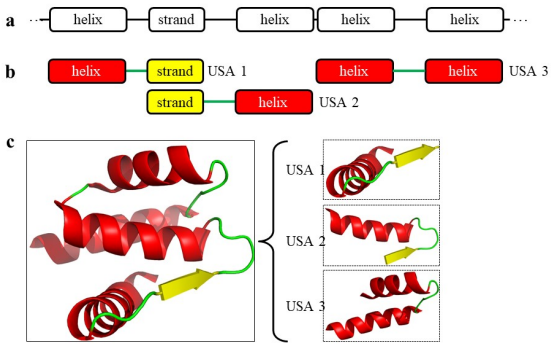

Figure 2. A schematic view of a USA. (a) A protein structure containing multiple secondary structures, such as α -helices and β -strands, (b) Two secondary structures (red) and a partially elastic structure in between (green) form a USA. A protein can contain multiple USAs, (c) As an example, the part of protein 2r0b_A is composed of 4 helices, 1 strand and loops. Based on the definition of a USA, there are 3 USAs in this protein structure.

-Random Walk with Restart

Random walks are a calculation method for determining the importance of each node based on the simulation of random walks throughout the entire network structure [24]. This method simulates the chances of any one node randomly walking to a directly connected node in a network and sorts the probability of each node finding a network node with a higher importance [2].

Other studies have noted that the consistency of protein networks and gene expression networks could reliably predict disease genes. Kohler, Bauer, Horn and Robinson [9], Navlakha and Kingsford [11], and Li and Patra [25] showed how to improve the random walk algorithm and developed the random walk with

restart (RWR) algorithm [26], a calculation network that can be applied to human genes for disease prediction. Their results identified previously unknown associations between genes and certain diseases. These studies noted that genes associated with the same or similar diseases are usually related to their directly connected neighbors within the molecular network. Thus, the current inferred or predicted genes associated with the illness are based on the interaction network and the integration of large amounts of genomic and gene expression data.

In this study, the RWR algorithm was used, but the basis for disease prediction was switched from protein interactions to protein networks with structural similarity. The foundation of the study was based on the alphabet structure of similar protein structures and used the relationship between the protein and diseases to identify possible disease-related proteins.

In the RWR algorithm, $G = (V, E)$ represents a complex USA network, wherein V represents a node in the network, E represents a non-directional connection (structural similarity between USAs), and $\vec{r_i}$ is defined as the probability of a node walking to an adjacent node, as shown in Table 1.

Items	Description
Algorithm	$\overrightarrow{r_{t+1}} = c\widetilde{W}\vec{r_t} + (1-c)\vec{e}$
Input	Similar network $G = (V,E)$, Starting node $\vec{r_0}$, Probability of random walk c, Return probability $(1-c)$
$\vec{r_i}$	Probability matrix for a random walk to a node
$\vec{e}$	Relationship matrix of USAs and diseases
$\widetilde{W}$	Column-normalized adjacency matrix of the graph
α	RWR operational termination conditions
Output	Return probability of each random walking node

Table 1. RWR Definitions

The RWR algorithm calculation steps are as follows:

Step 1: Load $\widetilde{W}$, a complex network matrix established from similar USA structures, to calculate the probability of connecting to neighboring nodes.

Step 2: Load $\vec{e}$, a relationship matrix of USAs and diseases, where 1 indicates a correlation between the node and the disease and 0 indicates no relevance to any disease.

Step 3: Return probability $(1-c)$ representing the random walk process of each node in terms of its chances to return to the starting point. Different return probability settings would lead to different results.

Step 4: In the initial calculation of the RWR algorithm, it substitutes $\overrightarrow{r_0}$ into $\vec{e}$ and then obtains $\overrightarrow{r_1}$.

Step 5: The reiteration of the RWR algorithm, where substituting the i-th times of $\overrightarrow{r_i}$ into the equation will yield $\overrightarrow{r_{i+1}}$. As long as the difference of the two scores, $\overrightarrow{r_{i+1}}$ minus $\overrightarrow{r_i}$, is still greater than α, $\overrightarrow{r_{i+1}}$ can be substituted into the calculation.

Step 6: When $\overrightarrow{r_{i+1}}$ minus $\overrightarrow{r_i}$ is less than the termination condition α, the calculation is ended. According to our observations, in the late RWR computation process, the difference between RWR scores is changed only slightly. Additionally, the difference decreases become stable at approximately 0.01. Therefore, we tentatively set the termination condition α to 0.01.

At this point, $\overrightarrow{r_i}$ represents the probability of each node being passed through during a random walk in a complex network, and this probability can be regarded as a score that signifies the importance of the node. In our USA structural similarity and disease association networks, the $\overrightarrow{r_i}$ score represents the probability of each USA being associated with a disease.

-Encoding Local Structure

We used the recall-precision verification method to determine the disease-causing threshold λ. Precision is the fraction defined by the number of correctly predicted disease-related proteins versus the total number of predictions. Recall (also known as True Positive Rate) refers to the fraction between the number of correctly predicted disease-related proteins and all correct protein– disease associations. This metric indicates the ability of the classifier to identify all the positive proteins in the dataset.

The formula for recall and precision are follows:

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Where TP, FP, TN and FN represent true positive, false positive, true negative, and false negative, respectively.

We further used F1-score to decide the disease-causing threshold λ . The F1-score were calculated using recall and precision with the following equation:

$$F1 = \frac{Recall \times Precision}{Recall + Precision} \times 2 \quad (3)$$

In order to assess the optimal Restart Probability (RP) parameter c, we calculated the ROC curve to test which RP parameter c was suitable for predicting the protein– disease correlation. When we are performing ROC curve calculations, FPR and TPR are represented in X- and Y-axis, respectively. FPR is defined as the ratio of false positive and true negative. TPR is proportion of true positive and false negative. In the ROC curve, we expect these points will be closer to the upper left corner for Ideal results. The FPR and TPR is calculated as:

$$FPR = \frac{FP}{FP + TN} \quad (4)$$

$$TPR = \frac{TP}{TP + FN} \quad (5)$$

In this study, we implemented the Leave One Out Cross Validation (LOOCV) method to were evaluated and verified our prediction performance. LOOCV is a kind of cross-validation which is only one of the samples is used as validating set at one time, while remain datasets are left as training set. This step continues until every single sample is used as a validation dataset, and finally the average model prediction performance is calculated.

RESULTS AND DISCUSSION

Assembling the USA complex network framework and its combination with RWR

In this study, after using USA-DB to establish a complex, local structural similarity network, the Random Walk with Restart (RWR) algorithm is used to make predictions of disease-protein associations. The analysis used LOOCV to find the most appropriate Restart Probability c when calculating RWR and used the recall-precision method to find the threshold λ needed for predicting the likelihood of a protein being disease-related. Finally, we discussed and analyzed the protein-disease association based on the predicted results.

The concept of establishing a significant association between the protein USAs and diseases is an innovative idea and the major conceptual contribution of this study. We needed to determine which USA-encoded protein fragments were associated with diseases in the previously established USA-DB [19]. First, we entered the corresponding 1603 SA protein names of the USA-DB into the OMIM database of human diseases for gene searches using protein names to determine which USAs were associated with a known causative gene. And then, we established an association matrix between the USA and diseases through the USA-disease complex networks. Figure 2 illustrates the definition of a USA.

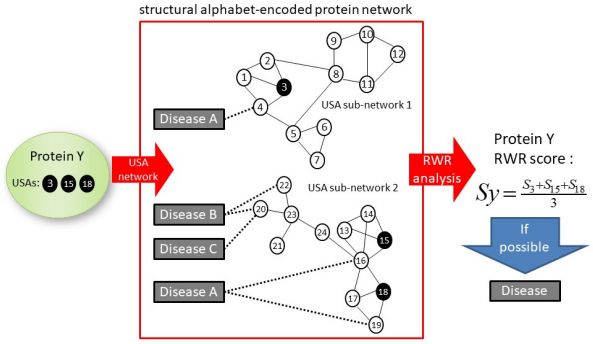

Figure 3. Illustration of the computation. First, protein Y is decomposed into three USAs. A local structural similarity is constructed using all USAs from all protein structures. Known diseases are mapped to USAs. RWR navigates the networks and generates a score for disease prediction for the given protein Y. Note that its score comes from the USAs 3, 15 and 18 that compose protein Y.

Structural protein information and a network-based diffusion algorithm identify protein-disease associations

After using USA-DB to construct a complex local structural similarity-based network, the RWR algorithm is used to prioritize the relationships between proteins and diseases to identify disease-related proteins. The RWR used in this study assumes that nodes are decomposed into USAs from a set of proteins. For example, let us consider a protein (protein Y) that contains three different USAs (nodes 3, 15 and 18 in Figure 3) and that is also associated with a known disease (disease A). We then use the RWR algorithm to iteratively calculate the most often visited USA on the network, which will exhibit the highest likelihood (or score) of being involved in the same disorder. Finally, each node in the complex network is assigned a score that represents the degree of association between each USA and the disease due to their structural similarity

(Figure 3). If the calculated probability of a node is higher than a certain threshold, then it can be assumed that the protein to which the USA belongs is likely to have relevance to the disease and may even be used to estimate the cause of the disease if it occurred in the position of an amino acid of the USA.

To determine the optimal Restart Probability (RP) parameter *c*, we calculated the scores of nodes when the RP parameter c values ranged from 0.1 to 0.9 and used LOOCV results and the ROC curve to test which RP parameter c was best for predicting the protein– disease correlation. In terms of error value, when the c value was greater than 0.7, random walk results would not exhibit significant changes. Thus, we then used the ROC curve to determine the merits of c values between 0.7 and 0.9. Our results show that c values at 0.7 constituted the largest area under the curve, indicating that the restart probability *c* equal to 0.7 had the best performance, as shown in Figure 4.

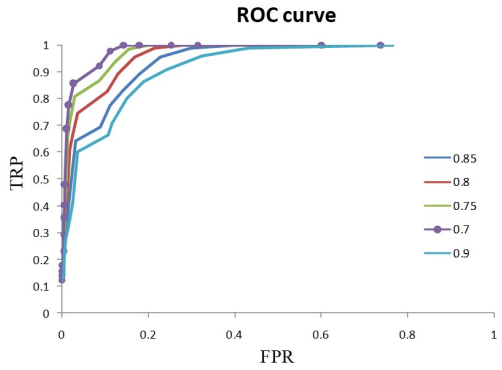

Figure 4. Optimal Parameter Analysis of the Restart Probability c.

-Determination of the disease-causing threshold λ

After determining the c value of the random walk, the threshold value λ for protein and disease relevance could be determined. When the average RWR score of a protein was greater than the threshold value, it could be assumed that such a protein might cause a disease. A score below the threshold value indicates a lower likelihood of causing the disease. We used the recall-precision verification method to determine the threshold value λ . The optimal threshold value λ corresponded to the intersecting value of recall-precision, and this result was between 0.4 and 0.5. Then, we used the F1-score for further evaluation, and the results showed that forecasts made when the threshold λ was at 0.45 were the most accurate. The F1-score was

0.8874, which meant that using 0.45 as the threshold value for predictions would yield accuracy of nearly 90%, as shown in Figure 5.

Experiments on prediction accuracy outperform previous results based on protein interaction networks

We obtained the interaction information for the 823 proteins used in our test data from the STRING protein interaction database [27, 28]. After constructing a protein interaction network, we also used the RWR algorithm to calculate and verify the prediction accuracy. As shown in Figure 5, the network established based on this protein interaction information had a lower accuracy after the RWR calculations, and the best F1-score was only 0.8474. This result showed that adding 3D protein structural information to the calculations for predicting the likelihood of disease-causing proteins provides predictions that are significantly more accurate than those using the traditional binary protein-protein interaction information.

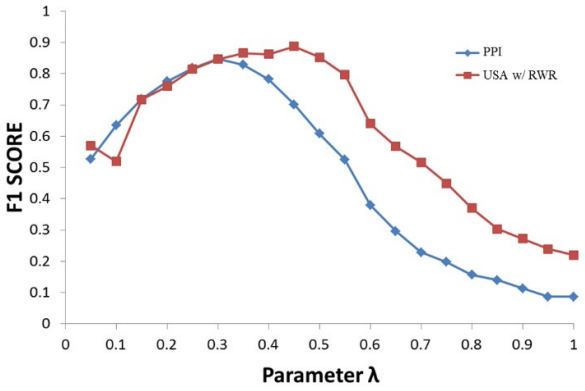

Figure 5. Prediction results of the F1-score graph. Prediction made at a threshold value λ of 0.45 were the most accurate, and the optimal F1-score was 0.8874. Networks constructed using traditional protein-protein interactions, after RWR verification, had an optimal F1-score of 0.8474, which was lower than that of our proposed method.

-False Negative Predictions

The complete prediction results obtained from the 1511 RWR scores used in this research are shown in Appendix section for table of results. Regarding known disease-related proteins that we predicted to be non -disease-related, we investigated these predictions further from a USA network perspective. Figure 6 shows that the incorrectly predicted proteins were all located at the boundary of the sub-network, with a degree of protein coupling of 1. In addition, the directly connected protein coupling in the sub-network was the highest,

and this phenomenon led to a reduced return probability, a low RWR score, and finally a prediction against the observed data. For example, node 837 (PDB ID: 2klz) in Figure 6(a) is located at the boundary of the network. Its degree is one, being connected only to a neighbor (node 1376) that is a high degree hub. Because of this topological reason, the return probability of node 837 is decreased when RWR navigates the entire USA network. In other words, node 837 has only one friend, which is very popular. Thus, node 837 will often feel excluded from the main group.

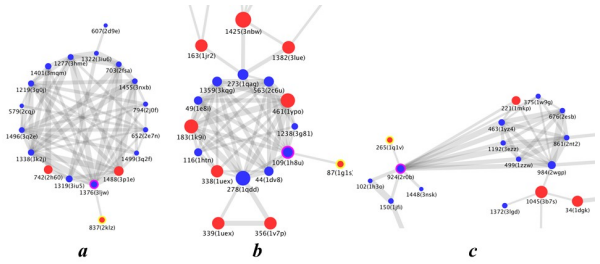

Figure 6. False negative predictions resulting from USA subnet maps. (a) A blue node with a purple border, 1376 is a hub node directly connected to USA number 837 (a red node with a yellow border), which is part of the protein 2klz, (b) The purple circle node 109 is a hub node directly connected to node 87 (PDB ID: 1g1s), (c) Node 265 belongs to the protein 1q1v and is connected to the hub node 924.

-False Positive Predictions (Case Study of PDB ID: 1qdd)

Conversely, we also observed USA cases in which a non-disease-related protein was predicted to cause disease. A common feature for these cases is that the USA structure is similar to that of the entire protein. Figure 7 shows the protein 1qdd, which is directly connected to structurally similar proteins with disease-causing likelihood. These proteins are 1ypo, 1uex, 1v7p and 1k9i, as shown in Figure 7. The structurally similar USA was located at the surface of each protein. For the structures of these USAs, whether they had α -helix similarity or an α -helix structure containing non-specific fragments, the bend and flip angles were very similar. Moreover, when we used the structural classification of proteins (SCOP) database [29, 30] to query each protein, we discovered that they were classified in the same domain (d.19.1.1), which suggests that our method could potentially be used to find protein blocks with similar structural functions. This finding indicates that the original non-disease-related protein known as protein 1qdd might in fact be a disease-

related protein, as predicted using our methods. By comparing the structural similarity to the already well-known disease-related proteins 1ypo, 1uex, 1v7p and 1k9i, we could assume that the key structure that makes protein 1qdd cause disease is that of the USA.

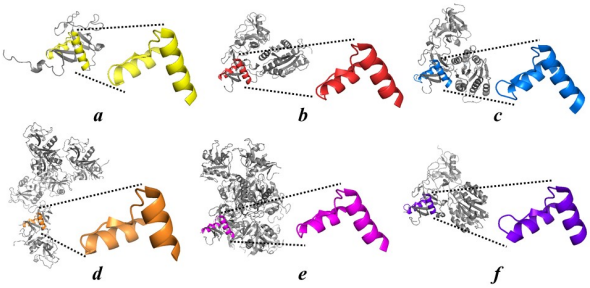

Figure 7. False positive predictions of the USA structural similarity network, considering protein 1qdd as an example. (a) The yellow portion of protein 1qdd is defined as a USA, (b) and (c) For the protein 1uex, the red and blue parts represent the two defined USAs, (d) The orange portion of the protein 1ypo is also a USA, (e) The pink portion of the protein 1k9i is the defined USA, (f) The purple portion of the protein 1v7p is the defined USA.

-False Positive Predictions (Case Study of PDB ID: 2fv7)

The overall structure of the protein 2fv7 was not structurally similar to the structures of the directly connected possibly disease-related proteins 2dae, 1j47, 2pe4, 2bxg and 1kw2, but the structures of the USAs, by our definition, were very similar in terms of the rotation or flip angle of the α -helix fragment or α -helix loop structure. Because the defining USA structure of the protein was also located at the surface of the protein structure, that structure could be deduced from the USA definition and RWR calculations at the said USA, and the protein to which it belonged, 2fv7, may cause disease.

In addition, we observed that the protein in the USA 2fv7 directly connected to a number of complex networks, including 613, 146, 902, 555, 187, 546, 188, 550, and 186, as shown in Table 2. The highest RWR scores of USA No. 555 were found through later screening points, and thus we might have inferred that the proteins 2fv7 and 2bxg were associated with similar diseases. Because our results indicate that the serum synthesis of protein 2bxg may cause damage that results in metabolic defects, the above reasoning, in addition to speculation that the protein 2fv7 may cause disease, appears to suggest that 2fv7 may lead to the same disease as that caused by the protein 2bxg.

The RWR score in this study indicates the level of correlation between the USA and the disease. If USAs have a significant relationship with one another, then they may have the same function. From this perspective, we could use the RWR scores to further infer what type of disease might be caused by a candidate disease-related protein.

	USA id	PDB id	Gene name	RWR score
Prediction	706	2fv7	RBKS	0.6779
Neighbors	613	2dae	TAB2	0.3869
	146	1j47	SRY	0.6016
	902	2pe4	HYAL1	0.9108
	555	2bxg	ALB	1.4451
	187	1kw2	GC	0.8472
	546	2bxg	ALB	1.0064
	188	1kw2	GC	0.9229
	550	2bxg	ALB	1.3051
	186	1kw2	GC	0.7050

Table 2. Predicting Disease Results of a USA (#706)

CONCLUSION

This work used the structurally similar USA network and a diffusion-based algorithm to predict whether proteins have disease-causing relationships. According our previous works, SA one dimension sequence derived from protein could be represented into a 3D structural shape. Therefore, it was used to develop the web service of structure database search [17, 18]. Moreover, we have established an automated server for identifying the protein domains and superfamilies of query structures [20]. Here, we expanded the application of SA to describe the local structure and elucidate the relationship between protein structure and disease from the point of view of network biology. The integration of both fields, network biology and structural biology, is therefore, one of the contributions of this work.

This study also provides a methodology for identifying proteins that, although classified as non-disease-related, could have a disease association based on the scores predicted by our methodology. The assembled USA network offers a fresh view that may assist to overcome the limitations of the currently available interactome maps, for which more than 80% of pairwise interactions (links) are missing, and offers a new avenue for network-based disease gene prioritization studies by considering the tertiary structure in the assembled network. Our work should be considered as a complementary study to those already successful network-based methodologies and offers a new perspective to address network medicine challenges [2].

The proposed USA network-based methodology is the first to integrate structural protein information with diffusion-based disease gene prioritization algorithms. Therefore, our work addresses an important question regarding the nature of the structural networks used by diffusion algorithms in their navigation. The network nodes are, instead of proteins as typically used in the literature, structural fragments of proteins, representing a huge conceptual and technical difference that also leads to better performance. Indeed, the results show that this approach offers the best predictive performance compared with methods in which tertiary protein structure was not included. The nodes in this complex network do not represent the entire protein abstractly but instead represent specific structural fragments with requirements for the stability of the secondary structure. However, there is still room for improvement and further exploitation of the presented methodology in both theoretical and experimental areas. For example, we predicted various proteins as likely to cause disease, but they have not yet been classified as disease-related proteins. Therefore, some of our predictions require further experimental confirmation. In addition, although we can predict whether a protein is disease-related, the method cannot identify the cause and structural relevance of a given USA in the overall disease-causing effect. It also does not provide information about whether a specific USA is responsible for the disease-causing activity of the protein. The method also cannot predict the possible specific human disorder a protein might cause. To predict a specific disease for a protein, a procedure should be added to annotate the disease that is linked to a high-scoring USA node. By doing so for every node, we could obtain a weighted proportion of each disorder for a specific node and finally properly assess the human disorder.

From the connection degree of nodes in the network, connectivity is not always positively correlated to pathogenicity. A possible reason is that when a node has more connections in the network graph, the options to choose a connected path for the random walk is increased. Hence, the probability of the random walk is reduced. In previous studies, it was pointed out that when a join weighting-adjustment scheme and a set of nodes defined by the node to its neighbors and neighbors' neighbors are considered, the recalculated RWR scores would lead to positive effects and could be identify proteins associated to multiple human disorders [31].

It should be noted that a single protein or structurally similar proteins can be responsible for multiple related to multiple human disorders, rather than single disorders, in completely sporadic events. To properly evaluate these cases, additional analyses are encouraged and left as future work.

In the future, we could expand this research to investigate the relevance among protein structures, disorders, drugs and side effects such that we could perform a comprehensive analysis based on the relationship between each major aspect that plays a role in the disease-therapy process [3] using the proposed networks of USA complexes. Such networks could certainly be helpful for practical clinical applications. For example, when a protein structure with an unknown function is considered, this method could not only predict what might cause a human disorder but also assist in estimating a suitable set of drugs for this particular disorder, along with its side effects. When researchers design a new drug molecule, our methodology could explore its associated molecular target and the possible side effects of the treatment.

REFERENCES

[1] Hirschhorn, J. N. 2009. Genomewide association studies--illuminating biologic pathways. N Engl J Med, 360, 17 (Apr. 2009), 1699-1701. DOI=http://dx.doi.org/10.1056/NEJMp0808934.

[2] Barabasi, A. L., Gulbahce, N. and Loscalzo, J. 2011. Network medicine: a network-based approach to human disease. Nat Rev Genet, 12, 1 (Jan. 2011), 56-68. DOI=http://dx.doi.org/10.1038/nrg2918.

[3] Csermely, P., Korcsmaros, T., Kiss, H. J., London, G. and Nussinov, R. 2013. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacol Ther, 138, 3 (Jun. 2013), 333-408. DOI=http://dx.doi.org/10.1016/j.pharmthera.2013.01.016.

[4] Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J. and Barabasi, A. L. 2015. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science, 347, 6224 (Feb. 2015), 1257601. DOI=http://dx.doi.org/10.1126/science.1257601.

[5] Krauthammer, M., Kaufmann, C. A., Gilliam, T. C. and Rzhetsky, A. 2004. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. Proc Natl Acad Sci U S A, 101, 42 (Oct. 2004), 15148-15153. DOI=http://dx.doi.org/10.1073/pnas.0404315101.

[6] Franke, L., van Bakel, H., Fokkens, L., de Jong, E. D., Egmont-Petersen, M. and Wijmenga, C. 2006. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet, 78, 6 (Jun. 2006), 1011-1025. DOI=http://dx.doi.org/10.1086/504300.

[7] Oti, M., Snel, B., Huynen, M. A. and Brunner, H. G. 2006. Predicting disease genes using protein-protein interactions. J Med Genet, 43, 8 (Aug. 2006), 691-698. DOI=http://dx.doi.org/10.1136/jmg.2006.041376.

[8] Iossifov, I., Zheng, T., Baron, M., Gilliam, T. C. and Rzhetsky, A. 2008. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. Genome Res, 18, 7 (Jul. 2008), 1150-1162. DOI=http://dx.doi.org/10.1101/gr.075622.107.

[9] Kohler, S., Bauer, S., Horn, D. and Robinson, P. N. 2008. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet, 82, 4 (Apr. 2008), 949-958. DOI=http://dx.doi.org/10.1016/j.ajhg.2008.02.013.

[10] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. and Sharan, R. 2010. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol, 6, 1 (Jan. 2010), e1000641. DOI=http://dx.doi.org/10.1371/journal.pcbi.1000641.

[11] Navlakha, S. and Kingsford, C. 2010. The power of protein interaction networks for associating genes with diseases. Bioinformatics, 26, 8 (Apr. 2010), 1057-1063. DOI=http://dx.doi.org/10.1093/bioinformatics/btq076.

[12] George, R. A., Liu, J. Y., Feng, L. L., Bryson-Richardson, R. J., Fatkin, D. and Wouters, M. A. 2006. Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic Acids Res, 34, 19 2006), e130. DOI=http://dx.doi.org/10.1093/nar/gkl707.

[13] Perez-Iratxeta, C., Bork, P. and Andrade-Navarro, M. A. 2007. Update of the G2D tool for prioritization of gene candidates to inherited diseases. Nucleic Acids Res, 35, Web Server issue (Jul. 2007), W212-216. DOI=http://dx.doi.org/10.1093/nar/gkm223.

[14] Wu, X., Liu, Q. and Jiang, R. 2009. Align human interactome with phenome to identify causative genes and networks underlying disease families. Bioinformatics, 25, 1 (Jan. 2009), 98-104. DOI=http://dx.doi.org/10.1093/bioinformatics/btn593.

[15] Tyagi, M., de Brevern, A. G., Srinivasan, N. and Offmann, B. 2008. Protein structure mining using a structural alphabet. Proteins, 71, 2 (May. 2008), 920-937. DOI=http://dx.doi.org/10.1002/prot.21776.

[16] Mahajan, S., de Brevern, A. G., Sanejouand, Y. H., Srinivasan, N. and Offmann, B. 2015. Use of a structural alphabet to find compatible folds for amino acid sequences. Protein Sci, 24, 1 (Jan. 2015), 145-153. DOI=http://dx.doi.org/10.1002/pro.2581.

[17] Yang, J. M. and Tung, C. H. 2006. Protein structure database search and evolutionary classification. Nucleic Acids Res, 34, 13 2006), 3646-3659. DOI=http://dx.doi.org/10.1093/nar/gkl395.

[18] Tung, C. H., Huang, J. W. and Yang, J. M. 2007. Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. Genome Biol, 8, 3 2007), R31. DOI=http://dx.doi.org/10.1186/gb-2007-8-3-r31.

[19] Tung, C. H. and Nacher, J. C. 2013. A Complex Network Approach for the Analysis of Protein Units Similarity Using Structural Alphabet. International Journal of Bioscience, Biochemistry and Bioinformatics, 32013), 433-437. DOI=http://dx.doi.org/10.7763/IJBBB.2013.V3.250.

[20] Tung, C. H. and Yang, J. M. 2007. fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies. Nucleic Acids Res, 35, Web Server issue (Jul. 2007), W438-443. DOI=http://dx.doi.org/10.1093/nar/gkm288.

[21] Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. and Hamosh, A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res, 43, Database issue (Jan. 2015), D789-798. DOI=http://dx.doi.org/10.1093/nar/gku1205.

[22] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. 2000. The Protein Data Bank. Nucleic Acids Res, 28, 1 (Jan. 2000), 235-242. DOI=http://dx.doi.org/10.1093/nar/28.1.235

[23] Rose, P. W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., Costanzo, L. D., Duarte, J. M., Dutta, S., Feng, Z., Green, R. K., Goodsell, D. S., Hudson, B., Kalro, T., Lowe, R., Peisach, E., Randle, C., Rose, A. S., Shao, C., Tao, Y. P., Valasatava, Y., Voigt, M., Westbrook, J. D., Woo, J., Yang, H., Young, J. Y., Zardecki, C., Berman, H. M. and Burley, S. K. 2017. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res, 45, D1 (Jan. 2017), D271-D281. DOI=http://dx.doi.org/10.1093/nar/gkw1000.

[24] Burioni, R. and Cassi, D. 2005. Random walks on graphs: ideas, techniques and results. Journal of Physics A: Mathematical and General, 38, 8 (Feb. 2005), R45-R78. DOI=http://dx.doi.org/10.1088/0305-4470/38/8/r01.

[25] Li, Y. and Patra, J. C. 2010. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics, 26, 9 (May. 2010), 1219-1224. DOI=http://dx.doi.org/10.1093/bioinformatics/btq108.

[26] Tong, H., Faloutsos, C. and Pan, J.-Y. 2007. Random walk with restart: fast solutions and applications. Knowledge and Information Systems, 14, 3 (Mar. 2007), 327-346. DOI=http://dx.doi.org/10.1007/s10115-007-0094-2.

[27] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J. and von Mering, C. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res, 39, Database issue (Jan. 2011), D561-568. DOI=http://dx.doi.org/10.1093/nar/gkq973.

[28] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C. and Jensen, L. J. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res, 41, Database issue (Jan. 2013), D808-815. DOI=http://dx.doi.org/10.1093/nar/gks1094.

[29] Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol, 247, 4 (Apr. 1995), 536-540. DOI=http://dx.doi.org/10.1006/jmbi.1995.0159.

[30] Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C. and Murzin, A. G. 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res, 36, Database issue (Jan. 2008), D419-425. DOI=http://dx.doi.org/10.1093/nar/gkm993.

[31] Le, D. H. and Kwon, Y. K. 2013. Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization. Comput Biol Chem, 44(Jun. 2013), 1-8. DOI=http://dx.doi.org/10.1016/j.compbiolchem.2013.01.001.

ACKNOWLEDGMENT

This research was supported by Ministry of Science and Technology, Taiwan, R.O.C. under grants MOST-105-2221-E-216-021 and NSC-101-2311-B-216-001.