A Complex Network Approach for the Analysis of Protein Units Similarity Using Structural Alphabet

Chi-Hua Tung and Jose C. Nacher

***Abstract*—In this paper we present a network approach based on the recent developed 3D-BLAST method of rapid protein structure search. We defined new local segments that represent structural feature of proteins named units of structural alphabet (USA). Each USA is composed of two protein secondary structures, and one loop located between these two secondary structures. We performed all-against-all structural comparison of USA and recognized the USA-based similarity network. The analytical result shows that the network with a power degree distribution is called scale free. These results not only suggest the existence of organizing principles in the local protein structure but also allow us to identify potential key fragments that could be useful for future new drug development and design.**

***Index Terms*—Local structure similarity network, network biology, protein modularity.**

I. Introduction

In the past few decades, genomics (DNA sequences), structural genomics (protein structures), and proteomics (protein expression and interactions) have rapidly enhanced knowledge on biological functions and systems. With structural models developed using genome-wide investigative strategies [1]–[3], the number of protein structures in the Protein Data Bank (PDB) has rapidly increased. By Dec. 25, 2012, there were already more than 87,090 known protein structures [4]. The increasing number of known protein structures with unknown/unassigned functions emphasizes the demand for effective bioinformatics methods for annotating the structural homology or evolutionary family of proteins and inferring their cellular functions.

The comparison and analysis of the relationship between new protein structures with unclear functions and well-known structures seeks to bridge the protein structure–function research gap. Given a query protein structure, we may search through the database and report similar protein structures. However, unlike one-dimensional sequence comparison, structural alignment for determining similarities is much more complex and computationally expensive. Some methods can be used for efficient pair-wise structural comparison [5], but these methods entail an exhaustive search to compare the query structure against all protein structures in the database.

To bridge the current protein structure–function research gap and address anterior questions, many approaches have been proposed for encoding 3D local structural fragments based on Cartesian coordinates into a one-dimensional representation using several letters called the structural alphabet [6]–[13]. The structural alphabet represents advantageous local structures and has been used to 1) compare/analyze 3D structures [14]–[16], 2) predict protein 3D structures from amino acid sequences [6], [9], 3) reconstruct protein backbones [11], and 4) model loops [17]. In addition, given that local structures are generally more evolutionary conserved than amino acid sequences, a series of research has been developed to explore protein structures [18]. The structural alphabet theory has already been utilized to compare protein structures, search for homologs, and assign protein families [19]-[21].

Earlier, we developed the kappa-alpha (κ, α) plot derived structural alphabet and a novel BLOSUM-like substitution matrix, called structural alphabet substitution matrix (SASM), which searches through the structural alphabet database (SADB). This structural alphabet was used in developing the fast structure database search method called 3D-BLAST, which is as fast as BLAST [22] and provides the statistical significance (*E*-value) of an alignment, indicating the reliability of a hit protein structure [19], [20]. Moreover, we developed an automated server called fastSCOP [21] for integrating a fast structure database search tool (3D-BLAST) and a detailed structural comparison tool, as well as for recognizing the SCOP domains and SCOP superfamilies of query structures [23].

Structural networks with complex topology are common in nature. Numerous network biology researchers have demonstrated that networks in many biological systems can be characterized [24]. Biological networks observed in epidemiology, metabolic pathways, gene regulation, protein domain interaction, drug–target binding, and protein structures have some similar topological properties [24]–[29]. In these networks, most nodes have only a few links, and a disproportionate number of nodes have high connections. Networks characterized by power-law degree distribution are called scale free [30]. Furthermore, the clustering coefficient of hierarchical modularity in the metabolic networks of 43 distinct organisms follows power-law scaling [27].

Protein fold and functional site similarity networks provide evidence of protein evolution and help in structure-based functional annotation [31], [32]. Moreover, one kind of structural similarity network was proposed a

Manuscript received January 15, 2013; revised March 15, 2013. This work was supported in part by the National Science Council (NSC), Taiwan, under Contract of NSC 101-2311-B-216-001 and 101-2221-E-216-041.

Chi-Hua Tung is with the Department of Bioinformatics, Chung-Hua University, Hsinchu 300, Taiwan (e-mail: chihua.tung@chu.edu.tw).

Jose C. Nacher is with Department of Information Science, Faculty of Science, Toho University, Miyama 2-2-1, Funabashi, Chiba 274-8510, Japan (e-mail: nacher@is.sci.toho-u.ac.jp).

framework for classifying the structures of protein segments and analyzing whether the degree distribution of this network obeys the power law. Proteins are divided according to their local structures using the specific length of sliding windows. The distribution of the structural diversity of local protein structures also shows a power-law property. However, the local structures of proteins, which consist of consecutive fixed numbers of amino acids, are not used for generating information on typical secondary structures [33].

II. Purposes and Major Claims

Only a small number of residues are often conserved in the functional active sites or binding regions of proteins with similar functions. Therefore, in this study, we look deeply into the core of proteins and evaluate their basic unit. Proteins are then divided into various fragments based on the location of secondary structures and loops. Moreover, similarities in the local structures of fragments are analyzed to acquire insights for bridging the protein structure–function research gap.

We develop a novel network biology approach based on the recently developed 3D-BLAST method of protein structure identification. With this method and using tertiary protein structures, we can conduct a fast protein similarity search and identify 23 states of structural alphabet (SA) sequences that represent the local structures of protein backbones. Additionally, we define new fragments that can describe local structural features called units of structural alphabet (USA). Each USA is composed of two secondary structures and one loop.

Subsequently, we develop a complex structural similarity network based on USAs and assess its degree distribution. All-against-all alignment of USA sequences is utilized to determine structural similarity. In our similarity network, each USA is taken as a node, and alignment is represented by the link between two USAs with similar structure. After building the complex network, we characterize its topological properties and determine whether it follows power-law degree distribution and is therefore scale free.

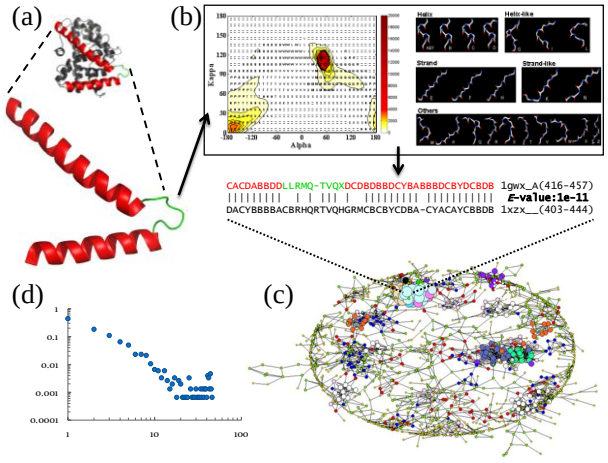

Fig. 1. Research methodology.

III. Materials and Methods

Fig. 1 illustrates this study's methodology. Every protein structure can be divided into USAs composed of two secondary protein structures and one loop located between these two secondary structures (Fig. 1 (a)). After determining USAs, protein units are translated into encoded SA sequences according to the kappa and alpha map [19], [20] (Fig. 1 (b)). A complex network is obtained, with nodes representing USAs and links representing structural similarity based on the results of all-against-all USA alignment (Fig. 1 (c)). Furthermore, the topological properties of the similarity network are analyzed to determine whether the network is scale free (Fig. 1 (d)).

Fig. 2 shows the flow of this study. First, a testing set is prepared from nr-PDB-50 dated April 8, 2011 (Fig. 2 (a)); only proteins from the source species *Homo sapiens* are selected (Fig. 2 (b)). Second, each protein structure is translated into SA sequences. Overall, 1603 proteins with SA encoding are included in the testing set called SADB-nrPDB50-HUMAN (Fig. 2 (c)). Third, protein chains chains are divided into many USAs with various kappa and alpha angles (Fig. 2 (d)), leading to a USA database with 5525 protein units (Fig. 2 (e)). Fourth, 3D-BLAST is used to search and align rapidly every USA against the whole database (Fig. 2 (f)). Based on E-values in alignment results, the USA-based similarity network is developed (Fig. 2 (g)). Finally, the characteristics and properties of this network are analyzed (Fig. 2 (h)).

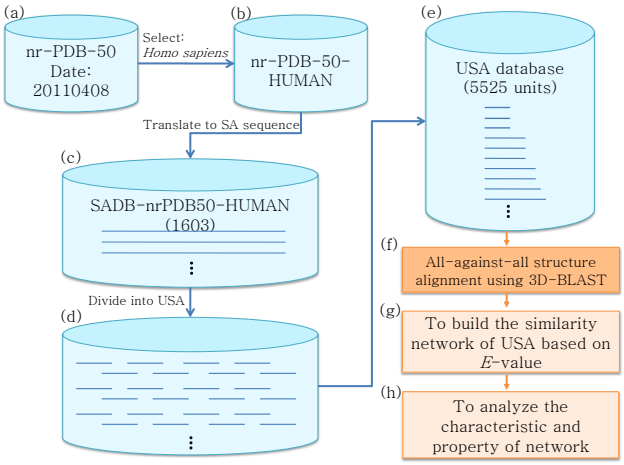

Fig. 2. Research flow.

A. Preparation

The date of PDB used as testing set is April 8, 2011. The testing set is collected based on certain principles. First, the selected database is nr-PDB-50, in which the sequence identities are lower than 50% among proteins. In addition, the species of protein are only chosen from Homo sapiens. Second, the length of each protein chain must be longer than 15 residues. Finally, each protein chain must have at least one USA. A total of 1603 protein chains are included in the testing data set.

After translating all structures in the testing data set into SA sequences, the USAs are divided based on the location of secondary structures and loops. The determination of USA is explained further in the next section. A total of 5525 protein units are obtained from 1603 proteins.

In this study, the unit of protein includes both secondary structures and random coils. These novel protein units can

maintain not only the flexibility of variable loops but also the stability of secondary structures. Fig. 3 demonstrates the USA in one protein and its SA sequence. This protein with chain named 1gwx_A belongs to one kind of all-helix proteins classified in the SCOP database [23]. It has 9 USAs shown as a short loop (green color) between two helical structures (red color).

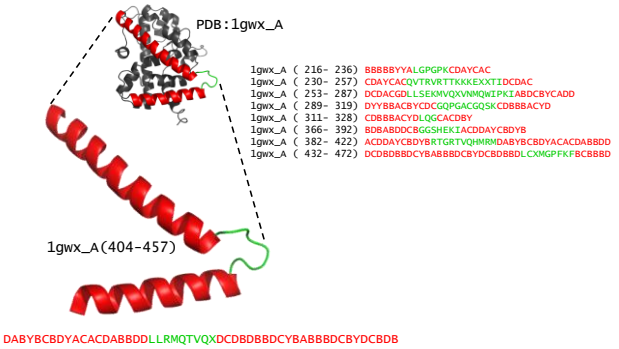

Fig. 3. USAs in protein 1gwx_A and its SA sequence.

B. Construction of the USA-Based Similarity Network Using E-Values

We use 3D-BLAST to align all 10291 USAs against all USAs. In 3D-BLAST results, *E*-values indicate the degree of similarity between query USAs and subject USAs. An *E*-value lower than the threshold suggests that the given two USAs are conformationally similar. Based on the results of all-against-all USA comparison, we can find the homology similarity among all USAs and thus build the similarity network. In this network, each node represents one USA and each link between nodes represents a homology relationship.

We use two kinds of *E*-values to determine homology relationships. The first kind considers whole-structure similarity between USAs, and the second kind called E^{loop}-value measures the conformation of variable loops in a very specific way. The threshold *E*-value, which is used to determine if two structures are homologous, has been evaluated in previous studies [19], [20]. This threshold value is set at 10^{-15}. However, the length of USA is usually smaller than that of the whole protein. Hence, the original *E*-value is not suitable for determining whether USAs are homologous. We try different threshold values to decide which is appropriate for determining homologs.

Furthermore, we modify the parameter of original *E*-values and re-compute *E*-values only in loop structures because if two USAs with long secondary structures align to each other, the resulting *E*-value becomes insignificant. In this situation, the alignment score for two secondary structures is high. Even if the two USAs are totally dissimilar, the *E*-value is still lower than the threshold.

To avoid the foregoing problem, we focus only on loop conformation and consider the score in the loop to modify *E*-values. We re-compute for the E^{loop}-value instead of using the original *E*-value. The E^{loop}-value is given as

$$E^{loop} = mn2^{-S} \tag{1}$$

m is total number of SA within loop coding, *n* is the length of alignment only in the loop region, and *S* is the bit score in the loop region. In our database, the total number (*m*) of SA within loop coding is 111922. Finally, the threshold *E*-value is set at 10^{-5} and the E^{loop}-value is set at 5.0.

C. Analysis of Network Characteristics and Properties

In this study, we mainly measure two quantifiable descriptions of complex networks: the power-law degree distribution and the clustering coefficient. Most biological networks are scale free. Their degree distribution approximates the power law, $P(k) \sim k^{-\gamma}$. Degree distribution, $P(k)$, is the probability of nodes with exactly k links, and γ is the degree exponent with a value usually between 2 and 3. In a network with a degree distribution following power law, the highly connected node is in very small fraction. Conversely, most nodes are only linked to a few neighbors.

Another quantifiable characteristic description is the clustering coefficient. The function $C(k)$ is defined as the average clustering coefficient over nodes with the same node degree k. The clustering coefficient can be defined for each node I as:

$$C_I(k_I) = 2n_I / k_I(k_I - 1) \tag{2}$$

where n_I is the number of links connecting k_I neighbors of node I to each other [24].

In hierarchical networks, the distribution of clustering coefficient, which follows $C(k) \sim k^{-1}$, is a straight line with a slope equals -1 on a log-log plot. The hierarchical network is one kind of a scale-free network. Unlike traditional scale free networks, a hierarchical architecture implies a central node connected to one or more other nodes that are two levels lower in the hierarchy with a link between each of the second-level nodes and the central node. Meanwhile, each of the second-level nodes that are connected to the central node also have one or more other nodes that are three levels lower in the hierarchy connected to it [24].

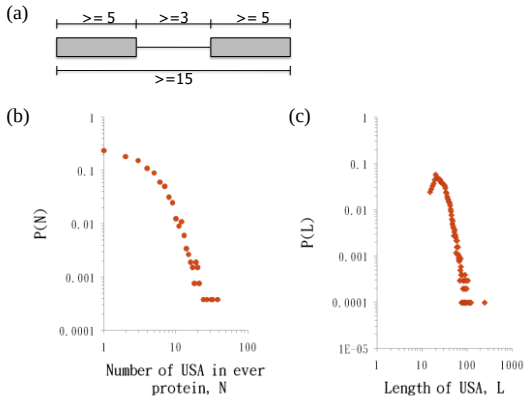

Fig. 4 (a) Criteria for the length of secondary structures, loops, and whole USAs. (b) Distribution of the number of USAs in each protein. (c) Distribution of length of each USA.

IV. Results and Discussion

A. Definition of USA

We tested various parameters for the length of secondary structures, loops, and whole USAs. The best results are shown in Fig. 4. The length of secondary structures must be

$\geqq 5$ residues, the limitation of loop length is set at $\geqq 3$ residues, and the total USA length must be $\geqq 15$ residues (Fig. 4a). These criteria are used for filtering short USAs because USAs smaller than 15 residues are not reliable for comparing conformation. Moreover, very short secondary structures and loops may lack structural information and biological meaning.

B. Distribution of USA Database

Fig. 4 (b) and Fig. 4 (c) present the distribution of the number of USAs in each protein and the length of each USA, respectively. The curves of both distributions are smooth in the log-log plot, suggesting that the set criteria satisfy the nature of local protein structures. Fig. 4 (b) shows that the highest number of USAs in a protein is 38. In addition, 627 proteins have only one USA, and $P(N=1)$ is 0.2374. Fig. 4c shows that the length of the longest USA is 247 residues, and the biggest $P(L)$ is 0.0578 (L=20).

C. USA-Based Similarity Network

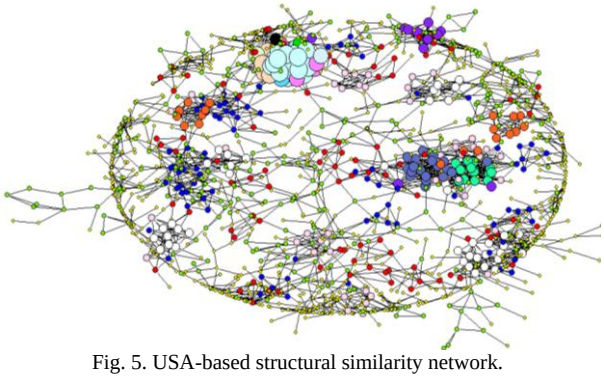

Fig. 5. USA-based structural similarity network.

Fig. 5 illustrates the USA-based structural similarity network. This figure is drawn using the software Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). In the network, every spot is a node representing one USA. The size and color of nodes are allocated for degree. The degree is higher, the size is bigger. And, the same number of degree would show as the same color. Two nodes are connected by an edge, and they are considered of similar structure if their E-value of alignment is less than 10^{-5} and E^{loop}-value is less than 5.0. The structural similarity network contains 1511 nodes with at least one neighbor.

D. Network Characteristics and Properties

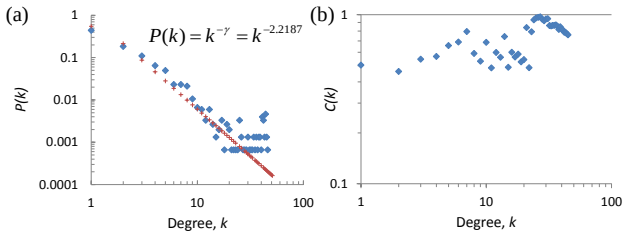

Fig. 6 (a) Log-log plot of the degree distribution of network and (b) clustering coefficient distribution of the USA-based similarity network.

We determined if our novel network is scale free and even hierarchical. We first analyze the degree distribution of the network. Fig. 6 (a) presents that log-log plot of the distribution. The network is approximately characterized by power law, where $P(k) \sim k^{-2.2187}$. Thus, there is no doubt that the USA-based structural similarity network is scale free. In addition, the highest degree of USA is 51, and $P(k=1)$ is 0.4421. However, the evaluation result of clustering coefficient in Fig. 6 (b) points out that $C(k)$ is independent of degree in our network. Therefore, the USA-based similarity network is scale free without hierarchical modularity.

V. Conclusion

In this study, we develop a novel local structural fragment called USA to describe unique features of the functional sites of protein structures. We extend the structural alphabet research by integrating another totally different research field, complex networks. Previous studies have proven that SA is robust and reliable for representing protein structures. Thus, we further use SA in describing local structures and designing USA. Moreover, we use 3D-BLAST to search for USA homologs rapidly and build our proposed similarity network.

Our structural similarity network is constructed using knowledge of complex networks. In addition, the analysis of the characteristics and behavior of the similarity network is based on the complex network literature. Results show that there is a highly uneven degree of distribution in the USA-based similarity network. Highly connected USAs, which are called hubs, constitute a small fraction of all USAs. In other words, the probability of having USAs with only a small number of neighbors is usually high.

We have combined two distinct research fields and provided a new and alternative viewpoint for investigating the relationship between protein structures and functions. Our approach constructs complex networks based on protein fragments. Our final goal is to use this network structure to predict novel target binding sites for drugs and also to aid in the development of new drugs to target specific binding sites. We could then compare our predictions with different methods, but currently, there is no point to compare the statistical results of the computed networks with any other approach.

In the future, we can further utilize USAs in drug development and design. We will identify possible key fragments that may be useful for new drug development and design. Drug-related databases, such as PDTD [34] and DrugBank [35], may be used to identify potential USAs in the set of known drug protein targets as new drugs.

References

[1] S. K. Burley *et al.*, "Structural genomics: beyond the human genome project," *Nat Genet*, vol. 23, pp. 151-7, Oct. 1999.

[2] S. K. Burley and J. B. Bonanno, "Structural genomics of proteins from conserved biochemical pathways and processes," *Curr Opin Struct Biol*, vol. 12, pp. 383-91, Jun. 2002.

[3] A. E. Todd, R. L. Marsden, J. M. Thornton, and C. A. Orengo, "Progress of structural genomics initiatives: an analysis of solved target structures," *J Mol Biol*, vol. 348, pp. 1235-60, May 20, 2005.

[4] N. Deshpande *et al.*, "The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema," *Nucleic Acids Res*, vol. 33, pp. D233-7, Jan. 1, 2005.

[5] A. R. Ortiz, C. E. Strauss, and O. Olmea, "MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison," *Protein Sci*, vol. 11, pp. 2606-21, Nov. 2002.

[6] C. Bystroff and D. Baker, "Prediction of local structure in proteins using a library of sequence-structure motifs," *J Mol Biol*, vol. 281, pp. 565-77, Aug. 21, 1998.
[7] A. C. Camproux, R. Gautier, and P. Tuffery, "A hidden markov model derived structural alphabet for proteins," *J Mol Biol*, vol. 339, pp. 591-605, Jun. 4, 2004.
[8] A. G. de Brevern, "New assessment of a structural alphabet," *In Silico Biol*, vol. 5, pp. 283-9, 2005.
[9] A. G. de Brevern, C. Etchebest, and S. Hazout, "Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks," *Proteins*, vol. 41, pp. 271-87, Nov. 15, 2000.
[10] J. S. Fetrow, M. J. Palumbo, and G. Berg, "Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme," *Proteins*, vol. 27, pp. 249-71, Feb. 1997.
[11] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt, "Small libraries of protein fragments model native protein structures accurately," *J Mol Biol*, vol. 323, pp. 297-307, Oct. 18, 2002.
[12] M. Levitt, "Accurate modeling of protein conformation by automatic segment matching," *J Mol Biol*, vol. 226, pp. 507-33, Jul. 20, 1992.
[13] M. J. Rooman, J. Rodriguez, and S. J. Wodak, "Automatic definition of recurrent local structure motifs in proteins," *J Mol Biol*, vol. 213, pp. 327-36, May 20, 1990.
[14] M. Tyagi, V. S. Gowri, N. Srinivasan, A. G. de Brevern, and B. Offmann, "A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications," *Proteins*, vol. 65, pp. 32-9, Oct. 1, 2006.
[15] M. Tyagi *et al.*, "Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet," *Nucleic Acids Res*, vol. 34, pp. W119-23, July 1, 2006.
[16] R. Unger and J. L. Sussman, "The importance of short structural motifs in protein structure analysis," *J Comput Aided Mol Des*, vol. 7, pp. 457-72, Aug. 1993.
[17] L. Fourrier, C. Benros, and A. G. de Brevern, "Use of a structural alphabet for analysis of short loops connecting repetitive structures," *BMC Bioinformatics*, vol. 5, pp. 58, May 12, 2004.
[18] C. Chotia and A. M. Lesk, "The relation between the divergence of sequence and structure in proteins," *EMBO J.*, vol. 5, pp. 823-6, 1986.
[19] C. H. Tung, J. W. Huang, and J. M. Yang, "Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database," *Genome Biol*, vol. 8, pp. R31, 2007.
[20] J. M. Yang and C. H. Tung, "Protein structure database search and evolutionary classification," *Nucleic Acids Res*, vol. 34, pp. 3646-59, 2006.
[21] C. H. Tung and J. M. Yang, "fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies," *Nucleic Acids Res*, vol. 35, pp. W438-43, July 2007.
[22] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J Mol Biol*, vol. 215, pp. 403-10, Oct. 5, 1990.
[23] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J Mol Biol*, vol. 247, pp. 536-40, Apr. 7, 1995.
[24] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet*, vol. 5, pp. 101-13, Feb. 2004.
[25] M. G. Grigorov, "Global properties of biological networks," *Drug Discov Today*, vol. 10, pp. 365-72, Mar. 1, 2005.
[26] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, pp. 910-3, May 3, 2002.
[27] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551-5, Aug. 30, 2002.
[28] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli," *Nat Genet*, vol. 31, pp. 64-8, May 2002.
[29] S. Wuchty, "Evolution and topology in the yeast protein interaction network," *Genome Res*, vol. 14, pp. 1310-4, July 2004.
[30] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509-12, Oct. 15, 1999.
[31] N. V. Dokholyan, B. Shakhnovich, and E. I. Shakhnovich, "Expanding protein universe and its origin from the biological Big Bang," in *Proc Natl Acad Sci U S A*, vol. 99, pp. 14132-6, Oct. 29, 2002.
[32] Z. Zhang and M. G. Grigorov, "Similarity networks of protein binding sites," *Proteins*, vol. 62, pp. 470-8, Feb. 1, 2006.
[33] Y. Sawada and S. Honda, "Structural diversity of protein segments follows a power-law distribution," *Biophys J*, vol. 91, pp. 1213-23, Aug. 15, 2006.
[34] Z. Gao *et al.*, "PDTD: a web-accessible protein database for drug target identification," *BMC Bioinformatics*, vol. 9, pp. 104, 2008.
[35] D. S. Wishart *et al.*, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Res*, vol. 34, pp. D668-72, Jan. 1, 2006.

Chi-Hua Tung was born in Kaohsiung, Taiwan. He received his bachelor degree in biology (2002) from National Sun Yat-Sen University. Also, He received the M.S. and Ph.D. degree in 2005 and 2009, respectively, all from the Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan. Since 2011, He has been with the Chung-Hua University, Taiwan, where he is currently an Assistant Professor with the Department of Bioinformatics. His research is focused on structure genomics and the studying of relationships between protein structures and functions using "Structural Alphabet." He is currently working to apply structural alphabet to structure motif discovery, peptide drug design and system biology.