行政院國家科學委員會專題研究計畫 成果報告

蛋白質結構基本單元網路於醫藥應用之研究
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 101-2311-B-216-001-
執 行 期 間 ：101 年 01 月 01 日至 101 年 07 月 31 日
執 行 單 位 ：中華大學生物資訊學系

計 畫 主 持 人：董其樺

計畫參與人員：碩士班研究生-兼任助理人員：葉融
碩士班研究生-兼任助理人員：楊凱傑

公 開 資 訊 ：本計畫可公開查詢

中 華 民 國 101 年 10 月 31 日

中 文 摘 要 ： 近年來由於結構基因體學蓬勃發展，使得蛋白質結構資料庫（PDB）中所記載的蛋白質結構資料，以驚人的速度增加。也因此在後基因體學時代，已知結構卻未知其功能的蛋白質數量不斷累積的情況下，便十分需要設計快速有效之生物資訊的方法，幫助研究者研究蛋白質之間的結構同源性以及分析其功能，進而探索蛋白質結構與功能的關聯性。

針對上述議題，我們近五年的研究已開發出「3D-BLAST」等方法論與工具，設計出利用 23 個結構字元所組成的結構字元集序列。此結構字元集之一級序列，包含蛋白質三級立體結構的資訊，可應用於快速的結構相似度搜尋。

本研究計畫根據結構字元集之研究，提出一全新定義之蛋白質局部結構片段，命名為「結構字元單位」，用以表現組成蛋白質結構上的基本單元。此結構字元單位乃是由兩個蛋白質二級結構以及一個位於兩二級結構之間的無特定結構片段所組合而成的序列，能同時呈現具高度變異的結構彈性及二級結構所提供的結構穩定性。接著，我們根據 3D-BLAST 快速結構比對的結果，建立一結構字元單位相似度網路。此網路中的每個節點代表一個結構字元單位，而節點之間的連線則表示兩局部結構具有高度相似性。最後整合複雜系網路學的方法論，探究此網路之拓樸特徵。初步研究結果顯示，用蛋白質局部結構之相似度所建立的複雜網路，符合無尺度網路的拓樸特徵，其連接度分佈呈現冪次函數曲線而非常態分佈，表示該網路中存在少數幾個擁有高度連接的節點，而大部分其餘的節點則只擁有很少的連線。

本研究計畫提出證據證實蛋白質具模組性，可由一定數量的基本單元組合而成。某些通用之基本單元頻繁出現於許多蛋白質中，而某些獨特之基本單元則用以決定蛋白質的特殊功能。本計畫未來可延伸應用於已知的藥物-標的蛋白資料庫，配合結構相似度網路所紀錄之資訊，幫助我們辨識可能執行重要功能的蛋白質片段，輔助開發胜肽類藥物、多標靶藥物或舊藥新用之應用。

中文關鍵詞： 局部結構相似度網路、蛋白質模組性、網路生物學

英 文 摘 要 ： With structural models developed using genome-wide investigative strategies, the number of protein structures with unknown or unassigned functions in the Protein Data Bank (PDB) has been rapidly increasing. Effective bioinformatics methods are therefore needed to annotate the structural homology and possible functions of these protein structures. In this study, we develop a new network approach to

identify protein structures based on the 3D-BLAST method. Using tertiary protein structures, this method enables not only a fast protein similarity search but also the identification of 23 states of structural alphabet (SA) sequences that represent the structural motifs of protein backbones.
Using SA, we define new local structural fragments called units of structural alphabet (USAs) that represent unique features of protein structures. Each USA is composed of two secondary protein structures and one loop located between these two secondary structures; USAs can maintain not only the flexibility of variable loops but also the stability of secondary structures. We conduct a similarity search and investigate the network formed by all-against-all USA sequence comparisons, where USAs represent nodes and links represent homology relationships.
Our findings show a highly uneven degree distribution characterized by a few and highly connected USAs (hubs) coexisting with many nodes having only a few links. Networks with such a power-law degree distribution are scale free. These findings not only suggest the existence of organizing principles for local protein structures but also allow us to identify key fragments that are potentially useful for new drug development and design. Of particular interest is the identification of USAs in the set of known drug protein targets.

英文關鍵詞： local structure similarity network, protein modularity, network biology

行政院國家科學委員會補助專題研究計畫 ☑成果報告 □期中進度報告

蛋白質結構基本單元網路於醫藥應用之研究

Networks of Protein Structural Units in Pharmaceutical Applications

計畫類別：☑個別型計畫　□整合型計畫
計畫編號：NSC 101-2311-B-216-001
執行期間：101 年 1 月 1 日至 101 年 7 月 31 日

執行機構及系所：中華大學 生物資訊系

計畫主持人： 董其樺 助理教授
共同主持人：
計畫參與人員：碩士班研究生-兼任助理人員：葉融
碩士班研究生-兼任助理人員：楊凱傑

成果報告類型(依經費核定清單規定繳交)：☑精簡報告　□完整報告

本計畫除繳交成果報告外，另須繳交以下出國心得報告：
□赴國外出差或研習心得報告
□赴大陸地區出差或研習心得報告
□出席國際學術會議心得報告
□國際合作研究計畫國外研究報告

處理方式：除列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

中　華　民　國 101 年　10 月　31 日

中文摘要

近年來由於結構基因體學蓬勃發展，使得蛋白質結構資料庫（PDB）中所記載的蛋白質結構資料，以驚人的速度增加。也因此在後基因體學時代，已知結構卻未知其功能的蛋白質數量不斷累積的情況下，便十分需要設計快速有效之生物資訊的方法，幫助研究者研究蛋白質之間的結構同源性以及分析其功能，進而探索蛋白質結構與功能的關聯性。

針對上述議題，我們近五年的研究已開發出「3D-BLAST」等方法論與工具，設計出利用 23 個結構字元所組成的結構字元集序列。此結構字元集之一級序列，包含蛋白質三級立體結構的資訊，可應用於快速的結構相似度搜尋。

本研究計畫根據結構字元集之研究，提出一全新定義之蛋白質局部結構片段，命名為「結構字元單位」，用以表現組成蛋白質結構上的基本單元。此結構字元單位乃是由兩個蛋白質二級結構以及一個位於兩二級結構之間的無特定結構片段所組合而成的序列，能同時呈現具高度變異的結構彈性及二級結構所提供的結構穩定性。接著，我們根據 3D-BLAST 快速結構比對的結果，建立一結構字元單位相似度網路。此網路中的每個節點代表一個結構字元單位，而節點之間的連線則表示兩局部結構具有高度相似性。最後整合複雜系網路學的方法論，探究此網路之拓樸特徵。初步研究結果顯示，用蛋白質局部結構之相似度所建立的複雜網路，符合無尺度網路的拓樸特徵，其連接度分佈呈現冪次函數曲線而非常態分佈，表示該網路中存在少數幾個擁有高度連接的節點，而大部分其餘的節點則只擁有很少的連線。

本研究計畫提出證據證實蛋白質具模組性，可由一定數量的基本單元組合而成。某些通用之基本單元頻繁出現於許多蛋白質中，而某些獨特之基本單元則用以決定蛋白質的特殊功能。本計畫未來可延伸應用於已知的藥物-標的蛋白資料庫，配合結構相似度網路所紀錄之資訊，幫助我們辨識可能執行重要功能的蛋白質片段，輔助開發胜肽類藥物、多標靶藥物或舊藥新用之應用。

關鍵字：局部結構相似度網路、蛋白質模組性、網路生物學

Abstract

With structural models developed using genome-wide investigative strategies, the number of protein structures with unknown or unassigned functions in the Protein Data Bank (PDB) has been rapidly increasing. Effective bioinformatics methods are therefore needed to annotate the structural homology and possible functions of these protein structures.

In this study, we develop a new network approach to identify protein structures based on the 3D-BLAST method. Using tertiary protein structures, this method enables not only a fast protein similarity search but also the identification of 23 states of structural alphabet (SA) sequences that represent the structural motifs of protein backbones.

Using SA, we define new local structural fragments called units of structural alphabet (USAs) that represent unique features of protein structures. Each USA is composed of two secondary protein structures and one loop located between these two secondary structures; USAs can maintain not only the flexibility of variable loops but also the stability of secondary structures. We conduct a similarity search and investigate the network formed by all-against-all USA sequence comparisons, where USAs represent nodes and links represent homology relationships.

Our findings show a highly uneven degree distribution characterized by a few and highly connected USAs (hubs) coexisting with many nodes having only a few links. Networks with such a power-law degree distribution are scale free. These findings not only suggest the existence of organizing principles for local protein structures but also allow us to identify key fragments that are potentially useful for new drug development and design. Of particular interest is the identification of USAs in the set of known drug protein targets.

Keyword: local structure similarity network, protein modularity, network biology

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 September 30, 2008, there were already more than 53384 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 (i) compare/analyze 3D structures [14–16], (ii) predict protein 3D structures from amino acid sequences [6, 9], (iii) reconstruct protein backbones [11], and (iv) 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 [19, 20], and assign protein families [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].

Random 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]. In this research, we propose a structural similarity network as a framework for classifying the structures of protein segments and analyze 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 [33]. However, in this research, the local structures of proteins, which consist of consecutive fixed numbers of amino acids, are not used for generating information on typical secondary structures.

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.

In the future, USA will be applied to peptide drug discovery and multi-target drug design for enhancing drug development efficiency and the biological diversity of targets. Potential important fragments are valuable for new drug development and prediction based on the complete networked system of binding interactions with proteins.

Materials and Methods

Figure 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. After determining USAs, protein units are translated into encoded SA sequences according to the kappa and alpha map. A complex network is obtained, with nodes representing USAs and links representing structural similarity based on the results of all-against-all USA alignment.

Furthermore, the topological properties of the similarity network are analyzed to determine whether the network is scale free.

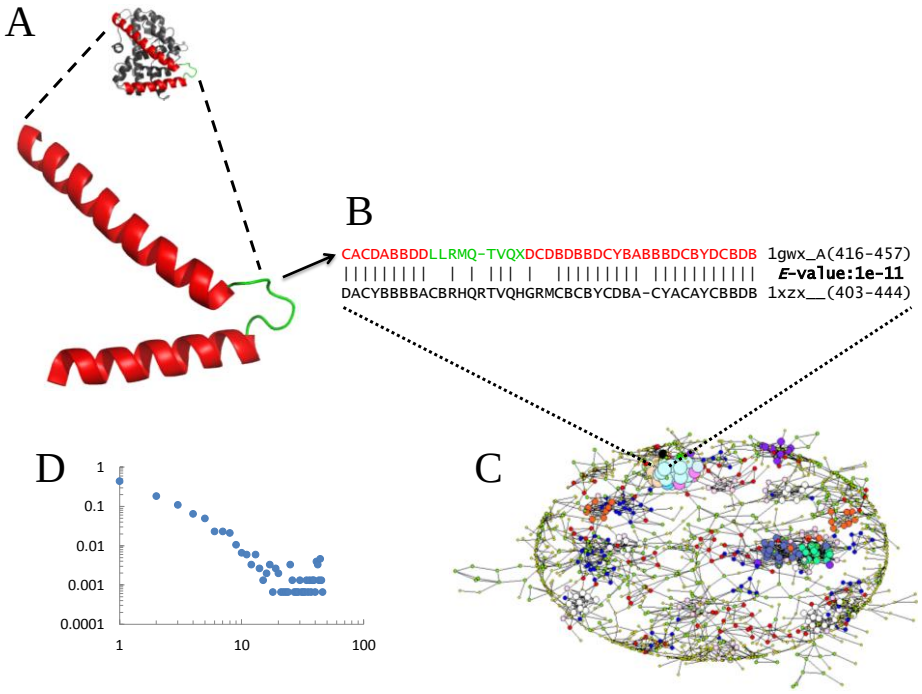

Figure 1. Research methodology.

Figure 2 shows the flow of this study. First, a testing set is prepared from nr-PDB-50 dated April 8, 2011; only proteins from the source species *Homo sapiens* are selected. 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. Third, protein chains are divided into many USAs with various kappa and alpha angles, leading to a USA database with 5525 protein units. Fourth, 3D-BLAST is used to search and align rapidly every USA against the whole database. Based on *E*-values in alignment results, the USA-based similarity network is developed. Finally, the characteristics and properties of this network are analyzed.

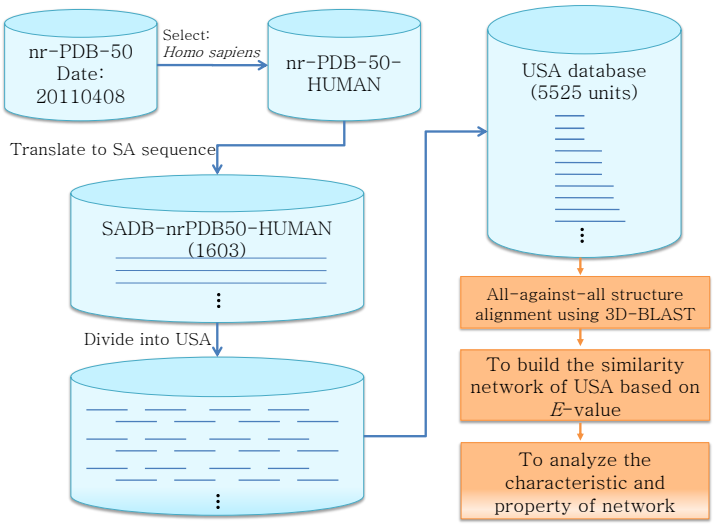

Figure 2. Research flow.

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. Figure 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).

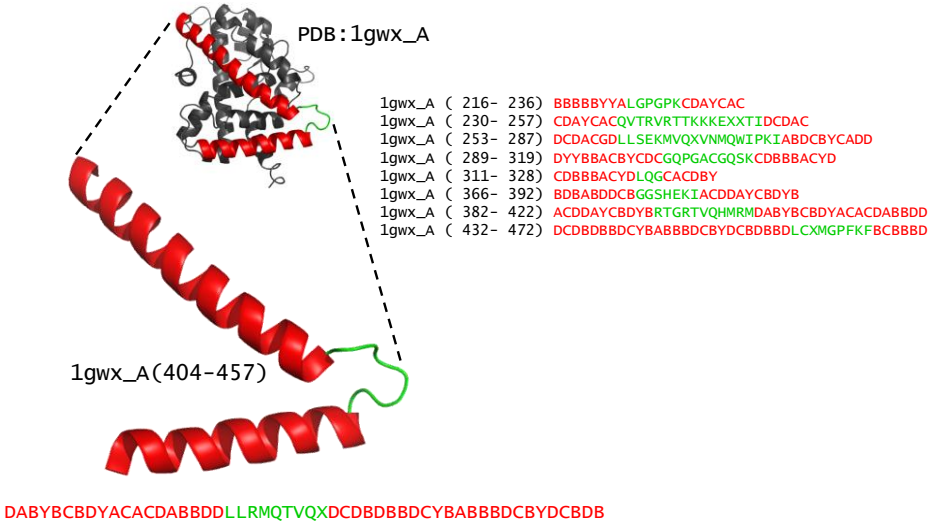

Figure 3. USAs in protein 1gwx_A and SA sequence.

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

$$E^{loop} = mn2^{-S} \quad \text{.....(1)}$$

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.

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 linked to a small fraction of all nodes in the network and most nodes are linked to a few neighbors.

Another quantifiable characteristic description is the clustering coefficient. The function $C(k)$ is defined as the average clustering coefficient of nodes with k links. In the clustering coefficient $C_I = 2n_I/k(k-1)$, 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].

We evaluate the statistical distribution of the USA-based similarity network and provide characteristics to show its power-law behavior. Results show that this new USA-based network is a scale-free and hierarchical network.

Results and Discussion

Definition of USA

We test various parameters for the length of secondary structures, loops, and whole USAs. The best results are shown in Figure 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 (Figure 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.

Distribution of USA database

Figures 4B and 4C 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 satisfies the nature of local protein structures. Figure 4B 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. Figure 4C shows that the length of the longest USA is 247 residues, and the biggest $P(L)$ is 0.0578 (L=20).

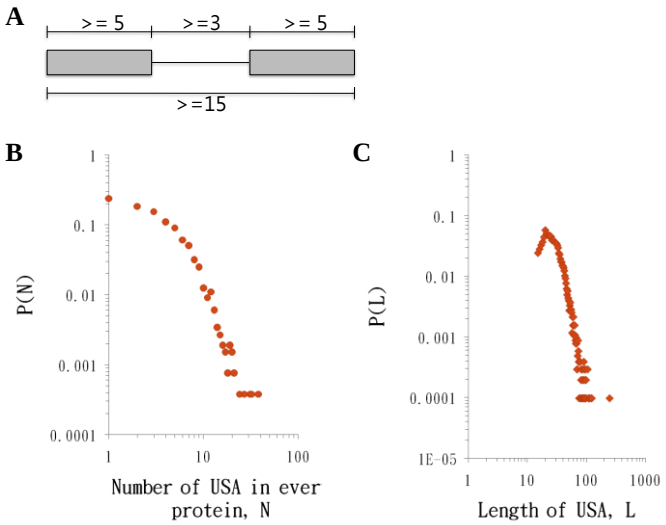

Figure 4. A) Criteria for the length of secondary structures, loops, and whole USAs. B) Distribution of the number of USAs in each protein. C) Distribution of length of each USA.

USA-based similarity network

Figure 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 red spot is a node representing one USA. Two nodes are connected by an edge (white color), 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.

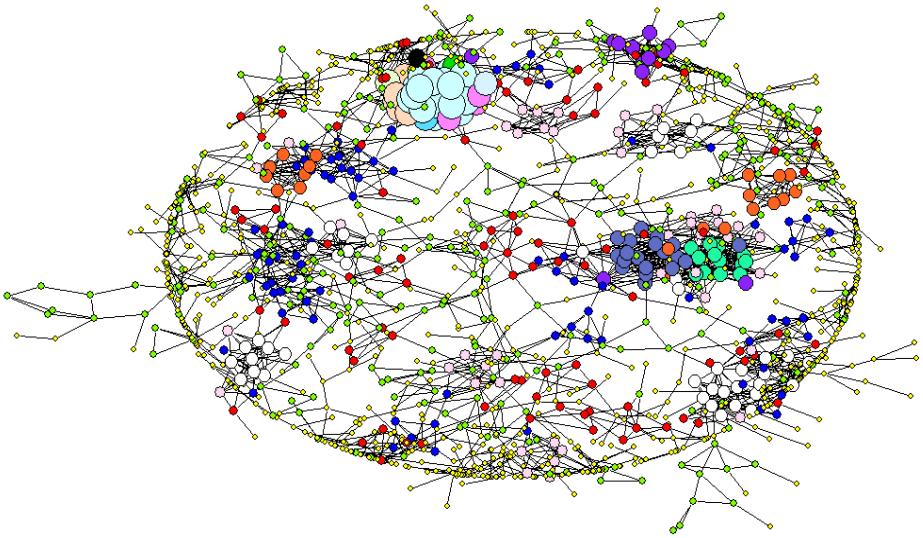

Figure 5. USA-based structural similarity network.

Network characteristics and properties

In this section, we determine if our novel network is scale free and even hierarchical. We first analyze the degree distribution of the network. Figure 6A presents that log-log plot of the distribution. The network is approximately characterized by power law, where $P(k) \sim k^{-1.34}$. 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 Figure 6B points out that $C(k)$ is independent of degree in our network. Therefore, the USA-based similarity network is scale free without hierarchical modularity.

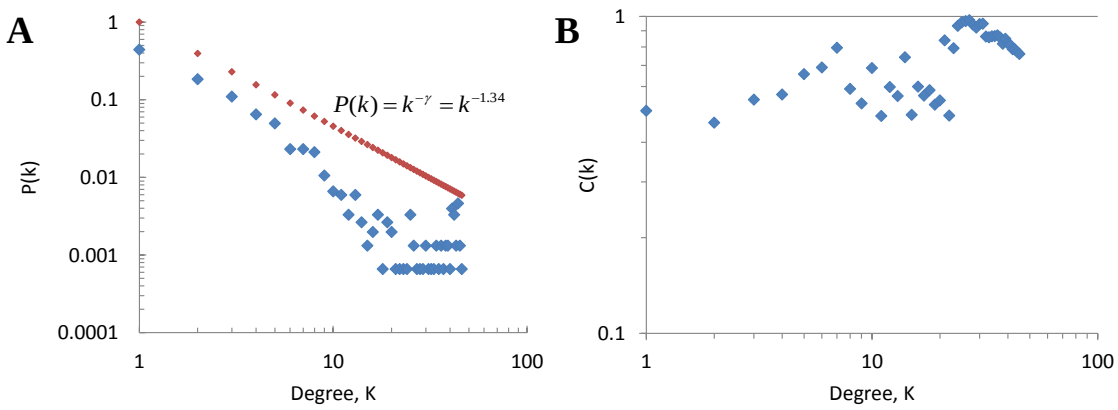

Figure 6. Log-log plot of the degree distribution of network and Clustering coefficient distribution of the USA-based similarity network.

Conclusions

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.

In contributing to the literature, this study combines two distinct research fields and provides a new and interesting viewpoint for investigating the relationship between protein structures and functions.

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. Burley, S.K., et al., *Structural genomics: beyond the human genome project.* Nat Genet, 1999. **23**(2): p. 151-7.
2. Burley, S.K. and J.B. Bonanno, *Structural genomics of proteins from conserved biochemical pathways and processes.* Curr Opin Struct Biol, 2002. **12**(3): p. 383-91.
3. Todd, A.E., et al., *Progress of structural genomics initiatives: an analysis of solved target structures.* J Mol Biol, 2005. **348**(5): p. 1235-60.
4. Deshpande, N., et al., *The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema.* Nucleic Acids Res, 2005. **33**(Database issue): p. D233-7.
5. Ortiz, A.R., C.E. Strauss, and O. Olmea, *MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.* Protein Sci, 2002. **11**(11): p. 2606-21.
6. Bystroff, C. and D. Baker, *Prediction of local structure in proteins using a library of sequence-structure motifs.* J Mol Biol, 1998. **281**(3): p. 565-77.
7. Camproux, A.C., R. Gautier, and P. Tuffery, *A hidden markov model derived structural alphabet for proteins.* J Mol Biol, 2004. **339**(3): p. 591-605.
8. de Brevern, A.G., *New assessment of a structural alphabet.* In Silico Biol, 2005. **5**(3): p. 283-9.
9. de Brevern, A.G., C. Etchebest, and S. Hazout, *Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks.* Proteins, 2000. **41**(3): p. 271-87.
10. Fetrow, J.S., M.J. Palumbo, and G. Berg, *Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme.* Proteins, 1997. **27**(2): p. 249-71.
11. Kolodny, R., et al., *Small libraries of protein fragments model native protein structures accurately.* J Mol Biol, 2002. **323**(2): p. 297-307.
12. Levitt, M., *Accurate modeling of protein conformation by automatic segment matching.* J Mol Biol, 1992. **226**(2): p. 507-33.
13. Rooman, M.J., J. Rodriguez, and S.J. Wodak, *Automatic definition of recurrent local structure motifs in proteins.* J Mol Biol, 1990. **213**(2): p. 327-36.
14. Tyagi, M., et al., *A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications.* Proteins, 2006. **65**(1): p. 32-9.
15. Tyagi, M., et al., *Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W119-23.
16. Unger, R. and J.L. Sussman, *The importance of short structural motifs in protein structure analysis.* J Comput Aided Mol Des, 1993. **7**(4): p. 457-72.
17. Fourrier, L., C. Benros, and A.G. de Brevern, *Use of a structural alphabet for analysis of short loops connecting repetitive structures.* BMC Bioinformatics, 2004. **5**: p. 58.
18. Chotia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* EMBO J., 1986. **5**: p. 823-6.
19. Tung, C.H., 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, 2007. **8**(3): p. R31.
20. Yang, J.M. and C.H. Tung, *Protein structure database search and evolutionary*

classification. Nucleic Acids Res, 2006. **34**(13): p. 3646-59.
21. Tung, C.H. and J.M. Yang, *fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W438-43.
22. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
23. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.
24. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
25. Grigorov, M.G., *Global properties of biological networks.* Drug Discov Today, 2005. **10**(5): p. 365-72.
26. Maslov, S. and K. Sneppen, *Specificity and stability in topology of protein networks.* Science, 2002. **296**(5569): p. 910-3.
27. Ravasz, E., et al., *Hierarchical organization of modularity in metabolic networks.* Science, 2002. **297**(5586): p. 1551-5.
28. Shen-Orr, S.S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nat Genet, 2002. **31**(1): p. 64-8.
29. Wuchty, S., *Evolution and topology in the yeast protein interaction network.* Genome Res, 2004. **14**(7): p. 1310-4.
30. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-12.
31. Dokholyan, N.V., B. Shakhnovich, and E.I. Shakhnovich, *Expanding protein universe and its origin from the biological Big Bang.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14132-6.
32. Zhang, Z. and M.G. Grigorov, *Similarity networks of protein binding sites.* Proteins, 2006. **62**(2): p. 470-8.
33. Sawada, Y. and S. Honda, *Structural diversity of protein segments follows a power-law distribution.* Biophys J, 2006. **91**(4): p. 1213-23.
34. Gao, Z., et al., *PDTD: a web-accessible protein database for drug target identification.* BMC Bioinformatics, 2008. **9**: p. 104.
35. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic Acids Res, 2006. **34**(Database issue): p. D668-72.

國科會補助專題研究計畫成果報告自評表

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

☑ 達成目標

☐ 未達成目標（請說明，以 100 字為限）

☐ 實驗失敗

☐ 因故實驗中斷

☐ 其他原因

2. 研究成果在學術期刊發表或申請專利等情形：

論文：☐已發表 ☐未發表之文稿 ☑撰寫中 ☐無

專利：☐已獲得 ☐申請中 ☑無

技轉：☐已技轉 ☐洽談中 ☑無

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

本研究內容與結果依照原計畫行程表執行，除原訂兩年期計畫之第二年部份未能在本次研究執行期間（七個月）內完成外，本研究皆相符於原計畫預定第一年之內容。

本研究第一年預期完成發展結構字元單位、設計新的計分方程式與適當的搜尋閥值、建構局部結構相似度網路等目標。最終研究結果顯示，結構字元單位之定義合理且具生物意義。此結構字元單位能將研究蛋白質結構與功能之關聯性的層次，從整體結構、功能性區域更進一步聚焦在更小的單元。同時，分析與觀察局部結構相似度網路後，驗證了原計畫之預期想法：「各種蛋白質由一定數量的基本單元組合而成，少部分基本單元頻繁出現在大多數蛋白質中，而另外某些基本單元則用來決定蛋白質之專一性功能」。

本研究計畫指導碩士級研究生與大學部專題生共四位，期間訓練學生程式設計、文獻搜尋與整理等資訊相關技能，並培養學生探討生物意義之生物相關能力。此亦符合原計畫預期目標。

本研究成果已準備發表期刊論文，將投稿於 *Biophysical Journal* 或 *BMC Bioinformatics* 等 SCI 期刊。

國科會補助計畫衍生研發成果推廣資料表

日期:2012/10/28

國科會補助計畫	計畫名稱: 蛋白質結構基本單元網路於醫藥應用之研究
	計畫主持人: 董其樺
	計畫編號: 101-2311-B-216-001- 學門領域: 生物學之生化及分子生物

無研發成果推廣資料

101 年度專題研究計畫研究成果彙整表

計畫主持人：董其樺	計畫編號：101-2311-B-216-001-
計畫名稱：蛋白質結構基本單元網路於醫藥應用之研究	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	2	2	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 □未發表之文稿 ■撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

本研究內容與結果依照原計畫行程表執行，除原訂兩年期計畫之第二年部份未能在本次研究執行期間（七個月）內完成外，本研究皆相符於原計畫預定第一年之內容。

本研究第一年預期完成發展結構字元單位、設計新的計分方程式與適當的搜尋閥值、建構局部結構相似度網路等目標。最終研究結果顯示，結構字元單位之定義合理且具生物意義。此結構字元單位能將研究蛋白質結構與功能之關聯性的層次，從整體結構、功能性區域更進一步聚焦在更小的單元。同時，分析與觀察局部結構相似度網路後，驗證了原計畫之預期想法：「各種蛋白質由一定數量的基本單元組合而成，少部分基本單元頻繁出現在大多數蛋白質中，而另外某些基本單元則用來決定蛋白質之專一性功能」。

本研究計畫指導碩士級研究生與大學部專題生共四位，期間訓練學生程式設計、文獻搜尋與整理等資訊相關技能，並培養學生探討生物意義之生物相關能力。此亦符合原計畫預期目標。

本研究成果已準備發表期刊論文，將投稿於 Biophysical Journal 或 BMC Bioinformatics 等 SCI 期刊。