BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 17 2014, pages 2440–2446
doi:10.1093/bioinformatics/btu327

Structural bioinformatics

Advance Access publication May 10, 2014

Exploring protein domain organization by recognition of secondary structure packing interfaces

Lizong Deng[1,2,†], Aiping Wu[1,†], Wentao Dai[1,2], Tingrui Song[1,2], Ya Cui[1,2] and Taijiao Jiang[1,*]

[1]Key Laboratory of Protein & Peptide Pharmaceuticals, National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101 and [2]University of the Chinese Academy of Sciences, Beijing 100049, China

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Protein domains are fundamental units of protein structure, function and evolution; thus, it is critical to gain a deep understanding of protein domain organization. Previous works have attempted to identify key residues involved in organization of domain architecture. Because one of the most important characteristics of domain architecture is the arrangement of secondary structure elements (SSEs), here we present a picture of domain organization through an integrated consideration of SSE arrangements and residue contact networks.

Results: In this work, by representing SSEs as main-chain scaffolds and side-chain interfaces and through construction of residue contact networks, we have identified the SSE interfaces well packed within protein domains as SSE packing clusters. In total, 17 334 SSE packing clusters were recognized from 9015 Structural Classification of Proteins domains of <40% sequence identity. The similar SSE packing clusters were observed not only among domains of the same folds, but also among domains of different folds, indicating their roles as common scaffolds for organization of protein domains. Further analysis of 14 small single-domain proteins reveals a high correlation between the SSE packing clusters and the folding nuclei. Consistent with their important roles in domain organization, SSE packing clusters were found to be more conserved than other regions within the same proteins.

Contact: taijiao@moon.ibp.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 25, 2014; revised on April 12, 2014; accepted on May 6, 2014

1 INTRODUCTION

The 3D structure of a protein is crucial for understanding its physiological function in living cells. Since 1958, numerous efforts have been spent on determination of protein structures, and the number of proteins with known structure is increasing with an exponential speed. By the end of 2013, over 96 000 protein structures have been determined and deposited in the Protein Data Bank (http://www.rcsb.org/pdb/) (Berman *et al.*, 2000). Given the complexity of protein structures, researchers have attempted to decompose them to compact and fundamental units, so called domains, and thus several databases including CATH (Sillitoe *et al.*, 2013) and Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995) have been generated to provide valuable resources for the understanding of protein structure and function at domain level.

As the fundamental units of proteins, understanding the organization of domain architectures would deepen our insight into the relationship between protein structure and function, which could further facilitate protein structure prediction, design and simulation. Previously, a number of efforts have been devoted to identifying key residues involved in organization of domain architecture, and these works could fall into two categories. One was to identify a specific group of residues based on structural features like hydrophobicity and compactness (Heringa and Argos, 1991; Swindells, 1995; Zehfus, 1995, 1997). For example, Zehfus (1995, 1997) examined the compactness of hydrophobic clusters within proteins and revealed a certain correlation between the hydrophobic clusters and protein folding units. The other attempted to characterize domain architecture by identifying tertiary packing motifs within domains based on geometric properties (Bandyopadhyay *et al.*, 2009a, b; Day *et al.*, 2010; Kannan and Vishveshwara, 1999). For example, Day *et al.* (2010) applied the Delaunay tessellation to define the tetrahedral packing motifs within proteins and found the general repetitive tetrahedral packing motifs in protein tertiary structures. All of these works mainly focused on residue-level contacts, but they did not explicitly consider the important roles of SSE packing in the organization of domain architectures.

As we know, one of the most important characteristics of domain architecture is the spatial arrangement of secondary structure elements (SSE), especially of these regular ones like α-helix and β-strand (Branden and Tooze, 1991; Lesk, 2001). The SSE arrangements have been well explored in many protein families, including helix-turn-helix motif in a class of DNA-binding proteins called BZIP family proteins (Brennan and Matthews, 1989; Harrison and Aggarwal, 1990), four-helix bundles present in many hormones (De Vos *et al.*, 1992) and the Rossmann fold ($\beta\alpha\beta\alpha\beta$ unit) present in nucleotide-binding proteins (Rao and Rossmann, 1973; Rossmann and Argos, 1976). Many works have been developed in mining SSE motifs in protein structures (Aung and Li, 2007; Comin *et al.*, 2008; Richards and Kundrot, 1988); however, most of them treated an SSE as a

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

simplified vector representation and did not uncover the details of underlying residues mediating SSE packing.

To gain a comprehensive picture of domain organization, our work integrated the previous efforts at both residue level and SSE level by considering the SSE interfaces as the basic units for packing of SSEs. We first constructed an effective contact network among side chains within a protein domain, and then identified the well-packed SSE interfaces within the network as an SSE packing cluster. Such SSE packing clusters not only characterize the domain organization but also shed lights into the molecular mechanisms underlying domain formation.

2 METHODS

2.1 Domain structure database

The ASTRAL SCOP40 database (Chandonia *et al*., 2004) was used for large-scale analysis of domain organization. It is a subset of SCOP database (v1.75) (Murzin *et al*., 1995) with at most 40% sequence identity between any two sequences. The domains with missing main chain atoms or low resolution (>3 Å) were excluded. The g-k classes defined in SCOP, namely small proteins, coiled-coil proteins, low resolution proteins, peptides and fragments and designed proteins, were also removed. In total, 9015 domains were retained in our final dataset.

2.2 Definition of scaffold and interfaces for regular secondary structure element

Regular SSEs play important roles in organizing local conformations of proteins, and the specific interactions between different SSEs constitute the structural basis for protein stability. Previously, Preißner *et al.* (1998) pointed out that SSE interfaces consist of pairs of matching molecular surface patches between neighboring SSEs. Based on this concept, in our study, an SSE was viewed to be composed of main-chain scaffold and side-chain interfaces (Fig. 1A). Usually, the side chains of an α-helix sit in two panels and formed two different interfaces. But when an α-helix is fully surrounded by other structure elements, such as within a β-barrel, all its side chains form a continuous cylindrical interface. As for a β-strand, it could offer two separate panel interfaces formed by alternate side chains at the flank of its main-chain scaffold. One interface consists of odd-numbered sites and the other consists of even-numbered sites.

2.3 Recognition of SSE packing clusters

Given the decomposition of SSEs to scaffolds and interfaces, the interfaces, rather than single residues, could be regarded as basic interaction units in the arrangement of structure elements within a protein structure. Based on such interface-mediated interaction model, those interfaces in close contacts in protein structure were identified as an SSE packing cluster, and the recognition of SSE packing clusters was carried out by the following four steps:

Step 1: Identification of buried sites. The residue solvent accessibility (RSA) for each residue in protein structure was calculated with NACCESS v2.1.1 (Hubbard, 1992). Those residues with RSA <25% (the widely adopted criteria as in Bloom *et al.* (2006) were identified as buried sites.

Step 2: Identification of SSE and its buried interfaces. A secondary structure state (helix, strand and coil) for each site was assigned by DSSP v2.1.0 (Kabsch and Sander, 1983). Only the two regular secondary structure states, helix and strand, were considered for definition of SSE. A fragment of ≥7 amino acids with helix state was assigned as an α-helix SSE, and a fragment of ≥4 amino acids with strand state was assigned as a β-strand SSE. The buried interfaces of an SSE were identified based on the structural properties of the SSE and the distribution of buried sites.

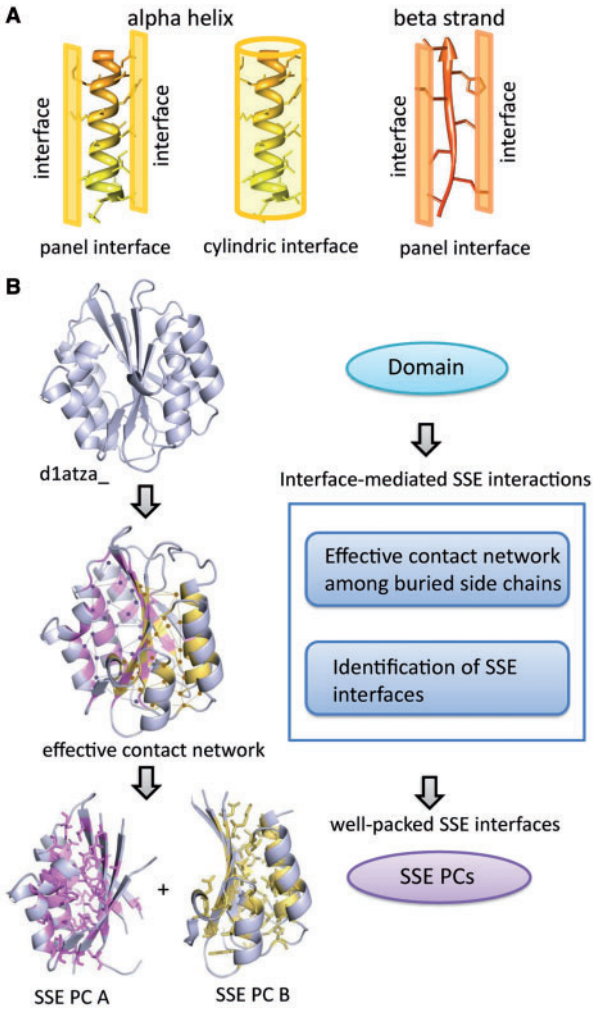

Fig. 1. Decomposition of protein domains into SSE packing clusters (SSE PCs). (**A**) The types of side-chain interfaces in α-helix and β-strand secondary structures. (**B**). Key steps in recognition of SSE packing clusters from protein domains

Briefly, for α-helix, if the i-th buried site and the (i + 1)-th buried site are within sequence separation <5, they were considered to be in the same buried interface; otherwise, they belong to different interfaces. For β-strand, one interface consisted of odd-numbered sites and the other consisted of even-numbered sites.

Step 3: Construction of an effective contact network. To characterize the direct interactions between residues, we attempted to identify the effective contact between side chains within a distance threshold 16.0 Å (distance between side chain centroids). Two side chains are considered to be in effective contact if the line connecting the centroids of two buried side chains does not pass through any atoms of the main chain. Such pairs of side chains constitute an effective contact network.

Step 4: Recognition of SSE packing clusters. Based on the effective contact network of side chains, the sites that mediate SSE packing were identified and grouped into SSE packing clusters. Supplementary Figure S1 illustrates the process of recognition of SSE packing clusters from SSE interfaces and effective contact network. In brief, we first selected an SSE interface with a site having the biggest number of contacts within the effective contact network as the initial SSE interface for a SSE packing cluster. Next, the interfaces of SSEs having sites with over two

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

contacts with the initial SSE interface were included in the SSE packing cluster. Then, the remaining SSE interfaces having sites with over two contacts with all the sites in the SSE packing cluster were included. The sites that do not belong to SSEs were also included in the SSE packing cluster, if they have over two contacts with the SSE packing cluster. This step continues until no such SSE interfaces were satisfied. If there remain sites in the effective contact network, new SSE packing clusters could be identified by following the above procedure. After all the SSE packing clusters have been identified, they were further merged as follows: If the number of contacts between two different SSE packing clusters exceeds the number of sites in smaller SSE packing cluster, the smaller SSE packing cluster was merged into the bigger one. These SSE packing clusters of less than seven sites that could not be merged were not considered in our analysis.

2.4 Structure comparison between SSE packing clusters and identification of SSE packing patterns

An SSE packing cluster identified is actually a cluster of sites that mediate the association of SSEs. To compare the geometric similarity between two SSE packing clusters, we first reconstructed their intact structures by including the structures of whole SSEs with an extension of one residue at both ends. If there existed sites that do not belong to an SSE, the structures of the extension of three residues at both sides were included. Then a non-sequential structure alignment method named MICAN (Minami *et al*., 2013) was used to compare the structures between two SSE packing clusters. Based on pairwise structure alignment, SSE packing clusters derived from SCOP domains with similarity score TM-score $\geq$0.4 (a statistically significant threshold for structural similarity) (Zhou and Skolnick, 2007) were linked to form a network. Then, based on the network, the complete-linkage clustering algorithm (Johnson, 1967) was used to identify groups of SSE packing clusters with similar geometry. Each group was regarded as an SSE packing pattern.

2.5 Conservation measurement of SSE packing clusters

The effect of substitution was used as an indicator for the degree of conservation of SSE packing clusters. For a pair of domains belonging to the same family, the structurally equivalent sites were derived by performing MICAN structure alignment. Given a site pair, the substitution effect between any two amino acids could be evaluated based on the overall physicochemical similarity between them, ranging from 0 (dissimilar) to 8 (identical), as described in the McLachlan matrix (McLachlan, 1971) (AAindex entry: MCLA710101). If the sites involved in SSE packing clusters are conservative, it was expected that they tended to substitute similar amino acids; thus, the average score of substitutions, named as conservation score, would be close to 8. The degree of conservation on SSE packing clusters was further compared with that of other regions within same domains.

3 RESULTS

3.1 Overview of the method for recognizing SSE packing clusters

To explore the organization of domain architecture, we viewed an SSE as a combination of a main-chain scaffold and several side-chain interfaces (Fig. 1A). For an α-helix, the definition of its side-chain interfaces depends on the structural context it locates in. Generally, the side chains of an α-helix form two panel-like interfaces, except that they form a cylindrical interface when sitting inside a β-barrel (Fig. 1A). For a β-strand, the even-numbered sites form one panel-like interface, and the odd-numbered sites form the other interface. By decomposing an SSE to a scaffold and several interfaces, the arrangement of SSEs was regarded as interactions among different interfaces; those interfaces in close contacts were identified as an SSE packing cluster. To effectively recognize the SSE packing clusters, we first constructed an effective contact network among side chains, then identified those interfaces in dense contacts and further decomposed domains into SSE packing clusters (Fig. 1B, see Section 2 for details).

3.2 SSE packing clusters could be regarded as common scaffolds for organization of domain architectures

In total, 17 334 SSE packing clusters were derived from 9015 non-redundant domains of ASTRAL SCOP40 database (see Section 2). Most domains (90.4%) contain one to three SSE packing clusters, and only few domains can contain up to 11 SSE packing clusters (Fig. 2A and see Supplementary Table S1 for details). These SSE packing clusters mostly (77.8%) consist of approximately two to eight regular SSEs (Supplementary Fig. S2A), with a peak at five SSEs. The size of SSE packing clusters (the number of residues) also exhibits a unimodal distribution with the peak appearing in the range of 17–24 sites (Supplementary Fig. S2B).

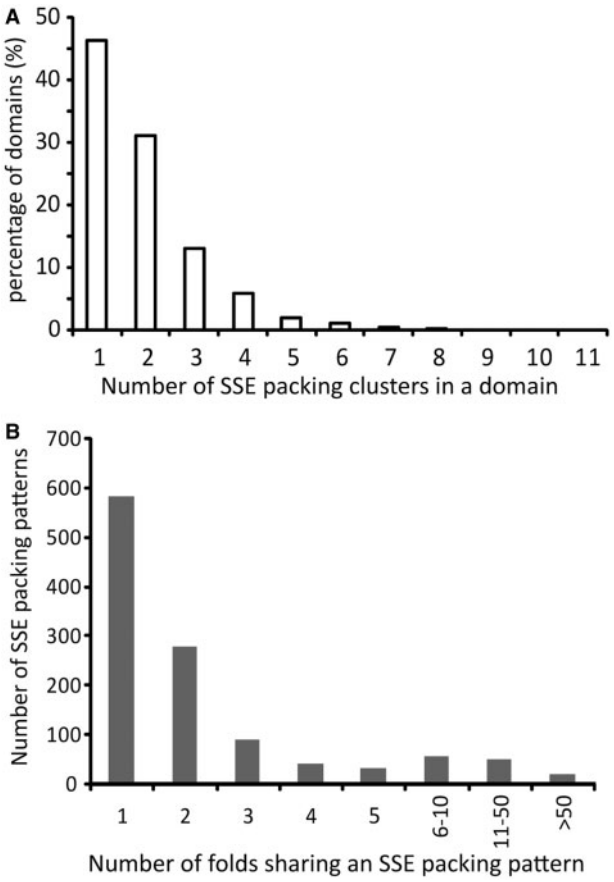

Fig. 2. The statistical characteristics of SSE packing clusters. (**A**) Histogram of the number of SSE packing clusters in a domain. (**B**) Histogram of the number of SSE packing patterns shared by different numbers of folds

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Based on their geometric similarity, the 17 334 SSE packing clusters were further clustered into 6497 classes, so called SSE packing patterns (see Section 2). Among the 6497 SSE packing patterns, 1160 SSE packing patterns contained two or more packing clusters, which covered 69% (11 997) of SSE packing clusters. In SCOP, the 9015 domains could also be classified into 1004 folds of distinct architecture. As the SSE packing patterns derived from domains are more basic structural units, we wondered whether they were able to capture organizational correlation between folds of different architecture. Indeed, 576 of 1160 SSE packing patterns were found across different folds (Fig. 2B). The most commonly used SSE packing pattern spanned 582 domains across 158 different folds. If we connect those pairs of folds sharing an SSE packing pattern, this will result in a network of high connectivity in protein fold space (see Supplementary Fig. S3 for details). Notably, >80% (837) of the folds were involved in the network. Moreover, we observed significant amino acid patterns of at least five amino acids ($P<0.05$) underlying the 21 SSE packing patterns with >100 packing cluster members (see Supplementary Methods and Supplementary Fig. S4 for details). Taken together, we argue that SSE packing clusters as lower level structure units than domains are common scaffolds for organization of domain architectures.

3.3 SSE packing clusters correspond well to folding nuclei in small protein domains

To find out the potential roles of SSE packing clusters in domain organization, a set of 14 small single-domain proteins, which have been studied with detailed Φ-value analysis, were retrieved from literatures (see Supplementary Materials for details). Φ-value analysis is an experimental protein engineering method used to study the structure of the folding transition state in small protein domains that fold in a two-state manner (Fersht and Sato, 2004; Ozkan *et al.*, 2001). Using this method, point mutations are performed and their effects are estimated by Φ-values. The Φ-values obtained are generally grouped into low (Φ-value <0.1), medium ($0.1 \leq$ Φ-value <0.4) and high (Φ-value ≥ 0.4) regions (Chiti *et al.*, 1999). Usually, sites with medium or high Φ-value are possibly essential for the formation of folding nucleus (Chiti *et al.*, 1999).

There are 170 residues in the SSE packing clusters, namely 170 PC sites, identified from the 14 small single-domain proteins with detailed Φ-value analysis. Remarkably, 81.8% (139) of them have either medium (89) or high Φ-values (50) (Fig. 3A). Further, the percentage of PC sites with medium/high Φ-values for each domain was calculated, and it was shown that most of the PC sites possess medium/high Φ-values in each domain, ranging from 52 (ubiqutin) to 100% (FKBP12) (Fig. 3B andsee Supplementary Table S2 for details). When comparing the location of the PC sites of medium/high Φ-values with those of low Φ-values, we found that PC sites with medium/high Φ-values located closer to the centroid of SSE packing clusters than the PC sites with low Φ-values (*t*-test, $P<0.05$) (Fig. 3C). Figure 3D–G showed the Φ-values of PC sites in four representative proteins belonging to different structure classes: all-α, all-β (β-sheet and β-barrel) and α/β. Clearly, the PC sites with medium/high Φ-values (in purple and red) mainly sit in the center of SSE packing clusters, whereas PC sites with low Φ-values (in blue) are scattered in the periphery. The high correlation between the packing cluster and the folding nuclei as determined by medium/high Φ-values suggest the important role of SSE packing cluster in formation of protein domains.

3.4 SSE packing clusters are more conserved within domains

To further shed lights into the important role of SSE packing clusters in domain organization, we looked into the conservation of SSE packing clusters in the evolution of domains. Here we selected 1437 pairs of domains, each of which belongs to the same family. Based on the structure alignments of these domain pairs, the conservation of SSE packing clusters could be evaluated by measuring the substitution effects on the aligned sites with McLanchlan matrix, and further compared with that of other regions in domains (see Section 2 for details). Figure 4 shows the distributions of conservation score for SSE packing clusters and other regions in domains. On average, the conservation score for SSE packing clusters (5.0) is significantly higher than that for other regions (4.1) (*t*-test, $P<2.2e\text{-}16$). Similar results were also observed when we attempted another two different amino acid substitution matrices, BLOSUM62 matrix (Henikoff and Henikoff, 1992) and CSSM matrices (context-specific amino acid substitution matrices) (Goonesekere and Lee, 2008) (Supplementary Fig. S5). The higher degree of conservation for SSE packing clusters further indicates their important role in domain organization.

3.5 Case study: implication of SSE packing clusters of kinase domains in human diseases

Kinases are essential enzymes in human proteome, which phosphorylate the serine, threonine or tyrosine residues of target proteins with phosphate groups from ATP. It is estimated that up to 30% of human proteins are modified by kinases (Hubbard and Cohen, 1993). Thus, mutations on kinases could underlie many human diseases, such as developmental and metabolic disorders, as well as certain cancers (Lahiry *et al.*, 2010). To shed lights into the roles of SSE packing clusters of kinase domains in human diseases, we collected 729 disease-related mutations for 16 kinase catalytic domains with high-resolution crystal structures (see Supplementary Materials for details).

We first identified SSE packing clusters from the 16 kinase domains, and it was found that kinase domains usually contained two SSE packing clusters of different size: the bigger one corresponds to the C-terminal lobe of kinase domain, and the smaller one corresponds to the N-terminal lobe (Fig. 5A). Further, we investigated the overlap between the catalytic regions and SSE packing clusters of kinase domains. Previous studies have identified five functional elements or subdomains as catalytic regions for kinases (Hanks and Hunter, 1995), namely SD I (P-loop, ATP-binding loop), SD III (α C-helix), SD VIB (catalytic loop), SD VII (activation loop) and SD VIII (P + 1 loop) (Fig. 5A). On average, of 1791 sites involved in the SSE packing clusters, 28.9% are found to locate at the catalytic regions; reversely, of 1472 sites belonging to the catalytic regions, 35.6% are found to participate in the formation of the SSE

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

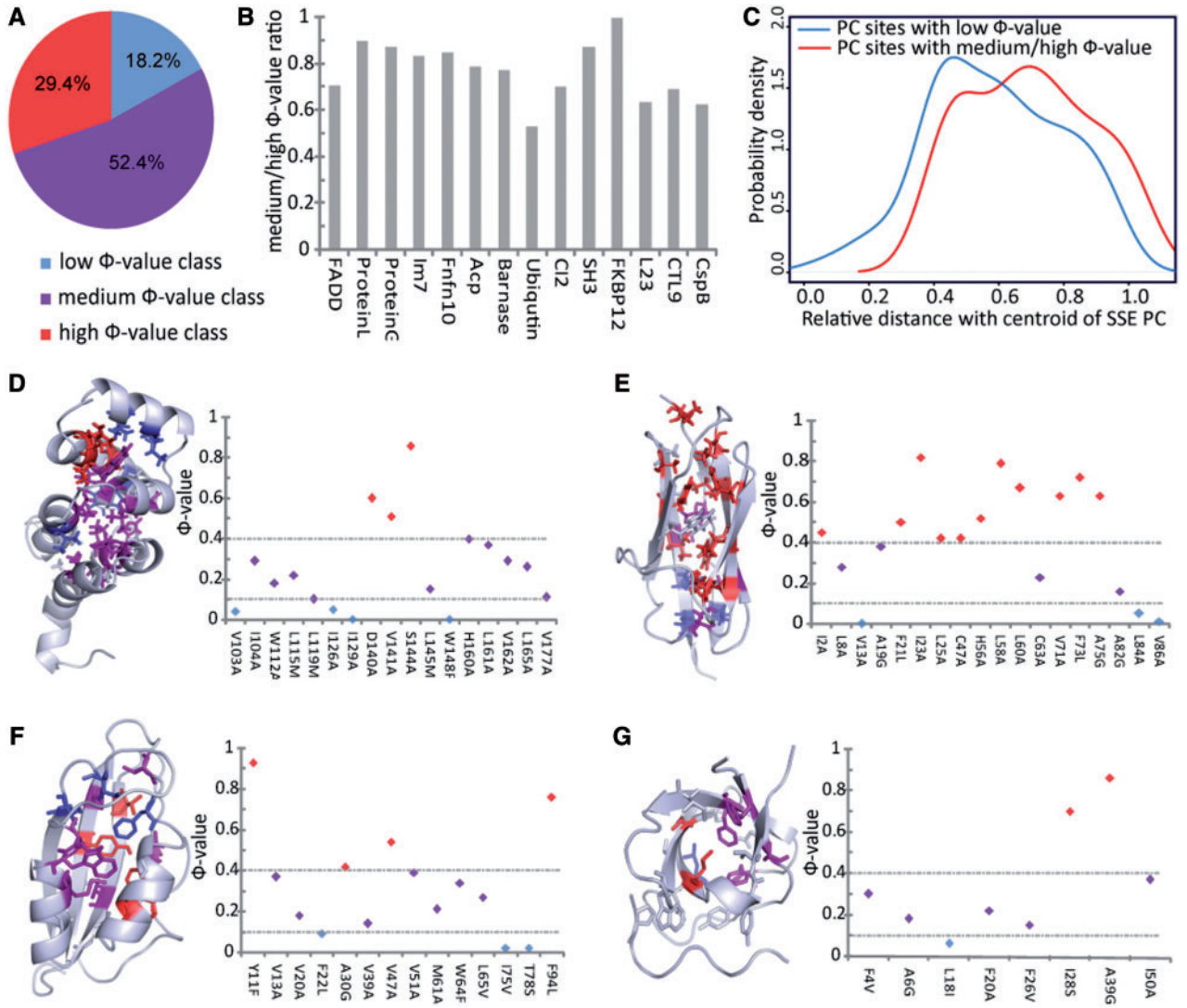

Fig. 3. Φ-value characteristics of the sites involved in SSE packing clusters (PC sites). (**A**) Pie chart showing the distribution of PC sites within different Φ-value classes. Three Φ-value classes (low with Φ-value <0.1, medium with Φ-value between 0.1 and 0.4 and high with Φ-value >0.4) were colored with blue, purple and red, respectively. (**B**) Histogram showing the percentages of PC sites with medium/high Φ-values in 14 different small single-domain proteins. (**C**) Comparison of the distribution of the relative distance to the centroid of SSE packing clusters for PC sites with low Φ-values (blue) and PC sites with medium/high Φ-values (red).(D–G) Correlations between Φ-values and PC sites for FADD domain (**D**), Im7 domain (**E**), Acp domain (**F**) and SH3 domain (**G**). PC sites with low Φ-value were colored with blue, medium Φ-value colored with purple and high Φ-value colored with red

packing clusters (Supplementary Table S3). Such a moderate overlap suggests that the catalytic regions and the scaffolds of the kinase domain are distinct, but with close spatial relationship. For example, in ACVRL1 kinase domain, SD VIB (catalytic loop), SD VII (activation loop) and SD VIII (P + 1 loop) were linked to the C-terminal SSE packing cluster; and SD I (P-loop) and SD III (α C-helix) were supported by the N-terminal SSE packing cluster (Fig. 5A).

Then, we looked into the distribution of the 729 disease-related mutations on these kinase domains. Only 281 were directly involved in kinase activity as they occur on the catalytic regions. But how the remaining 448 mutations affect kinase activity is unclear. Remarkably, we found that 213 disease-related mutations occurred in the packing clusters, indicating that the SSE packing clusters as scaffolds for kinase domains accounted for many disease-related mutations in kinase. Notably, for some kinase domains, such as ACVRL1 kinase domain and BTK kinase domain, the disease-related mutations are significantly enriched in packing cluster regions compared with that of other regions in domains (Fig. 5B, see Supplementary Table S4 for details).

4 DISCUSSION

In this study, through recognition of interfaces mediating SSE packing, we have attempted to identify the structural and physicochemical features underlying the organization of domain architecture. The packing clusters identified not only depicted the

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

arrangement of structure elements in domain architecture, but could also capture the underlying residues mediating SSE packing patterns in domain formation.

Previously, hydrophobic cores have been analyzed within domains (Alexeevski *et al.*, 2003; Swindells, 1995; Zehfus, 1995). We compared the structure model for recognition of hydrophobic core (left panel of Supplementary Fig. S6A) with that for identification of SSE packing clusters (right panel of Supplementary Fig. S6A). A hydrophobic core usually consists of densely packed hydrophobic residues within a protein domain. While in SSE packing clusters, only the residue pairs in effect contact are considered (see Section 2) and thus the hydrophobic residues involved are less densely packed (comparing left and right panels of Supplementary Fig. S6B). Moreover, as shown in Supplementary Figure S6C, for a protein domain with complex architecture, SSE packing clusters can give a better structural delineation than hydrophobic cores do.

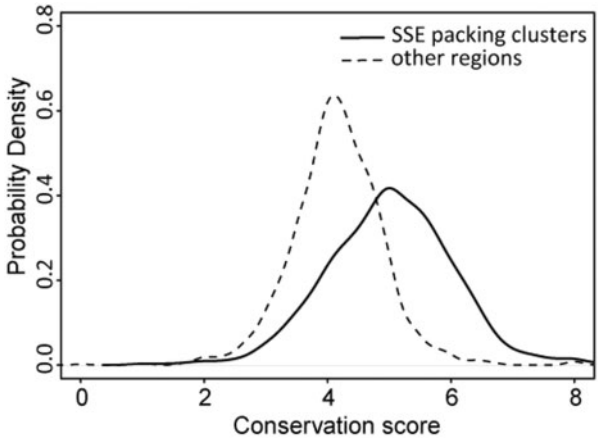

Fig. 4. SSE packing clusters are more conserved than other regions in domains. The distribution of conservation scores for SSE packing clusters within different families is displayed as a solid line, compared with that of other regions shown as a dash line

There arose an interesting phenomenon, when domain architectures were decomposed into SSE packing clusters. It was found that >80% of domains of different folds in SCOP could be related by sharing at least one SSE packing pattern, resulting in a high-connectivity network in protein fold space (Supplementary Fig. S3). The SSE packing clusters could also be derived based on CATH database, and we found that the recognition of SSE packing clusters is not drastically affected by different domain assign algorithm (see Supplementary Results for details). So, what puzzles us is why the SSE packing patterns are commonly shared among domains of different fold. Because the domains of different folds usually originate from different ancestors and perform different functions, the sharing of SSE packing clusters across domains of different fold could indicate the important role of physical constraints rather than evolutionary or functional constraints on formation of SSE packing clusters as common structural scaffolds.

Furthermore, we observed a high correlation between the SSE packing clusters and the folding nuclei for single-domain proteins. As most of the small single-domain proteins used in this study was found to fold via a nucleation-condensation mechanism (Nölting and Agard, 2008; Wetlaufer, 1973), several key residues would form a folding nucleus that serves as a scaffold for directing the rapid folding of the whole domain. The good correspondence between SSE packing clusters identified and folding nuclei for these single-domain proteins suggest the important roles of the SSE packing clusters in initiating folding and organizing structure elements. Moreover, through investigation of the distribution of 729 disease-related mutations on 16 different

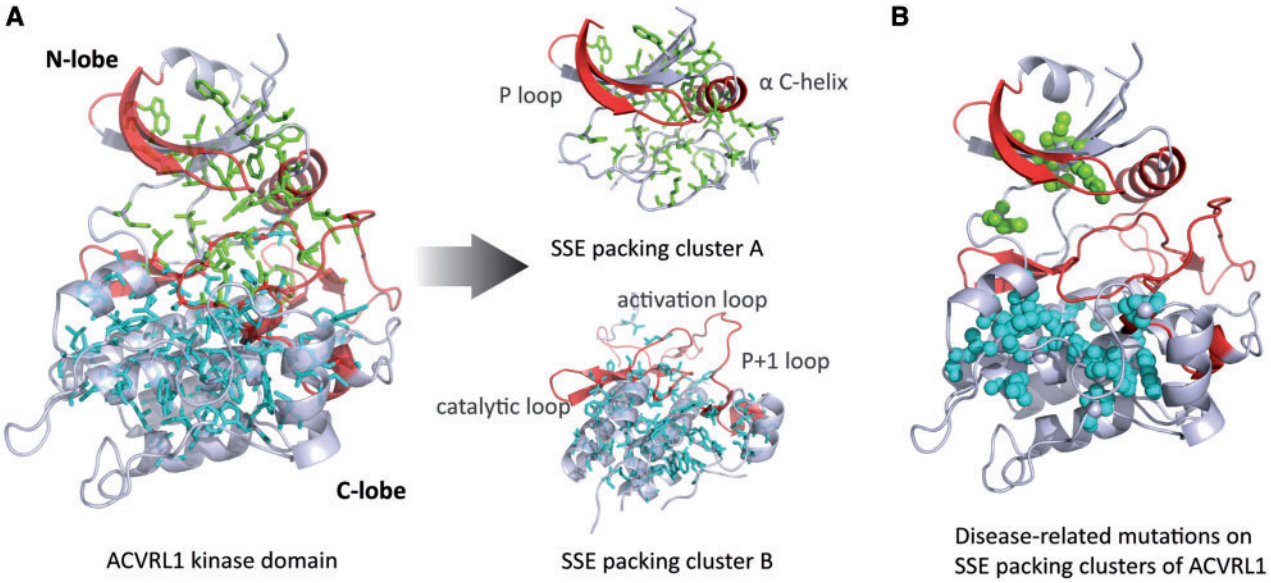

Fig. 5. Disease-related mutations on SSE packing clusters of kinase domains. (**A**) The two SSE packing clusters of ACVRL1 kinase domain. The SSE packing cluster at C-lobe is colored in cyan, and the SSE packing cluster at N-lobe is colored in green. Residues underlying SSE packing clusters are shown as side chains. The functional subdomains of kinases are colored in red. (**B**) The disease-related mutations on residues in SSE packing clusters (PC sites) of ACVRL1 kinase domain. The PC sites with disease mutations are shown as spheres

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

kinase domains, it was observed that SSE packing clusters could account for ~29.2% (213 of 729) of previously reported disease-related mutations on kinases. The association between SSE packing clusters and diseases reinforces the importance of SSE packing clusters in domain organization not only in terms of structural scaffolds but also in terms of functionality.

5 CONCLUSION

In this work, by considering the SSE interfaces as the basic interaction units for the packing of SSEs, we developed a new descriptor named SSE packing cluster to characterize domains of different fold. Moreover, our analysis showed that SSE packing clusters were more conserved in domain architectures and corresponded well with folding nuclei in small single-domain proteins. Therefore, SSE packing clusters could be regarded as common structural scaffolds underlying the diverse protein structure space.

ACKNOWLEDGMENTS

We thank all members of Jiang lab for help and discussions. We also thank HPC-Service Station in Center for Biological Imaging, Institute of Biophysics, Chinese Academy of Sciences for providing computation service.

Funding: This study was supported by National Program on Key Basic Research Project of China (973 Program) (2014CB910500), Major National earmark Project for Infectious Diseases (2013ZX10004611-002 to T.J.) and National Natural Science Foundation of China (31100950 to A.W.).

Conflict of Interest: none declared.

REFERENCES

Alexeevski,A. *et al.* (2003) CluD, a program for determination of hydrophobic clusters in 3D structures of protein and protein-nucleic acid complexes. *Biophysics*, **48**, 146–156.

Aung,Z. and Li,J. (2007) Mining super-secondary structure motifs from 3D protein structures: a sequence order independent approach. *Genome Inform.*, **19**, 15–26.

Bandyopadhyay,D. *et al.* (2009a) Identification of family-specific residue packing motifs and their use for structure-based protein function prediction: I. Method development. *J. Comput. Aided Mol. Des.*, **23**, 773–784.

Bandyopadhyay,D. *et al.* (2009b) Identification of family-specific residue packing motifs and their use for structure-based protein function prediction: II. *Case studies and applications. J. Comput. Aid. Mol. Des.*, **23**, 785–797.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bloom,J.D. *et al.* (2006) Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.*, **23**, 1751–1761.

Branden,C. and Tooze,J. (1991) *Introduction to Protein Structure*. Garland, New York.

Brennan,R. and Matthews,B.W. (1989) The helix-turn-helix DNA binding motif. *J. Biol. Chem.*, **264**, 1903–1906.

Chandonia,J.M. *et al.* (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.

Chiti,F. *et al.* (1999) Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.*, **6**, 1005–1009.

Comin,M. *et al.* (2008) Mining overrepresented 3D patterns of secondary structures in proteins. *J. Bioinform. Comput. Biol.*, **6**, 1067–1087.

Day,R. *et al.* (2010) Characterizing the regularity of tetrahedral packing motifs in protein tertiary structure. *Bioinformatics*, **26**, 3059–3066.

De Vos,A.M. *et al.* (1992) Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science*, **255**, 306–312.

Fersht,A.R. and Sato,S. (2004) Φ-Value analysis and the nature of protein-folding transition states, *Proc. Natl Acad. Sci.USA*, **101**, 7976–7981.

Goonesekere,N.C. and Lee,B. (2008) Context-specific amino acid substitution matrices and their use in the detection of protein homologs. *Proteins*, **71**, 910–919.

Hanks,S.K. and Hunter,T. (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.*, **9**, 576–596.

Harrison,S.C. and Aggarwal,A.K. (1990) DNA recognition by proteins with the helix-turn-helix motif. *Ann. Rev.Biochem.*, **59**, 933–969.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Heringa,J. and Argos,P. (1991) Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol.*, **220**, 151–171.

Hubbard,M.J. and Cohen,P. (1993) On target with a new mechanism for the regulation of protein phosphorylation. *Trends Biochem. Sci.*, **18**, 172–177.

Hubbard,S. (1992) NACCESS: A program for calculating accessibilities. Department of Biochemistry and Molecular Biology, University College of London. http://wolf.bms.umist.ac.uk/naccess/.

Johnson,S.C. (1967) Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kannan,N. and Vishveshwara,S. (1999) Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.*, **292**, 441–464.

Lahiry,P. *et al.* (2010) Kinase mutations in human disease: interpreting genotype–phenotype relationships. *Nat. Rev. Genet.*, **11**, 60–74.

Lesk,A.M. (2001) *Introduction to Protein Architecture: The Structural Biology of Proteins*. Oxford University Press, London.

McLachlan,A. (1971) Tests for comparing related amino-acid sequences. *Cytochrome c and cytochrome c 551. J. Mol. Biol.*, **61**, 409.

Minami,S. *et al.* (2013) MICAN: a protein structure alignment algorithm that can handle Multiple-chains, Inverse alignments, C α only models, alternative alignments, and Non-sequential alignments. *BMC Bioinformatics*, **14**, 24.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nölting,B. and Agard,D.A. (2008) How general is the nucleation–condensation mechanism? *Proteins*, **73**, 754–764.

Ozkan,S.B. *et al.* (2001) Transition states and the meaning of φ-values in protein folding kinetics. *Nat. Struct. Biol.*, **8**, 765–769.

Preißner,R. *et al.* (1998) Dictionary of interfaces in proteins (DIP). Data bank of complementary molecular surface patches. *J. Mol. Biol.*, **280**, 535–550.

Rao,S.T. and Rossmann,M.G. (1973) Comparison of super-secondary structures in proteins. *J. Mol. Biol.*, **76**, 241–256.

Richards,F.M. and Kundrot,C.E. (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins*, **3**, 71–84.

Rossmann,M.G. and Argos,P. (1976) Exploring structural homology of proteins. *J. Mol. Biol.*, **105**, 75–95.

Sillitoe,I. *et al.* (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.*, **41**, D490–D498.

Swindells,M.B. (1995) A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci.*, **4**, 93–102.

Wetlaufer,D.B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl Acad. Sci. USA*, **70**, 697–701.

Zehfus,M.H. (1995) Automatic recognition of hydrophobic clusters and their correlation with protein folding units. *Protein Sci.*, **4**, 1188–1202.

Zehfus,M.H. (1997) Identification of compact, hydrophobically stabilized domains and modules containing multiple peptide chains. *Protein Sci.*, **6**, 1210–1219.

Zhou,H. and Skolnick,J. (2007) Ab initio protein structure prediction using chunk-TASSER. *Biophys. J.*, **93**, 1510–1518.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016