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], Taijiao Jiang[1§]

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

*These authors contributed equally to this work
[§]Corresponding author

Correspondence should be addressed to TJ (tel/fax: 86-10-64888427, email: taijiao@moon.ibp.ac.cn or taijiao@aya.yale.edu).

Supplementary information: Supplementary materials, methods, results, figures and tables.

Supplementary Materials

Collected Φ-value dataset

In this research, we collected point mutations of 14 domains with detailed Φ-value studies in the literatures, including FADD DD (Steward, et al., 2009), protein L (Kim, et al., 2000), protein G (Hubner, et al., 2004), Im7 (Fowler and Clarke, 2001), Fnfn10 (Cota, et al., 2001), Acp (Chiti, et al., 1999), barnase (Serrano, et al., 1992), ubiquitin (Went and Jackson, 2005), CI2 (Itzhaki, et al., 1995), SH3 (Northey, et al., 2002), FKBP12 (Fulton, et al., 1999), L23 (Hedberg and Oliveberg, 2004), CTL9 (Li, et al., 2007) and CspB (Garcia-Mira, et al., 2004). This dataset covered most of now-existing data of Φ-value experiments. Besides, it covered representative structural classes of small protein, including all-α, all-β and α/β. In total, 170 PC sites with Φ-value records were collected (Table S5).

Collected kinase domain dataset

A set of 16 kinase domains in KinMutBase (Ortutay, et al., 2005), which have high-resolution crystal structures (resolution ≤ 3Å), were selected (Table S6). The disease-related missense mutations on these kinase domains were retrieved from KinMutBase, Swiss-Prot Variant Page (Yip, et al., 2004) and COSMIC (Catalogue Of Somatic Mutations In Cancer) (Forbes, et al., 2008). Finally, 729 non-redundant disease-related mutations on these kinase domains were collected (Table S7).

Supplementary Methods

Parametric testing

In our method, two parameters, namely residue contact cutoff (16 Å in our study) and minimum effective contacts (minimum two effective contacts in our study), may have influences on decomposing domain architectures into SSE packing clusters. Therefore, we performed a series of parameter testing on residue contact cutoff and minimum effective contacts.

To test the effect of residue contact cutoffs on SSE packing clusters, we performed domain-by-domain comparisons of SSE packing clusters at cutoffs 14-18 Å with an interval of 0.5 Å to those at 16 Å. In brief, the 9015 domains were decomposed into SSE packing clusters at different residue contact cutoff and the SSE packing clusters were compared to those at 16 Å at domain level. A domain was regarded to have similar packing clusters if the following three criteria were satisfied: (i) same number of derived SSE packing clusters; (ii) same type of secondary structures for the corresponding SSE packing cluster; (iii) 90% of residues overlapped between the corresponding SSE packing clusters. As shown in Fig. S7A, the cutoffs 14-18Å reproduced over 70% of the results by the cutoff 16 Å.

We also tested the effect of minimum effective contacts on the recognition of SSE packing clusters. It was found that the original results (by using minimum two site contacts) could be well reproduced when we applied minimum three (86.3%) or four site contacts (74.6%) in generating SSE packing clusters (Fig. S7B).

Taken together, the recognitions of SSE packing clusters were not drastically affected by the changes of the parameters used in our study.

Comparison between SSE packing clusters and hydrophobic cores

Here we compared our method with CluD (Alexeevski, et al., 2003), an excellent method used for recognition of hydrophobic cores in protein structures

(http://mouse.belozersky.msu.ru/npidb/cgi-bin/hftri.pl). In CluD, a protein structure is reduced into a set of points, with each points representing a non-polar group of a residue (Fig. S6A, left panel). The basic idea of CluD is that the non-polar groups in a cluster should have many more interactions inside the cluster than between different clusters. Compared to CluD, our method employs a more detailed structure model, which simultaneously takes into consideration of main chain, side chains and secondary structure elements (SSEs) of a protein (Fig. S6A, right panel). This detailed structure model possesses two potential advantages in exploring domain architectures: Firstly, the consideration of main chain makes the description of residue-residue interaction more accurate, which could discriminate effective contacts from non-effective contacts between side chains as shown in Fig. S1C; secondly, the concept of SSE interfaces, which makes use of the characteristics of regular SSEs in organizing local conformations, simplifies the complexity of domain architectures by focusing on the topological arrangement of SSEs.

Based on the comparisons of hydrophobic cores and SSE packing clusters for same domains, It is found that the recognized SSE packing clusters are similar to the recognized hydrophobic cores for simple domain architectures. For example, the SSE packing cluster of d12e8h1, whose main chain forms only one enclosure space from a topological perspective, is similar with the derived hydrophobic core (Fig. S6B). Of the totally 38 residues involved in SSE packing cluster of d12e8h1, 35 are also included in the hydrophobic core recognized by CluD. However, for domains with complex architectures such as kinase domains, SSE packing clusters seem to better characterize the domain organization compared to hydrophobic cores. As shown in Fig. S6C, the RET kinase domain is found to be with only one hydrophobic cluster; In contrast, three SSE packing clusters were derived from RET kinase domain. According to the functional map of kinase domain, the biggest SSE packing cluster (colored in green) corresponds well with the C-lobe of kinase domain, and may serve as the structure scaffold for SD VIB (catalytic loop), SD VII (activation loop) and SD VIII (P+1 loop); the two smaller SSE packing clusters correspond to the N-lobe, and may serve as structure scaffold for SD I (P-loop) and SD III (α C-helix). Therefore, SSE packing clusters seem to reveal more details about the domain organization of RET kinase domain than hydrophobic cores do.

Measurement of the effect size for different physicochemical features associated with SSE packing clusters

To further investigate the specific physicochemical features underlying the conservation of SSE packing clusters, four physicochemical properties, including side-chain volume (small, medium and big volume), side-chain shape (linear, branched and circled shape), polarity (polar and non-polar) and hydropathy (hydrophobic, hydrophilic and amphipathic) (see Table S8) were investigated. The charge property was not considered due to its low frequency in SSE packing clusters (see Figure S8). Given a pair of domains belonging to the same family, the structurally equivalent sites were derived by performing MICAN structure alignment.

For each pair of aligned sites, the identity score was assigned a value of 1 if the two residues belonged to the same state for a physiochemical property; otherwise, the identity score was assigned a value of 0. Then the identity score for the aligned regions of their SSE packing clusters was calculated as the average identity score of their aligned site pairs. Similarly, the identity score was calculated for the regions other than SSE packing clusters. To find out the dominant features associated with SSE packing clusters, we looked into the magnitude of identity difference (effect size) as measured by Cohen's d (Cohen, 1988). Cohen's d is defined as the difference between two means divided by a standard deviation of two groups of data (SDpooled), which is an appropriate effect size for the comparison between two means. It was calculated as follows:

$$d = (M_{group1} - M_{group2})/SD_{pooled}$$

$$SD_{pooled} = \sqrt{(SD^2_{group1} + SD^2_{group2})/2}$$

Usually, $d<0.2$ means a small effect size, $0.2 \leq d<0.8$ means a medium effect size and $d \geq 0.8$ means a large effect size. By calculating the Cohen's d for different physicochemical features, the feature with largest effect size would be regarded as the most dominant features associated with SSE packing clusters.

Detecting significant amino acid patterns underlying SSE packing patterns

In our study, 576 of 1160 SSE packing patterns were found across different folds. Since tertiary structure is encoded in primary sequence, we wonder whether there exist certain amino acid configurations for these common used SSE packing patterns. To answer this question, 21 SSE packing patterns with over 100 packing cluster members were selected. For each of these SSE packing patterns, the multiple sequence alignment of its members was derived by using MUSCLE program (v3.8.31) (Edgar, 2004); based on the multiple sequence alignment, well-aligned columns with at most 10% gap were retained and used as the foreground set for plogo program (version 1.2.0, http://plogo.uconn.edu/) (O'Shea, et al., 2013), a probabilistic approach to visualizing sequence motifs; By employing human proteome as background set, those amino acids with statistical significance ($p<0.05$) were combined into a sequence pattern. A significant sequence pattern should contain at least five amino acids with heights over the statistical significance ($p<0.05$).

Supplementary Results

Conservation measurement based on different types of substitution matrices

Based on the conservation measurement by McLachlan matrix, SSE packing clusters were found to be more conserved than other regions in domains in our study (Fig. S5A). To test the robustness of conservation analysis, we performed another two conservation analysis based on BLOSUM62 matrix (Henikoff and Henikoff, 1992) and CSSM matrices (context-specific amino acid substitution matrices) (Goonesekere and Lee, 2008).

When using BLOSUM62 matrix, the average conservation score for SSE packing clusters is 1.71, which is significant higher than the average conservation score (0.81) for other regions (t-test, p-value<2.2E-16) (Fig. S5B).

Different with McLachlan matrix and BLOSUM62 matrix, CSSM matrices belong to context-specific amino acid substitution matrices, with consideration of structure context in which the substitution takes place. CSSM matrices consist of a set of four matrices, each of which is specific for a specified range of the environment polarity (EP) of the residue being substituted. The environment polarity of the residue being substituted could be calculated by using SHEBA program (version 3.1.1) (Jung and Lee, 2000). An amino acid substitution matrix was selected separately for each EP range of 0%–25%, 25%–50%, 50%–75%, and 75%–100%. Fig. S5C shows the distributions of conservation score for SSE packing clusters and other regions. On average, the conservation score for SSE packing clusters (1.72) is still significantly higher than that for other regions in domains (0.96) (t-test, p-value<2.2E-16).

The volume and surface area of SSE packing clusters

To investigated the geometric characteristics of SSE packing clusters, the solvent excluded volume and surface area of an SSE packing cluster was calculated with MSMS program (version 2.6.1, probe radius was set as 1.5) (Sanner, et al., 1996). It is found that the preferential volume and surface area of SSE packing clusters locate at 9300.1 $Å^3$ and 3790.7 $Å^2$ (Fig. S9), and the peak of surface-to-volume ratio (SVR) of SSE packing clusters appear at 0.375 $Å^{-1}$.

Large-scale analysis of packing clusters on CATH domains

The two popular domain structure databases, SCOP (Murzin, et al., 1995) and CATH (Sillitoe, et al., 2013), adopt different criteria for defining protein structural domains. In our study, large-scale analysis of domain organization were performed on the ASTRAL SCOP 40 database. To assure that the recognition of SSE packing clusters is not drastically affected by different domain assignment algorithms, we repeated the

analysis of domain organization on CATH S35 , a non-redundant CATH domain database with at most 35% sequence identity between any two sequences. After removing the domains with sequence length below 30 amino acid, missing main chain atoms or low resolution (>3 Å) , 7703 domains were collected in final dataset and used for large-scale analysis of domain organization. Then we mainly compared the statistical characteristics of SSE packing clusters between CATH domain database and SCOP domain database.

12721 SSE packing clusters were derived from 7703 CATH domains. Most domains (96.1%) contain one to three SSE packing clusters (Fig. S10A) (compared to 90.4% in SCOP domains). These SSE packing clusters mostly (81.4%) consist of 2~8 regular SSEs with a peak at five SSEs (Fig. S10B) (compared to 77.8% and the same peak at five SSEs in SCOP domains). The number of residues also exhibits a unimodel distribution with the peak appearing in the range of 17 to 24 sites (Fig. S10C) (the peak is similar to that in SCOP domains). The peaks of solvent excluded volume and surface area for SSE packing clusters in domains appears at 9168.3 $Å^3$ (Fig. S10D) and 3746.0 $Å^2$ (Fig. S10E) respectively, which are also comparable to 9300.1 $Å^3$ and 3790.7 $Å^2$ for SCOP domains.

Based on the geometric similarity between SSE packing clusters, the 12721 SSE packing clusters of CATH domains were clustered into 5014 classes, so called SSE packing patterns. Among the 5014 SSE packing patterns, 892 SSE packing patterns contained 2 or more packing clusters, which covered 67.6% (8599) of SSE packing clusters (compared to 69% for SCOP40 domains). In CATH domains, the 7703 domains could also be classified into 996 topologies (the level topology in CATH corresponds to the level of fold in SCOP (Csaba, et al., 2009)), and 429 of 892 SSE packing patterns were found across different CATH topologies (Fig. S10F). The most commonly used SSE packing pattern spanned 504 domains across 91 different topologies. If we connect those pairs of topologies sharing an SSE packing pattern, this will also result in a network of high connectivity in protein fold space. Specifically, 647 different topologies were involved in the network.

Next, we also investigated the conservation of SSE packing clusters by randomly select 790 pairs of domains belonging to the same CATH homologous super-family. The degree of conservation was measured based on different amino acid substitution matrices. Specifically, when the conservation score was measured by McLachlan matrix, the average conservation score for SSE packing cluster is 4.91, which is significantly higher than the average conservation score of 4.10 for other regions in domains (t-test, p-value<2.2e-16). For BLOSUM62 matrix, the average conservation score for SSE packing clusters (1.57) is also significantly higher than that for other regions (0.83) (t-test, p-value<2.2e-16); and for CSSM matrices, SSE packing clusters are still found to be more conserved in domains (1.59 vs. 0.94, t-test, p-value<2.2e-16).

Taken together, the statistics, classification and conservation of packing clusters are quite similar between SCOP domains and CATH domains, indicating that our analysis of packing clusters is not drastically affected by different domain assignment algorithm.

The dominant physicochemical features associated with SSE packing clusters

We sought to identify more specific physicochemical features associated with SSE packing clusters by considering their physicochemical identity within SSE packing patterns compared to other regions (see Supplementary Methods). Four physicochemical properties were investigated, including side chain volume, side chain shape, polarity and hydropathy. As shown in Supplementary Figure S11, all the four properties demonstrated a higher identity in SSE packing clusters than other regions (see Table S9 for details). When looking into the magnitude of identity difference (effect size) as measured by Cohen's d (Cohen, 1988) (see Methods), we found that polarity, hydropathy, and side-chain shape showed a significantly large effect size with Cohen's d > 0.8 while side-chain volume only had a Cohen's d of 0.69 (see Table S10 for details). In particular, polarity was found to be with the largest Cohen's d of 1.52, suggesting its important role in the formation of SSE packing clusters.

Data availability

The SSE packing clusters derived from SCOP domains could be downloaded from http://jianglab.ibp.ac.cn/lims/download/ASTRALS40_DOMAIN_PC.tar.gz.

The SSE packing clusters derived from CATH domains could be downloaded from http://jianglab.ibp.ac.cn/lims/download/CATHS35_DOMAIN_PC.tar.gz.

Supplementary 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.

Chiti, F., *et al.* (1999) Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding, *Nature Structural & Molecular Biology*, **6**, 1005-1009.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciencies*. Routledge.

Cota, E., *et al.* (2001) The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold, *Journal of molecular biology*, **305**, 1185-1194.

Csaba, G., Birzele, F. and Zimmer, R. (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis, *BMC structural biology*, **9**, 23.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic acids research*, **32**, 1792-1797.

Forbes, S., *et al.* (2008) The catalogue of somatic mutations in cancer (COSMIC), *Current protocols in human genetics*, 10.11. 11-10.11. 26.

Fowler, S.B. and Clarke, J. (2001) Mapping the folding pathway of an immunoglobulin domain: structural detail from phi value analysis and movement of the transition state, *Structure*, **9**, 355-366.

Fulton, K.F., *et al.* (1999) Mapping the interactions present in the transition state for unfolding/folding of FKBP12, *Journal of molecular biology*, **291**, 445-461.

Garcia-Mira, M.M., Boehringer, D. and Schmid, F.X. (2004) The folding transition state of the cold shock protein is strongly polarized, *Journal of molecular biology*, **339**, 555-569.
Goonesekere, N.C. and Lee, B. (2008) Context - specific amino acid substitution matrices and their use in the detection of protein homologs, *Proteins: Structure, Function, and Bioinformatics*, **71**, 910-919.
Hedberg, L. and Oliveberg, M. (2004) Scattered Hammond plots reveal second level of site-specific information in protein folding: $\Phi'(\beta\ddagger)$, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 7606-7611.
Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks, *Proceedings of the National Academy of Sciences*, **89**, 10915-10919.
Hubner, I.A., Shimada, J. and Shakhnovich, E.I. (2004) Commitment and nucleation in the protein G transition state, *Journal of molecular biology*, **336**, 745-761.
Itzhaki, L.S., Otzen, D.E. and Fersht, A.R. (1995) The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding, *Journal of molecular biology*, **254**, 260-288.
Jung, J. and Lee, B. (2000) Protein structure alignment using environmental profiles, *Protein Engineering*, **13**, 535-543.
Kim, D.E., Fisher, C. and Baker, D. (2000) A breakdown of symmetry in the folding transition state of protein L, *Journal of molecular biology*, **298**, 971-984.
Li, Y.*, et al.* (2007) Mutational analysis of the folding transition state of the C-terminal domain of ribosomal protein L9: A protein with an unusual β-sheet topology, *Biochemistry*, **46**, 1013-1021.
Murzin, A.G.*, et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *Journal of molecular biology*, **247**, 536-540.
Northey, J.G., Di Nardo, A.A. and Davidson, A.R. (2002) Hydrophobic core packing in the SH3 domain folding transition state, *Nature Structural & Molecular Biology*, **9**, 126-130.
O'Shea, J.P.*, et al.* (2013) pLogo: a probabilistic approach to visualizing sequence motifs, *Nature methods*, **10**, 1211-1212.
Ortutay, C.*, et al.* (2005) KinMutBase: A registry of disease - causing mutations in protein kinase domains, *Human mutation*, **25**, 435-442.
Sanner, M.F., Olson, A.J. and Spehner, J.C. (1996) Reduced surface: an efficient way to compute molecular surfaces, *Biopolymers*, **38**, 305-320.
Serrano, L., Matouschek, A. and Fersht, A.R. (1992) The folding of an enzyme: III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure, *Journal of molecular biology*, **224**, 805-818.
Sillitoe, I.*, et al.* (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures, *Nucleic acids research*, **41**, D490-D498.
Steward, A., McDowell, G.S. and Clarke, J. (2009) Topology is the principal determinant in the folding of a complex all-alpha Greek key death domain from human FADD, *Journal of molecular biology*, **389**, 425-437.
Went, H.M. and Jackson, S.E. (2005) Ubiquitin folds through a highly polarized transition state, *Protein Engineering Design and Selection*, **18**, 229-237.
Yip, Y.L.*, et al.* (2004) The Swiss - Prot variant page and the ModSNP database: A resource for sequence and structure information on human protein variants, *Human mutation*, **23**, 464-470.

Supplementary Figures

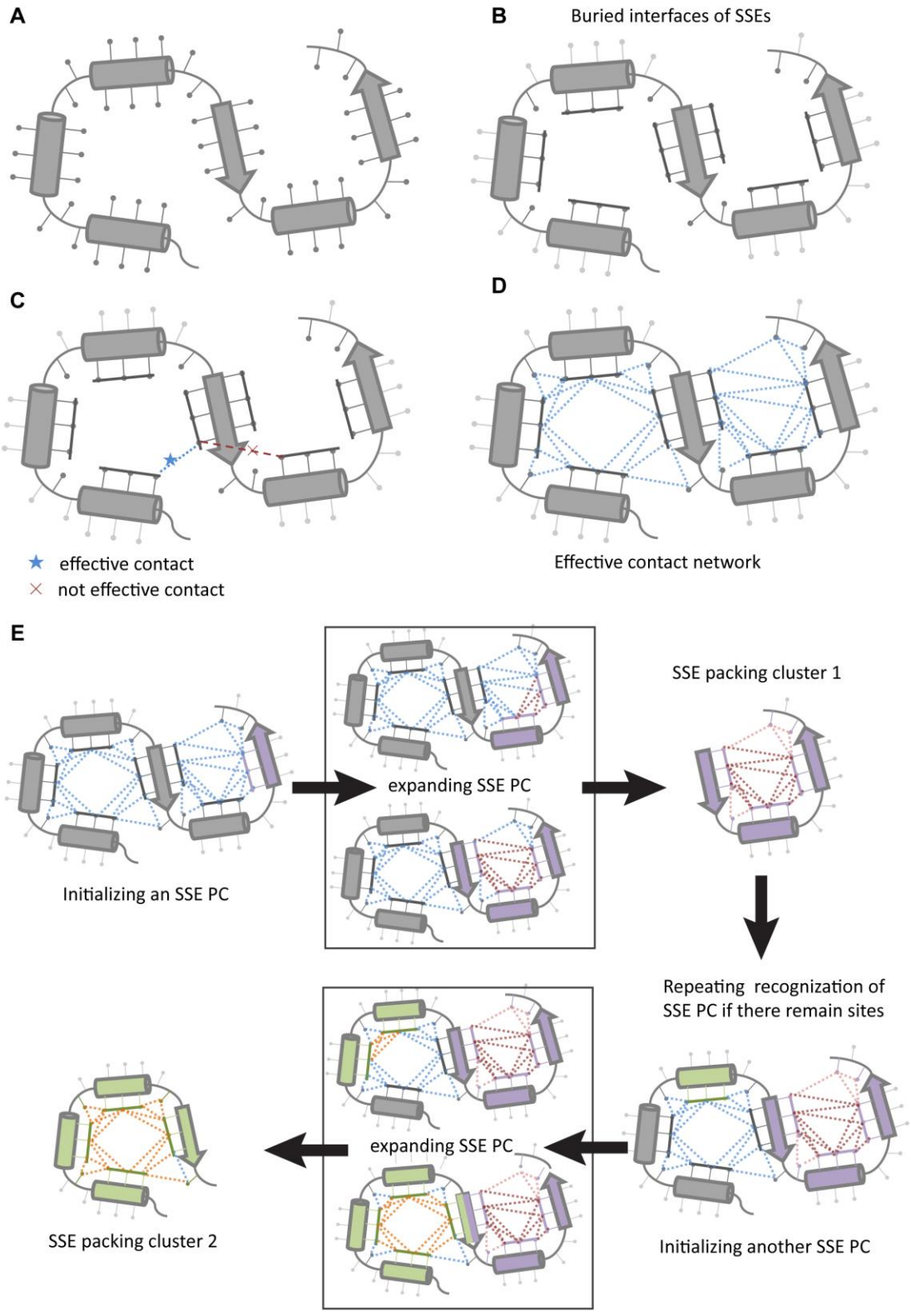


Fig. S1. The schematic diagram of decomposing protein domains into SSE packing clusters. (A) The simplified representation of the main chain, side chains and secondary structure elements (SSEs) of a domain. (B) Identification of buried interfaces of SSEs. The buried interfaces of regular SSEs (α-helix and β-strand) are shown as dark lines. (C) The definition of an effective contact between two buried sites. Two buried sites within a given distance (16Å used in our study) are considered to be in effective contact if the line connecting the centroids of their side chains does not pass through any atoms of the main chain. (D) The effective contact network among all buried sites. (E) The schematic diagram of constructing SSE packing clusters (SSE PCs) based on SSE interfaces and effective contact network (see Methods for details).

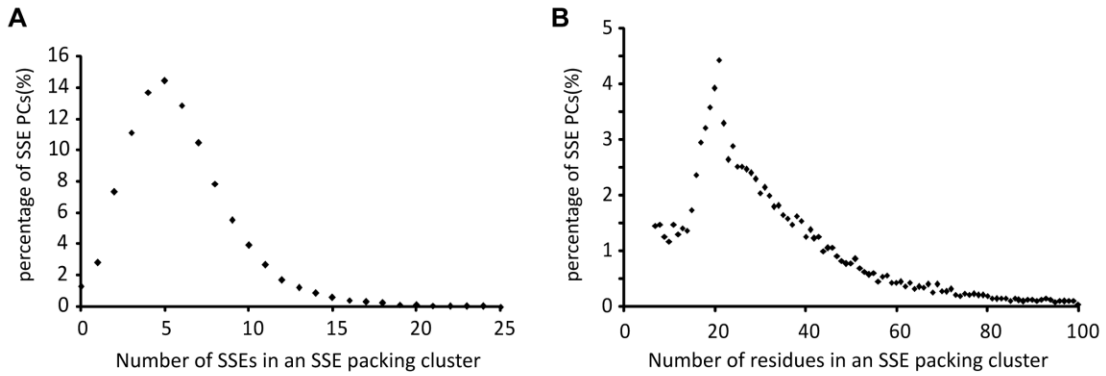


Fig. S2. The distributions of SSE number and residue number involved in an SSE packing cluster. (A) The distribution of SSE numbers within an SSE packing cluster. (B) The distribution of residue numbers within an SSE packing cluster.

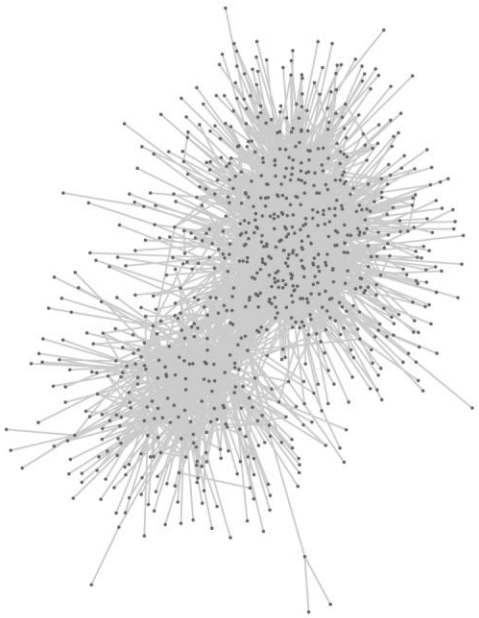
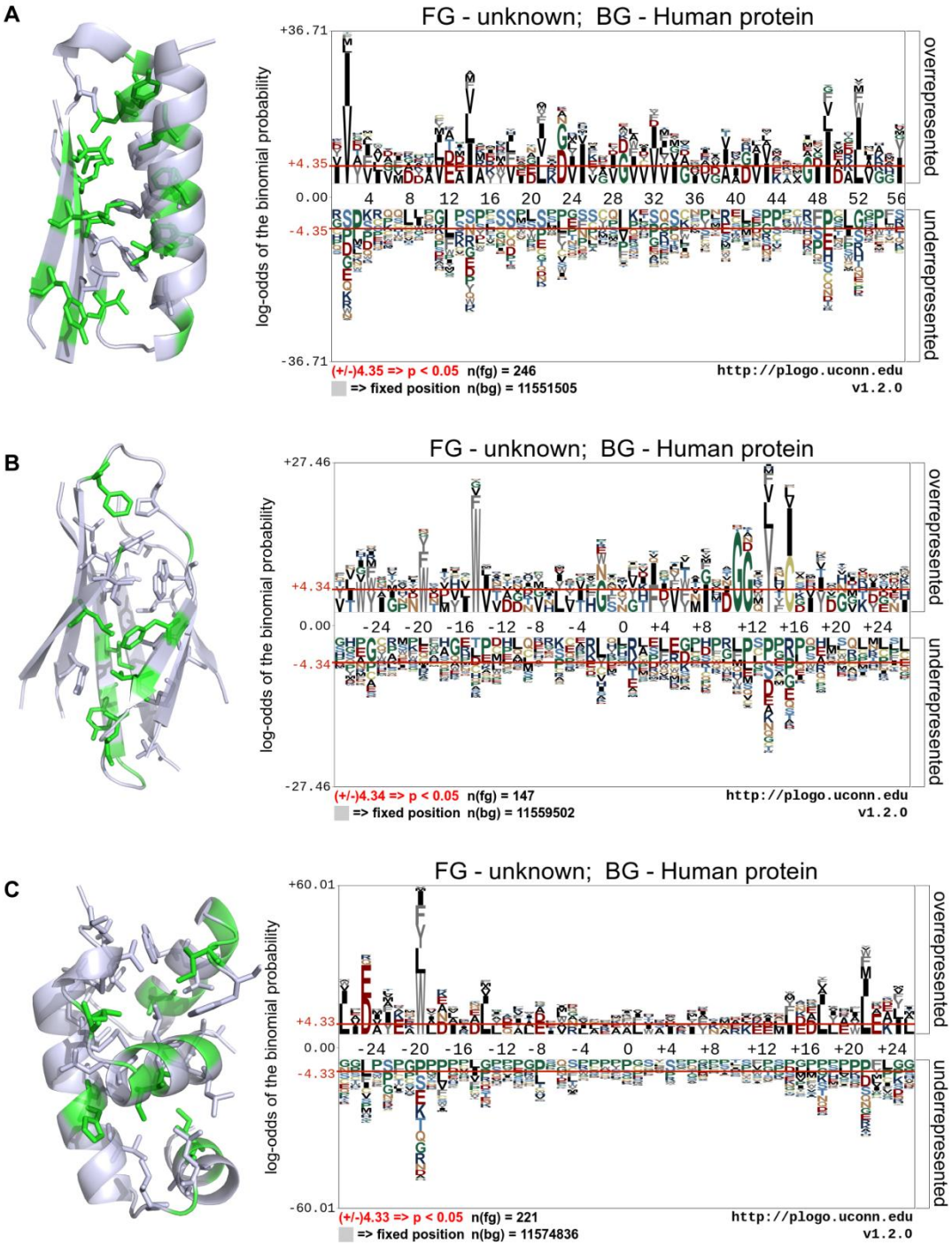


Fig. S3. The network diagram for those folds sharing an SSE packing pattern. Each point in the network represent a fold as defined in SCOP, and the edge between two folds means both of them share an SSE packing pattern as their lower structure units.


A
FG - unknown; BG - Human protein
log-odds of the binomial probability
overrepresented
underrepresented
+36.71
+4.35
0.00
-4.35
-36.71
4 8 12 16 20 24 28 32 36 40 44 48 52 56
(+/-)4.35 => p < 0.05 n(fg) = 246
=> fixed position n(bg) = 11551505
http://plogo.uconn.edu
v1.2.0
B
FG - unknown; BG - Human protein
log-odds of the binomial probability
overrepresented
underrepresented
+27.46
+4.34
0.00
-4.34
-27.46
-24 -20 -16 -12 -8 -4 0 +4 +8 +12 +16 +20 +24
(+/-)4.34 => p < 0.05 n(fg) = 147
=> fixed position n(bg) = 11559502
http://plogo.uconn.edu
v1.2.0
C
FG - unknown; BG - Human protein
log-odds of the binomial probability
overrepresented
underrepresented
+60.01
+4.33
0.00
-4.33
-60.01
-24 -20 -16 -12 -8 -4 0 +4 +8 +12 +16 +20 +24
(+/-)4.33 => p < 0.05 n(fg) = 221
=> fixed position n(bg) = 11574836
http://plogo.uconn.edu
v1.2.0

1

Fig. S4. The significant sequence patterns underlying popular SSE packing patterns. (A) (left panel) an alpha/beta SSE packing pattern which contains 318 SSE packing clusters within 305 domains belonging to 92 different folds. The residues involved in the SSE packing clusters are highlighted with side chains. The position with a significant amino acid found by plogo is colored in green. (right panel) visualization of sequence motifs in plogo: the red horizontal bar represents the p = 0.05 statistical significant threshold following Bonferroni correction. Residues are stacked from most to least significant. (B) an all-beta SSE packing pattern which contains 188 SSE packing clusters within 161 domains belonging to 54 different folds. (C) an all-beta SSE packing pattern which contains 274 SSE packing clusters within 271 domains belonging to 92 different folds. (See supplementary data for details)

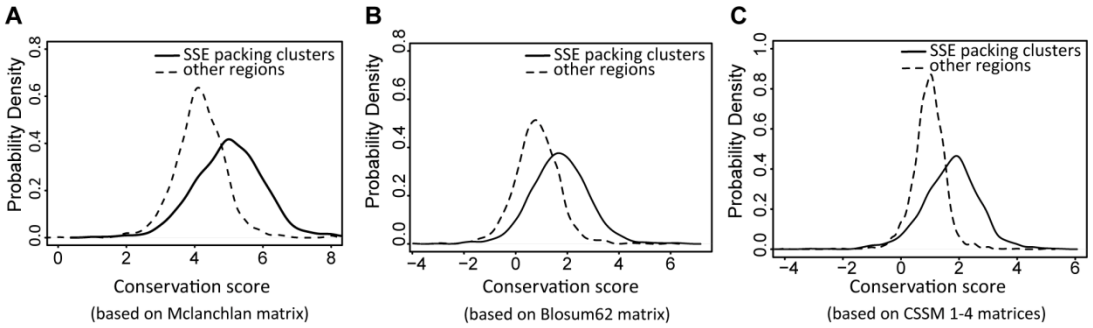


Fig. S5. SSE packing clusters are more conserved than other regions in domains. The conservation measurement were performed based on different amino acid substitution matrices, namely McLanchlan matrix (A), BLOSUM62 matrix (B) and CSSM matrices (C).

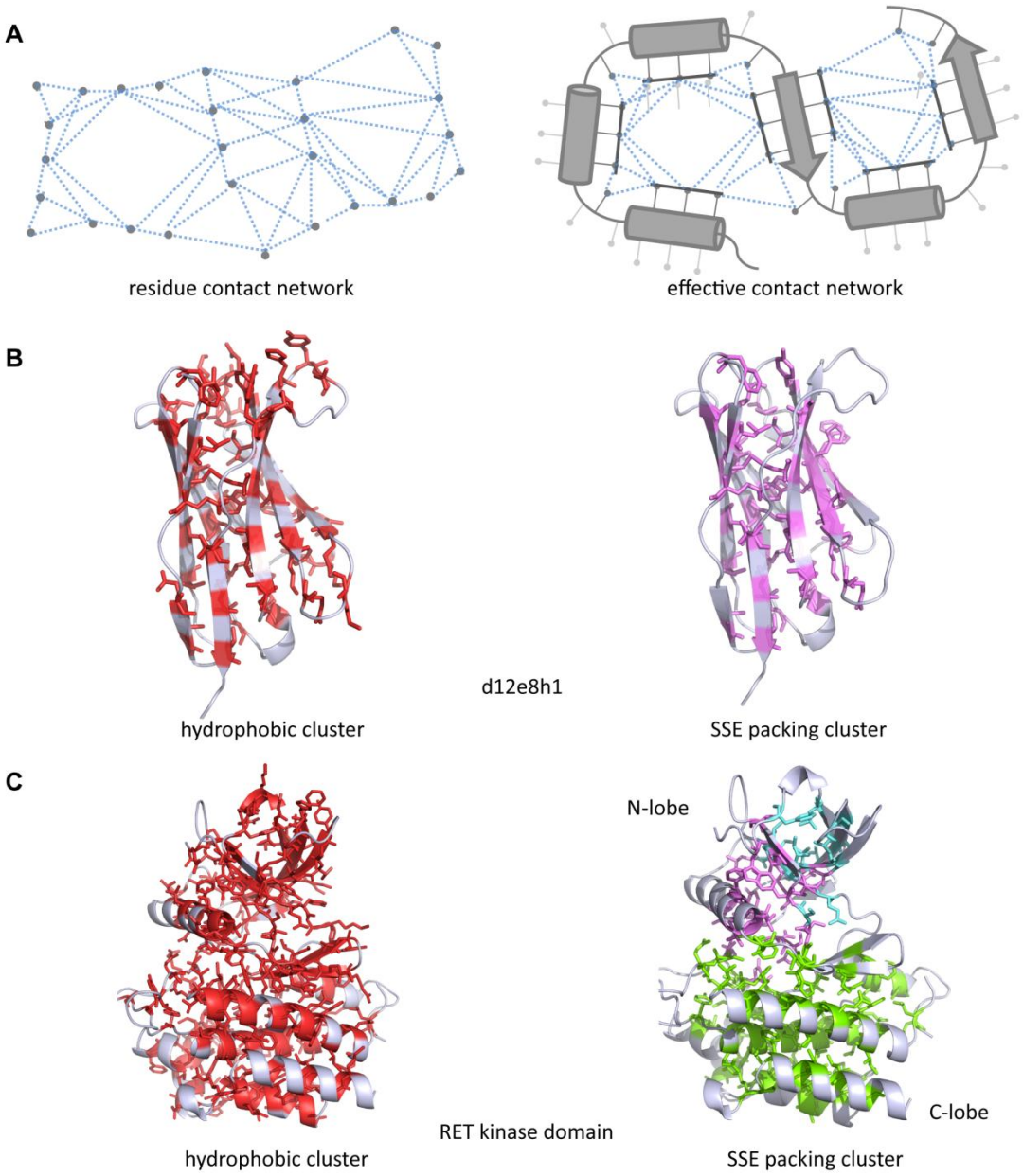


Fig. S6. Comparison between SSE packing clusters and hydrophobic clusters. (A) Comparison of the structure models used for recognition of hydrophobic core (left panel) and SSE packing clusters (right panel). (B) The recognized hydrophobic core and SSE packing cluster for a domain with simple architecture (d12e8h1). (C) The recognized hydrophobic core and SSE packing clusters for a domain with complex architecture (RET kinase domain).

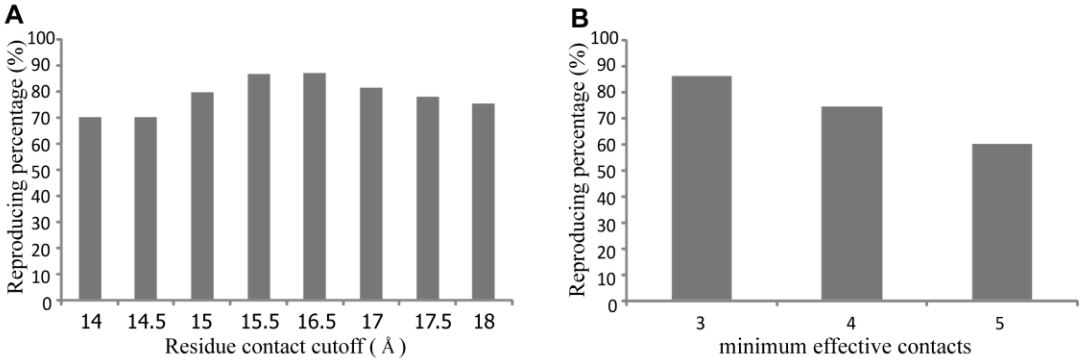


Fig. S7. The effect of residue contact cutoff and minimum site contacts on the recognition of SSE packing clusters. (A) The effect of residue contact cutoff on the recognition of SSE packing clusters. The x-axis is the residue contact cutoff (14-18 Å with an interval of 0.5 Å) used for the recognition of SSE packing clusters. The y-axis is the percentage of reproducing the results at the cutoff 16 Å. (B) The effect of minimum site contacts on recognition of SSE packing clusters. The x-axis is minimum site contacts used in generating SSE packing clusters. The y-axis is the percentage of reproducing the results by using minimum two site contacts.

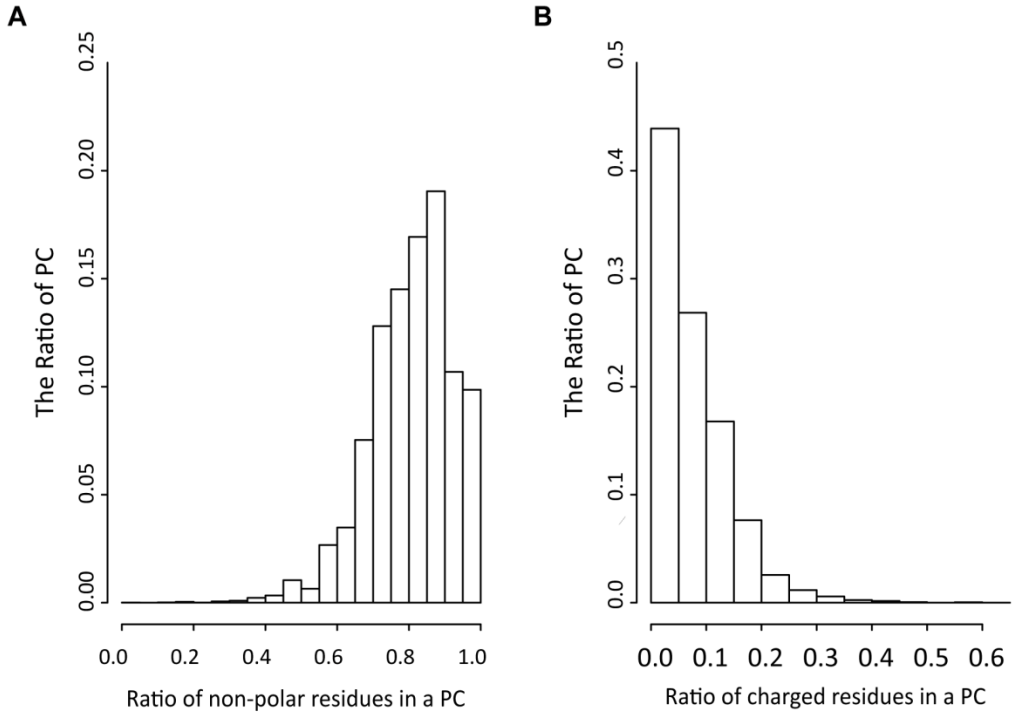


Fig. S8. Histograms for the distribution of non-polar residues and charged residues within SSE packing clusters. The ratio is calculated as the non-polar (or charged) residues against all the sites involved in an SSE packing cluster.

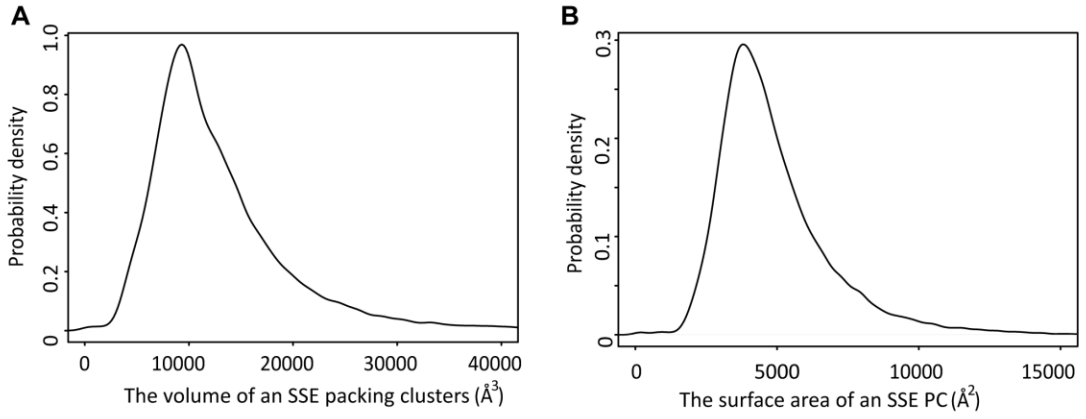


Fig. S9. The distribution of volume (A) and surface area (B) of SSE packing clusters.

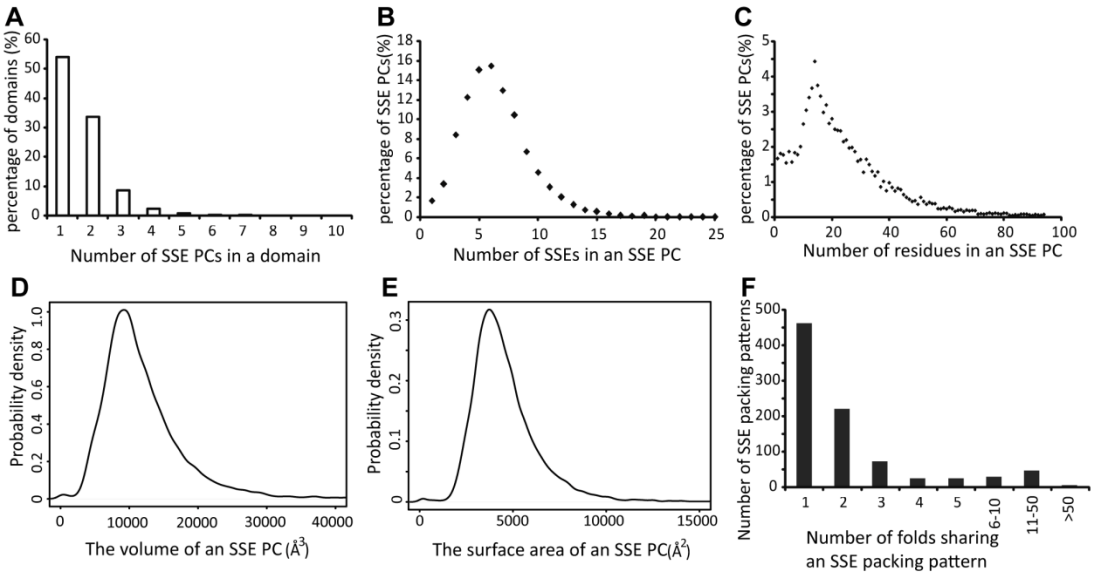


Fig. S10. The statistical characteristics of SSE packing clusters for CATH domains. (A) Histogram of the number of SSE packing clusters (SSE PCs) in a domain. (B) The distribution of SSE numbers within an SSE packing cluster. (C) The distribution of residue numbers within an SSE packing cluster. (D) The distribution of the volume of SSE packing clusters. (E) the distribution of the surface area of SSE packing clusters. (F) Histogram of the number of folds contained in an SSE packing pattern.

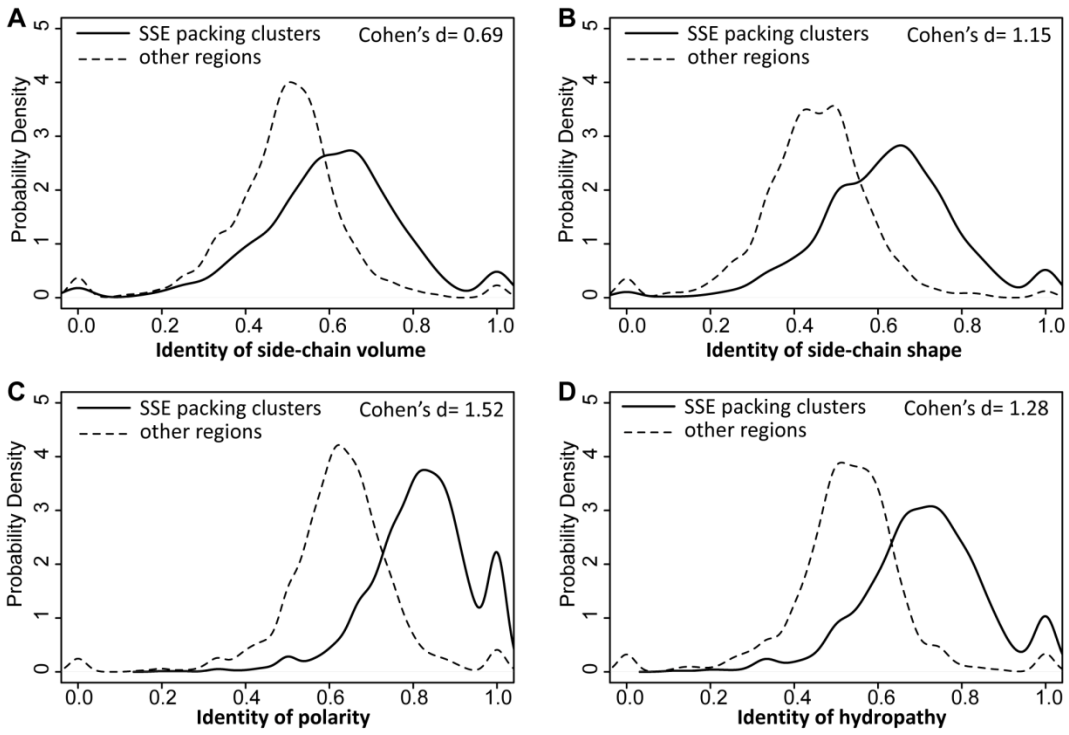


Fig. S11. The specific physicochemical features associated with SSE packing patterns and their effect sizes. Four physicochemical properties were considered, including side-chain volume (A), side-chain shape (B), polarity (C) and hydropathy (D). The distribution of specific property identities for SSE packing clusters within different families are displayed as solid lines, compared with that of other regions shown as dash lines.

Supplementary Tables

Table S1. The distribution of domains with different SSE packing clusters (SSE PC).

numbers of SSE PC within a domain	Count of domains	Percentage(%)
1	4180	46.367
2	2796	31.015
3	1175	13.034
4	537	5.957
5	170	1.886
6	94	1.043
7	41	0.455
8	15	0.166
9	3	0.033
10	3	0.033
11	1	0.011

Table S2. Φ-value distribution of the sites involved in SSE packing clusters (PC sites).

Index	Protein	PC Sites	PC Sites with Φ-value	PC Sites with Φ>0.1	Medium/high Φ-value Ratio
1	FADD DD	22	17	12	0.706
2	Protein L	12	10	9	0.900
3	Protein G	9	8	7	0.875
4	Im 7(TI 127)	21	18	15	0.833
5	Fnfn10	18	13	11	0.846
6	Acp	21	14	11	0.786
7	Barnase	34	13	10	0.769
8	Ubiquitin	18	17	9	0.529
9	CI2	12	10	7	0.700
10	SH3	17	8	7	0.875
11	FKBP12	25	10	10	1.000
12	L23	17	11	7	0.636
13	CTL9	22	13	9	0.692
14	CspB	15	8	5	0.625

Table S3. The overlap between kinase SSE packing clusters and the five functional subdomains for 16 kinase domains.

ID	PC sites[1]	Func sites[2]	overlap	Overlap ratio	
				Ratio against PC sites	Ratio against Func sites
ACVRL1	109	102	32	0.2935	0.3137
BTK	111	91	33	0.2972	0.3626
CHEK2	98	105	28	0.2857	0.2667
FGFR1	112	91	33	0.2946	0.3626
FGFR3	108	91	34	0.3148	0.3736
FLT3	136	91	48	0.3529	0.5275
JAK3	110	92	35	0.3181	0.3804
MERTK	100	70	24	0.2400	0.3429
MET	127	95	39	0.3070	0.4105
NTRK1	106	91	26	0.2452	0.2857
PHKG2	108	88	36	0.3333	0.4091
RET	115	91	38	0.3304	0.4176
ROR2	114	91	36	0.3158	0.3956
RPS6KA3	106	93	21	0.1981	0.2258
TEK	110	94	26	0.2364	0.2766
ZAP70	121	96	33	0.2727	0.3438

[1]Residues involved in SSE packing clusters are denoted as PC sites.

[2]residues belonging to functional subdomains of kinases (SD I, SD III, SD VIB, SD VII and SD VIII) are denoted as Func sites.

Table S4. The distribution of disease-related mutations (abbr. muts.) on kinase domains. The sites of a kinase domain are divided into three separate sets: i) sites within the five functional subdomains (catalytic regions); ii) sites belonging to SSE packing clusters but not belonging to catalytic regions (packing cluster regions); iii) sites beyond catalytic regions and packing cluster regions (other regions). For a specific region, the enrichment ratio of disease-relation mutations is calculated as the number of disease-relation mutations against the number of sites involved in this region.

ID	Total Muts.	Catalytic Regions			Packing cluster Regions			Other Regions		
		Muts.	Sites	Ratio	Muts.	Sites	Ratio	Muts.	Sites	Ratio
ACVRL1	73	27	102	0.27	28	77	0.36	18	116	0.16
BTK	143	46	91	0.50	66	78	0.85	31	95	0.33
CHEK2	44	24	105	0.23	7	70	0.10	13	108	0.12
FGFR1	38	17	91	0.19	11	79	0.14	10	108	0.09
FGFR3	27	16	91	0.18	3	74	0.04	8	128	0.06
FLT3	49	30	91	0.33	3	88	0.03	16	119	0.13
JAK3	17	7	92	0.08	5	75	0.07	5	114	0.04
MERTK	22	3	70	0.04	5	76	0.07	14	103	0.14
MET	60	29	95	0.31	15	88	0.17	16	118	0.14
NTRK1	54	20	91	0.22	11	80	0.14	23	89	0.26
PHKG2	9	3	88	0.03	5	72	0.07	1	124	0.01
RET	75	25	91	0.27	21	77	0.27	29	116	0.25
ROR2	42	12	91	0.13	9	78	0.12	21	100	0.21
RPS6KA3	30	10	93	0.11	9	85	0.11	11	121	0.09
TEK	15	4	94	0.04	2	84	0.02	9	121	0.07
ZAP70	31	8	96	0.08	13	88	0.14	10	98	0.10
Total	729	281	1472	0.191	213	1269	0.168	235	1778	0.132

Table S5. Details of 14 small single-domain proteins in collected Φ-value dataset.

Index	Protein	PDB ID	Class	Size	Mutation Sites
1	FADD DD	1E41	α	104	21
2	Protein L	2PTL	α/β	78	45
3	Protein G	1PGB	α/β	56	24
4	Im 7(TI 127)	1TIT	β	98	26
5	Fnfn10	1TTG	β	94	20
6	Acp	1APS	α/β	98	24
7	Barnase	1RNB	α/β	110	32
8	Ubiquitin	1UBQ	α/β	76	20
9	CI2	1CIQ	α/β	65	33
10	SH3	1FYN	β	62	8
11	FKBP12	1FKP	α/β	107	22
12	L23	1N88	α/β	96	16
13	CTL9	1DIVC	α/β	92	18
14	CspB	1CSP	β	67	16

Table S6. The structure information of 16 kinase domains used in this study. TK: Tyrosine protein kinase; STK: Serine/Threonine Protein Kinase.

ID	Type	PDB ID	Resolution (Å)
ACVRL1	STK	3MY0	2.65
BTK	TK	3GEN	1.6
CHEK2	STK	2CN5	2.25
FGFR1	TK	1FGK	2
FGFR3	TK	4K33	2.34
FLT3	TK	1RJB	2.1
JAK3	TK	3LXL	1.74
MERTK	TK	2P0C	2.4
MET	TK	1R0P	1.8
NTRK1	TK	4AOJ	2.75
PHKG2	TK	2Y7J	2.5
RET	STK	2IVS	2
ROR2	TK	3ZZW	2.9
RPS6KA3	TK	4D9T	2.4
TEK	STK	1FVR	2.2
ZAP70	TK	1U59	2.3

Table S7. The collected disease-related mutations on 16 kinase domains used in this study. The five functional subdomains of kinase domains include SD I (P-loop), SD III (α C-helix), SD VIB (catalytic loop), SD VII (activation loop) and SD VIII (P+1 loop).

ID	Disease-related mutations on kinase domains	
	Total mutations	On functional subdomains
ACVRL1	73	27
BTK	143	46
CHEK2	44	24
FGFR1	38	17
FGFR3	27	16
FLT3	49	30
JAK3	17	7
MERTK	22	3
MET	60	29
NTRK1	54	20
PHKG2	9	3
RET	75	25
ROR2	42	12
RPS6KA3	30	10
TEK	15	4
ZAP70	31	8
Total	729	281

Table S8. The classifications of amino acid based on different physicochemical properties.

Amino Acid		Amino Acid Property			
		R-Volume	R-Shape	Polarity	Hydropathy
Ala	A	Small	Linear	Non-polar	Hydrophobic
Pro	P	Small	Circle	Non-polar	Hydrophobic
Gly	G	Small	Linear	Non-polar	Hydrophobic
Ser	S	Small	Linear	Non-polar	Hydrophilic
Thr	T	Small	Linear	Polar	Hydrophilic
Asn	N	Small	Branch	Polar	Hydrophilic
Asp	D	Small	Branch	Polar	Hydrophilic
Glu	E	Medium	Branch	Polar	Hydrophilic
Gln	Q	Medium	Branch	Polar	Hydrophilic
Arg	R	Large	Branch	Polar	Hydrophilic
His	H	Medium	Circle	Polar	Hydrophilic
Lys	K	Large	Branch	Polar	Amphiphilic
Ile	I	Large	Branch	Non-polar	Hydrophobic
Leu	L	Large	Linear	Non-polar	Hydrophobic
Met	M	Large	Linear	Non-polar	Amphipathic
Val	V	Medium	Branch	Non-polar	Hydrophobic
Phe	F	Large	Circle	Non-polar	Hydrophobic
Trp	W	Large	Circle	Non-polar	Amphipathic
Tyr	Y	Large	Circle	Polar	Amphipathic
Cys	C	Small	Linear	Non-polar	Hydrophilic

Table S9. Significance of physicochemical property identity between SSE packing clusters and other regions.

Physicochemical Property	Average physicochemical identity		
	SSEPCs	**Other Region**	**p-value**
Identity for side-chain volume	0.605	0.494	2.20E-16
Identity for side-chain shape	0.624	0.449	2.20E-16
Identity for polarity	0.818	0.621	2.20E-16
Identity for hydropathy	0.712	0.528	2.20E-16

Table S10. The effect size (Cohen's d) for different physicochemical features.

Physicochemical Property	Average physicochemical identity		
	SSEPCs	**Other Region**	**Cohen's d**
Identity for side-chain volume	0.605	0.494	0.6948
Identity for side-chain shape	0.624	0.449	1.1525
Identity for polarity	0.818	0.621	1.5238
Identity for hydropathy	0.712	0.528	1.2794