Synthetic and Systems Biotechnology 6 (2021) 224–230

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.sciencedirect.com/journal/synthetic-and-systems-biotechnology

Classification of protein domains based on their three-dimensional shapes (CPD3DS)

Zhaochang Yang [a,1], Mingkang Liu [a,1], Bin Wang [b], Beibei Wang [a,c,*]

[a] School of Life Science and Technology, University of Electronic Science and Technology of China, China
[b] School of Information and Software Engineering, University of Electronic Science and Technology of China, China
[c] Centre for Informational Biology, University of Electronic Science and Technology of China, 2006 Xiyuan Road, Chengdu, Sichuan, 611731, China

ARTICLE INFO

Keywords:
3D-zernike descriptors
K-means
Shape similarity
Domain surface shapes
Structural similarity

ABSTRACT

Protein design has become a powerful method to expand the number of natural proteins and design customized proteins according to demands. Domain-based protein design spares the need to create novel elements from scratch, which makes it a more efficient strategy than scratch-based protein design in designing multi-domain proteins, protein complexes and biomaterials. As the surface shape plays a central role in domain-domain and protein-protein interactions, a global map of the surface shapes of all domains should be very beneficial for domain-based protein design. Therefore, in this study, we characterized the surface shapes of protein domains, collected from CATH and SCOP databases, with their 3D-Zernike descriptors (3DZDs). Then similarities of domain shape features were identified, and all domains were classified accordingly. The preferences of the combinations of domains between different clusters were analyzed in natural proteins from the Protein Data Bank. A user-friendly website, termed CPD3DS, was also developed for storage, retrieval, analyses and visualization of our results. This work not only provides an overall view of protein domain shapes by showing their variety and similarities, but also opens up a new avenue to understand the properties of protein structural domains, and design principles of protein architectures.

1. Introduction

Being involved in almost all of the physiological processes in living cells, proteins are nano-machines whose functions are determined, in principle, by their three-dimensional (3D) structures [1]. Proteins consist of structural domains, which are evolutionarily and functionally conserved units, and fold their tertiary structures independently from the rest of the protein chains [2]. Duplication, deletion or recombination of the genes of domains are the dominant mechanisms to increase the protein repertoire in the process of evolution [3–5]. Domain-based protein design, such as domain swapping [6,7], has been used to make chimeric proteins. In principle, domain-based protein design requires much less work than scratch-based protein design, and is more suitable for designing multi-domain protein systems and biomaterials [8].

Numerous domain databases have been developed to identify domains, such as CATH (Class, Architecture, Topology, Homology) [9], Structural Classification of Proteins (SCOP) [10], Pfam [11], DALI [12], 3Dee [13], SMART [14], CDD [15] and ProDom [16]. Among these databases, SCOP, CATH, and Pfam are the best in the maintenance and update. So far, there are more than 500,000 domains in CATH, more than 700,000 domains in SCOP and about 6400 domains in Pfam. Apparently, the number of domains [17] is far more than the number of amino acids, which is only 20. What is more, the understanding of properties of domains is far less than that of amino acids, which have been best characterized. Except for a few well-known domains (such as the PH, SH3, and PDZ domains), however, most of the domains were poorly understood [18]. It is very challenging to characterize such a large number of domains systematically. Consequently, the large number and complexity of domains make domain-based protein design more difficult to perform than scratch-based protein design currently. Therefore, an overall understanding of the properties of domains may make the domain-based protein design easier and more efficient.

Classification is a commonly used method to reduce the dimensionality of data. The domains were classified according to their sequences in

Peer review under responsibility of KeAi Communications Co., Ltd.
* Corresponding author. School of Life Science and Technology, University of Electronic Science and Technology of China, China.
E-mail address: bbwang@uestc.edu.cn (B. Wang).
[1] They contributed equally to this work.

https://doi.org/10.1016/j.synbio.2021.08.003
Received 28 June 2021; Received in revised form 23 August 2021; Accepted 30 August 2021

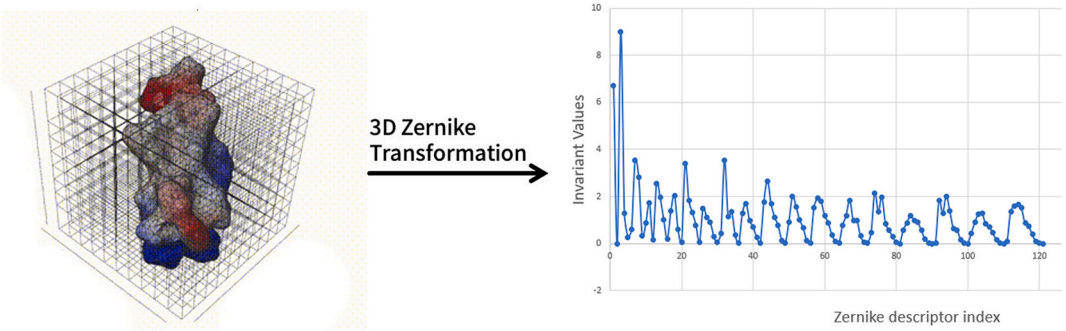

Fig. 1. A schematic of the 3DZD extraction.

Pfam and folds (secondary structural elements) in CATH and SCOP. The advantage of the sequence-based approach is that no structural information is required, but sequence similarities can be extremely low between proteins that share very similar structures [1]. The fold-based approach provides the structural similarity, but the conservation of folds makes it difficult to do a fine classification and to provide a global map of the domain universe.

In this study, we tried to characterize the surface shapes of all domains and classify domains with their global surface shape similarity. Protein surface shapes are more relevant to their functions and protein-protein interactions than their sequences and folds. The 3D surface shapes of proteins have been characterized and classified into a similarity space [19]. Protein functional surfaces, the surfaces of ligand-bound regions, have also been classified using their attributes, such as hydrophobic strength, charge concentration, and sphericity [20]. To our knowledge, there is no attempt to use the 3D surface shape for domain classification so far.

We presented the surface shapes of domains with the 3D-Zernike descriptors (3DZDs) [21], which has been used to compare protein shapes [22] and electron microscopy maps [23] efficiently. After de-redundancy, domains from SCOP and CATH databases were categorized based on their 3DZDs. The distribution of all clusters may provide a global map of the domain shapes. The frequencies of domain combinations in natural proteins were also analyzed to understand principles of natural protein architectures. A webserver, termed CPD3DS, was also developed to store, retrieve, and visualize our results.

2. Methods

2.1. Data acquisition

About 130,000 entries of domain information were collected from CATH [9], SCOP and SCOP 2 [10]. Since the dataset is highly redundant, we firstly removed invalid data, and then removed redundant domains with a similarity threshold of 80% using Cd-hit [24]. The procedure finally yielded 33,455 domains.

2.2. Extracting domain surface features

The domain surfaces were characterized by 3DZDs [21], performed by the 3D-surfer webserver [25]. Fig. 1 shows the principle of 3DZD. Firstly, the domain surface was voxelized and discretized. Then, 3D Zernike transformation was carried out and resulted in a 121 dimensional vector, which is independent of the translation and rotation of the domain. The 121 dimensional vector was used to describe the geometrical shape and group domains with similar shapes. The traditional Euclidean distance (d_E) was calculated to compare the domain surface shapes:

$$d_E = \sqrt{\sum_{i=0}^{i=120} (x_i - y_i)^2},$$

where x_i *and* y_i are the *i*th components of the extracted 121 dimensional vectors of domains x and y.

2.3. Clusterability

Clustering requires that the data are not evenly distributed. The Hopkins statistic (H) is used to test the randomness of the n dimensional dataset:

$$H = \frac{\sum_{i=1}^{n} z_i}{\sum_{i=1}^{n} d_i + \sum_{i=1}^{n} z_i},$$

where d_i is the distance between i and its nearest neighbor in the dataset, and z_i is the distance between i and its nearest neighbor in an artificially generated n dimensional dataset, in which the data are randomly distributed across the test data space. If the data are evenly distributed, H is close to 0.5 and clustering is not recommended. If H is close to 0, it indicates that the data are clustered and clustering is recommended. For our dataset, the calculated H is 0.14, indicating that clustering is feasible.

2.4. Clustering

The widely used clustering algorithms, K-means [26], was employed in this study to group domains with similar shapes. K-means calculates the distances between each sample and K cluster centers, and groups them into the clusters with the minimal distances. The number of clusters K is the most important parameter, directly determining the clustering quality. The "elbow" rule [27] was used to estimate the most appropriate K.

$$\mathrm{SSE} = \sum_{i=1}^{K} e^{\sum_{p \in C_i} |p - c_i|^2} + e^{max_{1<i<K} \sum_{p \in C_i} |p - c_i|^2},$$

where C_i is the set of all samples in the current cluster, c_i is the cluster center of the current cluster. Sum of squared errors (SSE) represents the degree of dispersion of all clusters when the number of cluster is K. The smaller the SSE value, the better the clustering effect. As K increases, SSE firstly decreases sharply, then stabilizes. The inflection point is considered as the most appropriate K. So K = 100 was used in our clustering analysis (Fig. 2A). The clustering analysis was carried out using the sklearn module of Python [28].

2.5. Domain combination frequencies in natural proteins

All available domain annotations in the protein data bank (PDB) [29] were used to calculate the frequencies of the domain combinations (p_{ij}).

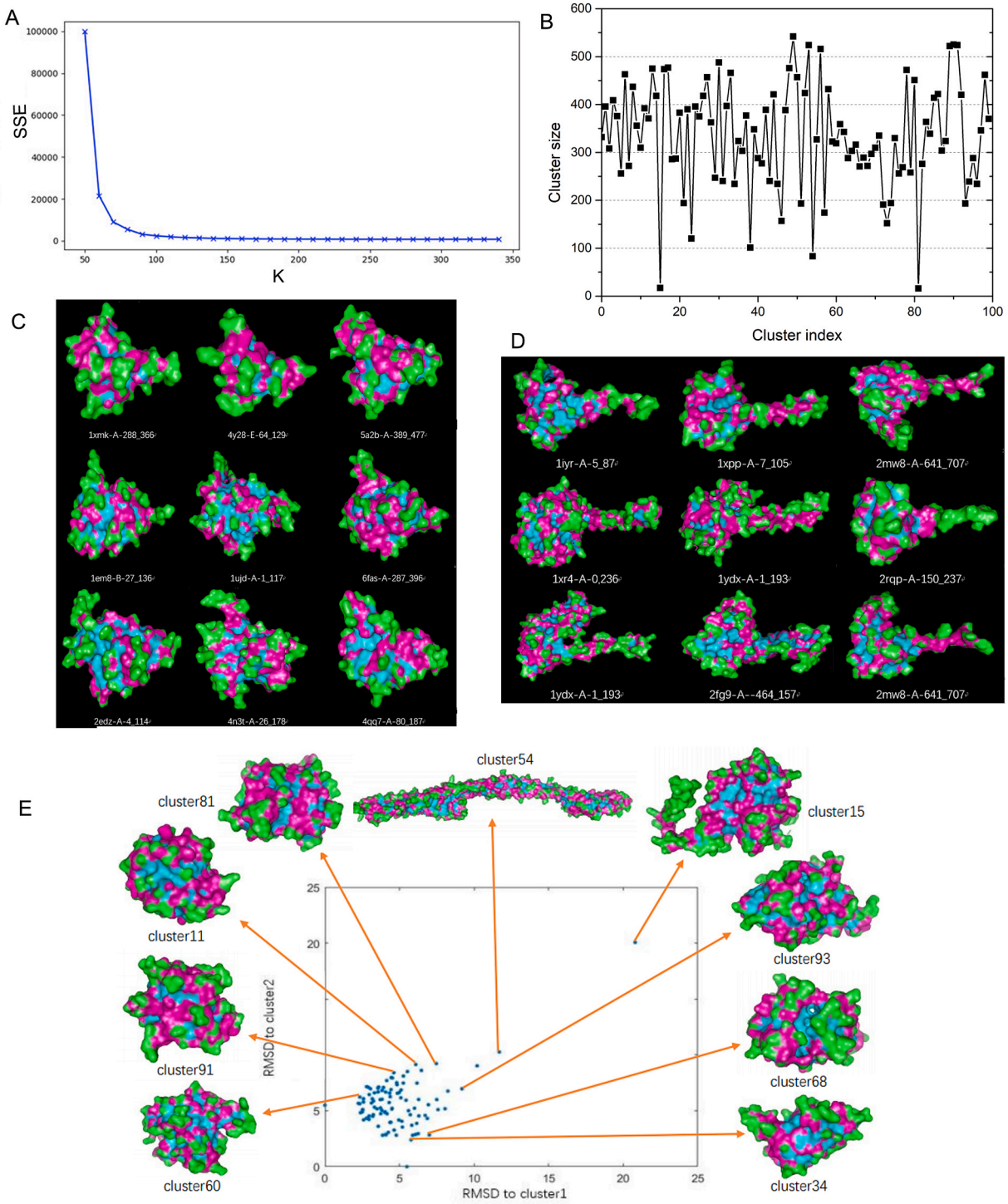

Fig. 2. The results of clustering analysis. (A) Variation of SSE values with the increase of K. (B) The cluster sizes of 100 clusters. (C–D) Nine randomly selected domain surface shapes in clusters 1 and 2. The surfaces were colored in blue for pockets, in red for protrusions, and green for flat regions respectively. The figures were generated by VisGrid [30]. The labels below the domains identify the source of the domains. For example, 1xmk-A-288_366 denotes the domain was taken from chain A of the protein with the PDB ID of 1xmk and the range of residue indices 288–366. (E) Projections of representative domains of 100 clusters on RMSDs of 3DZD with reference to the representative domains of clusters 1 and 2.

$$p_{ij} = \frac{N_{ij(i\in C_i, j\in C_j)}}{N_{max}} \times 10,$$

where N_{ij} is the number of times that two domains from clusters C_i and C_j respectively are present in one protein, and $N_{max}$is the maximum N_{ij}. So the value of p_{ij} is in the range of 0–10.

2.6. Database construction and interface

A webserver, CPD3DS (http://175.24.69.122:8880), was developed with HTML, CSS, and JavaScript on a Windows platform, for the storage, retrieval, and visualization of our results. Swagger2 (https://swagger.io/), a standard and complete framework for generating, describing, invoking and visualizing restful style web service, was used for all

Table 1
The intra-cluster structural similarity comparison of our clustering results and CATH classes.

	$\overline{\text{TMscore}_{ic}}$	$\overline{S_{FATCATic}}$ (%)
Our clustering results	0.305 ± 0.012	14.9 ± 0.6
CATH classes	0.432 ± 0.013	16.8 ± 0.7

Table 2
Correlation (upper right, in blue) and Reliability (lower Left, in orange) between the TM-score, $RMSD_{C\alpha}$ and $RMSD_{3DZD}$.

Correlation (r) / Reliability (p)	TM-score	$RMSD_{C\alpha}$	$RMSD_{3DZD}$
TM-score		-0.005	0.007
$RMSD_{C\alpha}$	0.008		0.026
$RMSD_{3DZD}$	4.41×10^{-4}	1.97×10^{-9}	

Application Programming Interfaces (APIs). Spring Cloud (https://spring.io/) was used in the back end. Bootstrap (https://v3.bootcss.com/) and Vue (https://v3.cn.vuejs.org/) were mainly used to build interactive pages of the front end. 3Dmol (http://3dmol.csb.pitt.edu) was implemented to show the domain structures. The data were stored in a MySQL database. We also packed all data, programs, and the operating environment, and uploaded the package to GitHub (https://github.com/igemsoftware2020/Team_UESTC_Software).

3. Results

To make a global view of domain shapes possible, clustering analysis based on the 3D surface shape was performed to reduce the dimensionality of large amounts of domains to a reasonable amount.

3.1. Classifying domain surface shapes

All nonredundant 33,455 domains were grouped base on their 3DZDs using the K-means clustering algorithm with the number of clusters K = 100 (Fig. 2A). The 100 clusters were used to build a basic set of domain surface shapes. The cluster size (the number of members in a cluster, N_c) varies greatly in the range of 10–550 (Fig. 2B). Among the 100 clusters, only 3 clusters have $N_c \leq 100$, 10 clusters have $100 < N_c \leq 200$, 23 clusters have $200 < N_c \leq 300$, 36 clusters have $300 < N_c \leq 400$, 22 clusters have $400 < N_c \leq 500$, and 6 clusters have $N_c > 500$. The distribution of the cluster sizes basically conforms to the normal distribution.

The surfaces of selected domains in clusters 1 and 2 (Fig. 2C–D) are characterized by pockets, protrusions and flat regions using VisGrid [30]. The overall shape of the domains in a cluster is relatively consistent. The structural domains in the cluster1 generally present a triangular shape with a small bulge, while the structural domains in the cluster2 are relatively elongated with a long protrusion. Therefore, in general, the clustering algorithm could group similar shapes into a cluster.

The representative domains of 100 clusters are mapped on the plane defined by Root-mean-square deviations (RMSDs) of 3DZD ($RMSD_{3DZD}$) with reference to representative domains of clusters 1 and 2 (Fig. 2E), to give an overview of the distribution of 100 clusters. The $RMSDs_{3DZD}$ of most clusters in both dimensions are less than 10. They all present a relatively compact structure, but the specific shapes are different. For example, the domain in the cluster11 appears a more spherical shape, while the domain in the cluster91 is closer to a rectangular shape. The $RMSDs_{3DZD}$ of clusters 15 and 54 are more than 10, especially the $RMSD_{3DZD}$ of the cluster15. The domain shape in the clusters54 is slender, while the domain shape in the cluster15 is similar to that of the cluster2, but with a rolled up protrusion. Therefore, in general, the cluster analysis based on the 3DZD could distinguish different shapes, and is also sensitive to local surface shapes.

3.2. Evaluation by comparison with CATH

To access the performance of the K-means clustering based on the 3DZD, we randomly selected 1000 domains to form a subset, and after de-redundancy, 727 were left. These 727 domains were assigned into

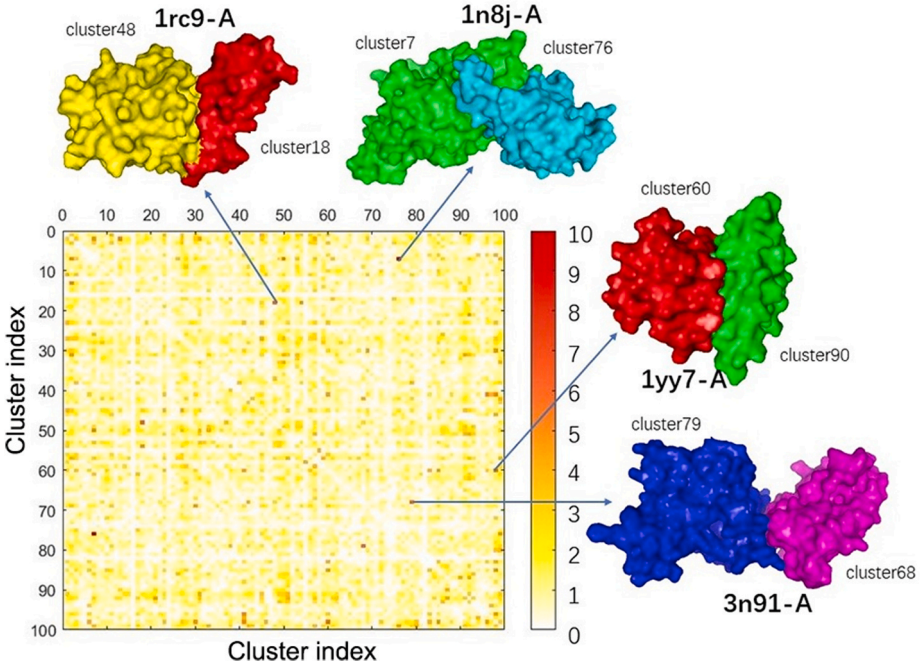

Fig. 3. Domain combination frequencies in the protein repertoire. Examples of domain combinations with large p_{ij} was denoted by their cluster indices and PDB IDs.

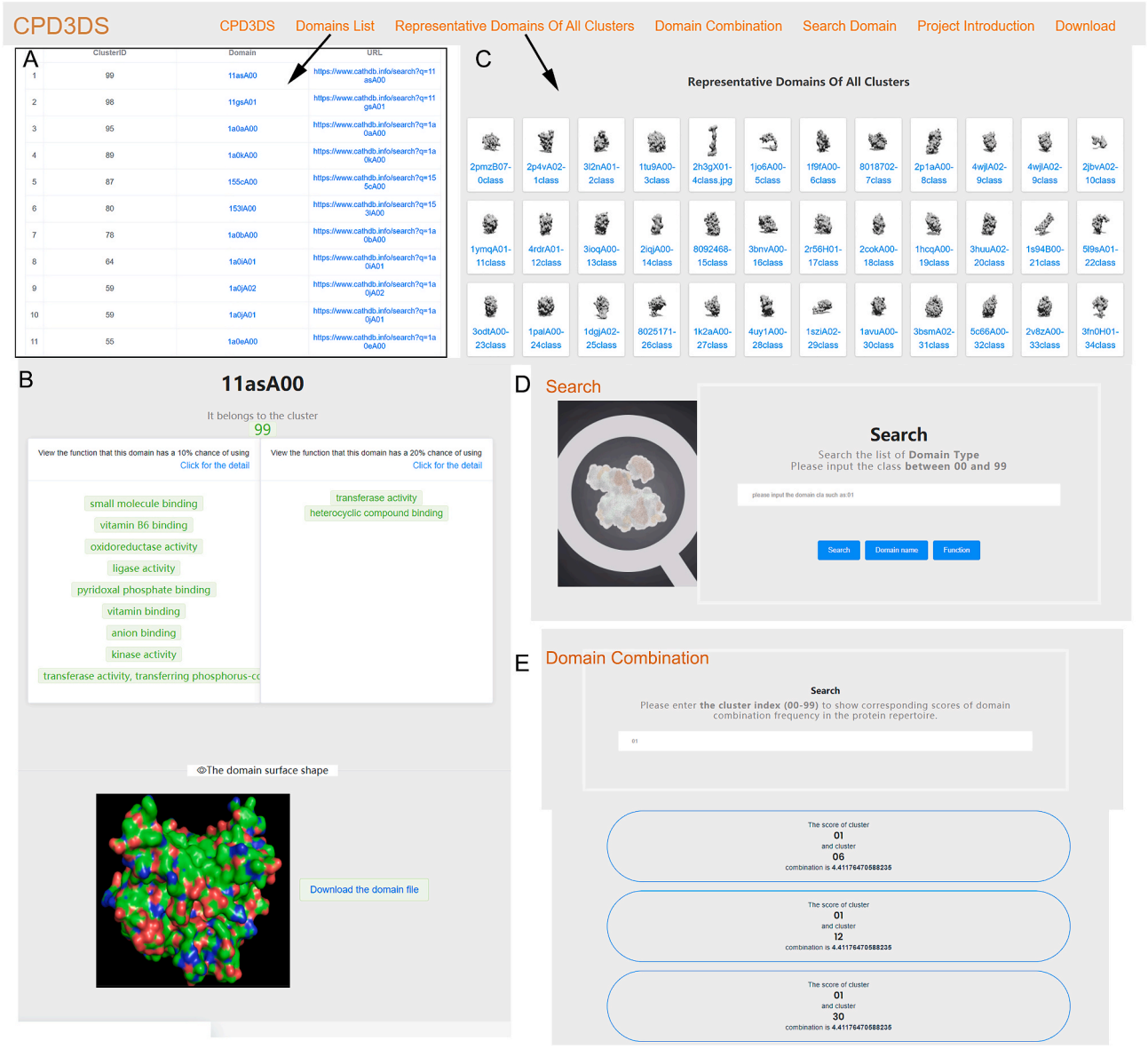

Fig. 4. Snapshots of the webserver CPD3DS.

different clusters according to our clustering results and CATH classes respectively. Then pairwise structural similarities within clusters were calculated by TM-align [31] and FATCAT [32]. TM-align uses the template modeling score (TM-score) rotation matrix [33], instead of the RMSD rotation matrix, to superimpose protein structures, and the TM-score is more sensitive to the global structural topology than local structural changes. The value of TM-score is between 0 and 1, with the larger value indicating the higher structural similarity. FATCAT has good performance for structure alignment of flexible proteins. FATCAT identifies the flexible regions in the protein before minimizing the overall RMSD.

The average intra-cluster TM-score ($\overline{\text{TMscore}_{ic}}$) of our clustering results is 0.305 ± 0.012. The P-value of a TM-score > 0.3 is less than 0.001, indicating that the similarity between structures is significantly different from randomly selected structures [34]. Comparatively, the $\overline{\text{TMscore}_{ic}}$ of CATH is 0.432 ± 0.013, about 0.13 larger than that of ours (Table 1). This is not surprising, as both the calculation of TM-score and the classification of CATH are based on folds of proteins. For a further comparison, we calculated the uncorrelated FATCAT similarity. It is found that the average intra-cluster FATCAT similarity ($\overline{S_{FATCATic}}$) is only slightly (about 2%) lower than that of CATH (Table 1).

The results of the comparisons indicate that the 3DZD-based clustering analysis is able to reflect the overall shape as well as the global topology. It also suggests that the shape-based classification could also reflect structural similarity to a large extent.

3.3. Correlations between $RMSD_{3DZD}$, TM-score and $RMSD_{C\alpha}$

TM-score scores domain similarity according to their global topology [33], 3DZD describes the domain surface shape, and the RMSD of all Cα atoms ($RMSD_{C\alpha}$) reflects local conformational changes. We calculated the pairwise TM-score, $RMSD_{3DZD}$ and $RMSD_{C\alpha}$ of all 727 domains. Their correlations are listed in Table 2. The correlation between $RMSD_{3DZD}$ and $RMSD_{C\alpha}$ is the strongest with a reliability of 1.97×10^{-9}, while TM-score has weak correlations with the other two. It indicates that 3DZD is sensitive to the local structural variations on the surface. It further demonstrates that the surface shape descriptor 3DZD could also reflect the structure details.

It is worth to notice that the 3DZD is rotation invariant, and rotation optimization is not necessary during the calculation. So the

computational complexity of the 3DZD is much less than the commonly used $RMSD_{C\alpha}$. The 3DZD may be a better choice for processing large-scale data.

3.4. The frequency of domain combination in the protein repertoire

A substantial fraction of proteins are composed of multiple domains. Some domains are involved in diverse proteins, and some are only present in specific combinations. The domain combination frequencies (p_{ij}) between clusters were calculated for proteins from the PDB database. The obtained 100 × 100 matrix was shown in Fig. 3 with examples of high frequency of combination mode. Most of the combinations have a $p_{ij} < 5$, consistent with the domain promiscuity and protein diversity [18]. Among all the possible combination modes, the combination between clusters 7 and 76 is most frequent. It is worth to note that domain combinations with large p_{ij} show good shape matching.

3.5. The CPD3DS webserver

All information of the 33,455 domains and our results were stored and can be searched on the webserver CPD3DS (http://175.24.69.122:8880). In the webpage of the domain list, information of all domains is listed, including the cluster index (ClusterID) of our clustering results, domain name in CATH or SCOP, and the URL link of its source (Fig. 4A). One can click on a domain name to view the detailed information of the domain (Fig. 4B), containing the possible functions taken from the gene product annotation in the PDB database and the picture of the domain surface shape which could be downloaded. The representative domains of 100 clusters are list to give a global view of the domain shapes (Fig. 4C). CPD3DS supports 3 search methods: cluster index, domain name, and function (Fig. 4D). The cluster index search (clusters 00–99) yields a list of domains in the cluster, similar to the list in Fig. 4A. Searching for the domain name only returns one entry, as the domain name is unique. The function search (for example, anion binding) returns a list of domains that contain the searched function, also similar to the list in Fig. 4A. Finally, the domain combination frequencies (p_{ij}) could be searched as well (Fig. 4E), the search results are arranged in descending order of p_{ij}. All data can be downloaded in SQL format.

Due to the using of K-means clustering analysis, its randomness makes the database difficult to extend and update automatically. If a new domain is identified, its $RMSDs_{3DZD}$ with reference to all cluster centers will be calculated first, it will be assigned to the cluster with the smallest $RMSD_{3DZD}$. If there are lots of new domains, we will re-do the cluster analysis then.

4. Conclusions

In this study, we constructed a map of the domain surface shape space by clustering domains based on their 3DZDs, to explore the variety and similarity of domain shapes. Our approach is not only powerful in detecting the domain similarity of global structural topology, but also sensitive to local structural variations. Therefore, coupled with the feature of the fast calculation speed, 3DZD may be an ideal parameter for comparison and retrieval of large-scale structural information. We also tried to analyze the inter-cluster domain combination frequencies of proteins in the PDB database. The domain combination in natural proteins may indicate a primary principle of protein organizations. All the results can be easily viewed through our CPD3DS webserver.

As shape matching between domains is one of the most important factors in protein architectures, this study may be helpful to the domain-based protein design. A global view of all domain shapes could enhance the understanding of protein domains and the domain constitution of proteins, and make the selection of the desired domain easier. Of course, this study is just a coarse beginning. Lots of more detailed work needs to be carried out in the future. The physicochemical properties of the surface could be considered in future work, different clustering methods could be performed and compared, and inter-protein domain interactions could also be analyzed with the data available in protein-protein interaction databases, such as STRING [35].

Notes

The authors declare no competing financial interests.

CRediT authorship contribution statement

Zhaochang Yang: Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft. **Mingkang Liu:** Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft. **Bin Wang:** Data curation, Formal analysis, Software. **Beibei Wang:** Conceptualization, Investigation, Formal analysis, Methodology, Supervision, Writing – review & editing.

Acknowledgments

Thanks to the contributions of all iGEM team members of UESTC-Software-2020. This work was supported by the National Natural Science Foundation of China (No. 31971176 and 31800616) and the Fundamental Research Funds for the Central Universities (No. A03018023601045).

References

[1] Orengo CA, Todd AE, Thornton JM. From protein structure to function. Curr Opin Struct Biol 1999;9:374–82.
[2] Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. Science 2003;300:1701–3.
[3] Weiner J, Beaussart F, Bornberg-Bauer E. Domain deletions and substitutions in the modular protein evolution. FEBS J 2006;273:2037–47.
[4] Bjorklund AK, Light S, Sagit R, Elofsson A. Nebulin: a study of protein repeat evolution. J Mol Biol 2010;402:38–51.
[5] Dohmen E, Klasberg S, Bornberg-Bauer E, Perrey S, Kemena C. The modular nature of protein evolution: domain rearrangement rates across eukaryotic life. BMC Evol Biol 2020;20:30.
[6] Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. Protein Sci 2002;11:1285–99.
[7] Zhu J, Avakyan N, Kakkis A, Hoffnagle AM, Han K, Li Y, Zhang Z, Choi TS, Na Y, Yu C-J, Tezcan FA. Protein assembly by design. Chem Rev 2021. https://doi.org/10.1021/acs.chemrev.1c00308.
[8] Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. Nature 2016;537:320–7.
[9] Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, Pang CSM, Woodridge L, Rauer C, Sen N, Abbasian M, Le Cornu S, Lam SD, Berka K, Varekova IH, Svobodova R, Lees J, Orengo CA. CATH: increased structural coverage of functional space. Nucleic Acids Res 2021;49:D266–73.
[10] Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res 2020;48:D376–82.
[11] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. Pfam: the protein families database in 2021. Nucleic Acids Res 2021;49:D412–9.
[12] Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. A fully automatic evolutionary classification of protein folds: dali Domain Dictionary version 3. Nucleic Acids Res 2001;29:55–7.
[13] Siddiqui AS, Dengler U, Barton GJ. 3Dee: a database of protein structural domains. Bioinformatics 2001;17:200–1.
[14] Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci USA 1998;95:5857–64.
[15] Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 2007;35:D237–40.
[16] Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D. ProDom: automated clustering of homologous domains. Briefings Bioinf 2002;3:246–51.
[17] Doolittle RF. The multiplicity OF domains IN proteins. Annu Rev Biochem 1995;64:287–314.
[18] Basu MK, Poliakov E, Rogozin IB. Domain mobility in proteins: functional and evolutionary implications. Briefings Bioinf 2009;10:205–16.
[19] Han X, Sit A, Christoffer C, Chen S, Kihara D. A global map of the protein shape universe. PLoS Comput Biol 2019;15:e1006969.

[20] Tseng YY, Li W-H. Classification of protein functional surfaces using structural characteristics. Proc Natl Acad Sci USA 2012;109:1170–5.
[21] Sael L, Li B, La D, Fang Y, Ramani K, Rustamov R, Kihara D. Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins 2008;72:1259–73.
[22] Kihara D, Sael L, Chikhi R, Esquivel-Rodriguez J. Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. Curr Protein Pept Sci 2011;12:520–30.
[23] Han X, Wei Q, Kihara D. Protein 3D structure and electron microscopy map retrieval using 3D-surfer2.0 and EM-SURFER. Curr. Protoc. Bioinform. 2017;60:3.14.1-3.14.15.
[24] Li W, Godzik A, Cd-hit. A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9.
[25] La D, Esquivel-Rodriguez J, Venkatraman V, Li B, Sael L, Ueng S, Ahrendt S, Kihara D. 3D-SURFER: software for high-throughput protein surface comparison and analysis. Bioinformatics 2009;25:2843–4.
[26] Selim SZ, Ismail MA. K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. IEEE Trans Pattern Anal Mach Intell 1984;6:81–7.
[27] Wang Jianren MX, Duan Ganglong. Improved K-means clustering k-value selection algorithm. Comput. Eng. Appl. 2019;55:27–33.
[28] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[29] Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, Dutta S, Feng Z, Ganesan S, Goodsell DS, Ghosh S, Green RK, Guranovic V, Guzenko D, Hudson BP, Lawson CL, Liang Y, Lowe R, Namkoong H, Peisach E, Persikova I, Randle C, Rose A, Rose Y, Sali A, Segura J, Sekharan M, Shao C, Tao Y-P, Voigt M, Westbrook JD, Young JY, Zardecki C, Zhuravleva M. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res 2021;49:D437–51.
[30] Li B, Turuvekere S, Agrawal M, La D, Ramani K, Kihara D. Characterization of local geometry of protein surfaces with the visibility criterion. Proteins 2008;71:670–83.
[31] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–9.
[32] Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A. Fatcat 2.0: towards a better understanding of the structural diversity of proteins. Nucleic Acids Res 2020;48:W60–4.
[33] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–10.
[34] Xu J, Zhang Y. How significant is a protein structure similarity with TM-score=0.5? Bioinformatics 2010;26:889–95.
[35] Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2017;45:D362–8.