OPEN ACCESS Freely available online

PLOS ONE

Effective Moment Feature Vectors for Protein Domain Structures

Jian-Yu Shi[1,2], Siu-Ming Yiu[2], Yan-Ning Zhang[3], Francis Yuk-Lun Chin[2]*

1 School of Life Science, Northwestern Polytechnical University, Xi'an, Shaanxi Province, China, **2** Department of Computer Science, The University of Hong Kong, Hong Kong, China, **3** School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi Province, China

Abstract

Imaging processing techniques have been shown to be useful in studying protein domain structures. The idea is to represent the pairwise distances of any two residues of the structure in a 2D distance matrix (DM). Features and/or submatrices are extracted from this DM to represent a domain. Existing approaches, however, may involve a large number of features (100–400) or complicated mathematical operations. Finding fewer but more effective features is always desirable. In this paper, based on some key observations on DMs, we are able to decompose a DM image into four basic binary images, each representing the structural characteristics of a fundamental secondary structure element (SSE) or a motif in the domain. Using the concept of moments in image processing, we further derive 45 structural features based on the four binary images. Together with 4 features extracted from the basic images, we represent the structure of a domain using 49 features. We show that our feature vectors can represent domain structures effectively in terms of the following. (1) We show a higher accuracy for domain classification. (2) We show a clear and consistent distribution of domains using our proposed structural vector space. (3) We are able to cluster the domains according to our moment features and demonstrate a relationship between structural variation and functional diversity.

Citation: Shi J-Y, Yiu S-M, Zhang Y-N, Chin FY-L (2013) Effective Moment Feature Vectors for Protein Domain Structures. PLoS ONE 8(12): e83788. doi:10.1371/journal.pone.0083788

Editor: Franca Fraternali, King's College, London, United Kingdom

Received July 2, 2013; **Accepted** November 8, 2013; **Published** December 31, 2013

Funding: This work was supported by Hong Kong Scholars Program (No. XJ2011028, http://www.hkscholars.org/) and Small Project Funding of the University of Hong Kong(No. 201209176053, http://www2.caes.hku.hk/research/research-funding/), and partially supported by Fundamental Research Foundation of Northwestern Polytechnical University in China (No. JC201164, http://www.nwpu.edu.cn/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: chin@cs.hku.hk

Introduction

The study of protein structures is no doubt a very important issue as structures are related to functions and can provide insights on the evolution of proteins. As protein structure is more conserved than its sequence throughout evolution, remote homologies should be detected in the universe of protein structures [1]. Proteins with similar structures are known to have similar functions. Thus, the classification of protein structures is a crucial step to capture functionality as well as evolutionary relationships of the proteins [2,3]. In the following, we focus our discussion on protein domains which are the functional units of proteins [4].

A straightforward approach to do classification is to align and compare the 3D structures of two protein domains (e.g. CE [5] and Dali [6]). Based on a similarity score (e.g. root-mean-square deviation (RMSD)), we can decide if they are in the same class. However, there are several drawbacks to this approach. First, the score usually captures very little biological context of functions and evolution [4]. Second, 3D structural alignment is computation-intensive, and the score may violate the triangular inequality [7]. To resolve this problem, another direction is to perform feature extraction, a common method in pattern recognition and computer vision [8]. In this approach, important features related to the structure are extracted and represented as a feature vector. Instead of aligning the 3D structures directly, we compare the vectors. Since we do not need to align 3D structures, the comparison can be done much faster and can be used for all-against-all comparison, clustering and classification, and even structure retrieval with a well-designed indexing system in large scale datasets [9]. Moreover, the feature vectors, with the help of tools such as principal component analysis, enable researchers to analyze and visualize the distribution of domains in a 3-dimensional structural universe [3,7,10,11] so as to deduce important factors related to evolution and functions.

There has been quite a lot of work following this feature extraction paradigm. Some characterize the domain structure directly as a topological object. Scaled Gauss Metric (SGM) [7] treated the protein backbone as a space curve, applied knot theory to describe such a curve by 29 topological invariants [12] and combined the length of domain into a 30-D feature vector for a domain. Recently, Penner et al. [11] exploited the idea of homeomorphism, a kind of topological isomorphism, to transform protein structure into a Fatgraph [13], and used two topological invariants of Fatgraph, the number of twisted alpha carbon linkages and the length of domain together as a feature vector of a domain structure. In general, topological characterization of domain demands complicated mathematical operations.

To speed up the feature extraction, as inspired from computer vision and pattern recognition, Choi et al. [10] considered the distance matrix (DM), a 2D matrix storing the pairwise distances of any two residues, as a representation of a domain structure. Different structures show different DMs. Simple thresholds were

employed to discretize the image and 100 overlapping submatrices were selected from DM by partitioning around the medoids [14]. The frequencies of these submatrices were counted and grouped into a 100-D vector called local feature frequency profile (LFF). On the other hand, Chi et al. [15] treated a DM as a textural image, and extracted 24 features by local histograms as well as 9 features by Haralicks texture descriptors to represent a domain. But the extraction of the histogram and textural features of an image involves time-consuming computation. FragBag [16] was inspired by bag-of-word (BOW) in text recognition. It regarded protein structure as a BOW in which the words were 400 short fragments of the structure, counted the occurring numbers of all words in a protein by local structure alignment and combined them together into a 400-D feature vector as a representation of a domain structure. To summarize, these existing methods either use a large feature vector or involve complicated mathematical operations. Thus, representing domain structures with meaningful and short feature vectors remains a challenge.

Motivated by the previous findings that different types of structures show different images of DM and common types of structures show similar images [10,15], in this paper, we also treat a DM as a textural image for effective and efficient analysis. We observe that, based on the inter-distance between residues (as given in DM) and the key patterns for different fundamental secondary structure elements (SSE), such as alpha helices and beta sheets, as well as basic structural motifs (e.g. $\beta\alpha\beta$ motif), different patterns exist in the DM image. Thus, using Gabor filter efficiently, we can decompose a DM into four binary contact images (BCM) representing alpha helix, parallel and anti-parallel beta sheet and other structure bondings. Each of these basic images represents the structural characteristics of one type of fundamental SSE or motifs. By using the concepts of moments in image processing, we calculate a series of image moments for each BCM from low order to high order (capturing structural characteristics from low to high granularity, such as the sizes of these SSEs and their inter-relationship). These moment series together with some simple counting statistics on BCM will form a 49-D feature vector (please refer to the Method section for a detailed description of our workflow). Since our feature vector is built from elementary SSE and motifs, it captures the structural information more effectively even with only 49 features.

We illustrate the effectiveness of our feature vectors in the following manner. (1) Based on two well-known protein domain classification databases, CATH [17] and SCOP [18], we compare the accuracy of prediction using our moment vector versus other representations. We show that we achieve a much higher accuracy at all levels. (2) Using our moment vectors, we construct a 3D domain structure universe. We are able to show and visualize a clear and consistent distribution of domains in this universe. (3) We cluster the domains according to our moment vectors and demonstrate a relationship between structural variation and functional diversity.

Methods

Distance Matrix (DM)

Given a protein domain, we represent the distances between any two residues (using their C_α atoms as representatives) in a 2D matrix. An example is shown in Fig. 1. Since this distance matrix is symmetric, we only need to focus on the upper triangular part. To identify the protein secondary structure elements (SSE), we mainly look at the regions in which the residues are close together, i.e., the regions in the DM with deep blue colors (the deeper the color is, the closer the residues are). The reason is that in these structures, residues are bounded by strong hydrogen bonds in alpha helix and beta sheets or non-hydrogen bonds (such as ionic bonds, disulfide bonds, van der Waals/hydrophobic interactions that exist between helix and other SSEs, e.g. helix-sheet interaction in $\beta\alpha\beta$ motif, $\alpha\alpha$ hairpin or closed α helix-coil part). These residues tend to be closer. In our approach, we do not need any input about what and where SSEs are in a given protein domain.

1. Alpha helix:

In a typical alpha helix, each spiral loop takes about 3–4 residues(3.6), say the i-th residue to (i+3)-th residue, with the (i+3)-th residue closest to the i-th residue and the (i+4)-th residue wrapped overshooting the i-th residue. Since these 4 residues are very close together, thus in terms of the DM, we should see a deep blue strip with width of 3–4 residues near the diagonal (T_α in Fig. 1). Furthermore, based on the findings in [19], the pairwise distances of these 4 residues exhibit a local minimum between the distance of the i-th and the (i+3)-th residues. Thus, capturing the strip and identifying the local minimum within the strip, which is 3 residues from and parallel to the diagonal, should enable us to identify alpha helix.

2. Beta sheets:

Two strands of a beta sheet are connected by hydrogen bonds. Depending on the orientations of these two strands (Fig. 2), the distances between these bounded residues (either$\{(p, q); (p+1, q-1); (p+2, q-2)\}$ in Fig. 2-A or $\{(p, q-2); (p+1, q-1); (p+2, q)\}$ in Fig. 2-B) are smaller compared to the distances between other nearby residues. Thus, in terms of DM, there again exists a blue strip either parallel (for parallel beta sheet) or perpendicular (for anti-parallel beta sheet) to the diagonal, but due to the extra residues in between the two sheets, this strip might not be near the diagonal ($T_{\beta\|}$ and $T_{\beta\perp}$ in Fig. 1). Identifying these strips parallel and perpendicular to the diagonal helps us to identify beta sheets.

3. Other structures:

There are other non-hydrogen bondings that occur between α helix and other SSEs such as the helix-sheet part in $\beta\alpha\beta$ motif, $\alpha\alpha$ hairpin or the closed α helix-coil part. Due to the periodic structure of alpha helix, we found that these bondings correspond to local minima in DM. For example, residue p of a sheet is close to residues q, $q+3$ or 4, $q+6$ or 7 or 8, ... of a helix, similarly for residues $p+1$, $p+2$ of the sheet, which are also close to the corresponding residues of the helix. Thus, there exists a narrow band of short blue horizontal or vertical strips of 3–4 residues apart ($T_{\alpha\#}$ in Fig. 1). When two helices interact, both horizontal and vertical strips exist, cross each other and form textural patterns comprising '+'-shaped units.

Gabor Filters and Binarization

Even though the above features are represented by relatively low values (since they are closer) in the DM, simple threshold filtering [10] cannot easily identify and separate these SSEs. We used the Gabor filter [20] to identify these SSEs from the images. Since simple cells in the visual cortex of mammalian brains can be modeled by Gabor functions, people tend to utilize Gabor functions to analyze image with the way similar to the perception in human visual system [20]. For example, human perceives static shapes by accumulating the information (orientations, boundaries or patterns, etc.) around them to enlarge contrast between shapes and their neighbors. Similarly, 2-dimensional Gabor function can enhance the contrast between local extrema (minima or maxima) and their neighboring entries in a matrix or an image by an

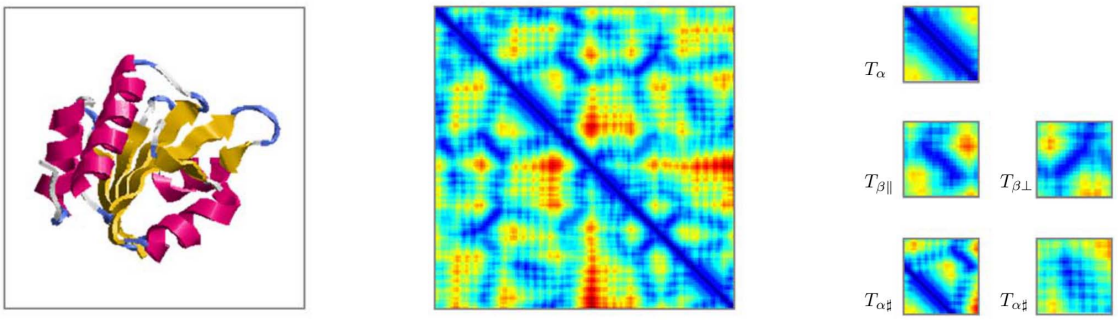

Figure 1. Representation of domain structure by its distance matrix, near contacts and their local image patterns. From left to right, SSE topology, distance matrix image and local patterns of SSEs and motifs are shown respectively. The example domain (d2c0ga2) selected from SCOP database belongs to c.47.1.7 category. T_α shows a deep blue strip near and parallel to main diagonal of DM, $T_{\beta\|}$ and $T_{\beta\perp}$ are deep blue slant strips parallel and perpendicular to main diagonal respectively, and $T_{\alpha\#}$ exhibits texture patches comprising horizontal and vertical blue strips. Entries of distance matrix are colored from blue to red according to their values from small to large.
doi:10.1371/journal.pone.0083788.g001

accumulating operation, such as convolution. Gabor functions have been found to be particularly appropriate for the representation and discrimination of lines(local extrema), edges (local extrema) and texture (a spatial pattern of line or curve combination) [20].

According to Fig. 1, in order to extract these SSE patterns from the images, we designed three Gabor templates: one for alpha helix and parallel beta sheet; one for anti-parallel beta sheet; and one for other structures. The match between Gabor templates and patterns in image is highly specific. The technical details are given below. Gabor functions $g_{\lambda,\sigma,\gamma,\theta}$ in four orientations $\theta = 0^\circ,90^\circ,45^\circ,135^\circ$ were selected, where wavelength $\lambda=4$, the spatial aspect ratio of major axis to minor $\gamma=1.7$ and the length of major axis in the Gaussian elliptical envelop $\sigma=7$. Then, three templates t_{135}, t_{45} and t_{0+90} were set up, corresponding to Gabor functions in orientation $135^\circ,45^\circ,0^\circ+90^\circ$, for filtering, in which the sizes were all 16×16, where $t_{135}=g_{135^\circ}$ was used for T_α and $T_{\beta\|}$, t_{45} for $T_{\beta\perp}$ and t_{0+90} for $T_{\alpha\#}$ (see the templates in Fig. 3).

We performed the Gabor filtering by the two-dimensional convolution of DM and three templates independently to obtain three groups of patterns of low values. Three binary contact matrix (BCM) were then constructed by assigning 1 to those elements whose convoluted values in the convoluted DM are less than or equal to zero and assigning 0 to the remaining positive values. The binary slant line (135°) patterns of α-helices and parallel β-sheets in the BCM derived from template t_{135} were further separated to form two BCMs according to their closeness to the main diagonal

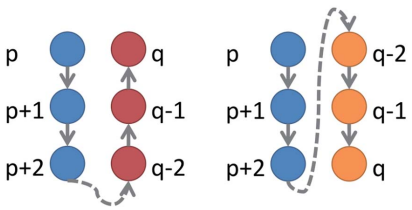

Figure 2. Structural diagram of anti-parallel and parallel β sheets. Each circle node represents an amino acid residue, and its position number in sequence is marked nearby. Arrows denote the direction of domain sequence from 5′ terminal to 3′ terminal. The dotted arrow linking node (p+2) and node (q-2) means that other residues exist between these two residues. Anti-parallel(blue and red nodes) and parallel(blue and orange nodes) β sheets are arranged side-by-side.
doi:10.1371/journal.pone.0083788.g002

(the slant line patterns of parallel β-sheets are further away from the diagonal than those of the α-helix). Finally, we have four BCMs marked as B_α, $B_{\beta_\|}$, $B_{\beta_\perp}$ and $B_{\alpha_\#}$ respectively (Fig. 3). The count and type of α helices and β-sheets from the BCMs are highly consistent with domain structure. More examples are shown in Fig. S1.

Structural Analysis by Higher-Order Moments

These four BCMs can be treated as four binary images containing different information of the SSEs. Each set of the adjacent 1's (edges or edge patterns) in each BCM represents one type of SSE or some SSEs' pairwise interactions. To capture the characteristics of these four BCMs, we define the first set of moments as follow and refer to them as Composition Moments:

$$\begin{aligned} &m_\alpha=\#(B_\alpha>0)/n,\ m_\beta=\#(B_{\beta_\|}>0)/N,\\ &m_{\beta_\perp}=\#(B_{\beta_\perp}>0)/N,\ m_{\alpha_\#}=\#(B_{\alpha_\#}>0)/N \end{aligned} \tag{1}$$

where n is the sequence length of the domain(domain size), $N=n(n-1)/2$ is the area of right-top part of BCM, and $\#$ denotes the count of elements that satisfy the specific condition in BCM. In terms of residue-residue contact, the moments m_α, $m_{\beta_\|}$, $m_{\beta_\perp}$ and $m_{\alpha_\#}$ are the average count of four types of near contacts occurring in the protein structure respectively. In particular, m_α is the composition of helix in a domain structure, $m_{\beta_\|}$ and $m_{\beta_\perp}$ are the composition of parallel and anti-parallel sheets, and $m_{\alpha_\#}$ is the composition of hydrophobic contacts.

It has been shown that any image, in principle, can be recovered by a sufficient number of image moments [21]. That is to say, image moments contains all information in image. Similar to orthogonal polynomials, orthogonal moments, e.g., Legendre moments, can be used to capture the image information. Legendre moments are derived from and equivalent to geometric moments but with considerably computational benefits, such as a stable and fast numerical implementation, the avoidance of the loss of precision caused by overflow or underflow, and a higher robustness to random noise [21]. According to geometric moments, for a given shape, the 0 order and the 1-order moments are its area and mass center respectively, the 2-order moments are its Newtons inertia moments which define shapes resistance to change of 3 axes of rotation, the 3-order moments measure the degree of asymmetry of shape (skewness), and the 4-order moments measure the degree of locally extreme irregularity in

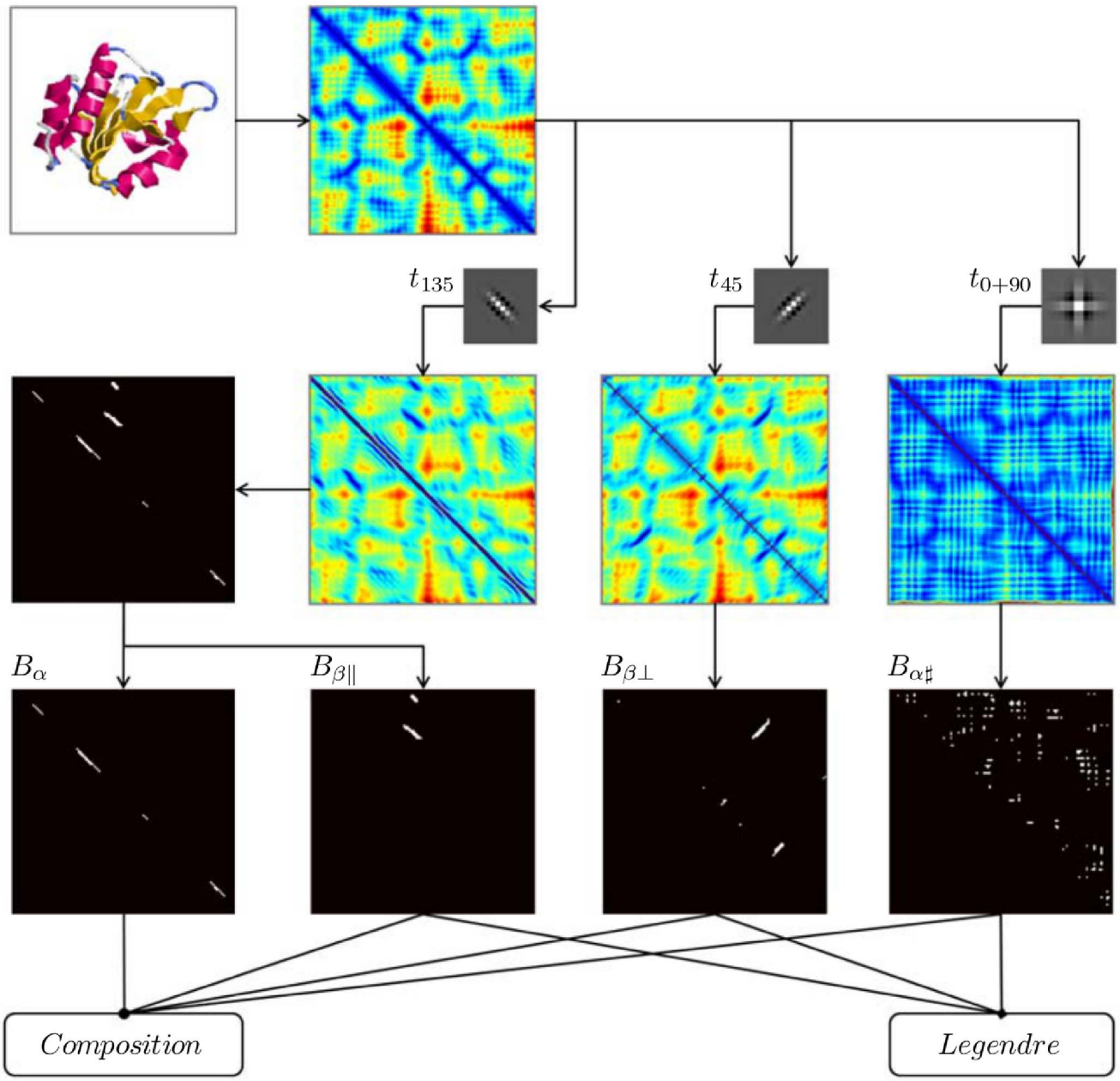

Figure 3. The work-flow of generating feature vector from domain structure(c_{α}-atom coordinates). First, the distance matrix of domain structure (d2c0ga2, c.47.1.7) is calculated. Then, its distance matrix is passed concurrently through three Gabor templates(small patches tagged by t_{135}, t_{45} and t_{0+90} respectively). After that, four binary contact matrices B_{α}, $B_{\beta\|}$, $B_{\beta\perp}$ and $B_{\alpha\#}$ are output by filtering. Last, composition and Legendre moments are calculated to form the feature vector.
doi:10.1371/journal.pone.0083788.g003

shape (kurtosis). Five-order or higher-order moments are also used in many application though they are still not easy to explain in a conceptual way [21].

Analogous to Fourier transform, the relationship between different orders of moments and BCM is similar to the relationship between different frequencies and the signal, where high frequencies represent higher-order information of the signal while low frequencies give the signal's outline. The lower-order Legendre moments depict the outlines of slant lines or maze-like patterns, and the higher-order Legendre moments capture their higher-order information. In the context of BCM, for example, the moments of layout of α-helices has the following meanings: (1) λ_{00} shows how many residues belonging to α-helix, (2) the 1-order moments show where the center of α-helices in domain is according to the sequence positions, (3) the 2-order moments define how helices span domain according to the sequence positions, (4) the 3-order moments measure the degree of asymmetry of helix distribution along the sequence and (5) the 4-order moment denote the deviating degree of the extreme helix from other helices according to the sequence positions. This is usually caused by the insertion of other SSEs between the extreme helix and other helices.

For anti-parallel β-sheets of a domain in BCM, the moments of their layout can be generally explained as follows. (1) The 0-order

moment shows how many residues involving in anti-parallel β-sheets. (2) The 1-order moments show the center of anti-parallel β-sheets in BCM according to participating residue pairs. (3) The 2-order moments measure how anti-parallel β-sheets span BCM. For example, two domains contain the same motif of meanders which have the same count of residue pairs in anti-parallel β-sheets but they may still have different 2-order moments if the intervals between their member strands in sheets differ. (4) The 3-order moments measure the layout asymmetry of anti-parallel β-sheets in BCM. For example, the difference of meander and Greek-key motifs in BCM can be represented by the 3-order moments because meander motif can show a symmetric layout in BCM but Greek-key motif surely shows an asymmetric layout along the diagonal of BCM. (5) The 4-order moments measure how deviating an outlier sheet is from other sheets in BCM and have large values when the deviation is large. The meaning of moments of other two kinds of near contacts can be explained in similar way. Therefore, the Legendre moments represent a gradual approximation and detailed characterization of domain structure through the slant lines or maze-like patterns, which correspond to the SSE, and their interactions. In addition, moments can be treated as pattern features in image and vision analysis in many applications because of the invariant properties, such as scaling and translation and the insensitivity of image noises and the avoidance of disadvantages on structural alignments [21].

The Legendre moments λ_{pq} of order $(p+q)$ are defined as

$$\lambda_{pq} = \frac{(2p+1)(2q+1)}{4} \sum_{r=-1}^{+1} \sum_{c=-1}^{+1} L_p(r) L_q(c) f(r,c) \qquad (2)$$
$$r,c \in [-1,+1], \quad p,q = 0,1,2,\ldots$$

where $p+q$ is the order of moment, $f(r,c)$ stands for an image intensity function at the normalized row value r and column value c, and $L_p(t)$ is the Legendre polynomial of degree p.

In our study, only the first five degrees of the Legendre polynomials are considered,

$$L_0(t)=1,\ L_1(t)=t,\ L_2(t)=\left(3t^2-1\right)/2, \qquad (3)$$
$$L_3(t)=\left(5t^3-3t\right)/2,\ L_4(t)=\left(35t^4-30t^2+3\right)/8$$

Thus only those moments λ_{pq} with the order $p+q \le 4$ are used in experiments.

Results

In this section, we shall demonstrate that moment features are superior in many aspects, such as protein domain classification, domain structure and function representation, when compared with other approaches.

Classification Performance

We first illustrate the classification power of the image moments. Our composition moments and all additional Legendre moments λ_{pq} with order $p+q \le 4$ were used in the experiments. In Table S1, we study the increase in effectiveness of classification by using Legendre moments with different higher orders. The same datasets of CATH domains and the same training and testing schemes as those in the recent work [11] were adopted. We downloaded the top 500 H-level domains(denoted by CATH_32 with 98,110 domains) from CATH v3.2 and the new domains in CATH v3.3(denoted by CATH_33New with 14,437 domains). The classification scheme adopted a 3-fold cross-validation that two-thirds of CATH_32 were randomly sampled for training and the remaining one-third of CATH_32 and all domains of CATH_33New were used for testing respectively. This scheme was repeated three times and its final accuracy is the average of accuracies resulted in three trials.

We compared the classification performance of the image moments with the algebra-topological description, a most recent approach proposed by [11]. The classification accuracy is defined as the percentage of correctly classified domains according to the classification in CATH. To have a fair comparison, we also used the same classifier (Random Forest, RF) as that proposed in their paper. In addition, we also tried Support Vector Machines (SVM) using our moments. The results in Table 1 show that moment features give better classification performance than the algebra-topological description for all classification levels. In particular, for the lowest classification level(H), the accuracies of using moment feature are more than 20% higher than others in both CATH_32 and CATH_33New testing datasets. This demonstrates that moment features can be more effective in capturing the structural characteristics of domains. As a by-product, we also see that SVM gives a slightly better result than RF. Because the datasets used in a recent work [11] were collected several years ago, we also downloaded the latest version (v3.4) of top 500 H-levels in CATH to validate the classification performance of our proposed features by the 3-fold cross-validation. The classification of total 129,739 domains is consistent with CATH according to the accuracies of 99.8412%, 99.4441%, 99.3763% and 99.1916% in C, A, T and H levels respectively.

We also evaluated the moment features using another popular classification database SCOP(v 1.75). Again, the top 500 superfamilies (totally 109, 533 domains, denoted as S_H500) were selected for evaluation. Following a similar training scheme as above, we found that the results of classification are also consistent with those of SCOP in all levels (Table 2).

We found that Legendre moments are powerful enough to distinguish folds within a class (see Fig. S2) and superfamilies within a fold (see Fig. S3). This explains why moment features can achieve much higher classification accuracies even for more fine-grained classification levels. Note that in both evaluations of CATH and SCOP, SVM seems to perform better, so we used SVM in our experiments. In the rest of the paper, we focus on SCOP classification because SCOP treats α/β and $\alpha+\beta$ domains separately whereas CATH merges them into mixed $\alpha-\beta$ class.

Domain Structure Universe

In this section and the next section, we demonstrate another two more useful applications of moment features. One approach to investigate the relationship between protein structure and function is to represent protein structure in a high dimensional space, e.g. the three-dimensional maps generated by dimension reduction, such as multidimensional scaling (MDS) [22–24], singular vector composition (SVD) [10] and principal component analysis (PCA) [3]. Such 3-D maps are not only good for human visualization, they are capable of (i)representing the distribution of domain size (peptide length), types of secondary structural elements in class level [10,11,22], (ii) inferring protein functions [23], tracing the "common structural ancestor" [24] and (iii) analyzing the distribution of functional diversity [3].

Though MDS, PCA and SVD (related to PCA strongly) give a good 3-D visualization to domain structure maps, it is difficult to visualize the spatial relationship between small-size and middle-size domains as they always cluster closely together in such space.

Table 1. Comparison of methods on the top 500 superfamilies of CATH.

Method	Classifier*	C† (%)	A† (%)	T† (%)	H† (%)
Ref. [11]	RF	96.1/92.8‡	84.6/74.1	78.4/63.0	74.9/55.3
Moment	RF	99.6291/96.5921	98.9437/87.6913	98.7644/79.8088	98.3783/72.2882
Moment	SVM	99.8022/95.9825	99.512/89.2429	99.3914/82.4686	99.1258/77.2321

*RF:Radom Forest, SVM:Support Vector Machines.
†The four levels, C, A, T and H contain 4, 33, 328 and 500 types respectively.
‡The value pair separated by "/" in each cell means the accuracies of classification on CATH_32 and CATH_33New.
doi:10.1371/journal.pone.0083788.t001

This problem becomes more eminent as most domains (84.51%) belong to these categories of sizes smaller than 300 peptides. The sizes of domains have been shown to follow approximately the power-law distribution (Fig. S4). In addition, these methods of dimension reduction cannot work well on non-linear data [25]. Kernel Principal Component Analysis(KPCA) has been designed to address dimension reduction for non-linear data [25], and the dimension reduction through KPCA provides an excellent data visualization in bioinformatics, such as gene expression [26,27] and protein phylogenetic profile [28]. Since both KPCA and SVM share the same kernel trick which is the core of all kernel methods, we used KPCA to visualize the feature space while keeping the classification consistency property of SVM.

To illustrate that the moment features can also perform well in this application, we chose 3,000 domains from non-redundant subset of S_H500 (with less than 40% sequence similarity). The feature vector of each domain is represented as a point in the 49-dimensional space. We applied KPCA to map these points non-linearly onto a reproducing kernel Hilbert space in which the original linear operations of PCA were performed [25]. The first three principal components of newly mapped points were taken to form domain structure universe.

To represent the universe, diverse maps of the selected dataset are shown in Fig. 4 and their more perspectives of different viewpoints are shown in Fig. S5. Three types of maps are rendered by different color schemes which are corresponding to their diverse attributes. (1) The first one is Class map shown in both Fig. 4-A and Fig. 4-B in 2 different viewpoints. All-α (red), all-β (green) and α/β (cyan) domains are clustered into three distinct zones which appear quite separately in three planar regions respectively. In particular, both α region and β region are sickle-shaped but nearly mirror-symmetric to each other (Fig. S5-A provides a better perspective.), while α/β region is wing-shaped. Moreover, the extensions of three planar regions intersect roughly at an axis(denoted as a long arrow) with an included angle of about 120° from each other. On the other hand, $\alpha+\beta$ (purple) domains appear between aforementioned regions and mix slightly with them. Some β domains, such as β-helix folds, deviate from β planar region because of more parallel β-sheets occurring in all-β domains (Fig. S5-B). (2) The map in Fig. 4-C shows the distribution of domain sizes in the same viewpoint of Fig. 4-A. Domains are gradually rendered from blue to red corresponding to their sizes from small to large. According to Class map, two joint points, S_0 and S_1, between three regions are found on the intersecting axis of planar regions. S_0 is only shared by the sickle points of α and β sickle-shaped regions while S_1 is the meeting point of all regions. Domains close to S_0 are small. When moving away from S_0 to S_1 along α, β and α/β regions respectively, the sizes of domains gradually increase. In other words, the intersecting axis of planar regions shows a significant trend of domain sizes from small(blue) to large(red). (3) Fig. 4-D, E, F and Fig. S5-R represent four composition maps of near contacts for domains respectively. Fig. 4-D is shown in the same viewpoint of Fig. 4-B while Fig. 4-E and Fig. 4-F are shown in the same viewpoint of Fig. 4-A. These composition maps are rendered by similar color schemes which label domains from blue to red according to their composition feature values from small to large. Fig. 4-D shows that domains in α and β regions have small values of $m_{\beta\|}$, domains deviating from α or β regions and domains in α/β regions have large values of $m_{\beta\|}$, and the larger the value is, the farther the domain deviates downward from α or β regions to the edge of α/β regions(see also Fig. S5-H). Fig. 4-E shows that domains in α region and β regions have large values and small values of m_α respectively. When moving from the leftmost point of α region to the intersecting axis and to the rightmost point of β region, the m_α values domains gradually decrease (see also Fig. S5-I from top to bottom). Domains in α/β region also follow the trend and they are just located on the intersecting axis according to the moving direction so as that they have median values of m_α (see also Fig. S5-J). Fig. 4-F shows that domains in α region and β regions have small values and large values of $m_{\beta\perp}$ respectively. When moving in the aforementioned direction, the $m_{\beta\perp}$ values domains gradually increase(see also Fig. S5-L from top to bottom). Again, domains in α/β region follow the trend(see also Fig. S5-M). Fig. S5-R shows the trend of $m_{\alpha\#}$ and a similar change as in Fig. 4-E, therefore it validates that most non-hydrogen bondings occur between α-helix and other SSE. Similar class maps and composition maps of the whole S_H500 and CATH_32 datasets ($\sim$100,000 domains) are also shown in Fig. S6 and Fig. S7 respectively.

Table 2. Agreement with SCOP classification using moment.

Classifier*	Class† (%)	Fold† (%)	Superfamily† (%)
RF	99.5131	98.5392	98.1515
SVM	99.7870	99.3166	99.2058

*RF:Radom Forest, SVM:Support Vector Machines.
†Class, Fold and Superfamily levels contain 4, 352 and 500 types respectively.
doi:10.1371/journal.pone.0083788.t002

Overall speaking, the architecture of our domain structure universe is consistent with those of former works [10,22] in terms of the following aspects. (1) Most of α, β and α/β domains are grouped into three separated Class clusters of which each shows roughly a plane(called Class plane), and $\alpha+\beta$ domains occur in their joints. (2) The size of domains increases gradually when moving away from the small-size point to the small-large point along with α, β and α/β planes respectively. Moreover, fewer α/β domains occur at the small-size region than other classes of

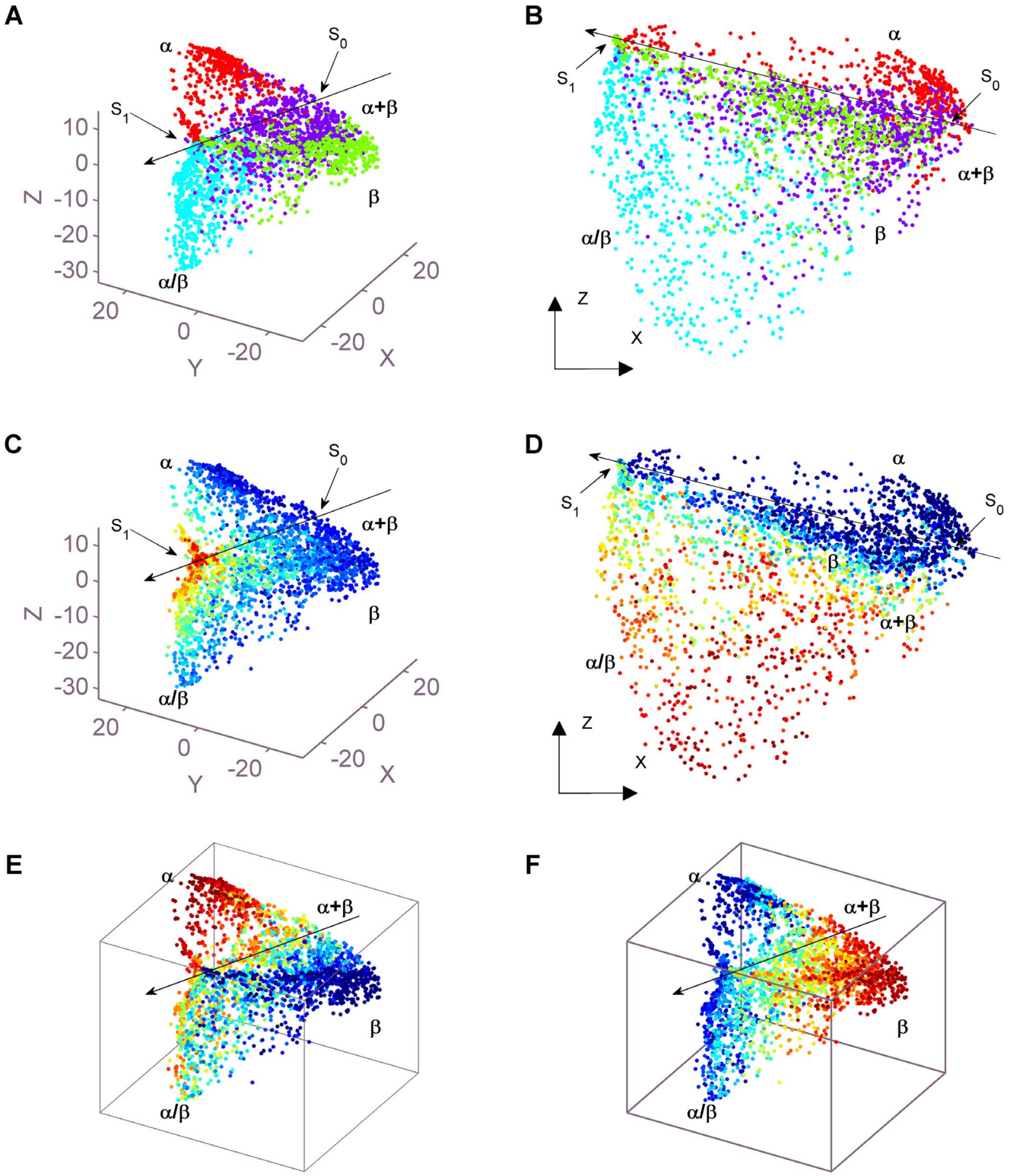

Figure 4. Maps of domain structure universe. (A)(B)Class map in two viewpoints. All-α (red), all-β (green) and α/β (cyan) domains are clustered into three distinct zones which appear quite separately in three planar region respectively. In particular, α region and β region are sickle-shaped but nearly mirror-symmetric, while α/β region is wing-shaped. Moreover, the extensions of three planar regions intersect approximately at an axis with an included angle of about 120° from each other. On the other hand, $\alpha+\beta$ (purple) domains appear between aforementioned regions and mix slightly with them. In (A)(B), red for all-α, green for all-*beta*, cyan forα/β, purple for $\alpha+\beta$. X, Y and Z denote the first three principal components respectively. (C)The distribution of domain sizes. This map is shown in the same viewpoint of (A). Domains are gradually rendered from blue to red corresponding to their sizes from small to large. According to Class map, two joint points, S_0 and S_1 between three regions are found on the intersecting axis(denoted as a long arrow) of planar regions. S_0 is only shared by the sickle points of α and β sickle-shaped regions while S_1 is the meeting point of all regions. Domains close to S_0 are small. When moving away from S_0 to S_1 along with α, β and α/β regions respectively, the sizes of domains

gradually increase. In other words, the intersecting axis of planar regions shows the significant trend of domain sizes from small to large. (D), (E) and (F) are the composition maps of $m_{\beta\|}$, m_{α} and $m_{\beta\perp}$ respectively. In (D)-(F), similar color schemes are adopted to label domains from blue to red according to their composition feature values from small to large. (D) is shown in the same viewpoint of (B) while (E) and (F) are shown in the same viewpoint of (A). (D) shows that domains in α and β regions have small values of $m_{\beta\|}$, domains deviating from α or β regions and domains in α/β regions have large values of $m_{\beta\|}$, and the larger the value is, the farther the domain deviates downward from α or β regions to the edge of α/β regions. (E) shows that domains in α region and β regions have large values and small values of m_{α} respectively. When moving from the leftmost point of α region to the intersecting axis and to the rightmost point of β region, the m_{α} values domains gradually decrease. Domains in α/β region also follow the trend and they are just located on the intersecting axis according to the moving direction so as that they have median values of m_{α}. (F) shows that domains in α region and β regions have small values and large values of $m_{\beta\perp}$ respectively. When moving in the aforementioned direction, the $m_{\beta\perp}$ values domains gradually increase. Again, domains in α/β region follow the trend.
doi:10.1371/journal.pone.0083788.g004

domains. (3) The type of SSE including α-helix and β-strand, and the parallelism of β-sheets between β and $\alpha+\beta$ domains are also strongly related to corresponding Class planes respectively [10]. (4) Some of β domains deviate greatly from β plane due to their types of β-roll or β-helix folds which contain more parallel β sheets than anti-parallel β sheets [22]. (5) Additionally, the extension of our α, β and α/β planes intersecting at a specific angle is similar to the universe in [10], but is slightly different from the universe in [22] which suggests that α and β domains are coplanar and α/β plane is perpendicular to the $\alpha-\beta$ plane. On the other hand, our universe has other advantages. First, a better visualization of domain distribution is provided for small- or middle-size domains rather than grouping them together (what other works did). This is important as most domains are less than 300 in size. Secondly, a quantitative visualization of near residue-residue contacts, including hydrogen bonds and non-hydrogen interaction, is provided by the corresponding composition features.

One Utility of Domain Structure Universe: Function Diversity of Superfamily

A set of domains in a superfamily has a common functional ancestor [29]. Therefore, the analysis of structural domain superfamilies provides an important insight for the exploration of the evolution of protein structure and function. Although domain undergoes structural changes during evolution, the structural diversity of domains in a superfamily is caused by their extensive structure embellishments and not by the common core which is shared by all members in the superfamily [2]. More importantly, structural diversity in a superfamily plays a crucial role in functional variations [2,30]. An evolutionary explanation can be found in [31].

The variations of domain combinations can make significant contribution to the evolution of organismal function [29]. The creation of new proteins is predominantly caused by frequent rearrangements of existing domain combinations, including duplication and permutation, rather than *de novo* creation [31–33]. Consequently, the functional diversity of superfamily is mainly caused by both the structural variation of domains [2] and the combination arrangement of domains [34]. Here we attempt to make a link between structural variation and combination diversity of domain within a superfamily. We measure structural variation by the number of clusters in our structural space of domains. We believe that our moment vectors can provide enough information for grouping domains with similar structures together, thus the number of clusters could reflect the number of structural variations(that is also related to the diversity of functions) among the domains. We denote the number of clusters for each superfamily as $\#C$, which is determined automatically by Mean Shift clustering method [35,36] (of which the whole procedure is described in Section S1).

Using $\#C$ as a measure of structural variation across a superfamily has also been used in previous work [37]. In order to validate such measurement, another independent measure of structural variation is calculated by a pairwise structural alignment algorithm, called jFATCAT (freely available in http://www.rcsb.org/pdb/workbench/workbench.do). We selected two superfamilies, a.60.9 and a.69.1, from SCOP (V1.75) and investigated the alignment scores between domains in each superfamily respectively. The higher the alignment score is, the less structural variations two domains have. The detailed scores are listed in Score S1. In the proposed structural space, the superfamily a.60.9 (SCOP Name: *lambda integrase-like, N-terminal domain*) shows 3 clusters which contain 48, 12 and 1 domains and are rendered by red, green and blue respectively (see Fig. 5-A). The ranges ([minima, maxima]) of alignment scores within each cluster and between any cluster pair are calculated. In details, the range of alignment scores within the red cluster is [292.48, 335.52], while the score range between red and green clusters is [106.41, 128.05] and that between red and blue clusters is [153.82, 177.25]. And the score range within green cluster is [313.57, 359.92] while the score range between green and blue clusters is [114.83, 123.53]. Obviously, the scores are higher within each cluster, the scores are lower between domains in different clusters, and there is even no overlap between score ranges of domains within-cluster and between-clusters. This provides an additional independent evidence showing that moment features are useful to classify structural difference of domains. More importantly, three clusters just represent three different types of tyrosine recombinases (*Flp recombinase, Cre recombinase and Recombinase XerD*) which share conserved DNA binding mechanism in recombination reaction, but also show apparent mechanistic and regulatory differences [38]. Another superfamily a.69.1(SCOP Name: *C-terminal domain of alpha and beta subunits of F1 ATP synthase*) also shows similar results of structural comparison. It has 2 spatial clusters in structural space, each of which is composed of 47 domains. One cluster denotes *alpha subunit* of *F1 ATP synthase* and another denotes its *beta subunit*. Three *beta subunits* in *F1* component are the ATP-ADP binding sites whereas *alpha subunits* are not sites but just support the *F1 ATP synthases* structure, even though they are placed together with the alternating arrangement in *F1 ATP synthase* [39]. Consequently, the above results demonstrates that structural variations of domains in a superfamily strongly related to the diversity of their functions according to the annotations in SCOP, and also illustrate that $\#C$ can capture the number of diverse functions, despite the fact that the domains can still have common conserved functional features, in a superfamily.

Combination diversity is investigated by SUPERFAMILY which provides a domain-based gene ontology at the superfamily level [40]. A protein consists of one or more domains and a multi-domain protein generally show a specific combination of domains according to domain counts and superfamily types [40]. Here, unique superfamily arrangement of domains is defined as the unique sequential combination of superfamily labels of member domains in a protein and a unique arrangement usually can be usually shared by multiple proteins. For example, considering 2 proteins, one contains sequentially three domains which belong to

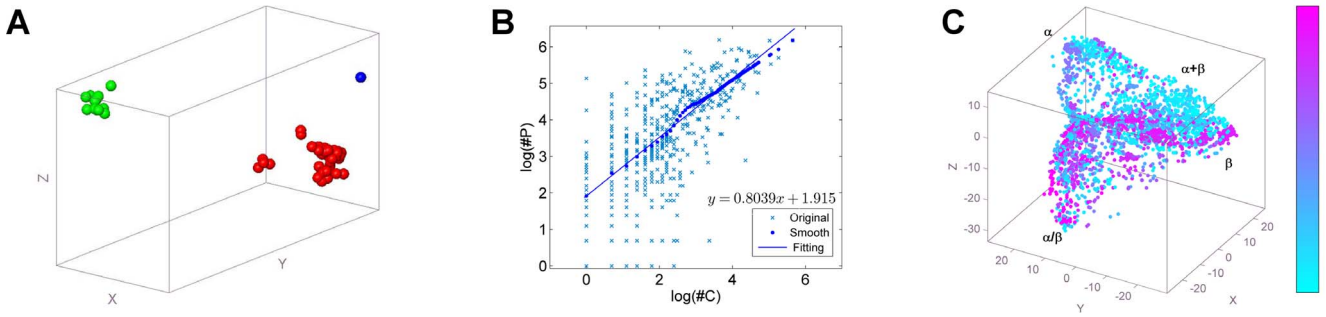

Figure 5. The relationships between structural variation and domain combination in superfamily. (A)The domain clusters of superfamily a.60.9. The superfamily shows three clusters in the proposed structural space. Red(48 domains), green(12 domains) and blue(1 domain) clusters represent exactly three types of tyrosine recombinases: Recombinase XerD, Flp recombinase and Cre recombinase respectively. (B)The relationship between the count of clusters ($\#C$) and the count of non-repeated partners ($\#P$). The logarithmic values of $\#C$ and $\#P$ in selected superfamilies, their smoothed values and the fitting line are drawn together to illustrate the significant linear relationship between $\#C$ and $\#P$ in logarithmic coordinates system. (C)Non-repeated partners map for superfamilies. Domains are labelled from cyan to purple according to $\#P$ values of their corresponding superfamilies from small to large. Most of superfamilies holding large $\#P$ appear in α/β plane, and some in β plane, especially near the edge of β plane.
doi:10.1371/journal.pone.0083788.g005

SF_p, SF_q and SF_p superfamilies respectively, so it holds the domain arrangement $[SF_p,SF_q,SF_p]$. If another protein has the same arrangement in terms of superfamily, then only one unique arrangement occurs between two proteins. If not, such as $[SF_p,SF_p,SF_q]$, two unique arrangements are counted. Here, $\#A$ is defined as the number of unique superfamily arrangements for a given superfamily. Moreover, the set of non-repeated SF_p-exclusive superfamilies in the unique arrangement A_i including superfamily SF_p is called SF_p's partners in A_i. For all unique arrangements $\{A_i\}$ including SF_p, a non-repeated set of its partners in $\{A_i\}$ can be obtained. For example, given three unique arrangements $[SF_p,SF_q,SF_p]$, $[SF_q,SF_p]$ and $[SF_u,SF_u,SF_p]$. The unique set of SF_p's partners is $\{SF_q,SF_u\}$, while SF_q's unique partner and SF_u's unique partner are only $\{SF_p\}$. Here, $\#P$ is defined as the number of non-repeated partner superfamilies in all unique arrangements for a specific superfamily. These two counts, $\#A$ and $\#P$, are able to approximate the function diversity of the superfamily in a large extent [40]. As the information in SUPERFAMILY is not complete, the two counts are missing in some superfamilies, such as, *viral protein domain* superfamily(b.19.1). In total, 475 superfamilies with both non-zero values of $\#A$ and $\#P$ from the dataset S_H500 are studied. The Spearmans correlation coefficients of $\#C$ to $\#A$ and $\#P$ are 0.6234 and 0.6785 respectively with $p-values<2e-16$, with larger Spearmans correlation coefficients 0.6763 and 0.7238($p-values<2e-16$) to $\#A$ and $\#P$ respectively in the case of bigger superfamilies ($size>150$). Therefore, the count of clusters ($\#C$) shows significant correlation with the count of non-repeated partners ($\#P$) and the count of unique arrangements ($\#A$).

To investigate the relations between $\#C$, $\#P$ and $\#A$ values, we applied a robust smoothing algorithm [41] to eliminate the noise in $\#C$, $\#P$ and $\#A$, and then provided a fitting. By considering potential power-law-distribution, we used natural logarithm of values in $\#C$, $\#P$ and $\#A$ instead of their original values. Smoothed data shows a linear relationship between $log(\#C)$ and $log(\#P)$ (Fig. 5-B) or $log(\#A)$ (Fig. S8), i.e. the relationships between $\#C$ and $\#P$ or $\#A$ follow the power-law-distribution. We studied the smoothed $\#P$ of each superfamily in domain structural space by rendering the points according to the smoothed values of $\#P$ (Fig. 5-C). Most of superfamilies with big $\#P$ appear in α/β plane, and some in β plane, especially in the edge of β plane. This result may be explained by the conformational stability of proteins, with the α/β folds most stable, followed by the all-β folds [42]. The more stable the protein is, the higher degree of mutation it can tolerate [31]. Accordingly, the superfamilies including stable proteins have more chances to produce diverse embellishments in evolution.

From the viewpoint of structural space, the more clusters a superfamily has, the bigger structural variation it has, the more diverse its embellishments are, and the more variant functions occur within superfamily [2,30]. From the viewpoint of evolution, the larger $\#P$ or $\#A$ of a superfamily are, the higher diversity of function the superfamily has. Our moment feature representation links up these two aforementioned views and supports the hypothesis that the evolution might be more disposed for selecting domains from a superfamily with a higher structural variation in order to produce new proteins with domain combination [31–33].

Discussion

Based on the image patterns of the hydrogen-bond contacts of α-helix and β-sheet and other hydrophobic contacts in distance matrix, we decompose the distance matrix image into basic binary images representing elementary SSEs and motifs of domain structures. By further deriving image moments from these basic images, we propose a moment feature vector to capture the structural characteristics of a protein domain. This feature vector was demonstrated to be useful in improving the domain classification accuracy, and provided a clear domain structure universe for the study of the distribution of domains. The findings of the distributions (e.g. length) are consistent with [22] and [10].

The same structure universe also demonstrates a positive relationship between the degree of structural variation (approximated by the number of clusters of the superfamily in the universe) and functional diversity (approximated by the number of non-repeated partners and the number of unique arrangements of the superfamily). Moreover, in the structure universe, the partner map of domain combination shows a significant distribution of superfamily combination diversity.

Our work can be explored for other applications. A popular utility is to query some newly determined structures with unknown functions. Mapping it into a structure universe and comparing to other domains will give its structural classification. The feature vector with existing retrieval techniques in computer vision

provides an efficient search of structural similarity. Furthermore, diverse moments can be expected to contribute to model domain structure mathematically, and can be used for structure prediction. Finally, the binary contact matrix can be used for other applications, for example, automatic detection of the common structural core within a superfamily and automatic multi-domain decomposition.

Supporting Information

Figure S1 Comparison of different classes of domains. The structures, distance matrices and four binary contact matrices including B_{α}, $B_{\beta\|}$, $B_{\beta\perp}$ and $B_{\alpha\#}$ are listed from left to right. The names of four domains and their lineage of SCOP classification are d1fnna1(a.4.5.11), d1beba_(b.60.1.1), d2c0ga2(c.47.1.7), and d1tdja2(d.58.18.2) from the top down. Only the upper triangular part of each BCM is used because of its symmetry.
(TIFF)

Figure S2 The classification of different folds within a class. Three folds, a.51 (Cytochrome c oxidase subunit h, 4 helices, irregular array, disulfide-linked,46 domains), a.52 (Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin, 4 helices, folded leaf, right-handed superhelix, 47 domains) and a.56(CO dehydrogenase ISP C-domain like, 4 helices, bundle, 45 domains) are investigated. For two given folds, one of moments of $B_{\alpha\#}$ can separate them according to the histogram of moment values. The height of each bar is the count of domains within a specific range of moment values.
(TIFF)

Figure S3 The classification of different superfamilies within a fold. Three superfamilies, a.4.7(Ribosomal protein L11, C-terminal domain, 105 domains), a.4.8(Ribosomal protein S18, 72 domains, and a.4.12 (TrpR-like, contains an extra shared helix after the HTH motif, 42 domains) are investigated. For two given superfamilies, one of moments of $B_{\alpha\#}$ can separate them according to the histogram of moment values. The height of each bar is the count of domains within a specific range of moment values.
(TIFF)

Figure S4 The distribution of domain sequence length. The fitting function is a power law function and the fitting performance is indicated by R-square = 0.9917 and RMSE = 9.93.
(TIFF)

Figure S5 Map of domain structure universe with more perspectives. Class map is shown in X-Y view (A) and Y-Z view(B). Size map is shown in X-Y view(C), Y-Z view(D), and X-Z view(E). Composition map of $m_{\beta\|}$ is shown in X-Y view(F), Y-Z view(G), and a 3-D view(H). Composition map of m_{α} is shown in X-Y view(I), Y-Z view(J), and X-Z view(K). Composition map of $m_{\beta\perp}$ is shown in X-Y view(L), Y-Z view(M), and X-Z view(N). Composition map of $m_{\alpha\#}$ is shown in X-Y view(O), Y-Z view(P), X-Z view(Q), and a 3-D view(R). The color schemes are the corresponding ones used in Fig. 4 and described in main text.
(TIFF)

Figure S6 The maps of domain structure universe of SCOP. Totally, 109,533 domains are drawn in maps. (A) In the map of class, red for all-α, green for all-β, cyan for α/β,purple for $\alpha+\beta$. (B) Map of 500 Superfamilies. (C)–(F) Map of composition moments: m_{α}, $m_{\alpha\#}$, $m_{\beta\perp}$ and $m_{\beta\|}$, and the values of composition moments go incrementally from blue to red.
(TIFF)

Figure S7 The maps of domain structure universe of CATH. Totally, 98,033 domains are drawn in maps. (A) In the map of class, red for mainly α, green for mainly β, cyan for mixed α-β,purple for small protein. (B) Map of 500 Superfamilies. (C)-(F) Map of composition moments: m_{α}, $m_{\alpha\#}$, $m_{\beta\perp}$ and $m_{\beta\|}$, and the values of composition moments go incrementally from blue to red.
(TIFF)

Figure S8 The relationship between the count of clusters ($\#C$) and the count of unique arrangements ($\#A$). The logarithmic values of $\#C$ and $\#A$ in selected superfamilies, their smoothed values and the fitting line are drawn together to illustrate the significant linear relationship between $\#C$ and $\#A$ in logarithmic coordinates system.
(TIFF)

Table S1 Comparison of classifications on S_H500 with different feature groups.
(PDF)

Section S1 Mean shift clustering.
(PDF)

Score S1 Alignment Scores of Domains.
(XLSX)

Acknowledgments

This work was supported by Hong Kong Scholars Program (No.XJ2011028) and Small Project Funding of the University of Hong Kong (No.201209176053), and partially supported by China Postdoctoral Science Foundation (No.2012M521803), Fundamental Research Foundation of Northwestern Polytechnical University in China (No.JC201164), National Natural Science Foundation of China (No.11171086), Shenzhen basic research project (No.JCYJ20120618143038947) and Hong Kong GRF grant(No.HKU719611E).

Author Contributions

Conceived and designed the experiments: JYS YNZ FYLC. Performed the experiments: JYS. Analyzed the data: JYS SMY FYLC. Contributed reagents/materials/analysis tools: YNZ. Wrote the paper: JYS SMY FYLC. Developed the codes used in analysis: JYS.

References

1. Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. Nature 372: 631–634.
2. Dessailly BH, Redfern OC, Cuff A, Orengo CA (2009) Exploiting structural classifications for function prediction: towards a domain grammar for protein function. Curr Opin Struct Biol 19: 349–356.
3. Osadchy M, Kolodny R (2011) Maps of protein structure space reveal a fundamental relationship between protein structure and function. Proc Natl Acad Sci USA 108: 12301–12306.
4. Valas RE, Yang S, Bourne PE (2009) Nothing about protein structure classification makes sense except in the light of evolution. Curr Opin Struct Biol 19: 329–334.
5. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11: 739–747.
6. Holm L, Rosenstrom P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38: W545–W549.
7. Røgen P, Fain B (2003) Automatic classification of protein structure by using Gauss integrals. Proc Natl Acad Sci USA 100: 119–124.
8. Aung Z, Tan KL (2007) Rapid retrieval of protein structures from databases. Drug Discov Today 12(17–18): 732–739.
9. Chi PH, Pang B, Korkin D, Shyu CR(2009) Efficient SCOP-fold classification and retrieval using index-based protein substructure alignments. Bioinformatics 25: 2559–2565.
10. Choi IG, Kwon J, Kim SH (2004) Local feature frequency profile: A method to measure structural similarity in proteins. Proc Natl Acad Sci USA 101: 3797–3802.

11. Penner RC, Knudsen M, Wiuf C, Andersen JE (2011) An Algebro-topological description of protein domain structure. PloS One 6: e19670.
12. Røgen P, Bohr H (2003) A new family of global protein shape descriptors. Math Biosci 182: 167–181.
13. Penner RC, Knudsen M, Wiuf C, Andersen JE (2010) Fatgraph Models of Proteins. Communications on Pure and Applied Mathematics 63: 1249–1297.
14. Kaufman L, Rousseeuw PJ (1990) in Finding Groups in Data: An Introduction to Cluster Analysis (Wiley, New York), 68163.
15. Chi PH, Shyu CR, Xu D (2006) A fast SCOP fold classification system using content-based E-predict algorithm. BMC Bioinformatics 7: 362.
16. Budowski-Tal I, Nov Y, Kolodny R (2010) FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. Proc Natl Acad Sci USA 107: 3481–3486.
17. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, et al. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35: D291–D297.
18. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, et al. (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36: D419–D425.
19. Shi JY, Zhang YN (2009) in Proceedings of 4th IAPR International Conference on Pattern Recognition in Bioinformatics, eds Kadirkamanathan V, Sanguinetti G, Girolami M, Niranjan M, Noirel J, (IEEE Computer Society, Los Alamitos, CA), pp 344–353.
20. Porat M, Zeevi YY (1988) The Generalized Gabor Scheme of Image Representation in Biological and Machine Vision. IEEE Trans Pattern Anal Mach Intell 10: 452–468.
21. Teh CH, Chin RT (1988) On Image-Analysis by the Methods of Moments. IEEE Trans Pattern Anal Mach Intell 10: 496–513.
22. Hou JT, Sims GE, Zhang C, Kim SH (2003) A global representation of the protein fold space. Proc Natl Acad Sci USA 100(5): 2386–2390.
23. Hou JT, Jun SR, Zhang C, Kim SH (2005) Global mapping of the protein structure space and application in structure-based inference of protein function. Proc Natl Acad Sci USA 102(10): 3651–3656.
24. Choi IG, Kim SH (2006) Evolution of protein structural classes and protein sequence families. Proc Natl Acad Sci USA 103(38): 14056–14061.
25. Schölkopf B, Smola A, Mller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5): 1299–1319.
26. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, et al. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer 8: 37–49.
27. Bartenhagen C, Klein HU, Ruckert C, Jiang X, Dugas M (2010) Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. BMC Bioinformatics 11: 567.
28. Vert JP (2002) A tree kernel to analyse phylogenetic profiles. Bioinformatics 18: S276–S284.
29. Ponting CP, Russell RR (2002) The natural history of protein domains. Annu Rev Biophys Biomol Struct 31: 45–71.
30. Dessailly BH, Redfern OC, Cuff AL, Orengo CA (2010) Detailed Analysis of Function Divergence in a Large and Diverse Domain Superfamily: Toward a Refined Protocol of Function Classification. Structure 18: 1522–1535.
31. Chothia C, Gough J (2009) Genomic and structural aspects of protein evolution. Biochem J 419(1): 15–28.
32. Yang S, Bourne PE (2009) The evolutionary history of protein domains viewed by species phylogeny. PloS One 4(12): e8378.
33. Bornberg-Bauer E, Huylmans AK, Sikosek T (2010) How do new proteins arise? Curr Opin Struct Biol 20(3): 390–396.
34. Bashton M, Chothia C (2007) The generation of new protein functions by the combination of domains. Structure 15(1): 85–99.
35. Cheng YZ (1995) Mean Shift, Mode Seeking, and Clustering. IEEE Trans Pattern Anal Mach Intell 17(8): 790–799.
36. Comaniciu D, Meer P (2002) Mean shift: A robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24(5): 603–619.
37. Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA (2006) Structural diversity of domain superfamilies in the CATH database. J Mol Biol 360(3): 725–741.
38. Swalla BM, Gumport RI, Gardner JF (2003) Conservation of structure and function among tyrosine recombinases: homology-based modeling of the lambda integrase core-binding domain. Nucleic Acids Res 31(3): 805–818.
39. Walker JE, Fearnley IM, Lutter R, Todd RJ, Runswick MJ (1990) Structural aspects of proton-pumping ATPases. Philos Trans R Soc Lond B Biol Sci 326: 367–378.
40. de Lima Morais DA, Fang H, Rackham OJ, Wilson D, Pethica R, et al. (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res 39: D427–434.
41. Cleveland WS (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. J Am Stat Assoc 74(368): 829–836.
42. Minary P, Levitt M (2008) Probing protein fold space with a simplified model. J Mol Biol 375(4): 920–933.