Modularity in protein structures: study on all-alpha proteins

Taushif Khan and Indira Ghosh*

School of Computational & Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India

Communicated by Ramaswamy H. Sarma

(Received 2 October 2014; accepted 30 December 2014)

Modularity is known as one of the most important features of protein’s robust and efficient design. The architecture and topology of proteins play a vital role by providing necessary robust scaffolds to support organism’s growth and survival in constant evolutionary pressure. These complex biomolecules can be represented by several layers of modular architecture, but it is pivotal to understand and explore the smallest biologically relevant structural component. In the present study, we have developed a component-based method, using protein’s secondary structures and their arrangements (i.e. patterns) in order to investigate its structural space. Our result on all-alpha protein shows that the known structural space is highly populated with limited set of structural patterns. We have also noticed that these frequently observed structural patterns are present as modules or “building blocks” in large proteins (i.e. higher secondary structure content). From structural descriptor analysis, observed patterns are found to be within similar deviation; however, frequent patterns are found to be distinctly occurring in diverse functions e.g. in enzymatic classes and reactions. In this study, we are introducing a simple approach to explore protein structural space using combinatorial- and graph-based geometry methods, which can be used to describe modularity in protein structures. Moreover, analysis indicates that protein function seems to be the driving force that shapes the known structural space.

Keywords: Topology space; contact string; secondary structure; prevalent pattern; modules; structural descriptor; protein function; enzyme reaction

1. Introduction

Biological systems are complex, but modular in nature. Modularity has been observed in almost all levels. Starting from biomolecules (e.g. protein structures), pathways (e.g. gene regulation, metabolic networks), and cellular to organism level, the concept of modularity has been explored (Hleap, Susko, & Blouin, 2013; Lammert, Schug, & Onuchic, 2009; Pereira-Leal, Levy, & Teichmann, 2006). Integration and complex interplay among different levels of cellular components (or modules) are crucial for survival of an organism (Rorick, 2012; Tasdighian et al., 2013). In generic terms, modules could be described as the mutually exclusive individual components of a system which act coherently (Pereira-Leal et al., 2006; Wagner, Pavlicev, & Cheverud, 2007). Concept of modularity is the key feature of design as it provides necessary robustness and flexibility to the system and are found in (or adopted) in various engineering principles and computer programming models (Kitano, 2004). But in case of biological systems, defining “module” becomes a challenging problem, especially in the context of protein structures (Banerji & Ghosh, 2011; Rorick, 2012). Identification of the smallest component of protein’s modular structural arrangement will provide an insight into its evolution and structural–functional space.

Modularity in proteins has been already known from genetic (e.g. exon shuffling, duplication, fusion [Broom et al., 2012; Chothia & Gough, 2009]) and folding studies (Efimov, 1998; Murzin et al., 1995; Todd, Orengo, & Thornton, 2001). While domain remains the evolutionary unit of protein structures, definition of modules has been approached from subdomain (Tasdighian et al., 2013), non-evolutionary or structural levels (Hleap et al., 2013; Wathen & Jia, 2013). Seminal work on protein symmetry (Myers-Turnbull et al., 2014) and identification of smaller fragments (or modules) to build complex folding architecture like β Trefoil, β propeller, and (βα)n-barrel (or TIM barrel) (Broom et al., 2012; Eisenbeis et al., 2012) emphasizes the presence of basic building blocks or subdomain modules (for recent review (Höcker, 2014)). In terms of protein folding, modularity has been described as close loops, foldons, and automatic folding units (Panchenko, Luthey-Schulten, & Wolynes, 1996). All definitions of modules have different average size (Rorick, 2012; Tasdighian et al., 2013), but the main essence revolves around combination of secondary structures which has been explored in this study.

Using component-based approach (by considering the arrangement of secondary structural elements (SSEs)), in this study, we have presented a systematic...
approach to identify structural modules in alpha (α) helical proteins. In strict geometrical sense, protein fold could be defined as architecture of SSEs with their chain connectivity, relative orientation, and packing; also referred as topology (Skolnick et al., 2009; Srinivasan & Rose, 1995; Taylor et al., 2009). It has been proposed that folds are explored by genetic mechanisms of gene duplication, insertion, deletion, and circular permutations, which are incorporated into protein structure (Kolodny et al., 2013) and can be accounted by analyzing protein structures at secondary structure level (Gordeev & Efimov, 2013; Przytycka, Srinivasan, & Rose, 2002). Hence, using topology, fold space can be explored aptly. There has been different attempts to identify structural modules in proteins without considering the evolutionary or functional relation (Efimov, 1997; Gordeev & Efimov, 2013; Skolnick & Gao, 2013; Taylor, 2002; Wathen & Jia, 2013). PCBOST is an emerging approach which tries to mimic the hierarchical folding by constructing a “structural tree” which can trace back to a “root motif” (Efimov & Kondratova, 2003; Gordeev & Efimov, 2013). In its recent addition of α/β structural tree (Gordeev & Efimov, 2013), it has been illustrated that occurrence of some structural motifs are higher than others whose reasons are yet to be acknowledged. Similar observation has been recorded for β –proteins. The theory of limited fold space and biased distribution of protein among different folds have been supported by recent studies (Cuff et al., 2009; Grainger, Sadowski, & Taylor, 2010; Hleap et al., 2013), which not only increases the perplexity of the problem but also makes it more interesting to understand the contributing factors shaping protein fold space. Limited use of structural space may be induced by different structural properties
(Paola et al., 2012; Tasdighian et al., 2013) or due to certain functional objective (Myers-Turnbull et al., 2014) or because of easy accessibility of folded structure (Cossio et al., 2010; Gianni et al., 2009; Onuchic & Wolynes, 2004).

Graph abstraction or network-based methods have been widely applied to explain and understand complex systems across disciplines (Kloczkowski et al., 2009; Kun & Scheuring, 2009; Lorenz, Jeng, & Deem, 2011; Tasdighian et al., 2013). Protein structures are also explained in terms of edges and nodes (contact networks), which can be described as a short molecular formula for macromolecules (Dokholyan et al., 2002; Hlep et al., 2013; Kamat & Lesk, 2007). In this study, we have proposed an efficient methodology to annotate and compare protein structures based on contact between SSEs (or tertiary contact). In present study, we have focused on all-alpha proteins by considering alpha helices (Kabsch & Sander, 1983) as the secondary structure and defined contact among them. By scanning known structural space, we have found that the protein structure space is modular and act as “building blocks” (or “legos”) for larger protein structures. In addition, it has been observed that some of the modules are repeatedly used to build proteins having diverse functionality.

2. Material and methods

Protein structures have been extracted from Protein Data Bank (PDB) (Berman et al., 2000). PDB, being the primary repository of biomolecular structural data, is subjected to multiple levels of redundancy (Berman et al., 2000; Thiruv et al., 2005) introduced mainly by evolutionary conserved proteins and different variants of same protein. These redundancies have been removed by clustering approach based on sequence identity cut-offs and selecting representative protein sequences among these clusters (Li & Godzik, 2006; Wang & Dunbrack, 2005).

To encounter rapid growth in structural data and for robustness of proposed result, we have formed four datasets of same structural quality (protein structure solved by X-ray diffraction method, resolution \(\leq 2.5\) Å and R-free <.30) but different sequence homology cut-offs and releases (Table S1). The main data-set D1 (#April, 2014) is composed of 977 protein chains with alpha helix ranging from 3 to 10. Similarly, data-set D2 (#June, 2013), D3 (Astral25 #June, 2009, SCOP v1.75), and D4 (with 25% sequence identity of D1) have been prepared for the study. Details of data selection process have been recorded in Table 1.

Table 1. Proteins and patterns of data-set D1.

| HS\(^a\) | # PDBs | Total Patterns | \(x_f\) \(^b\) | Prevalent \((x_{ob} >)\) | Others \((x_{ob} \leq x_f)\) |
|---|---|---|---|---|---|
| | # patterns | % patterns | % PDBs | # patterns | % patterns | % PDBs |
| 3 | 79 | 7 | .1428 | 3 | 42.85 | 66.6 |
| 4 | 135 | 41 | .2439 | 13 | 31.7 | 62.96 |
| 5 | 189 | 99 | .0101 | 39 | 39.39 | 68.25 |
| 6 | 117 | 98 | .0088 | 12 | 12.24 | 26.49 |
| 7 | 116 | 104 | .0096 | 8 | 7.69 | 17.24 |
| 8 | 92 | 91 | .0109 | 1 | 1.09 | 2.17 |
| Total | 977 | 660 | 94 | 566 |

Note: Data-set D1 has been grouped into different.

\(^a\)HS (Helix Set = number of alpha helices in a pdb chain) and number of protein chains (# PDBs) and corresponding total pattern count has been provided in the first three columns.

\(^b\)\(x_f\) is the expected pattern frequency i.e. fraction of protein chain with a particular pattern in each HS if all patterns have equal proportion of representation in the data-set \(\left(x_f = 1/\text{Total Pattern}(\text{HS})\right)\). In each HS, all patterns have been classified into two sets (“Prevalent” and “Others”) depending on their \(c\) observed pattern frequency \(\left(x_{ob} = 1/\text{number of representing pdb chain}\right)\). For both classes, number of patterns (# patterns); percentage of pattern in the set (% patterns), and corresponding percentage of protein chain (% PDBs) has been shown in corresponding sections. The last row shows the sum of protein chains, total number of patterns across HS, and number of prevalent as well as other patterns observed in data-set D1.
provided in the supplementary (Figure S1, S2, and Table S1). We have focused our analysis mainly on the latest data-set D1, whereas other data-sets (D2, D3, and D4) have been checked for consistency of results.

In each protein chain, contact between two helices has been considered if at least three contacting residues are established as per atom–atom contact (Chothia, 1981; Hleap et al., 2013; Konagurthu, Stuckey, & Lesk, 2008). An atom to atom contact is considered to be formed if distance between atoms is less than sum of their van der Waals radii with a threshold of .6 Å. If residues from different helices have at least one atom to atom contact,

Figure 3. Above figures illustrate pattern space and respective quartile scores in observed pattern space for (a) HS 3 and (b) HS 4. Relative percentile scores in each case has been shown by dark dotted line (Q = 75) for first quartile and gray dotted line for third quartile (i.e. Q < 25). For each case, pattern which is observed in respective HS has been shown by dark bars (labeled as “Pattern”) and as modules in higher HS has gray colored bars (labeled as “modules”).
we consider them as contacting residues. For a helix contact pair to form, it must have a minimum of three criteria for selection of contact residues (Chothia, 1981). The contact definition has been validated for different threshold values (Figure S3) and found that the used criteria is comparable to reported studies (Day et al., 2010; Konagurthu, Lesk, & Allison, 2012; Plaxco, Simons, & Baker, 1998; Yuan, Chen, & Kihara, 2012; Zuo, Wang, & Wang, 2006). Orientation of contacting helices have been assigned to be anti-parallel (a), parallel (p), and orthogonal (r) based on their relative helix-axis angles (Chothia, 1981; Chothia & Lesk, 1985) as illustrated in Figure 1.

For concise representation of contact among SSEs in proteins, we have developed “contact string”. Similar to “fingerprints”, given a set of SSEs, “contact string” can uniquely represent all different contact patterns. For efficiency in comparison and analysis, contact string have been comprised of contact information segments (separated by “-”), which describes the sequential distance in terms of SSEs (i.e. the first segment of contact string describes contact between SSEs which are sequentially adjacent and henceforth). Contact string is an abstraction of contact matrix (D), which represents all possible pairwise contact between SSEs as illustrated in Figure 1. This simplified representation can be utilized as a tool in various protein structure studies like 3D structure comparison, identifying folding patterns, and analyzing structural modules. However, increase in alpha helix content (h) simultaneously amplifies possible number of patterns \((2^h - 1)\) and length of contact string \(l = h \times (h - 1)/2\). This increases gap between observed and theoretically possible patterns (Figure 2). For systematic study, proteins in data-sets have been grouped as per number of alpha helix content (h) and patterns have been evaluated by constructing contact string. By considering the relative orientation of each contacting helix pair, the possible pattern space will further expand with the exponent of 4 (i.e. \(4^h\)). But, for statistical significance of protein distribution in each HS, pattern wise study has been undertaken (p-value <.01 Table S3, S4 for details) (Mann, 1999).

Table 2. Example of ranking and indexing of patterns observed in data-set D1.

| HS | Pattern | \(x_{ob}\) | Pattern index {HS}.P-{Rank} |
|----|---------|----------|-----------------------------|
| 3  (\(x_f = .142\)) | 01-0    | .3       | 3.P-1                       |
|    | 11-0    | .2       | 3.P-2                       |
|    | 10-0    | .16      | 3.P-3                       |
|    | 01-0    | .12      | 3.P-4                       |
|    | 11-0    | ...      | 3.P-7                       |
| 4  (\(x_f = .024\)) | 001-00-1 | .088     | 4.P-1                       |
|    | 111-00-1 | .066     | 4.P-2                       |
|    | 010-01-0 | ...      | 4.P-13                      |
|    | 110-00-1 | .022     | 4.P-14                      |
|    | 000-01-1 | ...      | 4.P-41                      |
| 5  (\(x_f = .0101\)) | 0100-000-00-0 | .0529 | 5.P-1 |
|    | 0001-000-00-0 | .04761 | 5.P-2 |
|    | 0000-001-00-0 | ...      | 5.P-39                       |
|    | 1111-001-10-0 | .0052 | 5.P-40                       |
|    | 0000-000-01-0 | ...      | 5.P-99                       |
| 6  (\(x_f = .0088\)) | 01010-0000-00-0-0 | .0291 | 6.P-1 |
|    | 1100-1010-0000-00-0 | .0218 | 6.P-2 |
|    | 0001-0000-0000-00-0 | ...      | 6.P-15                       |
|    | 11111-0101-0000-00-0 | .0072 | 6.P-16                       |
|    | 00000-0000-0000-01-0 | ...      | 6.P-113                      |

Notes: Example of ranking and indexing of patterns observed in data-set D1 in each HS has been shown in above table. For each HS (shown from 3 to 6 HS), corresponding expected pattern frequency \((x_f)\) has been reported in parenthesis of first column. Observed pattern frequency \((x_{ob})\) of each pattern in a HS has been provided. Patterns in each HS has been ranked in descending order of \(x_{ob}\). All observed patterns has been indexed as described in the last column representing helix set and corresponding rank associated to that pattern (pattern index = {HS}.P-{Rank}). In each HS, dotted rows show the discontinuity in pattern index (for detail index pattern see supplementary data).
Table 3. Illustration of Topological cartoon (prevalent patterns in data-set D1).

| Helix Set \( (X_f) \) | Cartoon of prevalent patterns \( (PP; X_f < X_{ob}) \) |
|-------------------------|--------------------------------------------------|
|                         | 3 (.1428)                                        |
|                         | ![Helix Cartoons 3](image1)                      |
|                         | 4 (.2439)                                        |
|                         | ![Helix Cartoons 4](image2)                      |
|                         | 5 (.0101)                                        |
|                         | ![Helix Cartoons 5](image3)                      |

(Continued)
Modules or substructures of a protein (with HS >3) could be explored using contact string. Similar to subgraph search method, for a protein with “h” alpha helix (i.e. HS = “h”), all possible substructures with HS (from 3 to h - 1 helix set) have been generated. Contact strings of helix subsets have been formed in sliding window-fashion (window size = 1 alpha helix), scanning from N to C terminal of protein chain. Formed module set is comprised patterns of all present substructures of helix subsets (i.e. HS = [3, h - 1]) in a protein with “h” alpha helix.

To compare different structural properties, we have analyzed structural descriptors of proteins within patterns of each HS. Structural descriptors like relative contact order (rCO) (Plaxco et al., 2000), normalized radius of gyration (nRg) (Ivankov et al., 2009), and loss in solvent

| Helix Set ($X_f$) | Cartoon of prevalent patterns (PP; $X_f < X_{ob}$) |
|------------------|-----------------------------------------------|
| 6 (.0088)        | ![Image of pattern set 6](image)               |
| 7 (.0102)        | ![Image of pattern set 7](image)               |

Notes: Illustration of cartoon representation of patterns that belongs to prevalent set in data-set D1. Cartoon represents alpha helices (in circles) with chain connectivity (bold lines) from N (in-arrow) to C terminal (out-arrow). Contact between alpha helices has been shown in dotted gray lines. In each HS, corresponding pattern index, contact string (in braces), and observed pattern frequency ($X_{ob}$) has been shown.
accessibility of helix (LSAH) of protein chains have been calculated and patternwise comparison has been performed. Used descriptors are reported to have the ability to describe higher variance in structural properties of proteins (Paola et al., 2012; Tasdighian et al., 2013).

Protein function has been analyzed based on homology (Engelhardt et al., 2011; de Lima Morais et al., 2011) and enzyme reaction comparison methods (Rahman et al., 2014). Function annotation has been assigned and compared based on latest release of functional data of superfamily database (de Lima Morais et al., 2011). Whereas, enzyme reaction similarity has been analyzed using EC-BLAST using similarity scores of bond change (BC), reaction center (RC), and reaction structure similarity (Holliday et al., 2014; Rahman et al., 2014). All similarity scores are reported in Jaccard coefficient ranging between 0 and 1. Using the reaction information from KEGG (Hattori et al., 2003) and (Alcántara et al., 2012) database, mapped reactions have been produced and their similarities have been evaluated.

All protocols (protein contact string generation, classify motif belongs to prevalent or not, calculation of protein descriptors and functional annotation) have been automated using python programming language, with SQL database (available as “ProLego” in http://ecbb.jnu.ac.in/proLego/index.html).

3. Results and discussion

We have investigated modularity in protein structure in a non-hierarchical approach, treating each HS (Helix Set) independently and analyzed observed pattern space. By clustering proteins in data-set as per number of alpha helix content, possible pattern (or topology) space for each HS has been explored (Figure 2). Comparing number of observed patterns to number of theoretically possible patterns from 3 to 10 HS, it can be observed that with current known structures only a fraction of possible topology space can be explored. Analyzing proteins from 2009 Astral data-set (i.e. D3) and current data-set (D1), a marginal expansion in observed patterns over the years can be noticed (from 524 in D3 to 660 in D1) (Table S2). However, study from all data-set suggests that observed pattern space is only fraction of possible pattern space (Figure S2).

Observed pattern space in each HS confirms the presence of certain patterns that have occurred more than expected frequency ($) (Table 1). For comparative analysis across HS, we have ranked observed patterns in each HS, as per fraction of occurrence in “relative percentile scores” (Q-score) and “observed frequency” (ob). These percentile scores are further classified into quartiles depending on Q-score (Figure 3). With corresponding ob each pattern has been assigned a “pattern index” which represents the relative ranking of that pattern in HS (Table 2). In each HS, only a small fraction of patterns have Q-score ≥ 75, which have observed frequency (ob) greater than expected frequency (e) (~12% in all HS). These high frequent pattern sets in each HS has been referred as “prevalent patterns”. Figure 3 shows the observed pattern space in 3 and 4 HS and their relative percentile scores. In case of 3 HS, prevalent patterns (like “01-0” and “11-1”) are more frequent (as Q-score ≥ 75), occurring in ~50% of structure, out of 7 possible patterns (Table 1 and Figure 3(a)). Similarly, as shown in Table 1 and Figure 3(b), frequent patterns for 4 HS proteins have been identified. The prevalent pattern sets are found to be consistent for all four data-sets with marginal variation in quartile scores. Across HSs, prevalent patterns could be identified for cases when available representative structure is statistically comparable to observed pattern space (results from $^2$-test for statistical significance Table S3, S4).

Limitation in observed pattern space can be explored by inclusion of subpatterns or “modules” in each HS. Using the contact string, structural modules in protein structure can be analyzed as described in “Material and methods” section. Inclusion of modules as patterns for each HS increases the observed pattern space by ~3.4 times in data-set D1 (from 660 to 2296) (Figure 2) and similar growth rate has been observed across data-sets (see Table S2, Figure S4). However, the expansion in observed pattern space is marginal as compared to possible space (as shown in Figure 2), although the 3-fold increase in pattern space is reflected in new patterns. Analyzing distribution of percentile rank, we have observed the relatively high frequency of same prevalent patterns (Figure 3) (i.e. up to 7 HS Table S4). Some changes have been noticed in the quartile ranking of prevalent patterns, but fraction of observed occurrence (ob) is always higher than expected (e). We have checked the presence of prevalent pattern sets in D3 by explicitly analyzing orthogonal and aligned helix orientation (Supplementary Section 1.2) and found similar sets of prevalent pattern sets with considerable deviation of observed frequency (ob ≥ e) (~12%). Combining these prevalent patterns and modules (statistically significant) across all HS, a prevalent pattern (PP) set has been proposed (Table 3). Cartoon depiction of representative protein structures from some of the prevalent pattern sets has been illustrated in Figure 4.

From observations (Tables 1 and 3 and Figure 3), it might be proposed that the prevalent patterns (of lower HS) has been reused by protein as building blocks for complex and large proteins (as higher HSs). The recurrence of common structural motifs (α-hairpin and αβ-motifs) (Broom et al., 2012) has recently been identified and discussed (Cuff et al., 2009; Gordeev & Efimov, 2013; Orengo et al., 1997). These motifs or super-secondary structures are smaller than a domain but could
be used as “Lego-blocks” or components to give more complex architecture (“Russian doll effect”) by duplication and (or) fusion events (Broom et al., 2012; Höcker, 2014; Orengo et al., 1997). It has also been shown that protein domain evolve their optimal structural design by exploring structure space at the level of contact between SSEs which can be referred as modules (Emmert-Streib & Mushegian, 2007). In the present study, we have vindicated the known hypothesis of modular architecture and identified prevalent patterns in case of mainly alpha proteins.

We attempted to analyze the rationale behind prevalence of certain pattern sets using different structural descriptors (described in “Material and methods”).
Pattern wise analyses are found to be within known average values and deviations (Ivankov et al., 2009) but have not provided significant difference between prevalent and other patterns (Figure 5, S6).

To understand why protein space has few prevalent pattern sets, it is crucial to study the underlying functional requirements that the protein scaffold fulfill. We have compared protein function with known functional terms (de Lima Morais et al., 2011) as well as by analyzing enzyme and its reaction (Tables 4, 5 and Figure 6). Data-set D1 has 224 distinct number of distinct scop superfamilies (sf) (SCOP v1.75, all-alpha has 507 sf). Different proportions of superfamily functional terms (from SUPERFAMILY) have been observed in data-set

![Figure 5](image-url)

Figure 5. Above figure shows normalized radius of gyration (norm_Rg), relative contact order (rCO) and Loss in helix accessible solvent area (%LSAS_H) from top to bottom, respectively, for proteins in (a) 3 and (b) 4 HS. X-axis shows the pattern index (as described in Table II) and corresponding average structural descriptor values are given in the y-axis. In all cases, prevalent pattern sets are marked as star, whereas, other patterns are marked as circle.
Table 4. Analyzing EC class similarity at substrate specificity level.

| HS  | # ECs (distinct) | Pattern sets (functional diversity score) |
|-----|----------------|------------------------------------------|
|     |                | Prevalent | Others                      |
| 3   | 9              | .148 (±.052) | .22 (±.090) |
| 4   | 13             | .11 (±.038)  | .076 (±.002) |
| 5   | 19             | .067 (±.032) | .053 (±.001) |
| 6   | 14             | .19 (±.089)  | .071 (±.017) |
| 7   | 12             | .167 (±.000) | .083 (±.00)  |
| 8   | 18             | .056 (±.000) | .055 (±.00)  |
| 9   | 16             | .062 (±.00)  | .062 (±.00)  |
| 10  | 23             | .044 (±.00)  | .043 (±.00)  |

Notes: Above table shows the average value of functional diversity score (=Number of distinct EC number in a pattern of a HS(i)/Total distinct ECs in a HS(i)) in prevalent pattern set and other pattern set. In each helix set (HS), total number of distinct ECs (#ECs) in data-set D1 has been given with corresponding functional diversity score for prevalent and other pattern set. Average value along with standard deviation has been provided for each set.

Table 5. Comparing EC-BLAST result for a case of prevalent and other pattern set of proteins from 5 helix set (HS).

| HS 5 (Example from “Prevalent pattern”) | E.C. | Substructure similarity (SS) | Bond change (BC) | Reaction center (RC) |
|----------------------------------------|------|-------------------------------|------------------|---------------------|
|                                        | 1.15.1.1 | 3.6.1.31 | 4.1.1.44 | 1.15.1.1 | 3.6.1.31 | 4.1.1.44 |
| Substructure similarity (SS)           | 1     | .01   | .04    | 1     | .25    | 1     |
| Bond change (BC)                       | .04   | 1     | .17    | .04   | 0      | 1     |
| Reaction center (RC)                   | 0     | .25   | 1      | 0     | .08    | .04   |

| HS 5 (Example from “Other Pattern set”) | E.C. | SS | BC | RC |
|----------------------------------------|------|----|----|----|
|                                        | 4.6.1.1 – 3.6.3.14 | .98 | .78 | .24 |

Notes: Enzymes are compared in three scores by EC-BLAST on balance reaction files of tabulated enzymes. The case has been extracted from HS 5, prevalent-patterns (a) while, (b) shows the similarity scores for “other” pattern of 5 HS.
This ability to perform diverse function by prevalent pattern sets has been identified across all HS’s and palpable from analyzing different functional annotation methods from homology-based to reaction-based methods (Figure S9, S10). This distinct difference leads to conclude that function might be the driving force for proteins to repeat (or reuse) “prevalent modules”.

4. Conclusion

It has been well known that evolution uses the preexisting parts (or modules) to invent new functionality or to conserve certain function (Hleaf et al., 2013; Lorenz et al., 2011). As observed in the present study, it can be proposed that protein structures evolve using limited set of patterns which act as “building block” to form more complex proteins and emerge newer modules, if necessary, mostly to satisfy cell’s functional requirements. These frequent or prevalent modules can be responsible for providing the required scaffold for wide spectrum of catalytic functions as present study shows. Influence of polarity between scaffold and active site regions on protein functional evolution has been reported in the context of enzyme activity with limited fold sets (Dellus-Gur et al., 2013; Khersonsky & Tawfik, 2010). The balance between scaffold and active sites not only increases the chance of evolution, but also influences organism’s fitness by improving functional efficiency (Tomatis et al., 2008). In present study, using non-hierarchical analysis of all-alpha proteins, we have reported prevalent patterns (total 82 among 5 types of HS from 3 to 7) which are ~.14 fraction of observed pattern space (Table 1).
These patterns (or modules) may provide necessary structure for diverse functional enzyme activities in proteins they occur and help in designing proteins with desired functions. It extends the support for the hypothesis that evolution shapes the protein structure space in the lights of function and prefers to reuse the existing known folds instead of exploring the astronomical available space, explaining the occurrence of the large gap between combinatorial possibility and observed patterns (Figure 2), and current limitations of unknown space (or “Dark matter”) of structure and function universe (Taylor et al., 2009). However, topological preference caused by easy kinetic accessibility to diverse structural space (Cossio et al., 2010; Leopold, Montal, & Onuchic, 1992; Wang et al., 2012) has not been addressed in the present study but in progress using advanced techniques like coarse-graining methods (Takada, 2012).

Author disclosure statement

The authors declare that no conflict of interests exist.

Abbreviation

| Abbreviation | Description         |
|--------------|---------------------|
| HS           | Helix Set           |
| SSE          | Secondary Structural Elements |
| CS           | Contact String      |
| PP           | Prevalent Pattern   |
| AFU          | Automated Folding Units |
| E.C.         | Enzyme Class        |

Acknowledgment

We acknowledge UGC-MANF fellowship and DBT-CoE for supporting this work. We will also like to acknowledge IG-group members for their comments.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2014.1003969.

References

Alcántara, R., Axelsen, K. B., Morgat, A., Belda, E., Coudert, E., Bridge, A., … Steinbeck, C. (2012). Rhea—a manually curated resource of biochemical reactions. Nucleic Acids Research, 40, D754–D760.

Banerji, A., & Ghosh, I. (2011). Fractal symmetry of protein interior: What have we learned? Cellular and Molecular Life Sciences, 68, 2711–2737.

Berman, H. M., Westbrook, J., Feng, Z., Gililland, G., Bhat, N. T., Weissig, H., … Bourne, E. P. (2000). The Protein Data Bank. Nucleic Acids Research, 28, 235–242.

Broom, A., Doxey, A. C., Lobanov, Y. D., Berthin, G. L., Rose, R. D., Howell, L. P., … Meiering, M. E. (2012). Modular evolution and the origins of symmetry: Reconstruction of a three-fold symmetric globular protein. Structure, 20, 161–171.

Chothia, C. (1981). Helix to helix packing in proteins. Journal of Molecular Biology, 145, 215–250.

Chothia, C., & Gough, J. (2009). Genomic and structural aspects of protein evolution. Biochemical Journal, 419, 15–28.

Chothia, C., & Lesk, A. (1985). Helix movements in proteins. Trends in Biochemical Sciences, 10, 116–118.

Cossio, P., Trovato, A., Pietrucci, F., Seno, F., Maritan, A., & Laio, A. (2010). Exploring the universe of protein structures beyond the protein data bank. PLoS Computational Biology, 6, e1000957.

Cuff, A., Redfern, O. C., Greene, L., Sillitoe, I., Lewis, T., Dibley, M., … Orenco, C. (2009). The CATH hierarchy revisited: structural divergence in domain superfamilies and the continuity of fold space. Structure, 17, 1051–1062.

Day, R., Lennox, K. P., Dahl, D. B., Vannucci, A., & Tañi, W. J. (2010). Characterizing the regularity of tetrahedral packing motifs in protein tertiary structure. Bioinformatics, 26, 3059–3066.

de Lima Morais, D. A., Fang, H., Rackham, O. J. L., Wilson, D., Pethica, R., Chothia, C., & Gough, J. 2011. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Research 39, D427–D434.

Dellus-Gur, E., Toth-Petroczy, A., Elias, M., & Tawfik, S. D. (2013). What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. Journal of Molecular Biology, 425, 2609–2621.

Dokholyan, N. V., Li, L., Ding, F., & Shakhnovich, I. E. (2002). Topological determinants of protein folding. Proceedings of the National academy of Sciences of the United States of America, 99, 8637–8641.

Efimov, A. V. (1997). A structural tree for proteins containing 3beta-corners. FEBS Letters, 407, 37–41.

Efimov, A. V. (1998). A structural tree for proteins containing S-like beta-sheets. FEBS Letters, 437, 246–250.

Efimov, A. V., & Kondratova, M. S. (2003). A comparative analysis of interhelical polar interactions of various alpha-helix packings in proteins. Molecular Biology, 37, 515–521.

Eisenbeis, S., Profitt, W., Coles, M., Truffault, V., Shamugaratnam, S., Meiler, J., & Hocker, B. (2012). Potential of fragment recombination for rational design of proteins. Journal of the American Chemical Society, 134, 4019–4022.

Emmert-Streib, F., & Mushegian, A. (2007). A topological algorithm for identification of structural domains of proteins. BMC Bioinformatics, 8, 237–247.

Engelhardt, B. E., Jordan, M. I., Srouji, J. R., et al. (2011). Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Research, 1969–1980.

Furnham, N., Holliday, G. L., de Beer, T. A. P., Jacobsen, J. O. B., Pearson, W. R., & Thornton, M. J. (2014). The Cata-lytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Research, 42, D485–D489.

Gianni, S., Brunori, M., Jemth, P., Mikael, O., & Mingjie, Z. (2009). Distinguishing between smooth and rough free energy barriers in protein folding. Biochemistry, 48, 11825–11830.

Gordeev, A. B., & Efimov, A. V. (2013). Modeling of folds and folding pathways for some protein families of (α+β)- and (α/β)-classes. Journal of Biomolecular Structure and Dynamics, 31, 4–16.
Takada, S. (2012). Coarse-grained molecular simulations of large biomolecules. *Current Opinion in Structural Biology*, 22, 130–137.

Tasdighian, S., Paola, L. D., Ruvo, M. D., Paci, P., Santoni, D., Palquale, P., ... Guillani, A. (2013). Modules identification in protein structures: The topological and geometrical solutions. *Journal of Chemical Information and Modeling*, 54, 159–168.

Taylor, W. R. (2002). A “periodic table” for protein structures. *Nature*, 416, 657–660.

Taylor, W. R., Chelliah, V., Hollup, S. M., MacDonald, T. J., & Jonassen, I. (2009). Probing the “dark matter” of protein fold space. *Structure*, 17, 1244–1252.

Thiruv, B., Quon, G., Saldanha, S. A., & Steipe, B. (2005). Nh3D: A reference dataset of non-homologous protein structures. *BMC Structural Biology*, 5, 12.

Todd, A. E., Orengo, C. A., & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, 307, 1113–1143.

Tomatis, P. E., Fabiane, S. M., Simona, F., Carloni, P., Sutton, B. J., & Vila, A. J. (2008). Adaptive protein evolution grants organismal fitness by improving catalysis and flexibility. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 20605–20610.

Wagner, G. P., Pavlicev, M., & Cheverud, J. M. (2007). The road to modularity. *Nature Reviews Genetics*, 8, 921–931.

Wang, G., & Dunbrack, R. L. 2005. PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Research*, 33, W94–W98.

Wang, J., Oliveira, R. J., Chu, X., Whitford, C. P., Chahine, J., Han, W., ... Leite, V. (2012). Topography of funneled landscapes determines the thermodynamics and kinetics of protein folding. *Proceedings of the National academy of Sciences of the United States of America*, 109, 15763–15768.

Wathen, B., & Jia, Z. (2013). A hierarchical order within protein structures underlies large separations between strands in β-sheets. *Proteins: Structure, Function, and Bioinformatics*, 81, 163–175.

Yuan, C., Chen, H., & Kihara, D. (2012). Effective inter-residue contact definitions for accurate protein fold recognition. *BMC Bioinformatics*, 13, 292–305.

Zuo, G., Wang, J., & Wang, W. (2006). Folding with downhill behavior and low cooperativity of proteins. *Proteins*, 63, 165–173.