Bioinformatics, 32(11), 2016, 1709–1715
doi: 10.1093/bioinformatics/btw032
Advance Access Publication Date: 30 January 2016
Original Paper

OXFORD

Structural bioinformatics

Library of binding protein scaffolds (LibBP): a computational platform for selection of binding protein scaffolds

Seungpyo Hong and Dongsup Kim*

Department of Bio and Brain Engineering, KAIST, Daejeon 305-338, South Korea

*To whom correspondence should be addressed.

Associate Editor: Burkhard Rost

Received on 31 August 2015; revised on 2 November 2015; accepted on 18 January 2016

Abstract

Motivation: Developments in biotechnology have enabled the *in vitro* evolution of binding proteins. The emerging limitations of antibodies in binding protein engineering have led to suggestions for other proteins as alternative binding protein scaffolds. Most of these proteins were selected based on human intuition rather than systematic analysis of the available data. To improve this strategy, we developed a computational framework for finding desirable binding protein scaffolds by utilizing protein structure and sequence information.

Results: For each protein, its structure and the sequences of evolutionarily-related proteins were analyzed, and spatially contiguous regions composed of highly variable residues were identified. A large number of proteins have these regions, but leucine rich repeats (LRRs), histidine kinase domains and immunoglobulin domains are predominant among them. The candidates suggested as new binding protein scaffolds include histidine kinase, LRR, titin and pentapeptide repeat protein.

Availability and implementation: The database and web-service are accessible via http://bcbl.kaist.ac.kr/LibBP.

Contact: kds@kaist.ac.kr

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

1 Introduction

The development of recombinant DNA technologies and binding protein screening methods has enabled researchers to imitate *in vivo* adaptive immune processes in the laboratory and thereby engineer-specific protein binders. This *in vitro* protein evolution is composed of two steps: (1) generation of a large span of diversified proteins and (2) selection of proteins with high affinity to the target molecules. Various combinations of protein diversification and selection technologies, such as phage display (Scott and Smith, 1990), yeast surface display (Boder and Wittrup, 1997) and ribosome display (Hanes and Plückthun, 1997), are currently applied to perform *in vitro* protein evolution.

Except for the development of peptide epitopes, these technologies require scaffold proteins where the epitope recognition sites are randomized. Antibodies are the key component of the adaptive immune system and have highly variable complementarity-determining regions; consequently, they have been widely engineered as the scaffolds for binding proteins (Hoogenboom, 1998, 2005; Jost and Plückthun, 2014; Ossipow and Fischer, 2014). However, the limitations of antibodies as scaffold proteins, such as their high molecular weight, multi-chain nature and overly large variable regions, have emerged, and alternative proteins have therefore been sought (Skerra, 2007). To date, more than 40 alternative binding proteins, composed of a structurally stable region and a variable region where diverse amino acids can be placed, have been identified (Binz *et al.*, 2005). Skerra (2007) specifically described three desirable properties for scaffolds in a binding protein library: they must have (1) a compact and rigid core, (2) a contiguous surface region where the amino acids can be replaced and (3) exposure of hydrophobic residues in the variable region. For many scaffold proteins, their loop regions are selected

as the variable regions (Binz *et al.*, 2005; Nuttall and Walsh, 2008), and structural components are rarely considered during scaffold protein selection. The exceptions are repeat proteins, in which repetitive patterns constitute-specific horseshoe shapes and their concave regions are used as the variable regions (Lee *et al.*, 2012; Main *et al.*, 2003; Stumpp *et al.*, 2003a, b). However, most binding protein scaffolds were selected for their native functions, and little attention has been paid to their diversity and structure. Along with exponential increase in protein sequence information (Acland *et al.*, 2014), recent advances in structural genomics have expanded our understanding of proteins and their structures (Khafizov *et al.*, 2014). This information can be utilized to systemically identify new candidates for binding protein scaffolds.

In this study, we computationally analyzed the structures and sequences of proteins to discover scaffolds with desirable properties for binding protein libraries. First, the sequences of evolutionarily-related proteins were collected and their amino acid diversity was evaluated to establish the variable residues. The spatial relationship among the variable residues was evaluated using the protein structures. A scoring function, combining the sequence and structure information, was employed to determine highly variable and contiguous regions. Consequently, we were able to identify new candidates for binding protein scaffolds along with previously developed binding proteins. In addition, we provide a web-based database for the binding protein scaffolds found in this analysis, as well as a web-based analysis tool that can be used to detect the variable region of any protein structure.

2 Methods

To be successfully engineered into a scaffold for binding proteins, a protein should possess a surface-exposed region that can be diversified without sacrificing structural integrity. In this study, a computational method was employed to find and evaluate highly diversifiable and contiguous surface regions. The overall scheme of the method is illustrated in Figure 1a, with specific details described in the following sections. In brief, residue variability was extracted from evolutionarily-related proteins (sequence variation analysis) and protein structure was analyzed to evaluate the external accessibility and internal connectivity among residues (structure analysis). This information was then combined, and the contiguous regions composed of variable residues, which are referred to as 'variable patches' here, were searched (i.e. variable patch analysis). Finally, the variable patches were combined into a larger structural unit, termed an 'extended patch', to represent the diversifiable region that can be utilized for further affinity maturation.

2.1 Structure set

Since we aimed to find the alternatives to immunoglobulin G whose dimeric structure is one of the weaknesses as a scaffold protein, only the monomeric protein structures were considered in this study. In this process, monomer protein structures that were elucidated by X-ray crystallography and had a resolution <2.0 Å were analyzed. Redundant structures were removed during collection from the Protein Data Bank (PDB) (Berman *et al.*, 2002) by setting a 40% sequence identity cutoff. As a result, 4818 biological monomer structures were collected and analyzed. Because protein size is an important factor for binding proteins, proteins were classified as either 'small', with <200 residues, or 'large', with ≥200 residues.

2.2 Sequence variation analysis

To obtain the sequence information of a PDB structure, evolutionarily-related sequences were retrieved from the NCBI non-redundant database (Sayers *et al.*, 2012), compiled on December 2014, by using BLASTP (Camacho *et al.*, 2009) with an *e*-value cutoff of 0.001. From the retrieved sequences, the amino acids that were mapped to the residues in the structure were collected. For each residue, sequence variation was evaluated by calculating the Shannon entropy (Shannon, 1951) of the amino acid composition using the following equation:

$$S = \sum_{i=1}^{20} p_i \ln p_i, \quad (1)$$

where p_i is the frequency of an amino acid i. The residues with entropy >2.0 were defined as variable residues; otherwise, they were defined as invariable residues. In addition, residues with structural importance were excluded from variable residues. Some residues were mainly composed of non-polar amino acids (FILMWY), and yet exposed to the solvent (Supplementary Fig. S1). These residues may be critical for stabilizing the hydrophobic cores of the proteins. Therefore, residues with the non-polar amino acid fraction >50% were designated as hydrophobic residues and they were excluded from the variable residues. Glycine and proline prefer a distinctive backbone conformation (Ho and Brasseur, 2005); therefore, residues comprising >50% glycine or proline were also regarded as invariable.

2.3 Structure analysis

Surface-exposed residues were identified by evaluating the solvent accessible surface area (SASA) relative to the fully exposed SASA of each amino acid (Chothia, 1976). SASA was calculated by using the AREAIMOL program in CCP4 suite (The CCP4 suite: Programs for protein crystallography, 1994) with probe radius of 1.4 Å. Residues with relative SASA >2.5% were defined as surface residues. For each residue, its accessible region in three-dimensional space was further evaluated considering the orientation of its side chain. Specifically, the protein was placed in space divided with three-dimensional grid points at 1-Å intervals. The grid points within 3 Å from atoms of invariable residues or main chain atoms were classified as inaccessible. Among the remaining grid points, those around a residue was classified as accessible if that grid point was within 7 Å of the α carbon of the residue and the angle between the grid, β and α carbons was >90° (Fig. 1b). The accessible grids were used for further evaluation of the accessible region of the residue. If the overlap between the accessible regions of two residues was larger than 10 $Å^3$, the two were considered to form a contiguous surface (Fig. 1c). By combining the amino acid variability from the sequence variation analysis and the spatial accessibility and residue contiguity information from the structure analysis, a variable residue network was constructed. This network was composed of nodes for variable residues, which contain the amino acid variation information, and edges for residue–residue connection information, which reflect the spatial overlap of accessible regions (Fig. 1d).

2.4 Variable patch analysis

For the purpose of library construction, the diversity of the available amino acids is more important than the mere variability score of the amino acid composition defined by Equation 1. For example, a residue composed of three amino acids with the compositions 0.8, 0.1 and 0.1 has smaller entropy than a residue with 0.5, 0.5 and 0 compositions for the same amino acids. However, because the three amino acids can be placed in the first residue, it should be considered more diversified. Therefore, the amino acid frequency was converted into a Boolean amino acid presence vector, and the number of available amino acids

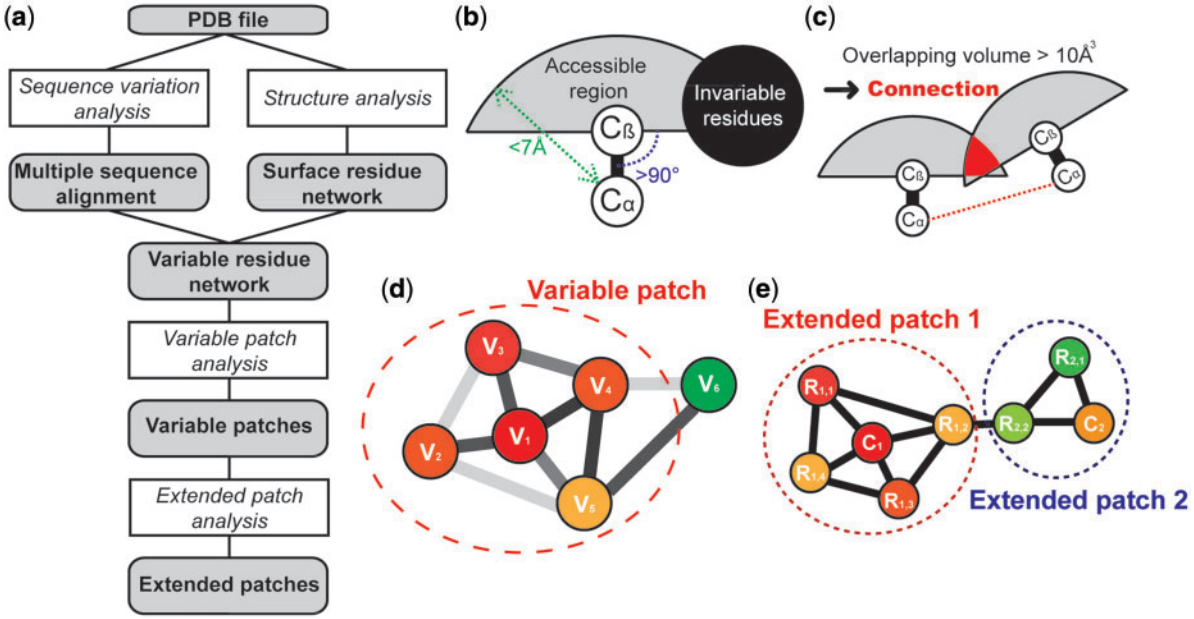

Fig. 1. Methods: the sequence variation and structural features of the proteins were analyzed and the variable patch regions were found. **(a)** The entire analysis is a collection of four analyses (each detailed in the 'Methods' section): (1) sequence variation analysis, (2) surface analysis, (3) surface patch analysis and (4) extended patch analysis. **(b)** The accessible region was defined by the distance and angle conditions, which were defined by the coordinates of α and β carbons. During evaluation, the regions occupied by the atoms of the invariable residues were excluded from the accessible region. **(c)** When the overlapping accessible region of two residues was larger than 10 Å^3, connections were established, with the overlapping volume representing the strength of the connection. **(d)** Consequently, a network comprised of variable residues and their connections was constructed. Then, the sub-networks with high variability (red nodes) and strong connection (darker edges) were sought with the patch finding algorithm (see 'Methods' section). **(e)** A network of variable patches was constructed and a community search algorithm was employed to determine the extended patches. Within each extended patch, the patch with the highest patch score was designated as the core (C) and the others as the rims (R). In this illustration, the patches with high patch score were colored in red

was used as a measure of the residue diversity. The amino acids observed significantly less often than expected were presumed to be absent at the residue, and the following equation was applied for the conversion:

$$B_{r,i} = \begin{cases} \text{False}, & \text{if} \quad \ln\left(f_{r,i}/f_i\right) < -2.0 \\ \text{True}, & \text{elsewhere} \end{cases}, \tag{2}$$

where $B_{r,i}$ is the Boolean presence variable, $f_{r,i}$ is the observed frequency of amino acid i at residue r and f_i is the frequency of amino acid i in the surface. The number of available amino acids for residue r (N_r) was calculated as the summation of the presence variables.

2.5 Variable patch score

A variable patch was defined as a set of interconnected variable residues. A large combination of the variable patches can be extracted from the variable residue network, but only those with desirable properties should be selected. The desirable properties were modeled with a variable patch score for which two factors were considered: (1) the chemical diversity and (2) the spatial accessibility and contiguity. As the interaction environment is provided by more than one residue in protein-mediated interaction (Reichmann *et al.*, 2005), the environment formed by two adjacent residues was evaluated. Thus, the chemical diversity was modeled as the product of the number of available amino acids in two adjacent residues. The spatial accessibility and contiguity was evaluated as the overlapping volume of the accessible regions of the two residues. Finally, the product of the two values was used as the variable patch score,

$$S_v = \sum_{i=1}^{n} \sum_{j=i+1}^{n} V_{i,j} N_i N_j, \tag{3}$$

where S_v represents the variable patch score for a variable patch with n residues, and $V_{i,j}$ is the overlapping volume of the accessible region between two residues. A large overlapping volume can limit spatial accessibility and therefore the effect of the overlapping volume would not be linear. In order to consider this effect, step functions and saturating linear functions with different parameters were tested. These tests produced similar results (Supplementary Fig. S3): thus, the simple linear equation was used here.

2.6 Variable patch search algorithm

In the variable residue network, sets of variable residues with high variable patch scores were sought (Fig. 1d). First, all possible mutually connected nodes of three were identified and used as an initial patch. Next, an adjacent node that produced the largest increase in variable patch score was determined and incorporated into the original patch to form a new patch. This patch extension process was repeated until further expansion was not possible or the number of residues in the patch reached seven. Consequently, multiple patches with corresponding variable patch scores were generated for each protein structure. In this study, the expansion of the patch was stopped at seven residues, because full randomization of these seven residues would result in $\sim 10^9$ variations, which is the size limit of library in a typical experimental condition. Note, however, that the number of residues to randomize can be increased by using the extended patches described below.

2.7 Extended patch analysis

High score patches were merged into larger structural units called extended patches (Fig. 1e). Patches with scores >700 000, which is the patch score cutoff for top 25 variable patch scores in small proteins, were collected and used as the nodes of a new graph representing the entire variable surface of the protein. Edges were established if the two patches shared any residues. Highly connected structures were identified using a community search algorithm, the fast greedy modularity optimization algorithm (Clauset *et al.*, 2004), and the resulting communities were regarded as extended patches. Within the extended patches, the residues in the highest scoring patch were defined as the interface core variable region and the remaining residues were defined as the interface rim variable region. The interface core region can be used as the initial library generation site, and the interface rim region can be used as a further library generation site for consecutive affinity maturation screens.

3 Results and discussion

3.1 Amino acid diversity in the variable residues

The variability of residues was highly dependent on their location in protein structure; surface exposed residues were highly variable, whereas buried residues were less variable (Fig. 2a). Variable residues, i.e. those with amino acid composition entropy >2.0, comprised ~30% of the surface residues. Because the diversity of the amino acids at a site is more important than the frequency, the amino acid frequency was converted into an amino acid presence vector. Compared with the amino acid composition on the surface region, the composition of the amino acid presence vector was relatively uniform, and all amino acids were well observed in the variable residues (Fig. 2b). In addition, the number of amino acid types observed on each variable residue tended to be large (Fig. 2c). These results suggest that our variable residue criteria have successfully detected the residues that can accommodate a variety of amino acids.

3.2 Accessibility of the variable residues

In addition to having high variability, residues should be spatially accessible to the binding partners. The spatial accessibility of the residues was determined from the structure of the protein and the accessible region around the residues. Interestingly, the variable residues on the surface were more accessible than the other invariable surface residues (Fig. 2d). This may be because the highly accessible residue is under less structural constraint and can accommodate different types of amino acids. Therefore, the variable residues can be utilized as the diversification sites for a binding protein library.

3.3 Connectivity to other variable residues

The physicochemical environment generated by a single residue is limited, but when several residues are located close to each other, they can produce a complex interaction environment. Hence, variable residues connected with a larger number of other variable residues are considered more desirable for construction of a binding protein library. Among the variable residues, ~24% were isolated, and the number of residues decreased as connectivity increased (Fig. 2e). About 60% of variable residues were connected with ≥2 variable residues and these could provide diverse spatially contiguous surfaces. Among all of these possible contiguous surfaces, those with high variability and accessibility were selected using the patch score.

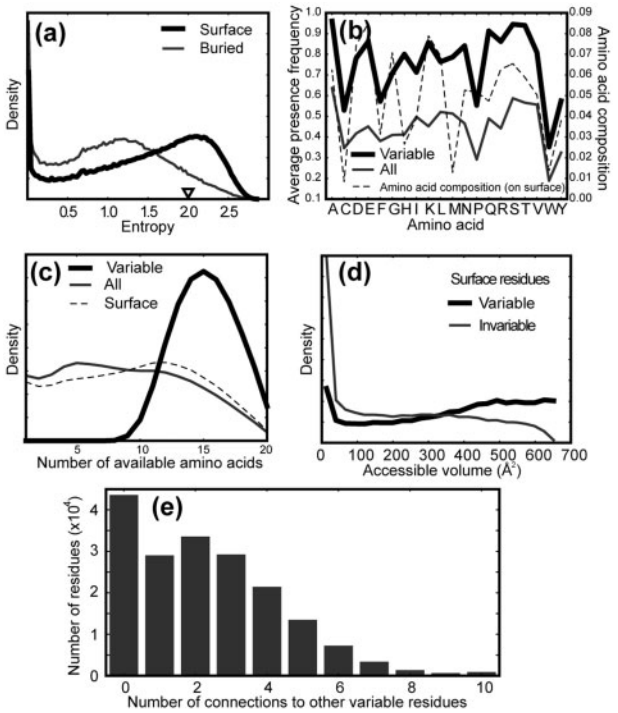

Fig. 2. Properties of variable residues. The properties of variable residues were compared with those of the other residues. **(a)** Residues on protein surface with high sequence diversity were designated as the variable residues. The surface residues were generally more variable than buried residues. Among the surface residues, those with amino acid composition entropy larger than 2.0, which is marked by a triangle, were defined as the variable residues. **(b)** Variable residues encompassed a greater variety of amino acids. In addition, the distribution of the amino acid presence vector showed increased uniformity compared to the amino acid composition. All amino acids except tryptophan had been found in more than half of the variable residues, and the amino acids E, K, N, Q, R, S, T and V were found >80% of the variable residues. **(c)** A large number of different amino acids were observed at the variable residues, compared with the residues in the protein surface. **(d)** Among surface residues, the variable residues were more accessible from outside of the protein than the invariable residues. **(e)** The variable residues were generally well connected to other variable residues. The plots in **(a)**, **(c)** and **(d)** are normalized histograms for each group

3.4 Variable patches

The residue variation, spatial accessibility and connectivity to other residues were collectively evaluated by using the 'variable patch score,' and high scoring variable patches were established for each protein. Consequently, protein scaffolds with exceptionally variable surface regions were identified (Fig. 3a). The top 25 small and large proteins with the highest patch scores are listed in Supplementary Tables S1 and S2, respectively (complete information is accessible via our web server).

3.5 Extended variable patches

The surface of a protein can accommodate more than one highly variable patch, and the size of the variable patch can be larger than seven residues. Such structures were defined here as 'extended variable patches.' In proteins with highly variable patches, the extended patch regions were successfully found. In addition, the highly variable 'core' patch was topologically located on the center of the extended variable patch surrounded by 'rim' patches. The extended patches for five protein structures are illustrated in Figure 3b–f.

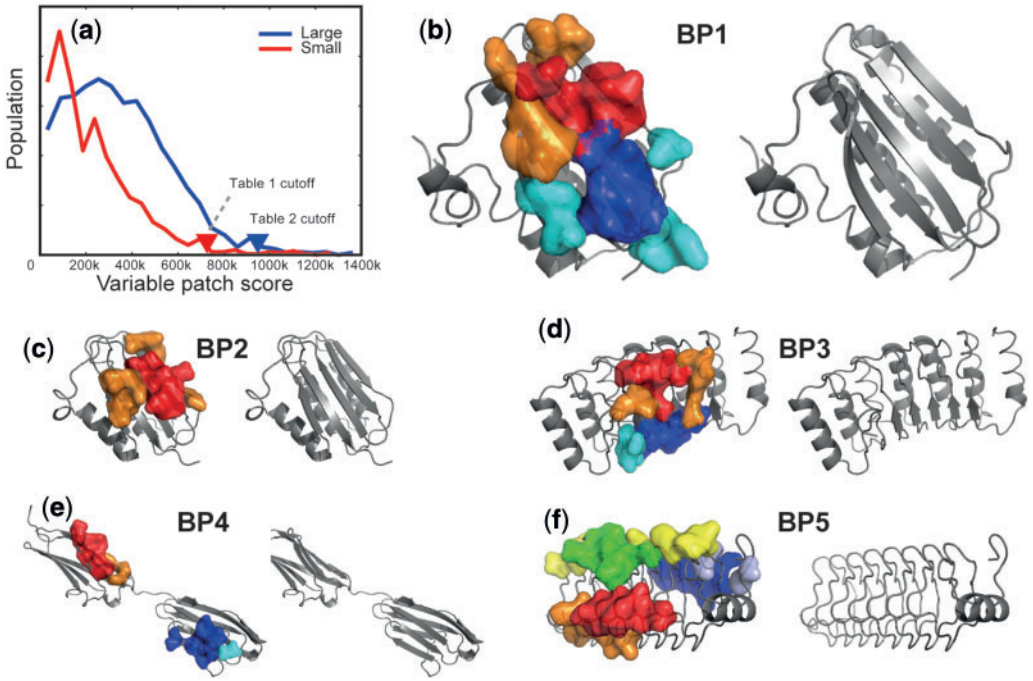

Fig. 3. Variable patches. **(a)** Variable patches were analyzed among thousands of monomer protein structures, and proteins with highly variable surfaces were identified. Among proteins with high patch scores, the extended patches were also established. Consequently, five proteins were selected as scaffolds for binding proteins: **(b)** histidine kinase (PDB 3A0Y); **(c)** histidine kinase (PDB 4GT8); **(d)** LRR domain in NALP1 (PDB 4IM6); **(e)** immunoglobulin domains in titin (PDB 2A38) and **(f)** pentapeptide repeats (PDB 2J8K). The core variable regions of extended patch 1, 2 and 3 are displayed as red, blue, and green surfaces and rim variable regions of extended patch 1, 2, and 3 are displayed as orange, pale blue and yellow surfaces. The protein structures are shown in gray. PyMol (Schrödinger, 2010) was used to produce these illustrations

3.6 Proteins with high patch scores

High scoring patches were predominantly found in three protein structures; the leucine rich repeat (LRR) fold, histidine kinase domain and immunoglobulin domain (Supplementary Tables S1 and S2). The LRR is a structure composed of repeats of short structures and is known to mediate protein–protein interactions (PPIs) (Kobe and Kajava, 2001). Because of their biological role in PPIs and unique modular structure, LRRs have been engineered as protein binder molecules (Lee *et al.*, 2012; Stumpp *et al.*, 2003a). In our analysis, LRR-based protein binder molecules had high patch scores (Supplementary Tables S1 and S2). The variable regions were located on the concave region of the protein (Fig. 3d); these regions were also used in a previous binder design (Lee *et al.*, 2012; Stumpp *et al.*, 2003a). Histidine kinase is a kinase domain of signal-transduction enzymes, which are widely used in bacteria and eukaryotes from outside the animal kingdom (Wolanin *et al.*, 2002). The proteins are composed of three α-helixes and five β-strands, and the variable sites are located on the β-sheet region (Fig. 3b and c). The immunoglobulin fold is a protein domain composed of a 2-layer sandwich of β-sheets, which is found in proteins with a diverse range of biological functions (Bork *et al.*, 1994). Proteins with immunoglobulin domains were widely used in the binding protein library and many immunoglobulin domains were retrieved with high patch scores. Unlike the previous research, where the loop region was randomized, the variable contiguous patch region found in this study is located on the β-sheet region of the protein (Fig. 3e).

3.7 Selection of binding protein scaffolds

The variable patches were analyzed for 4818 monomer proteins, among which five proteins were selected as examples for new binding protein scaffolds to illustrate the utility of the analysis (Table 1). In addition to the variable patch scores, other features, such as size, source, expression host and solubility, were considered. The first two structures were histidine kinase domains. The first binding protein scaffold (BP1) was derived from a histidine kinase domain originating from a hyperthermophilic organism. This protein was selected because it is functional at high temperatures and its thermal stability is expected to be high. To construct BP2, a histidine kinase predicted to be soluble in the ESPRESSO server (Hirose and Noguchi, 2013) was used.

LRR domains tended to have high patch scores. The LRR domain in NALP1 (NACHT, LRR and PYD domains-containing protein 1) was used to construct the third library, BP3. NACHT–LRR protein is a human-originated protein involved in the recognition of bacteria, and the LRR domain contributes toward recognition of the interaction partner (Kufer *et al.*, 2005). This property indicates that the domain may be a good binder.

Immunoglobulin domains were predicted as desirable scaffolds for binding proteins, with the β-sheet region of the immunoglobulin selected as a good library generation site. BP4 was constructed from domains in titin, a long elastic filament in muscle composed of a number of immunoglobulin and fibronectin domains (Labeit and Kolmerer, 1995). The fibronectin type III domain (FN3) has been used as a binding scaffold (Koide *et al.*, 2012) where the β-sheet region has been randomized as the binding interface. The structure of titin found in this study was slightly different from that of FN3 and the β-sheet on the opposite side was selected in this study. Titin originates from humans and may be less prone to cause immunogenicity problems when used as biopharmaceuticals. In addition, the domain used in library construction lacked disulfide bonds, which is a desirable property for expression in *Escherichia coli*. In addition, two immunoglobulin domains are linked with a short peptide. This

Table 1. Selected binding protein scaffolds

	Number of residue[a]	SASA(Å^2)[a]	Protein name (PDB ID)	Expression host	Solubility[b]
BP1	13	910	Histidine kinase (3A0Y)	*Thermotoga maritima*	Soluble (0.654)
BP2	15	858	Sensor protein VraS (4GT8)	*Staphylococcus aureus*	Soluble (0.871)
BP3	12	527	NALP1 (4IM6)	*Homo sapiens*	Soluble (0.693)
BP4	10	643	Titin (2A38)	*Homo sapiens*	Soluble (0.579)
BP5	13	633	NP275-NP276 (2J8K)	*Nostoc punctiforme*	Soluble (0.805)

[a]The number of residues and SASA had been calculated for the first extended patch.
[b]The solubility had been calculated by using ESPRESSO (Hirose and Noguchi, 2013) and the confidence was displayed in parentheses.

tandem repeat architecture could allow the design of proteins for multiple targets or proteins with multiple binding sites for a single target.

BP5 was derived from the pentapeptide-repeat structures. Pentapeptide-repeat proteins are composed of repeated amino acid sequences, and they are widely found in prokaryotes and eukaryotes (Vetting *et al.*, 2006). This protein possessed three highly variable patches that faced different regions of the protein (Fig. 3f). This unique structure could be exploited to develop a binding protein with three distinctive interaction interfaces (Table 1).

The sequences of the five proteins, including variable residue positions, are listed in Supplementary Table S3. Except for BP1, the variable residues that belonged to the highest scoring patch were suggested as the diversification sites. In BP1, the residues in the second highest scoring patch were selected as diversification sites because they were closely located in the sequence, which allows easier construction of the DNA library. In BP1, BP2 and BP4, the span for the randomization sites was short with 12, 16 and 12 amino acids, respectively. Thus, for construction of the DNA library, the randomization portion can be synthesized as a whole and inserted into the scaffold in a single step. Due to the repeat architecture, the randomization sites in BP3 and BP5 were separated in sequence space and their spans were 32 and 42 amino acids, respectively. Therefore, a more elaborate library generation scheme might be required.

3.8 Structural variation of core regions

A good scaffold protein possesses a stable core region. Here, the structural stability of the selected proteins was indirectly evaluated based on the structural similarity of evolutionarily-related proteins. For each selected protein, the structures of proteins with high-sequence similarity were retrieved from the PDB using an *e*-value cutoff of 0.001 (Berman *et al.*, 2002). These structures were structurally aligned by using TM-align (Zhang and Skolnick, 2005) and visualized with PyMol (Schrödinger, 2010). The structures of proteins related to the five proteins of interest were highly similar (Fig. 4), which suggests that the selected proteins possess stable core regions and, therefore, could be utilized as scaffolds for binding protein. This was also supported by their sequence profiles (Supplementary Fig. S4). The buried residues constituting the hydrophobic core were well preserved throughout evolutionarily-related sequences and were composed of hydrophobic amino acids. The sequence profile illustrated in Supplementary Figure S4 was drawn by using WebLogo (Crooks *et al.*,2004).

3.9 Comparison with previously suggested binding protein scaffolds

Variable patch analysis was also applied to 34 proteins that are already used as binding protein scaffolds. Variable patches with moderately high scores were found in transferrin, TPR protein and lipocalins (Supplementary Table S4). However, in general, variable patch scores were low and initial variable patches, i.e. three

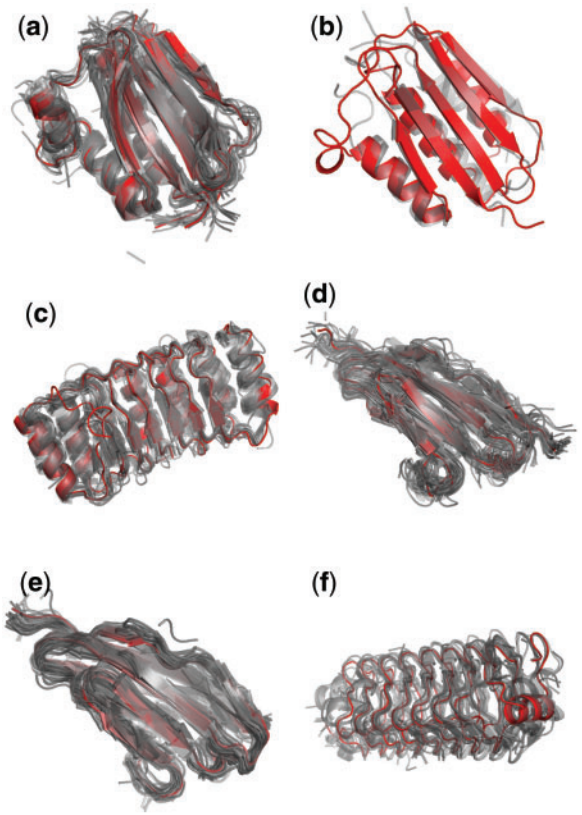

Fig. 4. Structural variation of five selected proteins. To estimate the structural stability of the selected proteins, the structural variation of related proteins was evaluated. The structures of the five selected proteins are illustrated in red, whereas the related proteins are illustrated in gray. **(a)** BP1. **(b)** BP2. **(c)** BP3. **(d, e)** Two immunoglobulin domains in BP4 are separately illustrated. **(f)** BP5. The structures of the related proteins were well conserved; therefore, the core region of the proteins is expected to be stable

interconnected variable residues, were not found in 20 cases. These results suggest that the previously-developed binding protein scaffolds examined here were generally derived from well-conserved proteins and that the randomization sites were selected relying on intuition (mainly under the assumption that the loop would be varied). Hence, the new protein scaffolds identified computationally in this research cover different proteins and could therefore expand the existing set of binding protein scaffolds.

3.10 Web server

The analysis results for all 4818 monomer structures can be accessed in the DATABASE section of our web server (http://bcbl.kaist.ac.kr/LibBP). The collective information for the proteins is displayed as a

table. The gene name, patch scores, source host, expression host, size of the protein, predicted solubility by ESPRESSO (Hirose and Noguchi, 2013) and figures of the structure and patches can be accessed. Detailed information on each protein can also be accessed via a hypertext link on the PDB identifiers or figures. This information includes the patch scores and a list of residues in the patches for the 20 highest scored patches, lists of residues for the extended patches, structures of the proteins and patches and the sequences marked with extended patches. Users can submit protein structures for variable patch analysis in the ANALYSIS section. When the analysis is complete, an e-mail will be sent to the user reporting the detail information described above.

4 Conclusion

In this study, we developed a computational method that utilized residue variability and connectivity to evaluate proteins as scaffolds for binding proteins. LRR folds, histidine kinase domains and immunoglobulin domains were found to possess desirable variable patches. The LRR fold and immunoglobulin domain have previously been reported as good scaffolds for binding proteins. The high rank of these structures in our study suggests that our method identified structures that could be used as binding protein scaffolds. Considering our results and several other relevant features of the proteins, we selected five new binding protein scaffolds. These five proteins can be used to construct binding protein libraries, and we expect that the systematic evaluation of protein structure, as conducted here, will facilitate the development of additional new binding protein libraries.

Funding

This study was supported by grants of the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea [HI12C0014], the Stem Cell Research Program [2012M 3A9B 4027957] and the KAIST Future Systems Healthcare Project, Ministry of Science, ICT and Future Planning, Republic of Korea.

Conflict of Interest: none declared.

References

Acland,A. *et al*. (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **42**, 7–17.

Berman,H.M. *et al*. (2002) The Protein Data Bank. *Acta Crystallogr. Sect. D Biol. Crystallogr*., **58**, 899–907.

Binz,H.K. *et al*. (2005) Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol*., **23**, 1257–1268.

Boder,E.T. and Wittrup,K.D. (1997) Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol*., **15**, 553–557.

Bork,P. *et al*. (1994) The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol*., **242**, 309–320.

Camacho,C. *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chothia,C. (1976) The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol*., **105**, 1–12.

Clauset,A. *et al*. (2004) Finding community structure in very large networks. *Phys. Rev. E*, **70**, 1–6.

Crooks,G.E. *et al*. (2004) WebLogo: a sequence logo generator. *Genome Res*., **14**, 1188–1190.

Hanes,J. and Plückthun, A. (1997) in vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl Acad. Sci. USA*, **94**, 4937–4942.

Hirose,S. and Noguchi,T. (2013) ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. *Proteomics*, **13**, 1444–1456.

Ho,B.K. and Brasseur,R. (2005) The Ramachandran plots of glycine and preproline. *BMC Struct. Biol*., **5**, 14.

Hoogenboom,H. (1998) Antibody phage display technology and its applications. *Immunotechnology*, **4**, 1–20.

Hoogenboom,H.R. (2005) Selecting and screening recombinant antibody libraries. *Nat. Biotechnol*., **23**, 1105–1116.

Jost,C. and Plückthun,A. (2014) Engineered proteins with desired specificity: DARPins, other alternative scaffolds and bispecific IgGs. *Curr. Opin. Struct. Biol*., **27**, 102–112.

Khafizov,K. *et al*. (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proc. Natl Acad. Sci. USA*, **111**, 3733–3738.

Kobe,B. and Kajava,A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol*., **11**, 725–732.

Koide,A. *et al*. (2012) Teaching an old scaffold new tricks: monobodies constructed using alternative surfaces of the FN3 scaffold. *J. Mol. Biol*., **415**, 393–405.

Kufer,T. *et al*. (2005) NACHT-LRR proteins (NLRs) in bacterial infection and immunity. *Trends Microbiol*., **13**, 381–388.

Labeit,S. and Kolmerer,B. (1995) Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science*, **270**, 293–296.

Lee,S.C. *et al*. (2012) Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. *Proc. Natl Acad. Sci. USA*, **109**, 3299–3304.

Main,E.R.G. *et al*. (2003) Design of stable alpha-helical arrays from an idealized TPR motif. *Structure*, **11**, 497–508.

Nuttall,S.D. and Walsh,R.B. (2008) Display scaffolds: protein engineering for novel therapeutics. *Curr. Opin. Pharmacol*., **8**, 609–615.

Ossipow,V. and Fischer,N. (2014) *Monoclonal Antibodies*. Humana Press, Totowa, NJ.

Reichmann,D. *et al*. (2005) The modular architecture of protein-protein binding interfaces. *Proc. Natl Acad. Sci. USA*, **102**, 57–62.

Sayers,E.W. *et al*. (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **40**, D13–D25.

Schrödinger,L. (2010) The PyMOL molecular graphics system, Version 1.3. (https://www.pymol.org/citing)

Scott,J.K. and Smith,G.P. (1990) Searching for peptide ligands with an epitope library. *Science*, **249**, 386–390.

Shannon,C.E. (1951) Prediction and entropy of printed English. *Bell Syst. Tech. J*., **30**, 50–64.

Skerra,A. (2007) Alternative non-antibody scaffolds for molecular recognition. *Curr. Opin. Biotechnol*., **18**, 295–304.

Stumpp,M.T. *et al*. (2003a) Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *J. Mol. Biol*, **332**, 471–487.

Stumpp,M.T. *et al*. (2003b) Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol*., **332**, 471–487.

The CCP4 suite: Programs for protein crystallography (1994) *Acta Crystallogr. Sect. D Biol. Crystallogr*, **50**, 760–763.

Vetting,M.W. *et al*. (2006) Pentapeptide repeat proteins. *Biochemistry*, **45**, 1–10.

Wolanin,P.M. *et al*. (2002) Histidine protein kinases: key signal transducers outside the animal kingdom. *Genome Biol*., **3**, Reviews3013.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*., **33**, 2302–2309.