Protein Engineering, Design & Selection vol. 26 no. 4 pp. 283–289, 2013
Published online January 21, 2013 doi:10.1093/protein/gzs108

A structural bioinformatics approach for identifying proteins predisposed to bind linear epitopes on pre-selected target proteins

Eun Jung Choi[1], Ron Jacak[1] and Brian Kuhlman[1,2,3]

[1]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599, USA and [2]Lineberger Comprehensive Cancer Care Center, University of North Carolina, Chapel Hill, NC 27599, USA

[3]To whom correspondence should be addressed.
E-mail: bkuhlman@email.unc.edu

Received September 24, 2012; revised December 18, 2012; accepted December 19, 2012

Edited by Valerie Daggett

We have developed a protocol for identifying proteins that are predisposed to bind linear epitopes on target proteins of interest. The protocol searches through the protein database for proteins (scaffolds) that are bound to peptides with sequences similar to accessible, linear epitopes on the target protein. The sequence match is considered more significant if residues calculated to be important in the scaffold–peptide interaction are present in the target epitope. The crystal structure of the scaffold–peptide complex is then used as a template for creating a model of the scaffold bound to the target epitope. This model can then be used in conjunction with sequence optimization algorithms or directed evolution methods to search for scaffold mutations that further increase affinity for the target protein. To test the applicability of this approach we targeted three disease-causing proteins: a tuberculosis virulence factor (TVF), the apical membrane antigen (AMA) from malaria, and hemagglutinin from influenza. In each case the best scoring scaffold was tested, and binders with K_ds equal to 37 μM and 50 nM for TVF and AMA, respectively, were identified. A web server (http://rosettadesign.med.unc.edu/scaffold/) has been created for performing the scaffold search process with user-defined target sequences.

Keywords: epitope/protein engineering/protein–protein interaction/protein scaffold/structural bioinformatics

Introduction

Protein–protein interactions are critical to most biological processes, and a diversity of diseases that are caused by mutations at binding interfaces, such as Von Hippel–Lindau syndrome and hypercholesterolemia (Pawson and Nash, 2003; Steward *et al.*, 2003). Reengineered protein–protein interactions can be used to rewire signaling networks, and novel protein binders can activate or inhibit medically important pathways.

In the field of protein engineering, the concept of protein scaffolds has become prevalent, where different protein folds are engineered to bind to a variety of ligands. Commonly used scaffolds are immunoglobulin folds and repeat proteins (Sidhu and Koide, 2007; Koide, 2009; Lee *et al.*, 2012). In most cases, the scaffolds have been redesigned to bind target proteins using techniques in combinatorial biology such as phage display and yeast display. Recently, there have also been promising results using computational sequence optimization protocols to create novel interactions (Jha *et al.*, 2010; Fleishman *et al.*, 2011; Karanicolas *et al.*, 2011; Stranges *et al.*, 2011; Der *et al.*, 2012). Both strategies have limitations. With screening techniques it is not always straightforward to dictate the location and orientation of binding, which can have important functional consequences. With computational design it is possible to pre-define binding geometry; however, these methods are not yet robust and often fail to produce tight binders or incorrectly predict the binding orientation (Stranges and Kuhlman, 2013).

One critical step in computational protein interface design is identifying a protein scaffold that can be docked on to the target patch of the protein of interest in such a way that the newly created interface can be stabilized with amino acid mutations on the protein scaffold surface. One solution to this problem is to computationally dock many alternative proteins on to the target patch and then perform sequence optimization simulations at the interface to identify which proteins and docked configurations are most designable, i.e. give rise to interfaces that are predicted to have more favorable binding energies. For instance, in their design of a novel interaction with the hemagglutinin (HA) protein from influenza, Fleishman *et al.* computationally screened through 865 alternative protein scaffolds in search of designable interfaces (Fleishman *et al.*, 2011). This process is computationally expensive and relies primarily on an energy function to identify the most favorable scaffolds. Here we develop a method for identifying protein scaffolds for target interactions that relies less on energy-based docking, but rather makes use of interactions already observed in the protein database (PDB). In particular, we focus on targeting linear epitopes.

To summarize the protocol (Fig. 1A), first, regions of the target protein are identified that are likely to be solvent exposed and disordered. These will be the target epitopes. They can be identified using nuclear magnetic resonance (NMR) data or computational predictions of disorder, and are likely to be found in long loops, termini or connections between folded domains. Second, the amino acid sequences of the target epitopes are aligned and scored against the sequences of peptides that have been co-crystallized with proteins in the PDB. Alternative alignments are scored using a weighting scheme that emphasizes conservation of residues predicted to be important to the crystallized peptide–protein interaction. Favorable alignments indicate that the protein from the crystal structure has potential to be a good scaffold for engineering binders against the target epitope. In support

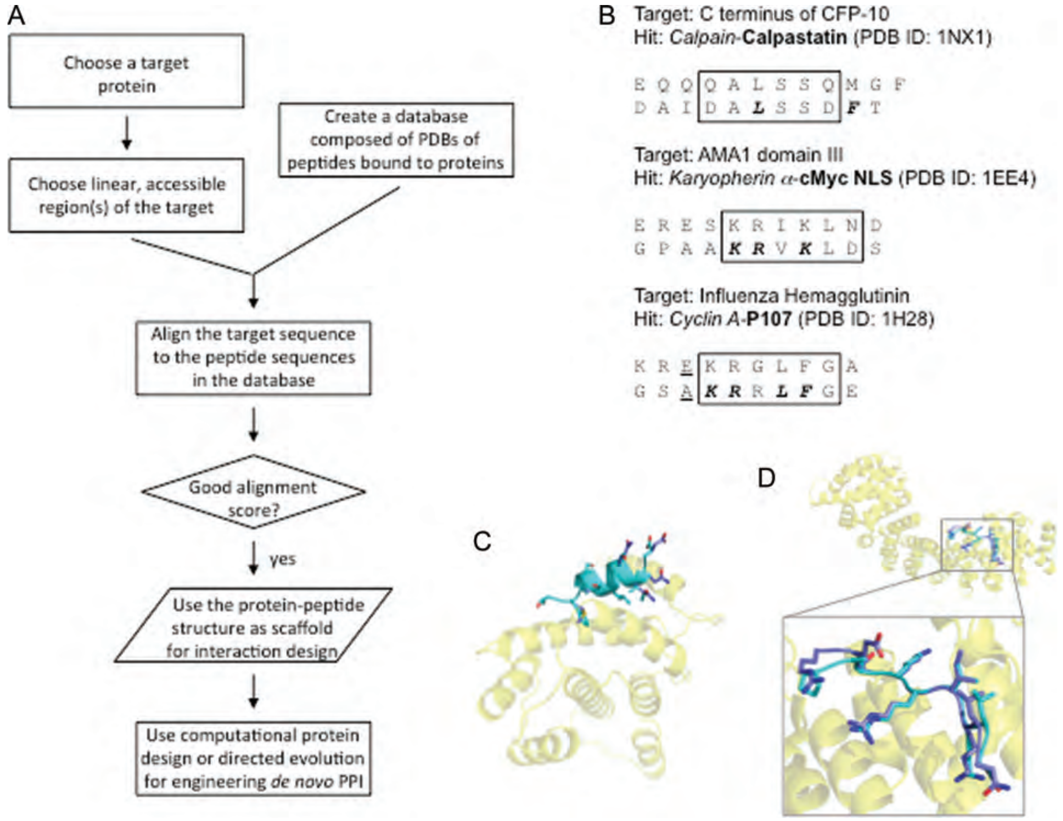
A
Choose a target protein
Choose linear, accessible region(s) of the target
Create a database composed of PDBs of peptides bound to proteins
Align the target sequence to the peptide sequences in the database
Good alignment score?
yes
Use the protein-peptide structure as scaffold for interaction design
Use computational protein design or directed evolution for engineering *de novo* PPI
B
Target: C terminus of CFP-10
Hit: *Calpain*-**Calpastatin** (PDB ID: 1NX1)
E Q Q Q A L S S Q M G F
D A I D A ***L*** S S D ***F*** T
Target: AMA1 domain III
Hit: *Karyopherin α*-**cMyc NLS** (PDB ID: 1EE4)
E R E S K R I K L N D
G P A A ***K R*** V ***K*** L D S
Target: Influenza Hemagglutinin
Hit: *Cyclin A*-**P107** (PDB ID: 1H28)
K R E K R G L F G A
G S A ***K R*** R ***L F*** G E
C
D

Fig. 1. (A) Diagram of the Scaffold Selection Protocol. (B) Result of scaffold selection protocol for target protein, TVF CFP-10 C-terminus (top), malaria AMA1 domain III (middle) and influenza HA (bottom). The top sequence is the target sequence of interest and the bottom sequence is the high scoring peptide sequence in the PDB file specified. The peptide binding protein used as the scaffold is italicized, the peptide is in bold font and the PDB ID is in parenthesis. The high scoring sequence alignment is boxed and the residues that are important for binding from the literature is in bold font italicized. Residue that is incompatible in the third example is underlined. (C) Structural representation of Calpain (yellow transparent)—Calpastatin peptide (cyan) and CFP-10 sequence threaded onto the backbone of Calpastatin peptide (blue). (D) Structural representation of Karyopherin α (yellow transparent)—c-*myc* NLS peptide (cyan) and AMA1 domain III sequence modeled using ROSETTA software (blue). Images made using PYMOL (Schrodinger, 2010).

of our strategy it has been shown that peptides derived from linear regions of proteins can be used as inhibitors of protein–protein interactions (Brunner *et al.*, 1973; Hashemzadeh *et al.*, 2008) and a recent study shows that many naturally occurring protein–protein interactions are dominated by one linear epitope from one of the binding partners (London *et al.*, 2010). Furthermore, antibodies often bind tightly to short linear epitopes. To identify which linear epitope an antibody recognizes, peptide libraries derived from the target protein are screened for binding to the antibody (Craig *et al.*, 1998). Our design strategy can be thought of as the reverse of this procedure. We specify the accessible linear region of the protein of interest that we want to target, then use sequence alignment to select a protein scaffold that has the potential to bind to our target epitope sequence. Our approach is similar in concept to protein threading methods that are used to evaluate whether a protein sequence is compatible with a pre-existing protein structure (Bowie *et al.*, 1991). However, our approach only uses the template structure to weight sequence conservation at the various residue positions in the alignment. Many protein-threading methods use additional structure-based scores including the environmental preference of amino acids and the preference of particular amino acid pairs to be near in space (David *et al.*, 2000; Xu *et al.*, 2000; Dunbrack, 2006).

Materials and methods

Scaffold Selection Protocol

A database composed of PDB files that contain structures of peptides bound to proteins was built by scanning the PDB database for PDB files with the following two criteria: files with multiple chains of proteins and files which has at least one chain with less than 30 amino acids. The final database used in this study was composed of 3137 peptide-bound protein structures.

To identify potential scaffolds from the database of peptide-bound protein structures, sequences from flexible regions within the target protein were aligned against the peptide sequences from the peptide-protein database. All possible alignments were scored using a sliding window with a user-specified sequence length, the default being 6. Alignments were scored based on a pairwise sequence comparison scoring scheme where the score for each residue pair was weighted based on the degree of burial for the residue at the interface in the protein–peptide structure. Exact matches for buried residues on the peptide were favored over those that were exposed. The scoring function was formulated as follows: the largest benefit was for buried residues with an exact match (4 points), the next largest benefit was for buried residues with amino acids of similar chemical type (hydrophobic, same charge, aromatic, etc.) or for exposed residues with an exact match (3 points) and the smallest benefit was for exposed residues with amino acids of similar chemical type (2 points). Penalties were assigned to residue pairs that did not match. The largest penalty was for mismatched buried residues (4 points), while a weaker penalty was assigned to mismatched exposed residues (3 points). The extent of residue burial was assessed by calculating the number of residues within 10 Å of the residue of interest. Residues from both the peptide and protein binding partner were included in this count. A residue was considered buried if it had over 18 neighbors.

Protein expression

The Calpain clone in pet-3d plasmid was provided by Dr Masatoshi Maki of Kyoto University. Protein purification was slightly modified from the original method (Takano *et al.*, 1995). Proteins were expressed in BL21 (DE3) pLysS cells. Cells were resuspended in lysis buffer (20 mM Tris-HCl pH 8.0, 50 mM NaCl, 5 mM β-mercaptoethanol) then lysed by sonication. The supernatant was filtered with a 0.2 μm filter and loaded onto a pre-packed anion exchange column (HiTrap Q HP, GE Healthcare). The protein was further purified by gel filtration chromatography (Superdex 200, GE Healthcare).

The Karyopherin α clone in pPROEX-Htb plasmid was provided by Dr Elena Conti of the Max Planck Institute for Biochemistry. Protein purification was slightly modified from the original method (Conti and Kuriyan, 2000). Cells were lysed by sonication and the supernatant was loaded onto a pre-packed Ni-NTA column (HisTrap HP, GE Healthcare). Peak fractions were dialyzed in the anion exchange equilibration buffer (20 mM Tris-HCl pH 7.6, 100 mM NaCl, 10% glycerol) overnight, filtered with a 0.2 μm filter and loaded onto a pre-packed anion exchange column (HiTrap Q HP, GE Healthcare) and eluted with a linear NaCl gradient from 100 to 600 mM. The protein was further purified by gel filtration chromatography (Superdex 200, GE Healthcare).

The six histidine-tagged AMA1 domain III clone in pQE-9 plasmid was provided by Dr Robin Anders of La Trobe University. Proteins were expressed in BL21 (DE3) pLysS cells and induction was carried out at 30°C for 5 h. Cells were resuspended in lysis buffer (20 mM Tris-HCl pH 7.0, 500 mM NaCl) and lysed by sonication. The cell lysate was mixed with Talon resin (ClonTech) and purified using the batch/gravity-flow purification protocol. The resin was washed twice with lysis buffer then once with wash buffer (lysis buffer with 8 mM imidazole), and eluted with elution buffer (lysis buffer with 500 mM imidazole). The eluate was dialyzed overnight in 20 mM Tris pH 8 to get rid of imidazole and salt. Precipitants, which accumulated during dialysis, were filtered away and the eluate was further purified by anion exchange column (HiTrap Q HP, GE Healthcare) using 0 to 500 mM NaCl gradient. Finally, peak fractions were ran on a gel filtration column (Superdex 75, GE Healthcare) for further purification.

Fluorescence polarization and isothermal titration calorimetric

Binding affinities between peptides and proteins were measured using fluorescence polarization (FP). All peptides were synthesized with fluorescein isothiocyanate (FITC) at the N-terminus and a β-alanine as the linker at the Tufts University peptide synthesis core facility. Lyophilized peptides were solubilized in their respective protein's elution buffers to 500 nM and purified proteins were titrated into it. Protein concentration was quantified by measuring the UV absorbance at 280 nm and using the theoretical molar extinction coefficient. Peptide concentration was quantified by measuring the UV absorbance at 494 nm and using molar extinction coefficient of 68 000 M^{-1} cm^{-1}. FP assays were carried out on a Jobin Yvon Horiba Spec FluoroLog-3 instrument (Jobin Yvon Inc.) performed in L-format with the excitation wavelength set at 494 nm for Karyopherin and 492 for Calpain experiments and emission wavelength set at 518 and 516 nm, respectively. Titrations were performed using a 3 mm × 3 mm quartz cuvette with a starting volume of 200 μl. Each experiment was done one to four times and each polarization readings consisted of three averaged measurements. The data acquired were analyzed by fitting it to a single-site binding equation using non-linear regression with SigmaPlot software.

The ITC measurements for Calpain and Karyopherin binding to their respective partners were carried out using a VP-ITC and Auto-ITC calorimeters, respectively (MicroCal Inc.). The purified proteins in their respective elution buffers were placed in the sample cell. Peptides or partner proteins were placed in the injection syringe at a concentration of at least 10 times or more than that of the protein in the sample cell. The injection volume was 5 μl for each titration. The data were processed by fitting the calorimetric data using a one-site binding model in the ITC sub-routine in Microcal Origin software.

Results and discussion

Target selection

Three pathogen proteins were selected as targets to test our Scaffold Selection Protocol: a tuberculosis virulence factor (TVF), the malaria apical membrane antigen (AMA) and HA from influenza. Crystal or NMR structures were available for all three proteins, which allowed us to select unstructured regions as the target epitopes, such as flexible loops, disordered regions or protein termini, based on whether the electron density is weak to non-existent or whether the region shows large fluctuations between different NMR models, respectively. However, accessible regions of proteins can also be determined by experimental methods such as protease digestion or deuterium exchange or by computational predictions of intrinsic disorder, and therefore the protocol is not restricted to proteins with known three-dimensional structures.

Calpain is a good scaffold candidate for CFP-10

The TVFs CFP-10 and ESAT-6 form a heterodimer complex, which has been shown to play an essential role in tuberculosis pathogenesis. Both CFP-10 and ESAT-6 form helix-turn-helix structures, which associate to form a four-helix bundle. NMR-derived structures of the complex (PDB ID: 1WA8) show that both the N- and C-termini of CFP-10 and ESAT-6 are disordered, forming long flexible arms (Renshaw *et al.*, 2005). The sequences of these disordered termini (residues 1–6 and 79–99 of CFP-10 and 601–607 and 680–695 of ESAT-6, PDB numbering) were input into the Scaffold Selection Protocol. High scoring hits were further evaluated by probing the literature to determine expression levels for the potential scaffold, checking for previously identified hot spot residues in the crystallized peptide that may be present in our target epitope, and checking the measured binding affinity of the template complex. We also examined the template structure to make sure that it was sterically compatible with our target epitope.

Consideration of these criteria resulted in the selection of the pdb file 1NX1 as a template, in which the domain VI (DVI) of Calpain is bound to a peptide from Calpastatin (Todd *et al.*, 2003). The highest scoring alignment of the C-terminus of CFP-10 and the Calpastatin peptide from our Scaffold Selection Protocol is shown in Fig. 1B. Structural representation of Calpastatin peptide bound to Calpain and CFP-10 threaded onto the Calpastatin peptide backbone is shown in Fig. 1C. Calpain DVI and CFP-10 C-terminus are an appealing design candidate pair for several reasons. (i) The C-terminus of CFP-10 is thought to be important in the recognition of the host cell target protein in tuberculosis infection, thus it is possible that a protein scaffold designed to bind to this region will block the virulence of tuberculosis. (ii) Although we did not explicitly include secondary structure information in our sequence alignment scoring scheme, both the CFP-10 peptide and the high scoring Calpastatin peptide have intrinsic propensity for forming helices. The Calpastatin peptide binds to Calpain DVI as a helix while the chemical shift and nuclear overhauser effect data for the TVFs show that C-terminus of CFP-10 has a propensity to adopt a helical conformation. It is hypothesized that when the C-terminus of CFP-10 binds to the host cell protein, the helical conformation is stabilized (Renshaw *et al.*, 2005). (iii) Mutation data show that Leu606 on the Calpastatin peptide is important in Calpain DVI and Calpastatin binding (Ma *et al.*, 1994). This residue is conserved in the alignment of our target CFP-10 sequence to the Calpastatin peptide (Fig. 1B).

To determine whether Calpain is a promising scaffold for generating binders against CFP-10, the binding of the wild-type Calpain DVI for our target linear epitope peptide (C-terminus of CFP-10), as well as for the Calpastatin peptide, was determined using FP and ITC. Peptides, N-terminally labeled with a fluorescent dye (FITC) (CFP-10 peptide: FITC-βA-DEEQQQALSSQMGF, Calpastatin peptide: FITC-βA-PDDAIDALSSDFTS) were used for both FP and ITC (Fig. 2). Wild-type Calpain DVI bound to the Calpastatin peptide with low micromolar affinity (the measured K_d was 2 μM by FP and 5 μM by ITC) (Table I). Despite not being evolutionarily optimized for binding, the target CFP-10 peptide also bound to Calpain with measured affinities of 37 μM (FP) and 147 μM (ITC) (Table I). These results suggest that Calpain can serve as a scaffold for generating binders against CFP-10.

Karyopherin α is a good scaffold candidate for AMA1

The second target protein we tested our protocol with was domain III of the malarial AMA1 (AMA1). AMA1 is an essential protein for malaria parasite invasion of erythrocytes. It is established as a major malaria vaccine candidate and recently a high-affinity AMA1 binding peptide, identified using phage display, was shown to inhibit merozoite invasion of malaria parasites cultured *in vitro* (Li *et al.*, 2002; Keizer *et al.*, 2003; Harris *et al.*, 2005). The structure of domain III of AMA1 has been solved with NMR (PDB ID: 1HN6), and has several flexible regions surrounding a structured core held together by three disulfide bonds. Sequences from three disordered regions of AMA1 domain III were submitted to our Scaffold Selection Protocol (residues 436–439, 453–479, 510–545, PDB numbering). The top-scoring hit that passed our manual inspection criteria was a nuclear localization signal (NLS) peptide from c-*myc* proto-oncogene bound to Karyopherin α in the PDB file 1EE4 (Conti and Kuriyan, 2000). The alignment of the second disordered segment of AMA1 domain III with the NLS peptide is shown in Fig. 1B. A structural representation of Karyopherin α with NLS peptide bound and domain III of AMA1 threaded onto the backbone of the NLS peptide is shown in Fig. 1D. Karyopherin α and domain III of AMA1 are an interesting design candidate pair for several reasons. (i) Domain III of AMA1, in addition to domain I, is thought to be a binding hot spot for inhibition by various inhibitory molecules such as antibodies and peptides (Todd *et al.*, 2003; Harris *et al.*, 2005). Thus it is possible that a protein designed to bind to the AMA1 domain III could be used as a therapeutic against Malaria. (ii) Karyopherin α is a repeat protein composed of Armadillo motifs. Armadillo repeat proteins are involved in a broad range of protein–protein interactions with high affinities. They are able to bind to different types of peptides but retain binding conservation by utilizing binding to the peptide backbone. They also bind to extended conformation of peptide targets. A recent study which engineered a well-expressed and stable Armadillo repeat protein scaffold describes the possibility of using this scaffold to design a protein which binds to any sequence of interest with high affinity (Parmeggiani *et al.*, 2008). (iii) All the consensus residues of a classical NLS motif (K–K/R–X–K/R, where X is any residue) are conserved in the alignment with our target sequence from the AMA1 domain III (Fig. 1B). The Leu327 and the Asp328 after the consensus sequence, which have also been shown by mutagenesis studies to be important in binding (Makkerh *et al.*, 1996), are also conserved with identical (in the case of Leu327) or similar amino acid (in the case of Asp328 it is Asn), respectively, in our alignment (Fig. 1B).

To determine whether Karyopherin α is a good scaffold for generating binders against AMA1, we measured the binding affinity of the wild-type Karyopherin α with our target linear epitope peptide (second disordered segment of AMA1 domain III) and its native binder, the c-*myc* NLS peptide using FP (AMA1 peptide: FITC-βA-IRESKRIKLND, c-*myc* NLS peptide: FITC-βA-GPAAKRVKLDS). Both the AMA1-derived peptide and the c-*myc* NLS peptide bound to Karyopherin α with a dissociation constant of 50 ± 0.2 nM (AMA1) or ± 10 (NLS) (Fig. 3A and B, Table I). The high affinity of the target sequence for the wild-type Karyopherin α is thought to be the result of the high identity between the native binding peptide and the target sequence, especially in the consensus residues known to be important in binding.

Because of the high affinity between the candidate protein scaffold, Karyopherin α and our target sequence peptide from AMA1 domain III, we were curious whether Karyopherin α would bind to the full-length AMA1 domain III. We expressed the full-length AMA1 domain III by using a native Ni-NTA purification protocol. Because our purification scheme was different from the original denaturing Ni-NTA purification and refolding protocol used by Nair *et al.*, we decided to compare the structure of our protein, especially the formation of the three disulfides bonds, with the published

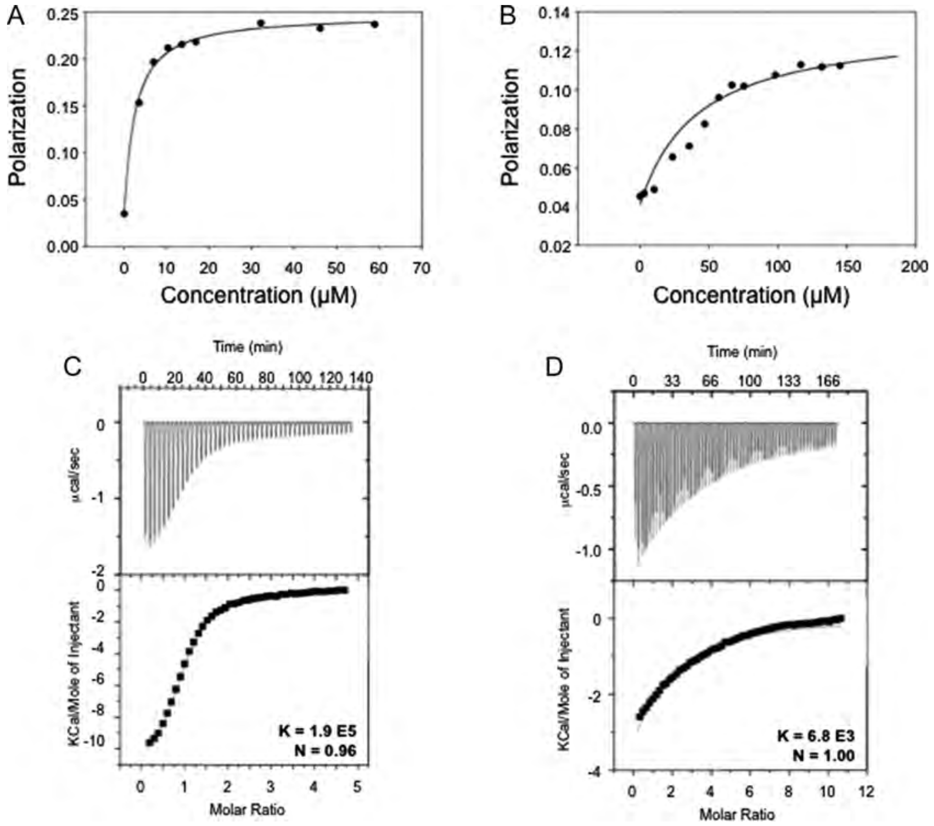

Fig. 2. FP data for Calpain and Calpastatin peptide (A) and Calpain and CFP-10 peptide (B). ITC data for Calpain and Calpastatin peptide (C) and Calpain and CFP-10 peptide (D).

Table I. Binding affinity from FP and ITC experiments

	K_d (FP)	K_d (ITC)
Calpain–Calpastatin peptide	2 ± 0.23 μM (4)	5.2 ± 0.1 μM (1)
Calpain–CFP-10 peptide	37 ± 5.7 μM (2)	147 ± 8 μM (1)
Karyopherin–*myc*NLS peptide	50 ± 10 nM (1)	
Karyopherin–AMA1 peptide	50 ± 0.2 nM (2)	
Karyopherin–AMA1 full length		760 ± 20 nM (2)

The number of replicates is marked in parentheses. Standard deviations are reported except in the case of single replicates where the error to the fit is reported instead.

NMR structure. Comparison of the ^{15}N NMR chemical shift of our sample with the published chemical shift of AMA1 domain III from the BioMagResBank confirms that the two proteins are identical in all of the residues where peak intensity is discernable (data not shown) (Nair *et al.*, 2001). The binding affinity of the full-length AMA1 domain III protein for Karyopherin α was measured using ITC and was determined to be 760 nM (Fig. 3C, Table I).

Limitations of the Scaffold Selection Protocol

The successful results from the two cases above plus an additional case in which the LOV2 domain of *Avena Sativa* phototropin 1 sequence was used as the target sequence. Lungu *et al.* (2012) show that our initial assumption was correct. By using sequence alignments and structural information to select an optimal protein scaffold for design, we have obtained proteins that bind to our target epitope peptides. These protein scaffolds can be used as the starting point for the generation of designed *de novo* protein–protein interactions for the target protein of interest. With computational protein design or directed evolution methodologies it should be possible in many cases to use these scaffolds to engineer higher affinity and specificity to the protein–protein interactions of interest.

In contrast to the three successful examples, the surface cleavage loop of the influenza HA did not bind to its predicted protein scaffold. The best scoring hit for HA from the Scaffold Selection Protocol was CDK2/cyclin protein bound to the p107 peptide (Fig. 1B, PDB file 1H28) (Lowe *et al.*, 2002). Although binding was observed between the p107 peptide and CDK2/cyclin when probed with FP experiments (K_d value is 37 μM), no binding was observed between the HA peptide and CDK2/cyclin. Closer examination of CDK2/cyclin and p107 peptide structure revealed a possible explanation for the negative result. One of the residues in the HA sequence that is immediately adjacent to the aligned region used to identify the match is structurally incompatible (alanine in p107 peptide vs. glutamate in HA) with CDK2/cyclin according to the energy function in the molecular modeling program ROSETTA. In the p107 CDK2/cyclin crystal the alanine is buried in a hydrophobic pocket that is not sterically or chemically compatible with a glutamate (Supplementary data, Fig. S1). This example shows one limitation of our Scaffold Selection Protocol. Because we are using a window of six residues to obtain hits with high alignment score, residues outside the window, which might be

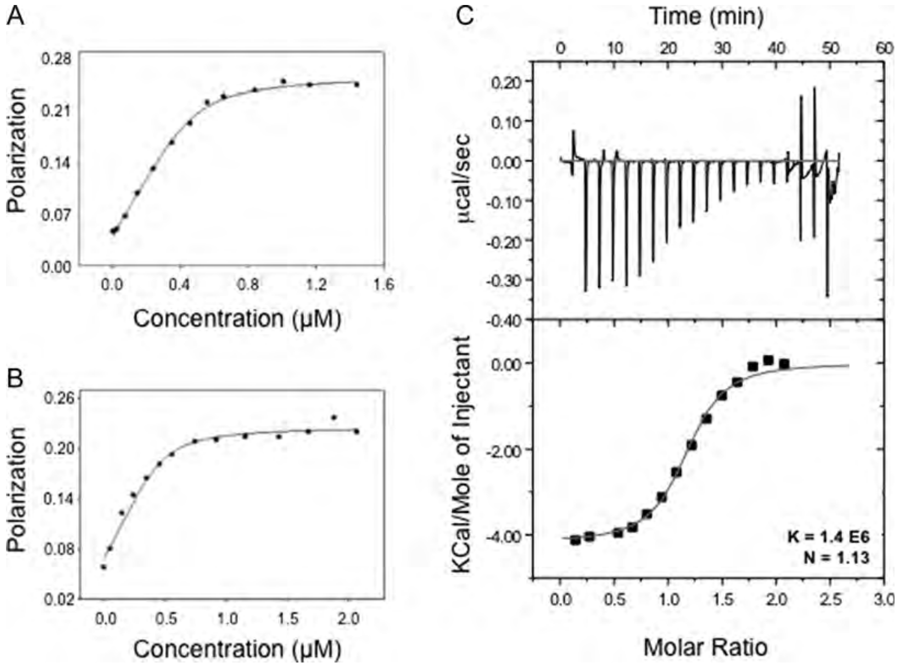

Fig. 3. FP data for Karyopherin and c-*myc* NLS peptide (A) and Karyopherin and AMA1 domain III peptide (B). ITC data for Karyopherin and full-length AMA1 domain III (C).

important in binding between the original protein and the peptide are ignored during alignment. Thus a more in-depth inspection of not only the aligned region but also the regions immediate to it should be performed before choosing proteins as scaffolds.

Another limitation of our method as currently constructed is that it is constrained by the number and variety of peptides that have been crystallized bound to a protein. For instance we tested several targets, malaria surface protein, dengue virus envelope protein and influenza neuraminidase, for which we did not identify any strong hits. This limitation will become less significant as the PDB grows, but could also be mitigated by including peptide-protein pairs in our database that are experimentally known to interact, but have not been co-crystallized. The most useful cases would be instances where there are mutational data on the peptide that indicates which residues are important for binding. This could be used when scoring alignments with the target epitope similar to the manner in which we currently use structural information.

Scaffold Selection Protocol server

To allow easy public access to our Scaffold Selection Protocol, we have created a web server: http://rosettadesign.med.unc.edu/scaffold/. The user submits the sequence of the target and a sliding window size for the alignment process (defaults to 6). Results are emailed to the user. The result, composed of six columns, is a list of matches, with the best match based on the overall score at the top. The first column shows you the matching residue number and the target sequence of the epitope from your input. The second column is the PDB ID of the high scoring match from the database, the chain of the matching peptide in the PDB, the peptide residue number in parenthesis and the matching peptide sequence. Next is the number of neighbors from the crystal structure for each of the peptide residues in the match. In the case of non-natural amino acids the number of neighbors are not counted. The fourth column is the complete peptide sequence of the match in the PDB file. And the last two columns show you the raw score and the overall score. The overall score is the raw score normalized by the number of natural amino acid in the match. The list of PDB files in our peptide–protein PDB database can be also found on the website.

Conclusion

We have developed a protocol that selects protein scaffolds that are well-suited for binding target proteins of interest. It uses the sequence information of the target protein plus structural information from the PDB database to select the scaffolds. We validated the protocol by using it to successfully identify novel binding partners for the TVF CFP-10 and for malaria AMA 1. There are many possible applications for the protocol, including the identification of novel regulators of disease-related proteins, the generation of crystallization agents that bind to the flexible regions of proteins and hold them rigid and the creation of biosensors that detect conformational changes or post-translational modifications in a target epitope.

Supplementary data

Supplementary data are available at *PEDS* online.

Acknowledgments

We thank A. Tripathy for help in ITC experiments and K. Y. Lee for editing the manuscript. Some computational calculations relevant to the work in this paper was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

Funding

This work was supported by the National Institutes of Health (grant numbers R01GM073960 and R01GM093208).

References

Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.

Brunner,H.R., Gavras,H. and Laragh,J.H. (1973) *Lancet*, **2**, 1045–1048.

Conti,E. and Kuriyan,J. (2000) *Structure*, **8**, 329–338.

Craig,L., Sanschagrin,P.C., Rozek,A., Lackie,S., Kuhn,L.A. and Scott,J.K. (1998) *J. Mol. Biol.*, **281**, 183–201.

David,R., Korenberg,M.J. and Hunter,I.W. (2000) *Pharmacogenomics*, **1**, 445–455.

Der,B.S., Machius,M., Miley,M.J., Mills,J.L., Szyperski,T. and Kuhlman,B. (2012) *J. Am. Chem. Soc.*, **134**, 375–385.

Dunbrack,R.L., Jr. (2006) *Curr. Opin. Struct. Biol.*, **16**, 374–384.

Fleishman,S.J., Whitehead,T.A., Ekiert,D.C., Dreyfus,C., Corn,J.E., Strauch,E.M., Wilson,I.A. and Baker,D. (2011) *Science*, **332**, 816–821.

Harris,K.S., Casey,J.L., Coley,A.M., *et al.* (2005) *Infect. Immun.*, **73**, 6981–6989.

Hashemzadeh,M., Furukawa,M., Goldsberry,S. and Movahed,M.R. (2008) *Exp. Clin. Cardiol.*, **13**, 192–197.

Jha,R.K., Leaver-Fay,A., Yin,S., Wu,Y., Butterfoss,G.L., Szyperski,T., Dokholyan,N.V. and Kuhlman,B. (2010) *J. Mol. Biol.*, **400**, 257–270.

Karanicolas,J., Corn,J.E., Chen,I., *et al.* (2011) *Mol. Cell.*, **42**, 250–260.

Keizer,D.W., Miles,L.A., Li,F., Nair,M., Anders,R.F., Coley,A.M., Foley,M. and Norton,R.S. (2003) *Biochemistry*, **42**, 9915–9923.

Koide,S. (2009) *Curr. Opin. Biotechnol.*, **20**, 398–404.

Lee,S.C., Park,K., Han,J., *et al.* (2012) *Proc. Natl Acad. Sci. USA*, **109**, 3299–3304.

Li,F., Dluzewski,A., Coley,A.M., Thomas,A., Tilley,L., Anders,R.F. and Foley,M. (2002) *J. Biol. Chem.*, **277**, 50303–50310.

London,N., Raveh,B., Movshovitz-Attias,D. and Schueler-Furman,O. (2010) *Proteins*, **78**, 3140–3149.

Lowe,E.D., Tews,I., Cheng,K.Y., Brown,N.R., Gul,S., Noble,M.E., Gamblin,S.J. and Johnson,L.N. (2002) *Biochemistry*, **41**, 15625–15634.

Lungu,O.I., Hallett,R.A., Choi,E.J., Aiken,M.J., Hahn,K.M. and Kuhlman,B. (2012) *Chem. Biol.*, **19**, 507–517.

Ma,H., Yang,H.Q., Takano,E., Hatanaka,M. and Maki,M. (1994) *J. Biol. Chem.*, **269**, 24430–24436.

Makkerh,J.P., Dingwall,C. and Laskey,R.A. (1996) *Curr. Biol.*, **6**, 1025–1027.

Nair,M., Hodder,A.N., Hinds,M.G., Anders,R.F. and Norton,R.S. (2001) *J. Biomol. NMR*, **19**, 85–86.

Parmeggiani,F., Pellarin,R., Larsen,A.P., Varadamsetty,G., Stumpp,M.T., Zerbe,O., Caflisch,A. and Pluckthun,A. (2008) *J. Mol. Biol.*, **376**, 1282–1304.

Pawson,T. and Nash,P. (2003) *Science*, **300**, 445–452.

Renshaw,P.S., Lightbody,K.L., Veverka,V., *et al.* (2005) *Embo. J.*, **24**, 2491–2498.

Schrodinger,L.L.C. (2010) *The PyMOL Molecular Graphics System*. Schrödinger, LLC.

Sidhu,S.S. and Koide,S. (2007) *Curr. Opin. Struct. Biol.*, **17**, 481–487.

Steward,R.E., MacArthur,M.W., Laskowski,R.A. and Thornton,J.M. (2003) *Trends Genet.*, **19**, 505–513.

Stranges,P.B. and Kuhlman,B. (2013) *Protein Sci.*, **22**, 74–82

Stranges,P.B., Machius,M., Miley,M.J., Tripathy,A. and Kuhlman,B. (2011) *Proc. Natl Acad. Sci. USA*, **108**, 20562–20567.

Takano,E., Ma,H., Yang,H.Q., Maki,M. and Hatanaka,M. (1995) *FEBS Lett.*, **362**, 93–97.

Todd,B., Moore,D., Deivanayagam,C.C., Lin,G.D., Chattopadhyay,D., Maki,M., Wang,K.K. and Narayana,S.V. (2003) *J. Mol. Biol.*, **328**, 131–146.

Xu,D., Xu,Y. and Uberbacher,E.C. (2000) *Curr Protein Pept Sci*, **1**, 1–21.