Trends in Biochemical Sciences

CellPress

Review

Protein Engineering by Combined Computational and *In Vitro* Evolution Approaches

Lior Rosenfeld,[1,3] Michael Heyne,[1,2,3] Julia M. Shifman,[2,*] and Niv Papo[1,*]

Two alternative strategies are commonly used to study protein–protein interactions (PPIs) and to engineer protein-based inhibitors. In one approach, binders are selected experimentally from combinatorial libraries of protein mutants that are displayed on a cell surface. In the other approach, computational modeling is used to explore an astronomically large number of protein sequences to select a small number of sequences for experimental testing. While both approaches have some limitations, their combination produces superior results in various protein engineering applications. Such applications include the design of novel binders and inhibitors, the enhancement of affinity and specificity, and the mapping of binding epitopes. The combination of these approaches also aids in the understanding of the specificity profiles of various PPIs.

Engineering Protein–Protein Interactions

PPIs are crucial for all essential processes in the cell, including transcription, translation, replication, intra- and intercellular signaling, and molecular transport. Thus, it is not surprising that aberrant PPIs have been implicated in several types of disease, including neurodegenerative diseases and cancers. Studies from many different laboratories have shown that it is possible to modify various characteristics of PPIs through mutations and even to 'create' novel PPIs. Therefore, PPI engineering presents an attractive strategy in synthetic biology and in the design of biosensors, imaging agents, and novel therapeutics.

Two different approaches are commonly used in PPI engineering: the combinatorial approach and computational protein design (CPD). In the first approach, also known as directed evolution, large libraries of protein mutants are constructed, proteins with certain binding characteristics are selected, and the sequences of the selected proteins are determined. This 'irrational' approach is used for affinity maturation [1], for identifying target-specificity profiles [2,3], and for producing high-affinity and high-specificity PPI inhibitors from antibodies [4], natural protein effectors [5,6], and unrelated protein scaffolds [7–9]. The main advantage of combinatorial methods is that they require only minimal knowledge of the PPI under study. However, combinatorial approaches do not provide much information about the nature of the created intermolecular contacts, which often hampers an understanding of the obtained results. An additional drawback is that the number of sequences that can be explored by such methods is limited to several million, thus allowing the exploration of only a small fraction of the protein sequence space.

The second approach, CPD, is a 'rational' methodology that relies on our understanding of the biophysical forces that govern protein binding. This method requires a detailed knowledge of the

Trends

Novel protein binders to various targets can be engineered by first applying computational approaches and then optimizing the binder with yeast surface display (YSD). Computational methods can narrow down the choices of possible mutations and combinations of mutations, thereby enabling the construction of smaller, more focused libraries.

Together, combinatorial and computational techniques can be used to map binding epitopes of poorly characterized PPIs, thus identifying binding hot-spots and affinity-enhancing mutations.

Binding-specificity profiles can be mapped with a combination of combinatorial approaches and next-generation sequencing (NGS). Computational methods contribute to understanding the nature of specific PPIs, determining their binding specificity, and predicting ligands for homologous targets.

[1]Department of Biotechnology Engineering and the National Institute of Biotechnology in the Negev, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[2]Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel
[3]These authors contributed equally to this work.

*Correspondence:
jshifman@mail.huji.ac.il (J.M. Shifman) and papo@bgu.ac.il (N. Papo).

structure and function of the PPI under study. As for the combinatorial approach, CPD has been successfully applied to manipulating PPI binding specificity [10–14] and binding affinity [15–20]. More recently, it has also been used to create novel binding interactions [21–23]. The advantage of CPD lies in its ability to explore a huge sequence space *in silico* and to select a few tens of protein sequences for experimental verification. Yet, CPD is impeded by the inaccuracy of the energy functions for calculating binding energetics and by current sampling methods that sometimes 'miss' the correct conformations of the binding interface residues.

Thus, each method has its particular advantages and disadvantages. Recent studies have shown that combining the two approaches could overcome the above limitations and produce superior results in PPI design. In this review, we describe the application of combined computational and combinatorial methodologies to problems in PPI characterization and engineering.

Increasing the Affinity and Specificity of Binding Interactions

Combinatorial and computational methods, separately or in combination, can be used to alter the binding characteristics of natural PPIs for various biomedical and synthetic biology applications [5,6,24–28]. All combinatorial approaches are limited in the size of the libraries that they can explore (a maximum of 10^{10} mutants; Box 1), which means that, in binding selection experiments, eight positions at most in a protein can be randomized to all 20 amino acids. Yet, the number of protein residues that can affect binding affinity and specificity, either through direct contacts or through allosteric effects, is usually larger than eight [29]. To overcome this limitation and to better exploit the randomization strategy, CPD can be used to design focused libraries of protein binders by identifying positions on the protein–protein interface at which mutations have the highest potential for affinity and specificity improvement and the lowest potential to compromise the protein structure [30].

Guntas *et al.* [31] used such an approach to generate a photoswitchable binding protein based on a naturally occurring photoswitch, the light-oxygen-voltage 2 (LOV2) domain, which partially unfolds upon exposure to light. The authors embedded the SsrA peptide into the LOV2 domain and engineered a light-sensitive binder for the natural ligand of SsrA, SspB. CPD was used to

Box 1. Principles of Combinatorial Approaches

The most commonly used combinatorial approaches for PPI engineering are phage display (PD), yeast surface display (YSD), and human surface display (HSD). In all these techniques, large combinatorial libraries of proteins are displayed on a cell surface, and a receptor protein is used as a 'bait' to select for binders. Since each cell contains the DNA for the displayed protein mutant, the sequence of the selected protein binders can easily be recovered.

PD

In PD, the library of interest is fused to a bacteriophage coat protein and displayed on the phage surface. Thereafter, to isolate specific binders, the pool of phages is mixed with a target protein that has been immobilized on either paramagnetic beads or microtiter plates [77]. The binding of the selected clones is verified by phage ELISA, in which the phage-displayed protein mutant is added to the plate-immobilized target and binding is detected via colorimetric output. The library size for PD, the largest of the cell display methods, can reach 10^{10} mutants. The limitation of this technique is that large proteins, proteins containing disulfide bonds, and proteins with post-translational modifications (PTMs) are frequently not compatible with the technology.

YSD

In YSD, combinatorial protein libraries have a maximal size of 10^8 mutants and are displayed on the surface of *Saccharomyces cerevisiae* cells [78,79] (Figure IA). The large size of yeast cells enables the selection of antigen-binding cells by flow cytometry (fluorescence-activated cell sorting; FACS) (Figure IB), thereby conferring a major advantage over the PD technology. YSD utilizes a two-color fluorophore labeling system (Figure IC), with one fluorophore detecting expression and the other detecting antigen binding. Thus, stability and affinity screenings are accomplished simultaneously. Once the selection process is complete, the binding affinity of the individual protein mutants can be estimated while the protein is displayed on the yeast surface, thereby enabling rapid screening of the clones without the need for lengthy expression and purification processes.

HSD

In HSD, a protein library (with a maximal size of 10^6 mutants) is expressed on the surface of human cells, and the cells carrying binding mutants are selected by FACS (Figure ID). The particular advantage of HSD is that it facilitates correct protein folding and the display of human proteins with PTMs [80–83]. The use of HCD greatly enhances the probability of selecting protein variants that function as agonists or antagonists of human proteins and could thus serve as future therapeutics.

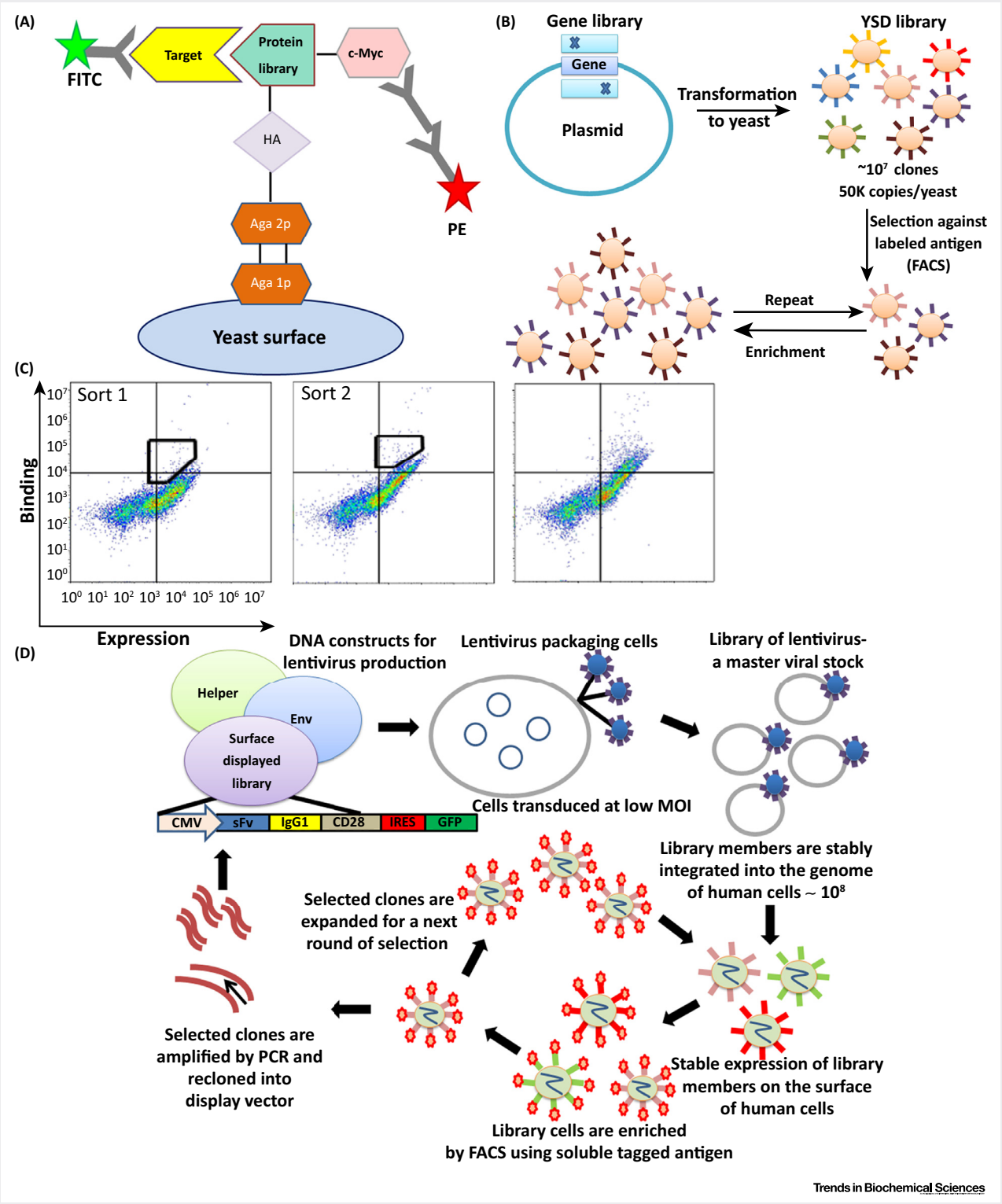
(A)
FITC
Target
Protein library
c-Myc
HA
Aga 2p
Aga 1p
PE
Yeast surface
(B)
Gene library
Gene
Plasmid
Transformation to yeast
YSD library
~10^7 clones
50K copies/yeast
Selection against labeled antigen (FACS)
Repeat
Enrichment
(C)
Sort 1
Sort 2
Binding
Expression
(D)
DNA constructs for lentivirus production
Helper
Env
Surface displayed library
CMV
sFv
IgG1
CD28
IRES
GFP
Lentivirus packaging cells
Library of lentivirus- a master viral stock
Cells transduced at low MOI
Library members are stably integrated into the genome of human cells ~ 10^8
Stable expression of library members on the surface of human cells
Library cells are enriched by FACS using soluble tagged antigen
Selected clones are expanded for a next round of selection
Selected clones are amplified by PCR and recloned into display vector
Trends in Biochemical Sciences

Figure I. Principles of the Experimental Display Approaches. (A) Yeast surface display (YSD) system. The library of the protein of choice is fused to the yeast cell wall and is monitored via a c-Myc epitope tag. Binding of the target protein to the displayed protein, followed via a secondary FITC-labeled antibody, enables detection of binding, while binding to an anti c-Myc antibody, followed via a secondary phycoerythrin (PE)-labeled antibody, allows detection of expression. (B) The process of YSD library sorting. First, the library is transformed into yeast to generate diversity of approximately 10^7 clones. The constructed library is expressed and several rounds of sorting (for affinity maturation) are performed using fluorescence-activated cell sorting (FACS). (C) FACS density dot plots of the libraries. Different receptor concentrations are used during the different sorting rounds. The X-axis represents the level of expression, while the Y-axis represents the level of binding. The cells with highest level of binding normalized to expression are collected in each round, as shown by a black gate. The first two panels show the sorting process and the rightmost panel shows FACS analysis of an affinity-matured library. (D) Human surface display (HSD) system. The desired protein library is inserted into human cells by using a lentivirus, with a single clone being integrated into the genome of each cell. The protein is then presented on the surface of the cell and is sorted in a similar way to that shown in (C). Adapted from [51] (A).

generate a focused library of LOV2 variants by identifying positions at which mutations did not disrupt protein folding. Phase display (PD) technology (Box 1) was applied to select LOV2 variants with the highest change in Kd for binding to SspB upon exposure to light. Among the most successful photoswitchable binders selected from the libraries, one changed the affinity for wild-type SspB from 4.7 μM to 132 nM and the other changed the affinity for an SspB mutant from 47 μM to 800 nM, following exposure to blue light.

Integrating combinatorial and computational tools may also predict specific positions that are likely to improve PPI binding affinity [9,16]. This methodology was pursued by Qiao *et al.* to achieve affinity maturation of the single-chain variable fragment of an antibody (scFv) of the therapeutic MIL5 antibody (M5scFv) against the human epidermal growth factor receptor 2 (HER2), a primary target for drug design in cancer. A focused library of M5scFv was designed by predicting positions at which mutations would produce higher affinity towards HER2. PD technology was then used to select M5scFv variants that gave a fourfold enhancement of *in vitro* binding affinity to HER2 and acted as potent inhibitors of tumor growth both in cell lines and *in vivo* [32].

Several studies have shown the advantage of combining computational and experimental techniques to select positions that are predicted to maintain protein structure while enhancing binding specificity [29,33–36]. For example, Dutta *et al.* [35] used combined computational-combinatorial approaches to design the Bim-BH3 variant that shows a marked preference for Bcl-x_L, its prosurvival binding partner, over other natural ligands, such as Bfl-1, Bcl-2, and Bcl-w. SPOT peptide arrays that measured relative binding affinities between 180 Bim-BH3 single mutants and five natural Bim-BH3 ligands were used to select Bim-BH3 positions for randomization on the basis of their potential to improve specificity [37]. Focused libraries of Bim-BH3 variants randomizing eight positions were constructed and selected by YSD (Box 1) for binding to Bcl-x_L, resulting in a Bim-BH3 variant with a 1000-fold preference for Bcl-x_L over other prosurvival ligands.

In summary, the design of focused combinatorial libraries greatly facilitates protein engineering aimed at improving PPI affinity and specificity. Thus, we expect that this approach will become widely used in future protein engineering studies.

Engineering of Novel Binders

Over the past decade, it has been shown that protein-based inhibitors or novel binding domains (NBDs) can be created from a range of protein scaffolds that display secondary and tertiary structures of different sizes. Moreover, similar to an antibody, a single scaffold protein can be evolved to bind to many different unrelated targets. Several NBDs have been evolved using combinatorial approaches [38–40], although not all attempts to isolate novel binders have been successful. CPD approaches to NBD design have proved to be challenging, mainly due to the difficulty of creating the initial structure of the NBD–target complex without prior knowledge of the NBD sequence. For this reason, the initial computationally designed NBDs suffered from low specificity and low affinity [41]. More recently, however, an attractive strategy has been proposed

in which CPD design is used to obtain an initial low-affinity NBD, which is subsequently subjected to affinity maturation via combinatorial approaches.

Several different computational approaches have been suggested for the creation of the initial structure of the NBD–target complex. In the hot-spot approach [42], a few energetically favorable interactions between the target and disembodied amino acids are first created. These interactions are then grafted onto candidate scaffold proteins, and the best scaffold is selected on the basis of its ability to interact with the target. Thereafter, the residues surrounding the binding hot-spot interactions on the NBD are redesigned. Using this approach, Fleishman *et al.* designed NBDs for a conserved region of influenza hemagglutinin and scanned them for binding by using YSD [22,43]. Of 88 experimentally tested designs, two showed a detectable affinity for hemagglutinin. The NBDs were further optimized by scanning the effect of all single mutations on the binding affinity with YSD and next-generation sequencing (NGS) and incorporating the best mutations into the NBD; two variants with Kd values of 900 and 600 nM were obtained. Karanicolas *et al.* [23] set their goal to design a novel binding interface between two non-interacting proteins. Initially, they searched for the best potential binding partner for an ankyrin repeat protein among 37 different thermostable proteins and chose PH1109, a *Pyrococcus horikoshii* coenzyme A-binding protein, as a scaffold for NBD design. Then, they grafted frequent hot-spot interactions between Trp or Tyr and Asp residues at various places in the novel PPI and redesigned the residues surrounding the hot-spot interactions. Their best PPI design exhibited a Kd of 100 nM and was further optimized through YSD to a Kd of 180 pM. The X-ray structure confirmed the designed hot-spot interaction in the novel PPI, but revealed that the actual binding interface was rotated by 180° with respect to the original model, pointing to the need to further improve the CPD methodology for designing peripheral interface contacts.

In an alternative grafting approach, a protein epitope known to bind to a specific target is grafted onto a scaffold protein. Then, the backbone structures of the grafted and the connecting regions are modeled, and the residues surrounding the grafted region are computationally optimized to enhance NBD stability and affinity [44,45]. Azoitei *et al.* [44] used this approach to computationally graft a noncontinuous, two-segment HIV gp120 epitope onto an unrelated protein scaffold, endoglucanase. Selection from the designed library by YSD identified several binding clones. Out of 62 designs that showed binding by YSD, 25 could be expressed as isolated soluble proteins. While the affinity of the computationally designed binder was weak (300 μM), affinity maturation by YSD yielded a mutant with a Kd of 10.3 nM. In a similar approach, Azoitei *et al.* [45] grafted the epitope of the HIV1-neutralizing antibody 2F5 onto an unrelated scaffold. The best-selected NBD exhibited a Kd of 400 pM.

Procko *et al.* [46] used two approaches to design an NBD for BHRF1, a homolog of the human Bcl-2 protein. In the first approach, they grafted Bim-BH3 side chains onto various three- and four-helical bundle proteins and optimized the interactions with BHRF1 to produce a BHRF1 NBD with a Kd of approximately 60–80 nM. In the second approach, they designed a BHRF1 NBD based on a helical bundle through several rounds of sequence design and structure minimization, resulting in a binder with a Kd of approximately 60 nM. The NBD was then subjected to affinity maturation using error-prone PCR and YSD, and specificity was optimized by incubating the NBD library with unlabeled Bcl-2 proteins as competitors during the YSD selection. The obtained NBD exhibited a Kd of 220 pM and high specificity towards BHRF1.

In a different study, Procko *et al.* [47] used two approaches to design an NBD to hen egg lysozyme. In the first approach, a scaffold with high surface complementarity to the target was chosen, and the interface was directly redesigned using the CPD approach. All interfaces resulting from this approach were largely hydrophobic and contained a few specific contacts. Among the 24 experimentally tested NBDs, only one showed binding to lysozyme by YSD, but

exhibited no specific binding when tested in isolation. Designs using the second, hot-spot-based approach yielded more polar binding interfaces compared with those designed with the first approach. Twenty-one variants were experimentally characterized, and one NBD showed specific binding to lysozyme as a purified protein. The initial hen egg lysozyme inhibitor, with a Kd of 7 μM, was further optimized by YSD, producing an inhibitor with a Kd of 1.4 nM.

In yet another approach for NBD design, a novel binding interface is created through an extension of a complementary secondary structure element. In such an approach, a homodimer was designed by Kuhlman and colleagues by pairing exposed beta strands of various candidate proteins, removing backbone clashes, and computationally redesigning interfacial positions [48]. The designed homodimer, based on the γ-adaptin appendage domain, exhibited a Kd of 1 μM. The X-ray structure revealed high similarity with the computational model, with a RMSD of 1.0 Å. In a recent study, an ∝-helix-mediated homodimer based on a monomeric *Drosophila* engrailed homeodomain scaffold was designed using the CPD methodology and confirmed by its NMR structure [49]. In that study, 128 computationally designed variants were screened using a Förster Resonance Energy Transfer assay between identical fluorophores (homo-FRET) to select soluble and stable homodimer variants.

In summary, several computationally designed NBDs have been reported over the past few years. Nonetheless, purely computational design of NBDs remains a difficult task with a relatively low success rate. All the above studies showed that combinatorial techniques, such as YSD, could be used to facilitate NBD engineering through quick assessment of the designed interactions and through affinity maturation.

Epitope Mapping and Analysis of PPIs

To better understand and manipulate PPIs, one first needs to identify the set of amino acids that are directly involved in the interactions; that is, the PPI binding interface or the binding epitope. It is possible to extract the binding epitope from a high-resolution PPI structure, but in many cases structure determination presents a major challenge. An additional problem is that, while residues in the direct binding interface can be easily identified from the structure, other residues that contribute to binding energetics through allosteric effects cannot be directly inferred. An efficient method for epitope mapping using YSD has been established [50]. In this method, a library of single point mutants that span the sequence of a particular ligand protein is screened for decreased affinity to the protein target. Sequencing of the selected clones allows identification of particular mutations that reduce binding affinity to the target. Computational modeling of experimentally identified mutations is then applied to explain the effect of such mutations. For example, this modeling may be used to predict whether these mutations reduce affinity through protein unfolding, through elimination of favorable intermolecular interactions, or through allosteric effects. Recently, several studies used such integrated approaches to map binding epitopes of a few therapeutically important PPIs [51–53].

Rosenfeld *et al.* used YSD to map residues important for interactions between the macrophage colony-stimulating factor (M-CSF) and its receptor c-FMS [51]. Since there is no available structure for the M-CSF•c-FMS complex but structures for the unbound components have been solved, the researchers used homology modeling to produce a structure of the M-CSF•c-FMS complex and to infer the location of the binding interface residues. They then computed the effect of experimentally identified mutations on M-CSF stability and on its binding affinity to c-FMS and divided the mutations into two groups; namely, those that significantly destabilized the protein and those that eliminated favorable interactions with the receptor. Three selected M-CSF mutants were expressed and purified and their *in vitro* affinities to c-FMS were measured, confirming both YSD selection results and the computational predictions regarding the effect of the mutations.

A similar approach was applied by Traxlmayr *et al.* to identify the residues responsible for the stability of the CH3 domain of human IgG1 [54]. In that study, a library of CH3 domain point mutants was screened by YSD for increased stability, which was determined according to the capability of the protein to bind its binding partners, an anti-CH2 antibody and FcγRI, following heat incubation. The resultant libraries were sequenced using NGS, enabling construction of the entire stability landscape of the CH3 domain and identifying the binding epitope of the CH2 domain to anti-CH2 and FcγRI. Experimentally observed stability landscapes agreed well with the evolutionary sequence conservation profiles and with CPD-based calculations of changes in protein stability due to mutation.

In another example, Mata-Fink *et al.* [53] used YSD to analyze the direct binding epitopes of the broadly neutralizing anti-gp120 antibody VRC01 and two VRC01-competitive antibodies, b12 and b13. In addition to using a random library of gp120 mutants, the authors constructed a rationally designed library of a stripped-core gp120 based on a homology model of this protein. The library contained gp120 mutants with highly disruptive amino acid substitutions based on differences in polarity, charge, and size, and restricted to mutations that are observed in HIV viruses with at least 0.1% frequency. The resulting epitope map was consistent with the map produced using the random stripped-core gp120 library and with crystallographic data available for the homologous complex. The study revealed energetic differences in b12 and b13 epitopes that had not been obvious from the existing crystal structures of the antibody•gp120 complexes.

As shown in these examples, combining computational and combinatorial methods permits more efficient mapping of PPI binding epitopes and reveals the effect of each protein position on binding energetics.

Computational Modeling for Understanding the Selection Results and for Substrate Prediction

Advances in PD and YSD technologies combined with NGS have enabled experimentalists to generate an unprecedented amount of data on protein sequences that are compatible with high-affinity binding. In such experiments, binding selection from a large combinatorial library of mutants is followed by sequencing of multiple binding clones. Alignment of the sequences is converted into a position weight matrix (PWM), in which each column contains an amino acid frequency observed at each ligand position, thus generating binding specificity profiles [55]. This experimental approach can be utilized for the prediction of ligands for homologous domains or for the design of proteins with altered binding specificity. However, to convert the large amounts of data that are generated with this approach into useful information, computational modeling is necessary. Here, two distinct approaches have been applied: one is based on machine learning (ML) techniques and the other on the CPD approach (Box 2).

It has been shown that ML approaches are especially powerful for predicting interactions of small domains with their linear peptide ligands, such as in the case of SH3 and PDZ domains [55–62]. Structure-based methods have been also used to predict binding specificity profiles for SH2 [63], SH3 [64], and PDZ domains [65,66]. In addition to predicting binding specificity profiles, computational modeling can assist in explaining binding selection results and in designing proteins with altered specificity. In many binding selection experiments, it is not clear why a particular amino acid is preferred at a certain position and what mutations are crucial for the desired function. Since it is frequently important in protein engineering studies to keep the number of mutations to a minimum, it is necessary to distinguish the truly beneficial mutations from neutral mutations. For these purposes, various computational methods have been utilized, including CPD, MD simulations, and homology modeling (Box 2).

Box 2. Principles of Computational Approaches

Computational protein design

The CPD methodology is implemented in several programs [84–86] that all use the same basic principles to compute the amino acid sequences most compatible with a given 3D protein structure. As the input for the CPD calculations, a high-resolution structure or an ensemble of structures for a PPI is required. Movements of the amino acid side chains are modeled by representing them using low-energy conformations or rotamers that are taken from a rotamer library [87]. In CPD calculations, the protein backbone is either fixed or allowed small movements [71] and side chain–side chain and side chain–backbone interactions are evaluated using an atomic-based energy function representing physical interactions, such as van der Waals energy, electrostatic interactions, hydrogen bonds, or solvation [88]. Various fast search algorithms are then used to predict the amino acid sequences corresponding to the lowest energy conformation [89]. For PPI engineering, the change in the free energy of binding ($\Delta\Delta G_{bind}$) is calculated and is then used to predict changes in PPI affinity and specificity (Figure I) [12].

Machine-learning methods

ML approaches use a set of experimental data to train an algorithm to predict a certain physical parameter (e.g., Kd of binding) or to discriminate between binding and nonbinding protein sequences [55]. The advantages of ML methods are that they are rapid and have the ability to predict binding based on sequence information alone. Thus, ML methods do not require knowledge of PPI structures, although incorporation of structure-based features frequently increases prediction accuracy [64,90]. The main disadvantage of these methods lies in the large amount of data required to train the algorithm, including data on nonbinding sequences. Such data are sometimes difficult to obtain in high-throughput experiments.

Molecular dynamics simulations

Molecular dynamics (MD) uses atomic-based molecular force fields and Newtonian equations of movement to simulate structural changes in proteins. In PPI engineering, MD is applied to model substantial backbone changes that are associated with binding, to generate structural ensembles of proteins for subsequent design calculations, or to evaluate the designed proteins [91,92]. Based on these structural ensembles, $\Delta\Delta G_{bind}$ values could be calculated using CPD or different methodologies (Figure I).

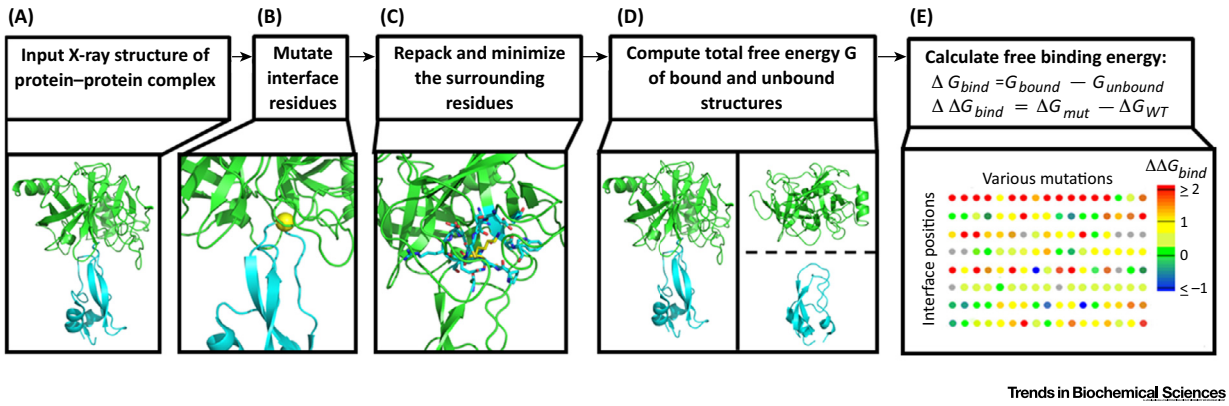

Figure I. Flowchart Illustrating the Main Steps of Computational Protein Design. (A) An X-ray structure of the protein–protein complex is used as input (Protein Data Bank: 1GVL). The interface residues are individually mutated (B) and the side chains surrounding the mutated position are repacked (C). The total energy of the bound and the unbound states is computed using an energy function (D). Finally, the change in free energy of binding due to all possible mutations is calculated and visualized (E).

Several studies have utilized slightly different CPD approaches to simulate observed sequence profiles obtained in PD binding selections [67–70]. In all the explored protocols, an ensemble of backbone structures was first generated, either using computational approaches [71] or taken from an NMR structure. The backbone ensemble was used to design multiple amino acid sequences compatible with binding, and sequence profiles were then constructed from these designed sequences. The above studies revealed that CPD calculations faithfully reproduce the most prominent features of the experimental sequencing profiles. Furthermore, they could explain certain selection features by focusing on specific intermolecular interactions. For example, a Thr at position –2 of peptides that bind to PDZ domains is highly conserved in all PD selections, since Thr forms a hydrogen bond to a His on the PDZ domain; removal of this His results in a loss of Thr conservation [70]. All the above CPD studies reproduced experimental results more accurately if backbone flexibility was incorporated and intermolecular interactions were emphasized in the scoring function. Minor disagreements between computational predictions and experimental results could be due to limitations of CPD methods (e.g., inaccuracy in

the energy function and limited sampling) as well as due to a bias in the PD selection towards highly expressed and stable clones.

CPD methods are limited in the degree of conformational changes they can sample and, thus, MD simulations are utilized when modeling of medium- to large-scale conformational changes is necessary. For example, Murciano-Calles *et al.* used short MD simulations to understand how the same residues in the vicinity of the C-terminal ligand residue in two different PDZ domains gave rise to different peptide specificity profiles observed from PD results. These differences were attributed to conformational changes caused by more distant mutations in the adjacent loop of the PDZ domains [70].

In a recent study, Ratnikov *et al.* used PD to design proteolytic substrates of eight homologous matrix metalloproteinases (MMPs) and measured catalytic parameters for more than 10 000 MMP–substrate pairs [72]. Structural modeling of MMP–substrate interactions together with analysis of sequential differences between different MMPs revealed a remarkable correlation between the MMP substrate preference and the sequence identity of 50–57 discontinuous residues surrounding the catalytic groove. Furthermore, structure-based modeling allowed the authors to convert one type of MMP into another type by grafting just a few key residues at positions that contacted specificity-determining residues on the substrate.

Overall, computational modeling could be useful in explaining complicated experimental selection results, predicting substrates, and redesigning binding specificities. Yet, it cannot replace experiments that produce highly valuable data on binding affinity preferences.

Concluding Remarks and Future Directions

In this review, we have demonstrated how computational and combinatorial techniques are not mutually exclusive and how their combination can solve PPI engineering problems that would be difficult to solve by either of the approaches alone [22,23]. In the combined methodology, computational modeling is applied to narrow down the choices (for the construction of smaller and focused combinatorial libraries) and *in vitro* evolution methods are used to quickly assay the binding of millions of protein variants and validate computational results [73] (Figure 1, Key Figure). In addition, computational methods can be applied after the experimental selection results are obtained to better understand how each mutation contributes to binding energetics. Furthermore, based on experimental results, predictions can be made for the design of proteins with altered affinity and specificity. Conversely, computational designs can be optimized through combinatorial selections, as was shown for NBD designs. Thus, we foresee that the integration of computational and combinatorial methodologies will become a common approach in future PPI engineering studies.

An example of the future application of combined approaches is the mapping of energetic binding landscapes for various PPIs, thereby providing a better understanding of PPI evolution and facilitating the design of high-affinity and highly specific PPIs. Such binding landscapes can be constructed *in silico* by scanning each binding interface position with all 20 amino acids and determining the change in free energy of binding due to all the single mutations (see Figure I in Box 2) [12,74]. The binding landscape can then be used for the selection of protein mutants with enhanced binding affinity and binding specificity. In addition to predicting single mutations, the generated binding landscapes can also be utilized for the construction of focused combinatorial libraries that contain multiple mutants with desired binding properties.

Similarly, binding landscapes can be generated experimentally [43] by constructing combinatorial libraries that contain all the single mutants for one binding partner, thereby probing all binding interface positions with 20 amino acids. These single-mutant libraries can then be

Outstanding Questions

How can we use large data sets from directed evolution experiments and computational modeling to better understand the principles of binding energetics?

How can we map binding landscapes that explore multiple simultaneous mutations and use these landscapes to study epistasis in binding? NGS of the selected pools of variants with multiple mutations could facilitate the construction of such landscapes. High-throughput structural data are also needed to better understand changes associated with mutations.

How can we quickly and accurately measure the binding affinity and specificity of millions of protein variants and, thus, facilitate PPI design?

How can computational and combinatorial methods contribute to understanding the evolution of natural PPIs and binding specificity?

How can we improve computational methods to design novel high-affinity binders from scratch without the help of directed evolution? Analysis of extensive binding affinity data would assist the optimization of the energy function for PPI design.

Key Figure

Schematic Representation of the Combined Computational-Combinatorial Approach

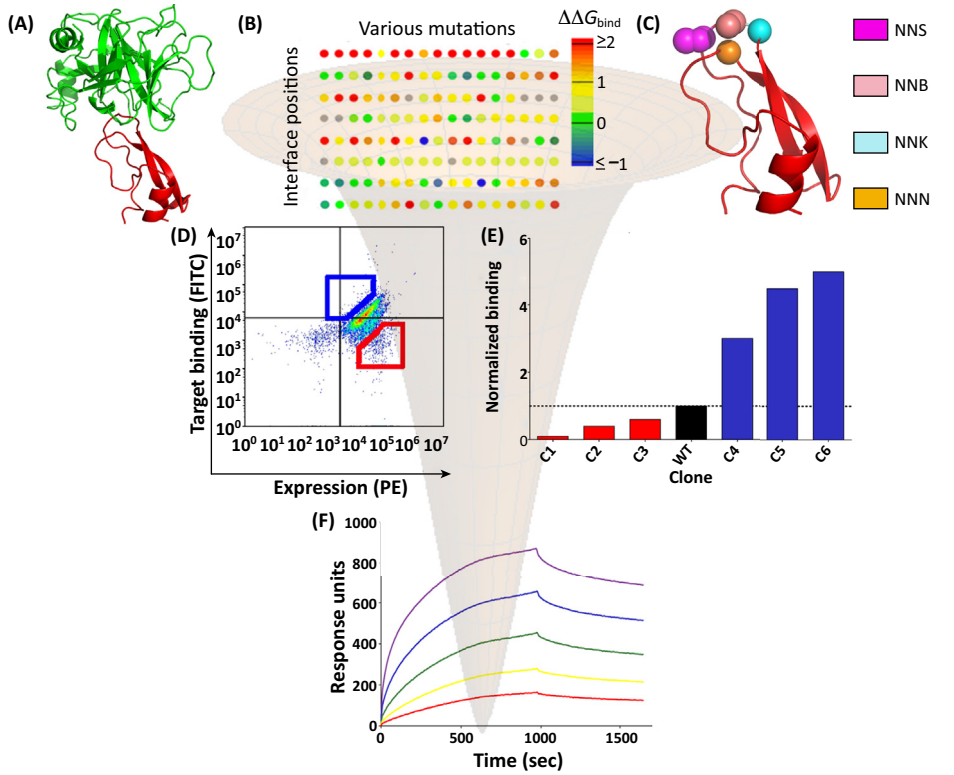

Figure 1. The funnel in the background represents the number of protein sequences explored at each stage. (A) The structure of the protein–protein interaction (PPI) is either taken from the Protein Data Bank (PDB) or modeled using structural information on homologous complexes and unbound components. The binding interface is defined according to the actual or modeled structure (PDB: 1GVL). (B) An *in silico* saturated mutagenesis protocol is performed on all the binding interface residues to compute the changes in the free energy of binding due to all possible mutations. (C) Based on the results obtained in (A) and (B), the positions for the construction of the combinatorial libraries are selected (N-A/G/C/T. S-C/G, K-G/T, B-C/G/T). Further reduction in library diversity is possible by restricting certain positions to the amino acids that produce the most favorable $\Delta\Delta G_{bind}$ according to (B). (D) The resultant library is then sorted according to the levels of expression and binding to the target. It is possible to enrich both the high-affinity (blue gate) and the low-affinity (red gate) clones. (E) After the sorting process, individual variants are isolated, expressed on the yeast surface and tested for binding using fluorescence-activated cell sorting (FACS; red, low-affinity clones; black, wild type; blue, high-affinity clones). (F) The best variants are purified as soluble proteins, and their binding affinity is determined *in vitro*. Alternatively, computational methods can be used to better understand the nature of the mutations selected in the directed evolution experiments.

displayed using any of the display technologies and selected for binding. In YSD, for example, the cells expressing mutants belonging to several affinity groups can be collected and sequenced through Sanger- or NGS-based methods [75] (Figure 2). Several studies have shown that the enrichment of each amino acid in the selected pool vis-à-vis the naive, nonselected library correlates with the energetic contribution of that amino acid to binding [76]. Energetic binding landscapes can be used for the selection of protein mutants with enhanced binding affinity and binding specificity, for optimization of protein therapeutics, and for identification of PPI spots that should be targeted in inhibitor design. Comparison of

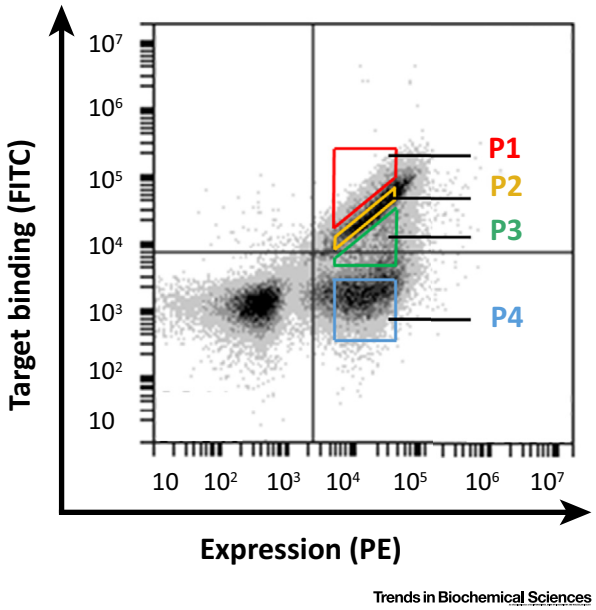

Figure 2. Construction of Experimental Binding Landscapes by Yeast Surface Display (YSD). Combinatorial libraries that contain all the single mutants for one binding partner are displayed on the yeast surface and selected for binding by using fluorescence-activated cell sorting (FACS). Four different categories of mutants are collected: those that bind to the target protein with higher than wild-type affinity (P1), those whose binding affinity is similar to that of the wild type (P2), those whose binding affinity is lower than that of the wild type (P3), and those showing significantly lower binding affinity than the wild type (P4). After selection, multiple clones belonging to each pool are sequenced by next-generation sequencing (NGS) and the sequences are compared with those of the naïve library. The enrichment of each amino acid in the selected pool with respect to the naïve, nonselected, library correlates with the energetic contribution of this amino acid to binding.

computational and experimental binding landscapes could be used to improve computational methods and to reveal some biases associated with the various display technologies (see Outstanding Questions).

To better understand how PPIs are created and destroyed in nature, future studies should be extended to generating binding landscapes for double and higher-order mutants, thereby revealing the nature of molecular epistasis. To conduct such studies, advances in both combinatorial and computational approaches are necessary. We suggest that progress in both computation and experimentation can be highly synergistic, greatly enhancing our understanding of PPI evolution and paving the way towards the all-rational design of novel binding molecules.

Acknowledgments

We are grateful to our students Yuval Zur, Valeria Arkadash, and Tomer Shlamkovich for helpful comments on the manuscript. The laboratory of N.P. is supported by the European Research Council (ERC) and the laboratory of J.M.S. is funded by the Israel Science Foundation (ISF).

References

1. Cho, S. *et al.* (2005) Structural basis of affinity maturation and intramolecular cooperativity in a protein-protein interaction. *Structure* 13, 1775–1787
2. Tonikian, R. *et al.* (2007) Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries. *Nat. Protoc.* 2, 1368–1386
3. Tonikian, R. *et al.* (2008) A specificity map for the PDZ domain family. *PLoS Biol.* 6, e239
4. Miersch, S. *et al.* (2015) Scalable high throughput selection from phage-displayed synthetic antibody libraries. *J. Vis. Exp.* 17, 51492
5. Ernst, A. *et al.* (2012) A strategy for modulation of enzymes in the ubiquitin system. *Science* 339, 590–595
6. Papo, N. *et al.* (2011) Antagonistic VEGF variants engineered to simultaneously bind to and inhibit VEGFR2 and αvβ3 integrin. *Proc. Natl. Acad. Sci. U.S.A.* 108, 14067–14072
7. Wojcik, J. *et al.* (2010) A potent and highly specific FN3 monobody inhibitor of the Abl SH2 domain. *Nat. Struct. Mol. Biol.* 17, 519–527
8. Gilbreth, R.N. *et al.* (2011) Isoform-specific monobody inhibitors of small ubiquitin-related modifiers engineered using structure-guided library design. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7751–7756
9. Sha, F. *et al.* (2013) Dissection of the BCR-ABL signaling network using highly specific monobody inhibitors to the SHP2 SH2 domains. *Proc. Natl. Acad. Sci. U.S.A.* 110, 14924–14929
10. Kortemme, T. *et al.* (2004) Computational redesign of protein-protein interaction specificity. *Nat. Struct. Mol. Biol.* 11, 371–379
11. Shifman, J.M. and Mayo, S.L. (2003) Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc. Natl. Acad. Sci. U.S.A.* 100, 13274–13279
12. Sharabi, O. *et al.* (2013) Computational methods for controlling binding specificity. *Methods Enzym.* 523, 41–59
13. Grigoryan, G. *et al.* (2009) Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* 458, 859–864
14. Yosef, E. *et al.* (2009) Computational design of calmodulin mutants with up to 900-fold increase in binding specificity. *J. Mol. Biol.* 385, 1470–1480

15. Filchtinski, D. *et al.* (2010) What makes Ras an efficient molecular switch: a computational, biophysical, and structural study of Ras-GDP interactions with mutants of Raf. *J. Mol. Biol.* 399, 422–435
16. Sharabi, O. *et al.* (2014) Affinity- and specificity-enhancing mutations are frequent in mulstispecific interaction between MMP14 and its inhibitor TIMP2. *PLoS ONE* 9, e93712
17. Sammond, D.W. *et al.* (2007) Structure-based protocol for identifying mutations that enhance protein-protein binding affinities. *J. Mol. Biol.* 371, 1392–1404
18. Lippow, S.M. *et al.* (2007) Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.* 25, 1171–1176
19. Selzer, T. *et al.* (2000) Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.* 7, 537–541
20. Sharabi, O. *et al.* (2009) Design, expression and characterization of mutants of fasciculin optimized for interaction with its target, acetylcholinesterase. *Protein Eng. Des. Sel.* 22, 641–648
21. Stranges, P.B. *et al.* (2011) Computational design of a symmetric homodimer using beta-strand assembly. *Proc. Natl. Acad. Sci. U. S.A.* 108, 20562–20567
22. Fleishman, S.J. *et al.* (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 332, 816–821
23. Karanicolas, J. *et al.* (2011) A de novo protein binding pair by computational design and directed evolution. *Mol. Cell* 42, 250–260
24. Lim, W.A. (2010) Designing customized cell signalling circuits. *Nat. Rev. Mol. Cell Biol.* 11, 393–403
25. Levin, A.M. *et al.* (2012) Exploiting a natural conformational switch to engineer an interleukin-2 "superkine". *Nature* 484, 529–533
26. Rao, B.M. *et al.* (2004) Interleukin 2 (IL-2) variants engineered for increased IL-2 receptor alpha-subunit affinity exhibit increased potency arising from a cell surface ligand reservoir effect. *Mol. Pharmacol.* 66, 864–869
27. Bahudhanapati, H. *et al.* (2011) Phage display of tissue inhibitor of metalloproteinases-2 (TIMP-2): identification of selective inhibitors of collagenase 1 (metalloproteinase 1 (MMP-1)). *J. Biol. Chem.* 286, 31761–31770
28. Kariolis, M.S. *et al.* (2013) Beyond antibodies: using biological principles to guide the development of next-generation protein therapeutics. *Curr. Opin. Biotechnol.* 24, 1072–1077
29. Chen, T.S. *et al.* (2013) Structure-based redesign of the binding specificity of anti-apoptotic Bcl-xL. *J. Mol. Biol.* 425, 171–185
30. Wang, W. and Saven, J.G. (2002) Designing gene libraries from protein profiles for combinatorial protein experiments. *Nucleic Acids Res.* 30, e120
31. Guntas, G. *et al.* (2015) Engineering an improved light-induced dimer (iLID) for controlling the localization and activity of signaling proteins. *Proc. Natl. Acad. Sci. U.S.A.* 112, 112–117
32. Qiao, C. *et al.* (2013) Affinity maturation of antiHER2 monoclonal antibody MIL5 using an epitope-specific synthetic phage library by computational design. *J. Biomol. Struct. Dyn.* 31, 511–521
33. Smith, S.N. *et al.* (2014) Changing the peptide specificity of a human T-cell receptor by directed evolution. *Nat. Commun.* 5, 5223
34. Dutta, S. *et al.* (2013) Peptide ligands for pro-survival protein Bfl-1 from computationally guided library screening. *ACS Chem. Biol.* 8, 778–788
35. Dutta, S. *et al.* (2015) Potent and specific peptide inhibitors of human pro-survival protein Bcl-xL. *J. Mol. Biol.* 427, 1241–1253
36. Foight, G.W. and Keating, A.E. (2015) Locating herpesvirus Bcl-2 homologs in the specificity landscape of anti-apoptotic Bcl-2 proteins. *J. Mol. Biol.* 427, 2468–2490
37. Dutta, S. *et al.* (2010) Determinants of BH3 binding specificity for Mcl-1 versus Bcl-xL. *J. Mol. Biol.* 398, 747–762
38. Binz, H.K. *et al.* (2005) Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.* 23, 1257–1268
39. Binz, H.K. and Pluckthun, A. (2005) Engineered proteins as specific binding reagents. *Curr. Opin. Biotechnol.* 16, 459–469
40. Gilbreth, R.N. and Koide, S. (2012) Structural insights for engineering binding proteins based on non-antibody scaffolds. *Curr. Opin. Struct. Biol.* 22, 413–420
41. Huang, P.S. *et al.* (2007) A de novo designed protein–protein interface. *Protein Sci.* 16, 2770–2774
42. Clackson, T. and Wells, J.a. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science* 267, 383–386
43. Whitehead, T.A. *et al.* (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* 30, 543–548
44. Azoitei, M.L. *et al.* (2011) Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. *Science* 334, 373–376
45. Azoitei, M.L. *et al.* (2014) Computational design of protein antigens that interact with the CDR H3 loop of HIV broadly neutralizing antibody 2F5. *Proteins Struct. Funct. Bioinform.*
46. Procko, E. *et al.* (2014) A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells. *Cell* 157, 1644–1656
47. Procko, E. *et al.* (2013) Computational design of a protein-based enzyme inhibitor. *J. Mol. Biol.* 425, 3563–3575
48. Stranges, P.B. *et al.* (2011) Computational design of a symmetric homodimer using beta-strand assembly. *Proc. Natl. Acad. Sci. U. S.A.* 108, 20562–20567
49. Mou, Y. *et al.* (2015) Computational design and experimental verification of a symmetric protein homodimer. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10714–10719
50. Chao, G. *et al.* (2004) Fine epitope mapping of anti-epidermal growth factor receptor antibodies through random mutagenesis and yeast surface display. *J. Mol. Biol.* 342, 539–550
51. Rosenfeld, L. *et al.* (2015) Combinatorial and computational approaches to identify interactions of macrophage colony stimulating factor (M-CSF) and its receptor c-fms. *J. Biol. Chem.* 290, 26180–26193
52. Makiya, M. *et al.* (2012) Structural basis of anthrax edema factor neutralization by a neutralizing antibody. *Biochem. Biophys. Res. Commun.* 417, 324–329
53. Mata-Fink, J. *et al.* (2013) Rapid conformational epitope mapping of anti-gp120 antibodies with a designed mutant panel displayed on yeast. *J. Mol. Biol.* 425, 444–456
54. Traxlmayr, M.W. *et al.* (2012) Construction of a stability landscape of the CH3 domain of human IgG1 by combining directed evolution with high throughput sequencing. *J. Mol. Biol.* 423, 397–412
55. Teyra, J. *et al.* (2012) Elucidation of the binding preferences of peptide recognition modules: SH3 and PDZ domains. *FEBS Lett.* 586, 2631–2637
56. Reimand, J. *et al.* (2012) Domain-mediated protein interaction prediction: from genome to network. *FEBS Lett.* 586, 2751–2763
57. Chen, J.R. *et al.* (2008) Predicting PDZ domain-peptide interactions from primary sequences. *Nat. Biotechnol.* 26, 1041–1045
58. Kalyoncu, S. *et al.* (2010) Interaction prediction and classification of PDZ domains. *BMC Bioinform.* 11, 357
59. Shao, X. *et al.* (2011) A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain-peptide interaction from primary sequence. *Bioinformatics* 27, 383–390
60. Zaslavsky, E. *et al.* (2010) Inferring PDZ domain multi-mutant binding preferences from single-mutant data. *PLoS ONE* 5, e12787
61. Ferraro, E. *et al.* (2006) A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity. *Bioinformatics* 22, 2333–2339
62. Tonikian, R. *et al.* (2009) Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol.* 7, e1000218
63. Sanchez, I.E. (2008) Protein folding transition states probed by loop extension. *Protein Sci.* 17, 183–186
64. Fernandez-Ballester, G. *et al.* (2009) Structure-based prediction of the *Saccharomyces cerevisiae* SH3-ligand interactions. *J. Mol. Biol.* 388, 902–916

65. Kaufmann, K. *et al.* (2011) A physical model for PDZ-domain/peptide interactions. *J. Mol. Model.* 17, 315–324
66. Hui, S. *et al.* (2013) Predicting PDZ domain mediated protein interactions from structure. *BMC Bioinform.* 14, 27
67. Smith, C.A. and Kortemme, T. (2011) Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. *PLoS ONE* 6, e20451
68. Humphris, E.L. and Kortemme, T. (2008) Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design. *Structure* 16, 1777–1788
69. Smith, C.A. and Kortemme, T. (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. *J. Mol. Biol.* 402, 460–474
70. Murciano-Calles, J. *et al.* (2014) Alteration of the C-terminal ligand specificity of the erbin PDZ domain by allosteric mutational effects. *J. Mol. Biol.* 426, 1–9
71. Davis, I.W. *et al.* (2006) The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure* 14, 265–274
72. Ratnikov, B.I. *et al.* (2014) Basis for substrate recognition and distinction by matrix metalloproteinases. *Proc. Natl. Acad. Sci. U.S.A.* 111, E4148–E4155
73. Chen, T.S. and Keating, A.E. (2012) Designing specific protein-protein interactions using computation, experimental library screening, or integrated methods. *Protein Sci.* 21, 949–963
74. Aizner, Y. *et al.* (2014) Mapping the binding landscape of a picomolar protein-protein complex through computation and experiment. *Structure* 22, 1–10
75. Reich, L. *et al.* (2015) SORTCERY - a high-throughput method to affinity rank peptide ligands. *J. Mol. Biol.* 427, 2135–2150
76. Pal, G. *et al.* (2006) Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. *J. Biol. Chem.* 281, 22378–22385
77. Barbas, C.F. *et al.* (1991) Assembly of combinatorial antibody libraries on phage surfaces: the gene III site. *Proc. Natl. Acad. Sci. U.S.A.* 88, 7978–7982
78. Boder, E.T. and Wittrup, K.D. (1997) Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* 15, 553–557
79. Gai, S.A. and Wittrup, K.D. (2007) Yeast surface display for protein engineering and characterization. *Curr. Opin. Struct. Biol.* 17, 467–473
80. Bowers, P.M. *et al.* (2011) Coupling mammalian cell surface display with somatic hypermutation for the discovery and maturation of human antibodies. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20455–20460
81. Ho, M. and Pastan, I. (2009) Display and Selection of scFv Antibodies on HEK-293T Cells. *Antib. Phage Disp.* 562, 99–113
82. Beerli, R.R. *et al.* (2008) Isolation of human monoclonal antibodies by mammalian cell display. *Proc. Natl. Acad. Sci. U.S.A.* 105, 14336–14341
83. Taube, R. *et al.* (2008) Lentivirus display: stable expression of human antibodies on the surface of human cells and virus particles. *PLoS ONE* 3, e3181
84. Dahiyat, B.I. and Mayo, S.L. (1997) De novo protein design: fully automated sequence selection. *Science* 278, 82–87
85. Leaver-Fay, A. *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* 487, 545–574
86. Schymkowitz, J. *et al.* (2005) The FoldX web server: an online force field. *Nucleic Acids Res.* 33, W382–W388
87. Dunbrack, R.L. (2002) Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* 12, 431–440
88. Li, Z. *et al.* (2013) Energy functions in de novo protein design: current challenges and future prospects. *Annu. Rev. Biophys.* 42, 315–335
89. Shifman, J.M. and Fromer, M. (2009) Search algorithms. In *Protein Engineering and Design* (Park, S. and Cochran, J., eds), pp. 293–312, Taylor & Francis
90. Li, N. *et al.* (2011) Characterization of PDZ domain-peptide interaction interface based on energetic patterns. *Proteins* 3208–3220
91. Kiss, G. *et al.* (2013) Molecular dynamics simulations for the ranking, evaluation, and refinement of computationally designed proteins. *Methods Enzymol.* 523, 145–170
92. Davey, J.A. and Chica, R.A. (2014) Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins* 82, 771–784