doi:10.1016/j.jmb.2006.01.045 J. Mol. Biol. (2006) 357, 917–927

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Recapitulation and Design of Protein Binding Peptide Structures and Sequences

Vanita D. Sood and David Baker*

Department of Biochemistry Box 357350, University of Washington, Seattle, WA 98195 USA

**Corresponding author*

An important objective of computational protein design is the generation of high affinity peptide inhibitors of protein–peptide interactions, both as a precursor to the development of therapeutics aimed at disrupting disease causing complexes, and as a tool to aid investigators in understanding the role of specific complexes in the cell. We have developed a computational approach to increase the affinity of a protein–peptide complex by designing N or C-terminal extensions which interact with the protein outside the canonical peptide binding pocket. In a first *in silico* test, we show that by simultaneously optimizing the sequence and structure of three to nine residue peptide extensions starting from short (1–6 residue) peptide stubs in the binding pocket of a peptide binding protein, the approach can recover both the conformations and the sequences of known binding peptides. Comparison with phage display and other experimental data suggests that the peptide extension approach recapitulates naturally occurring peptide binding specificity better than fixed backbone design, and that it should be useful for predicting peptide binding specificities from crystal structures. We then experimentally test the approach by designing extensions for p53 and dystroglycan-based peptides predicted to bind with increased affinity to the Mdm2 oncoprotein and to dystrophin, respectively. The measured increases in affinity are modest, revealing some limitations of the method. Based on these *in silico* and experimental results, we discuss future applications of the approach to the prediction and design of protein–peptide interactions.

Keywords: protein–peptide binding; protein design; computational modelling; flexible backbone design; fluorescence polarization

Introduction

The recognition of short, linear peptide sequences by receptor proteins is central to many essential cellular processes, such as signalling, regulation, and the formation of protein networks.[1] The ability to modulate protein interaction networks by designing proteins or peptides that bind specifically and with high affinity to any given target, and either activate or inhibit it, has promise for understanding the roles of each component in a network, and for the future of "individualized medicine", where specific disease causing protein complexes can be inhibited or circumvented.

Previous computational peptide design efforts have primarily used known structures with fixed backbones in attempts to change target specificity. Serrano and co-workers engineered novel PDZ domain–peptide pairs using the backbone from the crystal structure of PSD-95.[2] By designing complementary mutations in the PDZ binding groove and on the peptide ligand, they were able to generate specific pairs with high affinity. The scope of the designs was limited by the inability to model backbone changes as part of the design process: they could engineer a switch from Class

Abbreviations used: NR, nuclear receptor; ER, estrogen receptor; AR, androgen receptor; Dg, dystroglycan; DBR, dystroglycan binding region.

E-mail address of the corresponding author: dabaker@u.washington.edu

I to Class II specificity, but not to Class III, which is structurally more divergent from the first two types of PDZ domains. Shifman & Mayo used fixed backbone design to modulate the specificity of calmodulin peptide interactions.[3,4] Interestingly, specificity was achieved in this case not by increasing the affinity of calmodulin for the peptide for which they optimized the design, but by decreasing the affinity for peptides that were not included in the design calculation.

Fixed backbone design often yields one or a few optimal sequences for the backbone conformation used;[5] because of strict steric constraints, good solutions can be missed when using a fixed backbone. Introducing backbone flexibility can allow the exploration of a larger sequence space and can be critical in successful design and specificity prediction. Wollacott & Desjarlais showed that by introducing backbone perturbations in peptide ligands from crystal structures, it was possible to identify a larger number of interacting sequences for a given target than using only the crystal structure conformation of the peptide.[6] Allowing backbone flexibility is critical in the design of new protein structures, since there may be no sequence that precisely adopts an arbitrary desired target structure; the novel TOP7 protein was designed by simultaneously optimizing sequence and structure during the design process.[7]

In both the Wollacott & Desjarlais work, and the new fold design work from our group, the range of target structures was limited; in the former case, by starting from a specific peptide backbone and carrying out small perturbations, and in the latter, by using distance constraints to specify the overall topology of the designed protein. However, in some applications the goal is to find the best solution to a given problem for all possible structures. Loop modelling[8–15] and *de novo* structure prediction methods[16] have been developed that can generate new conformations, but to date these methods have been carried out in the context of structure prediction where the sequence is fixed, and have not to our knowledge been applied to protein design problems.

We explore a new computational method for *de novo* peptide design and prediction of interaction specificity that simultaneously optimizes backbone conformation and amino acid side-chain sequence and conformation. Our method combines features of Rosetta's *de novo* structure prediction and loop modelling protocols[14] with the sequence optimization of RosettaDesign.[7] This is a step forward over previous loop modelling efforts in that both sequence and structure are optimized. We first benchmark the efficacy of this method in recovering known peptide sequences and structures. We then investigate the ability of the method to increase the affinity of a dystroglycan peptide for the dystroglycan binding region (DBR) of dystrophin, and of a p53 peptide for Mdm2, by creating *de novo* extensions for these two peptides.

Results

Recovery of native peptide backbones and sequences

To test the ability of our protocol to predict the structure and sequence of protein binding peptides, we first conducted a benchmark test on five protein–peptide complexes of known structure. The peptide ligands in the structure were trimmed from one end by deleting both the backbone and side-chains of most of the affinity and specificity determining residues, while one or more residues were kept constant as an anchor for the extension. We then tested the ability of our extension protocol to recover the native peptide sequence and structure without using any information from the deleted residues (see Materials and Methods). We were able to recover native peptide structures and sequences for four out of five test cases.

Mdm2 ligand

The crystal structure of the p53 transactivation domain complexed with the N-terminal domain of human Mdm2[17] reveals a binding groove on Mdm2 into which p53 packs deeply (Figure 1(a)). We removed the seven N-terminal residues of the p53 peptide (including the important F3, L6 and W7, whose side-chains are shown in Figure 1(a)), leaving residue eight intact as the anchor residue. We used our protocol to reconstruct the amino acid sequence and conformation of the seven N-terminal residues of p53. After removing models that were predicted to have a higher free energy of binding than the native structure, and those that had an unfavourable score for the peptide extension (according to Rosetta's full-atom scoring potential), 48 models remained, all with the correct α-helical backbone conformation (C^{α}-RMSD 0.13–0.39 Å for residues 2–8 of the peptide). The model with the best predicted binding affinity recovers both the identity and conformation of the three important residues F3, L6 and W7 (Figure 1(a)), demonstrating the ability of our protocol to predict the native sequence and structure of a peptide ligand for a given binding site.

Unlike design using the fixed backbone derived from the crystal structure of Mdm2 with p53, our protocol recovers a variety of sequences that include both the native p53 sequence and phage display selection clones (Figure 1(b)). As discussed by Wollacott & Desjarlais,[6] flexibility in peptide backbone conformation results in a wider sequence profile than does fixed backbone design. For example, neither phage display nor our method shows any sequence bias at the largely solvent-exposed position 2. Fixed backbone design, however, displays a bias for threonine (Figure 1(b)), because the backbone torsion angles constrain the amino acid selected by RosettaDesign. At position 6 of the peptide, no phage display clone recovers the

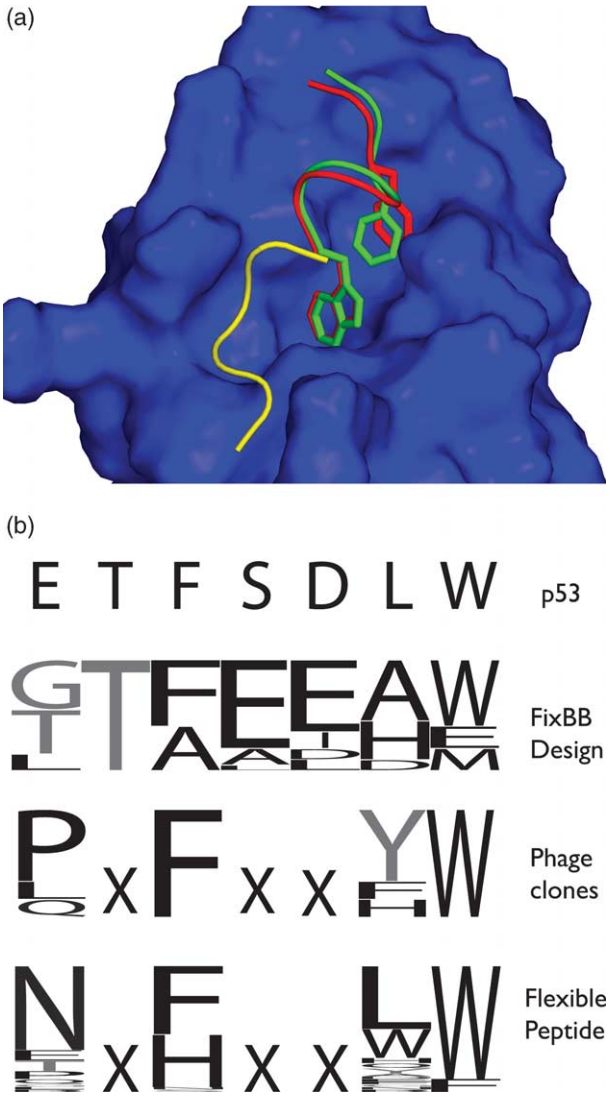

Figure 1. Recovery of p53 bound to Mdm2 using Rosetta peptide extension protocol. (a) The C-terminal six residues of the p53 peptide of 1YCR were kept fixed (yellow) and the N-terminal seven residues were recovered. The native peptide is shown in green and the reconstructed peptide in red. The conserved phenylalanine (F3), leucine (L6) and tryptophan (W7) are accurately recovered in the best ranked models. This and subsequent structural Figures were generated using PyMol (http://www.pymol.org/). (b) Comparison of different methods of sequence recovery. The actual sequence of p53 that was used in the co-crystallization of p53 and Mdm2 is shown on top. The sequence logos[44,45] for fixed backbone design, phage display clones[18] and the Rosetta peptide extension protocol are shown below.

native p53 leucine;[18] instead, only aromatic residues are seen. In contrast, our method recovers both the native leucine, as well as some aromatic residues at this position (Figure 1(b)). These data show that even in helical regions, modelling small changes in backbone tosion angles can have large effects on side-chain packing and improves recapitulation of ligand sequence variability.

Beta-catenin ligand

Beta-catenin is involved in cell adhesion[19] and is a mediator of the Wnt signalling pathway.[20] A number of ligands have been shown to bind beta-catenin, and many of these bind to a groove on the central armadillo repeats of beta-catenin.[21–23] We attempted to recover the C-terminal 11 residues of the beta-catenin ligand ICAT.[21] In this case, the native backbone structure was never sampled at the low resolution extension stage, and thus no models were generated at the design stage that had a predicted binding free energy or packing score as good as that of the native complex. Nevertheless, one of the models with the best predicted binding free energy had side-chains in the same hydrophobic pockets on the surface of beta-catenin that the ICAT peptide exploits for binding (Figure 2). More conformational sampling would be appropriate for cases such as this, where no models with energies as low as the native complex are generated with the standard amount of sampling.

SH2 ligand

The SH2 domain is a modular phospho-tyrosine binding domain found in most eukaryotes.[24] The human genome is predicted to contain 361 SH2 domains[25] and elucidating all ligands for each of these is an important task for understanding signal transduction and protein interaction networks. We used the SH2 domain of $p56^{lck}$ (1LKK[26]) with the bound phospho-tyrosine that is common to all SH2 ligands as the anchor to build three residue peptide extensions. A total of 992 models of diverse sequence were generated. These were pruned to

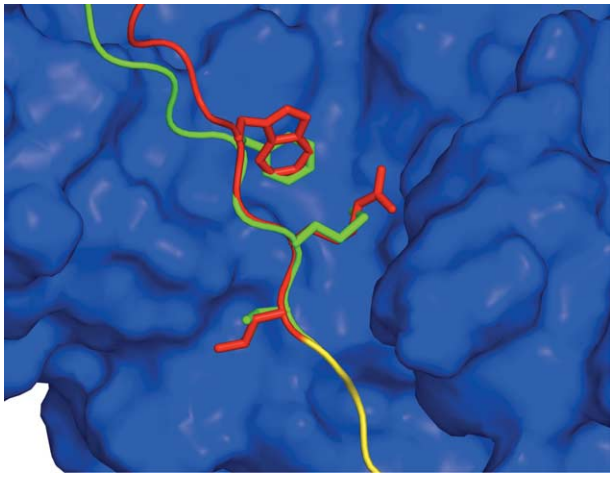

Figure 2. Rosetta peptide extension recovery of ICAT bound to beta-catenin. The 12 C-terminal residues of the peptide ligand of beta-catenin (1M1E) were reconstructed using our method. The fixed residues are shown in yellow, the native residues in green and the model in red. No native-like models were obtained; however, one model was obtained that substituted a tryptophan for an important phenylalanine and a methionine that makes similar hydrophobic contacts as an important arginine.

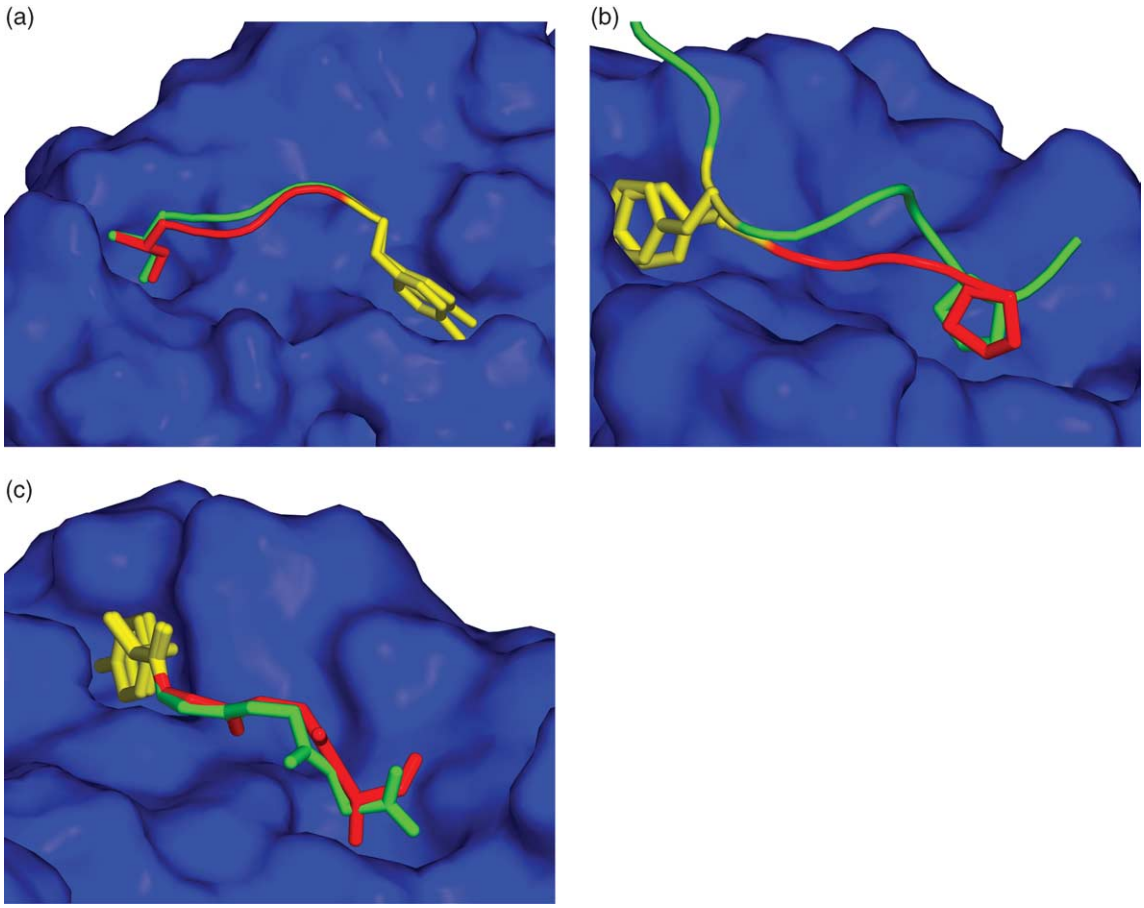

Figure 3. Rosetta peptide extension recovery of SH2 domain ligands. (a) The phospho-tyrosine residue of a high affinity ligand for the SH2 domain of $p56^{lck}$[26] was kept fixed (yellow) and a three residue extension built (red). The native residues are shown in green. The best ranked model superimposes well with the native peptide, and the important isoleucine at position +3 of the peptide is recovered. (b) Superposition of $p56^{lck}$ bound to a proline containing peptide with an extension model.[27] (c) Superposition of $p56^{lck}$ bound to a glycine containing peptide with an extension model.[26]

remove identical sequences and then clustered according to C^{α}-RMSD. The models in each cluster were sorted according to their packing score. The highest ranked model of the largest cluster superimposes very well (C^{α}-RMSD = 0.12 Å) on the native peptide (Figure 3(a)), and the conserved isoleucine at position pY+3 is recovered correctly, again demonstrating the ability of this method to recover both backbone and side-chain conformation of the native peptide ligand.

In contrast to fixed backbone design which uses only the native peptide backbone structure and which generates only one sequence for binding to the SH2 domain (Table 1), a number of different SH2 binding sequences are generated by our peptide extension protocol. Further examination of the best ranked models in clusters 3 and 5 reveals a remarkable correspondence with the available crystal structures of $p56^{lck}$ in complex with similar sequences (Figure 3(b) and (c)). A model from cluster 3 superimposes the pY+3 proline directly onto that in the crystal structure, despite having a different number of residues than 1LCK[27] (Figure 3(b)). The best ranked model of cluster 5 superimposes well on 1LKL,[26] although it does replace the glycine at pY+3 with an alanine (Figure 3(c)). This is not unexpected, as a water molecule that normally occupies space in that pocket is not present in the Rosetta models; indeed, although the absence of water results in a worse

Table 1. Selected SH2 domain ligands and Rosetta extension model sequences

	pY+1	pY+2	pY+3
Lck	E	E	I
FixedBB	E	Y	I
Cluster 1, #1	I	E	I
Crk	K	F	L
Cluster 2, #2	F	F	L
Cluster 7, #1	I	F	L
PLC-γ-C	I	L	P
Cluster 4, #1	I	L	P
GRB-2	Q	N	F
Cluster 19, #1	Q	D	F

Peptide ligands selected from a library[28] are shown along with the closest sequence detected with the Rosetta peptide extension protocol. The smaller cluster numbers refer to larger cluster sizes, and the ranking of a model within a cluster is based upon the packing score. For comparison, the single ligand sequence generated by fixed backbone (FixedBB) design is also given.

packing score, the third ranked model of that same cluster 5 places a glycine in good agreement with 1LKL (data not shown). Table 1 shows some selected ligands for various SH2 domains[28] along with highly ranked model sequences from our protocol; encouragingly, the extension protocol is able to detect a variety of biologically relevant ligands that are not detected using fixed backbone design. Taken together, these results again demonstrate the advantage of flexible backbone peptide design over fixed backbone design in recovering all ligand sequences for a modular binding domain.

Nuclear receptor cofactors

Nuclear receptors (NRs) are a large family of α-helical, steroid activated transcription factors. All NRs utilize a common peptide binding pocket (Figure 4(a) and (b)) to bind cofactors that regulate transcriptional activation by the NRs.[29] Many NRs, including the estrogen receptor (ER) bind cofactors containing a consensus sequence of LxxLL,[30] while the androgen receptor (AR) prefers a cofactors with aromatic residues at the first and last positions of the consensus.[31] We tested the ability of our protocol to predict the specific cofactors for the AR and the ER, starting from co-crystal structures 1XOW[32] and 3ERD,[33] respectively. In each case we deleted all but two N-terminal residues of the cofactor peptide (Figure 4(c) and (d), yellow) and used our protocol to generate C-terminal extensions. The method generated models with the correct alpha-helical backbone for both NRs. For the ER ligand, the C^{α}-RMSD for residues 247–253 of the cofactor models were between 0.09–0.31 Å for all but one model; for the AR ligand, the C^{α}-RMSD

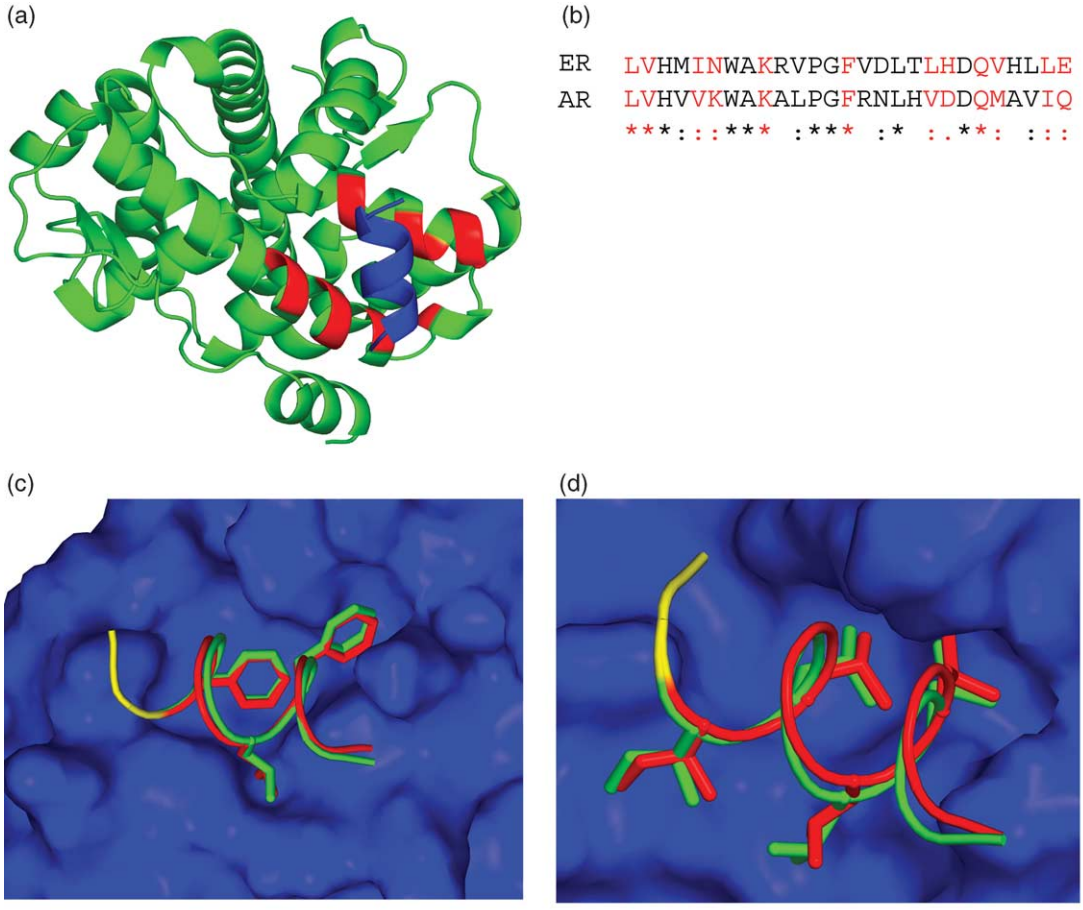

Figure 4. Rosetta peptide extension recovery of cofactor ligands of nuclear receptors. (a) The AR is shown in green, with residues closest to the FxxLF motif of the cofactor highlighted in red. The cofactor is shown in blue. (b) Clustal W[46] alignment of the ER and the AR ligand binding domains. Only a portion of the alignment is shown, with the residues that most closely contact the cofactor (corresponding to the red highlighted residues in (a)) highlighted in red. The asterisks indicate identities and the dots indicate similarities. (c) Two N-terminal residues of the cofactor for the androgen receptor ligand binding domain (1XOW[32]) were kept fixed (yellow) and the C-terminal eight residues were recovered. The native peptide from the crystal structure is shown in green and the reconstructed peptide in red. The two important aromatics and the important hydrophobic residue are recovered. (d) Two N-terminal residues of the cofactor for the estrogen receptor ligand binding domain (3ERD[33]) were kept fixed and the C-terminal nine residues were recovered. The leucine residues of the LxxLL motif are recovered in addition to the beta-branched amino acid at position −1. Colouring as in (c).

for residues 245–252 was between 0.16–0.31 Å for all models. Gratifyingly, the extension protocol correctly distinguished the FxxLF motif preferred by the AR and the LxxLL motif preferred by the ER (Figure 4(c) and (d)); this is especially remarkable considering that many of the receptor residues closest to the FxxLF or LxxLL have considerable sequence similarity (Figure 4(b)).

As with the p53 and SH2 ligand recovery described above, the peptide extension approach better recapitulates cofactor sequence requirements than does fixed backbone design, and is complementary to phage display data and mutagenesis data. The correspondence between the known experimental specificities of the ER and the AR and the results of our *in silico* peptide extension design are described below.

The best ranked models for the ER cofactor faithfully recovered the leucine residues at positions 1 and 5 (Figure 4(d)). Furthermore, the native isoleucine in the −1 position is also recovered in the same conformation (Figure 4(d)), consistent with mutagenesis data that suggest that a beta-branched amino acid is important at this position.[30] The whole range of selected models, however, displays a fairly degenerate sequence profile (data not shown). This is in partial concordance with phage display data,[34] which suggest that modest changes in receptor conformation can affect the optimal cofactor sequence. Together, our data and that of Norris *et al*.[34] suggest a wider range of cofactors for the ER than simply LxxLL containing proteins; it would be interesting to see if the additional ligands suggested by our protocol have biological significance.

In the case of the AR, markedly different sequence profiles are obtained from fixed backbone design and the extension method (Table 2). For example, fixed backbone design tends to converge on a single amino acid at several positions of the peptide, which does not reflect the true diversity of sequences that bind to the AR. Neither fixed backbone design nor phage display suggest that position 8 of the peptide has any influence on binding; our peptide extension protocol, however, strongly favours hydrophobic residues, especially valine, at this position; it would be interesting to see if this residue does indeed contribute to stable binding. At position 4 of the peptide, fixed backbone design is unable to recover the native leucine; in contrast, both our protocol and phage display recover the native leucine as well as other hydrophobic residues. At position 5, fixed backbone design recovers only phenylalanine; our protocol, on the other hand, recovers Phe, Tyr and Trp, in accordance with phage display data. 'These sequence profiles provide further evidence of how a computational search of sequence and structure space can capture differences in specificity even between two relatively similar proteins binding similar alpha-helical ligands.'

Table 2. Sequence profiles for AR cofactors

Residue	Crystal structure	FixedBB design	Flexible peptide	Phage display
1	**F**	F	F,S	F,W
2	Q	K	x	x
3	N	R	x	x
4	**L**	Q,M	L,M,G,A	L,Y,F
5	**F**	F	F,Y,W	F,W,Y
6	Q	L	x	x
7	N	N,S	x	x
8	**V**	D	V,A,I,L,T,H	x

Peptide extension and experimental determination of changes in binding affinity

We chose two protein–peptide complexes to experimentally test the use of the peptide extension protocol to increase the affinity of naturally occurring peptide-protein complexes: the N-terminal domain of Mdm2 complexed with the transactivation peptide of p53 (PDB code 1YCR[17]) and the DBR domain of dystrophin complexed with the C-terminal peptide of dystroglycan (1EG4[35]). We used our protocol to extend the p53 peptide by five amino acid residues and to replace the first four residues of the Dg peptide (which do not interact significantly with the DBR in the crystal structure) by an 11-residue extension. After generating the extended peptides and removing models that did not have a better predicted binding affinity than the native complexes, the remaining models were clustered according to the C^{α}-RMSD and the three models with the best calculated binding energy from each cluster were selected. At this point the models were visually inspected and 15% of the models with sub-optimal packing or hydrogen bonding were discarded. Next, the extension region was subjected to full-atom refinement[14,16] to optimize the backbone for the designed sequence. The refinement was followed by a second design optimization to ensure that the sequence of the peptide extension was still compatible with the refined backbone. Finally, a total of 13 Dg-based and 33 p53-based peptides were selected for *in vitro* characterization, based on their predicted binding affinity for their target proteins and favourable full-atom energy for the extension (see Figure 5).

To determine the affinity of the wild-type Dg–DBR complex and the p53–Mdm2 complexes, tetramethylrhodamine labelled peptides were incubated with increasing amounts of protein and the anisotropy of the complex was measured. Fitting the data yielded a K_d of 0.72(±0.02) μM for the p53–Mdm2 complex, similar to that obtained by isothermal calorimetry using an unlabeled peptide.[17] The K_d obtained for the Dg–DBR complex was 7.6(±1.5) μM, lower than the K_d of 40 μM obtained by isothermal calorimetry using an unlabelled peptide;[35] this discrepancy is most likely due to interaction of the tetramethylrhodamine with the dystrophin DBR.

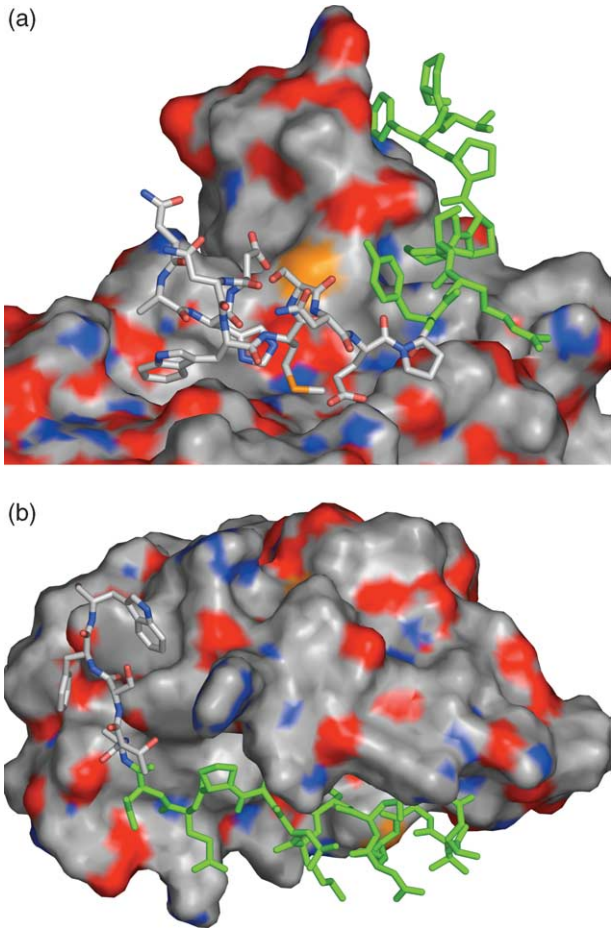

Figure 5. Increasing the affinity of protein–peptide interfaces using the Rosetta peptide extension protocol. (a) Peptide 5. Eleven residues were built onto the nine C-terminal residues of the dystroglycan peptide from 1EG4. The dystrophin DBR protein is shown as a surface, the native peptide residues in green and the extension residues in CPK colouring. (b) Peptide 19. Five residues were added to the C terminus of the p53 peptide from 1YCR. Colouring as in (a).

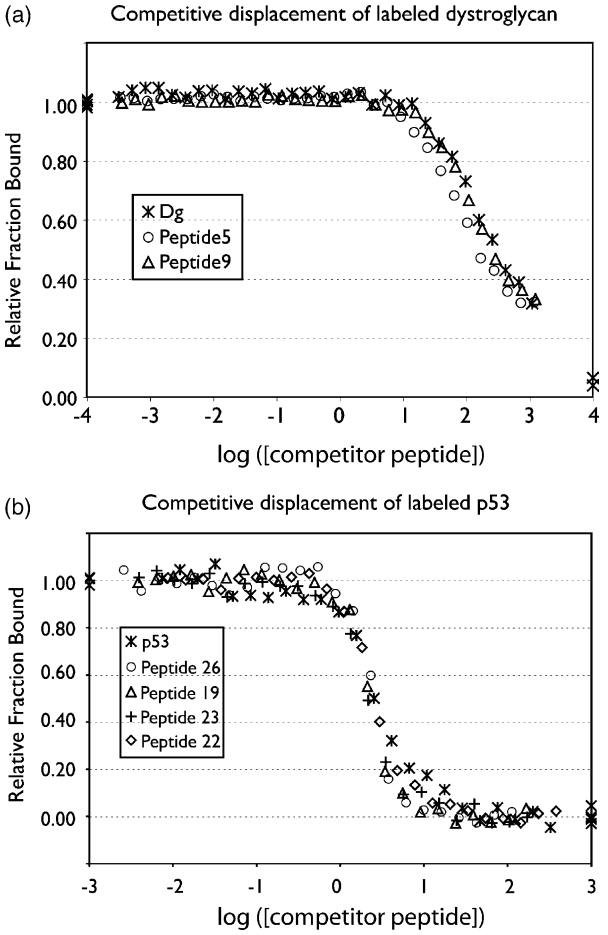

Figure 6. Effect of peptide extensions on affinity of protein–peptide interactions. Fluorescently labelled peptide was incubated with a low concentration of receptor protein and increasing amounts of unlabelled peptide to determine IC50 values for unlabelled peptides. (a) Labelled dystroglycan peptide bound to dystrophin DBR was competitively displaced with unlabelled dystroglycan peptide (Dg) or with extended peptides chosen from the Rosetta peptide extension protocol. The fraction of labelled peptide bound (relative to the fraction bound with no competitor peptide) is plotted against the concentration of competitor. (b) As in (a), except labelled p53 peptide bound to Mdm2 was competitively displaced by unlabelled p53 or with four extended peptides.

For this reason, we use a competitive displacement assay to measure relative affinities of different peptides (see below), avoiding potential complications arising from differential effects of the label on binding affinity in different contexts (see the Fluorescence Polarization Technical Resource Guide; Invitrogen).

Initial competition assays with crude peptides showed that most of the selected peptides did not differ significantly from the native peptide in affinity for the target protein (data not shown). Several of the most promising peptides were chosen for more careful analysis using HPLC purified peptides in competition assays, and the results are shown in Figure 6 and Table 3. The results were modest, with the best p53 extension a 1.3-fold better competitor than the native peptide, and a dystroglycan extension a 2.3-fold better competitor than the native peptide.

Discussion

We have developed a *de novo* peptide extension protocol that incorporates complete backbone flexibility to allow the design of peptide extensions targeted to specific receptor proteins. This method holds promise for the prediction of peptide binding specificity, and can be used to attempt to increase the affinity of peptide inhibitors of protein-protein complexes.

We used our method to extend two peptides in an attempt to increase their affinity for their target proteins, but observed only small effects on the IC50 of extended peptides relative to the native peptides.

Table 3. Relative IC50 values for native and extended peptides

	Sequence	Relative IC50
p53	SQETFSDLWKLLPEN	1
Peptide 26	SQETFSDLWKLLPEN	0.82±0.16
Peptide 19	SQETFSDLWKLLPEN ATSFW	0.78±0.01
Peptide 22	SQETFSDLWKLLPEN LQFGK	0.88±0.02
Peptide 23	SQETFSDLWKLLPEN ALDWG	0.80±0.07
Dg	KNMTP YRSPPPYVP	1
Peptide 5	NAHDNWMSNEP YRSPPPYVP	0.44±0.05
Peptide 9	NEEEQRRPTSV YRSPPPYVP	0.77±0.08

The native p53 and Dg peptide sequences are shown, along with designed peptide extension model sequences. The IC50 values are normalized to those of p53 or Dg, respectively.

Why were the affinity increases so modest? In the case of the dystrophin DBR target protein, the target binding site for the peptide extension is an extremely polar surface without a well defined binding pocket (Figure 5(a)), and all selected models were very polar in nature (Table 2, and data not shown). Consequently, the extension may interact with solvent rather than the dystrophin structure. In the case of Mdm2, the surface patch that we targeted is fairly hydrophobic (Figure 5(b)); however, it is relatively flat and lacks a deep pocket in which the peptide extension can bind and exclude solvent. In both cases, the entropic cost of binding the peptide extension in the conformation modelled may be greater than the enthalpic gains. When we compare the interfaces between the native peptides and the target protein binding pockets to the modelled interfaces between the extensions and the target proteins, a striking difference is the extent of burial of the peptide (Figure 5). This suggests that receptor proteins with well-defined hydrophobic pockets (outside the canonical peptide binding pocket) are likely to be more promising targets for this protocol. The lack of a well-defined binding pocket results in a relatively flat energy landscape, with no deep energy minimum for the "correct" peptide conformation. This is reflected in the lack of similarity between all the selected models, compared to the native recovery cases where there is a well-defined pocket and, presumably, well-defined energy minimum. This difference between the *de novo* extension models and the native recovery models suggests that structural and sequence convergence *in silico* can be used as a criterion to indicate those peptide extensions that are likely to bind well *in vitro*. For future peptide extension experiments, we suggest a combination of careful target choice and selection of models based upon packing and convergence as well as predicted binding free energy.

The accurate recovery of native peptides suggests that our peptide extension protocol is a promising approach to model backbone flexibility and sequence diversity in protein-peptide interface design and prediction. Using a peptide extension protocol that assumes no knowledge of either structure or sequence, we were able to recover low C^{α}-RMSD models with correct side-chain identity and conformation at key interface positions. The method currently requires an anchor residue or short peptide stub; this could be obtained, however, from a docking calculation with a short peptide fragment. In cases where the native peptide ligands have little sequence variation (Mdm2, AR), the protocol generates very few models, all with low C^{α}-RMSDs to the native; additionally, the consensus sequences of the models tend to recapitulate those observed in phage display or site-directed mutagenesis experiments. In cases where the target protein has less stringent ligand sequence requirements (SH2 domain), we recover many more models, representing both high and low affinity ligands; although the high affinity ligands are recovered more accurately, many of our *in silico* generated peptides correspond to *in vitro* validated peptide ligands. In the case of beta-catenin ligands, conformational sampling seems to be the bottleneck. Long peptide extensions with little regular secondary structure coupled with the lack of a well-defined binding pocket result in lack of sampling of the correct backbone structure at the low-resolution level. Successful prediction is thus dependent on adequate conformational sampling at the low-resolution level, and may be improved by applying more computing power to this problem. Taken together, our *in silico* and *in vitro* results suggest that the Rosetta loop extension protocol will be useful for designing or predicting peptide extensions in those cases where there exists a well-defined binding site for the extension.

Materials and Methods

Computational modelling and design of peptide extensions

The protein structure prediction and design program Rosetta[16,36] was used to design peptide extensions to augment the interactions between an existing peptide and its receptor protein. As a first step, an "anchor" residue for the extension was designated; the anchor residue was either the N or C-terminal residue of a peptide in a known protein–peptide complex crystal structure or, in some cases, an internal residue in the peptide. For native recapitulation experiments, the anchor residue was chosen such that most or all specificity and affinity-determining residues were eliminated. The exception was the SH2 ligand where the phospho-tyrosine, which contributes greatly to binding affinity, was retained as the anchor residue; this is because phosphorylated residues are not currently modelled by Rosetta. All residues preceding or following the anchor residue (depending on whether an N or C-terminal extension was to be built, respectively) were deleted, and no side-chain or backbone information from these deleted residues was used in subsequent steps of *de novo* peptide extension design.

Next, the length of the extension to be designed was decided based on a visual inspection of the receptor protein structure for potential interaction sites for the

extension, and the distance of these sites from the anchor residue. A library of about 500,000 peptide fragments of the appropriate length was then compiled from a non-redundant subset of the Protein Data Bank (PDB).[14] A peptide was chosen randomly from the fragment library and the backbone torsion angles were used to build an extension of the anchor peptide. Initially, the peptide fragment was modelled as a poly-alanine sequence; the native sequence of the fragment was not used here or in subsequent modelling steps. The phi and psi angles of the extension residue that overlaps the anchor residue were minimized using a low-resolution energy function which favours burial of non-polar residues and disfavours steric clashes.[16] In total, 60,000 peptide extensions were created; this large and diverse set of peptide extensions was pruned to eliminate extensions that resulted in steric clashes between backbone atoms and to eliminate expanded structures in which there was little chance of favourable interactions between the peptide extension and the target protein.

In the second step, RosettaDesign[7,37] was used to add side-chains to the peptide extensions. All 20 amino acids were allowed at each residue of the extension, while residues on the target protein in close proximity to the extension, as well as the anchor residue of the peptide, were fixed in sequence but allowed to change conformation (repack) to accommodate the extension. Side-chain conformations were from the Dunbrack backbone dependent rotamer library, with extra χ_1 and χ_2 rotamers;[38] side-chain conformations from the native peptide were not included. A Monte-Carlo search procedure was used to sample all rotamers of all residues at the protein–peptide interface and identify the sequence with the lowest energy according to a potential that includes a Lennard-Jones potential to describe atomic packing interactions, an implicit solvation model,[39] an orientation-dependent hydrogen-bonding potential,[40] statistical terms approximating the backbone-dependent amino acid-type and rotamer probabilities,[5] and an estimate of unfolded reference state energies.[5,41] The full-atom models thus generated were evaluated by calculating the energy of the peptide extension in the context of the complex, as well as the predicted energy of binding of the protein–peptide complex. Those models with negative energies and low predicted binding energies were selected for further analysis.

For those cases where there were more than 50 different models passing the filter, clustering[42] according to the C^{α}-RMSD was carried out and the three models with the best calculated binding energy or the best packing score (described below) from each cluster were compared to the native peptides (for *in silico* recovery experiments) or selected for further optimization (before *in vitro* characterization of peptide extensions). Packing was assessed by comparing the solvent-accessible surface area (SASA) of each residue when calculated with a small probe (0.5 Å) to the average of that value for the same residue type with a similar level of burial in a set of real proteins, and summing over all residues in the design model:

$$\text{packing\_score} = \sum_{\text{residues}} \text{SASA\_0.5(residue)} - \text{SASA\_0.5\_avg[residue\_type, SASA\_1.4 (residue)]}$$

where SASA_0.5 is the SASA calculated with a 0.5 Å probe, SASA_1.4 is the SASA calculated with a 1.4 Å probe and serves as a measure of burial, and SASA_0.5_avg is an average over residues of the same amino acid type with similar SASA_1.4 values in a large set of native proteins. Large accessible areas at 0.5 Å that are not accessible to a 1.4 Å water probe are indicative of poor packing: small voids are present within the interface that cannot be filled by solvent molecules (P. Bradley, personal communication).

Preparation of proteins and peptides

All peptides were obtained from Sigma-Genosys or GeneMed (San Francisco, CA). Peptides corresponding to the sequence of the p53 transactivation peptide[17] and the dystroglycan (Dg) peptide[35] from PDB structures 1YCR and 1EG4, respectively, were labelled on the N terminus with tetramethylrhodamine, purified by high performance liquid chromatography (HPLC) and used in fluorescence anisotropy experiments. The peptides were dissolved directly in phosphate buffered saline (pH 7.4) (PBS). Peptide concentration was determined by absorption of the tetramethylrhodamine at 554 nm. Unlabelled peptides used in competition experiments were dissolved in a small amount of dimethylsulfoxide and then slowly diluted into PBS. Concentrations of unlabelled peptides were determined by UV absorption at 280 nm.

The DBR of human dystrophin[35] was expressed as a glutathione-*S*-transferase (GST)-fusion protein and purified by glutathione affinity chromatography. Five hundred units of thrombin (Amersham) were loaded onto the glutathione column with DBR bound, the column sealed and incubated overnight at 4 °C to cleave the GST from the DBR. The DBR was washed off the column, concentrated and the buffer exchanged during concentration to buffer D (50 mM Mops (pH 6.5), 150 mM NaCl, 400 mM Na_2SO_4, 10 mM DTT).

A GST–Mdm2 fusion protein[43] was expressed and purified by glutathione affinity chromatography. The fusion protein was eluted from the glutathione column with 20 mM glutathione, followed by desalting and buffer exchange into buffer M (50 mM Tris (pH 8), 250 mM NaCl, 1 mM DTT) using a G-25 column.

In vitro measurement of binding affinities

Fluorescence anisotropy experiments were performed at 25 °C using a Wallac 1420 Victor3 (PerkinElmer). 100–200 nM tetramethylrhodamine-labelled peptide was incubated with increasing concentrations of the corresponding protein in buffer D for the DBR–Dg binding and buffer M for the p53–Mdm2 binding, in the presence of 100 μg/ml of BSA. Anisotropy values were measured at an excitation wavelength of 531 nm and an emission wavelength of 595 nm. Binding dissociation constants (K_d) were determined by plotting the anisotropy against the concentration of protein and fitting the data to the equilibrium binding equation:

$$fb = \frac{P + T + K_d - \sqrt{P^2 + T^2 + K_d - 2 \bullet P \bullet T + 2 \bullet T \bullet K_d + 2 \bullet P \bullet K_d}}{2 \bullet P}$$

where P is the total concentration of labelled peptide, T is the total concentration of target protein, fb is the fraction bound of the labelled peptide, and K_d is the apparent dissociation constant for the complex.

The IC50 values for the unlabelled wild-type and extended peptides were determined by competition with the labelled peptide. Labelled peptide, 100–200 nM was incubated with a concentration of protein close to the K_d of the complex and the concentration of unlabelled

inhibitor peptide (wild-type or extended) was titrated. The data were fit to the sigmoidal equation:

$$fb = a + \frac{b}{1 + e^{(c \ln x - d)}}$$

where fb is the fraction bound of the labelled peptide, normalizing to 1 for no competitor, a is the anisotropy in the absence of competitor and b is the anisotropy change over the course of the titration, c is a cooperativity coefficient, and d is the natural logarithm of the IC50. The competition of labelled dystroglycan by designed peptides yielded values of c close to 1, as expected. The competition of labelled p53 by unlabelled p53 fit with a value of $c = 1.5$, and the designed peptides yielded values of c between 2.3 and 3.6, suggesting some cooperativity in the competition of labelled p53 by the designed peptides. As the physical basis for this cooperativity increase is unclear, the IC50 values of the designed p53 peptides should be taken as approximate values.

Acknowledgements

We thank Philip Bradley for help modifying Rosetta source code and Dylan Chivian for help with perl scripting. We thank members of the Baker laboratory for comments on the manuscript and K. Laidig for system administration. We thank Michael Eck and Florence Poy for the GST-DBR clone and advice on the purification of DBR. V.D.S. was supported by a fellowship from the Canadian Institutes of Health Research. This work was supported by grants from the National Institutes of Health and from the Department of Defense (CDMRP CM030097 and PC040879 to D.B.). D.B. is also supported by the Howard Hughes Medical Institute.

References

1. Neduva, V. & Russell, R. B. (2005). Linear motifs: evolutionary interaction switches. *FEBS Letters*, **579**, 3342–3345.
2. Reina, J., Lacroix, E., Hobson, S. D., Fernandez-Ballester, G., Rybin, V., Schwab, M. S. *et al*. (2002). Computer-aided design of a PDZ domain to recognize new target sequences. *Nature Struct. Biol.* **9**, 621–627.
3. Shifman, J. M. & Mayo, S. L. (2003). Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc. Natl Acad. Sci. USA*, **100**, 13274–13279.
4. Shifman, J. M. & Mayo, S. L. (2002). Modulating calmodulin binding specificity through computational protein design. *J. Mol. Biol.* **323**, 417–423.
5. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
6. Wollacott, A. M. & Desjarlais, J. R. (2001). Virtual interaction profiles of proteins. *J. Mol. Biol.* **313**, 317–342.
7. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
8. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.
9. Al-Lazikani, B., Jung, J., Xiang, Z. & Honig, B. (2001). Protein structure prediction. *Curr. Opin. Chem. Biol.* **5**, 51–56.
10. Bruccoleri, R. E., Haber, E. & Novotny, J. (1988). Structure of antibody hypervariable loops reproduced by a conformational search algorithm. *Nature*, **335**, 564–568.
11. Deane, C. M. & Blundell, T. L. (2000). A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins: Struct. Funct. Genet.* **40**, 135–144.
12. Zhou, Y. & Abagyan, R. (1998). How and why phosphotyrosine-containing peptides bind to the SH2 and PTB domains. *Fold. Des.* **3**, 513–522.
13. Deane, C. M. & Blundell, T. L. (2001). CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.* **10**, 599–612.
14. Rohl, C. A., Strauss, C. E., Chivian, D. & Baker, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins: Struct. Funct. Genet.* **55**, 656–677.
15. van Vlijmen, H. W. & Karplus, M. (1997). PDB-based protein loop prediction: parameters for selection and methods for optimization. *J. Mol. Biol.* **267**, 975–1001.
16. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93.
17. Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J. & Pavletich, N. P. (1996). Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science*, **274**, 948–953.
18. Bottger, V., Bottger, A., Howard, S. F., Picksley, S. M., Chene, P., Garcia-Echeverria, C. *et al*. (1996). Identification of novel mdm2 binding peptides by phage display. *Oncogene*, **13**, 2141–2147.
19. Provost, E. & Rimm, D. L. (1999). Controversies at the cytoplasmic face of the cadherin-based adhesion complex. *Curr. Opin. Cell Biol.* **11**, 567–572.
20. Moon, R. T. & Kimelman, D. (1998). From cortical rotation to organizer gene expression: toward a molecular explanation of axis specification in *Xenopus*. *Bioessays*, **20**, 536–545.
21. Daniels, D. L. & Weis, W. I. (2002). ICAT inhibits beta-catenin binding to Tcf/Lef-family transcription factors and the general coactivator p300 using independent structural modules. *Mol. Cell*, **10**, 573–584.
22. Huber, A. H. & Weis, W. I. (2001). The structure of the beta-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by beta-catenin. *Cell*, **105**, 391–402.
23. Graham, T. A., Weaver, C., Mao, F., Kimelman, D. & Xu, W. (2000). Crystal structure of a beta-catenin/Tcf complex. *Cell*, **103**, 885–896.
24. Cohen, G. B., Ren, R. & Baltimore, D. (1995). Modular binding domains in signal transduction proteins. *Cell*, **80**, 237–248.
25. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
26. Tong, L., Warren, T. C., King, J., Betageri, R., Rose, J. & Jakes, S. (1996). Crystal structures of the human

p56lck SH2 domain in complex with two short phosphotyrosyl peptides at 1.0 Å and 1.8 Å resolution. *J. Mol. Biol.* **256**, 601–610.
27. Eck, M. J., Atwell, S. K., Shoelson, S. E. & Harrison, S. C. (1994). Structure of the regulatory domains of the Src-family tyrosine kinase Lck. *Nature*, **368**, 764–769.
28. Songyang, Z., Shoelson, S. E., Chaudhuri, M., Gish, G., Pawson, T., Haser, W. G. *et al.* (1993). SH2 domains recognize specific phosphopeptide sequences. *Cell*, **72**, 767–778.
29. Nettles, K. W. & Greene, G. L. (2005). Ligand control of coregulator recruitment to nuclear receptors. *Annu. Rev. Physiol.* **67**, 309–333.
30. Heery, D. M., Kalkhoven, E., Hoare, S. & Parker, M. G. (1997). A signature motif in transcriptional co-activators mediates binding to nuclear receptors. *Nature*, **387**, 733–736.
31. Hsu, C. L., Chen, Y. L., Yeh, S., Ting, H. J., Hu, Y. C., Lin, H. *et al*. (2003). The use of phage display technique for the isolation of androgen receptor interacting peptides with (F/W)XXL(F/W) and FXXLY new signature motifs. *J. Biol. Chem.* **278**, 23691–23698.
32. He, B., Gampe, R. T., Jr, Kole, A. J., Hnat, A. T., Stanley, T. B., An, G. *et al.* (2004). Structural basis for androgen receptor interdomain and coactivator interactions suggests a transition in nuclear receptor activation function dominance. *Mol. Cell*, **16**, 425–438.
33. Shiau, A. K., Barstad, D., Loria, P. M., Cheng, L., Kushner, P. J., Agard, D. A. & Greene, G. L. (1998). The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, **95**, 927–937.
34. Norris, J. D., Fan, D., Stallcup, M. R. & McDonnell, D. P. (1998). Enhancement of estrogen receptor transcriptional activity by the coactivator GRIP-1 highlights the role of activation function 2 in determining estrogen receptor pharmacology. *J. Biol. Chem.* **273**, 6679–6688.
35. Huang, X., Poy, F., Zhang, R., Joachimiak, A., Sudol, M. & Eck, M. J. (2000). Structure of a WW domain containing fragment of dystrophin in complex with beta-dystroglycan. *Nature Struct. Biol.* **7**, 634–638.
36. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225.
37. Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y. & Baker, D. (2002). Accurate computer-based design of a new backbone conformation in the second turn of protein L. *J. Mol. Biol.* **315**, 471–477.
38. Dunbrack, R. L., Jr & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661–1681.
39. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins: Struct. Funct. Genet.* **35**, 133–152.
40. Kortemme, T., Morozov, A. V. & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. *J. Mol. Biol.* **326**, 1239–1259.
41. Kortemme, T. & Baker, D. (2002). A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
42. Shortle, D., Simons, K. T. & Baker, D. (1998). Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl Acad. Sci. USA*, **95**, 11158–11162.
43. Schon, O., Friedler, A., Bycroft, M., Freund, S. M. & Fersht, A. R. (2002). Molecular mechanism of the interaction between MDM2 and p53. *J. Mol. Biol.* **323**, 491–501.
44. Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18**, 6097–6100.
45. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190.
46. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.

Edited by J. E. Ladbury

(Received 26 October 2005; received in revised form 3 January 2006; accepted 9 January 2006)
Available online 31 January 2006