PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Sub-angstrom modeling of complexes between flexible peptides and globular proteins

Barak Raveh,[1,2†] Nir London,[1†] and Ora Schueler-Furman[1*]

[1] Department of Microbiology and Molecular Genetics, Insitute for Medical Research Israel-Canada, Hadassah Medical School, The Hebrew University, Jerusalem, 91120 Israel

[2] The Blavatnik School of Computer Science, Tel-Aviv University, Ramat Aviv, 69978 Israel

ABSTRACT

A wide range of regulatory processes in the cell are mediated by flexible peptides that fold upon binding to globular proteins. Computational efforts to model these interactions are hindered by the large number of rotatable bonds in flexible peptides relative to typical ligand molecules, and the fact that different peptides assume different backbone conformations within the same binding site. In this study, we present Rosetta FlexPepDock, a novel tool for refining coarse peptide–protein models that allows significant changes in both peptide backbone and side chains. We obtain high resolution models, often of sub-angstrom backbone quality, over an extensive and general benchmark that is based on a large nonredundant dataset of 89 peptide–protein interactions. Importantly, side chains of known binding motifs are modeled particularly well, typically with atomic accuracy. In addition, our protocol has improved modeling quality for the important application of cross docking to PDZ domains. We anticipate that the ability to create high resolution models for a wide range of peptide–protein complexes will have significant impact on structure-based functional characterization, controlled manipulation of peptide interactions, and on peptide-based drug design.

Proteins 2010; 78:2029–2040.

Key words: peptide–protein interactions; peptide docking; linear binding motifs; Rosetta FlexPepDock; short peptides; peptide-mediated interactions.

INTRODUCTION

Interactions that are mediated by flexible peptides play a key role in major cellular processes, including signal transduction, transcription regulation, cell localization, protein degradation, and immune response.[1–3] These interactions were recently estimated to account for 15–40% of protein–protein interactions.[2] Peptides often lack a distinct fold in their unbound state, and go through simultaneous binding and folding upon encountering their target (the receptor), often as part of larger intrinsically unstructured domains (IUPs).[4–6] Due to their cardinal role in regulatory interactions, flexible peptides are in many cases implicated in human disease and cancer.[2,7] Peptides may also serve to block existing protein–protein interactions, and have been suggested as promising leads for drug design.[2,7–11] Structural models of peptide–protein complexes have been used to control the specificity of protein–protein interactions,[10,12] and to derive inhibitory peptides, peptoids, or small organic molecules that mimic the binding modes of peptides.[10,13–15] These challenges require accurate high-resolution models of the binding interface, to bridge the gap between the number of solved structures and the actual number of interactions.

The modeling of a peptide–protein interaction can be roughly divided into several consecutive steps, each representing a smaller subproblem, in line with prevalent approaches for modeling[16] and docking[17] of globular proteins: (1) model the receptor structure (2) predict potential binding sites on the receptor surface (3) model the peptide backbone in the binding site; and finally (4) refine the peptide–protein complex to high resolution. In practice, for many real life peptide docking problems the refining step alone is sufficient. Coarse-grained models can often be obtained from complexes with alternative peptides, unbound structures or homology models, for example in the cases of peptides that bind the MHC, SH3, and PDZ domains,[18–20] where existing structures provide approximate structural information about the receptor and the peptide[18–20] or the location of the

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: The Israel Science Foundation (founded by the Israel Academy of Science and Humanities); Grant number: 306/6; Grant sponsor: German–Israeli Foundation for Scientific Research and Development (Young Scientists Program); Grant number: 2159-1709.3/2006

[†]Barak Raveh and Nir London contributed equally to this work.

*Correspondence to: Ora Schueler-Furman, Department of Microbiology and Molecular Genetics, Institute for Medical Research Israel-Canada, Hadassah Medical School, The Hebrew University, POB 12272, Jerusalem 91120 Israel. E-mail: oraf@ekmd.huji.ac.il.

Received 14 December 2009; Revised 3 February 2010; Accepted 19 February 2010
Published online 18 March 2010 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22716

binding site.[18,21–23] In addition, a wide range of experimental methods such as H/D exchange,[24] combinatorial Alanine scanning by mutational analysis,[25] and NMR shift experiments, can be used to define key interacting residues and create plausible coarse starting models, without having to solve the entire structure in detail. These experimental methods have recently been complemented by computational methods for identifying molecular binding sites for small molecules,[26–28] for generating appropriate coarse models of the interaction *de novo*,[26] and for conformation sampling of unbound peptide backbones, as shown recently by Hu *et al.*[29] Accurate refinement of the resulting coarse starting models is the key for peptide design, specificity prediction, and molecular mimicry.

Nonetheless, refining peptide structures and scoring the correct structures accurately is far from trivial. First, peptides in general contain significantly more rotatable bonds than small molecules,[30] raising the bar for computational efforts to model the interactions accurately and efficiently. Second, slight changes in peptide sequences may result in significant backbone rearrangements of the peptide.[18–20] Finally, the receptor structure may differ between various bound, unbound (apo) or homologue conformations, making it harder to model and score the native structure correctly. Rubinstein and Niv[10] have recently reviewed the state-of-the-art in peptide docking and design. Several studies applied novel and existing ligand-docking protocols to globally dock peptides (e.g., Arun and Gautham[31]; Hetenyi and Spoel[32]), but were successful only when applied to very short peptides with a small number of rotatable bonds.[10] Others have addressed high-resolution peptide docking for specific, well characterized systems, most notably peptide binding to MHC, SH3, and PDZ domains,[19,20,33–36] and recently, the HIV-1 protease[37] and the YopE effector protein.[29] Liu *et al.*[38] performed peptide docking over a dataset of 25 peptide interactions with a large number of rotatable bonds, but using an energy scoring function that was explicitly biased towards the native backbone using a coarse-grained Gö potential. Finally, promising results for high-accuracy modeling of peptide–protein complex structures have been reported on a small set of 15 peptide–protein complexes, by using a molecular dynamics-based approach with a locally adjusted soft-core potential for very efficient sampling of conformations.[39]

In this study, we address the challenge of refining coarse models of peptide–protein complexes into high-resolution models. We implemented our protocol FlexPepDock within the well-established Rosetta modeling framework using its generically calibrated energy function.[40] The protocol incorporates full flexibility and rigid body orientation for the peptide backbone, as well as side chain flexibility for both the peptide and the receptor protein. We assess the modeling quality of FlexPepDock on an extensive and general benchmark of 89 different peptide–protein complexes: for each, we create a very large set of starting conformations that differ from the native conformations to a varying degree, apply FlexPepDock, and evaluate the quality of the final prediction. This allows us to assess the quality of the template that is necessary for successful application of our protocol. We also analyze the performance of the protocol over several practical applications, including unbound (apo) and cross docking to PDZ domains, as well as docking from extended backbone conformations, reaching towards the broader context of *ab initio* peptide docking.

To the best of our knowledge, this is the first non-biased modeling protocol to be assessed over an extensive dataset of peptide interactions, with a large number of rotatable bonds. We show that accurate models, often with sub-angstrom backbone quality, are obtained even from fairly remote starting structures. Importantly, residues of known binding-motifs are modeled with atomic accuracy in the top ranking models. We expect that this ability to create high-resolution models for a wide range of interactions, with full flexibility for the peptide backbone, will promote our understanding of key interactions and form the baseline for rational design of peptides and peptide-mimicking drugs.

RESULTS

Our protocol for high-resolution docking of flexible peptides is outlined in Figure 1. Its main components consist of two alternating modules that optimize the pep-

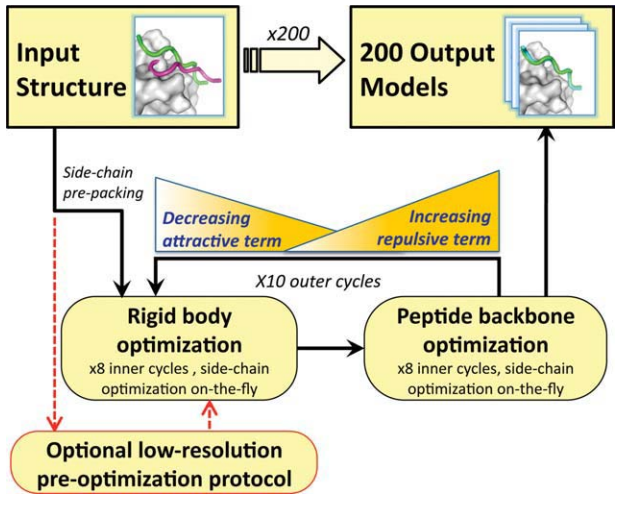

Figure 1

Outline of the Rosetta FlexPepDock refinement protocol. The starting structure is refined by 10 iterative cycles of rigid body and peptide backbone optimizations. These cycles start with a reduced VDW repulsive term and an increased VDW attractive term, which are gradually ramped back to their original value in the energy function. Output models from $n = 200$ independent simulations are ranked using the Rosetta energy score. See Methods for more details. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

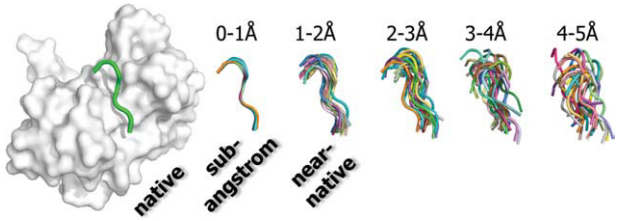

Figure 2

Figurative example of the range of starting structures that were assessed by FlexPepDock in this study. Values are given in peptide bb-RMSD from native. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

tide backbone and rigid body orientation, respectively, using the Monte-Carlo with Minimization approach.[41] Each starting structure is refined in n independent FlexPepDock simulations, producing n different candidate models (unless stated otherwise, we set $n = 200$). The resulting models are ranked based on their Rosetta generic full-atom energy score (See Methods for more details). In the following, we describe the performance of this protocol over a series of increasingly difficult challenges. As for terminology, throughout the Results section, we distinguish RMSD calculated over all peptide backbone atoms (bb-RMSD) and RMSD restricted to peptide backbone atoms in the interface (bb-iRMSD; see Methods for definition of the interface).

The effective range for bound and unbound docking: a benchmark of synthetic perturbations

To define the effective range for sampling of near-native conformations, we created for 89 different peptide–protein interactions a synthetic benchmark of systematically increasing perturbations of the native peptide conformation, both in rigid body and peptide backbone conformation space (see Methods). The resulting perturbed structures were then grouped into bins according to their peptide backbone bb-RMSD values, and used as starting points for the FlexPepDock protocol (see Fig. 2 for an illustration of starting conformations). The simulations allowed us to assess the effective range of our protocol, namely the maximal deviation from the native conformation from which relevant conformations are

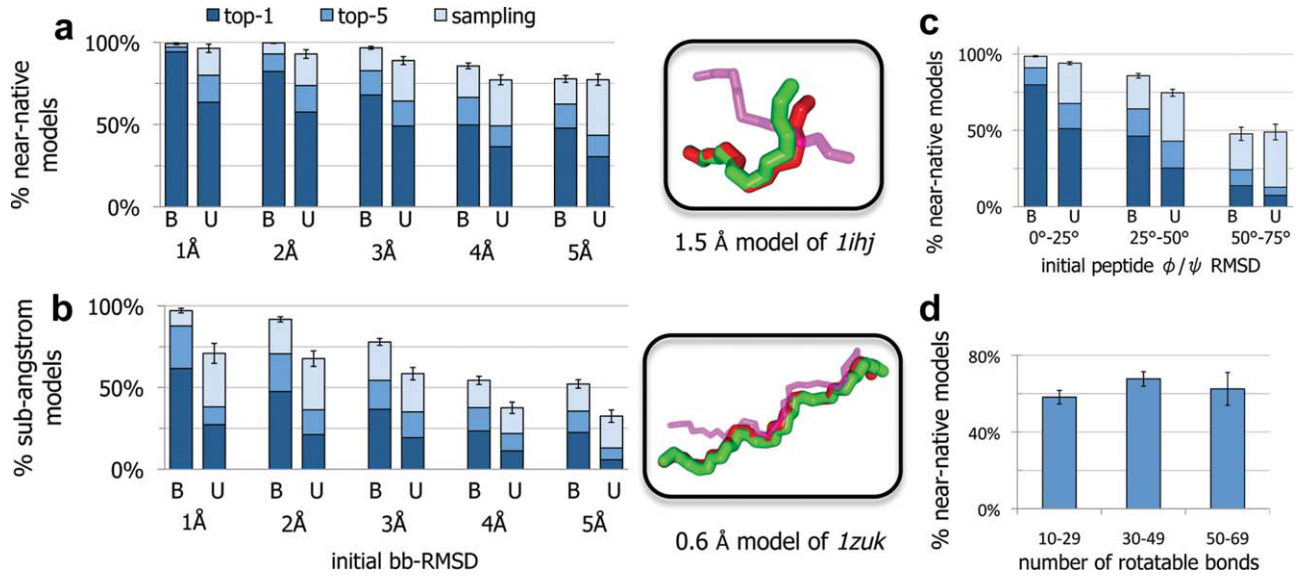

Figure 3

Quality of models created by the FlexPepDock protocol, as a function of the starting structure RMSD from the native peptide backbone conformation. (**a**) Percent of near-native models (<2Å bb-iRMSD, see Figure 2), for different bins of initial bb-RMSD. B = bound docking (holo); U = unbound (apo) docking. The stacked bars show the percent of starting structures in which a good model is the top scoring decoy (dark blue), ranked among the top five scoring decoys (blue), or sampled but not ranked among the top five decoys (light blue). Error bars are standard deviations, calculated assuming an underlying binomial distribution. The illustration shows a structure of the NorpA C-terminal peptide refined to near-native quality (native in green; model in red; taken from a complex with a PDZ domain: pdb-id *1ihj*). The refinement started from an approximate conformation (magenta, thin) that deviates from the native peptide backbone conformation by 5.3Å bb-RMSD and 37° RMSD in φ/ψ torsion space. (**b**) Same as (a), but the y-axis shows the percent of sub-angstrom models (<1Å bb-iRMSD; see Fig. 2). The illustration shows a sub-angstrom model of a peptide bound to an SH3 domain (pdb-id *1jwg*), refined from a starting structure with 3.4Å backbone bb-RMSD and 39° RMSD in φ/ψ torsion space from the native conformation. (**c**) Same as (a), but arranged according to bins of initial φ/ψ torsion-angles RMSD. (**d**) The plot shows the percent of near-native models for different number of rotatable bonds in the peptide (for starting structures in the 5Å bb-RMSD bin). The modeling quality does not depend on the number of rotatable bonds. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

sampled and identified. The model of the receptor protein backbone was taken either from the native bound complex (bound docking), or the apo monomer (unbound docking). We use two different measures for success: A model is defined as near-native, if the peptide residues lie within 2 Å backbone interface-RMSD (bb-iRMSD) of the native peptide conformation. A corresponding bb-iRMSD under 1 Å defines a sub-angstrom model (see Fig. 2 for illustration).

The effective range for near-native modeling is 5.5 Å bb-RMSD

Figure 3a shows the distribution of successful refinements to near-native quality, as a function of the initial distance from the native conformation. For bound docking of starting structures within 5.5 Å bb-RMSD from the native peptide, our standard protocol sampled near-native models in 91% of the cases, and ranked them among the top five models in 78% (in fact, even for starting structures that are as far as 6.5–7.5 Å from the native backbone, a near-native model is sampled in 48% of the cases). In the challenging task of unbound (apo) docking, near-native models were sampled in 85% of the cases and ranked correctly in 59% (for starting structures within 5.5 Å bb-RMSD from the native). Notably, performance is similar for peptides that adopt different types of secondary structures (Supporting information Fig. S1), although ranking improved slightly for peptides with β-strand conformations (probably due to the energetic contribution of hydrogen bonds between paired strands).

The effective range for sub-angstrom modeling is 3.5 Å bb-RMSD

For bound docking of starting structures within 3.5 Å bb-RMSD from the native peptide, sub-angstrom solutions were sampled in 87% of the cases and ranked among the top five models in 68% of the cases. Figure 3(b) shows the success rate of sub-angstrom models for each bin of starting bb-RMSD. Sub-angstrom modeling with unbound (apo) receptors was also successful for a substantial, albeit smaller portion of the starting structures (65% sampled; 37% ranked among the top five models). Note that in some cases, backbone models of sub-angstrom quality were successfully sampled even from very distant starting backbones (>10 Å starting bb-RMSD; Supporting information Fig. S2).

An iterative variant

Our results show that the effective "basin of attraction" for sampling near-native solutions (<2 Å bb-iRMSD) is broader than for sampling sub-angstrom solutions [compare Fig. 3(a) and 3(b)]. Baring this in mind, we tested an iterative variant of our protocol on a set of challenging cases from the bound benchmark that lie in the periphery of our protocol effective sampling range (starting structures with 4.5–5.5 Å bb-RMSD from the native structure). In the iterative variant, we selected the five top scoring solutions of the first iteration, reapplied the standard protocol to each of these solutions, and finally ranked the resulting 1000 decoys from the second iteration. Indeed, when compared to the standard protocol that uses only one iteration, the iterative protocol improves performance by additional 8% (both for sampling and model selection; Supporting information Fig. S3). These results may suggest that the effective sampling range is increased by iterative application of our protocol. Further calibration and testing of this variant over the entire benchmark are required, and are left outside the scope of this manuscript. In the following, we refer to the single-iteration protocol.

Modeling backbone flexibility

The measure of RMSD convolutes the effect of rigid body and torsion-angle deviations. Not surprisingly, the modeling performance of our protocol is high in cases where the RMSD from the native is the result of pure rigid-body perturbations, and it is harder to refine structures where the RMSD is a result of pure backbone torsion perturbations (Supporting information Fig. S4). Still, our protocol has an effective range of up to 50° between φ/ψ torsion angles, covering a 100° by 100° region within the Ramachandran plot, meaning that the initial fold of the peptide can be very different from native, and still be refined correctly [Fig. 3(c)]. Importantly, highly accurate models were obtained even for long peptides with as much as 60 rotatable bonds. In fact, the modeling quality did not depend on the number of rotatable bonds [Fig. 3(d)], and even improved slightly with peptide length (Fig. S5).

Sub-angstrom all-atom modeling

High quality backbone modeling allows rather accurate side-chain placement. Indeed, models with sub-angstrom bb-iRMSD have an average all-atom iRMSD of 1.45 Å (±0.51). Many peptides are known to bind their targets through short, highly conserved binding motifs.[42–44] Therefore, it is plausible that peptide side chains are modeled more accurately within short linear subsegments of the peptide. We first examined our ability to model four consecutive residues (4-mers) at sub-angstrom all-atom RMSD (heavy atoms). At this threshold, almost every atom of the 4-mer must be positioned very accurately. Indeed, for starting structures within 3.5 Å bb-RMSD, our protocol managed to sample all-atom sub-angstrom 4-mers for 82% of the bound cases and 62% of

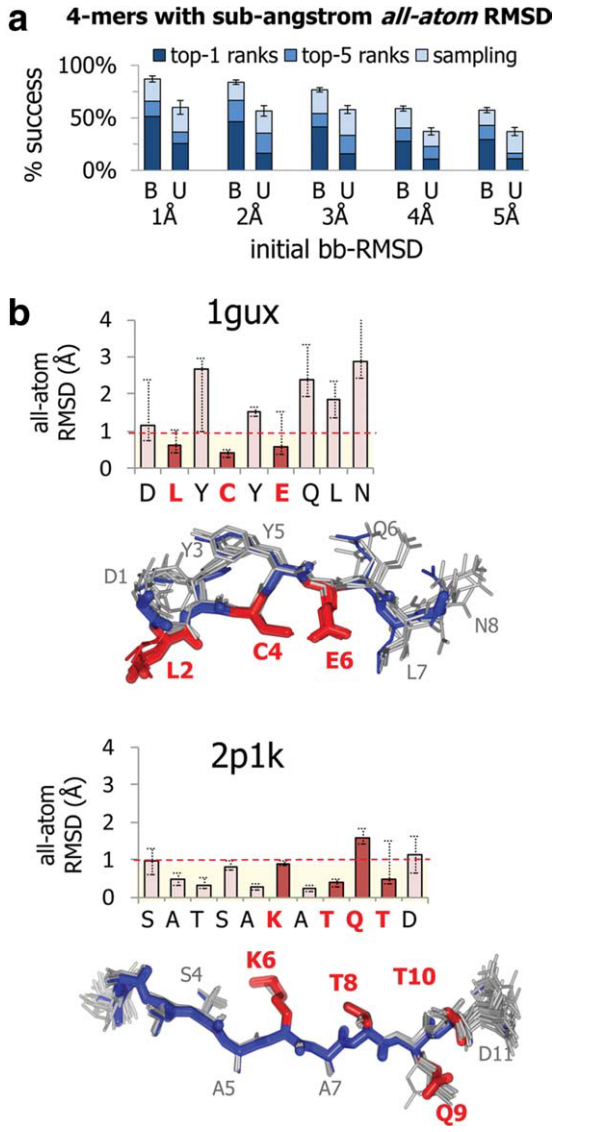

Figure 4

Sub-angstrom all-atom modeling of peptide parts. (**a**) Sub-angstrom all-atom modeling of 4-mers, depicted according to bb-RMSD values of the starting structures. The stacked bars show the percentage of cases in which a model with sub-angstrom all-atom RMSD over four consecutive residues is the top scoring model (dark-blue)/ranked among the top five scoring models (blue)/sampled but not ranked among the top five models (light blue). (**b**) Modeling of motif (red) and nonmotif (pink) residues for pdb-id *1gux* (Papilloma virus E7 oncoprotein) and pdb-id *2p1k* (peptide derived from Drosophila Swallow protein). The plots show for each residue the median all-atom RMSD of the set of top-30 ranking models from the perturbations benchmark that resulted in near-native backbone conformations. Error bars are placed at 0.15 and 0.85 percentiles. The illustrations show the top ranking models (gray) alongside the native peptide (blue). Motif side chains are marked in red (lines for models, sticks for native). For clarity, only the top five scoring models are shown for *1gux*. Note the increased convergence for motif residues (see Supporting information Table S1 and Fig. S6 for the full results on modeling quality of motif residues). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

the unbound cases, and to rank them among the top five models in 62 and 35% of the cases, respectively (see Fig. 4(a) for results according to starting RMSD bins).

All-atom modeling and ranking of known binding motifs

We next examined nine interactions with known binding motifs (listed in Supporting information Table S1), to see whether the Rosetta full-atom energy function favors models where the residues of known motifs are modeled accurately. For each of the nine bound complexes, we examined the top-30 ranking models restricted to near-native backbones (<2 Å bb-iRMSD; out of the thousands of models created in simulations starting from different perturbations of each interaction) and computed the median all-atom RMSD of every residue. Indeed, in top ranking models, the residues of known binding motifs are modeled with sub-angstrom all-atom RMSDs in most of the cases, even when nonmotif residues are not modeled well [Fig. 4(b); Supporting information Figs. S6 and S7]. Thus, the residues that matter most for the binding are favored by the Rosetta energy function, and modeled with superior resolution.

Application: cross docking and unbound docking to PDZ domains with a known anchor

In the practical application of "cross docking", the structure of a peptide–protein complex is modeled based on a solved structure of the receptor bound to a different peptide. This problem is becoming increasingly relevant, as often a diverse set of peptide sequences may bind the same receptor or a close homologue.[18,21–23] The family of PDZ domains is present in well over 500 proteins in the human genome,[19] and mediates a large number of protein–protein interactions by binding the C-terminal peptide of target proteins in a specific manner, using a shared binding mode. Here we compare our generic peptide docking protocol to PDZscheme by Niv and Weinstein[19] a state-of-the-art algorithm for docking peptides to PDZ domains. We assume that the well-conserved C-terminal (P0) anchoring position is known, and set the initial peptide backbone to ideal extended conformation. For this comparison, we used the same benchmark as reported in Niv and Weinstein.[19] (Supporting information Table S2). We generated 2000 models for each interaction. In all seven cases, sub-angstrom peptide backbones (<1 Å bb-iRMSD) were sampled by our algorithm, and in four cases, sub-angstrom models were ranked among the top five solutions (and between 1.2 to 1.5 Å in the other three).

Figure 5 compares the quality of top ranking solutions obtained by Rosetta FlexPepDock and PDZscheme. For consistency with Niv *et al.*,[19] the all-atom RMSD measure is used for this comparison. In six out of seven

cases, our top ranking solutions lie within 2 Å all-atom RMSD of the native structure. In three cases, our results show a significant improvement over PDZscheme, with comparable results in all other cases. In the cross docking to the erbin model pdb-id: 1mfg, the top ranking solution of both methods was incorrect, with an all-atom RMSD of approximately 3.5 Å from the native peptide. Intriguingly, both methods sampled much better solutions in this case (2.0 Å all-atom RMSD from the native in FlexPepDock, 2.3 Å in PDZscheme), but did not rank them first. The poor ranking might be attributed to the relatively large backbone deviation (1.7 Å) between the native receptor (1n7t) and the cross docking receptor (1mfg) (Supporting information Table S2).

Docking from extended backbone conformations with given anchor positions

Motivated by the PDZ example, we were interested in evaluating our protocol in a particularly challenging setting, namely, docking without any prior information about the native peptide backbone conformation. In this benchmark, the peptide was docked starting from an ideal extended backbone conformation ($\pm 135^{\circ}$ for all φ/ψ angles), based on a single anchor residue, and each residue was used as an anchor in turn once.

In this setup, near-native solutions could be sampled in 66% of the 71 nonhelical complexes (31% for sub-angstrom models), and ranked among the top five solutions in 49% of the cases (24% for sub-angstrom models). Representative examples of these simulations are shown in Figure 6. These results indicate that a great portion of nonhelical peptides can be modeled well even in this difficult test, in which no prior information about the peptide backbone is given. Not surprisingly, the 18 helical peptides in our dataset could not be modeled starting from an extended conformation, and very sharp coils or turns were also not modeled accurately. We note that these peptides nevertheless are modeled well in other benchmarks (Supporting information Fig. S1).

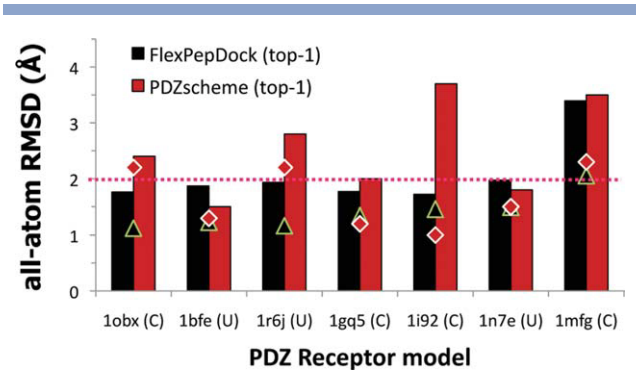

Figure 5

Cross/Unbound docking to PDZ domains. Comparison of sampling and top ranking models between FlexPepDock (black) and PDZscheme (red; as reported in Niv et al..[19] The lowest RMSD model generated by FlexPepDock (PDZscheme) is indicated by a triangle (diamond). The dotted line indicates the 2Å all-atom RMSD threshold. (C) = cross docking. (U) = Unbound docking (For more details on the dataset, see Supporting information Table S2). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

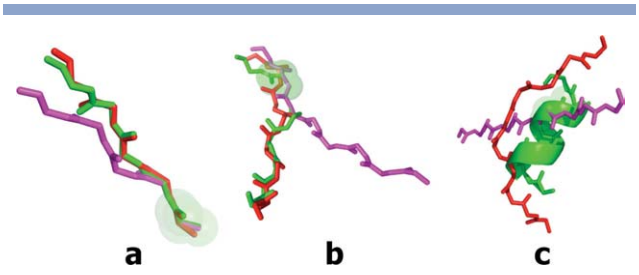

Figure 6

Examples of top scoring models created starting from an extended peptide conformation anchored at a known anchor residue. (**a**) β-strand peptide: Staurosporine, a kinase inhibitor in complex with checkpoint kinase chk1 (pdb-id *1nvr*). The bb-RMSD improved from 2.5 Å in the starting structure to 0.8 Å in the top scoring model. (**b**) coiled peptide: H-77, an inhibitor of aspartic proteinases (pdb-id *1er8*). The bb-RMSD improved from 7.8Å in the starting structure to 1.6Å in the top scoring model. (**c**) α-helical peptides are not modeled well when starting from an extended peptide conformation: An FxxLF motif peptide bound to the androgen receptor ligand binding domain (pdb-id *1t7r*). The bb-RMSD improved from 10.8 Å in the starting structure to 6 Å in the top scoring model, which is not a near-native model. Color code: magenta - starting structure with extended backbone; green - native backbone (sticks) and the anchor residue (spheres); red - top scoring model.

DISCUSSION

In this study, we have put much emphasis on the development of a generally applicable, stable protocol for modeling flexible peptide–protein interactions, by benchmarking its performance on a large and representative set of protein–peptide complex structures. The resulting protocol is able to account for the considerable diversity of peptide conformations within a given binding site. We are often able to produce models with sub-angstrom deviation from the native structure, and importantly, side chains of well-characterized binding motifs are modeled at nearly atomic accuracy. We now relate to some of the insights and future prospects that emerge from this study.

Accurate modeling of known binding motifs

Rosetta FlexPepDock was shown to model side chains of known binding motifs particularly well. Such short binding motifs have been found to play an increasingly important role in mediating protein–protein interactions.[5,42] These motifs are often disordered and gain structure only upon binding to their protein partner.[5,6]

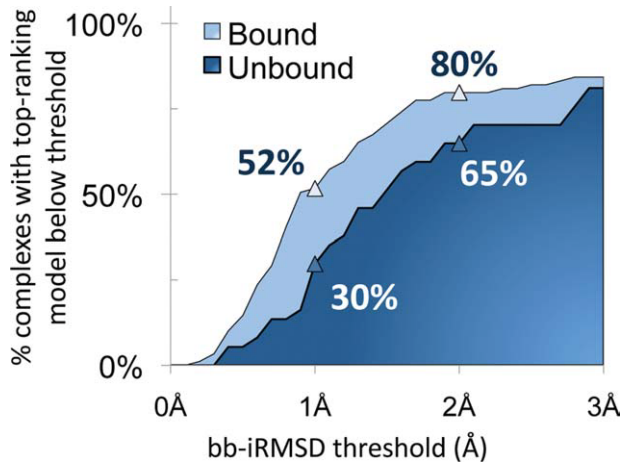

Figure 7

Quality of top ranking models. For each complex, the globally top ranking model was selected from among the thousands of models generated in the perturbations benchmark. The plot shows the percent of complexes (y-axis) whose globally top ranking solution is modeled better than a given RMSD threshold (x-axis). The results are evaluated on all 89 bound and 37 unbound (*apo*) complexes. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

In fact, many of the peptides in our benchmark contain such motifs. Hence the distinction between motif and nonmotif residues proves particularly important for assessing models of peptide interactions. As the energetic contribution of motif residues seems to be the key to high quality sampling and ranking of near-native models, we expect to be able to dock much longer peptides by first docking core binding motifs, and then relaxing the flanks. In addition, we may be able to combine information from top scoring ensembles with methods like computational Alanine scanning (e.g.,[45–47]), to improve the characterization of key binding residues (Supporting information Fig. S7).

Fitness of the energy function for bound and unbound peptide docking

Our protocol builds upon the default full-atom energy function of the well-established Rosetta modeling package,[40] which has not been optimized for peptides in particular. We examined the quality of the top scoring model (over all models in the perturbations benchmark) for each complex in our dataset (Fig. 7). In 80% of the complexes of the bound dataset, the top ranking model falls within the near-native (<2 Å bb-iRMSD) range of solutions, and the same holds for 65% of the complexes even in the unbound dataset. This agrees with our recent finding that peptide binding often induces only minor conformational changes on the protein partner,[48] and that our flexible side-chain scheme accounts for most of them. Nonetheless, the future incorporation of backbone flexibility for the receptor protein may improve the performance even further.

Towards ab initio peptide docking

Many peptide and protein interactions are centered on well conserved anchor positions.[18,21–23] In the extended backbone benchmark (Fig. 6), and to some extent in the example of PDZ docking (Fig. 5), we showed that given a correct anchor position, accurate models are often obtained even without prior information over the peptide backbone (at least for nonhelical peptides, which make up most of the dataset). Notwithstanding the importance of a refinement protocol *per se*, we also aim at the development of a stable and general *ab initio* peptide docking scheme, starting from a given receptor structure and peptide sequence only. Such a scheme can often be guided by previous information about the binding site. We are currently optimizing such an "assisted" *ab initio* protocol to incorporate experimental evidence about the binding site[24,25] as well as the approximate secondary structure of the peptide.[49] In this context, we are also working to increase the extent of backbone flexibility of both the peptide and the receptor, using the extensive set of Rosetta backbone sampling tools, such as backrub-like movements,[50] loop modeling[51,52] and backbone fragments library,[40] which were recently shown to cover the space of peptide backbones by Vanhee *et al.*[53]

As an illustrative example for "assisted" *ab initio* docking, we have attempted to dock *ab initio* the HIV-capsid protein-derived peptide (Sequence: HAGPIA) onto its cellular receptor, the human Proline isomerase cyclophilin A (pdb-id of the bound complex: 1awr). We used the FTmap server by Brenke *et al.*,[27] a computational tool originally designated for predicting potential binding sites of small molecules, to map potential binding sites for the peptide over the bound receptor surface (after removing the peptide). The binding position that was ranked second by the FTMap server roughly correlated with the native position of the central Proline residue. We manually posed an extended form of the peptide ($\pm 135^\circ$ for all φ/ψ angles) onto the binding position predicted by FTmap, and used FlexPepDock to refine the complex. In the third ranking solution provided by FlexPepDock, the starting structure was refined from 4.3 Å bb-RMSD to only 0.8 Å bb-iRMSD from the native, with sub-angstrom all-atom modeling for most interacting residues (Fig. 8). Further calibration is certainly needed before these initial promising results can be applied at large scale, and current computational work in our group is focused on the accurate location of potential peptide binding sites and anchor positions on proteins (along similar lines as recently published prediction schemes[26,27,54]), as well as the automated placement of peptide starting structures in anchor positions, to enable high resolution *ab initio* peptide modeling.

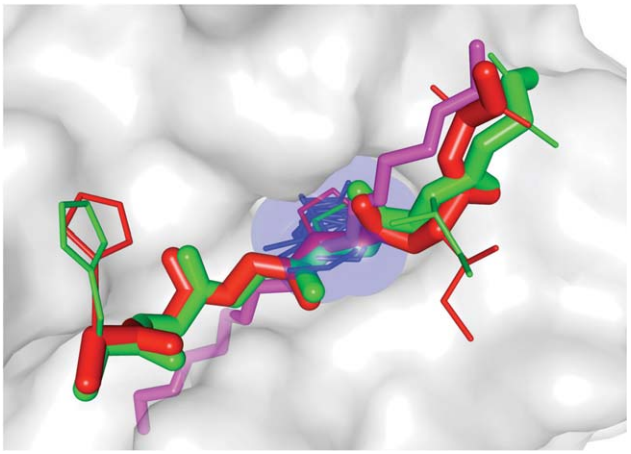

Figure 8

Assisted ab initio peptide docking solution for the HIV-capsid peptide derivative (pdb-id *1awr*). The backbone of the extended starting structure (magenta) was manually placed over the second ranked prediction from the FTMap server[27] (blue spheres/lines). The refined model (red) lies within 0.8Å bb-RMSD from the native peptide (green). All four N-terminal residues, which face the binding pocket, are modeled with sub-angstrom all-atom accuracy (left side of illustration).

Practical application of the method to real-world problems

As we mentioned in the introduction and throughout this article, a myriad of real world problems can be addressed using our method. If a structure is available for the desired interaction or for a close peptide sequence, this structure is an ideal starting structure for the simulation. If no such structure has been solved, but information about the binding site is known (from mutational data, conservation data, bioinformatics predictions, etc.), the peptide can be manually placed according to this data and computational constraints can be set to enforce the fulfillment of the known interactions. In cases where data is available regarding the backbone conformation of the bound peptide (e.g. from CD or NMR studies, or from close homologues sequences binding other proteins), this peptide structure can be used as a starting structure. Alternatively, conformational sampling of the peptide backbone may serve to create starting conformations.[29]

Towards sequence design and specificity prediction with flexible peptide backbones

Design of inhibitory peptides, peptoids, or small organic molecules that mimic the binding modes of peptides are only a few of the possible applications of rational peptide design.[10,13–15] Our ability to produce an ensemble of high quality peptide backbones with well positioned motif residues [Fig. 4(b)] can form the baseline for controlled manipulation of peptide–protein (and protein–protein) interactions and for specificity predictions, while accounting for backbone rearrangements within a given binding site, even in the absence of homologue structures. Our initial results for sequence design based on ensembles generated with FlexPepDock look promising in recapitulating motifs and sequence profiles. Indeed, recent work by Chaudhury and Gray[37] has shed light on the HIV-1 protease specificity mechanism using a similar approach to peptide docking.

CONCLUSIONS

We present a robust protocol for the high resolution modeling of peptide–protein interactions. Our protocol was put under extensive benchmarking, which demonstrated its utility in producing highly accurate results, reaching near-atomic modeling quality of the peptide backbone and key binding residues. We therefore anticipate that this study will significantly contribute to elucidating the pivotal role of peptide-mediated interactions in the living cell. The reported Rosetta FlexPepDock protocol will become publicly available in the next public release of the Rosetta modeling package.

METHODS

Rosetta infrastructure

We implemented our refinement protocols within the Rosetta modeling framework.[40] Rosetta provides well calibrated energy functions, efficient energy calculations and a battery of established conformational sampling protocols. In particular, we use the Rosetta full-atom energy function (Rosetta score12), the coarse-grained energy function that employs a unified spheres side chains model (Rosetta centroid score4), Monte-Carlo sampling with Energy Minimization,[41] and a side chain repacking protocol that relies on the Dunbrack rotamer library.[55]

Input model

The input to the protocol is an initial coarse model of the peptide-protein complex (approximate backbone coordinates for peptide in the receptor binding site). Initial side chain coordinates (such as the crystallographic side chains of an unbound receptor) can be optionally provided as part of the input model. In our analysis, we discarded the input side chains from native complexes (but included side chains of unbound receptors, or of bound receptors that were solved with alternative peptides, as they are also available for practical applications).

Step I: Prepacking the side chains of the input structure

The first step in our protocol involves the "prepacking" of the input structures, to remove internal clashes in the protein monomer and the peptide: Side chain conformations are optimized by determining the best rotamer combination for both the protein and the peptide separately, along similar lines to the RosettaDock protocol.[56–58] If the input structure contains side chain coordinates, they may be considered as additional optional rotamers. We note that in all tests over bound receptor structures, we omitted any information about the input side chains to prevent bias. We refer to the prepacked input structure as the starting structure.

Step II: Model generation

To create a single model, we conduct 10 outer cycles of optimization (Fig. 1). In the first cycle, the weight of the repulsive van der Waals term is reduced to 2% of its normal magnitude, and the attractive van der Waals term is increased by 225%. This allows significant perturbations within the binding pocket, while preventing the peptide and protein to separate during energy minimization. During refinement, the repulsive and attractive terms are gradually ramped back towards their original values (so that in the last cycle the energy function corresponds to the standard Rosetta score). Within each outer cycle, we first optimize the rigid body orientation between the protein and the peptide, and then optimize the peptide backbone for the new orientation. Side chain rotamers are recalculated for the interface on-the fly (Fig. 1).

Optimizing the rigid-body orientation

We apply eight inner cycles of Monte-Carlo search with energy minimization.[41] In each inner cycle, we apply a Gaussian rigid body perturbation (translational magnitude 0.2 Å; rotational magnitude 7°). The rotamers of all interface side chains are then repacked on-the-fly, followed by energy minimization (Davidon-Fletcher-Powell: DFP minimization),[59] with absolute tolerance of one energy unit). The Metropolis criterion is applied after energy minimization only. In the last inner cycle, the rotamer trials with minimization procedure[60] is also applied to optimize all side chains of both the receptor protein and the peptide. We note that energy minimization enables the inclusion of off-rotamer side chain orientations.

Optimizing the peptide backbone

The peptide backbone is optimized by applying the Monte-Carlo with energy minimization procedure, similar to the cycle of rigid-body optimization described above. We perform eight inner cycles of backbone perturbations. The backbone perturbations alternate between the Rosetta small and shear moves[61] with a magnitude of $\pm 6^\circ$.

Preoptimization in low resolution

Due to its sensitivity to mild atomic clashes, the energy landscape of the Rosetta full atom energy function is rugged. The large number of energetic local minima may restrict the sampling process. Similar to other Rosetta protocols,[55] we provide an optional fast, low resolution optimization step before the full atom optimization (Fig. 1). In this step, side chains are represented as spherical centroids of variable size. Similar to the high resolution protocol, the rigid body and the peptide backbone degrees of freedom are optimized alternately for several cycles. The major difference compared to full atom refinement is the use of the centroid-mode energy function (Rosetta score4), which results in a smoother energy landscape. In addition, energy minimization is omitted from Monte-Carlo optimizations steps. Results presented in this study were obtained from 100 runs with the low resolution preoptimizing protocol, and 100 runs without.

Rosetta revision

The protocol and tests described in this manuscript follow the FlexPepDocking protocol as implemented within revision 31671 of the Rosetta repository.

Running time

For each starting structure, we generate within a few minutes 200 candidate models, using an AMD Sun cluster (a single simulation takes 3–4 min on a single CPU).

Benchmark datasets

Datasets of peptide–protein interactions

The bound dataset used in this study includes 89 bound peptide–protein complex structures (Supporting information Table S3). This dataset is a subset of the peptiDB dataset,[48] a culled dataset of 103 high resolution (<2 Å), and nonredundant peptide–protein complexes (<70% sequence identity with respect to the receptor protein), representing a wide range of biological contexts. 14 interactions from peptiDB (chosen at random) have been used for calibration of the protocol, and the reported results regard the remaining 89. The lengths of peptides in the dataset vary between 5 and 15 amino acids, with up to 64 rotatable bonds. The bound dataset includes 18 helical peptides, 19 β-strand peptides, and 52 coiled-coils.

Unbound dataset

In 37 out of the 89 bound complexes, a high resolution (<2 Å) unbound receptor structure in its free form

was available in the PDB. These 37 unbound structures were extracted from peptiDB[48] and superimposed onto their bound counterparts, creating an unbound dataset of peptide–protein interactions.

Synthetic perturbations benchmark

For each native peptide in the dataset, a series of random perturbations was applied both in rigid body and in backbone torsion angle space. The rigid body perturbations were sampled from Gaussians with standard deviations of 0 Å/1 Å/3 Å for rigid body translation, and 0°/15°/30° for rigid body rotation. In addition, each φ/ψ torsion angle was altered by a uniform sample of range ±0, ±25, ±50, ±75, and ±100°. Overall, for each native complex the benchmark contains 45 random start models with increasing RMSD and backbone deviations from the native peptide. These structures were grouped into bins according to their distance from the starting structures, e.g. all structures with 0.5–1.5 Å bb-RMSD deviations belong to the 1 Å RMSD bin, etc. The same perturbations were used for both the bound and unbound protein partners, where applicable, resulting in 4005 synthetic bound structures and 1665 synthetic unbound structures. The entire perturbations benchmark is available upon request.

Dataset for cross docking of PDZ domains

A set of seven PDZ domains in complex with peptides was taken from Niv *et al.*[19] (Supporting information Table S2). For all of these complexes, the protein receptor has been solved with two alternative peptides (cross docking), or once with a peptide and once in the unbound (apo) form (unbound docking). To compare our performance with Niv *et al.*, we reproduced the reported starting conditions. For cross docking, the peptide C-terminal residue was positioned according to its location in the alternative peptide, and the peptide initial φ/ψ angles were set to ideal extended conformation (+135°/−135°). Unbound receptors were aligned to the native receptor, and the structurally well conserved C-terminal residue was positioned according to its native backbone coordinates. Native side chain information was discarded in all cases.

Dataset of extended peptide backbone

A dataset of extended peptide backbones was created from the 89 bound complexes. The peptide's φ/ψ angles were set to an ideal extended conformation (+135°/−135°). Each peptide position was considered in turn as an anchor, and its coordinates were extracted from the native complex, such that a peptide of length *k* resulted in *k* extended starting structures.

Secondary structure definitions for peptides

We used the program Stride[62] to assign secondary structures for each amino acid. Each peptide is labeled as a helix, a strand or a coil, if it contains more than three residues of the equivalent secondary structure, and these labels may overlap. We consider Stride labels H/G/I as helices, E/B/b as strands, and C/T as coils.

Assessment of modeling accuracy

We define solutions as near-native if the interface backbone atoms of the peptide deviate by up to 2 Å bb-iRMSD (backbone interface-RMSD) from the native peptide, and sub-angstrom solutions as those below 1 Å bb-iRMSD. According to Kontoyianni *et al.*,[63] the latter criterion is considered to be stringent (see Figure 2 for illustration of different peptide RMSDs). A peptide residue belongs to the interface if its Cβ atom is less than 8 Å from any Cβ atom on the other chain (Cα atom for Glycines). Unbound receptors were superimposed to native bound receptors using the MaxSub algorithm.[64]

ACKNOWLEDGMENTS

We thank the members of the Furman lab and the Rosetta developer community for stimulating discussions. Protein structure figures were created by PyMOL (http://www.pymol.org).

REFERENCES

1. Pawson T, Nash P. Assembly of cell regulatory systems through protein interaction domains. Science 2003;300:445–452.
2. Petsalaki E, Russell RB. Peptide-mediated interactions in biological systems: new discoveries and applications. Curr Opin Biotechnol 2008;19:344–350.
3. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
4. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 2002;12:54–60.
5. Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, Uversky VN, Dunker AK. Characterization of molecular recognition features. Mo RFs, and their binding partners. J Proteome Res 2007;6:2351–2366.
6. Fuxreiter M, Tompa P, Simon I. Local structural disorder imparts plasticity on linear motifs. Bioinformatics 2007;23:950–956.
7. Naider F, Anglister J. Peptides in the treatment of AIDS. Curr Opin Struct Biol 2009;19:473–482.
8. Hayouka Z, Rosenbluh J, Levin A, Loya S, Lebendiker M, Veprintsev D, Kotler M, Hizi A, Loyter A, Friedler A. Inhibiting HIV-1 integrase by shifting its oligomerization equilibrium. Proc Natl Acad Sci USA 2007;104:8316–8321.
9. Monfregola L, Vitale RM, Amodeo P, De Luca S. A SPR strategy for high-throughput ligand screenings based on synthetic peptides mimicking a selected subdomain of the target protein: a proof of concept on HER2 receptor. Bioorg Med Chem 2009;17:7015–7020.
10. Rubinstein M, Niv MY. Peptidic modulators of protein–protein interactions: progress and challenges in computational design. Biopolymers 2009;91:505–513.
11. Parthasarathi L, Casey F, Stein A, Aloy P, Shields DC. Approved drug mimics of short peptide ligands from protein interaction motifs. J Chem Inf Model 2008;48:1943–1948.

12. Grigoryan G, Reinke AW, Keating AE. Design of protein-interaction specificity gives selective bZIP-binding peptides. Nature 2009;458: 859–864.
13. Simon RJ, Kania RS, Zuckermann RN, Huebner VD, Jewell DA, Banville S, Ng S, Wang L, Rosenberg S, Marlowe CK, Spellmeyer DC, Tans R, Frankel AD, Santi DV, Cohen FE, Bartlett PA. Peptoids: a modular approach to drug discovery. Proc Natl Acad Sci USA 1992;89:9367–9371.
14. Sillerud LO, Larson RS. Design and structure of peptide and peptidomimetic antagonists of protein–protein interaction. Curr Protein Pept Sci 2005;6:151–169.
15. Eichler J. Peptides as protein binding site mimetics. Curr Opin Chem Biol 2008;12:707–713.
16. Zhang Y. Protein structure prediction: when is it useful? Curr Opin Struct Biol 2009;19:145–155.
17. Vajda S, Kozakov D. Convergence and combination of methods in protein-protein docking. Curr Opin Struct Biol 2009;19:164–170.
18. Cesareni G, Panni S, Nardelli G, Castagnoli L. Can we infer peptide recognition specificity mediated by SH3 domains? FEBS Lett 2002; 513:38–44.
19. Niv MY, Weinstein H. A flexible docking procedure for the exploration of peptide binding selectivity to known structures and homology models of PDZ domains. J Am Chem Soc 2005;127:14072–14079.
20. Bordner AJ, Abagyan R. Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. Proteins 2006;63: 512–526.
21. Sudol M. Structure and function of the WW domain. Prog Biophys Mol Biol 1996;65:113–132, 1996.
22. Songyang Z, Fanning AS, Fu C, Xu J, Marfatia SM, Chishti AH, Crompton A, Chan AC, Anderson JM, Cantley LC. Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. Science 1997;275:73–77.
23. Madden DR, Gorga JC, Strominger JL, Wiley DC. The three-dimensional structure of HLA-B27 at 2.1 A resolution suggests a general mechanism for tight peptide binding to MHC. Cell 1992;70:1035–1048.
24. Mandell JG, Falick AM, Komives EA. Identification of protein-protein interfaces by decreased amide proton solvent accessibility. Proc Natl Acad Sci USA 1998;95:14705–14710.
25. Morrison KL, Weiss GA. Combinatorial alanine-scanning. Curr Opin Chem Biol 2001;5:302–307.
26. Petsalaki E, Stark A, Garcia-Urdiales E, Russell RB. Accurate prediction of peptide binding sites on protein surfaces. PLoS Comput Biol 2009;5:e1000335.
27. Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, Mattos C, Vajda S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. Bioinformatics 2009;25:621–627.
28. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 2009;5: e1000585.
29. Hu X, Lee MS, Wallqvist A. Interaction of the disordered Yersinia effector protein YopE with its cognate chaperone SycE. Biochemistry 2009;48:11158–11160.
30. Sousa SF, Fernandes PA, Ramos MJ. Protein-ligand docking: current status and future challenges. Proteins 2006;65:15–26.
31. Arun Prasad P, Gautham N. A new peptide docking strategy using a mean field technique with mutually orthogonal Latin square sampling. J Comput Aided Mol Des 2008;22:815–829.
32. Hetenyi C, van der Spoel D. Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci 2002; 11:1729–1737.
33. Tong JC, Tan TW, Ranganathan S. Modeling the structure of bound peptide ligands to major histocompatibility complex. Protein Sci 2004;13:2523–2532.
34. Fagerberg T, Cerottini JC, Michielin O. Structural prediction of peptides bound to MHC class I. J Mol Biol 2006;356:521–546.
35. Staneva I, Wallin S. All-Atom Monte Carlo Approach to Protein-Peptide Binding. J Mol Biol 2009;393:1118–1128.
36. Fernandez-Ballester G, Beltrao P, Gonzalez JM, Song YH, Wilmanns M, Valencia A, Serrano L. Structure-based prediction of the Saccharomyces cerevisiae SH3-ligand interactions. J Mol Biol 2009;388: 902–916.
37. Chaudhury S, Gray JJ. Identification of structural mechanisms of HIV-1 protease specificity using computational peptide docking: implications for drug resistance. Structure 2009;17:1636–1648.
38. Liu Z, Dominy BN, Shakhnovich EI. Structural mining: self-consistent design on flexible protein-peptide docking and transferable binding affinity potential. J Am Chem Soc 2004;126:8515–8528.
39. Antes I. DynaDock: A new molecular dynamics-based algorithm for protein–peptide docking including receptor flexibility. Proteins 2010;78:1084–1104.
40. Das R, Baker D. Macromolecular modeling with rosetta. Annu Rev Biochem 2008;77:363–382.
41. Li Z, Scheraga HA. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci USA 1987;84:6611–6615.
42. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB. Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 2005;3:e405.
43. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 2003;31:3625–3630.
44. Stein A, Pache RA, Bernado P, Pons M, Aloy P. Dynamic interactions of proteins in complex networks: a more structured view. FEBS J 2009;276:5390–5405.
45. Massova I, Kollman PA. Computational alanine scanning to probe protein–protein interactions: a novel approach to evaluate binding free energies. J Am Chem Soc 1999;121:8133–8143.
46. Kortemme T, Kim DE, Baker D. Computational alanine scanning of protein–protein interfaces. Sci STKE 2004;2004:l2.
47. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–387.
48. London N, Movshovitz-Attias D, Schueler-Furman O. The structural basis of peptide–protein binding strategies. Structure 2009;18: 188–199.
49. Greenfield N, Fasman GD. Computed circular dichroism spectra for the evaluation of protein conformation. Biochemistry 1969;8:4108–4116.
50. Smith CA, Kortemme T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J Mol Biol 2008;380:742–756.
51. Canutescu AA, Dunbrack RL, Jr. Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci 2003;12:963–972.
52. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods 2009;6:551–552.
53. Vanhee P, Stricher F, Baeten L, Verschueren E, Lenaerts T, Serrano L, Rousseau F, Schymkowitz J. Protein–peptide interactions adopt the same structural motifs as monomeric protein folds. Structure 2009;17:1128–1136.
54. Henrich S, Salo-Ahen OM, Huang B, Rippmann FF, Cruciani G, Wade RC. Computational approaches to identifying and characterizing protein binding sites for ligand design. J Mol Recognit 2009;23:209–219.

55. Dunbrack RL, Jr., Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
56. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003;331:281–299.
57. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein–protein docking. Protein Sci 2005;14:1328–1339.
58. Schueler-Furman O, Wang C, Baker D. Progress in protein–protein docking: atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. Proteins 2005;60:187–194.
59. Davidon WC. Variable metric method for minimization. SIAM Journal on Optim 1991;1:1–17.
60. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97:10383–10388.
61. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
62. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
63. Kontoyianni M, McClellan LM, Sokol GS. Evaluation of docking performance: comparative data on docking algorithms. J Med Chem 2004;47:558–565.
64. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 2000;16:776–785.