CHAPTER

Modeling protein-protein interactions with intrinsically disordered proteins

6

Charles W. Christoffer[*], **Daisuke Kihara**[†,1]

Department of Computer Science, Purdue University, West Lafayette, IN, United States[*]

Department of Biological Sciences, Department of Computer Science, Purdue University, West Lafayette, IN, United States[†]

[1]*Corresponding author. E-mail: dkihara@purdue.edu*

6.1 Introduction

Intrinsically disordered proteins (IDPs) are proteins that do not fold into a fixed three-dimensional (3D) structure. Historically, structural biology has assumed that protein structures tend to be stable in the living cell. Indeed, the "thermodynamic hypothesis," perhaps better known as Anfinsen's Dogma, postulates that under physiological conditions, a protein will take on a unique and stable conformation *(1)*. Solid-state structure determination methods such as X-ray crystallography and cryo-electron microscopy have produced a wealth of solved 3D protein structures *(2, 3)*. However, it is often the case that sizable regions of structures from such methods are not resolved at all. While there is plenty of work to be done on solely ordered proteins, IDPs cannot be safely ignored. In fact, it is estimated that 33.0% of eukaryotic proteins contain disordered regions *(4)*. It is also discussed that the fraction of disordered residues in transcription factors, which play essential roles in the gene regulatory network of an organism, has a positive correlation to the organism complexity measured as the number of cell types in that organism *(5)*. Furthermore, IDPs are often hubs in the protein-protein interaction (PPI) network: it is estimated that 15%–45% of all PPIs involve IDPs *(6)*. Additionally, fundamental cellular phenomena such as posttranslational modification and alternative splicing are more common in disordered regions *(7, 8)*, and are important players in allosteric regulation *(9)*. Thus, elucidating IDP interactions is necessary for advancing the understanding of cellular processes in organisms.

While solving such structures experimentally may be difficult, detecting disordered regions from a protein sequence is a well-studied problem. Methods such as DISOPRED *(10)* detect disordered regions based on sequence profiles of proteins with crystal structures containing disordered regions and secondary structure prediction confidence of the target sequence. Precomputed and preannotated databases such as DisProt *(11)*, D2P2 *(12)*, and IDEAL *(13)* contain extensive disordered region prediction information.

Intrinsically Disordered Proteins. https://doi.org/10.1016/B978-0-12-816348-1.00006-5

The dynamics of IDPs have been actively studied. While structure determination is challenging due to the tendency of IDPs to form transient interactions, NMR has elucidated, for example, transient secondary and tertiary structures in the conformational ensembles of IDPs *(14)*. On the computational side, molecular dynamics force fields such as CHARMM36m *(15)*, AWSEM-IDP *(16)*, and ff14IDPSFF *(17)* have been designed specifically for the simulation of IDPs. However, if the goal is to computationally model a complex of an IDP with some binding partner, molecular dynamics is not a computationally feasible route due to the simulation time required. Rigid-body docking and template-based modeling are common approaches to modeling PPI structures, but the flexibility of IDPs precludes the immediate application of rigid-body docking, and template-based modeling does not work without a clear template. Coarse-grained and flexible docking approaches also exist, but they have generally been developed and tested using 2–16 residue peptides versus the 10–70 residue IDPs that commonly participate in PPIs *(18)*, and often require binding site information. These shortcomings motivated our development of IDP-LZerD, the first program capable of docking longer IDPs.

IDP-LZerD *(19)* is based on the dock-and-coalesce model of disordered PPIs *(20)*, wherein a small region of the IDP, possibly having already taken on secondary structure, initially binds to the interaction partner, after which the rest of the bound conformation forms. The first method of this sort, IDP-LzerD, combines fragment picking, our protein-protein docking software, LZerD *(21–23)*, and novel selection heuristics to reliably model PPIs involving IDPs.

6.2 Methods

The basic flow of IDP-LZerD is that small fragments are generated, docked to the receptor structure, combined based on their pairwise geometry, and finally refined using molecular dynamics. The fragment generation and docking correspond to the "dock" part of dock-and-coalesce while the combination and refinement correspond to the "coalesce" part. The flow of IDP-LZerD is illustrated in Fig. 6.1.

6.2.1 Fragment generation

In the first stage of IDP-LZerD, fragments of the IDP structure must be generated. IDP-LZerD uses the Rosetta fragment picker program *(24)* to generate fragments using secondary structure prediction input from PSIPRED *(25)*, JPRED *(26)*, Porter *(27)*, and SSPro *(28)*. The goal of incorporating secondary structure predictions is to bias the secondary structure distribution of the generated fragments because it is known that protein secondary structure can often be predicted well in IDPs *(18)*. The target sequence is divided into nine-residue windows with three-residue overlap, and 30 fragments are generated for each window. Because the fragment picker only outputs C_α atoms, full-atom models for each fragment are built using PULCHRA *(29)* and OSCAR-star *(30)*. The overall accuracy of modeling with IDP-LZerD depends partly on the quality of the secondary structure prediction. Note that although

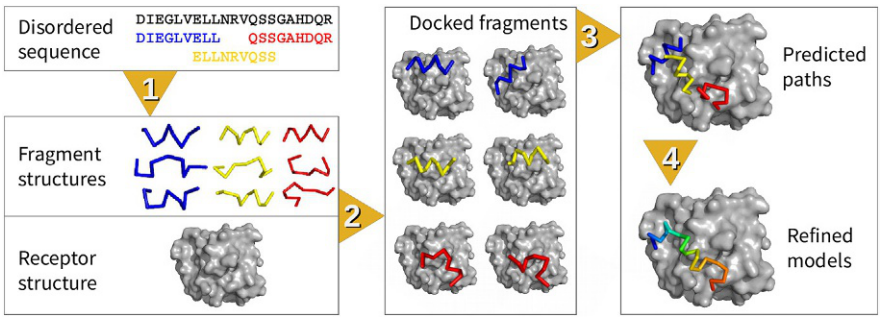

FIG. 6.1

Four steps in the IDP-LZerD algorithm. 1. fragment structure prediction, 2. fragment docking, 3. path assembly, and 4. refinement. Steps 1 and 2 correspond to "dock" and Steps 3 and 4 correspond to "coalesce."

This figure is taken from Fig. 1 of the original IDP-LZerD paper (19) (available under the Creative Commons CC0 public domain dedication).

in testing all methods predicted the same correct secondary structure 57% of the time, at least one method predicted the correct secondary structure 86.1% of the time. Naturally, the accuracy of fragment generation is also important. For the 30 fragments generated by the Rosetta fragment picker for each IDP window, the mean backbone root mean square deviation (RMSD) to the native structure was 1.8 Å for the training set and 1.6 Å for the testing set.

6.2.2 Fragment docking

Each generated fragment is docked to the receptor structure using our protein-protein docking software, LZerD *(21–23)*, which implicitly considers some flexibility through soft-shape representation. The top 50,000 output models are clustered with an RMSD cutoff of 4.0 Å to eliminate redundant models and reduce the problem size, and cluster centers are chosen using the LZerD score, the scoring function to select probable docking models by considering essentially surface shape matching. Subsequently, we used additional scoring functions that consider preferred residue and atom interactions in proteins to further narrow down candidate models. ITScorePro *(31)*, a potential derived from known protein tertiary structures, is used to score the cluster centers, and the top 1000 models are retained, leaving 30,000 models per window. For each window, the 4,500 models with the lowest DI score, defined as the ITScorePro *Z*-score plus the DFIRE *(32)* (another such potential) *Z*-score, are retained. Multiple potentials are used to score models based on the effectiveness of this strategy in other PPI modeling contexts *(23)*. Thus, the individual binding affinities of the fragments are modeled. Using bound receptor structures (those taken from experimental models where the IDP is present), the mean worst minimum L-RMSD among all windows was 3.7 Å for the training set and 4.1 Å for the testing set. Using unbound receptor structures (those taken from experimental models where the IDP is not present), these were 4.4 Å and 4.3 Å, respectively.

6.2.3 Fragment combination

In this step, one docked fragment model from each sequence window is combined to make a full-length docked conformation of the IDP. This combination is called a "path." Distance and angle cutoffs heuristically determined from observations in known protein structures in DisProt *(11)* are applied to prohibit the combination of incompatible fragments, as illustrated in Fig. 6.2. Fragment model pairs that are too close, that is, an atom distance less than 3Å or a fragment midpoint distance less than 6.5Å for neighboring windows and 3.8Å otherwise, are not combined. Pairs impossibly far apart, indicated by a midpoint residue distance of more than 18.5Å times the difference in window number, are also not combined. Also, to ensure that the refinement stage is able to generate realistic bond lengths and angles at fragment boundaries, pairs must satisfy the following criteria: the overlap residue distance (min. 5.2Å, max. 13.6Å), the overlap atom pair distances (max. 6Å for all atoms or 10Å for any atom), and the overlap angle ($\cos\theta \geq 0.1$ so that only smoothly connected turns are included). Paths are constructed starting from the first sequence window: initially, the set of paths is just the fragment models of the first window. Then, each path in the set is combined with all fragment models from the next window that do not violate any of the fragment pair constraints. The set of paths is then clustered with a 4.0Å RMSD cutoff. This cycle of extending and clustering is repeated until all windows have been exhausted.

6.2.4 Path scoring

Now we select paths by a scoring function. Each path is evaluated by its Path Score, which is designed to encode the binding affinity of the IDP to the receptor by

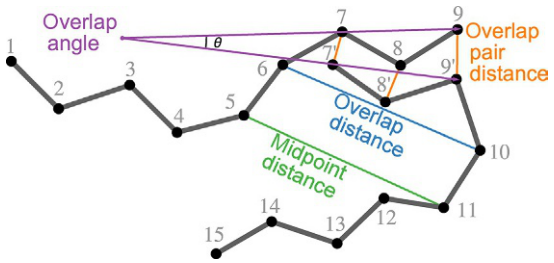

FIG. 6.2

Fragment geometry subject to cutoffs. Midpoint distance: between the C_α atoms of the middle residues of two fragments; overlap distance: between the C_α atoms of the residues before and after the overlapping residues; overlap pair distance: between the corresponding N, C_α, C_β, and C atoms of the three overlapping residues; overlap angle: formed by the vectors from the N atom of the first overlapping residue to the C atom of the third overlapping residue.

This figure is taken from Fig. 8 of the original IDP-LZerD paper (19) (available under the Creative Commons CC0 public domain dedication).

combining various knowledge-based and ensemble properties across all the fragments. The Path Score is defined as

$$S_{path} = w_5\left(w_1 Z\left(S_E\right) + w_2 Z\left(S_O\right) + w_3 Z\left(-S_C\right) + w_4 Z\left(-S_R\right)\right) + \left(1 - w_5\right)\min\left\{Z\left(S_E\right), Z\left(S_O\right), Z\left(-S_C\right), Z\left(-S_R\right)\right\}$$

where the energy score S_E is the mean of the DI scores (as defined in the "Fragment docking" section) of all fragment models in the path, which rewards high-affinity binding sites for each fragment; the overlap score S_O is the mean of the mean square distance between the overlapping residues of consecutive fragments, which rewards more contiguous paths; the cluster size S_C is the number of paths in a cluster, which rewards agreement with conformational consensus; and the receptor score S_R is the sum of N_{op} of all contacting residues of the path, where the N_{op} of a residue is the number of paths contacting that residue (contact is defined as having any heavy atoms within 5.0 Å), which rewards agreement with binding site consensus. The last term of the Path Score is intended to allow paths that score exceptionally well in one of the constituent scores to rise in the ranking. The weights w_i were trained to maximize the minimum recall as detailed in the IDP-LZerD paper. The top 1000 paths by Path Score are retained for refinement.

6.2.5 Refinement

Without refinement, the models produced by IDP-LZerD would not be physical. In fact, they would variously have clashes, broken bonds, and impossible torsion angles. Molecular dynamics refinement is used to turn these models into something physically plausible. The refinement is performed with molecular dynamics simulation using the CHARMM potential *(33)* with the FACTS implicit solvation *(34)*. During the initial energy minimization, all receptor atoms are held fixed. First, the ligand is placed under a harmonic constraint. During energy minimization, progressively weaker ligand constraints are applied. Next, minimization is performed with the constraints placed only on the backbone atoms of the ligand. The final minimization round is minimization with no ligand constraints. Finally, the structure is equilibrated with fixed hydrogen covalent bond lengths, and a harmonic constraint on all C_α atoms.

The refined models are reranked using the Model Score, which is designed to aggregate the rankings from several protein model scoring functions. The Model Score is defined as

$$S_{model} = w_5\left(w_1 Z\left(\text{ITScorePro}\right) + w_2 Z\left(\text{DFIRE}\right) + w_3 Z\left(\text{MolMech}\right) + w_4 Z\left(\text{GOAP}\right)\right) + \left(1 - w_5\right)\min\left\{Z\left(\text{ITScorePro}\right), Z\left(\text{DFIRE}\right), Z\left(\text{MolMech}\right), Z\left(\text{GOAP}\right)\right\}$$

where MolMech is a molecular mechanics score *(35)* and GOAP *(36)* is another statistical potential. The approach of combining several scoring functions has been successful in other PPI structure prediction contexts *(23)*. The weights w_i were trained to minimize the mean rank of the first hit as detailed in the IDP-LZerD paper, where hits are models that are "acceptable" or better according to the CAPRI protein docking evaluation criteria (Table 6.3) *(37)*.

6.3 Performance results summary

Tables 6.1 and 6.2 summarize prediction results on the data set from the original IDP-LZerD paper *(19)*. The various weights required by the IDP-LZerD scoring functions were trained using the 14 complexes detailed in Table 6.1, and the resulting pipeline was tested using the eight complexes detailed in Table 6.2. For both training and testing, a model was considered a hit if it was categorized as "acceptable" or better by the CAPRI criteria *(37)*. The CAPRI criteria (detailed in Table 6.3) are a measure of PPI structure modeling based on f_{nat}, the fraction of native residue contacts present in the model; I-RMSD, the RMSD of the backbone of the native interface residues; and L-RMSD, the RMSD of the ligand (which here is the IDP) to its native structure after the receptor in the model has been superimposed to the native structure. The results of the full modeling procedure, including fragment combination and refinement, are summarized in Tables 6.1 and 6.2.

Overall, IDP-LZerD was fairly successful in building correct models (i.e., acceptable quality models based on the CAPRI criteria and hits) in most of the target complexes. Moreover, the first hit usually has a reasonably accurate interface with f_{nat} much higher than the CAPRI-acceptable cutoff of 0.1. Of the training targets (Table 6.1), IDP-LZerD only failed to generate hits for a single bound case and three unbound cases. Of the testing targets (Table 6.2), hits were generated for all cases except a single unbound case. Although these hits are not always ranked in the top 10, the reliable presence of hits in the output reduces the IDP structure modeling problem to model ranking. Hits were present in the top 10 ranked models for nine bound and six unbound training cases (Table 6.1) as well as two bound and three unbound testing cases. Notable cases include both the bound and unbound cases corresponding to proteinase A with protease A inhibitor 3 in the testing set, where IDP-LZerD produced both the first-ranked hits and top 10-ranked CAPRI-medium models. Another notable case is the unbound case corresponding to YopE chaperone SycE with the outer membrane virulence protein YopE in the testing set. Despite this IDP containing 69 residues, IDP-LZerD generated a second-ranked hit. Overall, IDP-LZerD performed well even as the IDP length increases.

6.4 Examples

In this section, we show several examples of the modeled disordered PPIs. The first three examples of disordered PPI predictions were freshly computed for this chapter while the latter three show models built for the original paper. In the first three examples, secondary structure predictions were taken from SPIDER3 *(38)*, DeepCNF *(39)*, and DNSS *(40)* instead of the secondary structure prediction methods used in the original work. As shown in the figures, for all the cases, IDP-LZerD was able to build docking structures of IDPs in almost the correct conformations.

Table 6.1 Summary of Modeling Performance on the Training Set

	Bound				Unbound		
Protein Names	**PDB ID**	**Len**	**RFH**	f_{nat}	**PDB ID**	**RFH**	f_{nat}
MDM2 with P53	1ycrA	15	1	0.42	1z1mA	6	0.13
DRA/DRB5 with myelin basic protein	1fv1AB	20	6	0.31	4ah2AB	1	0.40
eIF4E with 4E-BP1	1wkwA	20	16	0.3l9	1ipbA	53	0.24
PKA C-alpha with PKI-alpha	2cpkE	20	4	0.56	1j3hA	15	0.17
CREB-binding protein with Transcriptional activator Myb	1sb0A	25	3	0.32	4i9oA	136	0.18
Actin, alpha skeletal muscle with Ciboulot, isoform A	1sqkA	25	14	0.36	1ijjA	9	0.55
Bcl2-L-1 with BAD	2bzwA	27	1	0.49	1pq0A	–	–
DNA-binding protein RAP1 with Regulatory protein SIR3	3owtAB	27	6	0.33	3cz6AB	52	0.13
MAD homolog 2 with Madh-interacting protein	1devA	41	2	0.60	1khxA	16	0.22
p300 HAT with Cbp/p300-interacting transactivator 2	1p4qB	44	5	0.27	1l3eB	3	0.25
Catenin beta-1 with Transcription factor 7-like 2	1jpwA	45	1	0.38	2z6hA	2	0.23
CREB-binding protein with HIF-1-alpha	1l8cA	51	33	0.26	1u2nA	16	0.32
Importin subunit alpha with Nucleoporin NUP2	2c1tA	51	–	–	1bk5A	–	–
BoNT/A with SNAP-25	1xtgA	59	5	0.17	1xtfA	–	–

Bound indicates that the structure from PDB was solved with the IDP present. Unbound indicates that the structure from PDB was solved without the IDP present. RFH: rank of first hit; f_{nat}: fraction of native contacts present in the first acceptable hit. Len: length of the IDP. A "–" indicates that no hits were found among the 1000 models generated. This table is a reduced version of Table 7 in the original IDP-LZerD paper (19) (available under the Creative Commons CC0 public domain dedication).

Table 6.2 Summary of Modeling Performance on the Test Set

	Bound				Unbound		
Protein Names	**PDB ID**	**L**	**RFH**	f_{nat}	**PDB ID**	**RFH**	f_{nat}
Peroxisomal membrane protein PEX14 with PTS1R	2w84A	20	3	0.54	5aonA	6	0.19
PCNA with cyclin-dependent kinase inhibitor 1	1axcA	22	104	0.18	1vymA	81	0.17
Activated RNA polymerase II transcriptional coactivator p15 with Tegument protein VP16	2pheAB	26	11	0.25	1pcfAB	15	0.29
Saccharopepsin with Protease A inhibitor 3	1g0vA	31	1	0.65	1fmxA	1	0.29
Kelch-like ECH-associated protein 1 with NFE2-related factor 2	3wn7A	35	111	0.15	1x2jA	343	0.28
PP-1B with protein phosphatase 1 regulatory subunit 12A	1s70A	39	252	0.31	4ut2A	–	–
PP-1G with IPP-2	2o8gA	40	17	0.32	1jk7A	37	0.19
YopE regulator with outer membrane virulence protein YopE	1l2wAB	69	321	0.25	1jyaAB	2	0.21

Bound indicates that the structure from PDB was solved with the IDP present. Unbound indicates that the structure from PDB was solved without the IDP present. RFH: rank of first hit; f_{nat}: fraction of native contacts present in the first acceptable hit. Len: length of the IDP. A "–" indicates that no hits were found. This table is a reduced version of Table 8 in the original IDP-LZerD paper (19) (available under the Creative Commons CC0 public domain dedication).

Table 6.3 The Four Quality Levels of Docking Models Used in the Critical Assessment of Predicted Interactions (CAPRI) and Their Criteria

Quality	f_{nat}	L-RMSD		I-RMSD
High	≥0.5	≤1.0	or	≤1.0
Medium	≥0.3	$1.0 < x \leq 5.0$	or	$1.0 < x \leq 2.0$
Acceptable	≥0.1	$5.0 < x \leq 10.0$	or	$2.0 < x \leq 4.0$
Incorrect	<0.1			

f_{nat}: the fraction of native contacts that is reproduced in a predicted docking model. Native contacts are defined as pairs of residues from interacting proteins with any heavy atoms within 5.0Å of each other. L-RMSD: Ligand RMSD, the backbone (N, C_α, C, O) RMSD of the ligand (here a docked IDP) after superimposition of a docking model to its native complex structure using the receptor structure. I-RMSD: Interface RMSD, the backbone (N, C_α, C, O) RMSD of the interface residues. An interface residue is defined as having any heavy atom within 10.0Å of any heavy atom in a docked partner protein.
This table is taken from the Supplementary Table S1 of the IDP-LZerD paper (19) (available under the Creative Commons CC0 public domain dedication).
From Peterson, L. X.; Roy, A.; Christoffer, C.; Terashi, G.; Kihara, D. Modeling disordered protein interactions from biophysical principles. PLoS Comput. Biol. 2017, 13 (4), e1005485. Originally adapted from Mndez, R.; Leplae, R.; De Maria, L.; Wodak, S. J. Assessment of Blind Predictions of Protein-Protein Interactions: Current Status of Docking Methods. Proteins 2003, 52 (1), 51–67.

6.4.1 DRA/DRB5 with myelin basic protein (PDB code 1FV1)

The first example is two major histocompatibility complex class II chains complexed with myelin basic protein, a 20-residue IDP, which was also part of the data set in the original work. This complex is associated with susceptibility to multiple sclerosis *(41)*. This final refined model set contained 16 hits among the 1000 models generated based on the CAPRI criteria, of which four were in the top 10 models by model score. The first-ranked model was the highest-ranked hit, consistent with the original work where the sixth-ranked model was the highest ranked hit. The five highest-ranked hits are shown with the native structure in Fig. 6.3.

6.4.2 PKA C-alpha with PKI-alpha (PDB code 2CPK)

The second example is the cAMP-dependent protein kinase catalytic subunit alpha complexed with cAMP-dependent protein kinase inhibitor alpha, a 20-residue IDP, which was also part of the data set in the original work. The final refined model set contained six hits, although none were in the top 10 by model score. The 366th-ranked model was the highest-ranked hit in this run. The five highest-ranked hits are shown with the native structure in Fig. 6.4.

6.4.3 PCNA with p15 (PDB code 4D2G)

The third example is the proliferating cell nuclear antigen complexed with PCNA-associated factor of 15 kDa, a 21-residue IDP; this IDP was not part of the original data set, although another PPI involving PCNA was. P15 regulates DNA replication and repair by forming a complex with PCNA *(42)*. The final refined model set contained 147 hits, of which two were in the top 10 models by model score. The seventh-ranked model was the highest-ranked hit. The five highest-ranked hits are shown with the native structure in Fig. 6.5.

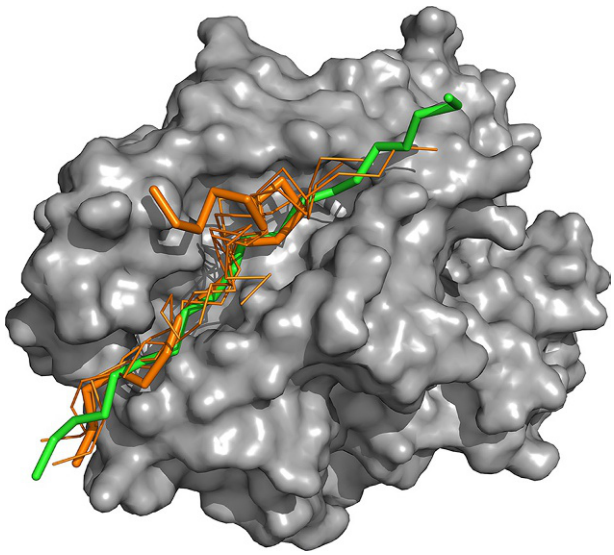

FIG. 6.3

Five highest-ranked hits for DRA/DRB5 with myelin basic protein (PDB ID 1FV1). Conformations are those ranked at 1, 2, 4, 5, and 7 by the scoring function. *Gray*: DRA/DRB5 (receptor protein); *green*: native bound myelin basic protein; *orange*: IDP-LZerD models of myelin basic protein. The extra-thick orange ribbon is the highest-ranked IDP-LZerD hit (meeting the CAPRI criteria).

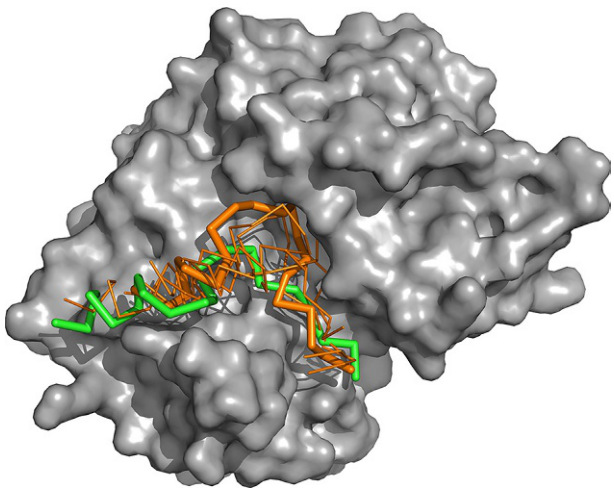

FIG. 6.4

Five highest-ranked hits for PKA C-alpha with PKI-alpha (PDB ID 2CPK). Score ranks of the models are 366, 431, 647, 691, and 861. *Gray*: PKA C-alpha; *green*: native bound PKI-alpha; *orange*: IDP-LZerD models of PKI-alpha. The extra-thick orange ribbon is the highest-ranked IDP-LZerD hit.

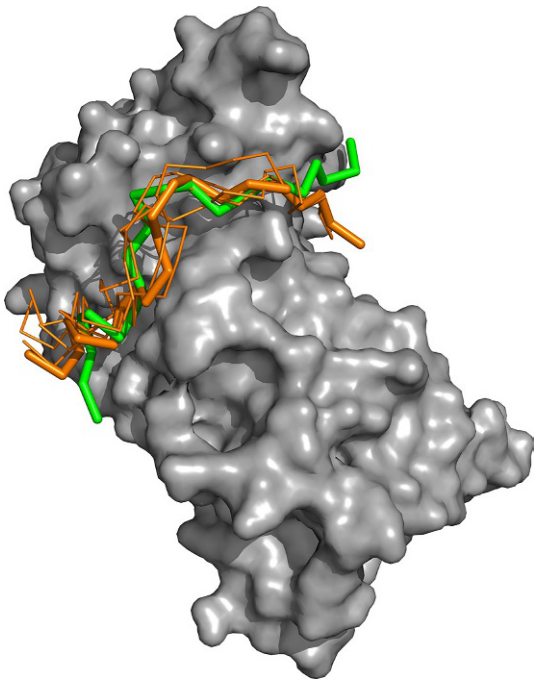

FIG. 6.5

Five highest-ranked hits for PCNA with p15 (PDB ID 4D2G). Score ranks of the models are 7, 9, 12, 13, and 14. *Gray*: PCNA; *green*: native bound p15; *orange*: IDP-LZerD models of p15. The extra-thick orange ribbon is the highest-ranked IDP-LZerD hit.

6.4.4 CREB-binding protein with HIF-1-alpha (PDB code 1L8C)

The latter three examples show docking with rather long IDPs of 51, 59, and 41 residues. This example is a CREB-binding protein with HIF-1-alpha, a 51-residue IDP. HIF-1 mediates cellular response to hypoxia *(43)*. The refined model set contained 99 hits, although none were in the top 10 by model score. The five highest-ranked hits are shown with the native structure in Fig. 6.6. As shown, the long IDP wraps around the receptor protein, which is well reproduced in the model.

6.4.5 BoNT/A with SNAP-25 (PDB code 1XTG)

This example is BoNT/A, a neurotoxin, with SNAP-25, a 59-residue IDP. The final refined model set contained seven hits, one of which was in the top 10 models by model score. The five highest-ranked hits are shown with the native structure in Fig. 6.7. Similar to models shown for other complexes, the models capture binding sites on the receptor very well, which would be sufficient to understand the residue-level interactions of these proteins.

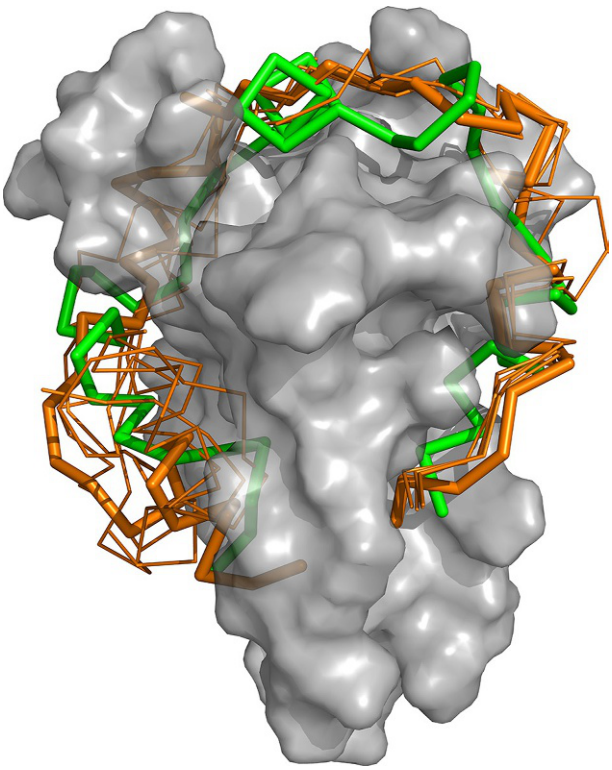

FIG. 6.6

Five highest-ranked hits for CREB-binding protein with HIF-1-alpha (PDB ID 1L8C). Score ranks of the models are 33, 58, 63, 88, and 97. *Gray*: CREB-binding protein; *green*: native bound HIF-1-alpha; *orange*: IDP-LZerD models of HIF-1-alpha. The extra-thick orange ribbon is the highest-ranked IDP-LZerD hit. Because HIF-1-alpha wraps around the receptor structure, the receptor is rendered partially transparent to show the entire binding pose.

6.4.6 MAD homolog 2 with SARA (PDB code 1DEV)

This example is MAD homolog 2 with SARA, a 41-residue IDP. SARA recruits MAD homolog 2 for phosphorylation *(44)*. The final refined model set contained six hits, of which one was in the top 10 models by model score. The 5five highest-ranked hits are shown with the native structure in Fig. 6.8. There is some discrepancy between the modeled IDP conformations and the correct structure (green) at the middle of the IDP; the two ends of the IDP were well modeled.

6.5 Comparison with related methods

Several protein-peptide complex modeling methods exist that can model relatively long peptides. For example, DynaDock *(45)* is a molecular dynamics-based method

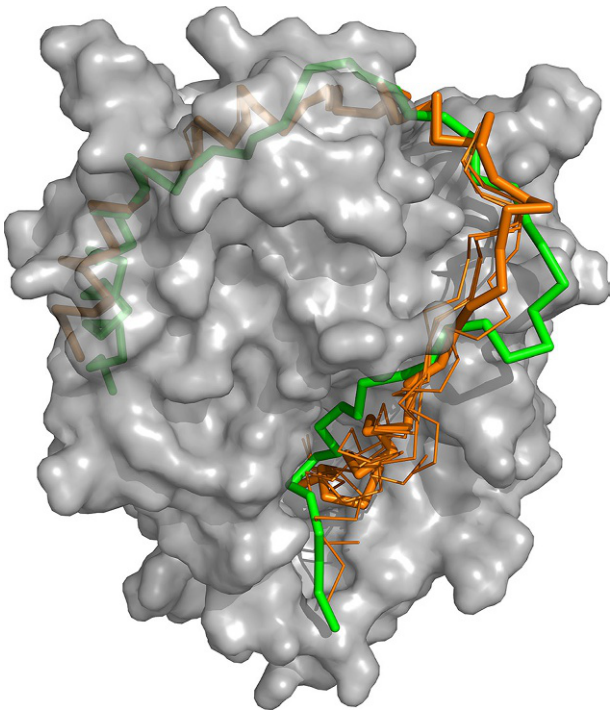

FIG. 6.7

Five highest-ranked hits for BoNT/A with SNAP-25 (PDB ID 1XTG). Score ranks of the models are 5, 11, 12, 16, and 25. *Gray*: BoNT/A; *green*: native bound SNAP-25; *orange*: IDP-LZerD models of SNAP-25. The extra-thick orange ribbon is the highest-ranked IDP-LZerD hit. Because SNAP-25 wraps around the receptor structure, the receptor is rendered partially transparent to show the entire binding pose.

that requires the binding site as input, and was tested on peptides of up to 16 residues. AnchorDock *(46)* is another molecular dynamics-based method, but it does not require binding site information. AnchorDock was tested on peptides of up to 15 residues. The Rosetta FlexPepDock ab initio protocol *(47, 48)* is a docking-based method that requires binding site information, and was tested on peptides of up to 15 residues (although the web server will allow up to 30-residue queries). pepATTRACT *(49)* and CABS-dock *(50)* are both docking-based methods that do not require binding site information; both were tested on peptides of up to 15 residues. However, none of these methods was designed to work with long IDPs.

These last two methods, pepATTRACT and CABS-dock, were compared with IDP-LZerD *(19)*. Only complexes where the peptide has at most 30 residues were used because CABS-dock limits the query length to 30; this limitation reduced the data set size to 11 bound and 11 unbound cases. Within the top 10 output models, CABS-dock had hits for six bound cases and four unbound cases while pepATTRACT had hits for three bound cases and one unbound case; IDP-LZerD had

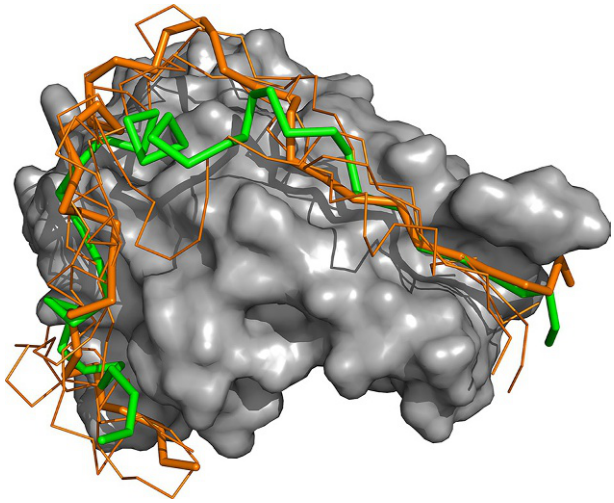

FIG. 6.8

Five highest-ranked hits for MAD homolog 2 with SARA (PDB ID 1DEV). Score ranks of the models are 2, 15, 62, 85, and 107. *Gray*: MAD homolog 2; *green*: native bound SARA; *orange*: IDP-LZerD models of SARA. The extra-thick orange ribbon is the highest-ranked IDP-LZerD hit.

hits for seven bound and four unbound cases. Furthermore, while CABS-dock and pepATTRACT found a hit within the top 10 for IDPs of up to 26 and 22 residues on this data set, respectively, IDP-LZerD found such hits for the longest IDP in this data set, which had 27 residues. Thus, IDP-LZerD is unique in its ability to dock long IDPs and showed better performance than the two methods for IDPs of at most 30 residues.

6.6 Limitations of the current method

While IDP-LZerD outperforms other existing methods, it has limitations. Perhaps the most obvious is that conformation changes in the receptor protein are not modeled; LZerD's soft surface representation can tolerate incorrect side chain conformations and small backbone deviations, but they are not considered in a way that permits large backbone changes at the interface to be modeled. Another limitation is the assumption that all segments of the IDP bind to the receptor; IDP-LZerD does not explicitly model disordered regions that "float" even when the IDP is bound. A related issue is the presence of stable globular regions in an otherwise disordered protein. IDP-LZerD assumes that the entire input sequence is disordered.

6.7 Summary

In this chapter, we introduced the workflow of IDP-LZerD and several examples of IDPs modeled with it. Although disordered PPI structures can be difficult to capture from experimental data, they can be modeled by following the dock-and-coalesce model using IDP-LZerD. Currently, IDP-LZerD may be limited by its assumption that the entire IDP chain is disordered and has contact with the receptor in the bound form, but it is nonetheless effective at predicting reasonably good PPI models. As shown in the results from the original work and the freshly computed examples, IDP-LZerD nearly always generates at least one hit, and is often able to raise hits to the top 10 ranks.

6.8 Availability

LZerD is available for download at http://www.kiharalab.org/proteindocking/lzerd.php. IDP-LZerD is available for download at http://www.kiharalab.org/proteindocking/idp_lzerd.tar.bz2.

Acknowledgments

This work was supported by the National Institutes of Health (R01GM123055). DK also acknowledges support from the National Science Foundation (DMS1614777, CMMI1825941) and the Purdue Institute for Drug Discovery. This research was supported in part through computational resources provided by Information Technology at Purdue, West Lafayette, Indiana.

References

1. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **1973**, *181* (4096), 223–230.
2. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.
3. Lawson, C. L.; Patwardhan, A.; Baker, M. L.; Hryc, C.; Garcia, E. S.; Hudson, B. P.; Lagerstedt, I.; Ludtke, S. J.; Pintilie, G.; Sala, R.; Westbrook, J. D.; Berman, H. M.; Kleywegt, G. J.; Chiu, W. EMDataBank Unified Data Resource for 3DEM. *Nucleic Acids Res.* **2016**, *44* (D1), D396–D403.
4. Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337* (3), 635–645.
5. Yruela, I.; Oldfield, C. J.; Niklas, K. J.; Dunker, A. K. Evidence for a Strong Correlation Between Transcription Factor Protein Disorder and Organismic Complexity. *Genome Biol. Evol.* **2017**, *9* (5), 1248–1265.

6. Petsalaki, E.; Russell, R. B. Peptide-Mediated Interactions in Biological Systems: New Discoveries and Applications. *Curr. Opin. Biotechnol.* **2008**, *19* (4), 344–350.
7. Zhou, J.; Zhao, S.; Dunker, A. K. Intrinsically Disordered Proteins Link Alternative Splicing and Post-Translational Modifications to Complex Cell Signaling and Regulation. *J. Mol. Biol.* **2018**, *430* (16), 2342–2359.
8. Darling, A. L.; Uversky, V. N. Intrinsic Disorder and Posttranslational Modifications: The Darker Side of the Biological Dark Matter. *Front. Genet.* **2018**, *9*, 158.
9. Berlow, R. B.; Dyson, J.; Wright, P. E. Expanding the Paradigm: Intrinsically Disordered Proteins and Allosteric Regulation. *J. Mol. Biol.* **2018**, *430* (16), 2309–2320.
10. Jones, D. T.; Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **2015**, *31* (6), 857–863.
11. Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; Oldfield, C. J.; Aspromonte, M. C.; Davey, N. E.; Davidović, R.; Dosztányi, Z.; Elofsson, A.; Gasparini, A.; Hatos, A.; Kajava, A. V.; Kalmar, L.; Leonardi, E.; Lazar, T.; Macedo-Ribeiro, S.; Macossay-Castillo, M.; Meszaros, A.; Minervini, G.; Murvai, N.; Pujols, J.; Roche, D. B.; Salladini, E.; Schad, E.; Schramm, A.; Szabo, B.; Tantos, A.; Tonello, F.; Tsirigos, K. D.; Veljković, N.; Ventura, S.; Vranken, W.; Warholm, P.; Uversky, V. N.; Dunker, A. K.; Longhi, S.; Tompa, P.; Tosatto, S. C. E. DisProt 7.0: A Major Update of the Database of Disordered Proteins. *Nucleic Acids Res.* **2017**, *45* (D1), D1123–D1124.
12. Oates, M. E.; Romero, P.; Ishida, T.; Ghalwash, M.; Mizianty, M. J.; Xue, B.; Dosztányi, Z.; Uversky, V. N.; Obradovic, Z.; Kurgan, L.; Dunker, A. K.; Gough, J. D2P2: Database of Disordered Protein Predictions. *Nucleic Acids Res.* **2013**, *41* (D1), D508–D516.
13. Fukuchi, S.; Sakamoto, S.; Nobe, Y.; Murakami, S. D.; Amemiya, T.; Hosoda, K.; Koike, R.; Hiroaki, H.; Ota, M. IDEAL: Intrinsically Disordered Proteins With Extensive Annotations and Literature. *Nucleic Acids Res.* **2012**, *40* (D1), D507–D511.
14. Konrat, R. NMR Contributions to Structural Dynamics Studies of Intrinsically Disordered Proteins. *J. Magn. Reson.* **2014**, *241*, 74–85.
15. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14* (1), 71–73.
16. Wu, H.; Wolynes, P. G.; Papoian, G. A. AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins. *J. Phys. Chem. B* **2018**, 122 (49), 11115–11125.
17. Song, D.; Luo, R.; Chen, H.-F. The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *J. Chem. Inf. Model.* **2017**, *57* (5), 1166–1178.
18. Mohan, A.; Oldfield, C. J.; Radivojac, P.; Vacic, V.; Cortese, M. S.; Dunker, A. K.; Uversky, V. N. Analysis of Molecular Recognition Features (MoRFs). *J. Mol. Biol.* **2006**, *362* (5), 1043–1059.
19. Peterson, L. X.; Roy, A.; Christoffer, C.; Terashi, G.; Kihara, D. Modeling disordered protein interactions from biophysical principles. *PLoS Comput. Biol.* **2017**, *13* (4), e1005485.
20. Zhou, H.-X. Intrinsic Disorder: Signaling via Highly Specific But Short-Lived Association. *Trends Biochem. Sci.* **2012**, *37* (2), 43–48.
21. Venkatraman, V.; Yang, Y. D.; Sael, L.; Kihara, D. Protein-Protein Docking Using Region-Based 3D Zernike Descriptors. *BMC Bioinform.* **2009**, *10* (407).
22. Esquivel-Rodríguez, J.; Filos-Gonzalez, V.; Li, B.; Kihara, D. Pairwise and Multimeric Protein–Protein Docking Using the LZerD Program Suite. *Methods Mol. Biol.* **2014**, 209–234.

23. Peterson, L. X.; Kim, H.; Esquivel-Rodriguez, J.; Roy, A.; Han, X.; Shin, W.-H.; Zhang, J.; Terashi, G.; Lee, M.; Kihara, D. Human and Server Docking Prediction for CAPRI Round 30-35 Using LZerD With Combined Scoring Functions. *Proteins* **2016**, *85* (3), 513–527.
24. Gront, D.; Kulp, D. W.; Vernon, R. M.; Strauss, C. E. M.; Baker, D. Generalized Fragment Picking in Rosetta: Design, Protocols and Applications. *PLoS One* **2011**, *6* (8), e23294.
25. Jones, D. T. Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *J. Mol. Biol.* **1999**, *292* (2), 195–202.
26. Drozdetskiy, A.; Cole, C.; Procter, J.; Barton, G. J. JPred4: A Protein Secondary Structure Prediction Server. *Nucleic Acids Res.* **2015**, *43* (W1), W389–W394.
27. Mirabello, C.; Pollastri, G. Porter, PaleAle 4.0: High-Accuracy Prediction of Protein Secondary Structure and Relative Solvent Accessibility. *Bioinformatics* **2013**, *29* (16), 2056–2058.
28. Magnan, C. N.; Baldi, P. SSpro/ACCpro 5: Almost Perfect Prediction of Protein Secondary Structure and Relative Solvent Accessibility Using Profiles, Machine Learning and Structural Similarity. *Bioinformatics* **2014**, *30* (18), 2592–2597.
29. Rotkiewicz, P.; Skolnick, J. Fast Procedure for Reconstruction Of Full-Atom Protein Models From Reduced Representations. *J. Comput. Chem.* **2008**, *29* (9), 1460–1465.
30. Liang, S.; Zheng, D.; Zhang, C.; Standley, D. M. Fast and Accurate Prediction of Protein Side-Chain Conformations. *Bioinformatics* **2011**, *27* (20), 2913–2914.
31. Huang, S.-Y.; Zou, X. Statistical Mechanics-Based Method to Extract Atomic Distance-Dependent Potentials From Protein Structures. *Proteins* **2011**, *79* (9), 2648–2661.
32. ZHou, H.; Zhou, Y. Distance-Scaled, Finite Ideal-Gas Reference State Improves Structure-Derived Potentials of Mean Force for Structure Selection and Stability Prediction. *Protein Sci.* **2009**, *11* (11), 2714–2726.
33. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217.
34. Haberthür, U.; Caflisch, A. FACTS: Fast Analytical Continuum Treatment of Solvation. *J. Comput. Chem.* **2007**, *29* (5), 701–715.
35. Esquivel-Rodríguez, J.; Yang, Y. D.; Kihara, D. Multi-LZerD: Multiple Protein Docking for Asymmetric Complexes. *Proteins* **2012**, *80* (7), 1818–1833.
36. Zhou, H.; Skolnick, J. GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. *Biophys. J.* **2011**, *101* (8), 2043–2052.
37. Méndez, R.; Leplae, R.; De Maria, L.; Wodak, S. J. Assessment of Blind Predictions of Protein-Protein Interactions: Current Status of Docking Methods. *Proteins* **2003**, *52* (1), 51–67.
38. Heffernan, R.; Abdollah, D.; Lyons, J.; Paliwal, K.; Sharma, A.; Wang, J.; Sattar, A.; Zhou, Y.; Yang, Y. Highly Accurate Sequence-Based Prediction of Half-Sphere Exposures of Amino Acid Residues in Proteins. *Bioinformatics* **2016**, *32* (6), 843–849.
39. Wang, S.; Peng, J.; Ma, J.; Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci. Rep.* **2016**, *6* (1).
40. Spencer, M.; Eickholt, J.; Cheng, J. A Deep Learning Network Approach to Ab Initio Protein Secondary Structure Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12* (1), 103–112.
41. Li, Y.; Li, H.; Martin, R.; Mariuzza, R. A. Structural Basis for the Binding of an Immunodominant Peptide From Myelin Basic Protein in Different Registers by Two HLA-DR2 Proteins. *J. Mol. Biol.* **2000**, *304* (2), 177–188.

42. De Blasio, A.; de Opakua, A. I.; Mortuza, G. B.; Molina, R.; Cordeiro, T. N.; Castillo, F.; Villate, M.; Merino, N.; Delgado, S.; Gil-Cartón, D.; Luque, I.; Diercks, T.; Bernadó, P.; Montoya, G.; Blanco, F. J. Structure of p15PAF–PCNA Complex and Implications for Clamp Sliding During DNA Replication and Repair. *Nat. Commun.* **2015**, *6* (1).
43. Dames, S. A.; Martinez-Yamout, M.; De Guzman, R. N.; Dyson, H. J.; Wright, P. E. Structural Basis for Hif-1α/CBP Recognition in the Cellular Hypoxic Response. *Proc. Natl. Acad. Sci.* **2002**, *99* (8), 5271–5276.
44. Wu, G.; Chen, Y.-G.; Ozdamar, B.; Gyuricza, C. A.; Chong, A.; Wrana, J. L.; Massagué, J.; Shi, Y. Structural Basis of Smad2 Recognition by the Smad Anchor for Receptor Activation. *Science* **2000**, *287* (5450), 92–97.
45. Antes, I. DynaDock: A New Molecular Dynamics-Based Algorithm for Protein-Peptide Docking Including Receptor Flexibility. *Proteins* **2009**, *78* (5), 1084–1104.
46. Ben-Shimon, A.; Niv, M. Y. AnchorDock: Blind and Flexible Anchor-Driven Peptide Docking. *Structure* **2015**, *23* (5), 929–940.
47. Raveh, B.; London, N.; Schueler-Furman, O. Sub-Angstrom Modeling of Complexes Between Flexible Peptides and Globular Proteins. *Proteins* **2010**, *78* (9), 2029–2040.
48. Raveh, B.; London, N.; Zimmerman, L.; Schueler-Furman, O. Rosetta FlexPepDock Ab-Initio: Simultaneous Folding, Docking and Refinement of Peptides Onto Their Receptors. *PLoS One* **2011**, *6* (4), e18934.
49. Schindler, C. E. M.; de Vries, S. J.; Zacharias, M. Fully Blind Peptide-Protein Docking With pepATTRACT. *Structure* **2015**, *23* (8), 1507–1515.
50. Kurcinski, M.; Jamroz, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. CABS-Dock Web Server for the Flexible Docking of Peptides to Proteins Without Prior Knowledge of the Binding Site. *Nucleic Acids Res.* **2015**, *43* (W1), W416–W424.