



# HHS Public Access

Author manuscript

*Nat Protoc.* Author manuscript; available in PMC 2018 February 01.

Published in final edited form as:

*Nat Protoc.* 2017 February ; 12(2): 255–278. doi:10.1038/nprot.2016.169.

## The ClusPro web server for protein-protein docking

Dima Kozakov<sup>a,b,d,\*</sup>, David R. Hall<sup>c</sup>, Bing Xia<sup>b</sup>, Kathryn A. Porter<sup>b</sup>, Dzmitry Padhorny<sup>a</sup>, Christine Yueh<sup>b</sup>, Dmitri Beglov<sup>b</sup>, and Sandor Vajda<sup>b,\*</sup>

<sup>a</sup>Department of Applied Mathematics and Statistics, Stony Brook University NY, USA

<sup>b</sup>Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

<sup>c</sup>Acpharis Inc., Holliston, MA 01746, USA

<sup>d</sup>Laufer Center for Physical and Quantitative Biology, Stony Brook University NY, USA

### Abstract

The ClusPro server (<https://cluspro.org>) is a widely used tool for protein-protein docking. The server provides a simple home page for basic use, requiring only two files in Protein Data Bank format. However, ClusPro also offers a number of advanced options to modify the search that include the removal of unstructured protein regions, applying attraction or repulsion, accounting for pairwise distance restraints, constructing homo-multimers, considering small angle X-ray scattering (SAXS) data, and finding heparin binding sites. Six different energy functions can be used depending on the type of proteins. Docking with each energy parameter set results in ten models defined by centers of highly populated clusters of low energy docked structures. This protocol describes the use of the various options, the construction of auxiliary restraints files, the selection of the energy parameters, and the analysis of the results. Although the server is heavily used, runs are generally completed in < 4 hours.

### INTRODUCTION

Protein-protein interactions are important for understanding cellular function and organization. Substantial progress has been made toward generating potential protein-protein interaction networks using high-throughput proteomics studies, primarily yeast two-hybrid assays<sup>1,2</sup> and mass spectrometry<sup>3,4</sup>. Mechanistic interpretation of the interactions frequently requires atom-level details, ideally obtained by X-ray crystallography. However, some of the biologically important interactions occur in transient complexes, and hence experimental structure determination may be very difficult, even when the structures of the component proteins are known. Therefore, computational docking methods have been developed that,

\*Corresponding authors: Sandor Vajda, vajda@bu.edu, 1-617-353-4757; Dima Kozakov, midas@laufercenter.org, 1-617-353-4742.

#### Author contributions statements

D.K., D.R.H., B.X., and D.B. developed the server; D.K., D.R.H., D.P., B.X., and K.P. performed experiments; S.V., K.P., D.R.H., and C.Y. prepared the manuscript.

#### Competing financial interests

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/nprot/index.html>.

TWEET: The ClusPro web server predicts models of the complex based on the structures of two interacting proteins.

starting from the structures of component proteins, attempt to determine the structure of their complexes targeting an accuracy close to that provided by X-ray crystallography<sup>5-7</sup>. Docking usually generates a number of detailed models that define the positions of all atoms, but the current scoring functions are usually not accurate enough for reliable model discrimination, and in most cases the model closest to the native structure cannot be identified solely by computational tools. However, model selection can be based on additional information obtained by lower resolution methods such as site-directed mutagenesis or chemical cross-linking, and the selected models generated by the docking provide atom-level details.

Docking methods can be classified as direct or template-based. Based on thermodynamics, direct methods attempt to find the structure of the target complex located at the minimum of Gibbs free energy in the conformational space, and thus require a computationally feasible free energy evaluation model and an effective minimization algorithm<sup>8</sup>. As will be discussed, direct docking methods may give good results if the conformational changes upon protein-protein association are moderate. Template-based docking is based on the observation that interacting pairs sharing above 30% sequence identity often interact in the same way, and hence the structure of the target complex can be obtained by homology modeling tools if an appropriate template complex of known structure is available<sup>9</sup>. Although the applicability of template-based docking has been extended based on the observation that partial structures representing the interface region can provide templates<sup>10</sup>, the coverage of the template space at present is still limited and hence direct methods are generally more useful in many applications.

This protocol describes ClusPro, a web based server for the direct docking of two interacting proteins. ClusPro was introduced in 2004<sup>11,12</sup> but since then has been substantially modified and expanded<sup>13-15</sup>. The server performs three computational steps as follows: (1) rigid body docking by sampling billions of conformations, (2) root-mean-square deviation (RMSD) based clustering of the 1000 lowest energy structures generated to find the largest clusters that will represent the most likely models of the complex, and (3) refinement of selected structures using energy minimization (Figure 1). The rigid body docking step uses PIPER<sup>16</sup>, a docking program based on the Fast Fourier Transform (FFT) correlation approach. The FFT approach, introduced by Katchalski-Katzir and co-workers<sup>17</sup> in 1992, led to major progress in rigid body protein-protein docking. In this method, one of the proteins (which we will call the receptor) is placed at the origin of the coordinate system on a fixed grid, the second protein (which we will call the ligand) is placed on a movable grid, and the interaction energy is written in the form of a correlation function (or as a sum of a few correlation functions). The numerical efficiency of the methods stems from the fact that such energy functions can be efficiently calculated using Fast Fourier Transforms, and results in the ability of exhaustively sampling billions of the conformations of the two interacting protein, evaluating the energies at each grid point. Thus, the FFT based algorithm enables docking of proteins without any *a priori* information on the structure of the complex. Katchalski-Katzir *et al.*<sup>17</sup> used a simple scoring function that accounted only for shape complementarity. However, subsequent methods based on the FFT correlation approach to docking introduced more complex and more accurate scoring functions that also included terms representing electrostatic interactions<sup>18,19</sup>, or both electrostatic and desolvation

contributions<sup>20</sup>. A key to the success of rigid body methods is that the shape complementarity term allows for some overlaps, and hence the methods are able to tolerate moderate differences between bound and unbound (separately crystallized) structures. As will be discussed, one of the distinguishing characteristics of PIPER, the docking program used in the current version of ClusPro, is that this implementation of the FFT correlation method employs a scoring function that includes a structure-based pairwise interaction term, and the combination with the other terms in the energy function substantially increases the accuracy of docking, resulting in more near-native structures<sup>16</sup>.

Rigid body methods including PIPER perform exhaustive sampling of the conformational space on a dense grid, and hence certainly sample some near-native structures. However, the need for tolerating some steric clashes due to docking unbound protein structures requires the use of approximate scoring functions, and reducing sensitivity to conformational differences also reduces specificity. Therefore, the docked conformations that are close to the native structure do not necessarily have the lowest energies, whereas low energy conformations may occur far from the X-ray structures. In PIPER we retain the 1000 lowest energy docked structures for further processing and hope that this set includes at least some that are close to the native structure of the complex. Another unique feature of ClusPro is that we select the centers of highly populated clusters of the low energy structures rather than simply the lowest energy conformations as predictions of the complex<sup>21</sup>. As will be further discussed, the size of each cluster represents the width of the corresponding energy well, and hence provides some information on entropic contributions to the free energy. While the largest clusters do not necessarily contain the most near-native structures, we have shown<sup>12</sup> that the 30 largest clusters include near-native structures for 92% of complexes in a protein-protein benchmark set<sup>22-25</sup>. However, the success rates are higher for certain classes of complexes such as enzyme-inhibitor pairs<sup>26,27</sup>, and hence it is generally sufficient to retain only ten or fewer highly populated clusters. Supplementary Table 1 lists some performance characteristics of ClusPro 2.0 for the different classes of complexes in benchmark 4.0. The characteristics shown are the number of complexes with at least one cluster of docked structures within 10 Å interface root mean square deviation (IRMSD) from the native complex, the average number of docked structures with less than 10 Å IRMSD within the 1000 lowest energy structures, and the average value of the lowest IRMSD achieved. Note that to define the IRMSD of a docked structure from the native complex we first select the interface residues as the ligand residues that have any atom within 10 Å of any receptor atom. Then we superimpose the receptors in the two structures, and calculate the C $\alpha$  RMSD for the selected interface residues. According to Supplementary Table 1, a large fraction of enzyme-inhibitor pairs are in the “easy” category, and in almost all cases we have a cluster with its center within 10 Å IRMSD, with the average IRMSD of 2.94 Å. In contrast, for the “other” type complexes such clusters are found only in about 50% of cases, and the average IRMSD is 7.54 Å even in the easy category. The success rate is similar for antibody-antigen pairs if both proteins are independently crystallized. However, as will be discussed, the performance is improved if the non-CDR regions of the antibody are masked and hence cannot be in the interface.

The current version, ClusPro 2.0, is heavily used; by June 2016 it had over 17,000 users (among them over 7,000 registered, which is not required), and completed over 172,000

docking calculations, currently adding about 3,000 per month. Models built by the server have been reported in over 400 publications. Later in this protocol we describe the types of the applications that frequently occur in these papers.

### The PIPER docking algorithm

As stated, the ClusPro server is based on PIPER, which performs the sampling. The center of mass of the receptor is fixed at the origin of the coordinate system, and the possible rotational and translational positions of the ligand are evaluated at the given level of discretization. The rotational space is sampled on a sphere-based grid that defines a subdivision of a spherical surface in which each pixel covers the same surface area as every other pixel<sup>28</sup>. The 70,000 rotations we consider correspond to about 5 degrees in terms of the Euler angles. The step size of the translational grid is 1 Å. It is easy to see that for an average size protein this amounts to sampling  $10^9$ – $10^{10}$  conformations.

PIPER represents the interaction energy between two proteins using an expression of the form  $E = w_1E_{rep} + w_2E_{attr} + w_3E_{elec} + w_4E_{DARS}$ , where  $E_{rep}$  and  $E_{attr}$  denote the repulsive and attractive contributions to the van der Waals interaction energy, and  $E_{elec}$  is an electrostatic energy term.  $E_{DARS}$  is a pairwise structure-based potential constructed by the Decoys as the Reference State (DARS)<sup>29</sup> approach, and it primarily represents desolvation contributions, i.e., the free energy change due to the removal of the water molecules from the interface<sup>16</sup>. The coefficients  $w_1$ ,  $w_2$ ,  $w_3$ , and  $w_4$  define the weights of the corresponding terms, and are optimally selected for different types of docking problems. Unless specified otherwise in Advanced Options, the server generates four sets of models using the scoring schemes called (1) balanced, (2) electrostatic-favored, (3) hydrophobic-favored, and (4) van der Waals + electrostatics (see Table 1). As will be described, for complexes that do not fit into these categories and are classified as “others” in the protein-protein docking benchmarks<sup>22–25</sup>, ClusPro generates structures using three different coefficient sets (Table 1). To understand the magnitude of these coefficients we note that  $E_{rep}$  and  $E_{attr}$ , while defined on the grid, are scaled to the van der Waals energy<sup>16</sup>, and hence  $w_1 < 1.0$  and  $w_2 < 1.0$  yield “softening” of both repulsive and attractive van der Waals terms. Such softening is necessary, since the bound and unbound conformations of the proteins to be docked generally differ, in some cases substantially. The DARS potential has been parameterized on a set of complexes that included a large number of enzyme-inhibitor pairs and multi-subunit proteins, and hence the resulting potential assumes good shape and electrostatic complementarity<sup>29</sup>. Since  $E_{DARS}$  is scaled to the magnitudes of protein-protein binding free energies,  $w_4 = 1.0$  is the “neutral” choice.  $E_{elec}$  is represented by a truncated and smoothed Coulomb expression, also defined on the grid<sup>16</sup>. Since the charges are given as multiples of the electronic charge and the interatomic distance is measured in Å, obtaining the electrostatic interaction energy in kcal/mol using the Coulomb expression and these units<sup>30</sup> would require the coefficient  $w_3 = 332$ . However, the truncation and smoothing reduce the magnitude of  $E_{elec}$  relative to the value based on the Coulomb expression, and hence we use  $w_3 = 600$  in the balanced option of the parameter set. This set was shown to generally provide very good results for enzyme-inhibitor complexes. If it is known or assumed that the association of two proteins is mainly driven by their electrostatic interactions, then we select results obtained by the electrostatic-favored weights, in which the weight of the

electrostatics is doubled relative to the balanced energy expression. In contrast, for complexes primarily stabilized by hydrophobic interactions we use the hydrophobic-favored potential, which thus doubles the weight of the desolvation term. In the fourth option, van der Waals + electrostatics, the pairwise potential  $E_{pair}$  is not used. The need for this option occurs for proteins that are very different from the ones used for the parameterization of  $E_{DARS}$ . The selection of the parameters in the Others Mode will be discussed further in this protocol. We recognize that the weights in the PIPER energy function are somewhat arbitrary, and were selected to optimize the success rates for different classes of protein complexes. However, as described below, the most likely models of the complex are selected on the basis of cluster population rather than energy value. In fact, PIPER does not aim to estimate the true interaction energy, and the score provided by ClusPro should not be considered as a measure of binding affinity.

### Model selection by cluster population

The second step of ClusPro is clustering the lowest energy 1000 docked structures using pairwise IRMSD as the distance measure.<sup>12,21</sup> We calculate IRMSD values for each pair among the 1000 structures, and find the structure that has the highest number of neighbors within 9Å IRMSD radius. The selected structure will be defined as the center of the first cluster, and the structures within the 9Å IRMSD neighborhood of the center will constitute the first cluster. The members of this cluster are then removed, and we select the structure with the highest number of neighbors within the 9Å IRMSD radius among the remaining structures as the next cluster center. These neighbors will form the next cluster. Up to 30 clusters are generated in this manner. After clustering we minimize the energy of the retained structures for 300 step with fixed backbone using only the van der Waals term of the Charmm potential<sup>31</sup>. The minimization removes the steric overlaps but generally yields only very small conformational changes. In its basic operation ClusPro outputs the structures at the centers of the 10 most populated clusters.

Considering the centers of the large clusters of low energy structures rather than simply low energy structures appears to be unique to ClusPro,<sup>32</sup> and may implicitly account for some entropic effects.<sup>33</sup> We have recently shown that under fairly natural assumptions the cluster populations are proportional to cluster probability. This argument is based on considering the approximate partition function  $Q = \sum_j \exp(-E_j/RT)$ , where  $E_j$  is the energy of the  $j$ th pose, and the summation includes all structures generated by the docking. Similarly, for the  $k$ th low energy cluster the partition function can be approximated by  $Q_k = \sum_j \exp(E_j/RT)$ , where now we consider only the structures in the cluster. In terms of these approximate partition functions the probability of the  $k$ th cluster is given by  $P_k = Q_k/Q$ . Furthermore, within the low energy cluster we introduce the approximation  $E_j = E$ , *i.e.*, we assume that the energy values are the same for all structures in the cluster, because the energies are from a narrow range that is comparable to the uncertainty of the energy function. Due to this assumption the probability of the cluster is given by  $P_k = \exp(-E/RT) \times N_k/Q$ , and thus it is proportional to the number  $N_k$  of the structures in the  $k$ th cluster. Therefore we suggest finding the most populated clusters rather than the lowest energy structures and using the centers of the clusters as putative models. Although this approach to model selection is somewhat unusual and has not been widely adopted, we believe that the consistent success

of ClusPro in the CAPRI experiment<sup>34–38</sup> supports its use (see the section Comparison to other methods for details on the performance of ClusPro). Since it was requested by a number of users, ClusPro also outputs the PIPER energies of cluster centers, as well as the lowest PIPER energy within each cluster. However, since these values do not include entropic contributions and the weights of the energy components are selected to yield near-native structures rather than correct thermodynamics of binding, the PIPER energy does not provide valid information on binding free energy, and should not be used for ranking the models. We repeatedly emphasize that the latter is based on cluster population rather than cluster energy score.

### Applications of the method

The usage of the ClusPro server has been growing beyond our expectations. By June 2016 the server had over 17,000 users (among them over 7,000 registered, which is not required), and completed more than 172,000 docking calculations, currently adding about 3,000 per month. Models built by the server have been reported in over 400 publications. These statistics demonstrate that there is substantial need for protein-protein docking, and that implementing our algorithms as a server expanded the availability of the method to parts of the research community without experience in computational structural biology. The large number of publication also implies that we can follow how the server is employed. We describe here some of the typical applications, each illustrated by examples.

**Docking X-ray or NMR structures of proteins**—This is the most basic and straightforward application of ClusPro. As an early example, Chance and co-workers constructed the three-dimensional structure of cofilin, an important actin binding protein, bound to monomeric actin<sup>39,40</sup>. The binding model was validated by hydroxyl radical-mediated oxidative protein footprinting, and identified key ionic and hydrophobic interactions at the binding interface<sup>39</sup>. Models can be frequently used for mechanistic interpretation. For example, Luxán *et al.*<sup>41</sup> docked JAG1, a cell surface protein that interacts with receptors in the mammalian Notch signaling pathway, to the dimers of the E3 ubiquitin-protein ligase MIB1, and used the results to show that certain mutations of MIB1 arrest chamber myocardium development, preventing trabecular maturation and compaction, and thereby cause left ventricular noncompaction cardiomyopathy.

**Modeling antibody-antigen interactions**—The analysis of antibody-antigen interactions is a particularly important modeling and docking problem, required in biotechnology and vaccine design applications. For example, Tran *et al.*<sup>42</sup> crystallized the Fab fragments of two vaccine-elicited monoclonal antibodies (mAbs) binding to HIV-1 gp120. Alanine scanning of their complementarity-determining regions, coupled with epitope scanning of their epitopes on gp120, revealed putative contact residues. Using this information they docked the Fabs to gp120. Coupled with EM reconstructions of the mAb-gp120 complexes, the docking results suggested that the antibodies use a distinct approach to the HIV-1 primary receptor binding site, and this information was used for vaccine redesign.

**Constructing the structure of large multi-domain proteins**—We reviewed the work<sup>43</sup> by Kuriyan and co-workers<sup>44</sup>, who combined computational docking, small-angle x-ray scattering, mutagenesis, and calorimetry to study the histone domain of the Ras-specific nucleotide exchange factor son of sevenless (SOS). They have shown that the domain folds into the rest of SOS and interacts with a helical linker that connects the pleckstrin-homology (PH) and Dbl-homology (DH) domains to the catalytic domain of SOS. Results suggested that the histone domain is a potential mediator of membrane-dependent activation signals<sup>44</sup>.

**Building homo-oligomers of proteins**—Examples are the construction of the human p53-controlled ribonucleic reductase (p53R2) homodimer, which was used to explain mutations that cause mitochondrial DNA depletion<sup>45</sup>, and the modeling of an L-type Ca<sup>2+</sup> channel used for the characterization of interactions with 1,4-dihydropyridines<sup>46</sup>. Determination of homo-oligomeric proteins by docking from monomeric structures solved by NMR spectroscopy is frequently required, because NMR determination of multimers by NMR is far from simple due to the lack of chemical shift perturbation data and the difficulty of obtaining a sufficient number of intermonomer distance restraints. For example, Zweckstetter and co-workers<sup>47</sup> determined the solution structure of the 15.4 kDa homodimer CylR2, the regulator of cytolysin production from *Enterococcus faecalis*, by combining the available experimental information with docking.

**Protein-peptide docking**—Although in the present form ClusPro is not an appropriate tool for docking very flexible peptides to proteins, the server can be used if some information on the protein-bound structure of the peptide is available. For example, to study activating mutations of STAT5B in lymphomas, Chan and co-workers<sup>48</sup> docked a phosphorylated self-peptide to the homology model of the SH2 domain of a STAT5B mutant. In spite of docking a flexible peptide to a homology model, which may have some structural deviations from the X-ray structure, this analysis was feasible, because highly homologous templates of the SH2 domain with bound peptides were available, and provided both a peptide structure and the key binding residues of STAT5B.

**Docking of homology models**—The ClusPro server is frequently used for docking homology models. For example, Williams and co-workers<sup>49</sup> studied the specificity of binding of KirCII, a trans-acting acyltransferase, to a panel of acyl carrier proteins, by docking a homology model of KirCII. In addition to ClusPro, the author used the PatchDock server<sup>50</sup>, and considered the convergence of the two methods on very similar models as part of the validation. Steeland *et al.*<sup>51</sup> studied the binding of small single domain antibodies to human tumor necrosis factor receptor (TNFR) using homology models of both proteins. Results were used for the design of selective inhibitors of the TNF/TNFR interaction<sup>51</sup>. However, while we explored relatively well the performance of ClusPro for docking X-ray structures<sup>16</sup>, we have somewhat limited experience with homology models, and our incomplete analysis suggests that even moderate errors in key side chains or loops may substantially reduce the accuracy of docking results. Therefore we have recently collected a benchmark set of proteins to facilitate the development of methods that integrate homology modeling and docking<sup>52</sup>.

## Experimental Design

We describe here a number of advanced options for use in ClusPro as follows.

**Structure Modification**—Structure modification is suggested if any of the proteins has an unstructured or uncertain terminal region. Such regions may occur as the result of chemical tagging in the purification process, or may be created by homology modeling programs due to the lack of templates for the given region. Removal of such regions is frequently advantageous, because they can interfere with rigid body docking. When using this option, the server removes terminal residues until a regular secondary structure (alpha helix, extended strand, 3-helix,  $\pi$ -helix, or a hydrogen bonded turn) would be reached within 2 amino acid residues along the sequence. The modification can be requested for the receptor, the ligand, or both, and applies to both ends of the chain.

**Attraction and Repulsion**—If experimental information is available that indicates certain residues are involved in binding, whereas other residues remain surface accessible upon complex formation, this can be used to influence the results of the docking by setting attraction and repulsion, respectively, on these residues. To specify attraction, the user must enter the residues in one or both sides of the interface that are believed to participate in the binding. In the docking calculations an extra attractive force is applied to the selected residues. Repulsion can be specified similarly, by selecting a number of residues that are not expected to be in the interface, and hence are the origins of repulsive interactions in the docking. Alternatively one can upload a protein data bank (PDB) file containing the residues that are not supposed to be in the interface and hence are “masked” during the docking calculations by adding a repulsive force component. The masking files specify the amino acids that do not participate in intermolecular interaction, and hence substantially restrict the conformational space that can be occupied by the complex.

**Restraints**—We have recently added to ClusPro the option to define pairwise distance restraints. Such restraints can be derived, e.g., from NMR Nuclear Overhauser Effect (NOE) experiments or by chemical crosslinking. A pairwise distance restraint can be defined by two sets of atoms,  $S_1$  and  $S_2$  and a distance range,  $d_{\min}$  to  $d_{\max}$ . The restraint is considered satisfied if there are at least one atom in  $S_1$  and at least one atom in  $S_2$  such that the distance between them falls in this range. While the implementation allows for arbitrary sets of atoms to be used to define a restraint, most frequently these involve a single atom or residue on each side of the interface. Given multiple restraints, users may want to require a certain number of restraints out of a group to be satisfied. In addition, restraints may be based on sources with varying reliability, requiring different cutoff values. Our implementation allows for grouping restraints into restraint groups, and restraint groups into restraint sets. Restraint groups are considered satisfied when more than a user specified number of restraints in the group are satisfied, and a restraint set is satisfied when more than a user specified number of its groups are satisfied. When a restraint set is provided, ClusPro will only report solutions that satisfy the restraints. To do this efficiently, for each rotation we first generate the set of translations that satisfy each individual restraint, called the feasible translation set for the particular restraint. We then consider the intersection of feasible translation sets for the restraints in each restraint group, and select the translations that appear more often than the



cutoff for the restraint group. The selected feasible translation sets for each restraint group are merged in a similar way to generate the feasible translation set for an entire restraint set. Providing restraints can actually decrease the running times, since only the van der Waals interaction energy is computed for each feasible translation, and translations that result in unacceptable clashes are skipped. After selecting the solutions that satisfy the restraints, 1000 structures with the lowest PIPER energies are clustered and minimized as customary in ClusPro.

**The Others Mode**—The Others Mode uses a special scoring scheme to target the complexes that are classified as “others” in the established protein-protein docking benchmark<sup>22–25,53</sup>. The motivation for developing a special scoring function is their diverse nature and their generally limited shape and electrostatic complementarity. To overcome these challenges we use three different sets of weighting coefficients, generate 500 structures with each, and cluster the resulting 1500 conformations (Table 1). This is the only case when we consider 1500 rather than 1000 docked structures. In view of the assumed weaker shape complementarity and higher structural uncertainty we reduce the coefficients of  $E_{attr}$  and  $E_{elec}$ , and use three different values for the coefficient of  $E_{DARS}$ . The 1500 structures are mixed and clustered together, and the centers of well-populated clusters are considered as models of the complex. Our docking studies<sup>15</sup> confirm that the strategy substantially improves the overall success rates for the “other” type of complexes, in agreement with the assumption that the interaction energy of these complexes cannot be well described by a single set of coefficients.

**The Antibody Mode**—In the “Antibody Mode” PIPER uses a potential developed for docking antibody and antigen pairs.<sup>54</sup> Analysis of antibody–protein antigen complexes has revealed inherent asymmetry between the two sides of the interface. Specifically, phenylalanine, tryptophan and tyrosine residues highly populate the paratope of the antibody but not the epitope of the antigen. Since this asymmetry cannot be adequately modeled using the symmetric pairwise potential generally used in PIPER, we have removed the usual assumption of symmetry. Interaction statistics were extracted from antibody–protein complexes under the assumption that the interaction preferences of an atom of a particular type on the antibody differ from the preferences of an atom of the same type on the antigen. The use of the new potential significantly improved the performance of docking for antibody–protein antigen complexes<sup>54</sup>. The method allows for the masking of non-paratope residues, either by using automated selection or by providing a masking file as described for the Attraction and Repulsion option.

**Multimer Docking**—A subclass of interactions between proteins is the case where two or more identical proteins interact to form a homomultimer. The construction of such multimers is frequently required, because a number of proteins have been solved as monomers but exist in a homomultimeric state in vivo. We have developed a special mode in our docking where we limit the rotations sampled by the docking algorithm to rotations that satisfy either  $C_2$  or  $C_3$  symmetry, respectively, for dimers and trimers<sup>55</sup>. The updated method we currently use is similar to the one developed by the Weng lab and implemented in the M-ZDOCK program<sup>56</sup>, with some differences that result in a slightly simpler

algorithm. The difference is that Pierce et al.<sup>56</sup> rotate the receptor protein in the process of generating symmetric structures, whereas ClusPro rotates the coordinate system. Since the new method has not been published, we provide the steps of the algorithm here. (1) Center the receptor (the input monomer) at the origin. (2) Select values for the Euler angles  $\psi$  and  $\theta$  to define an axis of rotation. (3) Copy the receptor, and rotate it by  $360^\circ/N$  around the axis of rotation to create the ligand, where  $N=2$  for a dimer and  $N=3$  for a trimer. (4) Discretize both the ligand and receptor, with a grid spacing of 1 Å. (5) Perform FFT search in the plane perpendicular to the axis of rotation for the best scoring multimer position. Since the search space is now restricted to  $S^2$  rather than the rotational group  $S^3$ , a set of points uniformly distributed on the sphere is generated using the S2 sequence code from <https://mitchell-lab.biochem.wisc.edu/SOI/index.php>. This set of points is used as the basis for symmetric rotations of 180 degrees (dimer) and 120 degrees (trimer). Symmetry is enforced during docking by considering only translations within 2Å of the plane defined by the axis of rotation and passing through the center of rotation. (6) Repeat Steps 2 through 5 for all  $\psi$  and  $\theta$  values on a grid with 5 degrees step size. The 1000 lowest energy structures are retained and clustered as in the general ClusPro docking.

**SAXS Profile**—Small Angle X-ray Scattering (SAXS) experiments are based on observing the X-ray scattering of a macromolecule in solution at different scattering angles, resulting in a one-dimensional scattering profile. While the SAXS profile provides information about the shape and size of the molecule<sup>57</sup>, the amount of such information is much lower than the one that can be obtained by X-ray crystallography, and on its own does not provide atom-level resolution. This makes docking a natural complement to SAXS for the determination of complex structures. ClusPro can account for SAXS experimental data by retaining a number of docked structures that best agree with the provided SAXS profile<sup>58</sup>. These structures are then ranked by the PIPER scoring function and clustered as usual in ClusPro, thus the information from SAXS is used only to filter the structures generated by PIPER and the search is not biased by modifying the scoring function as done in a number of other methods that account for SAXS results<sup>59,60</sup>. The advantage of this approach is that the docked structures are restrained rather than determined by the SAXS data, and hence the method can be used even with moderately informative SAXS profiles<sup>58</sup>.

**Heparin Ligand**—Many proteins bind the polysaccharides heparan sulfate and heparin. Heparan sulfate (HS) consists of alternating hexuronic acid and D-glucosamine disaccharide units. Variations in sulfation and hexuronic acid structure result in various HS molecules<sup>61–64</sup>. Heparin, a member of the HS family, consists of highly-sulfated disaccharides, and is frequently used as a model compound in studies of protein-HS interactions<sup>65</sup>. Since crystallizing protein-heparin complexes for structure determination is generally difficult, docking can be a useful approach for understanding specific interactions, and hence we have extended PIPER and ClusPro to heparin docking<sup>66</sup>. The method generates and evaluates close to a billion poses of a heparin tetrasaccharide. The docked structures are clustered using pairwise RMSD as the distance measure. However, since we use a generic heparin structure as a probe, and since there are not enough protein-heparin complex structures to improve the interaction potential, clustering the docked heparin structures and selecting the clusters with the highest protein-ligand contacts predict only the

heparin binding sites rather than bound heparin poses<sup>66</sup>. Nevertheless, the cluster centers can provide starting points for further refinement of heparin positions using methods that account for flexibility, e.g., molecular dynamics.

### Comparison to existing methods

Table 2 shows a classification of direct protein-protein docking methods based on the amount of the information that is required in addition to the structures of the component proteins<sup>8</sup>. Each class of methods has its own strengths and limitations. Global methods are the most useful when no *a priori* information on the complex is available, and hence the entire 6D conformational space must be sampled. Since such methods use rigid body approximation, they allow only for moderate conformational change upon binding, which is a major limitation. The medium-range methods such as RosettaDock<sup>67</sup> and ATTRACT<sup>68</sup> can sample only selected regions of the conformational space and hence require some information on the putative complex, but the Monte Carlo or similar search algorithms can account for some level of flexibility, primarily in side chains. Therefore these methods are particularly useful when side chain conformations are very uncertain, for example when one of the component protein structures is a homology model, but some information is available on the structure of the complex to identify the region of interest in the conformational space. Finally, restraint based docking, exemplified by the program HADDOCK (High Ambiguity Driven biomolecular DOCKing)<sup>69,70</sup>, incorporates interaction restraints into the scoring function to guide the search toward regions of the conformational space in which the restraints are satisfied. The method can work very well if a sufficient number of correct restraints are available, but performance may deteriorate without them.

The quality of docking methods is continuously monitored by CAPRI (Critical Assessment of Predicted Interactions), the ongoing communitywide experiment devoted to protein docking<sup>71</sup>. In the CAPRI challenge, initiated in 2003, participating research groups and automated servers are given prediction targets, each being an unpublished experimentally determined structure of a protein-protein complex. Given the atomic coordinates of the component proteins or of their homologues, the participants have to model the complexes. Servers must submit results within two days, whereas human predictor groups have several weeks and can use any available information. Each group can submit ten predictions for each target. The submitted models are evaluated by independent assessors. For each group or server, the assessors report the number of targets with acceptable or better quality predictions, and also note the number of targets for which highly accurate (\*\*\*) or medium accuracy (\*\*\*) models were submitted. These categories have been defined on the basis of the fraction of native contacts, the backbone root mean square deviation of the ligand (LRMSD) from the reference ligand structure after superimposing the receptor structures, and the backbone RMSD of the interface residues (IRMSD). The calculation of these measures and the exact definitions of categories were given in the first CAPRI evaluation paper<sup>72</sup>. Although the CAPRI quality score is based on a number of characteristics including several RMSD measures and the predicted fraction of native contacts, for simplicity we focus on LRMSD, and note that for the highly accurate, medium accuracy, acceptable, and incorrect models the ligand LRMSD is given by  $LRMSD < 1\text{\AA}$ ,  $1\text{\AA} < LRMSD < 5\text{\AA}$ ,  $5\text{\AA} < LRMSD < 10\text{\AA}$ , and  $LRMSD > 10\text{\AA}$ , respectively.

Since its release in 2004, ClusPro has consistently been the best automated server in the CAPRI challenge. Table 3 shows the performance of the four best servers based on the results of the last three CAPRI evaluation meetings in 2009<sup>36</sup>, 2013<sup>37</sup>, and 2016. These results confirm that the global systematic sampling of the entire conformational space as performed by ClusPro is useful when essentially *no a priori* information on the structure of the target complex can be used, which is generally the case for servers that must submit results within two days, directly produced by the server. The only freedom we had using ClusPro was the choice of the energy coefficient set. The comparison of different methods is more complex if the performance of both “human” predictor groups and servers is considered (Table 4). Because predictors can use additional information, medium-range methods such as ATTRACT by Zaccharias<sup>68,73</sup>, SWARMDOCK by Bates<sup>74</sup>, and FRODOCK by the Chacon group<sup>75</sup>, recently combined with a new scoring scheme by Guerois<sup>76</sup>, performed well. The restraint based HADDOCK method by Bonvin and co-workers also provided good predictions for many targets<sup>69,70</sup>. However, good results were also obtained by several methods that, similarly to ClusPro, perform global search assuming rigid proteins, including ZDOCK<sup>53,77</sup>, GRAMM<sup>78</sup>, and PatchDock<sup>50</sup>. Our own group’s “human” predictions are based on running ClusPro, followed by refinement and selection of most likely clusters<sup>79</sup>, in some cases involving Monte Carlo minimization runs either using RosettaDock<sup>67</sup> or our own implementation of the Monte Carlo algorithm<sup>80</sup>. For a few targets the refinement improved medium accuracy structures into highly accurate ones, but generally the impact was moderate, and in 2016 we actually lost an acceptable prediction. According to Table 4, the relative performance of ClusPro has been improving over the years, and based on the 2016 results the server appears to be competitive with the best “human” predictor groups. We note that the PIPER scoring function has been recently used in an algorithm that expands FFT-based sampling to five rotational coordinates<sup>81</sup>. Working in a space with spherical coordinates defined as a manifold, the new 5D algorithm retains the accuracy of PIPER for globular proteins while providing at least 10-fold speedup. However, moderate loss of accuracy may occur for proteins that are very elongated along some direction. An additional advantage of the method is that increasing the number of correlation function terms is computationally inexpensive, which enables using even more complex energy functions as well as accounting for any number of pairwise distance constraints<sup>81</sup>.

## Limitations

PIPER performs rigid body docking in the 6D space of rotations and translations. The rigid body assumption is a good approximation for several classes of protein complexes, e.g., for most enzyme-inhibitor complexes (Supplementary Table 1). In fact, PIPER uses a “soft” potential that allows for certain steric overlaps. However, the protein-protein docking benchmark set<sup>22–25</sup> also includes a number of “difficult” targets, i.e., complexes in which at least one of the proteins substantially changes backbone conformation upon binding. Neither PIPER nor any other rigid body method can produce acceptable docked structures for such complexes. However, we note that extending ClusPro to docking short flexible peptides to proteins is in preparation.

As discussed, unless requesting the “other” mode, ClusPro yields four sets of docked structures using the scoring schemes called (1) balanced, (2) electrostatic-favored, (3)

hydrophobic-favored, and (4) van der Waals + electrostatics. In the “other” mode ClusPro generates structures using three different scoring schemes. The differences in the weighting of energy coefficients represent real differences in the biophysical forces that dominate interactions between two proteins, as the association in some complexes is driven primarily by electrostatic interactions, while in others desolvation of hydrophobic interfaces may be the main driving force. Unfortunately at present we are unable to perform automated selection of the best scoring scheme. Thus, it is left to the user to make such selection based on the assumed properties of the particular complex. If no such information is available, we suggest using the balanced option, and the “other” mode for signal transduction complexes. It is generally a good sign if the selected parameters lead to a cluster of low energy structures that is substantially more populated than the others. It is also useful if different parameter sets yield the same model. Once a model is selected, the user can also explore whether the assumption used was correct by analyzing the properties of the interface. Nevertheless, we recognize that the problem of parameter selection is not resolved, and further research is needed.

Even assuming that a parameter set is selected, ClusPro returns ten clusters of low energy structures, the center of each cluster representing a putative model of the complex. Again, the existence of a large cluster provides some level of certainty, but generally additional information is needed for reliable model selection. As mentioned, we found over 400 publications that reported models constructed by ClusPro. In many applications the models were validated by experimental techniques, including site-directed mutagenesis with NMR, calorimetry, FRET, or surface plasmon resonance<sup>44,82–101</sup>, cross-linking<sup>102–106</sup>, various spectroscopic methods<sup>107–113</sup>, X-ray scattering<sup>114</sup>, electron self-exchange reaction<sup>115</sup>, radiolytic protein footprinting with mass spectrometry<sup>39,40</sup>, hydrogen/deuterium exchange<sup>116</sup>, or intermolecular Nuclear Overhauser Effect restraints<sup>47</sup>. ClusPro results can also be used to design low resolution and hence cost effective validation experiments. We believe that such combination of computational and experimental tools is the most meaningful use of docking.

In addition to the above fundamental limitations we have several technical shortcomings that we hope remove in the future. Some of these shortcomings are as follows: (1) At this point the molecules to be docked can include only standard amino acids, nucleic acids to define an RNA molecule as the receptor, and heparin as the ligand. Thus, the docking cannot account for the presence of co-factors or other small ligands that can be important in the modulation of protein-protein interactions. (2) Similarly, we are unable to account for non-standard or modified amino acids, e.g. phosphorylation, another important factor that can affect the interactions. (3) Currently ClusPro can create only dimers and trimers in the Multimer Docking mode, and does not provide an option for constructing more complex multimers. (4) There are two limitations on size of the proteins to be docked. First, uploaded files cannot be more than 10MB each. Second, the total grid size cannot be more than roughly  $40 \times 10^6 \text{ \AA}^3$ , which corresponds to a cube of  $\sim 350 \text{ \AA}$  in each dimension. This grid has to contain both the receptor and ligand in each potential relative orientation of the two molecules.

## Availability

The server can be used without registration, but in this case the results will be publicly accessible. The advantage of registering is that the job does not show up on the website, but this options is available only to users with educational or governmental email addresses. The server provides options to view the results online, but protein visualization tools allow for more convenient analyses. We use and recommend PyMOL, which was used to demonstrate the analysis of docking results in this Protocol.

## MATERIALS

### EQUIPMENT

- A computer with internet access and a web browser. The program is also available for download by contacting <https://structure.bu.edu/contact>.
- Atomic resolution structures of biomolecules under investigation, in PDB format. The PDB ID can be used to directly fetch the structure or the structure may be uploaded from the computer.
- Access to PyMOL or similar structure viewing software is recommended but not required. PyMOL can be downloaded [www.pymol.org](http://www.pymol.org). Alternatively, you can use any molecular viewer that supports the visualization of multiple structures in one PDB file.

## PROCEDURE

**CRITICAL** The current version of ClusPro provides a very simple home page for the basic use of the server, and a number advanced options that modify the docking process. Some options can be requested by simply clicking on a label, while others require providing additional information.

### Input data for docking TIMING ~1 to 2 min

1| Locate the server at <https://cluspro.org>. ClusPro can be used without a user account or with a user account if you have an educational or governmental email address. If you decide to use the server without a user account, click ‘Use the server without the benefits of your own account’. To create an account, register on the ClusPro server website. A password will be sent to the email address and can be later changed by clicking on the label Preferences. You can also request an e-mail to be sent to you when your job is completed. If you already have a username and password, fill in the boxes and click Login. Either with or without a user account you will get to the server home screen (Fig. 2). From this page you will be able to submit a new docking job and select a number of Advanced Options.

**CRITICAL STEP** Users who run ClusPro without an account will have their results publicly accessible.

2| (Optional) Provide a job name for this submission. If you choose to leave this blank, a unique ID will be created for this field.

3| Select the type of the computer the job will use. Selecting cpu will lead it to use computer clusters at the Massachusetts Green High Performance Computing Center (MGHPC). Selecting gpu will lead it to use the graphic computing units, also located at MGHPC. Since we generally cannot predict the actual usage on these systems, selecting cpu versus gpu has no predictable consequence for the user. Currently we use the cpu version, because ample computer time is available and usually the jobs start immediately after submission.

4| Input the coordinates of the receptor structure using PDB format. Only atoms of 20 standard amino acid residues and nucleotides will be retained. All HETATM (hetero atom) records, including waters, ligands, and cofactors, will be automatically removed. There are two options for inputting a structure: use Option A to import coordinates from the PDB, or option B to upload a structure.

**(A) Import coordinates from PDB**

i) Import coordinates directly from the PDB by typing the four digit PDB ID into the field PDB ID.

**(B) Upload structure**

i) Upload a structure from your computer by clicking on the Upload PDB option under the PDB ID field. Select Browse to upload a file containing a structure in PDB format.

**CRITICAL STEP** At this point only structures containing standard amino acid residues and nucleotides will be docked. ATOM records of nonstandard amino or nucleic acids will cause an error.

**CRITICAL STEP** Since the protein considered the receptor is placed on a fixed grid whereas the protein considered the ligand is placed on a rotating and translating grid, it is computationally advantageous to consider the larger protein as the receptor.

**CRITICAL STEP** If the target is a protein-RNA complex, the RNA must be defined as the receptor rather than the ligand.

(Optional) To see the supported RNA residue and atom names click on the label RNA.

**? TROUBLESHOOTING**

5| In the Chains field, enter the protein chains of the receptor that you wish to include in the docking. List chains using their chain id and separate multiple chains with a whitespace. If no chains are specified, then all chains in the PDB file will be used for docking.

6| Input the coordinates of the ligand structure using PDB format. Only atoms of 20 standard amino acid residues or a standard heparin structure can be used. If the ligand is a protein, you can use Option A to import coordinates from the PDB, or option B to upload a structure directly as in Step 4.

**CRITICAL STEP** If you will use the Advanced Option Multimer Docking, no ligand structure needs to be specified.

**CRITICAL STEP** If you will use the Advanced Option Heparin Docking, no ligand structure needs to be specified.

### ? TROUBLESHOOTING

7| In the Chains field, enter the chains of the ligand protein that you wish to include in the docking as in Step 5.

### Selection of Advanced Options TIMING ~5 to 45 min (see below)

8| (Optional) At this point, it is possible to define additional advanced options depending on your requirements. See the following table for a summary of these options. These options are accessible by clicking the “Advanced Options” label.

Step	Option	Description
8A	Structure Modification	Removal of unstructured terminal residues
8B	Attraction and Repulsion	Setting attraction or repulsion on selected residues
8C	Restraints	Selection of pairwise distance restraints
8D	Others Mode	Selecting a special scoring scheme for “other” type complexes
8E	Antibody Mode	Selecting a special scoring scheme for docking antibody-antigen pairs
8F	Multimer Docking	Constructing homodimers or homotrimers
8G	SAXS Profile	Accounting for experimental SAXS data
8H	Heparin Ligand	Global docking of a generic heparin molecule

#### A. Selecting Structure Modification TIMING ~5 min

- i. Structure modification is suggested if any of the proteins has an unstructured terminal region. Click on Advanced Options and select Structure Modification. You can then request removal of unstructured terminal residues from the receptor, the ligand, or both.

**CRITICAL STEP** This mode can be combined with other advanced options. However, editing the PDB files manually provides greater flexibility, e.g., the potential for removing flexible loops with uncertain structure that would interfere with docking.

#### B. Selecting Attraction and Repulsion TIMING ~30 min

- i. If *a priori* experimental information indicates that certain residues are in the binding interface, you can influence the results of the docking by setting attraction on these residues. Alternatively, if some information indicates that certain residues remain surface



accessible upon binding, you can influence the results of the docking by setting repulsion on these residues. Click on Advanced Options and select Attraction and Repulsion. You will then see the entry fields for attraction and repulsion.

- ii. Attracting and repulsing residues can be selected by typing whitespace separated “chain-residue” entries, e.g., a-23 a-25 a-26 a-27, in the appropriate boxes.
- iii. (Optional) Repulsing residues can also be selected by uploading a masking file that includes residues you do not want to be in the interface. Click on the Use PDB Masking File option, and select Browse to upload a pre-constructed masking file. To generate a PDB masking file, we recommend using PyMOL. Open in PyMOL the PDB file of the protein to be docked, and click on the button marked with “S” in the lower right corner to see the sequence of the protein. You can then select the residues to avoid in the interface by clicking on their one-letter amino acid code in the sequence. This will create a selection object called “sele” on the right side of your screen that holds these residues. You can also see the selection by what residues have dots placed on them in the viewer. You can then choose File->Save Molecule from the PyMOL menu. This will give you a window where you choose to save “sele”. You can then upload this PDB file as your masking file.

**CRITICAL STEP** This mode can be combined with other advanced options.

#### C. **Selecting Restraints** TIMING ~30 to 45 min

- i. Pairwise distance restraints may be available, e.g., from NMR Nuclear Overhauser Effect (NOE) experiments or from chemical crosslinking. To account for such restraints in docking, click on Advanced Options and select Restraints.
- ii. To use this option you need to prepare a restraint file. There are two options for restraint file format: either the AIR (Ambiguous Interaction Restraints) format, which is also used by the HADDOCK program, or our JSON format. To create a restraint file in AIR format, you can download a sample AIR type restraint file from [https://cluspro.org/examples/ja026939x\\_s2.txt](https://cluspro.org/examples/ja026939x_s2.txt). To provide restraints as a JSON file, you can download a sample JSON type restraint file from [https://cluspro.org/examples/ja026939x\\_s2.json](https://cluspro.org/examples/ja026939x_s2.json). We have developed a web based restraint set generator to help the preparation of the restraint file in JSON format (see Fig. 3), available at [https://cluspro.org/generate\\_restraints.html](https://cluspro.org/generate_restraints.html).

**CRITICAL STEP** Restraints can either be a JSON file with a .json suffix, or an AIR restraint file from HADDOCK with any other suffix.

- iii. Choose the restraint file.

**D. Selecting Others Mode** TIMING ~5 min

- i. The use of Others Mode is suggested for complexes that may have less perfect shape and electrostatic complementarity than the enzyme-inhibitor complexes. Examples are the “other” type complexes in the protein-protein docking benchmark set<sup>22–25</sup>. Click on Advanced Options and select Others Mode. As described, this mode uses three different sets of energy weight coefficients and selects 500 low energy structures from each. The 1500 conformations are clustered together to select the most populated clusters.

**E. Selecting Antibody Mode** TIMING ~10 to 45 min

- i. Using the Antibody Mode is suggested for antibody-antigen docking. Click on Advanced Options, and select Antibody Mode. Click on the label Use Antibody Mode. As described, this mode will use a scoring function specifically developed for antibody-antigen complexes.
- ii. (Optional) Avoiding non-CDR (non-complementarity determining region) residues in the interface may improve the docking result. There are two options for specifying CDR-residues: either automatic or manual selection. For automatic masking of non-CDR residues, click on the label “Automatically Mask Non-CDR Regions”. For manual masking of non-CDR residues using the Attraction and Repulsion option, you need to click on the label Attraction and Repulsion. You will then see the entry fields for attraction and repulsion. Repulsing residues can be selected by typing them as whitespace separated “chain-residue” entries as shown in Step 8B(ii). Alternatively, repulsing residues can be selected by uploading a masking file generated as described in Step 8B(iii). In both cases it is necessary to identify the CDRs of the specific antibody for selecting non-CDR residues as repulsive or listed in a masking file. To find CDRs, you can consult the website <http://www.bioinf.org.uk/abs/#cdrid>, or use the annotation tools in the abYsis server at <http://www.abysis.org>. On the home page of the abYsis server, select Sequence Input Key Annotation, enter the sequence of either the heavy or the light chain of the antibody in FASTA format, and click Submit. Click on the label Numbering and Regions. The resulting page provides three different annotations of CDR regions (Chothia, ABM, and Kabat). Select

the amino acid residues that are reliably not part of any CDR, and generate the masking file as described in Step 8B(ii).

**F. Selecting Multimer Docking** TIMING ~5 min

- i. Multimer docking is special in that the receptor and ligand are the same molecule. Therefore you only need to fill in the information for your molecule in the receptor field. Similarly, if you are specifying information such as attraction and repulsion, you only need to specify this information for the receptor. Click on Advanced Options and select Multimer Docking. You can then enter either 2 or 3 subunits for either a dimer or a trimer.

**CRITICAL STEP** The current multimer mode in ClusPro supports only dimers and trimers.

**G. Selecting SAXS Profile** TIMING ~10 to 30 min

- i. If you have SAXS data, you can submit it to ClusPro for use during the docking process. The server will fit the theoretical profiles generated from the docked structures against the uploaded SAXS profile using the chi scores to filter the structures before clustering. To use SAXS data click on Advanced Options and select SAXS Profile.
- ii. To use this option you need to prepare a file containing the SAXS profile. The file should contain the SAXS results in a 3-column format. The first column should be the angle ( $q$ ) in units of  $1/\text{\AA}$ . The second column should be the scattering intensity ( $I$ ). The third column is optional, and should be the experimental error if it is available. You can download a zip file with sample inputs from <https://cluspro.org/examples/saxs.zip>. The zip file contains two pdb files, rec.pdb and lig.pdb, as well as a sample SAXS profile file, SalGEc\_new\_autoRg.dat.
- iii. Choose the file containing the SAXS profile.

**H. Selecting Heparin Ligand** TIMING ~5 min

- i. Click on the labels Advanced Option and Heparin Ligand, and then select Use Heparin as Ligand.

**Running ClusPro** TIMING ~ 1 to 8 hours

9| Click on the DOCK button to begin the docking calculation. The status of the job can be immediately checked on the Queue page, which also shows the jobs that are ahead of yours. The jobs will run in order of their submission. Each job is listed with ID number, job name, user name, and a status update. To see a detailed Status page, click on the ID of your job. The Status page shows the job ID number, job name, job status, and submission time. See Box 1 for the list of status abbreviations. The Status

page also shows small pictorial representations of the uploaded and processed input structures (Fig. 4).

**10|** If requested, an email will be sent when the job has completed or if an error occurred (see Box 2 for a listing of possible errors and their meanings). The email provides a link to the Results page or to an error message. Click the link to get to the Results page. Alternatively, locate the results under the Results tab on the server. All user results will be listed in order of ID number.

**CRITICAL STEP** We will store the result files on the server for at least 2 months. After this time, the results may be deleted.

### ? TROUBLESHOOTING

#### Analyzing results from ClusPro TIMING ~ 30–45 min

**11|** View the results by clicking on the ID number of the job under the Results tab. On the Result page you can view and download the results of your docking. The page (see Fig. 5) starts with the job name. We recall that a single job actually performs 4 docking calculations with 4 different sets of energy parameters. By default the page shows results for the Balanced set (Table 1). Results for electrostatic-favored, hydrophobic-favored, or van der Waals + electrostatics sets can be viewed by clicking on the corresponding label. For each parameter set the Result page shows small pictures representing the top 10 models, but the user can request more models (up to 30 or less if fewer than 30 clusters have been created). Clicking on the number above a picture will download that model as a PDB file for your viewing. You can also download all displayed models or all models for all coefficients (see Step 13). Optionally, you can click on the label Job details at the top of the Results page opens the Status page that was available from the Queue page while the job was running. The Status page now shows the job ID number, job name, job status, submission time stamp, and the error message if there was an error. The page also shows pictorial representations of the uploaded and processed input structures.

**CRITICAL STEP** In the Heparin Ligand Mode, the Results page also shows a small picture of the target protein in surface representation with the putative heparin binding site indicated in atomic charge colors (blue for N, red for O, and white for C), whereas the non-contact region is dark. In addition, the page has the label Atom Contacts. Clicking on the label downloads the file atom\_contacts.csv, which lists the contact atom pairs on the two sides of the heparin-protein interface.

**12|** View model scores by clicking on the corresponding label on the results page. Depending on the selected set of energy parameters, the page shows the actual weighting coefficients of the PIPER energy terms, and a table that lists the requested number of clusters of docked structures in the order of cluster size (Fig. 6). For each cluster, the table shows the size (i.e., the number of docked structures), the PIPER energy of the cluster center (i.e., the structure that has the highest number of neighbor structures in the cluster), and the energy of the lowest energy structure in the cluster.

**CRITICAL STEP** Although we show energy values, it is emphasized that model selection is based on cluster size rather than on energy. In fact, the energy calculated by PIPER does not directly relate to binding affinity. However, low energy regions tend to generate large clusters of docked structures, and the size of a cluster is approximately proportional to its probability, and thus the energy landscape indirectly determines the most likely conformation of the complex.

**CRITICAL STEP** In the Other Mode, the table shows the size for each cluster (i.e., the number of docked structures). No energy values are shown, as the clusters may include structures obtained using different energy functions.

**13|** (optional) From the Results page, download PDB files of the displayed models, or also the models for all energy coefficients. This latter option downloads a large .tar file that includes 4 times 30 PDB files (or fewer if for some of the energy coefficients ClusPro generates fewer than 30 clusters). Each PDB file includes a structure at the center of a cluster. The naming conventions are as follows: files model.000.00.pdb through model.000.29.pdb include the structures generated using the balanced parameter set; files model.002.00.pdb through model.002.29.pdb are from the electrostatics favored calculations; files model.004.00.pdb through model.004.29 are hydrophobicity-favored; and finally, files model.006.00.pdb through model.006.29 have the structures obtained using the parameters we have defined as van der Waals + electrostatics. Since the models generated by the four different energy parameters have different names, structures selected from the four sets can also be opened simultaneously in PyMol. To do this, click on the first selected model to open it in PyMOL, and use the PyMOL command “load filename” to add further structures.

**CRITICAL STEP** In the Others Mode, the PDB files are named model.003.00.pdb through model.003.29.pdb.

**CRITICAL STEP** In the Antibody Mode, the PDB files are named model.000.00.pdb through model.000.29.pdb.

**14|** Use the appropriate software to visualize the protein structure files. PyMOL provides a convenient tool for the visualization of the structures generated by ClusPro. When you open a downloaded structure, e.g., model.000.00.pdb in PyMOL, it actually creates two structures named rec.pdb for the receptor and lig.000.00.pdb for the ligand. You can add further models using the PyMOL command load fname, where fname is the name of the file to be loaded. In these cases only a new ligand is opened, e.g., load model.000.01.pdb will create the structure lig.000.01.pdb, as the receptor is named rec.pdb in all result files. In test cases you can also load the X-ray structure of the complex (Fig. 7). Further examples of visualization by PyMOL are shown in the Anticipated Results section.

## ? TROUBLESHOOTING

### TIMING

Steps 1–7, inputting the coordinates of the component proteins: ~ 1 to 2 min

Step 8 (A), selecting Structure Modification: ~ 5 min

Step 8 (B), selecting Attraction and Repulsion: ~ 30 min

Step 8 (C), selecting Restraints: ~30 to 45 min

Step 8 (D), selecting Others Mode: ~5 min

Step 8 (E), selecting Antibody Mode: ~10 to 45 min

Step 8 (F), selecting Multimer Docking: ~ 5 min

Step 8 (G), selecting SAXS Profile: ~ 10 to 30 min

Step 8 (H), selecting Heparin Ligand: ~ 5 min

Steps 9–10, running ClusPro: ~ 1 to 8 hours; depends on the sizes of the proteins and the number of jobs in the queue. On average, results are returned in less than 4 hours.

Steps 11–14, analysis of ClusPro results for selecting putative models of the complex, considering result obtained by four different scoring functions: ~ 30–45 min

**TROUBLESHOOTING**—Troubleshooting advice can be found in Table 5.

## ANTICIPATED RESULTS

### Docking an enzyme-inhibitor pair

The first enzyme-inhibitor target in the protein-protein docking benchmark<sup>25</sup> is docking the x-ray structure of soybean trypsin inhibitor (PDB ID 1BA7) to the X-ray structure of porcine trypsin (PDB ID 1QQU). As shown in Fig. 2, we select the enzyme as the receptor. Since the only chain in the file is A, the Chain ID can be provided or omitted. The PDB file 1BA7 has two copies of the inhibitor structure, of which chain B is slightly more complete (169 rather than 165 residues) and hence it is used as the ligand, providing B as the Chain ID. For both proteins we import the coordinates from the PDB. As soon as the job shows up on the Queue page, we can check its status. The Status page shows the job ID number, job name, job status, submission time stamp, and PDB ID. The page also shows that both files have been processed by the server, and only one of the chains was used for the ligand (Fig. 3). Once the job is completed, by default the Results page (Fig. 4) shows pictures of the 10 models (i. e., the centers of the 10 most populated clusters). Clicking on the label View Model Scores opens a page that shows the weighting coefficients of the PIPER energy terms and a table of cluster scores (Fig. 5). For each cluster, this table shows the size (i.e., the number of members), the weighted energy score of the cluster center (i.e., the structure that has the highest number of neighbor structures in the cluster), and the energy score of the lowest energy structure in the cluster. Note that the top cluster, with 253 members, is substantially larger than the second largest cluster (122 members), indicating a well-defined set of encounter complexes. Such outcome generally provides some level of assurance that the clustering of docked structures occurs in the neighborhood of the native state. We suggest downloading selected models for visualization in PyMOL. Fig. 6 shows the receptor, porcine trypsin, as grey surface. The ligand at the center of the largest cluster (lig.000.00.pdb) is shown as cyan cartoon. We also loaded the X-ray structure of the enzyme-inhibitor complex (PDB ID 1AVX) into PyMol, aligned it with the structure of the receptor, and show the native pose of the inhibitor as magenta cartoon in Fig. 6. The agreement is good, resulting in

the Ca IRMSD of 3.3 Å. This is a typical outcome for many enzyme-inhibitor complexes, because both proteins are fairly rigid, the conformational change upon association is small, and the two molecules have excellent shape complementarity. Although the results are shown for the balanced set of energy coefficients, the top model remains the same for the electrostatic favored energy expression, with an even larger difference in the populations of the largest and second largest clusters, with 268 and 100 members, respectively. The stability of the model upon variation in the energy parameter further increases confidence in its correctness.

### Predicting an “others” type complex

Docking the component proteins of “other” type of complexes is the same as the basic ClusPro run, apart from selecting Others Mode in Advanced Options. As an example, we dock the X-ray structure of the cytoplasmic domain of the unphosphorylated type 1 TGF-beta receptor (PDB ID 1IAS, Chain A, 342 residues), which actually was defined here as the ligand, to the X-ray structure of the FK506 binding protein (FKBP, PDB ID 1D6O, Chain A, 107 residues). Fig. 8 shows the PyMOL visualization of the results. As in the previous example, the receptor (FKBP in this case) is shown as grey surface, the ligand lig.003.00.pdb at center of the largest cluster represented by the cyan cartoon, and the ligand extracted and superimposed from the complex (PDB ID 1B6C) is shown in magenta. We note that the IRMSD between free and bound structures of the TGF-beta receptor is 1.9 Å, and the structural differences between free and bound states are clearly seen in Fig. 8. Nevertheless, docking of the unbound structures yields a large cluster with 267 members, whereas the second largest cluster has only 100 members, again suggesting a stable native complex, occupying a well-defined energy well on the energy landscape. Indeed, the interface is predicted well, resulting in the IRMSD value of 2.96 Å, in spite of the overall conformational change in the ligand.

### Antibody-antigen docking

We docked the X-ray structure of the extracellular domain of the human tissue factor (PDB ID 1TFH) to the unbound X-ray structure of the FAB domain of the inhibitory antibody 5G9 (PDB ID 1FGN). Both the heavy and light chains were used to represent the receptor. We first used the Antibody Mode with automatic masking of non-CDR regions, selected in Advanced Options. As shown in Fig. 9, we have an acceptable model, lig.000.05.pdb, with IRMSD of 4.7 Å from the ligand position in the native complex (PDB ID 1AHW). However, this model is the center of the 6<sup>th</sup> largest cluster with only 48 members, whereas the largest cluster has 118 members, but much larger IRMSD from the native. Thus, while docking produces a good model, without *a priori* information we would not have been able to identify it. We also prepared a sequence-specific masking file to place repulsion on the non-CDR residues as suggested in Step 8E(ii). The abYsis server provides three different annotations of CDR regions (Chothia, ABM, and Kabat). We put repulsion only on residues that were not considered to be in a CDR by any of these annotations. The manual masking substantially improved the results, and a model (lig.000.01.pdb) with the IRMSD of 4.8 Å moved to the 2<sup>nd</sup> most populated cluster with 73 members. The largest cluster, with 190 members, was still a false positive, emphasizing the uncertainty of docking results for

antibody-antigen complexes. Nevertheless, the example also shows that manual preparation of a sequence-specific masking file can be useful.

### Accounting for pairwise distance restraint

As an example we present constructing the complex from the X-ray structures of the *Escherichia coli* glucose-specific phosphocarrier protein IIAGlc (E2A, PDB ID 1F3G, considered the receptor) and the signal transducing protein HPr (PDB ID 1POH, considered the ligand)<sup>117</sup>. ClusPro works very well for this problem even without any restraint, as the center of the second ranked cluster, has the IRMSD of 3.78 Å from the native state of the complex (PDB ID 1GGR). As shown in in Fig. 10A, this model (cyan cartoon) is oriented similarly to the native ligand (magenta cartoon), but its position is slightly shifted. However, this second cluster has only 221 members, whereas the top cluster has 424, but the ligand structures in this cluster are rotated relative to the native structure, resulting in the IRMSD of more than 20 Å. NOE (Nuclear Overhauser Effect) measurement were available for this complex<sup>117</sup>, resulting in 20 intermolecular distance restraints. Docking with these restraints was also performed by HADDOCK using a set of AIRs<sup>69</sup> ([https://cluspro.org/examples/ja026939x\\_s2.txt](https://cluspro.org/examples/ja026939x_s2.txt)). We have prepared a JSON file using [https://cluspro.org/generate\\_restraints.html](https://cluspro.org/generate_restraints.html), the on-line restrain set generator tool (see Fig. 3). The JSON file is shown at [https://cluspro.org/examples/ja026939x\\_s2.json](https://cluspro.org/examples/ja026939x_s2.json). Accounting for the restraints further improves the result, and the center of the largest cluster, with population 268, is shifted to 2.88 Å IRMSD. Indeed, as shown in Fig. 10B, the model (shown as blue cartoon) is now turned by a few degrees around an axis perpendicular to the middle of the receptor binding site, and this yields only small IRMSD. While accounting for the restraints improved the result, the identification of the best structure without additional information still would be difficult, as the second largest cluster has 235 members, thus it is almost as large as the top cluster.

### Accounting for experimental SAXS data

Target 58 of the CAPRI docking challenge was determining the complex between the salmon cold-active goose-type lysozyme and the *e. coli* PliG lysozyme inhibitor using SAXS data as restraints<sup>118</sup>. A model of salmon lysozyme was built using the MODELLER v9.0<sup>119</sup> program using the template of black swan goose lysozyme (PDB ID 1GBS, 57.8% sequence identity). Aromatic residues (Tyr, Phe, and Trp) that were not present in the template were placed in the most probable non-clashing rotamer positions. Other side chains that were not present in the template were not modeled since they have uncertain localization. The input files for this target are given at <https://cluspro.org/examples/saxs.zip>. The zip file contains two PDB files, rec.pdb, which provides coordinates of the model for the receptor, the salmon goose-type lysozyme, and lig.pdb, the *e. coli* PliG lysozyme inhibitor (PDB ID 4DY3). The zip file also includes the relevant SAXS profile file, SalGEc\_new\_autoRg.dat. Details of the SAXS profile were described in our report on extending ClusPro to accounting for SAXS data<sup>58</sup>.

ClusPro provided a near-native model even without the use of the SAXS restraints, but it was the center of the 6<sup>th</sup> largest cluster. Essentially the same near-native model was obtained when accounting for the SAXS data (see cyan cartoon in Fig. 11), but now as the center of



the 3<sup>rd</sup> largest cluster. The conformation of the ligand, the *e. coli* PliG lysozyme inhibitor, is shown as magenta cartoon for reference (Fig. 11). The Others Mode can be combined with the SAXS Profile Mode, and this was used for the calculation.

### Constructing a dimer in Multimer Mode

As an example of multimer modeling, we constructed the dimer of the sugar aminotransferase AtmS13 from *Actinomadura mellioura*. Since no monomeric structure of the protein was available, we have first constructed a homology model. The sequence of AtmS13 from *Actinomadura mellioura* (Uniprot entry Q0H2X1) was downloaded from the Uniprot website at <http://www.uniprot.org> in FASTA format. To build a homology model using the HHpred server <http://toolkit.tuebingen.mpg.de/hhpred>, the amino acid sequence was pasted into the box provided on the HHpred home page, and submitted for template selection and alignment using the pdb70 database for hidden Markov model (HMM) construction. All other parameters were left at their default values. HHpred returned a large number of potential template sequences aligned to the AtmS13 sequence, with the sugar aminotransferase CalS13 from *Micromonospora echinospora* (PDB ID 4ZAS) at the top of the list with the highest similarity score (63% sequence identity). The next step was clicking on the label Create model, which provided the options of either creating a model from the manually selected templates or automatically selecting the best templates. We have used the first option. Notice that building a homology model by HHpred requires the user to have a license to the MODELLER program<sup>119</sup>. The license is freely available for academic users and easily obtainable at <http://salilab.org/modeller/registration.shtml>. Entering the coordinates of the homology model constructed as the receptor, we used the Multimer Docking option of ClusPro with the number of subunits set to 2 and left all the other options at their default values. Once the docking was completed, the resulting models were compared to the available crystal structure of AtmS13 homodimer (PDB ID 4XAU, chains A and B). In agreement with our prior observation for dimers, the balanced coefficient provided the best ranking of near-native models. In particular, a near native docking model with IRMSD of 2.62 Å was ranked second (Fig. 12).

### Finding heparin binding sites

As an example, we docked the heparin tetramer probe to the ligand-free structure of the basic fibroblast growth factor (PDB ID 1BFG) that has also been co-crystallized with a heparin hexamer (PDB ID 1BFC). The docking and clustering yield eight clusters, with the two largest clusters having very similar populations, 387 and 378, respectively. Fig. 13 shows the center of the second largest cluster (cyan sticks), since some of the sulfate groups interacting with the protein have better overlap with the bound heparin (magenta sticks) than the sulfate groups in the top model, which is reoriented tail-to-head relative to the models shown. However, all eight models interact with the relatively deep pocket in the heparin binding site. Since we dock a generic heparin structure rather than the specific ligand, it is no surprise that the results have somewhat limited accuracy, but generally identify the most likely region of heparin binding.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This investigation was supported by grants R35 GM118078 and R01 GM061867 from the National Institute of General Medical Sciences, and grants DBI 1147082, DBI 1458509, and AF 1527292 from the National Science Foundation.

## References

1. Ito T, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*. 2001; 98:4569–4574. [PubMed: 11283351]
2. Gavin AC, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*. 2002; 415:141–147. [PubMed: 11805826]
3. Ho Y, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*. 2002; 415:180–183. [PubMed: 11805837]
4. Ewing RM, et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*. 2007; 3:89. [PubMed: 17353931]
5. Smith GR, Sternberg MJ. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol*. 2002; 12:28–35. [PubMed: 11839486]
6. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*. 2002; 47:409–443. [PubMed: 12001221]
7. Ritchie DW. Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci*. 2008; 9:1–15. [PubMed: 18336319]
8. Vajda S, Kozakov D. Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol*. 2009; 19:164–170. [PubMed: 19327983]
9. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. *J Mol Biol*. 2003; 332:989–998. [PubMed: 14499603]
10. Sinha R, Kundrotas PJ, Vakser IA. Protein docking by the interface structure similarity: how much structure is needed? *PLoS One*. 2012; 7:e31349. [PubMed: 22348074]
11. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res*. 2004; 32:W96–99. [PubMed: 15215358]
12. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*. 2004; 20:45–50. [PubMed: 14693807]
13. Comeau SR, et al. ClusPro: performance in CAPRI rounds 6–11 and the new server. *Proteins*. 2007; 69:781–785. [PubMed: 17876812]
14. Kozakov D, et al. Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13–19. *Proteins*. 2010; 78:3124–3130. [PubMed: 20818657]
15. Kozakov D, et al. How good is automated protein docking? *Proteins*. 2013; 81:2159–2166. [PubMed: 23996272]
16. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*. 2006; 65:392–406. [PubMed: 16933295]
17. Katchalski-Katzir E, et al. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A*. 1992; 89:2195–2199. [PubMed: 1549581]
18. Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol*. 1997; 272:106–120. [PubMed: 9299341]
19. Mandell JG, et al. Protein docking using continuum electrostatics and geometric fit. *Protein Eng*. 2001; 14:105–113. [PubMed: 11297668]

20. Chen R, Weng ZP. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins*. 2002; 47:281–294. [PubMed: 11948782]
21. Kozakov D, Clodfelter KH, Vajda S, Camacho CJ. Optimal clustering for detecting near-native conformations in protein docking. *Biophys J*. 2005; 89:867–875. [PubMed: 15908573]
22. Chen R, Mintseris J, Janin J, Weng Z. A protein-protein docking benchmark. *Proteins*. 2003; 52:88–91. [PubMed: 12784372]
23. Mintseris J, et al. Protein-Protein Docking Benchmark 2.0: an update. *Proteins*. 2005; 60:214–216. [PubMed: 15981264]
24. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein-protein docking benchmark version 3.0. *Proteins*. 2008; 73:705–709. [PubMed: 18491384]
25. Hwang H, Vreven T, Janin J, Weng Z. Protein-protein docking benchmark version 4.0. *Proteins*. 2010; 78:3111–3114. [PubMed: 20806234]
26. Vajda S, Camacho CJ. Protein-protein docking: is the glass half-full or half-empty? *Trends Biotechnol*. 2004; 22:110–116. [PubMed: 15036860]
27. Vajda S. Classification of protein complexes based on docking difficulty. *Proteins*. 2005; 60:176–180. [PubMed: 15981248]
28. Yerushova A, Jain S, LaValle SM, Mitchell JC. Generating Uniform Incremental Grids on SO(3) Using the Hopf Fibration. *Int J Robot Res*. 2010; 29:801–812.
29. Chuang GY, Kozakov D, Brenke R, Comeau SR, Vajda S. DARS (Decoys As the Reference State) potentials for protein-protein docking. *Biophys J*. 2008; 95:4217–4227. [PubMed: 18676649]
30. Gilson MK, Honig B. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins*. 1988; 4:7–18. [PubMed: 3186692]
31. Brooks BR, et al. Charmm - a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem*. 1983; 4:187–217.
32. Lorenzen S, Zhang Y. Identification of near-native structures by clustering protein docking conformations. *Proteins*. 2007; 68:187–194. [PubMed: 17397057]
33. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. *Proc Natl Acad Sci U S A*. 1998; 95:11158–11162. [PubMed: 9736706]
34. Mendez R, Leplae R, Lensink MF, Wodak SJ. Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. *Proteins*. 2005; 60:150–169. [PubMed: 15981261]
35. Lensink MF, Mendez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins*. 2007; 69:704–718. [PubMed: 17918726]
36. Lensink MF, Wodak SJ. Docking and scoring protein interactions: CAPRI 2009. *Proteins*. 2010; 78:3073–3084. [PubMed: 20806235]
37. Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. *Proteins*. 2013; 81:2082–2095. [PubMed: 24115211]
38. Lensink MF, et al. Prediction of homo- and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. *Proteins*. 2016
39. Kamal JK, Benchaar SA, Takamoto K, Reisler E, Chance MR. Three-dimensional structure of cofilin bound to monomeric actin derived by structural mass spectrometry data. *Proc Natl Acad Sci U S A*. 2007; 104:7910–7915. [PubMed: 17470807]
40. Kamal JK, Chance MR. Modeling of protein binary complexes using structural mass spectrometry data. *Protein Sci*. 2008; 17:79–94. [PubMed: 18042684]
41. Luxan G, et al. Mutations in the NOTCH pathway regulator MIB1 cause left ventricular noncompaction cardiomyopathy. *Nat Med*. 2013; 19:193–201. [PubMed: 23314057]
42. Tran K, et al. Vaccine-elicited primate antibodies use a distinct approach to the HIV-1 primary receptor binding site informing vaccine redesign. *Proc Natl Acad Sci U S A*. 2014; 111:E738–E747. [PubMed: 24550318]
43. Schwede T, et al. Outcome of a workshop on applications of protein models in biomedical research. *Structure*. 2009; 17:151–159. [PubMed: 19217386]

44. Sondermann H, Nagar B, Bar-Sagi D, Kuriyan J. Computational docking and solution x-ray scattering predict a membrane-interacting role for the histone domain of the Ras activator son of sevenless. *Proc Natl Acad Sci U S A*. 2005; 102:16632–16637. [PubMed: 16267129]
45. Bourdon A, et al. Mutation of RRM2B, encoding p53-controlled ribonucleotide reductase (p53R2), causes severe mitochondrial DNA depletion. *Nat Genet*. 2007; 39:776–780. [PubMed: 17486094]
46. Cosconati S, Marinelli L, Lavecchia A, Novellino E. Characterizing the 1,4-dihydropyridines binding interactions in the L-type Ca<sup>2+</sup> channel: Model construction and docking calculations. *J Med Chem*. 2007; 50:1504–1513. [PubMed: 17335186]
47. Rumpel S, Becker S, Zweckstetter M. High-resolution structure determination of the CylR2 homodimer using paramagnetic relaxation enhancement and structure-based prediction of molecular alignment. *J Biomol NMR*. 2008; 40:1–13. [PubMed: 18026911]
48. Kucuk C, et al. Activating mutations of STAT5B and STAT3 in lymphomas derived from gammadelta-T or NK cells. *Nat Comm*. 2015; 6:6025.
49. Ye Z, Musiol EM, Weber T, Williams GJ. Reprogramming acyl carrier protein interactions of an Acyl-CoA promiscuous trans-acyltransferase. *Chem Biol*. 2014; 21:636–646. [PubMed: 24726832]
50. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res*. 2005; 33:W363–367. [PubMed: 15980490]
51. Stealand S, et al. Generation and characterization of small single domain antibodies inhibiting human tumor necrosis factor receptor 1. *J Biol Chem*. 2015; 290:4022–4037. [PubMed: 25538244]
52. Bohnuud T, et al. A benchmark testing ground for integrating homology modeling and protein docking. *Proteins*. 2016
53. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins*. 2003; 52:80–87. [PubMed: 12784371]
54. Brenke R, et al. Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics*. 2012; 28:2608–2614. [PubMed: 23053206]
55. Comeau SR, Camacho CJ. Predicting oligomeric assemblies: N-mers a primer. *J Struct Biol*. 2005; 150:233–244. [PubMed: 15890272]
56. Pierce B, Tong W, Weng Z. M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. *Bioinformatics*. 2005; 21:1472–1478. [PubMed: 15613396]
57. Yang S. Methods for SAXS-based structure determination of biomolecular complexes. *Adv Materials*. 2014; 26:7902–7910.
58. Xia B, et al. Accounting for observed small angle X-ray scattering profile in the protein-protein docking server cluspro. *J Comput Chem*. 2015; 36:1568–1572. [PubMed: 26095982]
59. Pons C, et al. Structural characterization of protein-protein complexes by integrating computational docking with small-angle scattering data. *J Mol Biol*. 2010; 403:217–230. [PubMed: 20804770]
60. Schneidman-Duhovny D, Hammel M, Sali A. Macromolecular docking restrained by a small angle X-ray scattering profile. *J Struct Biol*. 2011; 173:461–471. [PubMed: 20920583]
61. Bernfield M, et al. Functions of cell surface heparan sulfate proteoglycans. *Annu Rev Biochem*. 1999; 68:729–777. [PubMed: 10872465]
62. Lindahl U. Heparan sulfate-protein interactions - A concept for drug design? *Thromb Haemostasis*. 2007; 98:109–115. [PubMed: 17598000]
63. Fugedi P. The potential of the molecular diversity of heparin and heparan sulfate for drug development. *Mini Rev Med Chem*. 2003; 3:659–667. [PubMed: 14529507]
64. Esko JD, Lindahl U. Molecular diversity of heparan sulfate. *J Clin Invest*. 2001; 108:169–173. [PubMed: 11457867]
65. Forster M, Mulloy B. Computational approaches to the identification of heparin-binding sites on the surfaces of proteins. *Biochem Soc Trans*. 2006; 34:431–434. [PubMed: 16709179]
66. Mottarella SE, et al. Docking server for the identification of heparin binding sites on proteins. *J Chem Inf Model*. 2014; 54:2068–2078. [PubMed: 24974889]
67. Gray JJ, et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*. 2003; 331:281–299. [PubMed: 12875852]

68. Zacharias M. ATTRACT: protein-protein docking in CAPRI using a reduced protein model. *Proteins*. 2005; 60:252–256. [PubMed: 15981270]
69. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*. 2003; 125:1731–1737. [PubMed: 12580598]
70. de Vries SJ, van Dijk M, Bonvin AM. The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc*. 2010; 5:883–897. [PubMed: 20431534]
71. Janin J, et al. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins*. 2003; 52:2–9. [PubMed: 12784359]
72. Mendez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins*. 2003; 52:51–67. [PubMed: 12784368]
73. de Vries SJ, Schindler CE, Chauvot de Beauchene I, Zacharias M. A web interface for easy flexible protein-protein docking with ATTRACT. *Biophys J*. 2015; 108:462–465. [PubMed: 25650913]
74. Moal IH, Bates PA. SwarmDock and the use of normal modes in protein-protein docking. *Int J Mol Sci*. 2010; 11:3623–3648. [PubMed: 21152290]
75. Ramirez-Aportela E, Lopez-Blanco JR, Chacon P. FRODOCK 2.0: fast protein-protein docking server. *Bioinformatics*. 2016
76. Yu J, et al. InterEvDock: a docking server to predict the structure of protein-protein interactions using evolutionary information. *Nucleic Acids Res*. 2016
77. Pierce BG, Hourai Y, Weng Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS One*. 2011; 6:e24657. [PubMed: 21949741]
78. Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res*. 2006; 34:W310–W314. [PubMed: 16845016]
79. Kozakov D, Schueler-Furman O, Vajda S. Discrimination of near-native structures in protein-protein docking by testing the stability of local minima. *Proteins*. 2008; 72:993–1004. [PubMed: 18300245]
80. Moghadasi M, et al. The impact of side-chain packing on protein docking refinement. *J Chem Inf Model*. 2015; 55:872–881. [PubMed: 25714358]
81. Padhorn D, et al. Protein-protein docking by fast generalized Fourier transforms on 5D rotational manifolds. *Proc Natl Acad Sci U S A*. 2016; 113:E4286–E4293. [PubMed: 27412858]
82. Gruschus JM, Greene LE, Eisenberg E, Ferretti JA. Experimentally biased model structure of the Hsc70/auxilin complex: substrate transfer and interdomain structural change. *Protein Sci*. 2004; 13:2029–2044. [PubMed: 15273304]
83. Liu J, Wang H, Zuo Y, Farmer SR. Functional interaction between peroxisome proliferator-activated receptor gamma and beta-catenin. *Mol Cell Biol*. 2006; 26:5827–5837. [PubMed: 16847334]
84. Lee DY, et al. Mutagenesis and modeling of the peroxiredoxin (Prx) complex with the NMR structure of ATP-bound human sulfiredoxin implicate aspartate 187 of Prx I as the catalytic residue in ATP hydrolysis. *Biochemistry*. 2006; 45:15301–15309. [PubMed: 17176052]
85. Watson AA, et al. The crystal structure and mutational binding analysis of the extracellular domain of the platelet-activating receptor CLEC-2. *J Biol Chem*. 2007; 282:3165–3172. [PubMed: 17132623]
86. Martin MC, Allan LA, Mancini EJ, Clarke PR. The docking interaction of caspase-9 with ERK2 provides a mechanism for the selective inhibitory phosphorylation of caspase-9 at threonine 125. *J Biol Chem*. 2008; 283:3854–3865. [PubMed: 18083711]
87. Liu C, et al. The SH3-like domain switches its interaction partners to modulate the repression activity of mycobacterial iron-dependent transcription regulator in response to metal ion fluctuations. *J Biol Chem*. 2008; 283:2439–2453. [PubMed: 18055464]
88. Pilpa RM, et al. Functionally distinct NEAT (NEAr Transporter) domains within the *Staphylococcus aureus* IsdH/HarA protein extract heme from methemoglobin. *J Biol Chem*. 2009; 284:1166–1176. [PubMed: 18984582]
89. Liang S, et al. Mapping of a microbial protein domain involved in binding and activation of the TLR2/TLR1 heterodimer. *J Immunol*. 2009; 182:2978–2985. [PubMed: 19234193]

90. Cohavi O, Tobi D, Schreiber G. Docking of antizyme to ornithine decarboxylase and antizyme inhibitor using experimental mutant and double-mutant cycle data. *J Mol Biol.* 2009; 390:503–515. [PubMed: 19465028]
91. Hofmann WA, et al. SUMOylation of nuclear actin. *J Cell Biol.* 2009; 186:193–200. [PubMed: 19635839]
92. Arthur CJ, et al. Structure and malonyl CoA-ACP transacylase binding of streptomyces coelicolor fatty acid synthase acyl carrier protein. *ACS Chem Biol.* 2009; 4:625–636. [PubMed: 19555075]
93. Nuth M, Cowan JA. Iron-sulfur cluster biosynthesis: characterization of IscU-IscS complex formation and a structural model for sulfide delivery to the [2Fe-2S] assembly site. *J Biol Inorg Chem.* 2009; 14:829–839. [PubMed: 19308466]
94. Stauch B, et al. Model structure of APOBEC3C reveals a binding pocket modulating ribonucleic acid interaction required for encapsidation. *Proc Natl Acad Sci U S A.* 2009; 106:12079–12084. [PubMed: 19581596]
95. Venkatachari NJ, et al. Human immunodeficiency virus type 1 Vpr: oligomerization is an essential feature for its incorporation into virus particles. *Virology.* 2010; 7:119. [PubMed: 20529298]
96. Fredericks WJ, et al. The bladder tumor suppressor protein TERE1 (UBIAD1) modulates cell cholesterol: Implications for tumor progression. *DNA Cell Biol.* 2011; 30:851–864. [PubMed: 21740188]
97. Lira-Navarrete E, et al. Structural insights into the mechanism of protein o-fucosylation. *PLoS One.* 2011; 6
98. Li D, et al. A comprehensive model of the spectrin divalent tetramer binding region deduced using homology modeling and chemical cross-linking of a mini-spectrin. *J Biol Chem.* 2010; 285:29535–29545. [PubMed: 20610390]
99. Zhang GF, et al. Ligand-independent antiapoptotic function of estrogen receptor-beta in lung cancer cells. *Mol Endocrinol.* 2010; 24:1737–1747. [PubMed: 20660297]
100. Naudin C, et al. The occluding loop of cathepsin B prevents its effective inhibition by human kininogens. *J Mol Biol.* 2010; 400:1022–1035. [PubMed: 20538006]
101. Guzman L, et al. Blockade of ethanol-induced potentiation of glycine receptors by a peptide that interferes with Gbetagamma binding. *J Pharmacol Exp Ther.* 2009; 331:933–939. [PubMed: 19773530]
102. Xi J, et al. Interaction between the T4 helicase-loading protein (gp59) and the DNA polymerase (gp43): A locking mechanism to delay replication during replisome assembly. *Biochemistry.* 2005; 44:4600–4600.
103. Nelson SW, Yang JS, Benkovic SJ. Site-directed mutations of T4 helicase loading protein (gp59) reveal multiple modes of DNA polymerase inhibition and the mechanism of unlocking by gp41 helicase. *J Biol Chem.* 2006; 281:8697–8706. [PubMed: 16407253]
104. Sobrado P, et al. Identification of the binding region of the [2Fe-2S] ferredoxin in stearoyl-acyl carrier protein desaturase: insight into the catalytic complex and mechanism of action. *Biochemistry.* 2006; 45:4848–4858. [PubMed: 16605252]
105. Kedlaya RH, Bhat KM, Mitchell J, Darnell SJ, Setaluri V. TRP1 interacting PDZ-domain protein GIPC forms oligomers and is localized to intracellular vesicles in human melanocytes. *Arch Biochem Biophys.* 2006; 454:160–169. [PubMed: 16962991]
106. Li D, Tang HY, Speicher DW. A structural model of the erythrocyte spectrin heterodimer initiation site determined using homology modeling and chemical cross-linking. *J Biol Chem.* 2008; 283:1553–1562. [PubMed: 17977835]
107. Krauss U, Losi A, Gartner W, Jaeger KE, Eggert T. Initial characterization of a blue-light sensing, phototropin-related protein from *Pseudomonas putida*: a paradigm for an extended LOV construct. *Phys Chem Chem Phys.* 2005; 7:2804–2811. [PubMed: 16189596]
108. Surolia I, Reddy GB, Sinha S. Hierarchy and the mechanism of fibril formation in ADan peptides. *J Neurochem.* 2006; 99:537–548. [PubMed: 17029605]
109. Alminaita A, et al. Oligomerization of hantavirus nucleocapsid protein: analysis of the N-terminal coiled-coil domain. *J Virol.* 2006; 80:9073–9081. [PubMed: 16940519]

110. Alminaité A, Backstrom V, Vaheiri A, Plyusnin A. Oligomerization of hantaviral nucleocapsid protein: charged residues in the N-terminal coiled-coil domain contribute to intermolecular interactions. *J Gen Virol.* 2008; 89:2167–2174. [PubMed: 18753226]
111. Juszczak P, et al. Binding epitopes and interaction structure of the neuroprotective protease inhibitor cystatin c with beta-amyloid revealed by proteolytic excision mass spectrometry and molecular docking simulation. *J Med Chem.* 2009; 52:2420–2428. [PubMed: 19317448]
112. Brown KA, Dayal S, Ai X, Rumbles G, King PW. Controlled assembly of hydrogenase-CdTe nanocrystal hybrids for solar hydrogen production. *J Am Chem Soc.* 2010; 132:9672–9680. [PubMed: 20583755]
113. Raab M, Parthasarathi L, Treumann A, Moran N, Daxecker H. Differential binding of ICln in platelets to integrin-derived activating and inhibitory peptides. *Biochem Biophys Res Commun.* 2010; 392:258–263. [PubMed: 20034469]
114. Nan R, et al. Zinc binding to the Tyr402 and His402 allotypes of complement factor H: possible implications for age-related macular degeneration. *J Mol Biol.* 2011; 408:714–735. [PubMed: 21396937]
115. Sato K, Crowley PB, Dennison C. Transient homodimer interactions studied using the electron self-exchange reaction. *J Biol Chem.* 2005; 280:19281–19288. [PubMed: 15743773]
116. Man P, et al. Accessibility changes within diphtheria toxin T domain when in the functional molten globule state, as determined using hydrogen/deuterium exchange measurements. *FEBS J.* 2010; 277:653–662. [PubMed: 20050921]
117. Wang GS, et al. Solution structure of the phosphoryl transfer complex between the signal transducing proteins HPr and IIA(Glucose) of the *Escherichia coli* phosphoenolpyruvate : sugar phosphotransferase system. *EMBO J.* 2000; 19:5635–5649. [PubMed: 11060015]
118. Leysen S, Vanderkelen L, Weeks SD, Michiels CW, Strelkov SV. Structural basis of bacterial defense against g-type lysozyme-based innate immunity. *Cell Mol Life Sci.* 2013; 70:1113–1122. [PubMed: 23086131]
119. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol.* 2003; 374:461–491. [PubMed: 14696385]
120. Viswanath S, Ravikant DV, Elber R. DOCK/PIERR: web server for structure prediction of protein-protein complexes. *Methods Mol Biol.* 2014; 1137:199–207. [PubMed: 24573483]
121. Fernandez-Recio J, Totrov M, Abagyan R. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins.* 2003; 52:113–117. [PubMed: 12784376]
122. Esquivel-Rodriguez J, Filos-Gonzalez V, Li B, Kihara D. Pairwise and multimeric protein-protein docking using the LZerD program suite. *Methods Mol Biol.* 2014; 1137:209–234. [PubMed: 24573484]
123. Terashi G, et al. The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation. *Proteins.* 2007; 69:866–872. [PubMed: 17853449]
124. May A, Zacharias M. Protein-protein docking in CAPRI using ATTRACT to account for global and local flexibility. *Proteins.* 2007; 69:774–780. [PubMed: 17803217]
125. Huang SY, Zou X. An iterative knowledge-based scoring function for protein-protein recognition. *Proteins.* 2008; 72:557–579. [PubMed: 18247354]
126. Eisenstein M, Ben-Shimon A, Frankenstein Z, Kowalsman N. CAPRI targets T29–T42: proving ground for new docking procedures. *Proteins.* 2010; 78:3174–3181. [PubMed: 20607697]
127. Andrusier N, Mashiah E, Nussinov R, Wolfson HJ. Principles of flexible protein-protein docking. *Proteins.* 2008; 73:271–289. [PubMed: 18655061]
128. Shen Y. Improved flexible refinement of protein docking in CAPRI rounds 22–27. *Proteins.* 2013; 81:2129–2136. [PubMed: 23996302]
129. Heo L, Lee H, Seok C. GalaxyRefineComplex: Refinement of protein-protein complex model structures driven by interface repacking. *Sci Rep.* 2016; 6:32153. [PubMed: 27535582]
130. Cheng TM, Blundell TL, Fernandez-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins.* 2007; 68:503–515. [PubMed: 17444519]
131. Zhou HX, Qin S. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics.* 2007; 23:2203–2209. [PubMed: 17586545]

**EDITORIAL SUMMARY**

ClusPro is a web server that performs rigid body docking of two proteins by sampling billions of conformations. Low energy docked structures are clustered, and centers of the largest clusters are used as likely models of the complex.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



**BOX 1****STATUS UPDATES FOR CLUSPRO RUNS**

The progress of the ClusPro run can be monitored in the “Queue” tab. Note that since the timings for each step are highly dependent on the input structure, they are not provided here.

**Processing pdb files**

Downloading PDB file from the [www.pdb.org](http://www.pdb.org) web site, processing chain information, extracting the chains the user specified

**Pre-docking minimization**

Running CHARMM to add missing atoms and polar hydrogens, minimizing the added atoms in the presence of the protein

**Copying to supercomputer**

Copying the PDB files to the cluster where ClusPro will run

**Held on supercomputer**

Files are on cluster, but job is not yet submitted

**In queue on supercomputer**

Job has been submitted on the cluster, but has not started running

**Running on supercomputer**

Job has begun running on the cluster

**Clustering and minimization**

Clustering top structures, selecting representative from top ranked clusters for minimization. Cluster are ranked by cluster size

**Finalizing job**

Compresses files on server for transfer

**Copying to local computer**

Results are being copied from the cluster back to the ClusPro server

**Finished**

Everything is complete

**Computing Saxes profiles**

Preliminary to docking (Saxes Profile mode only)

**Error on local system**

Processing PDB files fails. Check error messages

**Error on supercomputer**

Job fails to run on the cluster. Check error messages

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**BOX 2****ERROR MESSAGES AND THEIR MEANINGS**

When the calculation encounters an error, the job will be terminated and the user will receive an email with an error reason. The error will also be listed next to the corresponding job id in the “Results” tab. The error codes and explanations are listed below.

**xxxx not found in PDB**

xxxx is the four letter PDB ID. This error occurs when the computer is unable to download the entered PDB ID from the website, [www.pdb.org](http://www.pdb.org). Usually this error occurs when the PDB ID does not exist, but it can also occur when the PDB website is down.

**Unknown residue xxx in receptor. Please remove**

xxx is the three letter amino acid code. This error occurs when a residue in an atom record is not recognizable by ClusPro and thus ClusPro does not have parameters for it. Check to make sure that the proper three letter code is used and the amino acid is one of the 20 naturally occurring amino acids.

**Unknown residue xxx in ligand. Please remove**

xxx is the three letter amino acid code. This error occurs when a residue in an atom record is not recognizable by ClusPro and thus ClusPro does not have parameters for it. Check to make sure that the proper three letter code is used and the amino acid is one of the 20 naturally occurring amino acids.

**Receptor chains must be fewer than 20 characters**

Chain specification is incorrect.

**Receptor chains must be white space separated alphanumeric characters**

Incorrect or missing receptor chain specification.

**Receptor PDB id must be 4 alphanumeric characters**

Invalid PDB id.

**Receptor file too large**

PDB file exceeds maximum allowed size. May happen if the protein has multiple domains. Some domains can be docked separately.

**Receptor file only partially uploaded**

PDB file is either too large or a network error has occurred during upload.

**Ligand chains must be fewer than 20 characters**

Chain specification is incorrect.

**Ligand chains must be white space separated alphanumeric characters**

Incorrect or missing ligand chain specification.

**Ligand PDB id must be 4 alphanumeric characters**

Invalid PDB id.

**Ligand file too large**

PDB file exceeds maximum allowed size. May happen if the protein has multiple domains. Some domains can be docked separately.

**Ligand file only partially uploaded**

PDB file is either too large or a network error has occurred during upload.

**Copy of receptor failed**

File did not transfer to computing cluster. Check input for consistency.

**Copy of ligand failed**

File did not transfer to computing cluster. Check input for consistency.

**Processing failed on receptor**

This error occurs during the initial steps when Charmm is used to add missing hydrogen atoms and minimize the structure. Usually, this occurs when the protein structure has sterically clashing atoms or the structure generally does not make physical sense in terms of bonds.

**Processing failed on ligand**

This error occurs during the initial steps when Charmm is used to add missing hydrogen atoms and minimize the structure. Usually, this occurs when the protein structure has sterically clashing atoms or the structure generally does not make physical sense in terms of bonds.

**Not enough lines in output for xxx**

Restraints are wrong or cannot be satisfied.

**Job ran out of memory on server**

Receptor and/or ligand are too large.

**Ligand repulsion file too large**

If the protein is large, the repulsion file is even larger (Attraction and Repulsion mode only).

**Ligand attraction/repulsion file only partially uploaded. Please try again**

A network error may have occurred during upload. (Attraction and Repulsion mode only).

**Ligand attraction/repulsion must be whitespace separated chain-residue**

Incorrect format (Attraction and Repulsion mode only).

**Receptor repulsion file too large**

(Attraction and Repulsion mode only).

**Receptor attraction/repulsion file only partially uploaded. Please try again**

A network error may have occurred during upload. (Attraction and Repulsion mode only).

**Receptor attraction/repulsion must be whitespace separated chain-residue**

Incorrect format (Attraction and Repulsion mode only).

**Number of multimers must be expressed as a number**

Format error (Multimer mode only).

**Number of multimers must be an integer**

Format error (Multimer mode only).

**Number of multimers must be 2 or greater**

Format error (Multimer mode only).

**Copy of heparin failed**

(Heparin Ligand mode only). Restart job.

**'SAXS' profile only partially uploaded. Please try again**

File is either too large or a network error has occurred during upload (Saxs Profile mode only).

**Restraints file too large**

If the protein is large, the restraints file is even larger (Restraints mode only).

**Restraints file only partially uploaded. Please try again**

File is either too large or a network error has occurred during upload (Restraints mode only).

**Error converting AIR restraints to JSON. Please consult the help page**

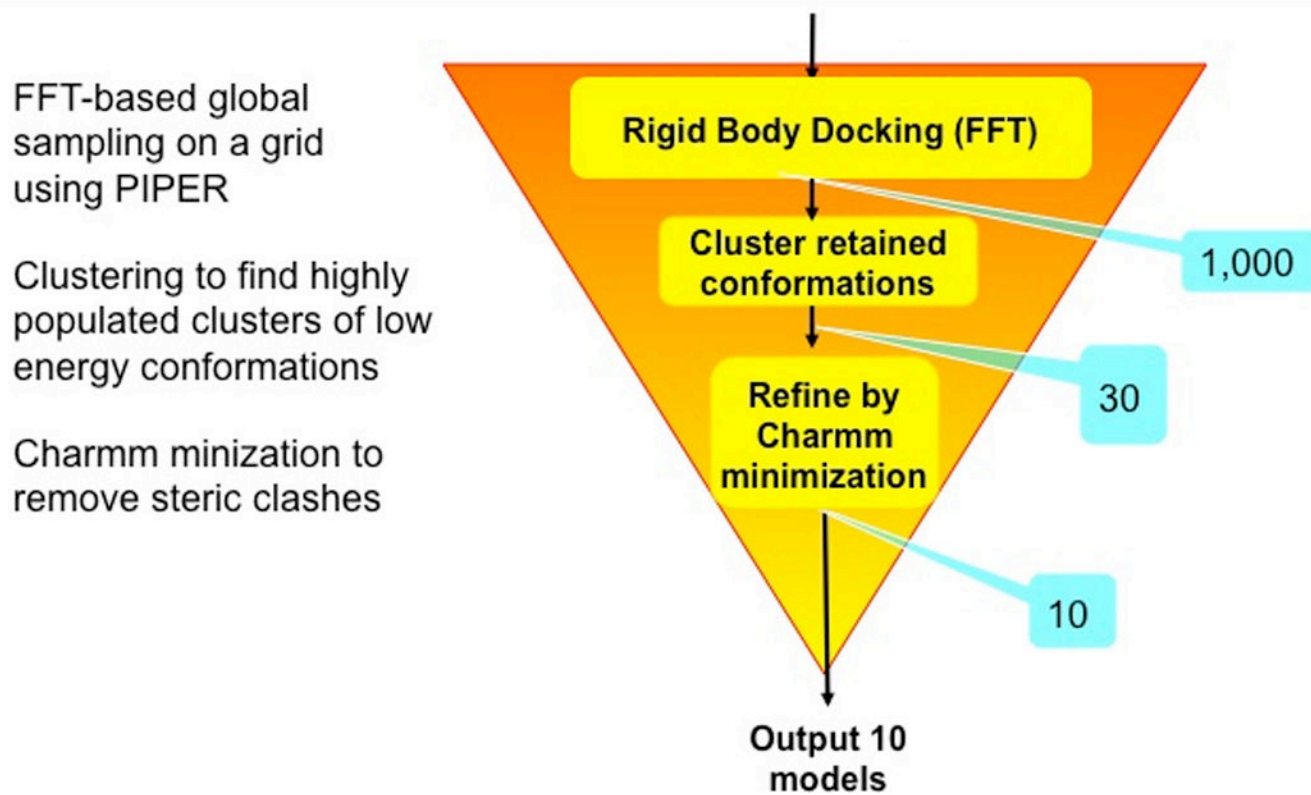
Submitted restraints may not be formatted properly. A sample AIR type restraint file can be downloaded from the help page (Restraints mode only).

**Restraints file has incorrect format. Please consult the help page**

Restraints files must be in either AIR format or in JSON format. See the help page to use tool for generating restraints in JSON format and for sample AIR and JSON formatting (Restraints mode only).

**Restraints file refers to unknown residues**

Check residue names (Restraints mode only).



**FIGURE 1.** Outline of the ClusPro algorithm. After each step, the number of structures retained is shown in a blue box.

**ClusPro**  
protein-protein docking

[Dock](#) [Queue](#) [Results](#) [Preferences](#) [Downloads](#) [Papers](#) [Help](#) [Contact](#)

[sign out](#)

## Dock

**Job Name:**

**Server:**

Accepted PDB Input:  
20 standard amino acids and RNA (as receptor only), ref: [RNA](#) Select Heparin Mode to use Heparin as Ligand.

Receptor	Ligand
<b>PDB ID:</b> <input type="text" value="1QQU"/>	<b>PDB ID:</b> <input type="text" value="1BA7"/>
<a href="#">Upload PDB</a>	<a href="#">Upload PDB</a>
<b>Chains:</b> <input type="text" value="A"/>	<b>Chains:</b> <input type="text" value="B"/>

Whitespace separate desired chains. Leave chains blank to use all chains.

► **Advanced Options**

**FIGURE 2.** Home screen of the ClusPro server. The example shows submission of the job to dock soybean trypsin inhibitor (ligand, PDB ID 1BA7) to the X-ray structure of porcine trypsin (receptor, PDB ID 1QQU).

## Restraint Set Generator

**Restraint Set**

Required percent of groups  %

**Restraint Group**

Required percentage of restraints  % Remove Group

Residue 1	Residue 2	Minimum Distance	Maximum Distance	
Receptor Residue <input type="text" value="A1"/>	Ligand Residue <input type="text" value="B1"/>	Min Distance <input type="text" value="1"/>	Max Distance <input type="text" value="5"/>	<span>Add Restraint</span>

Add Group Create Restraints

**Restraints**

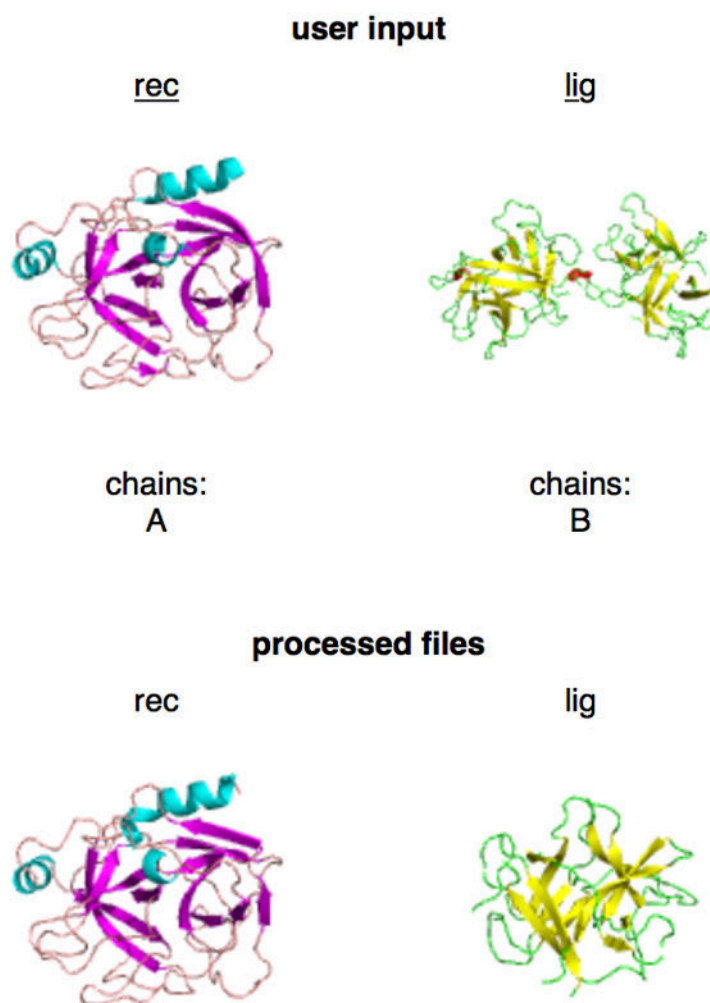
Save As...

**FIGURE 3.** Screen image of the restraint file generator web tool to prepare a JSON file. The generator is available at the URL [https://cluspro.org/generate\\_restraints.html](https://cluspro.org/generate_restraints.html).



**168361: 1AVX**

**Status** copying to supercomputer  
**Submitted** 2016-05-23 03:01:44  
**Errors** (none reported)  
**Rotations** 70000

**FIGURE 4.**

Screen image of the ClusPro Status page. The screen was obtained by clicking on the job Id on the queue page, for docking of soybean trypsin inhibitor (ligand) to the porcine trypsin (receptor). The Status page shows the job ID number, job name, job status, and submission time stamp. The page also shows pictorial representations of the uploaded and processed input structures. The PDB file for the ligand includes two chains, but only chain B is used for docking. Clicking on the labels `rec` and `lig` will download the submitted coordinate files as read by the server.

[sign out](#)

**Job Details: 1AVX**

**View Model Scores**

Download all Models for all Coefficients

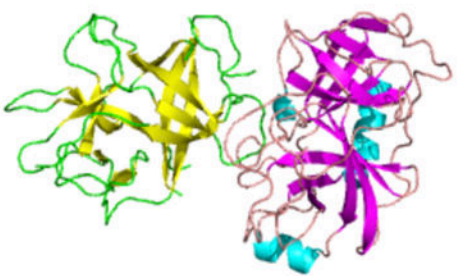
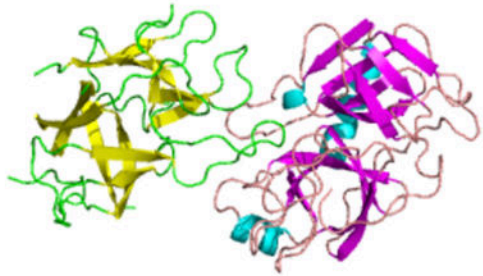

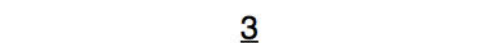
Balanced | Electrostatic-favored | Hydrophobic-favored | VdW+Elec

Display Models:

Download Displayed Models

**If you use these models in a paper, please cite our papers**

00

<p>0</p> 	<p>1</p> 
<p>2</p> 	<p>3</p> 

**FIGURE 5.**

Screen image of the ClusPro Results page for docking of soybean trypsin inhibitor (ligand) to the porcine trypsin (receptor). The page shows the job name and, by default, results for the balanced set of scoring function coefficients. Results for electrostatic-favored, hydrophobic-favored, or van der Waals + electrostatics sets can be viewed by clicking the corresponding labels. For each parameter set the result page shows small pictures representing the top 10 models, but the user can display more models (up to 30 or less if fewer than 30 clusters have been created). Clicking on the number above a picture will download that model as a PDB file for your viewing. The page also provides the label to download models for all coefficients.

**Job Details: 1AVX****View Models**Balanced | [Electrostatic-favored](#) | [Hydrophobic-favored](#) | [VdW+Elec](#)[Download Model Scores for this Coefficient](#)**Coefficient Weights**See *Kozakov et. al.* in [Papers](#) for a description of these terms

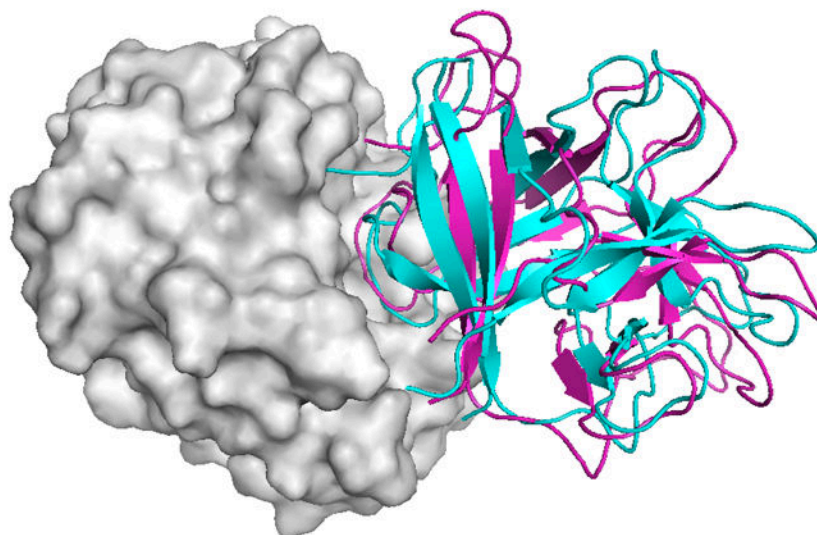
$$E = 0.40E_{rep} + -0.40E_{att} + 600E_{elec} + 1.00E_{DARS}$$

**Cluster Scores**We strongly encourage you to read the [FAQ related to these scores](#) before using them.

Cluster	Members	Representative	Weighted Score
0	253	Center	-793.1
		Lowest Energy	-902.6
1	122	Center	-706.1
		Lowest Energy	-802.8
2	97	Center	-698.7
		Lowest Energy	-919.9
3	56	Center	-744.2
		Lowest Energy	-842.0
4	46	Center	-687.5
		Lowest Energy	-795.3
5	38	Center	-694.3
		Lowest Energy	-804.9
6	33	Center	-748.1
		Lowest Energy	-818.0

**FIGURE 6.**

Screen image of the ClusPro Results page showing model scores for the balanced coefficient set when docking soybean trypsin inhibitor to porcine trypsin. The page shows the actual weighting coefficients of the energy terms, and a table that lists the clusters of docked structures in the order of cluster size. For each cluster, the table shows the size (i.e., the number of docked structures), the energy of the cluster center (i.e., the structure that has the highest number of neighbor structures in the cluster), and the energy of the lowest energy structure in the cluster.

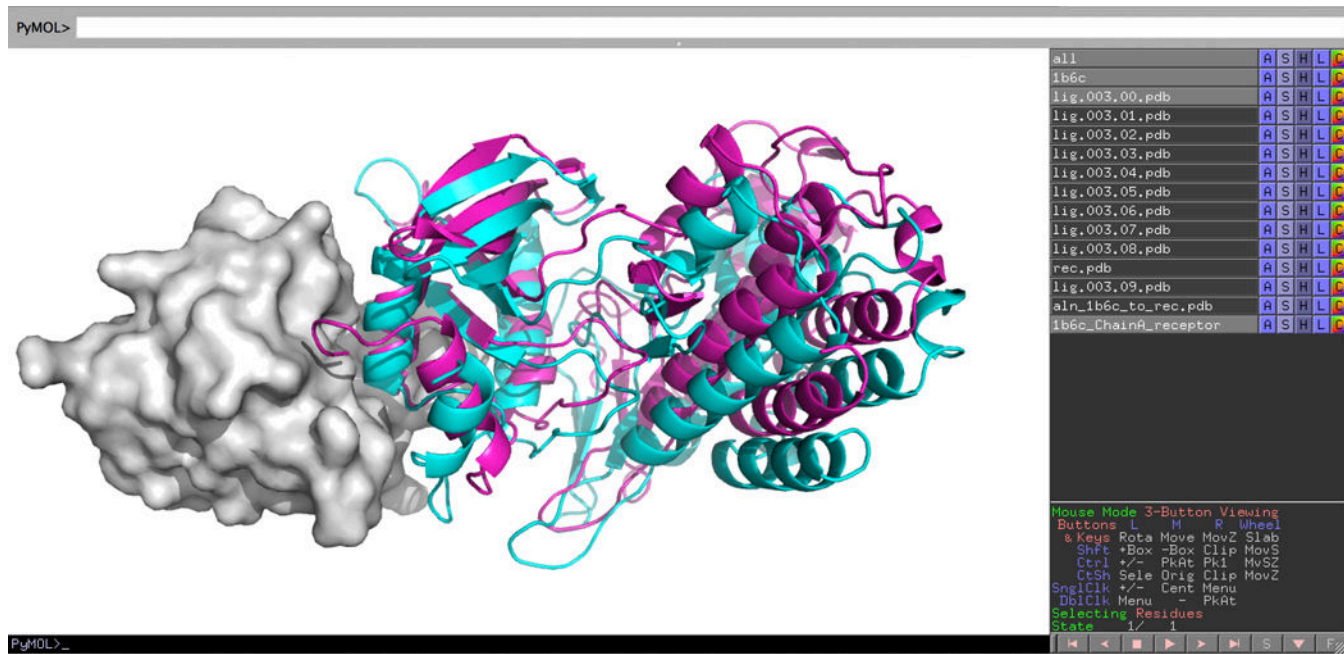


```
all A S H L C
lig.000.00.pdb A S H L C
1avx A S H L C
aln_1avx_to_rec.pdb A S H L C
1avx_lig A S H L C
lig.000.01.pdb A S H L C
lig.000.02.pdb A S H L C
lig.000.03.pdb A S H L C
lig.000.04.pdb A S H L C
lig.000.05.pdb A S H L C
lig.000.06.pdb A S H L C
lig.000.07.pdb A S H L C
lig.000.08.pdb A S H L C
rec.pdb A S H L C
lig.000.09.pdb A S H L C

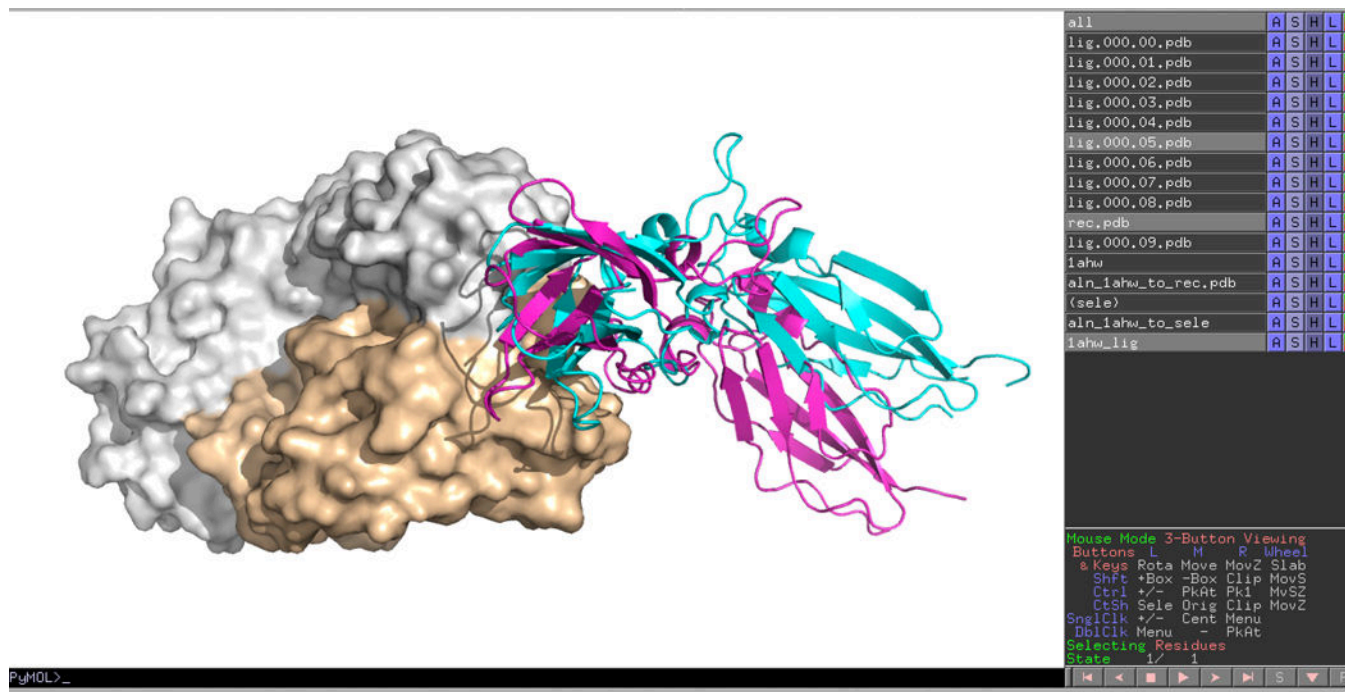
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
sKeys Rota Move Mov2 Slab
Shift +Box -Box Clip MovS
Ctrl +/- PKAt PK1 MvSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DbClk Menu = PKAt
Selecting Residues
State 1/ 1
```

**FIGURE 7.**

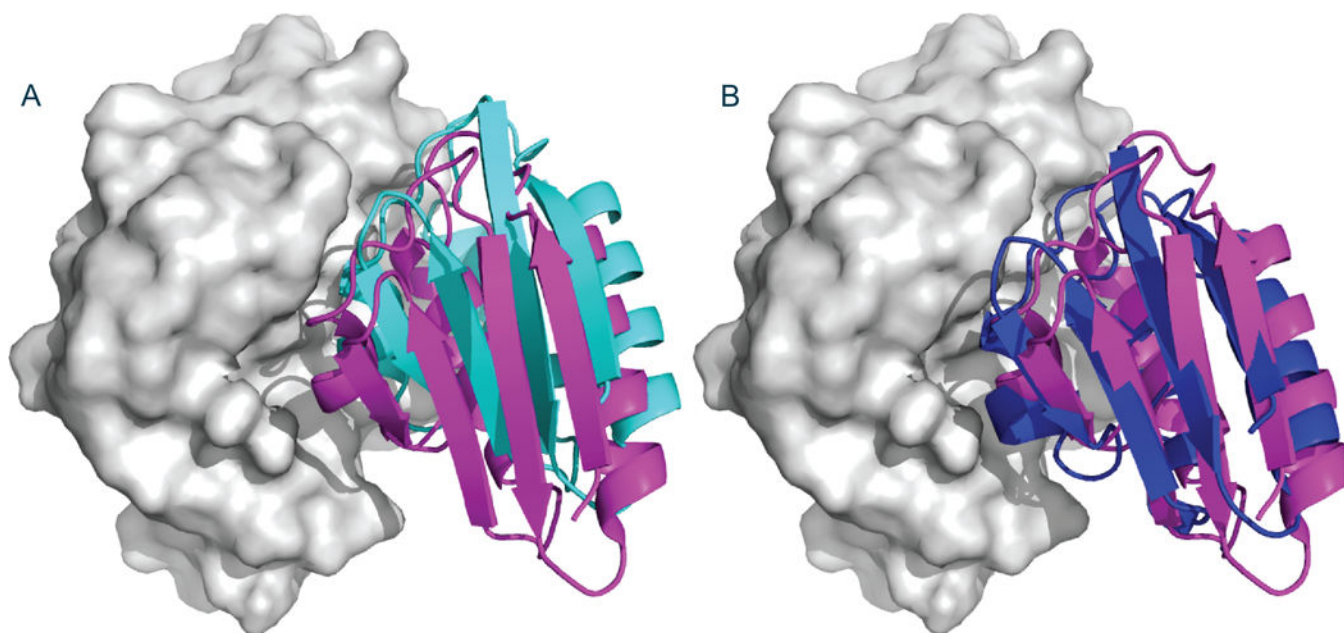
Visualization by PyMol of the structure at the center of the most populated cluster in docking soybean trypsin inhibitor (ligand) to porcine trypsin (receptor). The docked ligand structure (lig.000.00.pdb) is shown as cyan cartoon, whereas the receptor is shown as grey surface. For comparison we also loaded the X-ray structure of the enzyme-inhibitor complex (PDB ID 1AVX) into PyMol, aligned it with the structure of the receptor, and show the native pose of the inhibitor as magenta cartoon.

**FIGURE 8.**

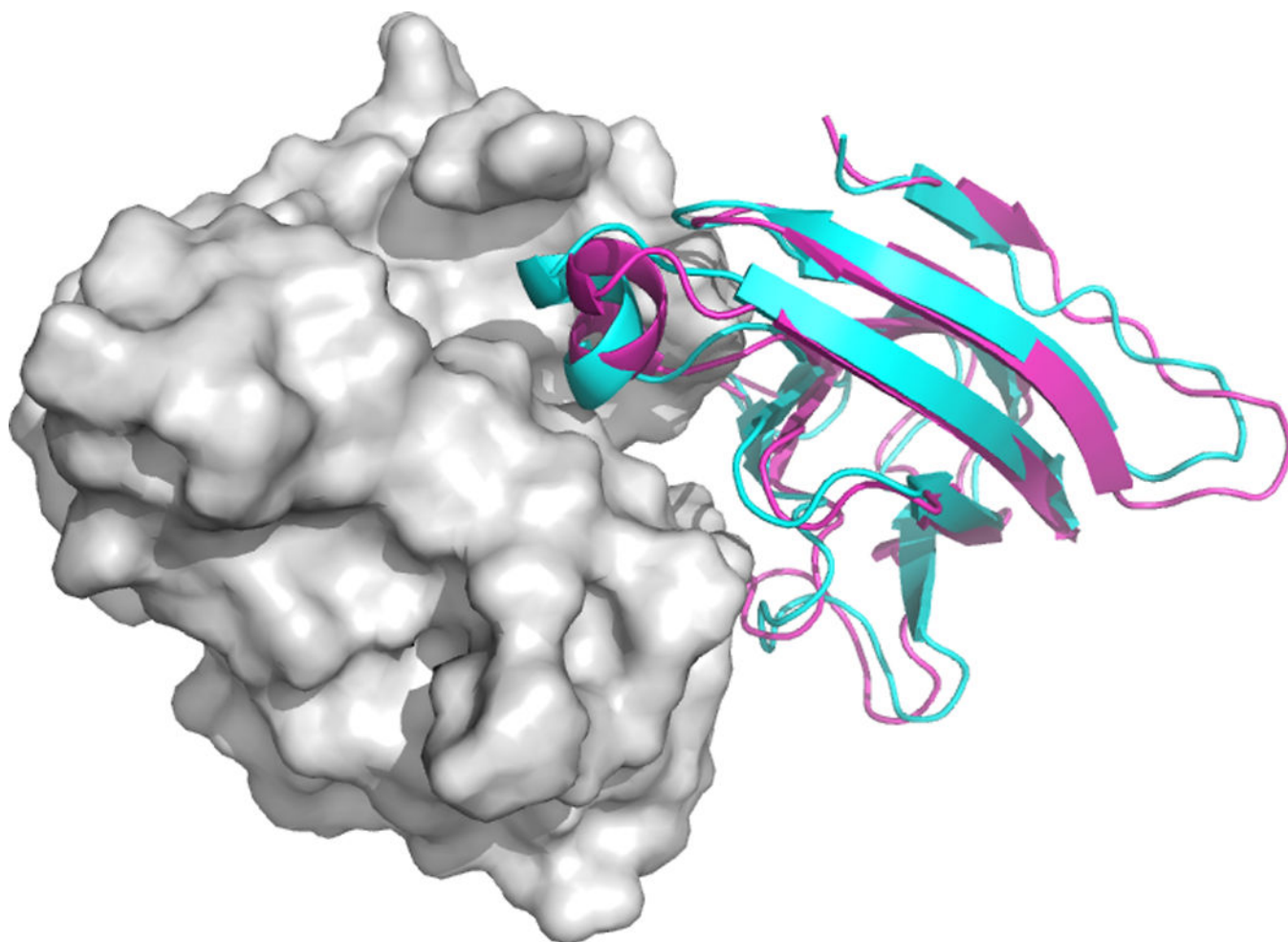
Screen image of exploring the results of running ClusPro in Others Mode. Results are shown using PyMOL from docking the X-ray structure of the ligand, to the X-ray structure of the FK506 binding protein (FKBP). The receptor, FKBP, is shown as grey surface, and the ligand (lig.003.00.pdb) at center of the largest cluster is shown as cyan cartoon. For comparison we superimposed the native complex (PDB ID 1B6C) on the receptor in the docked structure, and the corresponding ligand pose is shown as magenta cartoon.

**FIGURE 9.**

Screen image of the PyMOL visualization of the results of running ClusPro in Antibody Mode. We docked the X-ray structure of the extracellular domain of the human tissue factor (PDB ID 1TFH) to the unbound X-ray structure of the FAB domain of the inhibitory antibody 5G9 (PDB ID 1FGN). Both the heavy and light chains were used to represent the receptor. The center of the 6<sup>th</sup> most populated cluster, lig.000.05.pdb, is shown as cyan cartoon, whereas the antibody is shown in surface representation. The antigen in the native complex (PDB ID 1AHW) is shown as magenta cartoon. The IRMSD between native and predicted ligand poses is 4.7 Å.

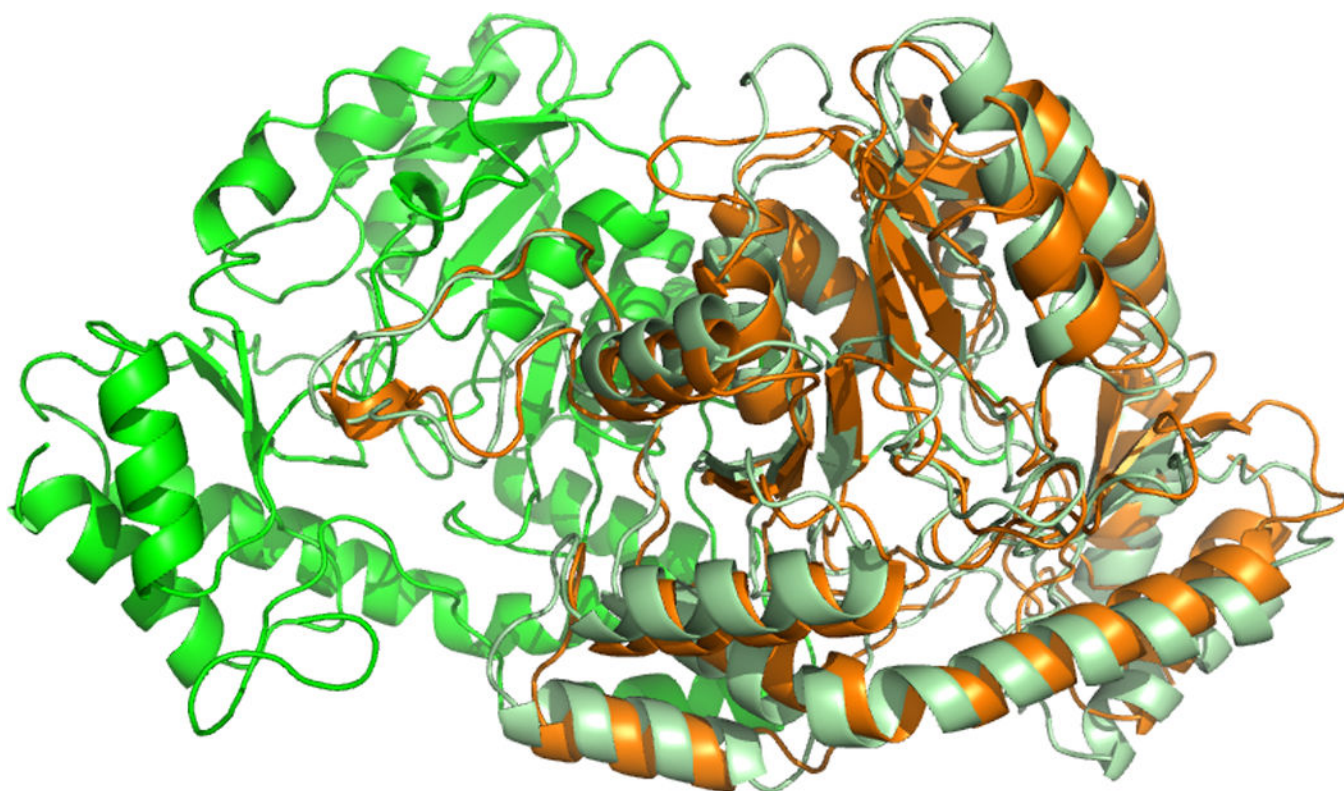
**FIGURE 10.**

Best models from docking the signal transducing protein HPr (PDB ID 1POH) to the glucose-specific phosphocarrier protein E2A (PDB ID 1F3G) without and with restraints. The receptor protein, E2A, is shown as grey surface. **A.** Ligand position at the center of the second largest cluster (shown as cyan cartoon) from docking without restraints. The position is slightly shifted relative to the native ligand binding position, shown as magenta cartoon. **B.** Ligand position at the center of the largest cluster (shown as blue cartoon) from docking with restraints. The ligand is now turned a few degrees around an axis perpendicular to the center of the receptor.

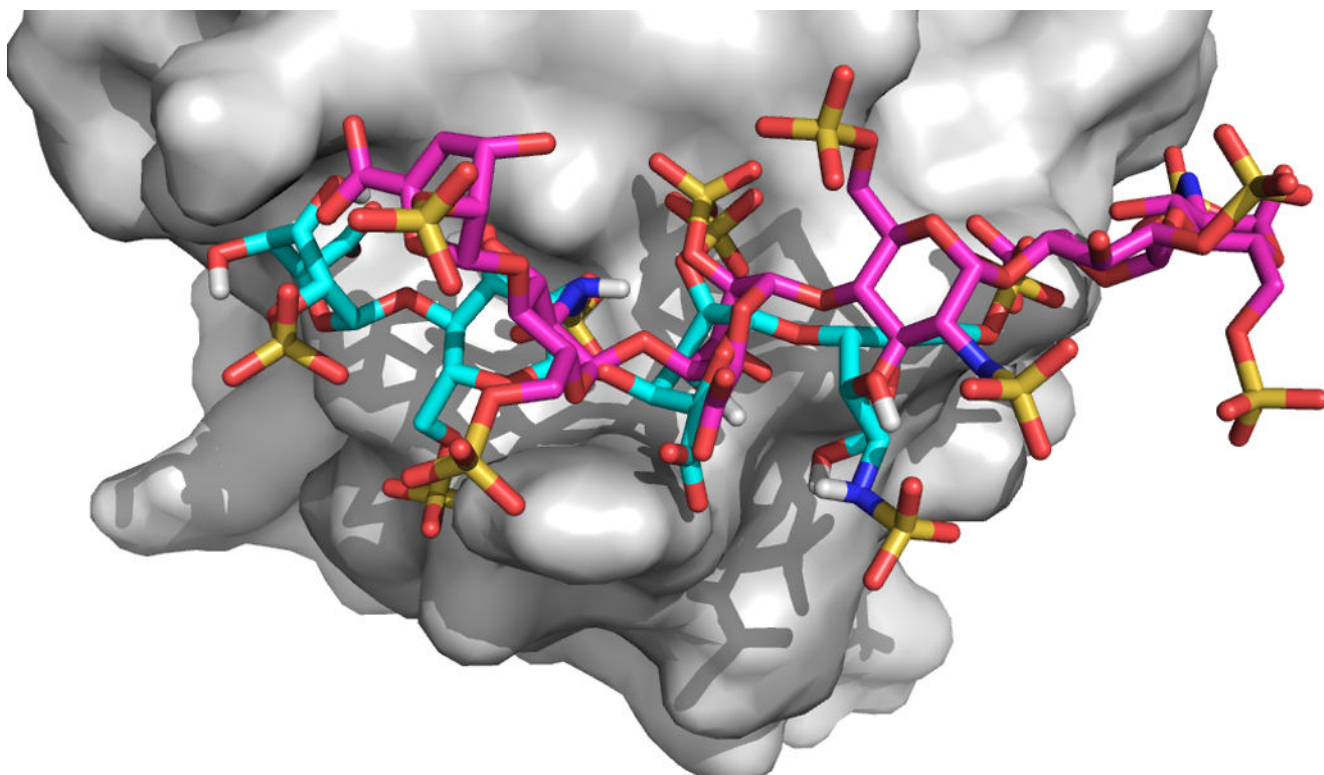


**FIGURE 11.** Docking the *E. coli* PliG lysozyme inhibitor to the salmon goose-type lysozyme using SAXS data as restraints. A near-native model (shown cyan cartoon) was obtained as the center of the 3<sup>rd</sup> largest cluster. The X-ray conformation of the inhibitor is shown as magenta cartoon.





**FIGURE 12.** Constructing the dimer of the sugar aminotransferase AtmS13 from *Actinomadura mellioura* by homology modeling and multimer docking. Chains A and B of the target dimer are shown in green and light green, respectively. Chain B, predicted from the homology model of chain A with the IRMSD of 2.62 Å, is shown as orange cartoon.



**FIGURE 13.** Docking of the heparin tetramer probe to the ligand-free structure of the basic fibroblast growth factor. The center of the second largest cluster is shown as cyan sticks. The X-ray structure of the bound hexamer shown in magenta.

**TABLE 1**

Weighting coefficients of PIPER energy terms in various docking modes

Coefficient set	<u>Energy term weight coefficients</u>			
	$E_{rep}$	$E_{attr}$	$E_{elec}$	$E_{DARS}$
Balanced	0.40	-0.40	600.0	1.0
Electrostatic-favored	0.40	-0.40	1200.0	1.0
Hydrophobic-favored	0.40	-0.40	600.0	2.0
Van der Waals + electrostatics	0.40	-0.40	600.0	0.0
Others Mode, Set 1	0.30	-0.30	300.0	1.25
Others Mode, Set 2	0.50	-0.20	300.0	0.50
Others Mode, Set 3	0.50	-0.20	300.0	0.0

TABLE 2

Important classes of direct docking algorithms

Method class	Properties		Examples
	Search Method	Protein Flexibility	
Global systematic rigid body docking	Fast Fourier Transform; Geometric matching	Minimal; smooth potential allows for some overlaps	ZDOCK <sup>53</sup> , GRAMM <sup>78</sup> , PIPER <sup>16</sup> , DOCK/PIERR <sup>120</sup>
Medium-range methods: Localized searches over selected regions	Monte Carlo minimization; Multi-start quasi-Newton minimizer with side chain search	Moderate, mostly side chains, some loops; motion along normal modes	RosettaDock <sup>67</sup> , ICM-DISCO <sup>121</sup> , ATTRACT <sup>68</sup> , SWARMDOCK <sup>74</sup>
Restraint-based docking	Supported by <i>a priori</i> information in the scoring function	Can be more substantial if restraints are available	HADDOCK <sup>69</sup>

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 3

Server performance based on the last three CAPRI evaluation meetings<sup>a</sup>

		CAPRI evaluation meeting year and number of targets					
		2009, 12 Targets		2013, 14 Targets		2016, 14 Targets	
Rank	Server	Success	Server	Success	Server	Success	
1	ClusPro <sup>14</sup> (Vajda/Kozakov)	5/1***/3***	ClusPro <sup>15</sup> (Vajda/Kozakov)	6/4**	ClusPro <sup>15</sup> (Vajda/Kozakov)	9/3***	
2	HADDOCK <sup>69</sup> (Bonvin)	4/1***/1**	HADDOCK <sup>69</sup> (Bonvin)	4/1***/1**	DOCK/PIERR <sup>120</sup> (Elber)	6/2***	
3	GRAMM-X <sup>78</sup> , (Vakser)	2/2**	SWARMDOCK <sup>74</sup> (Bates)	4/1**	LzerD <sup>122</sup> (Kihara)	4/1***/3**	
4	SKE-DOCK <sup>123</sup> (Uneyama)	2/1**	DOCK/PIERR <sup>120</sup> (Elber)	3/1**	HADDOCK <sup>69</sup> (Bonvin)	4/2**	

<sup>a</sup>Number of targets with acceptable or better quality predictions / number of targets with highly accurate (\*\*\*) predictions / number of target with medium accuracy (\*\*\*) predictions.

**TABLE 4**

Predictor group performance based on the last three CAPRI evaluation meetings<sup>a</sup>

CAPRI evaluation meeting, year and targets						
2009, 12 Targets		2013, 14 Targets		2016, 14 Targets		
Rank	Group	Success	Group	Success	Group	Success
1	Vajda/Kozakov <sup>16</sup>	6/4***/2**	Bonvin <sup>69</sup>	9/1***/3**	Guerois <sup>76</sup>	10/1***/8***
2	Zacharias <sup>124</sup>	6/4***/1**	Bates <sup>74</sup>	8/2**	Zacharias <sup>124</sup>	10/3***/2***
3	Zou <sup>125</sup>	6/3***/2**	Vakser <sup>78</sup>	7/1***	ClusPro <sup>15</sup>	9/3**
4	Eisenstein <sup>126</sup>	6/3***/1**	Kozakov/Vajda <sup>16</sup>	6/2***/3**	Kozakov/Vajda <sup>16</sup>	8/3***/2**
5	Wolfson <sup>127</sup>	6/3***/1**	Shen <sup>128</sup>	6/1***/3**	Seok <sup>129</sup>	8/3***/2**
6	Weng <sup>53</sup>	6/2***/2**	Fernandez-Recio <sup>130</sup>	6/1***/3**	Fernandez-Recio <sup>130</sup>	7/1***/3**
7	Zhou <sup>131</sup>	6/2***/2**	ClusPro <sup>15</sup>	6/4**	Zou <sup>125</sup>	7/1***/2**
8	Bonvin <sup>69</sup>	6/1***/4**	Zou <sup>125</sup>	6/1***/2**	Weng <sup>53</sup>	6/1***/4**
9	ClusPro <sup>14</sup>	5/1***/3**	Zacharias <sup>124</sup>	6/1***	Vakser <sup>78</sup>	6/2***/2**
10	Fernandez-Recio <sup>130</sup>	5/2**	Eisenstein <sup>126</sup>	5/1***/2**	Bates <sup>74</sup>	6/3**

<sup>a</sup>Number of targets with acceptable or better quality predictions / number of targets with highly accurate (\*\*\*) predictions / number of target with medium accuracy (\*\*) predictions.

TABLE 5

Troubleshooting table.

Step	Problem	Possible reason	Possible Solution
4	I uploaded a structure from my computer, but as soon as I submitted the job failed with the message: "Processing failed on receptor." What is wrong?	Inconsistency with the PDB format. Note that homology modeling tools frequently fail to produce correct PDB format, e.g., by not placing a chain identifier.	Make sure that your file is in PDB format and has the required spacing between columns. Documentation with respect to the PDB file format can be found at: <a href="http://www.wwpdb.org/docs.html">http://www.wwpdb.org/docs.html</a> . Also, make sure that all ATOM records the PDB file contain only standard amino or nucleic acids.
	I entered a HETATM record to include in the calculation, but it was not recognized. What do I do?	At this point there is no option for considering heteroatoms (HETATM entries).	Remove the HETATM records.
6	I uploaded a structure from my computer, but as soon as I submitted the job failed with the message: "Processing failed on ligand." What is wrong?	Inconsistency with the PDB format. Note that homology modeling tools frequently fail to produce correct PDB format, e.g., by not placing a chain identifier.	See TROUBLESHOOTING for Step 4.
10	My job keeps crashing even though my input file looks fine.	FFT grids are too large. This is typically caused by uploading a file output by homology modeling that has extremely large regions without a template, resulting in a tail that basically floats off into space.	Use the Structure Modification Option to remove the questionable parts of the model.
14	I cannot load the docking results into my molecular viewer.	Your viewer likely does not support multiple structures in one PDB file.	Try using PyMol, available at <a href="http://www.pymol.org">www.pymol.org</a>