이학박사 학위 논문

Prediction of Protein Interactions by Bioinformatics and Physical Chemistry Approaches

생물정보학과 물리화학적 접근법을 통한

단백질 상호작용 예측

2016년 2월

서울대학교 대학원

화학부 물리화학 전공

이 하 섭

Prediction of Protein Interactions by Bioinformatics and Physical Chemistry Approaches

지도교수 석 차 옥

이 논문을 이학박사 학위논문으로 제출함

2016년 2월

서울대학교 대학원

화학부 물리화학 전공

이 하 섭

이하섭의 이학박사 학위논문을 인준함

2015년 12월

위원장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

ABSTRACT

Prediction of Protein Interactions by Bioinformatics and Physical Chemistry Approaches

Hasup Lee

Department of Chemistry

The Graduate School

Seoul National University

Proteins play key roles in many biological systems through protein interactions. Research of protein interactions can help to understand protein functions and develop new drugs. Protein interactions can be classified into homo-oligomer interactions, protein-peptide interactions, and protein-protein interactions. Protein interactions can be studied based on co-crystallized complex structure determined by X-ray crystallography or Nucleic Magnetic Resonance method, but experimentally determined structures cover only small part of the known protein-protein interactions. Therefore, there are many interests to develop computational methods for predicting protein interactions. Predicting protein interactions can be classified into methods based on bioinformatics and physical chemistry approaches. According to bioinformatics approaches, proteins with high sequence similarity convey similar interfaces and similar interactions. According to physical chemistry

approaches, the funnel-like energy landscape is a general feature of protein interactions and protein interactions can be predicted by a global optimization method. In this thesis, I show bioinformatics and physical chemistry approaches for predicting homo-oligomer interactions, protein-peptide interactions, and protein-protein interactions. Both bioinformatics approaches and physical chemistry approaches played important roles to achieve improvement in predicting protein interactions.

Keywords: homo-oligomer interactions, protein-peptide interactions, protein-protein interactions, bioinformatics, physical chemistry, global optimization

Student Number: 2010-20290

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

Protein interactions play key roles in many biological systems. There are many interests to study protein interactions in biological system for controlling protein functions and developing new drugs (Ritchie 2008). Protein interactions can be classified into three categories: homo-oligomer interactions, protein-peptide interactions, and protein-protein interactions.

Homo-oligomer interactions are very important in many biological systems, because many proteins self-assemble into oligomers in order to perform their biological functions. For example, dimer interfaces of certain enzymes form as substrate-binding pockets. Also, antibodies form oligomers to create additional binding sites, increasing effective binding affinity via a “multivalent effect”. Many membrane proteins perform signal transduction by forming protein oligomers. There are many diseases related to mis-assembly of homo-oligomers (Levy *et al.,* 2008; Poupon and Janin 2010).

Protein-peptide interactions play important role in a broad range of biological processes, such as signaling pathways, immune system, apoptotic system, and post-translational modifications. The importance of such interactions is evident because of their involvement in critical human diseases, such as cancer and infections. Normally, protein-peptide interactions are mediated to small size of interface area. Because of the small sizes of protein-peptide interfaces, there have been many attempts to modulating protein-peptide interactions by small chemicals and synthetic peptides (London *et al.,* 2013; Petsalaki and Russell 2008).

Protein-protein interactions play key roles in various biological processes, such as cellular regulation, biosynthetic pathways, signal transduction, and DNA

replication. Also, protein-protein interactions are related to immune response, oligomer formation, and multi-molecular associations. To understand protein functions, it is essential to precisely describe protein-protein interactions in atomic details. (Keskin *et al.,* 2005; Perkins *et al.,* 2010)

Protein interactions can be studied by experimentally determined co-crystallized structure. However, despite the continuous increase in the number of deposited protein structures in the Protein Data Bank (PDB), the number of co-crystallized protein structures is still not sufficient to offer in-depth understanding of a majority of important biological processes. Furthermore, they cover less than 10% of the known protein-protein interactions in human. The large gap between the number of experimentally resolved structures for protein monomers and that for protein complexes in the PDB highlights the need to computational methods for predicting protein interactions that provide atomic structures using much less resources than experimental methods (Park *et al.,* 2015).

Computational methods for predicting protein interactions can be classified into two categories: bioinformatics approaches and physical chemistry approaches. For the bioinformatics approaches, sequence homologues convey similar interfaces and similar interactions. Some hotspot residues in interface regions guide to protein interactions. Theses residues are very conserved and called “interolog”. Therefore, searching good interolog is key to the success of predicting protein interactions by bioinformatics approach (Alsop and Mitchell 2015). For the physical chemistry approaches, funnel-like energy landscape is general feature of protein interactions, so native protein-peptide complexes and protein-protein complexes are the lowest free energy state. It is important to find global minimum in conformational space of energy landscape of protein-peptide complexes and

protein-protein complexes. In other words, study of predicting protein interactions can be classified as one of the global optimization problems (Lee *et al.,* 2005).

In this thesis, I will describe three computational methods: GalaxyGemini for predicting homo-oligomer interaction, GalaxyPepDock for predicting protein-peptide interactions, and GalaxyPPDock for predicting protein-protein interactions. GalaxyGemini generates oligomer models from input protein tertiary structure based on template information. First, GalaxyGemini searches homologues of query tertiary structure by sequence alignment method. Then, it predicts homo-oligomer interactions from database based on tertiary/quaternary structure similarity. Sequence similarity score, secondary structure similarity score, and alignment coverage of query sequence and template sequence are used to calculate tertiary structure similarity, and interface alignment score are used to calculate quaternary structure similarity. If oligomer template is found, the oligomer models are generated by superimposing query tertiary structure onto each subunits of selected oligomer template. The overall GalaxyGemini method is described in **chapter 2**. GalaxyPepDock generates protein-peptide complex models from input protein structure and peptide sequence. First, it searches co-crystallized protein-peptide template structures based on structural similarity of input protein structure and interaction similarity of input protein and peptide. Second, it performs energy-based optimization to generate more accurate models. The overall GalaxyPepDock method is described in **chapter 3**. GalaxyPPDock predicts protein-protein interactions based on physical chemistry approaches. It uses Cluster-Guided Conformational Space Annealing (CG-CSA), one of the most effective global optimization methods. The clusters are generated from initial structures and they evolved by communicating each other and changes number of members of each clusters. Instead of searching whole spaces of energy landscape, CG-CSA

concentrates on the nearby cluster regions. Effective sampling of CG-CSA can help to find global minimum and near-native structures. The overall GalaxyPPDock method is described in **chapter 4**.

2. GalaxyGemini: a program for protein homo-oligomer structure prediction based on similarity

2.1. Introduction

Many proteins self-assemble into oligomers in order to perform their biological functions (Poupon and Janin 2010). For example, certain enzymes form substrate-binding pockets at their dimer interfaces (Snijder *et al.,* 1999), whereas antibodies form oligomers to create additional binding sites, increasing effective binding affinity via a "multivalent effect" (Pluckthun and Pack 1997). Many membrane proteins also form oligomers for effective signal transduction (Heldin 1995). Knowledge of the protein oligomeric state is therefore crucial for understanding protein function at the molecular level.

In the case of experimental protein structures deposited in the Protein Data Bank (PDB), oligomeric states may be annotated by the authors or can be assigned from crystallographic information through the Protein Interfaces, Surfaces and Assembly (PISA) database (Krissinel and Henrick 2007). When such information is not available, e.g., for protein model structures, prediction of the oligomeric state is required. Recent studies have suggested that homology-based homo-oligomer prediction methods can be more powerful than *ab initio* methods (Morita *et al.*, 2012).

Methods for prediction of protein oligomeric structures were assessed in a blind fashion for the first time in the 9th Critical Assessment of Protein Structure

Prediction (CASP9) (Mariani *et al.,* 2011). In this experiment, participants were asked to predict homo-oligomer structures from amino acid sequences. Surprisingly, no method performed better than naïve predictors that take the top-ranking protein by HHsearch (Soding 2005) as a template, implying that the current methods for prediction of oligomeric structures are ineffective, with substantial room for improvement.

We developed a program named GalaxyGemini for predicting protein homo-oligomer structure, which shows clear improvement over other programs and naïve predictors tested on CASP9.

2.2. Methods

2.2.1. Overall procedure of GalaxyGemini

GalaxyGemini generates oligomer models from input protein tertiary structure based on template information. First, GalaxyGemini searches homologues of query tertiary structure using HHsearch (Soding 2005). Then, it determines whether query protein is monomer or oligomer using scoring function derived from HHsearch sequence score, HHsearch secondary structure score, alignment coverage of query sequence and template sequence, and interface alignment score. If query protein is determined as monomer, GalaxyGemini returns monomer. If query protein is determined as oligomer, clustering for oligomer templates is performed. Then, GalaxyGemini searches oligomer template based on scoring function and cluster sizes of oligomer templates, and subunit number prediction and contact prediction are performed based on selected oligomer template. Finally, the oligomer model is generated by superimposing query tertiary structure onto each subunits of selected oligomer template using TM-align (Zhang and Skolnick 2005) (**Figure 2.1**).

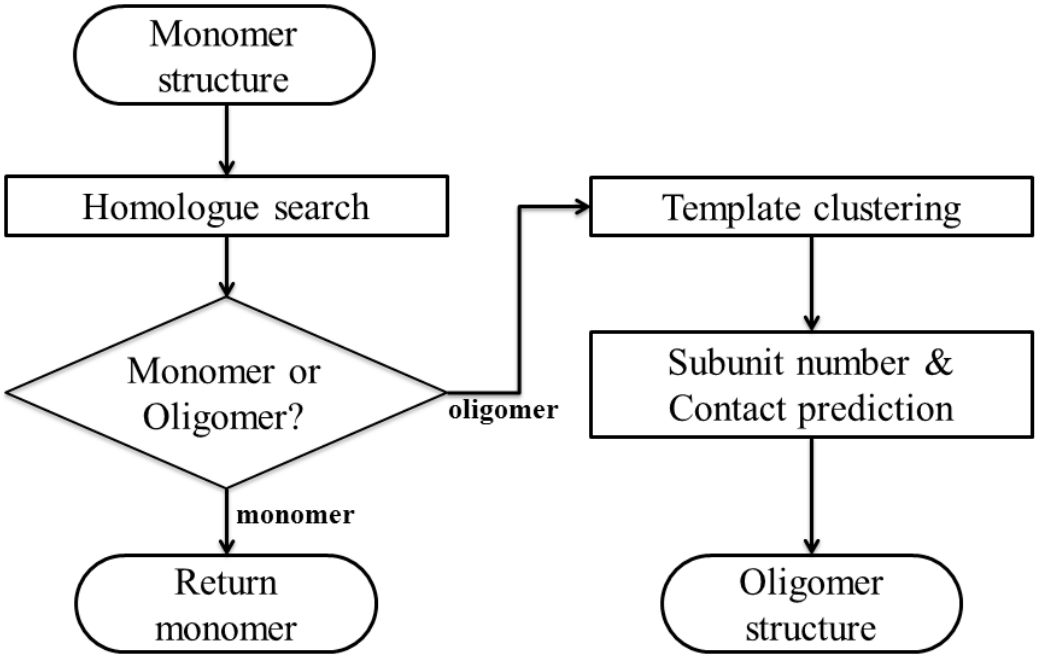


Figure 2.1. Flowchart of GalaxyGemini

2.2.2. Oligomer database and test sets

We constructed a database of known homo-oligomer structures containing 22,233 proteins with mutual sequence identity < 70% from all the structures deposited in the PDB (Apr 10, 2010). Oligomer templates are selected from this database. For each crystal structure, the oligomeric state was assigned as the biological unit determined by authors if "REMARK 350" in PDB was available and assigned by PISA otherwise. When PISA predicted multiple oligomeric states, the top oligomeric state was used, instead of being removed from the database, to increase the coverage of the database. According to the previous benchmark results, PISA assignments can be regarded reliable with a success rate of 80~90%. For protein structures solved by NMR, the oligomeric states were defined as the assembled chain structures in the PDB entry.

The database was generated before CASP9 experiment, so the current test results on the CASP9 set (96 proteins containing 43 monomers; Mariani *et al*., 2011) can be fairly compared with CASP9 predictors including Naïve predictors. For parameter training on the PISA benchmark set (195 proteins containing 55 monomers; Ponstingl *et al.*, 2003), target proteins were removed from the oligomer template lists.

2.2.3. Oligomer structure prediction

For a given input protein, HHsearch is first run on the oligomer database. Whether the query protein is oligomeric or not is then predicted by a scoring function $S1$. If the top-ranking protein is monomeric, the query protein is predicted to be monomeric. Otherwise, an oligomer template is selected by ranking with a

second function *S*2. Prediction of the oligomeric state corresponding to each template is obtained by superimposing the input monomer structure onto the subunits of the oligomer template using the structure alignment tool TM-align (Zhang and Skolnick 2005). Finally, rigid-body energy minimization is performed to remove steric clashes at the oligomer interface as explained in Supplementary Information.

The 2 scoring functions *S*1 and *S*2 are expressed as the weighted sums of Z-scores of 5 components. The first 4 components are derived from HHsearch: (i) HHsearch sequence score, (ii) HHsearch secondary structure score, (iii) ratio of aligned residues to the query sequence length and (iv) ratio of aligned residues to the sequence length of template candidate in the HHsearch alignment. These components account for sequence similarity to the query protein. The fifth component, called interface alignment score, accounts for tertiary and quaternary structure similarity by adding BLOSUM62 matrix scores (Henikoff and Henikoff 1992) between the interface residues of template candidate and the residues of the query protein aligned to them. Addition of this component is important because interface residues are more conserved than other surface residues (Caffrey *et al.,* 2004). The weight parameters for the 2 scoring functions were determined by training on the PISA benchmark set with a grid search.

2.2.4. Scoring function for predicting oligomer state

The function *S*1 used for scoring candidate proteins is expressed as a weighted sum of the five components described in the main text as follows:

$$S_1 = \begin{cases} 10\,Z_{\mathrm{Seq}} + 15\,Z_{\mathrm{SS}} + 15(Z_{\mathrm{Cov1}} + Z_{\mathrm{Cov2}}) + 0\,Z_{\mathrm{Interf}} & \text{if monomer ratio} > 0.6 \\ 10\,Z_{\mathrm{Seq}} + 10\,Z_{\mathrm{SS}} + 15(Z_{\mathrm{Cov1}} + Z_{\mathrm{Cov2}}) + 2\,Z_{\mathrm{Interf}} & \text{otherwise} \end{cases} \quad (2.1)$$

where Z_{Seq}, Z_{SS}, Z_{Cov1}, Z_{Cov2} and Z_{Interf} stand for the Z-scores of HHsearch sequence score, HHsearch secondary structure score, ratio of the aligned residues to the query sequence length, ratio of the aligned residues to the candidate sequence length and the interface alignment score defined as

Interface alignment score = $\sum_j^N s(i_j\ , j)$,

$$s(i_j, j) = \begin{cases} BLOSUM62\ (aa_i, aa_j) & \text{if } j \text{ is interface residue} \\ 0 & \text{otherwise} \end{cases} \quad (2.2)$$

where j is the residue index of the candidate protein, N is the total number of residues in the candidate protein, i_j is the residue index of the query protein aligned to the jth residue of the candidate protein, aa_i and aa_j are amino acid types of residues i_j and j, respectively. The Z-score for each component is calculated for a background pool of top 2000 proteins ranked by HHsearch sequence score.

The weight parameters of $S1$ depend on the 'monomer ratio' defined as

$$\text{monomer ratio} = \frac{\sum Z_{\mathrm{Seq}}(\text{monomer})}{\sum Z_{\mathrm{Seq}}(\text{monomer}) + \sum Z_{\mathrm{Seq}}(\text{oligomer})} \quad (2.3)$$

where ΣZ_{Seq} (monomer) and ΣZ_{Seq} (oligomer) are the sums of the Z-scores of HHsearch sequence scores for monomeric candidates and oligomeric candidates, respectively, with HHsearch probability > 90%. If there is no protein with HHsearch probability > 90%, the top ranking protein is selected as the oligomer template.

2.2.5. Scoring function for predicting homo-oligomer interactions

The second scoring function is used to select the oligomer template, which is used to predict the number and orientations of the subunits of oligomer. It has the same functional form as *S*1, but the weights are different as follows:

$$S_2 = \begin{cases} 10\,Z_{\mathrm{Seq}} + 10\,Z_{\mathrm{SS}} + 3(Z_{\mathrm{Cov1}} + Z_{\mathrm{Cov2}}) + 4\,Z_{\mathrm{Interf}} & if\ (CLC > 0.7) \\ 10\,Z_{\mathrm{Seq}} + 10\,Z_{\mathrm{SS}} + 3(Z_{\mathrm{Cov1}} + Z_{\mathrm{Cov2}}) + 7\,Z_{\mathrm{Interf}} & if\ (0.4 < CLC \leq 0.7) \\ 10\,Z_{\mathrm{Seq}} + 15\,Z_{\mathrm{SS}} + 3(Z_{\mathrm{Cov1}} + Z_{\mathrm{Cov2}}) + 10\,Z_{\mathrm{Interf}} & if\ (CLC \leq 0.4) \end{cases} \quad (2.4)$$

The weight factors for the second and the last terms vary depending on the target difficulty estimated by a parameter *CLC* (convergence of the largest cluster) defined as

$$CLC = \frac{\sum_{\text{candidates in the largest cluster}} Z_{\mathrm{Seq}}}{\sum_{\text{all candidates}} Z_{\mathrm{Seq}}} \quad (2.5)$$

which estimates the degree of convergence of the largest cluster of the template candidates. The summation is over proteins with HHsearch probability > 90%. Clustering is carried out by a greedy algorithm with similarity criterion (contact similarity) > 0.5. Contact similarity between two protein structures A and B are calculated as

$$\text{Contact similarity} = \frac{N(\text{Contacts in } A \cap \text{Contacts in } B)}{N(\text{Contacts in } A)} \quad (2.6)$$

where $N(\ )$ is the number of inter-subunit residue contacts (C_β distance < 12 Å).

The weight factors of scoring function *S*2 were determined by performing three-fold cross-validation. The sets for cross-validation was generated by randomly dividing the PISA benchmark set into three subsets maintaining approximate proportions of different oligomers, as reported in **Table 2.1**. Fixing the parameters for sequence score at 10, the number of trained parameters was three

for each of 3 difficulty ranges, as can be seen from **Eq. 2.4**. The parameters trained on the subsets were pretty robust, although variations in the third component, interface alignment score, were found. The final parameter set corresponds to that of the first fold, which shows the same average contact agreement score S_{agree} for both training and test sets.

Table 2.1. Weight factors for scoring function *S*2 determined by three-fold cross-validation

		Number of proteins		Average S_{agree}		Parameters
		Training	Test	Training	Test	(SS, Cov, Interf)
Fold1	1-mer	37	18	0.63	0.63	(10,3,4) (10,3,7) (15,3,10)
	2-mer	50	25			
	3-mer	16	8			
	4-mer	21	10			
	6-mer	6	4			
Fold2	1-mer	37	18	0.64	0.61	(10,3,4) (10,3,7) (15,3,6)
	2-mer	50	25			
	3-mer	16	8			
	4-mer	20	11			
	6-mer	7	3			
Fold3	1-mer	36	19	0.62	0.64	(10,3,2) (10,3,7) (15,3,10)
	2-mer	50	25			
	3-mer	16	8			
	4-mer	21	10			
	6-mer	7	3			
All	1-mer	55	-	0.63	-	(10,3,4) (10,3,7) (15,3,10)
	2-mer	75				
	3-mer	24				
	4-mer	31				
	6-mer	10				

2.2.6. Energy minimization

An oligomer structure generated by superimposition onto the template structure may have steric clashes at the oligomer interface because the input monomer structure at the interface may be different from that of template. To remove such steric clashes, rigid-body energy minimization by a Monte Carlo method is performed fixing the internal structure of monomer subunits. The objective energy function is a sum of physicochemical energy implemented in the GALAXY (Park and Seok 2012) and harmonic restraints for the distances between C_α atoms at the interface (C_α distance < 14 Å) of the oligomer template.

2.2.7. Assessment measures

Identification of the correct number of subunits in an oligomer was evaluated by measuring the "relative accuracy" (Acc_{Rel}). For more precise evaluation of the predicted structure, the "contact agreement score" (S_{agree}) was measured, which reflects the fraction of correctly modeled interface contacts in the complex.

"Relative accuracy" (Acc_{Rel}) is an accuracy measure for the number of subunits defined as

$$Acc_{\mathrm{Rel}} = \frac{\text{Number of targets with correctly predicted number of subunits}}{\text{Number of targets}} \times 100\ (\%) \qquad (2.7)$$

Contact agreement score (S_{agree}) is a measure for interface contact similarity between the native and predicted oligomer structures defined as

$$S_{\mathrm{agree}} = \frac{\sum f(x_{ij}, y_{ij})}{\sum g(x_{ij}, y_{ij})} \qquad (2.8)$$

$$f(x_{ij}, y_{ij}) = \begin{cases} 1 - \frac{|x_{ij} - y_{ij}|}{\max(x_{ij}, y_{ij})} & if\ \max(x_{ij}, y_{ij}) > 0 \\ 0 & if\ \max(x_{ij}, y_{ij}) = 0 \end{cases} \quad (2.9)$$

$$g(x_{ij}, y_{ij}) = \begin{cases} 1 & if\ \max(x_{ij}, y_{ij}) > 0 \\ 0 & if\ \max(x_{ij}, y_{ij}) = 0 \end{cases} \quad (2.10)$$

where x_{ij} and y_{ij} are the numbers of contacts (C_β distance < 12Å) between residue i and residue j that belong to different protein subunits for the native and the predicted oligomer structures, respectively. The number of residue i is same as the number of subunits. S_{agree} ranges from 0 to 1. S_{agree} = 1 corresponds to the exactly same contacts between the native and model structures, and S_{agree} = 0 to no match between contacts in the native and model structures.

2.3. Results and Discussion

2.3.1. Performance of GalaxyGemini on training set and test set

We tested on GalaxyGemini on PISA benchmark set and CASP9 oligomer set. GalaxyGemini increased relative accuracy from 75.4% (for the naïve predictor NaïveSeqScore that takes the HHsearch top ranker by sequence score) to 79.5% for the training set (PISA benchmark set) and from 69.8% to 77.1% for the test set (CASP9 set). The sum of S_{agree} over the targets increased from 74.7 to 88.0 for the training set and from 13.6 to 17.6 for the test set when "experimental" monomer structures were used as input (**Figure 2.2**). When tertiary structures predicted by GalaxyTBM (Ko *et al.,* 2012) were used as input for the CASP9 set, the sum of S_{agree} increased from 9.4 to 12.1. Sum of S_{agree} of NaïveCoverage is 9.8, the largest value among CASP9 predictors, but sum of S_{agree} of GalaxyGemini is also larger than that of NaïveCoverage (**Figure 2.3**). GalaxyGemini outperforms all other CASP9 predictors and naïve predictors by the two measures, Acc_{Rel} and S_{agree}, implying that GalaxyGemini may be successfully applied to "sequence-based" oligomeric structure prediction (**Figure 2.4**).

A successful example of CASP9 target T0576 (3na2) highlights the strength of GalaxyGemini (**Figure 2.5**). This protein forms a dimer through an inter-chain β-sheet. The best template determined by the NaïveSeqScore (2grg) is monomeric, but GalaxyGemini successfully found a dimer template (3fm2), which has an oligomer structure similar to the native structure, resulting in a high S_{agree} of 0.742. A tetramer target T0632 (3nwz) is also successful case. The best template selected by NaïveSeqScore (1vh9) is dimer, but the best template selected by GalaxyGemini (3f5o) is tetramer similar to the native structure, resulting in high S_{agree} of 0.708 (**Figure 2.6**). GalaxyGemini predicted inter-chain interactions of β-

strand of tetramer target T0632 based on selected template. These results showed that GalaxyGemini searches better templates than NaïveSeqScore on both dimer and tetramer targets.

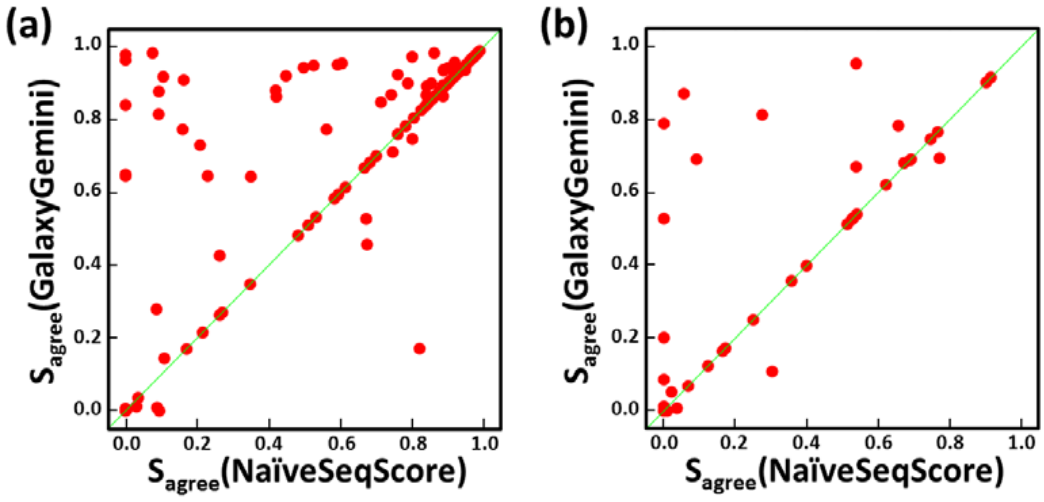


Figure 2.2. Target-based comparison of the performance of GalaxyGemini with that of a naïve predictors NaïveSeqScore as measured by S_{agree} for the (a) PISA benchmark set, (b) CASP9 set using the experimental monomer structure as input.

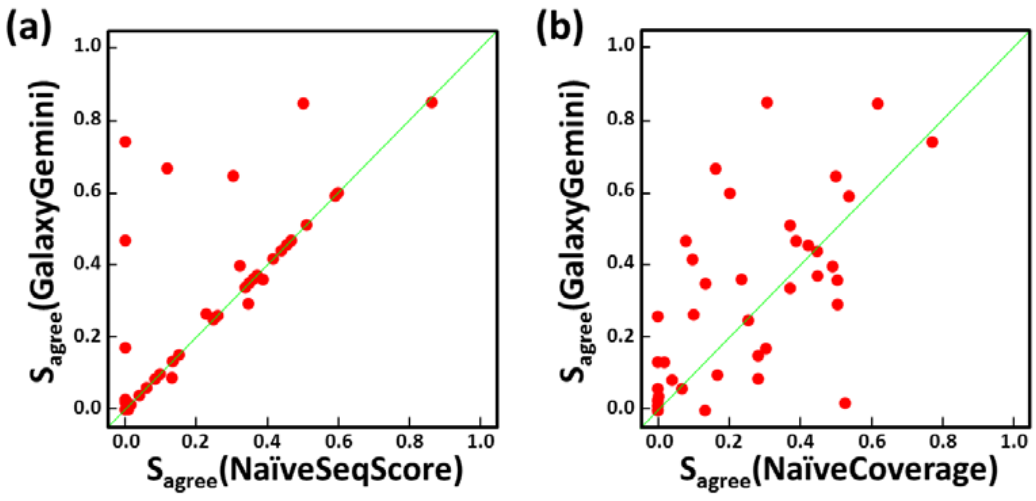


Figure 2.3. Target-based comparison of the performance of GalaxyGemini with that of a naïve predictors (c) NaïveSeqScore and (d) NaïveCoverage as measured by S_{agree} for the CASP9 set using the model structure as input.

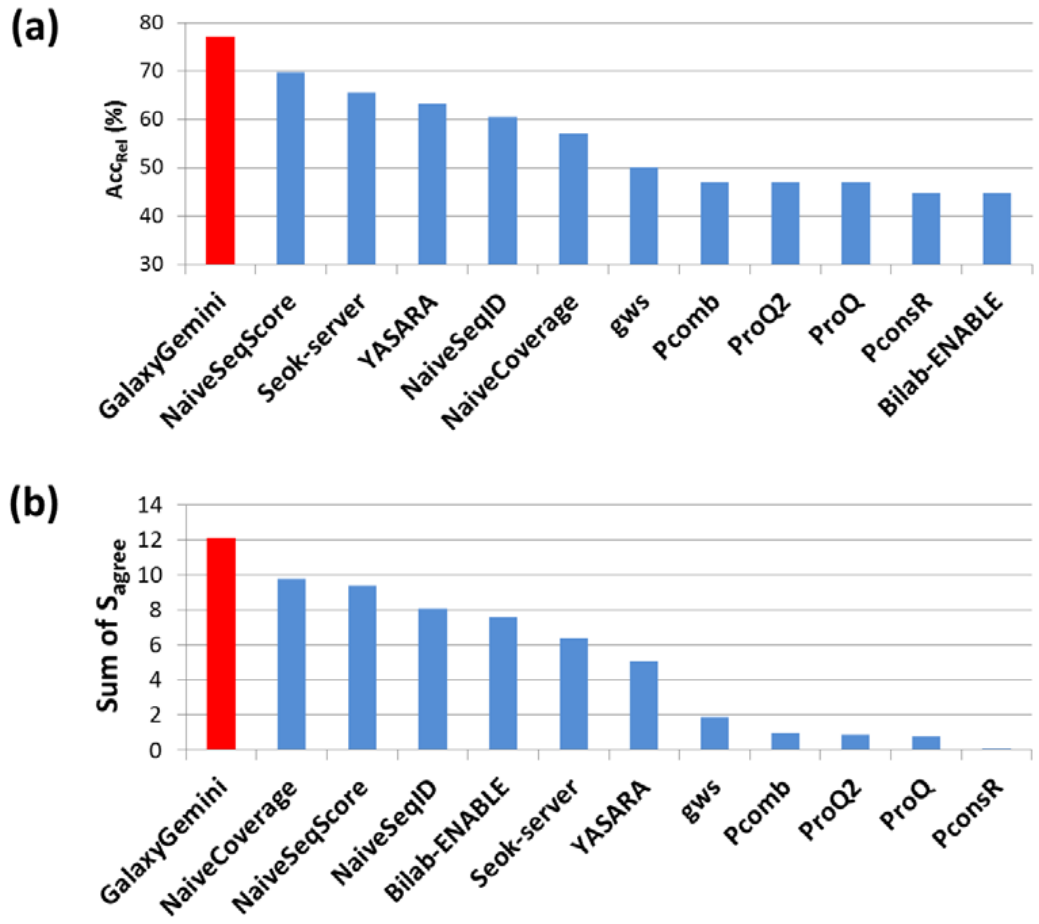


Figure 2.4. Comparison of the performance of GalaxyGemini as measured by (a) relative accuracy and (b) sum of S_{agree} for the CASP9 set with those of CASP9 predictors and 3 naïve methods which take the HHsearch top ranker by sequence score (NaïveSeqScore), sequence identity (NaïveSeqID), and coverage (NaïveCoverage).

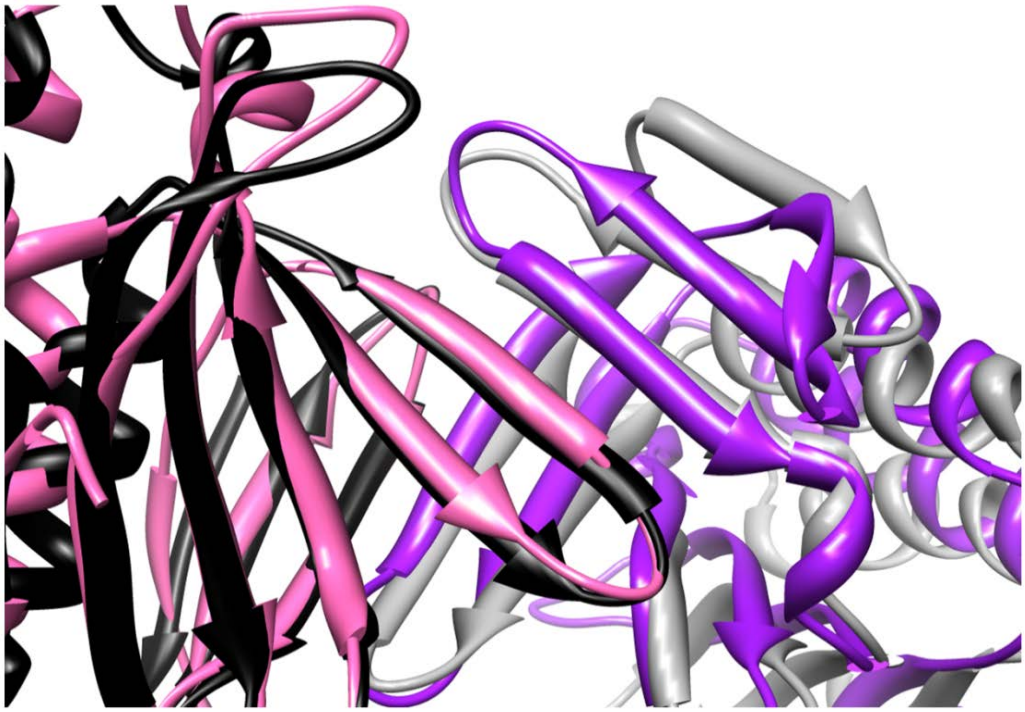


Figure 2.5. A successful dimer example (T0576, 3na2) of GalaxyGemini. Subunits of the native structure are shown in black and gray and those of the predicted structure in pink and purple.

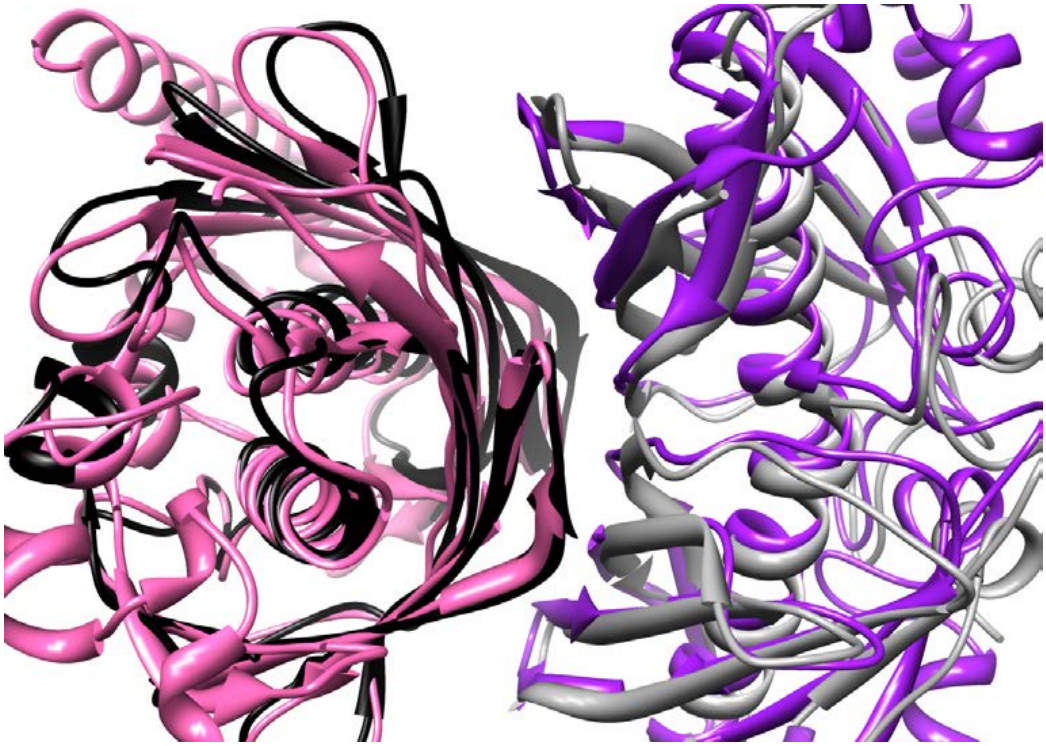


Figure 2.6. A successful tetramer examples (T0632, 3nwz) of GalaxyGemini. Subunits of the native structure are shown in black and gray and those of the predicted structure in pink and purple.

2.3.2. Contribution of score components

Among the five components of the GalaxyGemini scores *S*1 and *S*2, HHsearch sequence score contributes the most to the performance in terms of both relative accuracy and sum of contact agreement score. Contributions of the additional components were analyzed by successively adding more components to the sequence score, as shown in **Table 2.2**. Although improvement by adding three additional terms on the training set (PISA set) is rather small for the relative accuracy of subunit numbers (improved by 5.4%) which is already high (75%) with the sequence score alone, improvement is more significant for the contact agreement score (improved by 18%). Among the additional components, the interface alignment score contributes the most to the improved performance on the training set (PISA set). Interestingly, secondary structure score turned out to be important in increasing the relative accuracy for the test set (CASP9 set). This seems to be related to the fact that better templates for template-based modeling were obtained by including secondary structure score for more difficult targets in a previous study (Ko *et al.,* 2012). Overall, the weighted sum of all five energy components can maximize the performance for both training set and test set.

Table 2.2. Contribution of components of the GalaxyGemini scores

Components	Acc_{Rel}		Sum of S_{agree}	
	PISA Set	CASP9 Set	PISA Set	CASP9 Set
Seq	75.4%	69.8%	74.7	13.6
Seq + SS	72.8%	76.0%	72.4	15.1
Seq + Cov	76.4%	71.9%	78.8	15.4
Seq + Interf	78.5%	72.9%	86.9	14.9
Seq + SS + Cov	75.9%	74.0%	74.7	15.6
Seq + SS + Interf	76.9%	76.0%	87.3	17.7
Seq + Cov + Interf	78.5%	74.0%	87.5	16.7
Seq + SS + Cov + Interf	**79.5%**	**77.1%**	**88.0**	**17.6**

2.3.3. Oligomer states for improvement cases on CASP9 targets

We followed the assignments of oligomeric states made by the CASP9 assessors, as ex explained in Supplementary Table S1 of the CASP9 assessment paper (Mariani *et al.,* 2011). In **Table 2.3**, assignments for the CASP9 targets were showed for improved predictions of GalaxyGemini compared to NaïveSeqScore. All but one target had no ambiguities in the oligomer state assignment. The target T0632 for which both authors and PISA assigned two states was assigned to be a tetramer by CASP assessors after closer examination of PISA scores and structural details. Improvements are mostly on dimers for the CASP9 set (8 out of 12), but this fraction (67%) is smaller than that of dimers (78%) in CASP9 set, implying that GalaxyGemini may not be necessarily biased to dimers and GalaxyGemini also generate good models on tertiary or tetramer targets.

Table 2.3. Oligomeric state assignment of the CASP9 targets for which GalaxyGemini showed improved predictions over the naïve predictor NaïveSeqScore

TARGET	PDB ID	Assignment			Comment
		Author	PISA	CASP	
T0523	3mqo	2-mer	2-mer	2-mer	
T0536	3mxq	4-mer	4-mer	4-mer	
T0542	3n05	2-mer	2-mer	2-mer	
T0565	3npf	2-mer	2-mer	2-mer	
T0576	3na2	2-mer	2-mer	2-mer	
T0584	3nf2	2-mer	2-mer	2-mer	
T0586	3neu	2-mer	2-mer	2-mer	
T0592	3nhv	3-mer	3-mer	3-mer	
T0611	3nnr	2-mer	2-mer	2-mer	
T0632	3nwz	2,4-mer	2,4-mer	4-mer	Authors assigned different states, but the tetramer is confirmed as most stable complex.
T0635	3n1u	4-mer	4-mer	4-mer	
T0636	3p1t	2-mer	2-mer	2-mer	

2.4. Conclusions

We developed GalaxyGemini to predict homo-oligomeric structure from query protein tertiary structure. GalaxyGemini was successfully tested on both PISA benchmark set and CASP9 oligomer set. The performance of GalaxyGemini was better than other oligomer prediction methods tested in CASP9, implying wider applicability to oligomer state prediction from sequence.

3. GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization

3.1. Introduction

Protein-protein interactions that are mediated by short linear peptides of interacting partners are critical in a broad range of biological processes, such as signaling pathways, protein cellular localization and post-translational modifications (Miller *et al.,* 2008; Petsalaki and Russell 2008; Scott and Pawson 2009; Wen *et al.,* 1995). The importance of such interactions is evident because of their involvement in critical human diseases, such as cancer and infections (Maclaine and Hupp 2011). Because of the small sizes of protein-peptide interfaces, such interactions can be modulated by small chemicals or synthetic peptides (Vlieghe *et al.,* 2010; Yang *et al.,* 2005). Therefore, effective computational modeling of protein-peptide interactions can provide useful information for understanding complex biological processes in molecular detail and for modulating protein-protein interactions for disease treatment.

As in other areas of molecular modeling, it is very difficult to obtain reliable predictions by computational protein-peptide docking when prior knowledge of the interactions is not available. When there is no information on the binding site, putative binding sites must be searched for on the entire surface of the target protein. Such global docking methods show limited accuracy for predicting high-resolution complex structures, but successful predictions of at least part of the binding residues have been reported (Lavi *et al.,* 2013; Petsalaki *et al.,* 2009; Yan

and Zou 2015). When experimental or predicted data on binding site residues are available, such information can be used to constrain the docking to local regions of the protein surface (Trellet *et al.,* 2013). These local docking methods usually require a model protein-peptide complex structure as input, whereas global docking methods require only a protein structure and a peptide sequence. Among the various protein-peptide docking methods developed so far, only a small number of methods are available, such as PepSite (Trabuco *et al.,* 2012), PEP-SiteFinder (Saladin *et al.,* 2014), and CABS-dock (Kurcinski *et al.,* 2015) for global docking and Rosetta FlexPepDock (London *et al.,* 2011; Raveh *et al.,* 2010; Raveh *et al.,* 2011) and PepCrawler (Donsky and Wolfson 2011) for local docking.

As increasing number of protein-peptide complex structures are being deposited in the PDB, the probability of finding protein-peptide complexes similar to a given target complex in the structure database increases. For example, 87% of the non-redundant protein-peptide complexes in the PeptiDB set (London *et al.,* 2010) have similar proteins, with a protein TM-score > 0.6, among the experimentally resolved structures that were published prior to the given complex. Because protein-peptide interactions are usually stabilized through hot spot interactions (London *et al.,* 2010; London *et al.,* 2013), the observed hot spot interactions in known protein-peptide complex structures can be useful for predicting interactions that involve a range of new variations in target proteins and peptides.

The GalaxyPepDock utilizes information on protein-peptide interactions of similar proteins in the database of experimentally determined structures to generate high-resolution complex structures when reasonable template protein-peptide complex structures can be found. A further refinement by GALAXY

energy-based optimization (Heo *et al.,* 2013; Park *et al.,* 2011; Park and Seok 2012; Park *et al.,* 2014) enables the modeling of structural differences between the template and target complex structures by sampling the backbone and side chain flexibilities of both protein and peptide. GalaxyPepDock were successfully test on PeptiDB benchmark set, and showed good performance compared to other popular protein-peptide docking programs: PEP-SiteFinder, CABS-dock, and PepSite. Also, when tested on the CAPRI target 67, predictions of medium accuracy were made; this accuracy is among the best predictions made by human groups and superior to the best server predictions submitted during the CAPRI blind prediction experiment. For this target, the conformational change of the protein by peptide binding was also correctly predicted.

3.2. Methods

3.2.1. Overall procedure of GalaxyPepDock

GalaxyPepDock consists of two steps for protein-peptide docking. First, GalaxyPepDock searches crystallized protein-peptide template based on structural similarity of protein structure and interaction similarity of protein and peptide. Second step is energy-based optimization step. Protein-peptide models are generated based on molecular dynamics-based method using GalaxyTBM and GalaxyRefine. The energy function for energy-based optimization is summation of physics-based energy function used in GalaxyRefine and C_α restraints derived from selected template (**Figure 3.1**).

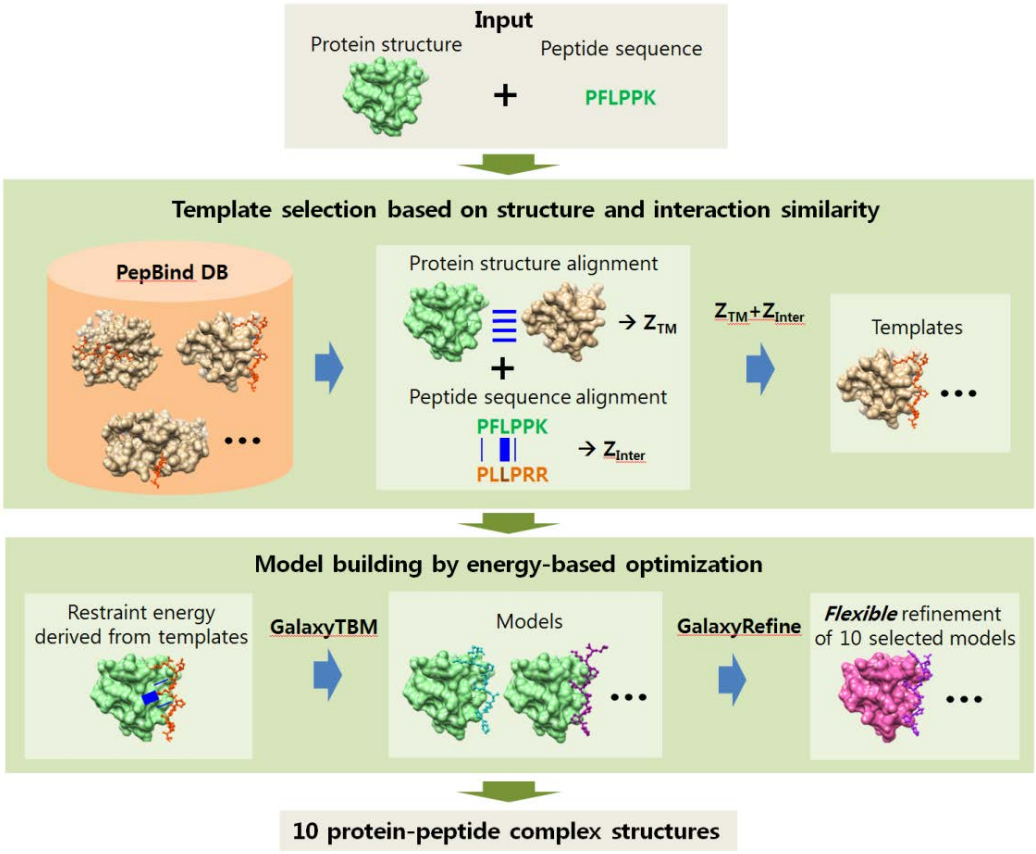


Figure 3.1. Flowchart of the GalaxyPepDock.

3.2.2. Template selection

Templates for protein-peptide complex structure prediction are selected from the PepBind (Das *et al.,* 2013) database with the following score for each complex structure in the database

$$S_{\text{complex}} = Z_{\text{TM}} + Z_{\text{Inter}} \quad (3.1)$$

where Z_{TM} measures the protein structure similarity by the Z-score of the TM-score of a database protein structure when aligned to the target protein structure by TM-align (Zhang and Skolnick 2005) and Z_{Inter} measures the interaction similarity of a database complex and the target complex when aligned to the former by the Z-score of the interaction similarity score S_{Inter} defined below. Up to 10 complexes with $S_{complex}$ > 90% of the maximum value are selected as templates and used in the model-building procedure described in the next subsection.

To measure the interaction similarity of a database complex and the target complex, the target complex is first aligned to the database complex by protein structure alignment and peptide sequence alignment. Peptide alignment is performed by gapless sequence alignment with a modified BLOSUM62 (Henikoff and Henikoff 1992) matrix score, by multiplying the weight of (1 + the number of hydrophobic or ionic protein residues contacting the given peptide residue in the template complex structure) to the BLOSUM62 matrix components with scores > 0. Hydrophobic (or ionic) protein-peptide residue pairs with at least one heavy atom pair within 5.0 Å (or 6.0 Å) are considered to be contacting following the PepBind criterion (Das *et al.,* 2013). In this way, more emphasis is put on the peptide residues contributing to hot spot interactions than on other residues during peptide alignment. An example case of peptide alignment is provided in **Figure 3.2**. The

interaction similarity score S_{Inter} is then calculated by summing the interaction pair similarity score $S_{i\text{-}j}$ for all of the protein-peptide residue pairs (*i-j*) in contact in the template complex, as illustrated in **Figure 3.3** for the example case. $S_{i\text{-}j}$ is measured by the similarities in the amino acids of the contacting pair (*i-j*) in the template complex and of the corresponding pair (*i'-j'*) in the target complex aligned to the template and is defined as $S_{i\text{-}j} = Max[\ B(i,i') + B(j,j'),\ B(i,j') + B(j,i')\ ]$, where $B(i,i')$ is the BLOSUM62 matrix component for the amino acid of residue *i* and that of residue *i*'.

j' =		1	2	3	4	5	6	7	8	
Query peptide	—	P	P	P	A	L	P	P	K	K
		|	|	|	|	|	|	|	|	
Template peptide	A	F	A	P	P	L	P	R	R	—
j =		1	2	3	4	5	6	7	8	
				Scores						
$B(j, j')$[1]	-8	-4	-1	7	-2	4	7	-2	2	-8
$N_{inter}(j)$[2]	0	1	2	2	1	4	3	0	1	—
Modified $B(j, j')$[3]	-8	-4	-1	21	-2	20	28	-2	4	-8

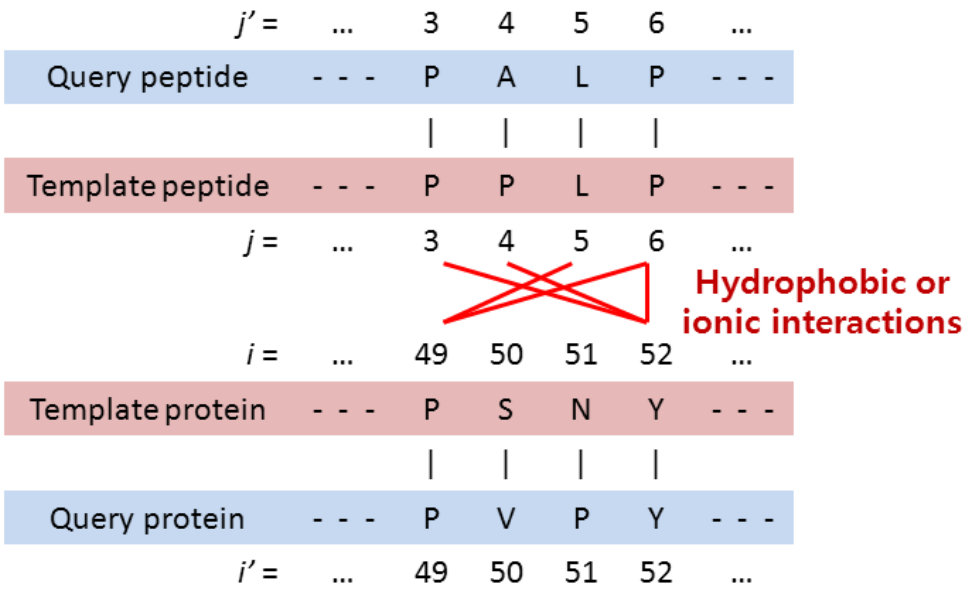
[1] $B(j, j')$ = BLOSUM62[amino acid (j), amino acid (j')]

[2] $N_{inter}(j)$ = Number of interacting hydrophobic or ionic protein residues for peptide residue j in the template complex structure

[3] Modified $B(j, j') = [1+N_{inter}(j) \times \Theta\{B(j, j')\}] \times B(j, j')$

$\Theta(B) = 0$ if $B \leq 0$, 1 if $B > 0$

Figure 3.2. Peptide alignment of GalaxyPepDock performed with a modified BLOSUM62 matrix



$$S_{\text{Inter}} = \sum_{\text{all } (i,j) \text{ pairs}} S_{i-j}$$

$$S_{i-j} = Max\left[B(i,i') + B(j,j'), B(i,j') + B(j,i')\right]$$

(i,j)	(49,5)	(49,6)	(52,3)	(52,4)	(52,6)	...
S_{i-j}	11	14	14	6	14	...

Figure 3.3. Calculation of interaction similarity score S_{inter} of GalaxyPepDock.

3.2.3. Model-building

For each template, 50 model complex structures are first generated with the model-building tool of GalaxyTBM (Ko *et al.,* 2012), using protein structure alignment and peptide sequence alignment. For the model-building optimization, restraints on the distances between interacting protein-peptide pairs are added to the GALAXY energy, with weights dependent on the interaction pair similarity score $S_{i\text{-}j}$ (**Figure 3.4**). Interaction pairs with higher similarities to the template tend to be conserved by stronger template-derived restraints, whereas the sampling of other parts of the structure is driven more by the physics-based energy than by template-derived information. Of the model structures generated by GalaxyTBM, 10 structures are selected by choosing the structures with the best energy values for each template and are further refined following the GalaxyRefine (Heo *et al.,* 2013) protocol. This refinement step allows for the adjustment of the backbone and side chain structures by repetitive molecular dynamics relaxations after side chain repacking.

$$E\left(\left\{\mathbf{r}_{i'},\mathbf{r}_{j'}\right\}\right)=E_{\mathrm{GALAXY}}\left(\left\{\mathbf{r}_{i'},\mathbf{r}_{j'}\right\}\right)+\sum_{\text{all }(i,j)\text{ pairs}}k_{i-j}\left(r_{i'j'}-r_{i'j'}^{(0)}\right)^2$$

(i,j)	(49,5)	(49,6)	(52,3)	(52,4)	(52,6)	...
S_{i-j}	11	14	14	6	14	...
k_{i-j}	6	6	6	4	6	...

S_{i-j}	Weight (k_{i-j})
$S_{i-j} < 0$	1
$0 < S_{i-j} \leq 5$	2
$5 < S_{i-j} \leq 10$	4
$10 < S_{i-j}$	6

Figure 3.4. GalaxyPepDock energy function for protein-peptide model building.

3.2.4. Evaluation measure

To evaluate the performance of GalaxyPepDock, four measures were used: LRMSD (peptide RMSD), IRMSD (interface RMSD), f_{nat} (fraction of native contact), and f_{site} (fraction of native binding site). For the definitions of acceptable/medium accuracy predictions, the following CAPRI criterion was used: acceptable prediction if (LRMSD < 4 Å or IRMSD < 2 Å) and f_{nat} > 0.2 and medium prediction if (LRMSD < 2 Å or IRMSD < 1 Å) and f_{nat} > 0.5 (Lensink and Wodak 2013). The values of LRMSD, IRMSD, and f_{nat} were used to compare GalaxyPepDock to PEP-SiteFinder and CABS-dock, and the value of f_{site} was used to compare to PepSite.

3.3. Results and Discussion

3.3.1. Performance compared to other protein-peptide docking programs

The performance of GalaxyPepDock was compared with those of three available protein-peptide docking programs, PEP-SiteFinder, CABS-dock, and PepSite, which perform global protein-peptide docking and thus do not require the protein-peptide structure as input. Because PEP-SiteFinder, CABS-dock, and PepSite are ab initio methods that do not rely on template information, the comparison of the results presented here demonstrate the extent to which a similarity-based method such as GalaxyPepDock can be useful compared with the ab initio methods for the benchmarking set. For a fair comparison, the complexes in the PepBind database that were released after each target complex were excluded during template search in GalaxyPepDock prediction. The accuracy of the best model of the 10 generated models was evaluated for each method.

The non-redundant set of PeptiDB (London *et al.,* 2010) was first employed for comparison. Peptide docking to unbound protein structures was performed on 57 of the 103 PeptiDB complexes for which unbound protein structures are available in the structure database because re-docking peptides to bound protein structures is only of theoretical interest. For the 40 PeptiDB targets that have ≤ 10 residue-long peptides that are accepted by PepSite, GalaxyPepDock identified 75.4% of the binding site residues on average, compared with the 66.2%, 64.1%, and 40.9% identified by PEP-SiteFinder, CABS-dock, and PepSite, respectively (**Table 3.1**). In terms of complex structure prediction, GalaxyPepDock generated structures with better than medium quality when measured by the CAPRI criterion (Lensink and Wodak 2013) for 27 of the 57 PeptiDB targets, compared with the 4 targets returned by PEP-SiteFinder and 0 targets returned by

CABS-dock. Also, GalaxyPepDock generated structures with better than acceptable quality for 37 of the 57 PeptiDB targets, compared with 9 targets returned by PEP-SiteFinder and 11 targets returned by CABS-dock (**Table 3.2**). These results showed that the performance of GalaxyPepDock is better than that of other ab-initio protein-peptide docking methods and template-based docking is very effective for many protein-peptide docking problems.

Table 3.1. Fraction of binding site residues correctly predicted by GalaxyPepDock, PEP-SiteFinder, CABS-dock, and PepSite on the 40 targets of the PeptiDB set that have available unbound protein structures and have ≤ 10 residue-long peptides.

PDB ID		Galaxy PepDock	PEP-Site Finder	CABS-dock	PepSite
Bound	Unbound				
1ER8_E:I	1OEW_A	0.969	0.813	0.719	0.313
1CKA_A:B	2DVJ_A	0.800	0.867	0.933	0.733
1AWR_C:I	2ALF_A	1.000	0.813	0.813	0.750
1CZY_C:E	1CZZ_C	1.000	0.000	0.318	0.000
1DDV_A:B	1I2H_A	0.900	0.500	0.900	0.000
1H6W_A:B	1OCY_A	0.742	0.742	0.613	0.677
1KL3_C:G	2RTM_A	1.000	0.647	0.706	0.000
1GYB_B:E	1GY7_B	0.125	0.250	0.250	0.250
1LVM_A:E	1LVB_B	0.000	0.385	0.308	0.000
1MFG_A:B	2H3L_A	1.000	0.941	0.882	0.647
1N7F_B:D	1N7E_A	0.938	0.875	0.563	0.000
1OAI_A:B	1GO5_A	0.200	0.933	0.800	0.333
1NVR_A:B	2QHN_A	0.000	1.000	0.667	1.000
1OU8_B:D	1OU9_A	0.000	0.810	0.667	0.000
1UJ0_A:B	1X2Q_A	0.933	0.733	0.867	0.933
1T4F_M:P	1Z1M_A	0.824	0.647	0.765	0.353
1T7R_A:B	2AM9_A	0.938	0.875	0.813	0.000
1VZQ_H:I	1JWT_A	1.000	0.231	0.615	0.000
1TP5_A:B	1PDR_A	1.000	0.722	0.278	0.944
1W9E_A:T	1R6J_A	0.875	1.000	0.750	0.000
1YWO_A:P	1Y0M_A	1.000	1.000	0.923	1.000
1X2R_A:B	1X2J_A	0.818	0.727	0.909	0.773
2AK5_B:D	2G6F_X	1.000	0.833	0.750	1.000
2B1Z_B:D	3ERT_A	0.692	0.000	0.231	0.000
2C3I_B:A	2J2I_B	0.905	0.952	0.429	0.000
2FGR_A:B	2FGQ_X	0.900	0.300	0.300	0.000
2FOJ_A:B	2F1W_A	0.867	0.800	0.667	0.667
2FVJ_A:B	2HWQ_A	1.000	0.933	0.933	0.400
2H9M_C:D	2H14_A	0.900	0.800	0.650	0.800
2DS8_B:P	2DS7_A	0.538	0.077	0.538	0.385
2HO2_A:B	2E45_A	0.875	1.000	0.750	1.000
2HPL_A:B	2HPJ_A	0.000	0.929	0.667	0.500
2O9V_A:B	2O9S_A	1.000	0.750	0.833	0.833
2P1T_A:B	1LBD_A	0.588	0.647	0.471	0.118
2PUY_B:E	2YQL_A	1.000	0.889	0.556	0.500
2R7G_C:D	1AD6_A	0.895	0.789	0.895	0.211
2VJ0_A:P	1B9K_A	0.643	0.143	0.500	0.643
2ZJD_A:B	1V49_A	0.750	0.500	0.792	0.000
3D1E_A:P	3D1G_A	0.818	0.500	0.500	0.545
3D9T_B:D	1QBH_A	0.714	0.143	0.143	0.071
Average		0.754	0.662	0.641	0.409
Median		0.885	0.770	0.667	0.369

Table 3.2. Performance of GalaxyPepDock compared to other docking programs on 57 peptiDB targets.

	Galaxy PepDock	PEP-Site Finder	CABS-dock
Medium Quality	47.3%	7.0%	0.0%
Acceptable Quality	64.9%	15.8%	19.3%
<LRMSD>	7.5	11.0	9.2
<IRMSD>	3.4	4.7	4.2
$<F_{nat}>$	0.545	0.256	0.227
$<F_{site}>$	0.763	0.625	0.640

3.3.2. Template search of GalaxyPepDock

GalaxyPepDock searches protein-peptide templates based on Z-score summation of TM-score and interaction similarity score. We compared template search method to naïve method which only uses TM-score for template search. The average and median LRMSD of templates searched by GalaxyPepDock are 8.25 Å and 0.99 Å, those of templates searched by naïve method are 8.52 Å and 1.44 Å. The fraction of targets with less than 1.0 Å, less than 2.0 Å, and less than 4.0 Å of GalaxyPepDock are 50.9%, 59.6%, and 63.2%, those of naïve method are 42.1%, 56.1%, and 59.6%. These results showed that TM-score contributes the most to the performance of template searching and adding interaction similarity score can help search better templates.

Table 3.3. LRMSD of template selected by highest TM-score and Z-score summation of TM-score and interaction similarity score on 57 PeptiDB targets

ID (Query)	LRMSD (TM-score)	LRMSD (Z_{TM}+Z_{Inter})
1ER8_E:I	0.72	0.56
1CKA_A:B	4.49	1.64
1AWR_C:I	1.75	0.26
1SFI_A:I	0.11	0.25
1CZY_C:E	23.72	0.34
1DDV_A:B	8.35	8.35
1EG4_A:P	25.51	40.60
1JBU_H:X	17.34	25.76
1H6W_A:B	30.90	6.82
1KL3_C:G	0.56	0.56
1GYB_B:E	14.65	26.49
1LVM_A:E	32.17	29.47
1MFG_A:B	0.78	0.78
1N7F_B:D	0.94	0.94
1OAI_A:B	19.72	19.99
1NVR_A:B	29.00	29.00
1NX1_A:C	30.36	30.36
1OU8_B:D	23.55	26.20
1UJ0_A:B	0.96	5.74
1RXZ_A:B	0.34	0.34
1SSH_A:B	2.90	0.47
1T4F_M:P	0.94	0.94
1T7R_A:B	1.18	1.18
1VZQ_H:I	0.28	0.27
1TP5_A:B	0.29	0.29
1W9E_A:T	9.33	0.34
1YUC_A:C	0.74	0.74
1YWO_A:P	0.99	0.99
1X2R_A:B	21.43	7.29
2A3I_A:B	1.95	0.52
2AK5_B:D	1.24	5.47
2B1Z_B:D	25.62	25.62
2C3I_B:A	0.19	0.19
2B9H_A:C	34.37	2.39
2FGR_A:B	0.19	0.19
2FMF_A:B	0.70	0.70
2FOJ_A:B	3.80	3.80
2CCH_D:F	0.49	0.41
2FVJ_A:B	0.26	0.45
2H9M_C:D	0.17	0.17
2DS8_B:P	13.99	13.99
2HO2_A:B	8.35	5.54
2HPL_A:B	22.88	66.81
2O02_A:P	8.13	0.39

2O4J_A:C	0.17	0.17
2O9V_A:B	1.13	1.13
2P1K_A:C	0.59	0.59
2P1T_A:B	0.37	0.12
2P54_A:B	0.29	0.48
2PUY_B:E	1.39	1.39
2QOS_C:A	4.32	4.32
2R7G_C:D	1.53	5.20
2VJ0_A:P	39.69	52.35
3BU3_A:B	1.44	1.73
2ZJD_A:B	7.98	7.98
3D1E_A:P	0.31	0.31
3D9T_B:D	0.27	0.66
Average	8.52	8.25
Median	1.44	0.99
Ratio (LRMSD<1.0Å)	42.1%	50.9%
Ratio (LRMSD<2.0Å)	56.1%	59.6%
Ratio (LRMSD<4.0Å)	59.6%	63.2%

3.3.3. Energy-based optimization of GalaxyPepDock

Flexible-structure energy-based model-building procedure of GalaxyPepDock improved the predictions beyond that of a simple method that superimpose the target onto the template structure. The improvement in prediction accuracy achieved by additional energy optimization compared with the template superimposition method can be observed from the increased number of high-accuracy/medium-accuracy/acceptable predictions from 5/22/36 to 6/27/37 and the improved average ligand-RMSD/interface-RMSD/(fraction of native contact) values from 8.6 Å/4.0 Å/0.485 to 7.6 Å/3.4 Å/0.545. These results showed that molecular dynamics-based optimization method with physicochemical energy functions can generated more accurate protein-peptide models compared to superimposition method (**Table 3.4**).

Table 3.4. Similarity of the query and the template protein structures measured by TM-score and ligand RMSD of the starting model and final model on the 57 PeptiDB targets.

ID (Query)	ID (Template)	TM-Score	Initial RMSD	Final RMSD
1ER8_E:I	3APR_E:I	0.917	1.57	0.84
1CKA_A:B	1PRM_C:A	0.750	1.94	2.80
1AWR_C:I	1FGL_A:B	0.989	1.44	1.37
1SFI_A:I	2BTC_E:I	0.996	2.13	2.81
1CZY_C:E	1QSC_A:D	0.926	3.28	1.04
1DDV_A:B	1QC6_A:C	0.751	8.16	7.23
1EG4_A:P	1BT6_A:C	0.617	51.13	42.96
1JBU_H:X	8GCH_G:C	0.875	31.12	26.03
1H6W_A:B	1FCH_A:C	0.150	16.88	13.92
1KL3_C:G	1RST_B:P	0.911	2.14	3.69
1GYB_B:E	1KL5_A:E	0.434	28.16	27.02
1LVM_A:E	1FN8_A:B	0.634	25.20	26.44
1MFG_A:B	2PDZ_A:B	0.796	6.04	2.68
1N7F_B:D	1BE9_A:B	0.683	3.08	1.23
1OAI_A:B	1H27_B:E	0.176	26.97	23.36
1NVR_A:B	1QMZ_A:E	0.737	29.04	28.90
1NX1_A:C	1NPQ_A:B	0.621	19.70	19.25
1OU8_B:D	3SEM_A:C	0.539	26.01	25.47
1UJ0_A:B	1OEB_B:C	0.850	0.88	1.00
1RXZ_A:B	1ISQ_A:B	0.907	2.99	1.71
1SSH_A:B	3GBQ_A:B	0.823	2.16	1.38
1T4F_M:P	1YCR_A:B	0.801	1.14	1.30
1T7R_A:B	1T5Z_A:B	0.992	1.18	1.13
1VZQ_H:I	1GHW_H:I	0.995	0.53	1.37
1TP5_A:B	1BE9_A:B	0.792	1.29	0.97
1W9E_A:T	1OBY_A:P	0.963	0.41	0.79
1YUC_A:C	1YOW_A:B	0.947	2.37	2.03
1YWO_A:P	1SSH_A:B	0.811	3.75	3.54
1X2R_A:B	1P22_A:C	0.493	11.27	8.18
2A3I_A:B	1KV6_A:C	0.926	2.98	3.91
2AK5_B:D	2SEM_A:C	0.856	1.49	1.19
2B1Z_B:D	1X7E_A:C	0.894	0.40	4.78
2C3I_B:A	2BZK_B:A	0.970	1.00	0.80
2B9H_A:C	1UKH_A:B	0.854	5.56	4.79
2FGR_A:B	1E54_A:B	0.990	0.99	1.49
2FMF_A:B	2FLW_A:B	0.968	0.33	1.31
2FOJ_A:B	1YY6_A:B	0.949	4.74	4.54
2CCH_D:F	1OKW_B:E	0.989	2.86	1.46
2FVJ_A:B	1ZGY_A:B	0.736	0.90	0.96
2H9M_C:D	2G9A_A:B	0.967	0.63	0.94
2DS8_B:P	2FSA_A:P	0.186	14.89	14.16
2HO2_A:B	1K9Q_A:B	0.565	14.05	16.73
2HPL_A:B	2AKA_A:L	0.095	69.46	41.79

2O02_A:P	2C23_A:P	0.659	7.85	2.06
2O4J_A:C	1RKG_A:C	0.817	0.67	0.91
2O9V_A:B	3GBQ_A:B	0.868	1.97	1.16
2P1K_A:C	1CMI_A:C	0.814	2.11	3.31
2P1T_A:B	1XIU_A:E	0.741	0.76	1.32
2P54_A:B	1K7L_A:B	0.754	1.78	1.98
2PUY_B:E	2G6Q_A:B	0.633	4.48	3.08
2QOS_C:A	1VWR_B:P	0.587	6.52	6.20
2R7G_C:D	1N4M_A:C	0.497	3.64	3.49
2VJ0_A:P	1KY6_A:P	0.949	11.72	10.88
3BU3_A:B	2Z8C_A:B	0.371	3.54	4.44
2ZJD_A:B	2ASQ_A:B	0.735	9.68	8.32
3D1E_A:P	1OK7_B:C	0.989	1.79	1.58
3D9T_B:D	1XB1_A:G	0.604	1.70	3.70
Average			8.60	7.57
Median			2.98	2.81

3.3.4. Performance of GalaxyPepDock on CAPRI target

GalaxyPepDock was also tested on the CAPRI target 67 (PDB ID: 4N7H), and a medium-accuracy prediction was made. Compared with template-superimposed models, the quality of the model was improved by energy optimization from acceptable to medium accuracy, with improvements in ligand-RMSD/interface-RMSD/(fraction of native contact) values from 2.9 Å/1.5 Å/0.500 to 1.8 Å/1.0 Å/0.688. Also, GalaxyPepDock predicted hydrophobic interaction of Leucine and polar interaction of Tryptophan and Histidine. In the CAPRI blind prediction experiment, 6 of the 44 registered groups submitted medium-accuracy models. The best server predictions were only of acceptable quality (**Table 3.5**; **Figure 3.5**).

Table 3.5. Prediction made by GalaxyPepDock on the CAPRI target 67 compared with those submitted by top 3 servers and top 6 human groups in the CAPRI blind prediction experiment.

	LRMSD	IRMSD	f_{nat}	Quality[1)]
GalaxyPepDock	1.80	1.01	0.688	**
Server Predictors				
SwarmDocK	2.92	1.37	0.625	*
HADDOCK	3.18	1.94	0.500	*
ClusPro	4.18	1.49	0.688	*
Human Predictors				
Bates	1.12	0.80	0.688	**
Furman	1.27	0.93	0.938	**
Zhou	1.40	1.11	0.688	**
Niv	1.43	0.99	0.688	**
Zacharias	1.62	0.80	0.875	**
Vajda	1.69	1.23	1.000	**

1) Model quality defined as CAPRI criterion (Medium quality (**), Acceptable quality (*)).

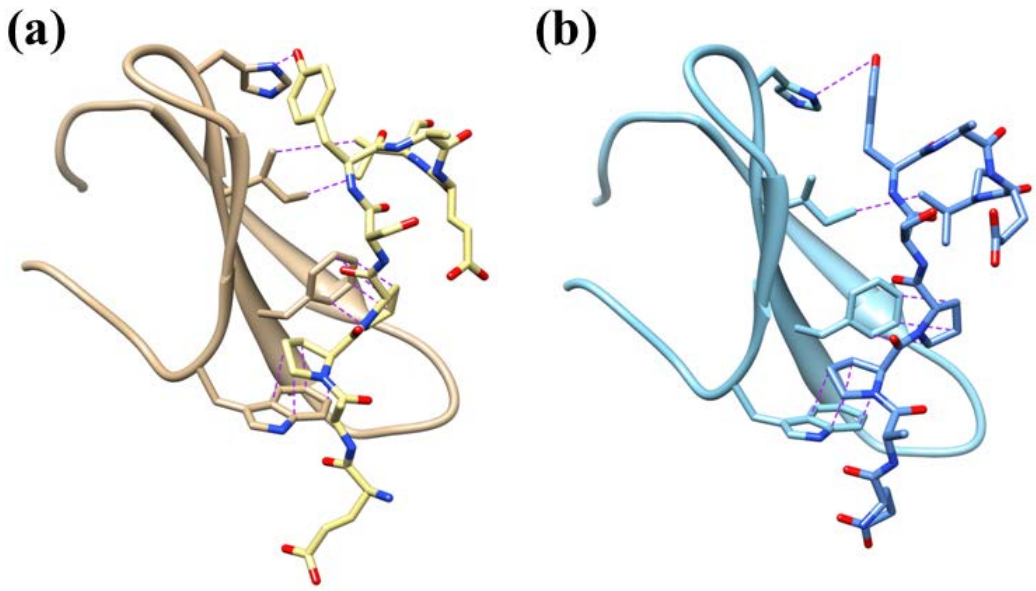


Figure 3.5. (a) Native structure and (b) GalaxyPepDock model on CAPRI target 67.

3.3.5. Limits of template-based docking

GalaxyPepDock is a template-based protein-peptide docking, so it means that the performace of GalaxyPepDock is influenced by the quality of template structure (**Table 3.6**). The success ratio of GalaxyPepDock was 64.9%, but the value was increased on targets having high structural similar templates. GalaxyPepDock failed to predict protein-peptide interactions on targets having low similar templates (TM-score < 0.6). These results showed that template-based protein-peptide docking is only effective on targets having high similar templates, and it is need to develop *ab initio* docking which performs well on targets having low similarity templates.

Table 3.6. Performance of GalaxyPepDock dependent on template quality

TM-score cut	Number of Success targets	Number of targets	Success ratio
TM-score > 0.0	37	57	64.9%
TM-score > 0.6	37	51	72.5%
TM-score > 0.7	35	44	79.5%
TM-score > 0.8	33	40	82.5%
TM-score > 0.9	25	29	86.2%

3.4. Conclusions

GalaxyPepDock is a similarity-based protein-peptide docking program that performs additional flexible-structure energy-based optimization. The effective combination of database search and physics-based optimization allows for a superior performance compared with the existing protein-peptide docking methods when complexes involving similar proteins can be found in the database.

4. GalaxyPPDock: a protein-protein docking program based on cluster-guided conformational space annealing

4.1. Introduction

Proteins play key roles in various biological processes, such as enzyme catalysis (Negri *et al.,* 2010) and signal transduction (Pawson and Nash 2000), through interactions with other proteins (Ozbabacan *et al.,* 2011; Perkins *et al.,* 2010). In order to understand protein functions, it is essential to precisely describe protein-protein interactions in atomic detail, which is the ultimate goal of protein-protein docking studies. For decades, many protein-protein docking programs have been developed to deliver atomic models of protein-protein interactions with various types of sampling approaches. There are many FFT-based docking program, including FTDock (Gabb *et al.,* 1997), ZDOCK (Chen *et al.,* 2003), PIPER (Kozakov *et al.,* 2006), DOT (Mandell *et al.,* 2001), and GRAMM (Vakser 1997). There are also methods using geometric hash, PatchDock (Schneidman-Duhovny *et al.,* 2005) and LZerD (Venkatraman *et al.,* 2009), Monte Carlo simulation, RosettaDock (Gray *et al.,* 2003), FireDock (Andrusier *et al.,* 2007), and FiberDock (Mashiach *et al.,* 2010), and molecular dynamics simulation, HADDOCK (Dominguez *et al.,* 2003). Despite their efforts, however, conformation sampling still remains as one of the most challenging problems in protein-protein docking study (Gray 2006; Huang 2014). Even with such diverse sampling approaches attempted to date, still searching conformation space in protein-protein docking problem - spanned by relative orientation and internal flexibility of the interacting

partners - is formidable (Bonvin 2006). A powerful global optimization method can therefore be indispensable to tackle this challenge.

Conformational Space Annealing (CSA) (Lee *et al.,* 1998) is regarded as one of the powerful global optimization methods that have been applied to general biological modeling studies. The key idea of CSA is to run a broad sampling in early stage and then to gradually focus on low-energy conformations. Sampling space is diverse in early stage and becomes gradually narrowed down. CSA has been successfully applied to many biological problems, such as protein structure prediction (Joo *et al.,* 2009; Ko *et al.,* 2012; Park *et al.,* 2011; Park and Seok 2012; Park *et al.,* 2014) and protein-ligand docking (Lee *et al.,* 2005; Shin *et al.,* 2011; Shin and Seok 2012; Shin *et al.,* 2013). Previously, Lee *et al* applied CSA to protein-protein docking study (Lee *et al.,* 2005) which was tested on round 5 of Critical Assessment of Prediction of Interactions (CAPRI), a community-wide experiment for evaluating the performance of protein-protein docking programs. However, at the moment the method was premature and only one of four targets have got acceptable result in CAPRI criteria. This suggests that applying CSA algorithm to protein-protein docking problem is not straightforward, but requires additional developments in order to take into account of specific features that the problem may possess.

Then what is particular aspect of protein-protein docking problem by understanding which we can systematically enhance the sampling performance? The main idea we took advantage of in this study is that conformational space in protein-protein docking problem can be dramatically reduced into a set of smaller sub-spaces with highest feasibilities. Feasibility of a model complex is strongly related to geometric or electrostatic complementarity between proteins unless

either protein undergoes huge conformational change upon binding. Moreover, those feasible spaces are not uniformly distributed but are found as discrete "patches" in whole space (Caffrey *et al.,* 2004; Jones and Thornton 1997; Malod-Dognin *et al.,* 2012). Therefore, based on this assumption, we can be reformulated the problem as more tractable one: to run global optimization on a limited conformational space described above.

In this work, we developed a protein-protein docking program named GalaxyPPDock using cluster-guided CSA (CG-CSA) sampling method for protein-protein docking. CG-CSA makes clusters from initially sampled structures and evolves them each cycle. Instead of annealing whole conformational space as in regular CSA, CG-CSA more focuses on annealing conformation space of each cluster. During the evolving step, these clusters communicate each other and changes number of members to gradually more concentrates on low-energy clusters. This idea makes high-energy clusters to survive and enables to search on multiple local minima efficiently at the same time. If energy function is relatively accurate, focusing on low-energy clusters can generate near-native predicted models. If energy function is inaccurate and global minimum is far from near-native still local minimum is close from near-native, high-energy clusters can find near-native structures. Accordingly, GalaxyPPDock can tolerate incorrectness of energy function to deliver correct solution as one of the clusters. Therefore, CG-CSA implemented in GalaxyPPDock can generate near-native protein complex models in cases that both energy functions is relatively accurate and energy function is relatively inaccurate.

4.2. Methods

4.2.1. Overall procedure of GalaxyPPDock

GalaxyPPDock consists of two steps for protein-protein docking. The first step is initial docking for find putative binding sites. In the initial docking step, rigid-body docking performed using ZDOCK. Then, complexes generated by ZDOCK are rescored by Z-score summation of ZDOCK score (Mintseris *et al.,* 2007), DFIRE score (Zhou and Zhou 2002), and electrostatic potential (MacKerell *et al.,* 1998). Then, 50 complexes are selected by clustering method by NMRCLUST (Kelley *et al.,* 1996) and are used to initial bank for next step. The second step is global optimization step for generate more accurate protein complex structure. In the second step, GalaxyPPDock uses CG-CSA sampling method for protein-protein docking. CG-CSA makes clusters from initially sampled structures and evolves them each cycle to find global minimum of energy land scape of protein-protein interaction. The energy used in GalaxyPPDock is hybrid energy of physics-based energy function and knowledge-based scoring function. After global optimization, 10 protein complex models are selected by their energy value and clustering method (**Figure 4.1**).

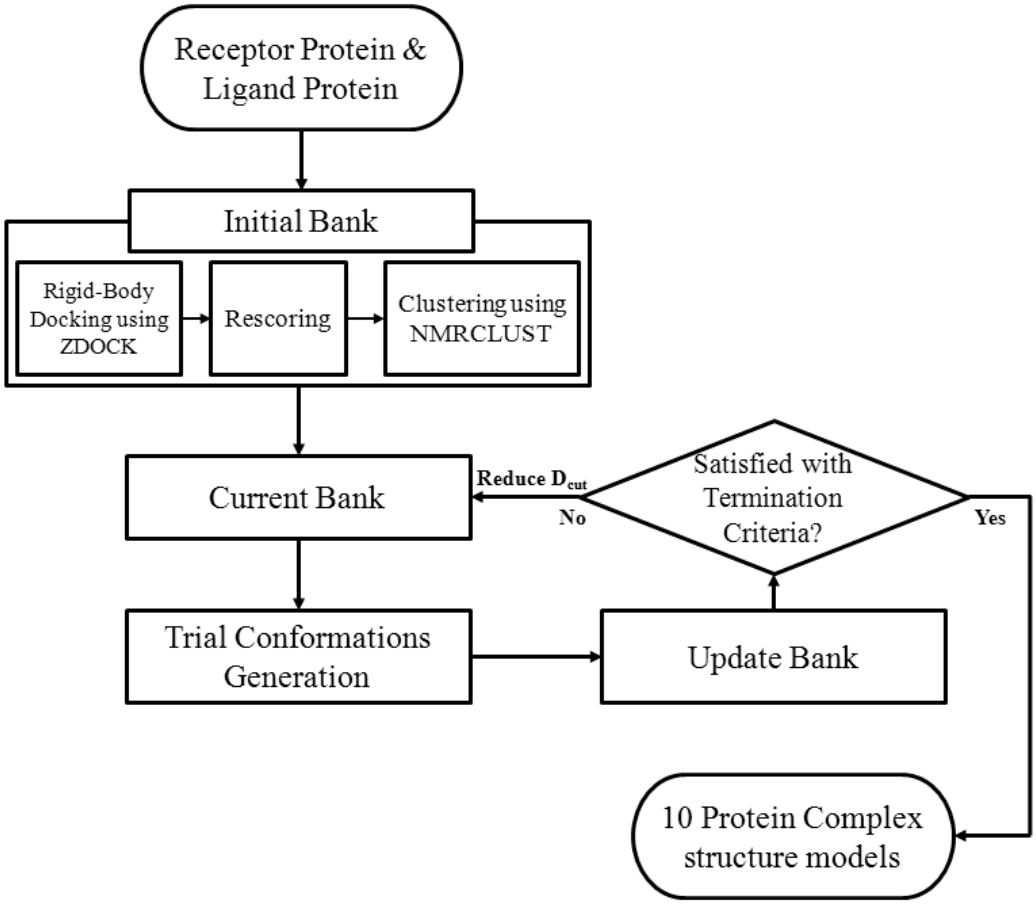


Figure 4.1. Flowchart of GalaxyPPDock

4.2.2. Sets of protein complexes used for method development

A set of 121 unbound/unbound complexes (rigid-body targets, classified by ZDOCK criterion) from ZDOCK benchmark set 4.0 (Hwang *et al.,* 2010) and 20 complexes (unbound/unbound and unbound/bound targets) from CAPRI round 1~19 (Janin *et al.,* 2003; Janin 2005; Janin 2007; Janin 2010) was used as a benchmark set to evaluate performance of GalaxyPPDock. Total 141 complexes were randomly divided into a training set of 35 complexes and a benchmark test set of 106 complexes. Conformational decoy sets for the 35 training set complexes generated by RosettaDock (500 decoy conformations for each complex) and another set of 80 complexes (Su *et al.,* 2009) with known structures and binding affinities were used to train energy parameters. The test set of 106 complexes was used to validate the performance of GalaxyPPDock by comparing with ZDOCK (Mintseris *et al.,* 2007), RosettaDock (Gray *et al.,* 2003), FireDock (Andrusier *et al.,* 2007), and FiberDock (Mashiach *et al.,* 2010). GalaxyPPDock was also compared with the previous CSA method by Lee *et al.* on four CAPRI targets (Lee *et al.,* 2005) and with other CAPRI predictors on 7 targets from the latest CAPRI rounds 22~27 (Janin 2013).

4.2.3. Training of energy parameters

GalaxyPPDock employs a hybrid energy function that combines physics-based energy E_{physics} and knowledge-based energy $E_{\text{knowledge}}$ as follows:

$$E_{\text{GalaxyPPDock}} = E_{\text{physics}} + E_{\text{knowledge}} \quad (4.1)$$

$$E_{\text{physics}} = w_{\text{LJ}}E_{\text{LJ}} + w_{\text{Coul}}E_{\text{Coul}} + w_{\text{SA}}E_{\text{SA}} \quad (4.2)$$

$$E_{\text{knowledge}} = w_{\text{DFIRE}}E_{\text{DFIRE}} + w_{\text{Hbond}}E_{\text{Hbond}} + w_{\text{cons}}E_{\text{cons}} + w_{\text{rot}}E_{\text{rot}} \quad (4.3)$$

where E_{LJ} and E_{Coul} are the Lennard-Jones energy and the Coulomb electrostatic potential energy, respectively, with the CHARMM22 force field parameters (MacKerell *et al.,* 1998), E_{SA} is the implicit solvation free energy described by solvent-accessible surface area with atomic solvation parameters (Zhou and Zhou 2002), E_{DFIRE} is the distance-dependent statistical pair potential DFIRE (Zhou and Zhou 2002), E_{Hbond} is the knowledge-based orientation-dependent hydrogen bond energy (Kortemme *et al.,* 2003), E_{cons} is the sequence conservation propensity score derived from the PSI-BLAST profile (Liang *et al.,* 2009), and E_{rot} is the statistical side chain rotamer energy derived from the backbone-dependent rotamer library (Eswar *et al.,* 2006). The energy parameters (w_{LJ}, w_{coul}, w_{SA}, w_{DFIRE}, w_{Hbond}, w_{cons}, w_{rot}) = (1.0, 0.15, 4.5, 8.0, 6.0, 3.0, 3.0) were determined as explained below.

The six out of seven energy weight parameters (w_{LJ}, w_{coul}, w_{SA}, w_{DFIRE}, w_{Hbond}, w_{cons}) were determined first, and the rotamer energy was added afterwards during our participation in the CAPRI experiments (after round 20) to improve the accuracy of local side chain structures. The six weights were searched for on parameter grids to maximize the product of (i) the Pearson correlation coefficient between the experimental binding free energy and the GalaxyPPDock energy for the binding affinity set of 80 complexes (Su *et al.,* 2009), (ii) the Pearson correlation coefficient for the energy-RMSD distribution of the 500 decoys averaged over the 35 training set complexes, and (iii) the absolute value of the Z-score of the average energy of the 20 decoy conformations closest to the experimental structure in the energy-RMSD distribution of the 500 decoys averaged over the 35 training set complexes (**Table 4.2**). 500 decoys were generated by RosettaDock starting from initial complex which unbound tertiary

structure superpose onto native complex. Fixing the six weights, the rotamer energy weight w_{rot} was finally determined to improve local side-chain accuracy of CG-CSA.

Table 4.1. Weight factors of GalaxyPPDock energy function

	E_{LJ}	E_{DFIRE}	E_{cons}	E_{SA}	E_{elec}	E_{Hbond}	$E_{LJ}+E_{DFIRE}+E_{cons}+E_{SA}$	$E_{GalaxyPPDock}$
BA set	0.663	0.685	0.526	0.606	0.003	0.104	0.726	0.724
Rosetta set	0.551 (-1.422)	0.610 (-1.371)	0.504 (-1.005)	0.486 (-1.312)	0.232 (-0.768)	0.193 (-0.626)	0.603 (-1.369)	0.603 (-1.381)

The values of the first row are Pearson correlation between RMSD and each energy component on 80 targets of Binding affinity set. The values out of bracket in the second row are Pearson correlation between RMSD and each energy component of 500 conformations generated by RosettaDock on training set. The values out of bracket in the second row are Z-score of the average energy of the 20 conformations closest to the native structures from 500 conformations generated by RosettaDock on training set.

4.2.4. Overview of the conformational space annealing

It is worthwhile to briefly go through the overall procedure of the general conformational space annealing (CSA) global optimization algorithm first before we describe the cluster-guided conformational space annealing (CG-CSA) algorithm in in detail the next subsection. Performance of the regular CSA (R-CSA) method is also compared with the CG-CSA method in a benchmark test.

In CSA, a fixed number of local minimum conformations called "bank" is evolved by gradually focusing on low-energy regions in the conformational space. Each bank member can be roughly considered as a representative low-energy conformation covering a conformational hyper-space of radius D_{cut}, where D_{cut} is a parameter used to control broadness of conformational search. Initial bank is desired to be composed of diverse conformations and may often be generated by random sampling. At each CSA step, new trial conformations are generated by crossovers and mutations of bank conformations, and the bank is updated by comparing each trial conformation with current bank members. If a trial conformation is $< D_{cut}$ from any bank conformation, it replaces the bank conformation if it has lower energy and is discarded otherwise. If a trial conformation is $> D_{cut}$ from all bank conformations, it replaces the highest-energy bank conformation. If D_{cut} is large, low-energy trial conformations tend to replace close-by bank conformations, leaving high-energy conformations at large distances. If D_{cut} is small, they tend to replace high-energy bank conformations leaving low-energy conformations at relatively close distances. By starting with a large value of D_{cut}, diverse high energy regions are allowed to be explored at the early stage, and low energy regions are searched more heavily as CSA iteration proceeds with gradually decreasing D_{cut}. The CSA iteration is considered converged if all bank

conformations have been used as seeds and are not further replaced by new conformations.

For CSA, a distance measure for comparing conformations thus has to be defined. In the current work, the distance between two docking conformations i and j is defined as

$$D_{ij} = \left|\mathbf{T}_i - \mathbf{T}_j\right| + w_{\mathrm{rot}}\left|\mathbf{R}_i - \mathbf{R}_j\right| \quad (4.4)$$

Where **T** is the values for the three translational degrees of freedom expressed as the center of C_α coordinates of the ligand protein (the smaller protein) when the center of the receptor protein (the larger one) is fixed at the origin, **R** is the values for the three rotational degrees of freedom expressed as the rotational angles of the current ligand pose relative to the reference pose about the x-, y-, and z-axis, and the weight w_{rot} is defined as the ratio between the average translational distance to the average rotational distance for the initial bank conformations $\langle\left|\mathbf{T}_i - \mathbf{T}_j\right|\rangle \;/\; \langle\left|\mathbf{R}_i - \mathbf{R}_j\right|\rangle$.

4.2.5. Cluster-guided conformational space annealing

In the current CG-CSA, clusters are defined from the initial bank generation stage. 200 complex conformations are selected from the 3,600 complexes generated by ZDOCK based on the Z-score summation of ZDOCK score, DFIRE potential, and Coulomb potential and are clustered by NMRCLUST (Kelley *et al.,* 1996), and 50 initial bank conformations are chosen by picking conformations from each cluster in proportion to the cluster size. In this work, the number of clusters ranged from 2 to 10.

At each CSA iteration step, 200 trial conformations are generated from 20 "seed" conformations selected from the clusters proportional to the cluster sizes. Seeds are selected to have large mutual distances to produce diverse conformations. For each seed conformation, (i) 5 trial conformations are generated by cross-over of **T** and **R** of the seed with those of 5 randomly selected partner conformations, (ii) 3 trial conformations by perturbation of **T** or **R** 3 times, and (iii) 3 trial conformations by cross-over of interface side-chain χ angles of the seed with 2 randomly selected partners. Partners are selected randomly from the current bank independent of cluster for generation of diverse low-energy conformations. After (i) and (ii), side-chain conformations are adjusted by removing clashes in the rotamer space (Dunbrack and Cohen 1997). All trial conformations are then energy minimized by gradient-based local minimization (Fuhrmann *et al.,* 2009) in the space of rigid-body translation/rotation and flexible interface side-chain χ angles. Rigid-body rotation is described by exponential mapping of quaternion (Fuhrmann *et al.,* 2009). Flexible interface residues are selected from the most common interface residues of the initial bank conformations (receptor and ligand residues with C_α distances < 10 Å), and the number of flexible residues is set to the average number of the interfaces residues in the initial bank.

With the new trial conformations generated as described above, the CG-CSA bank is updated within cluster (intra-cluster update) at each iteration, and inter-cluster update is allowed at every other iteration. Each trial conformation is assigned to the cluster that the closest bank member belongs to. In the intra-cluster update, the same update rule of general CSA is applied within each cluster, i.e., the closest bank conformation and the highest-energy conformation are selected within the cluster. In the inter-cluster update, a trial conformation that do not replace any bank conformation in the same cluster gets a chance to be compared with the

highest-energy conformation in other clusters. If the trial conformation has lower energy, it becomes a new bank member, increasing the size of the cluster by one, and decreasing the size of the other cluster. Changes in cluster sizes are limited to the maximum of 1 change at each iteration for slow change. In this way, the size of the low-energy cluster can become larger as CG-CSA proceeds except that the sizes of clusters > 20 or < 5 members are not allowed to change to keep sub-optimal clusters.

Finally, 10 structures are selected by clustering the structures of CSA final bank using greedy algorithm with ligand RMSD 5Å cutoff. The cluster having lowest energy complex structure is selected at first, then, other nine clusters are selected by cluster size. Finally, the lowest energy representatives from each cluster are selected.

4.2.6. Assessment measure

To evaluate the performance of GalaxyPPDock, three measures were used: LRMSD (peptide RMSD), IRMSD (interface RMSD), and f_{nat} (fraction of native contact). For the definitions of acceptable/medium/high accuracy predictions, the following CAPRI criterion was used: acceptable accuracy if (LRMSD < 10 Å or IRMSD < 4 Å) and $f_{nat} > 0.1$, medium accuracy if (LRMSD < 5 Å or IRMSD < 2 Å) and $f_{nat} > 0.3$, and high accuracy if (LRMSD < 1 Å or IRMSD < 1 Å) and $f_{nat} > 0.5$ (Lensink and Wodak 2013).

4.3. Results and Discussion

4.3.1. Performance of cluster-guided conformational space annealing

We first compare performances of CG-CSA with those of R-CSA on the 35 targets of training set and 106 targets of test set. Performances are compared in terms of the percentage of targets for which at least one docking conformation out of top 10 conformations are predicted with better than acceptable (or medium) quality.

According to **Table 4.2**, CG-CSA generated models with better than acceptable quality for 42.9%, compared to 25.7% for R-CSA on the training set. The average (LRMSD/IRMSD/f_{nat}) of CG-CSA is (18.0/6.6/0.30) and it is better than that of R-CSA (23.0/8.5/0.26) and initial bank (19.4/7.9/0.26). According to **Table 4.3**, CG-CSA generated models with better than acceptable quality for 43.4%, compared to 38.7% for R-CSA, and generated models with better than medium quality for 27.4% on the test set. The average (LRMSD/IRMSD/f_{nat}) of CG-CSA is (18.0 Å /6.6 Å /0.30) and it is better than that of R-CSA (23.0 Å /8.5 Å /0.26) and initial bank (19.4 Å /7.9 Å /0.26). These results showed that CG-CSA improved model quality from models of initial bank and improvement of CG-CSA is better than that of R-CSA. In **Figure 4.2** energy landscapes are shown for four representative examples for which CG-CSA was able to bring better predictions than R-CSA. For two targets, 1ay7 (**Figure 4.2(a)**) and 1r0r (**Figure 4.2(b)**), when the energy function relative accurate and low-energy structure is near-native structure, the lowest LRMSDs of 10 output complexes are 3.5 Å and 5.0 Å by CG-CSA, compared to 7.9 Å and 13.8 Å by R-CSA. CG-CSA showed better performance when energy function is relative accurate and low-energy structure is nearby native structure. Because region of low-energy cluster called main-optimal

cluster is nearby native structure and CG-CSA more focuses on main-optimal cluster increasing the number of main-cluster members, RMSD between native structure and predicted structure generated by CG-CSA is smaller than RMSD between native structure and predicted structure generated by R-CSA.

For the opposite cases when energy function is relatively inaccurate, CG-CSA also shows improved performances. For two targets, 1iqd (**Figure 4.2 (c)**) and 1r0r (**Figure 4.2 (d)**), when energy function is relatively inaccurate and global minimum is far from near-native structure, the lowest LRMSDs of 10 output complexes are 2.1 Å and 7.1 Å by CG-CSA, compared to 48.2 Å and 20.9 Å by R-CSA. R-CSA showed problems of converging into these false global minima. However, such a converged structural pool may not be the optimal as long as the correctness of energy function is not guaranteed. Instead of focusing on a single global minimum, CG-CSA also focuses on multiple sub-optimal conformational spaces at very distinct translational/rotational positions from global minimum. One may expect that even with incorrect energy function near-native conformation can be at one of the sub-optimal clusters.

We also compared CG-CSA to the first application of CSA to protein-protein docking by Lee *et al.* (**Table 4.4**). We call this previous approach as "CSA-Lee" here. The comparison is done on 4 targets in CAPRI round 5 for which "CSA-Lee" was tested. "CSA-Lee" succeeded to bring acceptable quality on only one target (target 15). In contrast, R-CSA predicted three targets to acceptable or better quality, and CG-CSA did four targets to acceptable or better. We also notice that the test set is not enough to derive statistically meaningful statement among different methods, as well as there can be other factors contributing to the difference such as energy function. However, these results showed that the

performance of CG-CSA is enough good compared to R-CSA and “CSA-Lee”.

Table 4.2. Ligand-RMSD (LRMSD), interface-RMSD (IRMSD), and fraction of native contact (f_{nat}) of initial bank results and final bank results of regular CSA (R-CSA) and cluster-guided CSA (CG-CSA) on 35 training set targets

Target	Initial Bank			Final Bank (R-CSA)			Final Bank (CG-CSA)		
	LRMSD	IRMSD	f_{nat}	LRMSD	IRMSD	f_{nat}	LRMSD	IRMSD	f_{nat}
1avx	6.7	1.6	0.71	7.1	1.6	0.74	10.2	3.8	0.51
1buh	23.5	13.8	0.04	32.3	13.4	0.00	13.8	3.5	0.30
1clv	4.6	2.2	0.35	3.7	1.8	0.43	3.6	1.8	0.50
1eaw	9.9	3.2	0.39	9.7	5.1	0.04	2.1	0.7	0.90
1fc2	32.2	11.6	0.00	28.5	14.0	0.00	28.2	14.1	0.00
1ghq	58.3	18.4	0.00	57.9	15.1	0.00	56.3	13.5	0.00
1gxd	38.1	12.5	0.02	39.8	12.7	0.02	39.7	12.7	0.02
1h9d	20.9	10.4	0.05	10.9	3.3	0.35	13.9	5.6	0.08
1j2j	9.2	3.3	0.33	5.0	2.0	0.64	6.6	3.1	0.58
1jps	39.3	18.6	0.00	40.6	17.7	0.00	32.6	14.4	0.00
1jwh	29.2	16.0	0.00	9.9	2.3	0.56	12.3	2.6	0.47
1k4c	31.6	9.3	0.00	63.8	18.2	0.00	35.2	10.6	0.00
1kxq	0.5	0.8	0.90	15.3	5.8	0.18	15.2	5.8	0.18
1mah	14.0	8.0	0.12	1.7	0.7	0.76	1.3	0.8	0.79
1mlc	20.4	11.0	0.00	52.2	20.5	0.00	16.5	8.7	0.00
1oc0	15.6	7.7	0.17	14.6	6.7	0.02	14.7	6.7	0.02
1oph	62.0	14.3	0.00	64.0	14.3	0.00	63.5	16.5	0.00
1s1q	26.8	9.4	0.05	26.8	9.4	0.05	26.9	9.4	0.05
1t6b	17.1	8.8	0.08	65.2	22.0	0.00	16.5	9.7	0.00
1us7	23.1	11.4	0.00	17.4	10.0	0.00	17.7	10.2	0.00
2ayo	3.3	2.0	0.41	3.5	1.9	0.58	3.5	1.9	0.58
2b4j	13.9	7.7	0.25	20.8	10.1	0.00	20.2	10.2	0.00
2o8v	29.3	14.7	0.00	25.6	10.1	0.00	18.2	8.9	0.14
2sni	16.0	7.5	0.01	9.8	2.5	0.56	9.5	2.4	0.58
2vdb	1.8	0.9	0.87	38.9	16.9	0.00	37.9	12.9	0.00
4cpa	3.0	1.2	0.74	5.9	2.7	0.40	5.6	2.3	0.45
9qfw	37.5	9.4	0.00	36.6	11.0	0.00	30.7	10.7	0.00
TA01	12.7	6.6	0.04	12.3	6.5	0.12	11.9	6.5	0.16
TA06	0.8	0.5	0.86	16.2	9.7	0.10	0.8	0.5	0.86
TA07	42.4	16.0	0.00	47.6	20.5	0.00	39.7	12.3	0.00
TA12	0.5	0.4	0.91	1.3	0.5	0.93	1.1	0.5	0.91
TA15	11.3	5.5	0.00	3.6	1.3	0.80	2.2	1.0	0.77
TA25	2.3	1.1	0.83	3.6	1.5	0.77	3.8	1.5	0.75
TA26	19.9	10.3	0.00	11.7	5.6	0.03	16.8	5.5	0.24
TA40	1.4	0.5	0.86	1.7	0.5	0.84	1.7	0.5	0.84
Average	19.4	7.9	0.26	23.0	8.5	0.26	18.0	6.6	0.30

Table 4.3. Ligand-RMSD (LRMSD), interface-RMSD (IRMSD), and fraction of native contact (f_{nat}) of initial bank results and final bank results of regular CSA (R-CSA) and cluster-guided CSA (CG-CSA) on 106 test set targets

Target	Initial Bank			Final Bank (R-CSA)			Final Bank (CG-CSA)		
	LRMSD	IRMSD	f_{nat}	LRMSD	IRMSD	f_{nat}	LRMSD	IRMSD	f_{nat}
1a2k	9.7	2.6	0.68	13.3	4.4	0.50	10.5	2.9	0.75
1ahw	27.6	9.3	0.00	32.3	15.5	0.00	33.4	10.5	0.00
1ak4	20.7	9.2	0.02	21.3	9.7	0.07	21.1	9.2	0.02
1akj	27.9	15.3	0.00	35.8	16.3	0.00	29.4	14.4	0.00
1ay7	11.2	3.5	0.40	7.9	1.8	0.62	3.5	1.5	0.80
1azs	63.3	10.4	0.00	43.1	16.9	0.00	61.3	14.9	0.00
1b6c	8.3	3.1	0.57	9.0	2.7	0.71	8.9	2.7	0.77
1bj1	22.6	11.4	0.00	20.9	10.2	0.00	7.1	1.0	0.86
1bvk	12.0	5.3	0.19	11.6	4.8	0.12	12.7	4.7	0.15
1bvn	2.8	1.4	0.70	2.5	1.4	0.55	1.9	1.1	0.68
1cgi	4.1	2.8	0.39	4.0	2.2	0.58	3.8	2.2	0.58
1d6r	11.6	5.1	0.03	18.3	7.8	0.02	18.3	7.8	0.02
1dfj	2.7	1.4	0.68	6.1	2.5	0.64	6.1	2.5	0.66
1dqj	20.5	11.5	0.00	11.3	5.8	0.27	19.0	11.4	0.00
1e6e	3.0	1.4	0.79	5.2	1.9	0.83	1.9	1.2	0.88
1e6j	12.7	4.9	0.16	12.7	4.9	0.16	12.9	5.7	0.10
1e96	30.5	6.6	0.05	28.9	12.0	0.00	25.3	13.2	0.00
1efn	32.4	9.5	0.00	28.9	10.8	0.00	27.6	8.4	0.03
1ewy	5.6	3.4	0.20	13.3	7.5	0.00	13.2	7.5	0.04
1ezu	37.8	21.4	0.00	37.9	21.2	0.00	37.9	17.1	0.00
1f34	43.3	16.1	0.00	40.9	16.3	0.06	33.1	18.9	0.03
1f51	3.3	1.7	0.55	4.1	2.3	0.63	4.1	2.3	0.63
1fcc	35.7	14.6	0.00	35.3	14.5	0.00	35.2	14.5	0.00
1ffw	9.3	5.1	0.42	7.4	3.6	0.50	8.6	3.3	0.50
1fle	22.0	10.0	0.01	22.5	10.1	0.01	22.3	10.0	0.01
1fqj	35.1	16.4	0.00	35.4	16.5	0.00	31.7	16.6	0.00
1fsk	2.4	0.9	0.91	2.1	0.9	0.89	2.2	0.9	0.86
1gcq	18.0	8.8	0.00	2.1	1.1	0.87	15.1	5.2	0.13
1gl1	2.9	1.5	0.69	7.0	3.0	0.70	2.6	1.4	0.56
1gla	52.9	20.7	0.00	52.0	19.5	0.00	52.3	19.5	0.00
1gpw	2.1	1.3	0.62	3.4	1.6	0.69	3.4	1.6	0.65
1hcf	22.5	8.0	0.07	24.2	10.0	0.07	24.4	10.0	0.04
1he1	7.0	3.8	0.19	2.9	1.7	0.76	2.8	1.7	0.73
1hia	9.6	4.1	0.11	8.6	3.5	0.22	9.9	4.6	0.13
1i4d	35.4	14.7	0.02	34.7	14.4	0.04	33.5	14.9	0.07
1i9r	9.9	4.9	0.09	12.9	9.0	0.00	12.8	9.0	0.00
1iqd	24.5	10.2	0.04	48.2	15.3	0.00	2.0	0.8	0.71
1jtg	3.6	1.1	0.63	6.0	2.3	0.40	6.2	2.5	0.41
1k74	3.5	1.3	0.73	6.7	1.9	0.54	6.1	2.0	0.63
1kac	33.9	12.5	0.00	32.1	9.9	0.00	36.8	12.6	0.00
1klu	41.5	13.5	0.00	33.3	11.6	0.00	33.4	11.7	0.00
1ktz	37.8	10.8	0.00	37.8	10.8	0.03	33.7	11.0	0.10
1kxp	6.3	1.9	0.44	7.0	1.9	0.50	3.3	1.6	0.53
1ml0	2.4	1.2	0.78	2.3	1.3	0.81	2.3	1.3	0.81
1n8o	9.8	1.0	0.78	10.6	1.4	0.71	10.0	1.1	0.82
1nca	26.5	18.3	0.00	25.7	17.0	0.00	26.4	17.9	0.00
1nsn	17.4	10.6	0.00	55.1	15.7	0.00	17.6	10.5	0.00
1ofu	15.7	6.6	0.00	36.5	18.9	0.00	26.0	15.7	0.00
1oyv	3.4	1.3	0.61	4.2	1.4	0.57	4.3	1.4	0.57
1ppe	0.9	0.6	0.85	3.3	1.2	0.80	2.8	1.0	0.86
1pvh	27.9	10.7	0.13	31.6	14.3	0.00	28.9	10.5	0.13
1qa9	46.2	16.5	0.00	46.4	16.5	0.00	46.5	16.7	0.00
1r0r	12.9	3.6	0.27	13.8	6.5	0.00	5.0	1.6	0.49

1rlb	12.2	2.3	0.63	18.0	9.0	0.00	18.1	9.1	0.00
1rv6	1.9	1.4	0.68	8.8	3.7	0.32	8.1	3.5	0.36
1sbb	54.9	14.1	0.00	56.6	13.9	0.00	55.2	14.6	0.00
1tmq	20.7	11.5	0.00	1.8	1.0	0.69	1.8	1.0	0.73
1udi	23.1	15.1	0.03	4.0	2.3	0.45	4.9	2.7	0.44
1vfb	21.3	7.5	0.00	8.0	4.1	0.19	7.3	3.7	0.23
1wdw	3.6	1.9	0.55	3.6	2.0	0.54	3.6	1.9	0.57
1wej	2.9	1.4	0.81	2.8	1.4	0.74	3.3	1.3	0.72
1xd3	6.9	4.0	0.23	7.4	2.8	0.35	7.4	2.4	0.40
1xu1	17.4	7.4	0.08	12.8	5.5	0.00	9.8	4.5	0.07
1yvb	5.3	1.6	0.52	9.4	1.7	0.74	9.9	1.8	0.76
1z0k	7.1	2.8	0.50	10.3	4.2	0.47	10.3	4.1	0.39
1z5y	26.7	10.5	0.02	29.1	10.1	0.04	16.7	5.0	0.36
1zhh	60.5	21.9	0.00	13.3	7.7	0.00	13.4	7.7	0.00
1zhi	34.1	14.4	0.00	32.3	13.6	0.00	36.0	8.7	0.02
2a5t	17.8	7.4	0.00	13.3	7.6	0.00	7.4	3.3	0.36
2a9k	37.7	15.8	0.00	23.3	11.3	0.00	32.5	12.9	0.00
2abz	16.3	7.3	0.00	14.5	7.4	0.03	11.5	5.9	0.07
2ajf	24.9	11.0	0.00	27.5	12.0	0.00	26.2	14.0	0.00
2b42	3.4	1.2	0.83	3.6	1.2	0.87	3.6	1.2	0.88
2btf	22.5	15.5	0.00	22.4	13.8	0.00	21.6	13.5	0.00
2fd6	13.0	3.5	0.28	13.1	3.8	0.26	13.0	3.9	0.23
2fju	83.1	0.6	0.00	82.5	0.6	0.00	82.3	0.6	0.00
2g77	17.0	7.6	0.07	16.0	9.5	0.00	22.2	11.3	0.00
2hle	14.3	4.4	0.26	4.2	2.3	0.44	4.2	2.2	0.43
2hqs	20.5	10.4	0.01	23.7	11.8	0.00	16.7	5.9	0.06
2i25	25.3	8.8	0.00	21.4	9.3	0.00	20.1	5.0	0.02
2j0t	20.3	9.4	0.03	21.4	8.0	0.02	16.8	5.6	0.05
2jel	5.8	1.4	0.68	16.1	9.3	0.00	8.1	3.2	0.29
2mta	12.7	4.2	0.26	8.3	3.5	0.28	13.2	7.4	0.00
2oob	29.6	7.6	0.04	29.1	7.6	0.11	29.1	7.6	0.11
2oor	16.7	7.0	0.22	21.6	12.9	0.07	22.9	14.5	0.04
2oul	2.0	0.8	0.83	3.7	1.0	0.81	3.7	1.1	0.82
2pcc	7.2	4.0	0.31	6.3	3.3	0.34	10.2	4.9	0.34
2sic	6.1	1.3	0.80	6.3	1.4	0.77	6.4	1.4	0.77
2uuy	17.7	7.0	0.00	16.3	6.7	0.00	16.0	6.7	0.00
2vis	35.7	18.0	0.00	35.8	14.4	0.00	31.9	17.0	0.00
3bp8	16.2	8.5	0.00	16.9	10.3	0.08	9.0	3.3	0.27
3d5s	3.5	1.3	0.56	5.3	2.4	0.54	5.4	2.4	0.52
3sgq	11.5	4.8	0.02	12.4	6.3	0.00	13.0	5.4	0.13
7cei	19.6	10.5	0.02	4.7	1.6	0.77	4.2	1.2	0.96
TA04	41.0	13.1	0.00	38.8	14.2	0.00	35.8	13.3	0.00
TA05	28.0	11.6	0.00	24.5	12.4	0.00	30.2	12.8	0.00
TA08	11.1	3.3	0.50	12.3	2.4	0.59	12.5	3.5	0.55
TA13	21.2	1.1	0.73	22.5	1.3	0.64	22.2	1.3	0.70
TA18	6.6	2.2	0.81	5.0	1.8	0.71	7.2	2.4	0.69
TA21	40.0	9.6	0.00	32.3	13.6	0.00	36.7	20.2	0.00
TA22	48.1	15.4	0.00	48.1	15.3	0.00	46.5	14.3	0.00
TA27	28.9	12.9	0.00	28.5	13.8	0.00	28.9	12.3	0.00
TA30	47.3	17.9	0.00	49.0	18.4	0.00	49.0	18.4	0.00
TA32	23.3	9.2	0.00	23.3	9.2	0.00	29.2	12.5	0.08
TA39	21.2	13.3	0.00	21.9	11.2	0.00	21.9	11.2	0.00
TA41	15.9	6.0	0.19	28.1	13.2	0.07	7.9	2.4	0.56
Average	20.0	7.6	0.24	19.7	7.8	0.25	18.2	7.1	0.29

Table 4.4. Performance of CG-CSA compared to R-CSA and "CSA-Lee" on CAPRI round 5 targets.

Targets	LRMSD / IRMSD / Fnat / Quality[1)]											
	CSA-Lee				**R-CSA**				**CG-CSA**			
TA14	54.4	20.1	0.00	-	50.7	13.3	0.01	-	3.7	2.2	0.30	*
TA15	8.8	3.3	0.18	*	3.6	1.3	0.80	**	2.2	1.0	0.77	***
TA18	32.4	15.2	0.00	-	5.0	1.8	0.71	**	7.2	2.4	0.69	*
TA19	26.1	14.6	0.00	-	9.9	3.3	0.40	*	9.8	3.4	0.35	*

1) Ligand RMSD, interface RMSD, fraction of native contacts, and model quality by CAPRI criterion (High quality(***), Medium quality(**), Acceptable quality(*)).

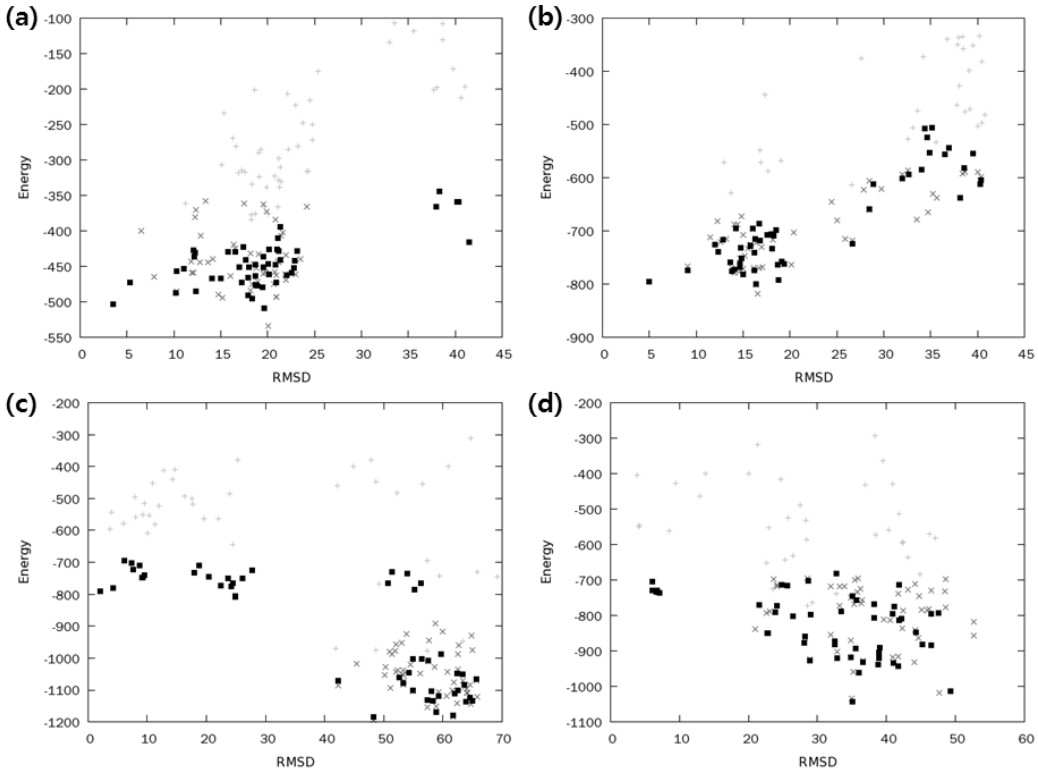


Figure 4.2. Ligand RMSD (LRMSD) versus energy plots for initial structures (+), final structures of R-CSA (×), and final structures of CG-CSA (■) on **(a)** 1ay7, **(b)** 1r0r, **(c)** 1iqd, and **(d)** 1bj1. Initial bank conformations brought from ZDOCK runs, shared by both CSA runs, are plotted as well in gray dots. X-axis is LRMSD between ligand protein of native complex and that of predicted complexes. Y-axis is energy value of predicted complexes.

4.3.2. Comparison to other protein-protein docking methods

For comparison with other protein-protein docking tools, we tested CG-CSA to ZDOCK (Mintseris *et al.,* 2007) which is one of most popular rigid-body docking programs and popular refinement docking programs such as RosettaDock (Gray *et al.,* 2003), FireDock (Andrusier *et al.,* 2007), and FiberDock (Mashiach *et al.,* 2010) (**Table 4.5**). To describe how the results were collected, ZDOCK result is collected by picking the best structure in 10 top-scoring structures ranked by ZDOCK score. Selected 10 structures were further refined by other refinement docking programs such as RosettaDock, FireDock, and FiberDock. RosettaDock generated 500 refined models for each selected structures and generated 5000 refined models totally. Then, lowest energy structures from 500 refined models for each of the 10 structures were selected (Pierce and Weng 2008). FireDock and FiberDock generated 10 refined structures from each 10 initial structures. The fraction of targets within “acceptable” quality in CAPRI measure for CG-CSA is 43.4% on test set, compared to 37.7% for ZDOCK, 32.1% for RosettaDock, 37.7% for FireDock, and 39.6% for FiberDock. The fraction of targets within “medium” quality in CAPRI measure for CG-CSA is 27.4% on test set, compared to 25.5% for ZDOCK, 17.9% for RosettaDock, 23.6% for FireDock, and 25.5% for FiberDock. CG-CSA also showed the best performance in terms of the predictions better than “acceptable” and “medium” accuracy. In case of top5 selection cases, the fraction of targets with better than “acceptable” quality for CG-CSA is 36.8% of the test targets, compared to 28.3% for ZDOCK, 25.5% for RosettaDock, 28.3% for FireDock, and 30.2% for FiberDock, and the fraction of targets with better than “medium” quality is 23.6% for CG-CSA, compared to 21.7% for ZDOCK, 15.1% for RosettaDock, 17.9% for FireDock, and 20.8% for FiberDock. CG-CSA also showed best performance at top5 selection cases.

According to the comparison above, we claim that regular CSA combined with current energy function is already good enough to be comparable to other methods, but adopting cluster-guided approach further improves it. We attribute success to both energy function and the sampling method. Using all-atom energy function combined with physics-based energy function and knowledge-based energy function can make better performance of CG-CSA. And success ratio of CG-CSA is higher than that of RosettaDock, FireDock, and FiberDock, because CG-CSA uses main concept of genetic algorithm rather than RosettaDock, FireDock, and FiberDock use Monte Carlo-based method. Crossover of translational and rotational degree of freedom can generate large perturbed conformations and search diverse local minima efficiently than mutation of translational and rotational degree of freedom. It makes sampling space of CG-CSA get broader and find global minimum efficiently. For example, the minimum LRMSD of initial bank on target 1udi is larger than 20 Å, but the minimum LRMSD of final bank of R-CSA and CG-CSA is smaller than 5 Å (**Figure 4.3**). CSA could generate better models by perturbing ligand proteins more than 15 Å. This result shows that large perturbation based on crossover of translational and rotational degree of freedom can generate successful models although structures of initial bank is so far from native structures.

Table 4.5. Performance comparison of CG-CSA, ZDOCK, RosettaDock, FireDock, and FiberDock on 106 benchmark test targets in terms of the percentage of targets predicted with better than acceptable/medium accuracy

Top10	CG-CSA	ZDOCK	Rosetta Dock	FireDock	FiberDock
> acceptable	43.4%	37.7%	32.1%	37.7%	39.6%
> medium	27.4%	25.5%	17.9%	23.6%	25.5%
Top5	CG-CSA	ZDOCK	Rosetta Dock	FireDock	FiberDock
> acceptable	36.8%	28.3%	25.5%	28.3%	30.2%
> medium	23.6%	21.7%	15.1%	17.9%	20.8%
Top1	CG-CSA	ZDOCK	Rosetta Dock	FireDock	FiberDock
> acceptable	12.3%	16.0%	14.2%	14.2%	15.1%
> medium	8.5%	11.3%	8.4%	10.4%	11.3%

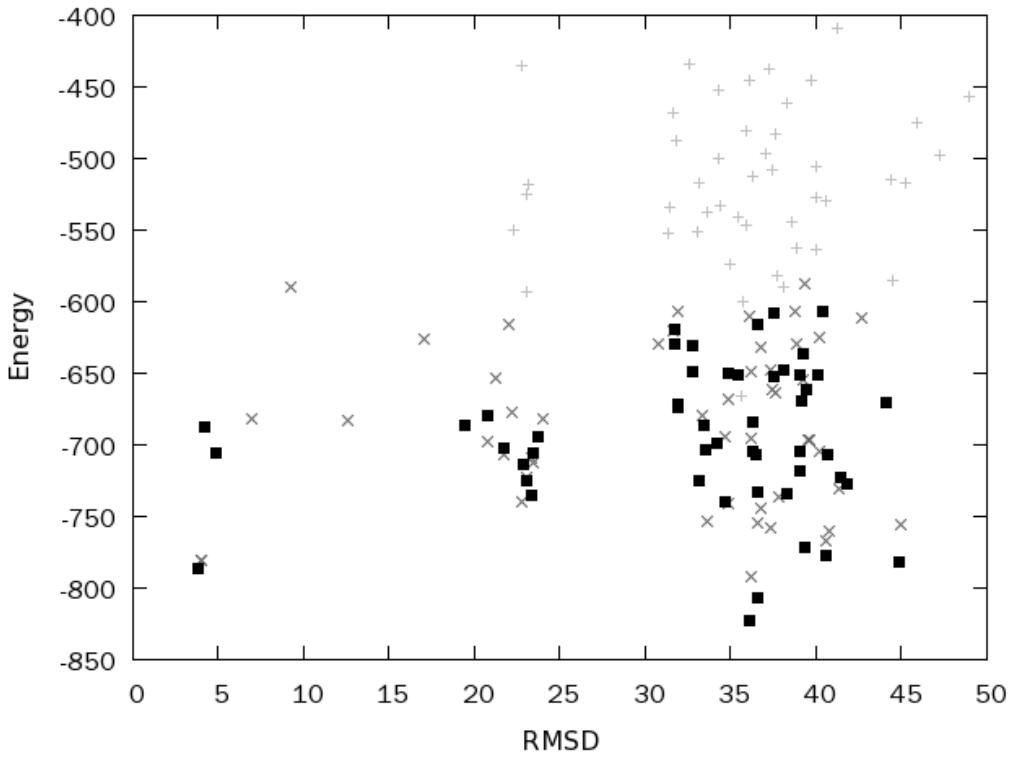


Figure 4.3. Ligand RMSD (LRMSD) versus energy plots for initial structures (+), final structures of R-CSA (×), and final structures of CG-CSA (■) on 1udi. Initial bank conformations brought from ZDOCK runs, shared by both CSA runs, are plotted as well in gray dots. X-axis is LRMSD between ligand protein of native complex and that of predicted complexes. Y-axis is energy value of predicted complexes.

4.3.3. Performance of GalaxyPPDock on recent CAPRI targets

We tested CG-CSA on 7 targets from recent CAPRI round. We compared the performance of CG-CSA to CAPRI predictors (Bonvin, Bates, Vakser) who did best on CAPRI from round 22 to round 27. This comparison will not only show the status of CG-CSA compared to state-of-the-art methods in the community, but also will show progress in the method during recent CAPRI rounds. CG-CSA predicted structures better than acceptable for all 7 targets and among them, models for target 53 and 58 showed medium quality. This overall result is better than any of top3 predictors' results.

CAPRI target 53 and target 58 are successful example of GalaxyPPDock (**Figure 4.3**). Especially the performance of GalaxyPPDock is better than other top3 CAPRI predictors. Target 53 (PDB ID: 4JW2) is designed Rep4/Rep2 α-repeat complexes and network of hydrophobic and aromatic residues is a key interaction of target 53. GalaxyPPDock predicted this target about 5.0Å and hydrophobic network of this target. Target 58 (PDB ID: 4G9S) is PilG/SalG lysozyme complex. Coulomb interaction of Aspartic acid, Glutamic acid and Arginine is key interaction of target 58. GalaxyPPDock predicted well about 3.0 Å and coulomb interaction of this target.

Table 4.6. Performance of CG-CSA compared to other top3 predictors on targets of CAPRI round 22-27.

Targets	LRMSD / IRMSD / Fnat / Quality[1]															
	CG-CSA (7/2**)				Bonvin (6/2**)[2]				Bates (5/1**)[2]				Vakser (5)[2]			
TA46	8.2	4.1	0.24	*	7.8	3.4	0.41	*	13.0	4.7	0.15	-	34.8	13.8	0.00	-
TA48	8.2	2.7	0.43	*	9.1	3.4	0.23	*	7.4	4.6	0.19	*	9.7	4.6	0.14	*
TA49	13.0	3.2	0.23	*	14.0	3.6	0.26	*	7.2	3.9	0.10	*	9.7	4.1	0.14	*
TA50	7.7	2.2	0.45	*	5.5	1.9	0.47	**	5.4	2.7	0.29	*	5.4	2.2	0.35	*
TA53	5.1	1.9	0.69	**	4.5	2.2	0.46	**	9.4	4.2	0.35	*	16.7	7.6	0.12	-
TA54	5.6	3.0	0.57	*	18.6	7.7	0.02	-	10.1	5.2	0.14	-	5.9	3.6	0.14	*
TA58	3.3	1.1	0.65	**	6.9	2.6	0.29	*	3.7	1.6	0.56	**	8.9	3.2	0.43	*

1) Ligand RMSD, interface RMSD, fraction of native contacts, and model quality by CAPRI criterion (High quality(***), Medium quality(**), Acceptable quality(*)).

2) Top 3 predictors in CARPI round 22-27.

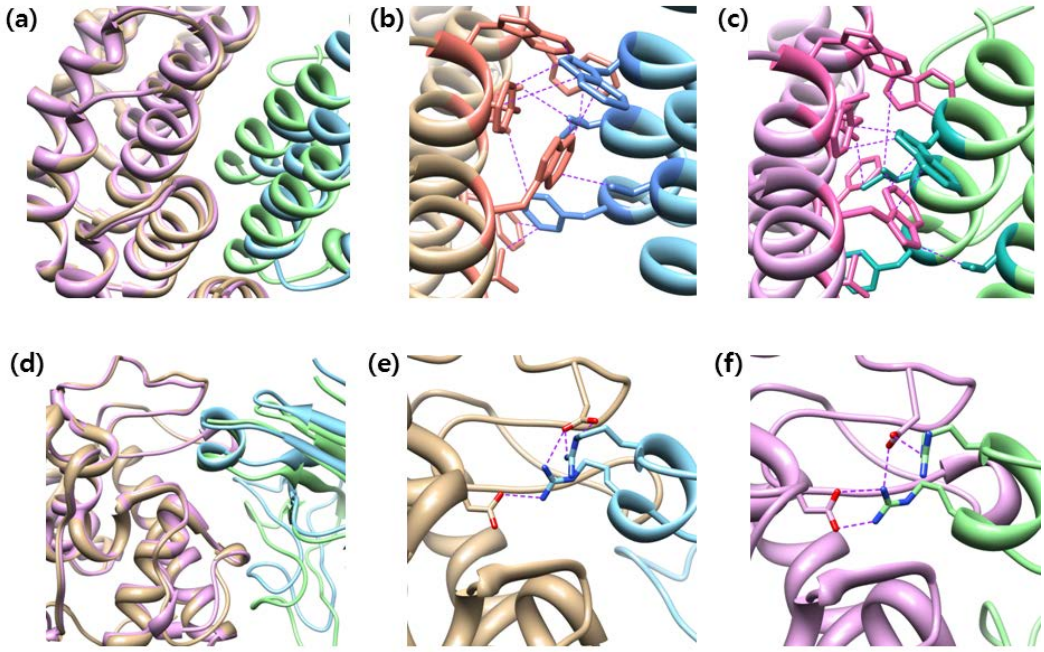


Figure 4.4. Successful examples of GalaxyPPDock on CAPRI target 53 (**(a)** to **(c)**, designed Rep4/Rep2 α-repeat complex, PDB ID: 4JW2) and CAPRI target 58 (**(d)** to **(f)**, PilG/SalG lysozyme complex, PDB ID: 4G9S). Structures colored in yellow and sky blue (panel **(b)** and **(e)**) are receptor and ligand proteins of the native structures, and plum and light green (panel **(c)** and **(f)**) are receptor and ligand proteins of predicted complex generated by GalaxyPPDock. There are hydrophobic interactions of **(b)** native structure and **(c)** GalaxyPPDock model on CAPRI target 53 and polar interactions of **(e)** native structure and **(f)** GalaxyPPDock model of CAPRI target 58.

4.3.4. Protein-protein docking with side-chain flexibility

Side-chain flexibility has an effect on interaction of receptor and ligand protein. We calculated fraction of native contact (f_{nat}) of GalaxyPPDock models and "unbound model" generated by superposing unbound subunit structure to GalaxyPPDock models. Fraction of native contact (f_{nat}) of GalaxyPPDock is slightly better than that of "unbound model" (**Table 4.7**). The different of each value is small, but chi-angle changes of key residues have a great effect on interactions of receptor proteins and ligand proteins. In target 53, side-chain flexibility of phenylalanine residue of receptor protein can generate hydrophobic interaction. In target 58, side-chain flexibility of arginine residue of ligand protein can generate coulomb interaction (**Figure 4.4**). These results show that protein-protein docking with side-chain flexibility more accurately predicts interaction of protein complexes and it derives generate more accurate protein complex models.

Table 4.7. Fraction of native contact (f_{nat}) for CG-CSA models and unbound complexes made by superimposing unbound structures on CG-CSA models. Better cases (37 targets on 106 targets). Worse cases (30 targets on 106 targets). Same cases (30 targets on 106 targets).

Target	Superposed Unbound structure	CG-CSA model
1a2k	0.7045	0.7500
1ahw	0.0000	0.0000
1ak4	0.0227	0.0227
1akj	0.0000	0.0000
1ay7	0.8750	0.8000
1azs	0.0000	0.0000
1b6c	0.7321	0.7679
1bj1	0.8429	0.8571
1bvk	0.1458	0.1458
1bvn	0.7260	0.6849
1cgi	0.4941	0.5765
1d6r	0.0172	0.0172
1dfj	0.6301	0.6575
1dqj	0.0000	0.0000
1e6e	0.8462	0.8846
1e6j	0.0980	0.0980
1e96	0.0000	0.0000
1efn	0.0000	0.0294
1ewy	0.0222	0.0444
1ezu	0.0000	0.0000
1f34	0.0345	0.0345
1f51	0.5968	0.6290
1fcc	0.0000	0.0000
1ffw	0.4444	0.5000
1fle	0.0282	0.0141
1fqj	0.0000	0.0000
1fsk	0.8939	0.8636
1gcq	0.1111	0.1333
1gl1	0.6406	0.5625
1gla	0.0000	0.0000
1gpw	0.6618	0.6471
1hcf	0.1111	0.0444
1he1	0.7460	0.7302
1hia	0.1587	0.1270
1i4d	0.0545	0.0727
1i9r	0.0000	0.0000
1iqd	0.6933	0.7067
1jtg	0.3978	0.4086
1k74	0.6269	0.6269
1kac	0.0000	0.0000
1klu	0.0000	0.0000
1ktz	0.0333	0.1000
1kxp	0.5283	0.5283
1ml0	0.7534	0.8082
1n8o	0.8052	0.8182
1nca	0.0000	0.0000
1nsn	0.0000	0.0000
1ofu	0.0000	0.0000
1oyv	0.5543	0.5652
1ppe	0.7887	0.8591
1pvh	0.1333	0.1333
1qa9	0.0000	0.0000

1r0r	0.4930	0.4930
1rlb	0.0000	0.0000
1rv6	0.4255	0.3617
1sbb	0.0000	0.0000
1tmq	0.7200	0.7333
1udi	0.4267	0.4400
1vfb	0.2083	0.2292
1wdw	0.5487	0.5664
1wej	0.8372	0.7209
1xd3	0.3750	0.4000
1xu1	0.0678	0.0678
1yvb	0.7000	0.7600
1z0k	0.3947	0.3947
1z5y	0.3774	0.3585
1zhh	0.0000	0.0000
1zhi	0.0244	0.0244
2a5t	0.3390	0.3559
2a9k	0.0000	0.0000
2abz	0.1017	0.0678
2ajf	0.0000	0.0000
2b42	0.8427	0.8764
2btf	0.0000	0.0000
2fd6	0.2128	0.2340
2fju	0.0000	0.0000
2g77	0.0000	0.0000
2hle	0.4268	0.4268
2hqs	0.0645	0.0645
2i25	0.0185	0.0185
2j0t	0.0517	0.0517
2jel	0.3036	0.2857
2mta	0.0000	0.0000
2oob	0.0370	0.1111
2oor	0.0435	0.0435
2oul	0.8333	0.8205
2pcc	0.3793	0.3448
2sic	0.7606	0.7746
2uuy	0.0000	0.0000
2vis	0.0000	0.0000
3bp8	0.2653	0.2653
3d5s	0.5000	0.5200
3sgq	0.1273	0.1273
7cei	0.8462	0.9615
TA04	0.0000	0.0000
TA05	0.0000	0.0000
TA08	0.5758	0.5454
TA13	0.7143	0.7000
TA18	0.7206	0.6912
TA21	0.0000	0.0000
TA22	0.0000	0.0000
TA27	0.0000	0.0000
TA30	0.0000	0.0000
TA32	0.0814	0.0814
TA39	0.0000	0.0000
TA41	0.6610	0.5593
Average	0.2873	0.2899

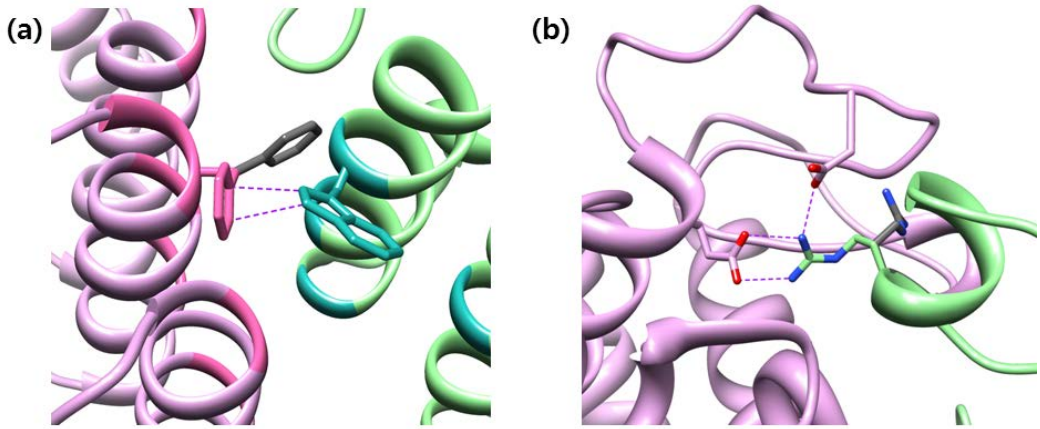


Figure 4.5. Interaction of models generated by GalaxyPPDock on CAPRI target 53 **(a)** and on CAPRI target 58 **(b)**. Residue colored in gray is side-chain of unbound structure.

4.3.5. Contribution of GalaxyPPDock energy components

We analyzed the performance and contribution of GalaxyPPDock energy components on ZDOCK benchmark set and CAPRI targets (Total 141 targets). We defined the success target when minimum LRMSD of selected 10 models ranked by each energy components among 50 final structures. The contribution of each energy components was calculated by average of standard deviation of final bank energy. Coulomb electrostatic interaction and hydrogen bond showed good performance for selecting near-native structures among structures of final bank, but their contribution smaller than other energy components (**Table 4.8**). These results imply that many proteins interact with other proteins through polar interactions. Therefore, it is need to consider electrostatic interaction and hydrogen interactions more importantly to generate more accurate energy function, and increasing weights of electrostatic interaction and hydrogen bond can be one of the methods to generate better protein-protein docking energy function.

Table 4.8. Performance and contribution of each energy components

	Number of Success targets	Success ratio	Contribution
$\mathbf{E}_{DFIRE}$	32	22.7%	26.9%
$\mathbf{E}_{vdw}$	39	27.7%	9.9%
$\mathbf{E}_{elec}$	49	34.8%	5.4%
$\mathbf{E}_{SA}$	36	25.5%	18.5%
$\mathbf{E}_{HBond}$	46	32.6%	5.0%
$\mathbf{E}_{cons}$	34	24.1%	17.5%
$\mathbf{E}_{rotamer}$	25	17.7%	16.8%

4.4. Conclusions

In this study, we introduced GalaxyPPDock which uses a new variant of CSA algorithm for protein-protein docking study. GalaxyPPDock focuses on regions on low-energy clusters, but keeps high-energy clusters and it helps to generate near-native predicted complexes not only energy function is relative accurate but also energy function is inaccurate. GalaxyPPDock generated more successful predicted complex than original CSA and other docking program ZDOCK and RosettaDock on benchmark set. Moreover, GalaxyPPDock shows good performance on recent CAPRI targets. Based on these results, it is concluded that GalaxyPPDock is good protein-protein docking program and efficient sampling of conformation space in protein-protein docking is very important. In spite of these achievements, developing accurate protein-protein docking program is still challengeable problem. Considering backbone flexibility can improve the performance of GalaxyPPDock by combining loop modeling using GalaxyLoop (Ko *et al.,* 2011; Lee *et al.,* 2010) or MD-based backbone refinement using GalaxyRefine (Heo *et al.,* 2013). Also, performance of GalaxyPPDock can be improved using experimental data such as small-angle X-ray scattering (SAXS) by selecting from initial structures generates to make better initial bank (Lensink and Wodak 2013).

5. Conclusions

We developed programs for predicting protein interactions based on bioinformatics and physicochemical approaches. For developing GalaxyGemini for predicting homo-oligomer structures and GalaxyPepDock for predicting protein-peptide interactions, we used bioinformatics approaches. GalaxyGemini searches good oligomer templates compared to other methods including naïve method using HHsearch, because GalaxyGemini uses both tertiary structure similarity and quaternary structure similarity by interface alignment score. GalaxyPepDock searches protein-peptide template based on protein structure similarity and protein-peptide interaction similarity. Picking oligomer-oriented bioinformatics feature can find good template and the great reason for success of GalaxyGemini and GalaxyPepDock. For developing GalaxyPPDock for predicting protein-protein interactions, we used physical chemistry approach. Both approaches are effective for generating good models. GalaxyPPDock uses Cluster-Guided Conformational Space Annealing, one of global optimization methods to finding global minimum effectively. Developing effective global optimization method is main reason of success of GalaxyPPDock. These results show that both bioinformatics method and physical chemistry method can be used to predict protein interaction.

Although, GalaxyGemini and GalaxyPepDock used bioinformatics approaches, and GalaxyPPDock used physical chemistry approaches, both bioinformatics approaches and physical chemistry approaches can be used for predicting homo-oligomer interactions, protein-peptide interactions, and protein-protein interactions. Hydrophobic interactions are key interactions of homo-oligomers, so native homo-oligomer is global minimum of energy landscape of homo-oligomer (Inbar *et al.,* 2005). Also, symmetry is very key point of sampling

homo-oligomer structure. Therefore, developing global optimization methods considering symmetric constraints can predict homo-oligomer interactions more accurately. For protein-peptide interactions, GalaxyPepDock used Molecular Dynamics-based refinement method, and it helps to improve the quality of protein-peptide complex models. For a few decades, *ab initio* methods were the majority of protein-protein docking methods, because of the database of protein-protein complexes were small. However, the number of experimentally resolved protein complex structures has been increasing, so data-driven protein-protein docking methods attract a lot of attention. HADDOCK, one of data-driven protein-protein docking method showed a good performance on the latest CAPRI experiments (Lensink and Wodak 2013).

In this research, I showed that bioinformatics approaches can help predict homo-oligomer interactions and protein-peptide interactions and physical chemistry approaches can help predict protein-protein interactions. Also, there are many studies that protein interactions can be predicted by both bioinformatics approaches and physical chemistry approaches. Therefore, combining bioinformatics approaches and physical chemistry approaches will help improve the performance of programs for predicting homo-oligomer interactions, protein-peptide interactions, and protein-protein interactions.

BIBLIOGRAPHY

Alsop, J. D., and Mitchell, J. C. (2015). "Interolog interfaces in protein-protein docking." Proteins 83, 1940-1946.

Andrusier, N., Nussinov, R., and Wolfson, H. J. (2007). "FireDock: fast interaction refinement in molecular docking." Proteins 69, 139-159.

Bonvin, A. M. (2006). "Flexible protein-protein docking." Curr Opin Struct Biol 16, 194-200.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., and Huang, E. S. (2004). "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?" Protein Sci 13, 190-202.

Chen, R., Li, L., and Weng, Z. (2003). "ZDOCK: an initial-stage protein-docking algorithm." Proteins 52, 80-87.

Das, A. A., Sharma, O. P., Kumar, M. S., Krishna, R., and Mathur, P. P. (2013). "PepBind: a comprehensive database and computational tool for analysis of protein-peptide interactions." Genomics Proteomics Bioinformatics 11, 241-246.

Dominguez, C., Boelens, R., and Bonvin, A. M. (2003). "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information." J Am Chem Soc 125, 1731-1737.

Donsky, E., and Wolfson, H. J. (2011). "PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors." Bioinformatics 27, 2836-2842.

Dunbrack, R. L., Jr., and Cohen, F. E. (1997). "Bayesian statistical analysis of protein side-chain rotamer preferences." Protein Sci 6, 1661-1681.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D.,

Shen, M. Y., Pieper, U., and Sali, A. (2006). "Comparative protein structure modeling using Modeller." Curr Protoc Bioinformatics Chapter 5, Unit 5 6.

Fuhrmann, J., Rurainski, A., Lenhof, H. P., and Neumann, D. (2009). "A new method for the gradient-based optimization of molecular complexes." J Comput Chem 30, 1371-1378.

Gabb, H. H., Jackson, R. M., and Sternberg, M. J. E. (1997). "Modelling Protein Docking using Shape Complementarity, Electrostatics and Biochemical Information." J Mol Biol 272, 106-120.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. (2003). "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations." J Mol Biol 331, 281-299.

Gray, J. J. (2006). "High-resolution protein-protein docking." Curr Opin Struct Biol 16, 183-193.

Heldin, C. H. (1995). "Dimerization of cell surface receptors in signal transduction." Cell 80, 213-223.

Henikoff, S., and Henikoff, J. G. (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A 89, 10915-10919.

Heo, L., Park, H., and Seok, C. (2013). "GalaxyRefine: Protein structure refinement driven by side-chain repacking." Nucleic Acids Res 41, W384-388.

Huang, S. Y. (2014). "Search strategies and evaluation in protein-protein docking: principles, advances and challenges." Drug Discov Today 19, 1081-1096.

Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010). "Protein-protein docking benchmark version 4.0." Proteins 78, 3111-3114.

Inbar, Y., Benyamini, H., Nussinov, R., and Wolfson, H. J. (2005). "Prediction of multimolecular assemblies by multiple docking." J Mol Biol 349, 435-447.

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I., Wodak, S. J., and Critical Assessment of, P. I. (2003). "CAPRI: a Critical Assessment of PRedicted Interactions." Proteins 52, 2-9.

Janin, J. (2005). "The targets of CAPRI rounds 3-5." Proteins 60, 170-175.

Janin, J. (2007). "The targets of CAPRI rounds 6-12." Proteins 69, 699-703.

Janin, J. (2010). "The targets of CAPRI Rounds 13-19." Proteins 78, 3067-3072.

Janin, J. (2013). "The targets of CAPRI rounds 20-27." Proteins 81, 2075-2081.

Jones, S., and Thornton, J. M. (1997). "Analysis of protein-protein interaction sites using surface patches." J Mol Biol 272, 121-132.

Joo, K., Lee, J., Seo, J. H., Lee, K., Kim, B. G., and Lee, J. (2009). "All-atom chain-building by optimizing MODELLER energy function using conformational space annealing." Proteins 75, 1010-1023.

Kelley, L. A., Gardner, S. P., and Sutcliffe, M. J. (1996). "An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies." Protein Eng 9, 1063-1065.

Keskin, O., Ma, B., Rogale, K., Gunasekaran, K., and Nussinov, R. (2005). "Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach." Phys Biol 2, S24-35.

Ko, J., Lee, D., Park, H., Coutsias, E. A., Lee, J., and Seok, C. (2011). "The FALC-Loop web server for protein loop modeling." Nucleic Acids Res 39, W210-214.

Ko, J., Park, H., and Seok, C. (2012). "GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions." BMC Bioinformatics 13, 198.

Kortemme, T., Morozov, A. V., and Baker, D. (2003). "An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for

proteins and protein-protein complexes." J Mol Biol 326, 1239-1259.

Kozakov, D., Brenke, R., Comeau, S. R., and Vajda, S. (2006). "PIPER: an FFT-based protein docking program with pairwise potentials." Proteins 65, 392-406.

Krissinel, E., and Henrick, K. (2007). "Inference of macromolecular assemblies from crystalline state." J Mol Biol 372, 774-797.

Kurcinski, M., Jamroz, M., Blaszczyk, M., Kolinski, A., and Kmiecik, S. (2015). "CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site." Nucleic Acids Res 43, W419-424.

Lavi, A., Ngan, C. H., Movshovitz-Attias, D., Bohnuud, T., Yueh, C., Beglov, D., Schueler-Furman, O., and Kozakov, D. (2013). "Detection of peptide-binding sites on protein surfaces: the first step toward the modeling and targeting of peptide-mediated interactions." Proteins 81, 2096-2105.

Lee, J., Scheraga, H. A., and Rackovsky, S. (1998). "Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing." Biopolymers 46, 103-116.

Lee, J., Lee, D., Park, H., Coutsias, E. A., and Seok, C. (2010). "Protein loop modeling by using fragment assembly and analytical loop closure." Proteins 78, 3428-3436.

Lee, K., Czaplewski, C., Kim, S. Y., and Lee, J. (2005). "An efficient molecular docking using conformational space annealing." J Comput Chem 26, 78-87.

Lee, K., Sim, J., and Lee, J. (2005). "Study of protein-protein interaction using conformational space annealing." Proteins 60, 257-262.

Lensink, M. F., and Wodak, S. J. (2013). "Docking, scoring, and affinity prediction in CAPRI." Proteins 81, 2082-2095.

Levy, E. D., Boeri Erba, E., Robinson, C. V., and Teichmann, S. A. (2008).

"Assembly reflects evolution of protein complexes." Nature 453, 1262-1265.

Liang, S., Meroueh, S. O., Wang, G., Qiu, C., and Zhou, Y. (2009). "Consensus scoring for enriching near-native structures from protein-protein docking decoys." Proteins 75, 397-403.

London, N., Movshovitz-Attias, D., and Schueler-Furman, O. (2010). "The structural basis of peptide-protein binding strategies." Structure 18, 188-199.

London, N., Raveh, B., Cohen, E., Fathi, G., and Schueler-Furman, O. (2011). "Rosetta FlexPepDock web server--high resolution modeling of peptide-protein interactions." Nucleic Acids Res 39, W249-253.

London, N., Raveh, B., and Schueler-Furman, O. (2013). "Peptide docking and structure-based characterization of peptide binding: from knowledge to know-how." Curr Opin Struct Biol 23, 894-902.

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). "All-atom empirical potential for molecular modeling and dynamics studies of proteins." J Phys Chem B 102, 3586-3616.

Maclaine, N. J., and Hupp, T. R. (2011). "How phosphorylation controls p53." Cell Cycle 10, 916-921.

Malod-Dognin, N., Bansal, A., and Cazals, F. (2012). "Characterizing the morphology of protein binding patches." Proteins 80, 2652-2665.

Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I., and Ten Eyck, L. F. (2001). "Protein docking using continuum electrostatics and geometric fit." Protein Eng 14, 105-113.

Mariani, V., Kiefer, F., Schmidt, T., Haas, J., and Schwede, T. (2011). "Assessment of template based protein structure predictions in CASP9." Proteins 79 Suppl 10, 37-58.

Mashiach, E., Nussinov, R., and Wolfson, H. J. (2010). "FiberDock: Flexible induced-fit backbone refinement in molecular docking." Proteins 78, 1503-1519.

Miller, M. L., Jensen, L. J., Diella, F., Jorgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S. A., Bordeaux, J., Sicheritz-Ponten, T., et al. (2008). "Linear motif atlas for phosphorylation-dependent signaling." Sci Signal 1, ra2.

Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R., and Weng, Z. (2007). "Integrating statistical pair potentials into protein complex prediction." Proteins 69, 511-520.

Negri, A., Rodriguez-Larrea, D., Marco, E., Jimenez-Ruiz, A., Sanchez-Ruiz, J. M., and Gago, F. (2010). "Protein-protein interactions at an enzyme-substrate interface: characterization of transient reaction intermediates throughout a full catalytic cycle of Escherichia coli thioredoxin reductase." Proteins 78, 36-51.

Morita, M., Kakuta, M., Shimizu, K., and Nakamura, S. (2012). “Blind prediction of quaternary structures of homo-oligomeric proteins from amino acid sequences based on template.” J Proteome Sci Comput Biol 1, 1.

Ozbabacan, S. E., Engin, H. B., Gursoy, A., and Keskin, O. (2011). "Transient protein-protein interactions." Protein Eng Des Sel 24, 635-648.

Park, H., Ko, J., Joo, K., Lee, J., Seok, C., and Lee, J. (2011). "Refinement of protein termini in template-based modeling using conformational space annealing." Proteins 79, 2725-2734.

Park, H., and Seok, C. (2012). "Refinement of unreliable local regions in template-based protein models." Proteins 80, 1974-1986.

Park, H., Lee, G. R., Heo, L., and Seok, C. (2014). "Protein loop modeling using a new hybrid energy function and its application to modeling in inaccurate structural environments." PLoS One 9, e113811.

Park, H., Lee, H., and Seok, C. (2015). "High-resolution protein-protein docking

by global optimization: recent advances and future challenges." Curr Opin Struct Biol 35, 24-31.

Pawson, T., and Nash, P. (2000). "Protein-protein interactions define specificity in signal transduction." Genes Dev 14, 1027-1047.

Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010). "Transient protein-protein interactions: structural, functional, and network properties." Structure 18, 1233-1243.

Petsalaki, E., and Russell, R. B. (2008). "Peptide-mediated interactions in biological systems: new discoveries and applications." Curr Opin Biotechnol 19, 344-350.

Petsalaki, E., Stark, A., Garcia-Urdiales, E., and Russell, R. B. (2009). "Accurate prediction of peptide binding sites on protein surfaces." PLoS Comput Biol 5, e1000335.

Pierce, B., and Weng, Z. (2008). "A combination of rescoring and refinement significantly improves protein docking performance." Proteins 72, 270-279.

Pluckthun, A., and Pack, P. (1997). "New protein engineering approaches to multivalent and bispecific antibody fragments." Immunotechnology 3, 83-105.

Postingl, H., Kabir, T., and Thornton, J. M. (2003). “Automatic inference of protein quaternary structure from crystals.” J Appl Cryst 36, 1116-1112.

Poupon, A., and Janin, J. (2010). "Analysis and prediction of protein quaternary structure." Methods Mol Biol 609, 349-364.

Raveh, B., London, N., and Schueler-Furman, O. (2010). "Sub-angstrom modeling of complexes between flexible peptides and globular proteins." Proteins 78, 2029-2040.

Raveh, B., London, N., Zimmerman, L., and Schueler-Furman, O. (2011). "Rosetta

FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors." PLoS One 6, e18934.

Ritchie, D. W. (2008). "Recent progress and future directions in protein-protein docking." Curr Protein Pept Sci 9, 1-15.

Saladin, A., Rey, J., Thevenet, P., Zacharias, M., Moroy, G., and Tuffery, P. (2014). "PEP-SiteFinder: a tool for the blind identification of peptide binding sites on protein surfaces." Nucleic Acids Res 42, W221-226.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). "PatchDock and SymmDock: servers for rigid and symmetric docking." Nucleic Acids Res 33, W363-367.

Scott, J. D., and Pawson, T. (2009). "Cell signaling in space and time: where proteins come together and when they're apart." Science 326, 1220-1224.

Shin, W. H., Heo, L., Lee, J., Ko, J., Seok, C., and Lee, J. (2011). "LigDockCSA: protein-ligand docking using conformational space annealing." J Comput Chem 32, 3226-3232.

Shin, W. H., and Seok, C. (2012). "GalaxyDock: protein-ligand docking with flexible protein side-chains." J Chem Inf Model 52, 3225-3232.

Shin, W. H., Kim, J. K., Kim, D. S., and Seok, C. (2013). "GalaxyDock2: protein-ligand docking using beta-complex and global optimization." J Comput Chem 34, 2647-2656.

Snijder, H. J., Ubarretxena-Belandia, I., Blaauw, M., Kalk, K. H., Verheij, H. M., Egmond, M. R., Dekker, N., and Dijkstra, B. W. (1999). "Structural evidence for dimerization-regulated activation of an integral membrane phospholipase." Nature 401, 717-721.

Soding, J. (2005). "Protein homology detection by HMM-HMM comparison." Bioinformatics 21, 951-960.

Su, Y., Zhou, A., Xia, X., Li, W., and Sun, Z. (2009). "Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction." Protein Sci 18, 2550-2558.

Trabuco, L. G., Lise, S., Petsalaki, E., and Russell, R. B. (2012). "PepSite: prediction of peptide-binding sites from protein surfaces." Nucleic Acids Res 40, W423-427.

Trellet, M., Melquiond, A. S., and Bonvin, A. M. (2013). "A unified conformational selection and induced fit approach to protein-peptide docking." PLoS One 8, e58769.

Vakser, I. A. (1997). "Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex." Proteins Suppl 1, 226-230.

Venkatraman, V., Yang, Y. D., Sael, L., and Kihara, D. (2009). "Protein-protein docking using region-based 3D Zernike descriptors." BMC Bioinformatics 10, 407.

Vlieghe, P., Lisowski, V., Martinez, J., and Khrestchatisky, M. (2010). "Synthetic therapeutic peptides: science and market." Drug Discov Today 15, 40-56.

Wen, W., Meinkoth, J. L., Tsien, R. Y., and Taylor, S. S. (1995). "Identification of a signal for rapid export of proteins from the nucleus." Cell 82, 463-473.

Yan, C., and Zou, X. (2015). "Predicting peptide binding sites on protein surfaces by clustering chemical interactions." J Comput Chem 36, 49-61.

Yang, Y., Ludwig, R. L., Jensen, J. P., Pierre, S. A., Medaglia, M. V., Davydov, I. V., Safiran, Y. J., Oberoi, P., Kenten, J. H., Phillips, A. C., et al. (2005). "Small molecule inhibitors of HDM2 ubiquitin ligase activity stabilize and activate p53 in cells." Cancer Cell 7, 547-559.

Zhang, Y., and Skolnick, J. (2005). "TM-align: a protein structure alignment algorithm based on the TM-score." Nucleic Acids Res 33, 2302-2309.

Zhou, H., and Zhou, Y. (2002). "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction." Protein Sci 11, 2714-2726.

Zhou, H., and Zhou, Y. (2002). "Stability scale and atomic solvation parameters extracted from 1023 mutation experiments." Proteins 49, 483-492.

국문초록

단백질은 생명체 내에서 서로 상호작용함으로써 기능을 수행한다. 단백질의 상호작용 연구를 통해 단백질의 기능을 보다 정확히 이해하는 것은 신약개발에 있어서 매우 중요하다. 단백질 상호작용은 호모-올리고머 상호작용, 단백질-펩타이드 상호작용, 단백질-단백질 상호작용으로 구분된다. 단백질 상호작용을 X-선 결정법이나 핵자기공명과 같은 실험적인 방법으로 알 수도 있으나, 현재까지 실험적으로 밝혀진 상호작용 수는 전체 단백질의 상호작용을 나타내기에는 많이 부족하기 때문에, 계산과학적인 방법을 통한 단백질 상호작용 예측 프로그램 개발에 큰 관심을 보이고 있다. 단백질 상호작용 예측은 크게 생물정보학적인 접근방법과 물리화학적 접근 방법을 통한 방법으로 구분할 수 있다. 생물정보학적인 접근방법에 따르면, 유사한 서열의 단백질은 유사한 상호작용 패턴을 지니고 있다. 물리화학적 접근방법에 따르면, 자연계에 존재하는 단백질 복합체는 에너지적으로 안정한 광역 최저점에 위치하고 있기 때문에, 광역최적화 방법을 통해 단백질 상호작용을 예측할 수 있다. 이 논문에서는 생물정보학 접근방법과 물리화학적 접근방법을 통해 새롭게 개발된 호모-올리고머 상호작용, 단백질-펩타이드 상호작용, 단백질-단백질 상호작용을 예측하는 방법에 대해 소개하고 있다. 생물정보학적인 접근방법과 물리화학적 접근방법 모두 단백질 상호작용 예측을 정확히 하는데 있어서 매우 큰 기여를 하였다.

주요어: 호모-올리고머 상호작용, 단백질-펩타이드 상호작용, 단백질-단백질 상호작용, 생물정보학, 물리화학, 광역최적화

학 번: 2010-20290