이학박사 학위논문

Protein Complex Structure Prediction by Template-Based and *Ab Initio* Docking

주형 기반 도킹과 *Ab Initio* 도킹을 이용한 단백질 복합체 구조 예측

2021 년 8 월

서울대학교 대학원

화학부 물리화학 전공

박 태 용

Protein Complex Structure Prediction by Template-Based and *Ab Initio* Docking

지도교수 석 차 옥

이 논문을 이학박사 학위논문으로 제출함

2021 년 6 월

서울대학교 대학원

화학부 물리화학 전공

박 태 용

박태용의 이학박사 학위논문을 인준함

2021 년 6 월

위원장 _____

부위원장 _____

위 원 _____

위 원 _____

위 원 _____

Abstract

Protein Complex Structure Prediction by Template-Based and *Ab Initio* Docking

Taeyong Park

Department of Chemistry

The Graduate School

Seoul National University

Protein-protein interactions play crucial roles in diverse biological processes, including various disease progressions. Atomistic structural details of protein-protein interactions that can be obtained from protein complex structures may provide vital information for the design of therapeutic agents. However, a large portion of protein complex structures is hard to be experimentally captured due to their weak and transient protein-protein interactions. Indeed, a limited fraction of protein-protein interactions happening in the human body has been experimentally determined. Computational protein complex structure prediction methods have been spotlighted for their roles in providing insights into protein-protein interactions in the absence of complete structural information by experiment. In this dissertation, three protein complex structure prediction methods are explained: GalaxyTongDock, GalaxyHeteromer, and GalaxyHomomer2. GalaxyTongDock performs *ab initio* docking for structure prediction of hetero- and homo-oligomers. GalaxyHeteromer and GalaxyHomomer2 predict heterodimer and homo-oligomer

structures, respectively, by template-based docking and *ab initio* docking depending on the template's availability. Lastly, examples of how these methods were utilized to predict protein complex structures in CASP and CAPRI, community-wide prediction experiments, are presented.

keywords: protein complex structure prediction, protein-protein docking, template-based docking, *ab initio* docking, CASP, CAPRI

Student Number: *2015-22613*

Table of Contents

List of Figures

List of Tables

1. Introduction

Protein-protein interactions play key roles in various biological processes such as cell division, maintenance of homeostasis, immunity, and various disease progressions[1-3]. Abnormal protein-protein interactions caused by gene mutations or environmental factors become the source of a wide range of diseases. Normal protein-protein interactions also can lead to diseases, for example, the PD-1-PD-L1 interaction in various cancers. Understanding the atomistic detail of protein-protein interactions is a prerequisite for identifying therapeutic molecules that regulate disease-related biological processes. Computational protein complex structure prediction methods have been a valuable tool for the understanding of protein-protein interactions due to the limited number of available protein complex structures obtained experimentally, especially for transient or weak protein complexes[4-7].

Protein complex structures are currently predicted using template-based or *ab initio* docking[8-14], depending on the availability of structural templates for the target complex in structure database, such as Protein Data Bank (PDB). Structural templates for a protein complex can be detected by exploiting sequence or structure similarities of consisting proteins to proteins in the database. When such similarity-based approaches are not reliable due to the lack of available structural templates, *ab initio* docking, based on the physical principles of protein binding, is used. *Ab initio* docking identifies the most stable binding pose in the conformational space of protein-protein complexes by conformational sampling and stability evaluation. The protein complexes can be divided into homo-oligomers, which are assemblies of identical proteins, and hetero-oligomers, which are assemblies of different

proteins. During the structure prediction of homo-oligomers, symmetricity may be additionally considered because homo-oligomers usually form symmetric conformations.

Here, we introduce three protein complex structure prediction methods: GalaxyTongDock[14], GalaxyHeteromer[15], and GalaxyHomomer2 (unpublished). GalaxyTongDock is an *ab initio* protein-protein docking method composed of GalaxyTongDock_A, GalaxyTongDock_C, and GalaxyTongDock_D, which performs asymmetric docking for heterodimer structure prediction, symmetric docking for C_n symmetric homo-oligomer structure prediction, and symmetric docking for D_n symmetric homo-oligomer structure prediction, respectively. Structures of proteins composing a protein complex should be provided as input for GalaxyTongDock. GalaxyHeteromer predicts structures of heterodimers from amino acid sequences or structures of two subunit proteins. Both template-based docking and *ab initio* docking are employed by automatically detecting the template's availability. When a subunit sequence is provided, GalaxyHeteromer utilizes GalaxyTBM[16] and GalaxyDBM (unpublished) for subunit structure prediction. GalaxyDBM employs inter-residue distance prediction by exploring the coevolution relationships among the homologous sequences via deep learning. GalaxyHomomer2, an upgraded version of GalaxyHomomer[10], predicts structures of homo-oligomers from an amino acid sequence or a monomer structure. As in GalaxyHeteromer, both template-based docking and *ab initio* docking are employed by automatically detecting the template's availability.

Utilizing the developed methods, we have been ranked 1st to 4th in multiple Critical Assessment of protein Structure Prediction (CASP)[17] and Critical Assessment of PRediction of Interactions (CAPRI)[18,19], which are community-wide prediction experiments for protein structure prediction and protein complex

structure prediction. Examples of how these methods were applied to predict protein complex structures in CASP and CAPRI are also presented.

2. GalaxyTongDock

GalaxyTongDock is an *ab initio* protein-protein docking method that performs rigid-body docking[20] just like ZDOCK[21], known to be one of the best-performing methods, but with improved energy parameters. The energy parameters were trained by iterative docking and parameter search so that more native-like structures are selected as top rankers. GalaxyTongDock performs asymmetric docking of two different proteins (GalaxyTongDock_A) and symmetric docking of homo-oligomeric proteins with C_n and D_n symmetries (GalaxyTongDock_C and GalaxyTongDock_D). Performance tests on an unbound docking benchmark set for asymmetric docking and a model docking benchmark set for symmetric docking showed that GalaxyTongDock is better or comparable to other state-of-the-art methods. In addition, experimental and evolutionary information on binding interfaces can be easily incorporated using interface and block options.

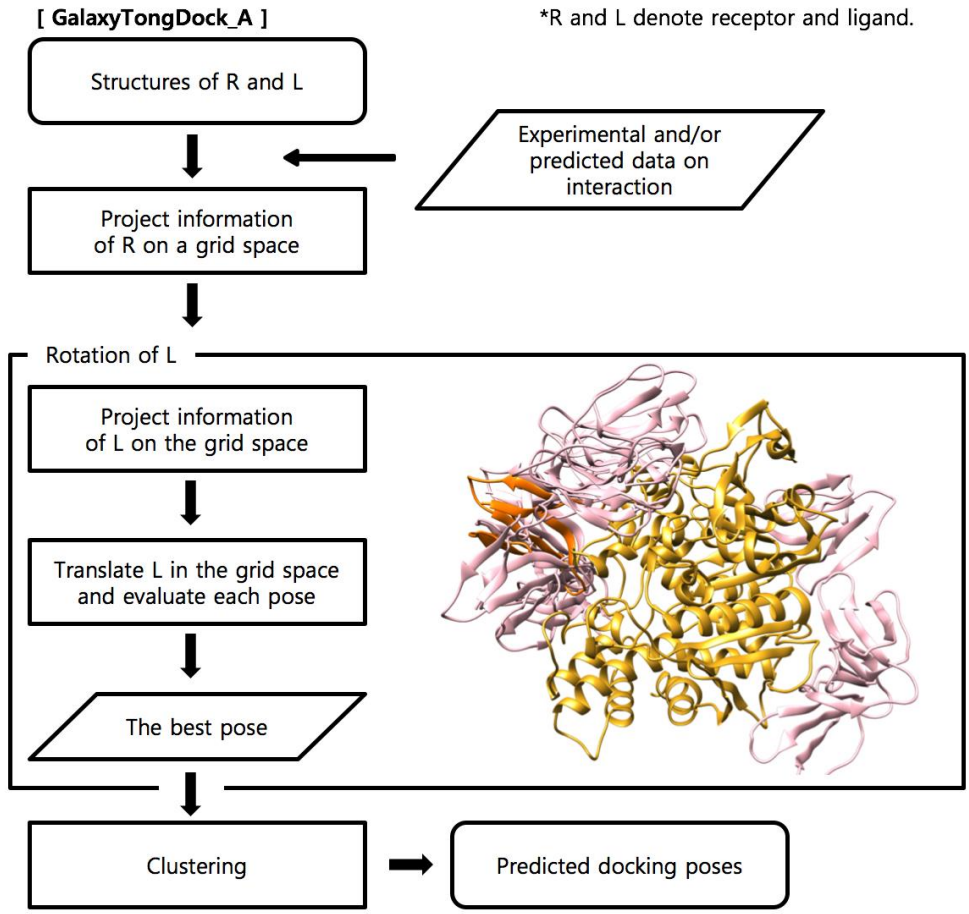
2.1. Methods

2.1.1. The prediction workflow of GalaxyTongDock_A

The prediction workflow of GalaxyTongDock_A is shown in **Figure 2.1**. Two proteins to be docked are conventionally referred to as a receptor and a ligand. GalaxyTongDock_C shares a similar workflow with that of GalaxyTongDock_A. The same proteins become the receptor and the ligand in GalaxyTongDock_C. Experimental and theoretically predicted data on interaction can be provided as interface and block options described in **Section 2.1.6**. In GalaxyTongDock_A, the

receptor structural information on a 3D continuous space is projected on a 3D lattice of grid size 1 Å. Then, translation of the ligand is sampled by FFT in the 3D lattice after rotating the Euler angles of the ligand in steps of 10° (or 6° in the case of symmetric docking). The best scoring docking pose from each rotated ligand undergoes clustering and model selection steps, described in **Section 2.1.5**.

Figure 2.1. The prediction workflow of GalaxyTongDock_A. The protein structures in the box represent an example of GalaxyTongDock_A output. The native receptor (yellow), native ligand (orange), and top 20 predicted ligands (pink) are shown.



2.1.2. Docking energy function: Functional form

GalaxyTongDock performs rigid-body docking utilizing fast Fourier transform (FFT)[22] with an energy function based on previously reported components. Using FFT, translation relative to a receptor of a pre-rotated ligand on a 3D grid space is performed much faster so that exhaustive sampling and evaluation of docking poses can be achieved. The GalaxyTongDock energy is composed of six elements as follows:

$$E_{\mathrm{GalaxyTongDock}} = E_{\mathrm{SC_{rep}}} + w_1 E_{\mathrm{SC_{att}}} + w_2 E_{\mathrm{Elec}} + w_3 E_{\mathrm{ACE}} + w_4 E_{\mathrm{IFACE}} + w_5 E_{\mathrm{consv}}$$

where the first five terms are from ZDOCK (repulsive and attractive parts of the shape complementarity score[23] $E_{\mathrm{SC_{rep}}}$ and $E_{\mathrm{SC_{att}}}$, Coulomb energy with distance-dependent dielectric constant[24] E_{Elec}, atomic contact energy[25] E_{ACE} of ZDOCK2.3.2[26], and interface atomic contact energy[23] E_{IFACE} of ZDOCK3.0.2[23]). The last term E_{consv} is the conservation score[27]. The energy components were implemented as described in detail below.

As in ZDOCK3.0.2[21], the energy value of the *i*th component, $E_i(\alpha, \beta, \gamma)$, which describes interaction energy between a fixed receptor structure and a pre-rotated ligand structure translated by (α, β, γ) from its initial position is expressed as a convolution $(R_i * L_i)(\alpha, \beta, \gamma)$ of a receptor grid $R_i(x, y, z)$ and a ligand grid $L_i(x, y, z)$ as follows:

$$E_i(\alpha, \beta, \gamma) = \sum_{x,y,z} R_i(x, y, z) L_i(x + \alpha, y + \beta, z + \gamma) \equiv (R_i * L_i)(\alpha, \beta, \gamma)$$

where (x, y, z) denotes a grid point in a 3-dimensional lattice, and definitions of the receptor and ligand grids, $R_i(x, y, z)$ and $L_i(x, y, z)$, depend on the energy

component. A grid size of 1 Å is used, which is slightly different from 1.2 Å of ZDOCK.

The energy components are calculated by FFT using the convolution theorem

$$R * L = F^{-1}\{F(R) \cdot F(L)\}$$

where F and F^{-1} refer to the Fourier and the inverse Fourier transformations.

(1) Repulsive part of the shape complementarity score $E_{\mathrm{SC_{rep}}}$

The repulsive energy component of the shape complementarity score is defined as

$$E_{\mathrm{SC_{rep}}}(\alpha,\beta,\gamma) = (R_{\mathrm{SC_{rep}}} * L_{\mathrm{SC_{rep}}})(\alpha,\beta,\gamma)$$

$$R_{\mathrm{SC_{rep}}}(x,y,z) = L_{\mathrm{SC_{rep}}}(x,y,z) = \begin{cases} 3.5 & \text{protein surface} \\ 3.5^2 & \text{protein core} \\ 0 & \text{otherwise} \end{cases}$$

Grid points (x, y, z) located within $0.8\times r_{\mathrm{vdW}}$, where r_{vdW} is van der Waals radius, from surface atoms are defined to be at the protein surface. Those atoms with the solvent accessible area, calculated by Naccess[28] with a probe radius of 1.4 Å, greater than 1 Å^2 are designated as surface atoms, and the rest as core atoms. Grid points located within $0.8\times r_{\mathrm{vdW}}$ from core atoms are defined to be in the protein core. The grid point value at the protein surface is set to be lower (3.5) than that in the core (3.5^2) to allow small clashes that may occur by rigid-body docking but could be relaxed by slight structure adjustments. The value 0.8 multiplied to r_{vdW} plays the same role. The same van der Waals radii as ZDOCK3.0.2 are used.

(2) Attractive part of the shape complementarity score $E_{\mathrm{SC_{att}}}$

This energy component is expressed in terms of two convolutions as follows:

$$E_{\mathrm{SC_{att}}}(\alpha,\beta,\gamma) = \frac{1}{2}(R_{\mathrm{SC_{att}}} * L_{\mathrm{atoms}} + R_{\mathrm{atoms}} * L_{\mathrm{SC_{att}}})(\alpha,\beta,\gamma)$$

$$R_{\mathrm{SC_{att}}}(x,y,z) = L_{\mathrm{SC_{att}}}(x,y,z) = -1.334n$$

$$R_{\mathrm{atoms}}(x,y,z) = L_{\mathrm{atoms}}(x,y,z) = m$$

where *n* is the number of atoms within 6 Å from the grid point (x,y,z), and *m* is the number of atoms for which (x,y,z) is the nearest grid point.

(3) Coulomb energy E_{Elec}

The Coulomb electrostatic interaction energy is calculated as

$$E_{\mathrm{Elec}}(\alpha,\beta,\gamma) = (R_{\mathrm{Elec\_pot}} * L_{\mathrm{charge}})(\alpha,\beta,\gamma)$$

$$R_{\mathrm{Elec\_pot}}(x,y,z) = \sum_i \frac{2500 q_i}{\varepsilon(r_i)\max(r_i, 2\ \text{Å})}$$

$$\varepsilon(r_i) = \begin{cases} 4 & r_i < 6 \\ 38r_i - 224 & 6 \le r_i < 8 \\ 80 & r_i \ge 8 \end{cases}$$

$$L_{\mathrm{charge}}(x,y,z) = \sum_{\mathrm{nearest}} q_i$$

where r_i is the distance of the grid point (x,y,z) from the *i*th receptor atom, r_i less than 2 Å is set to 2 Å to prevent the electric component from becoming too large upon clashes, 2500 is multiplied to balance the energy scale with other components, $\varepsilon(r_i)$ is the distance-dependent dielectric constant[24], and L_{charge} is

defined as a sum of the partial charges q_i of the atoms for which the nearest grid point is (x, y, z). The same partial charges as ZDOCK3.0.2 are used.

(4) Atomic contact energy E_{ACE}

This component is a sum of two convolutions:

$$E_{\mathrm{ACE}}(\alpha, \beta, \gamma) = \frac{1}{2}(R_{\mathrm{ACE}} * L_{\mathrm{atoms}} + R_{\mathrm{atoms}} * L_{\mathrm{ACE}})(\alpha, \beta, \gamma)$$

$$R_{\mathrm{ACE}}(x, y, z) = L_{\mathrm{ACE}}(x, y, z) = \sum_{r_i < 6\ \text{Å}} e_{\mathrm{type}(i)}$$

where $e_{\mathrm{type}(i)}$ is the average contact energy[25] for the atom type of the *i*th atom, and this is summed over atoms within 6 Å from the grid point (x, y, z).

(5) Interface atomic contact energy $E_{\mathbf{IFACE}}$

The IFACE (Interface Atomic Contact Energy) is expressed as a sum of convolutions for atom types *t* as follows:

$$E_{\mathrm{IFACE}}(\alpha, \beta, \gamma) = \sum_{t=1}^{12} \left(R_{\mathrm{IFACE},t} * L_{\mathrm{atoms},t}\right)(\alpha, \beta, \gamma)$$

$$R_{\mathrm{IFACE},t}(x, y, z) = \sum_{s=1}^{12} n_s e_{st}$$

$$L_{\mathrm{atoms},t}(x, y, z) = m_t$$

where n_s is the number of atoms of type *s* within 6 Å of the grid point (x, y, z), e_{st} is the interface atomic contact energy[23] between atom types *s* and *t*, m_t is the number of atoms of type *t* for which the nearest grid is (x, y, z). The atom types of IFACE are different from those of ACE.

(6) Conservation score $E_{\mathbf{consv}}$

The conservation score is defined as follows:

$$E_{\mathrm{consv}}(\alpha,\beta,\gamma) = \frac{1}{n_{\mathrm{surf}}}(R_{\mathrm{consv}} * L_{\mathrm{surf}} + R_{\mathrm{surf}} * L_{\mathrm{consv}})(\alpha,\beta,\gamma)$$

$$n_{\mathrm{surf}} = \left(R_{\mathrm{C}_\alpha} * L_{\mathrm{surf}} + R_{\mathrm{surf}} * L_{\mathrm{C}_\alpha}\right)(\alpha,\beta,\gamma)$$

$$R_{\mathrm{consv}}(x,y,z) = L_{\mathrm{consv}}(x,y,z) = \sum_{\text{nearest } l} E_{\mathrm{consv}}(l)$$

$$E_{\mathrm{consv}}(l) = \max\left[M_{l,aa(l)} - B_{aa(l),aa(l)}, 0\right]$$

$$R_{\mathrm{surf}}(x,y,z) = L_{\mathrm{surf}}(x,y,z) = \begin{cases} 1 & \text{if any } \mathrm{C}_\alpha \text{ atom exists within 15 Å from } (x,y,z) \\ 0 & \text{otherwise} \end{cases}$$

$$R_{\mathrm{C}_\alpha}(x,y,z) = L_{\mathrm{C}_\alpha}(x,y,z) = n_{\mathrm{C}_\alpha}$$

where $\sum_{\text{nearest } l}$ means a sum over amino acid residues at sequence position *l* for which (x, y, z) is the nearest grid point from its C_α position, $M_{l,aa(l)}$ is the self-substitution score in the position-specific substitution matrix generated from PSIBLAST[29] for the amino acid type $aa(l)$ at sequence position *l*, $B_{aa(l),aa(l)}$ is the diagonal element of BLOSUM62[30] for the amino acid type $aa(l)$, and n_{C_α} is the number of C_α atoms for which (x, y, z) is the nearest grid point. The conservation score is defined to be zero if the denominator n_{surf} is less than 9 to

prevent it becoming too large when there are only a small number of C_α-C_α contacts.

2.1.3. Docking energy function: Parameter training

The linear weights, w_1 to w_5, of the energy components were balanced through a training process. A training set for energy parameter optimization was constructed by selecting 120 asymmetric complexes randomly among the 196 protein complexes that belong to the "rigid-body" and the "medium difficulty" categories in the Docking benchmark 5[31]. These complexes have interface RMSD (IRMSD)[32] between the bound complex structure (receptor and ligand structures determined together by experiment) and the unbound complex structure (receptor and ligand structures determined separately by experiment and superposed onto the bound structure) less than 2.2 Å. The training set was again randomly split into 80 complexes (Set 1) and 40 complexes (Set 2). Set 1 was used as a training set and Set 2 as an independent validation set to prevent overtraining. The remaining 76 complexes were used as a performance test set. The full list of the targets in each set is provided in **Tables 2.1 and 2.2**. For all training and test set targets, unbound docking was performed with two unbound protein structures.

Table 2.1. List of PDB IDs for the targets in Set 1 (80 targets) and Set 2 (40 targets), used for energy parameter training.

Set 1									
1A2K	**1AHW**	**1AZS**	**1D6R**	**1DQJ**	**1EZU**	**1F51**	**1FFW**	**1GHQ**	**1GLA**
1HIA	**1IB1**	**1IQD**	**1J2J**	**1JPS**	**1JTD**	**1JWH**	**1K4C**	**1K5D**	**1KAC**
1KTZ	**1KXP**	**1LFD**	**1M27**	**1ML0**	**1MQ8**	**1NW9**	**1OC0**	**1OFU**	**1OYV**
1PPE	**1QFW**	**1R6Q**	**1RLB**	**1RV6**	**1T6B**	**1UDI**	**1WDW**	**1WEJ**	**1XQS**
1Z0K	**1Z5Y**	**2A1A**	**2A5T**	**2A9K**	**2BTF**	**2FJU**	**2G77**	**2GAF**	**2H7V**
2HQS	**2HRK**	**2JEL**	**2NZ8**	**2O8V**	**2OZA**	**2SIC**	**2UUY**	**2VXT**	**2W9E**
2X9A	**3BIW**	**3BX7**	**3CPH**	**3EOA**	**3H2V**	**3LVK**	**3S9D**	**3SZK**	**4CPA**
4DN4	**4FZA**	**4G6M**	**4GXU**	**4HX3**	**4JCV**	**4LW4**	**7CEI**	**BAAD**	**BOYV**
Set 2									
1B6C	**1BUH**	**1CLV**	**1E6J**	**1E96**	**1EAW**	**1EWY**	**1FCC**	**1FQJ**	**1GPW**
1H9D	**1HE1**	**1JIW**	**1JTG**	**1KLU**	**1M10**	**1MAH**	**1N2C**	**1NSN**	**1OPH**
1QA9	**1R0R**	**1SBB**	**1TMQ**	**1US7**	**1VFB**	**1WQ1**	**1ZHI**	**1ZM4**	**2I25**
2VDB	**3D5S**	**3EO1**	**3G6D**	**3K75**	**3L5W**	**3P57**	**3VLB**	**4M76**	**CP57**

Table 2.2. List of PDB IDs for the targets in the asymmetric docking test set (76 targets).

1AK4	**1AKJ**	**1AVX**	**1AY7**	**1BJ1**	**1BVK**	**1BVN**	**1CGI**	**1DFJ**	**1E6E**
1EFN	**1EXB**	**1F34**	**1FC2**	**1FLE**	**1FSK**	**1GCQ**	**1GL1**	**1GP2**	**1GRN**
1GXD	**1HCF**	**1HE8**	**1I2M**	**1I4D**	**1I9R**	**1IJK**	**1K74**	**1KKL**	**1KXQ**
1MLC	**1NCA**	**1PVH**	**1S1Q**	**1SYX**	**1XD3**	**1XU1**	**1YVB**	**1ZHH**	**2ABZ**
2AJF	**2AYO**	**2B42**	**2B4J**	**2CFH**	**2FD6**	**2GTP**	**2HLE**	**2J0T**	**2MTA**
2OOB	**2OOR**	**2OUL**	**2PCC**	**2SNI**	**2VIS**	**2YVJ**	**2Z0E**	**3A4S**	**3AAA**
3BP8	**3DAW**	**3HI6**	**3HMX**	**3MXW**	**3PC8**	**3R9A**	**3RVW**	**3SGQ**	**3V6Z**
4FQI	**4G6J**	**4H03**	**4IZ7**	**9QFW**	**BP57**				

The docking poses for the training set complexes were used to find the weight parameters that maximize the objective function SR200. SR200, Success Rate for top 200, evaluates docking performance of an energy parameter set during parameter training. Top 200 was chosen because subsequent clustering after conformational sampling by FFT can usually reduce 200 predictions to around 50, which is an affordable number for observation or a later structural refinement. SR200 is defined as

$$\mathrm{SR200} = \frac{n_{\text{top 200 success}}}{n_{\text{target}}} \times 100\ (\%)$$

where $n_{\text{top 200 success}}$ is the number of targets for which at least one successful prediction is obtained within the top 200 predictions, and n_{target} is the number of targets in the set. A successful prediction is defined as a predicted conformation obtained by conformational sampling for which ligand RMSD (LRMSD)[32] from the crystal structure is less than 10 Å. Ligand RMSD was calculated after superposing the receptor structure of the docking pose to the bound receptor structure. The criterion of LRMSD < 10 Å was used as an accuracy range that can be obtained by rigid-body docking and may be improved by flexible refinement docking methods such as RosettaDock[33] or GalaxyRefineComplex[34]. Because docking poses depend on the energy parameters used to generate them, parameter optimization involved an iterative procedure that alternated parameter search and docking pose generation.

Four initial weight parameter sets, $(w_1, w_2, w_3, w_4, w_5) =$ (0.34, 0.51, 0.29, 0.60, 250), (1.22, 2.04, 0.81, 1.92, 200), (1.00, 1.00, 1.00, 1.00, 250), and (0.60, 00.60, 0.60, 0.60, 400), were chosen considering the scales of the energy components. For each initial parameter set, docking poses were generated and a grid search in the parameter space was performed.

At each parameter training round that was performed with a fixed set of docking poses, a grid search method in the 5D parameter space $\{w_1\}$ $(i = 1, 2, 3, 4, 5)$ was employed with SR200 for Set 1 as the objective function. Since change in the energy landscape by parameter change was ignored by parameter search at fixed conformations for computational efficiency, special care was taken so that parameter search does not become too broad. A large parameter change would involve a large change in the energy landscape, so the optimal parameters found at fixed conformations would not be optimal anymore in the changing energy landscape. The grid size and the search range were therefore defined in terms of a factor m_i multiplied to the parameter $w_{i,\mathrm{prev}}$ of the previous round, i.e. $w_i = m_i w_{i,\mathrm{prev}}$, and the m_i values around one were searched at each round, as summarized in **Table 2.3**. The parameter search range at each round was not set too broad because the gap between the actual performance evaluated after conformational sampling with the changed parameter set and the expected performance evaluated with fixed conformations can be huge for a large parameter change. The two parameter sets with the highest SR200 on Set 2, the validation set, among the five sets with the highest SR200 on Set 1, the main training set, were selected for the next round. Parameter sets that showed high performance on the validation set were chosen to avoid overtraining on the main training set. Parameter search was then performed again with docking poses generated with the new parameters. After four rounds of iteration, the parameter set {0.36, 0.36, 0.12, 0.48, 60}, which showed the highest SR200 on both Set 1 and Set 2 among the 124 (4 + 8 + 16 + 32 + 64) parameter sets, was finally chosen as the optimal weight set. Since the magnitudes of the parameters do not directly represent contributions of the corresponding components, a separate analysis of contributions of the components to the overall energy distribution is provided in **Table 2.4**.

Table 2.3. Grid sizes and search ranges of the parameter training rounds at fixed conformations, defined in terms of a factor $\boldsymbol{m_i}$ which is multiplied by each weight parameter $\boldsymbol{w_i}$ $(\boldsymbol{i = 1, 2, 3, 4, 5})$ that was selected in the previous training round.

Energy component (*i*)	Range of $\boldsymbol{m_i}$ (Grid size in $\boldsymbol{m_i}$)			
	Round 1	Round 2	Round 3	Round 4
SC_{att} (1)	0.1 ~ 2 (0.1)	0.55 ~ 1.5 (0.05)	0.82 ~ 1.2 (0.02)	0.82 ~ 1.2 (0.02)
Elec (2)	0.1 ~ 2 (0.1)	0.55 ~ 1.5 (0.05)	0.82 ~ 1.2 (0.02)	0.82 ~ 1.2 (0.02)
ACE (3)	0.2 ~ 2 (0.2)	0.6 ~ 1.5 (0.1)	0.8 ~ 1.25 (0.05)	0.8 ~ 1.25 (0.05)
IFACE (4)	0.2 ~ 2 (0.2)	0.6 ~ 1.5 (0.1)	0.8 ~ 1.25 (0.05)	0.8 ~ 1.25 (0.05)
Consv (5)	0.2 ~ 2 (0.2)	0.6 ~ 1.5 (0.1)	0.8 ~ 1.25 (0.05)	0.8 ~ 1.25 (0.05)

Table 2.4. The contribution of each energy component defined as the standard deviation of the energy values for top 200 docking poses before clustering, multiplied by the weight factor, $\{\boldsymbol{w_1}, \boldsymbol{w_2}, \boldsymbol{w_3}, \boldsymbol{w_4}, \boldsymbol{w_5}\} = \{\mathbf{0.36}, \mathbf{0.36}, \mathbf{0.12}, \mathbf{0.48}, \mathbf{0.60}\}$, and averaged over 120 complexes in the training set and normalized over the energy components.

Energy Component	Contribution
$E_{\mathrm{SC_{rep}}} + w_1 E_{\mathrm{SC_{att}}}$	0.3310
$w_2 E_{\mathrm{Elec}}$	0.3096
$w_3 E_{\mathrm{ACE}}$	0.0445
$w_4 E_{\mathrm{IFACE}}$	0.2935
$w_5 E_{\mathrm{consv}}$	0.0214

2.1.4. Symmetric and asymmetric docking protocols

In GalaxyTongDock, symmetric docking methods that generate protein oligomers of C_n and D_n symmetry are also available along with asymmetric docking. A protein oligomer of C_n symmetry belongs to the cyclic C_n point group, which has a principal axis of n-fold rotational symmetry C_n but no C_2 axes perpendicular to the principal axis. A protein oligomer of D_n symmetry belongs to the dihedral D_n point group, which has a C_n principal axis and nC_2 axes perpendicular to the principal axis. GalaxyTongDock performs docking with the same scoring function for symmetric (GalaxyTongDock_C and GalaxyTongDock_D) and asymmetric docking (GalaxyTongDock_A). A separate scoring function was not derived because a large unbound set was not available for symmetric complexes.

GalaxyTongDock_C samples docking poses of two-neighboring monomers by 2D FFT. GalaxyTongDock_D generates poses of D_n symmetry by 1D FFT of the top 30 poses of C_n symmetry prepared with GalaxyTongDock_C. The conservation score is excluded in 1D FFT of GalaxyTongDock_D for computational efficiency. All of GalaxyTongDock_A, GalaxyTongDock_C, and GalaxyTongDock_D employ the model selection method described in **Section 2.1.5**.

2.1.5. Model selection

Before model selection, the top poses (maximum of 1000 and 800 poses for asymmetric and symmetric docking, respectively) are clustered to remove structural redundancy. A clustering radius of $\sqrt[3]{N}$, where N is the number of amino acid residues in the complex, is used considering the dependency of RMSD

on the complex size. After ranking the clusters, the lowest-energy cluster representatives are reported. In asymmetric docking, clusters are ranked by cluster size, considering their potential relationship with conformational entropy. In symmetric docking, clusters are ranked by energy because ranking by cluster size is not practical due to higher structural redundancy resulting from decreased degrees of freedom. When the interface option (explained in **Section 2.1.6**) is used for asymmetric docking, clusters are also ranked by energy as higher redundancy emerges due to the additional restraints.

2.1.6. Prior information in the form of interface and block options

Information on the binding interface can be provided as input in the form of interface and block options. If interface residues are designated, docking poses that have those residues within 8 Å from the partner protein are prioritized during docking. If non-interface regions are designated by the block option, docking poses with those regions at the interface are strongly disfavored by a large penalty (positive energy) during docking. The GalaxyTongDock score is defined to have the same absolute value as the GalaxyTongDock energy but with the opposite sign to make the better poses have the higher scores.

2.2. Performance of GalaxyTongDock

Docking performance was evaluated in terms of success rate, which is defined as the percentage of targets for which at least one model has ligand RMSD < 10 Å when a receptor is superposed to the bound structure. In the case of models from symmetric docking, one subunit was treated as the ligand, and the others were treated as the receptor to calculate the ligand RMSD. Success rates of GalaxyTongDock for the top 1, top 10, and top 50 models were compared with other top-performing *ab initio* docking methods: ZDOCK3.0.2 for asymmetric docking and M-ZDOCK[13] and SAM[35] for symmetric docking. The compared methods were run with default options. The results are presented in **Table 2.5**. GalaxyTongDock_A showed increased success rates than ZDOCK3.0.2 for top 10 and top 50 models on both training and test sets. Here, the unbound receptor and ligand structures were randomly rotated before docking to remove the dependency on the initial orientation.

In the case of symmetric docking, unbound monomer structures of homo-oligomers resolved experimentally were not available in most cases; hence, a large-scale unbound docking test using experimentally resolved unbound structures was not possible. Unbound docking with model structures was performed instead. Model docking would be also more relevant in actual applications. GalaxyTBM was used for template-based modeling using templates with sequence identity < 40%. The inaccurately modeled loops or termini detected by GalaxyTBM were deleted before docking because they often interrupt docking. The PISA benchmark set[36] was used to evaluate symmetric docking performance. Among the 142 homo-oligomer proteins in the PISA benchmark set, those predicted to have multiple subunits for modeling by GalaxyTBM were excluded. Final test sets include 83 and

29 complexes with C_n and D_n symmetries, respectively. The lists of the symmetric docking targets are provided in **Tables 2.6 and 2.7**.

The performance of GalaxyTongDock_C was higher than that of M-ZDOCK and SAM, except that SAM showed a higher success rate for the top 1 model (by 1 target). GalaxyTongDock_D showed higher performance than SAM when the top 1 and top 10 models are considered.

A two-sample z-test analyzed the statistical significance of the performance comparison presented in **Table 2.5**. The *p*-values against the null hypothesis that GalaxyTongDock performs equal to or worse than each compared method for each top 1, top 10, and top 50 predictions are presented in **Table 2.8**. The training set and test set of the asymmetric docking were combined for the analysis. The table shows that the performance of GalaxyTongDock is statistically better than other methods with *p*-values 0.05~0.14 when the top 10 and top 50 predictions are considered, except in the comparison of GalaxyTongDock_D with SAM on top 50 predictions.

ZDOCK3.0.2 and M-ZDOCK do not perform clustering after sampling. Performance after applying the same clustering method to ZDOCK3.0.2 and M-ZDOCK is shown in **Table 2.9**. GalaxyTongDock still shows higher performance. The performance of bound docking is also compared in **Table 2.10**, although bound docking is not of practical importance.

Table 2.5. Success rates of GalaxyTongDock and other methods for the top 1, top 10, and top 50 models in the cases of asymmetric and symmetric docking, where the success rate is defined as the percentage of the targets for which at least one model is within ligand RMSD < 10 Å from the bound structure.

			Top 1 (%)	Top 10 (%)	Top 50 (%)
Asymmetric docking	Test set (76 targets)	GalaxyTongDock_A	17.1	32.9	48.7
		ZDOCK3.0.2	9.2	31.6	42.1
	Training set (120 targets)	GalaxyTongDock_A	10.0	34.2	55.8
		ZDOCK3.0.2	14.2	25.8	47.5
Symmetric docking	C_n set (83 targets)	GalaxyTongDock_C	10.8	36.1	54.2
		M-ZDOCK	9.6	24.1	36.1
		SAM	12.1	26.5	45.8
	D_n set (29 targets)	GalaxyTongDock_D	10.3	27.6	41.4
		SAM	3.5	10.3	41.4

Table 2.6. List of PDB IDs for the targets in the C_n symmetric docking test set (83 targets).

1A3C	**1AA0**	**1AD3**	**1AF5**	**1AJS**	**1ALK**	**1AMK**	**1AQ6**	**1AUO**	**1B77**
1BAM	**1BSR**	**1BUO**	**1CA4**	**1CB0**	**1CE0**	**1CHM**	**1CJD**	**1CMB**	**1CP2**
1CSH	**1CUK**	**1CZJ**	**1DAA**	**1DPT**	**1E2A**	**1FGJ**	**1FIP**	**1FRO**	**1GVP**
1HJR	**1HSS**	**1ICW**	**1IMB**	**1ISA**	**1ISO**	**1JHG**	**1JSG**	**1KBA**	**1KPF**
1LYN	**1MJL**	**1MKA**	**1MOQ**	**1NHP**	**1NIF**	**1NKS**	**1NOX**	**1NSY**	**1OPY**
1OTP	**1PGT**	**1PPR**	**1PRE**	**1PUC**	**1QLM**	**1RLA**	**1RPO**	**1SES**	**1SLT**
1SMN	**1SMT**	**1TOX**	**1TRK**	**1TYS**	**1UBY**	**1UTG**	**1XSO**	**2CCY**	**2CHS**
2PII	**2RSP**	**2STD**	**2TCT**	**2TGI**	**3CLA**	**3EOJ**[a]	**3GRS**	**3SDH**	**3TDT**
4KBP	**5TMP**	**9WGA**							

[a] 4BCL was substituted to 3EOJ which has a higher resolution.

Table 2.7. List of PDB IDs for the targets in the D_n symmetric docking test set (29 targets).

1A0L	**1A2Z**	**1A3G**	**1A4E**	**1ADO**	**1BUC**	**1BVQ**	**1CG2**	**1CS1**	**1DCI**
1DCO	**1DXE**	**1ETA**	**1EUH**	**1FTR**	**1GP1**	**1GSH**	**1ITH**	**1MPY**	**1MXB**
1NDC	**1NHK**	**1UOX**	**1XVA**	**2CEV**	**2EIP**	**2IZG**	**4PGA**	**5PGM**	

Table 2.8. The *p*-values obtained by a two-sample z-test against the null hypothesis that GalaxyTongDock performs equal to or worse than each compared method for each top 1, top 10, and top 50 predictions.

Compared methods	*p*-values		
	Top 1	Top 10	Top 50
GalaxyTongDock_A vs ZDOCK3.0.2	0.4404	0.1151	0.0643
GalaxyTongDock_C vs M-ZDOCK	0.3974	0.0455	0.0096
GalaxyTongDock_C vs SAM	0.5948	0.0901	0.1379
GalaxyTongDock_D vs SAM	0.1492	0.0475	0.5000

Table 2.9. Success rates for the top 1, top 10, and top 50 models of asymmetric and symmetric docking by ZDOCK3.0.2 and M-ZDOCK with clustering.

			Top 1 (%)	Top 10 (%)	Top 50 (%)
Asymmetric docking	Test set (76 targets)	ZDOCK3.0.2 with clustering	5.3	26.3	40.8
	Training set (120 targets)	ZDOCK3.0.2 with clustering	11.7	30.8	47.5
Symmetric docking	C_n set (83 targets)	M-ZDOCK with clustering	9.6	27.7	42.2

Table 2.10. Success rates for the top 1, top 10, and top 50 models of asymmetric and symmetric "bound" docking by GalaxyTongDock and those by other methods.

			Top 1 (%)	Top 10 (%)	Top 50 (%)
Asymmetric docking	Docking benchmark 5 (230 targets)	GalaxyTongDock_A	49.6	78.3	89.6
		ZDOCK3.0.2	54.8	75.7	81.7
		ZDOCK3.0.2 with clustering	19.6	53.5	77.0
Symmetric docking	C_n set (83 targets)	GalaxyTongDock_C	95.2	98.8	100.0
		M-ZDOCK	86.8	95.2	100.0
		M-ZDOCK with clustering	86.8	97.6	100.0
		SAM	74.7	86.8	95.2
	D_n set (29 targets)	GalaxyTongDock_D	44.8	72.4	82.8
		SAM	58.6	79.3	86.2

3. GalaxyHeteromer

GalaxyHeteromer is a method that predicts protein heterodimer structures from two subunit protein sequences or structures. When subunit structures are unavailable, they are predicted by template- or distance-prediction-based modeling methods. Heterodimer structure can be predicted by both template-based and *ab initio* docking, depending on the template's availability. Structural templates are detected from the protein structure database based on both the sequence and structure similarities. The templates for heterodimers may be selected from monomers and homo-oligomers, as well as from hetero-oligomers, owing to the evolutionary relationships of heterodimers with domains of monomers or subunits of homo-oligomers. In addition, the method employs one of the best *ab initio* docking methods when heterodimer templates are unavailable. The multiple heterodimer structure models and the associated scores provided by the method can be further examined to test or develop functional hypotheses or to design new functional molecules.

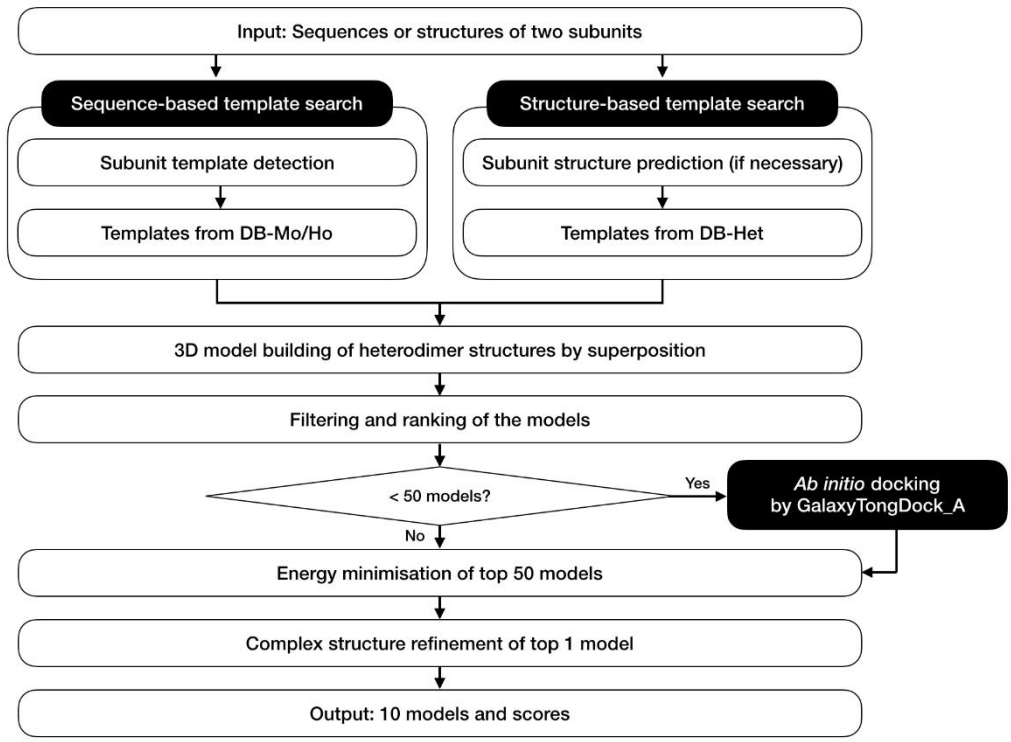
3.1. Methods

3.1.1. The overall pipeline of GalaxyHeteromer

The prediction pipeline of GalaxyHeteromer for predicting heterodimer structure is shown in **Figure 3.1**. In GalaxyHeteromer, template-based docking is performed by detecting templates for heterodimer structure building based on subunit sequence similarities (sequence-based template search) and subunit structure

similarities (structure-based template search), as described in detail below. If subunit structures are not provided as input, they are predicted from subunit sequences using a recently developed protein structure prediction method explained below. Then, 3D models for heterodimer structures are generated by superposing the subunit structures on the template structures. The models are filtered based on physical criteria, such as steric clashes, inter-subunit contacts, and interface area. After removing redundancy (of TM-score[37] > 0.8) among the heterodimer models, the models are ranked according to a template score, which consists of subunit and interface structure similarities measured in TM-score to the template structures. If < 50 models are left, *ab initio* docking is performed using GalaxyTongDock_A to generate more models so that a total of 50 models can be obtained. The best scoring model is further refined by using GalaxyLoop[38] and GalaxyRefineComplex.

Figure 3.1. GalaxyHeteromer pipeline for protein heterodimer structure prediction.



3.1.2. Subunit structure prediction

When only subunit sequences are provided as input, subunit structures are predicted from sequences by the protein structure prediction method of the GALAXY group who participated in CASP14 (2020) as the Seok-server. The protein structure prediction method selects a model through a random forest classifier from the models that were predicted by the template-based structure prediction method, GalaxyTBM, and from those that were predicted by a distance-prediction-based structure prediction method, GalaxyDBM. The accuracy of this method is comparable to that of AlphaFold[39] on CASP13 targets (https://predictioncenter.org/casp13/), in terms of the CASP measure GDT-TS[40] (average GDT-TS = 62.3, whereas that of AlphaFold is 62.9). The selected model is used for template search and heterodimer modeling.

GalaxyDBM predicts the probability distributions over distances between C_β atoms (C_α atom for GLY) of different residues using a deep residual convolutional neural network based on MSA-based features, including sequence profile and raw co-evolutionary coupling features from CCMPred[41], following AlphaFold. Thereafter, 3D backbone structures are predicted by the global optimization method, which is known as conformational space annealing[42], maximizing the likelihood of probability distributions and satisfying local stereochemistry controlled by GALAXY energy function[38,43]. The predicted structures are then refined by GalaxyRefine[44] for optimizing side-chain conformations.

3.1.3. Sequence- and structure-based template search

The sequence-based template search is performed on the database of monomer and homo-oligomer proteins, named DB-Mo/Ho, as shown in **Figure 3.1**. DB-Mo/Ho is the protein structure database of monomers and homo-oligomers, which has a maximum mutual sequence identity of 70%. Subunit templates are first detected by HHsearch[45]. Proteins in DB-Mo/Ho with high sequence similarity (in terms of GalaxyTBM template score[16] within the top 200) and high structure similarity (TM-score > 0.4) to both subunits in different parts of the same protein (e.g., different domains of a monomer or different subunits of a homo-oligomer) are selected as templates for building the heterodimer structure. Proteins with interface structure similarity of less than TM-score < 0.4 are discarded for homo-oligomer templates. The interface region for interface structure similarity comparison is defined as residues whose C_α atoms are placed closer than 20 Å from any of the other chain's C_α atoms.

The structure-based template search is performed on the database of heterodimers, named DB-Het, as shown in **Figure 3.1**. DB-Het was prepared by collecting non-redundant heterodimer structures from protein complex structures that have resolution better than 4.0 Å and consist of more than two distinct proteins in PDB. The detailed procedures constructing DB-Het are as follows:

1) Structures satisfying following conditions from PDB were downloaded; a) Resolution < 4.0 Å, b) Experimental method = X-RAY or EM, c) Number of Entities (Protein) > 1.

2) The structures were disassembled into dimer-wise structures.

3) Dimers having less than 5 C_α atoms within 10 Å from C_α atoms of another chain were excluded.

4) Monomers in dimers were clustered using CD-HIT[46] with a sequence identity cutoff of 70%.

5) Dimers that consist of monomers in the same clusters were clustered if TM-score between corresponding chains is larger than 0.8 and TM-score between interface regions is larger than 0.8. The dimer with the best resolution in a cluster became the representative dimer of the cluster.

6) DB-Het consisted of 45,267 representative dimers from step 5 (Date: 20201215).

From DB-Het, structure-based templates are detected by searching heterodimers with subunits with high structure similarities (TM-score > 0.4) to the corresponding subunit. Proteins with interface structure similarity less than TM-score < 0.4 are discarded.

3.1.4. Heterodimer modeling

Heterodimer modeling is performed by superposing the subunit structures provided as input or modeled by the modeling method described in **Section 3.1.2** on the templates searched by the method described in **Section 3.1.3**. Each subunit structure is superposed on the interface region of the corresponding subunit (or domain) of a selected template, resulting in a heterodimer model. Heterodimer models having more than 15 steric clashes or less than five contacts or an interface area less than 100 $Å^2$ are excluded. The steric clashes and contacts are defined by

the number of C_α atoms within 4 Å from another chain and the number of C_α atoms within 15 Å from another chain. The remaining heterodimer models are ranked by a template score that is defined as follows:

Template score for heterodimer and homo-oligomer templates

= [2 × (sum of subunit TM-scores between target and template for receptor and ligand) + 5 × (sum of interface TM-scores between target and template for receptor and ligand)] / 14

Template score for monomer templates

= (sum of subunit TM-scores between target and template for receptor and ligand) / 2

The template score for monomer templates is defined differently because the interface region of monomer templates cannot be clearly defined. The heterodimer models are then clustered by TM-score > 0.8 criteria. If < 50 models are left, GalaxyTongDock_A, described in **Section 2**, is used to generate more models so that a total of 50 models can be obtained. After energy minimization, the best scoring model is further refined by re-modeling interfacial loop structures which are detected as inaccurate by the loop modeling method GalaxyLoop and relaxed by repetitive side-chain perturbations and molecular dynamics simulations using the complex structure refinement method GalaxyRefineComplex.

3.2. Performance of GalaxyHeteromer

The protein complex structure prediction method, which contains GalaxyHeteromer and GalaxyHomomer2, described in **Section 4**, participated in the assembly category of CASP14 and CASP14-CAPRI challenges as a group named Seok, and ranked as fourth and first, respectively.

The performance of GalaxyHeteromer was compared to that of the *ab initio* docking method GalaxyTongDock_A on a test set of 143 heterodimers of the Docking benchmark 5, to evaluate the combined effect of template-based and *ab initio* docking compared to *ab initio* docking alone. The same monomer models generated by GalaxyHeteromer were used as input subunit structures for *ab initio* docking. To simulate a rather difficult prediction case, the subunit templates with sequence identities > 70% were excluded for monomer modeling, and the protein templates with sequence identities of any subunits > 70% to the corresponding subunits of the test proteins were excluded for heterodimer modeling. The generated models were assessed by CAPRI criterion[32] that classifies the accuracy of protein complex models as follows:

Acceptable: $F_{nat} > 0.1 \; and \; (5\text{Å} < LRSMD \leq 10\text{Å} \; or \; 2\text{Å} < IRMSD \leq 4\text{Å})$

Medium: $F_{nat} > 0.3 \; and \; (1\text{Å} < LRSMD \leq 5\text{Å} \; or \; 1\text{Å} < IRMSD \leq 2\text{Å})$

High: $F_{nat} > 0.5 \; and \; (LRSMD \leq 1\text{Å} \; or \; IRMSD \leq 1\text{Å})$

As shown in **Table 3.1**, GalaxyHeteromer and GalaxyTongDock_A generated models with better than acceptable accuracy in 30% and 5% of cases, respectively, as the top 1, and in 50% and 34% of cases, respectively, within the top 50.

Next, the performance of GalaxyHeteromer was compared to that of HDOCK[9], which is one of the best methods, on the 54 heterodimers used previously for benchmarking HDOCK[9]. The protein templates with sequence identities to the target complex greater than 30% were excluded, and unbound subunit structures were used as input. As shown in **Table 3.3**, GalaxyHeteromer outperformed HDOCK except for the case of the top 1 prediction. Top N (N=1, 5, 10, and 50) success rates (percentage of the cases in which models better than acceptable qualities obtained within the N models) are 33.3%, 53.7%, 55.6%, and 68.5%, respectively, for GalaxyHeteromer, whereas those for HDOCK are 38.9%, 40.7%, 44.4%, and 59.3%, respectively.

Table 3.1. Performance comparison of GalaxyHeteromer, which combines template-based and *ab initio* docking, with that of GalaxyTongDock_A, which employs *ab initio* docking, in terms of CAPRI criterion of model accuracy on a test set of 143 protein heterodimers.

% of the cases with medium/acceptable accuracy models within top N		
N	**GalaxyHeteromer**	**GalaxyTongDock_A**
1	13.3/30.1	1.4/4.9
5	18.2/39.2	5.6/13.3
10	19.6/41.3	7.0/16.8
50	22.4/49.7	9.8/34.3

Table 3.2. Performance comparison of GalaxyHeteromer with that of HDOCK in terms of CAPRI criterion on a test set of 54 protein heterodimers.

% of the cases with acceptable accuracy models within top N		
N	**GalaxyHeteromer**	**HDOCK**
1	33.3	38.9
5	53.7	40.7
10	55.6	44.4
50	68.5	59.3

GalaxyHeteromer puts more emphasis on providing multiple alternative solutions for possible complex structures by exploring multiple templates compared to HDOCK. The provided multiple models may be combined with separate experimental information to select more feasible complex structures.

In CASP13, our automatic prediction server submitted no acceptable models on the heterodimer targets of CASP13 (H0957, H0968, H0974, H0980, and H0986) in the top 5 predictions. On the other hand, GalaxyHeteromer made acceptable models on three targets, H0968, H0974, and H0986, in the top 5 predictions. Here, for a fair comparison, DB-Het reconstructed with structures released before 20180425, which is before CASP13, was used in GalaxyHeteromer. An example of a successful prediction for H0986 is shown in **Figure 3.2**. GalaxyHeteromer made a model having medium accuracy at sixth prediction for H0974, shown in **Figure 3.3**. This model was generated using a homo-oligomer template (PDB ID: 1Y7Y), which shows the importance of utilizing the evolutionary relationship of heterodimers to different forms of quaternary complexes, i.e., homo-oligomers and monomers.

Figure 3.2. An example of a successful prediction on H0986. **Left)** Each subunit structure modeled by GalaxyHeteromer (in sky-blue and yellow) is superposed on the corresponding subunit structure in the native structure of H0986 (in blue and red). The two interfacial loops which were inaccurately modeled are in the circles. **Right)** The 5th prediction model of GalaxyHeteromer (in sky-blue and yellow), which has an acceptable accuracy [F_{nat}^{32}: 0.232, IRMSD: 4.89 Å, LRMSD: 8.45 Å], is superposed on the native structure (in blue and red). Although there were inaccurately modeled loops in the interface region of subunit structures, GalaxyHeteromer successfully generated an acceptable accuracy model by template-based docking.

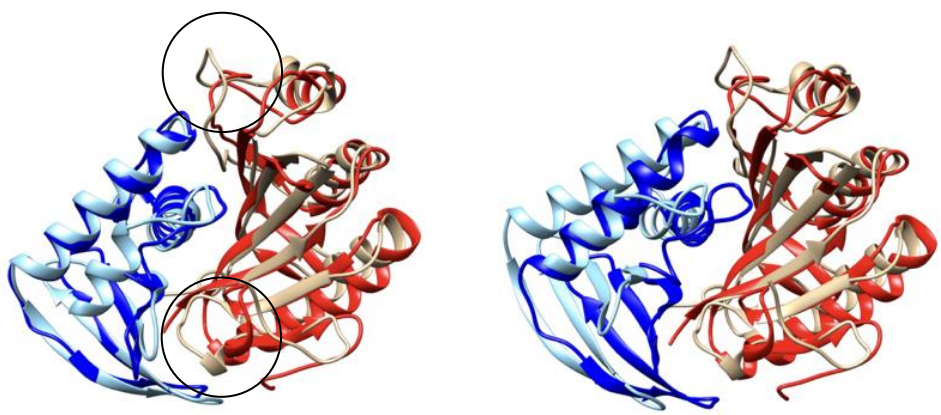
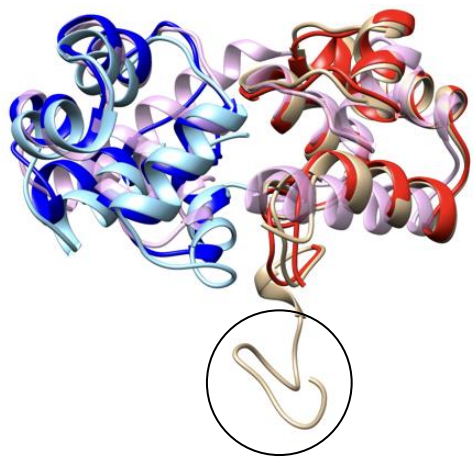


Figure 3.3. Model 6 of GalaxyHeteromer (in sky-blue and yellow) and the template, a homodimer protein (PDB ID: 1Y7Y), that was used to generate the model (in pink) are superposed on the native structure of H0974 (in blue and red). The model generated based on the template has medium accuracy [F_{nat}: 0.667, IRMSD: 2.60 Å, LRMSD: 3.33 Å]. The histidine tag of the model, not captured in the crystal structure, is in the circle.



4. GalaxyHomomer2

GalaxyHomomer2, an upgraded version of GalaxyHomomer[10], predicts a homo-oligomer structure from a monomer sequence or structure composing the homo-oligomer. When a monomer structure is unavailable, the monomer structure is predicted by the subunit structure prediction method, described in **Section 3.1.2**. The homo-oligomer structure can be predicted by both template-based and *ab initio* docking, depending on the template's availability. Structural templates are detected from the protein structure database based on both the sequence and structure similarities. The method employs one of the best *ab initio* docking methods when homo-oligomer templates are unavailable.

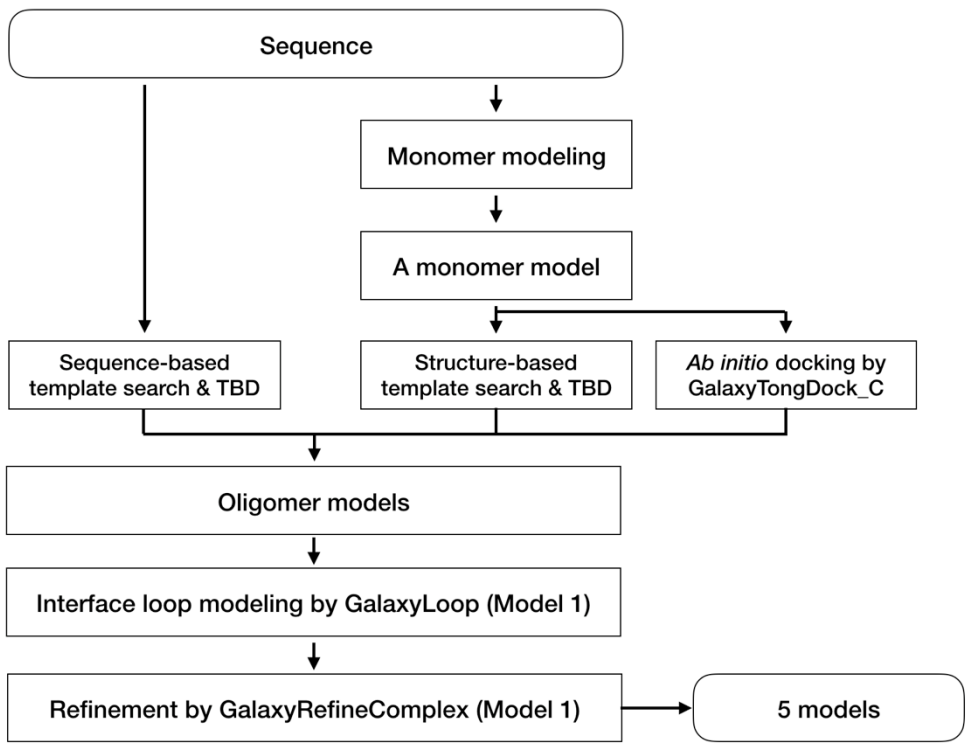
GalaxyHomomer was designed to prioritize sequence-based template search over structure-based template search when a protein sequence is provided as input. The structure-based template search was used only when less than five templates were found by sequence-based template search. Since the time that GalaxyHomomer was developed, the accuracy of the protein structure prediction method has been significantly improved[39]. Therefore, structure-based template search and modeling based on structural superposition, the performance of which primarily depends on the monomer model quality, is more emphasized in GalaxyHomomer2.

4.1. Methods

4.1.1. The overall pipeline of GalaxyHomomer2

The prediction pipeline of the GalaxyHomomer2 for predicting homo-oligomer structures is shown in **Figure 4.1**. In GalaxyHomomer2, template-based docking is performed by detecting templates for homo-oligomer structure modeling based on monomer sequence similarity (sequence-based template search) and monomer structure similarity (structure-based template search). The sequence-based template search and structure-based template search are performed on the DB-Ho, which is the same database that is used in GalaxyHomomer. Homo-oligomer templates are detected by HHsearch. Homo-oligomers in DB-Ho with high sequence similarity (in terms of GalaxyTBM template score within the top 200) become the templates detected from the sequence-based template search. If a monomer structure is not provided as input, it is predicted from a monomer sequence using a protein structure prediction method explained in **Section 3.1.2**. Among the top 200 templates from the sequence-based template search, templates with high structure similarity (TM-score > 0.4) become the templates detected from the structure-based template search. The template detected from the sequence-based template search and the structure-based template search are ranked based on a template score, described in **Section 4.1.3**. Homo-oligomer models are generated by the sequence threading method of GalaxyTBM if a template is from sequence-based template search and structural superposition if a template is from structure-based template search. If < 5 models are generated from template-based docking, *ab initio* docking is performed using GalaxyTongDock_C to generate more models so that a total of 5 models can be obtained. The best scoring model is further refined by using GalaxyLoop and GalaxyRefineComplex.

Figure 4.1. GalaxyHomomer2 pipeline for homo-oligomer structure prediction. TBD stands for template-based docking. The TBD corresponds to the sequence threading method of GalaxyTBM if a template is from sequence-based template search and structural superposition if a template is from structure-based template search.



4.1.2. Benchmark set preparation

A benchmark set for GalaxyHomomer2 was prepared by extracting homo-oligomers satisfying the following conditions from PDB:

a) Containing proteins only

b) Oligomeric state is between 2 to 6

c) Molecular weight is between 5,000 to 100,000

d) Has no free ligands

e) Has no modified polymeric residues

f) Resolution is between 0 Å to 2.5 Å

The homo-oligomers satisfying the conditions were clustered to have maximum mutual sequence identity < 30%, resulting in 2,410 homo-oligomers. Homo-oligomers composed of multi-domain proteins were detected by GalaxyDom (unpublished), an in-house domain detection method, and excluded from the benchmark set, resulting in a benchmark set of 2244 homo-oligomers. From the benchmark set, 600 homo-oligomers, Set 1, were randomly selected by maintaining the portion of each oligomeric state (dimer: 458, trimer: 52, tetramer: 63, pentamer: 5, hexamer: 22). In the same way, 600 homo-oligomers, Set 2, were randomly selected from the remaining homo-oligomers, maintaining the portion of each oligomeric state.

4.1.3. Parameter tuning for ranking templates

Scoring and ranking templates detected by different template search methods are tricky because the templates were selected based on different criteria. First, the GalaxyHomomer2 template score, S_{hom2}, for templates from the sequence-based template search and templates from the structure-based template search was formulated as follows:

$$S_{\mathrm{hom2}} = S_{\mathrm{seq}} = S_{\mathrm{GalaxyTBM}}\left(a \times TM_{\mathrm{pred}} + b \times id_{\mathrm{seq}}\right)$$

$$S_{\mathrm{hom2}} = S_{\mathrm{str}} = S_{\mathrm{GalaxyTBM}} \times TM_{\mathrm{pred}}\left(c + d \times TM_{\mathrm{mono}} + e \times TM_{\mathrm{iface}}\right)$$

where S_{seq} is a template score for a template from sequence-based template search, S_{str} is a template score for a template from structure-based template search, $S_{\mathrm{GalaxyTBM}}$ is the template score of GalaxyTBM, TM_{pred} is predicted TM-score that is calculated by GalaxyTBM based on HHsearch result and has a single value for a target homo-oligomer, id_{seq} is sequence identity between a target and a template, TM_{mono} is structure similarity between a monomer model and a template, TM_{iface} is interface structure similarity between a monomer model and an interface region of a template. The interface region was defined as residues whose C_α atoms are placed closer than 20 Å from any of the other chain's C_α atoms.

For parameter tuning of a, b, c, d, and e values, 20 models from each sequence-based template search followed by modeling based on sequence threading method and structure-based template search followed by modeling based on structural superposition were generated for homo-oligomers in Set 1. IRMSD of the generated model was calculated. c, d, and e values minimizing the best IRMSD of top 5 models from structure-based template search were first determined to be 1,

0.2, and 0.5 by rough grid search in parameter space. And then, a and b values minimizing the best IRMSD of top 5 models from both template searches were determined to be 1.65 and 0 by rough grid search in parameter space. Therefore, S_{hom2} is determined as follows:

$$S_{\mathrm{hom2}} = S_{\mathrm{seq}} = 1.65 \times S_{\mathrm{GalaxyTBM}} \times TM_{\mathrm{pred}}$$

$$S_{\mathrm{hom2}} = S_{\mathrm{str}} = S_{\mathrm{GalaxyTBM}} \times TM_{\mathrm{pred}}(1 + 0.2 \times TM_{\mathrm{mono}} + 0.5 \times TM_{\mathrm{iface}})$$

In both S_{seq} and S_{str}, TM_{pred}, which estimates the accuracy of a monomer model, is necessary to define a cutoff value for deciding whether use the templates for homo-oligomer modeling or not. *Ab initio* docking, by GalaxyTongDock_C, may be a better choice than template-based docking using templates with low S_{hom2}. It turned out that *ab initio* docking performs better than template-based docking when $\frac{S_{\mathrm{hom2}}}{\max(S_{\mathrm{GalaxyTBM}})} < 1.0$. Therefore, templates having $\frac{S_{\mathrm{hom2}}}{\max(S_{\mathrm{GalaxyTBM}})}$ lower than 1.0 are excluded in the homo-oligomer modeling step.

By applying the above GalaxyHomomer2 template score and template exclusion criterion, the usage of models from structure-based template searches in the top 1 and top 5 models was increased compared to that of GalaxyHomomer, as shown in **Table 4.1**. The usage of models from *ab initio* docking is also increased.

Table 4.1. The number of models in the top 1 and top 5 models generated for Set 1 depending on their modeling methods.

Set 1	**# of models (sequence-based template search)**	**# of models (structure-based template search)**	**# of models (*ab initio* docking)**
GalaxyHomomer (top 1 model)	525	28	47
GalaxyHomomer2 (top 1 model)	302	218	80
GalaxyHomomer (top 5 models)	2038	400	562
GalaxyHomomer2 (top 5 models)	1671	477	852

4.2. Performance of GalaxyHomomer2

The performance of GalaxyHomomer2 was compared to that of GalaxyHomomer. Templates having more than 40% sequence identity to targets were excluded for the performance test. Even though some parameters of GalaxyHomomer2 were tuned using Set 1, the performance of GalaxyHomomer2 was shown to be better on Set 2 in terms of CAPRI criterion, as shown in **Tables 4.2 and 4.3**. The same tendency was observed in terms of LRMSD, IRMSD, and F_{nat}[32]. Furthermore, the performance gap between GalaxyHomomer2 and GalaxyHomomer was more noticeable on Set 2. This result could happen because Set 1 and Set 2 have different properties and the grid search of parameters was rough enough not to be over-trained on Set 1. The performance of GalaxyHomomer2 was analyzed in detail on the combined set of Set 1 and Set 2 that is composed of 1,200 homo-oligomers. The results in **Tables 4.2 and 4.3** are slightly different from those in the following analysis because there were minor changes in the GalaxyHomomer2 pipeline after **Tables 4.2 and 4.3** were made.

Table 4.2. The performance test result of GalaxyHomomer2 and GalaxyHomomer on Set 1 in terms of CAPRI criterion.

Set 1	High	Medium	Acceptable
GalaxyHomomer (top 1 model)	0	102	86
GalaxyHomomer2 (top 1 model)	1	104	90
GalaxyHomomer (top 5 models)	5	154	88
GalaxyHomomer2 (top 5 models)	4	157	88

Table 4.3. The performance test result of GalaxyHomomer2 and GalaxyHomomer on Set 2 in terms of CAPRI criterion.

Set 2	High	Medium	Acceptable
GalaxyHomomer (top 1 model)	3	93	100
GalaxyHomomer2 (top 1 model)	4	104	91
GalaxyHomomer (top 5 models)	6	165	89
GalaxyHomomer2 (top 5 models)	7	174	93

The performance of the final pipeline of GalaxyHomomer2 was compared to that of GalaxyHomomer in terms of GDT-TS of monomer models (**Figure4.2**), CAPRI criterion (**Table 4.4**), LRMSD (**Figure 4.3**), IRMSD (**Figure 4.4**), and F_{nat}, (**Figure 4.5**). The average values of GDT-TS of monomer models, LRMSD, RMSD, and F_{nat} are summarized in **Table 4.5**. There were moderate improvements in GDT-TS of monomer models, CAPRI criterion, IRMSD, and LRMSD. GalaxyHomomer2 showed similar performance in terms of F_{nat} compared to GalaxyHomomer.

In summary, GalaxyHomomer2 generated better quality models and made successful predictions on more targets compared to GalaxyHomomer. However, the extent of improvement was moderate. There is still large room for improvement because GalaxyHomomer2 failed many targets that were successfully predicted by GalaxyHomomer, as can be seen in **Figures 4.3, 4.4, and 4.5**.

There was a considerable improvement of the monomer structure prediction method on CASP14 (2020). AlphaFold2 successfully predicted the monomer structures for most of the targets of CASP14 with remarkable precision. Surprisingly, the monomer structures of the obligate homo-oligomers were correctly predicted without explicitly considering their oligomeric states. If structure-based template search is combined with an advanced monomer structure prediction method, such as AlphaFold2, the performance of template-based docking combined with structure-based template search, which is more emphasized in GalaxyHomomer2, is expected to be significantly improved.

Figure 4.2. The performance test result of GalaxyHomomer2 and GalaxyHomomer on the combined set (1,200 homo-oligomers) in terms of GDT-TS of monomer models.

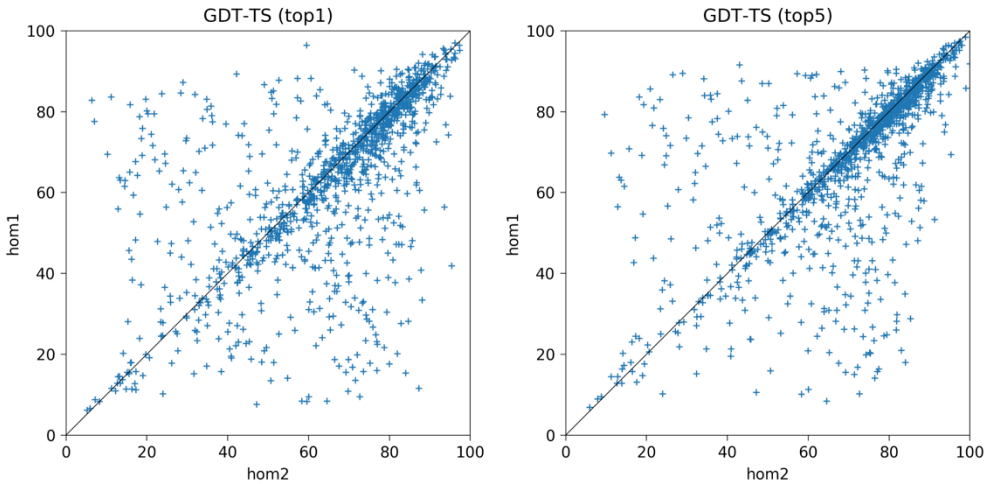


Table 4.4. The performance test result of GalaxyHomomer2 and GalaxyHomomer on the combined set in terms of CAPRI criterion.

Combined set	High	Medium	Acceptable
GalaxyHomomer (top 1 model)	3	195	186
GalaxyHomomer2 (top 1 model)	5	207	183
GalaxyHomomer (top 5 models)	11	319	177
GalaxyHomomer2 (top 5 models)	11	334	185

Figure 4.3. The performance test result of GalaxyHomomer2 and GalaxyHomomer on the combined set in terms of LRMSD.

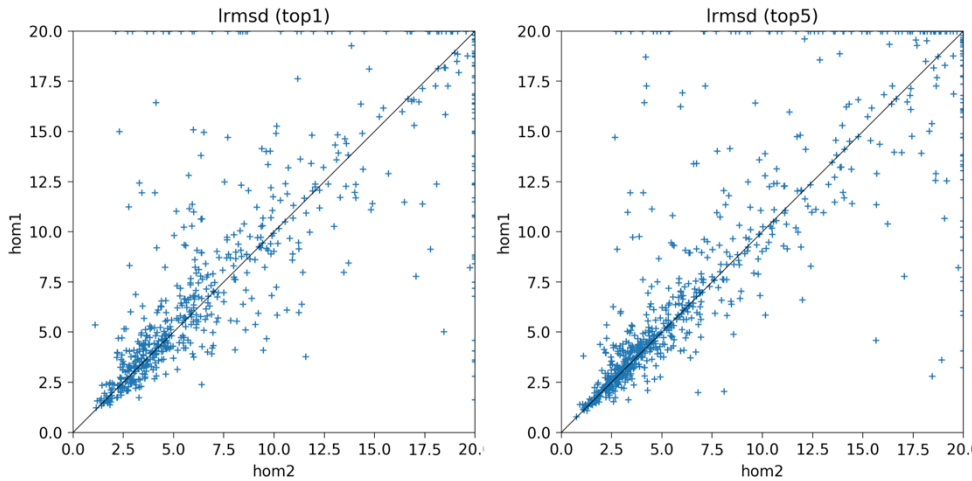


Figure 4.4. The performance test result of GalaxyHomomer2 and GalaxyHomomer on the combined set in terms of IRMSD.

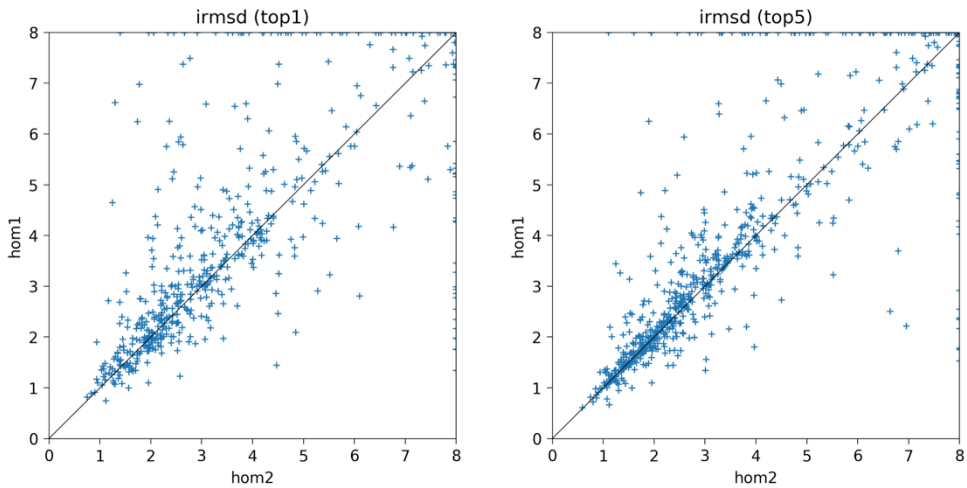


Figure 4.5. The performance test result of GalaxyHomomer2 and GalaxyHomomer on the combined set in terms of F_{nat}.

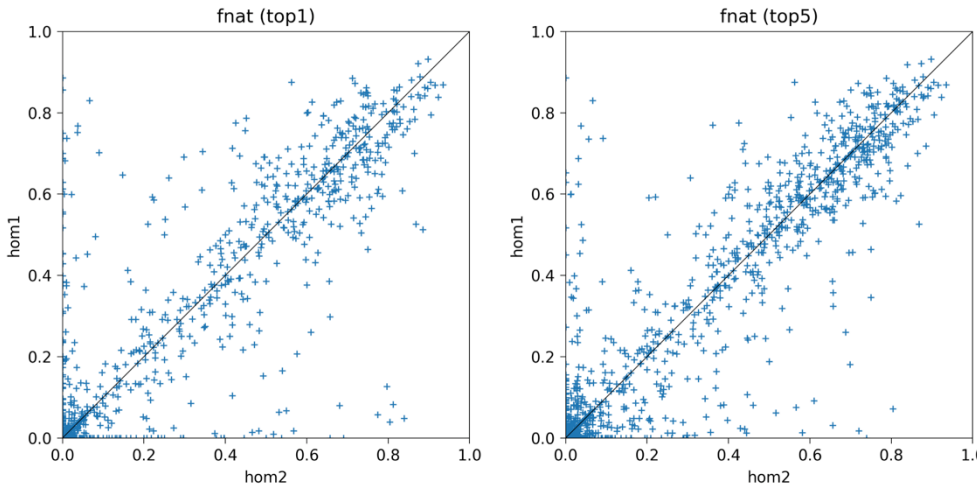


Table 4.5. The performance test result of GalaxyHomomer2 and GalaxyHomomer on the combined set in terms of multiple accuracy measures (average GDT-TS of monomer models, LRMSD, IRMSD, and F_{nat} of the top 1 and top 5 models). For the top 5 models, the model with the best value for each accuracy measure was used for computing the average. LRMSD over 20 Å was considered to be 20 Å, and IRMSD over 8 Å was considered to be 8 Å.

Combined set	GalaxyHomomer (top 1 model)	GalaxyHomomer2 (top 1 model)	GalaxyHomomer (top 5 model)	GalaxyHomomer2 (top 5 model)
GDT-TS of monomer models	68.2	70.3	71.6	73.6
LRMSD	7.12	6.82	6.76	6.28
IRMSD	3.48	3.24	3.21	2.97
F_{nat}	0.436	0.424	0.424	0.423

5. CASP and CAPRI

By utilizing the developed methods and other GALAXY software[47], we have achieved high performance in CASP and CAPRI, second place in the CASP13 assembly category[17], third place in CASP13-CAPRI[18], fourth place in CASP14 assembly, first place in CASP14-CAPRI, and second place in CAPRI 7th edition (round 38-45)[19]. CASP and CAPRI are community-wide prediction experiments for protein structure prediction and protein complex structure prediction in a blind fashion. In CASP and CAPRI, there are server and human predictions that participants should submit their models within three days and target-dependent time, respectively. In our group, the models of the server predictions were automatically generated by a prediction pipeline, and the models of the human predictions were manually generated with human intervention, such as literature search and template search. In this section, examples of how the developed methods have been utilized to predict protein complex structures in CASP and CAPRI are presented.

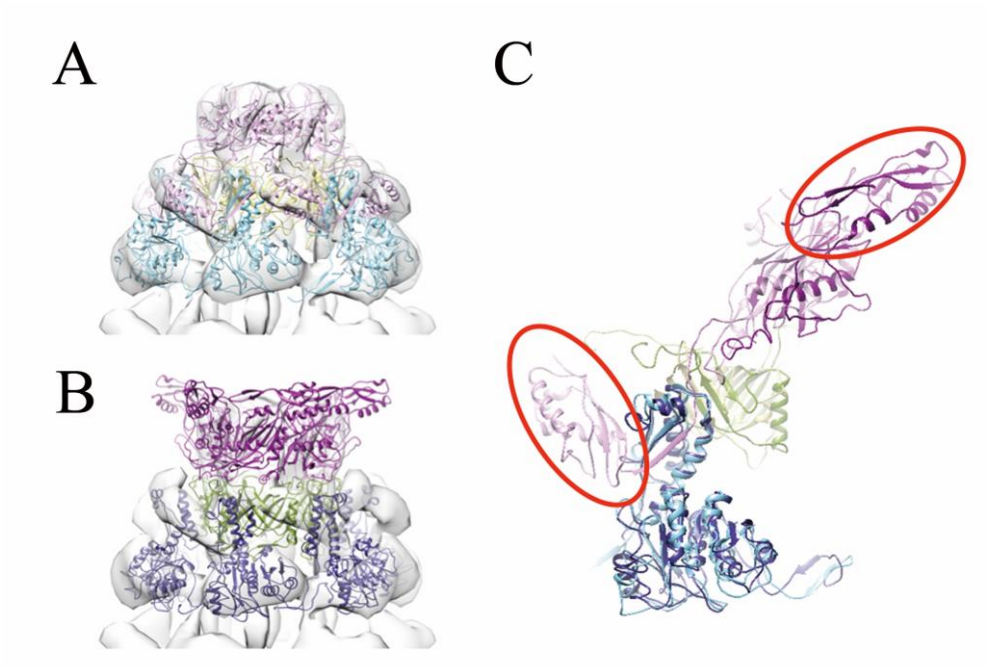
5.1. CASP13

5.1.1. H1021 ($A_6B_6C_6$)

H1021 has $A_6B_6C_6$ stoichiometry. H1021 was a challenging target due to its large assembly. In the human prediction of the target H1021, low-resolution EM data was utilized to predict the oligomer structure. We first built homo-hexamer structures of each subunit based on the templates (5W5F for A_6, 3J9Q for B_6, and 4HUH for C_6). The templates were detected by the template score of GalaxyTBM. Two homo-hexamer subunit structures, A6 and B6, were assembled by rigid-body

fitting into a 20 Å resolution EM map (Electron Microscopy Data Bank [EMDB]: EMD-2419)[48] using UCSF Chimera[49]. Due to the wrongly modeled C-terminal domain orientation in the third subunit, we failed to fit the last homo-hexamer subunits (C_6) into the EM data. Instead, the last subunits were docked into assembled A_6B_6 structure by GalaxyTongDock. Our model 5 turned out to have the best ICS score[50], and it showed quite good arrangements of the subunits except for the orientation of the last subunit's C-terminal domain, as shown in **Figure 5.1**. Our model 5 for the target H1021 was highlighted at the CASP13 conference for its accuracy.

Figure 5.1. Multimeric structure of model 5 for H1021. (A) The native structure and (B) model 5 are shown with a low-resolution EM map (Electron Microscopy Data Bank [EMDB]: EMD-2419). (C) A part of the complex (chains A, B, C) is shown in clarity. Native chain structures are in lighter colors, while those of model 5 are in darker colors. Except for the circled domain, the overall predicted structure is quite similar to the native structure.

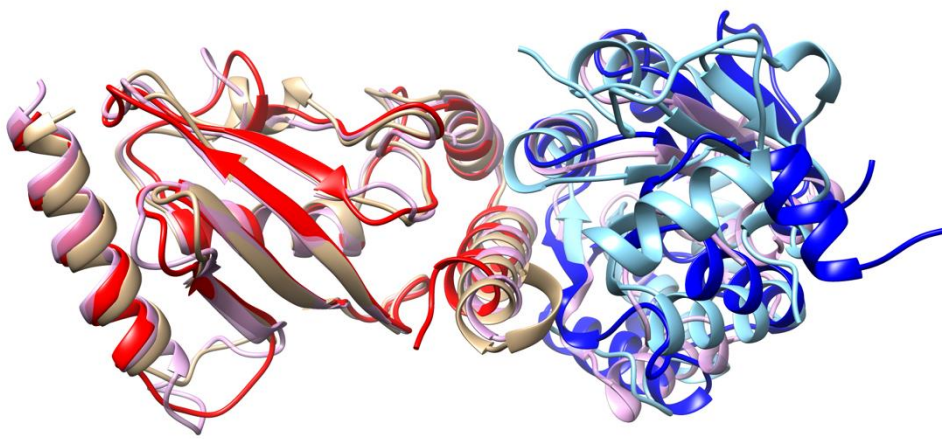


5.2. CASP14

5.2.1. H1045 (AB)

H1045 is a heterodimer protein. Model 1 of our server prediction, which is automatically predicted by GalaxyHeteromer, was of medium accuracy in terms of CAPRI criterion. GalaxyHeteromer generated subunit structure models of GDT-TS = 82 and 64. An available heterodimer template (PDB ID: 2Y9M) was detected via structure-based template detection. A heterodimer structure was built by superposing the models of the subunits onto the template, and energy minimization was performed to remove steric clashes. This model structure was refined by the complex structure refinement method GalaxyRefineComplex, resulting in a heterodimer structure with medium accuracy [F_{nat} = 0.630, IRMSD = 2.00 Å, LRMSD = 11.8 Å] (colored yellow and sky-blue in **Figure 5.2**). LRMSD was relatively large because of the low-quality monomer model of subunit B. The heterodimer model could have medium accuracy because the interface regions of both subunits were modeled accurately.

Figure 5.2. Model 1 of GalaxyHeteromer (in yellow and sky-blue), which has a medium accuracy [F_{nat} = 0.630, IRMSD = 2.00 Å, LRMSD = 11.8 Å], is superposed on the native structure of H1045 (in red and blue). The template (PDB ID: 2Y9M) used for the heterodimer modeling is colored pink. Even though the monomer model of the subunit B (in sky-blue) was relatively inaccurate, the heterodimer model could be successfully predicted because the interface regions of both subunits were accurately modeled.



5.2.2. T1070 (A_3)

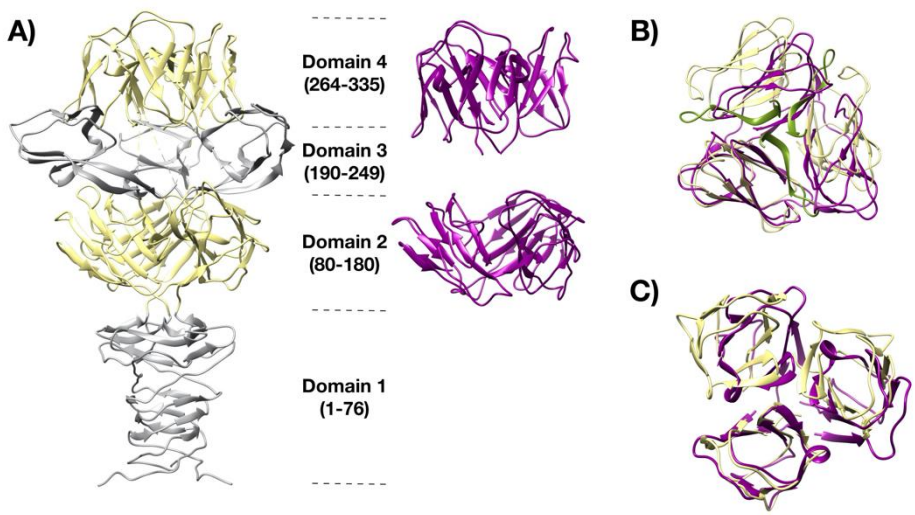
Our server prediction for the homo-trimer structure of the four-domain protein T1070 was unsuccessful due to the failure of monomer structure prediction: no proper oligomeric template was found when the target was treated as a single-domain protein. In the human prediction, the T1070 trimeric structure prediction was performed domain-wise after manually dividing the structure into four domains (residues 1-76, 80-165, 190-249, and 264-335). The GalaxyTongDock_C was used to predict the trimer structures for each domain using domain structures selected from the CASP14 TS server models. Indeed, the crystal structure of T1070 could be divided into four structural domains, with slightly different domain assignments (1-76, 80-180, 181-256, and 265-335), as shown in **Figure 5.3A**.

Figure 5.3 shows the results of our human predictions for only the second and fourth domains since predictions for the first and the third domains were unsuccessful (crystal structures are shown in gray color). The structures of the trimers of the first domain, an inter-twined domain, and the third domain, which do not include inter-subunit contacts, were challenging to be predicted via *ab initio* docking. The trimer structure of the second domain of our model 5 (**Figure 5.3B**) was of acceptable quality in terms of the CAPRI accuracy criterion [F_{nat} = 0.18, IRMSD = 5.26 Å, LRMSD = 5.31 Å]. The monomer structure for this domain was provided by Seok-server_TS1, which was generated by GalaxyDBM [GDT-TS = 76.19, RMSD = 4.283 Å]. The predicted trimer structure was more compact (see magenta-colored structure in **Figure 5.3B**) than the crystal structure (yellow) since the beta-strand residues (166-180, green) were tightly interacting with each other in the crystal structure. This region was not included in the monomer structure for docking due to domain mis-splitting. The domain boundary of the fourth domain was predicted accurately, and our human prediction model 1 for this domain

(**Figure 5.3C**) was of medium quality with respect to the CAPRI criterion [F_{nat} = 0.42, IRMSD = 3.22 Å, LRMSD = 2.06 Å]. The monomer structure used for docking was obtained from Zhang-CEthreade_TS1 [GDT-TS = 83.45, RMSD = 2.423 Å].

These results demonstrated that homo-oligomer structures can be predicted by symmetric *ab initio* docking when monomer structures of reasonable quality are available. For multi-domain proteins, an accurate domain assignment is crucial for successful structure prediction. We did not attempt assembly of the four trimer domains for T1070, as the first and third domain structures were deemed inaccurate.

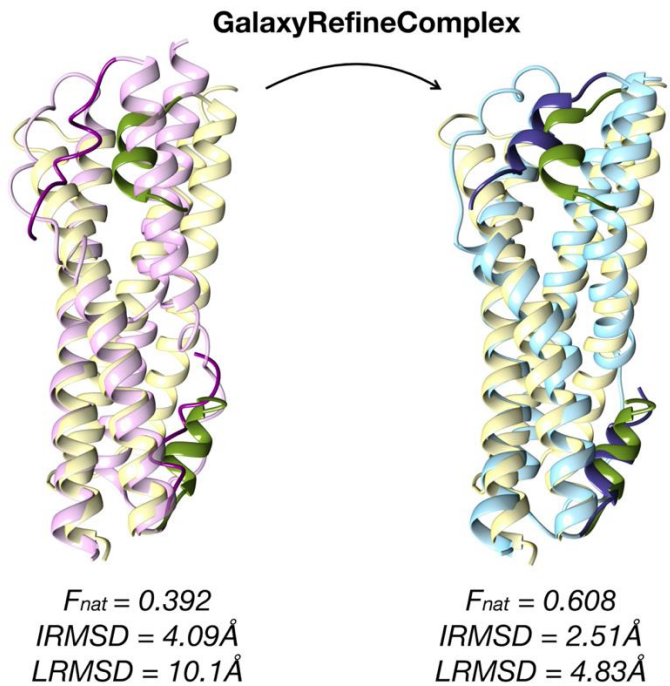
Figure 5.3. Comparison of the crystal and modeled structures of T1070. (A) The crystal structure of T1070 is divided into four domains. Domains 1 and 3 are colored in gray, and domains 2 and 4 are colored in yellow. The model structures of domains 2 and 4 extracted from our human models 5 and 1, respectively, are depicted in purple. (B) The model structure of domain 2 (magenta) is superposed onto the crystal structure (yellow and green) and shown from a different perspective. The beta-strand residues 166-180 (green), missing in the model structure, interact tightly with each other in the crystal structure. The subunits of the model structure are closer to each other in the model structure due to the missing region. (C) The model structure of domain 4 (magenta) is superposed onto the crystal structure (yellow).



5.2.3. T1083 (A_2)

T1083 is a homodimer protein. Model 1 of our server prediction, which is automatically predicted by GalaxyHomomer2, was of medium accuracy in terms of CAPRI criterion. GalaxyHomomer2 detected an available oligomer template [PDB ID = 3GWK, TM-score = 0.61] via the structure-based template detection using the monomer model generated by GalaxyDBM [GDT-TS = 87.5, RMSD = 2.82 Å]. A homodimer structure was built by superposing the monomer model onto the template, and energy minimization was performed to remove steric clashes. Local energy minimization was insufficient to induce a conformational change from the superposed structure (colored in pink in **Figure 5.4**) to the crystal structure (yellow). This model structure was incorrect [F_{nat} = 0.392, IRMSD = 4.09 Å, LRMSD = 10.1 Å]. This model structure was refined by GalaxyRefineComplex, resulting in an improved structure with medium accuracy [F_{nat} = 0.608, IRMSD = 2.51 Å, LRMSD = 4.83 Å] (colored sky-blue in **Figure 5.4**). This model was submitted as model 1. GalaxyRefineComplex refined the loose N-terminal portions of each subunit of the homodimer (magenta in **Figure 5.4**) to a helix structure (dark blue), which packed against the helix bundle, similarly to the N-terminus of the crystal (green). The relative orientation of the two subunit helices was also improved upon refinement (structures before and after refinement are colored pink and sky-blue, respectively).

Figure 5.4. The crystal (yellow) and modeled structures before (pink) and after refinement (sky-blue) of T1083. The loose N-terminal structures of the two subunits before refinement (magenta) were well-packed upon refinement (dark blue) and approached the crystal structure (green). The relative orientation between the two subunits was also improved by the refinement.



5.3. CAPRI

5.3.1. T131 (AB)

This target is a hetero-dimer protein. The crystal structures of both subunits were available at the time of prediction (PDB ID: 5LP2 and 4WHD). No acceptable or higher accuracy models were submitted within the top 10 by either server or human predictions for this target. Our server model 66 was of high quality with [F_{nat} = 0.92, IRMSD = 1.5 Å, LRMSD = 2.7 Å], which was the best among all models submitted by all predictors. Interestingly, the model was selected as top 1 by GalaxyTongDock. However, the model was evaluated to be wrong by GalaxyPPDock (unpublished), which performs global optimization starting with the models generated by GalaxyTongDock. We attribute this failure to the limitation of GalaxyPPDock that does not sample backbone structure flexibility. The rigid-body docking method GalaxyTongDock considers backbone flexibility implicitly with a low penalty for steric clashes. GalaxyPPDock employs high-resolution energy, which imposes a high penalty for steric clashes.

6. Conclusion

Three protein complex structure prediction methods, GalaxyTongDock, GalaxyHeteromer, and GalaxyHomomer2 were developed. GalaxyTongDock performs asymmetric and symmetric *ab initio* docking utilizing FFT. GalaxyHeteromer and GalaxyHomomer2 employ both template-based docking and ab initio docking for hetero- and homo-oligomer structure prediction, respectively. The methods have been being freely serviced for global users as web servers at

GalaxyTongDock web server (http://galaxy.seoklab.org/tongdock)

GalaxyHeteromer web server (http://galaxy.seoklab.org/heteromer)

GalaxyHomomer2 web server (http://galaxy.seoklab.org/homomer)

Our group has been ranked on top in multiple CASP and CAPRI using the methods, second place in the CASP13 assembly category, third place in CASP13-CAPRI, fourth place in CASP14 assembly category, first place in CASP14-CAPRI, and second place in CAPRI 7th edition (round 38-45). The methods have been actively utilized in various joint research with pharmaceutical companies and academic research groups regarding protein complex structure prediction, protein-peptide docking, and protein drug design.

Near future, we expect there will be meaningful progress in protein complex structure prediction from two aspects. First, there was a remarkable

improvement of the protein structure prediction in CASP14 by AlphaFold2. Accurate monomer model structures are a prerequisite for accurate protein complex structure prediction. A meaningful improvement would be followed as protein complex structure prediction methods are combined with advanced monomer structure prediction methods. Second, many researchers are trying to apply deep learning to protein complex structure prediction. Deep learning is thought to have massive potential in protein complex structure prediction. We are trying to apply deep learning to protein complex structure prediction. The performance of GalaxyTongDock has been highly increased by a deep learning-based docking pose rescoring method. Binding affinity between proteins could also be more accurately predicted by deep learning-based energy than conventional GALAXY energy.

Improvement in protein structure prediction and protein complex structure prediction will change the field of drug discovery. Computational methods are potent tools for rationally designing protein-based therapeutic agents, like antibody and cytokine drugs. Accurately predicted protein-protein interface makes it possible to determine where to target and how to target for discovering protein-protein interaction inhibitors. The era of precision medicine does not seem too far from where we are standing.

7. References

1 Cheng, F. *et al.* Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat Genet* **53**, 342-353, doi:10.1038/s41588-020-00774-y (2021).

2 Ryan, D. P. & Matthews, J. M. Protein-protein interactions in human disease. *Curr Opin Struct Biol* **15**, 441-446, doi:10.1016/j.sbi.2005.06.001 (2005).

3 Gonzalez, M. W. & Kann, M. G. Chapter 4: Protein interactions and disease. *PLoS Comput Biol* **8**, e1002819, doi:10.1371/journal.pcbi.1002819 (2012).

4 Vaynberg, J. & Qin, J. Weak protein-protein interactions as probed by NMR spectroscopy. *Trends Biotechnol* **24**, 22-27, doi:10.1016/j.tibtech.2005.09.006 (2006).

5 Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G. & Orengo, C. Transient protein-protein interactions: structural, functional, and network properties. *Structure* **18**, 1233-1243, doi:10.1016/j.str.2010.08.007 (2010).

6 Wang, Q., Zhuravleva, A. & Gierasch, L. M. Exploring weak, transient protein--protein interactions in crowded in vivo environments by in-cell nuclear magnetic resonance spectroscopy. *Biochemistry* **50**, 9225-9236, doi:10.1021/bi201287e (2011).

7 Acuner Ozbabacan, S. E., Engin, H. B., Gursoy, A. & Keskin, O. Transient protein-protein interactions. *Protein Eng Des Sel* **24**, 635-648, doi:10.1093/protein/gzr025 (2011).

8 Porter, K. A., Desta, I., Kozakov, D. & Vajda, S. What method to use for protein-protein docking? *Curr Opin Struc Biol* **55**, 1-7, doi:10.1016/j.sbi.2018.12.010 (2019).

9 Yan, Y., Zhang, D., Zhou, P., Li, B. & Huang, S. Y. HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res* **45**, W365-W373, doi:10.1093/nar/gkx407 (2017).

10 Baek, M., Park, T., Heo, L., Park, C. & Seok, C. GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. *Nucleic Acids Res* **45**, W320-W324, doi:10.1093/nar/gkx246 (2017).

11 Huang, S. Y. Exploring the potential of global protein-protein docking: an overview and critical assessment of current programs for automatic ab initio docking. *Drug Discov Today* **20**, 969-977, doi:10.1016/j.drudis.2015.03.007 (2015).

12 Szilagyi, A. & Zhang, Y. Template-based structure modeling of protein-protein interactions. *Curr Opin Struct Biol* **24**, 10-23, doi:10.1016/j.sbi.2013.11.005 (2014).
13 Pierce, B. G. *et al.* ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771-1773, doi:10.1093/bioinformatics/btu097 (2014).
14 Park, T., Baek, M., Lee, H. & Seok, C. GalaxyTongDock: Symmetric and asymmetric ab initio protein-protein docking web server with improved energy parameters. *J Comput Chem* **40**, 2413-2417, doi:10.1002/jcc.25874 (2019).
15 Park, T., Won, J., Baek, M. & Seok, C. GalaxyHeteromer: protein heterodimer structure prediction by template-based and ab initio docking. *Nucleic Acids Res*, doi:10.1093/nar/gkab422 (2021).
16 Ko, J., Park, H. & Seok, C. GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. *BMC Bioinformatics* **13**, 198, doi:10.1186/1471-2105-13-198 (2012).
17 Guzenko, D., Lafita, A., Monastyrskyy, B., Kryshtafovych, A. & Duarte, J. M. Assessment of protein assembly prediction in CASP13. *Proteins* **87**, 1190-1199, doi:10.1002/prot.25795 (2019).
18 Lensink, M. F. *et al.* Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins* **87**, 1200-1221, doi:10.1002/prot.25838 (2019).
19 Lensink, M. F., Nadzirin, N., Velankar, S. & Wodak, S. J. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins* **88**, 916-938, doi:10.1002/prot.25870 (2020).
20 Katchalski-Katzir, E. *et al.* Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A* **89**, 2195-2199 (1992).
21 Pierce, B. G., Hourai, Y. & Weng, Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS One* **6**, e24657, doi:10.1371/journal.pone.0024657 (2011).
22 Frigo, M. & Johnson, S. G. in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on.* 1381-1384 vol.1383.
23 Mintseris, J. *et al.* Integrating statistical pair potentials into protein complex prediction. *Proteins* **69**, 511-520, doi:10.1002/prot.21502 (2007).
24 Gabb, H. A., Jackson, R. M. & Sternberg, M. J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* **272**, 106-120, doi:10.1006/jmbi.1997.1203 (1997).
25 Zhang, C., Vasmatzis, G., Cornette, J. L. & DeLisi, C. Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol* **267**, 707-726, doi:10.1006/jmbi.1996.0859 (1997).
26 Chen, R., Li, L. & Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**, 80-87, doi:10.1002/prot.10389 (2003).

27 Liang, S., Meroueh, S. O., Wang, G., Qiu, C. & Zhou, Y. Consensus scoring for enriching near-native structures from protein-protein docking decoys. *Proteins* **75**, 397-403, doi:10.1002/prot.22252 (2009).
28 NACCESS (Department of Biochemistry and Molecular Biology, University College London, 1993).
29 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
30 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919 (1992).
31 Vreven, T. *et al.* Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol* **427**, 3031-3041, doi:10.1016/j.jmb.2015.07.016 (2015).
32 Lensink, M. F. & Wodak, S. J. Docking and scoring protein interactions: CAPRI 2009. *Proteins* **78**, 3073-3084, doi:10.1002/prot.22818 (2010).
33 Lyskov, S. & Gray, J. J. The RosettaDock server for local protein-protein docking. *Nucleic Acids Res* **36**, W233-238, doi:10.1093/nar/gkn216 (2008).
34 Heo, L., Lee, H. & Seok, C. GalaxyRefineComplex: Refinement of protein-protein complex model structures driven by interface repacking. *Sci Rep* **6**, 32153, doi:10.1038/srep32153 (2016).
35 Ritchie, D. W. & Grudinin, S. Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J Appl Crystallogr* **49**, 158-167, doi:10.1107/s1600576715022931 (2016).
36 Ponstingl, H., Kabir, T. & Thornton, J. M. Automatic inference of protein quaternary structure from crystals. *J Appl Crystallogr* **36**, 1116-1122, doi:10.1107/S0021889803012421 (2003).
37 Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-2309, doi:10.1093/nar/gki524 (2005).
38 Park, H., Lee, G. R., Heo, L. & Seok, C. Protein Loop Modeling Using a New Hybrid Energy Function and Its Application to Modeling in Inaccurate Structural Environments. *Plos One* **9**, doi:ARTN e11381110.1371/journal.pone.0113811 (2014).
39 Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-+, doi:10.1038/s41586-019-1923-7 (2020).
40 Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* **31**, 3370-3374, doi:10.1093/nar/gkg571 (2003).
41 Seemayer, S., Gruber, M. & Soding, J. CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30**, 3128-3130, doi:10.1093/bioinformatics/btu500 (2014).
42 Lee, J., Scheraga, H. A. & Rackovsky, S. New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *J Comput Chem* **18**, 1222-1232, doi:Doi 10.1002/(Sici)1096-987x(19970715)18:9<1222::Aid-Jcc10>3.0.Co;2-7 (1997).

43 Lee, G. R., Heo, L. & Seok, C. Simultaneous refinement of inaccurate local regions and overall structure in the CASP12 protein model refinement experiment. *Proteins-Structure Function and Bioinformatics* **86**, 168-176, doi:10.1002/prot.25404 (2018).

44 Heo, L., Park, H. & Seok, C. GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res* **41**, W384-388, doi:10.1093/nar/gkt458 (2013).

45 Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960, doi:10.1093/bioinformatics/bti125 (2005).

46 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).

47 Ko, J., Park, H., Heo, L. & Seok, C. GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Res* **40**, W294-297, doi:10.1093/nar/gks493 (2012).

48 Heymann, J. B. *et al.* Three-dimensional structure of the toxin-delivery particle antifeeding prophage of Serratia entomophila. *J Biol Chem* **288**, 25276-25284, doi:10.1074/jbc.M113.456145 (2013).

49 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).

50 Lafita, A. *et al.* Assessment of protein assembly prediction in CASP12. *Proteins* **86 Suppl 1**, 247-256, doi:10.1002/prot.25408 (2018).

국문초록

단백질 사이의 상호작용은 세포분열, 항상성 유지, 면역반응, 질병의 발생 등 많은 생물학적 과정에서 핵심적인 역할을 한다. 단백질 복합체 구조로부터 얻을 수 있는 단백질 상호작용에 대한 구조적 이해는 효과적인 항체 신약, 단백질 상호작용 저해제 등의 약물 설계를 위해 필수적인 요소이다. 그러나 단백질 복합체는 대체로 약한 상호작용에 의해 일시적으로 형성되어 실험을 통해 결정하기가 어렵다. 실제로 우리 몸에서 일어나는 수많은 단백질 상호작용 중 극히 일부에 대해서만 복합체 구조가 알려져 있다. 컴퓨터를 이용한 단백질 복합체 구조 예측 방법은 실험에 의해 결정된 단백질 복합체 구조가 없는 경우에 단백질 상호작용에 대한 정보를 제공하는 중요한 역할을 해왔다. 이 논문에서는 단백질 복합체 구조 예측 방법인 GalaxyTongDock 과 GalaxyHomomer2, GalaxyHeteromer 에 대해서 소개한다. GalaxyTongDock 은 *ab initio* 도킹을 통해 동종 올리고머 단백질과 이종 올리고머 단백질의 구조를 예측한다. GalaxyHomomer2 와 GalaxyHeteromer 는 각각 동종 올리고머 단백질과 이종 올리고머 단백질의 구조를 주형 기반 도킹과 *ab initio* 도킹을 모두 이용하여 예측한다. 마지막으로, 이 방법들이 국제 단백질 구조 및 복합체 구조 예측 대회인 CASP 과 CAPRI 에서 단백질 복합체 구조를 예측하기 위해 어떻게 활용되었는지 몇 가지 예시를 통해 소개한다.

주요어: 단백질 복합체 구조 예측, 단백질-단백질 도킹, 주형 기반 도킹, *ab initio* 도킹, CASP, CAPRI

학 번: 2015-22613

감사의 글

돌아보니 수학의 길에 올라 박사학위를 취득하기까지 15 년 이상의 시간이 흘렀습니다. 상당히 즐겁고 행복한 시간이었습니다. 이는 모두 제 곁에서 함께해 준 많은 분의 덕분입니다. 표현이 서툴러 전하지 못한 감사를 박사학위 논문의 지면을 빌려 전하고자 합니다.

지도해주시고 이끌어주신 스승님들께 감사합니다.

대학원에서 저를 지도해주신 **석차옥 교수님**. 보통 대학원생들은 대학원에서의 삶이 상당히 고단하고 힘들다고 합니다. 교수님께서 지도해주신 저의 대학원에서의 삶은 그러한 삶과 동떨어진 즐겁고 보람찬 삶이었습니다. 지난 6 년 반 동안 잘 돌봐주시고 성장시켜주셔서 감사합니다. 그리고 저를 믿어주시고 앞으로 펼쳐질 무궁무진한 미래에 함께할 기회를 주셔서 감사합니다. 한 치 앞을 알 수 없는 삶이지만, 교수님께서 함께해주시니 이보다 든든할 수 없습니다. 앞으로도 잘 부탁드리겠습니다.

학부 때 저를 지도해주신 **김우연 교수님**. 갓 스물이 된 천둥벌거숭이 같은 저를 연구실에 받아주시고 지도해주셔서 감사합니다. 교수님 오피스에서 1 시간이든 2 시간이든 앉아 가르침을 받던 때가 가끔 떠오릅니다. 한 번은 교수님께서 제게 연구를 하는 데 있어 중요한 자질인 강한 상상력을 가지고 있다고 칭찬해주신 적이 있습니다. 그때 해주신 칭찬이 아직도 마음에 남아 큰 힘이 됩니다.

최영호 선생님. 칠판에 하얀 분필로 경사로 위에 상자를 그리시던 모습이 기억납니다. 제자들의 성장에 함께 진심으로 기뻐해 주시는 선생님과 과학 공부를 하며 과학이 참 재미있다고 느끼기 시작했습니다. 자랑스러운 제자라고 해주실 때마다 기뻤습니다. 선생님이야말로 제가 자랑하고 싶은 선생님이십니다.

대학원생으로서 함께 동고동락한 선후배들께 감사합니다. **하섭이 형**, **민경 누나**, 제가 여러 부사수를 가르쳐보며 저는 형, 누나에게 손이 많이 가는 부사수였겠다는 생각이 들었어요. 잘 가르쳐주셔서 감사합니다. **범창이 형**, 저는 형이 없었으면 사뭇 다른 대학원 생활을 했을 것 같아요. 형 덕분에 즐거운 일도 많았고 유쾌하게 지냈습니다. **현욱아**, 너처럼 똑똑한 부사수를 두어 영광이었어. 영국도 가고, CASP, CAPRI 도 하고, 너는 내게 말 그대로 동고동락하는 동료였어, 고맙다. **민재**, 항상 문제를 스스로 해결해보려 노력하는 모습이 멋지다고 생각해. 앞으로 현욱이에게 많이 배우고, 꾸준히 노력해서 멋진 연구자가 된 모습 기대할게.

종훈이 형, **진솔이 형**, 우리 셋은 연구실을 함께 이끌었던 세대이죠. 무슨 문제가 생기면 바로 문제를 해결해주는 형들 덕분에 저는 참 편하게 연구할 수 있었어요. 졸업한 이후에도 함께하기로 어려운 결정 해줘서 고맙습니다. 이렇게 똑똑하고 듬직한 형들이 함께해주니 어떤 고난이든 헤쳐나갈 수 있을 것 같다는 생각이 듭니다. 앞으로 펼쳐질 많은 일들 같이 경험하고 합심해서 이겨내면서 끝까지 같이 할 수 있으면 좋겠습니다.

박사학위 과정 동안 힘이 되어준 친구들에게 감사합니다. **상헌이 형**, 형 덕분에 새로운 세상을 경험하고 다양한 사람을 만나 많이 성장할 수 있었어요. 학부 때부터 지금까지 긴 시간 동안 잘 챙겨주셔서 항상 고마운 마음을 가지고 있어요. 앞으로도 자주 보고, 즐거운 일들 함께하면 좋겠어요. **수형이**, 너는 어린 시절부터 같이 성장해온 가장 믿음직스러운 친구야. 네가 같이 창업하자고 제안했던 것이 이어져 내가 지금 창업이라는 길을 선택할 수 있었다고 생각해, 고맙다. 지금 당장은 같이하진 않지만, 나중에 좋은 아이디어 기획해서 같이 세상을 바꾸자. **재욱이**, 내가 대학원에 입학하고 거의 달에 1 번은 꾸준히 본 것 같네. 네가 없었으면 참 심심한 대학원 생활이 됐을 것 같다. 아무 이유 없이도 꾸준히 계속 만나는 우리가 정말 친구 같은 친구라고 느껴진다. **민영아**, 너의 긍정적이고 밝은 에너지가 나를 더 긍정적인 사람으로 변하게 하는 것 같아. 덕분에 힘들거나 어려운 상황에 부닥쳐도 잘 이겨낼 수 있었어. 항상 응원해줘서 고마워. 앞으로도 많은 즐거운 일들 같이 할 수 있으면 좋겠어. 잘 부탁할게. **마르틴**, 함께 연구실 생활할 때 같이 자주 어울렸었는데, 너는 어느새 교수가 됐네. 네가 교수가 되고 나서도 우리 관계는 변함없지만, 만나서 나누는 대화는 한층 더 깊어진 것 같다. 대화할수록 네가 정말 생각이 깊고 배울 점이 많은 사람이라는 것을 느껴. 네가 서울대로 취직해서 자주 볼 수 있어 기쁘다. 앞으로도 자주 볼 수 있으면 좋겠어. **제민이 형**, 형과 어울렸던 그 시간 시간이 굉장히 강렬해서 여운이 길어요. 저도 나름대로 마음이 자유로운 사람이라고 자부했는데, 형은 정말.. 많이 배웠어요. **라헬아**, 과학고와 카이스트를 나온 내게 예술을 하는 네가 사는 세상은 마치 다른 세상 같아 신선한 충격을 줬어. 덕분에 다양한 사람들과 만나고 좋은 시간

보낼 수 있었어, 고마워. **세진 누나**, 처음 만났을 때 팬으로써 참 신기하고 기뻤답니다. 그 이후로 인연이 이어져, 지금은 참 좋은 누나가 생긴 것 같이 느껴져 더 좋아요. 누나가 저에게 사업하면 딱 맞을 사람이라고 이야기하곤 했었는데, 어쩌다 보니 그 말이 현실이 됐네요. 누나가 그렇게 말해줘서 뭐 하나 익숙한 것 없는 일에 대해 '딱 맞겠지. 뭐!'하고 용기 낼 수 있어요, 고마워요.

마지막으로 항상 응원해주고 사랑해준 가족에 감사합니다.

아버지. 어릴 때는 눈에 보이는 것만 볼 줄 알아 몰랐지만, 나이가 들어 아버지가 얼마나 가족을 위해 헌신해왔는지 알았습니다. 아버지가 은퇴하시니 제가 일을 시작하네요. 아버지가 오랜 세월 들고 계셨던 바통을 넘겨받는 느낌이 들어 기분이 좋습니다. 과거에도 지금도 앞으로도 아버지는 제 마음속의 지주입니다. 항상 든든한 우리 아버지, 감사합니다.

어머니. 맨날 엄마한테 전화해서 퇴근하는 길에 과자 사 와 줄 수 있는지 물어보던 아이가 서울대학교 박사로 성장할 수 있었던 것은 모두 어머니 덕분입니다. 힘들고 어려웠던 시절부터 많은 우여곡절을 이겨내고 지금까지 정말 고생 많으셨어요. 하나부터 열까지 잘 챙겨주고도 항상 더 챙겨주려고 해주셔서 감사합니다. 호강시켜드리고 다양한 경험 함께 할 수 있도록 두 분 모두 오래오래 건강하게 계셔주세요. 언젠가 옛날처럼 한집에서 살면서 같이 시간 보내고 싶습니다.

형. 형은 내게 어릴 때도 든든했지만 시간이 갈수록 더 든든해지네. 언제부터인가 형이 나를 인정해주는 말들을 하기 시작했는데 그게 참 기뻤던 게 기억난다. 나와는 많이 다른 형의 모습들을 동경했었는데, 그런 형한테 인정을 받으니 스스로에 대해 더 자부심을 가지고 계속 열심히 할 수 있었어. 서로 다른 길을 쭉 걸어왔는데 어째 지금 형하고 내가 굉장히 닮아있다는 생각이 드네. 내 행동과 생각하는 방식을 보면 어렸을 적부터 봐온 형의 모습이 많이 녹아있다는 게 느껴진다. 든든한 형이 있어서 다행이야.

앞으로도 여러분들의 자랑스러운 아들, 동생, 제자, 친구, 동료가 될 수 있도록 더 노력하겠습니다. 감사합니다.