이학박사 학위논문

Development of protein modeling methods for structure refinement in the context of unreliable environments

신뢰도가 낮은 구조 환경에서의

단백질 모델 정밀화 방법 개발

2017 년 2 월

서울대학교 대학원

화학부 물리화학 전공

이 규 리

Abstract

Development of protein modeling methods for structure refinement in the context of unreliable environments

Gyu Rie Lee

Department of Chemistry

The Graduate School

Seoul National University

The number of experimentally determined protein structures is increasing exponentially. Based on this abundant structural information, homology modeling is now the most popular method for protein structure prediction. Still however, knowledge of high resolution structures is critical for applications using the protein structure such as drug discovery and protein design. By realizing this, protein structure refinement methods have been developed to improve the structure quality of low resolution experimental structures or model structures. Another realm of protein structure refinement is to predict the protein structure in the environment of interest, such as binding to a specific partner, when only structures resolved in different conformational states or model structures are provided.

In this thesis, four modeling methods (GalaxyLoop-PS2, GalaxyRefine2, GalaxyVoyage, and Galaxy7TM) developed in the scope of refining predicted protein structures are introduced. The methods were evolved by either extending the range of structure targeted for refinement or considering the interaction with a particular binding partner. The shared problem of these methods was that the environment of modeling was unreliable due to errors embedded in model structures. Commonly, two approaches were taken to tackle this problem. These were initially searching the conformational space in low resolution and developing a hybrid energy function less sensitive to environmental error. The development and application results of the approaches taken for each modeling method will be addressed in detail.

keywords: Protein model refinement, protein structure prediction, loop modeling, flexible protein-ligand docking, hybrid energy function, G-protein-coupled receptor

Student Number: 2011-23233

Table of contents

List of tables

List of figures

Chapter 1

1. Introduction

Knowledge of protein structure is crucial to study the function of proteins. As a result of continuous development in experimental protein structure determination techniques, the number of protein structures revealed is on the rise (Rose et al., 2016). The enriched structure database made template-based modeling a still widely-used method for protein structure prediction. However, the structural information of the homologous protein can have limits in predicting a protein structure detailed enough to be used for applications. Even between proteins included in the same family, the functional difference can originate from structurally variant regions (Katritch et al., 2013). Therefore, *ab initio* refinement of protein models is necessary to provide high resolution information.

In **Chapter 2**, a method for protein loop modeling in unreliable structure environment is introduced (Park et al., 2014). As addressed above, local regions which are not conserved can be variable in sequence and structure. Refinement of these regions of template-based models can be done by applying protein loop modeling. However, majority of the existing protein loop modeling methods are optimized in accurate, high resolution experimental structures (Arnautova et al., 2011; de Bakker et al., 2003; Holtby et al., 2013; Jacobson et al., 2004; Liang et al., 2012). These can be sensitive to the inaccurate environment of protein loops and impede prediction. To tackle this problem, the loop conformational space was searched initially using a geometry-based method. Then, a hybrid energy function developed to be tolerant on environmental error was developed and applied.

The necessity of protein structure refinement for further application is not

confined to homology models. Protein residue contact prediction methods based on sequence information were enlightened recently (Jones et al., 2015; Ovchinnikov et al., 2014). Using the contact prediction results, the structure of proteins without templates could be predicted (Adhikari et al., 2015; Kim et al., 2014; Marks et al., 2011). The development in these template-free modeling methods is also increasing the targets of protein structure refinement.

Furthermore, recent improvement in experimental methods such as cryo-EM resulted in determinations of protein structures which could not be resolved by using other methods such as X-ray crystallography (Banerjee et al., 2016; von der Ecken et al., 2016). Yet, the resolution of these structures can yet be low. Protein structure refinement can also be applied to improve the quality of these structures.

Protein structure refinement methods (Lee et al., 2016) developed to improve the overall structure quality are introduced in **Chapter 3**. Global structure refinement also faces the difficulty originated by the unreliable structure environment. Protein modeling methods or energy functions optimized in the accurate structure cannot be effective. To tackle this problem, likewise the approaches taken for protein loop modeling in unreliable environments, conformational space search in low resolution followed by optimization with a hybrid energy function were tested.

In nature, proteins undergo conformational change as its environment changes. For example, for a same protein, its conformation would be different as it binds to different ligands (Ding and Dokholyan, 2013; Katritch et al., 2013). For a protein structure to be practically useful in applications such as drug discovery or protein design, protein modeling methods should be able to predict the protein structure in the desired state (Ding and Dokholyan, 2013). Up to now, most of the

protein structure refinement methods focused on improving the structure accuracy defined as similarity to the high resolution experimental structure (Chopra et al., 2008; Fan and Mark, 2004; Gront et al., 2012). Yet, it is impossible to experimentally determine all the possible conformational states of the protein. Therefore, refining the structure of a protein resolved in different states or model structures to predict the structure in a particular state becomes another important problem which current protein structure refinements should face.

Flexible docking of G-protein-coupled receptor and ligands was tackled by developing a method based on complex structure refinement (Lee and Seok, 2016), as described in **Chapter 4**. For a given single GPCR model structure, different complex structures are predicted as a result of docking different ligands. This can be one method particularly targeting the GPCR family to tackle the problem addressed above. The same problem of utilizing the unreliable model structure as a ligand docking environment appears in this study. To overcome this difficulty, low resolution ligand docking was performed on an ensemble of receptor conformations and further complex structure refinement was achieved by using a hybrid energy function.

As stated above, the protein modeling methods developed by this study share a common problem, which is that the modeling environment is unreliable. Common approaches were taken to tackle this: initial conformational space search in low resolution, less dependent on the energy and development of a hybrid energy function tolerant on environmental errors. Discussing the development and application results of these approaches is the scope of this thesis.

Chapter 2

2. Protein loop modeling in unreliable structural environments

2.1. Introduction

Protein loops are usually involved in the protein functional regions (Decanniere et al., 1999; Fiser et al., 2000; Saraste et al., 1990). A high-performing loop modeling method can be an invaluable tool for *de novo* protein design or designing small molecules involved in interaction with the protein loops. Template-based modeling methods have limits in tackling the protein loop modeling problem since the functional regions of proteins tend not to be conserved. Therefore, *ab initio* loop modeling methods have been developed for these purposes. Most of these methods were developed in the environment of high resolution crystal structures (Holtby et al., 2013; Jacobson et al., 2004; Liang et al., 2012). However, the practical applications which current loop modeling methods face are often situated in non-ideal conditions (Amaro et al., 2007; DiMaio et al., 2011; Mas et al., 1992). These conditions include the loop modeling environment in low resolution experimental structures, ensemble of NMR structure, structures in different states, or predicted model structures.

Some recent loop modeling methods extended the sampling region to the environment of the loop such as neighbor side chains (Mandell et al., 2009; Sellers et al., 2008; Stein and Kortemme, 2013). However these methods can target only a limited number of situations, which can still be non-realistic. The loop modeling environment can contain error originating from structures perturbed in backbone level, or the framework itself can be a predicted model structure. To target these

problems, two approaches were taken in this study. The large conformational space of the loop was initially searched geometrically, less dependent on the energy function used (Lee et al., 2010). Also, instead of extending the sampling region of loop modeling, a hybrid energy function composed of physics-based components and knowledge-based components was designed (Park et al., 2014). Due to the smooth energy landscape of knowledge-based energy functions, this hybrid energy function resulted to be more tolerant on loop environmental errors. This loop modeling method developed to target loops in unreliable environments was tested to refine template-based models.

2.2. Methods

2.2.1. Development of a new hybrid energy function

To perform protein loop modeling in unreliable structural environments, a new hybrid energy function composed of molecular mechanics based energy terms and knowledge-based energy terms was developed. Knowledge-based energy functions such as dipolar-DFIRE (Yang and Zhou, 2008a) have smooth energy landscapes. Due to this nature, it is widely used in applications such as quality assessment of decoy conformations. In this work, to develop an energy function which is less sensitive to environmental errors, knowledge-based energy functions were used together with molecular mechanics based energy functions.

The energy function for loop modeling is consisted of the energy terms as described in the following. The molecular mechanics based energy part consists of the bonded terms, van der Waals, and Coulomb interactions from the CHARMM22 force field (Best et al., 2012; MacKerell et al., 1998) mapped on to a unified

topology and the FACTS GB/SA solvation free energy (Haberthur and Caflisch, 2008; Qiu et al., 1997). The FACTS solvation free energy is a simplified version of the Generalized Born solvation model. The Born radius of each atom is dependent on the environment of each atom. The knowledge-based terms used in the energy function consists of the dipolar-DFIRE term (Yang and Zhou, 2008a), hydrogen-bond energy developed by Kortemme *et al*. (Kortemme et al., 2003), and torsion angle correction terms (Dunbrack and Karplus, 1993).

The relative weights of each energy terms were decided by training. Initially, the weight parameters that give the best decoy discrimination performance were searched by grid search. Decoy conformations were constructed for the training set consisting of 28 targets. These targets and decoy conformations were adopted from the previous work (Park and Seok, 2012).

2.2.2. Initial loop conformation sampling

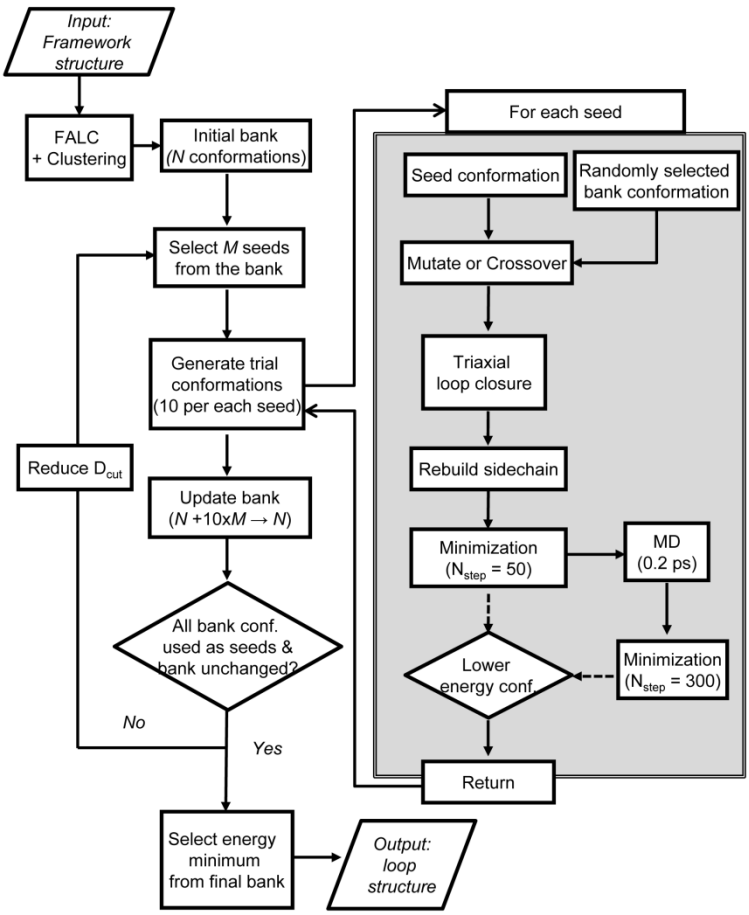
The whole process of protein loop modeling is summarized in **Figure 2.1**. It can be divided into two stages. The first stage is for initial loop conformation sampling. Before optimizing the models with the developed energy function, diverse loop conformations were generated using the fragment assembly and loop closure (Lee et al., 2010). By comparing to the fragment library generated based on query sequence, the fragments which the torsion angle less deviate from the current C-terminal torsion angle of the growing loop were randomly picked and added. The loops generated by assembling the selected fragments were closed by using the triaxial loop closing algorithm (Coutsias et al., 2004). This algorithm analytically solves to give 6 torsion angles of the loop which guarantees the loop to be closed towards the C-terminal framework. The generated conformations are filtered into quarter, first half with Ramachandran energy and another half with the knowledge-

based potential, dDFIRE. The remaining conformations were clustered into the number of bank conformations to be used in the following global optimization step. Typically, 2000 conformations were generated by using the fragment assembly and loop closure method and they were clustered into 30 or 50 conformations.

2.2.3. Global optimization using conformational space annealing

The initial loop conformations generated by using fragment assembly and loop closure were further optimized using the global optimization method using conformational space annealing (Lee et al., 1997; Park et al., 2011). For each iteration, the representative loop conformations, called 'bank' conformations were updated. New loop conformations were generated by mutating or crossing over the loop backbone torsion angles of the bank members. Representative conformations with lower energy were selected as 'seeds' and used to generate new conformations in the next iteration. The new bank was constructed by selecting the conformation with lower energy but is distant from each other, when the distance was defined by using the hamming difference of the torsion angles. The cutoff of this distance measure to define the other conformation as similar was decreased gradually for convergence.

Figure 2.1. Flowchart of the GalaxyLoop-PS2 protocol



2.2.4. Generation of test sets with various range of structural error

Different test sets with wide range of loop environment error were collected or generated to test the performance of the protein loop modeling method.

One type of the test set is based on the loop modeling benchmark set tested in Sellers *et al*. (Sellers et al., 2008), and Mandell *et al*. (Mandell et al., 2009). This set consists of high resolution crystal structures with 8-resdiue or 12-resiue loops. Applying protein loop modeling on this set is to test the ability of native reconstruction, or loop modeling in highly accurate structural environments.

The second type of the test set is also taken from Sellers *et al*. Same crystal structures from the first type of the set was used to deliberately perturb the loop environment sidechains.

The third type of the test set was generated by using the crystal structures from the first type set. To generate protein loop modeling targets with bigger environment error, the crystal structures underwent 2-ns molecular dynamics simulations with the AMBER package (Case et al., 2005). The loop environment including the backbone atoms were perturbed, inducing larger error.

The fourth type of test set has the largest environment error. This test consists of template-based models of protein sequences from the HOMSTRAD set (Mizuguchi et al., 1998). With the multiple sequence alignments taken from the HOMSTRAD benchmark study (Braberg et al., 2012), template-based models were made using MODELLER (Sali and Blundell, 1993). From the HOMSTRAD set, only the targets with model quality between GDT-TS (Kopp et al., 2007) 70 and 90 were considered. The loop regions to model were selected with unreliable local region prediction method described in Park *et al*. (Park and Seok, 2012). To

eliminate the effect of considering other interacting counterparts, loop regions interacting with ligands or other protein chains and those having crystal contacts were neglected. Finally, this resulted in 23 targets.

2.3. Results and discussion

2.3.1. Energy function for protein loop modeling

The energy function used for protein loop modeling in this study is composed like the following. This hybrid energy function is a combined form of physics-based energy terms and knowledge-based energy terms.

$$E_{PS2loop} = E_{physics-based} + E_{knowledge-based}$$

$$E_{physics-based} = E_{bonded} + E_{vdw} + w_{electrostatics}(E_{Coulomb} + E_{FACTS,GB}) + w_{FACTS,SA}E_{FACTS,SA}$$

$$E_{knowledge-based} = w_{\phi/\varphi}E_{\phi/\varphi} + w_{\chi}E_{\chi} + w_{hbond}E_{hbond} + w_{dDFIRE}E_{dDFIRE}$$

The weight parameters for each energy term were determined to be 0.16, 0.05, 1.2, 1.0, 4.0, and 12.0 for $w_{electrostatics}$, w_{SA}, $w_{\phi/\varphi}$, w_{χ}, w_{hbond}, and w_{dDFIRE}, respectively.

The contributions of each energy term were evaluated and the result is reported in **Table 2.1**. Contribution of the energy function was defined by multiplying each energy weight and the standard deviation of energy for the decoy conformations. Also, the Pearson correlation coefficient between the energy values and RMSD values of the protein loop decoy conformations were evaluated to

check the contribution. This is shown in **Table 2.2**. The same protein loop decoy conformations used for energy function weight parameter training was used.

Table 2.1. Contribution of each energy component

Energy component	$E_{coulomb}$	$E_{FACTS,GB}$	$E_{FACTS,SA}$	$E_{\varphi/\psi}$	E_{χ}	E_{hbond}	E_{dDFIRE}
Weight[1)]	0.16	0.16	0.05	1.2	1.0	4.0	12.0
Contribution(%)[2)]	7.6	9.6	6.8	19.9	10.0	7.9	38.1

1) Energy weight used for each component

2) Contribution is defined as the standard deviation of each energy term for the training set decoy conformations averaged over training loop target multiplied by the energy weight. These values are then normalized.

Table 2.2. Correlation between energy and decoy loop RMSD calculated using different subsets of energy components

1) Energy weight used for each component to generate different subsets of the energy components

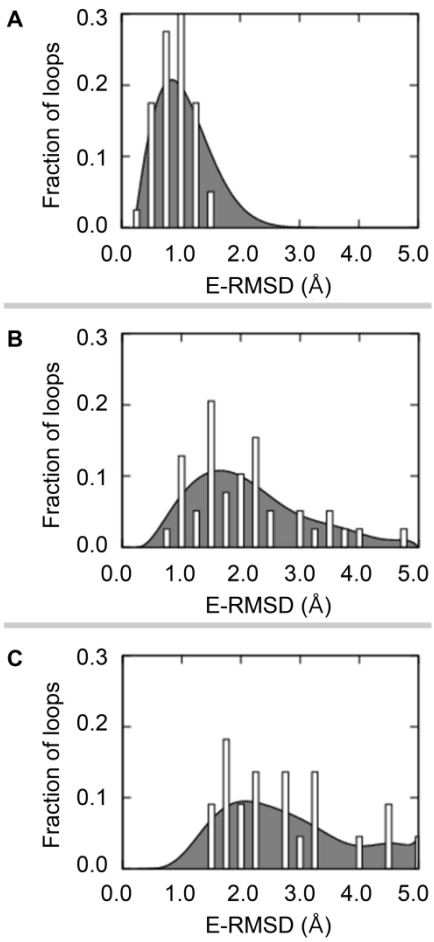
Energy function subset								Correlation[2]
Components	$E_{\text{Coulo-mb}}$	$E_{\text{FACTS,GB}}$	$E_{\text{FACTS,SA}}$	$E_{\varphi/\psi}$	E_{χ}	E_{hbond}	E_{dDFIRE}	
Weight combinations[1]	-	-	-	-	-	-	-	0.03
	0.16	0.16	-	-	-	-	-	0.10
	-	-	0.05	-	-	-	-	0.12
	0.16	0.16	0.05	-	-	-	-	0.22
	-	-	-	-	-	-	12.0	0.52
	-	-	-	1.2	-	-	12.0	0.48
					1.0	-	12.0	0.52
	-	-	-	-	-	4.0	12.0	0.48
	-	-	-	1.2	1.0	4.0	12.0	0.47
	0.16	0.16	-				12.0	0.52
	0.16	0.16	-	-	1.0	-	12.0	0.52
	0.16	0.16	0.05	-	1.0	-	12.0	0.50
	0.16	0.16	0.05	1.2	1.0	4.0	12.0	0.46

2) Pearson correlation coefficient calculated between the training set decoy loop Cα RMSD and energy and averaged over the training set loops

2.3.2. Environmental error of the test set

The degree of loop environment error of each test set could be evaluated by calculating the all-atom RMSD to the crystal structure. The environment of the loop was defined as a set of residues which have at least one atom within 10Å to the loop C_β atoms. The deviation of the loop environmental structure to the crystal structure was calculated after superimposing the two structures using TM-score (Zhang and Skolnick, 2004). The average E-RMSD values were 0.9 Å, 2.1 Å, and 2.8 Å for the side chain-perturbed set (**Figure 2.2A**), the backbone-perturbed set (**Figure 2.2B**), and the template-based model set (**Figure 2.2C**), respectively.

Figure 2.2. Distributions of environmental errors for the three types of test sets employed in the study. The test sets were generated by perturbing (A) side chains or (B) all-atoms of the crystal structure and (C) building template-based models.



2.3.3. Loop reconstruction in the crystal structure framework

The performance of the loop modeling method compared to other methods for three types of test set is summarized in **Table 2.3**. The method introduced in this chapter, named GalaxyLoop-PS2 is compared to another method HLP (Jacobson et al., 2004), which the energy function used is mostly based on molecular mechanics energy terms. In Sellers *et al.* (Sellers et al., 2008), the extended version of the method HLP, named HLP-SS is introduced. This protein loop modeling method was developed to tackle the problem of performing loop modeling in inaccurate structural environments. They applied extensive sampling by extending the sampling region to the environmental sidechains. Also, the protein loop modeling method was compared to the state-of-the art loop modeling method, NGK (Stein and Kortemme, 2013). This method also utilizes a hybrid type energy function but does extensive sampling involving environment sidechains.

The protein loop modeling method was first tested on the test set with targets having crystal structure environments. The result can be found in the 'Crystal' row of **Table 2.3**. For the crystal loop reconstruction test, GalaxyLoop-PS2 performs comparably to other methods.

Table 2.3. Comparison of loop modeling results by the average RMSD of main chain atoms (N, $C_{\square}$, C, and O) of loops in angstroms (Å) on test sets of varying environmental accuracies measured by E-RMSD. Standard deviations are also reported.

Frame-work	Loop set (No. residue)	E-RMSD (Å)	Loop Sampling[1]			Extended Sampling[1]		
			GalaxyLoop		HLP[2]	HLP-SS[2]	Rosetta KIC[3]	NGK[4]
			PS2	PS1				
Crystal	Set 1 (8)[5]	0	0.9	1.3	1.2	1.4	-	0.5
	Set 1 (12)[6]	0	1.6	2.4	1.2	1.4	1.9	1.7
	Set 2 (12)[7]	0	2.5	3.2	-	-	2.2	2.0
SCpert	Set 1 (8)[5]	0.9	1.3	1.8	2.4	1.3	-	0.5
	Set 1 (12)[6]	1.0	2.1	3.0	2.6	1.7	1.6	1.7
BBpert	Set 1 (8)[5]	1.9	2.0	2.2	-	-	-	2.1
	Set 1 (12)[6]	2.2	2.1	3.2	-	-	-	2.3

1) Loop sampling methods sample only the loop region, while extended sampling methods sample surrounding side chains in addition to the loop.
2) Taken from Sellers *et al*. (Sellers et al., 2008)
3) Taken from Mandell *et al*. (Mandell et al., 2009)
4) Results of the best-score models out of 500 models sampled for each target following the protocol provided by Stein *et al*.(Stein and Kortemme, 2013) with Rosetta v3.5. The results for the crystal structure set and the side chain-perturbed set are the same for NGK because extended sampling of loop environment was used for both sets.
5) Loop sets taken from Jacobson *et al. (Jacobson et al., 2004)*

6) Loop sets from Zhu *et al*. (Zhu et al., 2006b)
7) Loop set from Fiser *et al*. (Fiser et al., 2000)

2.3.4. Loop modeling in sidechain perturbed environment

Here, how the perturbed loop environment affected the loop modeling performance will be discussed. The details of the result are reported in the 'SCpert' row in **Table 2.3**. Interestingly, the performance of GalaxyLoop-PS2 for the sidechain perturbed test set is not noticeably worse than the result achieved for the crystal reconstruction test set. Increases in average main chain RMSD values compared to those obtained from the native reconstruction set are 0.4 Å (from 0.9 Å to 1.3 Å) and 0.5 Å (from 1.6 Å to 2.1 Å). This can be compared to the loop environment RMSDs of the sidechain perturbed sets, which are 0.9 Å and 1. Å. This shows that the hybrid energy function developed for this purpose is tolerant to loop environment errors.

The protein loop modeling method HLP, which the energy function used is based on molecular mechanics, performed worse in this set compared to the crystal structure reconstruction set. It resulted in RMSD values of 2.4 Å and 2.6 Å compared to 1.2 Å and 1.2 Å got from the crystal set, respectively. The different results of the two methods, GalaxyLoop-PS2 and HLP can be pinpointed by observing the examples illustrated in **Figure 2.3A**. GalaxyLoop-PS2 could generate lowest-energy models with loop RMSD 0.4 Å and 0.5 Å for 1oyc and 1c5e with perturbed sidechains. However, for HLP, it resulted in loop RMSD 2.2 Å and 1.8 Å. The perturbed Arginine sidechain in the loop environment breaks the salt bridge between the loop and framework. The electrostatics energy described by physics-based energy terms only in HLP seemed to be strongly dependent on short range local geometry. Also, Generalized Born solvation model included in the HLP energy function may over-stabilize the salt bridge interaction, as reported in some other studies before (Geney et al., 2006). The GalaxyLoop-PS2 energy function

also employs a GB solvation model but other knowledge-based energy terms seem to compensate the sensitivity of the physics-based energy terms towards local error.

The performance of GalaxyLoop-PS2 was slightly worse than the methods employing extensive sampling including the environment sidechains for this test set. A typical case for GalaxyLoop-PS2 resulting in inaccurate prediction is illustrated in **Figure 2.3B**. The perturbed sidechains in the loop environment would result in steric clashes when the native-like loop conformation is generated. Therefore, loop modeling without sampling the environment would be impossible to model loops in these situations.

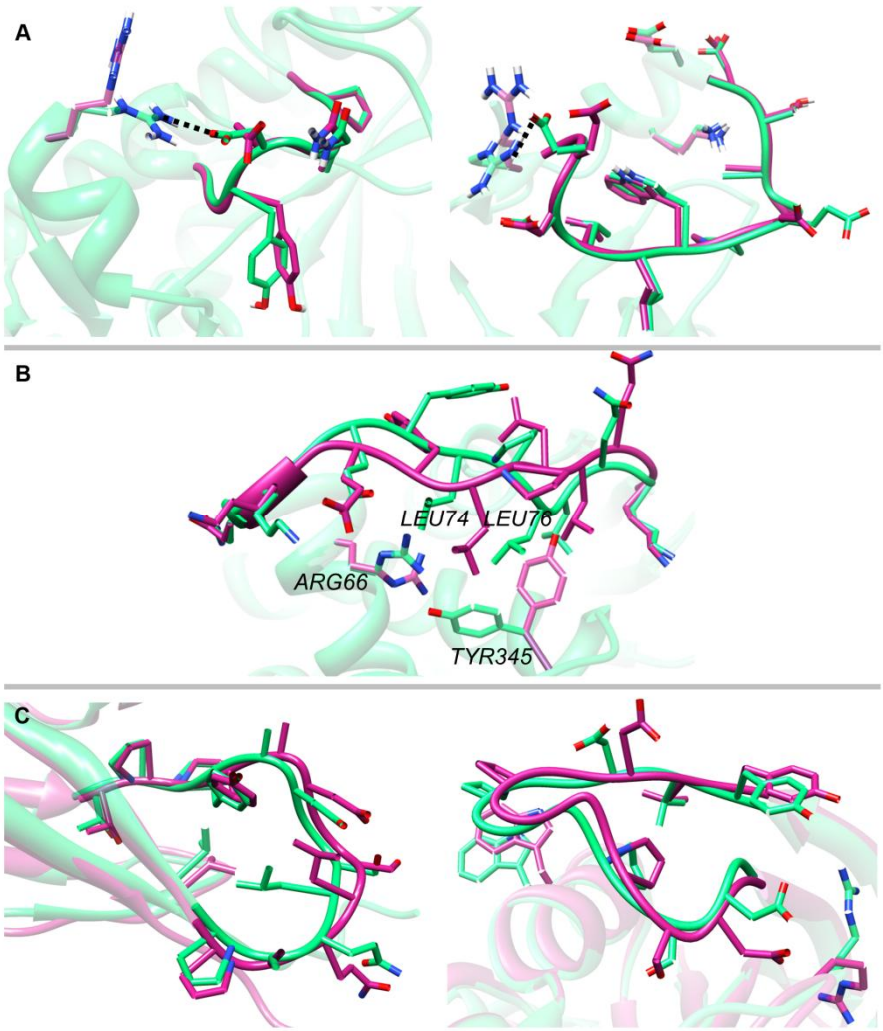
2.3.5. Loop modeling in backbone perturbed environment

Protein loop modeling method was tested on the test set with more environmental error, including errors in the backbone atoms. The test set targets include errors from thermal fluctuation generated by running molecular dynamics simulations. There were previous studies which the negative effects of distorted neighbor regions of the loop on loop modeling were shown (Fiser et al., 2000; Fiser and Sali, 2003; Subramani and Floudas, 2012). In this test, the performance of GalaxyLoop-PS2 was compared to that of the state-of-the art method, NGK (Stein and Kortemme, 2013). The average results of each method can be compared in the 'BBpert' row in **Table 2.3**.

Loop environmental error of this test set is 2 Å, increased from 1 Å of the sidechain perturbed test set. The loop environment of this test set includes the perturbation of the sidechain and the backbone atoms. However, increases of the prediction errors are smaller than this increase for GalaxyLoop-PS2, which are 0.7 Å and 0.0 Å for the 8 residue and 12 residue sets, respectively. When this result is compared to that of the NGK method, it is comparable even when NGK employs

extensive sidechain sampling of the erroneous loop environment. In **Figure 2.3C**, two examples of the GalaxLoop-PS2 predictions on the backbone perturbed test set are shown. While the loop environment errors of these two targets are high, 4.0 Å and 2.7 Å, the protein loop modeling results were relatively accurate, resulting in loop RMSD values of 1.0 Å and 0.9 Å, respectively. Likewise the cases observed in the sidechain perturbed test set, cases with environment error due to broken short-range salt bridge tolerated could be observed in this test set, also. However, for the targets which large environmental perturbations that can cause steric clash, the GalaxyLoop-PS2 could not predict the loop structures well.

Figure 2.3. Examples of loops modeled in inaccurate environmental structures. In all panels, the crystal structures are colored in green and the models in magenta. Framework structures are shown transparent for clarity.



2.3.6. Loop modeling on template-based models

The test set generated by template-based models is the set containing largest error. The loop environment RMSD of this test set ranges from 1.6 Å to 5.3 Å. The test result on this test is compared to other modeling methods in **Table 2.4**. From the twenty three targets, seven targets could be predicted with loop RMSD <3 Å and the other three with loop RMSD <2 Å. The loop structures predicted by using GalaxyLoop-PS2 resulted in average loop RMSD 3.7 Å. This is more accurate than the loop structures from MODELLER (4.2 Å) (Sali and Blundell, 1993) and the modeled loops using ModLoop (4.0 Å) (Fiser and Sali, 2003). The result of GalaxyLoop-PS2 was comparable to the method NGK which incorporates extensive sampling of the environment (Stein and Kortemme, 2013). Loop modeling using NGK resulted in average loop RMSD of 3.9 Å.

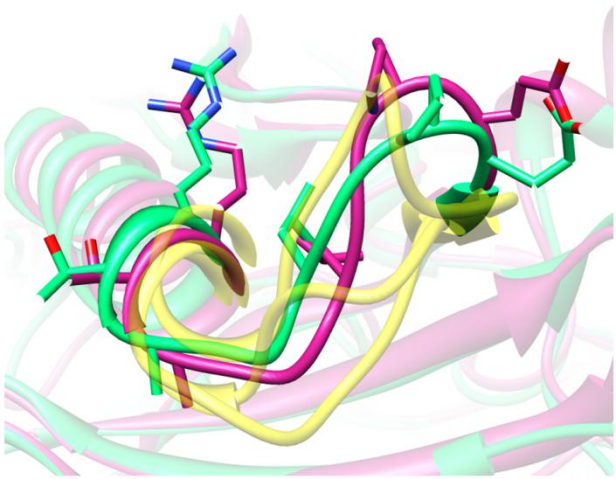
One successful example of performing loop modeling to the template-based model test set is illustrated in **Figure 2.4**. In this figure, transparent loops are from template proteins. The light colored structure is the crystal structure of the target protein and the darkest structure is the template-based model of the target sequence. It can be seen in the figure that while the template protein structures share almost identical folds with the crystal structure of the target protein, structures corresponding to the detected loop region is variable. This shows that ab initio modeling of the unreliable local region is necessary to generate highly accurate model structures. For the template-based model targets with more inaccurate environments could not be tolerated by using the hybrid energy function. For these cases, extending the sampling region towards the loop environment will be necessary.

Table 2.4. Comparison of loop modeling results on the test set of template-based models. The average RMSD and its standard deviation are reported in Å. The Loop RMSD is calculated as the root-mean-square deviation of the main-chain atoms N, Cα, C, and O.

Framework	E-RMSD (Å)	Loop RMSD (Å)				
		GalaxyLoop		MODELLER[1)]	Mod-Loop[2)]	NGK[3)]
		PS2	PS1			
Template-based model	3.0	3.7	3.9	4.2	4.0	3.9

1) Loop conformations generated by MODELLER (Sali and Blundell, 1993)
2) Loop conformations generated by loop refinement using ModLoop of MODELLER (Fiser and Sali, 2003)
3) Results of the best-score models sampled by Next-generation KIC (NGK) using the protocol provided by Stein *et al*. (Stein and Kortemme, 2013). 500 models were generated for each target as in Stein *et al*. The Rosetta program v3.5 was used.
4) Loop set constructed in this study.

Figure 2.4. A successful example of loop modeling in the framework of a template-based model. The crystal structure is colored in green and the model in magenta (PDB ID: 1avk, RMSD = 1.5 Å). Framework structures are shown transparent for clarity. Loops of three templates (used for template-based modeling) are shown with yellow transparent ribbons for comparison.



2.3.7. Comparison of using hybrid energy, physics-based energy, and knowledge-based energy for loop modeling

The development of a hybrid energy function to be used for protein loop modeling initiated with the idea that combining the physics-based and knowledge-based energy functions will result in the energy function containing advantages from the both. To critically compare the performance originating from different combinations of the energy terms, global optimization part of loop modeling using the same initial bank but with different energy functions, the hybrid, only physics-based, and only knowledge-based, was performed. The weight parameters for the energy function using knowledge-based terms only were 0.0, 0.0, 1.2, 1.0, 4.0, and 12.0 for $w_{\text{electrostatics}}$, w_{SA}, $w_{\varphi/\psi}$, w_{χ}, w_{hbond}, and w_{dDFIRE}, respectively. For the energy function using physics-based terms only, the weights used were 0.25, 0.02, 0.0, 0.0, 0.0, and 0.0 for $w_{\text{electrostatics}}$, w_{SA}, $w_{\varphi/\psi}$, w_{χ}, w_{hbond}, and w_{dDFIRE}, respectively.

The test sets with 12-residue loops from the same benchmark test sets were used for comparing the different energy functions. The results of using different energy functions are shown in detail in **Table 2.5**. When tested on the 12-residue crystal structure set, protein loop modeling using the hybrid energy function performed best or comparable to using the knowledge-based energy function for all the test sets with different range of error. Protein loop modeling using the hybrid energy function resulted in average loop RMSD of 1.6 Å, 2.1 Å, and 2.1 Å. Using the energy function with knowledge-based energy terms only resulted in loop RMSD of 1.7 Å, 2.3 Å, and 2.0 Å. When using the energy function based on molecular mechanics, loop RMSD of 2.1 Å, 3.0 Å, and 2.8 Å.

It is interesting to see that using the energy function consisting only of

CHARMM22 force field-based energy terms showed the worst result (loop RMSD 2.1 Å, 3.0 Å, and 2.8 Å) among the three compared energy functions for loop modeling. Also, this performance is worse compared to that of HLP (loop RMSD 1.2 Å and 2.6 Å for the 'Native' and the 'SCpert' set, respectively), which can be found in **Table 2.3**. The molecular mechanics-based energy terms we are using are based on united atom topology which only considers the hydrogen atoms bound to polar atoms. Although the parameters are taken from the CHARMM22 force field in all-atom topology and mapped to the topology we are using. We also use the FACTS solvation model, which is an approximate Generalized Born solvation model. Also, it has to be noted that for the energy functions using only physics-based or knowledge-based terms, the weight parameters have to be optimized again, but for this study, the weight parameters from the hybrid energy function were simply adopted or modified slightly.

Table 2.5. Performance of loop modeling using energy functions composed of physics-based terms, knowledge-based terms, and all terms. The crystal structure reconstruction results and modeling results on perturbed crystal structures for the 12-residue loop are shown.

Frame-work	Set number (No. of residue)	E-RMSD (Å)	Loop RMSD (Å)		
			Galaxy-Loop-PS2	PS-know-ledge[1)]	PS-physics[2)]
Crystal	Set 1 (12)	0.0	1.6	1.7	2.1
SCpert	Set 1 (12)	1.0	2.1	2.3	3.0
BBpert	Set 1 (12)	2.2	2.1	2.0	2.8

1) Results of the lowest-energy model structures obtained by the same procedure as GalaxyLoop-PS2 but using the energy function mainly composed of knowledge-based energy terms, named here as PS-knowledge. The weight parameters are set to ($w_{\text{electrostatics}}$, w_{SA}, $w_{\varphi/\psi}$, w_{χ}, w_{hbond}, w_{dDFIRE}) = (0.0, 0.0, 1.2, 1.0, 4.0, 12.0).

2) Results of the lowest-energy model structures obtained by the same procedure as GalaxyLoop-PS2 but using the energy function composed of only physical energy terms, named here as PS-physics. The weight parameters are set to ($w_{\text{electrostatics}}$, w_{SA}, $w_{\varphi/\psi}$, w_{χ}, w_{Hbond}, $w_{\text{atom-pair}}$) = (0.25, 0.02, 0.0, 0.0, 0.0, 0.0).

2.4. Conclusion

Protein loop modeling method which can be applied to targets in unreliable structural environments was developed. The conformational space of the loop residues were searched using fragment assembly and loop closure. This initial stage of loop conformational sampling allows conformational space search which is less dependent on the energy function while in the unreliable environment. To optimize the initially searched loop conformations, a hybrid energy function composed of physics-based and knowledge-based energy terms was developed.

When tested on the set constructed with environmental error, this method was comparable or better than other loop modeling methods which the search space is extended to the environment side chains. Still, there were environmental errors which could not be tolerated by using the hybrid energy function only. For these cases, extension of the conformational search space to the loop environment should be pursued.

Chapter 3

3. Global refinement of protein model structures

3.1. Introduction

3.1.1. Extension of the range of structure targeted for refinement

In the previous chapter, development of a protein loop modeling method was introduced. We tried to tackle the problem of performing loop modeling in unreliable environments (Park et al., 2014). This modeling condition realistically reflects the situation faced when protein loop modeling is applied in practical problems. We tested our method on different test sets covering wide range of environmental error. Our test set with the biggest error consisted of template-based models. Initially searching the conformational space of the loop less dependent on the energy and developing a new hybrid energy function were adopted to tackle this problem.

The hybrid energy function we developed to perform protein loop modeling in unreliable environments was particularly successful on problems where local interactions observed in crystal structures were disturbed by geometry difference. This effect was induced by using knowledge-based energy terms in addition to physics-based energy terms. However, we found that there are cases which the environment error cannot be tolerated by using the hybrid energy function. We assumed that the error of these cases will be recovered when the conformational search space of modeling is extended to the loop environment. Also, when this approach can be extended as a global structure refinement method, it will have potential in improving the structure quality of the protein model.

In this chapter, two modeling methods aiming for global structure refinement will be introduced. The two methods were developed sequentially. The first refinement method employs global refinement of the model structure based on molecular dynamics relaxation with structure perturbation (**Section 3.2**) (Lee et al., 2016). Unreliable local regions of the given initial model structure were refined with the protein loop modeling method introduced in **Chapter 2** and the locally refined conformation underwent global refinement. When this scheme is applied, structure refinements focused in local region and overall region are not concurrent. The idea underlying this approach was that the change of conformation resulted by protein loop modeling will be propagated to the global refinement stage. The global structure refinement method based on local structure perturbation and overall molecular dynamics relaxation resulted in improving the structure quality in both global and local measures. However, significant change in backbone level could not be induced just by applying this method.

The model structure refinement method described in **Section 3.2** targets the unreliable local region first and applies global refinement. Although the conformational search space for refinement was extended to the overall structure in this method, the conformational search space for protein loop modeling was not directly extended. This could have limited refinement because the protein loop modeling performance is dependent on the inaccuracy of the environment. To tackle this problem, another type of protein model structure refinement method was developed. This method is described in **Section 3.3**. Here, protein loop modeling was incorporated as a part of the overall refinement method after the unreliable local regions were automatically predicted. This global refinement method also incorporates other various types of modeling methods with different degree of conformational change. Also, an improved hybrid energy function was developed

to be used with this method. This method succeeds in producing more diverse conformations compared to the previous refinement method.

3.1.2. Conformational search methods and energy functions for global refinement

Currently, molecular dynamics simulation with molecular mechanics force fields such as AMBER, CHARMM, and OPLS-AA is the mainstream approach for protein structure refinement (Best et al., 2012; Chen and Brooks, 2007; Cornell et al., 1996; Ishitani et al., 2008; Jorgensen et al., 1996; Lee et al., 2001; Mirjalili and Feig, 2013; Raval et al., 2012). These MD simulations in all-atom representation with explicit solvent molecules in consideration can effectively optimize local interactions in a protein structure and produce physically reliable structures. However, massive computation is necessary to search the large conformational space of a protein (Fan and Mark, 2004). Also, due to flaws in current force fields and high energy barriers, the refined models can drift away, being inaccurate, as simulations proceed. This problem has been alleviated by applying additional restraints to the initial structure and proven to be successful in consistently improving the qualities of models (Gront et al., 2012; Mirjalili et al., 2014; Raval et al., 2012).

Another stream of refinement is using knowledge-based potentials. These include DFIRE (Zhou and Zhou, 2002; Zhu et al., 2008; Zhu et al., 2006a), dDFIRE (Mirjalili and Feig, 2013; Yang and Zhou, 2008a), and RAPDF (Rodrigues et al., 2012; Samudrala and Moult, 1998; Summa and Levitt, 2007). The knowledge-based potentials were used for sampling and/or scoring for refinement. These potentials have smooth energy landscapes compared to physics-based energy functions. Therefore, protein conformational search using the

knowledge-based energy function is more efficient. Also, some physical aspects which is hard to be described using non-polarizable force fields such as hydrogen bonding are described better (Kortemme et al., 2003). They are also used in combination with physics-based energy functions (Heo et al., 2013; Jagielska et al., 2008; Zhang et al., 2011).

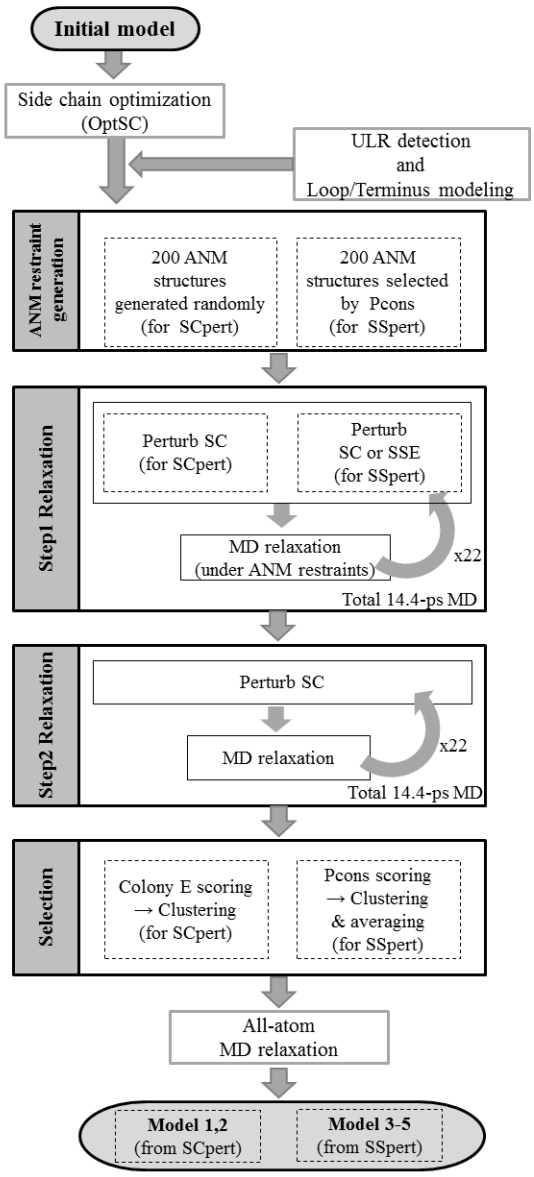
The refinement method introduced in **Section 3.2**, GalaxyRefine2 (Lee et al., 2016) is based on loop modeling and overall molecular dynamics relaxation. The method introduced in **Section 3.3**, GalaxyVoyage incorporates diverse types of protein modeling methods followed by short molecular dynamics relaxation and gradually search lower energy conformations. These both methods use a hybrid energy function composed of physics-based energy terms and knowledge-based energy terms. These conformational space searching method and the energy function were developed to result in a refinement method which can overcome the deficiency of current force fields and yet be computationally effective.

3.2. Global refinement based on loop modeling and overall relaxation

3.2.1. Methods

The refinement method described in this chapter consists of six steps. The method is illustrated in **Figure 3.1**. Initially, only the sidechains are optimized. When unreliable local regions exist in the given protein model, loop modeling method described in **Chapter 2** was applied. Two methods based on molecular dynamics relaxation which use different types of local structure perturbations were developed. Conformations were randomly generated using anisotropic network model (ANM) (Bakan et al., 2011) and used as guiding structures for the two methods. The selected conformations among the sampled conformations finally underwent optimization in full atom representation.

Figure 3.1. Flowchart of the refinement method



3.2.1.1. The energy function used for MD relaxation

In **Chapter 2**, a hybrid energy function was developed for protein loop modeling in unreliable environments. Following this approach, the energy function for global refinement was developed likewise. This energy function consists of the same energy terms use for loop modeling. It is a hybrid energy function based on knowledge-based energy terms and physics-based energy terms. It also includes a restraint energy function. The energy function composition is like the following.

$$E = E_{physics-based} + E_{knowledge-based} + E_{restraint}$$

$$E_{physics-based} = E_{bonded} + E_{vdw} + w_{electrostatics}(E_{Coulomb} + E_{FACTS,GB}) + w_{FACTS,SA}E_{FACTS,SA}$$

$$E_{knowledge-based} = w_{\phi/\varphi}E_{\phi/\varphi} + w_{\chi}E_{\chi} + w_{hbond}E_{hbond} + w_{dDFIRE}E_{dDFIRE}$$

$$E_{restraint} = w_{restraint,dist}E_{restraint,dist} + w_{restraint,cart}E_{restraint,cart}$$

Like the energy function used for loop modeling, the physics-based energy function consists of molecular mechanics term from CHARMm22 force field mapped on to CHARMm19 topology (Best et al., 2012). FACTS implicit solvation model is used (Haberthur and Caflisch, 2008). The knowledge-based function is mainly contributed by the dipolar-DFIRE potential (Yang and Zhou, 2008a). The other terms include the bias potentials which favor the backbone or side chain torsion angles that frequently occur in nature (Dunbrack and Karplus, 1993). These potentials are generated based on the data collected from the structure database. Also, the knowledge-based potential describing the hydrogen bond geometrically is included (Kortemme et al., 2003). The weight parameters used in

GalaxyRefine (Heo et al., 2013) were adopted.

The additional term included in the energy function compared to that used for loop modeling is the potential for restraint energy. The restraint energy functions used for refinement can be written in detail as follows.

$$E_{restraint,dist} = \sum_{i-j>1, d^0_{ij,C_\alpha-C_\alpha}<10\text{Å}} \left(\frac{d_{ij,C_\alpha-C_\alpha} - d^0_{ij,C_\alpha-C_\alpha}}{2} \right)^2 + \sum_{|i-j|>1, d^0_{ij,N-O}<10\text{Å}} \left(\frac{d_{ij,N-O} - d^0_{ij,N-O}}{2} \right)^2$$

$$E_{restraint,cart} = \sum_i \left(\frac{\left| r_{i,C_\alpha} - r^0_{i,C_\alpha} \right|}{2} \right)^2$$

The relative weights of the two energy functions were determined to show best performance on the training set (data not shown). As a result, the energy weights were used like the following: (0.2, 0.04, 2.0, 2.0, 2.0, 5.0, 5.0, and 5.0) for ($w_{electrostatics}$, $w_{FACTS,SA}$, $w_{\varphi/\psi}$, w_{χ}, w_{hbond}, w_{dDFIRE}, $w_{restraint,dist}$, and $w_{restraint,cart}$).

3.2.1.2. Optimization of initial side chain conformation

A side chain optimization method named Galaxy-optSC was applied in this step. This method solves the combinatorial problem of rotamers to optimize the given energy function. The backbone atoms are all fixed and only the different rotameric states are searched for all side chain torsion angles. In this step, the same energy function used for molecular dynamics relaxation except the restraint energy was used. To solve this combinatorial problem, this method breaks the whole protein into subsets of residues and iteratively gives solutions part by part. In detail, the method first randomly selects a third of the residues from the whole sequence. Starting from the residue that has the biggest number of neighbor residues which

are within 8 Å C_β distance, each residue subset undergoes single round of optimization. The solution is cumulated at every round until we get the final solution. This optimization step is implemented by following the method of Scwrl3 (Krivov et al., 2009). Scwrl3 is a fast sidechain optimization method based on graph theory and dead-end elimination. The best combination of the subset rotamers is searched. Rotamers that have rotamer probability higher than 1% and cumulative rotamer probability lower than 90% were all treated as candidate rotamer states. The method for building and solving the interaction graph of sidechains were adopted from Scwrl3. The energy function used for optimization and the energy cutoff to define the node connection of a graph is different between Scwrl3 and our method.

3.2.1.3. Generating conformations using anisotropic network model

Anisotropic network model (ANM) was used to generate conformations. These conformations were used to guide molecular dynamics relaxation described in the following subsections. Anisotropic network model calculation was done by using the Prody python module (version 1.5.1) (Bakan et al., 2011) with Cα- Cα distance cutoff of 12 Å. After the normal modes are calculated we can choose which mode to use and the degree of amplitude to extrapolate the atom coordinates. In this method, only one mode from the twenty lowest frequency modes was selected every time. The extrapolation amplitude was randomly selected within 20Å. Thousand conformations in total were generated by using ANM.

3.2.1.4. Overall relaxation guided by ANM (Step 1 relaxation)

Two types of overall relaxation guided by ANM were adopted to sample conformations for refinement. The first type was performing molecular dynamics relaxation with side chain perturbations. From the thousand conformations

generated using ANM, two hundred conformations were selected randomly. The other type was molecular dynamics relaxation with secondary structure element perturbations. For this type, two hundred ANM conformations were selected after scoring the conformations with Pcons (Wallner and Elofsson, 2005), a consensus method. More converged ANM conformations were chosen for this type because secondary structure perturbation induced bigger degree of movement.

For both two types of the step 1 relaxation, repetitive structural perturbations were followed by molecular dynamics relaxation of 0.6-ps. For each trajectory, twenty two perturbations were applied and resulted in a 14.4-ps MD simulation including the 1.2-ps pre-relaxation. For the first type relaxation ('SCpert'), randomly selected side chain clusters were perturbed. This type of relaxation is similar to the method used in the published version of GalaxyRefine (Heo et al., 2013). The second type relaxation ('SSpert') is applied by perturbing the secondary structure elements of the given structure. We intended to search the conformational space more diversely by giving large structural perturbations for easier energy barrier crossing. Each time of structural perturbation in 'SSpert', secondary structure element to perturb was randomly selected. If more than one secondary structure elements were interacting via hydrogen bonding, the elements were perturbed together. The type of perturbation to be applied, which is translation or rotation of the secondary structure, the axis of perturbation, and the amplitude of perturbation were also chosen randomly. The Relaxation was applied with the refinement energy function but by updating the Cα-Cα distance and Cα cartesian restraints with the values taken from the selected two hundred ANM conformations. The simulation temperature for each trajectory was slowly annealed. The MD relaxation was done in 300K for the two-thirds of one simulation and it was cooled down to be 50K during the remaining time.

3.2.1.5. Self-guided overall relaxation (Step 2 relaxation)

The conformations sampled from the first step relaxation were further optimized by undergoing another round of molecular dynamics relaxation. A total of four hundred conformations from Step 1 relaxation were relaxed with the refinement energy function. This time, the restraint was applied on each self-conformation. By biasing the global structure to itself, which is generated in Step 1, the global structure was nearly maintained while more relaxation in this step allows optimization of local structures. Overall relaxation applied in this step was same with the one used for 'SCpert' in Step 1 relaxation except for the restraint energy function. Small sets of neighbor side chains were perturbed repeatedly, likewise.

3.2.1.6. Selection of the sampled conformations

From the conformations sampled from the overall relaxation steps, representative structures were selected. The refinement energy function we used for molecular dynamics relaxation is a hybrid energy function. The largest contribution comes from the knowledge-based energy function, dipolar-DFIRE (Yang and Zhou, 2008a). This composition of relative contributions was determined to optimize the sampling performance mostly. Also, the refinement based on molecular dynamics relaxation with different degree of perturbations is not aimed for global optimization. Therefore, for the conformations generated in the previous relaxation steps, the energy minimum conformation is not guaranteed to have high quality. A different method had to be developed with the purpose of scoring and selecting the representative conformations. First we tried to select the conformations from the four hundred conformations generated by two relaxation methods with a single scoring method such as colony energy (Xiang et al., 2002). However, the conformations originated from two types of relaxations were inhomogeneous in the

curve drawn using the refinement energy function. Therefore, we had to come up with independent selection schemes for the two groups of conformations.

For the conformations generated by the first type of relaxation ('SCpert'), colony energy (Xiang et al., 2002) was applied for scoring. Colony energy is a method to evaluate free energy accounting the shape of the potential curve. Colony energy is composed of the standard potential energy term and the term favoring conformations with more neighbors. The structural similarities were evaluated using TM-score (Zhang and Skolnick, 2004). Colony energy was evaluated like the following in this study.

$$CE_i = \sum_j \left\{ \exp\left[-\alpha \left(\frac{1 - TM_{ij}}{1 - \overline{TM}} \right) \right] \exp\left[-\beta \left(\frac{E_j - \overline{E}}{\sigma(E)} \right) \right] \right\}$$

The first term represents the similarity of the other conformations compared to conformation in state i. Parameter to scale the similarity, the pairwise similarity, and the average similarity of all conformation pairs are α, TM_{ij}, and $\overline{TM}$, respectively. The second term represents the normalized energy value. After sorting the two hundred conformations using colony energy, they were clustered by NMRclust (Kelley et al., 1996). The same term used to define similarity for colony energy was used to define distance between conformations. Two lowest colony energy cluster centers were selected finally.

For the other two hundred conformations originated from the second type of relaxation ('SSpert'), the conformations were clustered and structure-averaged for each cluster. To cluster the conformations, they were first ranked with the consensus-based Pcons score (Wallner and Elofsson, 2005). To increase the

structural diversity of the scored conformations, two hundred ANM conformations used as guiding structures in the first step were added to the evaluation pool. The number of ANM conformations to include was determined by training (data not shown). The ANM conformations were only used to affect scoring and not targeted for selection. Starting from the top Pcons conformation, other conformations are compared using normalized TM-score (Zhang and Skolnick, 2004) (defined in the first term of the colony energy function). Only the conformations dissimilar (normalized TM-score < 1.0) were treated as the members of the cluster initiated by the reference conformation. Up to twenty conformations were included in the cluster. Three clusters with the highest Pcons scoring conformation as references were selected. For each cluster, the Cartesian coordinates of the member conformations were averaged. This scheme resulted in three structure-averaged conformations, finally.

3.2.1.7. Final optimization in all-atom representation

During the relaxation steps ahead of this final optimization step, the protein was represented in the united atom topology consisting of heavy atoms and hydrogen atoms which are connected to polar heavy atoms. This was to efficiently search the conformational space while also being fast in computation compared to using all-atom representation. However, if we want to focus on optimizing the local structure quality in atomic detail, using the all-atom topology is necessary. Therefore, in this last step of the whole refinement process, the molecule representation is converted into all-atom topology and the finally selected five conformations were re-optimized based on overall relaxation. Like the second round of relaxation described in the previous subsection, to maintain the global structure, strong restraint is applied during relaxation. The strength of the restraint applied in this

step is nearly twice of that applied in Step 2 relaxation. The energy function used for this step consists of the same energy terms used for the relaxations done in united atom topology. The weight parameters were re-optimized for this purpose. The weights were determined to be (0.15, 0.02, 4.0, 2.0, 2.0, 6.0, 10.0, and 10.0) for ($w_{electrostatics}$, $w_{FACTS,SA}$, $w_{\varphi/\psi}$, w_{χ}, w_{hbond}, w_{dDFIRE}, $w_{restraint,dist}$, and $w_{restraint,cart}$). Simulation time was 7.2-ps for one trajectory. For each representative structure, eight conformations were generated and the energy minimum structure was evaluated as the final structure.

3.2.1.8. The benchmark test set

This refinement method was benchmarked on a set consisting of past CASP refinement targets. This set has fifty three targets from CASP8-10 (MacCallum et al., 2009; MacCallum et al., 2011; Nugent et al., 2014). The method was also tested on the second test set consisting of 69 submitted server models to the CASP10 template-based modeling category. The server models generated by Zhang-server (I-TASSER (Yang and Zhang, 2015)) and BAKER-ROSETTASERVER (Robetta (Kim et al., 2004)) were chosen. The last set consists of 131 FG-MD benchmark set targets (Zhang et al., 2011). All targets have GDT-HA higher than 0.4.

3.2.1.9. Model refinement with unreliable local regions: loop modeling and relaxation

The subsections written above describe the components of the refinement pipeline for a single given initial protein model. This refinement method is for overall structure refinement. When unreliable local regions could be detected for the initial model, loop modeling was preceded on the side chain-optimized model structure. Loop modeling method is described in **Chapter 2**. After performing loop modeling, the energy minimum conformation was targeted for overall refinement, starting

from Step 1 relaxation.

This refinement scheme based on loop modeling and overall relaxation was benchmarked on 54 previous CASP refinement targets. Unreliable local region for these initial models were manually assigned, assuming that we already know the erroneous regions.

This method was tested in a blind manner in CASP11 refinement category (Modi and Dunbrack, 2016). There were 35 targets. For the given model structures, unreliable local regions were predicted by observing the structures.

3.2.2. Results and discussion

3.2.2.1. Overall refinement performance without loop modeling

The overall refinement method result on four benchmark test sets can be seen in **Table 3.1**. The result of the method described in this chapter (GalaxyRefine2 (Lee et al., 2016)) is compared to the result of applying GalaxyRefine (Heo et al., 2013). The degree of refinement is measured in improvements in GDT-HA (Kopp et al., 2007), GDC-SC (Keedy et al., 2009), and MolProbity (Chen et al., 2010). For GalaxyRefine2, one model each from the two types of relaxation ('SCpert' and 'SSpert') was evaluated. For GalaxyRefine, one energy minimum conformation was evaluated ('Mild' version). For all benchmark test sets and for all measures, the conformations generated by 'SCpert' of GalaxyRefine2 showed larger improvement compared to GalaxyRefine. The major difference of the 'SCpert' type relaxation used in GalaxyRefine2 and GalaxyRefine is the type of restraint applied. The effect of this difference will be analyzed in the following parts. Also, GalaxyRefine2 is different from GalaxyRefine in that it is

composed of several relaxation steps. The improvement gained in each step will also be discussed in the following.

When the best conformations among the five predictions from GalaxyRefine2 are compared to GalaxyRefine, they are also better in all measures for all sets, on average. However, for the conformations generated by 'SSpert', for three test sets out of four, the performance is worse than GalaxyRefine when evaluated in GDT-HA. When compared in GDT-HA measure, the conformations generated by 'SSpert' are not improved as much as they are by 'SCpert'. For 'SSpert' predictions, more diverse structures were sampled by perturbing the secondary structure elements. However, there were many targets which the conformations drift away from the experimental structure or the selection method didn't work well. The detailed analysis of sampling and selection results of both types ('SCpert' and 'SSpert') of relaxation will be discussed.

Table 3.1. Performance of two refinement methods on benchmark test sets. Structure quality is evaluated in GDT-HA, GDC-SC, and MolProbity.

Test set	GalaxyRefine			GalaxyRefine2 (SCpert / SSpert)		
	ΔG DT-HA	**ΔGDC-SC**	**-ΔMol-Probity**	**ΔGDT-HA**	**ΔGDC-SC**	**-ΔMol-Probity**
CASP TR set	0.10 (1.27)	1.22 (2.27)	0.53 (0.71)	0.68/0.20 (1.48)	2.41/2.03 (3.35)	1.27/1.33 (1.46)
CASP10 I-TASSER[1]	-0.20 (0.66)	2.65 (3.56)	0.54 (0.87)	0.37/-0.29 (1.12)	3.28/2.69 (4.51)	1.34/1.38 (1.53)
CASP10 ROSETTA[2]	0.00 (0.55)	-0.47 (0.02)	-0.35 (-0.28)	0.46/-0.07 (1.09)	0.05/-0.42 (0.99)	0.21/0.19 (0.31)
FG-MD bench-mark set	0.55 (1.66)	1.74 (2.91)	0.82 (1.05)	0.91/0.40 (1.96)	1.99/1.74 (3.22)	1.52/1.59 (1.73)

1) Server models generated by Zhang-server (I-TASSER (Yang and Zhang, 2015))

2) Server models generated by BAKER-ROSETTASERVER (Robetta (Kim et al., 2004))

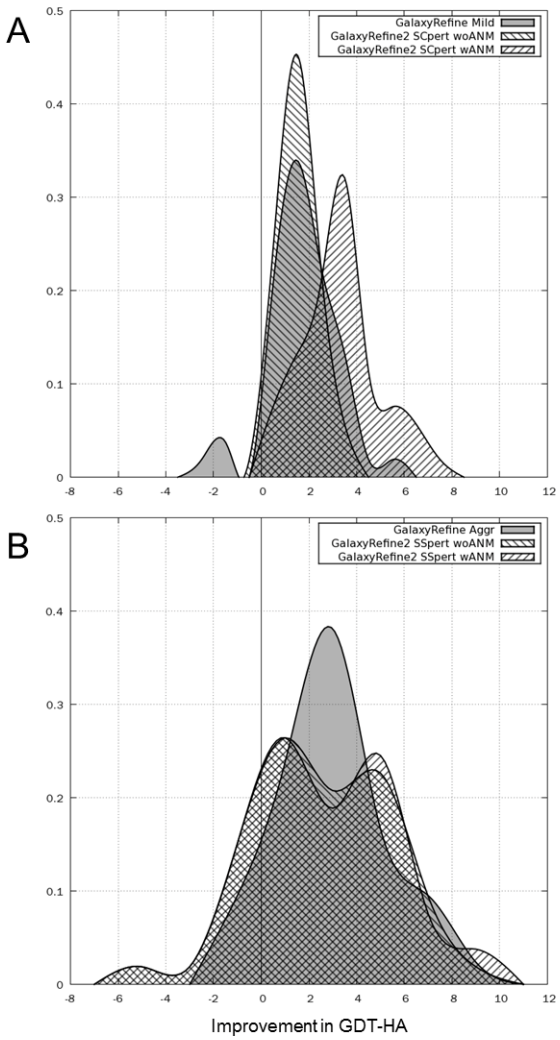
3.2.2.2. Increased sampling performance using ANM and secondary structure perturbation

The improvement in sampling performance by using conformations generated by anisotropic network model to guide MD relaxation was studied. This was done by comparing the results of structural sampling with and without ANM restraints. The results were analyzed separately for 'SCpert' and 'SSpert' methods. For each method, the distribution of maximum improvement in GDT-HA for the fifty three previous CASP refinement targets were evaluated and illustrated in **Figure 3.2**. From **Figure 3.2A,** it is clear that it was effective to use ANM-guided restraints for generating conformations with big improvemnet for 'SCpert'. The distribution peak of applying 'SCpert' without ANM-guided sampling is narrower than the 'Mild' version of GalaxyRefine (Heo et al., 2013) because cartesian coordinate restraints were additionally used.

The change of the range of conformational space sampled by perturbing secondary structures ('SSpert') with or without ANM can be seen in **Figure 3.2B**. It shows a wider distribution compared to 'SCpert'. This difference is also observed when the 'Aggressive' version of GalaxyRefine is compared to the 'Mild' version. Relaxation of 'SSpert' type shows wider distribution than the 'Aggressive' GalaxyRefine. The 'Aggressive' version of GalaxyRefine also perturbes secondary structures, but more diverse conformations could be sampled by the new method because a group of secondary structures could be assigned to be perturbed together when they were interacting with hydrogen-bonding. Also, for the first step relaxation using secondary structure perturbation, when the new conformation generated by perturbing secondary structures was accepted to be used by evaluating with the refinement energy function, the restraint was updated to the new conformation. Updating the restraints to the new conformations generated

could have gradually decreased the restraints adopted from ANM conformations, and thus the distributions of ‘SSpert’ relaxation with and without ANM was almost similar.

Figure 3.2. The change of distribution of maximum improvement in GDT-HA by applying ANM-guided restraints. Data were collected for conformations generated with (A) side chain perturbation and (B) secondary structure perturbation

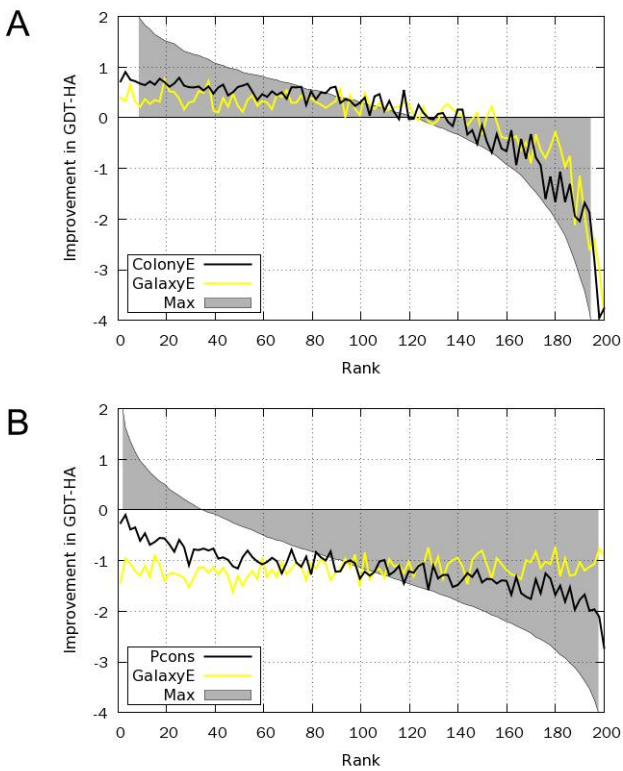


3.2.2.3. Selection of representative conformations

After the conformations were generated by two steps of relaxation, the representative conformations were selected by different methods for each type of relaxation, 'SCpert' and 'SSpert'. For the conformations generated by relaxation with side chain perturbation, colony energy including the conformational entropy term was adopted. For the conformations generated by relaxation with secondary structure perturbation, a quality assessment method Pcons (Wallner and Elofsson, 2005) was used. This method evaluates the conformational entropy of the conformation based on model consensus.

In **Figure 3.3**, the performance of each scoring method compared to using the refinement energy function is shown. In **Figure 3.3A**, for the conformations sampled by 'SCpert', colony energy is better at selecting conformations with bigger improvement in GDT-HA compared to using the refinement energy function. In **Figure 3.3B**, Pcons fails in selecting the conformation with improved GDT-HA on average, but still performs better than using the refinement energy function. Due to the difficulty of scoring conformations generated by 'SSpert', structure-averaging after clustering was adopted as a method for selection. There is room for improvement in scoring the sampled conformations.

Figure 3.3. GDT-HA improvements of conformations selected by each method averaged on all targets. The conformations were generated by (A) 'SCpert' and (B) 'SSpert'.



3.2.2.4. Overall performance of the refinement method with loop modeling

When unreliable local regions of the model structure could be detected, loop modeling was performed before applying overall relaxation. The results of applying refinement without loop modeling and after loop modeling were analyzed. These two methods were blind-tested on the thirty-five CASP11 refinement category targets (Modi and Dunbrack, 2016). The result is summarized in **Table 3.2**. The unreliable local regions were manually assigned by observing the structure. Improvement in quality compared to the given initial model was measured by using GDT-HA (Zemla, 2003), Cα-RMSD, SphereGrinder score (Kryshtafovych et al., 2014), and the Molprobity Score.

In **Table 3.2**, the median improvements and percentages of improved targets of all targets are shown for the first models and best models. The major difference between the performances of two methods is in improving the Cα–RMSD and SphereGrinder score. By applying loop modeling before performing overall relaxation resulted in improvements of 0.05 and 1.81 for Cα –RMSD and SphereGrinder, respectively. GalaxyRefine2 resulted in 0.04 and 0.53. Based on this analysis we can assume that SphereGrinder can measure the conformation change resulted by performing loop modeling. SphereGrinder is a score developed to evalute the closeness between two structures based on local similarity of their substructures (Kryshtafovych et al., 2014).

For GDT-HA, improvements achieved by both methods for the first models are comparable but the improved percentage is smaller (69% to 77%) when loop modeling is performed. In **Figure 3.4,** the first model improvements in GDT-HA for all CASP11 refinement targets are drawn versus the initial model qualities in GDT-HA. When comparing **Figure 3.4B** to **3.4A**, there are more targets that

could not be refined in GDT-HA by peforming loop modeling, which was also shown as the percentage of improvement in **Table 3.2**. However, when targets with initial model quality of GDT-HA higher than 60.0 are considered, all targets except one could be refined when performing loop modeling before overall relaxation. This can be related to the dependancy of loop modeling performance to the initial model quality and its propagation to global structure quality improvement.

Some successful examples of the refinement results are shown in **Figure 3.5**. The refined model shown in **Figure 3.5A** was the third best model in GDT-HA measure among the all submitted predictions for target TR228 in CASP11. Improvements of 5.21, 2.07, 14.1, and 1.54 were achieved for GDT-HA, Cα-RMSD, SphereGrinder, and MolProbity, respectively. The refined model using GalaxyRefine2 only resulted in 1.18, 0.21, 0.60, and 1.94 for the same evaluation measures. The refined model by using loop modeling and GalaxyRefine2 (**Figure 3.5B**) was the best model in GDT-HA of target TR760 in CASP11. Improvements were 1.87, 0.09, 4.48, and 2.25 for the four measures, respectively. The improvements were 0.00, -0.03, 0.75, and 2.12, respectively when appyling GalaxyRefine2 only. These successful examples show that using the loop modeling result to apply further relaxation were effective in improving Cα-RMSD and SphereGrinder. For these cases, the refinement in backbone level could be propagated to refinement in overall structure, which is observed by the improvements in GDT-HA.

Table 3.2. Overall performance of GalaxyRefine2 with and without loop modeling

Method	GalaxyRefine2		Loop modeling and GalaxyRefine2	
	First models[1]	Best models[2]	First models	Best models
ΔGDT-HA	1.04 (77%[3])	1.82 (83%)	0.95 (69%)	2.05 (83%)
-ΔCα-RMSD	0.04 (80%)	0.07 (97%)	0.05 (63%)	0.10 (80%)
ΔSphGr	0.53 (63%)	1.36 (77%)	1.81 (74%)	2.88 (83%)
-ΔMolPrb	1.56 (97%)	1.73 (97%)	1.51 (97%)	1.73 (97%)

1) Median values on the first models of all targets
2) Median values on the best out of five models
3) The percentage of improved targets in parentheses

Figure 3.4. GDT-HA improvement versus initial model quality in GDT-HA for the first models. The models were refined by applying (A) GalaxyRefine2 and (B) applying loop modeling and GalaxyRefine2.

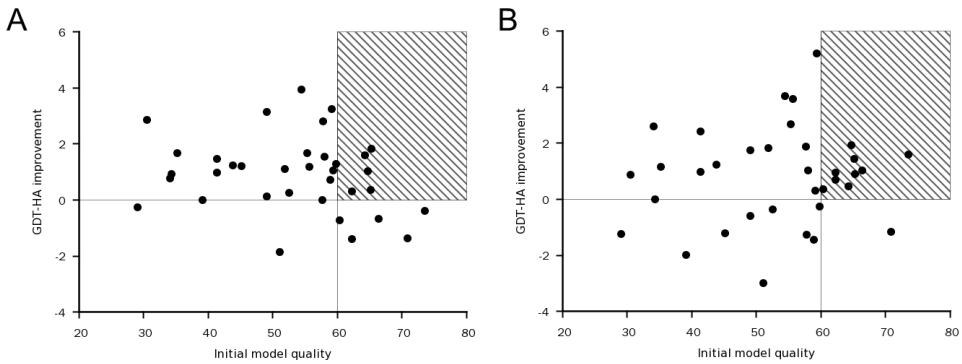
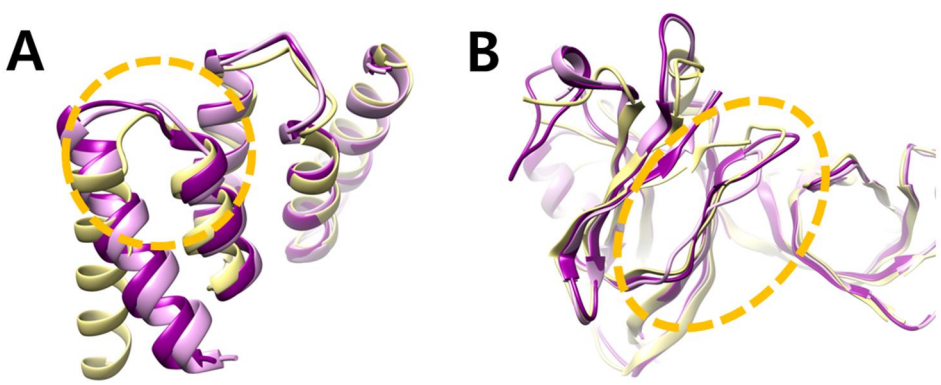


Figure 3.5. Successful examples of applying GalaxyRefine2 with protein loop modeling



3.2.2.5. Improvement in structure quality gained at each refinement step

Average model quality improvements in GDT-HA and Cα-RMSD gained at each refinement step were analyzed for the first models of thirty-five CASP11 refinement category targets. The result is summarized in **Figure 3.6**. **Figure 3.6A** describes the changes of structure qualities obtained by applying GalaxyRefine2 and **Figure 3.6B** describes the improvements achieved by apply loop modeling and GalaxyRefine2.

For the models refined by applying overall relaxation only, the largest contribution to the backbone structure improvement measured by GDT-HA and Cα-RMSD came from the first step relaxation ('Step 1') guided by ANM (**Figure 3.6A**). For refined models using additional loop modeling, the large amount of RMSD improvement was achieved at the loop modeling step (**Figure 3.6B**). In comparison, improvement in GDT-HA was not as pronounced in the loop modeling step.

As the last step of refinement, the selected sturctures from the second relaxation step were optimized in all-atom representation. In **Table 3.3**, the improvements in structure quality evaluation measures gained by optimization are shown. Due to applying strong restraints on the given backbone conformation, not much gain is obatined for global structure accuracy measures such as GDT-HA and lDDT (Mariani et al., 2013). However, improvements in local structure accuray evaluted by GDC-SC, χ^{1+2}, and MolProbity could be observed. It was found that this optimization step was the most effective in improving the MolProbity score.

Figure 3.6. Improvement of the first models in each refinement step. Improvements evaluated in GDT-HA (black lines) and Cα-RMSD (grey lines) averaged over targets for (A) GalaxyRefine2 and (B) GalaxyRefine2 with loop modeling are illustrated.

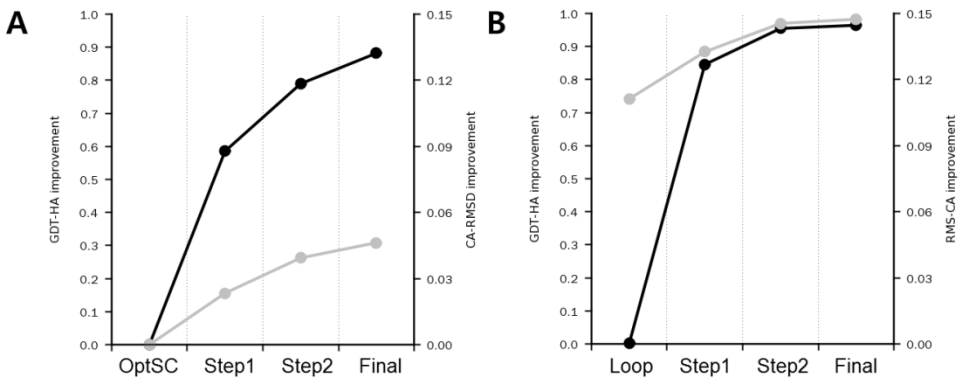


Table 3.3. Structure quality improvements gained after optimization in all-atom representation

Measures	Compared to SCpert	Compared to SSpert
ΔGDT-HA	0.05	0.06
ΔGDC-SC	0.20	0.50
ΔlDDT	0.0002	-0.0035
$\Delta\chi_{1+2}$	0.27	4.32
-ΔMolProbity	0.58	0.43

3.2.2.6. Dependency of the loop modeling performance on model quality

In section 3.2.2.5, we could guess that the different loop modeling performance depending on initial model quality was propagated to global structure quality improvement as shown in **Figure 3.7**. In this section, to backup this idea, the dependency of loop modeling performance on model quality will be analyzed. In **Figure 3.7,** the change in backbone (N, Cα, C) RMSD values for the 54 re-modeled loop regions of 31 CASP11 refinement targets are plotted against the initial model qualities in GDT-HA. Loop RMSD was calculated after superimposing the environment to the experimental structure, as it was done in Park *et al*. (Park et al., 2014). Decrease in loop RMSD means the local structure could be refined by performing loop modeling. Because loop modeling was performed on the fixed initial structure, this distribution can show how the initial model quality affects the loop modeling performance. Although the purpose of developing the loop modeling method was to cope the error originating from the unreliable loop environment, we found that the limit of error which cannot be tolerated exists. When we assume that initial model quality in GDT-HA represents the loop environment error, we can see from **Figure 3.7** that loop RMSD is less expected to be improved when the loop environment is less reliable.

Figure 3.7 shows three examples of applying loop modeling which are marked as A, B, and C in **Figure 3.8**. Initial quality of the models depicted in **Figure 3.8A** and **3.8B** were relatively high (GDT-HA = 59.4 and 73.5, respectively). The backbone loop RMSD could be decreased from 8.95 to 2.96 and 1.97 to 1.22 for each case, respectively. Overall relaxation of these models resulted in increase in GDT-HA (GDT-HA increase from 59.4 to 64.6 and 73.5 to 75.1). However, the initial quality of the model illustrated in **Figure 3.8C** was 45.2 and

loop RMSD rather increased from 2.58 to 3.50. In the figure, a phenylalanine residue in the loop environment which the incorrect orientation can hinder the correct positioning of the loop is drawn.

Figure 3.7. Loop RMSD change versus initial model quality in GDT-HA. Points A, B, and C represent the example cases illustrated in **Figure 3.8**.

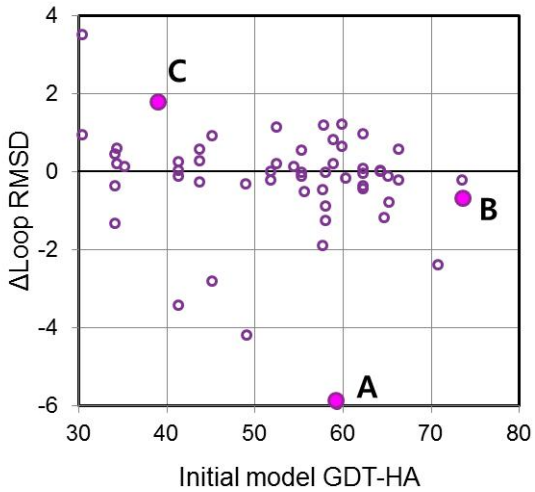
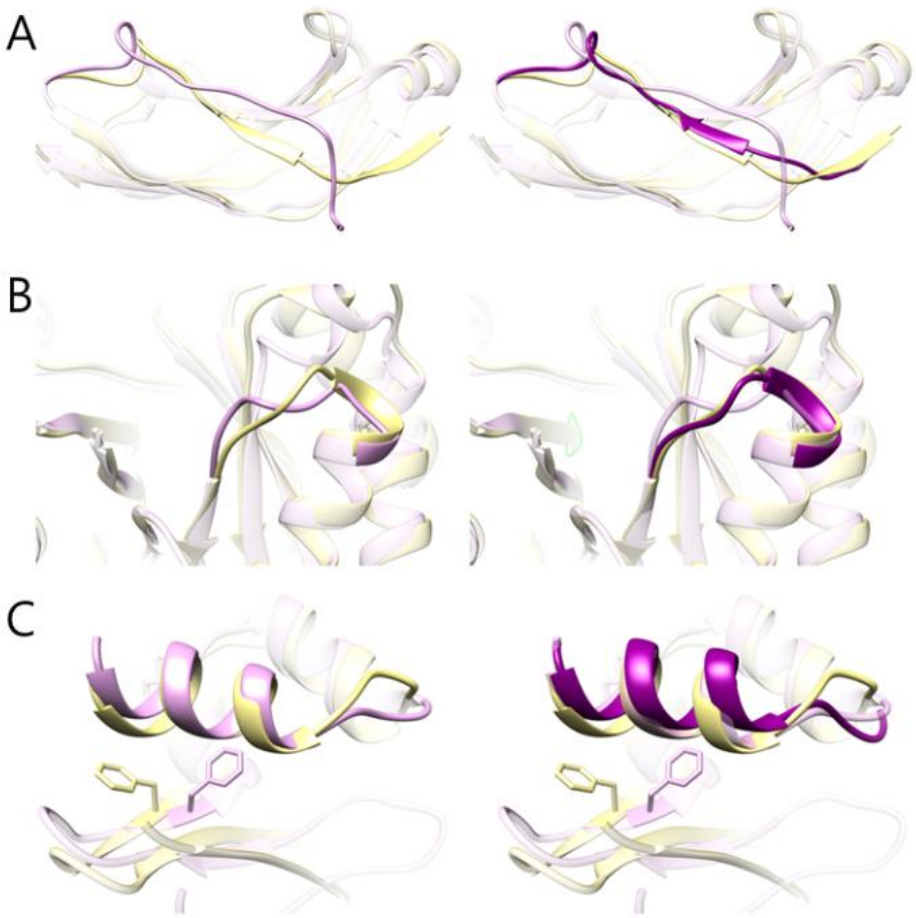


Figure 3.8. Examples of loop modeling applied for global refinement



3.3. Global refinement with diverse modeling methods applied on unreliable local regions

3.3.1. Methods

In **Section 3.2**, loop modeling was performed on manually detected unreliable local regions of a predicted model before applying overall relaxation. This limited the performance of loop modeling and overall refinement. To overcome this, an improved structure refinement method was developed incorporating diverse protein conformation sampling methods which model the protein structure simultaneously with overall relaxation. Unreliable local regions of the protein model were automatically predicted.

3.3.1.1. Residue level quality assessment of a protein model

A method to predict residue level accuracy when a protein model is given was developed. The relative quality score predicted for the protein residues was further used to detect unreliable local regions based on clustering.

First, the initial model was relaxed using molecular dynamics simulation with the energy function developed for refinement. This generated 96 conformations. Restraints on the initial model were weakly applied, with weight 1.0. Root mean square fluctuation of the Cα atoms was evaluated for the 60% of the conformations with low energy. Second, fragment library was generated using Rosetta (Gront et al., 2011). A match score was evaluated by comparing the initial model torsion angles to the torsion angle data derived from the fragment library. Finally, based on the fold recognition result for the sequence of the protein model, PSSM score was evaluated (Stormo et al., 1982). A weighted sum of these scores for each residue was used as residue level quality assessment score.

$$Score_i = Score_{RMSF} + w_{FRAG} \times Score_{FRAG} + w_{PSSM} \times Score_{PSSM}$$

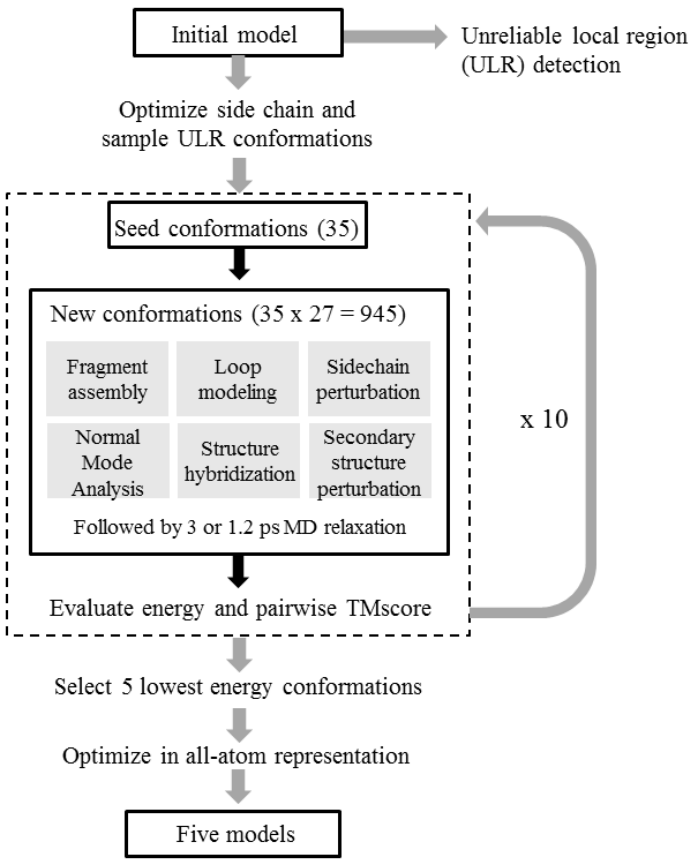
3.3.1.2. Outline of the global refinement method, GalaxyVoyage

The improved version of global refinement method can be summarized as in **Figure 3.9**. First, using the initial protein model, unreliable local regions (ULR) are predicted by using the residue level QA score obtained by using the method described above. When ULRs were detected, initial sampling using fragment assembly and loop closure was performed on these regions to generate the initial pool of seed conformations.

The conformational space search was done by performing up to ten iterations of successively deeper sampling of the low-energy regions. Diverse sampling methods which are explained in the **subsection 3.3.1.4** were implemented or developed to generate 945 conformations for each iteration. Residue level QA score was used to determine the probability of each modeling methods to be applied at each residue. After each iteration step, 35 conformations with the lowest energies that are distant from each other were selected as seed conformations for the next iteration.

After iteration, all conformations were collected and scored with the energy function without restraints. The five lowest-energy models were selected after reducing redundancy. These models were converted into all-atom representation and were re-optimized.

Figure 3.9. Summary of GalaxyVoyage



3.3.1.3. Energy function used in GalaxyVoyage

3.3.1.3.1. The hybrid energy function with the new knowledge-based potential

We took most of the components from the energy function described in **subsection 3.2.1.1**. The new energy function also maintains the formation of a hybrid energy function. It is composed of physics-based energy terms and knowledge-based energy terms. Instead of dipolar DFIRE (Yang and Zhou, 2008a), we used KGB, the knowledge-based potential developed by our lab (Heo et al., *in preparation*). Same types of restraint functions described in subsection **3.2.1.1** were adopted.

$$E = E_{physics-based} + E_{knowledge-based} + E_{restraint}$$

$$E_{physics-based} = E_{bonded} + E_{vdw} + w_{electrostatics}(E_{Coulomb} + E_{FACTS,GB}) + w_{FACTS,SA}E_{FACTS,SA}$$

$$E_{knowledge-based} = w_{\phi/\varphi}E_{\phi/\varphi} + w_{\chi}E_{\chi} + w_{hbond}E_{hbond} + w_{KGB}E_{KGB}$$

$$E_{restraint} = w_{restraint,dist}E_{restraint,dist} + w_{restraint,cart}E_{restraint,cart}$$

The energy function KGB is a knowledge-based function derived in a similar way to DFIRE using the high resolution crystal structures (Samudrala and Moult, 1998) using the following formalism. KGB considers dependence of the atom pair potential on the solvation states of the interacting atoms as well as their pair distance. The weight parameters of this hybrid energy function were optimized for both relaxing and scoring conformations.

3.3.1.3.2. Partial restraints determined by Bayesian inference

When searching the conformational space for structure refinement, it has been known that applying restraints on the initial model structure can consistently improve the structure qualities (Mirjalili et al., 2014; Raval et al., 2012). Especially when the conformational searching method is based on molecular dynamics simulation, it has been shown that without restraints, the sampled conformations tend to drift away and the conformations become worse compared to the experimental structure due to not perfect conformational search methods and the force fields (Mirjalili et al., 2014; Raval et al., 2012). However, applying restraints to the given initial structures impede conformational change in large-scale and decreases the degree of refinement.

To search the conformational space more diversely, we use a subset of restraints collected from the initial model. We adopted a method based on Bayesian inference to determine which 90% of partial restraints to use at every time step during molecular dynamics relaxation (MacCallum et al., 2015).

3.3.1.4. Diverse protein modeling methods used to generate new conformations

To generate new conformations, diverse structure operators are used. The operators can be divided by the degree of conformational change resulted by applying each modeling method. The operators that give global structure change were those which use anisotropic network model, those that hybridize the conformation with other structures, those that perform overall MD relaxation with iterative side chain or secondary structure perturbation. Experimentally resolved structures of homologous proteins searched by using HHsearch (Soding, 2005) were filtered with structural similarity to the initial model using TM-align (Zhang and Skolnick,

2005). These were used to detect locally variable regions and generate hybrid structures. When these structures were not available, hybridization of the conformations generated during one iteration was also performed. The generated conformations underwent 1.2 or 3-ps of overall MD relaxation depending on the magnitude of conformational change.

3.3.1.5. Benchmark test set

The energy function of GalaxyVoyage was optimized by performing decoy discrimination or MD relaxation on a training set. The training set composed of 91 best server models submitted during CASP 10 and 11 tertiary structure prediction categories (Huang et al., 2014). The targets redundant to CASP refinement targets were removed. Models of single domain targets were collected. Decoy conformations were generated by running the preliminary version of the refinement method with the non-optimized version of the energy function. MD relaxation was tested by using the 'Mild' version of GalaxyRefine (Heo et al., 2013) with restraints on the initial model.

The test set of the refinement method consisted of 63 CASP 10 and 11 refinement category targets (Modi and Dunbrack, 2016; Nugent et al., 2014).

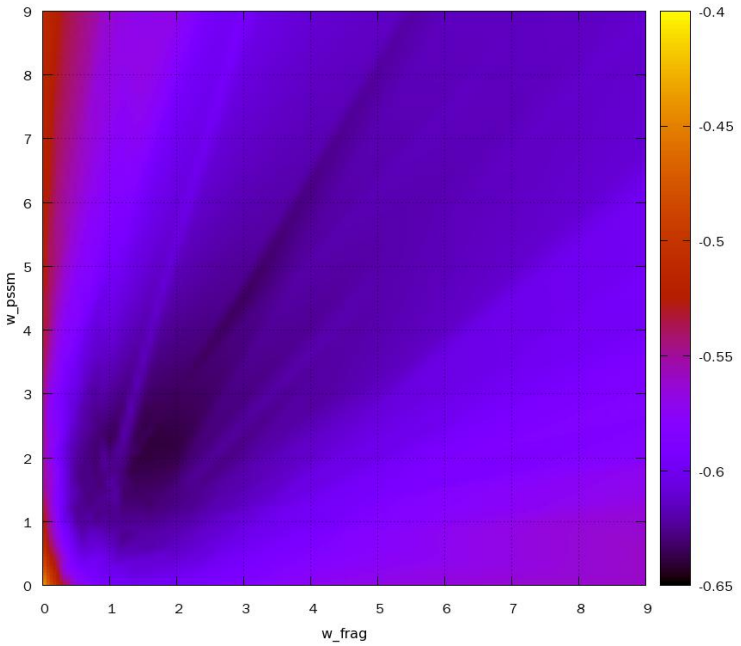
3.3.2. Results and discussion

3.3.2.1. Optimization of the residue level QA scoring method

Weight parameters of the scores to predict residue level QA score were determined to be 0.15 and 0.15 for w_{FRAG} and w_{PSSM}. This scoring function was optimized by performing grid search on the parameter space to increase the correlation between the predicted QA score and the local structure similarity to the

experimental structure. Local structure similarity to the experimental structure was evaluated similarly to calculating lDDT (Mariani et al., 2013). The grid search result is summarized in **Figure 3.10**. By dividing the training set into 2, cross validation approach was taken for parameter optimization. The figure shows the result on the median correlations on all training set targets.

Figure 3.10. Correlation between local structure accuracy measure and residue QA score for different combinations of weight parameters. Median correlations of the training set were used.



3.3.2.2. The performance of protein modeling methods in generating improved conformations

Diverse types of protein modeling methods which can result in different magnitudes of conformational change were adopted to generate new conformations. The effect of each modeling operator in producing conformations with improved structure accuracy measured in GDT-HA and Cα-RMSD was analyzed. This result is shown in **Table 3.4**. From the conformations generated using each modeling method, the percentage of conformations which showed improvement in different structure quality measures were evaluated. For both measures, overall relaxation with side chain perturbation resulted in the biggest fraction of improved conformations. However, this consistency could have been achieved because the magnitude of conformation change resulted by only perturbing side chain is small. For other protein modeling methods, no distinct characteristics could be observed.

Table 3.4. The percentage of structures improved in each measure from the structures generated by using each modeling operator. Median of percentages on sixty benchmark test set targets are shown.

Operator	GDT-HA improved (%)	Cα-RMSD improved (%)
ANM	35.0	44.0
Hybridize with template	23.5	22.5
Hybridize with partner	42.5	49.5
Fragment assembly	28.0	35.0
Loop mutation	27.5	33.5
Loop cross over	31.0	36.5
SSpert	26.5	33.0
SCpert	56.0	58.5

3.3.2.3. Energy function optimization with the new knowledge-based potential

Weight parameters of the energy function consisting of the new knowledge-based potential, KGB were optimized by performing decoy discrimination and overall relaxation. Initially, the weight parameters which resulted in selecting the decoy conformation with big improvement were chosen. By using this energy function, overall molecular dynamics relaxation with side chain perturbation was tested on the models of the training set. Weight parameters were modified slightly to increase the refinement performance. Using this weight set, GalaxyVoyage method was run to generate new conformations. These conformations were used as the next round decoy conformations.

As a result, the weights were determined to be (0.2, 0.04, 0.6, 2.0, 2.0, 2.0, 5.0, and 5.0) for ($w_{electrostatics}$, $w_{FACTS,SA}$, $w_{\varphi/\psi}$, w_{χ}, w_{hbond}, w_{KGB}, $w_{restraint,dist}$, and $w_{restraint,cart}$). It should be noted here that the Ramachandran map-biasing potential for backbone torsion angles was derived again. It was developed to consider the secondary structure prediction results using PSIPRED (McGuffin et al., 2000). The result of testing the new energy function for decoy discrimination and overall relaxation is given in the next two tables compared with the result of applying the energy function used in GalaxyRefine2.

Table 3.5. Quality of a decoy conformation selected with the energy function. Median of improvements compared to the initial model structure for the 91 training set targets are shown.

Energy function	ΔGDT-TS	ΔGDT-HA	ΔCα-RMSD	ΔGDC-SC
GalaxyRefine	+0.83	+1.32	+0.006	+1.74
GalaxyVoyage	+1.47	+1.47	-0.075	+2.02

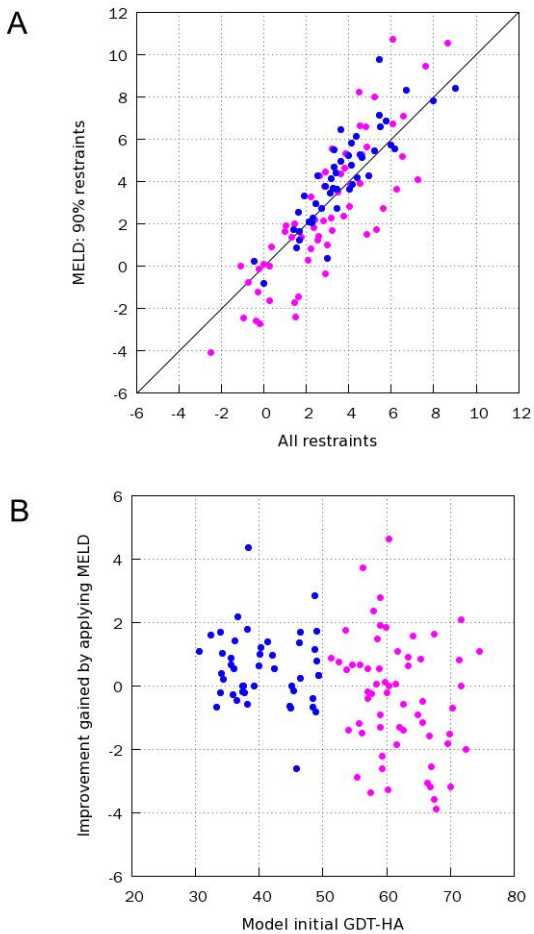
Table 3.6. Qualities of energy minimum conformations generated by GalaxyRefine with different energy functions. Median improvements compared to the initial model structure for the 91 training set targets are shown

Energy function	ΔGDT-HA	ΔCα-RMSD	ΔGDC-SC
GalaxyRefine	+1.05	-0.03	+1.05
GalaxyVoyage	+1.63	-0.04	+1.63

3.3.2.4. The effect of applying partial restraints determined by Bayesian inference

When applying overall MD relaxation to the new conformations generated during the conformational space search, restraints on initial structures were applied. Restraints were applied to the distance between Cα atoms or N and O atoms and Cartesian coordinates of Cα atoms. Because the aim of the new refinement method was to allow more large-scale movements, the restraints were partially applied. The subset of restraints to be applied for MD relaxation was determined based on Bayesian inference (MacCallum et al., 2015). In every time step, the energy of all restraints was evaluated and the 90% of restraints with lower energy were applied. The proportion of restraints to apply (90%) was chosen after testing the performance of GalaxyRefine with different amount of partial restraints. A subset of restraints composed of 70, 80, and 90% lower energy restraints were tested. As a result, 90% was chosen to give stable performance in refining the model structures with GalaxyRefine (Heo et al., 2013). The maximum improvement in GDT-HA obtained from the conformations generated by applying 90% partial restrains was compared to using all restraints in **Figure 3.11A**. The improvements are comparable. However, when the improvements gained by applying partial restraints are observed in relation to the model quality in GDT-HA (**Figure 3.11B**), it could be seen that applying 90% of restraints was effective for the targets with relatively low structure accuracy (GDT-HA < 50.0). Allowing more movements by neglecting 10% of restraints could result in sampling more improved structures for these targets.

Figure 3.11. Performance of GalaxyRefine affected by using partial restraints determined by Bayesian inference. Result of applying GalaxyRefine with 90% of restraints determined by Bayesian inference compared to the result of refinement with all restraints. The maximum improvements in GDT-HA for the generated models are compared in (A) and the improvements gained by using partial restraints is are drawn versus initial model quality in (B).



3.3.2.5. Overall performance of GalaxyVoyage

The overall performance of GalaxyVoyage in refining the model structures of the benchmark test set is summarized in **Table 3.7**. The performance was evaluated with average and median improvements of the first models of the targets. The result was compared to that of applying GalaxyRefine2, discussed in **section 3.2**. From **Table 3.7**, it can be seen that the average structure improvements were comparable to GalaxyRefine2 (Lee et al., 2016). For global structure accuracy measures GDT-TS and GDT-HA, the average improvements of GalaxyVoyage outperformed GalaxyRefine2 by 30% each.

Although the overall performances of two methods were comparable, when head to head comparisons of the improvements were analyzed, the distribution of the improvements were significantly different. This difference is illustrated in **Figure 3.12**. From the figure it can be noted that the improvements in GDT-HA or RMSD when applying GalaxyVoyage are widely spread than applying GalaxyRefine2. Especially in **Figure 3.12B**, the large deviation of the RMSD improvements of GalaxyVoyage results can be compared to the results of GalaxyRefine2, which the RMSD improvements are near 0.0. This implies that GalaxyVoyage could generate conformations with more conformational change but not always resulted in improvement.

Targets which were successful by applying GalaxyVoyage are illustrated in **Figure 3.13**. For the target illustrated in **Figure 3.13A** (PDB ID: 4ic1), the improvements in GDT-HA, GDC-SC, Cα-RMSD, and lDDT were 2.64, -0.20, 1.43, and 0.018, respectively. This could be compared to the result of applying GalaxyRefine2 depicted in **Figure 3.13B** with improvements of 0.88, 0.43, 0.00, and 0.003. Although the side chain accuracy became worse, improvements in

global structure accuracy measures could be increased. From the figure, the refined terminal region of the model structure which is similar to the crystal structure can be noted. Similarly, for the target illustrated in **Figure 3.13C** (PDB ID: 4r03), the improvements obtained by applying GalaxyVoyage were 1.85, 2.36, 2.39, and 0.046. In the other hand, the improvements were -1.39, -0.38, -0.01, and 0.0 for the conformation in **Figure 3.13D**. The terminal region of the initial model could be refined in this target, also.

Based on these analyses, the strength of GalaxyVoyage can be seen as generating conformations with more conformational change compared to GalaxyRefine2. However, there still are limits regarding the energy function. In **Figure 3.14**, an unsuccessful result by applying GalaxyVoyage to the target (PDB ID: 4fr9) is illustrated. The terminal region of the initial model was structurally changed but it became worse in terms of structure accuracy. This conformation was selected to be low in energy as the Tryptophan buried in the protein curvature was favored and the terminal part constructed a helix. Targets with problems like these imply that the balance of the energy terms composing the energy function still needs to be optimized. The problem currently noted is the over-estimation of short-range hydrogen bonds.

Table 3.7. Overall performance of GalaxyVoyage compared to GalaxyRefine 2 on the benchmark test set. Average improvements on first models of 63 previous CASP refinement targets are shown with median improvements in parentheses.

Energy function	ΔGDT-TS	ΔGDT-HA	ΔCα-RMSD	ΔMolProbity
GalaxyRefine2	+0.4 (0.5)	+0.7 (+0.9)	-0.06 (-0.05)	+1.2 (+1.4)
GalaxyVoyage	+0.7 (+1.0)	+1.0 (+0.8)	-0.01 (-0.07)	+1.2 (+1.4)

Figure 3.12. Head to head comparison of the improvements achieved by GalaxyRefine2 and GalaxyVoyage. Improvements were evaluated in increase in (A) GDT-HA and (B) Cα-RMSD.

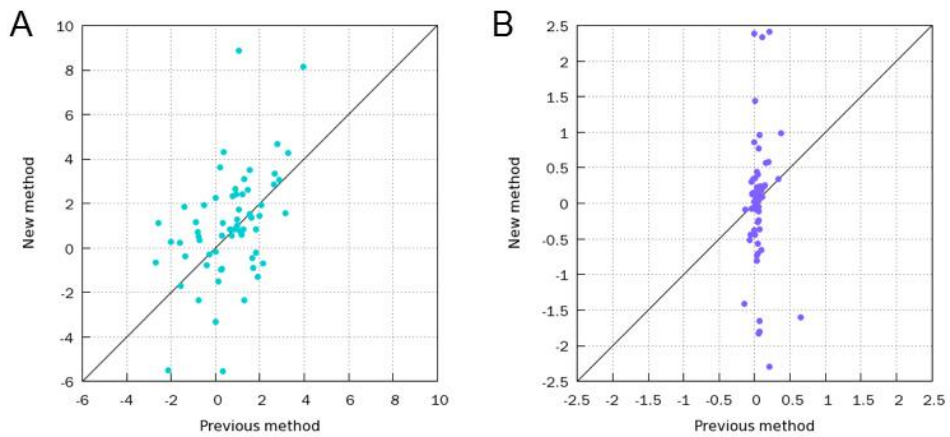


Figure 3.13. Successful examples of GalaxyVoyage. (A, C) Results of applying GalaxyVoyage (dark green) to the initial structure (light blue) are compared to (B, D) results of applying GalaxyRefine2 (blue) to the same initial structure (light blue). The crystal structure is drawn in yellow.

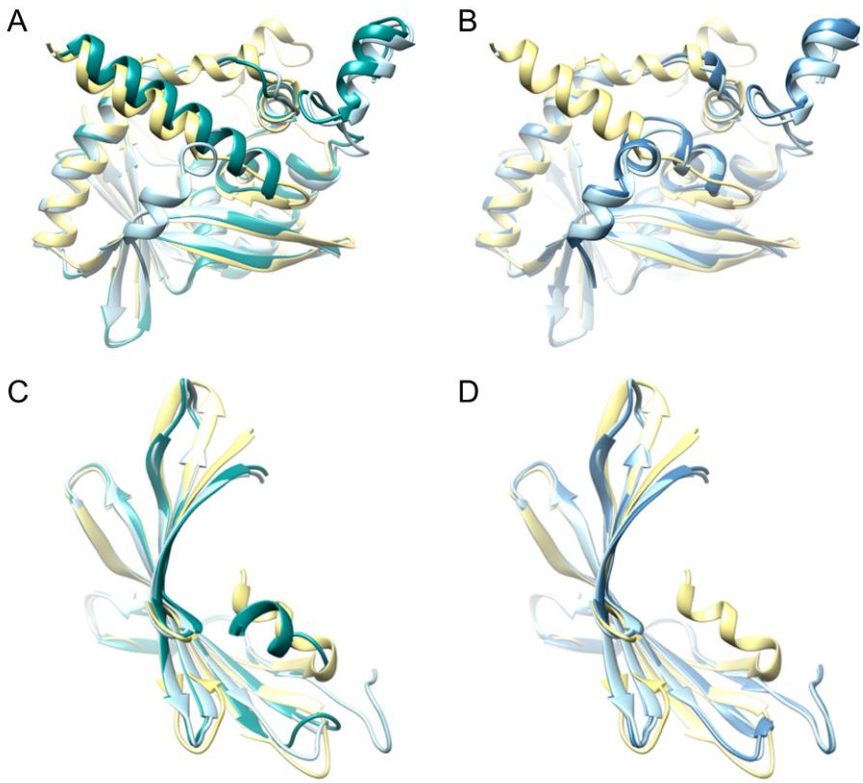
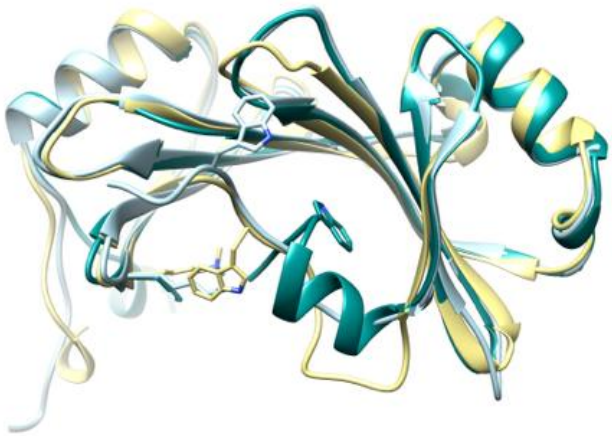


Figure 3.14. An example revealing the limits of GalaxyVoyage. A result of applying GalaxyVoyage (dark green) to the initial structure (light blue) is compared to the crystal structure (yellow)



3.4. Conclusion

To tackle the problem of refining protein model structures, two global refinement methods were developed. In the first approach, protein loop modeling was initially applied to the manually assigned unreliable local regions of the model. It was anticipated that the refinement in local region would be propagated globally by performing overall relaxation on the model. The method was successful at refining the models consistently. However, the degree of conformational change and improvement were small due to strong restraints applied. Also, loop modeling performance was limited by the model quality.

Another method based on simultaneous conformational search using diverse modeling methods and overall relaxation was developed to overcome the problems addressed by the first method. Partial restraints were applied to allow more conformational change during refinement. Also, an improved energy function incorporating a new knowledge-based potential was developed. The method could generate more diverse conformations, but optimization of the energy function to accurately evaluate the conformations is still necessary.

Chapter 4

4. Flexible docking of ligands to G-protein-coupled receptors based on structure refinement

4.1. Introduction

4.1.1. Ligand docking to model structures

Protein-ligand docking methods have been developed to predict the structures, interactions, and binding affinities of the protein-ligand complexes (Andrusier et al., 2008). Many protein-ligand docking methods were developed as treating the receptor rigid (Kitchen et al., 2004). However, the protein conformation undergoes change upon binding to different ligands. Therefore, in recent years, the protein-ligand docking methods have been developed to consider the flexibility of the protein. The flexibility of assigned sidechains to full receptor including backbone has been incorporated. The docking methods that incorporate full flexibility of receptor are those that try to mimic induced fit docking (Meiler and Baker, 2006; Sherman et al., 2006) and those based on ensemble docking (Ding and Dokholyan, 2013). Although protein-ligand docking methods have evolved to increase the dimensionality of the protein-ligand conformational space, due to the increased computational time and inaccurate energy function, it still remains to be a highly challenging problem (Totrov and Abagyan, 2008).

The protein-ligand docking methods developed to incorporate backbone flexibility were mostly tested on predicting the ligand-bound complex structure when the unbound structure of the receptor was given. When there is no experimentally determined receptor structure in any state, protein ligand docking should be applied to predicted model receptor structures. However, to our

knowledge, there is no generalized docking program benchmarked to predict the complex structure when model receptor structures are given. In this chapter, development of a flexible docking method based on complex structure refinement when no experimental receptor structure is known will be addressed (Lee and Seok, 2016; Shin et al., 2015).

Docking ligands to the inaccurate receptor structure was tackled by taking two approaches in this study. The energy functions of currently used docking programs are optimized to the relatively reliable docking environment such as experimental structures. These energy functions can be sensitive to environment error. Therefore, at the initial stage of docking, ligand docking was performed on an ensemble of receptor conformations generated by anisotropic network model. Also, initial docking on the ANM conformations was done with less optimization compared to deep global optimization for docking on experimental structures. The second stage was applying overall refinement to the complex structures. Hybrid energy function which is less sensitive to environmental error was developed for refinement.

4.1.2. Predicting the ligand-bound G-protein-coupled receptor structure

Galaxy7TM, the flexible docking method for model receptor structures introduced in this chapter (Lee and Seok, 2016), was developed particularly for G-protein-coupled receptors (GPCR). G-protein-coupled receptors are membrane proteins involved in important signal transduction pathways in the cell (Dorsam and Gutkind, 2007; Katritch et al., 2013). GPCR makes a major group of drug targets as these pathways regulate important physiological processes (Lagerstrom and Schioth, 2008). The information of the structures or interactions of GPCR-ligand complexes are crucial for understanding the function of this protein family and

further applications such as drug discovery.

Refinement of membrane protein model structures is usually done by running molecular dynamics simulation with explicit lipid molecules (Mortier et al., 2015; Wolf et al., 2008). However, using molecular dynamics simulation to refine GPCR models would be computationally inefficient and case-specific (Kufareva et al., 2014; Sandal et al., 2013). Computational time of the developed method could be saved by introducing an implicit solvation model describing the lipid environment. Also, the conformational space could be searched more efficiently by refining GPCR-ligand complexes after performing initial docking. These approaches could make the method more generalized and fast.

4.2. Methods

The Galaxy7TM method consists of two stages, initial docking and subsequent refinement docking. The initial docking stage is for performing low resolution docking on an ensemble of receptor conformations. Overall molecular dynamics relaxation is applied to the complex structures obtained from the initial docking stage.

4.2.1. Initial docking of ligand to receptor conformations generated by ANM

Initially, ligand docking was performed to an ensemble of receptor conformations generated by anisotropic network model (ANM) (Bakan et al., 2011) to incorporate receptor flexibility and overcome the error embedded in the model structure.

An ensemble of receptor conformations was generated by perturbing the initial receptor model in the direction of normal modes calculated by anisotropic network model. The normal mode was randomly selected from the twenty lowest frequency modes to generate two hundred conformations. Two hundred conformations were generated to have RMSD range of 0.5-2.5 Å from the initial structure. The sidechain structures of these conformations were optimized using SCWRL4 (Krivov et al., 2009). The ANM conformations optimized using SCWRL4 resulted in higher initial docking performance compared to the conformations with sidechain structures optimized using Galaxy-optSC (data not shown). This may have resulted because SCWRL4 uses a much simpler energy function than the high resolution energy function used by Galaxy-optSC (the same energy function used for overall relaxation). The two hundred conformations were clustered into thirty by using the all-atom RMSD of binding pocket residues as a measure of similarity. K-means clustering algorithm was used.

While developing Galaxy7TM, it was assumed that the binding pocket residues are not known. The binding pocket prediction results were used to cluster the ANM conformations and position the docking box for initial docking. Ligand binding pocket residues of a given model receptor structure were predicted by aligning the structure to the experimentally determined structures of GPCRs bound to small molecules (Zhang and Skolnick, 2005). The residues structurally aligned to the ten most similar binding pocket structures were detected by consensus.

Initial docking of ligands to the ANM conformations was done by applying a slightly modified version of GalaxyDock (Shin et al., 2013). The current version of GalaxyDock generates initial bank ligand conformations using BetaDock and applies global optimization of BP2-score (Baek et al., in preparation) using conformational space annealing (CSA) (Lee et al., 1997). To apply this in the initial stage of Galaxy7TM, the initial ligand conformations to be used as CSA initial bank were generated with random docking mode of GalaxyDock. Also, to alleviate the effect of inaccurate structure environment to ligand docking with high resolution BP2-score, CSA with modified parameters to do much less optimization compared to the original version was applied.

4.2.2. Energy function for complex structure refinement

For each thirty ANM conformation, four lowest docking energy ligand conformations were selected. A total of 120 GPCR-ligand complex structures were targeted for overall refinement. Complex structure refinement was applied with molecular dynamics relaxation, similar to GalaxyRefine (Heo et al., 2013). The energy function used for relaxation was composed of energy terms similarly to GalaxyRefine but an implicit solvation model extended to describe the lipid environment replacing the generalized Born-based solvation free energy term.

The GPCR decoy discrimination performance of two implicit solvation models describing the lipid environment, IMM1 (Lazaridis, 2003) and FACTSMEM (Carballo-Pacheco et al., 2014) was evaluated. IMM1 is an extended version of an empirical solvation free energy evaluation method, EEF1 (Lazaridis and Karplus, 1999). FACTSMEM is an extended version of FACTS (Haberthur and Caflisch, 2008), a solvation free energy function with the generalized Born formalism. In both models, the screening effect of solvent to the protein atom depending on its position on the axis which is vertical to the lipid membrane is represented by scaling the dielectric constant. The dielectric constant is empirically scaled depending on the atom position on the axis.

After choosing the implicit solvation model to describe the lipid environment, weight parameters of the energy function were optimized. The energy function is a hybrid energy function designed to be less sensitive in unreliable structural environment. The other energy terms except the terms related to solvation free energy were adopted from the energy function of GalaxyRefine (Heo et al., 2013). For the interactions involving ligand atoms, only physics-based energy terms were applied. The CGenFF (Vanommeslaeghe et al., 2010) parameters implemented in the GALAXY program were used after mapping the given ligand atom MOL2 types into charmm22 atom types. The restraint energy functions with same types as described in **section 3.2.1.1**, distance restraints and Cartesian restraints, on the initial receptor structure were applied to the ANM conformations. Also, restraints on Cartesian coordinates of the each initially docked ligand conformation were additionally applied. This approach was taken to refine the complex structures by biasing the ligand conformation from the initial docking results and partially restoring the receptor structure to the initial state because ANM conformations could have been generated with unphysical

perturbations. The entire energy function can be written in detail as follows.

$$E = E_{receptor,physics-based} + E_{ligand,physics-based} + E_{receptor-ligand,physics-based} + E_{receptor,knowledge-based} + E_{restraint}$$

$$E_{receptor,physics-based} = E_{receptor,bonded} + E_{receptor,vdw} + w_{electrostatics}(E_{receptor,Coulomb} + E_{FACTSMEM,GB}) + w_{FACTSMEM,SA}E_{FACTSMEM,SA}$$

$$E_{ligand,physics-based} = E_{ligand,bonded} + E_{ligand,vdw} + w_{electrostatics}E_{ligand,Coulomb}$$

$$E_{receptor-ligand,physics-based} = E_{receptor-ligand,vdw} + w_{electrostatics}E_{receptor-ligand,Coulomb}$$

$$E_{receptor,knowledge-based} = w_{\phi/\varphi}E_{\phi/\varphi} + w_{\chi}E_{\chi} + w_{hbond}E_{hbond} + w_{dDFIRE}E_{dDFIRE}$$

$$E_{restraint} = w_{receptor\_rsr,dist}E_{receptor\_rsr,dist} + w_{receptor\_rsr,cart}E_{receptor\_rsr,cart} + w_{ligand\_rsr,cart}E_{ligand\_rsr,cart}$$

4.2.3. Complex structure refinement and final model selection

A total of 120 complex structures by taking 4 lowest docking energy ligand conformations for 30 ANM conformations each underwent overall refinement using the energy function described above. The complex structure refinement was done by applying molecular dynamics relaxation with iterative side chain perturbation like GalaxyRefine. For each complex structure, a single trajectory 14.4-ps molecular dynamics simulation was run.

A scoring method was developed to select ten conformations from the refined complex structures. A sum of the rank by the refinement energy function and the rank by the docking energy function was adopted to score the conformations. By the result of training (data not shown), only the empirical

energy terms adopted from AutoDock4 energy function were used to rank the conformations. Finally, ten conformations with the lowest score evaluated as follows were evaluated.

$$Score_i = Rank_{E_{refinement},i} + 0.5 \times Rank_{E_{docking},i}$$

This sum of ranking scheme was adopted because the refinement energy function and the docking energy function describe different range of receptor and ligand interactions which made it hard to optimize the weight parameters for linearly combining them.

4.2.4. Benchmark test set construction

4.2.4.1. Training set for refinement energy function optimization

Experimentally determined GPCR crystal structures bound to small molecules were used to construct the benchmark test set. To select the implicit solvation model describing the lipid molecule and optimize the weight parameters of the whole refinement energy function, randomly selected 9 GPCR sequences were used after removing redundancy (pairwise sequence identity < 90%). Weight parameters of the energy function were optimized to show best performance in decoy discrimination. GPCRs binding to ligands were not considered in this stage. Decoy structures for the selected training set of GPCR sequences were generated by applying molecular dynamics relaxation and conformational space annealing. Decoy conformations including those which are close to the crystal structure were generated.

4.2.4.2. Benchmark test set of Galaxy7TM

The inputs of Galaxy7TM method are the GPCR model structure and the three-

dimensional ligand conformation. GPCR model structures of 23 sequences were constructed by three different methods. Two sets of models are generated by template-based modeling. By comparing to the crystal structure of the target sequence, the closest GPCR experimental structure in terms of RMSD was selected as a template. Global sequence alignment was done by HHalign (Soding, 2005). Using this same template and sequence alignment result, the two methods, MODELLER (Sali and Blundell, 1993) and GalaxyTBM (Ko et al., 2012), were applied for homology modeling. The sequence identity range of target and template was 9.2-59.9%. The last set is constructed by collecting the models provided in the GPCR-HGmod database (Zhang et al., 2015) generated using GPCR-I-TASSER (Zhang et al., 2015). Only the models of sequences provided in the database were used.

The input ligand conformations were taken from ligand bound GPCR crystal structures. This resulted in a total of 47 ligands. The ligand atom and coordinate information was converted into MOL2 format using OpenBabel (O'Boyle et al., 2011). Gasteiger charge assigned by OpenBabel was used. It has to be noted that the bond, bond angle, and ring conformation of the initial ligand conformation were not changed during the initial docking stage. A docking method to incorporate flexibility on these interactions will be dealt in the future work.

In total, 125 combinations of GPCR models and ligand structures were gathered as a benchmark test set.

4.3. Results and discussion

4.3.1. Energy function for complex structure refinement

Two implicit solvation models describing the lipid environment, IMM1 (Lazaridis, 2003) and FACTSMEM (Carballo-Pacheco et al., 2014) were tested for GPCR decoy discrimination on the energy function training set. Decoy discrimination performance of each energy function was evaluated by the Pearson correlation coefficient between the energy and structure quality and the quality of the conformation selected by energy. The structure quality of the decoy conformations were evaluated using Cα-RMSD, GDC-SC (Keedy et al., 2009), and local distance difference test (lDDT) (Mariani et al., 2013). Cα-RMSD was used to evaluate the global structure accuracy and GDC-SC was used to evaluate side chain accuracy. The other measure lDDT is used for evaluating local structure similarity without superposition. The result comparing these two implicit solvation models with other two energy functions, dDFIRE (Yang and Zhou, 2008b) and FACTS (Haberthur and Caflisch, 2008) which is consisted in the energy function of GalaxyRefine (Heo et al., 2013) is summarized in **Table 4.1**. The results were averaged on the targets composing the energy function training set.

From **Table 4.1**, it can be seen that the FACTSMEM energy was superior in most of the measures. Also, because the energy function used for complex structure refinement was designed to have the similar structure with the energy function used in GalaxyRefine (the composing terms are written in **subsection 3.X.X**), replacing FACTS with FACTSMEM made it easier for weight parameter optimization.

Table 4.1. GPCR decoy discrimination performance of different energy functions. The values were averaged on the training set targets.

Energy function	Pearson correlation coefficient[1]			Quality of the best out of 10 lowest energy conformations		
	Cα-RMSD	**GDC-SC**	**lDDT**	**Cα-RMSD**	**GDC-SC**	**lDDT**
dDFIRE[2]	0.21	-0.19	-0.28	2.10	35.6	70.6
IMM1[3]	0.22	-0.23	-0.25	1.85	37.8	71.2
FACTS[4]	0.24	-0.26	-0.38	2.03	36.8	70.1
FACTSMEM[5]	0.24	-0.26	-0.39	1.86	39.5	71.7

1) Pearson correlation coefficient between the energy values and structure qualities in each measure

2) Energy evaluated by dDFIRE (Yang and Zhou, 2008a)

3) Energy evaluated by IMM1 (Lazaridis, 2003)

4) Energy evaluated by FACTS (Haberthur and Caflisch, 2008)

5) Energy evaluated by FACTSMEM (Carballo-Pacheco et al., 2014)

The weight parameter optimization of the energy function for GPCR-ligand complex structure refinement written in **subsection 4.2.2**. resulted in the weights $w_{electrostatics}$, $w_{FACTSMEM,SA}$, $w_{\phi/\varphi}$, w_{hbond}, w_{dDFIRE}, $w_{receptor\_rsr}$, and w_{ligand_rsr} to be 0.4, 0.03, 2.5, 3.0, 5.0, 2.0, and 0.5, respectively. The relative strengths were determined by optimizing the decoy discrimination performance on the GPCR training set. Same decoy conformations and same measures to evaluate decoy discrimination performance described in the previous part for selecting the implicit solvation model were applied. **Table 4.2** shows the test result of applying two energy functions, one of GalaxyRefine and the other optimized for Galaxy7TM for simple MD relaxation on GPCR model structures. Overall MD relaxation with iterative side chain perturbation, same as the 'Mild' version of GalaxyRefine, was applied with different energy functions. In both cases, same weight parameter of 5.0 for the restraint energy function applied on the initial structure itself was applied. This resulted in small change in two global structure accuracy measures, Cα-RMSD and GDT-HA. The results obtained by two energy functions were also comparable. However, when the relaxed conformation was evaluated for its side chain accuracy, refinement with Galaxy7TM energy function resulted in bigger improvement, from 33.4 to 40.3 compared to 38.8.

Table 4.2. Molecular dynamics relaxation performances using the energy function of GalaxyRefine and Galaxy7TM. The qualities were averaged on the training set targets. Qualities of the energy minimum conformation and the best out of five lowest energy conformations (in parentheses) are reported.

Model generation method	Evaluation measure		
	Cα-RMSD	GDT-HA	GDC-SC
Initial model[1]	2.79	59.2	33.4
GalaxyRefine[2]	2.76 (2.59)	59.5 (62.2)	38.8 (40.5)
GalaxyRefine with the Galaxy7TM energy function[3]	2.74 (2.59)	59.3 (62.5)	40.3 (42.3)

1) Structure quality of the initial GPCR model built by GalaxyTBM (Ko et al., 2012)

2) Result of applying GalaxyRefine (Heo et al., 2013) with overall MD relaxation and side chain perturbation

3) Result of applying overall MD relaxation and side chain perturbation but with the refinement energy function optimized for Galaxy7TM complex structure refinement stage

4.3.2. Overall performance of Galaxy7TM

The Galaxy7TM method (Lee and Seok, 2016) was tested on the 125 GPCR-ligand inputs. The result of predicting the structure and interactions of GPCR and ligand complex was evaluated in two perspectives. First, it was evaluated in terms of predicting the ligand docking accuracy. Second, it was evaluated in terms of refining the receptor model structure.

4.3.2.1. Performance of Galaxy7TM in terms of docking accuracy

By focusing on ligand docking performance of the method, ligand RMSD of the final conformations and the fraction of correctly predicted receptor-ligand contacts compared to the crystal structure were evaluated. The result was compared to AutoDock Vina (Trott and Olson, 2010), which is a widely-used, open source docking program. The summary of the results are shown as **Table 4.3**. Galaxy7TM predicted the ligand binding pose with accuracy measured in ligand RMSD less than 2 Å for 20.8% of the cases. AutoDock Vina could predict accurate ligand binding poses for 16.0% of the cases. The best of ten selected predictions were considered for each target for this analysis. When the docking accuracy was defined as predicting more than 30% of native receptor-ligand contacts (defined as two atoms are within 4Å), Galaxy7TM succeeded in 24.8% of the cases. Autodock Vina was successful at predicting 15.2% of the cases. The conformation with the best contacts out of ten final conformations was also considered for this measure. The head-to-head comparisons of the docking accuracy of Galaxy7TM and AutoDock Vina for all benchmark test set targets are illustrated in **Figure 4.1**. For both measures in ligand RMSD and correctly predicted fraction of native contacts, Galaxy7TM outperforms AutoDock Vina. **Figure 4.2** illustrates the case of Galaxy7TM outperforming AutoDock Vina. Both methods were applied on a

model structure built by MODELLER for human orexin receptor type 2 and a ligand, suvorexant. Autodock Vina resulted in contact ratio of 22.0 % (**Figure 4.2A**) and Galaxy7TM in 41.0% (**Figure 4.2B**).

In **Table 4.3**, the performance of AutoDock Vina applied to the refined receptor model structure using Rosetta MPrelax (Alford et al., 2015) is additionally shown. This test was performed to evaluate the effect of receptor structure refinement on protein ligand docking. MPrelax is a method for membrane protein model structure refinement without considering the ligand. Interestingly, the success rates of AutoDock Vina became worse. This implies that receptor model refinement without considering the ligand can have limit in predicting accurate protein-ligand complex structures and interactions.

To incorporate receptor flexibility in protein ligand docking, the widely used docking programs provide options to treat receptor sidechains flexible. AutoDock Vina also has an option allowing receptor side chain flexibility. However, the side chains to be flexible have to be assigned before running docking. Sidechain-flexible version of Autodock Vina was applied to the benchmark test set after assigning the residues with native contacts to be flexible. This resulted in success ratios of 16.0 and 16.0% for two docking accuracy measures, ligand RMSD and correctly predicted receptor-ligand contacts.

Table 4.3. Comparison of Galaxy7TM with Autodock Vina in terms of docking accuracy on a benchmark test set of 125 GPCR-ligand inputs. Performacne of Autodock Vina was tested by applying to the input receptor model ('Input-Vina') and to the refined receptor model using MPrelax ('MPrelax-Vina'). Autodock Vina with flexible receptor sidechains was also tested on input receptor model ('Input-VinaFlex').

Docking accuracy	**Percentage of successful cases for the best of 10 predictions**			
	Galaxy7TM	**Input-Vina**	**Input-VinaFlex**	**MPrelax-Vina**
Ligand RMSD (≤ 2.0Å)	20.8	16.0	16.0	6.4
Contact ratio (≥ 30.0%)	24.8	15.2	16.0	12.0

Figure 4.1. Head-to-head comparison of Galaxy7TM and AutoDock Vina on their docking accuracy. Docking accuracy was measured by (A) ligand RMSD (Å) and (B) fraction (%) of correctly predicted native contacts.

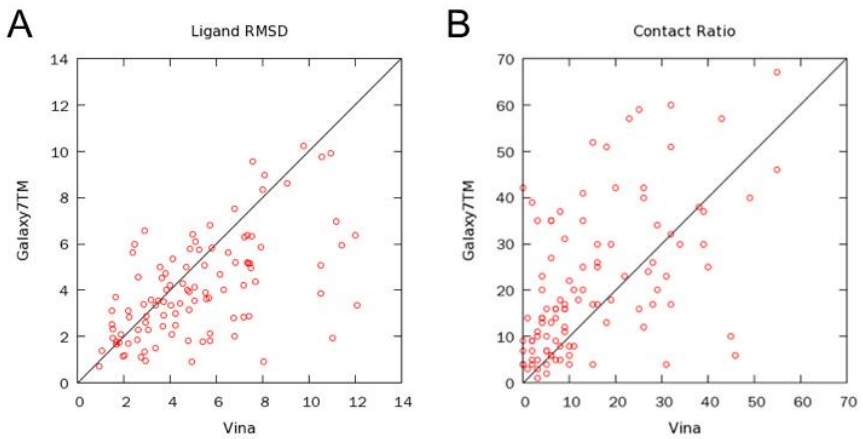
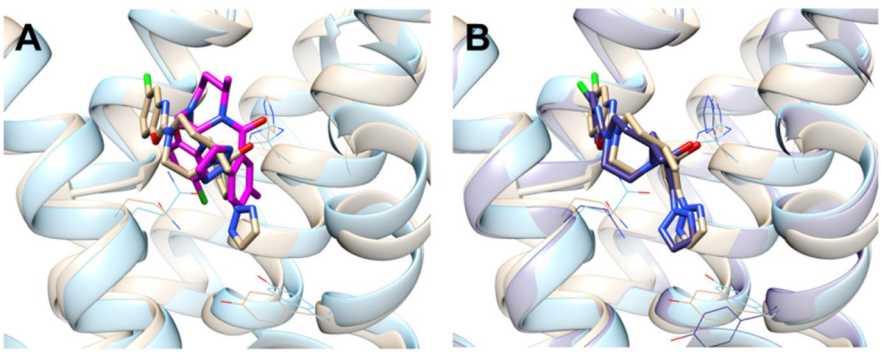


Figure 4.2. Successful example of applying Galaxy7TM (B) compared to applying AutoDock Vina (A) to a GPCR model structure built by MODELLER for human orexin receptor type 2 (light blue) and a ligand, suvorexant. The result conformations (magenta for AutoDock Vina and dark purple for Galaxy7TM) are compared to the crystal structure in light brown.



4.3.2.2. Performance of Galaxy7TM in terms of improving the receptor structure quality

Improvement in receptor structure accuracy is reported in **Table 4.4**. Improvement of structure quality was measured by GDT-HA, GDC-SC, and Cα-RMSD. Galaxy7TM resulted in 78.4, 93.6, and 74.4% of cases improved in each measure, respectively when best out of 10 final conformations were considered. The success rates were higher than those of GalaxyRefine (Heo et al., 2013) developed for soluble proteins, which are 77.6, 84.0, and 71.2%, respectively. Rosetta MPrelax (Alford et al., 2015) developed for membrane proteins resulted in 46.4, 84.8, and 62%. Similar trends were observed for the same evaluations targeting ligand-binding residues. Although the success rates are high, the magnitude of improvement is limited. Development of structure refinement methods targeting large-scale refinement will be pursued in future works. Another challenge of Galaxy7TM is that although it can predict the interaction between GPCR and ligand, it cannot discriminate between agonists and antagonists/inverse agonists. This is an important issue to be tackled from now on, as it is related to predicting the activity or efficacy of the GPCR when a specific ligand is bound.

Table 4.4. Comparison of Galaxy7TM with GalaxyRefine and MPrelax in terms of improvement in receptor structure quality on a benchmark test set of 125 GPCR-ligand inputs. The accuracy of the receptor structure was evaluated for full structure and for binding pocket residues (reported in parentheses).

Improvement in receptor structure quality	Percentage of successful cases for the best of 10 predictions		
	Galaxy7TM	GalaxyRefine	MPrelax
ΔGDT-HA (> 0.0)	78.4 (68.0)	77.6 (55.2)	46.4 (52.8)
ΔGDC-SC (> 0.0)	93.6 (88.8)	84.0 (52.8)	84.8 (80.0)
ΔCα-RMSD (< 0.0)	74.4 (75.2)	71.2 (64.8)	62.4 (60.8)

4.3.3. The relation between the docking accuracy of Galaxy7TM predictions and the receptor model quality

Although Galaxy7TM incorporates full receptor flexibility, due to the limit in sampling conformations with large-scale movements, it was expected that the performance in docking would be dependent on the quality of the initial GPCR model. To inspect this correlation, docking accuracy measured in ligand RMSD was observed with the initial receptor quality. The receptor model quality was assumed to be correlated with sequence identity between query and template GPCR sequence. In **Figure 4.3**, this result is illustrated for the targets with initial GPCR models generated by MODELLER or GalaxyTBM. From the figure, if we focus on the worst of the predictions on targets within certain range of sequence identity, the ligand RMSD tends to increase as the sequence identity decrease. Still however, the docking accuracies of the predictions within close sequence identity range had broad distribution, implying that there are more factors affecting docking accuracy.

Seven examples of Galaxy7TM predictions marked in **Figure 4.3** were closely observed to pinpoint the factors affecting the different docking performance. The results are illustrated in **Figure 4.4**. The prediction result of μ-opioid receptor model bound to the agonist BU72 (PDB ID: 5c1m) is shown in **Figure 4.4A**. The sequence identity between query and template was 59.2 and resulted in high model accuracy, 46.9 in GDT-HA. A very accurate prediction (ligand RMSD = 1.31 Å and contact ratio = 30.3%) was achieved. The accurate docking mainly resulted from correctly captured charge interaction between the positive N atom of the ligand and a negative O atom of Asp147. However, when Galaxy7TM is applied to the model of nociceptin receptor with a peptide-mimetic antagonist compound-24 (PDB ID: 4ea3), althouth the sequence identity and GDT-HA of the input receptor were much

higher (58.9% and 74.2), a low-accuracy prediction (ligand RMSD = 4.82 Å and contact ratio = 16.1%) was obtained. This was a result of wrong interaction between a positive N atom of ligand and a negative O atom of Asp130 in the refined model. Also, sidechain structure of an important binding pocket residue Met134 was inaccurately predicted. In addition, although a bound water molecule is observed in the crystal structure, this effect could not be explicitly considered with the current method.

Examples illustrated in **Figure 4.4C**, **D**, and **E** are predictions on targets with sequence identity within 30-40%. For the receptor model of 5-Hydroxytryptamine receptor 2 and the ligand ergotamine (PDB ID: 4ib4), GDT-HA of the receptor model was 52.2. It resulted in accurate docking with ligand RMSD of 1.60 Å and 44.7% contact ratio (**Figure 4.4C**). For this case, hydrophobic interactions between binding site residues (Trp337, Phe340, and Phe341) and hydrophobic moiety of the ligand was accurately predicted. Charge interaction between a positive N atom of the ligand and a negative O atom of Asp135 could be correctly captured. In the other hand, corticotropin-releasing factor receptor model with the ligand CP-376395 (PDB ID: 4k5y) resulted in low-accuracy prediction (ligand RMSD = 16.6 Å and contact ratio = 3.3%, **Figure 4.4D**). Although the receptor had intermediate structure quality (GDT-HA of 46.0) the docking grid box position was inaccurately positioned due to failure in binding residue prediction. Inaccurate result was also achieved for the model of β-2 adrenergic receptor bound to the agonist BI167107 (PDB ID: 4lde, **Figure 4.4E**). This receptor model was built with sequence identity of 35.5% and the GDT-HA was 56.9. The ligand RMSD of the prediction was 6.39 Å and the contact ratio was 11.1%. This inaccurate prediction was due to the wrong sequence alignment near the extracellular loop region. While the loop structure was blocking the native ligand-

binding position in the input structure, it could not be refined using Galaxy7TM. In the future, loop modeling will be included as a part of refinement to improve Galaxy7TM performance.

Examples of **Figure 4.4F** and **G** are those with low sequence identity, lower than 30%. Orexin receptor type 2 model bound to suvorexant (PDB ID: 4s0v) had relatively low sequence identity (22.5%) and GDT-HA of input structure (60.9). However, it resulted in very high-accuracy prediction (ligand RMSD = 0.68 Å and contact ratio = 40.7%). This may have resulted because the input structure quality in GDT-HA is relatively high compared to the sequence identity. Also, local interactions such as those between a ligand O atom and N atom of Asn324 could be captured (**Figure 4.4F**). However, for the model of Proteinase-activated receptor 1 bound to vorapaxar (PDB ID: 3vw7, **Figure 4.4G**) low-accuracy prediction with ligand RMSD = 8.90 Å and contact ratio = 8.6% was obtained. The receptor model with GDT-HA 45.8 was built with sequence identity 21.1%. The different binding pocket shape compared to the crystal structure resulted by inaccurate transmembrane helix orientations could not be refined using Galaxy7TM. Galaxy7TM has limitations for input structures generated by using templates with sequence identity lower than 20% and refinement methods incorporating large-scale movements will be pursued in the future.

Figure 4.3. A plot of ligand RMSD of the complex structure predicted by Galaxy7TM versus the sequence identity of the query to the template GPCR. Black filled dots marked as A–G represent the specific targets illustrated in **Figure 4.4**. Only the results of targets with input structure built by GalaxyTBM or MODELLER were shown.

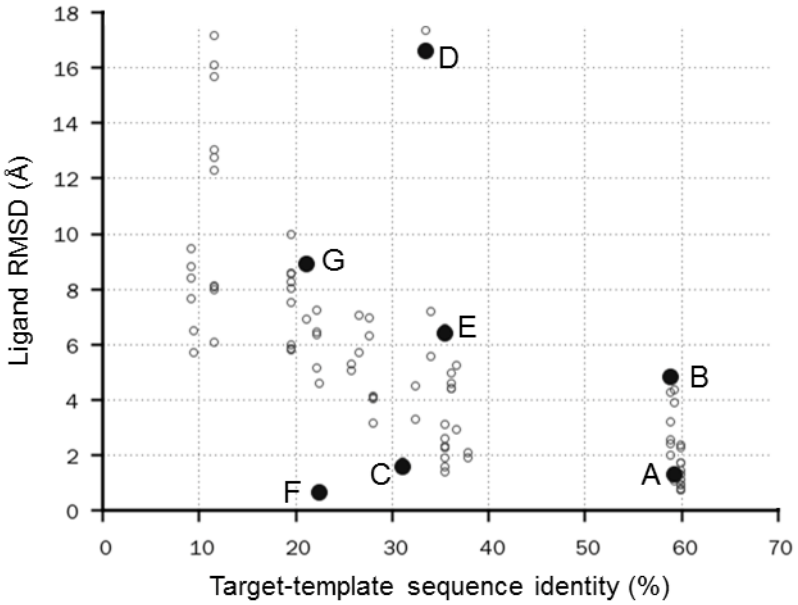
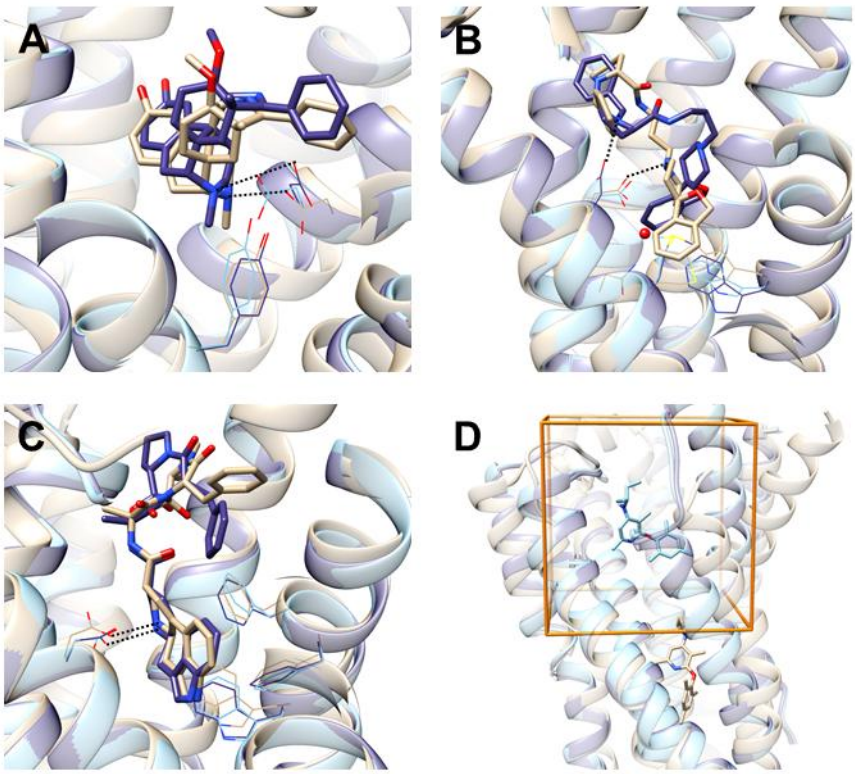
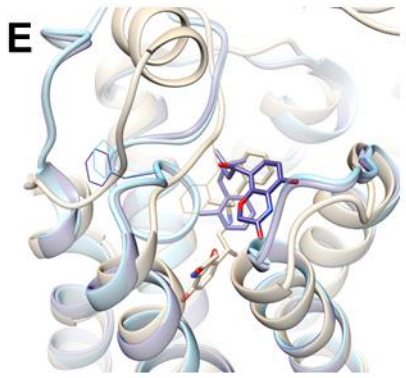
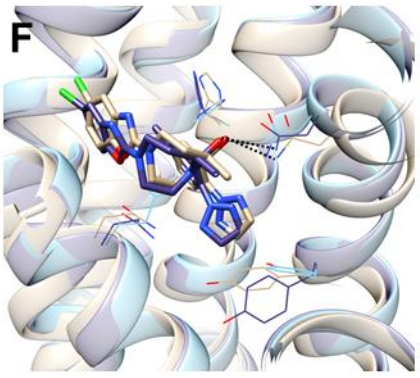
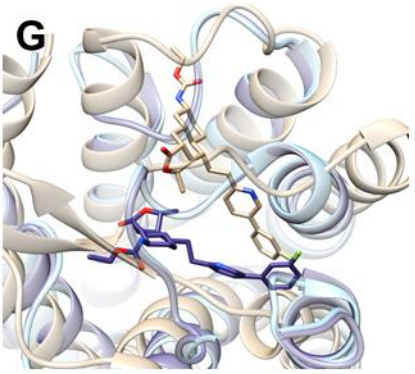


Figure 4.4. Galaxy7TM prediction result examples. Input GPCR structure is shown in light blue, Galaxy7TM result in purple, and the crystal structure in light brown (continued in the next page).




E


F


G

4.4. Conclusion

A GPCR-ligand docking method incorporating receptor flexibility through complex structure refinement was developed. The method especially targets problems of predicting the complex structure when the model receptor structure and ligand structure are given. The unreliable docking environment decreases the docking performance of methods developed in the high resolution structural environment. To overcome the error embedded in the receptor model, ligands were initially docked in low resolution to an ensemble of receptor conformations generated with ANM. Complex structure refinement was applied to the initially docked structures with a hybrid energy function including an implicit solvation model describing the lipid environment and knowledge-based energy terms.

When the method was tested on a set of GPCR models and ligands, it could predict the complex structures and interactions with higher accuracy than the widely used programs, AutoDock Vina and Rosetta MPrelax. The method particularly outperformed AutoDock Vina in predicting native receptor-ligand contacts. This method has potential to be applied for GPCR functional studies or drug discovery.

In terms of complex structure refinement, the method currently does not incorporate large-scale movements such as loop modeling or secondary structure perturbations. This restrained the degree of change and improvement of the receptor model structure quality. Also, the method cannot predict the activity of the receptor that will be resulted by the given ligand binding. These limitations will be tackled in future works.

Chapter 5

5. Conclusion

Most of the protein modeling methods widely used these days were optimized using high resolution experimental structures. The performance of these methods developed by assuming an ideal situation will be impeded when they are to be used in unreliable structural environments. The practical protein modeling situations can include cases of modeling low resolution structures, structures in different states, and predicted model structures.

In this research, protein modeling methods tackling the problem of structure refinement in unreliable structural environments were developed. Protein loop modeling in inaccurate structural environments was first developed. This was extended to refining the overall structure. Finally, predicting the GPCR-ligand structure based on complex structure refinement was pursued. These methods were developed with common approaches to overcome the effect of unreliable structure environments in refinement. Initial conformational search stage in low resolution and development of a hybrid energy function containing knowledge-based energy function with smooth energy landscapes were shared.

The methods succeeded in refining the target problems. However, they are still limited in overcoming large degrees of environmental error. This can be tackled in the future by developing a refinement method incorporating large-scale movements and an energy function which can accurately discriminate the conformations distributed in wide range.

Bibliography

Adhikari, B., Bhattacharya, D., Cao, R.Z., and Cheng, J.L. (2015). CONFOLD: Residue-residue contact-guided ab initio protein folding. Proteins *83*, 1436-1449.

Alford, R.F., Koehler Leman, J., Weitzner, B.D., Duran, A.M., Tilley, D.C., Elazar, A., and Gray, J.J. (2015). An Integrated Framework Advancing Membrane Protein Modeling and Design. PLoS Comput Biol *11*, e1004398.

Amaro, R.E., Minh, D.D.L., Cheng, L.S., Lindstrom, W.M., Olson, A.J., Lin, J.H., Li, W.W., and McCammon, J.A. (2007). Remarkable loop flexibility in avian influenza N1 and its implications for antiviral drug design. J Am Chem Soc *129*, 7764-+.

Andrusier, N., Mashiach, E., Nussinov, R., and Wolfson, H.J. (2008). Principles of flexible protein-protein docking. Proteins *73*, 271-289.

Arnautova, Y.A., Abagyan, R.A., and Totrov, M. (2011). Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling. Proteins *79*, 477-498.

Bakan, A., Meireles, L.M., and Bahar, I. (2011). ProDy: protein dynamics inferred from theory and experiments. Bioinformatics *27*, 1575-1577.

Banerjee, S., Bartesaghi, A., Merk, A., Rao, P., Bulfer, S.L., Yan, Y., Green, N., Mroczkowski, B., Neitz, R.J., Wipf, P.*, et al.* (2016). 2.3 A resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition. Science *351*, 871-875.

Best, R.B., Zhu, X., Shim, J., Lopes, P.E.M., Mittal, J., Feig, M., and MacKerell, A.D. (2012). Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. J Chem Theory Comput *8*, 3257-3273.

Braberg, H., Webb, B.M., Tjioe, E., Pieper, U., Sali, A., and Madhusudhan, M.S. (2012). SALIGN: a web server for alignment of multiple protein sequences and structures. Bioinformatics *28*, 2072-2073.

Carballo-Pacheco, M., Vancea, I., and Strodel, B. (2014). Extension of the FACTS Implicit Solvation Model to Membranes. J Chem Theory Comput *10*, 3163-3176.

Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The Amber biomolecular simulation programs. J Comput Chem *26*, 1668-1688.

Chen, J.H., and Brooks, C.L. (2007). Can molecular dynamics simulations provide

high-resolution refinement of protein structure? Proteins *67*, 922-930.

Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta crystallographica Section D, Biological crystallography *66*, 12-21.

Chopra, G., Summa, C.M., and Levitt, M. (2008). Solvent dramatically affects protein structure refinement. P Natl Acad Sci USA *105*, 20239-20244.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1996). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). J Am Chem Soc *118*, 2309-2309.

Coutsias, E.A., Seok, C., Jacobson, M.P., and Dill, K.A. (2004). A kinematic view of loop closure. J Comput Chem *25*, 510-528.

de Bakker, P.I.W., DePristo, M.A., Burke, D.F., and Blundell, T.L. (2003). Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the generalized born solvation model. Proteins-Structure Function and Genetics *51*, 21-40.

Decanniere, K., Desmyter, A., Lauwereys, M., Ghahroudi, M.A., Muyldermans, S., and Wyns, L. (1999). A single-domain antibody fragment in complex with RNase A: non-canonical loop structures and nanomolar affinity using two CDR loops. Structure *7*, 361-370.

DiMaio, F., Terwilliger, T.C., Read, R.J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H.L*., et al.* (2011). Improved molecular replacement by density- and energy-guided protein structure optimization. Nature *473*, 540-U149.

Ding, F., and Dokholyan, N.V. (2013). Incorporating Backbone Flexibility in MedusaDock Improves Ligand-Binding Pose Prediction in the CSAR2011 Docking Benchmark. J Chem Inf Model *53*, 1871-1879.

Dorsam, R.T., and Gutkind, J.S. (2007). G-protein-coupled receptors and cancer. Nature reviews Cancer *7*, 79-94.

Dunbrack, R.L., Jr., and Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol *230*, 543-574.

Fan, H., and Mark, A.E. (2004). Refinement of homology-based protein structures by molecular dynamics simulation techniques. Protein science : a publication of the Protein Society *13*, 211-220.

Fiser, A., Do, R.K., and Sali, A. (2000). Modeling of loops in protein structures. Protein science : a publication of the Protein Society *9*, 1753-1773.

Fiser, A., and Sali, A. (2003). ModLoop: automated modeling of loops in protein structures. Bioinformatics *19*, 2500-2501.

Geney, R., Layten, M., Gomperts, R., Hornak, V., and Simmerling, C. (2006). Investigation of salt bridge stability in a generalized born solvent model. J Chem Theory Comput *2*, 115-127.

Gront, D., Kmiecik, S., Blaszczyk, M., Ekonomiuk, D., and Kolinski, A. (2012). Optimization of protein models. Wires Comput Mol Sci *2*, 479-493.

Gront, D., Kulp, D.W., Vernon, R.M., Strauss, C.E., and Baker, D. (2011). Generalized fragment picking in Rosetta: design, protocols and applications. PloS one *6*, e23294.

Haberthur, U., and Caflisch, A. (2008). FACTS: Fast analytical continuum treatment of solvation. J Comput Chem *29*, 701-715.

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

Holtby, D., Li, S.C., and Li, M. (2013). LoopWeaver: Loop Modeling by the Weighted Scaling of Verified Proteins. J Comput Biol *20*, 212-223.

Huang, Y.J.P., Mao, B.C., Aramini, J.M., and Montelione, G.T. (2014). Assessment of template-based protein structure predictions in CASP10. Proteins *82*, 43-56.

Ishitani, R., Terada, T., and Shimizu, K. (2008). Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations. Mol Simulat *34*, 327-336.

Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J.F., Honig, B., Shaw, D.E., and Friesner, R.A. (2004). A hierarchical approach to all-atom protein loop prediction. Proteins *55*, 351-367.

Jagielska, A., Wroblewska, L., and Skolnick, J. (2008). Protein model refinement using an optimized physics-based all-atom force field. P Natl Acad Sci USA *105*, 8268-8273.

Jones, D.T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics *31*, 999-1006.

Jorgensen, W.L., Maxwell, D.S., and TiradoRives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and

properties of organic liquids. J Am Chem Soc *118*, 11225-11236.

Katritch, V., Cherezov, V., and Stevens, R.C. (2013). Structure-function of the G protein-coupled receptor superfamily. Annu Rev Pharmacol Toxicol *53*, 531-556.

Keedy, D.A., Williams, C.J., Headd, J.J., Arendall, W.B., 3rd, Chen, V.B., Kapral, G.J., Gillespie, R.A., Block, J.N., Zemla, A., Richardson, D.C.*, et al.* (2009). The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. Proteins *77 Suppl 9*, 29-49.

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.

Kim, D.E., Chivian, D., and Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res *32*, W526-531.

Kim, D.E., DiMaio, F., Wang, R.Y.R., Song, Y.F., and Baker, D. (2014). One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. Proteins *82*, 208-218.

Kitchen, D.B., Decornez, H., Furr, J.R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. Nature reviews Drug discovery *3*, 935-949.

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.

Kopp, J., Bordoli, L., Battey, J.N.D., Kiefer, F., and Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. Proteins *69*, 38-56.

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.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778-795.

Kryshtafovych, A., Monastyrskyy, B., and Fidelis, K. (2014). CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins *82 Suppl 2*, 7-13.

Kufareva, I., Katritch, V., Participants of, G.D., Stevens, R.C., and Abagyan, R.

(2014). Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: meeting new challenges. Structure *22*, 1120-1139.

Lagerstrom, M.C., and Schioth, H.B. (2008). Structural diversity of G protein-coupled receptors and significance for drug discovery. Nature reviews Drug discovery *7*, 339-357.

Lazaridis, T. (2003). Effective energy function for proteins in lipid membranes. Proteins *52*, 176-192.

Lazaridis, T., and Karplus, M. (1999). Effective energy function for proteins in solution. Proteins *35*, 133-152.

Lee, G.R., Heo, L., and Seok, C. (2016). Effective protein model structure refinement by loop modeling and overall relaxation. Proteins *84* (S1), 293-301.

Lee, G.R., and Seok, C. (2016). Galaxy7TM: flexible GPCR-ligand docking by structure refinement. Nucleic Acids Res *44*, W502-506.

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, J., Scheraga, H.A., and Rackovsky, S. (1997). New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. J Comput Chem *18*, 1222-1232.

Lee, M.R., Tsai, J., Baker, D., and Kollman, P.A. (2001). Molecular dynamics in the endgame of protein structure prediction. J Mol Biol *313*, 417-430.

Liang, S.D., Zhang, C., Sarmiento, J., and Standley, D.M. (2012). Protein Loop Modeling with Optimized Backbone Potential Functions. J Chem Theory Comput *8*, 1820-1827.

MacCallum, J.L., Hua, L., Schnieders, M.J., Pande, V.S., Jacobson, M.P., and Dill, K.A. (2009). Assessment of the protein-structure refinement category in CASP8. Proteins *77 Suppl 9*, 66-80.

MacCallum, J.L., Perez, A., and Dill, K.A. (2015). Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. Proc Natl Acad Sci U S A *112*, 6985-6990.

MacCallum, J.L., Perez, A., Schnieders, M.J., Hua, L., Jacobson, M.P., and Dill, K.A. (2011). Assessment of protein structure refinement in CASP9. Proteins *79 Suppl 10*, 74-90.

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.

Mandell, D.J., Coutsias, E.A., and Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods *6*, 551-552.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics *29*, 2722-2728.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D Structure Computed from Evolutionary Sequence Variation. PloS one *6*.

Mas, M.T., Smith, K.C., Yarmush, D.L., Aisaka, K., and Fine, R.M. (1992). Modeling the Anti-Cea Antibody Combining Site by Homology and Conformational Search. Proteins-Structure Function and Genetics *14*, 483-498.

McGuffin, L.J., Bryson, K., and Jones, D.T. (2000). The PSIPRED protein structure prediction server. Bioinformatics *16*, 404-405.

Meiler, J., and Baker, D. (2006). ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins *65*, 538-548.

Mirjalili, V., and Feig, M. (2013). Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. J Chem Theory Comput *9*, 1294-1303.

Mirjalili, V., Noyes, K., and Feig, M. (2014). Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. Proteins *82*, 196-207.

Mizuguchi, K., Deane, C.M., Blundell, T.L., and Overington, J.P. (1998). HOMSTRAD: A database of protein structure alignments for homologous families. Protein Science *7*, 2469-2471.

Modi, V., and Dunbrack, R.L., Jr. (2016). Assessment of refinement of template-based models in CASP11. Proteins *84 Suppl 1*, 260-281.

Mortier, J., Rakers, C., Bermudez, M., Murgueitio, M.S., Riniker, S., and Wolber, G. (2015). The impact of molecular dynamics on drug design: applications for the characterization of ligand-macromolecule complexes. Drug Discov Today *20*, 686-702.

Nugent, T., Cozzetto, D., and Jones, D.T. (2014). Evaluation of predictions in the CASP10 model refinement category. Proteins *82 Suppl 2*, 98-111.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., and Hutchison, G.R. (2011). Open Babel: An open chemical toolbox. J Cheminform *3*, 33.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife *3*, e02030.

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

Qiu, D., Shenkin, P.S., Hollinger, F.P., and Still, W.C. (1997). The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. J Phys Chem A *101*, 3005-3014.

Raval, A., Piana, S., Eastwood, M.P., Dror, R.O., and Shaw, D.E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. Proteins *80*, 2071-2079.

Rodrigues, J.P.G.L.M., Levitt, M., and Chopra, G. (2012). KoBaMIN: a knowledge-based minimization web server for protein structure refinement. Nucleic Acids Res *40*, W323-W328.

Rose, P.W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z.*, et al.* (2016). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol *234*, 779-815.

Samudrala, R., and Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol *275*, 895-916.

Sandal, M., Duy, T.P., Cona, M., Zung, H., Carloni, P., Musiani, F., and Giorgetti, A.

(2013). GOMoDo: A GPCRs Online Modeling and Docking Webserver. PloS one *8*.

Saraste, M., Sibbald, P.R., and Wittinghofer, A. (1990). The P-loop--a common motif in ATP- and GTP-binding proteins. Trends Biochem Sci *15*, 430-434.

Sellers, B.D., Zhu, K., Zhao, S., Friesner, R.A., and Jacobson, M.P. (2008). Toward better refinement of comparative models: Predicting loops in inexact environments. Proteins *72*, 959-971.

Sherman, W., Beard, H.S., and Farid, R. (2006). Use of an induced fit receptor structure in virtual screening. Chem Biol Drug Des *67*, 83-84.

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.

Shin, W.H., Lee, G.R., and Seok, C. (2015). Evaluation of GalaxyDock Based on the Community Structure-Activity Resource 2013 and 2014 Benchmark Studies. J Chem Inf Model.

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

Stein, A., and Kortemme, T. (2013). Improvements to Robotics-Inspired Conformational Sampling in Rosetta. PloS one *8*.

Stormo, G.D., Schneider, T.D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res *10*, 2997-3011.

Subramani, A., and Floudas, C.A. (2012). Structure Prediction of Loops with Fixed and Flexible Stems. J Phys Chem B *116*, 6670-6682.

Summa, C.M., and Levitt, M. (2007). Near-native structure refinement using in vacuo energy minimization. P Natl Acad Sci USA *104*, 3177-3182.

Totrov, M., and Abagyan, R. (2008). Flexible ligand docking to multiple receptor conformations: a practical alternative. Current opinion in structural biology *18*, 178-184.

Trott, O., and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem *31*, 455-461.

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I.*, et al.* (2010). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM

all-atom additive biological force fields. J Comput Chem *31*, 671-690.

von der Ecken, J., Heissler, S.M., Pathan-Chhatbar, S., Manstein, D.J., and Raunser, S. (2016). Cryo-EM structure of a human cytoplasmic actomyosin complex at near-atomic resolution. Nature *534*, 724-728.

Wallner, B., and Elofsson, A. (2005). Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics *21*, 4248-4254.

Wolf, S., Bockmann, M., Howeler, U., Schlitter, J., and Gerwert, K. (2008). Simulations of a G protein-coupled receptor homology model predict dynamic features and a ligand binding site. FEBS Lett *582*, 3335-3342.

Xiang, Z., Soto, C.S., and Honig, B. (2002). Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci U S A *99*, 7432-7437.

Yang, J., and Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Res *43*, W174-181.

Yang, Y., and Zhou, Y. (2008a). Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins *72*, 793-803.

Yang, Y.D., and Zhou, Y.Q. (2008b). Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins *72*, 793-803.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res *31*, 3370-3374.

Zhang, J., Liang, Y., and Zhang, Y. (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure *19*, 1784-1795.

Zhang, J., Yang, J.Y., Jang, R., and Zhang, Y. (2015). GPCR-I-TASSER: A Hybrid Approach to G Protein-Coupled Receptor Structure Modeling and the Application to the Human Genome. Structure *23*, 1538-1549.

Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins *57*, 702-710.

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 science : a publication of the Protein Society *11*, 2714-2726.

Zhu, J., Fan, H., Periole, X., Honig, B., and Mark, A.E. (2008). Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. Proteins *72*, 1171-1188.

Zhu, J., Xie, L., and Honig, B. (2006a). Structural refinement of protein segments containing secondary structure elements: Local sampling, knowledge-based potentials, and clustering. Proteins *65*, 463-479.

Zhu, K., Pincus, D.L., Zhao, S.W., and Friesner, R.A. (2006b). Long loop prediction using the protein local optimization program. Proteins *65*, 438-452.

국문초록

실험으로 밝혀진 단백질 구조의 수가 기하급수적으로 늘어나고 있다. 이렇게 제공된 구조 정보를 바탕으로, 서열이 비슷한 단백질의 실험 구조를 주형으로 단백질 구조를 예측하는 방법이 보편화되었다. 하지만 단백질 구조를 신약 개발이나 단백질 디자인과 같은 응용 문제에 이용하기 위해서는 고해상도의 구조 정보가 필요하다. 이러한 상황 하에 저해상도의 단백질 구조나 예측된 단백질 구조의 정확도를 향상시킬 수 있는 단백질 구조 정밀화 방법들이 개발되었다. 단백질 구조 정밀화가 다루어야 하는 또 하나의 문제 영역은 이용자가 원하는 특정 환경 상에서의 단백질 구조를 예측하는 것이다. 예를 들어, 단백질이 특정 대상과 상호 작용할 때의 구조를 다른 상태의 단백질 실험 구조나 예측된 단백질 구조를 이용하여 예측할 수 있어야 한다.

이 논문은 각기 다른 문제 영역을 다루는 단백질 구조 정밀화 방법 네 가지를 개발한 내용과 그 결과를 담고 있다. 이 방법들은 정밀화하는 구조의 영역을 확장시키거나 단백질이 상호 작용하는 다른 대상을 고려하는 방향으로 발전되었다. 예측된 단백질 구조는 엄밀히 파악되지 않는 정도의 부정확함을 가진다. 이러한 구조를 대상으로 계산을 수행할 때 구조 환경에 대한 신뢰도가 낮으므로 정확한 구조 환경에서 최적화된 구조 예측 방법이나 에너지 함수를 적용하기 어렵다. 본 연구에서는 공통적으로 두 가지의 접근 방법을 이용하여 이 문제를 해결하기 위해 노력하였다. 한 가지는 먼저 낮은 해상도에서 구조

공간을 효율적으로 탐색하는 것이다. 다른 한 가지는 구조 환경의 오류에 덜 민감한 에너지 함수를 개발하는 것이다. 물리 기반 에너지 함수와 통계 기반 에너지 함수를 혼합하는 방식으로 이러한 에너지 함수를 개발할 수 있었다. 각 단백질 구조 정밀화 방법이 이와 같은 접근 방식을 어떻게 차용했는지 논문에서 구체적으로 다룰 것이다.

주요어: 단백질 구조 정밀화, 단백질 구조 예측, 단백질 고리 구조 모델링, 수용체 구조 변화를 동반한 단백질 리간드 도킹, 물리 기반과 통계 기반 에너지가 혼합된 에너지 함수, G 단백질 연결 수용체

학 번: 2011-23233