이학박사 학위논문

Protein Structure Modeling at the Interface of Physical Chemistry and Bioinformatics

물리화학 및 생물정보학적 접근을 통한

단백질 구조 모델링 기법 개발

2015 년 8 월

서울대학교 대학원

화학부 물리화학 전공

허 림

Protein Structure Modeling at the Interface of Physical Chemistry and Bioinformatics

지도교수 석 차 옥

이 논문을 이학박사 학위논문으로 제출함

2015 년 8 월

서울대학교 대학원

화학부 물리화학 전공

허 림

허림의 이학박사 학위논문을 인준함

2015 년 8 월

위원장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

ABSTRACT

Protein Structure Modeling at the Interface of Physical Chemistry and Bioinformatics

Lim Heo

Department of Chemistry

The Graduate School

Seoul National University

Proteins play key roles in numerous biological processes. The protein functions are related to their structure; therefore protein structures can give invaluable information to their functional studies. Protein structures can be determined by experimental methods such as X-ray crystallography, NMR, and electron microscopy. However, these methods are time consuming, require high costs, and sometimes hard to determine due to experimental limitations. To complement these problems, there have been numerous protein structure predictions with computational methods. Among these methods, template-based modeling has been successful. The method utilizes known protein structures to model unknown protein structures from their sequences. For this approach, identification of similar known protein structures is the most important step. However, only with the template-driven information, it is hard to predict accurately on the regions where the sequence deviates from the templates, or template-driven information is not sufficient for modeling. In this thesis, a new template-based modeling method

named GalaxyTBM is introduced. The method adopted best performing existing methods for template identifications based on bioinformatics. In addition, we developed several methods based on physical chemistry to complement the problems described above. By combining existing methods based on bioinformatics and newly developed methods based on physical chemistry, it was possible to achieve improvements beyond the methods based on a single approach.

keywords: protein structure prediction, template-based modeling, protein structure refinement, bioinformatics, physical chemistry

Student Number: 2010-20300

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUTION

Proteins are involved in numerous biological processes. Proteins play their roles by interacting with their partner biomolecules such as DNAs, RNAs, the other proteins, small organic compounds, and metal ions (Glusker, 1991; Holm *et al*., 1996; Kristiansen, 2004; Negri *et al*., 2010; Pawson and Nash, 2000; Ren *et al*., 2000). In addition, they are also functioning with themselves (Pawson and Nash, 2000; Poupon and Janin, 2010). The protein structures are highly related to their functions. Therefore, protein structures can give invaluable information to know their functions. About 100,000 protein structures have been deposited in the protein databank (PDB) with various experimental methods such as X-ray crystallography, NMR, and electron microscopy (EM) (Bernstein *et al*., 1977). Most of these experimental structures provide atomic resolution information on protein structures. Though experimental structure determination methods supply the exact protein structures under their experimental conditions, they are time consuming and require high costs. In addition, they are sometimes hard to determine the structures with several experimental limitations (Berman *et al*., 2000; Bill *et al*., 2011).

To complement these problems, protein structure prediction methods have been studied more than two decades (Kryshtafovych *et al*., 2011, 2014; Marti-Renom *et al*., 2000; Zhang, 2008). Protein structure prediction methods can be classified into two categories (Kinch *et al*., 2011b; Taylor *et al*., 2014): template-based modeling (TBM) and free modeling (FM). The biggest difference between TBM and FM is that TBM utilizes known protein structures as templates to predict unknown protein structures, while FM does not use known protein structures directly. Protein structures are determined by their sequences, and homologous proteins have similar structures. From this observation, it is possible to predict

unknown protein structure by using known homologous protein structures with template-based modeling methods. On the other hand, unknown protein structures are predicted only with their sequences with free modeling methods. In biophysics, the folded protein structures are considered to have the minimum free energy, and there are folding funnels near to the folded structure to fold the protein structures (Bryngelson *et al*., 1995; Leopold *et al*., 1992). Free modeling methods are based on these idea, they predicts unknown protein structures by sampling various protein conformations and evaluating their energies. Progress in the number of known protein structures, protein structure prediction with template-based modeling has been boosted with their accurate predictions.

Template-based modeling methods commonly consist of following several steps (Marti-Renom *et al*., 2000). First, they identify template protein structures (fold recognition), and align their sequences or structures with the target protein sequence (sequence alignment). From the sequence alignment and related protein structures, protein tertiary structure models are built using template information. During the template-based modeling processes, results for each step can be assessed (quality assessment; QA). Some template-based modeling methods adopt protein structure refinement processes. Numerous methods have been developed for each step, and the full template-based modeling method can be constructed by combining each methods.

The existing template-based modeling methods highly rely on modeling steps related to bioinformatics such as template identifications and sequence alignments (Ma *et al*., 2014; Soding *et al*., 2005; Xu *et al*., 2011). These steps are important because the predicted protein structures greatly depend on the used template structures. There have been numerous studies on template identifications and sequence alignments, and there are huge improvements in these methods (Ma *et al*.,

2014; Soding, 2005; Yang *et al*., 2011). In contrast, there are few studies on how to use those identified template information. In addition, most of the previous template-based modeling methods are hard to reflect structure changes upon sequence changes. Because the target protein sequence is different from the template sequences, the protein structures deviate from the template structures (Ko *et al*., 2012; Park *et al*., 2011; Park *et al*., 2014; Park and Seok, 2012). Those methods solely rely on the template information; it is hard to overcome these changes.

In this thesis, I will describe a new template-based modeling method named GalaxyTBM. Several existing methods and newly developed methods are combined together to construct a new method for structure prediction. The method adopts some best performing methods for template identifications which are based on bioinformatics. From the detected templates, protein structures are modeled with a new protein tertiary structure building method, GalaxyCassiopeia. It uses not only template-driven information, but also physicochemical energy functions to overcome structure deviations from the sequence changes. The predicted models are further refined by using *ab initio* local structure refinement methods. Regions without template or with unreliable information (ULRs) are detected, and they are re-modeled without template information. In addition to protein structure prediction, two fundamental applications with the predicted protein structures are included in the GalaxyTBM method. The whole GalaxyTBM method is organized by me, some parts of the work were performed by collaboration with my colleagues, and some methods are adopted from the state-of-the-art methods or done by my colleagues due to its complexity. The overall GalaxyTBM method is described in **chapter 2**, and detailed explanation on major part of GalaxyTBM are followed after: protein chain building method, GalaxyCassiopeia, in **chapter 3**; ligand binding site prediction method, GalaxySite (Heo *et al*., 2014), in **chapter 4**;

protein structure refinement method, GalaxyRefine (Heo *et al.*, 2013), in **chapter 5**.

2. GalaxyTBM:
Protein structure prediction using GALAXY programs

2.1. Introduction to protein structure prediction using template-based modeling approaches

Progress in the number of experimental structures of homologous proteins, protein structure prediction using template-based modeling (TBM) has been boosted. (Kryshtafovych *et al*., 2011, 2014; Zhang, 2008) With methodological improvements in genome sequencing using next generation sequencing (NGS) techniques (Mardis, 2008) and in structural determination using various experimental methods, the number of unknown sequences and known structures became flourish; therefore, template-based modeling has been and become more promising method in proteomics era.

Template-based modeling generally follows several steps (Marti-Renom *et al*., 2000): (1) template protein structure identification (fold recognition); (2) sequence alignment between the query sequence and selected templates; (3) tertiary structure model building from the sequence alignment and template information; (4) structure model quality assessment (QA); (5) tertiary structure model refinement. There have been numerous methods developed for each step separately, and the full template-based modeling method can be constructed by combining each methods.

Several methods were proven to be successful in detecting reliable protein templates based on bioinformatics approaches (Ma *et al*., 2014; Soding, 2005), and this made some progresses in structure prediction field. However, the progress with

development of fold recognition methods has been almost converged. In addition, there is few study on how to use identified template structure information and overcome regions which cannot be predicted by templates (so called unreliable local regions (ULRs)) (Ko *et al*., 2012; Park *et al*., 2011; Park and Seok, 2012). In this chapter, several methods available in GALAXY programs are described that they are proposed to deal those problems in template-based modeling approach: (1) utilization methods of successful fold recognition tools which adopted re-scoring scheme using physicochemical energy functions; (2) a new protein chain building method named GalaxyCassiopeia which uses not only template-driven information, but also physicochemical energy functions; (3) a unreliable local regions (ULRs) detection method and two distinct ULR re-modeling methods. Additionally, two practical application of protein structure prediction are described: ligand binding site prediction (Heo *et al*., 2014) and oligomer structure prediction (Lee *et al*., 2013). The method described in here is extensively tested in CASP11, and the results also would be dealt in this chapter.

2.2. Methods

2.2.1. Overall structure prediction method using GALAXY programs

Overall structure prediction method based on template-based modeling approach using GALAXY programs are described in **Figure 2.1**. For a given target sequence, domains were first detected by GalaxyDom which utilizes HHsearch (Soding, 2005) alignments to SCOP domains and PDB structures. For each domain, multiple templates were selected by ranking related proteins detected by HHsearch and RaptorX (Ma *et al*., 2014) based on the fold recognition scores, similarity among templates, and energies of the model structures built from single templates. The top

HHsearch template was always included for easy targets (predicted TM-score > 0.8) and the top RaptorX template was included for harder targets. Multiple sequence alignment produced by Promals3D (Pei *et al*., 2008) was used for subsequent model-building by GalaxyCassiopeia. In the first stage of GalaxyCassiopeia, 48 models were constructed by short VTFM MD simulations with restraints derived from templates and CHARMM22-based energy (MacKerell *et al*., 1998). The main purpose of this stage was to construct global structure for the regions covered by templates. The lowest energy model of the first stage was refined by *ab initio* loop/terminus modeling (Ko *et al*., 2012; Park *et al*., 2014) of unreliable local regions (ULRs) detected based on structural fluctuations of the first-stage models, multiple sequence alignment, and deviations from backbone torsion angles in the fragment library. Long loop ULRs (>15 residues) and terminus ULRs were rebuilt by FALC (Ko *et al*., 2011; Lee *et al*., 2010) and short MD relaxation. Short loop ULRs (≤ 15 residues) were rebuilt by GalaxyLoop (Ko *et al*., 2012; Park and Seok, 2012). After this ULR modeling, the overall structure was further refined by repetitive side chain perturbations and short MD relaxations.

The predicted monomer structure was used as an input to the prediction of homo-oligomer structure and ligand-binding pose by GalaxyGemini (Lee *et al*., 2013) and GalaxySite (Heo *et al*., 2014), respectively. GalaxyGemini predicts protein oligomer structures by selecting oligomer templates from HHsearch results. After superimposing the monomer structure onto the oligomer template, steric clashes were removed by rigid-body energy minimization. Oligomer interfaces of high-confidence predictions were further refined by GalaxyRefine (Heo *et al*., 2013). GalaxySite predicts the protein-ligand complex structure by molecular docking of predicted ligands detected from similar proteins onto the predicted monomer structure. GalaxySite has been improved by using the BioLiP database (Yang *et al*., 2013) and introducing a confidence score. Side chain flexibility in

metal binding sites was also considered during docking.

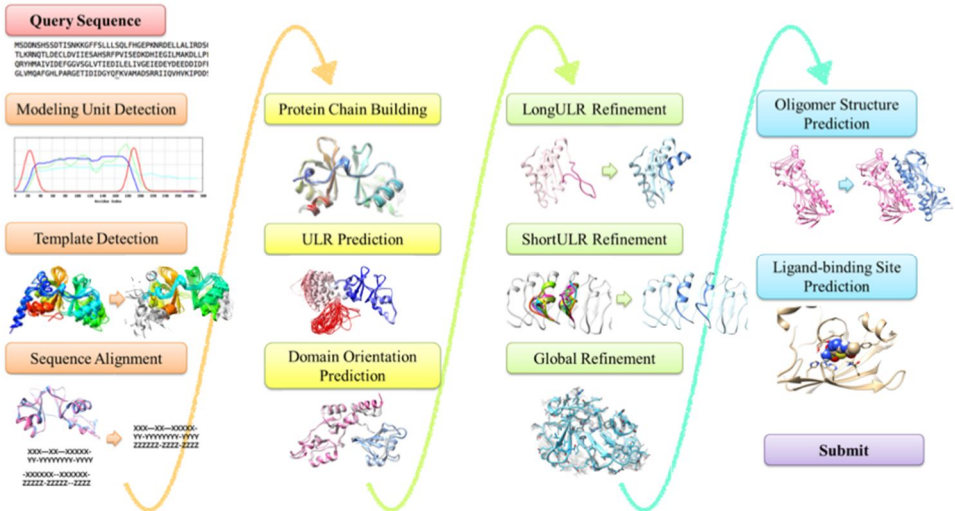

Figure 2.1. Overall flow chart for structure prediction method using GALAXY programs. Fold recognition and sequence alignment, protein structure buildings, protein structure refinement, and applications of predicted protein structures are colored in orange, yellow, green, and blue, respectively.

2.2.2. Fold recognition and sequence alignment

Identification of the target sequence regions that can be modeled reliably by template-based modeling, it is required to obtain high-quality alignments to potential templates of known structure. GalaxyDom detects modeling units which are the regions that are possible to detect reliable protein structure templates by using HHsearch (Soding, 2005). In GalaxyDom-PDB, the first version of GalaxyDom, templates are selected from the PDB (Bernstein *et al*., 1977) only. However, we found that multiple templates can also obscure domain boundaries when they are not similar enough to the target. We therefore combine sequence alignments with biological domains collected in the SCOP database (Murzin *et al*., 1995). This approach is implemented in GalaxyDom-PDB/SCOP, the second version of GalaxyDom. In CASP11, I adopted GalaxyDom-PDB/SCOP to modeling unit detection.

Both versions of GalaxyDom utilize global and local sequence alignments from the HHsearch. Global alignment gives information with higher coverage, while local alignment gives more accurate information on domain core regions. The “chunk score”, a measure of alignment quality, is determined for each residue of the target sequence for each set of templates derived from PDB and SCOP, resulting in a PDB chunk score profile and a SCOP chunk score profile, respectively. While GalaxyDom-PDB sets domain boundaries by applying a fixed cut-off value to the PDB chunk score profile, GalaxyDom-PDB/SCOP combines both chunk score profiles to predict modeling units.

The PDB70 and SCOP70 databases are searched to detect homolog structures related to the target sequence by using HHsearch. The searching process is performed only once on PDB70 database, and iteratively on SCOP70 until the lengths of unaligned regions are shorter than 35 residues. For each searching

process, it selects multiple templates after rescoring HHsearch results as described in GalaxyTBM (Ko et al., 2012). Using global and local alignments for the selected templates are then used to determine modeling units with following chunk score calculation method.

The chunk score, C_i, estimates whether the i-th residue belongs to a contiguous chunk, which can be modeled at once. The chunk score is a window averaged score with triangular scheme (window size=15) of the product of global alignment score G_j and local alignment score L_j, which are calculated using the global and local sequence alignment, respectively (**Eq. 2.1**).

$$C_i = \left\langle G_j L_j \right\rangle_{j \in (i-7, i+7)} \tag{2.1}$$

The global alignment score, G_j, is the weighted sum of neighbor alignment score:

$$G_j = \sum_{t}^{N_t} w^{(t)} \cdot g_{j,j+1}^{(t)} \cdot gl_j^{(t)} \tag{2.2}$$

where N_t is the number of selected templates, and $w^{(t)}$ is the weight for the t-th template, which is the fold recognition score ratio between t-th template and the top template. The neighbor alignment score $g_{j,j+1}^{(t)}$ is the probability that the consecutive residues j and $j+1$ in the target sequence belong to the same chunk, considering the global alignment to the t-th template. If both residues are aligned, it is calculated as $1 \Big/ \left\{ 1 + \left(d_{j,j+1}^{(t)} / 30 \right)^2 \right\}$, where $d_{j,j+1}^{(t)}$ is the distance in sequence between the two template residues aligned to the target residues j and $j+1$, otherwise it is set to 0. The final term $gl_j^{(t)}$ is set to 1 or -1 whether the j-th residue is aligned or not. The local alignment score, L_j, is calculated using **Eq. 2.3**.

$$L_j = \sum_{t}^{N_t} \sum_{a}^{N_{LA}^{(t)}} 1 \Big/ \left\{ 1 + w^{(t)} \cdot w_a^{(t)} \cdot l_{j,a}^{(t)} \right\} \tag{2.3}$$

In the equation, $N_{LA}^{(t)}$ is the number of suboptimal local alignments for each template *t*, $w_a^{(t)}$ is the weight assigned to the *a*-th local alignment of the *t*-th template, and $l_{j,a}^{(t)}$ is assigned 0 or 1 whether *j*-th residue is or not.

In GalaxyDom-PDB, modeling units are determined by using the chunk score from PDB with a fixed chunk score cutoff. The whole sequence is divided into multiple chunks with the chunk score cutoff. For a chunk longer than 35 residues are designated as a modeling unit, and if the length is shorter than 35 residues it is merged into the neighboring chunk. The chunk score cutoff is trained using CASP9 TBM targets (Mariani *et al*., 2011) to give maximum agreement to the CASP assessment units (MUs). However, a fixed cutoff approach sometimes showed boundary errors; it tended to give enlarged modeling unit regions for a region with higher chunk score. To tackle this issue, the slope of the chunk score is used for PDB chunk score and SCOP chunk score is introduced. Differently from the PDB chunk score, SCOP chunk score is calculated only with a single domain sequences, it tends to have sharp boundaries, and modeling units can be determined more precisely. In GalaxyDom-PDB/SCOP, modeling unit candidates are determined by using SCOP chunk score first. The same fixed chunk score approach is applied to SCOP chunk score to assign biological domains. For the PDB chunk score, a finite difference is calculated, and multiplying the derivative with a Gaussian function which has peak at the cutoff value to capture locations with derivatives near the cutoff. Among the modeling unit candidates, candidate boundaries with higher than 0.35 PDB chunk score derivative are selected as modeling unit boundaries.

For each detected modeling unit, HHsearch (Soding, 2005) and RaptorX (Ma

et al., 2014) are used for detecting homolog structures. I benchmarked several fold recognition methods, and HHsearch and RaptorX showed good performance on my benchmark set (data not shown). HHsearch searches homologs on 'pdb70', structure database with maximum mutual sequence identity 70%, which was built on Jul. 14th, 2012 (for the benchmark), and Apr. 24th, May 17th, and Jun. 26th, 2014 (for the CASP11 experiment). With HHsearch, there are four possible combinations for running HHsearch: global/MAC, global/Viterbi, local/MAC, and local/Viterbi. Among these running methods, global/MAC algorithm showed the best performance on easy TBM targets (data not shown). RaptorX searches homologs in two steps. For the first step, it searches on 'pdb40' (maximum sequence identity 40%) and select top 100 template candidates. For the second step, it searches on un-clustered structure database ('pdb100') with the selected top 100 template candidates. Structure databases for RaptorX were built on May 1st, 2012 (for the benchmark) and Apr. 30th, 2014 (for the CASP11 experiment). HHsearch showed better performance in easy template-based modeling regions (TM-score > 0.8), while RaptorX showed better performance in medium-to-hard template-based modeling regions (TM-score < 0.8) (**Figure 2.2**).

A target difficulty prediction method is developed to estimate target difficulty. It predicts TM-score (Zhang and Skolnick, 2004) between target protein structure and the most similar protein structure in the structure databases. It is composed of sigmoidal function of linear regression results with various features come from the fold recognition methods (**Eq. 2.4**, **Eq. 2.5**, and **Eq. 2.6**). It is trained on the benchmark set by using 2-fold cross validation (**Figure 2.3**). In **Eq. 2.5**, features with superscript 'HHsearch' and 'RaptorX' are come from the top-ranked template candidate for HHsearch and RaptorX search, respectively. The difficulty estimation method shows a correlation coefficient 0.881 (for training) and 0.775 (for test), and root-mean square error 0.054 (for training) and 0.081 (for test).

$$\mathrm{TM}_{\mathrm{pred}} = \left[1 + \exp\left[LR_{TM}\right]\right]^{-1} \tag{2.4}$$

$$\begin{aligned} LR_{TM} = & -6.493\times10^{-3}\left(\mathrm{N}_{\mathrm{res}}\right)^3 \\ & +6.521\times10^{-2}\left(f_{\mathrm{Helix}}^{\mathrm{PSIPRED}}\right)+0.3311\left(f_{\mathrm{Strand}}^{\mathrm{PSIPRED}}\right) \\ & +0.771\left(\mathrm{TM}_{\mathrm{templ}}^{\mathrm{HHsearch}}+\mathrm{TM}_{\mathrm{templ}}^{\mathrm{RaptorX}}\right) \\ & +2.780\times10^{-3}\left(\mathrm{Score}^{\mathrm{RaptorX}}\right)-3.691\times10^{-3}\left(-\log \mathrm{P}_{\mathrm{value}}^{\mathrm{RaptorX}}\right) \\ & +2.900\times10^{-4}\left(\mathrm{L}_{\mathrm{templ}}^{\mathrm{RaptorX}}\right)-3.606\times10^{-3}\left(\mathrm{N}_{\mathrm{templ\_gap}}^{\mathrm{RaptorX}}\right) \\ & +4.457\times10^{-3}\left(\mathrm{SeqID}^{\mathrm{RaptorX}}\right) \\ & +1.346\times10^{-3}\left(\mathrm{Score}^{\mathrm{HHsearch}}\right)+1.079\times10^{-2}\left(\mathrm{SS}^{\mathrm{HHsearch}}\right) \\ & +7.028\times10^{-3}\left(\mathrm{Prob}^{\mathrm{HHsearch}}\right)+4.014\times10^{-3}\left(\mathrm{L}_{\mathrm{align}}^{\mathrm{HHsearch}}\right) \\ & -0.897 \end{aligned} \tag{2.5}$$

$$\mathrm{TM}_{\mathrm{templ}}^{\mathrm{Method}} = \frac{\sum_{x=1}^{20} e^{-(x-1)/2}\mathrm{TM}(\text{rank } 1^{st}, \text{ rank } x^{th})}{\sum_{x=1}^{20} e^{-(x-1)/2}} \tag{2.6}$$

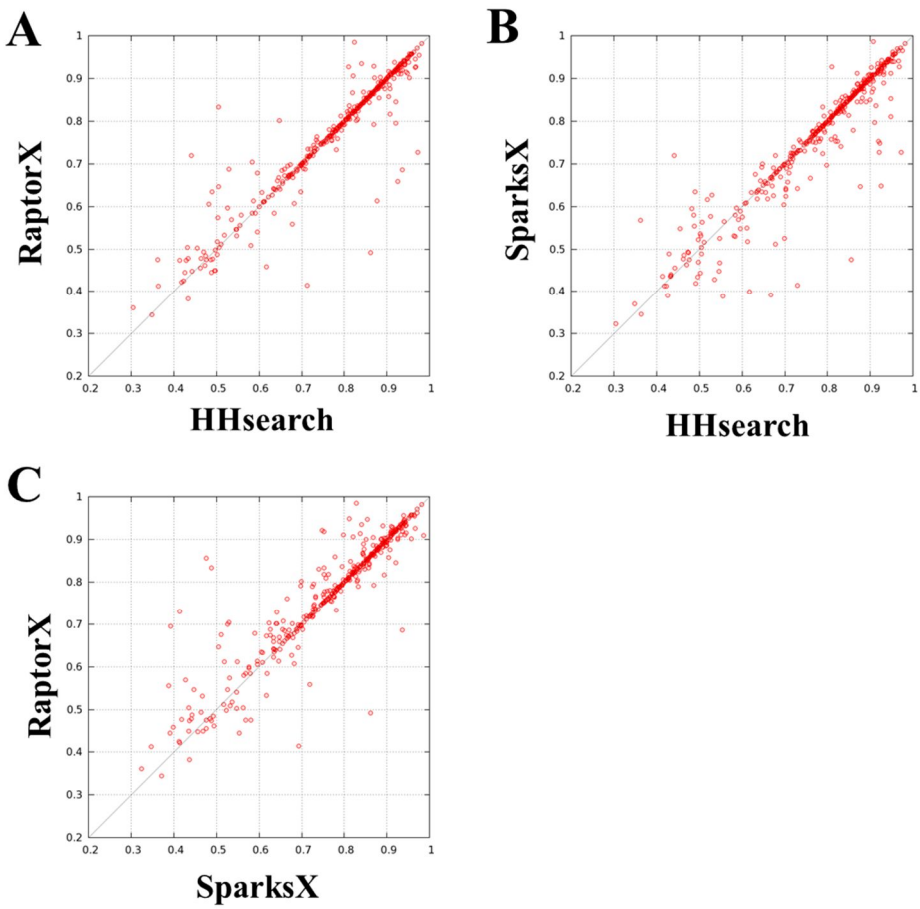

Figure 2.2. Fold recognition performance benchmark test result. The highest TM-score among top 30 fold recognition result for each target is plotted. Comparison between (A) RaptorX and HHsearch, (B) SparksX and HHsearch, and (C) RaptorX and SparksX. When RaptorX and HHsearch is compared, HHsearch showed better results for easy targets (TM-score > 0.8) than RaptorX, while RaptorX was better for medium-to-hard targets (TM-score < 0.8). SparksX showed worse results on overall ranges of targets than RaptorX and HHsearch.

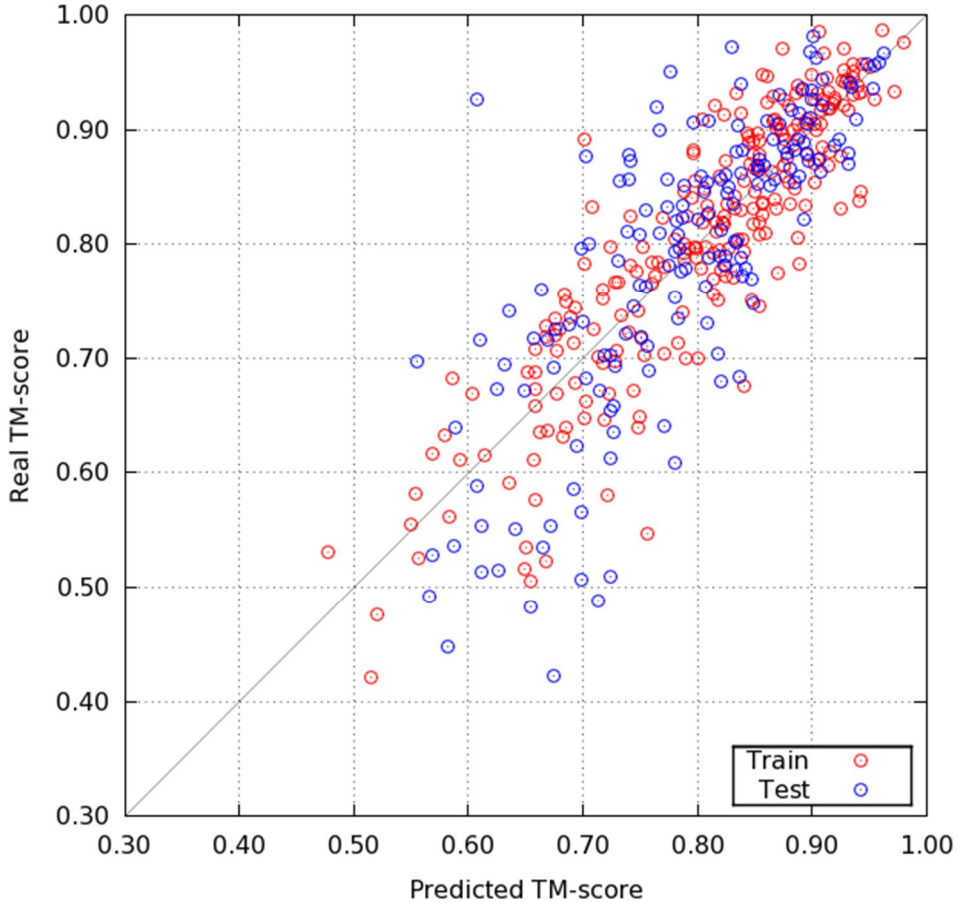

Figure 2.3. Two-fold cross validation results for developing target difficulty estimation method.

To maximize the fold recognition performance, HHsearch, RaptorX is applied on easy TBM targets (predicted TM-score > 0.8) and medium-to-hard TBM targets (predicted TM-score < 0.8), respectively. HHsearch tends to detect good templates within top 30 template candidates, but sometimes failed to rank among them. And as I adopted HHsearch for easy TBM targets, not for overall targets, it should be re-ranked to focus on easy TBM targets and resolve ranking issues. To re-rank the HHsearch result, it is sorted by using the score **Eq. 2.7**, which incorporates the predicted target difficulty and sequence identity.

$$H_{fr} = S_{seq} + (6.0 - 2.25\mathrm{TM}_{\mathrm{pred}}) \times S_{SS} + 0.375 \times (\mathrm{SeqID} - 20\%)^2 \qquad (2.7)$$

With the sorted HHsearch result, single model is generated for each high-ranked template. For a template having higher H_{fr} score than 0.9 H_{fr} for the top-ranked template candidate, single model is generated with the sequence alignment from HHsearch by using GalaxyCassiopeia structure optimization step I (See **section 3.2.4**). With the generated single models, GalaxyCassiopeia energies are evaluated and they are re-rescored by using the score **Eq. 2.8**.

$$H^*_{fr} = H_{fr} - 0.25E_{Elec} - 0.20E_{HBond} \qquad (2.8)$$

where E_{Elec} and E_{HBond} denotes electrostatics energy (Haberthur and Caflisch, 2008; MacKerell *et al*., 1998) and hydrogen bond energy (Kortemme *et al*., 2003), respectively. For RaptorX, it tends to give better rankings than HHsearch, only energy-based re-scoring scheme is applied to re-rank template candidates. Single model generation and GalaxyCassiopeia energy calculations are performed with the same method for RaptorX high-ranked template candidates, and template candidates are re-ranked by using the score **Eq. 2.9**.

$$R^*_{fr} = R_{fr} - 0.10E_{Elec} - 0.05E_{HBond} \qquad (2.9)$$

where R_{fr} is raw score for each template from RaptorX.

With the re-ranked template candidates for the both fold recognition method, templates are selected for the multiple sequence alignment with the method outlined in **Figure 2.4**. The top-ranked template is selected from the top-ranked template candidate from HHsearch if the predicted TM-score is higher than 0.8, otherwise it is selected from that of RaptorX. To construct template candidate pools, TM-scores are calculated between template candidates from the both fold recognition methods and the top-ranked template by using their experimental structures (TM_{PDB}) and their generated single models (TM_{model}). If a template candidate has higher TM_{PDB} than $0.3TM_{pred}+0.4$ or higher TM_{model} than $0.3TM_{pred}+0.2$, it is added to the pool up to 10 template candidates. Among these template candidates in the pool, template candidates meet tighter condition ($TM_{PDB} > 0.3TM_{pred}+0.5$ or $TM_{model} > 0.3TM_{pred}+0.3$) are designate as reference template candidates up to 5 template candidates. To identify structurally diverse residues, residue-wise deviations from the reference template candidates are calculated for template candidates in the pool. The average deviations for all residues in reference template candidates are designated as a deviation cutoff, and residues having higher deviations than the deviation cutoff are selected as structurally diverse residues.

$$Z^{(i)} = \sum_{j \notin \{\text{structurally diverse residues}\}} \frac{d_j^{(i)} - \overline{d_j^{ref}}}{\max(0.1, s_j^{ref})} \tag{2.10}$$

For *i*-th template candidate in the pool, the template selection score (**Eq. 2.10**) is evaluated where $\overline{d_j^{ref}}$ and s_j^{ref} denotes average and standard deviation for *j*-th residue deviation, respectively. Finally, according to the template selection score, templates having higher template selection score than the cutoff are selected as

templates. The cutoff value is the bigger value between -0.3 or the highest template selection score -0.3.

For selected templates, multiple sequence alignment is generated for the selected templates and the target sequence. In the previous GalaxyTBM (Ko *et al*., 2012), multiple sequence alignment is obtained by using Promals3D (Pei *et al*., 2008). Similarly, in the current GalaxyTBM, Promals3D is used to generate multiple sequence alignment, but two types of additional restraints are applied to build multiple sequence alignment. The first type of restraint is pairwise sequence alignments generated by fold recognition methods because pairwise sequence alignment generated by fold recognition methods usually shows better performance than that of Promals3D (data not shown). The second type of restraint is structure alignments between templates. Multiple sequence alignment generated by Promals3D without the second type of restraint sometimes gives conflicts between multiple templates, and this multiple sequence alignment generates weird protein tertiary structure models.

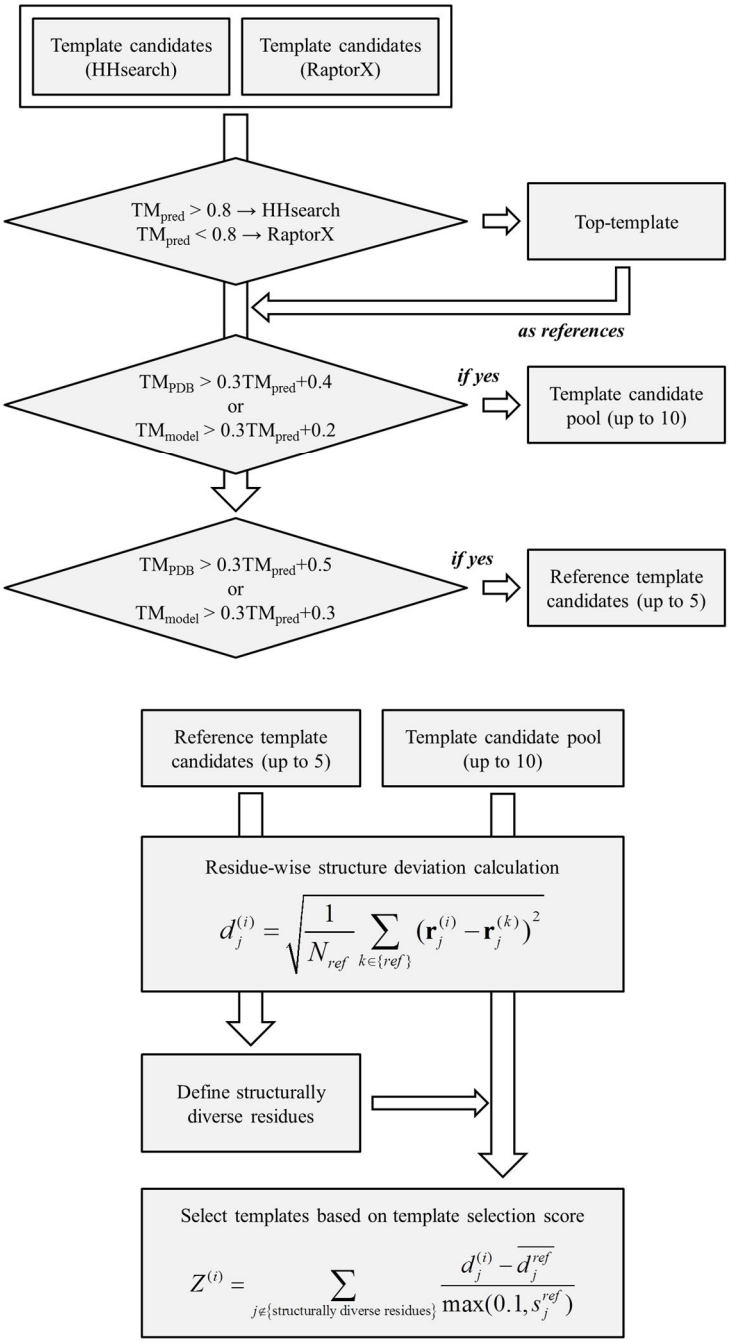

Figure 2.4. Flow chart for template selection method in GalaxyTBM.

2.2.3. Protein chain building using GalaxyCassiopeia and GalaxyLoop

From the multiple sequence alignment obtained from the previous section (**section 2.2.2**), protein tertiary structure models are generated by using GalaxyCassiopeia and GalaxyLoop (Ko *et al*., 2012; Park *et al*., 2011; Park *et al*., 2014; Park and Seok, 2012). GalaxyCassiopeia focuses on global structure generation and optimization, while GalaxyLoop focuses on local structure refinement such as loop and terminal. In the previous GalaxyTBM (Ko *et al*., 2012) (used in CASP9), tertiary structure models were generated by using MODELLERCSA and the generated structures were optimized by using GalaxyLoop-PS1. In the current GalaxyTBM used in CASP11, they are generated by using GalaxyCassiopeia. GalaxyCassiopeia performs two steps of optimization. In the current GalaxyTBM, tertiary structure models generated by using GalaxyCassiopeia optimization step I, and unreliable local regions (ULRs) are assigned for further refinements with the generated structure models. Unreliable local regions (ULRs) are refined in two steps; long loops (loop length > 15 residues) and terminals are refined first and short loops (loop length ≤ 15 residues) are refined later. After ULR refinements, global structure refinement is performed by using GalaxyCassiopeia optimization step II. Detailed explanations are described in the following paragraph.

From the multiple sequence alignment, protein tertiary structure models are generated by using GalaxyCassiopeia. GalaxyCassiopeia consists of five steps: initial model generation, template-driven restraint generation, two steps of global structure optimization, and unreliable side-chain assignment. Initial tertiary structure model is generated from the multiple sequence alignment and the selected templates by threading sequences on the templates (See **section 3.2.2**). Template-driven restraints are also generated from the multiple sequence alignment and the selected template structures (See **section 3.2.3**). There are five types of template

restraints: two types for backbone distance restraints, two types for side-chain distance restraints, and one type for backbone dihedral (Ramachandran) angle restraints. Distances and dihedral angles in the template structures are transferred with consideration of their importance as restraints. The importance estimation method was trained as described in **section 3.2.3**. The initial tertiary structure model is optimized using GalaxyCassiopeia optimization step I with the generated template-driven restraints and molecular mechanics energy (See **section 3.2.4**). In this step, 48 structure models are generated through 48 different trajectories of short molecular dynamics simulation with variable target function method (VTFM) for the energy function **Eq. 2.11**. Because the structure optimization is performed with different random numbers for each trajectory, this step generates diverse structures especially on regions where the template information are deficient. Among the generated structures, the minimum energy structure is selected for the next step.

$$E_{\text{opt1}} = E_{\text{rsr}} + E_{\text{MM}} + E_{\text{vdW}} \tag{2.11}$$

Before GalaxyCassiopeia optimization step II, unreliable local regions (ULRs) are predicted by using three features with the procedure described in **Figure 2.5**. For the first feature, residue-wise structure deviations from the lowest energy model to the generated models are used (**Eq. 2.12**). Residues with high structure deviations are came from deficient template information. These regions have little template information, they are hard to converge. Different from the previous GalaxyTBM ULR prediction, I only included only 30 low energy models out of 48 models because high energy models are not optimized enough, and they are resulted in high deviations. Backbone dihedral angle deviation from the fragment library (Gront *et al*., 2011) is used as the second feature (**Eq. 2.13**). The fragment score, S^{frag}, is smoothed by using triangular smoothing scheme with window size 9.

Uncommon backbone dihedral angle compositions are regarded as errors during the structure optimizations. The final feature is obtained from the multiple sequence alignment by evaluating local sequence alignment qualities (**Eq. 2.14**). The position specific scoring matrix (PSSM) is generated by using PSI-BLAST (Altschul *et al*., 1990) with options "-j 2 -h 0.001" on nr70 database. The MSA score, S^{MSA}, is smoothed by using triangular smoothing scheme with window size 11. Unreliable sequence alignment information is included into the ULR prediction with this feature.

$$S_i^{RMSF} = \sqrt{\frac{1}{30}\sum_{k=1}^{30}(\mathbf{r}_i^{(k)} - \mathbf{r}_i^{(0)})^2} \tag{2.12}$$

$$S_i^{frag} = \frac{1}{30}\sum_{k=1}^{30}\min\left(\sum_{j=-4}^{4}\left(\left|\varphi_{i+j} - \varphi_{i+j}^{frag}\right| + \left|\psi_{i+j} - \psi_{i+j}^{frag}\right|\right)\right)_{fraglib.} \tag{2.13}$$

$$S_i^{MSA} = \frac{1}{N_{templ}}\sum_{j=1}^{N_{templ}}\mathrm{PSSM}_i(aa_i^{(j)}) \tag{2.14}$$

With those three features, unreliable local regions are predicted as follows. ULR prediction score (**Eq. 2.15**) is calculated and it is standardized by using **Eq. 2.16**.

$$S_i^{ULR} = S_i^{RMSF} + 0.42S_i^{frag} - 0.9S_i^{MSA} \tag{2.15}$$

$$Z_i^{ULR} = \frac{S_i^{ULR} - \max\left(\overline{S^{ULR}},\ 2.8\right)}{\min\left(\sigma_{S^{ULR}},\ 1.5\right)} \tag{2.16}$$

Residues with higher standardized ULR prediction score (Z^{ULR}) than 2.0 are selected as unreliable residues. To remove fragmented ULR regions, two ULR regions with less than 2 residue separations are connected into single ULR regions.

ULR region boundaries are extended toward both sides until the standardized ULR prediction score is higher than 1.5. Among these ULR regions, up to 5 ULR regions are selected to be refined based on the average standardized ULR prediction score. In addition, per-residue error estimation is performed with ULR prediction score (S^{ULR}) by using **Eq. 2.17**.

$$Error_i = \exp\left[0.1S_i^{ULR} + 0.8\right] - 1 \tag{2.17}$$

The ULR prediction procedure and all the parameters are trained to our benchmark test set to predict true ULRs defined by *Park, et al.* (Park *et al.*, 2011).

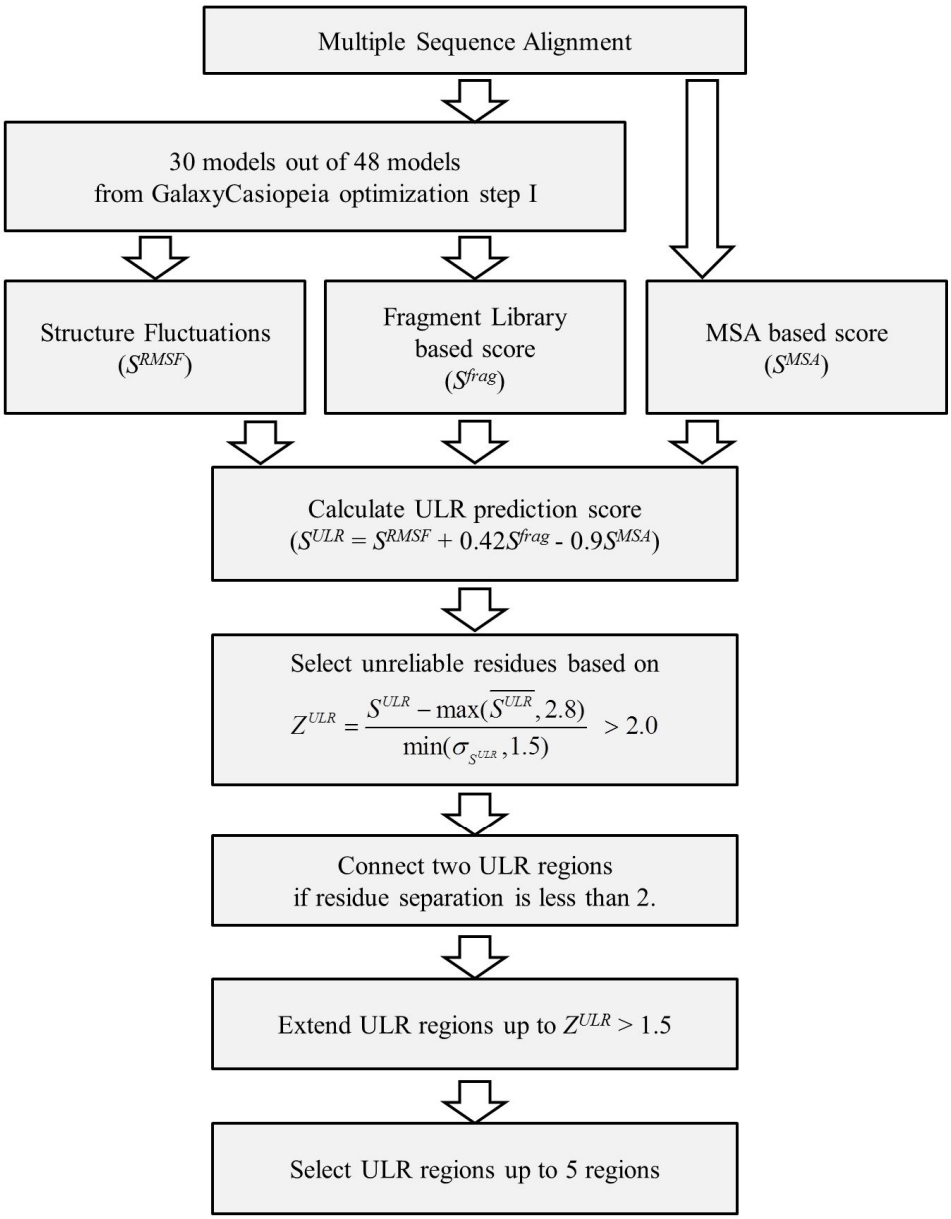

Figure 2.5. Flow chart for unreliable local region (ULR) prediction method.

The predicted unreliable local regions (ULRs) are refined by applying two steps of local structure refinement method. ULRs for terminals and long loops (loop length > 15 residues) are refined by using FALC (fragment assembly and analytical loop closure) (Ko *et al*., 2011; Lee *et al*., 2010) and short molecular dynamics relaxations. For these types of ULR, 2,000 conformations are generated by using FALC. If there are more than two ULRs to be modeled, they are modeled simultaneously. For each conformation generated by FALC, 1-ps of molecular dynamics relaxation and followed local energy minimization are performed. The energy function for these processes composed of weighted summation of physicochemical energy functions and statistical potentials (**Eq. 2.18**).

$$\begin{aligned} E_{\mathrm{ULR}} &= E_{\mathrm{MM}} + w_{\mathrm{vdw}}E_{\mathrm{vdw}} + w_{\mathrm{EEF1}}E_{\mathrm{EEF1}} \\ &\quad + w_{\mathrm{dDFIRE}}E_{\mathrm{dDFIRE}} + w_{\mathrm{HBond}}E_{\mathrm{HBond}} + w_{\mathrm{Rotamer}}E_{\mathrm{Rotamer}} + w_{\mathrm{Rama}}E_{\mathrm{Rama}} \end{aligned} \tag{2.18}$$

Physicochemical energy functions are constructed to consider molecular geometry, van der Waals interactions (MacKerell *et al*., 1998), and solvation energy described by using EEF1 (Lazaridis and Karplus, 1999). Statistical scoring functions are consisted of atomic pairwise interactions by using dDFIRE (Yang and Zhou, 2008), Rosetta hydrogen bond terms (Kortemme *et al*., 2003), and rotamer preference (Canutescu *et al*., 2003). The relative energy weights are determined and iteratively optimized manually by refining ULRs defined on CASP9 TBM targets (Mariani *et al*., 2011). For the left ULRs, modified GalaxyLoop method is applied to refine short loop regions. By using FALC (Ko *et al*., 2011; Lee *et al*., 2010), 2,000 conformations are generated for these regions, and they are clustered into 45 clusters using k-means clustering method. In addition to these loop structures, 5 loop structures are taken from the 5 lowest energy structures generated by GalaxyCassiopeia optimization step I. Loop conformations are optimized by using conformational space annealing (CSA) with those 50 loop structures as the initial

bank. The same energy function (**Eq. 2.18**) is used for the optimization.

After ULR regions are reconstructed, the global structure is refined by using GalaxyCassiopeia optimization step II. This step is highly similar to GalaxyRefine (Heo *et al*., 2013); the only different feature is this step uses template-driven restraints while GalaxyRefine uses restraints from the initial structure. Template-driven restraints are updated for the reconstructed regions with biasing restraints toward the reconstructed structures. To optimize the structure, repetitive side-chain perturbations and followed short molecular dynamics relaxation are performed. During this optimization step, the global structure is refined driven by side chain re-packing. A hybrid energy function is used for the optimization (**Eq. 2.19**).

$$\begin{aligned} E_{\text{opt2}} = {} & E_{\text{MM}} + w_{\text{vdw}}E_{\text{vdw}} + w_{\text{Coulomb}}E_{\text{Coulomb}} + w_{\text{FACTS}}E_{\text{FACTS}} \\ & + w_{\text{dDFIRE}}E_{\text{dDFIRE}} + w_{\text{HBond}}E_{\text{HBond}} + w_{\text{Rotamer}}E_{\text{Rotamer}} + w_{\text{Rama}}E_{\text{Rama}} \\ & + w_{\text{rsr}}E_{\text{rsr}} \end{aligned} \quad (2.19)$$

It is composed of weighted summation of template-driven restraints, physicochemical energy functions and statistical potential functions. For the physicochemical energy functions, molecular mechanics force field, van der Waals interactions (MacKerell *et al*., 1998), electrostatic interactions such as Coulomb interaction, solvation enthalpy and de-solvation entropy are considered (Haberthur and Caflisch, 2008). For the statistical potential functions, pairwise interaction preferences described by dDFIRE (Yang and Zhou, 2008), Rosetta hydrogen bond terms (Kortemme *et al*., 2003), side-chain and backbone torsion angle preference are included (Canutescu *et al*., 2003). The relative energy weights are determined manually to optimize the step I models well with CASP9 (Mariani *et al*., 2011) and CASP10 (Huang *et al*., 2014) single domain targets by considering improvements of global accuracy measured by GDT-HA (Zemla, 2003), local accuracy measured by GDC-SC (Keedy *et al*., 2009), lDDT (Mariani *et al*., 2013), hydrogen bond

accuracy, coverage (Keedy *et al*., 2009) and MolProbity (Chen *et al*., 2010). In the current GalaxyTBM, 48 models are generated during this step, and the lowest energy model is designated as the final model for GalaxyTBM. For the detailed explanation on GalaxyCassiopeia optimization step II, see **section 3.2.5**.

2.2.4. Applications of protein structure models

Prediction of interactions between proteins is one of the basic studies on their functions. Many proteins act their roles in the form of multiple domains or oligomer structures. For these proteins, domain orientations or oligomer structures are conserved across homologs, as in template-based modeling approaches. Recently, GalaxyGemini is proposed for oligomer structure prediction using oligomer template information (Lee *et al*., 2013). It searches oligomer templates from the structure database and predicts oligomer state and structures using oligomer template information. In CASP11, I adopted GalaxyGemini for oligomer structure prediction, which is developed by my colleagues.

Prediction of ligand binding sites of proteins is another important application for the protein structure prediction. It can be invaluable information for further studies such as protein functional studies and drug design (Campbell *et al*., 2003; Kinoshita and Nakamura, 2003). Recently, there have been progresses in ligand binding site prediction on the predicted protein structures using protein-ligand complex template information (Gallo Cassarino *et al*., 2014; Lopez *et al*., 2009; Schmidt *et al*., 2011; Wass and Sternberg, 2009). With this approach, it detects proteins complexed with ligand molecule, and ligand binding site is predicted using those protein-ligand complex structures. GalaxySite is a kind of ligand-binding site prediction method using the template-based approach (Heo *et al*., 2014).

Differently from the other template-based methods, it performs ligand docking simulations using not only template-driven information, but also physicochemical docking scoring functions to predict ligand binding structures with atomic details. In CASP11, I adopted GalaxySite for predicting ligand binding sites; it was first developed by myself and improved by my colleagues.

2.2.5. Test sets

For the preparation for CASP11 experiment, two different sets of protein structure prediction targets are used. To compare the performance for each step, these targets are modified to suitable for each step. The first test set is constructed by using CASP 9 (Mariani *et al*., 2011) and 10 (Huang *et al*., 2014) tertiary structure (TS) prediction targets, which is composed of 230 targets. In addition to the first set, a set of protein structures are constructed as the second test set. The structures are gathered after CASP10 experiment and before Jan. 3rd, 2013. And these structures are clustered by using PISCES server (Wang and Dunbrack, 2003) with 40% sequence identity, R-free < 0.25, and resolution < 3.0Å cutoff. I excluded structure having higher symmetry than 8-mers. Among these structures, structures having maximum sequence identity less than 40% are selected for the benchmark set. Both test sets are used as the benchmark test set for preparation for CASP11 experiments. Finally, CASP11 tertiary structure (TS) prediction targets are used as purely blind protein structure prediction targets, which are composed of 259 targets.

2.3. Result and Discussions

2.3.1. Modeling unit detection in GalaxyTBM

Performance of the modeling unit detection method is benchmarked and compared with the other existing domain prediction methods. For the comparison, DOMpro (Cheng *et al*., 2006), first network and second network PPRODO (Sim *et al*., 2005), and ThreaDom (Xue *et al*., 2013) are used. DOMpro and PPRODO are neural network based methods, while ThreaDom is similar to GalaxyDom in that it uses sequence alignments with the available protein structures. DOMpro uses sequence profiles, predicted secondary structure, and predicted relative solvent accessibility, and PPRODO uses sequence profiles (first network model) and predicted secondary structure as well as sequence profiles (second network model). Though ThreaDom is similar to GalaxyDom, it uses multiple sequence alignments generated by meta fold recognition method, LOMETS (Wu and Zhang, 2007). The major difference between GalaxyDom and the other domain prediction method is that GalaxDom was optimized to reproduce CASP assessment units (AUs) of CASP targets (118 CASP9 targets) (Mariani *et al*., 2011), while the other methods were trained to reproduce biological domain assignments (Murzin *et al*., 1995; Sillitoe *et al*., 2015). As noted, CASP AUs are set according to all information available, including results from domain parsers, available templates, and performance of participating methods (Kinch *et al*., 2011b; Taylor *et al*., 2014). Therefore, the comparisons presented here are meant to measure how well the methods for detecting biological domains perform in structure prediction.

Table 2.1 summarizes the accuracy of domain prediction by different methods for 187 targets from CASP10 and CASP11. In this table, the number of targets is listed for which a given method finds the correct or incorrect number of CASP AUs. For targets with correct number of units, we subdivide results into whether the correct boundaries were set or not. Domain boundaries are considered correct if they are within 10 residues of CASP AU boundaries (Tai *et al*., 2005). For targets with incorrect number of units, it is subdivided into over-split into too many

domains, or under-split into too few. Note that the test set contains 54 multi-unit targets, results for which are indicated in parentheses.

GalaxyDom predicted the correct number of units for 142 targets with more precise boundary predictions. In contrast, both versions of PPRODO tended to over-split targets, and locate boundaries incorrectly. ThreaDom also tended to over-split or under-split targets compared to GalaxyDom, but predicted boundaries precisely when it found the correct number of units, failing to locate the correct boundaries in only four targets. Results from DOMpro are the most similar to those from GalaxyDom, although the method tended to over-split more, and set less precise boundaries. Overall, DOMpro, PPRODO, and ThreaDom show a tendency to over-split, as expected, because these approaches do not consider experimentally determined multi-domain structures.

To assess the effect of modeling unit detection on the structure prediction accuracy, fold recognition results after domain assignment are compared. HHsearch (Soding, 2005) is used to search templates for each domain, and the similarity of the top-ranked template to the experimental structures is measured in terms of TM-score (Zhang and Skolnick, 2004) normalized with the length of the CASP AU.

The fold recognition results are summarized in **Table 2.2**. Templates with higher than 0.5 TM-score is considered as correct fold, and the number of found correct fold is compared between domain prediction methods. As expected, with the CASP assessment units (AUs), it is possible to find the most correct folds. The next best results are achieved with GalaxyDom. Therefore, I can conclude that more similar protein templates can be found with accurate modeling unit detections, and this would result in more accurate protein structure predictions.

Table 2.1. Modeling unit detection results on CASP10 and CASP11 TBM targets. Results in parentheses are for 54 multi-unit targets, which are also included in overall counts.

	Correct number of units		Incorrect number of units	
	Correct boundaries	Incorrect boundaries	Over-splitting	Under-splitting
GalaxyDom	**136 (11)**	**6 (6)**	**10 (2)**	**35 (35)**
DOMpro	120 (1)	16 (16)	16 (2)	35 (35)
PPRODO (first network)	86 (5)	23 (23)	52 (0)	26 (26)
PPRODO (second network)	48 (8)	28 (28)	93 (0)	18 (18)
ThreaDom	112 (5)	4 (4)	29 (3)	42 (42)

Table 2.2. Effect of domain prediction on template selection.

	Cases in which the TM-score between selected template and experimental structure > 0.5	
	out of 232 CASP AUs from all targets	out of 105 CASP AUs from multi-unit targets
CASP AU	*149*	*54*
GalaxyDom	**146**	**52**
DOMpro	139	45
PPRODO (first network)	131	46
PPRODO (second network)	123	50
ThreaDom	128	45
Without domain splitting	140	44

2.3.2. Fold recognition methods in GalaxyTBM

Fold recognition is the most important part of the template-based modeling approach for protein structure prediction. With a single fold recognition method, it sometimes failed to find proper templates with correct fold. And different fold recognition methods are trained to targeting on different difficulties of fold recognition problems. Many of the protein structure prediction methods are using several fold recognition methods due to these reasons. For the purpose, I tested three fold recognition methods as listed in **section 2.2.2** on my protein structure benchmark set to adopt suitable fold recognition methods, and the result is shown in **Figure 2.2**. HHsearch (Soding, 2005) showed better performance on easy TBM targets (TMscore > 0.8) than the other methods. RaptorX (Ma *et al*., 2014) showed better performance on medium-to-hard TBM targets (TMscore < 0.8) than the other methods. SparksX (Yang *et al*., 2011) showed poor performance than the other methods. As a result, I only adopted HHsearch for easy targets and RaptorX for hard targets.

Since I adopted HHsearch for easy targets and RaptorX for hard targets, they should be optimized for their purpose. I introduced a new re-ranking scheme for HHsearch, and it is tested on the GalaxyTBM benchmark set. The results for the comparison between before and after re-ranking scheme, and comparison to the previous re-ranking scheme which was targeting on overall ranges of target difficulties are shown in **Figure 2.6**. For overall ranges of target difficulties, average TMscore for found or re-ranked templates to the target structure was improved from 72.08 to 73.73. With the previous re-ranking method (S_{fr} score) (Ko *et al*., 2012) it was only improved to 73.13. For easy targets, the differences became bigger. With the new re-ranking scheme, it was improved from 83.14 to 84.94, while it was only to 83.94 with the previous scheme. HHsearch or the

previous re-ranking scheme for HHsearch are targeting on overall ranges of target difficulties, however the new re-ranking scheme is targeting only on easy TBM targets. By considering sequence identity explicitly to the re-ranking scheme, many of easy TBM targets were possible to find better templates.

In addition to fold recognition result-based re-ranking scheme, energy-based re-ranking methods are introduced for both fold recognition methods, and the results are described in **Figure 2.7**. Energy-based re-ranking method was effective for HHsearch, it selected better templates from 73.73 to 74.45. For RaptorX, it was only improved from 73.51 to 73.93. The improvement for RaptorX is relatively smaller than that for HHsearch because RaptorX showed better rankings in the original results (data not shown).

With those re-ranked template candidates, a new multiple template selection method is introduced, and the result is compared as shown in **Figure 2.8**. For each target, protein tertiary structure models are built with single and multiple templates, and their global structure accuracy measured by TMscore is compared to the target protein structure. The overall structure quality was improved from 80.13 to 80.98. The new multiple template selection method selects multiple templates having similar structures on the protein core regions and various structures on the unreliable regions. With this approach it was possible to maintain structure qualities on protein core regions, and improves structure qualities on the unreliable regions.

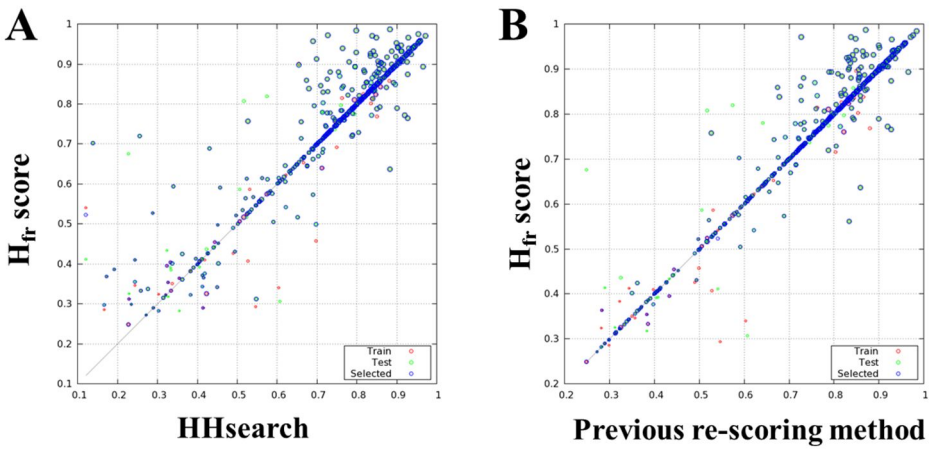

Figure 2.6. Template re-ranking method comparison for HHsearch. (A) Comparison between before and after the re-ranking step. (B) Comparison between the previous and current re-ranking method.

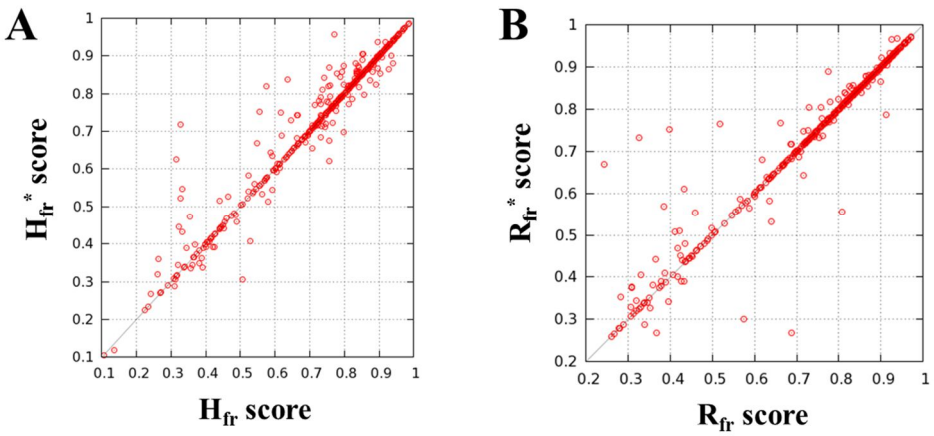

Figure 2.7. Fold recognition results before and after energy-based re-ranking schemes. Re-ranking results for (A) HHsearch and (B) RaptorX.

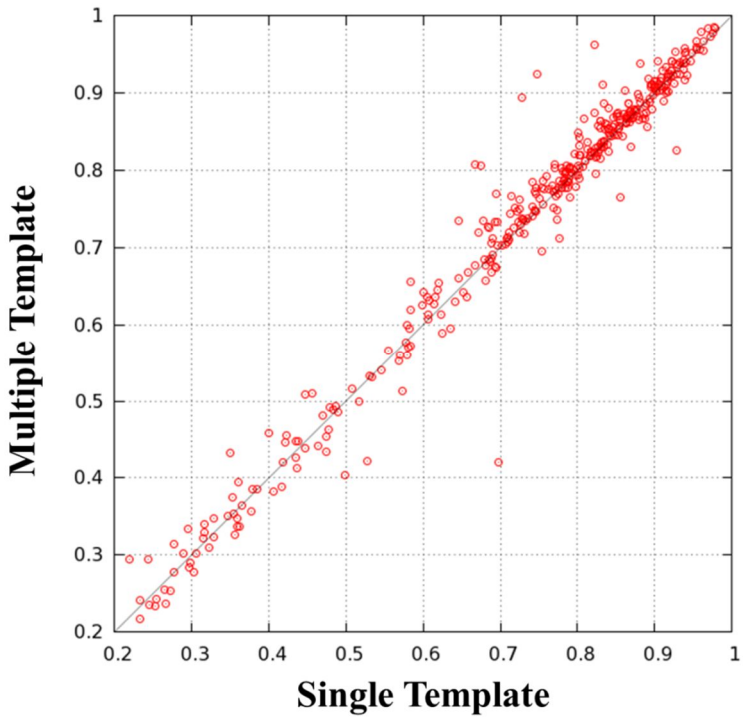

Figure 2.8. Model quality comparison between single and multiple templates.

2.3.3. Building protein tertiary structure models

The overall protein tertiary structure model building results by using GalaxyTBM are summarized in **Table 2.3**. Protein tertiary structure building performance is compared to the previous GalaxyTBM method (Ko *et al*., 2012) used in CASP10 with 44 CASP10 TBM targets, and average values in GDT-TS (a global structure accuracy measure) and GDC-SC (a local structure accuracy measure) are shown in the table. When the models from GalaxyCassiopeia optimization step I, the new method is superior to the previous GalaxyTBM method. There are two major reasons for these improvements: improvements in fold recognition methods (methods and related results are described in **section 2.2.2**, **2.3.2**, and **2.3.3**.) and improvements in GalaxyCassiopeia energy functions by re-optimization of energy parameters (methods and related results are not shown here.).

The detailed performance comparisons for GalaxyCassiopeia are described in **section 3.3**. In that section, protein tertiary structure chain building performances are compared to the state-of-the-art method, MODELLER, on CASP11 TBM blind test set and GalaxyTBM benchmark set. In sum, GalaxyCassiopeia showed slightly better accuracy in global structure accuracy measures, pretty much better accuracy in local structure accuracy and physicochemical correctness measures.

From the models for GalaxyCassiopeia optimization step I, they are locally refined on several unreliable local regions (ULRs). In the previous GalaxyTBM method, long ULRs (ULR length > 20 residues) are refined first, global structure optimization is performed next by using GalaxyCassiopeia optimization step II, and refinements on short ULRs (ULR length ≤ 20 residues) followed. Different from the previous GalaxyTBM, long ULRs (termini and loops with more than 15 residues) are refined first with simple refinement method by using FALC and short molecular dynamics relaxations. Short ULRs (only including short loops with less

than 15 residues) are refined after long ULR refinement by using GalaxyLoop, and global structure optimization is performed for the last step. The overall performance for the refinement processes are enhanced by changing the order of refinement processes. In addition to the changes, improvements in both local structure refinement methods contributed also for the improvement. The detailed analyses on both local structure refinement methods are followed.

Performances for both local structure refinement methods are described in **Figure 2.9**. ULR RMSDs are compared for before and after ULR refinements are shown in the figure where the circle sizes are proportional to the ULR environment RMSD (Park *et al*., 2014). The median ULR RMSD for this refinement is changed from 8.86 Å to 8.49 Å. ULR refinement method for long loops and termini showed relatively big changes than that for short loops. And it was also less sensitive with the ULR environment. The refinement method used for long ULRs performs FALC which can cover broader conformational spaces by using fragment assembly without energy functions. And the method perform short molecular dynamics relaxations rather than global optimization methods, this makes it less sensitive to errors in environment structures.

ULR refinement method for short loops performs global optimizations by using conformational space annealing (CSA). For this refinement step, the median ULR RMSD is changed from 7.87 Å to 7.64 Å. Due to it performs global structure optimizations of the physicochemical energy function, it is important to describe the energetics for the optimization system. However, refinement of loops with unreliable their environment is difficult because of their uncertain energetics. We have developed a new loop modeling method with less sensitive to their environment errors, but there are intrinsic limitations. For short loops having relatively accurate environments are succeeded in refinement where the most of

these ULRs having less than 10 Å, while the others are failed to be refined.

Global structure optimization is performed after two steps of local structure refinement by using GalaxyCassiopeia optimization step II. The global, local structure accuracy and physicochemical correctness improvements are describe in **section 3.3**. In addition to the analysis, refinements on unreliable local regions (ULRs) by global structure optimization are analyzed, and the related data are shown in **Figure 2.10**. Though they are already refined by using each local structure refinement method, they are refined a bit more. Especially, for long loops and termini, the median ULR RMSD is improved from 9.33 Å (before local structure refinement), 8.08 Å (after local structure refinement) to 7.40 Å (after global structure optimization). For short loops, no improvement in RMSD measure was observed (from 5.57 Å, 5.39 Å to 5.45 Å). The local structure refinement for long loops and termini generates less packed structures due to their simple sampling method, and they can be re-optimized by generating more packed structures using molecular dynamics relaxations. However, the method for short loops performs extensive global optimizations, they generate well packed structures during the refinement step and it is hard to be optimized with relaxations.

Table 2.3. Overall chain building results on 44 CASP10 TBM targets with both GalaxyTBM methods.

GalaxyTBM in CASP11			GalaxyTBM in CASP10		
Step	GDT-TS	GDC-SC	Step	GDT-TS	GDC-SC
Cassiopeia opt. step I	73.40	28.09	Cassiopeia opt. step I	70.93	27.72
Long ULR	73.82 (+0.42)	28.20 (+0.11)	Long ULR	70.62 (-0.31)	27.63 (-0.09)
Short ULR	74.16 (+0.76)	28.54 (+0.45)	Cassiopeia opt. step II	71.84 (+0.91)	29.82 (+2.10)
Cassiopeia opt. step II	75.14 (+1.74)	31.93 (+3.84)	Short ULR	71.70 (+0.77)	30.01 (+2.29)

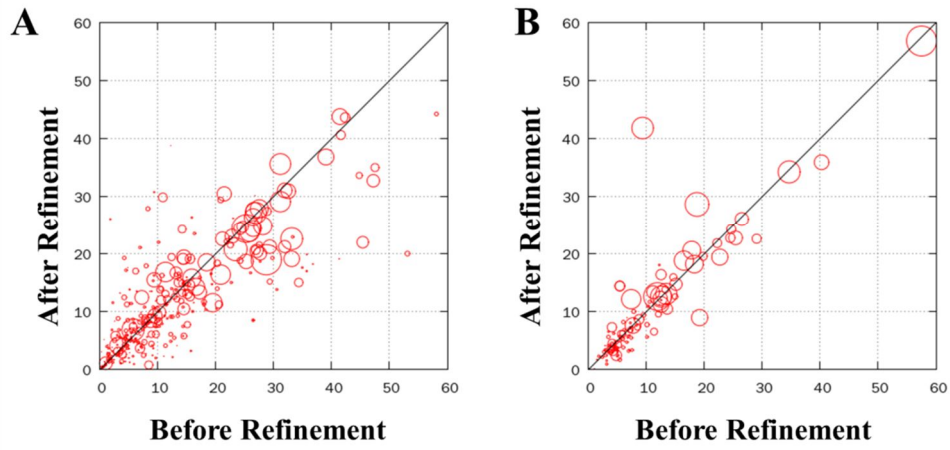

Figure 2.9. ULR refinement results for both steps. (A) Long ULR (termini and long loops) refinement results. (B) Short ULR (short loops) refinement results. Circle sizes are proportional to environment RMSD for each ULR.

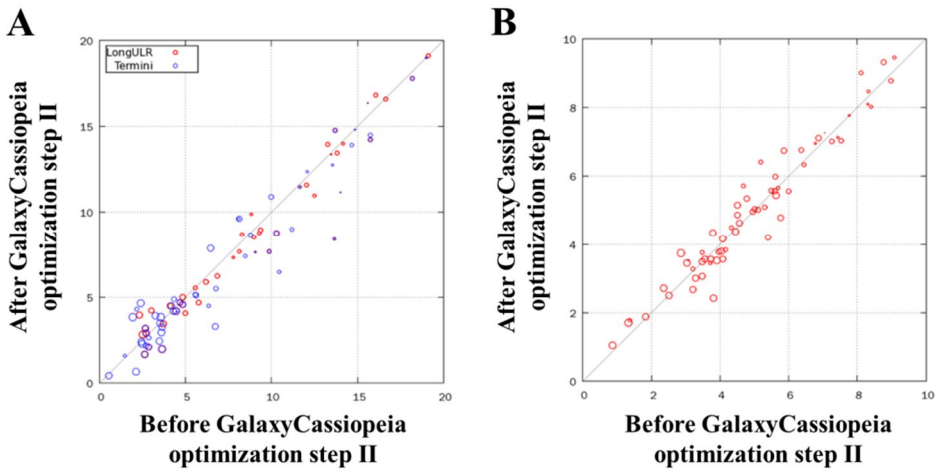

Figure 2.10. GalaxyCassiopeia optimization step II results for refined ULR regions. (A) ULRs with long ULR (termini and long loops) refinement method. Data for long loops and termini are shown in red and blue circles, respectively. (B) ULRs with short ULR (short loops) refinement method.

2.4. Conclusions

A new GalaxyTBM method is developed, which includes modeling unit detections, fold recognitions, sequence alignment, protein tertiary structure model buildings, and applications of the predicted tertiary structure models. A new concept for protein domain, modeling unit, is introduced; it is suitable domain definition for protein structure prediction. Different from the fold recognition method in previous GalaxyTBM, two different fold recognition methods are adopted to cover different ranges of target difficulties. In addition, new re-ranking methods for each fold recognition method are introduced by using structure features and physicochemical energetics. For the protein tertiary structure building processes, structure models are built by using GalaxyCassiopeia, and they are refined with two different local structure refinement method and global structure relaxations. Protein oligomer structure prediction and ligand binding site prediction are also included as fundamental applications of the protein structure prediction. The method is successfully benchmarked on GalaxyTBM benchmark set, and it is also applied on CASP11 TS category, which performs blind structure predictions.

3. GalaxyCassiopeia:
A protein chain building method by using physicochemical energy and template-driven restraint

3.1. Introduction to protein chain building

In template-based modeling (TBM) approach for the protein structure prediction, building protein chain models are one of the key issues (Marti-Renom *et al*., 2000). There have been huge improvements in template detection and sequence alignment methods for the last two decades (Huang *et al*., 2014; Kryshtafovych *et al*., 2014; Ma *et al*., 2014; Soding, 2005). However, there are few studies on how to generate protein tertiary structure models from the sequence alignment and its related template structures. MODELLER was first developed in 1993 by Sali *et al.* (Sali and Blundell, 1993), which uses template-driven restraints for the structure generations. Structural features in template structures are extracted in form of restraints based on conditional probability theory. The template-driven restraints are optimized to generate protein tertiary structure models by applying variable target function method (VTFM).

However, template-driven restraints used in MODELLER are hard to reflect sequence changes from the template to the target sequence. For tertiary structure generation with low sequence identity template structures, the structure is usually different from the template structures: both global structures and local structure packing. Another problem in MODELLER template-driven restraints is that they have lots of local minima and broad energy wells. With this type of restraints, it is hard to optimize structures with computationally cheap optimization methods (Joo

et al., 2009). In addition, this makes it hard to select better structures among the sampled structures with the template-driven restraints.

To tackle these problems, I developed a new protein chain building method named GalaxyCassiopeia. GalaxyCassiopeia uses different types of template-driven restraints. First, different from the restraint for MODELLER, restrains consider the probability of errors. Not only structural variability in the structural features, but also errors in structure features are trained together for GalaxyCassiopeia restraint generation method. Second, the restraint energy landscape is less rugged than that of MODELLER. Similar structural features from the multiple templates are averaged to have narrow and less rugged energy landscape. In addition to the different template-driven restraints, GalaxyCassiopeia is composed of two steps of structure optimization steps. For the first optimization step, it samples global structures only with the template-driven restraints. But for the second optimization step, it samples around the sampled structure with both template-driven restraints and physicochemical energy functions. During this optimization step, it extensively samples side chains to sample well-packed structures with physicochemical energy functions. This sampling methods and physicochemical energy functions enable better local structure packing, and this make is possible to reflect sequence changes from the template structure to the target sequence.

3.2. Methods

3.2.1. Overall method of GalaxyCassiopeia

The GalaxyCassiopeia builds protein structures from a protein sequence, template protein structures, and their multiple sequence alignment. An initial protein

structure is built based on the given set of template protein structures by threading sequence on to the template protein structures. It is further optimized through subsequent two steps of relaxation by molecular dynamics simulations. For the first optimization step, template-driven spatial restraints are mainly used for the global structure optimization. For the second optimization step, a hybrid energy function composed of physicochemical energy function, statistical scoring function, and the restraints is used for the local structure optimization and the global structure refinement. Each step is described below in more detail.

3.2.2. Generation of initial models

For a protein sequences, an initial protein structure is built with a selected set of template structures. CHARMM19 topology is used for the overall modeling procedure to reduce the number of atoms with having detail molecular descriptions (MacKerell *et al*., 1998). A template structure having the highest sequence identity with respect to the query sequence is designated as the primary template. Backbone Cartesian coordinates are copied for the sequences aligned to the primary template from the primary template. Side chain Cartesian coordinates are also copied for the same amino acid residues. The other templates are used for the left coordinates information: coordinates for the unaligned sequences, and side chain coordinates for the aligned to the primary template structure, but having different amino acids. To avoid problems from structure alignment, internal coordinates system is used for the other templates coordinates information. Chain cleavages in Cartesian coordinates system are connected using tri-axial loop closure (Coutsias *et al*., 2004) with unaligned residue backbone angles. Missing side chain Cartesian coordinates and both unaligned termini coordinates are built based on the default CHARMM19 topology (MacKerell *et al*., 1998).

3.2.3. Generation of template-driven restraints

Template-driven spatial restraints are generated from the template protein structures. The spatial restraints are similar to the MODELLER restraints (Sali and Blundell, 1993), but there are several differences: types of restraints, functional forms, use of side chain information, and treatment of multiple template information. There are 5 types of restraints are used for GalaxyCassiopeia: 4 distance restraints (between backbone Cα atoms, backbone N and O atom, side-chain and main-chain atom, and side-chain atoms) and 1 backbone dihedral angle restraints.

There are four types of distance restraints are used. Distance restraints between backbone Cα atoms are used for transferring global topology information. Distance restraints between backbone N and O atoms are used for transferring backbone hydrogen information. Both types of backbone atom restraints are expressed in double Gaussian functions (**Eq. 3.1**).

$$f^{BB}_{\text{Cassiopeia,Single}}(d \mid d_0^{(i)}) = -\log\left[\begin{array}{l} w^{(i)} N_0^{(i)} \exp\left(-\left(\dfrac{d-d_0^{(i)}}{s_0^{(i)}}\right)^2\right) \\ +(1-w^{(i)}) N_e^{(i)} \exp\left(-\left(\dfrac{d-(d_0^{(i)}+d_e^{(i)})}{s_e^{(i)}}\right)^2\right) \end{array}\right] \quad (3.1)$$

The first Gaussian function reflects template information, while the second Gaussian function considers possibility on distance errors. Relative weights ($w^{(i)}$), both standard deviations (s_0 and s_e), and distance error correction (d_e) are calculated based on four features (global sequence identity (i), distance from the nearest gap (g), relative solvent accessibility (a), and pairwise distance (d)), and $N_0^{(i)}$, $N_e^{(i)}$

are normalization factors for each Gaussian functions. Distance restraints are evaluated only for $d_e^{(i)}$ < 15.0 Å (restraints between Cα) or 10.0 Å (restraints between N and O), and restraints between adjacent residues are excluded.

To derive parameters for backbone template restraints, a non-redundant structure database having 3,786 protein structures was constructed by using PISCES server (Wang and Dunbrack, 2003) on June 24th, 2011. The non-redundant structure database was constructed with 20% maximum mutual sequence identity, resolution < 2.0 Å, R_{free} < 0.25, and sequence lengths were more than 40 and less than 1,000 residues, while non X-ray structures, structures only with Cα atoms were excluded. For a protein in the database, homolog proteins were searched by using HHsearch (Soding, 2005) with pdb70 HHsearch database released on February 24th, 2011. A protein having the highest HHsearch score was selected as template, and its sequence was aligned to the query protein sequence by using Promals3D (Pei *et al*., 2008). All spatial restraint parameters were determined based on features extracted from those pairwise alignments and their related protein structures. Four features were extracted from this information; global sequence identity and distance from the nearest gap were calculated from the pairwise alignment, and relative solvent accessibility and pairwise distance were obtained from the template structure. In addition, pairwise distance errors were also evaluated.

All the obtained data were categorized into hundreds of bins. For global sequence identity and relative solvent accessibility (Lee and Richards, 1971), they were binned uniformly into 5 bins, respectively; for pairwise distance, it was binned for every 5 Å from 5 Å to 30 Å; for distance from the nearest gap was binned into 6 bins ($0 \le g \le 2$, $2 < g < 4$, $4 \le g \le 5$, $5 < g \le 9$, $9 < g \le 16$, $g > 16$). For every bin, distribution of pairwise distance errors was fitted using **Eq. 3.1** to

obtain parameters ($w^{(i)}$, s_0, s_e, and d_e). And the fitted parameters were fitted using **Eq. 3.2** with four features.

$$\begin{aligned} p(i,a,g,d) = c_0 & \\ & + \sum_{x\in\{i,a,g,d\}} c_x x \\ & + \sum_{x\in\{i,a,g,d\}} \sum_{y\in\{i,a,g,d\}} c_{xy} xy \\ & + \sum_{x\in\{i,a,g,d\}} \sum_{y\in\{i,a,g,d\}} \sum_{z\in\{i,a,g,d\}} c_{xyz} xyz \end{aligned} \tag{3.2}$$

With the **Eq. 3.2**, 35 parameters were obtained for $w^{(i)}$, s_0, s_e, and d_e, respectively.

In addition to the backbone restraints, there are two types of distance restraints related to side chain atoms: between a side chain atom and a backbone atom, and between two side-chain atoms. For the side chain restraints, distances between two atoms in the templates are simply transferred using **Eq. 3.3**.

$$f^{SC}_{\text{Cassiopeia}}(d \mid d_0^{(i)}) = \frac{1}{2}\left(\frac{d - d_0^{(i)}}{s}\right)^2 \tag{3.3}$$

Different parameter s is used for the **Eq. 3.3** whether the aligned residue pairs have identical residues or not, and the parameters were using the same database used for backbone restraints.

In addition to the distance restraints, backbone dihedral angles in the templates are also used as restraints. For each residue, backbone dihedral angles are calculated, and the background Ramachandran map is modified with the calculated dihedral angles by using (**Eq. 3.4**). The background Ramachandran maps are built up for glycine, proline, and the other standard amino acids by using the structure database, and the standard deviations for template dihedral angles ($\sigma_\varphi^{(i)}$, $\sigma_\psi^{(i)}$) are

determined by their secondary structures. This process yields different Ramachandran maps for every residue. Example for this position-specific Ramachandran map is illustrated in **Figure 3.1**.

$$f_{\text{Rama}}^{(i)}(\varphi,\psi \mid \text{aa}^{(i)}) = f_{\text{Rama}}^{0}(\varphi,\psi \mid \text{aa}^{(i)}) \times \sum_{j}^{N_{templ}} \left[\frac{N^{(i)}(\sigma_{\varphi}^{(i)},\sigma_{\psi}^{(i)})}{N_{templ}} \exp\left[-\left(\frac{1-\cos(\varphi-\varphi^{(i)})}{\left(\sigma_{\varphi}^{(i)}\right)^2} + \frac{1-\cos(\psi-\psi^{(i)})}{\left(\sigma_{\psi}^{(i)}\right)^2} \right) \right] \right]$$

(3.4)

For use of multiple templates, a different approach is applied to the backbone restraints to simplify the energy landscape, as illustrated in **Figure 3.2**. In MODELLER, it uses multiple wells to consider all template information by using (**Eq. 3.5**). It is quite simple, but it yields rugged and broad energy landscape. As a result, MODELLER sometimes generate frustrated model because of their weak optimization power (Joo *et al*., 2009). In GalaxyCassiopeia, it uses complicated functional form (**Eq. 3.6**), but it generates much simple and narrow energy landscape to stand with simple optimization method. In this functional form, more important and accurate template information could be emphasized, and the other information would be less weighted. First, template information are grouped by using distances ($d_0^{(i)}$) and their standard deviations ($s_0^{(i)}$). An ungrouped distribution is joined into a group, if there are some distribution overlaps: any distance difference between ungrouped and grouped distribution is less than two mean standard deviations. Then, every grouped distribution is multiplied to reduce energy landscape complexity, and those multiplied distributions are summed up. For the side chain restraints, average distance is used for multiple templates.

$$f_{\text{MODELLER}}(d \mid \{d_0^{(i\}}\}) = -\log\left[\sum_i^{N_{templ}} w^{(i)} N^{(i)} \exp\left(-\left(\frac{d - d_0^{(i)}}{s_0^{(i)}}\right)^2\right)\right] \qquad (3.5)$$

$$f_{\text{Cassiopeia,Multiple}}(d \mid \{d_0^{(i)}\}_1, \{d_0^{(i)}\}_2, ..., \{d_0^{(i)}\}_{N_{group}})$$
$$= -\log\left[\sum_j^{N_{group}} \left[\prod_i^{N_{templ}^{(j)}} f_{\text{Cassiopeia,Single}}(d \mid d^{(i,j)})\right]\right] \qquad (3.6)$$

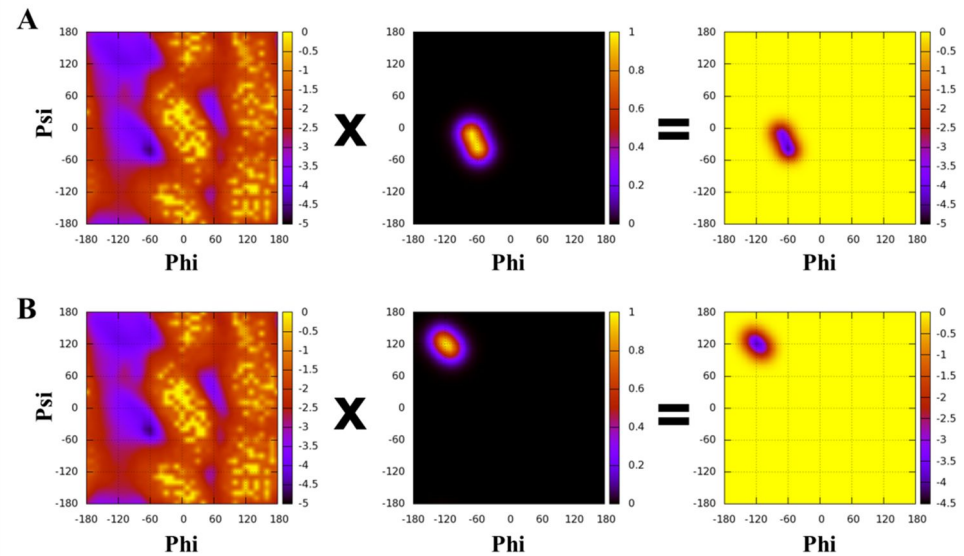

Figure 3.1. Examples of position-specific Ramachandran map. Ramachandran map for standard amino acids are shown in the left column. Ramachandran angles taken from templates are shown in the middle, and position-specific Ramachandran map is shown in the right column. Ramachandran angles from templates are (A) (-71.4,-9.8) and (-57.5, -40.1), (B) (-125.9, 128.2) and (-109.1, 110.4).

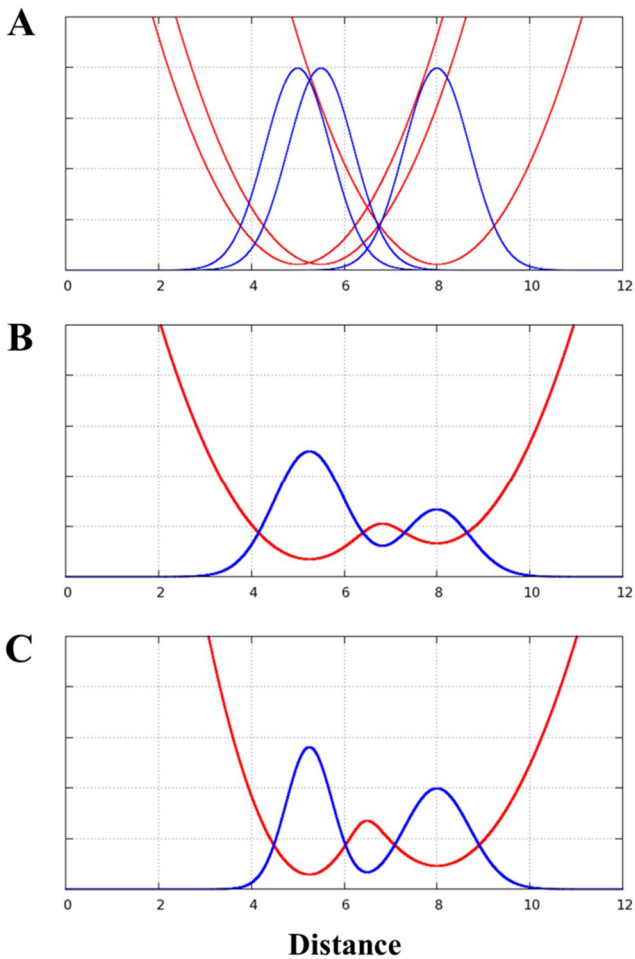

Figure 3.2. Illustrative example for distance restraint treatment from multiple templates. Restraint energy functions and its probability distribution are shown in red and blue, respectively. (A) Three single harmonic restraints from three templates; minimums are located at 5.0, 5.5, and 8.0 and standard deviations for each probability distribution is 0.5. Treatment of multiple template restraint in (B) MODELLER and (C) GalaxyCassiopeia.

3.2.4. Structure optimization step I

For the first optimization step, the GalaxyCassiopeia optimizes protein model structures mainly focusing on their global structures. In this step, a simple hybrid energy function (**Eq. 3.7**) which is composed of template-driven restraints, molecular mechanics energy functions, van der Waals non-bonded energy is used for the optimization.

$$E_{\text{opt1}} = E_{\text{rsr}} + E_{\text{MM}} + E_{\text{vdW}} \tag{3.7}$$

The GalaxyCassiopeia performs short molecular dynamics simulations to optimize the initial structure model for 1,000 steps of integration with 4 fs long time steps. Local energy minimizations are performed before and after the molecular dynamics simulations by using l-BFGSB method. During the molecular dynamics simulations, variable target function method (VTFM) is used for efficient samplings; weights for template-driven restraints are established prior to the other terms when overall weights are increasing from 0.0 to 1.0 during the simulations. Before and after every target function changes, local energy minimizations are performed with changed target functions. For a typical run of this step, it generates 48 different model structures starting from randomly perturbed initial structures with different random number seeds, and the lowest energy model is selected for the next optimization step.

3.2.5. Structure optimization step II

In the second optimization step, the GalaxyCassiopeia performs local and global sampling using a quite similar method of the GalaxyRefine (Heo *et al.*, 2013). Proceeding to the second optimization step, perturbing side chains for extensive

side chain samplings are selected based on three features: side chain fluctuations in the first optimization step models, evolutionary sequence conservations and side chain solvent exposures. Residue-wise side chain root-mean-square fluctuations are calculated with 10 lowest energy structures in the first optimization step models. Average root-mean-square fluctuation less than 5.0 Å are calculated and the bigger value between the average root-mean-square fluctuation and 1.0 Å, and is selected as the cut-off value for selecting highly fluctuating residues. Side chain solvent exposures are calculated by using Naccess (version 2.1) (Lee and Richards, 1971), and if side chain relative solvent exposure for a residue exceeds 20 % the residue is considered as exposed residue. Sequence conservation is measured by self-substitution score in position-specific scoring matrix which is obtained by three iteration of PSI-BLAST (Altschul *et al*., 1990) with 0.001 E-value cutoff using nr70 database. For a residue having lower than 8 self-substitution score is considered as non-conserved residue. Residues which meet two of these three measures are selected as side chain perturbing residues.

The energy function for the second optimization step is composed of weighted summation (**Eq. 3.8**) of template-driven restraints, physicochemical energy functions and statistical scoring functions. Template-driven restraints are modified to remove side chain restraints for the side chain perturbing residues. Physicochemical energy functions are constructed to consider molecular geometry, van der Waals interactions, electrostatic interactions such as Coulomb interaction (MacKerell *et al*., 1998), solvation enthalpy and de-solvation entropy (Haberthur and Caflisch, 2008). Statistical scoring functions are consisted of atomic pairwise interactions by using dDFIRE (Yang and Zhou, 2008), Rosetta hydrogen bond terms (Kortemme *et al*., 2003), rotamer preference (Canutescu *et al*., 2003). The relative energy weights are determined and iteratively optimized manually to refine the first step models well with CASP9 (Mariani *et al*., 2011) and CASP10 (Huang

et al., 2014) single domain targets. Global accuracy measured by GDT-HA (Zemla, 2003), local accuracy measured by GDC-SC (Keedy *et al*., 2009), lDDT (Mariani *et al*., 2013), hydrogen bond accuracy, coverage (Keedy *et al*., 2009) and MolProbity (Chen *et al*., 2010) are considered for the energy weight optimization.

$$\begin{aligned} E_{\text{opt2}} = {} & E_{\text{MM}} + w_{\text{vdw}} E_{\text{vdw}} + w_{\text{Coulomb}} E_{\text{Coulomb}} + w_{\text{FACTS}} E_{\text{FACTS}} \\ & + w_{\text{dDFIRE}} E_{\text{dDFIRE}} + w_{\text{HBond}} E_{\text{HBond}} + w_{\text{Rotamer}} E_{\text{Rotamer}} + w_{\text{Rama}} E_{\text{Rama}} \\ & + w_{\text{rsr}} E_{\text{rsr}} \end{aligned} \quad (3.8)$$

The second optimization step of the GalaxyCassiopeia performs structure relaxation by using molecular dynamics simulations. From the first optimization step model, the side chains for the side chain perturbing residues are optimized first. All those side chains are removed, and new rotamers having the highest rotamer probability without clashes are rebuilt starting from the protein center of the mass (Heo *et al*., 2013). Structure samplings using molecular dynamic relaxation are performed under 300 K and 4 fs is used for time integration intervals. After 300 steps of relaxation with 4 fs time steps (1.2 ps), repetitive side chain perturbations and followed structure relaxation (0.6 ps) are performed. A side chain perturbing residue is randomly selected, and residues within 8 Å are selected to construct a side chain perturbation cluster. Side chains for a selected side chain perturbation cluster are replaced to have a set of randomly selected non-clash rotamers based on their rotamer probability. A perturbation is accepted only if a trial meets Metropolis Monte Carlo criterion. During the overall simulation, 22 times of repetitive side chain perturbations and relaxations are performed. Typically GalaxyCassiopeia generates 48 models with different random number seeds, and the lowest energy model is selected for the final model.

3.3. Results and Discussion

3.3.1. Protein chain building benchmark test result

GalaxyCassiopeia is compared with MODELLER, the state-of-the-art protein chain building program. For each target, templates and their multiple sequence alignment with the target protein sequence are generated by GalaxyTBM method described in **section 2.2.2**. For each method, 48 protein tertiary structure models are generated with this information. Among these targets, targets with wrong fold recognition results are discarded from the analysis after model quality evaluations. The targets having greater than 40.0 GDT-TS score models from either GalaxyCassiopeia or MODELLER are considered that they are succeed in fold recognition. In this section, the minimum MODELLER energy function structure out of 48 models from MODELLER is compared with GalaxyCassiopeia energy minimum structure for each optimization step.

Several structure quality assessment measures are applied for the comparison. To compare global structure qualities, TM-score and RMSD using TM-score, GDT-TS and GDT-HA using LGA (Zemla, 2003) are used. To compare local structure qualities, GDC-SC using LGA and local distance difference test (lDDT) using lDDT (Mariani *et al*., 2013) are used. To measure physicochemical correctness of the models, the MolProbity score (Chen *et al*., 2010) is calculated which is composed of clash score, rotamer outlier ratio, and Ramachandran angle favored region ratio.

The overall structure qualities for CASP11 TBM targets and the benchmark test set are described in **Table 3.1** and **Table 3.2**, respectively. For the global structure accuracy measures, GalaxyCassiopeia showed better performance than MODELLER. Especially, the model qualities had been improved extensively

during the GalaxyCassiopeia optimization steps II. Template-driven information for structure generation is quite useful to generate similar folds, but it is hard to reflect changes in the sequence. Structural features from templates can be different slightly in the target protein structure. During the optimization step II in GalaxyCassiopeia, it uses not only template-driven restraints, but also physicochemical energy functions. These physicochemical energy functions induce correction of the template-driven information for the target protein by finding a new equilibrium between template-driven restraints and physicochemical energy functions. Structures from the GalaxyCassiopeia optimization step I are also better than that from MODELLER. Protein chain buildings using MODELLER with multiple template-driven restraints often suffer from finding minimum structures. Multiple local minima for each atomic pairs are resulted in numerous numbers of local energy minima, and these make it hard to find the global energy minimum. In addition to this problem, similar structural information generates broader energy wells in the energy surface, and this induces broader structure fluctuations. Different from the MODELLER restraints, the number of local energy minima is reduced in template-driven restraints used in GalaxyCassiopeia by grouping template information. And by multiplication of template restraints for grouped information, the energy well is become sharp rather than broad. These two features in GalaxyCassiopeia restraints made it easy to optimize with less expensive sampling method.

GalaxyCassiopeia also showed better performance in local structure accuracy measures. Different from the global structure accuracy comparison, structures from GalaxyCassiopeia optimization step I showed poor accuracy than MODELLER. The number of template-driven restraints for side chains is pretty smaller than MODELLER, because I only included highly reliable information. And unreliable side chains should be found with the GalaxyCassiopeia optimization step I for the

extensive side chain samplings during the step II. With the extensive side chain samplings which consisted of random rotamer perturbations and repacking induce huge improvement in local structure accuracies.

GalaxyCassiopeia generated structures with quite better physicochemical correctness. When physicochemical correctness measured by using the MolProbity score, the score is reduced by more than 1.0 with the reduced number of atomic clashes, rotamer outliers, and the increased ratio of Ramachandran angle favored regions. The number of clashes out of 1,000 atomic pairs is reduced by using more detailed representation of protein structure models and better description of atomic interactions. The number of rotamer outliers is high in structures from GalaxyCassiopeia optimization step I with the same reason for poor local structure qualities, but it showed pretty low for models from step II. In the optimization step II, the scoring function includes rotamer preference, and this effectively reduces rotamer outliers. Ratio of Ramachandran angle favored regions is also improved from MODELLER with better description on Ramachandran angles.

Table 3.1. Blind test result on CASP11 TBM targets. The best among three methods is in bold character.

Measures	MODELLER	GalaxyCassiopeia	
		Opt. step I	Opt. step II
Global structure accuracy			
TM-score	0.7716	0.7743	**0.7786**
GDT-TS	68.65	68.97	**69.57**
GDT-HA	50.68	51.13	**51.94**
RMSD	5.36	5.58	**4.98**
Local structure accuracy			
GDC-SC	27.52	26.85	**31.27**
lDDT	0.6017	0.5938	**0.6167**
Physicochemical correctness			
MolProbity	3.36	3.43	**2.27**
Clash score	124.9	30.5	**20.5**
Rotamer outlier	3.3%	29.6%	**1.5%**
Rama favored	92.3%	94.3%	**94.7%**

Table 3.2. Benchmark test result on benchmark set. The best among three methods is in bold character.

Measures	MODELLER	GalaxyCassiopeia	
		Opt. step I	Opt. step II
Global structure accuracy			
TM-score	0.8020	0.8066	**0.8099**
GDT-TS	74.20	74.70	**75.18**
GDT-HA	57.31	57.97	**58.69**
RMSD	4.61	4.65	**4.41**
Local structure accuracy			
GDC-SC	31.82	30.85	**34.90**
lDDT	0.4894	0.4856	**0.5001**
Physicochemical correctness			
MolProbity	3.14	3.29	**2.10**
Clash score	104.7	29.2	**18.2**
Rotamer outlier	3.2%	28.3%	**1.5%**
Rama favored	94.3%	95.7%	**96.0%**

3.3.2. Comparison of conformational samplings

Similar to the previous analysis in **section 3.3.1**, conformational sampling distributions are analyzed for each method. Structure qualities for each evaluation measure are evaluated for 48 protein tertiary structure models, and the average and standard deviation are compared. The results for CASP11 TBM targets and the benchmark set are summarized in **Table 3.3** and **Table 3.4**.

For global structure accuracies, models from MODELLER and GalaxyCassiopeia optimization step I showed similar average, but standard deviations for MODELLER is bigger than that of GalaxyCassiopeia optimization step I. As mentioned in **section 3.2.3**, the template-driven restraints for GalaxyCassiopeia have narrow distributions than MODELLER. And they have small number of local minima; therefore they are easy to be minimized. As a result, the distributions of global structure quality for GalaxyCassiopeia have narrow distributions. When structure quality distributions are compared with the results for the lowest energy structures, the model selection by MODELLER energy seems almost random. The structure qualities for the lowest energy model are similar to the average values for each of the distribution. In contrast, the lowest energy model qualities for GalaxyCassiopeia optimization step I are better than the average values for each of the distribution. This implies that template-driven restraints for GalaxyCassiopeia are good for picking up better structures among the sampled structures.

When the model quality distributions for GalaxyCassiopeia optimization step I and step II are compared, optimization step II showed even narrow distributions. The optimization step II performs global structure relaxations with local structure samplings to build well-packed protein structures. Therefore the purpose of this step is not a global structure sampling but a structure sampling around the initial

structure. In addition, the temperature used for the simulation is same for the both optimization steps, but the energy function is different; the energy function for the optimization step II has higher energy barriers than step I. As a result, the effective temperature is lower for optimization step II than step I, and this makes the optimization step II less variable.

Table 3.3. Structural quality distributions for 48 generated models on CASP11 TBM targets.

Measures	MODELLER	GalaxyCassiopeia	
		Opt. step I	Opt. step II
Global structure accuracy			
TM-score	0.7721±0.0085	0.7718±0.0058	0.7774±0.0032
GDT-TS	68.76±1.00	68.61±0.72	69.44±0.45
GDT-HA	50.85±1.06	50.61±0.82	51.82±0.51
RMSD	5.33±0.34	5.69±0.18	5.09±0.20
Local structure accuracy			
GDC-SC	27.70±1.45	26.03±1.40	30.76±0.86
lDDT	0.6025±0.0056	0.5891±0.0066	0.6145±0.0030
Physicochemical correctness			
MolProbity	3.38±0.17	3.52±0.09	2.32±0.11
Clash score	121.4±10.7	33.2±3.5	20.8±1.8
Rotamer outlier	3.7±1.5%	31.0±3.5%	1.8±0.7%
Rama favored	92.6±1.5%	93.5±1.1%	94.6±0.6%

Table 3.4. Structural quality distributions for 48 generated models on benchmark set.

Measures	MODELLER	GalaxyCassiopeia	
		Opt. step I	Opt. step II
Global structure accuracy			
TM-score	0.8020±0.0077	0.8054±0.0042	0.8097±0.0032
GDT-TS	74.21±0.86	74.51±0.52	75.17±0.42
GDT-HA	57.31±1.03	57.73±0.61	58.63±0.51
RMSD	4.56±0.29	4.70±0.14	4.44±0.14
Local structure accuracy			
GDC-SC	31.83±1.68	29.67±1.68	34.64±0.96
lDDT	0.4904±0.0043	0.4799±0.0056	0.4989±0.0025
Physicochemical correctness			
MolProbity	3.15±0.18	3.38±0.10	2.16±0.12
Clash score	104.2±9.5	31.1±3.5	18.6±1.7
Rotamer outlier	3.3±1.5%	29.7±3.5%	1.8±0.7%
Rama favored	94.3±1.3±	95.0±1.0%	95.9±0.5%

3.4. Conclusion

GalaxyCassiopeia is a protein chain building method that generates protein tertiary structure models from sequence alignment and its related template protein structures. Different from the state-of-the-art protein chain building program, MODELLER (Sali and Blundell, 1993), GalaxyCassiopeia uses not only template-driven information, but also physicochemical energy functions for both sampling and scoring processes. The method is composed of two steps of structure optimization steps. For the first optimization step, it focuses on sampling global structures only with template-driven restraints. For the second optimization step, it optimizes global structures from the lowest energy model from the first optimization step. In this step, the structure models are refined by side chain repacking. When the method is compared to MODELLER, the model qualities are better than that of MODELLER. Especially, GalaxyCassiopeia generates models with quite better local structure qualities, while global structure qualities are slightly better. This method may be useful for further structure based studies such as molecular docking studies by generating locally more accurate structure models.

4. GalaxySite:
A protein ligand binding site prediction method using molecular docking with homolog information

4.1. Introduction to ligand binding site prediction

Proteins perform their biochemical functions by interacting with other biomolecules such as small ligands, other proteins, or nucleic acids. The detection of binding site on a protein makes it possible to infer the protein function and to provide information on binding pockets crucial for computer-aided drug discovery (Campbell *et al*., 2003; Kinoshita and Nakamura, 2003; Laurie and Jackson, 2006; Sotriffer and Klebe, 2002; Thornton *et al*., 2000) Ligand binding site prediction from protein sequence only, for example, using a protein 'model' structure, has important implications regarding protein function prediction from sequence. Binding site prediction on 'experimental' protein structures is also important for drug design applications. Various evolutionary information-based methods, geometry-based methods, energy-based methods, and combined methods have been reported (Tripathi and Kellogg, 2010).

Recently, methods using the experimental structures of similar proteins have been successful in binding site predictions in CASP (Lopez *et al*., 2009; Lopez *et al*., 2007; Oh *et al*., 2009; Schmidt *et al*., 2011; Wass and Sternberg, 2009). In such methods, binding site information in homologous proteins of known structures is utilized by assuming that similar protein-ligand contacts occur in the target protein. In this paper, we introduce a new method that uses such information in the context of protein-ligand docking. Molecular docking could provide detailed atomic

interactions between protein and ligand as well as the overall binding site if successful, and such information would be extremely useful for the prediction of specific functions and for applications to drug discovery.

Although specific binding of protein and ligand occurs due to favorable physicochemical interactions, binding site prediction based on physical chemistry using molecular docking has rarely been used so far because of several difficulties. Typical docking methods are generally too sensitive to structural details, and docking accuracy may fall off dramatically if the protein structure is not accurate or conformational changes occur upon binding (Bordogna *et al*., 2011; Brylinski and Skolnick, 2008, 2009). However, binding site prediction is of interest when experimental ligand-bound protein structures are not available, and thus prediction has to be performed on 'inaccurate structures', for example, experimental structures containing no ligand, or on model structures. Therefore, it may be expected that molecular docking would not be successful if the unbound protein structure is very different from the bound form or if model structures are not accurate enough. In addition, the docking results usually depend heavily on which ligand is docked, but it is not clear whether it is possible to predict binding ligands accurately.

In this chapter, a successful molecular docking method for ligand binding site prediction called GalaxySite is presented. Binding ligands are predicted using a similarity-based method. The current binding site prediction method is successful even when only chemically similar ligands are predicted. The energy function for docking is designed not to be too sensitive to structural details by adapting a combination of physics-based terms of AutoDock3 docking energy function (Morris *et al*., 1998) and restraint terms derived from homologous protein-ligand complexes of known experimental structures. Tests on a nucleotide set (Kasahara *et*

al., 2010), a bound/unbound set (Huang and Schroeder, 2006), and CASP8 (Lopez *et al*., 2009) and CASP9 (Schmidt *et al*., 2011) targets containing non-metal ligands show that GalaxySite can be successfully applied to binding site predictions of both experimental protein structures (either bound or unbound to ligands) and model protein structures.

4.2. Methods

4.2.1. Overall method of GalaxySite

GalaxySite predicts the ligand binding site of a given protein by protein-ligand docking, as shown schematically in **Figure 4.1** (Heo *et al*., 2014). Three elements required for protein-ligand docking are a protein structure, a ligand, and a docking algorithm. GalaxySite takes a protein structure as input, and the structure may be either an experimental structure (with or without ligand) or a model structure. The ligand is predicted from the complex structures of homologues detected by HHsearch (Soding, 2005). A protein-ligand complex structure is then predicted by a docking algorithm called LigDockCSA (Shin *et al*., 2011). A distinct feature of GalaxySite is that it uses a hybrid docking energy of AutoDock3 energy (Morris *et al*., 1998) and ligand-specific restraints derived from protein-ligand interactions observed in the complex structures of homologues. Each step is described below in more detail.

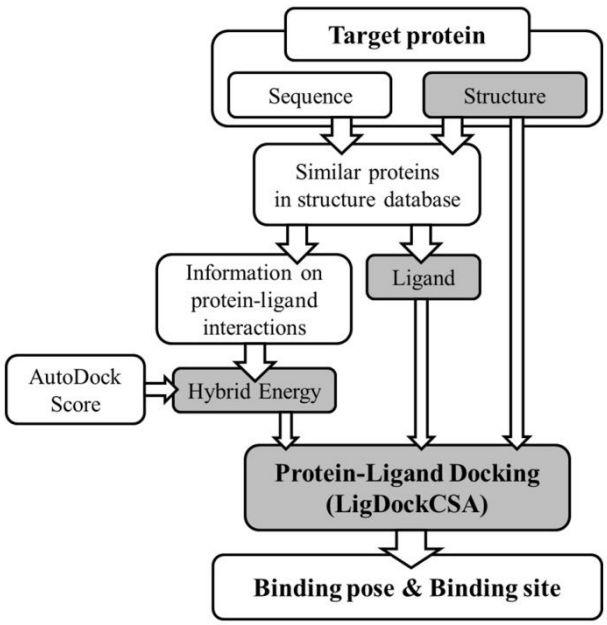

Figure 4.1. Overall procedure of GalaxySite. Ligand binding site of protein is predicted by protein-ligand docking. Protein structure needed for docking is provided as input, and ligand is predicted from similar proteins in structure database. A docking algorithm called LigDockCSA (Shin *et al*., 2011) is used with hybrid energy of AutoDock 3 (Morris *et al*., 1998) energy and template restraints derived from similar proteins. Ligand binding site is predicted from the docking pose.

4.2.2. Selection of template proteins

Proteins similar to the target protein and with known experimental structures are first selected to predict ligands and to derive restraints for docking by local alignment using HHsearch (Soding, 2005). The protein structure database 'pdb70' with maximum mutual sequence identity of 70% is used. Out of 30 proteins with highest re-ranking score of HHsearch results introduced previously by Ko *et al.* (Ko *et al*., 2012), proteins with dissimilar structures from the input structure of the target protein are filtered out using TM-score (Zhang and Skolnick, 2005) as a similarity measure. Proteins with TM-score < 0.5, 0.4, and 0.3 to input structure are filtered out for 'easy', 'medium', and 'hard' targets, respectively. The target difficulty is determined by the maximum TM-score to input structure out of those for the top 30 HHsearch rankers; targets with maximum TM-score > 0.8, > 0.6, and ≤ 0.6 are assigned to be easy, medium, and hard, respectively. Up to this point, homologue structures without bound ligands may be selected.

4.2.3. Ligand selection

Ligands to be docked to the query protein structure are selected from the ligands bound to homologues in the experimental structures. Non-biological ligands (such as sulfate ion, glycerol, and polyethylene glycol added to facilitate crystallization) are filtered out first. Ligands with high positional variation (> 10 Å) among the homologues are also screened out. Positional variation is measured by the average of all pair distances between the ligand center atoms (atoms closest to the ligand geometric center) when all homologue structures containing the same ligand are superimposed. The remaining ligands are ranked by summing the HHsearch score of the homologues containing the same ligand. Protein-ligand docking and binding

site prediction are executed for the top ligand. When alternative binding sites are desired, additional ligands of lower ranking are selected, and the corresponding binding sites are predicted.

4.2.4. Hybrid energy function for protein-ligand docking simulations

The following hybrid energy of AutoDock 3 energy and restraint is used for docking:

$$E = E_{\text{AutoDock}} + 1.1E_{\text{rsr}} \tag{4.1}$$

The AutoDock 3 energy (Morris *et al.*, 1998) is used for E_{AutoDock} except that the maximum energy value for each interacting atom pair is set to 1. This is to tolerate clashes that may be caused by inaccurate protein model structures or ligand-unbound structures.

The restraint term E_{rsr} is derived from the experimental protein-ligand complex structures of similar proteins (selected from the database, as described above) containing the selected ligand. These similar proteins are referred to as 'templates' from now on. Restraint is applied to each ligand atom i, imposing a penalty on r_{ij} (the distance between ligand atom i and protein atom j) deviating from $r_{ij}^{(k)}$ (the corresponding distance in the k-th template) with template-dependent weight factor ω_{ijk} as follows:

$$E_{\text{rsr}}(\{r_{ij}\}) = -\sum_i \ln\left[\sum_j \sum_k \omega_{ijk} \exp\left\{-\left(\frac{r_{ij} - r_{ij}^{(k)}}{d_{jk}}\right)^2\right\}\right] \tag{4.2}$$

where d_{jk} is the position deviation of the Cα atom of the residue to which the j-th

atom belongs in the input structure from that in the *k*-th template when the input and template structures are superimposed. The weight factor is expressed as

$$\omega_{ijk} = (\text{TM-score})_k \, (\text{residue score})_{jk} \, \frac{E_{\text{AutoDock},ij}(r_{ij}^{(k)})}{E_{\text{AutoDock},ij}(r_{\min})} \qquad (4.3)$$

where $(\text{TM-score})_k$ is the structural similarity between the *k*-th template and the input structure measured by TM-score. The second term $(\text{residue score})_{jk}$ is 0 if the corresponding template residue is not of the same amino acid type as the target residue or if $d_{jk} > 2$ Å. Otherwise, the residue score represents side-chain orientation similarity measured by the dot product of the normalized vectors connecting Cα atom and the side-chain centroid (of the residue to which the jth atom belongs) for the input and the *k*-th template structure. The third term accounts for the optimality of the template distance estimated by the ratio of AutoDock 3 van der Waals energy (or AutoDock 3 hydrogen bond energy for hydrogen bonding pairs) at that distance to the optimal energy. The relative weight of the AutoDock 3 energy to the restraint in (**Eq. 4.1**) is set to 1.1 after testing on the CASP7 targets in the function prediction category (Lopez *et al*., 2007).

4.2.5. Protein-ligand docking simulations

A docking algorithm used in LigDockCSA (Shin *et al*., 2011) is employed for protein-ligand docking. LigDockCSA performs global optimization of AutoDock 3 docking energy using the conformational space annealing (CSA) algorithm (Joo *et al*., 2009; Lee *et al*., 1997). In LigDockCSA, the protein structure is fixed at the input structure, and the ligand is considered fully flexible. In the current docking algorithm using CSA, a pool of 100 conformations is first generated by perturbing the initial conformations obtained by copying template ligand poses after

superimposition onto the query protein structure. The conformation pool is evolved by generating trial conformations and updating the pool repeatedly, gradually focusing on narrower areas of lower energy in the conformational space as iteration proceeds. Details on the algorithm can be found elsewhere (Shin *et al*., 2011). Out of the final pool of 100 structures, the energy minimum pose in the largest cluster is selected as the representative binding pose.

4.2.6. Test sets and accuracy measures for binding site prediction

Three test sets are employed to assess the performance of the current method. The 'nucleotide set' from Kasahara *et al.* (Kasahara *et al*., 2010) consisting of 644 nucleotide-derived ligand-protein complexes is used for method validation for various nucleotide-derived ligands, especially to verify the effectiveness of docking compared to simple 'mapping' of the template ligand pose. The 'bound/unbound set' from Huang and Schroeder (Huang and Schroeder, 2006), consisting of 46 pairs of ligand-bound/ligand-free protein structures, is employed to test the applicability of the method to binding site prediction of the protein structure in the ligand unbound form. Finally, 19 CASP8 and CASP9 binding site prediction targets with non-metal ligands are employed to test the performance of the method when model structures are used (Lopez *et al*., 2009; Schmidt *et al*., 2011). If the binding site prediction is successful with the model structures, application of the method to binding site prediction from protein sequence only would be possible.

We evaluate the performance of binding site prediction using three types of measures: minimum distances, centroid distance, and contact residue-based measures, following previous works (Huang and Schroeder, 2006; Kasahara *et al*., 2010; Lopez *et al*., 2009; Schmidt *et al*., 2011). To calculate the first two kinds of

distance measures, the predicted binding pose has to be first superimposed onto the native structure. 'Minimum distance' refers to the closest distance between any ligand atom in the predicted pose and that in the native pose (used for the nucleotide set) or the closest distance between the predicted binding pocket center and any ligand atom in the native structure (used for the bound/unbound set). The second definition is used to compare different methods used to predict binding pocket centers, but not ligand poses. A complementary measure, 'centroid distance' is the distance between the centroids of predicted and native poses. A prediction is considered successful if the distance measures are within given distance criteria (< 3 and < 4 Å for the two minimum distances, respectively, and < 5 Å for centroid distance (Huang and Schroeder, 2006; Kasahara *et al*., 2010).

The measures based on contact residues do not rely on structure superimposition. A residue is considered to be contacting ligand if distance between any atom in the residue and any ligand atom is less than the sum of the van der Waals radii plus 0.5 Å (Lopez *et al*., 2009; Lopez *et al*., 2007; Schmidt *et al*., 2011). Comparing the lists of contact residues in the predicted and the native structures, three measures, accuracy, coverage, and Matthew's correlation coefficient (MCC), are calculated as:

$$\text{Accuracy} = \frac{TP}{TP+FP} \tag{4.4}$$

$$\text{Coverage} = \frac{TP}{TP+FN} \tag{4.5}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{4.6}$$

where *TP*, *TN*, *FP*, and *FN* denote the number of true positive, true negative, false

positive, and false negative predictions for contact residues, respectively. MCC considers both accuracy and coverage, not biased to only one of the two measures.

4.3. Results and Discussion

4.3.1. Method validation test on the nucleotide set

When applied to the nucleotide set, GalaxySite shows an overall success rate of 92% with the minimum distance criterion of 3 Å and 86% with the centroid distance criterion of 5 Å. As shown in Table 1, this result is comparable to or better than an 'ideal mapping' method (success rates of 91% and 86% with the two criteria) that copies the ligand pose from the 'best' template structure after superimposition of the template and the query protein structures. Noting that the best template, with the closest structure to the native complex, is not known in advance in real prediction, this result implies that the docking strategy of GalaxySite is fairly successful both in predicting bound ligand and its docking pose.

Table 4.1. Success rate of binding site prediction on the nucleotide.

Ligand	Success[1] rate of binding site prediction (%)				Success rate of ligand prediction[2] (%)	
	Centroid Distance		Minimum Distance			
	Ideal Mapping	Galaxy-Site	Ideal Mapping	Galaxy-Site	Exact	Similar
AMP	66	69	76	79	24	41
ADP	75	78	86	86	49	74
ANP	88	88	93	93	23	83
ATP	79	79	89	92	22	64
FAD	94	92	95	95	91	91
FMN	95	97	95	97	95	95
GDP	95	98	98	98	90	90
GNP	100	100	100	100	11	100
GTP	83	83	87	83	0	73
NAD	90	87	94	95	75	89
NAP	96	96	98	100	45	93
Total	**86**	**86**	**91**	**92**	**52**	**80**

1) Definition for success depends on the measure; < 5Å for centroid distance and < 3Å for minimum distance.

2) Percentage of the cases in which exact or similar ligands are predicted. Ligands are considered to be similar if they are in the same group listed as follows: (AMP, ADP, ANP, ATP), (GDP, GNP, GTP), (NAD, NAP).

Although the exact ligands are predicted in only 52% of targets (**Table 4.1**), the success rate of predicting 'similar' ligands (for example, AMP, ADP, ANP, and ATP are considered similar to each other) is 80%. This suggests that detecting similar ligands can be enough to identify the correct binding sites. For example, targets containing GNP and GTP show very low accuracy in predicting ligands (11% and 0%, respectively), but binding sites are predicted accurately (with success rate of 100% and 83%, respectively) because templates that contain related nucleotides are detectable. We found that the success rate of binding site prediction is more strongly correlated with the identification of similar ligands rather than with the identification of the exact ligand, with Pearson's correlation coefficients of 0.78 and 0.21, respectively.

The minimum and centroid distances for individual targets are plotted in **Figure 4.2** for a more detailed comparison of GalaxySite and the ideal mapping method. Although the rates of predictions within criterion distances are not improved significantly by GalaxySite docking, the quality of binding site prediction measured by the absolute values of the distance measures is notably enhanced by docking, as can be seen from **Figure 4.2**. GalaxySite shows smaller average minimum distance (1.19 Å) than ideal mapping (1.44 Å) with a P-value of 3.2×10^{-9} and smaller average centroid distance (2.85 Å) than ideal mapping (3.11 Å) with a P-value of 1.2×10^{-5}. Contact residue measures are also improved. Accuracy, coverage, and MCC improvements are 2.4% (from 79.9% for mapping to 82.3% for GalaxySite), 7.7% (from 71.3% to 79.0%), and 0.054 (from 0.739 to 0.793), respectively, with P-values of 7.2×10^{-9}, 4.2×10^{-40}, and 1.3×10^{-31}, respectively. In summary, the GalaxySite docking method is superior to an ideal mapping method for the test set, producing higher-resolution predictions for ligand binding sites.

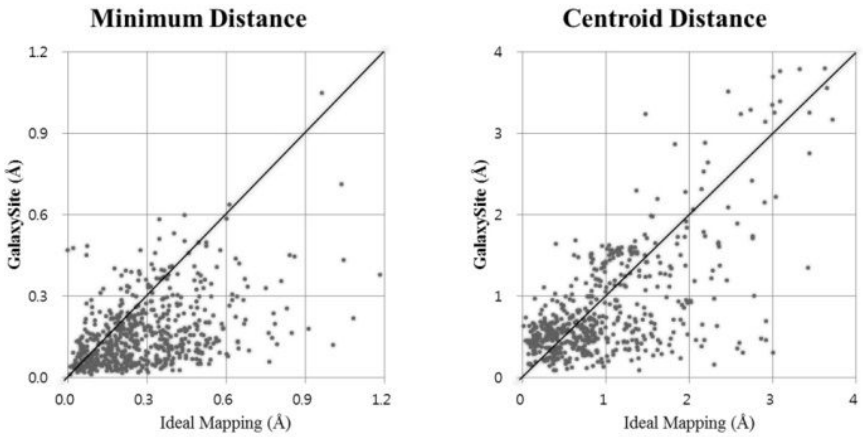

Figure 4.2. Performance comparison between GalaxySite and ideal mapping for individual targets in the nucleotide set. Ideal mapping refers to taking the ligand pose of the best template selected by structure superimposition on the native protein structure. GalaxySite shows improved performance both in terms of minimum distance measure (left) and centroid distance measure (right), providing higher-resolution prediction as a result of docking. The two measures, minimum distance and centroid distance, are defined in **section 4.2.6**.

4.3.2. Comparison with other methods on the bound/unbound set

Performance comparison of different binding site prediction methods is available for the bound/unbound set (Huang and Schroeder, 2006). In **Table 4.2** the performance of GalaxySite on the same set is compared with those of other methods. Success rates (with the minimum distance criterion of 4 Å following (Huang and Schroeder, 2006)) of GalaxySite are 92% (bound) and 90% (unbound) when the top ligand is docked, and 100% and 98% when the best predictions out of the docking results of the top three ligands are considered. This performance of GalaxySite is superior to those of other methods compared in the table.

Accuracy of ligand prediction is only 26% (top ligand) and 34% (top three ligands) for the bound/unbound set. The prediction of chemically similar ligands also contributes to the higher success rate for binding site prediction than the rate of exact ligand prediction, as in the nucleotide set. Notably, the top three ligands can cover actual ligand type in most of the cases for this test set, resulting in a success rate of 100%.

Failures of binding site prediction with the top ligand are mainly due to unsuccessful ligand prediction. For example, for the target 4dfr (bound)/5dfr (unbound), NDP (NADPH) was selected as the top ligand, but MTX (methotrexate), the second ligand, is actually bound. Docking of MTX locates the binding site with minimum distances of 0.26 Å and 0.75 Å for the bound and unbound structures, respectively. However, it is hard to exclude the possibility of the existence of alternative NDP binding sites since 5 templates have both ligands (NDP and MTX) on two different binding sites. This case suggests that the predicted binding sites may provide additional information on other putative ligands.

It is interesting to note that GalaxySite is not very sensitive to inaccuracies in

the input protein structure, showing only 2% decrease in success rate when an unbound protein structure is used for docking. Only FINDSITE shows lower sensitivity (no decease in success rate with unbound structure) than GalaxySite. Typical molecular docking results are expected to be sensitive to protein structure change and ligand type, but the current docking energy greatly reduces dependence on detailed protein and ligand structures by a proper combination of restraint energy and physicochemical energy. The overall energy landscape represented by template restraints is broad and generous so that docking tends to be easily driven toward the binding pocket, and the physicochemical docking function (AutoDock3 here) locates the ligand more precisely in the binding pocket.

The performance of GalaxySite is compared with that of FINDSITE (Brylinski and Skolnick, 2008, 2009) for individual targets to understand the characteristics of the current method better. Prediction with FINDSITE was carried out using a locally installed version, using the same template lists as GalaxySite for a fair comparison. GalaxySite generates a template list automatically, but FINDSITE requires an input template list. In **Figure 4.3** (a) and (b), the minimum and centroid distances are plotted. GalaxySite shows better performance than FINDSITE with the minimum distance measure (with smaller average minimum distance of 1.84 Å, compared to 2.98 Å for FINDSITE), and comparable performance with the centroid distance measure. When contact residue measures are employed, GalaxySite is better in average accuracy (70.5% vs 50.4% for FINDSITE) with a P-value of 7.3×10^{-10} but worse in average coverage (65.8% vs 81.8% for FINDSITE) with a P-value of 3.5×10^{-5}. Overall, GalaxySite shows a slightly better average MCC value (0.646 vs 0.611 for FINDSITE), but this is not statistically significant (P-value = 0.14).

FINDSITE is similar to GalaxySite in that it uses information from similar

proteins, but it may be classified as a mapping method. GalaxySite is expected to provide more precise predictions with docking when the protein input structure is more accurate, and FINDSITE is expected to be less sensitive to detailed protein structure. **Figure 4.3** shows that GalaxySite performs better when more precise prediction is possible (when the minimum or centroid distance is smaller).

Table 4.2. Comparison of success rates of different binding site prediction methods on the bound/unbound set using the best (top 1) and the best out of top 3 predictions.

Method[1)]	Top1		Top3	
	Bound	Unbound	Bound	Unbound
GalaxySite	**92**	**90**	**100**	**98**
FINDSITE	90	90	94	94
VICE	85	83	94	90
DoGSite	83	71	92	92
Fpocket	83	69	94	92
PocketPicker	72	69	85	85
LIGSITE	69	58	87	75
CAST	67	58	83	75
PASS	63	60	81	71
SURFNET	54	52	78	75

1) Results of GalaxySite and FINDSITE were obtained from our own calculations, and those of other methods are taken from (Tripathi and Kellogg, 2010).

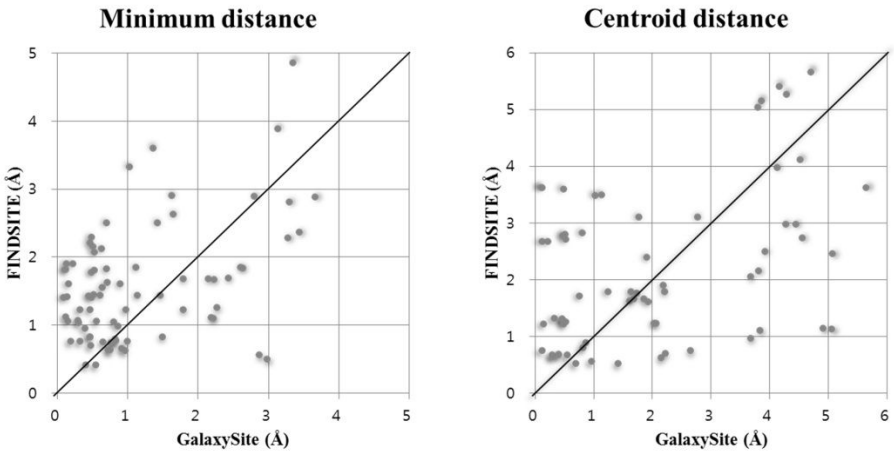

Figure 4.3. Comparison of GalaxySite and FINDSITE on the individual targets of the bound/unbound set. GalaxySite shows better performance when the minimum distance measure is used (left) and comparable performance when the centroid distance measure is used (right). GalaxySite performs better when more precise prediction is possible (with smaller distances).

4.3.3. Binding site prediction on protein model structures for CASP experiments

Here, we further show that the docking method of GalaxySite is sufficiently effective to be applied to predictions from protein sequences when experimental structures are not available. GalaxySite has been tested on CASP9 (Schmidt *et al*., 2011) and CASP10 (Gallo Cassarino *et al*., 2014) in a blind fashion. Predictions on CASP9 targets were performed under group names 'Seok-server' for server targets and 'Seok' for human/server targets. The prediction methods for 'Seok-server' and 'Seok' are identical, except that model structures of 'Seok-server' and 'Seok' were used, respectively.

In **Table 4.3** and **Table 4.4**, the results of CASP9 and CASP10 are summarized using the three contact-based measures, accuracy, coverage, and Matthew's correlation coefficient (MCC) of contacting residues. Distance measures cannot be compared because only the predictions of contact residues are provided in the CASP experiments. Only the results for template-based modelling targets with non-metallic ligands are presented here (9 for CASP9 and 5 for CASP10). We also made predictions for metal-containing proteins in CASP, but the results are not included here because a different prediction strategy was used. Unlike the official CASP9 assessment in which several targets lacking ligands in the experimental structures were included as putative targets, we only considered biologically relevant ligands that were observed in experimental structures, as in the CASP8 official assessment. 'Seok-server' and 'Seok' that use GalaxySite outperformed other methods, except 'ZHANG' and 'I-TASSER_FUNCTION' (which are from the same group), with regard to CASP9 targets. For CASP10 targets, the ranks of median and average MCC were very different, probably because of the small number of targets; the 'Seok-server' ranks among top servers in terms of median

MCC measure. These results suggest that GalaxySite is one of the best available binding-site prediction methods that can be successfully applied to real binding-site predictions from protein sequences.

However, for a number of reasons, the above results have to be interpreted with caution. First, the number of test cases is not large enough to draw statistically meaningful conclusions. Second, assessment results may vary depending on the definition of binding-site residues, for example, whether putative binding-site residues showing strong evolutionary conservation are treated as positive, negative, or neutral.

Table 4.3. Comparison of different binding-site prediction methods on CASP9 binding-site prediction targets with non-metal ligands in terms of median values (average in parentheses) of MCC, accuracy, and coverage.

Predictor	No. of Pred.[1)]	MCC[2)]	Accurac[2)]	Coverage[2)]
ZHANG	8	0.710 (0.671)	72 (63)	81 (77)
SEOK	7	0.709 (0.620)	76 (65)	67 (62)
I-TASSER_FUNCTION[3)]	9	0.697 (0.665)	68 (66)	80 (74)
SEOK-SERVER[3)]	8	0.682 (0.623)	81 (69)	67 (60)
INTFOLD-FN[3)]	8	0.640 (0.570)	82 (69)	58 (52)
JONES-UCL	8	0.585 (0.602)	63 (63)	62 (62)
ATOME2_CBS[3)]	6	0.584 (0.558)	44 (43)	87 (82)
FIRESTAR[3)]	8	0.581 (0.562)	66 (58)	58 (61)
MCGUFFIN	8	0.575 (0.559)	69 (61)	62 (56)
FAMSSEC	9	0.523 (0.568)	40 (49)	75 (75)
STERNBERG	9	0.520 (0.513)	65 (51)	60 (58)
CNIO-FIRESTAR	7	0.508 (0.559)	67 (58)	50 (59)
LEE	8	0.476 (0.506)	51 (57)	51 (53)
GWS[3)]	9	0.464 (0.432)	50 (54)	37 (41)
KIHARALAB	9	0.431 (0.501)	50 (51)	50 (55)
LOVELL_GROUP	9	0.412 (0.363)	40 (44)	43 (36)
TASSER	9	0.396 (0.414)	41 (40)	58 (52)
FINDSITE-DBDT[3)]	8	0.394 (0.346)	35 (34)	46 (45)

3DLIGANDSITE2[3)]	9	0.388 (0.343)	38 (38)	38 (41)
MN-FOLD[3)]	9	0.387 (0.373)	24 (27)	62 (68)
3DLIGANDSITE1[3)]	9	0.342 (0.346)	38 (37)	38 (41)
BILAB-ENABLE[3)]	9	0.326 (0.326)	21 (29)	40 (49)
HHPREDA[3)]	9	0.310 (0.291)	50 (52)	16 (22)
MASON[3)]	9	0.299 (0.284)	29 (31)	36 (33)
3DLIGANDSITE4[3)]	9	0.239 (0.293)	19 (30)	33 (38)
3DLIGANDSITE3[3)]	9	0.239 (0.284)	19 (31)	33 (35)
SAMUDRALA	8	0.227 (0.317)	30 (38)	29 (33)

1) Targets: T0516, T0533, T0547, T0597, T0604, T0609, T0632, T0636, T0641.
2) Contact-based measures: MCC, accuracy, and coverage. A residue is considered contacting ligand if the distance between any atom in the residue and any ligand atom is less than the sum of the van der Waals radii plus 0.5 Å. By comparing the lists of contact residues in the predicted and the native structures, accuracy, coverage, and MCC are used.
3) Server predictors.

Table 4.4. Comparison of different binding-site prediction methods on the CASP10 binding-site prediction targets with non-metal ligands in terms of median values (average in parentheses) of MCC, accuracy, and coverage.

Predictor	No. of pred.[1)]	MCC	Accuracy	Coverage
MCGUFFIN	5	0.850 (0.792)	88 (88)	82 (75)
INTFOLD2[2)]	5	0.845 (0.802)	86 (85)	82 (79)
FIRESTAR[2)]	5	0.821 (0.792)	80 (81)	79 (81)
HHPREDA[2)]	5	0.821 (0.752)	80 (77)	79 (77)
SEOK	5	0.814 (0.750)	94 (85)	71 (69)
SEOK-SERVER[2)]	5	0.814 (0.723)	94 (81)	73 (68)
3DLIGANDSITE[2)]	4	0.785 (0.772)	78 (78)	88 (81)
CNIO	5	0.790 (0.787)	77 (75)	87 (87)
SP-ALIGN[2)]	5	0.780 (0.744)	72 (70)	85 (84)
FNGUSHAK	4	0.779 (0.777)	75 (74)	84 (86)
COFACTOR_HUMAN	5	0.772 (0.768)	81 (78)	77 (78)
COFACTOR[2)]	5	0.772 (0.768)	81 (78)	77 (78)
3DLIGANDSITE2	5	0.772 (0.763)	75 (72)	84 (85)
ATOME2_CBS[2)]	5	0.755 (0.722)	72 (70)	86 (79)
CONPRED-UCL[2)]	5	0.604 (0.524)	56 (52)	68 (61)
CHUO-BINDING-SITES	5	0.467 (0.514)	32 (39)	86 (86)
BINDING_KIHARA[2)]	5	0.352 (0.368)	60 (63)	19 (27)

1) Targets: T0652, T0697, T0721, T0737, T0744.

2) Server predictors.

4.3.4. Binding site prediction from sequences of CAMEO targets

We tested GalaxySite on 480 targets of the ligand binding-site prediction category from the continuous automated model evaluation server released between August 16 and November 8, 2013. The protein structure database 'pdb70' released on 12 July 2012 is used. In **Table 4.5**, results of GalaxySite for these targets were compared with available results of other servers in terms of contact-based measures. In CAMEO, a confidence score is reported for each residue rather than the two-state assignment in CASP. To calculate contact-based accuracy measures, residues with a confidence score of >0.95 were considered to be contacting ligands. Small changes in the cut-off value of the confidence score did not affect the values of accuracy measures a lot. A confidence score for GalaxySite could be developed using the distances to ligand atoms or energy contributions, which can be obtained by docking calculations; however, further investigations in this direction were not carried out. The overall results show that the performance of GalaxySite is consistently comparable or superior to other available server methods for a larger set of prediction targets.

Table 4.5. Comparison of different binding-site prediction methods on CAMEO ligand binding-site prediction targets with non-metal ligands in terms of median values (average in parentheses) of MCC, accuracy, and coverage.

Servers	No. of common targets[1)]	MCC	Accuracy	Coverage	GalaxySite MCC	GalaxySite Accuracy	GalaxySite Coverage
Naïve Homology[2)]	45	0.575 (0.500)	100 (86)	41 (34)	0.801 (0.723)	85 (77)	78 (72)
Naïve Pocket[2)]	250	0.153 (0.148)	50 (45)	6 (7)	0.646 (0.511)	70 (57)	61 (51)
Naïve Conservation[2)]	142	0.142 (0.149)	17 (24)	15 (22)	0.656 (0.525)	71 (60)	61 (52)
FunFOLD[3)]	223	0.526 (0.471)	69 (60)	45 (44)	0.672 (0.555)	75 (62)	64 (56)
HHfunc[4)]	110	0.631 (0.607)	57 (53)	89 (81)	0.648 (0.501)	78 (58)	55 (49)
COACH[5)]	297	0.654 (0.585)	86 (76)	53 (51)	0.652 (0.542)	72 (61)	61 (54)

1) The numbers of predicted targets are different for different servers; thus, only common targets were considered for comparison.

2) http://www.schwedelab.org/.

3) http://www.reading.ac.uk/bioinf/FunFOLD/.

4) http://www.soeding.genzentrum.lmu.de/.

5) http://zhanglab.ccmb.med.umich.edu/COACH/.

4.4. Conclusion

GalaxySite is a binding site prediction method that employs molecular docking guided by evolutionary information. Unlike previous binding site prediction methods, GalaxySite predicts specific binding ligands and binding poses. Such specific information would be very useful for computer-aided drug discovery, for example when homology models are used for virtual ligand screening. Overall the success rate of GalaxySite is superior or comparable to an ideal mapping method when a medium-resolution criterion is used. It performs better with higher-resolution criteria, meaning that more precise predictions are possible when overall binding sites can be predicted more accurately. When tested on ligand-unbound protein structures and model protein structures, GalaxySite performs better than or comparable to existing methods, showing that molecular docking can be successfully applied to binding site prediction problems. This method may be further applied to optimizing conformations of binding site residues to provide more refined homology models for virtual screening.

5. GalaxyRefine:
A protein structure refinement method using GALAXY programs

5.1. Introduction to protein structure refinement

The structure of a protein can be predicted accurately from its sequence by template-based modeling when the sequence identity is sufficiently high (for example, > 30%) (Kryshtafovych *et al*., 2011; Marti-Renom *et al*., 2000). However, even at a high sequence identity, side-chain structure may be less accurate than the backbone structure, whereas at a lower sequence identity, predicted structures may have significant errors in both side-chain and backbone structures. Although *ab initio* protein structure predictions from sequences are notoriously difficult (Ben-David *et al*., 2009; Kinch *et al*., 2011a; Tai *et al*., 2014), *ab initio* refinement starting from a reasonable initial model structure is expected to be less difficult. Successful refinement can increase the applicability range of template-based models by providing more precise structures for functional study, molecular design, or experimental structure determination (MacCallum *et al*., 2009; MacCallum *et al*., 2011; Nugent *et al*., 2014).

To improve the local and global accuracy of template-based models, many types of structure refinement methods have been applied (Heo *et al*., 2013; Khoury *et al*., 2014; Ko *et al*., 2011; Mirjalili and Feig, 2013; Mirjalili *et al*., 2014; Park *et al*., 2011; Park *et al*., 2014; Rodrigues *et al*., 2012; Xu *et al*., 2011; Zhang *et al*., 2011). One type of these methods is based on molecular dynamics simulation (Mirjalili and Feig, 2013; Mirjalili *et al*., 2014). Molecular dynamics simulation

with typical energy functions based on molecular mechanics has often been adopted because it shows high performance in improving protein structure accuracy in terms of physical and chemical features (Nugent *et al*., 2014). To sample diverse conformations during the MD-based refinement process, a large amount of computation time is needed to run long trajectories of simulation. Recently, Mirjalili *et al.* has reported a refinement method based on 600 ns MD simulation which was successful in the recent CASP refinement experiment (Mirjalili and Feig, 2013; Mirjalili *et al*., 2014). They utilized the snapshots of each trajectory to generate an ensemble of diverse conformations. Also, some studies showed that when using MD simulation for structure refinement, some type of restriction is needed and otherwise the structure model will just drift away compared to the experimental structure (Mirjalili and Feig, 2013; Raval *et al*., 2012). Therefore, using structural restraints during the simulation has been adopted and shown success.

Features like these that are known to be required for refinement of protein structures based on MD simulation however actually hinder efficient conformational space sampling. To cross thermal barriers is difficult if simulation time is not long enough and applied restriction confines the sampling range. In this chapter, I introduced some features to the typical molecular dynamics-based refinement method to tackle the aforementioned problems and improve conformational space search for refinement.

The first version of GalaxyRefine method performs perturbations on a patch of amino acid side-chains and relaxation based on short molecular dynamics simulations, iteratively (Heo *et al*., 2013). With the new sampling method, a hybrid energy function is used for the simulation; it is composed of not only energy functions based on molecular mechanics, but also statistical potentials which

smooth the energy landscape. The second version of GalaxyRefine (GalaxyRefine2) additionally performs secondary structure element perturbations for larger conformational change. In addition to the local perturbations, a new type of structural restraint utilizing anisotropic network model (ANM) (Atilgan *et al*., 2001; Bakan *et al*., 2011) is implemented. Structure information of ANM conformations are used as restraints for relaxation based on MD simulation. The refinement methods were benchmarked on previous CASP experiment targets (MacCallum *et al*., 2009; MacCallum *et al*., 2011; Nugent *et al*., 2014) and had been tested to improving other structure models provided by different protein structure prediction methods (Leaver-Fay *et al*., 2011; Xu *et al*., 2011). The methods are computationally much efficient compared to other MD simulation based methods and results in explorative sampling which still contains conformations closer to the native structure. Briefly, details on developed strategies to enhance conformational sampling for protein structure refinement and its results are addressed in this thesis.

5.2. Methods

5.2.1. Overall method for GalaxyRefine

The GalaxyRefine method is composed of two steps, and it is outlined in **Figure 5.1**. The side-chains in the initial models are optimized first with a simple side-chain optimization method implemented in GALAXY programs (Heo *et al*., 2013; Ko *et al*., 2012). Two different relaxation methods (mild and aggressive) are applied, which performs repetitive local structure perturbation and followed short molecular dynamics relaxations. The lowest energy model out of 32 models generated by the mild relaxation is returned as model 1, and four additional models closest to the four largest clusters of 32 models generated by aggressive relaxation

are returned as models 2–5.

5.2.2. Overall method for GalaxyRefine2

The GalaxyRefine2 method is mainly composed of five steps, and it is outlined in **Figure 5.2**. In comparison with the previous GalaxyRefine method (Heo *et al*., 2013), several features have been added or elaborated to enhance global and local conformational sampling and model selections. The initial side chain optimization process is newly introduced to start with accurate local conformations. To cover broader conformational space, two methods are developed which sample global structures by using guiding structures generated by anisotropic network model (ANM) (Atilgan *et al*., 2001; Bakan *et al*., 2011). Local structure samplings are performed for every sampled global structure after ANM-guided sampling by using molecular dynamics-based relaxations. Optimized structures are ranked by colony energy-based method (Lee and Seok, 2008; Xiang *et al*., 2002) or protein model quality assessment method (Wallner and Elofsson, 2005). Finally, selected models are further optimized with full atom representation.

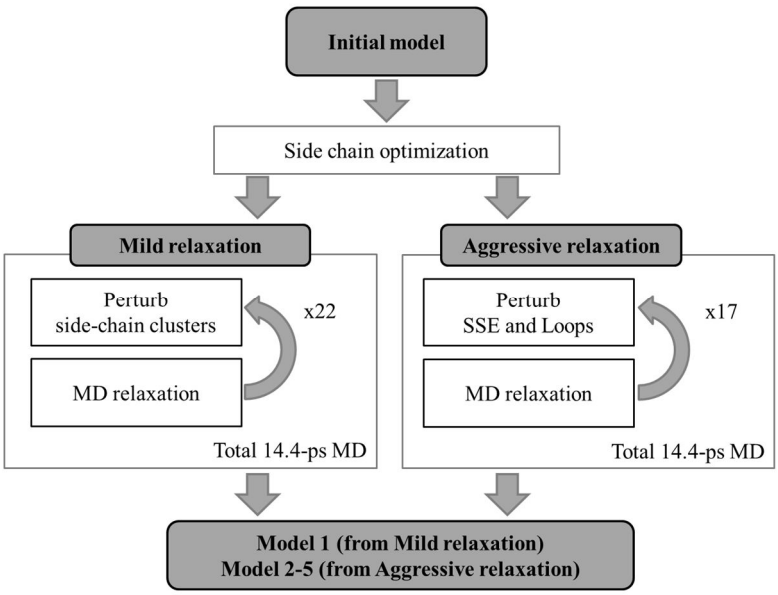

Figure 5.1. Flowchart for GalaxyRefine.

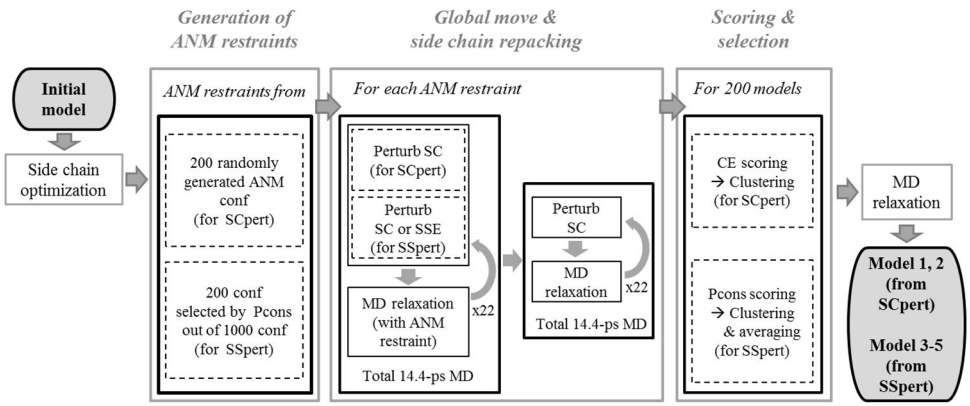

Figure 5.2. Flow chart for GalaxyRefine2.

5.2.3. Physicochemical energy functions used for conformation samplings

Energy functions used for conformational sampling in both GalaxyRefine are composed of physicochemical energy functions, statistical scoring functions, and structure-guiding spatial restraints (Heo *et al*., 2013). The weighted linear summation of these terms is used for energy calculation (**Eq. 5.1**).

$$\begin{aligned} E = {} & E_{\mathrm{MM}} + w_{\mathrm{vdw}} E_{\mathrm{vdw}} + w_{\mathrm{Coulomb}} E_{\mathrm{Coulomb}} + w_{\mathrm{FACTS}} E_{\mathrm{FACTS}} \\ & + w_{\mathrm{dDFIRE}} E_{\mathrm{dDFIRE}} + w_{\mathrm{HBond}} E_{\mathrm{HBond}} + w_{\mathrm{Rotamer}} E_{\mathrm{Rotamer}} + w_{\mathrm{Rama}} E_{\mathrm{Rama}} \\ & + w_{\mathrm{rsr}} E_{\mathrm{rsr}} \end{aligned} \tag{5.1}$$

Except for the optimization steps with all atom representation (based on CHARMM22 (MacKerell *et al*., 1998)) in GalaxyRefine2, the other conformational sampling processes are performed with simplified molecular representation (based on CHARMM19 (MacKerell *et al*., 1998)), and different weights are used for each topologies. For the energy function based on simplified molecular representation, weights for each energy terms except for the restraints are used likewise in GalaxyCassiopeia. These terms were manually optimized to give good energy-structure quality correlations for template-based modeling decoy structures. For the all atom representation using energy function, the energy weights are modified from that using simplified representation to show similar energy fluctuations during the GalaxyRefine simulations.

The only difference in the energy function between GalaxyRefine and GalaxyRefine2 is structure-guiding spatial restraints. In the GalaxyRefine, the spatial restraints are consisted of pairwise distance restraints between backbone Cα atoms or backbone nitrogen and oxygen atoms (**Eq. 5.2**). With these restraints, it sometimes failed to maintain global structures, if there was few numbers of contacts. To overcome the weakness of the previous spatial restraints, Cartesian

coordinates based spatial restraints are newly introduced (**Eq. 5.3**).

$$E_{rsr}^{\text{GalaxyRefine}} = \sum_{i-j>1, d_{ij,\text{C}\alpha\text{-C}\alpha}^{0}<10\text{Å}} \left(\frac{d_{ij,\text{C}\alpha\text{-C}\alpha} - d_{ij,\text{C}\alpha\text{-C}\alpha}^{0}}{2} \right)^2 + \sum_{|i-j|>1, d_{ij,\text{N-O}}^{0}<10\text{Å}} \left(\frac{d_{ij,\text{N-O}} - d_{ij,\text{N-O}}^{0}}{2} \right)^2 \quad (5.2)$$

$$\begin{aligned} E_{rsr}^{\text{GalaxyRefine2}} &= wE_{\text{rsr,Dist}} + (1-w)E_{\text{rsr,Cart}} \\ &= w\left(\sum_{i-j>1, d_{ij,\text{C}\alpha\text{-C}\alpha}^{0}<10\text{Å}} \left(\frac{d_{ij,\text{C}\alpha\text{-C}\alpha} - d_{ij,\text{C}\alpha\text{-C}\alpha}^{0}}{2} \right)^2 + \sum_{|i-j|>1, d_{ij,\text{N-O}}^{0}<10\text{Å}} \left(\frac{d_{ij,\text{N-O}} - d_{ij,\text{N-O}}^{0}}{2} \right)^2 \right) \\ &+ (1-w)\left(\sum_{i} \left(\frac{|\mathbf{r}_{i,\text{C}\alpha} - \mathbf{r}_{i,\text{C}\alpha}^{0}|}{2} \right)^2 \right) \end{aligned} \quad (5.3)$$

Balance between both types of restraints in **Eq. 5.3**, relative weights between these restraints and the other energy terms are trained with 53 refinement problems from CASP8–10 refinement targets (MacCallum *et al.*, 2009; MacCallum *et al.*, 2011; Nugent *et al.*, 2014). The GalaxyRefine mild relaxation method (Heo *et al.*, 2013) is applied to determine the balance between two types of restraints and relative weight for the restraints to the other energy terms. The spatial restraints are utilized to guide global structure sampling around biased to target structures. These restraints are generally used for various global structures during the overall refinement processes in GalaxyRefine2, whereas they were used only for the initial global structure in the previous GalaxyRefine method (Kryshtafovych *et al.*, 2011).

5.2.4. Optimization of initial side chain conformation

For GalaxyRefine method, it performs a simple side chain optimization method, before molecular dynamics relaxations. The method rebuilds all side chain by

placing the highest-probability rotamers (Canutescu *et al*., 2003; Heo *et al*., 2013), starting from the core and then extending to the surface layer by layer. On detecting steric clashes, rotamers of the next highest probabilities are attached. After attaching all side chains, the number of neighboring Cβ atoms is counted around each side chain, and the initial side-chain conformation is recovered if the number deviates from the canonical distribution for the amino acid under the same degree of surface exposure.

For GalaxyRefine2 method, a side chain optimization method named Galaxy-optSC is introduced to optimize the initial side chain conformations. The method optimizes the refinement energy function by replacing rotamers with fixed backbone structures by solving rotamer combinatorial problems. To make the problem simple, it does not optimize the whole protein at once, but it iteratively solves some selected part of proteins at once. It first selects a third of residues randomly to define side chain sets to be optimized at a time. For each selected residue, residues within 8 Å Cβ distance are included in a residue set. These selected residues are only considered for a single optimization process because the method uses more complicated energy function to be optimized. In a descending order of the number of neighbor residues, the best rotamer combinations are applied to a residue set. The solution finding procedure for a residue set is similar to that of Scwrl3 (Canutescu *et al*., 2003), which adopts graph theory and dead-end elimination approach to find out the best combination of rotamers. Rotamers having higher than 1% rotamer probability and less than 90% cumulative rotamer probability are incorporated into rotamer candidates. A side chain interaction graph is built and solved as in Scwrl3; different energy cutoff for connecting nodes is used because the energy function differs from that of Scwrl3. The cumulative solution for each iteration of optimization becomes the final solution for the global side chain optimization process.

5.2.5. Conformational sampling by repetitive perturbation and molecular dynamics (MD) relaxations

For GalaxyRefine method (Heo *et al*., 2013), it performs two relaxation methods, a mild relaxation and an aggressive one. Both of the methods are based on repetitive relaxations (22 and 17 for mild and aggressive relaxations, respectively) by short molecular dynamics simulations (0.6 ps and 0.8 ps for mild and aggressive relaxations, respectively) with 4 fs time step after structure perturbations. Structure perturbations are applied only to clusters of side chains in the mild refinement, whereas more forceful perturbations to secondary structure elements and loops are applied in the aggressive refinement. The triaxial loop closure method (Coutsias *et al*., 2004) is employed to avoid breaks in model structures caused by perturbations to internal torsion angles. Typically, GalaxyRefine method generates 32 models with each relaxation method.

5.2.6. Conformational sampling by anisotropic network model (ANM)-guided relaxations

For GalaxyRefine2 method, it performs two steps of relaxation; the first one is ANM-guided relaxation and the second one is similar to the mild relaxation described in **section 5.2.5**. In the previous GalaxyRefine method, it was successful in refining both global and local structures, but the amount of refinement was quite small for global structure improvement. Anisotropic network model (ANM) (Atilgan *et al*., 2001) is incorporated to enhance the global structure sampling by guiding structures for the global structure sampling. Prody python module (version 1.5.1) is used for the ANM calculations (Bakan *et al*., 2011). Several parameters

were determined for better ANM calculations: the number of the lowest frequency modes to be utilized, Cα-Cα distance cutoff, the maximum ANM mode extrapolation amplitudes for the upper limit of random amplitude, and the number of structure generation. These parameters were determined to maximize the efficacy of sampled structures by benchmarking on the CASP refinement category targets. The extrapolation amplitude is randomly selected within the maximum value.

Two types of ANM-guided relaxation methods were adopted to sample conformations with different range of structural perturbations, local and global. For relaxation with perturbations only on side chains, 200 of the 1,000 ANM conformations were selected randomly as guiding conformations. For relaxation with perturbations on secondary structure elements on the other hand, after evaluating the 1,000 ANM conformations using the Pcons score (Wallner and Elofsson, 2005), top 200 were selected. As secondary structure element perturbation gives more change to initial structures during relaxation, to start with more converged structures, guiding ANM conformations selected based on a consensus approach were used. For each method, 200 conformations are sampled and all the generated structures are used for the next sampling step.

The first ANM-guided relaxation method mainly samples side chains for a given ANM conformation. Starting from the side chain optimized initial structure, conformational space is searched by applying repetitive perturbations on side chain clusters (named SCpert) followed by structural relaxations based on short molecular dynamics. During the two-thirds of the simulation time from the start, high temperature (300 K) is used for efficient samplings, and it is annealed to low temperature (50 K) smoothly during the remaining time to let the structures converge towards a stable state. To briefly state, the overall sampling procedure

based on SCpert is similar to the mild relaxation method used in the previous GalaxyRefine method except that the searched conformational space has been broadened with various spatial restraints biased to the guiding conformations generated by ANM.

With bias to the guiding conformation as a result of the use of restraint energy, the energy barriers cannot be crossed easily without large structural perturbations. Therefore, for the second type of relaxation method which was developed with intention to sample more diverse conformations, secondary structure elements were selected to be perturbed (named SSpert) after each relaxation steps. When more than one secondary structure elements are contained in the structure, secondary structure elements with hydrogen bond between them are grouped to be perturbed together. Secondary structure elements were assigned to be chosen with more probability for perturbation when they are farther from the core of the structure. After the elements are chosen, one axis penetrating the secondary structure element closest to the structure core and two other axes making three of them vertical to each other are defined geometrically and one of them is selected randomly as the perturbation axis. The elements are perturbed with translation or rotation moves directed by the selected axis with the move size also determined randomly following the Gaussian distribution function with assigned average and standard deviation values.

The models generated by ANM-guided relaxations are further optimized by applying SCpert relaxations again. Two methods for ANM-guided relaxations, used in the previous step, mainly focus on searching the global conformational space, so the models from each method still have rooms for local structure improvements. Starting from single structure, its global structure deviates from the initial structure with structure guiding conformations, and these processes sample global structure

efficiently. However, the local structures for the sampled global structure could not be sampled enough, so they should be more sampled and optimized while the global structures are not to be changed too much. While the global structure is biased towards each starting model, their conformational spaces are searched again mainly focusing on improving their local structure qualities. Perturbations on side chain clusters and structural relaxations by short molecular dynamics (SCpert) are applied repeatedly for each starting structure to sample local conformational spaces.

5.2.7. Model selection methods

For GalaxyRefine method, it simply selects 5 refined models based on the refinement energy. The first model is taken from mild relaxation and the four models are taken from aggressive relaxation as model 2–5. However, this approach often failed to select better models among the samples structures (Heo *et al*., 2013; Nugent *et al*., 2014). Complicate model selection methods are introduced in GalaxyRefine2. Two different approaches are used to select the representative structures for different relaxation methods. After model selections with each method, two models are taken from SCpert sampling method as model 1 and 2, and three models are taken from SSpert sampling method as model 3–5.

For the structures generated by SCpert method, they are re-scored by colony energy (CE) (Lee and Seok, 2008; Xiang *et al*., 2002), which can consider both energy and structural similarities to the other conformations. For each generated structure, colony energy (**Eq. 5.4**) is evaluated.

$$CE_i = \sum_j \exp\left[-\alpha \left(\frac{1 - \mathrm{TM}_{ij}}{1 - \overline{\mathrm{TM}}} \right) - \beta \left(\frac{E_j - \overline{E}}{\sigma(E)} \right) \right] \tag{5.4}$$

In the formula, the first term in the exponential is related to its structural similarity to the other conformations. Parameter α, TM_{ij}, and $\overline{TM}$ stand a parameter to define similar structure, structure similarity measured by TM-score (Zhang and Skolnick, 2004), and their average on all structure pairs, respectively. The second term considers energy contributions of the other conformations. The energies are normalized because energy fluctuations vary depending on the system size. In ascending order of the colony energy, structures are picked up to construct a structure pool for clustering until 10 dissimilar structures are included. The structural similarity is measured by the first term in (**Eq. 5.4**). A structure is added to the pool as a dissimilar structure if any structure in the pool as dissimilar structures has less than 0.2 similarity measure, or it is just added to the pool. After the construction of structural pool, the structures are clustered by using NMRclust (Kelley *et al*., 1996) and the two lowest colony energy conformations are selected as representative structures.

On the other hand, for the structures generated by SSpert method, due to the diversity of the conformations sampled, a selection method based on the use of quality assessment method with consensus approach was employed. The 200 structures generated were evaluated with the Pcons score (Wallner and Elofsson, 2005). Though the evaluation result of the 200 conformations was used, 200 initial conformations selected from ANM-generated models were also included in the scoring pool to increase the diversity of the population, resulting in 400 conformations to be evaluated using the consensus method. The additionally included ANM models are only used for score evaluation and the top 100 structures among the 200 relaxed structures are subjected to further clustering process. Starting from the top Pcons score conformation, the structural similarities of the other conformations compared to the reference structure (initially the top Pcons score structure) are measured and if the normalized TM-score (Zhang and Skolnick,

2004) is smaller than 1.0, it is defined to be similar to the reference. With the maximum number of cluster members 20, similar conformations are included to the cluster containing the highest Pcons-scored conformation which is a representative of the cluster and also the reference of measuring structural similarity. When a conformation is not similar to any of the existing cluster representatives, it becomes a representative of a new cluster. This process is repeated until all the 100 conformations are sorted into any one of the clusters. Starting from the cluster containing the top Pcons-scored conformation as a representative, three clusters are selected. For each cluster, all the cluster member conformations are geometrically structure-averaged to generate a new cluster representative structure. Since a naively coordinate-averaged structure makes physical nonsense, the structure is optimized afterwards in all-atom representation based on short molecular dynamics simulation which will be described in detail in the **section 5.2.8**.

5.2.8. Molecular representation conversion into all-atom topology

The molecule representation used for the conformational space sampling steps consists of heavy atoms and hydrogen atoms only on polar heavy atoms. This simplified representation rather than the all-atom representation is used for the relaxation steps to make the energy function less sensitive to local atomic changes and hence enhance sampling efficiency. However, to optimize the model quality in the atomic level and increase the local structure accuracy, as the last step of the whole refinement process in GalaxyRefine2, the selected conformations from the prior relaxation steps are globally optimized with full molecule representation. A similar energy function used for the relaxation steps but with all-atom molecule representation and relative energy weights for each energy terms slightly modified is used. To be optimized, the selected initial structures undergo relaxation based on

short molecular dynamics simulation. The weight for self-structural restraint energy is set higher than the relaxation steps to not let the conformations move away from the initial structures. Eight simulation trajectories of relaxation are run for each selected initial structure and the lowest energy conformation from all trajectories becomes our final refined model.

5.2.9. Benchmark and test set

The GalaxyRefine method tested on 27 CASP10 refinement category targets (Nugent *et al*., 2014). The method also has been benchmarked in blind fashion on protein structure prediction server models for CASP10 template-based modeling (TBM) category targets: Zhang-server (I-TASSER) (Xu *et al*., 2011) models (35 proteins) and BAKER-ROSETTASERVER (Robetta) (Leaver-Fay *et al*., 2011) models (34 proteins). For both CASP10 TBM server models, only single domain targets with initial model GDT-HA > 0.4 were considered and targets for refinement category were excluded. In addition, it has been tested on FG-MD benchmark set targets (131 proteins), which have initial model GDT-HA > 0.4 were only considered.

The GalaxyRefine2 method also has been benchmarked in blind fashion on 53 CASP8–10 refinement category targets: 12 CASP8 targets (MacCallum *et al*., 2009), 14 CASP9 targets (MacCallum *et al*., 2011), and 27 CASP10 targets (Nugent *et al*., 2014). In addition to the CASP refinement category targets, the method has been benchmarked on the same CASP10 TBM benchmark sets and FG-MD benchmark set (Zhang *et al*., 2011), also. Finally, the GalaxyRefine2 method participated in CASP11 refinement category (37 targets). The initial model information for the benchmark and test set are summarized in **Table 5.1**.

Table 5.1. Summary of benchmark and test set.

Set	No. Targets	Initial Model Quality		
		GDT-HA	GDC-SC	MolProbity
CASP8	12	50.84	31.34	2.84
CASP9	14	53.04	30.76	2.49
CASP10	27	58.54	33.57	2.46
CASP11	35	53.12	29.60	2.60
CASP10 I-TASSER[1)]	35[3)]	58.13	31.59	2.77
CASP10 ROSETTA[2)]	34[3)]	61.18	38.40	1.34
FG-MD benchmark set	135[4)]	58.03	32.40	2.88

1) Zhang-server models submitted for the CASP10 TS category targets.

2) ROSETTA-BAKERSERVER models submitted for the CASP10 TS category targets.

3) Single domain targets with GDT-HA > 0.4 and CASP refinement category targets were excluded for CASP10 server model test sets.

4) Targets with GDT-HA > 0.4.

5.3. Results and Discussion

5.3.1. Benchmark and test results of GalaxyRefine and GalaxyRefine2

The overall refinement results on CASP refinement category targets (MacCallum *et al*., 2009; MacCallum *et al*., 2011; Nugent *et al*., 2014) are summarized in **Table 5.2**. In addition, the overall results on CASP10 TBM server models including I-TASSER server models (Xu *et al*., 2011) and BAKER-ROSETTASERVER models (Leaver-Fay *et al*., 2011), and FG-MD benchmark set (Zhang *et al*., 2011) are also summarized in **Table 5.3**. Three model quality evaluation measures are used: GDT-HA (Zemla, 2003) and GDC-SC (Keedy *et al*., 2009) for global and local model accuracy measure, respectively, and MolProbity score (Chen *et al*., 2010) for physical correctness. For clearance, all results are shown in their improvements from the initial structure, so improvements in each measurement have positive values.

In global model accuracy, GalaxyRefine2 method outperforms the previous GalaxyRefine method (Heo *et al*., 2013). When the model 1 is compared, GalaxyRefine2 shows much better results on average on overall sets. When cherry-picked model is compared, GalaxyRefine2 is still better than GalaxyRefine, but the gap is smaller than that of model 1. ANM-guided sampling and secondary structure element perturbation enriches the conformational samplings, and the new model selection method improves the refinement methods. In local model accuracy, GalaxyRefine2 method did much better than GalaxyRefine method. Local model refinement result on CASP10 ROSETTA set was relatively smaller than on the other test set because the initial local model accuracies are much higher than the other set. It is possible to improve side chain packing initially by Galaxy-optSC, a new side chain optimization method. Finally, improvement in physical correctness of the refined models was enhanced also. Physical correctness improvement on

CASP10 ROSETTA set was also smaller than on the other set and it was also due to their high physical correctness of the initial models. Due to the final optimization with all-atom representation, most of atomic clashes could be removed efficiently, and this gives dramatic improvement in MolProbity measure.

Table 5.2. Refinement results on CASP refinement category targets for model 1 and the best model out of model 1–5 in parentheses.

Set	GalaxyRefine Improvement			GalaxyRefine2 Improvement		
	GDT-HA	**GDC-SC**	**MolP**[1)]	**GDT-HA**	**GDC-SC**	**MolP**[1)]
CASP8	0.54 (1.36)	2.49 (3.27)	0.89 (1.19)	0.42 (1.44)	3.06 (3.86)	1.55 (1.79)
CASP9	0.52 (1.99)	0.82 (1.38)	0.53 (0.72)	1.53 (2.12)	1.64 (2.66)	1.16 (1.38)
CASP10	-0.32 (0.88)	0.86 (2.29)	0.36 (0.49)	0.35 (1.16)	2.53 (3.48)	1.20 (1.35)
CASP11	-[2)]	-[2)]	-[2)]	0.91 (1.68)	1.90 (3.09)	1.25 (1.41)

1) MolProbity Score Improvement.

2) GalaxyRefine was not applied to CASP11 TR targets.

Table 5.3. GalaxyRefine2 test results on CASP10 server models and FG-MD benchmark set for model 1 and the best model out of model 1–5 in parentheses.

Set	GalaxyRefine Improvement			GalaxyRefine2 Improvement		
	GDT-HA	**GDC-SC**	**MolP**[3)]	**GDT-HA**	**GDC-SC**	**MolP**[3)]
CASP10 I-TASSER[1)]	-0.20 (0.66)	2.65 (3.56)	0.54 (0.87)	0.37 (1.12)	3.28 (4.51)	1.34 (1.53)
CASP10 ROSETTA[2)]	0.00 (0.55)	-0.47 (0.02)	-0.35 (-0.28)	0.46 (1.09)	0.05 (0.99)	0.21 (0.31)
FG-MD benchmark set	0.55 (1.66)	1.74 (2.91)	0.82 (1.05)	0.91 (1.96)	1.99 (3.22)	1.52 (1.73)

1) Zhang-server models submitted for the CASP10 TS category targets.

2) ROSETTA-BAKERSERVER models submitted for the CASP10 TS category targets.

3) MolProbity Score Improvement.

5.3.2. Energy function parameterization

The GalaxyRefine method (Heo *et al*., 2013) was applied on 53 CASP8–10 refinement category targets (MacCallum *et al*., 2009; MacCallum *et al*., 2011; Nugent *et al*., 2014) to determine relative weights between pairwise distance restraints and Cartesian coordinates based restraints, and restraints weight with respect to the other energy terms. For each target, 32 models were generated by using the GalaxyRefine mild relaxation method with different weights on restraints, and the lowest energy model was selected as the final model. The restraint weights were determined by grid search. The model quality improvements were measured in GDT-HA (Zemla, 2003) and GDC-SC (Keedy *et al*., 2009). The results are described in **Figure 5.3**. The model quality improvements vary with restraint weights, but it showed considerably better results when both types of restraints are equally used than single type of restraints are solely used. The model quality improvements trends with respect to the overall restraint weight vary with the model quality measures. With restraint weight 5.0 showed the best performance measured in GDT-HA and there was no significant changes measured in GDC-SC. So, we had selected restraint weight 5.0 for further developments because the weight showed the best performance in GDT-HA measure, and it showed quite high improvements in the other measures.

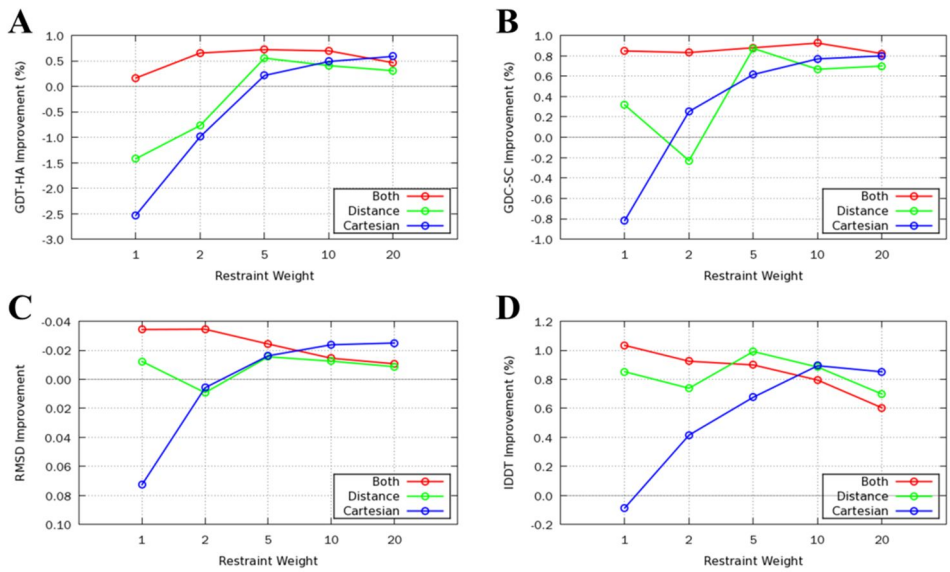

Figure 5.3. Restraint weight optimization results. (A) GDT-HA improvements, (B) GDC-SC improvement, (C) RMSD improvement, (D) lDDT (Mariani *et al*., 2013) improvement.

5.3.3. Initial side chain optimization

A new side chain optimization method named Galaxy-optSC has been introduced in GalaxyRefine2 to replace initial side chain placement method in the previous GalaxyRefine method. The new side chain structure prediction accuracy was compared with the other side chain structure prediction methods such as Scwrl4 (Krivov *et al*., 2009) and Rosetta FastRelax (Tyka *et al*., 2011), and it was also compared with the simple side chain optimization method used in GalaxyRefine (Heo *et al*., 2013). For the benchmark, 53 CASP refinement category targets were used; all methods were tested on the experimental backbone structure with randomly perturbed side chain structures and the initial structures for the refinement targets. To compare side chain prediction accuracy, GDC-SC (Keedy *et al*., 2009), lDDT (Mariani *et al*., 2013), $\chi 1+2$ accuracy, and rotamer outlier ratio (Chen *et al*., 2010) were evaluated. Benchmark results on the experimental structures are shown in the **Table 5.4.** Galaxy-optSC performs much better than the previous method, but the other prediction methods show slightly better results. Although Galaxy-optSC shows the highest rotamer outlier ratio, it is still acceptable ratio in Molprobity measurement (lower than 1%). Benchmark results on the initial structures for refinement targets are shown in the **Table 5.5**. On this benchmark set, Galaxy-optSC outperforms the other methods and improves the initial structures in overall measures. The energy function used in Galaxy-optSC has been trained on protein model structures, so it performs better than the other methods on protein model structures.

Table 5.4. Galaxy-optSC benchmark on 53 experimental structures for CASP refinement category targets.

Measure	Initial Structure[1)]	Galaxy-optSC[2)]	Galaxy-Naïve[3)]	Scwrl4	Rosetta-FastRelax
GDC-SC	56.19	70.98	64.42	69.61	**71.18**
lDDT	0.8568	0.9200	0.8911	0.9206	**0.9241**
χ_{1+2}	39.9%	60.1%	53.0%	**63.5%**	63.1%
Rotamer Outlier	0.33%	0.87%	0.21%	**0.01%**	0.56%

1) Experimental backbone structure with randomly perturbed side-chain structures.

2) The side-chain placement method used in the GalaxyRefine2 method.

3) The side-chain placement method used in the GalaxyRefine method.

Table 5.5. Galaxy-optSC benchmark on 53 initial structures for CASP refinement category targets.

Measure	Initial Structure	Galaxy-optSC[1)]	Galaxy-Naïve[2)]	Scwrl4	Rosetta-FastRelax
GDC-SC	32.32	**33.52**	31.88	32.81	32.99
lDDT	0.6542	**0.6573**	0.6495	0.6533	0.6557
χ_{1+2}	37.9%	**42.5%**	40.6%	40.6%	40.1%
Rotamer Outlier	2.89%	0.93%	0.77%	**0.01%**	2.04%

1) The side-chain placement method used in GalaxyRefine2 method.

2) The side-chain placement method used in GalaxyRefine method.

5.3.4. ANM model generation

With 53 CASP refinement category targets, ANM conformation generation parameters were determined to maximize efficacy of structure samplings. First, the number of low frequency ANM modes was determined. We measured the average correlation between structure deformation vector from model structure to the experimental structure and each calculated ANM modes. As illustrated in the **Figure 5.4** (A), the correlation decreases understandably. We have selected the 20 lowest frequency ANM modes for the ANM model generation based on the correlation trends. Next, the parameters for Cα-Cα distance cutoff, the most important parameter for ANM generation, and extrapolation amplitude were determined. The distance cutoff was tested from 10.0 Å to 20.0 Å for every 2.0 Å, and the maximum extrapolation amplitude was tested from 2.0 to 20.0 for every 2.0. For each parameter combinations, 1000 ANM conformations were generated using the 20 lowest frequency modes. The maximum improvement and percent of improved conformations in GDT-HA measure (Zemla, 2003) were evaluated, as illustrated in the **Figure 5.4** (C) and (D). Parameters with 12.0 Å distance cutoff and bigger than 10.0 for the maximum extrapolation amplitude gave better performance in the maximum improvement measure than the other parameters though there were no significant differences between the top performing parameters. Percent of improved conformation depended only on the maximum extrapolation amplitude; smaller extrapolation amplitudes gave better results. Considering both results, we have selected 12.0 Å distance cutoff and 10.0 for the maximum extrapolation amplitude. The maximum GDT-HA improvement was evaluated for these selected parameters with different number of conformation generations. In the **Figure 5.4** (B), the more conformation generation gave the better results though it was almost converged after 200 conformation generation. Therefore, we determined to select 200 conformations as guiding structures among

1000 generated conformations.

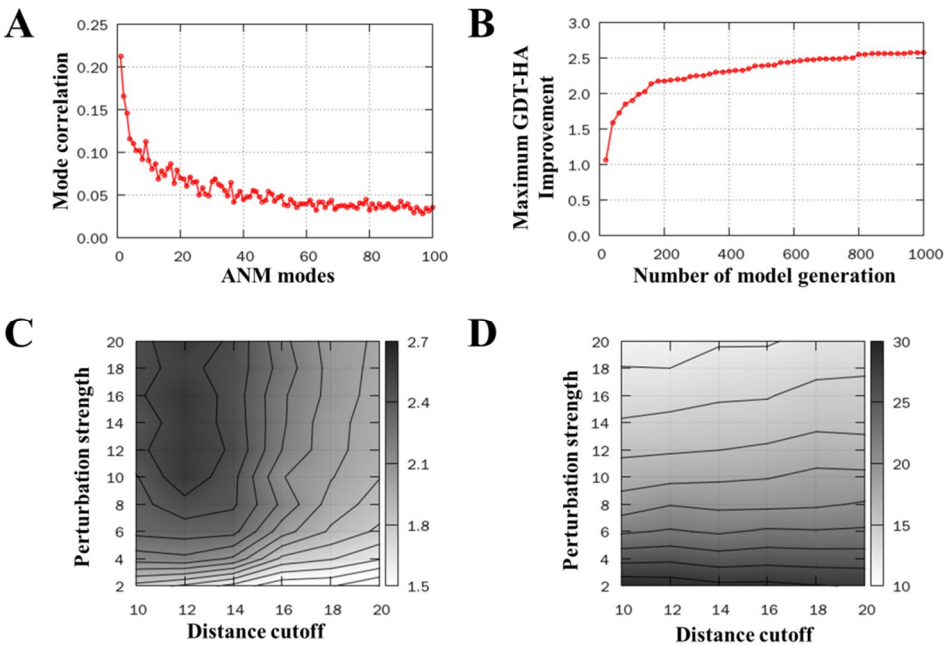

Figure 5.4. ANM model generation method training result. (A) Number of ANM mode use, (B) Number of ANM model generation, (C) and (D) Cα-Cα distance cutoff and perturbation strength.

5.3.5. Global sampling performance comparison for various relaxation methods

There have been two approaches introduced to enhance global samplings for GalaxyRefine method: ANM-guided sampling and secondary structure perturbation (SSpert) method. To measure the effects of new approaches for the GalaxyRefine method, they have been tested on the previous 53 CASP refinement category targets (MacCallum *et al*., 2009; MacCallum *et al*., 2011; Nugent *et al*., 2014). Structural samplings with and without ANM-guiding sampling were performed to test the effect of ANM, and structural samplings with SSpert and without structural perturbation were performed to test the effect of SSpert method. Three measures were used to compare structural diversity of the sampled structures with different sampling methods, and averaged distributions for the all target are shown in **Figure 5.5**. First, maximum global structure improvement in GDT-HA measure (Zemla, 2003) were compared. It is possible to sample much better conformations with ANM-guided sampling method than without ANM-guided sampling and the previous GalaxyRefine method (Heo *et al*., 2013). With SSpert method, it also samples better conformations than without perturbation and the previous GalaxyRefine method, though it sometimes failed to sample better conformations. Second, global structure deviation from the initial structure were compared. In the middle of **Figure 5.5**, both new sampling methods samples more deviating structures than the other methods. The conformational spaces sampled by the previous GalaxyRefine method and without ANM or perturbation were highly biased toward the initial structure with spatial restraints, while two new approaches made it possible to sample more different structures than the previous GalaxyRefine method. Finally, pairwise TM-score (Zhang and Skolnick, 2004) between the generated conformations were evaluated to measure the covered conformational space for each sampling method, and this also shows that the both

new sampling methods samples more diverse conformations.

Conformational spaces sampled by the previous GalaxyRefine method were localized because spatial restraints penalized for diverging conformations. Sampling with weaker or without restraints could sample more diverse conformations, but they often failed to maintain good features in the initial structures. Introduced two approaches diversify the conformational samplings without or less penalties in maintaining structural qualities of the initial structure. For ANM-guided sampling, spatial restraints were not biased toward the initial structure, but targeted new conformation sampled by ANM. The ANM-guided sampling method enables to search more diverse conformations by making it easy for global transitions. For secondary structure perturbation method, spatial restraints were initially biased toward the initial structure. But they were repeatedly updated after secondary structure perturbations, and these made it possible not to trap in the initial structure and accept energetically favored conformations.

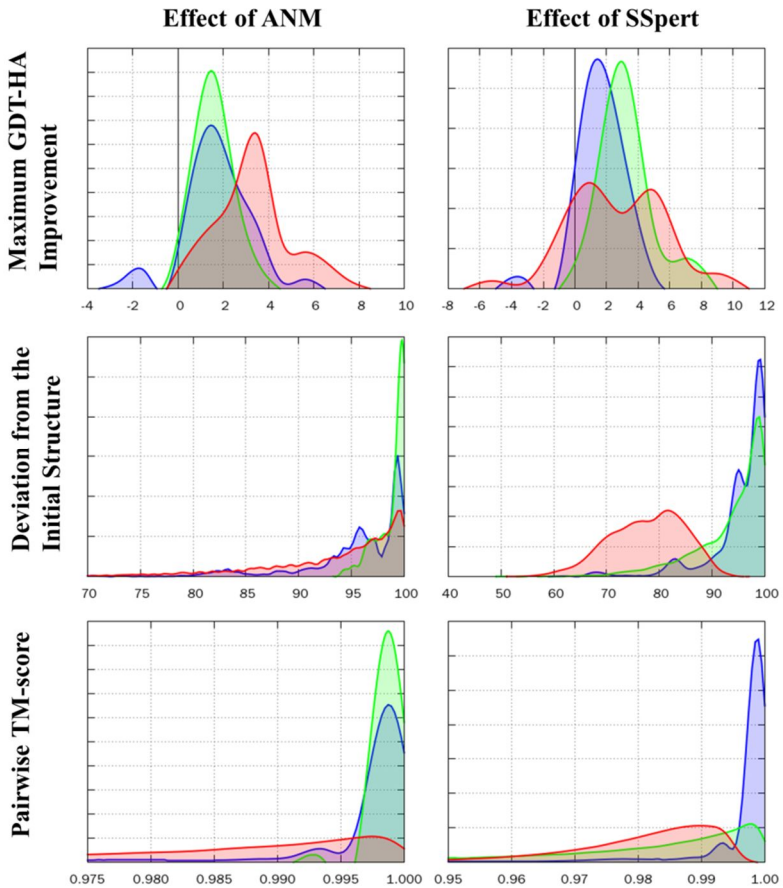

Figure 5.5. Global sampling performance comparison. Data for GalaxyRefine are shown in blue. For the "Effect of ANM" column, with and without ANM restraint are shown in red and green, respectively. For the "Effect of SSpert" column, with and without SSpert are shown in red and green, respectively.

5.3.6. Model selection methods

After the sampling steps, representative conformations are selected. In this section, model selection methods in GalaxyRefine and GalaxyRefine2 are compared. In GalaxyRefine, scoring the sampled conformations using the same energy function used for sampling was tried. In GalayRefine2, the sampled conformations are evaluated with two different model selection methods for different relaxation methods. For the structures generated by using SCpert sampling method, colony energy (Lee and Seok, 2008; Xiang *et al*., 2002) which includes the conformational entropy term to estimate the conformational free energy was adopted. On the other hand, for the structures sampled by using SSpert method, as the generated structures are more diverse, using a scoring function consisting of mostly conformational entropy terms was necessary. Therefore, a consensus quality assessment method Pcons (Wallner and Elofsson, 2005) was adopted. The compared ranking results for each adopted scoring function and the energy functions are depicted in **Figure 5.6**

As shown in **Figure 5.6** (A), for the structures sampled by using SCpert method, colony energy shows higher performance in ranking high quality structures compared to the energy function based selection method. In **Figure 5.6** (B), for the structures sampled by SSpert method, Pcons cannot still pick out the most accurate structure but shows slightly better performance compared to the energy function based method. As even the newly adopted scoring functions cannot discriminate the highest quality structure, representative structure within each cluster was selected finally after clustering based on the scores was done. For the structures generated by SSpert especially, due to the problem of scoring, structure-averaging the conformations from each cluster was adopted. While improving in terms of global conformational space sampling was mainly targeted in this work,

more study on scoring will be done in the future to improve performance in scoring.

Data showing the effect of structure-averaging for the conformations sampled using SSpert method is reported in **Table 5.6**. The result is shown only for the largest cluster. It compares the quality of the structure-averaged representative conformation to that of the cluster center structure with highest Pcons score (Wallner and Elofsson, 2005), the best quality among the cluster members, and the average quality value. Except for the $\chi 1+2$ accuracy measure, it shows improvement for all structure accuracy measures. As reported in previous works, in a global structure level, coordinate-averaged conformation can represent the sampled ensemble (Mirjalili and Feig, 2013). For the side chains however, coordinate-averaging obviously results in unrealistic structures and shows decrease in $\chi 1+2$ accuracy. However, due to reduced fluctuation and increase in packing density, it rather results in increase in other local accuracy measures.

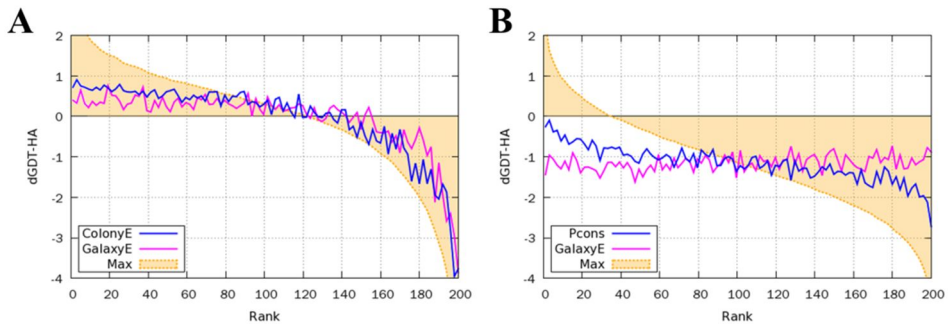

Figure 5.6. Refined model selection method comparison. (A) Colony energy based ranking method for SCpert relaxation and refinement energy based method are compared. (B) Consensus based ranking method for SSpert relaxation and refinement energy based method are compared.

Table 5.6. Model qualities of the largest cluster members and the effect of structure averaging in the selection step after SSpert sampling. Values are averaged on 53 CASP refinement category targets.

Measures	**Cluster center**[1)]	**Cluster best**[2)]	**Cluster average**[3)]	**Structure average**[4)]
GDT-HA	56.70	58.34	56.48	57.12
GDC-SC	33.97	36.13	33.91	34.86
lDDT	0.6633	0.6716	0.6626	0.6689
χ1+2	44.4%	48.3%	43.9%	42.4%
MolProbity	1.84	1.66	1.82	1.66

1) The quality of cluster center, with highest Pcons score (Wallner and Elofsson, 2005) from the largest cluster.

2) The best quality from that of the largest cluster members.

3) Quality values averaged for all cluster members.

4) Quality of the final representative structure which is structure-averaged.

5.4. Conclusion

In this study, I developed GalaxyRefine (Heo *et al.*, 2013) and GalaxyRefine2 which perform protein structure refinement. GalaxyRefine performs repetitive local structure perturbations and short molecular dynamics relaxations with a hybrid energy function. It was proven to be the most successful fully automatic refinement method in CASP10 refinement category (Nugent *et al.*, 2014). The method is particularly successful in improving local structure quality. However, it shows moderate improvement in backbone structure quality on average. To tackle this problem, I developed GalaxyRefine2, recently. It performs more efficient global conformational samplings with successful model selection methods.

In GalaxyRefine2, I adopted anisotropic network model (ANM) (Atilgan *et al.*, 2001; Bakan *et al.*, 2011) and secondary structure element based sampling methods to sample more diverse global structures. Structure sampling with those components allows more diverse sampling with not losing model structural qualities, but sampling much better conformations than the previous GalaxyRefine method. In GalaxyRefine2, it also deals model selection method based on colony energy and model quality assessment method. These approaches gave better model selection results than GalaxyRefine energy based method, but there were much room for improvements in model selection, so development of better model selection method is promising.

6. Conclusion

In this thesis, a new template-based protein structure prediction method, GalaxyTBM, is described. The method successfully combined several existing methods and newly developed methods. For the template identification method, the method adopted two best performing methods based on bioinformatics, and they are combined with a new physicochemical re-ranking scheme. Protein tertiary structure models are generated from sequence alignments and their related template information with consideration of their physicochemical properties. The regions where they are hard to be modeled only with the template information are modeled in *ab initio* fashions by using local structure refinement methods. To identify the functions for the predicted protein structures, ligand binding sites are predicted by using template search and molecular docking methods.

Throughout the thesis, methods from both bioinformatics and physical chemistry are used to complement each other. Methods based on bioinformatics are essential for predicting protein structures with template-based modeling approaches. Identification of the closest known structures and sequence alignment between the target sequence and identified known structures are the most important part of template-based modeling. They are hard to overcome with physicochemical approaches. However, only with those methods, it is hard to predict protein structures with high accuracy. The target protein structure deviates from template structures with the sequence differences. Local side chain packing is changed by point mutations, protein loop structures are varying for their functioning, and global structures are different if they are distant homologs. To reflect and make up these changes, protein structure modeling with consideration of physical chemistry

should be introduced. By combining methods based on bioinformatics and physical chemistry, protein structure modeling can go further beyond the methods based on a single approach.

BIBLIOGRAPHY

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410.

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal* **80**, 505-515.

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

Ben-David, M., Noivirt-Brik, O., Paz, A., Prilusky, J., Sussman, J.L., and Levy, Y. (2009). Assessment of CASP8 structure predictions for template free targets. *Proteins* **77 Suppl 9**, 50-65.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic acids research* **28**, 235-242.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of molecular biology* **112**, 535-542.

Bill, R.M., Henderson, P.J., Iwata, S., Kunji, E.R., Michel, H., Neutze, R., Newstead, S., Poolman, B., Tate, C.G., and Vogel, H. (2011). Overcoming barriers to membrane protein structure determination. *Nature biotechnology* **29**, 335-340.

Bordogna, A., Pandini, A., and Bonati, L. (2011). Predicting the accuracy of protein-ligand docking on homology models. *Journal of computational chemistry* **32**, 81-98.

Brylinski, M., and Skolnick, J. (2008). A threading-based method (FINDSITE) for

ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 129-134.

Brylinski, M., and Skolnick, J. (2009). FINDSITE: a threading-based approach to ligand homology modeling. *PLoS computational biology* **5**, e1000405.

Bryngelson, J.D., Onuchic, J.N., Socci, N.D., and Wolynes, P.G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167-195.

Campbell, S.J., Gold, N.D., Jackson, R.M., and Westhead, D.R. (2003). Ligand binding: functional site location, similarity and docking. *Current opinion in structural biology* **13**, 389-395.

Canutescu, A.A., Shelenkov, A.A., and Dunbrack, R.L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein science : a publication of the Protein Society* **12**, 2001-2014.

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.

Cheng, J.L., Sweredoski, M.J., and Baldi, P. (2006). DOMpro: Protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. *Data Min Knowl Disc* **13**, 1-10.

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

Gallo Cassarino, T., Bordoli, L., and Schwede, T. (2014). Assessment of ligand binding site predictions in CASP10. *Proteins* **82 Suppl 2**, 154-163.

Glusker, J.P. (1991). Structural aspects of metal liganding to functional groups in proteins. *Advances in protein chemistry* **42**, 1-76.

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. *Journal of computational chemistry* **29**, 701-715.

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

Heo, L., Shin, W.H., Lee, M.S., and Seok, C. (2014). GalaxySite: ligand-binding-site prediction by using molecular docking. *Nucleic acids research* **42**, W210-214.

Holm, R.H., Kennepohl, P., and Solomon, E.I. (1996). Structural and Functional Aspects of Metal Sites in Biology. *Chemical reviews* **96**, 2239-2314.

Huang, B., and Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC structural biology* **6**, 19.

Huang, Y.J., Mao, B., Aramini, J.M., and Montelione, G.T. (2014). Assessment of template-based protein structure predictions in CASP10. *Proteins* **82 Suppl 2**, 43-56.

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

Kasahara, K., Kinoshita, K., and Takagi, T. (2010). Ligand-binding site prediction of proteins based on known fragment-fragment interactions. *Bioinformatics* **26**, 1493-1499.

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 engineering* **9**, 1063-1065.

Khoury, G.A., Tamamis, P., Pinnaduwage, N., Smadbeck, J., Kieslich, C.A., and Floudas, C.A. (2014). Princeton_TIGRESS: protein geometry refinement using simulations and support vector machines. *Proteins* **82**, 794-814.

Kinch, L., Yong Shi, S., Cong, Q., Cheng, H., Liao, Y., and Grishin, N.V. (2011a). CASP9 assessment of free modeling target predictions. *Proteins* **79 Suppl 10**, 59-73.

Kinch, L.N., Shi, S., Cheng, H., Cong, Q., Pei, J., Mariani, V., Schwede, T., and Grishin, N.V. (2011b). CASP9 target classification. *Proteins* **79 Suppl 10**, 21-36.

Kinoshita, K., and Nakamura, H. (2003). Protein informatics towards function identification. *Current opinion in structural biology* **13**, 396-400.

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

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

Kortemme, T., Morozov, A.V., and Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of molecular biology* **326**, 1239-1259.

Kristiansen, K. (2004). Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. *Pharmacology & therapeutics* **103**, 21-80.

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., Fidelis, K., and Moult, J. (2011). CASP9 results compared to those of previous CASP experiments. *Proteins* **79 Suppl 10**, 196-207.

Kryshtafovych, A., Fidelis, K., and Moult, J. (2014). CASP10 results compared to those of previous CASP experiments. *Proteins* **82 Suppl 2**, 164-174.

Laurie, A.T., and Jackson, R.M. (2006). Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Current protein & peptide science* **7**, 395-406.

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

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W*., et al.* (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology* **487**, 545-574.

Lee, B., and Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology* **55**, 379-400.

Lee, H., Park, H., Ko, J., and Seok, C. (2013). GalaxyGemini: a web server for protein homo-oligomer structure prediction based on similarity. *Bioinformatics* **29**, 1078-1080.

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. *Journal of computational chemistry* **18**, 1222-1232.

Lee, J., and Seok, C. (2008). A statistical rescoring scheme for protein-ligand docking: Consideration of entropic effect. *Proteins* **70**, 1074-1083.

Leopold, P.E., Montal, M., and Onuchic, J.N. (1992). Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 8721-8725.

Lopez, G., Ezkurdia, I., and Tress, M.L. (2009). Assessment of ligand binding residue predictions in CASP8. *Proteins* **77 Suppl 9**, 138-146.

Lopez, G., Rojas, A., Tress, M., and Valencia, A. (2007). Assessment of predictions submitted for the CASP7 function prediction category. *Proteins* **69 Suppl 8**, 165-174.

Ma, J., Wang, S., Wang, Z., and Xu, J. (2014). MRFalign: protein homology detection through alignment of Markov random fields. *PLoS computational biology* **10**, e1003500.

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., 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. *The journal of physical chemistry B* **102**, 3586-3616.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG* **24**, 133-141.

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.

Mariani, V., Kiefer, F., Schmidt, T., Haas, J., and Schwede, T. (2011). Assessment

of template based protein structure predictions in CASP9. *Proteins* **79 Suppl 10**, 37-58.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure* **29**, 291-325.

Mirjalili, V., and Feig, M. (2013). Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. *Journal of chemical theory and computation* **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 Suppl 2**, 196-207.

Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., and Olson, A.J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of computational chemistry* **19**, 1639-1662.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* **247**, 536-540.

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

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

Oh, M., Joo, K., and Lee, J. (2009). Protein-binding site prediction based on three-dimensional protein modeling. *Proteins* **77 Suppl 9**, 152-156.

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.

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

Pei, J., Kim, B.H., and Grishin, N.V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research* **36**, 2295-2300.

Poupon, A., and Janin, J. (2010). Analysis and prediction of protein quaternary structure. *Methods in molecular biology* **609**, 349-364.

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.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E.*, et al.* (2000). Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-2309.

Rodrigues, J.P., Levitt, M., and Chopra, G. (2012). KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic acids research* **40**, W323-328.

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

Schmidt, T., Haas, J., Gallo Cassarino, T., and Schwede, T. (2011). Assessment of ligand-binding residue predictions in CASP9. *Proteins* **79 Suppl 10**, 126-136.

Shin, W.H., Heo, L., Lee, J., Ko, J., and Seok, C. (2011). LigDockCSA: protein-

ligand docking using conformational space annealing. *Journal of computational chemistry* **32**, 3226-3232.

Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G., *et al.* (2015). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research* **43**, D376-381.

Sim, J., Kim, S.Y., and Lee, J. (2005). PPRODO: prediction of protein domain boundaries using neural networks. *Proteins* **59**, 627-632.

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

Soding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **33**, W244-248.

Sotriffer, C., and Klebe, G. (2002). Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmaco* **57**, 243-251.

Tai, C.H., Bai, H., Taylor, T.J., and Lee, B. (2014). Assessment of template-free modeling in CASP10 and ROLL. *Proteins* **82 Suppl 2**, 57-83.

Tai, C.H., Lee, W.J., Vincent, J.J., and Lee, B. (2005). Evaluation of domain prediction in CASP6. *Proteins* **61 Suppl 7**, 183-192.

Taylor, T.J., Tai, C.H., Huang, Y.J., Block, J., Bai, H., Kryshtafovych, A., Montelione, G.T., and Lee, B. (2014). Definition and classification of evaluation units for CASP10. *Proteins* **82 Suppl 2**, 14-25.

Thornton, J.M., Todd, A.E., Milburn, D., Borkakoti, N., and Orengo, C.A. (2000). From structure to function: approaches and limitations. *Nature structural biology* **7 Suppl**, 991-994.

Tripathi, A., and Kellogg, G.E. (2010). A novel and efficient tool for locating and

characterizing protein cavities and binding sites. *Proteins* **78**, 825-842.

Tyka, M.D., Keedy, D.A., Andre, I., Dimaio, F., Song, Y., Richardson, D.C., Richardson, J.S., and Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of molecular biology* **405**, 607-618.

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

Wang, G., and Dunbrack, R.L., Jr. (2003). PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-1591.

Wass, M.N., and Sternberg, M.J. (2009). Prediction of ligand binding sites using homologous structures and conservation at CASP8. *Proteins* **77 Suppl 9**, 147-151.

Wu, S., and Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research* **35**, 3375-3382.

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. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 7432-7437.

Xu, D., Zhang, J., Roy, A., and Zhang, Y. (2011). Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins* **79 Suppl 10**, 147-160.

Xue, Z., Xu, D., Wang, Y., and Zhang, Y. (2013). ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* **29**, i247-256.

Yang, J., Roy, A., and Zhang, Y. (2013). BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic acids research* **41**, D1096-1103.

Yang, Y., Faraggi, E., Zhao, H., and Zhou, Y. (2011). Improving protein fold

recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **27**, 2076-2082.

Yang, Y., and Zhou, Y. (2008). 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 research* **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, Y. (2008). Progress and challenges in protein structure prediction. *Current opinion in structural biology* **18**, 342-348.

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 research* **33**, 2302-2309.

국문초록

많은 생물학적 과정들에 있어서 단백질은 중요한 역할을 수행한다. 단백질의 기능은 그 구조와 밀접한 관련이 있어서, 단백질의 구조를 결정함으로써 단백질 기능관련 연구에 중요한 기여를 할 수 있다. 단백질의 구조는 X-ray 결정학, 핵 자기공명 분광학, 전자 현미경 등의 다양한 실험적 방법을 통해 결정할 수 있다. 하지만, 이러한 방법들은 실험에 있어서 많은 시간과 비용이 소모되거나, 실험적인 한계에 의해 구조 결정이 힘든 경우도 있다. 이러한 점들을 극복하기 위해서 계산과학적 접근 방식을 통해 단백질의 구조를 예측하려는 연구가 많이 이루어져왔고, 이들 중 주형 단백질을 이용한 단백질 구조 예측 방법이 성공적이었다. 이러한 접근 방식에서는 기존에 알려져 있는 단백질의 구조를 주형으로 사용해 서열로부터 구조가 알려져 있지 않은 단백질의 구조예측을 수행한다. 이 방법의 요소들 중에서 주형으로 사용할 유사한 단백질의 구조를 찾는 과정이 가장 중요하다. 하지만, 주형 단백질의 구조적 정보만을 이용해서 모델링을 수행할 경우, 단백질의 서열이 주형 단백질과 다른 부분이나, 주형 단백질의 정보가 부족하거나 없는 부분에 있어서 정확한 모델링을 수행하기 힘들다. 이 논문에서는 기존의 생물정보학적 접근방식의 주형 탐색 방법과 물리화학적 접근방식의 단백질 모델링 방법을 결합한 새로운 단백질 구조 예측 방법을 소개하고 있다. 이러한 생물정보학과 물리화학적 접근 방식의 조합을 통해, 기존의 접근 방식들에서 얻기 힘들었던, 보다 고해상도의 단백질 구조 예측을 수행할 수 있었다.

주요어: 단백질 구조 예측, 주형기반 단백질 모델링, 단백질 구조 정밀화, 생물정보학, 물리화학

학 번: 2010-20300