博士論文

Theory and Applications of Computational Peptide Design

（計算機支援によるペプチド設計の理論と応用）

山岸 純也

Contents

Overview

Peptides play key roles in many biological processes. For example, the antigenic peptides are presented on the surface of the cell by the major histocompatibility complex (MHC) molecule and induce the immunological response. Peptides also work as hormones and transport many kinds of signals to target cells. Moreover, peptides cause a kind of disease known as amyloidosis, where peptides become insoluble and are deposited in organs. All physiological phenomena mentioned above are involved in peptide-protein (or peptide-peptide) interactions. It means that regulating specific peptide-protein interactions using an artificial high-affinity peptide would result in the control of the specific biological process. Angiotensin II receptor blocker (ARB) is one of examples based on this idea. ARB prevents the binding of angiotensin II to its receptor proteins, resulting in the reduction of the blood pressure. Peptide itself is also a potent inhibitor of peptide-protein (or protein-protein) interactions. Phan et al. succeeded to design 12-mer peptides binding to MDM2 and MDMX with high affinities [1]. Peptide inhibitors against various Src Homology 2 (SH2) domains were also identified [2-4]. Recently, therapeutics targeting protein-protein interactions are expected to fulfill unmet medical needs. There are increasing interests in rational design techniques of high affinity peptides.

Bioinformatic approaches have been widely used for the prediction of the amino-acid sequence of the binding peptide especially for the MHC class I and II molecules [5-7]. In these approaches, some scoring functions (such as a score matrix) were constructed based on known amino-acid sequences of binding peptides. These sequence-based approaches are useful in reducing candidate sequences in a short time, but the reliability highly depends on the quality and the quantity of the available experimental data. Thus, these applications are limited to well-known protein targets.

A structure-based approach is also a rational technique to design binding molecules. This approach has received strong attentions as increasing the number of information concerning three-dimensional (3D) structures of biomolecules. Today, numerous

techniques, including molecular simulations, are available for utilizing the 3D structures of biomolecules. Structure-based molecular design techniques use the 3D structures of the complex of receptor proteins and the ligand molecule to predict binding affinities. Because structure-based approaches do not require experimental data concerning known binding molecules, they are expected to be applicable to a wide variety of therapeutic targets.

In this study, we discuss the structure-based design of peptides based on molecular docking. Structure-based molecular design consists of several stages according to their computational costs. Molecular docking is used in the early stage of the molecular design because of its computational efficiency. Therefore, molecular docking is used with thousands of compounds in order to discriminate binders from non-binders. Detailed binding affinities will be further investigated in the next stages.

Molecular docking has two main purposes: predictions of the binding conformations and of the binding affinities. The important purpose of molecular docking is to predict binding conformations of the ligand molecule to its receptor proteins. Programs for molecular docking generate numerous conformations of the ligand molecule and judge them using a scoring function, called docking score. Docking scores are also used as the binding affinities in order to rank ligand molecules. Sometimes, other scoring functions are used for re-evaluating the binding affinities using binding conformations predicted by molecular docking (known as rescoring).

Molecular dockings have been widely used in traditional drug discoveries. They have been supposed to be used with drug-like small molecules. Because peptides have different characteristics from drug-like small molecules, conventional molecular docking cannot be applied to the peptide design. In chapter 1 of this study, we demonstrate the inability of conventional programs to predict correct binding conformations of peptides. To solve this problem, we developed a program for molecular docking of peptide. We incorporated the potential energy function of molecular mechanics and an implicit solvent model to

our scoring function. We used a GPGPU technology [8] to accelerate computations of our molecular docking program. We show performances of our program on predictions of binding conformations of peptides using various peptide-protein complexes.

Conventional programs for molecular docking have inabilities to predict not only the binding conformations but also the binding affinities of peptides. We demonstrate those in chapter 3. We tried to solve this problem by applying Molecular Mechanics and Poisson Boltzmann Surface Area (MM-PBSA) method as rescoring of the binding affinities [9-11]. However, Poisson Boltzmann (PB) implicit solvent, which is used in MM-PBSA, has low estimation accuracy of the polar contribution of the solvation free energy. In chapter 2, we improve the accuracy of PB by modifying PB radii, which are important parameters for PB calculations. Our PB radii gave high performances on estimations of both the solvation free energies of single molecules and the binding affinities of peptide-ligands predicted by MM-PBSA.

In chapter 1 and 2, we improved the estimation accuracies of the binding conformations and the binding affinities of peptides to their receptor proteins. In chapter 3, we combined and applied our improved methods to *in silico* screening of peptides. We measured the performances of our method on discriminating the binding peptides of several SH2 domains from a small set of peptides. In this chapter, we also examined the effect of the reorganization of ligand molecules on the performances of molecular-docking based approaches. Our results provide useful information for more large-scale screening of peptides.

Chapter 1
Program for
Molecular Docking
with GPU Acceleration

1.1. Introduction

As increasing the number of three-dimensional (3D) structures of biomolecules, many computational techniques have been developed to utilize these 3D structural data. Structure-based drug design (SBDD) is one of these techniques to design binding molecules using the 3D structures of the target proteins.

The ligand-based drug design (LBDD) techniques, such as a structure-activity relationship (SAR), highly depends on the quality and the quantity of the available experimental data. Thus, their applications were limited to the known therapeutic targets. Moreover, LBDD is not suited to find molecules having different scaffolds from known binding molecules.

SBDD predicts the binding affinities using the complex structures of the ligand molecules and the receptor proteins. The complex structures are usually predicted by molecular simulations, called molecular docking. Molecular docking predicts the binding conformations of ligand molecules to their target protein according to the score function, called docking score. Docking scores are also used as the binding affinities to rank ligand molecules. Because the computational time of molecular docking is very short, molecular docking-based approaches are used with thousands of compounds in chemical databases.

Conventional docking programs have been supposed to be used with drug-like small molecules. It is reasonable because many studies of the traditional drug discovery have devoted to find the small molecules binding to their target proteins. Recently, biomolecular drugs, such as vaccines, hormones, and antibodies, are also known to be effective in many types of diseases. Some of these diseases are involved in protein-protein interactions. Due to the large contact interface of protein-protein interactions, small molecules, whose typical molecular weight is smaller than 500 Da, is not suitable for inhibition of the binding between proteins. Instead of small molecules, larger amino-acid based molecules, peptides or small proteins, have received considerable attentions in recent years.

We can classify these amino-acid based molecules into two types: molecules having stable secondary structures of proteins and those having no stable secondary structures of proteins. The former molecules maintain their secondary structures through their bindings to receptor proteins. The same behaviors are expected in molecular docking: most part of the conformations of peptides will be kept during the simulation. To apply molecular docking to these peptides, an alternative docking algorithm has been developed and known as protein docking [12-14]. On the other hand, the other type of molecules has no stable secondary structures of proteins. They are expected to change their conformations through binding to their target proteins as small molecules do. Therefore, similar algorithms to conventional molecular docking can be applied to these peptides. However, peptides have different characteristics from small molecules. It is necessary to take into account the characteristics of peptides to molecular docking.

Peptides have many rotatable bonds and many polar functional groups. Many rotatable bonds enlarge the conformational search space in molecular docking. It results in the increase of the number of the evaluations of binding conformations until getting optimal binding conformations, as compared with drug-like small molecules.

For practical use, it may be necessary to reduce the conformational search space artificially by adding some positional restraints on several atoms of the ligand molecule. For example, positional restraints based on backbone atoms are efficient if the reference structure of the protein-peptide complex is available. Thus, it is preferred to be able to set up flexible positional restraints easily into simulations.

The second characteristic of peptides is many polar functional groups that affect the scoring function of molecular docking. As described in the proposal by Lipinski [15], drug-like small molecules are lipophilic and contain a few polar functional groups. For the drug-like small molecules, a few polar interactions, such as hydrogen bonds, between the ligand- and the receptor atoms are dominant in the formation of the ligand-receptor complex. On the other hand, peptides have many polar functional groups. Peptides can

form many polar interactions with not only receptor atoms but also solvent atoms. It is crucial to take into account the solvation effect in the scoring function for peptide docking. Implicit solvents can estimate the solvation free energy of the solute without explicit conformational sampling of water molecules. However, the computational cost of implicit solvents is still high for molecular docking.

In this study, we developed our program for molecular docking. We incorporated the implicit solvent into the scoring function. We attempted to solve the problem of the high computational cost of the implicit solvent by accelerating the computations using general-purpose computing on graphics processing units (GPGPU) technology. GPGPU utilizes an extreme computational power of GPU for general-purpose computations, not only for image processing. GPU consists of many computing units, and we can consider it as a parallel machine for single instruction, multiple data (SIMD) operations. To bring out the full performance of GPU, computational algorithms must be highly optimized for SIMD parallel processing.

Molecular docking is highly suited to SIMD computing. Procedures of molecular docking consist of two parts: the generation and the evaluation of the binding conformations. In the generation of binding conformations, the program enumerates candidate-conformations of the ligand molecule in a binding site of the receptor protein. In the evaluation of binding conformations, docking scores of each conformation are calculated from the scoring function. These two procedures are iterated until docking scores are well converged. In the two parts, the procedure for the evaluation of the binding conformations constitutes a large portion of the total computational cost. Therefore, it is reasonable to accelerate this part of calculations. This procedure can be easily parallelizable because the evaluation of each conformation is data-independent on each other. It enables to process each evaluation in parallel. Furthermore, computations of the scoring function are also easily parallelizable. A large part of the computational cost in the scoring function is attributed to the calculation of pairwise interactions between atoms

in the ligand-receptor complex molecules. Each pairwise interaction is also data-independent on each other and can be calculated in parallel.

For these reasons, molecular docking is highly suited to parallel computations using GPU. It is expected to reduce the computational time significantly and to enable us to apply molecular docking to peptides in a practical time scale.

In this chapter, we first describe the specification of our program and how we accelerate the computations of molecular docking using GPU. In following sections, we measured the performance of our program in the respect of the computational time and the estimation-accuracy of binding conformations using several proteins.

1.2. Theory and Specification

We developed our program for molecular docking using the GPU acceleration. The performance of molecular docking is highly dependent on the scoring function and the algorithm for the generation of binding conformations. First, we describe these two features implemented in our program. Next, we give the brief explanation of each procedure of our program.

1.2.1. Scoring Functions

We described that it is important to take into account the solvation effect of the solute in molecular docking of peptides. We incorporated the generalized born implicit solvent (GB) into our scoring function [16, 17], which is commonly used in molecular simulations. GB estimates the polar contribution of the solvation free energy (G_{GB}), and it is familiar with the potential energy function (Force Field) of Molecular Mechanics (MM). Our scoring function *DS* is represented as follows:

$$DS = V_{MM} + G_{GB} + V_{restraint}$$

where V_{MM} is the potential energy function of AMBER force field (ff99SB) [18], and $V_{restraint}$ is a user-defined harmonic penalty described below. The standard form of the V_{MM} is represented as follows:

$$
\begin{aligned}
V_{MM} = & \sum_{bonds} k_R\left(R - R_{eq}\right)^2 \\
& + \sum_{angles} k_\theta\left(\theta - \theta_{eq}\right)^2 \\
& + \sum_{dihedrals} \frac{k_\phi}{2}\{1 + \cos(n\phi - \gamma)\} \\
& + \sum_{\substack{all\ atom \\ pairs(i,j)}} \epsilon_{ij}\left[\left(\frac{r_{eq}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{eq}}{r_{ij}}\right)^{6}\right] \\
& + \sum_{\substack{all\ atom \\ pairs(i,j)}} \frac{q_i q_j}{\epsilon r_{ij}}
\end{aligned}
$$

where k_R is a bond force constant; R, a bond-distance; R_{eq}, an equilibrated bond-distance, k_θ, an angle force constant; θ, the angle; θ_{eq}, an equilibrated angle; k_ϕ, a dihedral force constant; n, a multiplicity of the dihedral function; ϕ, a dihedral angle; γ, a phase shift; ϵ_{ij}, a force constant for the Lennard-Jones (LJ) potential; r_{eq}, an equilibrated LJ distance; r, a distance between atom i and j; q, a partial charge; ϵ, a relative dielectric coefficient. MM potential energy function includes bond, angle, dihedral, van-der Waals (Lennard-Jones potential), and electrostatic (coulomb potential) terms. In our docking program, the effects from bond-stretching and angle-bending are neglected (fixed during the simulation) in order to simplify the problem.

In the Lennard-Jones potential function, the 12th order of the repulsive term is ordinary used due to a good approximation for the Pauli repulsion and a computational efficiency. In the docking simulation, it is better to use more soft repulsion term because it is difficult to generate conformations of the ligand molecule without any steric crashes. We used 8th order instead of 12th one. Fourth term of the equation above is replaced by $V_{LJ8\text{-}6}$:

$$V_{LJ8-6} = \sum_{\substack{all\ atom \\ pairs(i,j)}} \epsilon_{ij} \left[3\left(\frac{r_{eq}}{r_{ij}}\right)^8 - 4\left(\frac{r_{eq}}{r_{ij}}\right)^6 \right]$$

where the coefficients were determined to have the same depth of the potential well at the same distance as those of 12th order (Figure 1.1).

Additionally, we used scaled distance for non-bonded interactions to ease serious steric crashes. Scaled distance r'_{ij} was determined based on the equilibrated distance r_{eq} of Lennard-Jones potential of each atom pair. r'_{ij} was represented as follows:

$$r'_{ij} = \begin{cases} \dfrac{\beta - \alpha}{\beta^2 r_{eq}} {r_{ij}}^2 + \alpha r_{eq} & (r_{ij} < \beta r_{eq}) \\ r_{ij} & (r_{ij} \geq \beta r_{eq}) \end{cases}$$

where α and β are scaled factors. We used 0.45 and 0.55 for α and β, respectively.

Positional Restraints

Positional restraint is a reasonable solution to reduce the conformational search space in the molecular docking of peptides. In our program, two types of positional restraints are available and easy-to-use. One option is to fix a fragment of the ligand molecule at their input positions. It is effective if a portion of the ligand molecule works like an anchor. Another option adds the harmonic penalty ($V_{restraint}$) according to the position of the specified atoms of the ligand molecule. The harmonic penalty $V_{restraint}$ is represented as follows:

$$V_{restraint} = \begin{cases} 0 & (r \leq r_{cutoff}) \\ k(r - r_{cutoff})^2 & (r > r_{cutoff}) \end{cases}$$

where k is the force constant, and r is the distance from the reference position. The harmonic penalty is added only if r is longer than r_{cutoff}. User can set all of parameters for the positional restraint arbitrary.

1.2.2. Genetic Algorithm

We used the genetic algorithm (GA) to optimize binding conformations of the ligand molecule in the binding site of the receptor protein.

In our program, the conformation of the ligand molecule is represented in the two manners: the Z-matrix and the Cartesian coordinate. The Z-matrix represents the position of the atom as a relative position to other atoms. The position of the atom is determined by the list of the bond-length, the bond-angle, and the dihedral angles. Furthermore, additional six parameters are used to determine the relative orientations between molecules. In our program, binding conformations are represented in the form of the Z-matrix because it can handle the conformational change of the molecule easier than the Cartesian coordinate. On the other hand, the Cartesian coordinate is the appropriate description for handling the non-bonded interactions between separated atoms. Therefore, it is used in the evaluation of the binding conformations: the scoring function are calculated using the Cartesian coordinate which are converted from Z-matrix.

In our GA, new conformations are constructed by applying genetic operators to existing conformations: the single-point crossover and the single-point mutation. The default ratio for the crossover and mutation operator is 0.7 and 0.3, respectively. Partners of the crossover operator are limited to similar conformations of each conformation, which are determined by a pairwise Root Mean Square Deviation (RMSD) matrix between every conformation. It results in the reduction of the probability of trapping in a local minima in the conformational search space. In default, genetic operators generate five new conformations (called child-conformations) from each conformation (called parent-conformation). The Best conformations in each parent and its child-conformations become new parent-conformation in the next iteration (the survival stage of GA).

1.2.3. Procedures of Our Docking Program

Figure 1.2 is a flowchart of the docking procedures implemented in our program. In this subsection, we will give brief explanations of each procedure.

- **Preparation**

First, input structures of the ligand and the receptor molecules are read in the PDB format [19]. Because the conformation of the molecules are represented by the Cartesian coordinate in the PDB format, the Z-matrix of the ligand molecule are built from the Cartesian coordinate. The parameters for the scoring function (the force field and positional restraints) are set in this procedure.

- **Initial Pose Generation**

All conformations in the first iteration are generated by random numbers. The default population is 3,000.

- **Conversion of 3D Coordinate Systems**

The representation of the 3D coordinate system of each conformation are translated to the Cartesian coordinate from the Z-matrix. We used self-normalizing natural extension reference frame method for this conversion[20].

- **Pose Evaluation**

Figure 1.3 illustrates detailed procedures for the evaluation of the binding conformation. Before the calculation of the scoring function, we prepared a distance matrix between all pairs of atoms. This helps to reduce the computational cost because most of calculations of distances are duplicated, however the GPU code skips this procedure.

The conformation of the receptor molecules is fixed during the simulation of molecular docking. This approximation enables to reduce a large portion of the computation of docking score. Our code calculates the intramolecular- and the intermolecular contributions of non-bonded interactions separately. This treatment made our codes simple and easy-to-read.

- **Selection for Survival**

For the first iteration, conformations having top docking scores in all conformations are selected as parent-conformations of the next iteration. For the second or later iteration, the best conformations in each family, which is formed from each parent- and its child-conformations, are selected as new parent-conformations of the next iteration.

- **Convergence Test**

This procedure measures the replacement rate of the parent-conformations by new ones in next iteration. If the rate is under than the criterion, the optimization are considered to be well converged.

- **Next Pose Generation**

If the optimization is not converged, GA generates new conformations using genetic operators. First, the distance matrix between all pairs of conformations are calculated to determine the partners of the crossover operation in GA. Each parent-conformation produces several (five in default) new child-conformations using genetic operators.

- **Clustering and Output**

If the optimization is well converged, clustering analysis is performed using all parent-conformations of the final iteration. Representative conformations of each cluster are written in the PDB format.

1.2.4. GPU Acceleration

We developed our program using CUDA programming environment [8]. In this subsection, we describe our GPU computing that accelerates many calculation parts of the molecular docking. At first, we summarize CUDA programming model. Next, we describe our GPU computing.

- **CUDA Programming Model**

Thread Hierarchy

A thread is a basic execution unit in the CUDA programing model. Threads are grouped into thread-blocks, and thread-blocks are grouped into a grid. A function executed on GPU is called a kernel function. A kernel function is executed on all threads in parallel. Only threads in the same thread-block can synchronize. All thread-blocks are distributed to multiprocessors of GPU. Multiprocessors can execute the kernel function on several thread-blocks concurrently, but there are severe memory restrictions to increase the number of concurrently running threads. Because the resource per one multiprocessor is limited, we have to reduce the usage of the resource per thread or per thread-block for the rapid computing. The maximum number of the concurrently running threads is 1536 and 2048 for Fermi and Kepler architecture, respectively.

Memory Hierarchy

CUDA provides several types of the memory and they have different properties to use efficiently. We will summarize those as follows:

Register is the fastest on-chip memory. All variables declared in the kernel function generally resides in the registers. Only threads can read or write registers. The number of

registers per multiprocessor is 32K and 64K for Fermi and Kepler architecture, respectively. Therefore, to maximize the number of the concurrently running threads, the number of registers per thread must be under 20 and 31 for Fermi and Kepler, respectively. As increasing the number of required resisters per thread, the number of the concurrently running threads per multiprocessor decreased.

Shared memory is as fast as the register if an optimal memory access is achieved. Shared memory is divided into 32 banks (each bank is 32-bit word). If all threads access different banks concurrently, threads can read or write the memory as fast as the resisters. By contrast, if several threads access the same bank, these accesses are serialized (the one exception is that the access to the same address can be broadcasted to threads simultaneously). User can determine the amount of the shared memory per multiprocessor from 16 to 48 KB. This is shared by thread-blocks executed on the same multiprocessor concurrently. Therefore, the amount of the shared memory per thread-block determines the number of thread-blocks executed on single multiprocessor concurrently. Only threads can read or write shared memory.

Global memory is only memory which both CPU and GPU can read and write. Access to the global memory is slower several hundred times than that of registers even if the optimal memory accesses are achieved. Optimal accesses, known as the coalesced access, are required for practical use: the *k*-th thread accesses the *k*-th word of the array in the global memory. Some of the memory latencies can be hidden by any arithmetic operation on different threads. Cache for read from the global memory is available.

Constant memory is the read-only cache and suitable for broadcasting. The constant memory is as fast as the register if all threads read the same address. The size of the constant memory is limited to 64 KB.

Our GPU programming

- **Preparation for Docking Simulation**

Before the beginning of the optimization process of GA, some preparations for GPU computing are required: memory spaces are allocated on the global memory, and some parameters are transferred to the global memory and the constant memory in advance. This process requires an extra computational time, but it is a little.

- **Conversion of 3D-Coordinate System**

We tried to accelerate the function for the conversion of the 3D-coordinate system. Due to the data-independence of the conversions of each conformation, we can execute these calculations in parallel. However, in the Z-matrix representation, the positions of atoms are described as the relative positions to other atoms. Therefore, it is required to determine the Cartesian coordinates of atoms sequentially in the order listed in Z-matrix. For these reasons, one thread calculates the Cartesian coordinates of all atoms in the one conformation sequentially. The threads in the same thread-block calculates the Cartesian coordinate of the same atom in the different conformation at the same time. The number of threads is the same as that of conformations.

Before a kernel call, the list of parameters regarding dihedral angles and the relative orientation to the receptor molecules are transferred to the global memory. They are read by each thread in the coalesced manner. The parameters for Z-matrix (1^{st}, 2^{nd}, and 3^{rd} indexes of the reference atoms, the bond-length, and the bond-angles) are loaded from the constant memory. The Cartesian coordinates of each conformations are saved to the global memory and they are used in following functions.

- **Calculation of dihedral potential energy**

In this kernel function, one thread calculates all dihedral-angle potential energies of

one conformation. The threads in the same thread-block calculates the potential energy of the same dihedral angles in different conformations at the same time. All parameters for dihedral-angle terms (four atom indexes, a force constant, etc.) are read from the constant memory. Results are saved to the global memory and transferred to CPU immediately.

- **Realignment of Memory containing Cartesian Coordinates**

At this point, the alignment of the Cartesian coordinates is suited for the case that each thread reads the coordinate of the same atom but in different conformations at the same time. In the following functions, one thread-block handles one conformation i.e. each thread in the same thread-block reads the coordinate of different atoms in the same conformation at the same time. To achieve the optimal memory access from the global memory in the following function, the memory-realignment are needed for the array containing the Cartesian coordinates. This realignment corresponds the transposition of the matrix, and the optimal implementation on GPU is well known (included in the sample code of a CUDA toolkit).

After this realignment, the Cartesian coordinates of all conformations are transferred to CPU. Computational cost of this function is not discussed because it is too low.

- **Calculation of non-bonded interactions**

Calculations of the non-bonded term of MM on GPU are also divided into several functions in the similar manner to CPU codes. At first, the effective born radii of molecules are calculated, and next the potential energy of MM and GB are calculated. Both calculations can be represented as the sum of each pairwise contribution of all pairs of atoms. Because these pairwise interactions are data-independent on each other, they are easily parallelizable on GPU. Here, each thread in the thread-block associates with one particular atom in one particular conformation, and calculates non-bonded

interactions with other atoms in the conformation. All threads in the same thread-block calculates the interactions with the same atom at the same time. This permits to access to the same addresses in the shared memory and the constant memory from all threads.

There is a slight difference between the CPU and the GPU code. In the CPU code, the distances between all atom pairs are calculated in advance, and used these values several times. On the other hand, GPU cannot retain such a large amount of data and requires to calculate the distances each every time. It results in additional computational cost for GPU, but the high computational power of GPU overcomes such a weak point.

- **Make Neighbor List**

In this function, each thread in the thread-block associates with one particular atom in the one particular conformation. Each thread reads the coordinate of the associated atom in different conformations, and calculates the distance between two atoms. Results of each thread are summarized on the thread-block as the RMSD value between two conformations.

1.2.5. Conformational Search of Single Molecule

This program can be used for the conformational search of single molecule. Actually, our program was used to predict stable conformations of peptides in the unbound state in chapter 3.

1.3. Methods

1.3.1. Comparison of Computational Time

We compared the computational time of our program executed on GPU with those on CPU-only. We measured the computational time of each calculation part separately. We used the computational time for the first iteration of GA optimization. The number of parent-conformations was 3,000 and the number of child-conformations per parent conformations was 10, i.e., the computational time for the calculations of the docking score of 30,000 conformations and the calculation of RMSDs for 4,498,500 pairs of conformations were measured. All computational times were measured 10 times and averaged.

Benchmark tests were employed on a PC with Intel® Core™ i7 4770K (3.4GHz, Haswell architecture) and two GPUs with different architectures: nVidia GeForce® GTX 580 (Fermi) and GTX 780 (Kepler). The CPU-only code was executed using single core of CPU. Considering the situation where the molecular docking is carried out, it is more efficient to execute multiple programs with different ligand molecules on different CPU cores than to accelerate one program by several cores of CPUs. Namely, we can consider that the parallelization efficiency of the program for the molecular docking is 100 %. The CPU codes were compiled using Intel® C++ Compiler version 13.1.1 with the optimization option (-O3 –xHOST –no-prec-div). The GPU codes were compiled by *nvcc* installed in CUDA 5.5 with the optimization option (-O3 –use_fast_math). The gettimeday() function in C language was used to measure elapsed times on CPU codes. The functions managing the CUDA event were used for GPU codes.

We used two peptide-protein complexes to compare the computational times: a complex of Grb2 SH2 domain with its 8-mer binding peptide (PDB: 1TZE) [21] and GIP PDZ domain with its 8-mer binding peptide [22]. The number of the receptor atoms of GIP PDZ domain is 1.2 times larger than those of Grb2 SH2 domain, while the number

of atoms of ligand molecules are almost same. Structural preparation were carried out as follows: three-dimensional structures were downloaded from the Protein Data Bank [19]. All protonation states of the solutes were determined by the Protonate3D module of MOE [23]. Energy minimizations were carried out in the box of TIP3P waters using the sander module of AMBER11 [24]. Then, all waters and ions were removed.

1.3.2. Comparison of Prediction Accuracy of Binding Conformations

We compared the prediction accuracies of the binding conformations of the ligand molecules to their receptor proteins between our program and a widely-used program, GOLD [25]. Here, self-docking was performed using four protein-ligand complexes: Crk SH2 domain (PDB: 1JU5) [26], Grb2 SH2 domain [27], Src SH2 domain [28], and GIP PDZ domain (same as above). All lengths of the binding peptides of SH2 domains were adjusted to 8-mer (XXpYXXXXX: where pY denotes the phosphorylated-tyrosine and X denotes any amino acids) in order to maintain consistency with the next chapter. The 3D-structures of each protein were downloaded from the PDB web site. All protonation states of solutes were determined by the Protonate3D module of MOE. The solutes were soaked in the box of TIP3P waters and energy-minimizations were employed using the sander module of AMBER11. Then, energy-minimized structures were used for self-docking (named "min"). In addition, we prepared another structure of each complex structure using molecular dynamics (MD) simulations. We performed 2 ns MD simulations at 310K for the equilibration. The equilibrated structures were also used for self-docking (named "MD").

The condition for our molecular docking follows: the number of parent-conformations was 3,000 and the number of the child-conformations per parent-conformation was 10. Optimizations by GA were iterated until the replacement rate of the conformations was under 15 %. No cut-off schemes was employed for the non-bonded

interactions.

For GOLD, the search efficiency was set to 200%. The option for the early termination was off, and the search radius was 20 Å. ChemPLP [29] was used as the scoring function. Other parameters were set as default.

We also examined the efficacy of the positional restraint on molecular docking of peptides. In our program, positional restraints were added on every Cα atom of all residues of each ligand molecule. Moreover, we added positional restraints on the phosphorus atom of the phosphorylated tyrosine of SH2-binding peptides and the nitrogen atom of the side chain of the lysine for GIP-binding peptide, because these atoms form the strong ionic interactions with the receptor atoms. The phosphorus atom in SH2-peptide was fixed, while the nitrogen atom was harmonically restrained at the reference position. Positional restraints were added if the distance from the reference position was longer than 2.0 Å. The force constant was 10.0 kcal/mol/$Å^2$. In GOLD, similar restraints were archived by Region Constraints. In this algorithm, extra user-defined score (named weight) was added to the docking score if the specified atom located within the specified distance (named radius) from the reference position. In this study, we examined several weights of 10, 20, and 30, and set the radius of 2.0 Å. The molecular dockings with no positional restraints were also performed by both programs.

1.4. Results

1.4.1. Comparison of Computational Time

Computational times of each calculation part are listed in Table 1-1 and Table 1-2 for Grb2 and GIP, respectively. Our GPU codes could calculate more than 100 times faster than CPU code. Moreover, GTX780 was about 1.7 times faster than GTX580 as the overall performance. In every calculation part, GTX780 was the fastest and single core of CPU consumed the longest computational times. The total computational time executed on GPU for Grb2 was about 2,920 ms and 1,484 ms for GTX580 and GTX780, respectively. Computational time executed on CPU in the GPU code was similar between GTX580 and GTX780, and it was about 550 ms. The computational time for the memory allocations and transfers was under 20 ms per one iteration of GA.

The computational times of each calculation part increased in almost direct proportion to the number of interactions in the peptide-protein complex. Because the number of atoms of the ligand molecules in two complexes was almost the same, computational times involved with only ligand molecule were similar in the two simulations. By contrast, the number of atoms of the receptor molecule in GIP was 1.2 times as many as those of Grb2. This influence was shown in the computational times involved with receptor atoms. The functions handling the intermolecular interactions of GIP consumed 1.2 times longer than those for Grb2. The function handling the intramolecular non-bonded energy within the receptor atoms consumed about 1.4 (1.2 * 1.2) times longer computational time for GIP than those of Grb2.

1.4.2. Comparison of Prediction Accuracy of Binding Conformations

Table 1-3 lists ligand RMSDs of the best solutions from each molecular docking using two programs. In every case, both programs could not predict the correct binding conformations without any positional restraints. On the other hand, both programs could

achieve lower RMSD values by applying positional restraints. In GOLD, the constraint weight of 30 was required to accomplish the good predictions in all protein-peptide complexes. In this case, the contribution from the constraint term in the total docking score raised to three times larger than the native docking score in average.

In three of four cases, the uses of equilibrated structures from MD simulations improved the prediction accuracies of the binding conformations. Although RMSD values for Grb2 became worse by using the equilibrated structures, the conformations of the central four residues of the binding peptide, which are known as the core binding motif of SH2-binding peptides (pYxNx), were predicted correctly. Both the N- and C-terminal portions of the binding peptide were exposed to the solvents, and their conformations were highly fluctuated in the MD simulation.

1.4.3. Total Computational Time

Total computational times of each program with each protein were measured. The averaged total computational times of our program for Crk, Grb2, Src, and GIP were 3m15s, 3m56s, 3m55s, and 6m50s, respectively. Those of GOLD were 1m55s, 1m58s, 2m7s, and 2m9s, respectively. Our averaged computational times were varied depending on proteins, however GOLD were similar to each other. Our computational times were about 1.5 - 3.5 times longer than those of GOLD.

The computational times of our program were fluctuated in several runs. It was caused by the varied number of the iterations of GA optimizations in each run. Because our program generates initial conformations using random numbers, the degree of the convergence of the docking scores are varied in each run. By contrast, GOLD fixed the number of the iterations for optimizations. This resulted in similar computational times in every run.

1.4.4. Single Precision of Floating Points

GPU can process the operations using single precision floating points several times faster than those using double precision floating points. We investigated the influence from single precision floating points on the docking score by comparing calculated values by GPU (single precision) and by CPU (double precision). As a result, there were only slight differences on total docking score. The error was under 0.001 kcal/mol in most cases. In the molecular docking, docking scores were used only to compare with those of other conformations, and these small errors do not influence the comparisons of docking scores. For this reason, we concluded that operations using single precision floating points are not problematic in the molecular docking.

1.5. Discussions

1.5.1. GPU Architectures

Although the ratio of theoretical arithmetic capacity of GTX780 to those of GTX580 is about 2.7, the performance on GTX780 was about 3 times better than those on GTX 580 in two calculation parts. It resulted from not only the arithmetic capacities but also the difference of the architectures between the two GPUs: Fermi and Kepler. Kepler is the newer architecture of nVidia GPUs. Kepler not only has the improved arithmetic capacity, but also eases the restriction on the memory usage to utilize the full performance of GPU. One of significant differences between two architectures is in the restriction for the number of registers per thread. A remarkable example is the calculation part of the intermolecular non-bonded energy term of MM. This function requires many parameters for MM and GB calculations like partial charges, equilibrated distances for Lennard-Jones potential, etc. This increases the number of registers required for each thread: 45 registers were needed for one thread to execute this kernel function. Under this restriction, only 37.5% of threads per multiprocessor can be executed concurrently on Fermi GPU. On the other hand, 56.2% of threads per multiprocessor can be executed concurrently on Kepler GPU. The number of the registers usually becomes the bottleneck for utilizing the full performance of GPU. Improvements on the number of available threads were observed in most kernel functions, and contributed to the accelerations beyond the ratio of the arithmetic capacities of the two GPUs.

1.5.2. Prediction Accuracy of Binding Conformations

Positional restraints significantly impacted on prediction accuracies of the binding conformations for the two programs, but implementations of the positional restraints are different in them. For our program, native docking scores (the docking score excluding a restraint term) were improved compared to those with no positional restraints. The

positional restraint acted like a guide, leading binding conformations to more stable ones. This result indicated that the search space occupied by the stable conformations was very small under our scoring function. Our program cannot find this space alone, but positional restraints can help it.

By contrast, only subtle difference in native docking scores with or without positional restraints was observed in GOLD. For GOLD, the insufficient searching ability was not only the reason to predict the conformations having higher RMSD values. GOLD cannot distinguish the correct binding conformations in terms of the docking score. ChemPLP evaluates the binding conformations by the shape complementary and the formation of hydrogen bonds between the ligand and the receptor atoms. The shape complementary can be a good indicator for small molecules, because they have low internal degrees of the conformational freedom. It is difficult for small molecules to fit their conformations into the shape of the binding pocket of proteins. On the other hand, the high conformational flexibility of peptides permits to change their conformations to fit in anywhere in the binding site of proteins. In addition, the high conformational flexibility of peptides also allows to form many hydrogen bonds with receptor atoms. The scoring functions designed for small molecules may not make significant differences between binding conformations of peptides.

We examined the binding conformations predicted using two programs by means of the rescoring of the binding affinities using MM energy functions. Table 1-4 shows the binding affinities of the top solutions of each docking run calculated by MM-GB (not including SA term) method. Obviously, our program predicted more stable conformations than GOLD. This result indicated that our program could detect the key interactions in the ligand-protein complex properly, but GOLD cannot. MM-based rescoring schemes were often used as a post process of the molecular docking [9, 30]. In this situation, the problems for selecting binding poses predicted by the molecular docking would arise. Typical programs for the molecular docking output several binding conformations of the

ligand molecule in each execution. Because there was no consistency between the docking scores and MM-based binding affinities, all binding conformations predicted by the molecular docking have to be evaluated by MM-based method to find the most stable conformations in terms of the MM. On the other hand, our program uses the potential energy of MM and GB as the scoring function in the molecular docking. This is the great advantage for our program to be able to find the stable conformations in terms of MM in the molecular docking, and to get the consistent results with the rescoring scheme.

1.5.3. Computational Cost

We showed the high performance of our program to predict stable conformations in previous subsections. However, our computational time was too long, though extreme acceleration by GPU has already been accomplished. The computational cost for intramolecular non-bonded energy term within receptor atoms is especially high, though these interactions are not calculated in ordinary docking program. However, we have to calculate them because GB was incorporated to our scoring function: the effective born radii of receptor atoms are changed according to the conformation of ligand molecule, and it results in the change of the GB energy of intramolecular interactions within receptor atoms.

We could confirm the superiority of our scoring function in previous subsections. We have to make more efforts to reduce computational time while keeping the prediction accuracy of our scoring function.

One approach is to neglect interactions between the ligand atoms and the receptor atoms far from the any ligand atoms. This is the similar approach to the conventional programs. We examined this approach using the GIP-peptide complex. At that time, only receptor atoms within 15 Å from any ligand atoms were used as an input structure. It resulted in computational time of 4m6s and retained the RMSD of 2.31 Å. This approach seems to be useful, but the influence from the receptor atoms unnaturally exposed to

solvents remains unknown. Another approach is more reasonable to exclude the interactions between the ligand atoms and distant receptor atoms from the calculations of the effective born radii, instead of removing the distant receptor atoms. This classifies the receptor atoms into two layers according to the distance from the binding interface. This treatment can neglect the calculations of effective born radii between ligand atoms and distant receptor atoms, and the calculations of intramolecular non-bonded energy term within distant receptor atoms. It would work effectively if the target protein is large and highly charged.

An appropriate library design may compensate the deficiency of high computational cost of our program at another level. Because the number of combinations of amino acid sequences is numerous, the efficient enumeration of candidate peptides helps to reduce the computational time in total. Evolutional algorithm can be used to optimize amino acid sequences [31]. However, such optimization protocols are highly dependent on the prediction accuracy of the binding affinity. It is essential to use the accurate method for prediction of the binding affinity like the rescoring.

1.5.4. Recommended Usage

We describe the recommended usage of our program. First, our program requires the 3D-structure of the receptor proteins. If available, it is preferred to use the equilibrated structures generated by MD simulations. Because our scoring function is based on the potential energy function on MM, the equilibrated structure optimized to MM is compatible with our scoring function. On the other hand, the structures determined experimentally are not optimized for MM. MD simulation is also helpful to find optimal conditions for positional restraints on the ligand molecule. The use of positional restraints is highly recommended for the accurate prediction.

1.6. Conclusion

We demonstrated the high ability of our docking program to predict binding conformations of the ligand molecule. Although conventional docking program can also predict accurate binding conformations in terms of RMSD, only our program can predict stable conformations with the lower ΔMM-GB energy. MM-based predictions of the binding affinity are often used as a post process of the molecular docking. The consistency of the evaluation functions between the molecular docking and the post processes will be great advantages. On the other hand, our results revealed the limitation of conventional scoring functions for the molecular docking of peptides.

We accomplished the acceleration of the molecular docking using GPU. We showed the good example of GPU-acceleration for not only MM-based functions but also for the functions utilizing the 3D coordinate system. These techniques are applicable to other molecular simulations like the homology modeling. This can enhance the usability of applications and encourage the research on computational molecular design.

.

1.7. Figures

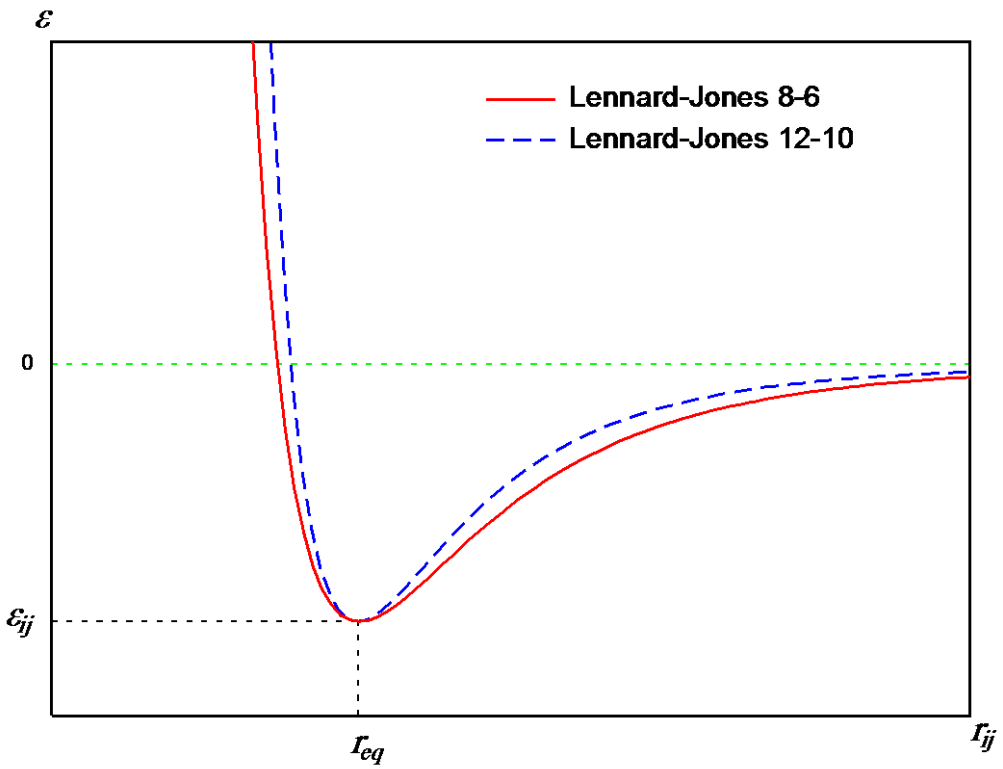

Figure 1.1 Soft Lennard-Jones Potential

Red line represents soft Lennard-Jones potential implemented in our scoring function. The standard 12-6 Lennard-Jones potential are represented by a blue dash line.

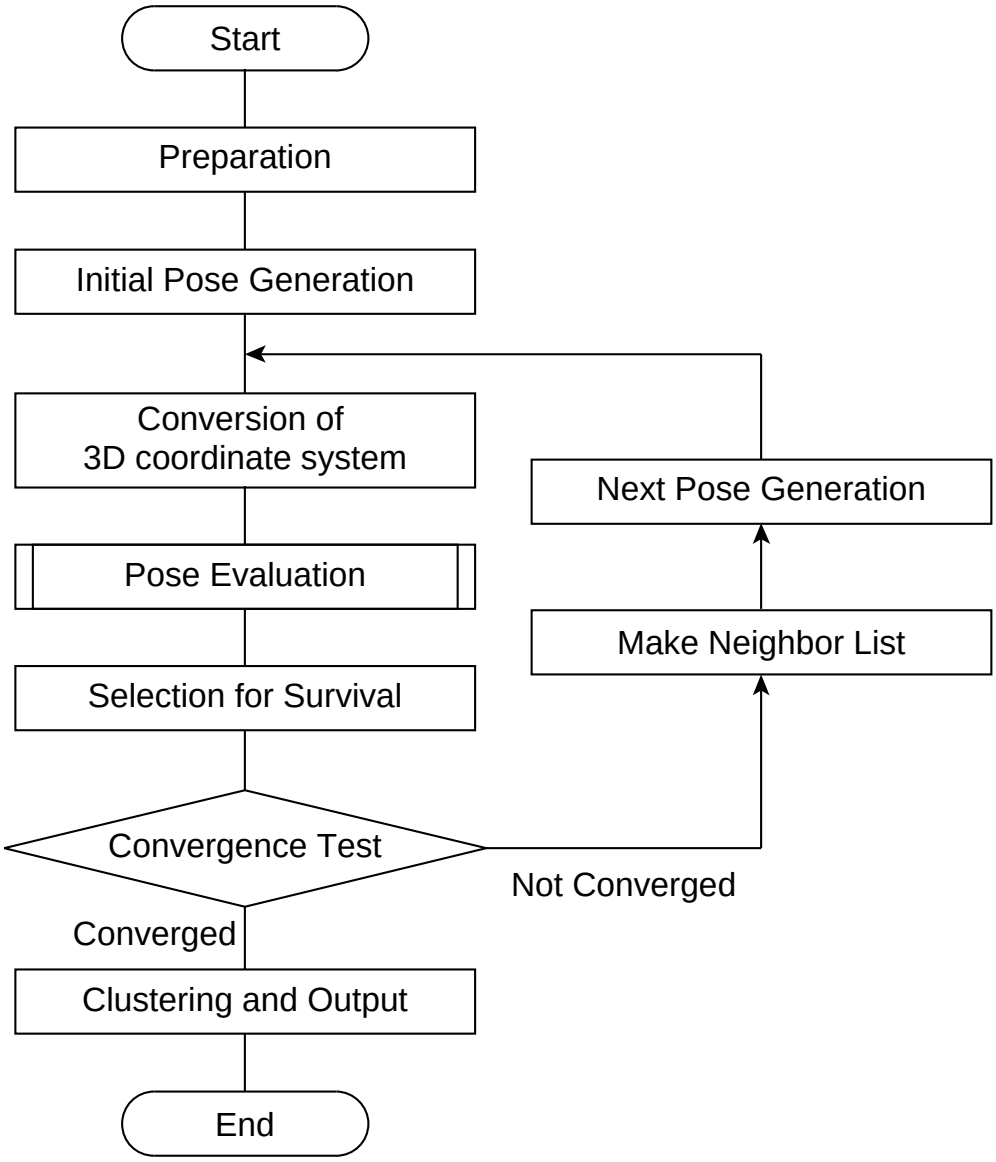

Figure 1.2 Flowchart of Procedures of Our Program

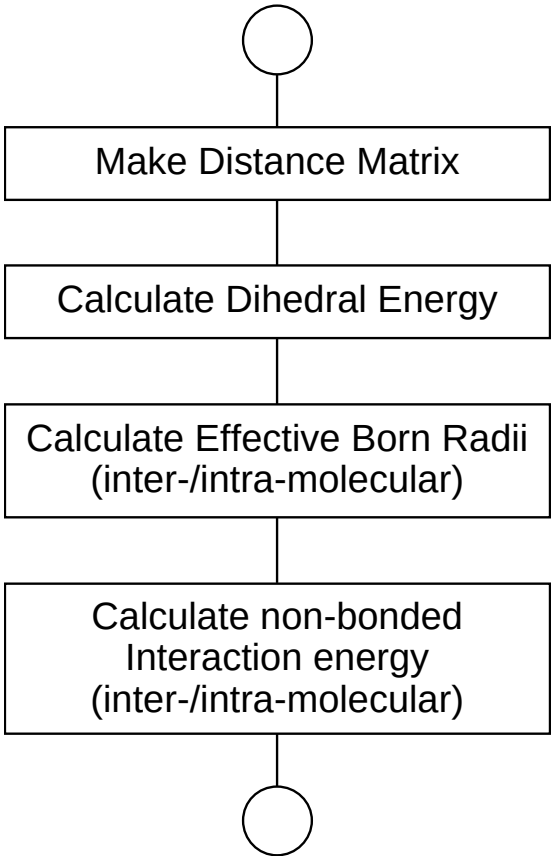

Figure 1.3 Flowchart of Calculation of Docking Score

1.8. Tables

Table 1-1 Computational Time of each Calculation Part for Grb2

Rates of computational time compared to CPU are listed in brackets.

	1 core of core i7 4770K	GTX580 (Fermi)	GTX780 (Kepler)
Conversion of 3D Coordinates System	288 ms	2.2 ms (x130.9)	1.5 ms (x189.5)
Calculation of Distance Matrix (CPU only)	42,829 ms	-	-
Calculation of Dihedral potentials	963 ms	2.3 ms (x427.2)	1.9 ms (x514.3)
Effective Born Radii (within ligand atoms)	3,626 ms	47 ms (x76.0)	16 ms (x224.2)
Effective Born Radii (between ligand-receptor)	54,964 ms	910 ms (x60.4)	471 ms (x116.7)
Non-bonded Energy (within ligand atoms)	2,030 ms	41 ms (x49.3)	17 ms (x117.8)
Non-bonded Energy (within receptor atoms)	207,645. ms	1350 ms (x153.8)	731 ms (x283.8)
Non-bonded Energy (between ligand-receptor)	44,865 ms	416 ms (x107.8)	135 ms (x332.4)
Make Neighbor List	1607 ms	152 ms (x10.5)	111 ms (x14.5)
Total Computational Time	359,576 ms	3497 ms (x102.8)	2053 msec (x175.2)

Table 1-2 Computational time of each Calculation Part for GIP

Rates of computational time compared to CPU are listed in brackets.

	1 core of core i7 4770K	GTX580 (Fermi)	GTX780 (Kepler)
Conversion of 3D Coordinates System	293 ms	2.4 ms (x122.1)	1.6 ms (x179.5)
Calculation of Distance Matrix (CPU only)	52,244 ms	-	-
Calculation of Dihedral potentials	945 ms	2.2 ms (x433.4)	1.9 ms (x506.4)
Effective Born Radii (within ligand atoms)	3,730 ms	49 ms (x76.4)	17 ms (x224.5)
Effective Born Radii (between ligand-receptor)	67,520 ms	1108 ms (x60.9)	572 ms (x118.0)
Non-bonded Energy (within ligand atoms)	2,010 ms	41 ms (x48.7)	17 ms (x116.6)
Non-bonded Energy (within receptor atoms)	326,471 ms	1941 ms (x168.2)	1091 ms (x299.3)
Non-bonded Energy (between ligand-receptor)	57,153 ms	502 ms (x114.0)	161 ms (x355.1)
Make Neighbor List	1626 ms	155 ms (x10.5)	114 ms (x14.3)
Total Computational Time	512,753 ms	4372 ms (x117.3)	2554 ms (x200.2)

Table 1-3 Comparison of RMSD values for Pose Predictions

P.R. off is RMSD values of the molecular docking without positional restraints.

P.R. on is those with positional restraints. Constraint weight are listed in brackets for GOLD.

min is those using minimized structure.

MD is those using equilibrated structure by MD.

	Crk		Grb2		Src		GIP	
	min	MD	Min	MD	min	MD	min	MD
Our Program								
P. R. off	5.45	15.50	10.69	19.2	16.09	9.66	10.48	10.68
P. R. on	1.81	1.55	2.04	2.21	3.46	1.60	3.38	2.58
GOLD								
P. R. off	10.11	11.15	4.55	7.43	5.93	11.45	19.28	16.16
P. R. on(10)	2.39	1.89	2.14	2.31	9.29	10.88	2.84	13.95
P. R. on(20)	1.95	1.69	2.40	2.59	6.78	3.29	2.45	2.40
P. R. on(30)	1.93	1.86	2.32	2.37	3.18	2.16	2.94	2.19

[Å]

Table 1-4 ΔMM-GB Energy of Top Solution

ΔMM-GB energy were calculated after the energy minimizations of the top solution of the molecular dockings, where the receptor conformation is fixed. Positional restraints were applied on every molecular docking (constraint weight is 30 for GOLD).

	Crk		Grb2		Src		GIP	
	min	MD	min	MD	min	MD	min	MD
Our Program	-48.8	-80.0	-81.3	-101.7	-68.5	-96.0	-23.8	-35.9
GOLD	-37.9	-55.7	-89.2	-61.5	-58.5	-86.0	-11.4	-29.1

[kcal/mol]

Chapter 2

New Radii for Poisson-Boltzmann Implicit Solvent

2.1. Introduction

The Poisson Boltzmann (PB) implicit solvent is commonly used to estimate the polar contribution of the solvation free energy of biological molecules. PB is used in the Molecular Mechanics and Poisson Boltzmann surface area (MM-PBSA) method [11], which estimates the free energy of the molecule in a solution. MM-PBSA is used to estimate the binding free energy (ΔG_{bind}) of a ligand to a receptor molecule. ΔG_{bind} is calculated as follows:

$$\Delta G_{bind} = G_{com} - (G_{rec} + G_{lig})$$

where G_{com}, G_{rec}, and G_{lig} denotes the free energy calculated by MM-PBSA method using the complex, receptor-only, and ligand-only structure, respectively. MM-PBSA is applied over the trajectory generated by molecular dynamics (MD) simulations or to a single snapshot, e.g., a docked structure (rescoring) [9, 30]. Many researchers have successfully designed inhibitors of various proteins using MM-PBSA [10, 32]; however, the low accuracy of PB has been pointed out. In MM-PBSA, ΔG_{bind} is expressed as follows:

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{PB} + \Delta G_{SA} - T\Delta S$$

where E_{MM} is the potential energy of MM in a gas phase; G_{PB}, a polar contribution of the solvation free energy calculated by PB; G_{SA}, a nonpolar contribution of the solvation free energy calculated by a surface-area based approach; T, an absolute temperature; and S, the entropy. G_{PB} is calculated by solving the Poisson equation:

$$\nabla \cdot \varepsilon(r)\phi(r) = -4\pi\rho(r)$$

where r is a given position; $\varepsilon(r)$, a dielectric constant at r; $\phi(r)$, an electrostatic potential at r; and $\rho(r)$, a solute charge distribution at r.

In this chapter, we discuss the dielectric boundary in PB solvents. The dielectric boundary defines a pseudo-volume of the solute. The dielectric constant at a given point is determined by whether this point is inside or outside the volume of the solute. Because

the result of PB is highly dependent on the distribution of dielectric constants, it is important to provide the appropriate definition of the dielectric boundary of the solute. Definitions based on the atomic radii have been well studied [33-36]. Sitkoff et al. developed PARSE radii to obtain an agreement between experimentally-determined solvation free energies and those calculated by PB using small organic molecules [33]. Tan et al. developed their PB radii to obtain agreement with the solvation free energy calculated using TIP3P explicit solvents. They used template molecules of amino-acid analogues and nucleic acids [36]. They designed atom-type specific radii using partial charges and atom types in AMBER force fields. Meanwhile, an abrupt and discontinuous transition of dielectric constants at the dielectric boundary causes large differences on the solvation free energy and the solvation forces between subtly different conformations. A smooth dielectric function alters the dielectric constants smoothly and continuously over the dielectric boundary [37, 38]. It can avoid large fluctuations of solvation free energies and forces; however, the smooth function also alters the optimal location of the dielectric boundary. An alternative set of atomic radii specific for the smooth dielectric function is then required. Im et al. introduced a spline-smoothed dielectric function [38]. Afterward, Swanson et al. developed PB radii [34, 35] specific for the smooth function developed by Im. Swanson et al. used template molecules of amino acids in the AMBER force field and parameterized PB radii to obtain agreement with the explicit solvent simulations using TIP3P waters.

The PB methods developed by both Tan et al. and Swanson et al. were based on simulation results using TIP3P explicit solvents and defined using the AMBER force field. This means that the two methods should provide consistent results for the solvation free energy; however they do not (Figure 2.3). From the differences between their methods for development of the PB radii set, we propose the reasons why their methods provide the different results. One is the difference between the boundary conditions of the explicit solvent simulations. Tan et al. used the periodic boundary condition with the Particle

Mesh Ewald (PME) method [39], whereas Swanson et al. used the spherical boundary with the spherical solvent boundary potential (SSBP) [40]. SSBP was used to approximate the influence of the bulk waters surrounding the spherical cap. Under the periodic boundary condition with PME, the net charge of the system must be zero. In the usual case, the system is neutralized by adding several ions with an opposite charge to the solute. In this case, however, any ions cannot be added to the system in order to calculate the solvation free energy and forces in pure waters. This results in the modification of the partial charges of the solute, if the solute has non-zero net charge. This effect must be carefully considered in the simulation results. On the other hand, under the spherical boundary condition, we have to consider the influence of abnormal distributions of waters at the edge of the solvent cap, even though SSBP is applied.

Here, we present results concerning the system-size dependence of the solvation free energies under the different boundary conditions (Table 2-1). We used one conformation of an N-terminal lysine (NLYS) as a template molecule because of its net charge (+2): the distribution of waters strongly influences the solvation free energy of NLYS. Table 2-1 indicates the system-size dependence of solvation free energies for the spherical boundary condition, and the system-size independence for the periodic boundary condition. Figure 2.1 indicates the system-size dependence for the spherical boundary through another observation. Figure 2.1 shows the radial solvent charge distribution around a Cα atom of NLYS calculated with solvent caps of various sizes. The peak shapes around 5 Å are similar to each other. We observe bulk-like distributions of waters beyond 10 Å in every graph and also unnatural peaks around each edge of water spheres. It suggested that these unnatural peaks caused the difference between the free energies for different system sizes. Therefore, the results of the explicit solvent simulations using the spherical boundary are less reliable. Likewise, the PB radii optimized by Swanson et al. are also less reliable. These influences are more remarkable for charged molecules (Figure 2.3). On the other hand, we observe the system-size independence of the solvation free energies of NLYS

for the periodic boundary condition. However, how neutralizing treatments influence the free energies of larger molecules remains uncertain.

The selection of template conformations to parameterize PB radii is also important. Tan et al. and Swanson et al. parameterized their PB radii using amino-acid-based molecules. Swanson used dipeptides of each non-terminal amino acid and several poly-alanines to consider the secondary structures of proteins, but the N- and the C-terminal amino acids were not included in the templates. We were forced to assign PB radii of non-terminal residues to terminal residues, although a large difference in solute-solvent interactions is expected. On the other and, Tan et al. used three dipeptides of alanine, proline, and glycine for backbone atoms. They also used the side-chain analogues of each amino acid for side chain atoms. Their PB radii were parameterized without considerations of the existence of backbones and side chains between one another. In addition, they did not consider the secondary structures of proteins.

In this chapter, we propose new PB radii for the accurate estimation of the polar contribution of the solvation free energy. Our work was based on the considerations of the previous studies described above. We parameterized our PB radii to obtain agreement with the explicit solvent simulations using TIP3P solvents. We used the spherical boundary condition for explicit solvent simulations. A special cut-off scheme was applied only to the calculation of the solvation free energy. It enabled us to exclude the influence of the abnormal distribution of waters on the edge of the spherical boundary.

Furthermore, we reduced excessive atom-type grouping in assignments of PB radii. In previous studies, all atoms in amino acids were grouped into several atom types. Groups were determined according to their physiochemical characteristics such as the radial solvent charge distribution around each atom. Identical PB radii were assigned to atoms in the same groups. Atom-type grouping is useful for enhancing compatibility between multiple conformations. However, atoms with the same PB radius but different partial charges may cause significant errors in the PB results owing to the sensitivity of

the PB calculations. In our work, we grouped only backbone atoms of non-terminal residues. Instead, we increased the number of template conformations of each amino acid to maintain the compatibility between various conformations.

In this chapter, we first present our parameterization of PB radii using a training set. Next, we measure the performance of our PB radii beyond the training set: we examined the prediction accuracy using larger molecules in a test set, and evaluated the prediction performance of the binding affinities by MM-PBSA method.

2.2. Methods

2.2.1. Atom-Type Grouping

We designed our PB radii specific for the AMBER protein force field (ff99SB or later) [18]. We assigned common PB radii only to each N, H, C, and O atom forming a peptide bond on each amino acid. They could be classified into several groups according to the charge states of the side chain atoms and terminal ends of the backbone atoms. Because terminal residues does not associate with the formation of secondary structures of proteins, we assigned common radii only to atoms in non-terminal residues. Thus, the PB radius of each N, H, C, and O atom consists of four patterns: three were determined by the charge states of the side chain of each amino acid, and the other is for an exceptional residue, a proline.

Each particular PB radius was assigned to all of the other atoms.

2.2.2. Training Set

All of our PB radii were parameterized using molecules in a training set. We first optimized PB radii using 12 poly-alanines used in Swanson's work. Poly-alanines include atoms forming the peptide-bond in non-terminal and non-charged amino acids and atoms in protein caps. In other words, we took account of secondary structures of proteins only for atoms forming the peptide-bond in non-charged amino acids. The structures of poly-alanines were prepared as follows: specific regions of proteins were extracted from PDB entries, 1AKI [41] and 1EJG [42]. Both the N-terminal and C-terminal ends of backbones were capped with N-acetyl (ACE) and N-methyl amide (NME) groups. Then, all amino acids were mutated to alanines. Energy minimizations were carried out in the box of TIP3P waters. After minimizations, all solvents and ions were removed.

Next, we parameterized the PB radii of atoms in each amino acid using multiple conformations as templates. Most conformations were selected from trajectories of MD

simulations. In addition, to increase sampling efficiency, conformations generated by systematic conformational search implemented in MOE [43] were used as necessary. Molecular simulations were carried out as follows: all non-terminal ends of the backbone of amino acids were capped with ACE and NME groups. After soaking solutes in the box of TIP3P waters, we generated 10 ns MD trajectories at 310 K using the pmemd module of AMBER [24]. Then, clustering analysis was performed to obtain the representative conformations of each molecule. Some of them were selected as template conformations. A systematic conformational search was carried out with the generalized born implicit solvent [16], and additional conformations were selected manually. Finally, all template conformations were energetically minimized in the box of TIP3P waters. Details of the number of conformations of each amino acid are illustrated in Figure 2.4 - Figure 2.6.

2.2.3. Test Set

To measure the performance of our PB radii beyond the training set, we used 23 structures of 13 peptides of various lengths and compared with the solvation free energies calculated by the explicit solvent simulations. We selected experimentally determined structures of 13 peptides, and downloaded from the PDB web site [19]. Detailed information of these molecules are listed in Table 2-2. The other 10 conformations were a wide variety of conformations of Chignolin generated by Replica Exchange MD [44] with the generalized born solvent from PDB entry 1UAO [45]. All conformations were energetically minimized in the box of TIP3P waters.

2.2.4. MM-PBSA Test

Our objective is to improve the accuracy of *in silico* screening based on the MM-PBSA method regardless of whether MD trajectories or single snapshots are used. We used experimental data of systematic alanine-mutational analysis of PMI peptides with MDM2 protein [46, 47]. We selected 12 peptides of the same length from experimental

data and estimated the binding affinities by MD trajectory-based MM-PBSA.

We calculated the binding affinity ΔG_{bind} by MM-PBSA in two manners: "one trajectory method" and "two trajectory method". The one trajectory method is the simple and standard application of MM-PBSA in which MD simulation is performed only for the complex structure. The structures of the receptor and ligand molecule are extracted from the complex structure from each snapshot of the complex structure. In the one trajectory method, ΔG_{bind} is calculated as follows:

$$\Delta G_{bind} = G_{com} - (G_{rec,bound} + G_{lig,bound})$$

where G_{com} is the free energy calculated using the complex structure from the MD trajectory, and $G_{rec,bound}$ and $G_{lig,bound}$ are the free energies calculated using the receptor and the ligand structures extracted from the complex structure, respectively. On the other hand, the two trajectory method uses additional MD trajectories of the ligand molecules in the unbound state. In the two trajectory method, ΔG_{bind} is calculated as follows:

$$\Delta G_{bind} = G_{com} - (G_{rec,bound} + G_{lig,unbound})$$

where $G_{lig,unbound}$ is the free energy calculated using the ligand structure from the MD trajectory in the unbound state. We compared the performances of our PB radii for two MM-PBSA methods.

All MD simulations were carried out as follows: the 3D structure of MDM2 and a N8A mutated PMI peptide complex was downloaded from the PDB web site (PDB ID: 3LNZ) [46]. The 3D structure of an unmutated peptide and MDM2 complex was predicted by a homology model module implemented in MOE. The complex structures with all other mutated peptides were generated by removing atoms in the side chain of relevant residues. Structures of PMI analogues in the unbound state were extracted from each complex structure. All solutes were soaked in the cube box of TIP3P waters and energetically minimized. We carried out 10 ns MD simulations for equilibration and 15 ns ones for production runs at 300K. Snapshots were sampled every 10 ps in the

production run. MM-PBSA was performed using these snapshots. Receptor atoms far from the binding interface were harmonically restrained at their initial positions with a force constant of 10 kcal/mol/Å^2 during MD simulations. Entropic contributions were not included in these calculations. All MD simulations were executed by the pmemd module of AMBER 12.

2.2.5. Explicit Solvent Simulations

We used the thermodynamic integration (TI) method [23] to estimate the polar contribution of the solvation free energy of the solute with TIP3P explicit solvents. We used 15 lambda points to scale the electrostatic interactions between the solute and the solvents. All lambda values were derived from the Gaussian quadrature equation. Initial structures were set up by locating the solute at the center of a spherical cap of TIP3P waters. The radius of the solvent cap for the training set and the test set was 45 Å and 53 Å, respectively. All atoms of the solute were harmonically restrained at their initial positions with a force constant of 50 kcal/mol/Å^2. We ran 15 MD simulations with the same initial coordinates but different lambda values. We performed 500 ps MD simulation for equilibration and another 500 ps one for the production run at 300 K at each lambda point. No cut-off schemes were employed for MD runs. Snapshots were sampled every 20 fs in the production run and a total of 25,000 snapshots of each lambda point was used for calculating the solvation free energy. To remove the influence of an abnormal distribution of waters at the edge of the solvent sphere, we employed a cut-off scheme in the calculation of the free energy. Because a simple scheme using a single cut-off distance results in the large fluctuation of the calculated free energy, we set multiple cut-off distances over a given range with a desired step size and averaged the calculated free energies at each cut-off distance. The cut-off distances ranged from 25 to 30 Å and from 28 to 33 Å for the training set and the test set, respectively. We observed the behaviors of bulked waters in these ranges of radial solvent charge distribution functions. The step size

was 0.1 Å.

We also calculated the polar solvation forces using the TIP3P waters. As described by Wagoner [48], a polar term of the solvation force F^p is represented by:

$$F^p = \overline{F^{p+np}} - \overline{F^{np}}$$

where $\overline{F^{np}}$ is an averaged force acting on each atom in the solute over an ensemble where only nonpolar interactions between the solute and the solvents are worked (electrostatic interactions are off), and $\overline{F^{p+np}}$ is an averaged force where full interactions between the solute and the solvents are worked. Thus, we carried out two MD simulations with full- and zero-charges of the solute. The simulation condition was the same as that for the solvation free energy except for the solute charges.

All explicit solvent simulations were carried out using AMBER 10 modified for use on the special-purpose computer MD-GRAPE3 [49].

2.2.6. Implicit Solvent Simulations

Our implicit solvent calculation was based on Swanson's work. We solved the non-linear Poisson-Boltzmann equation using the Adaptive Poisson-Boltzmann Solver (APBS version 1.3) [50]. The spacing of a PB grid was 0.20 Å. The number of grid points in each dimension was determined to become more than 10 Å larger than the size of solutes. Solute charges were distributed to PB grids using a cubic B-spline discretization. The dielectric functions were calculated by the smooth functions developed by Im et al. [38] with a half window of 0.3 Å. The dielectric constant inside and outside the solute was 1.0 and 78.4, respectively. The probe size of water was 1.4 Å. To reduce the dependence on the orientation of the solute to the grid, we prepared another orientation of each solute and averaged each result. We calculated the per-atom solvation free energy and forces, and compared them with explicit ones.

We compared the performance of our PB radii with three other implicit solvents that

are commonly used with the AMBER force field. The generalized born (GB) implicit solvent model is most common for molecular simulations. We employed the OBC model of GB [17] using AMBER 12. The dielectric constant inside and outside the solute was 1.0 and 80.0, respectively. The probe radius of water was set to 1.4 Å. The other two methods were Tan's and Swanson's methods described above. Tan's PB was carried out using an mm_pbsa.pl script of AMBER 12. The dielectric constant was the same as in the GB calculation, but the probe size of water was 1.6 Å. The grid spacing was set to 0.333 Å. Swanson's method was carried out using APBS program. Swanson's method used the same condition as ours.

2.2.7. Optimization of PB Radii

We optimized our PB radii using a genetic algorithm (GA) to obtain good agreement between the implicit and the explicit solvents. We first searched for an optimal combination of PB radii in poly-alanines and, next, we searched for those of each amino acid. The PB radii were optimized to minimize a fitness function, *f*.

$$f = \sum_{conformation\ i} f_i$$

where f_i is a fitness score of each conformation of optimized amino acids. We defined f_i as follows:

$$f_i = RMS(\Delta G_{PB,residue}) + aRMS(\Delta G_{PB,atom}) + bRMS(\Delta F^p)$$

where $\Delta G_{PB,residue}$ and $\Delta G_{PB,atom}$ are errors of the polar contribution of the solvation free energy between implicit and explicit solvents on a per-residue and per-atom basis, respectively. ΔF_p is the error of polar solvation forces for each dimension of each atom. We incorporated atom-based and residue-based terms to reduce the dependence of the PB radii on amino acid sequences. *a* and *b* are scaling factors for each component. Scaling factors *a* and *b* were set to 0.5 and 1.0, respectively. The parameters of GA are as follows:

the population size is 600. The gene is represented as the combination of PB radii. Initial combinations were determined by random numbers. The PB radii were optimized by using genetic operators: an uniform crossover and a single-point mutation. The rate of the crossover and the mutation operator is 0.7 and 0.3 respectively. The step size of the mutation operator is 0.01. Five new genes were generated from each of 600 genes, and a gene having the best fitness score becomes the gene in the next iteration.

2.3. Results

2.3.1. Explicit Solvent Simulations

We first validated our protocol of the explicit solvent simulation. Figure 2.2 shows the relation between the solvation free energy of a C-terminal asparaginic acid (CASP) and the cut-off distance applied to the calculation of the free energy. We also compared those characteristics between spherical caps of waters of different sizes. We observe almost perfect agreement between the free energies for different sphere sizes except for regions around the edge of each water sphere. The free energies around 20-30 Å are very similar in different sphere sizes but fluctuated. This is why multiple cut-off distances were used in this study. The free energies calculated using the multiple cut-off scheme show good agreement between solvent caps of different sizes (Table 2-3). We also observed different free energies between solvent caps of different sizes when using an infinite cut-off distance.

Similar results were obtained for the N-terminal lysine mentioned in the introduction of this chapter (Table 2-1). It is important for our averaged free energy to be consistent with the free energies calculated under the periodic boundary condition. This result provided strong evidence to support the validity of our explicit solvent simulations.

We also performed additional 500 ps MD for the production (total of 1000 ps) at each lambda point on the sphere size of 45 Å and obtained the free energy of -270.3 kcal/mol (standard error was ±1.03 kcal/mol). There was little difference between the calculated free energies for different time lengths of production runs. This result confirmed that our simulation time was long enough for convergence. This was expected because our simulation times were several times longer than those of previous studies.

2.3.2. Performance on Training set

The statistical performances of our PB method and the other implicit solvents on molecules in the training set are listed in Table 2-4 (a). From this table, it is indicated that

both the average of the absolute errors and the root mean square errors of our PB method are quite small, which means that our optimization of PB radii using GA was successfully accomplished. (It should be noted that we cannot compare these values with other implicit solvents because our PB radii were designed to minimize these values.) Figure 2.3 shows comparison between the polar contribution of solvation free energies of molecules in the training set calculated by explicit and implicit solvents. All implicit solvents had strong correlations with explicit solvents where the correlation coefficients of all implicit solvents were more than 0.99; however, other implicit solvents tended to have low accuracies in some particular types of molecules (Figure 2.4-Figure 2.6). GB tended to overestimate the solvation free energy in most molecules, and most of these errors were the largest, except for negatively charged amino acids. Tan's method tended to underestimate the solvation free energy of molecules in the training set, but showed good performances for charged amino acids. In addition, Tan's method had relatively low accuracy for poly-alanines considering their low polarities. Swanson's method had significant errors for charged molecules, especially for both N- and C-terminal amino acids. Swanson's method tended to underestimate the solvation free energy for negatively charged molecules and to overestimate those for positively charged molecules. All the other implicit solvents had problems with regard to the estimation accuracy of the polar contribution of the solvation free energy.

2.3.3. Performance on Test set

We measured the performances of our PB method and other implicit solvents beyond the training set using larger molecules in the test set. Our PB method showed high estimation accuracies of solvation free energies of molecules in the test set (Table 2-4 (b) and Figure 2.7). On the other hand, other implicit solvents had strong correlations with explicit solvents, but large errors of solvation free energies were observed. GB tended to overestimate the solvation free energies in most molecules. However, GB seems to have

good accuracies for 11 conformations of Chignolin: the averaged absolute errors of solvation free energies was 2.28 kcal/mol and the root mean square errors was 1.33 kcal/mol. GB certainly had good agreements with explicit solvents for the total solvation free energy, but errors in the solvation free energies on a per-residue basis were quite large. The fact suggests that the fairly accuracy of GB for Chignolin was due to an accidentally good balance between the overestimated and the underestimated solvation free energies. GB was useless for obtaining details of the solvation free energies at the residue-level. Swanson's method also had large errors on a per-residue basis. It tended to overestimate the free energies for positively charged molecules and to underestimate those for negatively charged molecules. Tan's method underestimated the free energies of all molecules. For all PB methods, the errors of solvation free energies on a per-residue basis were similar to those of the training set, but large root mean square errors were observed (Figure 2.8 - Figure 2.10).

2.3.4. Performance on MM-PBSA

Table 2-5 lists the binding free energies of 12 MDM2-peptide complexes calculated by the one trajectory method of MM-PBSA using various implicit solvents. Figure 2.11 shows the comparison between the calculated and the experimental binding free energies of 12 peptides to MDM2. The shifts of absolute calculated binding free energies occurred between different implicit solvents, but the relative binding free energies (compared to unmutated TSFAEYWNLLSP peptide) were similar to each other, especially for Swanson's method and our method. These two methods were the same except for the PB radii, and therefore, they tended to show similar binding free energies. The good correlations between calculated and experimental binding affinities were observed for all MM-PBSA method. Our method had subtly higher prediction accuracy of the binding affinities of 12 peptides in terms of the correlation coefficient.

Table 2-6 lists the binding free energies of 12 MDM2-peptide complexes calculated

by the two trajectory method of MM-PBSA with various implicit solvents. Figure 2.12 shows the comparison between the calculated and the experimental binding affinities. In the two trajectory method, large differences in the relative binding free energies of each peptide were observed as contrasted with the one trajectory method. Quite differences between Swanson's and our PB method were also observed in **A**SFAEYWNLLSP and T**A**FAEYWNLLSP. Large differences were also observed in the correlation coefficients between MM-PBSA methods. The MM-PBSA method using our PB radii had best prediction performance for binding affinities, but correlation coefficients decreased in all MM-PBSA methods compared to those in the one trajectory method.

2.4. Discussions

2.4.1. Performance of Our PB Radii

Swanson's method was inconsistent with the solvation free energy of charged molecules calculated by explicit solvent simulations. This was caused by the improper boundary conditions of the explicit solvent simulations. On the other hand, Tan's method shows good accuracies for charged residues including N- and C-terminal residues. However, Tan's method has lower accuracies as the length of peptides increases. This may be caused by insufficient selections of template molecules. Considerations of the secondary structures of proteins greatly affected to the accuracy of larger molecules. Our method could solve these problems by designing new PB radii. Our method showed the best performances for both estimations of the polar contribution of the solvation free energies of single molecule and of the binding free energies with MM-PBSA methods.

Our method may be further improved by adding other template conformations. To improve the estimation accuracy, it is necessary to increase the number of template conformations according to the number of rotatable bonds in the side chain of each amino acid. For example, we observed relatively lower accuracy for the solvation free energy of methionine in the test set. Methionine has many rotatable bonds, but only three template conformations were included in our training set. Only three conformations cannot cover various conformations of methionine. Therefore, adding other conformations is expected to improve the accuracy of our PB method.

2.4.2. Limitation of Modification of PB radii

We observed great differences between Swanson's and our PB radii. A typical example of these differences is the PB radius of the Cα atom in the backbone of amino acids. Our average PB radius of the Cα atom in non-terminal, N-terminal, and C-terminal amino acids is 2.428 Å, 2.827 Å and 1.431 Å, respectively. On the other hand, Swanson assigned a common PB radius of 2.353 Å and 2.428Å for the Cα atom of glycine and all

other amino acids, respectively. Because the PB radii represent the distances to the solvent-accessible surface, it is considered that atoms having large PB radii interact less with solvent molecules and atoms having small PB radii interact strongly with solvent molecules. Considering the results of Swanson's method, where overestimations for positively charged molecules and underestimations for negatively charged molecules of the solvation free energies occurred, our PB radii of the Cα atoms were quite reasonable.

Our PB radii provide an opportunity to consider the limitations of current PB methods. Too large or too small PB radii may work well in simple molecules such as those in our training set, but it is unknown how they would influence the solvation free energies of more complicated and larger molecules. Too small PB radii are especially problematic because they tend to generate small gaps in the interior of molecules. These gaps are regions having high dielectric constants, but they are not accessible by the explicit waters. It is easy to expect that these gaps cause errors in the calculated solvation free energies. In addition, small gaps increase the dependence of the free energy on the orientation of the solute to the PB grid. This is because wide grid spacing is insufficient to describe the distributions of dielectric constants around atoms having a small PB radius. We observed the fluctuation of the calculated solvation free energies between different orientations of the molecules in our training set to PB grid. A finer grid spacing is required to describe the small gaps well, but this increases the computational costs of PB calculations and requires a large amount of memory space on the computer resource. This is therefore no longer suitable for practical use. For these reasons, errors in the solvation free energy can be corrected to only a limited extent by modifying PB radii. Other approaches dealing with interactions between the charged moiety of the solute and solvents properly must be further studied.

2.4.3. Toward Further Improvement of MM-PBSA

In the one trajectory method of MM-PBSA, we observed similar relative binding

free energies for each peptide (Table 2-5). This seems to be caused by the cancellation of the errors in the solvation free energies of each complex, receptor, and ligand molecule in the calculations of the binding free energies. On the other hand, there were large differences in the binding free energies of each peptide calculated by the two trajectory method of MM-PBSA (Table 2-6). In the two trajectory method, these cancelling effects were decreased by half. The differences between implicit solvents were well reflected to the binding free energies.

Although the one trajectory method provides a rougher approximation of the process of the ligand binding, its correlation coefficients are higher than those of the two trajectory method. One plausible explanation for this result is that errors in the one trajectory method are consistent with some contributions of the binding free energy lacking in the one trajectory method such as entropic contributions. This must be a special case for MDM-peptide complexes; it does not always work well in other protein-ligand complexes. Nevertheless, the one trajectory method still remains the standard protocol of MM-PBSA. Our results revealed one reason for this: poor implicit solvents produce additional errors in the two trajectory method. Our results indicated that accurate implicit solvents showed good estimation accuracy for the binding free energy in the two trajectory method, but they create the need to describe lacking contributions of the binding free energies more precisely, e.g., entropic contributions. We did not include the entropic contribution from the normal mode calculation [51] in this study because of their large deviations. It is obvious that the protocol of normal mode calculations implemented in AMBER has a problem in an energy-minimization scheme. Recently, an improved method for the energy-minimization scheme in normal mode calculations was published [52]. It is worth applying this method to our studies. Furthermore, intermediate waters are also sources of errors in MM-PBSA calculations. We often observed in some receptor-ligand complexes that waters went into an interspace between ligand and receptor atoms and resulted in overestimated binding free energies. This will be more problematic when long MD

simulations are carried out, because the chances for waters to go into the interspace are increased. Although long MD simulation should be needed for the more precise description of biological processes, it results in producing the errors in MM-PBSA calculations. Often, more realistic treatments are problematic for MM-PBSA calculations. However, we believe these treatments are required for accurate estimations of binding free energies available for a wide range of protein-ligand complexes in the future. Our accurate PB radii are one of the steps toward that final goal.

Finally, we add some considerations of the LR MM-PBSA method [53]. The LR MM-PBSA method is one of the modified MM-PBSA methods. This method adjusts the magnitudes on each component of MM-PBSA (bonded, electrostatic, and van der Waals terms for the E_{MM}, E_{PB}, and E_{SA} terms) by the use of scaling factors. The original purpose of LR MM-PBSA is to force the calculated binding affinities to be consistent with experimental binding free energies. A similar strategy may be useful for correcting imbalances between the implicit and the explicit solvation free energies for previous PB solvents. However, this attempt ended unsuccessfully in our study. Significant improvements in correlations were not observed in all MM-PBSA methods. LR MM-PBSA will be useful only when similar ligand molecules are compared, because the optimal scaling factor is different according to the physiochemical characteristics of ligand molecules. Because peptides change their characteristics easily by the replacement of just one amino acid, LR MM-PBSA method would not be effective for peptide-protein complexes.

2.5. Conclusion

We developed novel PB radii set to improve the estimation accuracy of Poisson Boltzmann implicit solvents. The use of our PB radii showed the good accuracies of estimations of the solvation free energies on single peptide molecules. The accuracy maintained stable if the length of amino acid residues of peptides was up to 24. Unfortunately, we cannot examine the performances on larger molecules, because the explicit solvent simulations cannot be performed due to the limit of computational memory. However, it is expected that our PB will show better performances than other implicit solvents.

In the one trajectory method of MM-PBSA, the use of our PB radii set showed the best performance although the correlation coefficients between calculated and experimental binding free energies were similar in different MM-PBSA protocols. This is because a large portion of the errors in the solvation free energies were canceled in binding affinity calculations. In the two trajectory method of MM-PBSA, only the MM-PBSA method using our PB radii set showed high estimation performances for the binding free energies. The two trajectory method also revealed inaccuracies of conventional implicit solvents. To improve the accuracy of MM-PBSA methods, more precise descriptions of other components such as entropies are also required in addition to our PB radii set.

Our PB radii showed the limitation of optimizing the PB radii for accurate estimation of the solvation free energy. A too small radius on negatively charged groups tends to generate small gaps between solute atoms, which would cause the errors of the solvation free energy. Other approaches must be studied to describe appropriate interactions between the solutes and the solvents involved in charged moieties.

2.6. Figures

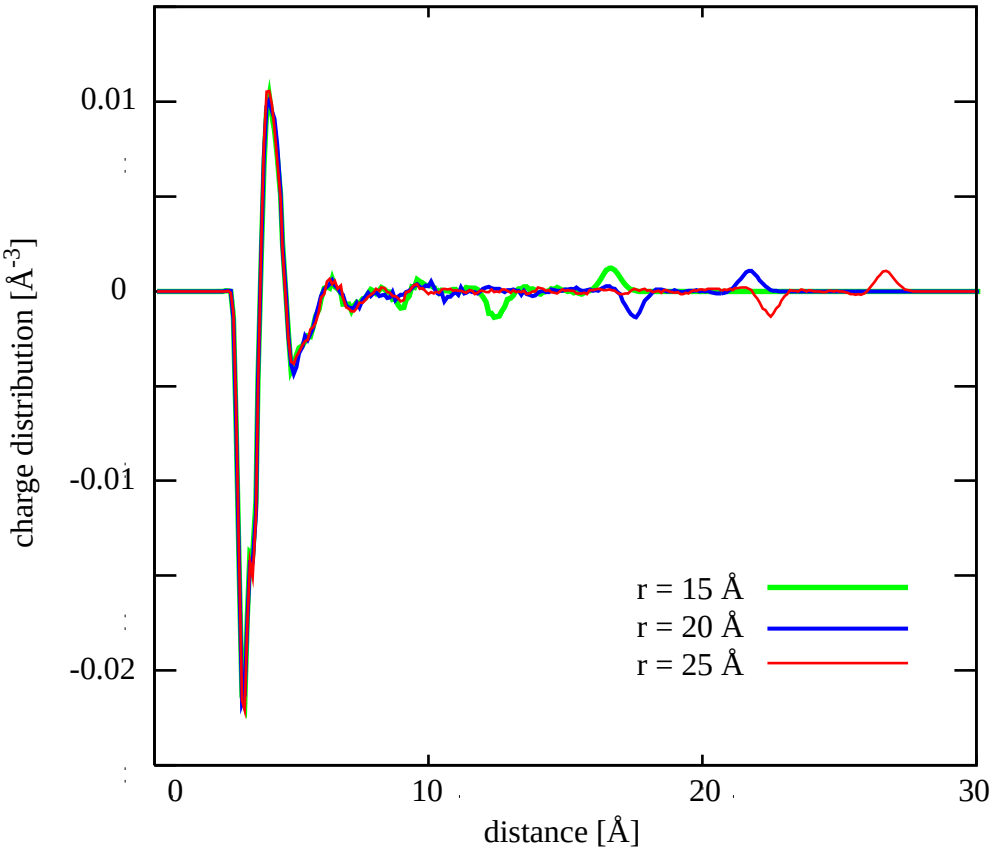

Figure 2.1 Radial Solvent Charge Distributions for Solvent Caps of Various Sizes

Radial solvent charge distributions around a Cα atom of an N-terminal lysine are calculated from explicit solvent simulations using solvent caps of different sizes (*r*: sphere radius).

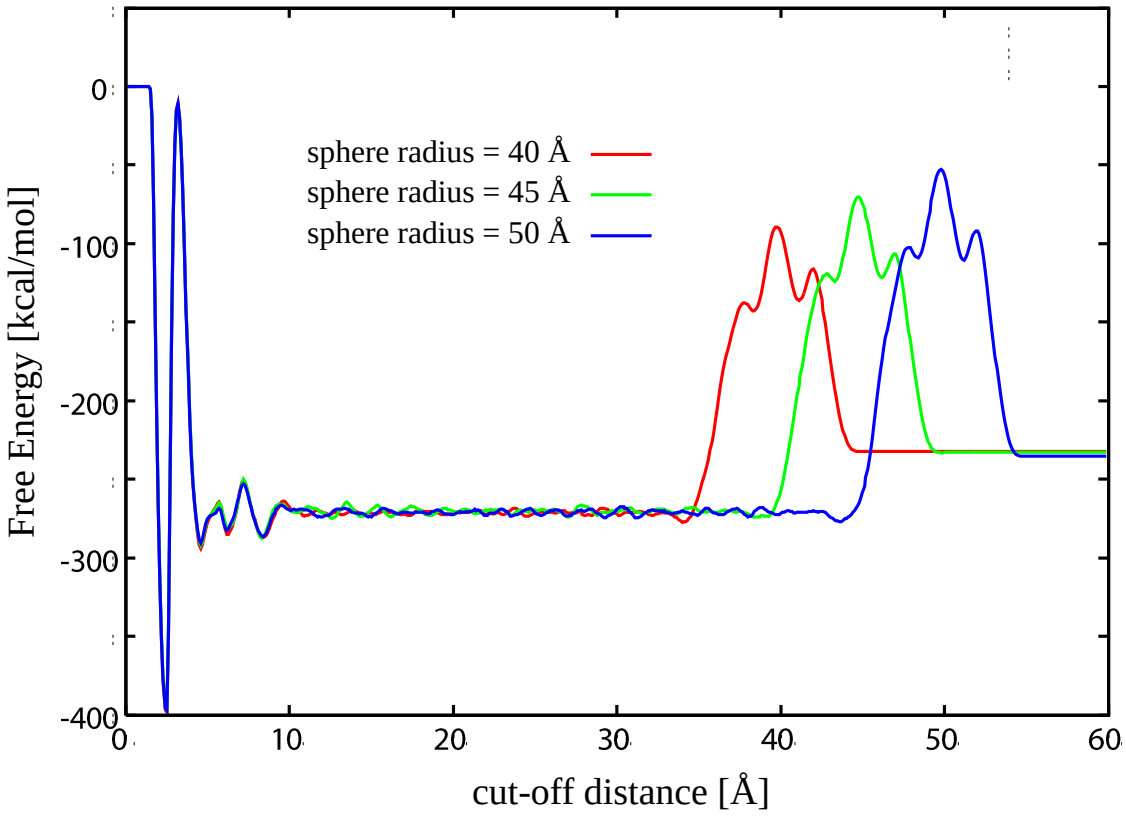

Figure 2.2 Calculated Free Energy vs. Cut-off Distance

The graph illustrates the relation of the polar contribution of the solvation free energy of a C-terminal asparaginic acid and the cut-off distance. We calculated the free energy with spherical water caps of different sizes.

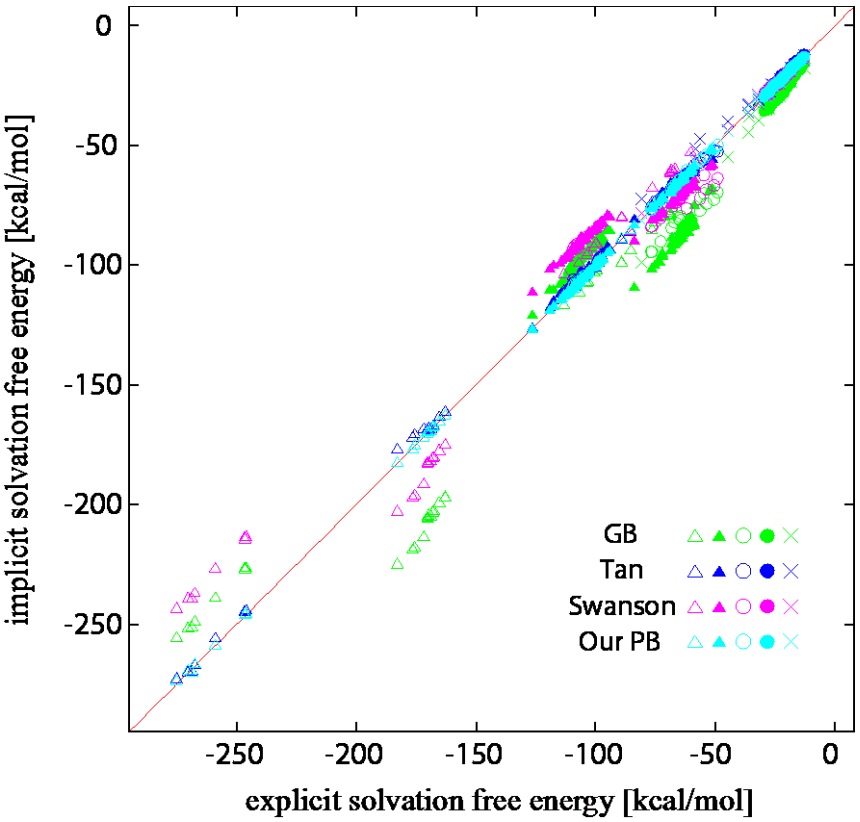

Figure 2.3 Comparison between Implicit and Explicit Solvent in Training Set

The relation of the polar contribution of the solvent free energy between explicit and implicit solvents model are illustrated as follows:

Circles: non-terminal amino acids.

Triangles: N- or C- terminal amino acids.

Filled marks: non-charged amino acids.

Blank marks: charged amino acids.

x-marks: poly-alanines.

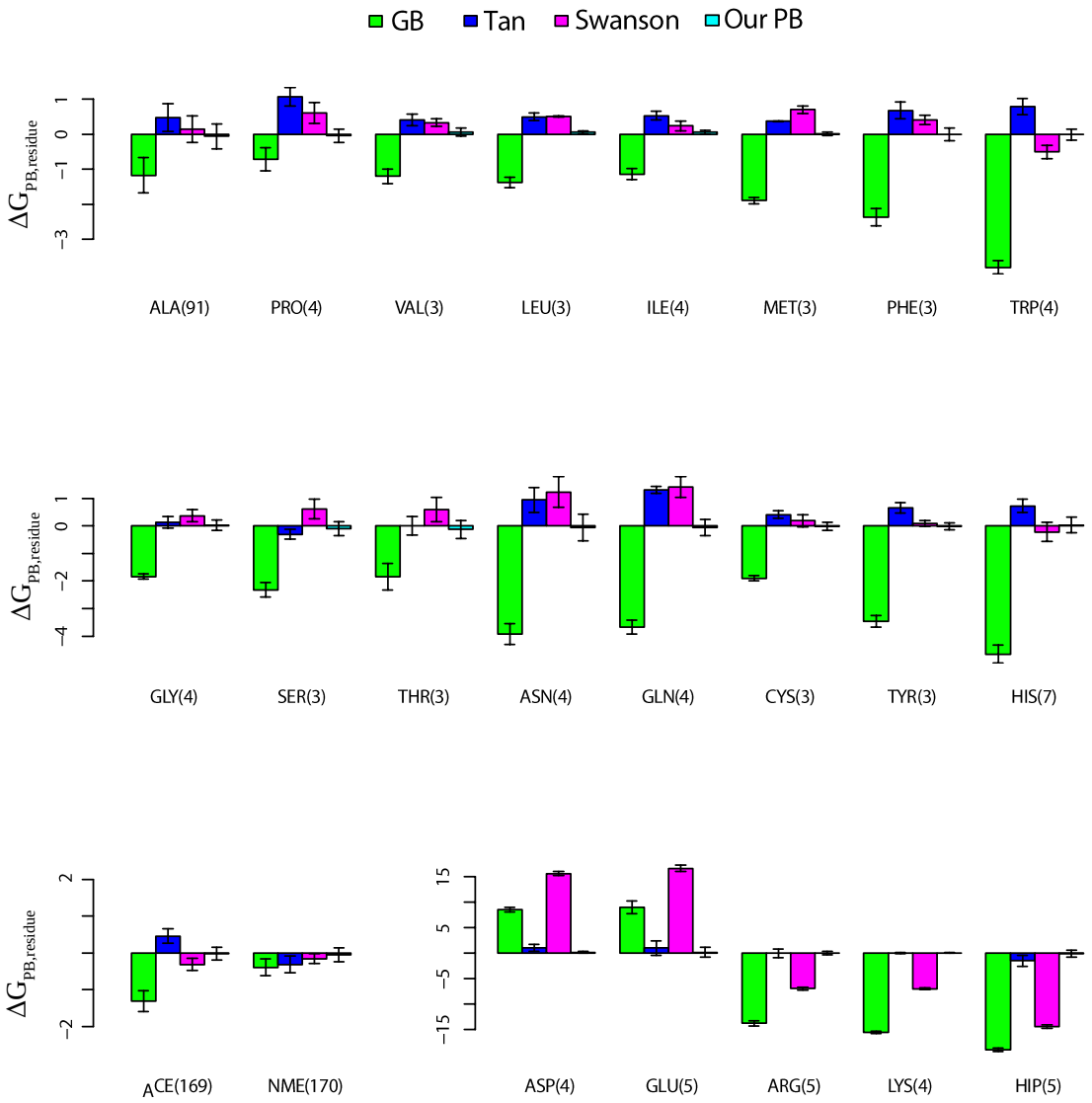

Figure 2.4 Errors in Solvation Free Energy per Residue (Training Set, Non-terminal Residues)

Averaged errors of the polar contribution of the solvation free energy on a per-residue basis $\Delta G_{PB,residue}$ [kcal/mol] for non-terminal residues in the training set. The number of residues included in molecules is listed in parentheses.

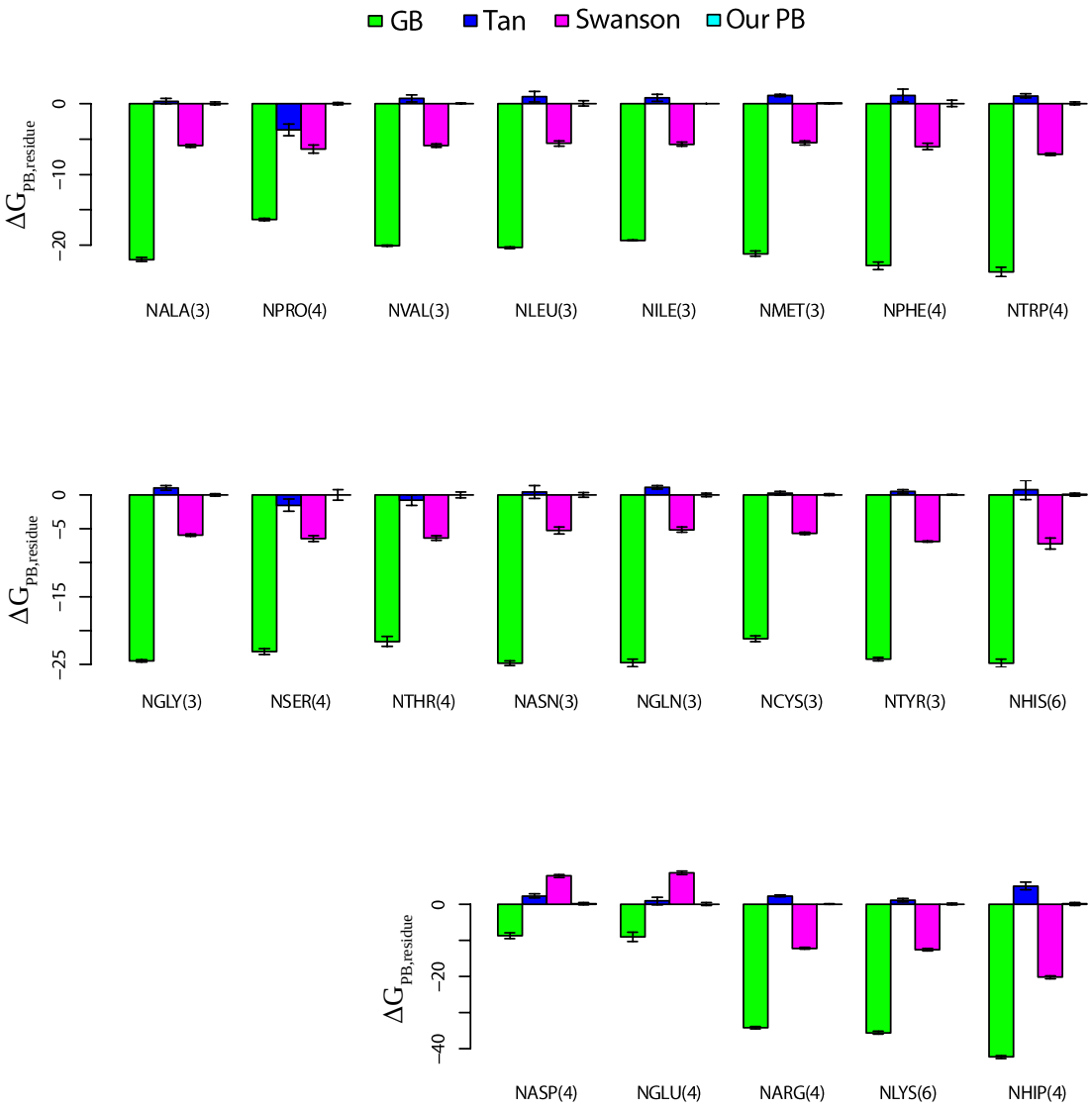

Figure 2.5 Errors in Solvation Free Energy per Residue (Training Set, N-terminal Residues)

Averaged errors of the polar contribution of the solvation free energy on a per-residue basis $\Delta G_{PB,residue}$ [kcal/mol] for N-terminal residues in the training set. The number of residues included in molecules is listed in parentheses.

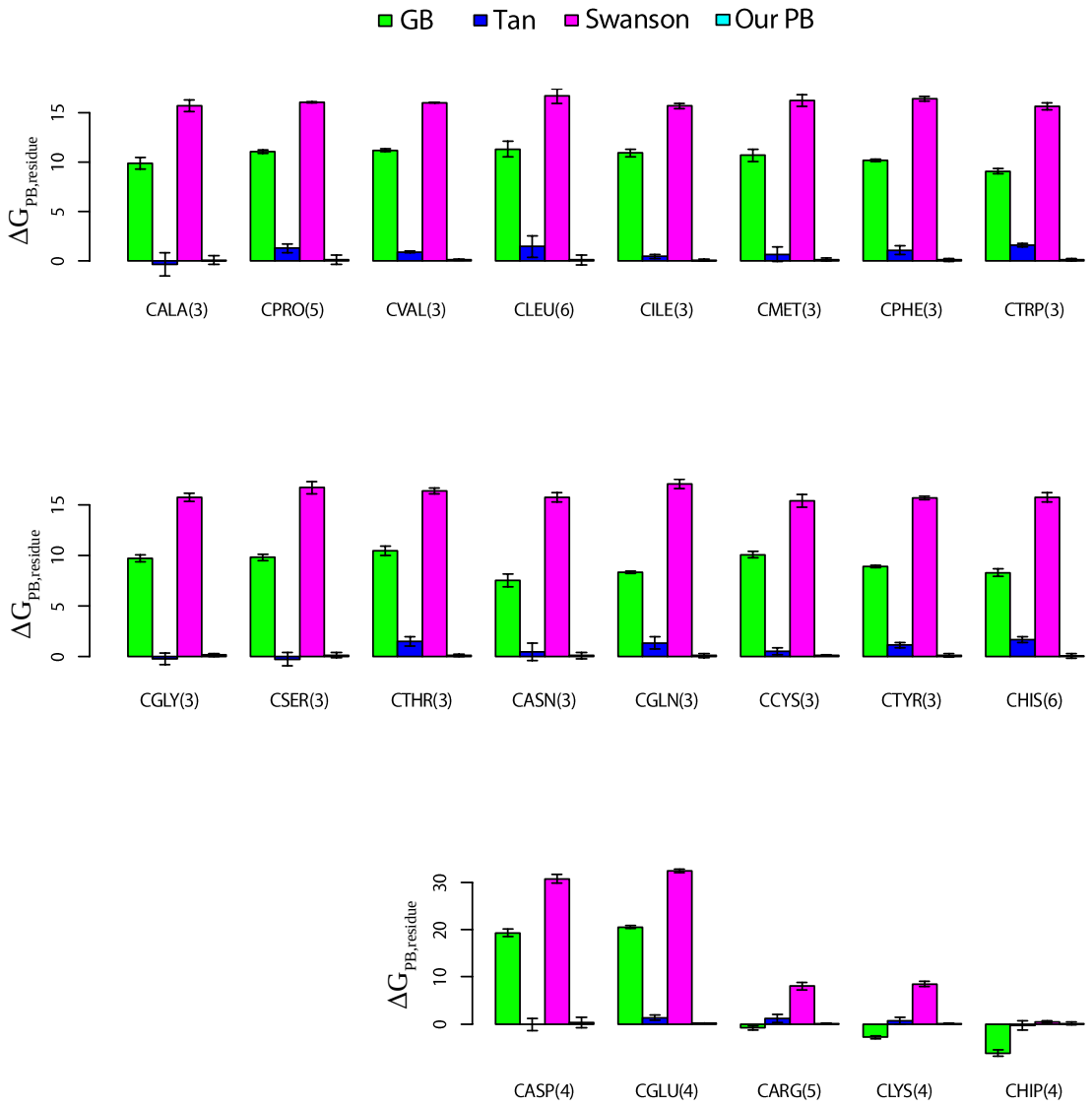

Figure 2.6 Errors of Solvation Free Energy per Residue (Training Set, C-terminal Residues)

Averaged errors of the polar contribution of the solvation free energy on a per-residue basis $\Delta G_{PB,residue}$ [kcal/mol] for C-terminal residues in the training set. The number of residues included in molecules is listed in parentheses.

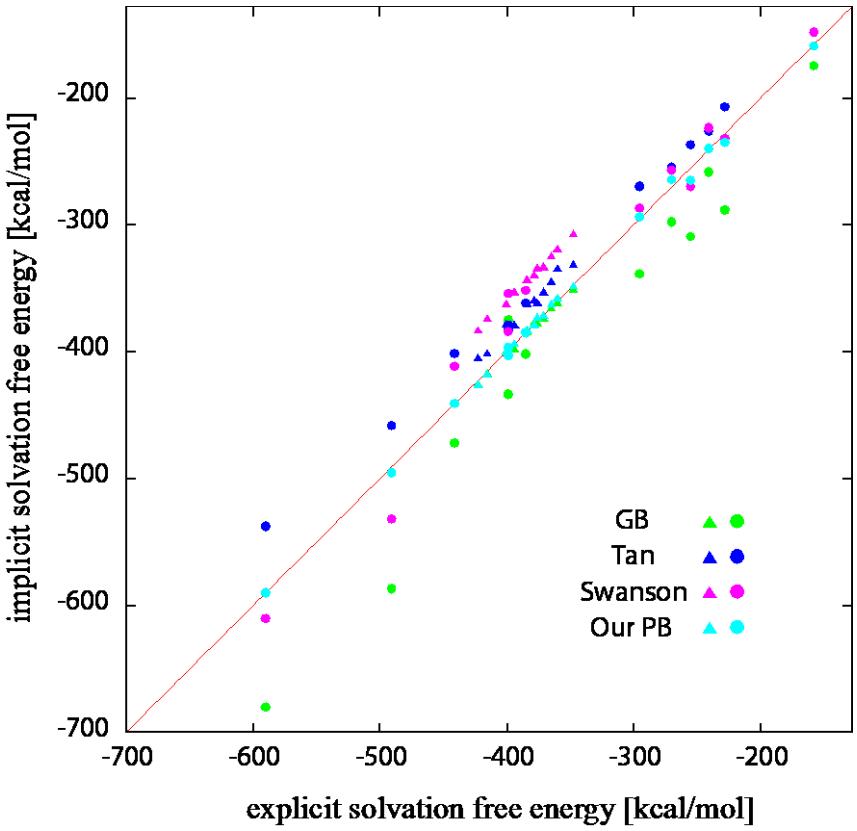

Figure 2.7 Comparison between Implicit and Explicit Solvents in Test Set

The relation of the polar contribution of the solvent free energy between explicit and implicit solvents model is illustrated as follows:

Circles: 12 different peptides

Triangles: 11 conformations of Chignolin (including a pdb structure)

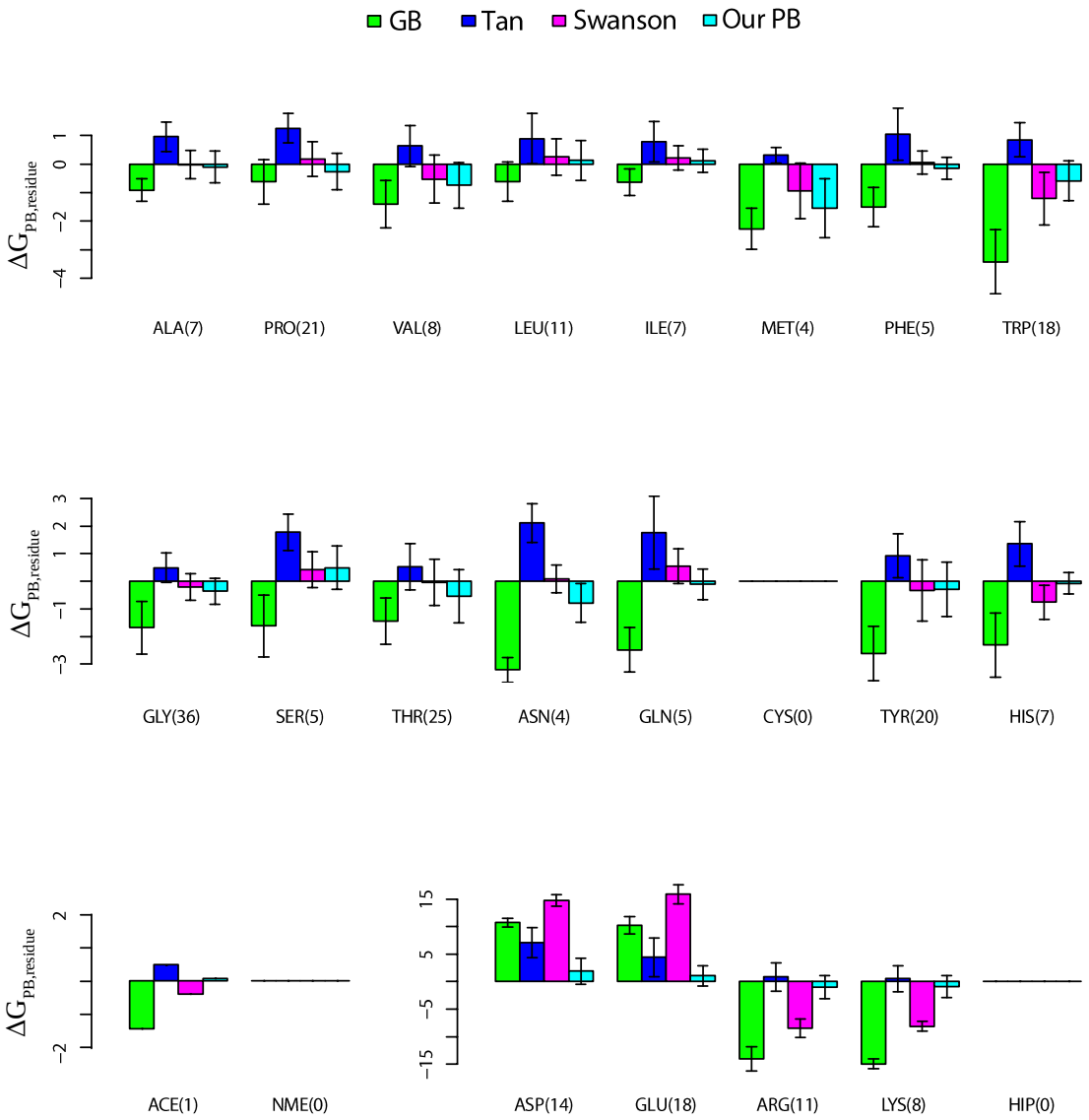

Figure 2.8 Errors of Solvation Free Energy per Residue (Test Set, Non-terminal Residues)

Averaged errors of the polar contribution of the solvation free energy on a per-residue basis $\Delta G_{PB,residue}$ [kcal/mol] for non-terminal residues in the test set. The number of residues included in molecules is listed in parentheses.

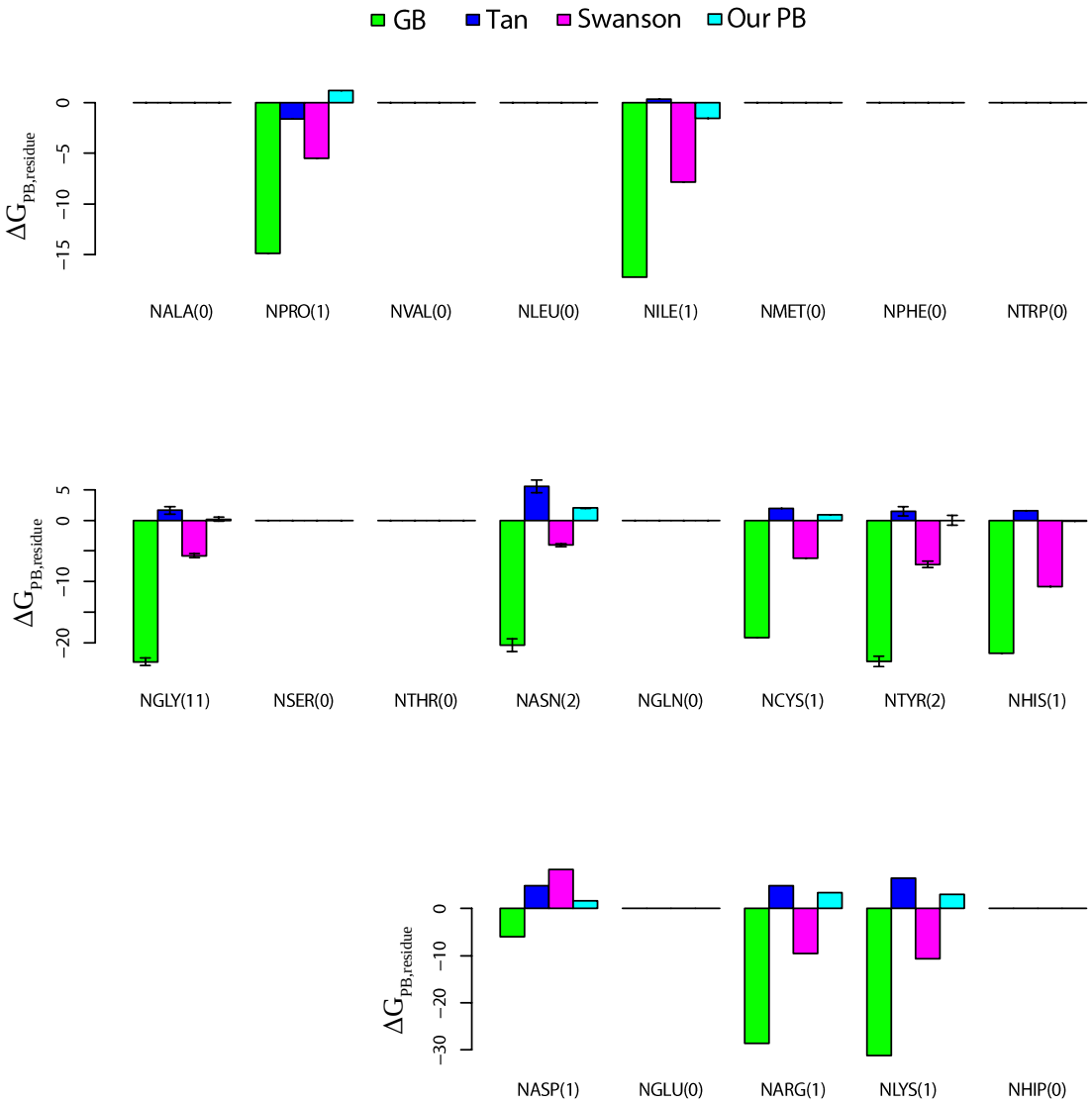

Figure 2.9 Errors of Solvation Free Energy per Residue (Test Set, N-terminal Residues)

Averaged errors of the polar contribution of the solvation free energy on a per-residue basis $\Delta G_{PB,residue}$ [kcal/mol] for N-terminal residues in the test set. The number of residues included in molecules is listed in parentheses.

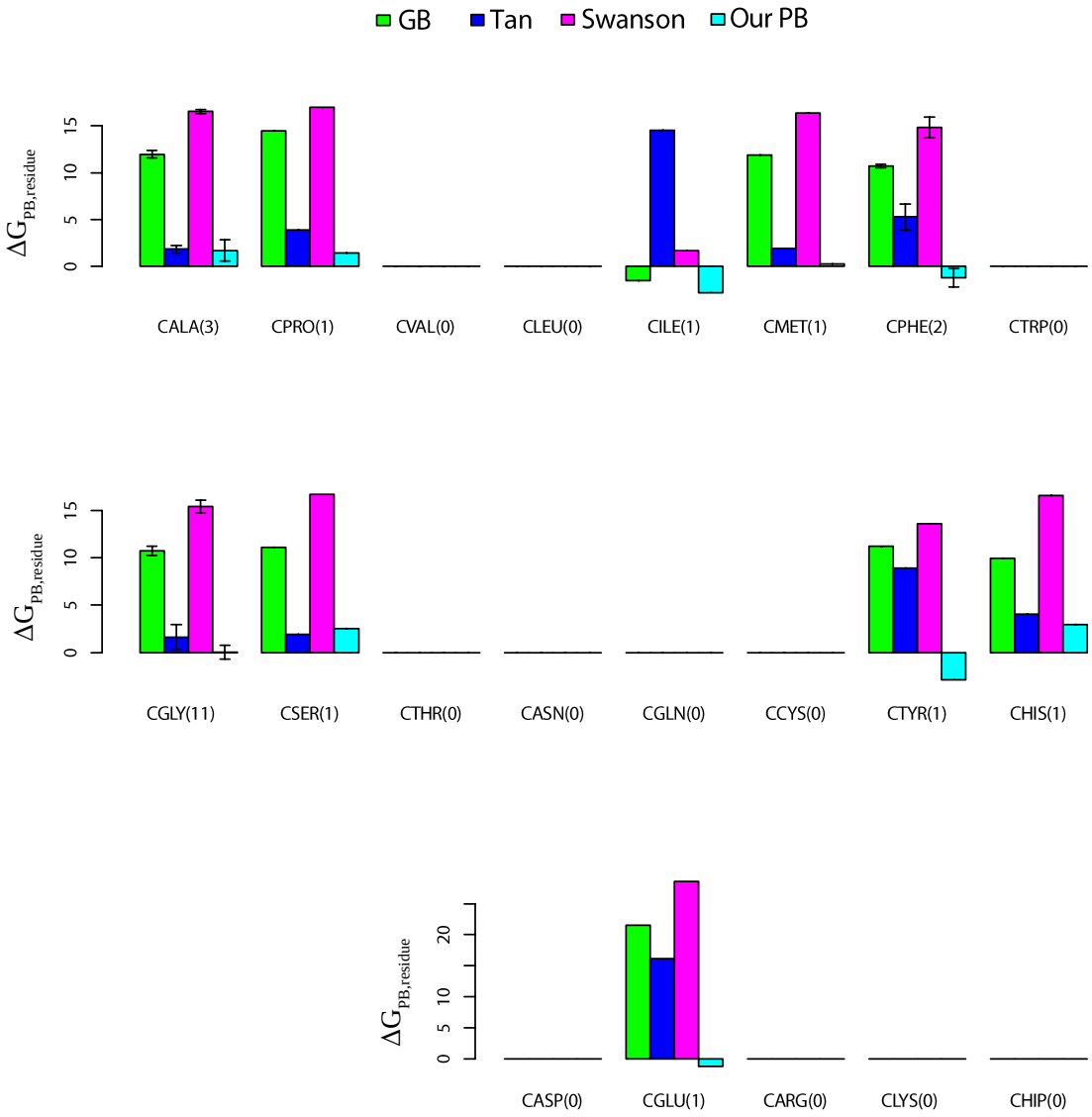

Figure 2.10 Errors of Solvation Free Energy per Residue (Test Set, C-terminal Residues)

Averaged errors of the polar contribution of the solvation free energy on a per-residue basis $\Delta G_{PB,residue}$ [kcal/mol] for C-terminal residues in the test set. The number of residues included in molecules is listed in parentheses.

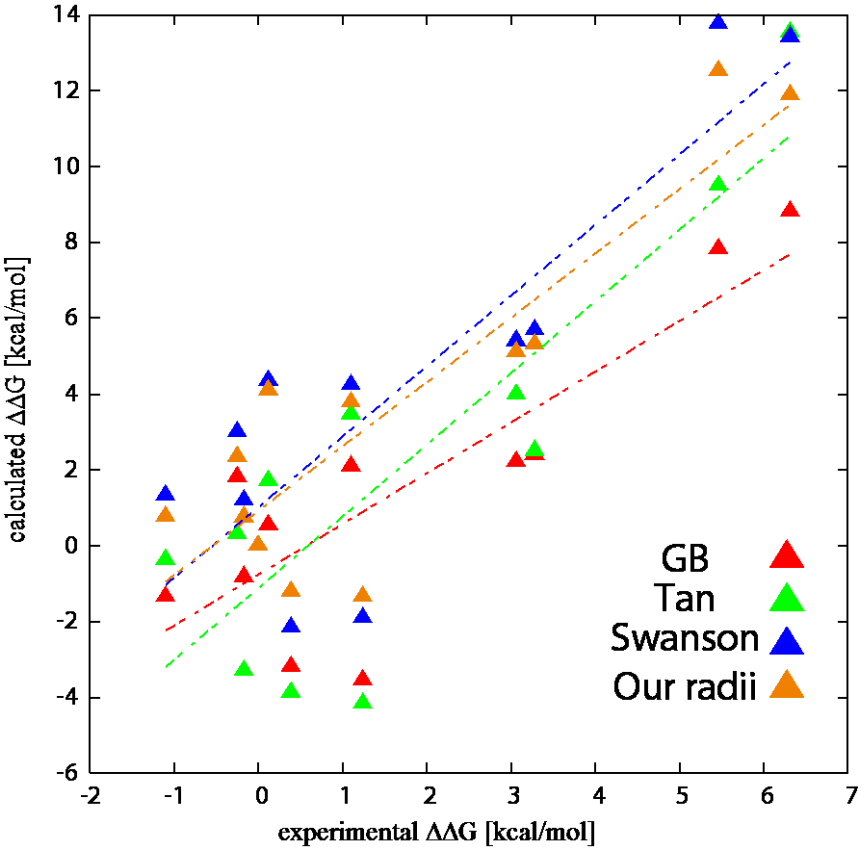

Figure 2.11 Performance on MM-PBSA using One Trajectory Method

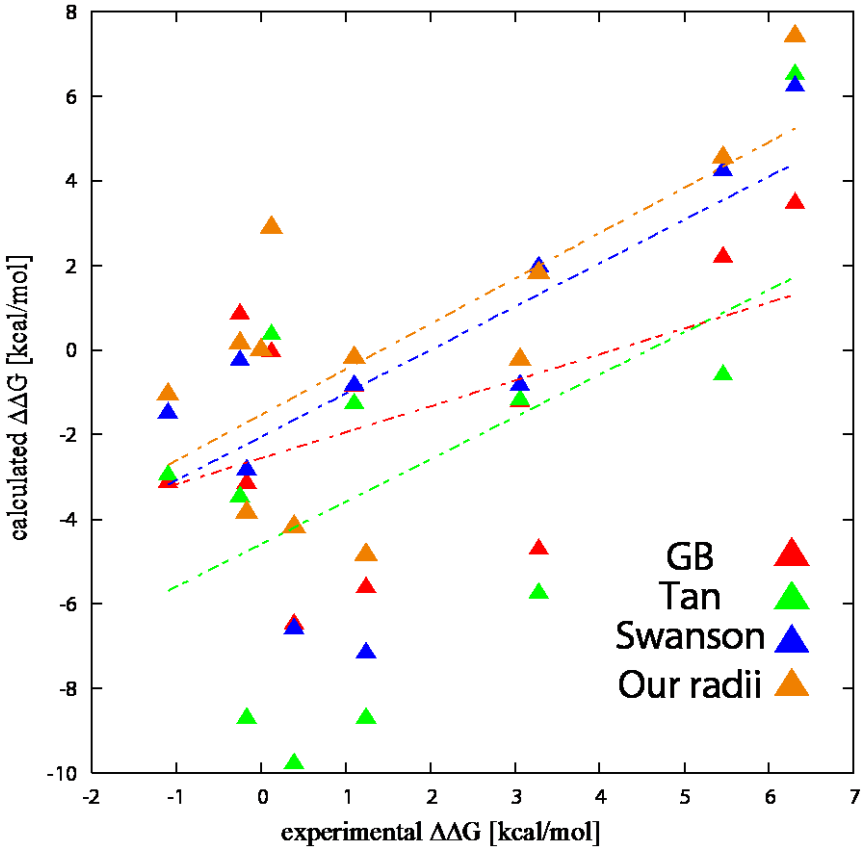

Figure 2.12 Performance on MM-PBSA using Two Trajectory Method

2.7. Tables

Table 2-1 Dependence of Solvation Free Energy on System Size

Polar contributions of the solvation free energies calculated by each method are listed. An N-terminal lysine is used as a template. Simulation protocol of "Spherical Boundary" and "Periodic Boundary" was based on Swanson's paper [35] and Tan's paper [36], respectively.

Spherical Boundary (Swanson)				
sphere radius (Å)	15	20	25	30
solvation free energy (kcal/mol)	-185.88	-186.78	-187.11	-187.45

Periodic Boundary (Tan)				
length of each edge (Å)	50	60	70	
solvation free energy (kcal/mol)	-168.12	-167.73	-167.74	

Our Method: Spherical Boundary with Infinite Cutoff Scheme				
sphere radius(Å)	20	30	40	45
solvation free energy (kcal/mol)	-159.53	-169.37	-174.52	-176.37

Our Method: Spherical Boundary with Multiple Cutoff Scheme				
sphere radius(Å)			40	45
solvation free energy (kcal/mol)			-168.04	-168.00

Table 2-2 Information of Molecules in Test Set

Protein Name	PDB ID	Length	Sequence	Reference
Met-enkephalin	1PLW	5	YGGFM	[54]
A fragment of ribonucleotide reductase	1AFT	7	Ac-FTLDADF	[55]
Angiotensin II	1N9V	8	DRVYIHPF	[56]
Histon H3 analogue	1CS9	9	CGGIRGERA	[57]
Chignolin	1UAO	10	GYDPETGTWG	[45]
Designed peptide	2O0S	12	YVLWKRKRMIFI	[58]
A fragment of staphylococcal nuclease	2FXZ	13	KMVNEALVRQGLA	[59]
mab198 bound peptide	2JRV	15	PMTLPENYFSERPYH	[60]
GCN4 trigger peptide	2OVN	17	NYHLENEVARLKKLVGE	[61]
Designed peptide	2DX4	18	INYWLAHAKAGYIVHWTA	[62]
TRP-cage	1L2Y	20	NLYIQWLKDGGPSSGRPPPS	[63]
Phosphopeptide P140 (non-phosphorylated form)	2L5I	21	RIHMVYSKRSGKPRGYAFIEY	[64]
Prion protein	1OEI	24	HGGGWGQPHGGGWGQPHGGGWGQP	[65]

Table 2-3 Polar Solvation Free Energy Calculated by Our Method

Radius of Spherical Cap	Calculated Free Energies [kcal/mol]	
	Infinite Cut-off Distance	Multiple Cut-off Scheme (Range)
40Å	-232.35	-271.16 ± 1.12 (20-25 Å)
45Å	-232.99	-270.32 ± 1.86 (25-30 Å)
50Å	-235.31	-271.42 ± 1.56 (25-30 Å)

Table 2-4 Statistical Performance of Implicit Solvents

Averaged absolute error (AAE) and the root mean square errors (RMSE) of the polar contribution of the solvation free energies between implicit and explicit solvents are listed. RMSE of polar solvation forces are listed.

(a) Performance on **training set** molecules

Implicit Solvents	AAE and RMSE of Solvation Free Energy [kcal/mol]	RMSE of Forces [kcal/mol/Å^2]
GB	13.09 ± 9.33	1.377
Tan	1.34 ± 1.79	1.977
Swanson	8.72 ± 7.40	0.616
Our PB	0.32 ± 0.31	0.740

(b) Performance on **test set** molecules

Implicit Solvents	AAE and RMSE of Solvation Free Energy [kcal/mol]	RMSE of Forces [kcal/mol/Å^2]
GB	23.27 ± 27.63	1.384
Tan	21.24 ± 9.32	1.715
Swanson	29.85 ± 13.04	0.767
Our PB	2.43 ± 2.39	0.958

Table 2-5 Binding Free Energy Calculated by One Trajectory Method of MM-PBSA

Experimental and calculated binding free energies of each peptide are listed. Relative binding free energy compared to TSFAEYWNLLSP is also listed in brackets. All experimental data were derived from ref. [46].

	Experimental Relative Binding Free Energy [kcal/mol]	Calculated Binding Free Energy [kcal/mol]			
		GB	Tan	Swanson	Our PB
ASFAEYWNLLSP	0.39	-46.41 (-3.19)	-56.63 (-3.87)	-72.96 (-2.16)	-67,39 (-1.21)
T**A**FAEYWNLLSP	1.24	-46.78 (-3.55)	-56.92 (-4.16)	-72.69 (-1.90)	-67.52 (-1.34)
TS**A**AEYWNLLSP	5.46	-35.39 (+7.83)	-43.26 (+9.49)	-57.03 (+13.76)	-53.66 (+12.52)
TSF**A**EYWNLLSP	0	-43.23 (0)	-52.75 (0)	-70.79 (0)	-66.18 (0)
TSFA**A**YWNLLSP	1.10	-41.13 (+2.10)	-49.29 (+3.47)	-66.56 (+4.23)	-62.40 (+3.78)
TSFAE**A**WNLLSP	3.06	-41.01 (+2.22)	-48.77 (+3.99)	-65.38 (+5.41)	-61.08 (+5.10)
TSFAEY**A**NLLSP	6.31	-34.41 (+8.82)	-39.22 (+13.54)	-57.37 (+13.42)	-54.29 (+11.89)
TSFAEYW**A**LLSP	-1.10	-44.57 (-1.34)	-53.13 (-0.37)	-69.46 (+1.33)	-65.41 (+0.77)
TSFAEYWN**A**LSP	-0.17	-44.05 (-0.82)	-56.05 (-3.30)	-69.59 (+1.20)	-65.42 (+0.76)
TSFAEYWNL**A**SP	3.28	-40.83 (+2.40)	-50.25 (+2.50)	-65.10 (+5.69)	-60.87 (+5.31)
TSFAEYWNLL**A**P	0.12	-42.67 (+0.55)	-51.05 (+1.71)	-66.44 (+4.35)	-62.06 (+4.11)
TSFAEYWNLLS**A**	-0.25	-41.41 (+1.82)	-52.44 (+0.31)	-67.79 (+3.00)	-63.83 (+2.34)
Correlation Coefficient		0.838	0.850	0.858	0.881

Table 2-6 Binding Free Energy Calculated by Two Trajectory Method of MM-PBSA

Experimental and calculated binding free energies of each peptide are listed. Relative binding free energy compared to TSFAEYWNLLSP is also listed in brackets. All experimental data were derived from ref. [46].

	Experimental Relative Binding Free Energy [kcal/mol]	Calculated Binding Free Energy [kcal/mol]			
		GB	Tan	Swanson	Our PB
ASFAEYWNLLSP	0.39	-47.07 (-6.49)	-59.03 (-9.78)	-72.75 (-6.59)	-66.80 (-4.18)
T**A**FAEYWNLLSP	1.24	-46.20 (-5.62)	-57.97 (-8.71)	-73.33 (-7.17)	-67.45 (-4.84)
TS**A**AEYWNLLSP	5.46	-38.40 (2.19)	-49.86 (-0.60)	-61.92 (4.24)	-58.07 (4.55)
TSF**A**EYWNLLSP	0	-40.59 (0.00)	-49.26 (0.00)	-66.16 (0.00)	-62.62 (0.00)
TSFA**A**YWNLLSP	1.10	-41.45 (-0.87)	-50.54 (-1.28)	-66.99 (-0.84)	-62.81 (-0.19)
TSFAE**A**WNLLSP	3.06	-41.81 (1.22)	-50.44 (-1.18)	-66.98 (-0.83)	-62.85 (-0.23)
TSFAEY**A**NLLSP	6.31	-37.12 (3.47)	-42.75 (6.51)	-59.93 (6.23)	-55.20 (7.42)
TSFAEYW**A**LLSP	-1.10	-43.73 (-3.14)	-52.22 (-2.96)	-67.64 (-1.49)	-63.66 (-1.05)
TSFAEYWN**A**LSP	-0.17	-43.73 (-3.15)	-57.97 (-8.71)	-68.98 (-2.83)	-66.46 (-3.84)
TSFAEYWNL**A**SP	3.28	-45.31 (-4.72)	-55.01 (-5.75)	-64.19 (1.97)	-60.79 (1.82)
TSFAEYWNLL**A**P	0.12	-40.64 (-0.05)	-48.90 (0.36)	-63.25 (2.90)	-59.72 (2.90)
TSFAEYWNLLS**A**	-0.25	-39.74 (0.85)	-52.73 (-3.47)	-66.41 (-0.26)	-62.46 (0.16)
Correlation Coefficient		0.470	0.509	0.615	0.708

Chapter 3

in silico Peptide Screening against SH2 domains

3.1. Introduction

In chapter 1, we improved the prediction accuracy of binding conformations of peptides to their target proteins by developing our original program for molecular docking. In chapter 2, we parameterized new PB radii and showed high performances on estimations of polar contributions of the solvation free energies of single molecules and on predictions of binding affinities by Molecular Mechanics and Poisson-Boltzmann Surface Area (MM-PBSA) method [51]. In this chapter, we combined our improved methods and applied them to *in silico* screening of peptides against various Src Homology 2 (SH2) domains.

SH2 domain is one of modules of adaptor proteins. Adaptor proteins basically have no catalytic activities, but they have several protein-binding modules. Each module of adaptor proteins physically associates with upstream or downstream signaling proteins in a signaling pathway and enhances the formation of protein complexes. SH2 domains recognize phosphorylated states of the specific tyrosine of upstream proteins and bind only to the phosphorylated state of the tyrosine (pY). On the other hand, SH3 domains binds proline-rich amino acid sequences of downstream signaling proteins. Adaptor proteins recruit upstream and downstream signaling proteins via binding of specific regions of signaling proteins to each protein-binding module of adaptor proteins.

Adaptor proteins are involved in some cancer cell activities, therefore, they are attractive therapeutic targets [66, 67]. For example, an activity level of Src proteins increased in many types of tumors. An activity level of Crk proteins is also elevated in many types of tumors, especially in the colon and lung cancers [68]. Grb2 proteins are involved in inappropriate cell proliferations in some leukemia [69] and in breast and ovarian cancers [70]. Then, preventing the signals mediated by adaptor proteins are promising approaches for several cancer therapies. Many researchers studied the design of peptide or small molecule inhibitors binding to SH2 or SH3 domains [71-73].

SH2 domains are optimal systems to examine the ability to discriminate binding

peptides for *in silico* screening method. SH2 domains are highly structurally conserved modules, but they selectively bind to signaling proteins. Each SH2 domain has a certain preferential binding sequence, called binding motif [74]. For example, pYxxI, pYxxP and pYxNx is the binding motif for Src, Crk and Grb2 SH2 domains, respectively (x indicates any amino acids). Recently, Liu et al. investigated the binding selectivities of 50 SH2 domains by SPOT analysis [75]. In this chapter, we utilized this experimental data to measure the performances of our screening method.

In this chapter, we tried to discriminate binding peptides of several SH2 domains from a small peptide library using our screening method. Our goal in this chapter is to find the optimal condition of peptide screening for each protein toward the large scale of screening. Our screening method was based on the molecular docking and MM-PBSA rescoring. Structures of ligand-receptor complexes were predicted by our molecular docking program, and binding affinities were estimated by MM-PBSA method using our PB radii. We compared our method with conventional methods in terms of the performances on peptide screening. Furthermore, we examined the dependency of the conformations of the receptor proteins on the screening performance.

In addition, we investigated the effects of the reorganization of ligand molecules on peptide screening. In other words, we applied the two trajectory method to the docking-based screening. The reorganization effects are explained as a free energy difference caused by a conformational change through the binding process. Both the ligand and the receptor molecules usually change their conformations into suitable conformations according to their binding partners. The free energy difference associating these conformational changes are unfavorable for each molecule (it is also referred as a restraint energy), however free energies obtained from the binding partner overcomes these free energy loss and lead to the formation of the complex structure. In the one trajectory method, the unstable binding conformations of the ligand molecule in the complex structure are permitted because the restraint energies are completely neglected. Only (so-

called) interaction energies are the interests in the one trajectory methods. However, it is questionable whether these binding conformations can represent the correct binding conformations.

The reorganization effects are more effective for molecules having the high conformational flexibility, like peptides. In chapter 2, we applied the two trajectory method of MM-PBSA to 12 MDM2-peptide complexes. Because binding peptides used in chapter 2 were in the forms of α-helixes, the reorganization effects were potentially small. Peptides used in this chapter, which are adjusted to 8-mer length, has no stable secondary structures of proteins. Thus, the reorganization effects of peptides may influence strongly the performances on peptide screening.

In this chapter, we performed additional conformational search of peptides in the unbound state. We used the same program described in chapter 1 to search stable conformations of peptides. Because the conformations of unbound peptides are highly fluctuated in solvents, just one stable conformation predicted by our program does not reflect the actual conformations of peptides in waters. However, the energy difference between the bound- and the unbound state of peptides may be useful as a rough estimation of the restraint energy.

3.2. Methods

3.2.1. Experimental Data

Our study was based on experimental data by Liu [75]. Liu examined the interactions between 192 phosphorylated peptides and 50 SH2 domains by SPOT analysis. We selected four SH2 domains, Crk, Grb2, Nck1, and Src SH2 domains, as target proteins of our peptide screening from 50 SH2 domains, because their 3D structures of peptide-SH2 complexes were available. We used 100 peptides illustrated in Figure 2 of Liu's paper as a small set of the peptide library. Peptides having more than 3 times greater binding intensity than the average intensity of 100 peptides were regarded as binding peptides: 16 peptides for Crk, 11 peptides for Grb2, 14 peptides for Nck1, and 14 peptides for Src SH2 domains were selected as the binding peptides respectively.

3.2.2. Procedures of Screening

Our protocol to predict the binding affinity of each peptide is as follows: we first prepared linearly extended structures of the peptide. We carried out the molecular docking using the extended structure as the input structure and obtained 30 candidate binding conformations of the peptide-receptor complexes. We also carried out the conformational search of the peptide in the unbound state and obtained 30 candidate-conformations. All predicted structures were energetically minimized in the box of TIP3P waters. After minimizations, all solvents and ions were removed. Receptor conformations were fixed in the whole processes.

We calculated the binding affinities by MM-PBSA method as the rescoring of the docked structures. First, we calculated the free energy of 30 complex structures and 30 ligand structures of each peptide by MM-PBSA method. Next, we selected the most stable structures of the complex and of the peptide structure in terms of the calculated free energies. We calculated binding affinities in two manners: the one trajectory method and the two trajectory method. In the one trajectory method, conformations of the receptor

and ligand molecule were extracted from the complex structure. The binding affinity ΔG was calculated as follows:

$$\Delta G = G_{com} - (G_{rec,bound} + G_{lig,bound})$$

where G_{com} is the free energy calculated by MM-PBSA using the complex structure predicted by molecular docking, $G_{rec,bound}$ and $G_{lig,bound}$ is the free energy calculated by MM-PBSA using the receptor and ligand structure extracted from the complex structure. On the other hand, two trajectory methods uses two predicted structures of the complex and the peptide. The binding affinity ΔG was calculated as folllows:

$$\Delta G = G_{com} - (G_{rec,bound} + G_{lig,unbound})$$

where $G_{lig,unbound}$ is the free energy calculated by MM-PBSA with the ligand structure predicted by conformational search in the unbound state.

3.2.3. Structural Preparation

In chapter 1, we demonstrated the efficacy of Molecular Dynamics (MD) simulations on structural preparations for our molecular docking. In this chapter, we prepared five conformations of each peptide-receptor complexes by MD simulations. The initial structures were downloaded from the PDB web site. PDB IDs are 1JU5 [26] for Crk, 1JYR [27] for Grb2, 2CI9 [76] for Nck1, and 1KC2 [28] for Src SH2 domains. The length of amino acids of all ligand peptides were adjusted to 8-mer (X_{-2}-X_{-1}-pY_0 -X_{+1}-X_{+2}-X_{+3}-X_{+4}-X_{+5}: each residue was named after the relative positions from pY for convenience). All protonation states of the solute were determined by the protonate3D module of MOE [43]. All solutes were soaked in the box of TIP3P waters. A total of 13 ns MD simulations was performed on each complex molecules. We used the receptor conformations at 5, 7, 9, 11, and 13 ns of MD simulations (named MD5, MD7, MD9, MD11, and MD13 for convenience). The position of ligand peptides of each

conformations were also used as reference positions of positional restraints. MD simulations were carried out using the pmemd module of AMBER 12 [24].

3.2.4. Molecular Docking

We carried out the molecular docking using two software: our program accelerated by GPU described in chapter 1 and GOLD [77]. In our program, the number of parent-conformations was set to 1,000 and the number of child-conformations per parent-conformation was 30. Positional restraints were applied as follows: the position of the phosphorus atom in the phosphorylated tyrosine were fixed during simulations. Cα atoms at -1, 0, +1, +2, +3, and +4 residues were harmonically restrained with the force constant of 10.0 kcal/mol/$Å^2$ when the atoms are located more than 3.0 Å away from the reference positions. We carried out the molecular docking three times, and each 10 conformations from top 10 clusters were used for following processes.

The force field for the phosphorylated tyrosine (pY) was derived from the work of Homeyer et al [78].

For GOLD, positional restraints were applied to the phosphorus atom of the phosphorylated tyrosine and every Cα atom. Constraint weights were 30 and the constraint radius is 3.0 Å from reference positions. The binding sites were determined with the center point of the reference ligand structure with sphere radius 20Å. A searching efficiency was set to 200%. The number of docking runs was 30, and the top conformation of each docking run were used for following processes. All other parameters remain as defaults.

3.2.5. Rescoring by MM-PBSA

We calculated the binding affinities using various MM-PBSA methods developed by Tan et al. [36], Swanson et al. [35] and us. Simulation conditions was the same described in chapter 2. We designed PB radii for the phosphorylated tyrosine in the same manners

described in Chapter 2 and used for our PB method. For Swanson's PB, the BONDI radii with optimal offset for smoothing dielectric functions were used [35, 79].

3.2.6. Performance Metric

We used the area under the receiver operating characteristics curve (ROC AUC) to measure the performances of peptide screening. ROC AUC was often used as a metric of the performance to discriminate binders from non-binders in virtual screening [80]. ROC AUC ranges from 0 to 1. ROC AUC of 0.5 corresponds to the random selection. Higher ROC AUC indicated a better performance of the screening method. We used ROC AUC of 0.7 as a criterion for good performances.

3.3. Results

3.3.1. Screening Performance of GOLD

Before discussing our method, we discuss the performance of conventional docking program, GOLD, on peptide screening. We described in previous chapter that GOLD is incapable of predicting correct binding poses without any positional restraints. Here, we discuss the performance of GOLD with the positional restraints: Table 3-1 lists the ROC AUCs of each screening using only GOLD: the conformations of the peptide-receptor complexes were predicted by GOLD, and docking scores were used as binding affinities. ROC AUCs higher than 0.7 were observed in only 1 of 4 proteins. This result indicated that GOLD was incapable of predicting correct binding affinities of peptides.

3.3.2. GOLD with MM-PBSA Rescoring

We measured performances of combined methods of GOLD and MM-PBSA rescoring where the binding conformations of the peptide-receptor complexes were predicted by GOLD and binding affinities were predicted various MM-PBSA rescoring.

In all proteins, the best ROC AUC values were higher than those using only GOLD. The deviations of ROC AUCs are also increased in all proteins. It suggested the MM-PBSA rescoring is sensitive to the binding conformations.

MM-PBSA method using Swanson's PB showed the highest performance in 2 of 3 proteins. MM-PBSA method using our PB method showed the highest performance in 1 of 3 proteins. The ROC AUCs of these two methods are similar, because these two PB methods was the same except for the PB radii set. It resulted in the high correlation coefficients, 0.873, between the binding affinities of peptides on the Crk MD9 structure calculated by our and Swanson's method. On the other hand, the correlation coefficient of Tan's PB with Swanson's and our PB is 0.654 and 0.709, respectively. This indicated the PB methods strongly influence the screening performances.

3.3.3. Our Molecular Docking with MM-PBSA Rescoring

We measured performances of screening methods using our program for molecular docking and various MM-PBSA rescoring (Table 3-3). For all proteins, the best ROC AUC values are higher than those using combined methods of GOLD and MM-PBSA rescoring. This result indicated the superiority of our molecular docking.

In most structures of Crk, Nck1, and Src, the ligand reorganization affected positively to the screening performance. The impact of the reorganization effects seems to be relatively small for MM-PBSA rescoring using Tan's PB method. The screening performances of Grb2 were decreased by including the reorganization effects except for Swanson's PB. By including the reorganization effects, MM-PBSA rescoring based on Swanson's and our PB methods accomplished the ROC AUC almost higher than 0.7 in all proteins.

3.4. Discussions

3.4.1. Ligand Reorganization Effects

Including reorganization effects of ligand molecules improved the screening performances on Crk, Nck, and Src SH2 domains. This effect decreased the ROC AUCs in some cases, but this losses were quite small in most cases. The reorganization effects were less effective for peptide screening of Grb2 SH2 domain. One plausible explanation for this results is the conformational flexibility of the ligand molecule in the bound state. The N-terminal and C-terminal regions of Grb2-binding peptides are exposed to solvents. We observed highly fluctuations of these regions (Figure 3.1). If ligand peptides change their conformations freely even in the bound state, it is unreasonable to represent binding conformations of peptides using just a single stable conformation predicted by molecular docking. Furthermore, considering this situation, the reorganization effects cannot be represented because the two trajectory method estimate the energy loss between only two stable conformations in bound and unbound states. Multiple conformations may be required for both the bound and the unbound state of ligand molecules to describe the conformational change in the fluctuated structures. One trajectory method seems to be rather appropriate for the highly fluctuated peptides, because it ignore the conformational change of peptides completely. The reorganization effects should be applied after the careful considerations of the conformational flexibility of peptide in the bound state.

The reorganization effects seems to work favorably to molecules having relatively less conformational flexibility, such as peptides including a proline residue. Because the conformational change between in the bound and the unbound state are less small for these molecules, the expected restraint energies tended to be small. There is a potential bias to increase the binding affinities for specific kinds of molecules. In our study, the binding motif of Crk SH2 domain, pYxxP, is relevant to this problem. We examined the high performances for Crk SH2 domain was caused by such biases or not. We measured the discrimination performances of known *non*-binding peptide sequences having pYxxP

motif. In our peptide library, 8 from100 peptides have pYxxP sequence but do not bind to Crk SH2 domain. ROC AUC for discriminating non-binding pYxxP peptides are listed in Table 3-4. This result indicated that our screening can discriminate pYxxP binding peptides from pYxxP non-binding peptides. The reorganization effects worked unfavorably to pYxxP non-binding peptides. MM-PBSA using Swanson's PB method showed subtly high ROC AUCs compared to other implicit solvents, which may be more problematic at the large scale of the peptide screening.

3.4.2. Negative Effects of the Use of MD Structure on Src SH2

The screening performances on Src SH2 domains were relatively low compared to other proteins. It was caused by structures used in the screening and the selection of the peptide library.

We described the importance of MD simulations to generate structures used in molecular docking in chapter 1. MD simulations can equilibrate molecular systems and generate stable conformations of molecules. As a result, these conformation of the receptor proteins were optimized according to their ligand molecules. It is known as an induced fit. Because we used the induced fitted conformations of the receptor proteins for peptide screening, there are some biases on discrimination of binding peptides. We did not get rid of these biases because the conformation of the receptor proteins were fully fixed in the whole process. In the case of peptide screening for Src SH2 domain, the amino acid sequences of the ligand peptide is PQpYEEIPI. The conformations of Src SH2 domain were optimized to its sequence. The binding motif of Src SH2 domain is known as pYxx(I/M/L); however, only 4 of 14 binding sequences from our peptide library fulfill this binding motif. Especially, the binding sequences satisfying pYxxI was just one sequence: EDpYGDIEI. This should be a major reasons for relative low ROC AUCs for Src SH2 domain. Perhaps, experimental data for Src SH2 domains did not meet the requirement to measure the performances of peptide screening. We will confirm to

screening performances of Src SH2 by rebuilding the peptide library for screening.

3.4.3. Best Implicit Solvents for MM-PBSA

MM-PBSA using Swanson's PB method showed totally high performances on four proteins. However, there should be any biases considering the results in chapter 2. We described the underestimations of the solvation free energies for negatively charged molecules for Swanson's PB method in chapter 2. In general, the binding peptides to SH2 domains are negatively charged. Therefore, the use of Swanson's method is inadvisable. Furthermore, the reorganization effects are useful for peptides as long as the peptides are less free in the bound state. However, the errors of the solvation free energy were increased for Swanson's PB by including the reorganization effects. We consider the only our PB method can estimate the solvation free energy of the solute correctly.

3.5. Conclusions

We measured our screening performances on four SH2 domains compared to conventional methods. We demonstrated the inability of conventional docking score to discriminate the binding peptides from the other. MM-PBSA rescoring with predicted structures by GOLD improved on the screening performances, however, our molecular docking showed further improvements in ROC AUCs.

It needs careful considerations for including the ligand reorganization effect in peptide screening. It may be useless if ligand peptides have high conformational flexibility even in bound states. These characteristics can be investigated in advance using MD simulations of the peptide-protein complex structure. MD simulations are also useful for generation the structure used in screenings. Because MM-based binding affinity predictions are highly dependent on the receptor conformations, the pre-screening using a small library against several conformations is essential.

MM-PBSA rescoring using Swanson's PB method showed high performances on peptide screening. However, the use of their PB radii set is not good idea because SH2-binding peptides are generally charged because of the inaccuracy of Swanson's PB for charged residues. The errors of the solvation free energies affect unfavorably to the large scale of peptide screening.

3.6. Figures

Figure 3.1 Superimposed Structures of Grb2-peptide complexes

Snapshots extracted every 1 nsec from 13 nsec MD simulations of Grb2-peptide complexes are superimposed. Backbones are represented by the ribbons. The phosphorylated tyrosine and the asparagine in the binding motif of Grb2 SH2 domain are represented as sticks. Receptor molecules are illustrated in green and ligand molecules are illustrated in cyan.

3.7. Tables

Table 3-1 Screening Performance of GOLD

The values of ROC AUC are listed. Best ROC AUC for each protein is highlighted in **bold**.

	Crk	Grb2	Nck1	Src
MD5	0.746	0.763	0.458	0.432
MD7	0.520	0.557	0.654	0.469
MD9	0.545	0.715	0.470	0.535
MD11	0.516	0.779	0.586	0.561
MD13	0.496	0.404	0.643	0.420
Average	0.565	0.644	0.562	0.483

Table 3-2 Screening Performance with GOLD and MM-PBSA

The values of ROC AUC are listed. Best ROC AUC for each protocol is highlighted in **bold**.

(a) Crk SH2 domain

	Tan	Swanson	Our PB
MD5	0.623	**0.663**	**0.637**
MD7	0.569	0.354	0.407
MD9	0.575	0.524	0.521
MD11	**0.664**	0.530	0.523
MD13	0.593	0.503	0.501
Average	0.605	0.515	0.518

(b) Grb2 SH2 domain

	Tan	Swanson	Our PB
MD5	0.733	0.740	0.655
MD7	0.810	**0.787**	**0.830**
MD9	0.669	0.740	0.647
MD11	**0.886**	0.730	0.813
MD13	0.633	0.662	0.709
Average	0.746	0.732	0.731

(c) Nck1 SH2 domain

	Tan	Swanson	Our PB
MD5	0.363	0.660	0.601
MD7	0.508	**0.724**	**0.647**
MD9	0.446	0.609	0.605
MD11	**0.545**	0.576	0.581
MD13	0.438	0.673	0.624
Average	0.460	0.648	0.612

(d) Src SH2 domain

	Tan	Swanson	Our PB
MD5	0.525	**0.608**	0.587
MD7	**0.584**	0.535	0.536
MD9	0.428	0.590	0.581
MD11	0.513	0.599	**0.622**
MD13	0.547	0.583	0.556
Average	0.519	0.583	0.576

Table 3-3 Screening Performance with Our Program and MM-PBSA Rescoring

The values of ROC AUC are listed. Left values in each PB method are ROC AUCs of screening not including the reorganization effects of peptides. Right values are those of screening including the reorganization effects of peptides. Best ROC AUC for each protocol is highlighted in **bold**.

(a) Crk SH2 domain

	Tan		Swanson		Our PB	
MD5	0.693	0.719	0.635	0.705	0.609	0.759
MD7	0.507	0.651	0.379	0.632	0.502	0.646
MD9	0.580	0.616	**0.725**	**0.806**	**0.750**	0.773
MD11	**0.722**	0.706	0.589	0.792	0.554	0.734
MD13	0.633	0.696	0.583	0.727	0.532	0.804
MD15	0.644	**0.766**	0.529	0.742	0.592	**0.841**
Average	0.630	0.692	0.573	0.734	0.590	0.760

(b) Grb2 SH2 domain

	Tan		Swanson		Our PB	
MD5	**0.888**	**0.866**	**0.901**	**0.813**	**0.890**	0.764
MD7	0.816	0.719	0.698	0.755	0.862	0.734
MD9	0.811	0.783	0.723	0.772	0.845	0.640
MD11	0.867	0.832	0.753	**0.813**	0.854	**0.799**
MD13	0.635	0.627	0.673	0.741	0.863	0.660
MD15	0.668	0.558	0.717	0.653	0.717	0.609
Average	0.781	0.731	0.744	0.758	0.839	0.701

(c) Nck1 SH2 domain

	Tan		Swanson		Our PB	
MD5	0.414	0.473	0.665	0.733	0.653	0.613
MD7	0.568	0.621	0.679	0.728	0.528	0.661
MD9	0.571	0.586	0.699	0.706	0.638	0.625
MD11	0.470	0.605	0.747	0.790	0.689	0.725
MD13	0.686	0.660	0.666	0.717	0.648	0.642
MD15	**0.716**	**0.822**	**0.797**	**0.819**	**0.792**	**0.824**
Average	0.571	0.628	0.709	0.749	0.658	0.681

(d) Src SH2 domain

	Tan		Swanson		Our PB	
MD5	0.453	0.538	**0.620**	0.680	0.576	0.609
MD7	**0.605**	**0.606**	0.603	0.699	0.547	0.585
MD9	0.524	0.537	0.540	0.633	**0.610**	0.602
MD11	0.578	0.564	0.567	**0.703**	0.591	**0.693**
MD13	0.453	0.496	0.488	0.601	0.488	0.497
MD15	0.474	0.535	0.496	0.631	0.442	0.639
Average	0.515	0.546	0.552	0.657	0.542	0.604

Table 3-4 Screening Performance for nonbinding pYxxP sequences

The values of ROC AUC are listed. Left values in each PB method are ROC AUCs of screening not including the reorganization effects of peptides. Right values are those of screening including the reorganization effects of peptides. Best ROC AUC for each protocol is highlighted in **bold**.

Crk SH2 domain

	Tan		Swanson		Our PB	
MD5	0.558	0.450	0.572	0.635	0.760	0.557
MD7	0.504	0.484	0.482	0.473	0.486	0.467
MD9	0.552	0.486	0.620	0.537	0.573	0.500
MD11	0.554	0.527	0.490	0.500	0.427	0.493
MD13	0.440	0.440	0.628	0.654	0.709	0.592
MD15	0.484	0.598	0.550	0.561	0.476	0.554
Average	0.515	0.498	0.557	0.560	0.572	0.527

Conclusion

Molecular docking-based approaches on *in silico* screening have been developed for the design of drug like small molecules. Because peptides have different characteristics from drug like small molecules, these approaches cannot be applied for peptide design.

Molecular docking have two main purposes: prediction of binding conformations and binding affinities. These two purposes are accomplished using the scoring functions. We developed our docking program for peptide design. We incorporated molecular mechanics (MM) into scoring functions, because MM have been well studied using proteins and peptides. We also incorporated implicit solvent model (generalized born model) into our scoring functions because many polar functional groups of peptides require the precise descriptions of interactions with solvents. Our program showed high performances on prediction of binding conformations of the peptide to its receptor proteins. In addition, our program was accelerated by the GPGPU technology. We could process the computing for the molecular docking more than 100 times faster than a single core of CPU.

We also tried to improve an accuracy of the molecular mechanics and Poisson-Boltzmann surface area (MM-PBSA) method used in rescoring of binding affinities. We improved the accuracy of Poisson-Boltzmann (PB) implicit solvents by modifying PB radii, which are important parameter for PB calculations. Our PB method showed high performances on the estimation the polar contributions of solvation free energy of single molecules. We also demonstrated improved accuracies for prediction of binding affinities by MM-PBSA method.

Combining our improved methods showed high performances on peptide screening of several SH2 domains. We incorporated the reorganization effects of ligand molecules into docking-based approaches. The reorganization effects are effective for Crk, Nck1, and Src SH2 domains but less effective for Grb2 SH2 domain. These efficiencies may

relevant with the conformational flexibility of the ligand molecules in the bound state. We must take careful considerations whether the reorganization effects are included or not to docking-based approches. Molecular dynamics simulations are useful to determine the screening protocols such as positional restraints and the reorganization effects.

We showed the beneficial information for docking-based peptide screening through this study. Our two improved methods are first important steps for accurate prediction of binding affinities of peptides.

Acknowledgements

This thesis was supervised by Dr. Makoto Taiji in the Laboratory for Computational Molecular Design, RIKEN Quantitative Biology Center (QBiC) and the Department of Computational Biology, the University of Tokyo. I am deeply grateful to Dr. Noriaki Okimoto (RIKEN QBiC). He gave insightful comments and suggestions for my research. I would really like to thank all members in the Laboratory for Computational Molecular Design, RIKEN QBiC for their constructive comments and discussion.

Support was provided by the RIKEN Advanced Science Institute (ASI), RIKEN QBiC, the Protein 3000 Project. Main part of our calculations were performed by using the RIKEN Integrated Cluster of Clusters (RICC) and TSUBAME 2.0 and 2.5 at Global Scientific Information And Computing Center of Tokyo Institute of Technology. The development of GPU program was supported by IPA 未踏. This work was supported by Grant-in-Aid for JSPS Fellows.

References

1. Phan, J., et al., *Structure-based design of high affinity peptides inhibiting the interaction of p53 with MDM2 and MDMX*. The Journal of biological chemistry, 2010. **285**(3): p. 2174-83.
2. Long, Y.-Q., F.-D.T. Lung, and P.P. Roller, *Global optimization of conformational constraint on non-phosphorylated cyclic peptide antagonists of the Grb2-SH2 domain.* Bioorganic & Medicinal Chemistry, 2003. **11**(18): p. 3929-3936.
3. Coleman, et al., *Investigation of the Binding Determinants of Phosphopeptides Targeted to the Src Homology 2 Domain of the Signal Transducer and Activator of Transcription 3. Development of a High-Affinity Peptide Inhibitor.* Journal of medicinal chemistry, 2005. **48**(21): p. 6661-6670.
4. Porter, C., et al., *Grb7 SH2 domain structure and interactions with a cyclic peptide inhibitor of cancer cell migration and proliferation.* BMC Structural Biology, 2007. **7**(1): p. 58.
5. Rammensee, H.G., et al., *SYFPEITHI: database for MHC ligands and peptide motifs.* Immunogenetics, 1999. **50**(3-4): p. 213-219.
6. Singh, H. and G.P.S. Raghava, *ProPred: prediction of HLA-DR binding sites.* Bioinformatics, 2001. **17**(12): p. 1236-1237.
7. Donnes, P. and O. Kohlbacher, *SVMHC: a server for prediction of MHC-binding peptides.* Nucleic acids research, 2006. **34**(Web Server issue): p. W194-7.
8. *https://developer.nvidia.com/category/zone/cuda-zone*.
9. Thompson, D.C., C. Humblet, and D. Joseph-McCarthy, *Investigation of MM-PBSA Rescoring of Docking Poses.* Journal of chemical information and modeling, 2008. **48**(5): p. 1081-1091.
10. Kuhn, B. and P.A. Kollman, *Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of Their Relative Affinities by a Combination of Molecular Mechanics and Continuum Solvent Models.* Journal of medicinal chemistry, 2000. **43**(20): p. 3786-3791.
11. Kollman, P.A., et al., *Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models.* Accounts of Chemical Research, 2000. **33**(12): p. 889-897.
12. Dominguez, C., R. Boelens, and A.M.J.J. Bonvin, *HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information.* Journal of the American Chemical Society, 2003. **125**(7): p. 1731-

1737.

13. Chen, R., L. Li, and Z. Weng, *ZDOCK: An initial-stage protein-docking algorithm.* Proteins: Structure, Function, and Bioinformatics, 2003. **52**(1): p. 80-87.
14. Strynadka, N.C., et al., *Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase.* Nature structural biology, 1996. **3**(3): p. 233-9.
15. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Advanced Drug Delivery Reviews, 2001. **46**(1–3): p. 3-26.
16. Still, W.C., et al., *Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics.* Journal of the American Chemical Society, 1990. **112**(16): p. 6127-6129.
17. Onufriev, A., D. Bashford, and D.A. Case, *Exploring protein native states and large-scale conformational changes with a modified generalized born model.* Proteins-Structure Function and Bioinformatics, 2004. **55**(2): p. 383-394.
18. Ponder, J.W. and D.A. Case, *Force fields for protein simulations.* Advances in protein chemistry, 2003. **66**: p. 27-85.
19. Berman, H.M., et al., *The Protein Data Bank.* Nucleic acids research, 2000. **28**(1): p. 235-42.
20. Parsons, J., et al., *Practical conversion from torsion space to Cartesian space for in silico protein synthesis.* Journal of computational chemistry, 2005. **26**(10): p. 1063-1068.
21. Rahuel, J., et al., *Structural basis for specificity of Grb2-SH2 revealed by a novel ligand binding mode.* Nature structural biology, 1996. **3**(7): p. 586-9.
22. Zoetewey, D.L., et al., *Promiscuous Binding at the Crossroads of Numerous Cancer Pathways: Insight from the Binding of Glutaminase Interacting Protein with Glutaminase L.* Biochemistry, 2011. **50**(17): p. 3528-3539.
23. Straatsma, T.P. and J.A. McCammon, *Multiconfiguration thermodynamic integration.* The Journal of chemical physics, 1991. **95**(2): p. 1175-1188.
24. Case, D.A., et al., *AMBER*, 2012, University of California.
25. Jones, G., et al., *Development and validation of a genetic algorithm for flexible docking.* Journal of molecular biology, 1997. **267**(3): p. 727-748.
26. Donaldson, L.W., et al., *Structure of a regulatory complex involving the Abl SH3 domain, the Crk SH2 domain, and a Crk-derived phosphopeptide.* Proceedings of the National Academy of Sciences, 2002. **99**(22): p. 14053-14058.
27. Nioche, P., et al., *Crystal structures of the SH2 domain of grb2: highlight on the*

binding of a new high-affinity inhibitor. Journal of molecular biology, 2002. **315**(5): p. 1167-1177.

28. Lubman, O.Y. and G. Waksman, *Dissection of the energetic coupling across the src SH2 domain-tyrosyl phosphopeptide interface.* Journal of molecular biology, 2002. **316**(2): p. 291-304.
29. Korb, O., T. Stützle, and T.E. Exner, *Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS.* Journal of chemical information and modeling, 2009. **49**(1): p. 84-96.
30. Kuhn, B., et al., *Validation and Use of the MM-PBSA Approach for Drug Discovery.* Journal of medicinal chemistry, 2005. **48**(12): p. 4040-4048.
31. Belda, I., et al., *ENPDA: an evolutionary structure-based de novo peptide design algorithm.* Journal of computer-aided molecular design, 2005. **19**(8): p. 585-601.
32. Ji, F.-Q., et al., *Computational Design and Discovery of Conformationally Flexible Inhibitors of Acetohydroxyacid Synthase to Overcome Drug Resistance Associated with the W586L Mutation.* ChemMedChem, 2008. **3**(8): p. 1203-1206.
33. Sitkoff, D., K.A. Sharp, and B. Honig, *Accurate Calculation of Hydration Free-Energies Using Macroscopic Solvent Models.* Journal of Physical Chemistry, 1994. **98**(7): p. 1978-1988.
34. Swanson, J.M.J., S.A. Adcock, and J.A. McCammon, *Optimized radii for Poisson-Boltzmann calculations with the AMBER force field.* Journal of chemical theory and computation, 2005. **1**(3): p. 484-493.
35. Swanson, J.M.J., et al., *Optimizing the Poisson dielectric boundary with explicit solvent forces and energies: Lessons learned with atom-centered dielectric functions.* Journal of chemical theory and computation, 2007. **3**(1): p. 170-183.
36. Tan, C.H., L.J. Yang, and R. Luo, *How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis.* Journal of Physical Chemistry B, 2006. **110**(37): p. 18680-18687.
37. Grant, J.A., B.T. Pickup, and A. Nicholls, *A smooth permittivity function for Poisson-Boltzmann solvation methods.* Journal of Computational Chemistry, 2001. **22**(6): p. 608-640.
38. Im, W., D. Beglov, and B. Roux, *Continuum Solvation Model: computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation.* Computer Physics Communications, 1998. **111**(1-3): p. 59-75.
39. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N ·log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.

40. Beglov, D. and B. Roux, *Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations.* The Journal of chemical physics, 1994. **100**(12): p. 9050.
41. Artymiuk, P.J., et al., *The structures of the monoclinic and orthorhombic forms of hen egg-white lysozyme at 6 Å resolution.* Acta Crystallographica Section B, 1982. **38**(3): p. 778-783.
42. Jelsch, C., et al., *Accurate protein crystallography at ultra-high resolution: Valence electron distribution in crambin.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(7): p. 3171-3176.
43. *Molecular Operating Environment (MOE)*, 2013, Chemical Computing Group Inc.
44. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chemical Physics Letters, 1999. **314**(1–2): p. 141-151.
45. Honda, S., et al., *10 Residue Folded Peptide Designed by Segment Statistics.* Structure, 2004. **12**(8): p. 1507-1518.
46. Li, C., et al., *Systematic mutational analysis of peptide inhibition of the p53-MDM2/MDMX interactions.* Journal of molecular biology, 2010. **398**(2): p. 200-13.
47. Pazgier, M., et al., *Structural basis for high-affinity peptide inhibition of p53 interactions with MDM2 and MDMX.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(12): p. 4665-70.
48. Roux, B. and T. Simonson, *Implicit solvent models.* Biophysical chemistry, 1999. **78**(1-2): p. 1-20.
49. Narumi, T., et al. *A 185 Tflops simulation of amyloid-forming peptides from Yeast Prion Sup35 with the special-purpose computer System MD-GRAPE3.* in *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*. 2006. Tampa, U.S.
50. Baker, N.A., et al., *Electrostatics of nanosystems: application to microtubules and the ribosome.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(18): p. 10037-41.
51. Srinivasan, J., et al., *Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices.* Journal of the American Chemical Society, 1998. **120**(37): p. 9401-9409.
52. Kongsted, J. and U. Ryde, *An improved method to predict the entropy term with the MM/PBSA approach.* Journal of computer-aided molecular design, 2009. **23**(2): p. 63-71.
53. Zhou, Z., M. Bates, and J.D. Madura, *Structure modeling, ligand binding, and*

binding affinity calculation (LR-MM-PBSA) of human heparanase for inhibition and drug design. Proteins: Structure, Function, and Bioinformatics, 2006. **65**(3): p. 580-592.

54. Marcotte, I., et al., *A Multidimensional 1H NMR Investigation of the Conformation of Methionine-Enkephalin in Fast-Tumbling Bicelles.* Biophysical journal, 2004. **86**(3): p. 1587-1600.
55. Fisher, A., P.B. Laub, and B.S. Cooperman, *NMR structure of an inhibitory R2 C-terminal peptide bound to mouse ribonucleotide reductase R1 subunit.* Nature structural biology, 1995. **2**(11): p. 951-5.
56. Spyroulias, G.A., et al., *Comparison of the solution structures of angiotensin I & II. Implication for structure-function relationship.* European journal of biochemistry / FEBS, 2003. **270**(10): p. 2163-73.
57. Phan-Chan-Du, A., et al., *Structure of Antibody-Bound Peptides and Retro−Inverso Analogues. A Transferred Nuclear Overhauser Effect Spectroscopy and Molecular Dynamics Approach†,‡.* Biochemistry, 2001. **40**(19): p. 5720-5727.
58. Bhattacharjya, S., et al., *High-Resolution Solution Structure of a Designed Peptide Bound to Lipopolysaccharide: Transferred Nuclear Overhauser Effects, Micelle Selectivity, and Anti-Endotoxic Activity†,‡.* Biochemistry, 2007. **46**(20): p. 5864-5874.
59. Wang, M., L. Shan, and J. Wang, *Two peptide fragments G55–I72 and K97–A109 from staphylococcal nuclease exhibit different behaviors in conformational preferences for helix formation.* Biopolymers, 2006. **83**(3): p. 268-279.
60. Jung, H.H., et al., *Structural Analysis of Immunotherapeutic Peptides for Autoimmune Myasthenia Gravis†,‡.* Biochemistry, 2007. **46**(51): p. 14987-14995.
61. Steinmetz, M.O., et al., *Molecular basis of coiled-coil formation.* Proceedings of the National Academy of Sciences, 2007. **104**(17): p. 7062-7067.
62. Araki, M. and A. Tamura, *Transformation of an alpha-helix peptide into a beta-hairpin induced by addition of a fragment results in creation of a coexisting state.* Proteins, 2007. **66**(4): p. 860-8.
63. Neidigh, J.W., R.M. Fesinmeyer, and N.H. Andersen, *Designing a 20-residue protein.* Nature structural biology, 2002. **9**(6): p. 425-30.
64. Page, N., et al., *The spliceosomal phosphopeptide P140 controls the lupus disease by interacting with the HSC70 protein and via a mechanism mediated by gammadelta T cells.* PloS one, 2009. **4**(4): p. e5273.
65. Zahn, R., *The octapeptide repeats in mammalian prion protein constitute a pH-*

dependent folding and aggregation site. Journal of molecular biology, 2003. **334**(3): p. 477-88.

66. Vidal, M., V. Gigoux, and C. Garbay, *SH2 and SH3 domains as targets for anti-proliferative agents.* Critical Reviews in Oncology/Hematology, 2001. **40**(2): p. 175-186.
67. Cody, W.L., et al., *Progress in the development of inhibitors of SH2 domains.* Current pharmaceutical design, 2000. **6**(1): p. 59-98.
68. Nishihara, H., et al., *Molecular and immunohistochemical analysis of signaling adaptor protein Crk in human cancers.* Cancer letters, 2002. **180**(1): p. 55-61.
69. Pendergast, A.M., et al., *BCR-ABL-induced oncogenesis is mediated by direct interaction with the SH2 domain of the GRB-2 adaptor protein.* Cell, 1993. **75**(1): p. 175-85.
70. Janes, P.W., et al., *Activation of the Ras signalling pathway in human breast cancer cells overexpressing erbB-2.* Oncogene, 1994. **9**(12): p. 3601-8.
71. Davidson, J.P., et al., *Calorimetric and Structural Studies of 1,2,3-Trisubstituted Cyclopropanes as Conformationally Constrained Peptide Inhibitors of Src SH2 Domain Binding.* Journal of the American Chemical Society, 2001. **124**(2): p. 205-215.
72. Migliaccio, A., et al., *Inhibition of the SH3 domain-mediated binding of Src to the androgen receptor and its effect on tumor growth.* Oncogene, 2007. **26**(46): p. 6619-6629.
73. Gilmer, T., et al., *Peptide inhibitors of src SH3-SH2-phosphoprotein interactions.* The Journal of biological chemistry, 1994. **269**(50): p. 31711-9.
74. Songyang, Z., et al., *SH2 domains recognize specific phosphopeptide sequences.* Cell, 1993. **72**(5): p. 767-78.
75. Liu, B.A., et al., *SH2 domains recognize contextual peptide sequence information to determine selectivity.* Molecular & cellular proteomics : MCP, 2010. **9**(11): p. 2391-404.
76. Frese, S., et al., *The Phosphotyrosine Peptide Binding Specificity of Nck1 and Nck2 Src Homology 2 Domains.* Journal of Biological Chemistry, 2006. **281**(26): p. 18236-18245.
77. Kang, L., et al., *An improved adaptive genetic algorithm for protein-ligand docking.* Journal of computer-aided molecular design, 2009. **23**(1): p. 1-12.
78. Homeyer, N., et al., *AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine.* Journal of Molecular Modeling, 2006.

12(3): p. 281-289.
79. Bondi, A., *van der Waals Volumes and Radii.* The Journal of Physical Chemistry, 1964. **68**(3): p. 441-451.
80. Cross, J.B., et al., *Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy.* Journal of chemical information and modeling, 2009. **49**(6): p. 1455-1474.