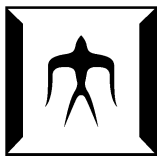
論文 / 著書情報
Article / Book Information

題目(和文)	立体構造情報に基づくタンパク質間相互作用ネットワーク予測
Title(English)	Protein-Protein Interaction Network Prediction Based on Tertiary Structure Data
著者(和文)	大上雅史
Author(English)	Masahito Ohue
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9553号, 授与年月日:2014年3月26日, 学位の種別:課程博士, 審査員:秋山　泰,佐藤　泰介,関嶋　政和,瀬々　潤,杉山　将
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9553号, Conferred date:2014/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Protein–Protein Interaction Network Prediction Based on Tertiary Structure Data

Masahito Ohue



Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology

Supervisor: Yutaka Akiyama

A Thesis Submitted for the Degree of *Doctor of Engineering*

February 21, 2014

This dissertation is dedicated
to my wife and my parents

Abstract

Protein–protein interactions (PPIs) are fundamental in the majority of cellular processes and their study is of enormous biotechnological and therapeutic interest. The computational prediction for elucidation of PPI networks is crucial in biological fields. However, the development of an effective method to conduct exhaustive PPI screening represents a computational challenge.

In this dissertation, we proposed a novel PPI network prediction system called MEGADOCK based on protein–protein docking calculation with protein tertiary structure information. MEGADOCK reduced the calculation time required for docking by using new score functions, rPSC and RDE, and was implemented on recent parallel high-performance computing environments by employing a hybrid parallelization with MPI and OpenMP and general-purpose graphics processing unit technique.

We showed that MEGADOCK is capable of exhaustive PPI screening and completed docking calculations 9.8 times faster than the conventional method (Mintseris, *et al.* 2007) while maintaining an acceptable level of accuracy. When MEGADOCK was applied to a subset of a general benchmark dataset to predict 120 relevant interacting pairs from 14,400 protein combinations, an F-measure value of 0.231 was obtained. Moreover, the system was scalable as shown by measurements carried out on two supercomputing environments, TSUBAME 2.0 and K computer.

It is now feasible to search and analyze PPIs while taking into account three-dimensional structures at the interactome scale. We demonstrated the applications to pathway analyses, bacterial chemotaxis, human apoptosis, and RNA binding proteins by using our system. As an example of the results, when analyzing the positive predictions of bacterial chemotaxis pathway from MEGADOCK, all the core signaling interactions were correctly predicted with the exception of interactions activated by protein phosphorylation.

Large-scale PPI prediction using tertiary structures is an effective approach that has a wide range of potential applications. This method is especially useful for identifying novel PPIs of new pathways that control cellular behavior.

Acknowledgements

I would like to express my sincere gratitude to Professor Yutaka Akiyama for providing me an excellent research environment and continuous guidance and encouragement throughout my research work. I also thank Assistant Professor Takashi Ishida for his valuable comments and discussions on my research. Dr. Yuri Matsuzaki is one of the most important co-authors in my research articles. She gave insightful comments and suggestions. Assistant Professor Nobuyuki Uchikoga at Chuo University also provided valuable suggestions and supports. I also thank all members of Professor Akiyama's laboratory for helpful discussions and for the wonderful time spent together. I wish to express my appreciation to Mr. Takehiro Shimoda and Mr. Takayuki Fujiwara for considerable suggestions to my research work. I would also like to thank Ms. Kanako Ozeki, the secretary of the Akiyama laboratory, for her much appreciated support during my six years of laboratory life.

For investigating the human apoptosis pathway estimation (Chapters 6 and 9), Ms. Saliha Ece Ozbabacan at Koç University, Istanbul, Turkey provided the results of PRISM experiments with details and Dr. Vachiranee Limviphuvadh at the Agency for Science, Technology and Research (A*STAR), Singapore provided integrated PPI database information. For the study on protein-RNA interaction predictions (Chapter 7), I thank Dr. Junichi Iwakiri at University of Tokyo for providing the protein–RNA complex dataset and information on related study. I also would like to thank Associate Professor Daisuke Kihara at Purdue University, West Lafayette, USA for helpful discussions about my research related to protein–protein interactions. I wish to express my gratitude to my dissertation committee, Professor Taisuke Sato, Associate Professor Masashi Sugiyama, Associate Professor Masakazu Sekijima and Associate Professor Jun Sese at the Tokyo Institute of Technology for their inspiring feedback and valuable comments on my research and the dissertation. Associate Professor Fumikazu Konishi and Associate Professor Hiroyuki Ogata both at the Tokyo Institute of Technology also gave valuable suggestions to my research.

I would like to express my sincerest appreciation to Japanese Society for the Pro-

motion of Science (JSPS) Research Fellow (DC1), JSPS Global COE program "CompView" led by Professor Osamu Watanabe at Tokyo Institute of Technology, The Next-Generation Integrated Life Simulation Software (ISLiM) project led by Professor Koji Kaya at RIKEN, and Ministry of Education, Culture, Sports, Science and Technology (MEXT) Program for Leading Graduate Schools "Education Academy of Computational Life Sciences (ACLS)" led by Professor Yutaka Akiyama, for financial support throughout my research works. Moreover, this study was supported in part by a Grant-in-Aid for Scientific Research (A) 24240044, a Grant-in-Aid for Scientific Research (B) 19300102, and a Grant-in-Aid for JSPS Fellows 23·8750, all of which were from the MEXT Program. Some of the results were obtained by using the K computer at the RIKEN Advanced Institute for Computational Science (AICS) through early access and access granted as a High Performance Computing Infrastructure (HPCI) Systems Research Program (proposal number hp120131).

I was provided with valuable discussions and encouragement by four young researchers' communities (wakate-no-kai), Young Researches Society for Bioinformatics (bioinfowakate), Young Researches Society for Biophysics (bpwakate), Young Researches Society for Biochemistry (seikawakate), and Study Meeting on Theory of Protein and Nucleic Acid Structure. I would like to thank all the young researchers who belong to the young researchers' communities.

Last but certainly not the least, I would like to extend a special thanks to my wife, Kana Ohue, and to my parents, Shunsuke Ohue and Namiko Ohue. Without their kind support and constant encouragement, this work would have not been possible. Finally, I dedicate this work to my wife and parents.

Contents

List of Figures

List of Tables

Part I

General Introduction

Chapter 1

Introduction

1.1 Protein–Protein Interaction (PPI)

In the field of life sciences and medical/pharmaceutical sciences, elucidation of regulatory relationships among the millions of protein combinations that function in living cells is crucial for understanding the mechanisms underlying diseases and for the development of medicines [1]. Predicting protein–protein interaction (PPI) networks at the genome scale is one of the main topics of interest in systems biology [2].

PPIs have been extensively investigated from the perspectives of biochemistry, quantum chemistry, and molecular dynamics. Several methods to determine PPIs have been developed. One of the main goals of proteome and interactome analyses is to identify proteins with the potential to bind and interact with each other; this is called PPI screening. High-throughput but noisy biological experiments, such as the yeast two-hybrid system [3], and precise but low-throughput methods, such as fluorescence resonance energy transfer [4], have been frequently used as experimental methods for PPI screening.

There are also computational methods for PPI prediction [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. Some successful methods include those based on protein sequences [5, 6, 7], evolutionary information [8, 9], and domain interaction information [10, 11]. Because protein structure provides fundamental information about function, computational PPI screening methods based on the known structures of protein complexes are also being considered [12, 13, 14]. Tertiary structural information also provides powerful features for recognition [15, 16], and is therefore useful for predicting binding affinity [17] in protein–protein complexes. However, the performance of these computational methods is highly dependent on known PPI information. These methods only detect interacting

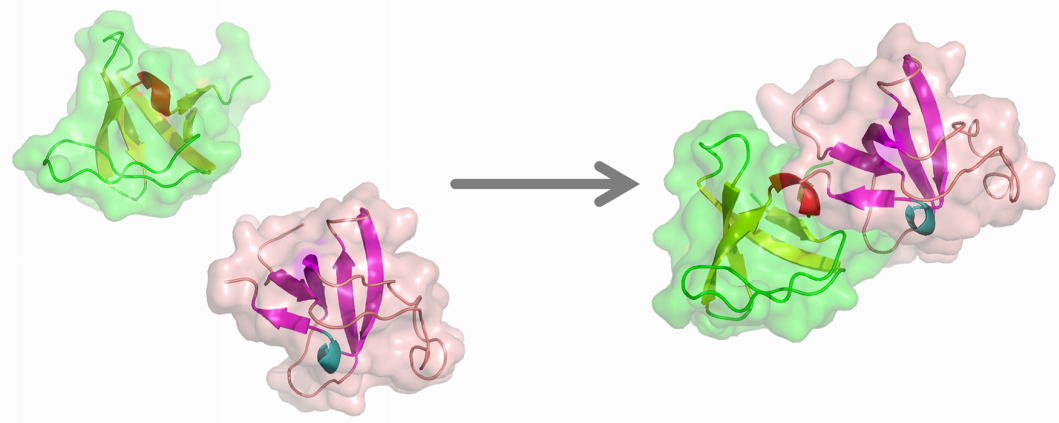


Figure 1.1: Protein-protein docking between two proteins (generated using PyMOL [21])

protein pairs resembling those of known protein complexes. Therefore, they do not completely reflect the structural basis of PPIs.

In structural biology, computational methods such as atomic-level molecular-dynamics simulations have been primarily applied to analyze in detail the mechanisms of individual protein interactions based on the physical behavior of atoms [18]. However, these methods are not applicable to the large-scale analyses required in systems biology, because the analyses are computationally expensive to perform. To fully utilize many protein tertiary structures deposited in the public database [19] that has continued to increase in recent years [20], we focused our attention on a rigid-body protein–protein docking method.

1.2 Rigid-Body Protein–Protein Docking

Rigid-body protein–protein docking is one of the effective solutions to predict a large-scale PPI network in realistic computation time (Fig. 1.1). Since PPIs mostly provoke conformational changes and treat protein structures with less flexibility, rigid-body protein–protein docking cannot conduct accurate calculations. Nevertheless, rigid-body docking can be calculated much faster than other methods that allow structural flexibility. It is by far the only effective method to introduce structural data for analysis at the proteome scale.

Rigid-body protein–protein docking methods have been applied as the initial stage for small-scale PPI network prediction [22, 23, 24]. Besides providing a useful technique

to help study fundamental biomolecular mechanisms, docking tools to predict PPIs are emerging as promising complementary approaches to rational drug design [25].

Rigid-body protein–protein docking has been implemented in various ways, including fast Fourier transform (FFT) convolution of 3D voxel space as proposed by Katchalski-Katzir [26] (MolFit [26, 27], FTDock [28], PIPER [29], ZDOCK [30, 31, 32, 33, 34], and pyDock [35, 36]), and others consider shape complementarity of local surface structure (PatchDock [37], LZerD [38], and Hex [39, 40]). RosettaDock [41, 42], BiGGER [43], FireDock [44], FiberDock [45], and EigenHex [46] take flexibility of main- and side-chains into account. Some of these flexible docking methods have successfully predicted protein complexes of targets used in the protein-complex structure prediction community-wide experiment called critical assessment of prediction of interactions (CAPRI). CAPRI is a blind prediction competition that does not release the structure of the protein complex judged by CAPRI assessors until after the submission of a target [47, 48, 49, 50]. However, the rigid-body docking methods are still used in situations such as pre-processing for considering flexibility, required calculation speed, and application to a large-scale problem.

Wass, *et al.* reported that the score distribution generated by the rigid-body protein–protein docking tool Hex showed significant difference between known interacting pairs and non-binding pairs when they used only shape complementarity for the scoring function [22]. Nonetheless, more investigation is required on the features of the computational methods, such as the scoring functions that best fit the problem and parameter spaces that produce predictions. Here, we propose a novel score function for rigid-body docking by taking into account electrostatic forces as well as shape complementarity. Such docking-based prediction of PPI has an advantage because it also produces several candidates for presumable docking poses. This provides insight into how the two predicted proteins undergo interactions according to their structural properties.

ZDOCK [30, 31, 32, 33] has been by far the most successful among the rigid docking tools [51]. ZDOCK employs voxel models in which protein complexes are divided into three-dimensional (3D) voxels and scored by the correlation functions of each discrete function. The ZDOCK scoring function comprises pairwise shape complementarity (PSC), electrostatics, and interface atomic contact energy score (IFACE) [33] for estimating desolvation free energy; in total, eight correlation functions are calculated by FFT. Generally, FFT-based docking tools that search the entire 3D grid space for presumable docking positions perform better than local search-based tools. With more correlation functions, it is possible to incorporate more features to evaluate docking pose, although the number of the correlation functions linearly affects calculation speed.

Matsuzaki, *et al.* applied ZDOCK to PPI screening and predicted whether two proteins interact by analyzing the high-scoring decoys produced by a rigid docking process [23]. Yoshikawa, *et al.* also developed a PPI screening method and used ZDOCK and their original post-docking process called affinity evaluation and prediction (AEP) [24]. However, to search the entire interactome space using these methods involves combinations of 1,000 proteins (1 Mega combinations). Thus, ZDOCK has limitations regarding computation time, increasing its flexibility is also unrealistic. Therefore, increasing the speed of rigid-body docking calculations is crucial.

1.3 High–Performance Computing

To realize large-scale PPI network prediction using tertiary structures, efficient execution by supercomputing environments is crucial. In recent years, the field of high-performance computing has been rapidly evolving. For example, Japan has powerful supercomputers such as the K computer [52] at RIKEN and TSUBAME 2.5 [53] at Tokyo Institute of Technology ranked 4th and 11th in the TOP500 list, respectively, in November 2013 [54]. Fully utilizing these large scale calculation environments makes large-scale PPI network prediction of the proteome scale possible.

In addition, the performance gained by use of accelerators has also attracted attention in recent years. The number of supercomputers equipped with accelerators, such as the graphics processing unit (GPU) of NVIDIA and many integrated core (MIC) architecture of Intel, is increasing [54]. An advantage of GPUs is that they consume power more efficiently. In the Green500 list (November 2013) [55] that ranks the TOP500 supercomputers by Flops/Watt, the top 10 machines were all equipped with NVIDIA Tesla GPUs. Taking advantage of the acceleration features available with these accelerators is important to fully utilize the supercomputers that will evolve in the future.

1.4 Purpose of Study

In the present study, we describe the development of a rigid-body docking-based method for PPI screening based on exhaustive calculations of pseudo-binding energies among pairs of target proteins that can be applied to PPI prediction problems of megaorder data. To enable applications to 1 megaorder combinations, we developed efficient FFT-based protein–protein docking software called MEGADOCK that is exe-

cutable on current supercomputing environments and makes it possible to conduct exhaustive PPI screening. MEGADOCK searches the relevant interacting protein pairs by conducting protein–protein docking between the tertiary structures of the target proteins and then analyzes the distributions of high-scoring decoys (candidate protein complexes).

Applications of MEGADOCK to real biological PPI network predictions are also one of the purposes of this study. We apply MEGADOCK to several pathway reconstruction problems, and then we evaluate our prediction performance and detect new PPI candidates for enrichment of known biological pathways.

1.5 Summary of Contributions

The contributions of this thesis are classified into three categories (i) development of a novel protein–protein docking method that is 9 times faster with the same level of accuracy than a conventional tool, (ii) parallelization and acceleration of PPI prediction calculations compatible with modern supercomputing environments, and (iii) broader applications of the proposed system to real biological networks. We now describe these in more detail.

- We proposed a novel shape complementarity score function called real Pairwise Shape Complementarity (rPSC) for FFT-based rigid-body protein–protein docking calculations. The rPSC function that uses only real number representations for shape complementarity was correlated with a conventional score function represented by a complex number. We also proposed a novel desolvation free energy function called Receptor Desolvation Free Energy (RDE). Therefore, it is possible to calculate a total energy score that includes shape complementarity, electrostatic interactions and desolvation effects with only one FFT correlation. As a result, the proposed method was shown to be 9.8 times faster than the conventional tool ZDOCK 3.0 while maintaining acceptable docking prediction accuracies.

- We implemented our protein–protein docking method to be suitable for running on supercomputers by using hybrid parallelization with Message Passing Interface (MPI) and Open Multi-Processing (OpenMP), where a number of docking processes are distributed among the nodes by MPI with each docking process that is also calculated in parallel by threads using OpenMP within one node. This implementation has significant advantages that (i) save memory space and

(ii) avoid a large overhead because of handling data communication on numerous core systems such as the K computer running a flat MPI implementation. As a result, we obtained a strong scaling value that is a type of evaluation value for parallel efficiency, of over 0.95 out of a maximum of 1.00 in both the K computer and TSUBAME 2.0.

- We enabled the use of recent computing systems by taking advantage of GPU features. We implemented not only FFT calculations but also generated grid (voxelization) and rotation of protein structures on GPUs to reduce the cost of data transfers. As a result, the system achieved 13.9-fold acceleration using 1 CPU core and 1 GPU, and 37.0-fold acceleration using 12 CPU cores and 3 GPUs by making full use of heterogeneous computing resources.

- We developed the MEGADOCK system for exhaustive PPI screening, that conducts protein–protein docking and post-analysis with reranking technique on protein tertiary structural data. For the detection of the relevant interacting protein pairs, we obtained better accuracy than the prediction without reranking technique. when our method was applied to a subset of a general benchmark dataset.

- We performed real applications in the field of systems biology. In this study, we applied MEGADOCK to (i) a bacterial chemotaxis pathway and (ii) a human apoptosis pathway to reconstruct pathways and determine unknown interactions. In the chemotaxis pathway analysis, all core signaling interactions were correctly predicted with the exception of interactions activated by protein phosphorylation. In the apoptosis pathway analysis, the prediction results included several new PPI candidates that might be suitable targets for drug discovery.

- We compared MEGADOCK with other structure-based PPI screening tools: (i) ZDOCK [33] that has similar scoring functions to MEGADOCK and (ii) PRISM [14] that is a template-based PPI prediction tool. The predicted interactions generated from MEGADOCK and ZDOCK in chemotaxis pathway analysis were slightly different; however when the positive predictions from both tools were combined, the vast majority of relevant interactions were represented. Indeed, there were only two exceptions, both requiring phosphorylation to activate the corresponding interaction. The consensus between template-based and non-template-based methods successfully predicted the PPI network more accurately than the conventional single template-/non-template-based methods. Because such precise prediction reduces biological screening costs, it should further

promote interactome analysis.

1.6 Thesis Organization

The remaining chapters of this thesis are organized as follows: Chapter 2 reviews the protein–protein docking study focusing mainly on FFT-based rigid-body protein–protein docking methods. Chapter 3 describes a new protein–protein docking method, called MEGADOCK, with a novel shape complementarity model called rPSC and simple hydrophobic interaction model. Chapter 4 presents a new PPI prediction method by using our protein–protein docking tool and its application to pathway analyses. Case studies on specific pathways are described in Chapter 5 for bacterial chemotaxis, and in Chapter 6 for human apoptosis. In Chapter 7, we apply our PPI prediction method to protein-RNA interaction predictions by extending atomic parameters for ribonucleic molecules. In Chapters 8 and 9, we discuss the combination of our method and other structure-based information. In Chapter 8, we apply two different rigid-body docking tools, MEGADOCK and ZDOCK [33], with different scoring models. In Chapter 9, we combine a template-based PPI prediction tool (PRISM [14]) and a non-template-based PPI prediction tool (MEGADOCK). Conclusions are presented in Chapter 10 together with future work and discussion. In addition, we report MEGADOCK with high-performance computing in Appendices A and B. Appendix A describes our implementation by MPI/OpenMP hybrid parallelization and execution results on two supercomputing environments, K computer and TSUBAME. Appendix B reports GPU implementation and execution results using TSUBAME GPU computing.

This thesis is based on the following publications by the author: [56, 57, 58, 59, 60, 61, 62, 63].

Part II

Protein–Protein Docking

Chapter 2

Overview of Protein–Protein Docking

2.1 Introduction

Practically every process in the living cell requires molecular recognition and formation of complexes that may be stable or transient assemblies of two or more molecules with one molecule acting on the other, or may be promoting intra- and inter-cellular communication, or representing permanent oligomeric ensembles [27]. The rapid accumulation of data on protein–protein interactions, protein sequences, and tertiary structures requires the development of advanced computational methods to help in our understanding of living cells. One of the methods involves the prediction of the protein complex structure from its components. Typically protein–protein docking methods are investigated in an attempt to predict the protein complex structures given the protein structures of components. Over the past 30 years, many docking approaches have been proposed, ranging from thermodynamic approaches to correlation approaches and from rigid-body docking to flexible docking [64, 65].

Docking algorithms operate on the atomic coordinates of two individual proteins usually considered as rigid bodies and generate a large number of candidate association models between them. These candidates are then ranked by using various scoring functions, used independently or in combination. The scoring functions generally include geometric and chemical complementarities measures, electrostatics, hydrogen-bonding interactions, van der Waals interactions, and some empirical potential functions. A number of algorithms and many different scoring functions have been developed in the last 20 years, as recently reviewed by Eisenstein, *et al.* (2004) [27], Ritchie, *et*

al. (2008) [66], Janin, *et al.* (2010) [67], Vakser, *et al.* (2013) [68] and Vajda, *et al.* (2013) [65], and the field has become extremely active.

2.2 Rigid-Body Protein–Protein Docking Approach

In the rigid-body docking approaches, the proteins are considered rigid and this inflexibility is taken into account. Here, an overview is provided to describe the different steps involved in rigid-body protein–protein docking:

1. First, we start with the simulated 3D structures of the two unbound component proteins. Assuming that the formed complex has limited conformational changes, the two component proteins are regarded as rigid bodies.

2. A 3D rotational and 3D translational search (6D search) is performed over all possible associations because in most cases of unbound-unbound complexes there is no biological information regarding what parts of the proteins will interact. This search will sample the space of all possible associations and consequently there will be a lower limit applied to the difference in conformations between two docked predicted complexes that determine the global solution of the search procedure.

3. A large number of different complexes (decoys) are generated after the global search procedure. Then a function is developed to score the quality of these decoys. At this stage, geometric and electrostatic complementarity are often used because it is very fast to compute. Ideally, the docking algorithm will then identify several complexes that are close to the native complex based on these complexes having the best scores.

4. Then a reranking of the resultant complexes may be undertaken possibly using computationally intensive calculations. Finally, conformational flexibility may be introduced into the algorithm to refine the few remaining decoys when there are only a limited number of complexes to consider.

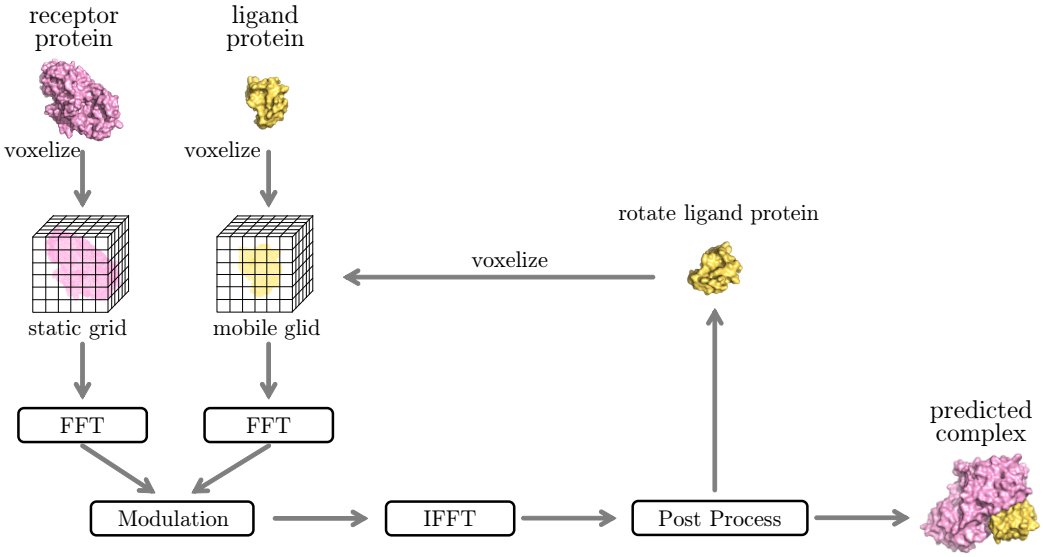


Figure 2.1: Typical FFT-based protein–protein docking procedure using the Katchalski-Katzir algorithm.

2.3 FFT-based Rigid-Body Protein–Protein Docking

In the first step of many docking methods, an attempt is made to represent the protein structures in an efficient manner. One of the major methods is the Katchalski-Katzir algorithm by Katchalski-Katzir, *et al.* (1992) [26], that applies a 3D grid representation and FFT correlation approach. Fig. 2.1 illustrates the procedure followed by the Katchalski-Katzir algorithm for protein–protein docking. In this method, the protein structure is projected onto a 3D grid. The pseudo interaction energy score (called the docking score) S between two proteins (here we call them the "receptor" and "ligand", apart from the typical biological definition, to indicate two docked proteins) is calculated by discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT) using the correlation of two discrete functions, as follows:

$$S(\alpha,\beta,\gamma) = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n)L(l+\alpha,m+\beta,n+\gamma) \tag{2.1}$$

$$= \mathrm{IDFT}[\mathrm{DFT}[R(l,m,n)]^{*}\mathrm{DFT}[L(l,m,n)]] \tag{2.2}$$

where R and L are the discrete score function of the Receptor (R) and Ligand (L) proteins, respectively, (l, m, n) is a coordinate in the 3D grid space, and (α, β, γ) is the parallel translation vector of the ligand protein. The asterisk operator * indicates the complex conjugate of a complex number. DFT and IDFT are defined below:

$$\begin{aligned}\mathrm{DFT}[R(l,m,n)] &= \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n)\exp\left(\frac{-2\pi i(lo+mp+nq)}{N}\right) \qquad (2.3)\\ &= \mathcal{R}(o,p,q),\\ \mathrm{IDFT}[\mathcal{R}(o,p,q)] &= \frac{1}{N^3}\sum_{o=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} \mathcal{R}(o,p,q)\exp\left(\frac{2\pi i(lo+mp+nq)}{N}\right) \qquad (2.4)\\ &= R(l,m,n)\end{aligned}$$

Proof Apply the equation for DFT to both sides of eq. (2.1), then

$$\begin{aligned}&\mathrm{DFT}[S(\alpha,\beta,\gamma)]\\ &= \mathrm{DFT}\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n)L(l+\alpha,m+\beta,n+\gamma)\right]\\ &= \sum_{\alpha=1}^{N}\sum_{\beta=1}^{N}\sum_{\gamma=1}^{N}\left(\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n)L(l+\alpha,m+\beta,n+\gamma)\right)\exp\left(\frac{-2\pi i(\alpha o+\beta p+\gamma q)}{N}\right)\\ &= \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n)\left(\sum_{\alpha=1}^{N}\sum_{\beta=1}^{N}\sum_{\gamma=1}^{N} L(l+\alpha,m+\beta,n+\gamma)\times\right.\\ &\qquad \left.\exp\left(-\frac{2\pi i\{(l+\alpha)o+(m+\beta)p+(n+\gamma)q\}}{N}\right)\exp\left(-\frac{2\pi i\{(-l)o+(-m)p+(-n)q\}}{N}\right)\right)\\ &= \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n)\exp\left(-\frac{2\pi(-i)\{lo+mp+nq\}}{N}\right)\times\\ &\qquad \sum_{\alpha=1}^{N}\sum_{\beta=1}^{N}\sum_{\gamma=1}^{N} L(l+\alpha,m+\beta,n+\gamma)\exp\left(-\frac{2\pi i\{(l+\alpha)o+(m+\beta)p+(n+\gamma)q\}}{N}\right)\\ &= \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n)\exp\left(-\frac{2\pi i\{lo+mp+nq\}}{N}\right)^{*}\mathrm{DFT}[L(l,m,n)]\\ &= \mathrm{DFT}[R(l,m,n)]^{*}\mathrm{DFT}[L(l,m,n)]\end{aligned}$$

□

To find the best docking poses, possible ligand orientations are exhaustively examined at n_θ rotation angles for a given stepsize θ. For each rotation, the ligand protein is translated into $N \times N \times N$ patterns in the $\mathbb{N}^3$ grid space (where $N = |\mathbb{N}|$ is the grid size

in each dimension). The decoy that yields the highest value of S for each rotation is recorded. In this manner, a total of $n_\theta \times N^3$ docking poses are evaluated for one protein pair. To directly execute the simple convolution sums in eq. (2.1), $\mathcal{O}(N^6)$ calculations are required; however, this is reduced to $\mathcal{O}(N^3 \log N)$ using the FFT in eq. (2.2).

2.4 Scoring Function $R(l, m, n)$ and $L(l, m, n)$

The Katchalski-Katzir algorithm has been further developed by several authors ([28, 29, 30, 32, 33, 69, 70, 71, 72, 73]) especially in terms of scoring functions.

2.4.1 Shape complementarity function

Katchalski-Katzir score

The original scoring function by Katchalski-Katzir, *et al.* [26] is based on the shape complementarity. The scoring functions are given below.

$$R_{\mathrm{KK}}(l,m,n) = \begin{cases} 1 & \text{(surface voxel)} \\ \rho & \text{(interior voxel)} \\ 0 & \text{(otherwise)} \end{cases} \tag{2.5}$$

$$L_{\mathrm{KK}}(l,m,n) = \begin{cases} 1 & \text{(surface voxel)} \\ \delta & \text{(interior voxel)} \\ 0 & \text{(otherwise)} \end{cases} \tag{2.6}$$

$$S_{\mathrm{KK}}(\alpha,\beta,\gamma) = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n)L(l+\alpha, m+\beta, n+\gamma) \tag{2.7}$$

Thus, for the receptor protein, surface grid points are given the value 1, those in the interior are given the value ρ (usually -15), and grid points outside the protein are given a value of 0. For the ligand protein, grid points on the surface are given the value 1, interior grid points are given the value δ (usually 1), and grid points outside the protein are given a value of 0.

Pairwise Shape Complementarity (PSC)

Chen, *et al.* proposed another shape complementarity score called PSC [31] that computes the total number of receptor-ligand atom pairs within a distance cutoff, minus

a geometric clash penalty. PSC uses a complex function representation as follows:

$$\Re[R_{\mathrm{PSC}}(l,m,n)] = \begin{cases} \#\text{ of receptor atoms within } (3.6\text{ Å} + r_{\mathrm{vdW}}) & \text{(open space)} \\ 0 & \text{(otherwise)} \end{cases} \tag{2.8}$$

$$\Im[R_{\mathrm{PSC}}(l,m,n)] = \begin{cases} 3 & \text{(solvent excluding surface of the receptor)} \\ 9 & \text{(core of receptor)} \\ 0 & \text{(open space)} \end{cases} \tag{2.9}$$

$$\Re[L_{\mathrm{PSC}}(l,m,n)] = \begin{cases} 1 & \text{(if this grid is the nearest grid of a ligand atom)} \\ 0 & \text{(otherwise)} \end{cases} \tag{2.10}$$

$$\Im[L_{\mathrm{PSC}}(l,m,n)] = \begin{cases} 3 & \text{(solvent excluding surface of the receptor)} \\ 9 & \text{(core of ligand)} \\ 0 & \text{(open space)} \end{cases} \tag{2.11}$$

$$S_{\mathrm{PSC}}(\alpha,\beta,\gamma) = \Re\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R_{\mathrm{PSC}}(l,m,n)L_{\mathrm{PSC}}(l+\alpha,m+\beta,n+\gamma)\right] \tag{2.12}$$

where $\Re[\cdot]$ and $\Im[\cdot]$ denote the real and imaginary parts of a complex function, and r_{vdW} represents the van der Waals atomic radius.

In eqs. (2.8)–(2.11), $\Im[R]$ and $\Im[L]$ are used to compute the unfavorable component of PSC. A core–core, surface–core, and surface–surface grid point overlap result in a penalty of $-9 \times 9 = -81$, $-3 \times 9 = -27$, and $-3 \times 3 = -9$, respectively. Overlaps involving surface grid points are only moderately penalized, allowing PSC to tolerate some structural flexibility. $\Re[R]$ and $\Re[L]$ are used to compute the favorable component of PSC. $\Re[R]$ denotes the number of receptor atoms within the distance cutoff (3.6 Å+r_{vdW}) of each grid point in the open space, and $\Re[L]$ records the nearest grid point for each ligand atom. The multiplication of these two terms results in the total number of receptor and ligand atom pairs within the distance cutoff. Eq. (2.12) computes both the favorable and unfavorable components of PSC, and sums them into one score, with a higher score indicating better shape complementarity. Fig. 2.2 is a 2D schematic illustration for computing PSC. PSC was used with ZDOCK and obtained better predictions compared to the Katchalski-Katzir function.

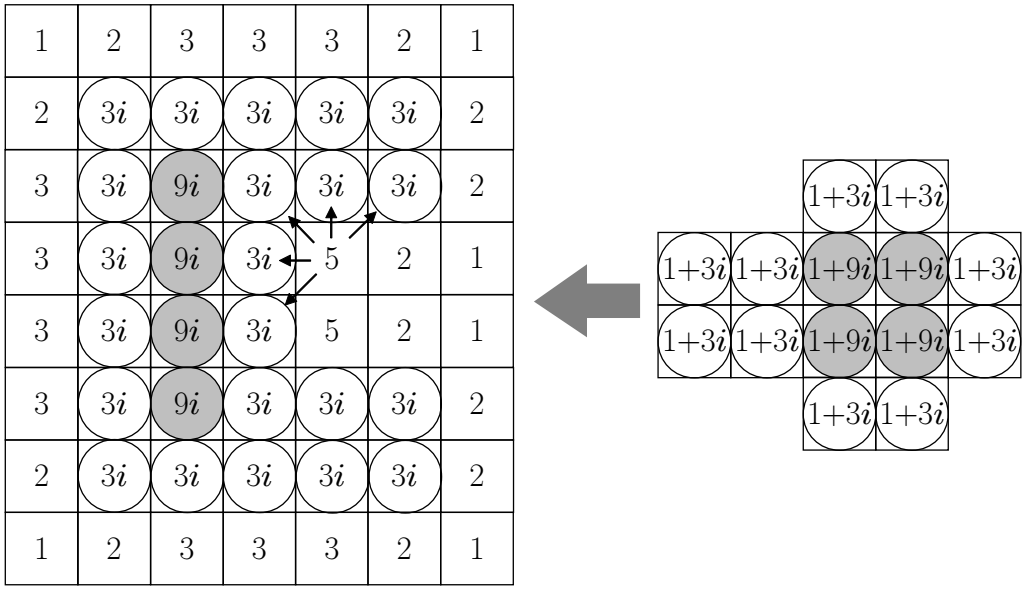


Figure 2.2: 2D schematic illustration for the discrete functions R and L for PSC. Protein atoms are indicated using circles, with open circles indicating surface atoms and shaded circles indicating core atoms. For clarity, we use a grid spacing that equals atom diameter and grid points whose values are 0 have been omitted from the figure. The value assigned to each grid point is indicated. Grid points with open circles are in the solvent excluded surface layer. The block arrow indicates the direction of translation for the ligand in order to achieve the optimal shape complementarity score. For each grid point in the open space of R, we record the number of atoms within a distance cutoff. Small arrows point out the five atoms that are within the distance cutoff of a grid and thus contribute to its score of 5.

2.4.2 Electrostatic function

Shape complementarity is not the only factor involved in protein–protein docking. Electrostatic attraction, particularly the specific charge-charge interactions in the binding interface, is also important.

For the FFT correlation approach, Gabb, *et al.* proposed a Coulombic model represented by a correlation function [28]. The electrostatic calculations proceed in a manner very similar to those of shape complementarity. Charges are assigned to the atoms of the receptor protein, and the protein is placed in a grid. An electric field $\varphi(l, m, n)$ is assigned to each grid point (excluding those of the protein core) and is calculated as follows:

$$\varphi(l,m,n) = \sum_{l'=1}^{N}\sum_{m'=1}^{N}\sum_{n'=1}^{N} \frac{q(l',m',n')}{\varepsilon(r)r} \tag{2.13}$$

$$\varepsilon(r) = \begin{cases} 4 & (r \leq 6 \text{ Å}) \\ 38r - 224 & (6 \text{ Å} < r < 8 \text{ Å}) \\ 80 & (8 \text{ Å} \leq r) \end{cases} \tag{2.14}$$

$$r = \|(l,m,n) - (l',m',n')\| \tag{2.15}$$

where $q(l', m', n')$ is the charge at grid point (l', m', n'), r is the Euclidean distance between grid points (l, m, n) and (l', m', n') (a minimum cutoff distance of 2 Å is imposed to avoid artificially large values of $\varphi(l, m, n)$), and $\varepsilon(r)$ is a distance-dependent dielectric function based on the work by Hingerty, *et al.* [74]. The electrostatic term S_{ES} is defined as follows:

$$S_{\text{ES}}(\alpha,\beta,\gamma) = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R_{\text{ES}}(l,m,n)L_{\text{ES}}(l+\alpha, m+\beta, n+\gamma) \tag{2.16}$$

$$R_{\text{ES}}(l,m,n) = \begin{cases} \varphi(l,m,n) & \text{(entire grid excluding core)} \\ 0 & \text{(core of protein)} \end{cases} \tag{2.17}$$

$$L_{\text{ES}}(l,m,n) = q(l,m,n) \tag{2.18}$$

where R_{ES} and L_{ES} represent the electrostatic grid values of receptor/ligand proteins, determined according to the charge of each grid point $q(l, m, n)$ in which matching atoms in the residues are assigned Gabb's potential [28]. FTDock was used with the Katchalski-Katzir function to analyze shape complementarity, and the electrostatic

function above was used for protein-pair analysis [28].

ZDOCK also used Gabb's electrostatic function except that it used the partial charges in the CHARMM19 potential [75]. In addition, grid points in the core of the receptor were assigned a value of 0 for the electric potential, to avoid any contributions from non-physical receptor–core/ligand contacts.

2.4.3 Desolvation free energy function

ZDOCK 2.3

Chen, *et al.* implemented a desolvation free energy term to ZDOCK by using the atomic contact energy (ACE) [30]. ACE, developed by Zhang, *et al.* [76], is defined as the free energy obtained by replacing an atom-water contact, with an atom-atom contact. The ACE scores were obtained for all pairs of 18 atom types. The total desolvation free energy of the complex formation is calculated by summing the ACE scores of all near atom pairs between the receptor and ligand. Expressed in the form of correlations, the computation of desolvation score requires 18 FFTs. To speed up the calculation, ZDOCK version 2.3 used 18 non-pairwise ACE scores, representing the score between one protein atom of a specific type and another protein atom of any type. Chen's desolvation free energy term S_{DE} [30] is defined as follows:

$$\Re[R_{\mathrm{DE}}(l,m,n)] = \begin{cases} \text{the sum of the ACE scores of all near receptor} \\ \quad \text{atoms that are within } (3.6\ \text{Å} + r_{\mathrm{vdW}}) & \text{(open space)} \\ 0 & \text{(otherwise)} \end{cases} \tag{2.19}$$

$$\Im[R_{\mathrm{DE}}(l,m,n)] = \begin{cases} 1 & \text{(if the grid point is the nearest grid point of an atom)} \\ 0 & \text{(otherwise)} \end{cases} \tag{2.20}$$

$$\Re[L_{\mathrm{DE}}(l,m,n)] = \begin{cases} \text{the sum of the ACE scores of all near ligand} \\ \quad \text{atoms that are within } (3.6\ \text{Å} + r_{\mathrm{vdW}}) & \text{(open space)} \\ 0 & \text{(otherwise)} \end{cases} \tag{2.21}$$

$$\Im[L_{\mathrm{DE}}(l,m,n)] = \begin{cases} 1 & \text{(if the grid point is the nearest grid point of an atom)} \\ 0 & \text{(otherwise)} \end{cases} \tag{2.22}$$

$$S_{\mathrm{DE}}(\alpha,\beta,\gamma) = \frac{1}{2}\Im\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R_{\mathrm{DE}}(l,m,n)L_{\mathrm{DE}}(l+\alpha,m+\beta,n+\gamma)\right] \tag{2.23}$$

ZDOCK 3.0

Mintseris, *et al.* introduced another pair-wise statistical potential called IFACE suitable for docking and showed that this potential could be incorporated into ZDOCK (version 3.0) [33].

In ZDOCK 3.0, Mintseris, *et al.* defined 6 discrete functions for each atom of type i $(= 1, 3, 5, 7, 9, 11)$ in a ligand:

$$\Re[L_{\mathrm{IFACE}:i}(l,m,n)] = \begin{cases} 1 & \text{(if grid cell is occupied by a ligand atom of type } i\text{)} \\ 0 & \text{(otherwise)} \end{cases} \tag{2.24}$$

$$\Im[L_{\mathrm{IFACE}:i}(l,m,n)] = \begin{cases} 1 & \text{(if grid cell is occupied by a ligand atom of type } (i+1)\text{)} \\ 0 & \text{(otherwise)} \end{cases} \tag{2.25}$$

and 6 discrete functions for each possible atom type i in contact with a receptor atom of type j:

$$\Re[R_{\mathrm{IFACE}:i}(l,m,n)] = \begin{cases} \sum e_{\mathrm{IFACE}:(i+1),j} & \text{(Neighbor atoms within 6 Å)} \\ 0 & \text{(Non-neighbor atoms)} \end{cases} \tag{2.26}$$

$$\Im[R_{\mathrm{IFACE}:i}(l,m,n)] = \begin{cases} \sum e_{\mathrm{IFACE}:i,j} & \text{(Neighbor atoms within 6 Å)} \\ 0 & \text{(Non-neighbor atoms)} \end{cases} \tag{2.27}$$

The sum of the resulting FFT correlations on a grid will give the total desolvation

energy summed over all atom types

$$S_{\mathrm{IFACE}}(\alpha,\beta,\gamma) = \Im\left[\sum_{i=1,3,5,\ldots}^{11}\left\{\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R_{\mathrm{IFACE}:i}(l,m,n)L_{\mathrm{IFACE}:i}(l+\alpha,m+\beta,n+\gamma)\right\}\right] \tag{2.28}$$

$$= \sum_{i=1,3,5,\ldots}^{11}\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\{\Re[R_{\mathrm{IFACE}:i}(l,m,n)]\Im[L_{\mathrm{IFACE}:i}(l+\alpha,m+\beta,n+\gamma)] + \Im[R_{\mathrm{IFACE}:i}(l,m,n)]\Re[L_{\mathrm{IFACE}:i}(l+\alpha,m+\beta,n+\gamma)]\} \tag{2.29}$$

where the imaginary part of the complex product evaluated as a result of the correlation accomplishes the summation of the energy components over atoms in contact with each other. In total, ZDOCK 3.0 used 8 FFT correlations (a PSC term (eq. (2.12)), an Electrostatics term (eq. (2.16)) and six IFACE terms (eq. (2.29))).

2.5 Refinement and Rescoring Tools

In protein–protein docking prediction, there are two necessary sequential stages conducted because of the complexity of the problem. The initial stage is rigid-body protein–protein docking analysis generating many predictions (10,000 or more), and the second stage is the refinement and rescoring stage that performs any combination of detailed scoring, energy minimization, side-chain or backbone searches, clustering, etc. on these predictions. In this section, we introduce some refinement and rescoring tools used in recent CAPRI assessments.

2.5.1 RDOCK

The refinement program RDOCK was developed by Li, *et al.* [77]. It uses CHARMM19 to perform energy minimization on the top ZDOCK predictions (the top 2,000 are recommended), and reranks these predictions using desolvation and electrostatics. However, there are some limitations to RDOCK. The energy minimization step takes roughly 1 min per test case, and therefore RDOCK is only feasible for a limited subset of ZDOCK predictions.

2.5.2 FireDock

FireDock is another refinement program developed by Andrusier, *et al.* [44]. FireDock optimizes the binding of each candidate by allowing flexibility in the side-chains and adjustments of the relative orientation of the molecules. Most of the interface residues that are important to the binding recognition, remain in the near-unbound conformations upon complexation. Andrusier, *et al.* used this observation and restricted the side-chain flexibility to include only the clashing interface residues. In addition, the atomic radii of the partners are smoothed in the rigid-body optimization and scoring stages. This coarse refinement approach is the key to the efficiency of FireDock. Scoring of the refined candidates is based on softened van der Waals interactions, ACE, electrostatic, and additional binding free energy estimations.

2.5.3 FiberDock

FireDock models only side-chain movements and keeps the backbone rigid. Mashiach, *et al.* proposed another refinement program, called FiberDock, that allows both backbone and side-chain flexibility [45]. The side-chain flexibility is modeled by a rotamer library, and the backbone flexibility is modeled by an a priori unlimited number of normal modes. Their comparison of FireDock and FiberDock showed that the modeling of backbone flexibility in the refinement process is often critical for creating near-native models with low energy values.

2.5.4 ZRANK

ZRANK is a rescoring tool developed by Pierce, *et al.* [78]. ZRANK quickly and accurately reranks the rigid-body docking results. It uses a more detailed potential than ZDOCK, but is fast enough to quickly process and rerank over 10,000 predictions produced by the ZDOCK sampling search. It significantly improves the accuracy of ZDOCK, and thus it is able to rerank predictions on its own, or may be used as a preprocessing step for refinement programs such as RDOCK, FireDock, and FiberDock.

2.6 CAPRI: Critical Assessment of PRedicted Interactions

CAPRI is a community-wide experiment to assess the capacity of molecular interaction prediction methods applied mainly to protein–protein docking [50]. About 50 groups participated in rounds 1–28 and submitted blind structure predictions based on the known structure of the component proteins. The predictions were compared with the unpublished X-ray structures of the complexes. CAPRI has already proven itself as a powerful driver for the community of computational biologists who develop docking algorithms. Each participating group is allowed to submit 10 models per target, and these models are compared to newly obtained X-ray structures of the complexes that crystallographers make available for the evaluation. The CAPRI experiments are hosted by the European Bioinformatics Institute (EBI). The website is `http://capri.ebi.ac.uk/`. In each round, one or more targets are presented and participants have to submit their predictions before a given deadline. After the submission deadline, the results are published on the CAPRI website and classified into the following: "Incorrect", "Acceptable", "Medium", or "High quality", based on several criteria such as fraction of native residue-residue contact and the root mean square deviation (RMSD) values of the ligands after superimposing the receptors of the prediction and the native complex structures. Until December 2013, there have been about 70 targets evaluated by CAPRI. Some of these targets were used as a benchmark data set, complementary to the ZLAB docking benchmark dataset [79, 80, 81, 82].

2.7 Summary

In this chapter, we introduced typical rigid-body protein–protein docking methods that primarily use FFT correlations with the Katchalski-Katzir algorithm. We also introduced basic discussions of the refinement and rescoring tools generally used for docking protocols.

Chapter 3

Development of a Rapid Protein–Protein Docking Method

3.1 Introduction

To realize the network level PPI predictions by fully utilizing protein tertiary structures required to sample some millions of protein dockings, a rapid protein–protein docking method is needed. The one of the key of calculation speedup is improvement of score functions to reduce the number of FFT correlations. In this chapter, we introduce a novel shape complementarity score function and a novel desolvation free energy score function. These score functions can calculate three effects, shape complementarity, electrostatics interaction and desolvation free energy, at the same time with only one FFT correlation function.

3.2 Materials and Methods

We propose a novel shape complementarity function with only real number representation (rPSC) and a novel desolvation free energy function by using non-pairwise ACE. In this section, we describe the MEGADOCK scoring functions: rPSC, electrostatics and desolvation free energy.

3.2.1 real Pairwise Shape Complementarity (rPSC)

Each receptor and ligand proteins are first allocated on a 3D grid space $\mathbb{N}^3$ with grid point spacing of 1.2 Å. The scores are then assigned to each grid point $(l, m, n) \in \mathbb{N}^3$

according to the location in a protein, such as the surface or core.

We introduce the following novel scoring function called real Pairwise Shape Complementarity (rPSC) for the shape complementarity term R_{rPSC} and L_{rPSC}:

$$R_{\mathrm{rPSC}}(l,m,n) = \begin{cases} \#\text{ of receptor atoms within } (3.6\text{ Å} + r_{\mathrm{vdW}}) & \text{(open space)} \\ -45 & \text{(inside of receptor)} \end{cases} \tag{3.1}$$

$$L_{\mathrm{rPSC}}(l,m,n) = \begin{cases} 1 & \text{(inside of ligand)} \\ 0 & \text{(otherwise)} \end{cases} \tag{3.2}$$

$$S_{\mathrm{rPSC}}(\alpha,\beta,\gamma) = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R_{\mathrm{rPSC}}(l,m,n) L_{\mathrm{rPSC}}(l+\alpha, m+\beta, n+\gamma) \tag{3.3}$$

where r_{vdW} represents the van der Waals atomic radius of receptor atoms in the grid point (l,m,n), and (α,β,γ) is a vector of the ligand translation. The parameter of these functions are optimized using the structural data of 176 complexes from a general docking benchmark dataset (protein–protein docking benchmark 4.0 [82]). Fig. 3.1 is a 2D schematic illustration for computing rPSC.

R_{rPSC} denotes the number of receptor atoms within the distance cutoff (3.6 Å+r_{vdW}) of each grid point in the open space. The multiplication of R_{rPSC} and L_{rPSC} results in the total number of receptor/ligand atom pairs within the distance cutoff. Compared to a similar score function, PSC by ZDOCK, the rPSC function uses only real number representation for shape complementarity. Therefore, we can place a physicochemical parameter in the imaginary part (see next section). As a result, it is possible to calculate a total energy score with only one complex number for each grid point. By decreasing the number of required DFT/IFT operations, the docking calculation is expected to be faster than other tools like ZDOCK.

3.2.2 Combination of rPSC and electrostatics

In addition to shape complementarity scores, we used the electrostatic interactions of each amino acid as a physicochemical score. We used Coulombic model by Gabb, *et al.* [28] and the CAHRMM19 potential [75] like ZDOCK.

The electric field $\varphi(l,m,n)$ is calculated by eq. (2.13) and the electrostatic terms R_{ES} and R_{ES} are decided by eq. (2.17) and eq. (2.18) respectively. The electrostatics score S_{ES} is represented by eq. (2.16).

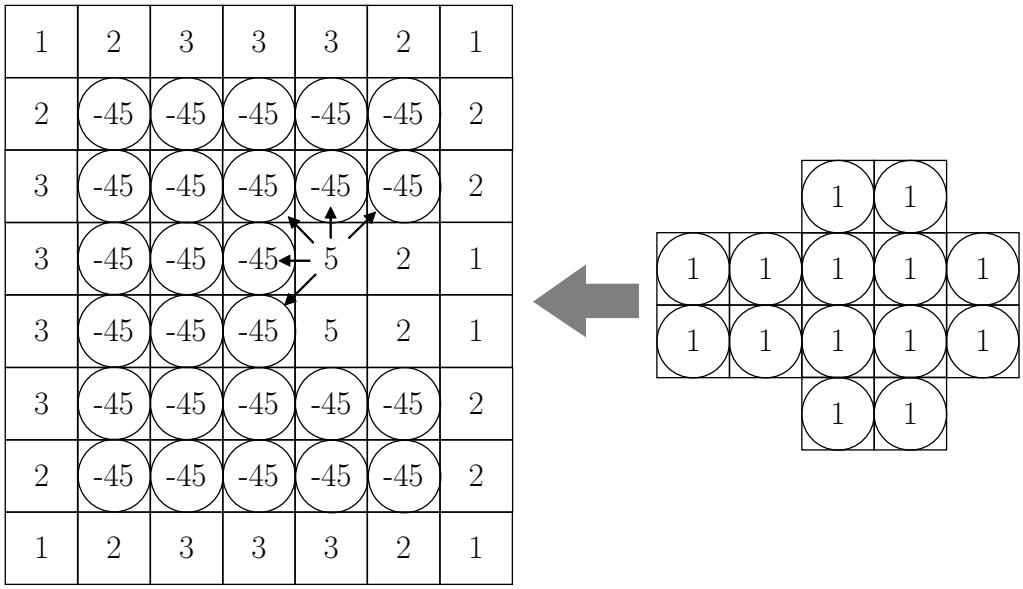


Figure 3.1: 2D schematic illustration for the discrete functions R and L for real Pairwise Shape Complementarity (rPSC). Protein atoms are indicated using circles, with open circles indicating surface atoms and shaded circles indicating core atoms. For clarity, we use a grid spacing that equals atom diameter and grid points whose values are 0 have been omitted from the figure. The value assigned to each grid point is indicated. Grid points with open circles are in the solvent excluded surface layer. The block arrow indicates the direction of translation for the ligand in order to achieve the optimal shape complementarity score. For each grid point in the open space of R, we record the number of atoms within a distance cutoff. Small arrows point out the five atoms that are within the distance cutoff of a grid and thus contribute to its score of 5.

Considering these two terms, the combination score S is represented as follows:

$$R(l,m,n) = R_{\mathrm{rPSC}}(l,m,n) + iR_{\mathrm{ES}}(l,m,n) \tag{3.4}$$

$$L(l,m,n) = L_{\mathrm{rPSC}}(l,m,n) + iwL_{\mathrm{ES}}(l,m,n) \tag{3.5}$$

$$S(\alpha,\beta,\gamma) = \Re\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n)L(l+\alpha,m+\beta,n+\gamma)\right] \tag{3.6}$$

$$= \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\{R_{\mathrm{rPSC}}(l,m,n)L_{\mathrm{rPSC}}(l+\alpha,m+\beta,n+\gamma) - wR_{\mathrm{ES}}(l,m,n)L_{\mathrm{ES}}(l+\alpha,m+\beta,n+\gamma)\} \tag{3.7}$$

$$= S_{\mathrm{rPSC}}(\alpha,\beta,\gamma) - wS_{\mathrm{ES}}(\alpha,\beta,\gamma) \tag{3.8}$$

where, w is the weight parameter which is set as w = 2,800, a value obtained by optimization by conducting pre-experiments using protein–protein docking benchmark 4.0 dataset.

3.2.3 Combination of rPSC, electrostatics and desolvation free energy

To improve performance as docking accuracy and calculation speed of MEGADOCK, we should introduce a hydrophobic interaction effect in our scoring model. However, using conventional score model used ZDOCK increase in number of FFT correlations. Therefore we need a new score model for varied applications.

In our proposed method, we used a non-pairwise-type atomic contact energy (ACE) score [76] to incorporate a desolvation free energy effect. For the current study, we introduce a simple model that considers only the receptor protein because, when both the receptor and ligand are taken into consideration, an increase in the number of correlation functions is unavoidable.

We modify the receptor rPSC function R_{rPSC} in eq. (3.1) in order to introduce the ACE score. The new receptor function $R_{\mathrm{rPSC+RDE}}$ is defined as follows:

$$R_{\mathrm{rPSC+RDE}}(l,m,n) = R_{\mathrm{rPSC}}(l,m,n) + w_{\mathrm{DE}}R_{\mathrm{RDE}}(l,m,n) \tag{3.9}$$

$$R_{\mathrm{RDE}}(l,m,n) = \begin{cases} \text{the sum of the ACE scores of all near} \\ \quad \text{receptor atoms within } (3.6\ \text{Å} + r_{\mathrm{vdW}}) & \text{(open space)} \\ 0 & \text{(inside of receptor)} \end{cases} \tag{3.10}$$

where w_{DE} is the weight parameter of S_{RDE}. w_{DE} is set as $w_{\mathrm{DE}} = 0.8$, a value obtained by optimization by conducting pre-experiments using protein–protein docking benchmark 4.0 dataset. The ligand rPSC function is not modified. The desolvation free energy term S_{RDE} is defined as follows:

$$S_{\mathrm{RDE}}(\alpha,\beta,\gamma) = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R_{\mathrm{RDE}}(l,m,n) L_{\mathrm{rPSC}}(l+\alpha, m+\beta, n+\gamma) \tag{3.11}$$

Fig. 3.2 shows a pattern diagram of the proposed model. We use the ACE values given in Table 3.1.

Finally, the combination score S with rPSC, electrostatics and our desolvation free energy is represented as follows:

$$R(l,m,n) = R_{\mathrm{rPSC+RDE}}(l,m,n) + iR_{\mathrm{ES}}(l,m,n) \tag{3.12}$$

$$= R_{\mathrm{rPSC}}(l,m,n) + w_{\mathrm{DE}} R_{\mathrm{RDE}}(l,m,n) + iR_{\mathrm{ES}}(l,m,n) \tag{3.13}$$

$$L(l,m,n) = L_{\mathrm{rPSC}}(l,m,n) + iw_{\mathrm{ES}} L_{\mathrm{ES}}(l,m,n) \tag{3.14}$$

$$S(\alpha,\beta,\gamma) = \Re\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n) L(l+\alpha, m+\beta, n+\gamma)\right] \tag{3.15}$$

$$\begin{aligned} = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} \{ & R_{\mathrm{rPSC}}(l,m,n) L_{\mathrm{rPSC}}(l+\alpha, m+\beta, n+\gamma) \\ & + w_{\mathrm{DE}} R_{\mathrm{RDE}}(l,m,n) L_{\mathrm{rPSC}}(l+\alpha, m+\beta, n+\gamma) \\ & - w_{\mathrm{ES}} R_{\mathrm{ES}}(l,m,n) L_{\mathrm{ES}}(l+\alpha, m+\beta, n+\gamma)\} \end{aligned} \tag{3.16}$$

$$= S_{\mathrm{rPSC}}(\alpha,\beta,\gamma) + w_{\mathrm{DE}} S_{\mathrm{RDE}}(\alpha,\beta,\gamma) - w_{\mathrm{ES}} S_{\mathrm{ES}}(\alpha,\beta,\gamma) \tag{3.17}$$

This score model attains a value of $R_{\mathrm{rPSC}} + w_{\mathrm{DE}} R_{\mathrm{RDE}}$ when the open space near the receptor surface is superposed on the ligand surface and core. ZDOCK 2.3 [32] uses three correlation functions, and ZDOCK 3.0 [33] uses eight correlation functions to consider three effects—shape complementarity, electrostatics, and desolvation free energy—our score model can calculate docking scores under consideration of three effects with only one FFT correlation, while maintaining an advantage in terms of

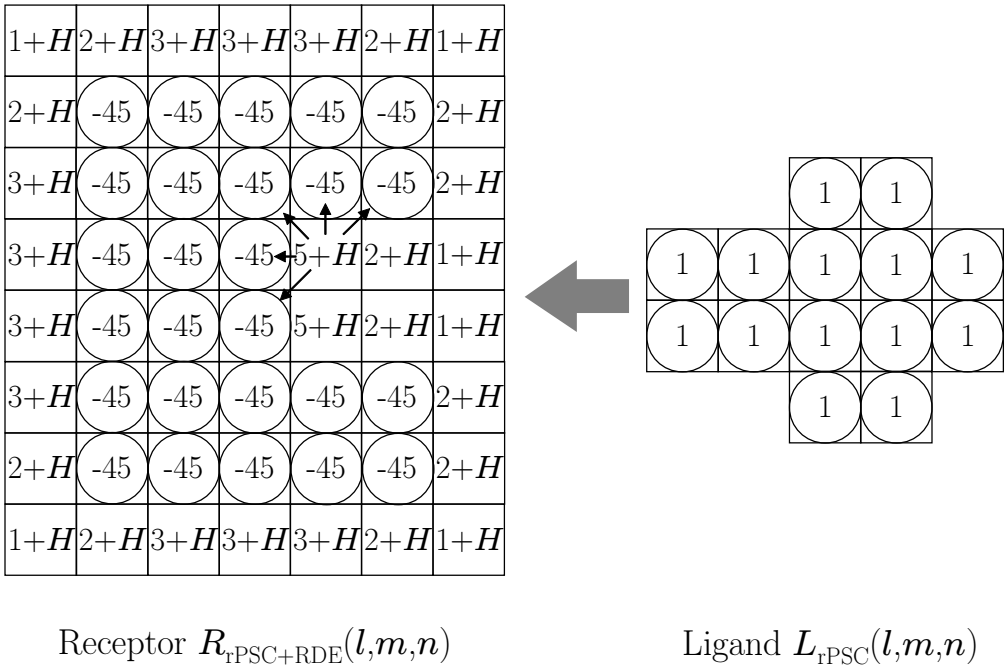


Figure 3.2: Proposed scoring model $R_{\mathrm{rPSC+RDE}}(l, m, n)$ and $L_{\mathrm{rPSC}}(l, m, n)$. The model consists of 3D grid, but here we show only two dimensions for simplicity. For clarity, grid points with a value of 0 have been omitted. Small arrows indicate the five atoms that are within the cutoff distance of a grid, and thus contribute to its score of $5 + H$, where H means $w_{\mathrm{DE}} R_{\mathrm{RDE}}(l, m, n)$.

Table 3.1: Non-pairwise ACE scores. The atom types are defined in below table.

atom type	N	C^{α}	C	O	GC^{α}	C^{β}	KN^{ζ}	KC^{δ}	DO^{δ}
ACE score	−0.495	−0.553	−0.464	−0.079	0.008	−0.353	1.334	1.046	0.933
atom type	RN^{η}	NN^{δ}	RN^{ε}	SO^{γ}	HN^{ε}	YC^{ζ}	FC^{ζ}	LC^{δ}	CS^{γ}
ACE score	0.726	0.693	0.606	0.232	0.061	−0.289	−0.432	−0.987	−1.827

atom type	amino acid	atom
N	BB	N
C^{α}	BB	C^{α}
C	BB	C
O	BB	O
GC^{α}	Gly	C^{α}
	Ala	C^{β}
	Arg	C^{β}
	Asn	C^{β}
	Asp	C^{β}
	Cyc	C^{β}
	Gln	C^{β}
	Glu	C^{β}
	His	C^{β}
	Ile	C^{β}
C^{β}	Leu	C^{β}
	Lys	C^{β}
	Met	C^{β}
	Phe	C^{β}
	Pro	C^{β}
	Pro	C^{γ}
	Pro	C^{δ}
	Thr	C^{β}
	Trp	C^{β}
	Tyr	C^{β}
	Val	C^{β}
KN^{ζ}	Lys	C^{ε}
	Lys	N^{ζ}
KC^{δ}	Lys	C^{δ}
	Asp	C^{γ}
	Asp	$O^{\delta 1}$
DO^{δ}	Asp	$O^{\delta 2}$
	Glu	C^{δ}
	Glu	$O^{\varepsilon 1}$
	Glu	$O^{\varepsilon 2}$

atom type	amino acid	atom
	Arg	C^{ζ}
RN^{η}	Arg	$N^{\eta 1}$
	Arg	$N^{\eta 2}$
	Asn	C^{γ}
	Asn	$O^{\delta 1}$
NN^{δ}	Asn	$N^{\delta 2}$
	Gln	C^{δ}
	Gln	$O^{\varepsilon 1}$
	Gln	$N^{\varepsilon 2}$
RN^{ε}	Arg	C^{δ}
	Arg	N^{ε}
	Ser	C^{β}
SO^{γ}	Ser	C^{γ}
	Thr	$C^{\gamma 1}$
	Tyr	O^{η}
	His	C^{γ}
	His	$N^{\delta 1}$
HN^{ε}	His	$C^{\delta 2}$
	His	$C^{\varepsilon 1}$
	His	$N^{\varepsilon 2}$
	Trp	$N^{\varepsilon 1}$
	Tyr	$C^{\varepsilon 1}$
YC^{ζ}	Tyr	$C^{\varepsilon 2}$
	Tyr	C^{ζ}

atom type	amino acid	atom
	Arg	C^{γ}
	Gln	C^{γ}
	Glu	C^{γ}
	Ile	$C^{\gamma 1}$
	Leu	C^{γ}
	Lys	C^{γ}
	Met	C^{γ}
	Met	S^{δ}
	Phe	C^{γ}
	Phe	$C^{\delta 1}$
	Phe	$C^{\delta 2}$
	Phe	$C^{\varepsilon 1}$
FC^{ζ}	Phe	$C^{\varepsilon 2}$
	Phe	C^{ζ}
	Thr	$C^{\gamma 2}$
	Trp	C^{γ}
	Trp	$C^{\delta 1}$
	Trp	$C^{\delta 2}$
	Trp	$C^{\varepsilon 2}$
	Trp	$C^{\varepsilon 3}$
	Trp	$C^{\zeta 2}$
	Trp	$C^{\zeta 3}$
	Trp	$C^{\eta 2}$
	Tyr	C^{γ}
	Tyr	$C^{\delta 1}$
	Tyr	$C^{\delta 2}$
	Ile	$C^{\gamma 2}$
	Ile	C^{δ}
	Leu	$C^{\delta 1}$
LC^{δ}	Leu	$C^{\delta 2}$
	Met	C^{ε}
	Val	$C^{\gamma 1}$
	Val	$C^{\gamma 2}$
CS^{γ}	Cyc	S^{γ}

Note: Atom names are taken from the typical PDB files. For convenience, a side-chain atom type (except for CB) is assigned the name of one of the atoms of that type prefixed by the single-letter amino acid code, e.g. KN^{ζ}=N^{ζ} of Lys. The amino acid 'BB' denotes backbone atoms.

calculation speed.

3.2.4 Other settings

Sampling number of decoys per rotation

Conventional software typically records the highest-scoring decoy obtained by all the translation patterns for each ligand rotation because it is well known that analyses of more than one decoy per rotation do not contribute significantly to an improvement of docking pose predictions. In contrast, we assumed that in the PPI screening problem, the distributions of high-scoring decoys provided important information for the analyses. Hence, MEGADOCK allows the user to input the number of decoys t that should be recorded per ligand rotation in order to obtain a larger number of high-scoring decoys.

FFT grid size

An FFT-based docking tool firstly reads the atom coordinates of a receptor and a ligand, and determines the grid size fitted for the receptor and ligand. The FFT size N is proportional to the grid size, which was automatically calculated from the single grid unit size and the size of proteins. FFTW algorithms [83], which we used in MEGADOCK, are optimized for sizes that represented as a multiple of $\{2, 3, 5, 7, 11, 13\}$. Thus, our algorithm to decide the grid size searches the smallest composite number consisted of those prime factors.

3.2.5 Dataset

For the evaluation of our new scoring function, the protein complex structures used in this study were retrieved from a standard protein–protein docking benchmark set (protein–protein docking benchmark 4.0 [82]). This benchmark set comprises 176 known complexes and included both a “bound” and “unbound” set. The “bound” set is composed of protein structures prepared by separating individual proteins from the crystal structures of 176 protein complexes. The “unbound” data means that each protein structure is taken from the isolated form of crystals rather than complex form. The “unbound” dataset includes protein structural data corresponding to the same set of proteins in the “bound” dataset. Structural differences in bound and unbound form in RMSD are shown in the reference [82].

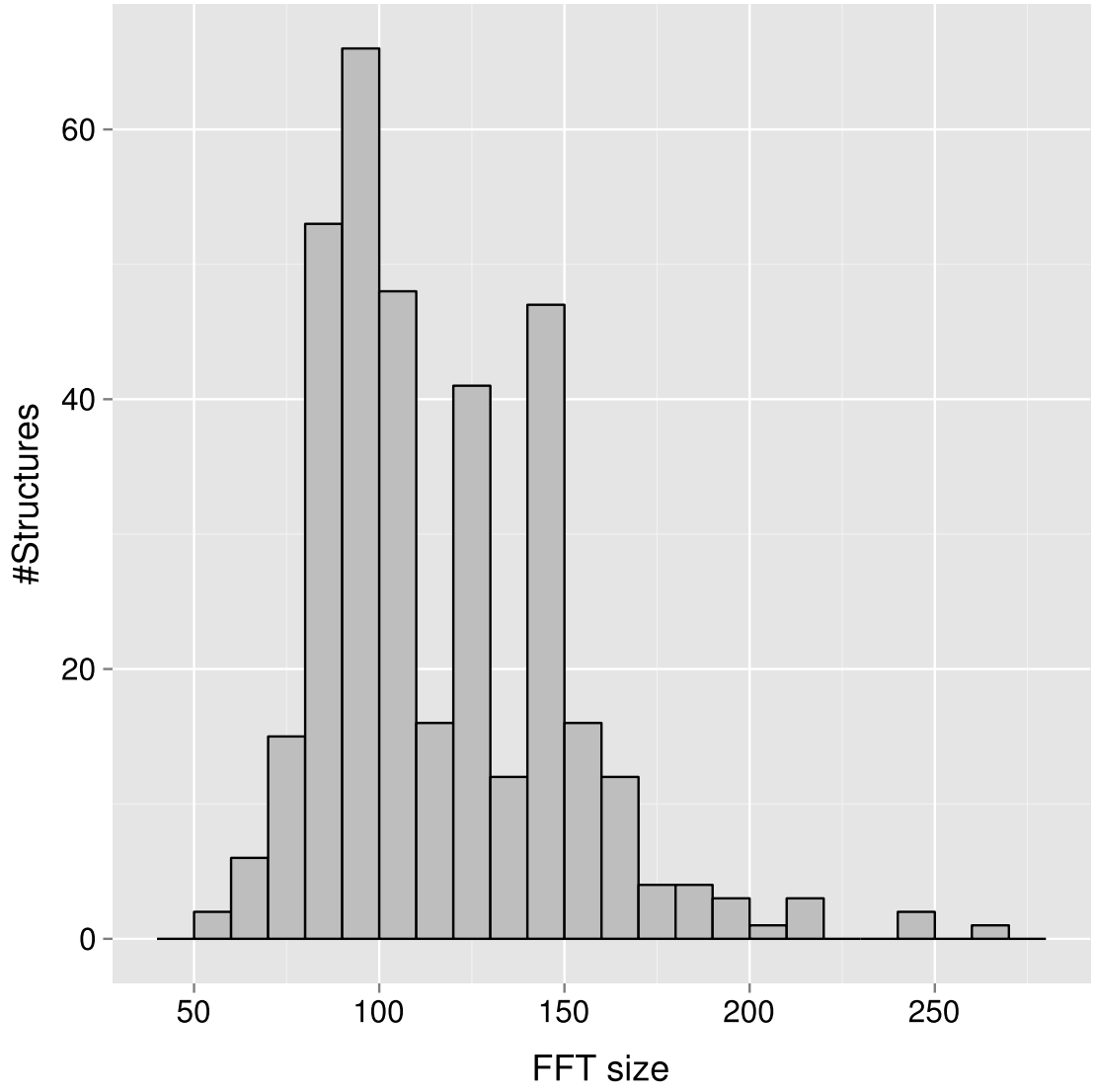


Figure 3.3: FFT size of various proteins in protein–protein docking benchmark 4.0 (176 protein complexes, 352 structures).

Distribution of size of FFT for protein in this dataset is shown in Fig. 3.3. Time consumed for FFT is expected to be longer with larger size of FFT.

3.2.6 Evaluation of the MEGADOCK approximation capability to the ZDOCK

To confirm the approximation capability of MEGADOCK score function, we assessed the correlation between the MEGADOCK score functions and ZDOCK ones:

(a) MEGADOCK rPSC vs. ZDOCK PSC (implemented in MEGADOCK) [30],

(b) MEGADOCK rPSC+ES+RDE vs. ZDOCK PSC+ES+DE (ZDOCK ver. 2.3) [32],

(c) MEGADOCK rPSC+ES+RDE vs. ZDOCK PSC+ES+IFACE (ZDOCK ver. 3.0) [33].

For comparison, we set parameters of $n_\theta = 3{,}600$ decoys per case and $\theta = 15°$ for the ligand rotation step. We calculated Spearman's rank correlation coefficient ρ and P-value for t-test between the 3,600 sequence values of MEGADOCK and ZDOCK. The way of construction of FFT grids on ZDOCK and MEGADOCK is the same. However, it is not possible to confirm that the grid made by ZDOCK and MEGADOCK is the same because ZDOCK is distributed in binary form and it does not support output of FFT grid. Thus, we allow MEGADOCK the difference of distance of 1.2 Å from ZDOCK predictions ((b), (c)). We calculated the MEGADOCK score on 7 position of $(l \pm 1, m, n)$, $(l, m \pm 1, n)$, $(l, m, n \pm 1)$, (l, m, n) based on the predicted position (l, m, n) by ZDOCK and the maximum value of these position was considered as the MEGADOCK score on (l, m, n). Fig. 3.4 shows this comparison method.

3.2.7 Evaluation of docking performance

To evaluate the docking pose prediction performance, we conducted a re-docking and unbound docking experiment using the protein–protein docking benchmark 4.0 dataset. In order to determine the accuracy of the docking predictions, we used the root mean square deviation (RMSD) of the ligand (L-RMSD), which is the RMSD of the predicted ligand position and that of the crystal complex structure calculated for all the atoms when the receptor positions are superimposed. The RMSDs of the unbound structures were calculated for residues that were aligned by pairwise alignment of the amino acid sequences between the bound and unbound structures. We defined a "near-native decoy" as that for which L-RMSD was less than or equal to 5 Å. We compared the performance of the following docking methods:

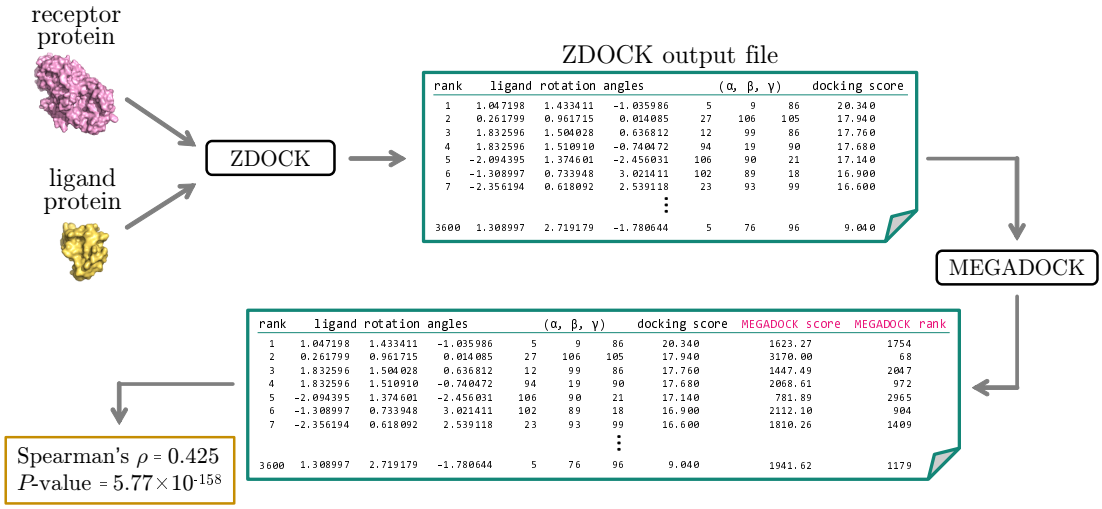


Figure 3.4: The method of Spearman's correlation coefficient calculations.

- MEGADOCK rPSC
- MEGADOCK rPSC+ES
- MEGADOCK rPSC+ES+RDE
- ZDOCK PSC (ZDOCK ver. 2.1) [31]
- ZDOCK PSC+ES+DE (ZDOCK ver. 2.3) [32]
- ZDOCK PSC+ES+IFACE (ZDOCK ver. 3.0) [33]

For comparison with ZDOCK, we set parameters of n_θ = 3,600 decoys per case and $\theta = 15°$ for the ligand rotation step.

We compared the following three values to determine the docking performance produce by the methods listed above.

- **#NND**: The number of near-native decoys (L-RMSD< 5Å) in 3,600 highest scoring decoys.
- **Best Rank**: The rank of the first near-native decoy.
- **RMSD**$_{best}$: The L-RMSD of the 'Best Rank' decoy.

In addition, we compared the following widely used value [30, 33, 78, 84] to determine the overall docking performance:

- **Success Rate**: The percentage of cases with near-native decoys for a given number of top-ranked predictions per test case.

3.3 Results and Discussion

3.3.1 rPSC approximation capability to PSC

Table 3.2 shows the average of Spearman's correlation coefficient between rPSC and PSC on 176 bound and unbound cases. The result shows a strong correlation between rPSC and PSC and rPSC has some approximation capability to PSC.

3.3.2 MEGADOCK approximation capability to ZDOCK 2.3/3.0

Table 3.3 and Table 3.4 show the average of Spearman's correlation coefficient between MEGADOCK and ZDOCK on 176 bound and unbound cases. The result shows a weak correlation between MEGADOCK and ZDOCK 2.3/3.0. A number of reasons can be given for these results. The one of the reasons is the difference of implementations and hidden internal parameters, and another one is the difference of the desolvation free energy functions. Because our desolvation free energy term RDE is simplified for faster calculation, this weak correlation is thought as the inevitable issue.

3.3.3 Docking prediction accuracy

Table 3.5 and Table 3.6 show the results of all docking predictions both bound and unbound set.

As a result of the incorporation of the rPSC score for the shape complementarity representation, we achieved almost same #NND and smaller Best Rank values in many complexes using MEGADOCK than in the case of PSC representations (ZDOCK 2.1). Moreover, by adding the electrostatic force and desolvation free energy to the score function with rPSC, we achieved better Best Rank and #NND values. Here, we show the sum of #NND values and the number of cases with at least one near-native decoy in the top 100 scored decoys in Table 3.7. MEGADOCK rPSC+ES+RDE gave #NND values of 661 in the bound set and 155 in the unbound set. Both values were higher than those obtained with rPSC (545 in the bound set and 103 in the unbound set) and rPSC+ES (593 in the bound set and 116 in the unbound set). In addition,

Table 3.2: The Spearman's correlation coefficient between rPSC and PSC. ρ_{mean} is the average value of coefficients of 176 complexes and s.d. is the standard deviation. P-value is calculated from t-distribution with $(3{,}600-2)$ degrees of freedom and ρ_{mean}.

	bound	unbound
Spearman's ρ_{mean}	0.450	0.438
s.d.	0.028	0.065
P-value of ρ_{mean}	3.082×10^{-179}	5.385×10^{-169}

Table 3.3: The Spearman's correlation coefficient between MEGADOCK rPSC+ES+RDE and ZDOCK 2.3 (PSC+ES+DE). ρ_{mean} is the average value of coefficients of 176 complexes and s.d. is the standard deviation. P-value is calculated from t-distribution with $(3{,}600-2)$ degrees of freedom and ρ_{mean}.

	bound	unbound
Spearman's ρ_{mean}	0.242	0.239
s.d.	0.058	0.060
P-value of ρ_{mean}	4.291×10^{-49}	5.318×10^{-48}

Table 3.4: The Spearman's correlation coefficient between MEGADOCK rPSC+ES+RDE and ZDOCK 3.0 (PSC+ES+IFACE). ρ_{mean} is the average value of coefficients of 176 complexes and s.d. is the standard deviation. P-value is calculated from t-distribution with $(3{,}600-2)$ degrees of freedom and ρ_{mean}.

	bound	unbound
Spearman's ρ_{mean}	0.166	0.163
s.d.	0.080	0.077
P-value of ρ_{mean}	9.059×10^{-24}	6.051×10^{-23}

MEGADOCK rPSC+ES+RDE achieved the same level of accuracy of ZDOCK 2.3 (Table 3.7).

By looking at the Best Rank values, we observed that MEGADOCK rPSC+ES+RDE successfully predicted at least one near-native decoy for 165 protein complexes in the bound set and 30 complexes in the unbound set in the top 100 scored decoys. This result gave higher values than those obtained by rPSC (149 in the bound set and 22 in the unbound set) and rPSC+ES (156 in the bound set and 22 in the unbound set). With MEGADOCK rPSC+ES+RDE, we obtained near native decoys that were not achieved with only shape complementarity scoring (MEGADOCK rPSC and ZDOCK PSC), such as in 1E6J (bound) or 1XD3 (unbound).

Fig. 3.5 shows the docking success rate. The vertical axis shows the ratio of the number of successfully predicted protein complexes; in total 176 benchmark 4.0 pairs. Here, we define the docking as successful when at least one near-native decoy was found in the top n scoring decoys. The number of decoys n is shown along the horizontal axis. A docking method was working well when the area is larger in the left part of the graph. While MEGADOCK was less successful when compared to ZDOCK 3.0, incorporation of the electrostatic term and desolvation free energy term clearly improved the docking success rate. We think that S_{rPSC} and S_{RDE} require further tuning using more complex structures in the PDB.

Fig. 3.6 shows examples of docking predictions by MEGADOCK. The proteins used as the receptor are shown by the surface representations, whereas ligands are shown by ribbons. The ligands colored red are placed in the predicted coordinates whereas those colored green are positioned in the original crystal structures. The structure on the left of Fig. 3.6 corresponds to the PDB data 1CGI, for which we obtained a ligand RMSD value of 1.02 Åfor the highest-ranked decoy. The structure on the right of Fig. 3.6 shows the highest-ranked decoy generated by the re-docking of 2BTF (the ligand RMSD value = 1.33 Å).

3.3.4 Calculation time

Table 3.8 shows the total time consumed for docking the benchmark 4.0 dataset. All the calculations were conducted on the TSUBAME 2.0 supercomputing system, Tokyo Institute of Technology, Japan, which consisted of two Intel Xeon processor 2.93 GHz (6 cores × 2) and 54 GB RAM, operational nodes connected via an InfiniBand and Gigabit Ethernet. An average of 11.93 min was required for each docking calculation with rPSC, electrostatics and desolvation free energy using one CPU core.

Table 3.5: Docking prediction performance of MEGADOCK and ZDOCK for the bound docking test cases in protein–protein docking benchmark 4.0. #NND denotes the number of near-native decoy in the top 3,600 predictions, Best Rank is the rank of first near-native decoy, and RMSD is the L-RMSD of first near-native decoy ($RMSD_{best}$).

	rPSC			rPSC+ES			rPSC+ES+RDE			ZDOCK 2.1 (PSC)			ZDOCK 2.3 (PSC+ES+DE)			ZDOCK 3.0 (PSC+ES+IFACE)		
PDB ID	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD
1A2K	1	5	1.75	2	3	1.75	3	2	1.75	2	11	2.30	1	2	2.30	3	1	2.30
1ACB	7	1	1.25	6	5	1.25	9	1	1.25	6	1	1.68	9	1	1.68	11	4	4.52
1AHW	1	1	1.92	1	1	1.92	1	1	1.92	1	6	1.89	1	2	1.99	1	3	1.99
1AK4	1	300	2.31	1	570	2.31	1	149	2.31	1	2274	1.57	1	307	1.57	2	56	1.57
1AKJ	1	9	1.22	1	1	1.22	1	1	1.22	2	1397	1.96	2	13	2.12	2	4	1.48
1ATN	2	21	2.13	2	10	2.13	2	1	2.13	2	66	1.59	2	1	1.59	2	1	1.59
1AVX	3	1	2.13	3	1	1.57	4	2	1.87	3	3	2.48	4	2	2.48	4	2	2.47
1AY7	4	1	1.50	4	1	1.40	4	3	1.50	3	15	2.05	5	2	2.05	7	16	2.05
1AZS	2	3	1.78	2	2	1.78	2	1	1.78	2	143	1.88	2	2	1.88	2	1	2.34
1B6C	2	1	2.29	2	1	2.29	2	1	1.41	1	8	1.76	2	1	1.76	1	1	1.76
1BGX	1	1	2.97	1	1	3.76	1	1	2.97	1	1	3.10	1	1	3.10	1	1	3.10
1BJ1	2	2	2.32	2	1	2.32	3	2	2.32	3	1	2.70	3	1	2.70	3	2	2.70
1BKD	3	1	1.94	4	1	1.94	4	1	1.94	4	1	1.94	5	1	1.94	4	1	1.94
1BUH	8	1	0.99	6	2	0.99	7	3	0.99	3	87	1.14	2	191	2.06	7	5	2.02
1BVK	4	8	1.75	4	38	1.75	3	14	1.75	3	291	2.07	2	608	1.61	3	85	2.07
1BVN	6	1	1.24	5	1	1.24	8	1	1.24	10	1	4.32	12	1	1.91	15	1	1.91
1CGI	8	1	1.02	9	1	1.02	10	1	1.02	11	1	1.33	12	1	1.33	11	1	1.33
1CLV	12	1	0.79	12	1	0.79	20	1	0.79	24	1	1.78	28	1	1.78	36	1	1.78
1D6R	7	1	1.38	9	1	1.54	9	1	1.54	5	20	1.51	4	28	2.45	2	1512	1.51
1DE4	1	6	2.89	1	11	2.89	1	1	2.89	1	245	2.98	1	10	2.98	1	1	3.61
1DFJ	2	1	2.54	2	1	1.95	2	2	2.54	1	22	2.14	2	2	2.14	2	1	4.56
1DQJ	6	1	1.57	5	1	1.57	4	1	1.57	4	1	1.51	4	1	1.51	4	1	1.51
1E4K	1	142	2.42	1	36	2.42	1	1	2.42	0	-	-	1	188	2.93	1	671	2.55
1E6E	2	1	1.33	5	1	1.33	5	1	1.33	4	3	1.15	5	1	1.15	6	1	1.15
1E6J	0	-	-	0	-	-	2	28	1.92	0	-	-	1	290	2.77	2	158	2.33
1E96	1	155	1.81	1	155	1.81	1	1218	4.97	1	1369	1.66	2	368	1.66	1	1790	4.73
1EAW	6	1	1.46	5	1	1.46	6	1	1.46	7	1	2.02	8	1	1.55	3	3	2.02
1EER	1	1	2.54	1	1	2.80	1	1	2.54	1	2	2.95	1	1	2.70	1	1	2.70
1EFN	5	1	2.86	9	1	2.86	6	1	2.86	8	7	3.03	13	1	3.03	16	1	3.03
1EWY	2	282	1.18	4	67	1.18	4	15	1.18	3	29	1.71	5	2	1.71	6	5	1.71
1EZU	1	1	2.76	1	1	2.76	1	1	2.76	2	1	2.58	1	1	2.58	1	1	2.58
1F6M	3	1	1.02	2	4	1.02	5	1	1.02	4	75	1.58	4	3	4.49	8	1	2.16
1F34	2	1	1.79	2	1	1.79	2	1	1.74	2	1	1.79	2	1	1.79	2	1	1.79
1F51	3	1	1.02	4	1	1.02	4	1	1.02	4	103	4.25	5	7	3.28	6	8	1.23
1FAK	1	1	1.93	2	1	2.17	2	1	2.17	1	1	1.95	3	1	1.95	3	1	1.95
1FC2	2	1	3.80	2	4	3.80	2	1	3.80	2	32	3.68	2	33	3.68	2	11	2.09
1FCC	2	27	0.79	2	25	0.79	1	93	0.79	2	323	3.80	1	853	3.80	1	495	1.29
1FFW	2	155	0.89	7	10	1.42	5	19	1.42	2	1417	1.50	6	16	1.50	11	2	3.30
1FLE	5	1	1.30	7	1	1.30	6	1	1.30	6	2	4.41	6	1	1.44	6	2	1.44
1FQ1	1	1	1.34	1	2	1.34	1	1	1.34	0	-	-	1	84	1.98	1	15	1.98
1FQJ	1	17	1.81	1	1	1.81	2	7	1.81	1	550	1.18	2	6	1.68	2	176	1.68
1FSK	5	1	1.33	4	1	1.33	5	1	1.33	4	5	1.71	4	1	1.71	3	1	1.71
1GCQ	7	1	1.35	7	1	1.35	6	1	1.35	4	18	1.37	4	9	1.37	8	3	1.37
1GHQ	0	-	-	0	-	-	1	3202	2.83	0	-	-	0	-	-	0	-	-
1GL1	12	1	0.95	13	1	0.99	14	1	0.99	15	1	1.64	14	1	1.64	15	1	1.37
1GLA	3	659	3.82	3	122	3.82	4	144	3.82	0	-	-	2	297	2.18	8	12	2.18
1GP2	2	69	2.22	2	1	2.22	2	1	2.22	1	41	2.22	2	1	2.22	2	1	2.22
1GPW	3	2	1.63	3	1	1.37	3	1	1.63	3	2	1.99	3	2	1.99	3	2	1.99
1GRN	1	1	1.89	2	1	1.42	2	1	1.42	3	2	1.74	4	1	1.74	4	1	1.74
1GXD	1	20	1.72	1	2	1.72	1	9	1.72	2	351	4.77	2	3	2.47	1	3	2.47
1H1V	1	43	1.67	1	164	1.67	1	42	1.67	0	-	-	0	-	-	0	-	-
1H9D	4	1	1.44	3	1	1.44	4	1	1.44	4	1	2.36	4	1	2.36	5	1	2.36
1HCF	2	1	1.31	2	2	1.31	2	1	1.31	3	15	1.42	4	5	1.42	4	4	1.89
1HE1	4	1	1.44	3	1	1.44	5	1	1.44	3	1	1.77	4	1	1.77	4	1	1.77
1HE8	0	-	-	5	354	4.83	3	208	4.27	0	-	-	0	-	-	1	393	4.29
1HIA	10	1	1.05	10	1	0.86	11	1	1.05	8	1	1.39	8	1	1.39	9	3	1.39
1I2M	2	1	1.69	2	1	1.69	1	1	1.69	1	1	1.73	1	1	1.73	1	1	1.73
1I4D	2	5	2.00	2	29	2.00	2	1	2.00	3	55	1.80	2	23	1.56	4	1	1.56
1I9R	1	225	3.84	1	88	3.84	1	93	3.84	0	-	-	1	227	3.07	1	442	3.07
1IB1	1	9	2.21	1	4	2.68	1	4	1.40	2	9	2.72	2	8	2.72	2	33	1.62
1IBR	1	1	2.45	1	1	2.50	1	1	2.45	1	5	2.13	1	1	2.13	1	1	2.13
1IJK	1	87	1.70	1	1	1.58	2	1	1.58	1	212	1.96	3	1	1.96	3	1	1.79
1IQD	1	1	1.33	1	1	1.33	1	1	1.33	1	11	1.93	3	1	1.74	3	1	1.74
1IRA	1	1	1.09	2	1	1.09	2	1	1.09	2	1	1.22	2	1	1.50	3	1	1.50
1J2J	4	32	0.99	4	12	0.99	8	1	0.99	1	1215	3.52	6	39	1.57	9	1	1.91
1JIW	1	1	1.22	1	1	1.22	1	1	1.22	1	16	1.28	4	1	1.28	5	1	1.28
1JK9	2	2	1.33	3	5	1.33	3	2	1.13	4	16	1.23	4	3	2.10	4	1	1.41
1JMO	2	1	1.54	2	1	1.54	3	1	1.54	3	1	1.71	3	1	1.71	3	1	1.71
1JPS	3	1	2.28	3	1	2.28	3	1	2.28	2	39	2.34	2	7	2.05	2	36	2.41
1JTG	2	1	1.35	3	1	1.35	3	1	1.35	5	1	1.23	5	1	1.23	5	1	1.23
1JWH	1	1778	1.93	1	2737	1.39	2	39	4.75	0	-	-	1	597	1.84	1	81	1.84
1JZD	1	1	1.12	1	1	1.12	1	1	1.24	1	20	1.98	2	2	1.98	4	1	1.98
1K4C	3	3	2.44	4	2	2.44	3	1	2.44	3	5	2.25	3	1	2.25	3	2	2.25
1K5D	2	1	1.69	2	1	1.69	2	1	1.69	2	1	1.73	2	1	1.73	2	1	1.73
1K74	1	1	2.33	1	1	2.33	1	1	2.33	1	3	3.00	2	1	3.00	2	1	2.39
1KAC	2	32	1.21	3	6	1.21	3	59	1.76	1	858	1.27	3	19	2.09	5	3	3.42
1KKL	2	196	1.33	2	463	1.33	1	702	3.60	2	251	1.55	2	198	1.30	6	510	3.54
1KLU	1	172	2.47	1	30	2.47	1	87	2.47	0	-	-	0	-	-	1	4	2.26
1KTZ	4	145	1.11	8	10	1.11	6	71	1.11	0	-	-	6	5	1.62	7	63	1.62
1KXP	1	1	2.33	1	1	2.33	1	1	2.33	1	3	2.88	1	1	2.14	1	1	2.14
1KXQ	4	1	1.33	5	1	1.33	6	1	1.33	5	3	1.36	7	1	1.36	5	1	1.36
1LFD	4	1	1.32	5	1	1.32	4	1	1.32	3	48	1.10	5	1	1.71	5	4	1.10
1M10	2	3	2.35	2	1	2.35	2	1	2.93	2	107	2.13	2	1	2.52	2	1	3.40
1MAH	5	1	0.80	6	1	0.80	7	1	0.80	6	1	1.52	8	1	1.52	7	1	1.52
1ML0	7	1	1.03	6	1	1.03	8	1	1.03	8	2	1.78	11	1	1.78	12	1	1.78
1MLC	3	75	1.54	2	43	4.65	4	8	4.65	2	358	2.39	3	62	1.37	6	20	2.39
1MQ8	3	9	1.97	2	71	1.97	2	338	1.97	2	427	2.14	1	922	2.14	2	274	2.27
1N2C	2	3	1.95	2	4	2.38	2	3	1.95	0	-	-	2	1	2.08	2	1	2.08
1N8O	3	1	2.44	3	1	2.44	3	1	2.44	3	1	2.99	3	1	2.99	2	1	2.99
1NCA	2	1	2.04	1	1	2.04	2	1	2.04	1	24	1.85	1	1	1.85	1	5	1.85
1NSN	3	502	1.11	3	93	1.11	3	51	1.36	3	742	3.32	3	77	2.16	4	34	2.16
1NW9	7	1	0.94	7	1	0.94	8	1	0.94	8	1	1.10	8	1	1.10	7	1	2.07
1OC0	8	2	0.93	11	2	0.93	6	1	0.93	3	92	1.11	7	3	1.11	10	3	1.11
1OFU	1	71	2.18	1	27	2.18	2	1	2.18	2	446	1.74	2	5	2.70	3	1	1.74
1OPH	3	3	1.67	4	12	1.67	5	3	1.53	3	169	1.42	5	18	1.42	6	71	1.42
1OYV	4	1	1.69	3	1	1.69	4	1	1.69	5	5	1.23	3	1	1.86	4	1	1.86
1PPE	14	1	1.31	14	1	1.31	16	1	1.38	19	1	1.33	20	1	1.33	21	1	2.30
1PVH	1	61	2.03	1	327	2.03	3	1	2.03	3	124	1.81	3	34	1.81	2	69	1.81
1PXV	2	1	1.01	3	1	1.01	3	1	1.01	6	1	1.25	6	1	1.25	5	1	2.05
1QA9	1	499	1.26	2	1	1.26	3	44	1.26	0	-	-	3	24	1.46	5	23	1.46
1QFW	1	68	3.88	1	30	3.88	1	29	3.88	1	56	3.65	1	20	3.65	1	345	3.65
1R0R	10	1	1.01	9	1	1.01	11	1	1.01	9	1	1.29	15	2	1.29	14	2	1.29
1R6Q	6	1	1.40	5	1	1.40	5	1	1.40	3	372	3.70	4	21	1.39	2	66	1.39
1R8S	2	1	1.47	4	1	1.47	4	1	1.47	4	1	1.81	4	1	1.81	4	1	1.81
1RLB	1	618	1.30	1	174	1.30	1	22	1.30	0	-	-	1	89	1.24	2	12	1.79
1RV6	3	4	1.55	2	3	1.55	2	2	1.55	2	359	1.72	3	17	1.72	8	1	1.72
1S1Q	4	4	0.99	4	8	0.99	4	5	0.99	3	771	1.60	4	41	1.60	3	43	1.60
1SBB	1	194	2.74	1	210	2.74	1	584	2.74	1	3114	2.94	1	1707	2.94	0	-	-
1SYX	5	71	1.45	8	1	1.85	7	3	1.85	3	128	1.28	7	1	3.38	5	416	3.38
1T6B	2	2	1.66	2	5	1.66	2	3	1.66	1	62	2.66	3	12	2.66	3	1	2.66
1TMQ	2	1	0.95	2	1	0.95	2	1	0.95	2	2	2.09	2	2	2.09	5	1	2.07
1UDI	9	1	1.35	9	1	1.53	11	1	1.03	6	1	2.15	10	1	2.15	10	1	1.18
1US7	1	144	2.32	1	58	2.32	1	36	2.32	0	-	-	1	381	2.59	1	296	2.81

Table 3.5 (continue)

PDB ID	rPSC			rPSC+ES			rPSC+ES+RDE			ZDOCK 2.1 (PSC)			ZDOCK 2.3 (PSC+ES+DE)			ZDOCK 3.0 (PSC+ES+IFACE)		
	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD
1VFB	3	3	1.03	2	8	1.59	2	1	1.59	2	196	1.65	1	198	1.65	1	64	1.65
1WDW	1	1	1.49	2	1	1.49	2	1	1.49	1	1	2.53	3	1	2.53	2	1	2.53
1WEJ	4	16	1.44	3	10	1.44	3	2	1.44	3	523	1.57	3	162	1.57	5	17	1.57
1WQ1	2	1	1.45	4	1	1.45	4	1	1.04	2	1	1.82	2	1	1.82	3	1	1.82
1XD3	4	1	1.05	3	1	1.05	5	1	1.05	2	1	1.77	5	1	1.66	7	1	1.77
1XQS	1	1	2.12	2	1	2.12	2	1	2.12	1	2	1.79	1	1	1.79	1	1	1.79
1XU1	4	1	1.29	4	1	0.98	5	1	0.98	5	7	3.15	9	2	1.48	16	3	3.15
1Y64	1	61	1.61	1	2	1.83	1	1	1.83	1	29	2.28	1	2	2.28	1	109	2.28
1YVB	1	1	2.06	1	1	2.06	1	1	2.06	1	31	2.03	1	2	2.76	2	1	2.76
1Z0K	5	2	4.29	5	2	1.18	6	1	1.69	6	4	1.57	6	1	0.92	6	1	0.92
1Z5Y	4	3	1.64	4	4	1.67	3	3	1.64	2	21	2.11	3	5	2.11	3	1	2.11
1ZHH	1	4	1.66	1	1	1.52	2	1	1.52	2	31	2.28	3	1	2.08	3	1	2.28
1ZHI	2	25	1.13	3	1	1.13	4	3	1.13	1	75	2.15	2	6	2.14	2	3	2.14
1ZLI	4	1	1.33	4	1	1.33	4	1	1.35	5	1	1.36	5	1	1.36	6	1	4.29
1ZM4	1	737	4.60	3	289	1.76	2	76	4.60	0	-	-	1	318	1.54	1	11	1.54
2A5T	2	1	1.66	2	1	1.66	2	1	1.66	1	7	2.03	2	2	2.03	2	2	2.03
2A9K	3	1	2.20	2	1	2.20	2	1	2.20	1	4	2.51	2	1	2.51	2	1	2.03
2ABZ	8	1	1.53	6	10	1.53	8	16	2.57	8	42	3.15	9	8	3.26	8	13	3.26
2AJF	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	1	143	1.88
2AYO	5	1	1.48	6	1	1.48	6	1	1.48	7	1	2.15	8	1	2.15	7	1	2.15
2B4J	4	1	1.04	4	2	1.04	3	1	1.04	2	761	1.88	4	14	1.88	5	71	3.65
2B42	4	1	1.42	3	1	1.42	5	1	1.42	5	1	1.68	5	1	1.68	5	1	1.68
2BTF	4	1	1.33	5	1	1.33	5	1	1.33	4	4	1.84	4	1	1.24	5	6	1.24
2C0L	1	1	2.01	1	1	2.01	1	1	2.01	1	5	1.51	2	1	1.51	2	1	1.51
2CFH	5	1	1.79	5	1	1.79	5	1	1.79	4	1	1.54	6	1	1.54	6	1	1.57
2FD6	2	228	1.98	1	877	4.39	2	157	1.98	0	-	-	2	249	1.74	1	21	1.74
2FJU	1	488	1.13	1	166	1.13	2	78	1.13	0	-	-	1	496	1.39	2	3	1.39
2G77	4	1	1.43	4	1	1.43	4	1	1.43	4	1	1.37	5	1	2.18	4	1	2.18
2H7V	1	3	1.56	1	2	1.56	1	1	1.56	1	371	2.29	2	18	2.29	2	1	2.29
2HLE	3	1	1.23	3	1	1.23	3	1	1.23	4	2	1.86	4	1	1.86	3	2	1.86
2HMI	2	310	4.46	1	481	4.46	0	-	-	0	-	-	0	-	-	1	10	4.53
2HQS	3	1	1.35	4	1	1.35	4	1	1.35	2	4	1.71	5	1	1.71	3	3	1.71
2HRK	1	1	1.17	3	1	1.17	6	1	1.17	4	20	2.05	4	7	2.05	5	5	2.54
2I9B	2	1	1.07	2	1	1.07	2	1	1.07	2	1	1.37	3	1	1.37	3	1	1.83
2I25	6	1	1.15	6	1	1.15	6	1	1.15	4	45	3.16	6	1	1.44	6	1	3.16
2IDO	7	1	1.20	7	1	1.20	8	1	1.20	7	1	1.60	7	1	1.60	7	1	1.60
2J0T	2	1	1.59	1	1	1.59	2	1	1.59	1	2	1.43	2	6	1.43	2	66	1.43
2J7P	1	1	1.82	2	1	1.82	2	1	1.82	2	1	2.08	2	1	2.08	1	1	1.45
2JEL	4	1	1.09	3	2	1.09	7	1	1.09	2	469	2.75	4	209	2.75	8	6	1.26
2MTA	4	5	0.99	4	1	0.99	8	1	0.99	3	226	1.48	10	5	1.87	11	1	2.57
2NZ8	2	2	2.07	2	1	2.07	1	1	2.07	4	10	1.93	4	2	1.93	3	1	1.82
2O3B	2	2	1.29	2	1	1.29	2	7	1.29	1	289	1.48	2	2	1.48	2	1	1.39
2O8V	8	1	1.59	8	1	1.59	9	1	1.59	5	4	1.30	7	1	1.30	9	1	1.47
2OOB	0	-	-	1	2081	4.63	1	887	1.13	0	-	-	0	-	-	0	-	-
2OOR	3	1	1.73	4	1	1.73	4	1	1.73	3	3	1.68	3	1	1.68	4	1	1.68
2OT3	4	1	1.84	4	1	1.84	4	1	1.84	4	1	1.31	6	1	1.31	4	1	1.31
2OUL	2	1	1.24	2	1	1.24	2	1	1.24	3	8	2.05	3	2	2.05	2	2	2.05
2OZA	1	1	1.74	1	1	1.74	1	1	1.66	1	1	1.98	1	1	1.98	1	1	1.98
2PCC	1	1169	1.03	1	91	1.03	1	81	1.03	0	-	-	1	216	1.64	2	32	1.79
2QFW	2	9	2.43	3	6	3.67	2	14	2.43	1	1313	2.52	1	314	1.81	1	45	1.81
2SIC	2	1	1.85	2	1	1.85	3	1	1.85	3	6	1.72	3	2	1.98	5	1	1.98
2SNI	6	2	1.71	7	2	1.71	7	1	1.61	9	1	1.10	9	1	1.10	7	1	1.10
2UUY	6	1	1.64	6	1	1.64	9	1	1.64	8	1	1.46	8	1	2.89	9	10	2.89
2VDB	5	1	1.02	5	1	1.02	5	1	1.02	7	8	1.55	6	1	1.56	10	1	3.72
2VIS	1	938	3.74	1	1165	3.74	1	1247	3.74	0	-	-	1	1804	2.23	2	213	3.27
2Z0E	1	1	1.57	3	1	1.57	3	1	1.57	4	1	1.55	3	1	1.55	3	1	1.55
3BP8	2	21	0.99	4	57	1.11	7	20	0.99	1	1183	1.32	2	563	1.32	3	388	3.10
3CPH	1	5	1.74	2	1	1.74	3	1	1.74	1	24	2.06	3	1	2.06	3	1	2.06
3D5S	3	5	0.88	5	3	1.48	5	1	1.48	3	18	2.55	7	1	2.55	5	4	1.28
3SGQ	8	1	0.78	9	1	1.28	9	1	0.78	8	5	1.43	7	8	1.43	5	56	3.51
4CPA	17	2	1.03	16	1	1.03	18	4	1.03	14	3	3.60	16	7	3.61	20	2	4.10
7CEI	2	1	1.49	2	1	1.49	3	1	1.49	4	7	4.99	6	1	4.99	5	2	4.77
BOYV	1	26	2.37	1	138	2.37	1	53	2.37	0	-	-	0	-	-	0	-	-

Table 3.6: Docking prediction performance of MEGADOCK and ZDOCK for the unbound docking test cases in protein–protein docking benchmark 4.0. #NND denotes the number of near-native decoy in the top 3,600 predictions, Best Rank is the rank of first near-native decoy, and RMSD is the L-RMSD of first near-native decoy ($RMSD_{best}$).

PDB ID	rPSC			rPSC+ES			rPSC+ES+RDE			ZDOCK 2.1 (PSC)			ZDOCK 2.3 (PSC+ES+DE)			ZDOCK 3.0 (PSC+ES+IFACE)		
	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD
1A2K	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1ACB	0	-	-	1	1640	4.94	1	2480	4.94	1	1666	4.78	3	948	4.08	4	204	4.78
1AHW	1	67	3.23	1	63	3.23	1	97	3.23	1	878	2.52	1	1149	2.52	1	1550	2.52
1AK4	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1AKJ	0	-	-	1	1872	4.99	1	2885	2.21	1	2029	4.48	2	173	3.71	1	1383	3.33
1ATN	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1AVX	1	1466	4.18	2	992	4.18	1	1021	4.18	1	2047	4.76	1	2435	4.76	1	228	3.16
1AY7	0	-	-	1	1483	3.69	0	-	-	2	1434	4.92	0	-	-	3	1510	4.96
1AZS	0	-	-	0	-	-	0	-	-	0	-	-	1	624	2.55	2	88	2.55
1B6C	1	2640	3.02	1	1965	3.02	1	157	3.02	1	1566	3.18	1	154	3.18	1	24	2.97
1BGX	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1BJ1	2	131	4.59	1	361	4.59	1	16	4.59	2	179	4.29	2	69	4.29	2	2	4.29
1BKD	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1BUH	1	1759	4.31	0	-	-	0	-	-	0	-	-	0	-	-	6	45	4.65
1BVK	0	-	-	0	-	-	0	-	-	1	3125	4.60	0	-	-	2	256	4.32
1BVN	0	-	-	1	466	4.83	5	437	4.83	4	91	4.23	7	16	4.28	8	2	4.28
1CGI	0	-	-	0	-	-	0	-	-	3	467	4.95	2	46	4.78	5	151	4.86
1CLV	0	-	-	0	-	-	0	-	-	4	258	3.22	6	64	4.51	16	1	3.22
1D6R	0	-	-	0	-	-	0	-	-	1	1564	4.48	1	1047	4.48	0	-	-
1DE4	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1DFJ	0	-	-	0	-	-	0	-	-	1	9	3.73	1	4	3.73	1	3	2.83
1DQJ	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1E4K	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1E6E	0	-	-	2	89	2.57	3	269	2.49	1	3375	4.81	6	240	3.92	7	15	3.16
1E6J	2	314	3.74	2	352	3.74	4	17	3.74	0	-	-	4	519	4.28	2	140	4.24
1E96	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1EAW	2	82	4.47	3	49	4.47	4	94	4.47	8	3	4.64	8	5	4.64	6	39	4.64
1EER	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1EFN	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1EWY	3	746	3.69	5	69	4.68	6	60	4.68	1	2993	4.37	6	24	4.78	6	11	4.78
1EZU	0	-	-	0	-	-	0	-	-	1	710	3.02	1	1096	3.02	0	-	-
1F6M	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1F34	0	-	-	0	-	-	1	1637	2.41	1	681	3.16	1	147	3.16	1	634	3.12
1F51	0	-	-	1	2831	3.10	2	1049	4.38	1	1103	4.46	1	2344	4.46	1	1438	3.06
1FAK	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1FC2	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1FCC	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1FFW	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	2	65	4.86
1FLE	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1FQ1	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1FQJ	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1FSK	4	1	1.83	4	1	1.83	4	1	1.83	4	14	2.12	4	1	2.12	4	1	2.12
1GCQ	1	2538	2.84	1	708	3.26	2	109	3.26	0	-	-	0	-	-	0	-	-
1GHQ	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1GL1	2	2493	4.04	0	-	-	0	-	-	8	149	3.69	5	953	3.60	2	262	3.69
1GLA	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1GP2	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1GPW	1	68	2.25	0	-	-	0	-	-	2	5	2.40	2	13	2.40	2	80	2.40
1GRN	3	78	3.26	2	281	3.26	3	1417	3.26	1	443	4.56	0	-	-	1	2653	4.56
1GXD	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1H1V	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1H9D	1	475	4.75	0	-	-	1	45	4.75	0	-	-	0	-	-	2	392	4.58
1HCF	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	3	103	4.29
1HE1	0	-	-	0	-	-	0	-	-	2	182	4.46	0	-	-	0	-	-
1HE8	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1HIA	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1I2M	0	-	-	0	-	-	0	-	-	0	-	-	1	2373	2.57	1	43	2.57
1I4D	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	1	1020	4.85
1I9R	1	136	3.59	1	186	3.59	0	-	-	1	348	4.47	0	-	-	0	-	-
1IB1	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1IBR	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1IJK	0	-	-	0	-	-	1	526	2.61	0	-	-	0	-	-	1	462	2.33
1IQD	1	2853	3.64	0	-	-	0	-	-	0	-	-	1	432	4.53	3	46	4.48
1IRA	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1J2J	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1JIW	0	-	-	0	-	-	0	-	-	1	2763	4.75	0	-	-	0	-	-
1JK9	1	1014	4.68	1	1371	4.68	1	602	4.68	0	-	-	2	394	3.78	1	1544	4.81
1JMO	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1JPS	2	283	2.78	2	246	2.78	2	20	2.78	1	2231	3.38	0	-	-	1	1180	3.53
1JTG	2	3	3.26	4	1	3.68	4	1	3.68	3	1	3.83	4	1	3.83	5	1	3.83
1JWH	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	1	170	2.48
1JZD	1	2994	4.96	1	2265	4.96	1	1594	4.96	0	-	-	1	1181	4.95	1	82	4.95
1K4C	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1K5D	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1K74	1	465	3.80	1	1	3.81	1	3	4.42	1	849	4.10	1	1	3.50	1	4	3.49
1KAC	3	175	4.95	3	1301	4.27	1	538	4.95	1	2586	4.30	0	-	-	1	1554	4.26
1KKL	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1KLU	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1KTZ	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1KXP	1	30	3.04	1	1	3.04	1	4	3.04	1	172	3.39	1	2	3.63	2	1	3.63
1KXQ	1	80	2.93	1	386	2.93	1	268	2.93	2	38	1.37	2	25	1.37	2	1	1.37
1LFD	1	2737	4.27	1	2720	4.27	1	2073	4.27	0	-	-	0	-	-	1	684	4.24
1M10	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1MAH	3	461	4.05	4	209	4.05	6	10	4.04	4	41	3.97	7	2	3.97	8	1	3.97
1ML0	1	534	3.12	1	150	3.12	3	19	3.12	3	152	3.62	4	20	4.10	7	35	3.59
1MLC	1	64	4.33	1	60	4.33	2	33	3.90	0	-	-	1	388	4.49	3	86	3.54
1MQ8	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1N2C	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1N8O	1	36	2.67	1	139	2.67	1	6	2.67	1	144	3.09	1	8	2.86	2	3	2.86
1NCA	2	8	5.00	2	28	4.76	2	9	4.76	1	339	1.97	1	58	2.75	1	183	1.97
1NSN	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1NW9	0	-	-	1	2070	4.49	0	-	-	0	-	-	0	-	-	0	-	-
1OC0	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1OFU	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	1	33	2.41
1OPH	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1OYV	1	1662	4.90	2	1029	4.90	1	1568	4.15	2	157	3.74	3	128	4.40	2	111	4.40
1PPE	8	3	4.75	9	1	4.75	15	1	3.07	17	1	4.23	16	1	2.58	17	1	2.58
1PVH	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	3	1519	4.70
1PXV	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1QA9	1	2011	2.90	1	1452	2.95	0	-	-	0	-	-	0	-	-	0	-	-
1QFW	1	94	3.73	1	119	3.73	1	191	3.73	1	491	3.63	1	65	3.63	1	326	3.26
1R0R	5	81	2.79	4	67	2.79	5	118	2.79	6	365	2.50	6	207	2.31	6	470	2.42
1R6Q	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1R8S	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1RLB	0	-	-	0	-	-	1	421	4.52	0	-	-	0	-	-	1	101	4.49
1RV6	1	2847	4.77	2	1418	2.50	1	903	2.50	0	-	-	1	422	2.46	3	1	2.45
1S1Q	3	341	2.86	2	547	2.86	6	76	2.86	1	1610	4.85	3	1397	4.27	4	481	4.48
1SBB	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1SYX	1	63	4.96	3	104	4.96	1	2829	4.96	2	108	4.99	2	84	4.99	1	17	4.90
1T6B	0	-	-	0	-	-	0	-	-	1	2133	3.16	1	495	3.16	1	196	3.16
1TMQ	1	657	4.59	0	-	-	0	-	-	1	709	2.69	2	171	4.89	4	12	4.89
1UDI	0	-	-	0	-	-	0	-	-	1	280	4.35	2	52	4.35	3	61	3.93
1US7	0	-	-	0	-	-	1	2901	4.53	0	-	-	0	-	-	2	874	4.24

Table 3.6 (continue)

PDB ID	rPSC			rPSC+ES			rPSC+ES+RDE			ZDOCK 2.1 (PSC)			ZDOCK 2.3 (PSC+ES+DE)			ZDOCK 3.0 (PSC+ES+IFACE)		
	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD	#NND	Best Rank	RMSD
1VFB	2	393	2.07	0	-	-	1	2169	4.39	1	3050	4.99	0	-	-	1	2261	4.93
1WDW	2	11	2.80	2	43	2.15	2	6	2.15	2	12	2.87	2	1	2.80	2	4	2.87
1WEJ	1	1583	1.64	3	183	1.64	3	170	1.79	1	3307	4.67	3	620	1.86	6	156	1.88
1WQ1	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	1	1618	4.34
1XD3	0	-	-	0	-	-	2	611	4.97	0	-	-	2	1662	4.31	8	6	4.85
1XQS	0	-	-	1	718	4.47	0	-	-	0	-	-	1	563	4.46	1	139	4.46
1XU1	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1Y64	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1YVB	1	9	3.94	1	34	3.94	1	4	3.94	0	-	-	0	-	-	0	-	-
1Z0K	1	1888	4.87	0	-	-	1	208	4.87	0	-	-	1	601	4.83	0	-	-
1Z5Y	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	3	87	4.99
1ZHH	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1ZHI	0	-	-	1	869	4.36	2	925	4.46	0	-	-	1	1424	4.47	2	85	4.47
1ZLI	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
1ZM4	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	1	25	3.29
2A5T	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2A9K	0	-	-	0	-	-	0	-	-	0	-	-	1	859	2.28	0	-	-
2ABZ	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2AJF	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2AYO	2	483	4.11	2	114	3.40	1	56	3.40	3	171	4.95	4	6	3.05	6	61	4.72
2B4J	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2B42	0	-	-	0	-	-	0	-	-	2	1	1.83	2	1	1.83	2	1	1.83
2BTF	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	4	170	3.93
2C0L	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2CFH	1	938	4.98	2	112	4.98	2	2	4.98	2	700	4.74	2	64	4.74	2	6	4.74
2FD6	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2FJU	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	1	1276	4.84
2G77	0	-	-	0	-	-	0	-	-	0	-	-	1	1362	3.83	5	19	3.83
2H7V	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2HLE	0	-	-	1	21	4.35	2	32	4.35	0	-	-	1	164	3.60	2	217	4.17
2HMI	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2HQS	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2HRK	0	-	-	0	-	-	0	-	-	1	416	4.92	1	923	4.92	1	779	4.92
2I9B	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2I25	0	-	-	1	1105	3.35	0	-	-	0	-	-	2	871	3.68	6	9	1.82
2IDO	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2J0T	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2J7P	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2JEL	1	2607	4.33	0	-	-	3	937	4.77	1	2621	3.69	1	1591	4.34	10	119	4.38
2MTA	2	195	4.93	1	167	4.93	4	20	4.93	0	-	-	2	2867	2.07	8	196	4.65
2NZ8	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2O3B	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2O8V	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2OOB	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2OOR	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	1	924	4.55
2OT3	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2OUL	2	1	1.65	2	1	1.65	3	1	1.65	2	3	2.28	3	1	2.25	3	1	2.25
2OZA	0	-	-	0	-	-	0	-	-	1	1074	4.81	1	2234	4.81	0	-	-
2PCC	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2QFW	1	40	3.61	1	26	3.61	1	118	3.61	0	-	-	1	702	3.41	1	98	4.15
2SIC	2	547	3.33	2	1173	3.33	2	548	3.33	1	706	2.18	1	235	2.18	2	4	2.18
2SNI	0	-	-	0	-	-	0	-	-	0	-	-	1	2427	4.68	3	1376	3.63
2UUY	0	-	-	0	-	-	1	2633	4.81	1	1349	4.61	1	1511	4.61	1	3411	4.61
2VDB	3	27	4.31	3	42	4.31	3	89	4.31	4	600	4.24	3	50	1.75	4	17	1.75
2VIS	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
2Z0E	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
3BP8	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
3CPH	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
3D5S	2	862	3.40	4	44	3.40	6	11	3.40	2	615	3.94	6	3	2.12	5	60	2.12
3SGQ	4	19	4.73	2	17	4.73	6	8	4.73	2	293	3.94	4	45	4.06	3	675	2.11
4CPA	4	257	4.03	6	61	4.03	6	34	4.03	10	8	4.76	13	6	3.56	17	1	3.55
7CEI	2	43	3.60	2	7	3.60	2	3	2.84	2	429	3.94	2	10	3.94	3	23	3.94
BOYV	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-

Table 3.7: The sum of #NND values (Σ#NND) and the number of cases with at least one near-native decoy in the top 100 scored decoys ($\#successes_{100}$).

	bound					
	rPSC	rPSC+ES	rPSC+ES+RDE	ZDOCK 2.1	ZDOCK 2.3	ZDOCK 3.0
Σ#NND	545	593	661	537	693	783
$\#successes_{100}$	149	156	163	116	144	154
	unbound					
	rPSC	rPSC+ES	rPSC+ES+RDE	ZDOCK 2.1	ZDOCK 2.3	ZDOCK 3.0
Σ#NND	103	116	155	143	193	299
$\#successes_{100}$	22	22	30	13	31	47

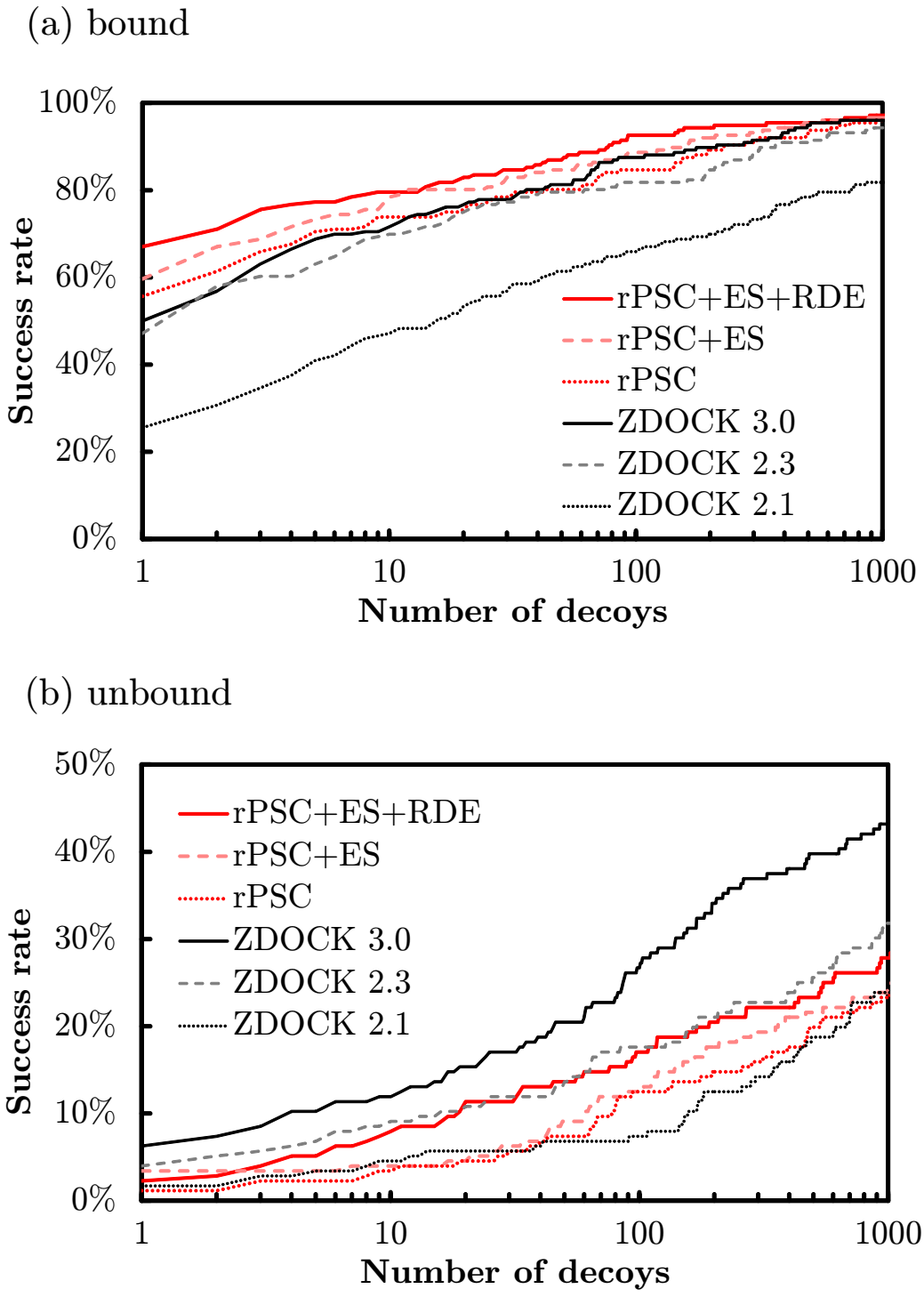


Figure 3.5: Success Rate for all test cases of benchmark dataset. The Success Rate was defined as the percentage of cases with near-native decoys for a given number of top-ranked docking predictions per test case.

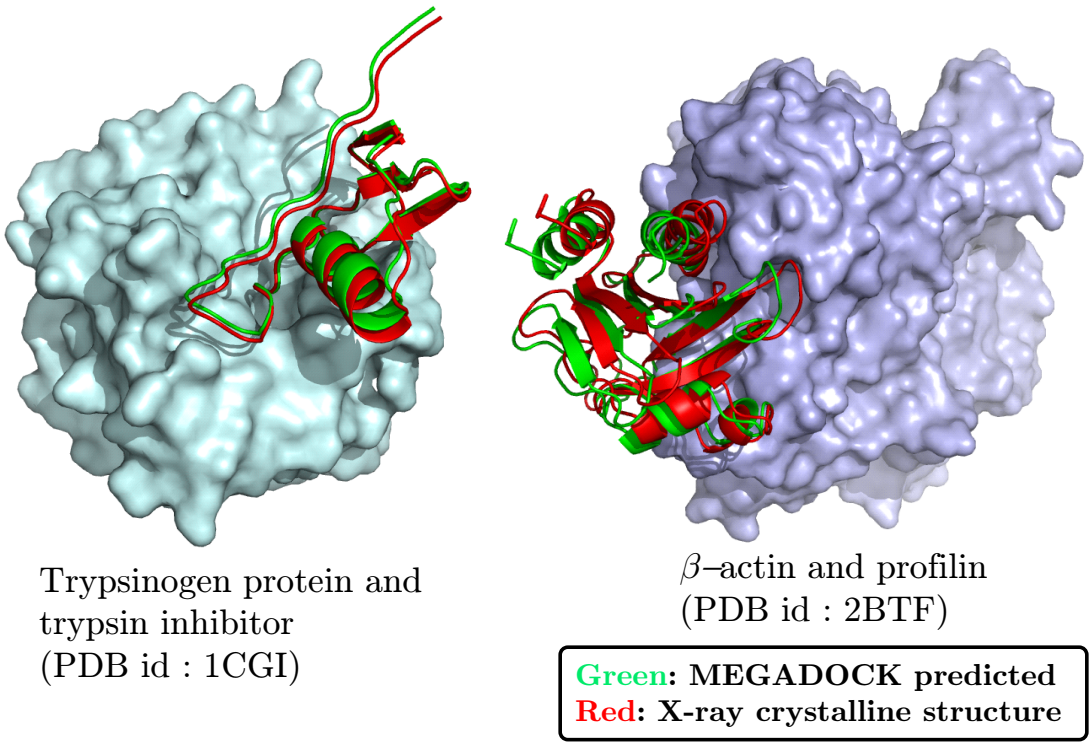


Figure 3.6: Complex structure predicted by docking (left: 1CGI; right: 2BTF). Proteins shown by the surface correspond to receptors whereas those shown by ribbon representations correspond to ligands both from bound structures. Green colored ligands show the prediction by MEGADOCK, whereas red colored ligands are X-ray structures.

Table 3.8: Total time for 352 docking calculations using the benchmark dataset.

	rPSC	rPSC+ES	rPSC+ES+RDE	ZDOCK 2.1	ZDOCK 2.3	ZDOCK 3.0
time (hour)	69.0	69.7	70.0	85.5	309.2	684.3
speedup from ZDOCK 3.0	9.91	9.82	9.78	8.00	2.21	(1.0)

Table 3.9: Ratio of time spent for each process in the total docking time (average of 352 dockings of protein–protein docking benchmark 4.0 [82], calculated with single thread setting)

Calculation	Ratio of time spent for the process [%] (mean ± s.d.)
Receptor voxelization and FFT	1.19 ± 0.62
Ligand rotation and voxelization	6.41 ± 3.13
Ligand FFT	40.38 ± 2.79
Inverse FFT	45.99 ± 2.25
Post processes	6.02 ± 1.46

Table 3.9 shows the ratio of time spent for each MEGADOCK process in the total docking time (average of 352 dockings of protein–protein docking benchmark 4.0. Ligand FFT and inverse FFT consists large part (on average 86.4% for 352 dockings) of the calculation time.

FFT consumed approximately 86.4% of the total docking time. In our case, the scoring function with only rPSC consumed 11.76 min on average, and hence, the time for FFT was estimated to be 10.16 min. The addition of a correlation function using the FFT results led to calculation times that were 1.86 times longer than the simple scoring function, or, in other words, an 10.16 min increase. In the proposed rPSC+ES+RDE function, by avoiding the addition of FFT, the time increase was suppressed to approximately 0.17 min; that is a 98.3% reduction in time than the simple FFT addition. Table 3.8 also shows that MEGADOCK was approximately 4.4 times faster than ZDOCK 2.3 and 9.8 times faster than ZDOCK 3.0. Since FFT takes most of the execution time of MEGADOCK, if we increase the FFT correlation function to two or three to get better performance of docking, calculation time will also increase 2- or 3-fold.

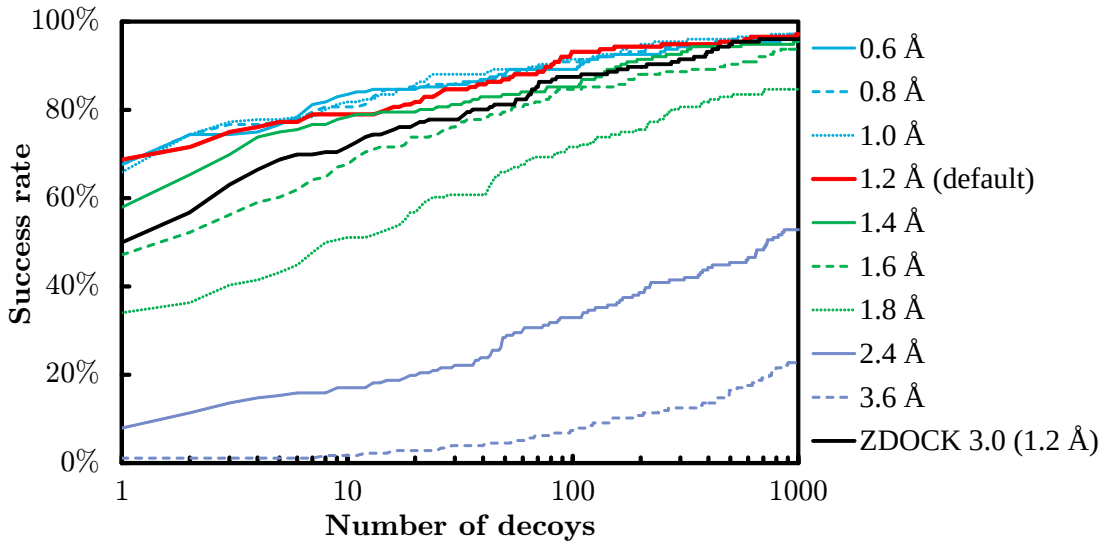
3.3.5 Parameter of grid width

Commonly, the FFT grid-based protein–protein docking methods used the grid width (spacing) of 1.2 Å. In comparison to ZDOCK, we used same parameter of grid width. In this subsection, we show the results of our method with other values of grid width. Fig. 3.7 shows the docking success rate in various grid width and Table 3.10 shows the total time consumed for docking the benchmark 4.0 dataset.

The calculation time get greater with smaller grid width. The theoretical ratio from the grid width of 1.2 Åin Table 3.10 were estimated using a protein with the FFT size of $N = 128$ as an example. When the grid width is changed to $v = 0.8$ Å, the FFT size is changed to $N = 192$ ($128 \times 1.2/0.8$). In theory FFT takes the order of $\mathcal{O}(N^3 \log N)$ for calculation. Therefore calculations involving a size of $N = 192$ ($v = 0.8$ Å) FFT should take 0.27 times ($(128^3 \log 128)/(192^3 \log 192) = 0.273...$) the elapsed time of a corresponding calculation involving a size of $N = 128$ ($v = 1.2$ Å) FFT. Although Table 3.10 shows the total time of various size of proteins, the speed up ratio from $v = 1.2$ Å is closed to theoretical value.

In summary of these results, although using the grid width parameters of $v = 1.4$–1.6 Å is efficient for faster calculation (26.2 times faster than ZDOCK 3.0), usually we should use the default grid width parameter of $v = 1.2$ Å.

(a) bound



(b) unbound

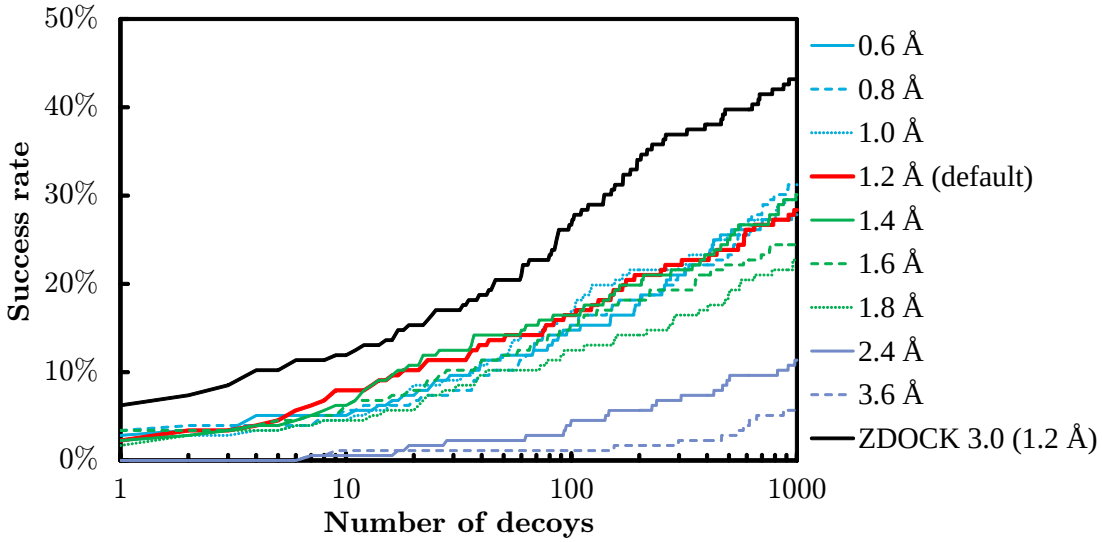


Figure 3.7: Success Rate for all test cases of benchmark dataset with various grid width parameters. The Success Rate was defined as the percentage of cases with near-native decoys for a given number of top-ranked docking predictions per test case.

Table 3.10: Total time for 352 docking calculations with various grid width parameters using the benchmark dataset. 1.2^z represents ZDOCK 3.0 (used $v = 1.2$ Å). The theoretical ratio is the case of a protein with FFT size of $N = 128$ at the grid width of $v = 1.2$ Å.

grid width v (Å)	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.4	3.6	1.2^z
time (hour)	660.1	272.0	133.9	70.0	40.8	26.1	18.0	7.2	2.0	684.3
speed up from ZDOCK 3.0	1.04	2.52	5.11	9.78	16.77	26.19	37.99	94.39	341.20	(1.00)
speed up from $v = 1.2$ Å	0.11	0.26	0.52	(1.00)	1.71	2.68	3.89	9.65	34.89	0.10
theoretical ratio from $v = 1.2$ Å	0.11	0.27	0.56	(1.00)	1.64	2.52	3.68	9.33	34.90	-

3.3.6 Large-scale parallel computing

MEGADOCK was parallelized with the MPI and OpenMP library (see Appendix A for more information) and implemented on GPUs (see Appendix B for more information). Because the calculations for each pair are almost independent, we can parallelize an all-to-all exhaustive protein–protein docking calculation task using several methods on hundreds of thousands of CPU cores. The user can specify the numbers of receptor and ligand protein data to be assigned to a single processor after considering the memory capacity. We tested this data parallelization using about 700,000 cores. When a processor is assigned for data comprising n_R receptors and n_L ligands, it calculates FFT for the first ligand with each possible rotation. The FFT results are repeatedly employed for docking with all n_R receptors to avoid redundant calculations. Subsequently, the process is repeated n_L times. MEGADOCK has an option to avoid DFT calculations and upload precalculated DFT results from the "FFT protein structure library" onto the hard disk drives. This approach is effective in a system with high I/O performance. The FFT routine in MEGADOCK uses FFT bases of {2, 3, 5, 7, 9, 11} to minimize the volume of the target 3D grid. However, if we choose too many FFT bases, it is necessary to prepare many precalculated FFT models in the library, because protein pairing is unknown a priori. In contrast, if we use GPU acceleration, it is better to simply repeat FFT calculations on a GPU with the most adequate combinations of FFT bases. We considered this in our study when we implemented our system with the aim of high computing power rather than I/O performance.

As a result, our GPU and parallel implementation achieved 37.0-fold acceleration using one computing node with three GPUs and worked in high-performance computing environments equipped with over ten thousands nodes ($\sim$ 25,000 nodes). We described more details in Appendices A and B.

3.4 Summary

In this chapter, we introduced a novel shape complementarity function rPSC and a novel desolvation free energy function RDE. rPSC and RDE represents shape complementarity, electrostatics interaction and desolvation free energy between target proteins with only one FFT correlation function without increasing the calculation time. MEGADOCK was shown to be 9.8 times faster than the conventional software ZDOCK 3.0 while maintaining acceptable docking prediction accuracies. However, to enhance the accuracy of the proposed model, further tuning of some system parameters is nec-

essary. For example, ACE was introduced only into the receptor side in the study. We are attempting to develop a new score model with both receptor and ligand ACE term using only one correlation function.

Part III

Protein–Protein Interaction Network Prediction and Its Applications

Chapter 4

Development of an Exhaustive Protein–Protein Interaction Prediction System

4.1 Introduction

In the present study, we describe the development of a rigid-body docking-based method for PPI screening based on exhaustive calculations of pseudo-binding energies among pairs of target proteins that can be applied to PPI prediction problems of mega-order data. Further, to enable applications to megaorder combinations, we developed efficient FFT-based protein–protein docking software called MEGADOCK, which is designed for exhaustive PPI screening. MEGADOCK searches the relevant interacting protein pairs by conducting protein–protein docking between the tertiary structures of the target proteins and then analyzing the distributions of high-scoring decoys.

4.2 Materials and Methods

MEGADOCK predicts the relevant PPIs according to the affinity scores calculated by the post-processing of all the docking results. The components and outline of MEGADOCK system is shown in Fig. 4.1

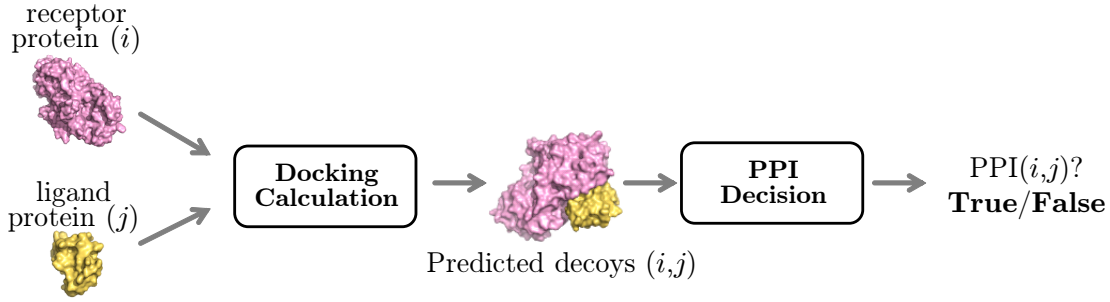


Figure 4.1: Process flow for MEGADOCK, the PPI prediction system proposed in this chapter. This system calculates FFT-based rigid-body docking by using the given receptor protein i and ligand protein j pair, generates 10,800 high-ranked decoys, and detects the interacting (i, j) pair from docking score distributions.

4.2.1 Reranking of decoys

By default, the docking part of the system outputs $3,600 \times t$ high-scoring decoys from $3,600 \times N^3$ ligand rotations and translations. In this study, we conducted docking with $t = 3$; the output was 10,800 decoys. However, some decoys with high docking scores often exhibit high binding energies when examined in more detailed methods. To reduce such unwanted structures, we applied re-ranking of the high-scoring decoys. This process collects near native decoys with high ranking, thereby excluding decoys with unrealistic, high binding energies. We used ZRANK [78] because it calculates the binding energy of each decoy based on the van der Waals energy, electrostatic energy and desolvation energy among the atoms in close contact.

4.2.2 PPI decision

From the results of docking and reranking calculations, we predicted whether a protein pair can interact or not. The PPI(i, j) of protein i and j evaluation value E is defined as follows:

$$E = \frac{S_1 - \mu}{\sigma}, \tag{4.1}$$

$$\mu = \frac{1}{D}\sum_{k=1}^{D} S_k, \tag{4.2}$$

$$\sigma^2 = \frac{1}{D}\sum_{k=1}^{D} (S_k - \mu)^2, \tag{4.3}$$

Table 4.1: The selected 44 complex structures from the protein–protein docking benchmark 2.0 dataset (small dataset)

1ACB	1AK4	1ATN	1AVX	1AY7	1B6C	1BUH	1BVN
1CGI	1D6R	1DFJ	1E6E	1E96	1EAW	1EWY	1F34
1FC2	1FQ1	1FQJ	1GCQ	1GHQ	1GRN	1H1V	1HE1
1HE8	1I2M	1IBR	1KAC	1KTZ	1KXP	1KXQ	1M10
1MAH	1PPE	1QA9	1SBB	1TMQ	1UDI	1WQ1	2BTF
2PCC	2SIC	2SNI	7CEI				

where S_1 is the top-ranked decoy's docking score for a protein pair, S_k is the k-th ranked decoy's docking score, and D is the number of decoys. In this study, we generated 10,800 decoys ($= D$) by using MEGADOCK. We concluded that a pair interacts if E is larger than threshold E^*:

$$\mathrm{PPI}(i,j) = \begin{cases} \text{True} & \text{if } E > E^* \\ \text{False} & \text{otherwise} \end{cases} \tag{4.4}$$

4.2.3 Dataset

Data for protein complexes were selected from protein–protein docking benchmark 2.0 [80] (Table 4.1) and protein–protein docking benchmark 4.0 [82] (Table 4.2), both from bound structures and used to evaluate the performance of our system. Each of the selected 44 complexes (called the small dataset) for optimization of the parameter t and 120 complexes (called the large dataset) for the evaluation of larger datasets consisted of a pair of monomer proteins (this selection of data was prepared based on a personal communication with Dr. Ryotaro Koike and Dr. Motonori Ota. All the complexes selected consisted of two monomers).

We conducted docking and PPI prediction processes on all combinations of all receptors and all ligand structures ($44 \times 44 = 1{,}936$ and $120 \times 120 = 14{,}400$ combinations) according to the prediction procedure.

4.2.4 Prediction accuracy measure

Each prediction of the possibilities of interactions in the given protein pair was evaluated as true positive (TP), false positive (FP), true negative (TN) and false negative (FN). For the benchmark data, we assumed 44 TP interac-

Table 4.2: The selected 120 complex structures from the protein–protein docking benchmark 4.0 dataset (large dataset)

1ACB	1AK4	1ATN	1AVX	1AY7	1B6C	1BKD	1BUH
1BVN	1CGI	1CLV	1D6R	1DFJ	1E6E	1E96	1EAW
1EFN	1EWY	1F6M	1F34	1FC2	1FFW	1FLE	1FQ1
1FQJ	1GCQ	1GHQ	1GL1	1GLA	1GPW	1GRN	1GXD
1H1V	1H9D	1HE1	1HE8	1I2M	1IBR	1IRA	1J2J
1JIW	1JK9	1JTG	1KAC	1KTZ	1KXP	1KXQ	1LFD
1M10	1MAH	1MQ8	1N8O	1NW9	1OC0	1OPH	1OYV
1PPE	1PVH	1PXV	1QA9	1R0R	1R6Q	1R8S	1S1Q
1SBB	1SYX	1T6B	1TMQ	1UDI	1US7	1WQ1	1XD3
1XQS	1Y64	1YVB	1Z0K	1Z5Y	1ZHH	1ZHI	1ZLI
1ZM4	2A5T	2A9K	2ABZ	2AJF	2AYO	2B42	2BTF
2C0L	2CFH	2FJU	2G77	2H7V	2HLE	2HQS	2HRK
2I9B	2I25	2IDO	2J0T	2J7P	2NZ8	2O3B	2O8V
2OOB	2OT3	2OUL	2OZA	2PCC	2SIC	2SNI	2UUY
2VDB	2Z0E	3CPH	3D5S	3SGQ	4CPA	7CEI	BOYV

tions in the small dataset and 120 TP interactions in the large dataset, where each protein has exclusively one interacting partner from the same crystal structure as the protein complex. The overall performance of the screening system was evaluated by employing the F-measure, the harmonic mean of the Precision (#TP/(#TP+#FP)) and the Recall (#TP/(#TP+#FN)). We also show the Accuracy ((#TP+#TN)/(#TP+#FN+#FP+#TN)) to show comparison of PPI prediction performance with previous works, however the Accuracy value is not appropriate to evaluate the all-to-all PPI prediction with small positives and large negatives.

4.3 Results and Discussion

4.3.1 Screening of relevant interacting protein pairs by all-to-all docking

Table 4.3 shows the performance of the PPI prediction with the small dataset. The performance was improved by introducing the re-ranking process rather than using the docking results alone. Moreover, the docking parameter ($t = 3$) that led to the best performance had the F-measure value of 0.415 (Precision 0.447, Recall 0.386, Accuracy

0.975). The performance of the application to the large dataset gave an F-measure value of 0.231 (Precision 0.500, Recall 0.150, Accuracy 0.992) using this setting ($t = 3$) with the PPI prediction parameter $E^* = 7.3$. In related work of Yoshikawa *et al.* [24], they used ZDOCK and their original post-docking process named affinity evaluation and prediction (AEP). Yoshikawa *et al.* have shown the prediction performance of their method as F-measure value of 0.063, Accuracy value of 0.902 on the 84×84 bound dataset. Our result (F-measure value of 0.231 with 0.992 Accuracy on the 120×120 bound dataset) performed significantly better than theirs.

The receiver-operator characteristics (ROC) curve [85] with the large dataset and this setting ($t = 3$) is shown in Fig. 4.2. The ROC curve is a plot of TP and FP fractions and shows the trade-off between them. A completely random prediction would lead to a diagonal line from the left-bottom to the top-right corners in the plot. The points above the diagonal line represent the scenario that the prediction is better than random. The ROC curve in Fig. 4.2 clearly shows that our method (magenta line in Fig. 4.2 is better than random predictions. The green line in Fig. 4.2 shows the result of MEGADOCK without using reranking method. Decoy reranking performed to improve the prediction accuracy (we obtained the area under the ROC curve (AUC) value of 0.824 with using reranking and 0.703 without using reranking). In addition, our PPI decision performance was validated by using ZDOCK. The dashed line in Fig. 4.2 represented the prediction results when we swapped MEGADOCK docking engine for ZDOCK 3.0. As a result, the reranking method also performed to improve the prediction accuracy if we used ZDOCK (AUC value of 0.796 with using reranking and 0.772 without using reranking). To evaluate the improvements by introducing ZRANK, we used another dataset derived from dockground 3.0 benchmark data [86]. Table 4.4 shows our dataset which is a subset of dockground 3.0; the subset consists of only monomer protein pairs. The ROC curve with the dockground 3.0 dataset is shown in Fig. 4.3. From these two ROC curves, the effect of improvement by ZRANK on ZDOCK is smaller than on MEGADOCK. One of the reasons is that the ZDOCK scoring functions is more accurate than the MEGADOCK scoring functions.

Fig. 4.4 shows a heat map obtained from our PPI prediction method ($t = 3$). We used the threshold value E^* as 7.3, and the cells wherein the corresponding pair was predicted as positive are colored red. The TPs are those on the diagonal cell with red values, FN are green squares on the diagonal line and FPs are high scoring squares off of the diagonal.

Table 4.3: Results of 44 × 44 protein–protein interaction predictions

Decoys recorded per rotation t		1	2	3	5	10	20
Predictions without reranking	Precision	0.563	0.435	0.474	0.429	0.409	0.450
	Recall	0.205	0.227	0.205	0.205	0.205	0.205
	F-measure	0.300	0.299	0.286	0.277	0.273	0.281
Predictions with reranking	Precision	-	0.375	0.447	0.320	0.347	0.318
	Recall	-	0.409	0.386	0.364	0.386	0.318
	F-measure	-	0.391	0.415	0.340	0.366	0.318

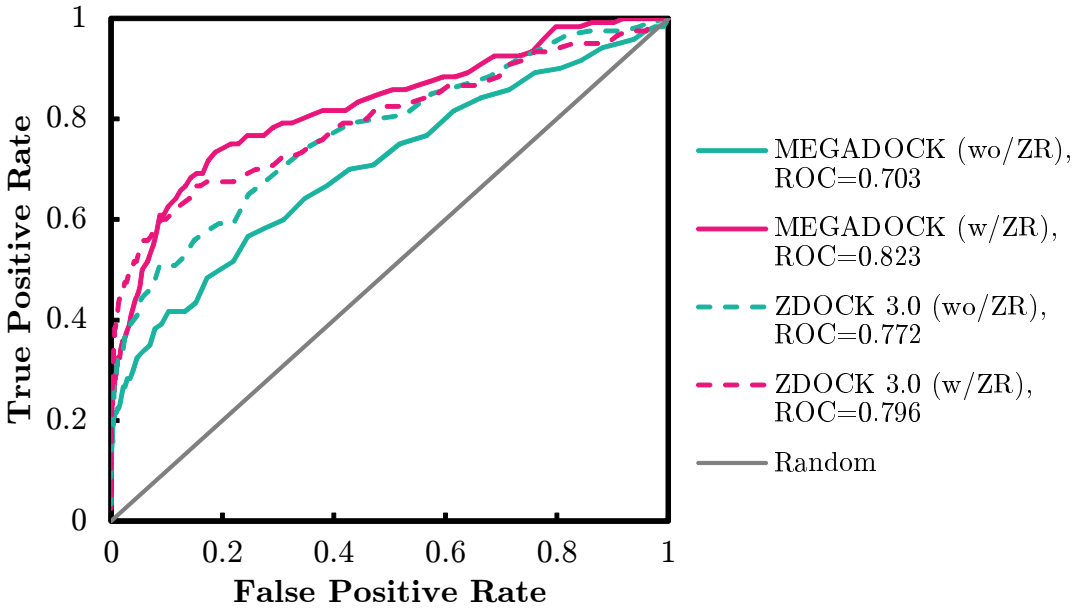


Figure 4.2: Evaluation of the docking post-processing system (large dataset, $t = 3$). The ROC curves for varying the threshold E^* values are shown. The x-axis represents the false-positive fraction (#FP/(#FP+#TN)) and the y-axis represents the true-positive fraction (#TP/(#TP+#FN)). Random predictions are indicated by the diagonal.

Table 4.4: The selected 102 complex structures from the dockground 3.0 benchmark dataset

1A2X	1AGR	1ARO	1AVA	1AVW	1BND	1BRS	1BZQ
1C9P	1CGJ	1CSE	1CXZ	1D4X	1DF9	1DHK	1DKF
1DP5	1EAI	1EFU	1F02	1F5Q	1F7Z	1FFG	1FM9
1H59	1I8L	1IAR	1JTD	1JTP	1K8R	1K93	1KPS
1KTK	1KU6	1L4D	1M27	1MA9	1MBX	1MR1	1MZW
1NMU	1NPE	1NU9	1OIU	1P9M	1PPF	1QAV	1QBK
1R1K	1RZR	1S3S	1SGP	1SHW	1SQ0	1SQ2	1STF
1TA3	1TE1	1TK5	1TX4	1U0S	1U7E	1UEA	1UJW
1UL1	1UUZ	1UZX	1V5I	1W98	1WPX	1WR6	1WRD
1X86	1XDT	1Z3G	1Z92	1ZLH	1ZM2	2A19	2A41
2A42	2A5D	2AUH	2B12	2B3T	2B5I	2BH1	2BKH
2BKK	2C1M	2C5D	2GY7	2HDI	2IW5	2J0M	2JB0
2OMZ	2P8W	2PAV	3BP5	3SIC	3YGS		

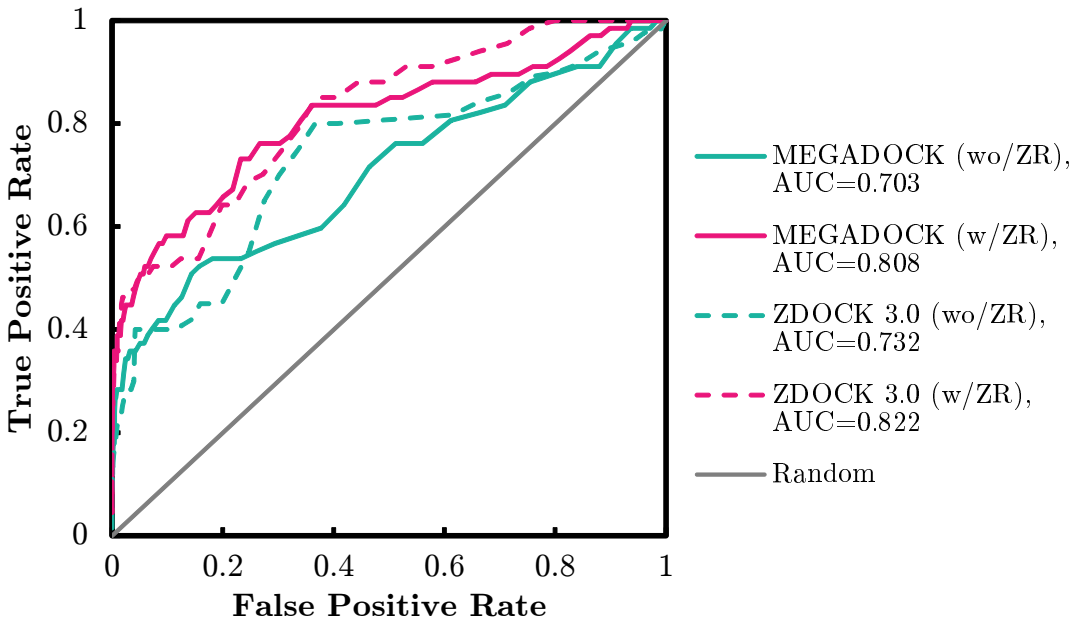


Figure 4.3: Evaluation of the docking post-processing system (dockground 3.0 dataset, $t = 3$). The ROC curves for varying the threshold E^* values are shown. The x-axis represents the false-positive fraction (#FP/(#FP+#TN)) and the y-axis represents the true-positive fraction (#TP/(#TP+#FN)). Random predictions are indicated by the diagonal.

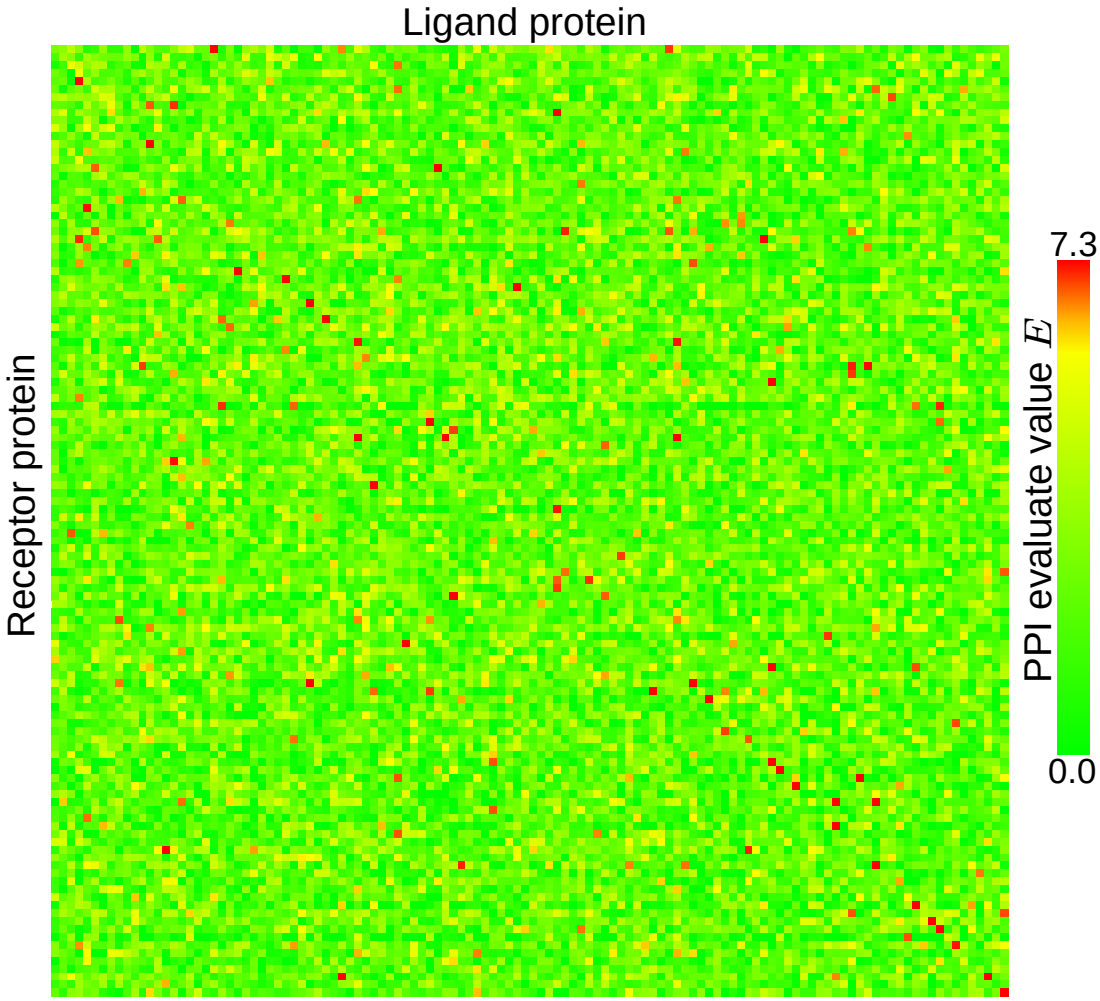


Figure 4.4: 120×120 map of protein–protein interaction prediction results. The red cells are those for which E is more than $E^*(= 7.3)$.

4.3.2 Toward developing a method applicable to unbound data

We also conducted all-to-all docking and PPI predictions to the unbound 120 × 120 large dataset. The performance of the application to this dataset gave an F-measure value of 0.0390 (Precision 0.0471, Recall 0.0333, Accuracy 0.981) with the PPI prediction parameter $E^* = 6.3$. In this result, the F-measure value was much worse when compared to the bound large dataset, whereas it was slightly better than the random prediction's F-measure value of 0.0164. We conducted the same analysis using ZDOCK but also failed to get a better F-measure value (0.0415).

This poor performance on application to unbound data is because of the high dependence of our current method on the docking score function. It assumes that the correct binding structure has significantly high scores when compared to the incorrect docking forms. Nevertheless we should say from the results that such significance of the docking score might be difficult to achieve with the unbound structures, because the unbound structures are not expected to have exact shape complementarity, which is expected in the re-docking of the bound structures.

To improve the PPI prediction of unbound structures, some additional analysis is required such as: (i) including not only the best decoy's score but also use a group of highly ranked decoys to calculate E; and (ii) analyzing the distributions of the high scoring decoys with respect to the interaction residues while improving the docking score function.

Another promising approach to the PPI prediction using the unbound dataset is to use cross docking using the ensemble structures. In unbound pair, shape complementarity based docking scores are not significantly high in high ranked decoys because the receptor and ligand protein structures do not have the exact shape complementarity. By sampling some possible structures of proteins and if successful, make a structure that is closer to the bound form, the PPI prediction process can be improved. There are some successful outcomes that uses ensemble docking and much efforts were put on getting better structure sampling starting from unbound form of the proteins [87, 88]. As such efforts that try to eventually make unbound docking problems to similar problems to bound docking matures, out method can provide the link to the structural docking to predicting possible binding pairs.

It should be noted that the datasets used contain much larger number of 'False' pairs against 'True' pairs (14,280 False pairs and 120 True pairs). It makes difficult to achieve high performance of PPI prediction. As an example of the application on smaller

Table 4.5: Divided dataset located to the Nucleus subcellular location

PDBID	UniprotID	Subcellular location
1FQ1_r	P24941	Cytoplasm, Nucleus, Cytoplasm, Endosome
1GXD_r	P08253	Secreted, Membrane, Nucleus
1H9D_r	Q03347	Nucleus
1I2M_r	P62826	Nucleus, Cytoplasm, Melanosome
1IBR_r	P62825	Nucleus, Cytoplasm, Melanosome
1S1Q_r	Q99816	Cytoplasm, Membrane, Nucleus, Late endosome membrane
1SYX_r	P83876	Nucleus
1ZHI_r	P54784	Nucleus
2OZA_r	Q16539	Cytoplasm, Nucleus
1ATN_l	P00639	Secreted, Nucleus envelope
1FQ1_l	Q16667	Cytoplasm perinuclear region
1H9D_l	Q08024	Nucleus
1I2M_l	P18754	Nucleus, Cytoplasm
1IBR_l	Q14974	Cytoplasm, Nucleus envelope
1S1Q_l	P0CG48	Cytoplasm, Nucleus
1XD3_l	P0CG48	Cytoplasm, Nucleus
1ZHI_l	P21691	Nucleus, Chromosome
2AYO_l	P0CG48	Cytoplasm, Nucleus
2OOB_l	P0CH28	Cytoplasm, Nucleus
2OZA_l	P49137	Cytoplasm, Nucleus

dataset we tried dividing our data according to the subcellular location information obtained from Uniprot database (Table 4.5–4.7). The performance of our method was varied according to the sub-datasets. While we did not see major improvement in the case of nucleus data (Fig. 4.5), higher F-measure value was observed in other cases of mitochondrion (Fig. 4.6) and Golgi apparatus (Fig. 4.7). Although our method aims at primary screening of PPI from large protein structure data, we think that we can improve the performance of our method using additional feature information.

4.4 Summary

In this chapter, we describe here the development of an exhaustive PPI screening system called “MEGADOCK” that conducts docking and post-analysis on protein tertiary structural data. For the detection of the relevant interacting protein pairs, we obtained an F-measure value of 0.231 when our method was applied to a subset of a

Table 4.6: Divided dataset located to the Mitochondrion subcellular location

PDBID	UniprotID	Subcellular location
1E6E_r	P08165	Mitochondrion matrix
1JK9_r	P40202	Cytoplasm, Mitochondrion intermembrane space
2PCC_r	P00431	Mitochondrion matrix
1E6E_l	P00257	Mitochondrion matrix
1JK9_l	P00445	Cytoplasm, Mitochondrion intermembrane space
2C0L_l	O62742	Cytoplasm, Mitochondrion, Peroxisome
2PCC_l	P00044	Mitochondrion intermembrane space

Table 4.7: Divided dataset located to the Golgi apparatus subcellular location

PDBID	UniprotID	Subcellular location
1HE8_r	P01112	Cell membrane, Golgi apparatus, Golgi apparatus membrane
1R8S_r	P84077	Golgi apparatus, Cytoplasm
1WQ1_r	P01112	Cell membrane, Golgi apparatus, Golgi apparatus membrane
2AJF_r	Q9BYF1	Processed angiotensin-converting enzyme, Secreted, Cell membrane
2CFH_r	O43617	Golgi apparatus, Endoplasmic reticulum
2G77_r	Q08484	Golgi apparatus
2OT3_r	Q9UL25	Endoplasmic reticulum membrane, Golgi apparatus membrane, Early endosome membrane, Cytoplasmic vesicle membrane, Cleavage furrow
1J2J_l	Q9UJY5	Golgi apparatus, Endosome membrane
2AJF_l	P59594	Virion membrane, Host endoplasmic reticulum-Golgi intermediate compartment membrane, Host cell membrane
2CFH_l	Q86SZ2	Golgi apparatus, Endoplasmic reticulum
2G77_l	O35963	Golgi apparatus membrane

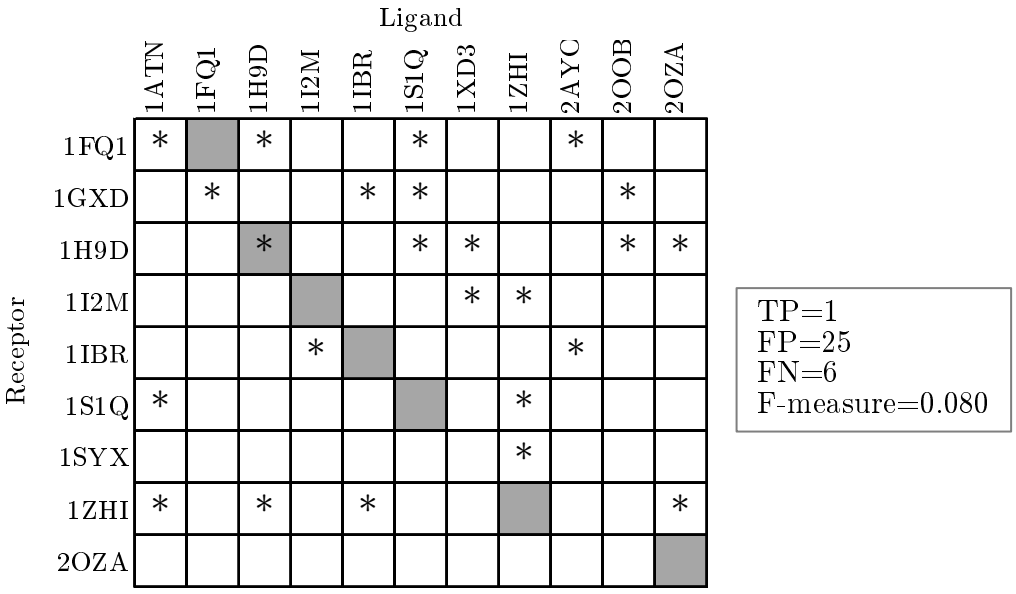


Figure 4.5: Result of the PPI predictions with nucleus sub-dataset. The interactions estimated as positive are marked with asterisks. The gray colored cells correspond to the known interactions.

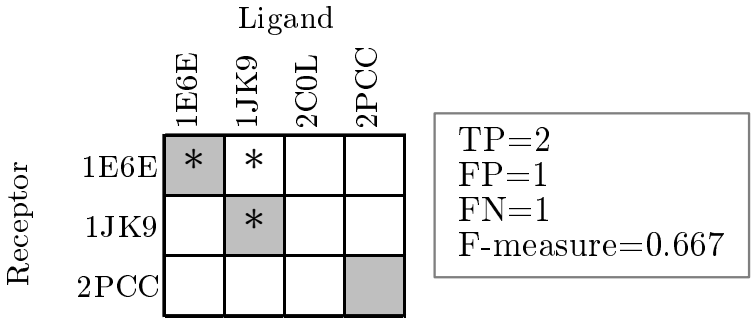


Figure 4.6: Result of the PPI predictions with mitochondrion sub-dataset. The interactions estimated as positive are marked with asterisks. The gray colored cells correspond to the known interactions.

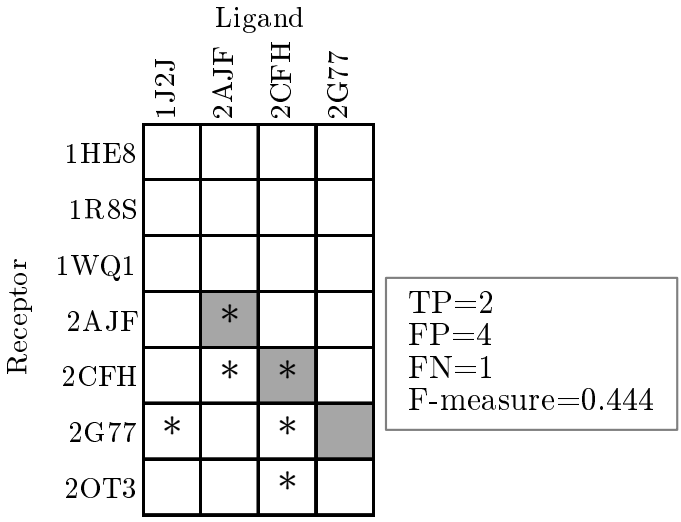


Figure 4.7: Result of the PPI predictions with Golgi apparatus sub-dataset. The interactions estimated as positive are marked with asterisks. The gray colored cells correspond to the known interactions.

general benchmark dataset.

Our future work will include the quantitative representation of the reliability of the prediction for each detected PPI. Moreover, we believe that integrating our prediction approach into conventional bioinformatics methods, such as those based on nucleotide sequencing should be useful.

Chapter 5

Application to Bacterial Chemotaxis Pathway Analysis

5.1 Introduction

Enteric bacteria like *Escherichia coli* control their locomotion by sensing changes of chemicals in the environment to move to more nutrient-rich areas and away from harmful conditions—the phenomenon called chemotaxis. Cell motility and chemotaxis are essential for the pathogenicity of many pathogenic bacteria, which must swim toward host cells to invade them [89]. In addition, bacteria protect themselves from phagocytosis by inhibiting host cell chemotaxis [90].

The bacterial chemotaxis pathway has been studied for several decades and most of the functional relationships among the proteins involved in this signal process have been identified especially those involving the core part of the signaling system [92, 93] (Fig. 5.1, Table 5.1). However there are still uncertainties concerning how flagellar motor proteins are assembled and operate (reviewed in [94]). Recent simulation studies with dynamic models and molecular imaging studies have suggested possible mechanisms for signal amplification and robustly accurate adaptation [95, 96, 97].

In this chapter we applied MEGADOCK to a pathway reconstruction problem of bacterial chemotaxis. Pathway reconstruction is a major goal of large-scale PPI prediction but currently there are only a few assessment studies testing the ability of the method to reconstruct an actual biological pathway. To further demonstrate its potential, we evaluated our PPI prediction system by applying it to the data from a real biological pathway. A small, well-studied pathway of bacterial chemotaxis was chosen as the system to reconstruct.

The MEGADOCK to computational PPI detection is a promising methodology for mediating between structural studies and systems biology by utilizing cumulative protein structure data for pathway analysis.

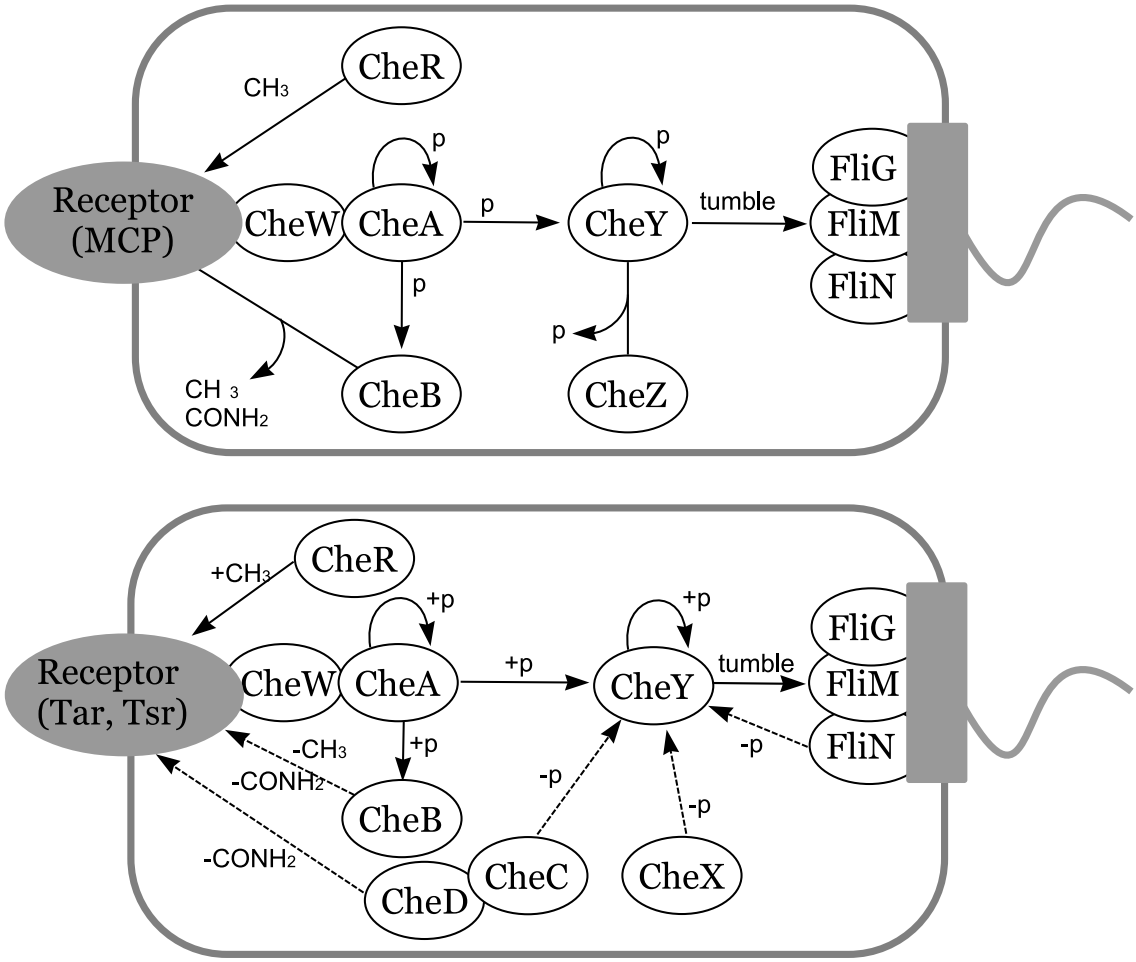


Figure 5.1: Chemotaxis pathway for *E. coli* (above) and *T. maritima* (below). The motion of these bacteria are controlled by the rotation direction of their flagellar motor. The phosphorylation state of CheY is responsible for the rotation direction. When the receptors (Methyl-accepting Chemotaxis Proteins, MCP) sense favorable signals such as those indicating nutrition molecules in the environment, CheA autophosphorylation is inhibited. Then the phosphorylation level of CheY will be reduced because of the repression of phosphotransfer from CheA. That low phosphorylation level of CheY reduces its affinity to the flagellar motor, which causes more frequent counterclockwise rotation and longer periods of smooth swimming of the cell. In addition, the stimulated receptors also undergo a gradual change in the methylation level controlled by CheR and CheB. That causes adaptation to the signal. The MCP family comprises Tar, Tsr, Trg, Tap and Aer, each of which senses distinct signals.

Table 5.1: Proteins that constitute the chemotaxis system.

Protein	Definition
MCP	Chemoreceptors (Methyl-accepting Chemotaxis Proteins, MCP) Tsr (serine receptor) is used in this paper.
CheA	Autophosphorylation capable histidine kinase which transfers the phosphate group to response regulators (CheY, CheB)
CheB	MCP methylesterase, deamidase whose demethylation activity increases in the phosphorylated form
CheC†	CheY phosphatase whose activity increases in the binding form with CheD
CheD†	MCP deamidase
CheR	MCP methyltransferase
CheW	Scaffold protein for MCP and CheA
CheX†	CheY phosphatase
CheY	Stimulate clockwise rotation of the flagellar motor, leading to tumbling behavior of the cell in the phosphorylated state
CheZ‡	CheY phosphatase
FliM	Flagellar motor protein
FliN	Flagellar motor protein
FliG	Flagellar motor protein

Note: † Proteins seen only in *T. maritima*, ‡ *E. coli* or *S. typhimurium*.

5.2 Materials and Methods

5.2.1 Collection of protein structural data

We selected the well characterized species *E. coli*, *S. typhimurium* and *T. maritima* as the targets for this study of bacterial chemotaxis.

The following procedures were performed to retrieve the data: get the PDB ID list that corresponded to the proteins in the bacterial chemotaxis pathway (Table 5.2). Pathway data were obtained from KEGG [98] (KEGG pathway ID: *eco02030, stm020230, tma02030* for each species). PDB IDs were obtained through LinkDB [99].

The collected protein structure files were prescreened according to the following criteria, as in the recently published protein–protein docking benchmark version 3.0 [81] (Table 5.2):

1. Experimental method: X-ray diffraction, resolution better than 3.25 Å,
2. Polypeptides consisting of more than 30 residues.

In principle, mutant data and synthetic objects were excluded with the one exception of CheZ, for which only mutant data was available (Table 5.2). Structure data for only the ligand binding domain of the membrane proteins, which is located in the periplasm, were also excluded. The protein structure data used in this study is shown in Table 5.2.

5.2.2 Known PPI information

Relevant PPIs are defined based on published data [100, 101, 102, 103, 104]. The interactions of short form CheA [105] were not considered because its structure was unavailable. In addition, interactions based on genetic observations alone were excluded. FliG, FliM and FliN were considered as binding to the protein species because they make solid flagellar motor machinery. For *in vitro* studies, large numbers of interactions are listed in public databases such as the STRING database [106]. However, these data sets were not included in this study because the physical interactions for those PPIs are not characterized.

5.2.3 PPI prediction

We conducted all-to-all PPI prediction by using MEGADOCK with the PDB structures in Table 5.2. Each PDB file was divided into data for each polypeptide chain, which for most cases in this dataset, corresponded to a single protein species.

Table 5.2: Chemotaxis dataset derived from PDB.

PDB ID	Chain	Organism	Molecule	Domain
1FFG	B,D	*E. coli*	CheA	P2
1FFS	B,D	*E. coli*	CheA	P2
1FFW	B,D	*E. coli*	CheA	P2
1A0O	A,C,E,G	*E. coli*	CheY	
1BDJ	A	*E. coli*	CheY	
1CHN	A	*E. coli*	CheY	
1F4V	A,B,C	*E. coli*	CheY	
1FFG	A,C	*E. coli*	CheY	
1FFS	A,C	*E. coli*	CheY	
1FFW	A,C	*E. coli*	CheY	
1FQW	A,B	*E. coli*	CheY	
1HEY	A	*E. coli*	CheY	
1JBE	A	*E. coli*	CheY	
1KMI	Y	*E. coli*	CheY	
1ZDM	A,B	*E. coli*	CheY	
2B1J	A,B	*E. coli*	CheY	
3CHY	A	*E. coli*	CheY	
1KMI†	Z	*E. coli*	CheZ	
1QU7	B	*E. coli*	MCP(Tsr)	Cytoplasmic domain
1I5N	A,B,C,D	*S. typhimurium*	CheA	P1
1A2O	A,B	*S. typhimurium*	CheB	
1CHD	A	*S. typhimurium*	CheB	C-terminal catalytic domain
1AF7	A	*S. typhimurium*	CheR	
1BC5	A	*S. typhimurium*	CheR	
2CHE	A	*S. typhimurium*	CheY	
2CHF	A	*S. typhimurium*	CheY	
2FKA	A	*S. typhimurium*	CheY	
2FLK	A	*S. typhimurium*	CheY	
2FLW	A	*S. typhimurium*	CheY	
2FMF	A	*S. typhimurium*	CheY	
2FMH	A	*S. typhimurium*	CheY	
2FMI	A	*S. typhimurium*	CheY	
2FMK	A	*S. typhimurium*	CheY	
2PL9	A,B,C	*S. typhimurium*	CheY	
2PMC	A,B,C,D	*S. typhimurium*	CheY	
1TQG	A	*T. maritima*	CheA	P1
1U0S	A	*T. maritima*	CheA	P2
2CH4	A,B	*T. maritima*	CheA	P4, P5 (Residues 355–671)
1XKR	A	*T. maritima*	CheC	
2F9Z	A,B	*T. maritima*	CheC	
2F9Z	C,D	*T. maritima*	CheD	
2CH4	W,Y	*T. maritima*	CheW	
1SQU	A,B	*T. maritima*	CheX	
1XKO	A,B	*T. maritima*	CheX	
1TMY	A	*T. maritima*	CheY	
1U0S	Y	*T. maritima*	CheY	
2TMY	A	*T. maritima*	CheY	
3TMY	A,B	*T. maritima*	CheY	
4TMY	A,B	*T. maritima*	CheY	
1LKV	X	*T. maritima*	FliG	C-terminal domain (Residues 104–335)
1QC7	A,B	*T. maritima*	FliG	C-terminal domain
2HP7	A	*T. maritima*	FliM	CheC-like domain
1O6A	A,B	*T. maritima*	FliN	C-terminal domain (Residues 59–154)
1YAB	A,B	*T. maritima*	FliN	Residues 68–154
2CH7	A,B	*T. maritima*	MCP	Cytoplasmic domain

Note: CheA comprises five domains: P1 (Histidine phosphotransfer domain), P2 (Response regulator binding domain), P3 (Histidine kinase-like homodimeric domain), P4 (Histidine kinase-like ATPases) and P5 (Receptor coupling domain). † Includes a mutation in residue 134 (Glu → Lys).

For the chemotaxis pathway we often obtained more than one structural data element for a protein species. In such a case we calculated affinity scores using all the data we had. When we found at least one positive evaluation between relevant protein pairs, we evaluated the pair as interacting.

5.3 Results

5.3.1 PPI detection performance

Fig. 5.2 shows the ROC curve for varying the threshold E^* values and Fig. 5.3 and Table 5.3 show PPI detection for the chemotaxis dataset with $E^* = 7.3$. This parameter is the same as those that produced the best F-measure value for the benchmark data. Gray-colored cells indicate the protein pairs known to interact with each other. We obtained an E-measure of 0.464 for this system, which is similar to that found in the previous study using ZDOCK 3.0 by Matsuzaki *et al.* [23].

5.3.2 Predicted interactions

The "false-positive" interactions include some interactions that are not possible, considering the localization of the proteins, such as the interaction between flagellar motor proteins and receptor proteins. Precluding these apparently false detections, we can restrict the "false-positive" interactions to those that are worth further analysis. One of the suggestions of currently unknown PPI, CheY–CheD, is shown in Fig. 5.4.

5.4 Discussion

Although a primary purpose of this study was the assessment of the computational PPI screening performance on the real biological pathway, some "false-positive" interactions detected in the chemotaxis pathway seemed to be worth further analysis.

One such interaction was CheY–CheD. CheC is known as a phosphatase of CheY. CheC activity is known to be enhanced by the existence of CheD [104, 107]. Although interactions between CheY–CheC, CheC–CheD are already known [107, 108], no evidence for direct binding of CheY–CheD has been found until now. Still, seeking the possibility of this direct interaction might be interesting.

In the complex form, CheC is in active state and CheD is inactivated. Chao *et al.* have suggested a mechanism by which MCP and phosphorylated CheY (CheYp)

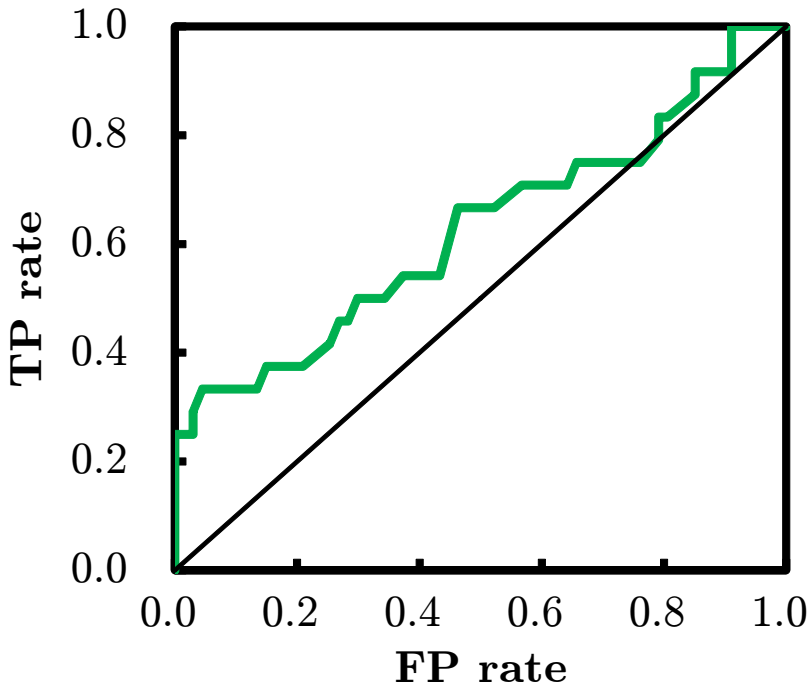


Figure 5.2: Evaluation of the prediction system in chemotaxis dataset. The ROC curves for varying the threshold E^* values are shown. x-axis is for the false positive rate ($\frac{\mathrm{TP}}{\mathrm{TP+FN}}$) and y-axis is for the true positive rate ($\frac{\mathrm{FP}}{\mathrm{FP+TN}}$). Random prediction is indicated by the diagonal.

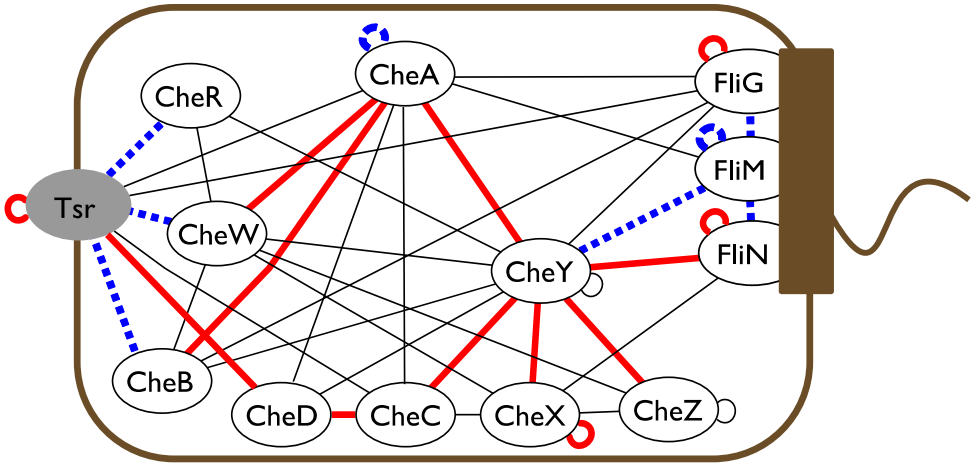


Figure 5.3: Results of the PPI predictions from the proposed system with $E^* = 7.3$. The red bold lines (true positives), blue dashed lines (false negatives) and thin lines (false positives) representing the predicted or known PPIs show the relevance of the predictions.

Table 5.3: Results of the PPI predictions using the proposed system with $E^* = 7.3$. The interactions estimated as positive are marked with asterisks. The gray colored cells correspond to the known interactions.

	A	B	C	D	R	W	X	Y	Z	FliG	FliM	FliN	Tsr
CheA		-	-	-	-	-	-	-	-	-	-	-	-
CheB	*		-	-	-	-	-	-	-	-	-	-	-
CheC	*			-	-	-	-	-	-	-	-	-	-
CheD	*		*		-	-	-	-	-	-	-	-	-
CheR						-	-	-	-	-	-	-	-
CheW	*	*			*		-	-	-	-	-	-	-
CheX			*			*	*	-	-	-	-	-	-
CheY	*	*	*	*	*	*	*	*	-	-	-	-	-
CheZ						*	*	*	*	-	-	-	-
FliG	*	*						*		*	-	-	-
FliM	*											-	-
FliN							*	*				*	-
Tsr	*		*	*						*			*

competitively control the free CheD availability through CheC-CheD complex [109]. When CheY phosphorylation level is high, it increases CheC-CheD complex and thus reduces CheD interaction with MCP molecules. Assuming that it actually occurs in the living cell, the mechanism may include a direct interaction between CheY and CheD.

The fact that CheC activity increases when CheD is present can also be explained by a model in which CheD first binds to CheY and then recruits CheC to lead to the correct binding pose with CheY. Fig. 5.4 shows a hypothetical CheY–CheD complex and CheC docking. In this preliminary result we couldn't determine whether the CheC and the CheY phosphorylation site were in close proximity. To seek the possibility of unknown PPIs, we need to exhaustively search the space of high scoring docking decoys of the protein pairs under consideration.

To further investigate the possibility of the detected but currently unknown interactions, validation by experiment is crucially important. It would be interesting to restrict the targets by further expensive calculations involving surface electric charge or interaction surface analysis and thereby obtain some strongly possible interactions for experimental validation. It would also be interesting to obtain the crystal structure of the CheY–CheC–CheD complex and see if we find similar structure in the hypothetical docking decoys.

5.5 Summary

In this chapter, we applied an all-to-all PPI prediction system, MEGADOCK, to bacterial chemotaxis pathway reconstruction problem. The results showed better performance when compared to those from previous research and random predictions. The proposed PPI detection method will enable the large scale PPI screening that is useful to restrict the search space before utilizing expensive PPI analysis methods.

Among the predicted PPIs for the chemotaxis proteins, we discussed an example of an unknown interaction (CheY–CheD) that is worthy of further analysis. Validation after the PPI screening is a problem to be explored in future research, as well as seeking to improve prediction performance.

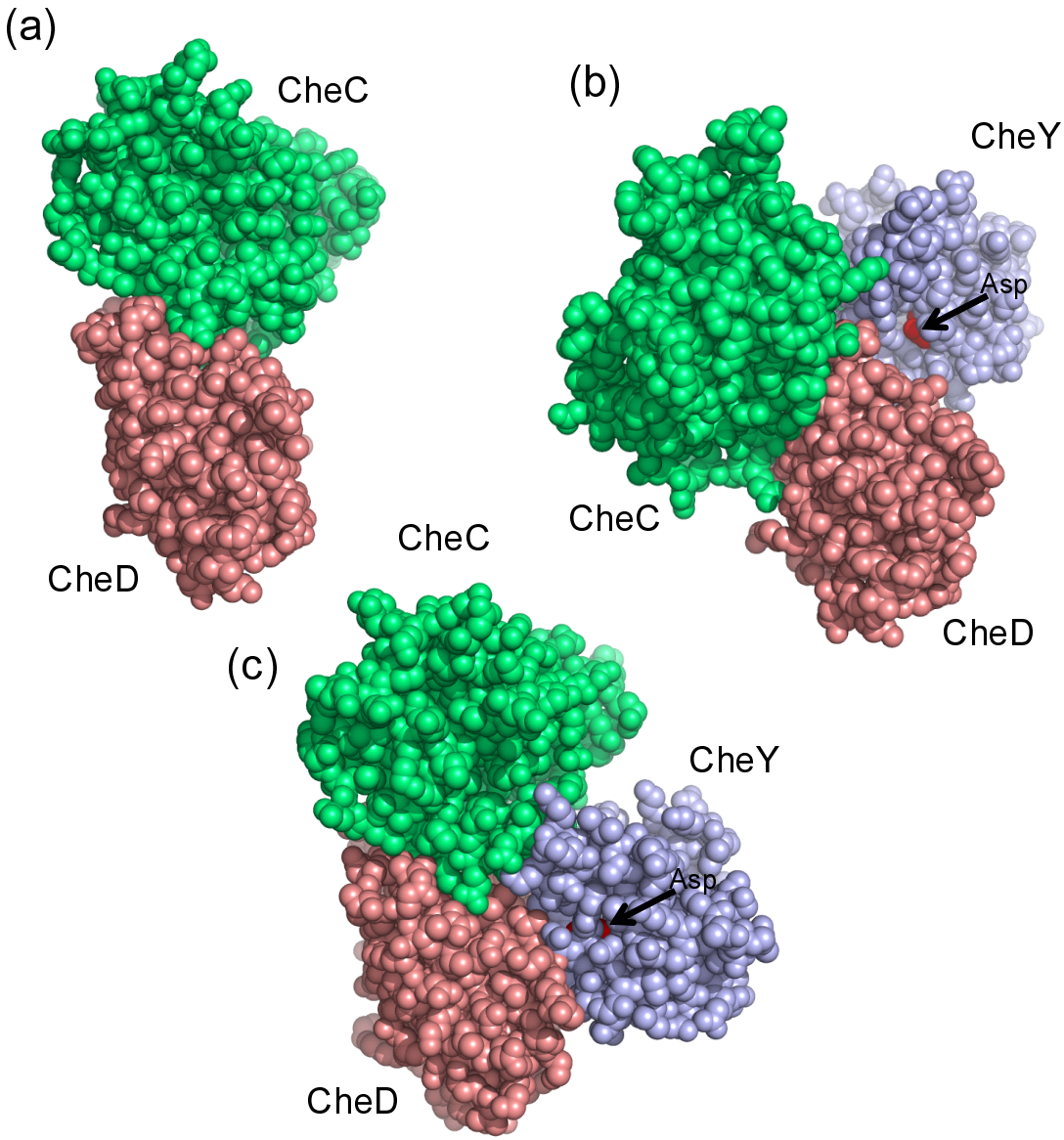


Figure 5.4: (a) Known structure of the CheC–CheD complex (PDB ID: 2F9Z, chains A, C). (b) Docking of CheY (PDB ID: 1A0O, chain C)–CheD (PDB ID: 2F9Z, chain C) hypothetical complex and CheC (PDB ID: 1XKR, chain A). The phosphorylation site of CheY is colored red. The hypothetical complex was constructed from the representative data with the highest E value among all combinations of CheY–CheD docking and clustering results. The docking prediction with the highest E value among all the combinations of the hypothetical complex and CheC structure data is shown. (c) Docking of a known structure of the CheC–CheD complex (PDB ID: 2F9Z, chains A, C) and CheY (PDB ID: 1F4V, chain C). The phosphorylation site of CheY is colored red. This hypothetical complex is also constructed using the representative data among all combinations of the CheC–CheD complexes and CheY structures.

Chapter 6

Application to Human Apoptosis Pathway Analysis

6.1 Introduction

Apoptosis is the process of programmed cell death that may occur in multicellular organisms; that is, cells committing suicide by activating an intracellular death program; getting engulfed and digested by macrophages without harming their neighbors. Apoptosis helps in regulation of cell number and size, such as the differentiation of fingers in a developing embryo by the programmed death of cells between them; or removal of infected or damaged cells. Apoptotic processes are regulated by extrinsic and intrinsic pathways [110] (Fig. 6.1).

In this chapter, we applied MEGADOCK to a pathway reconstruction problem of human apoptosis as larger problem than bacterial chemotaxis pathway (Chapter 5). In addition, we aim to predict the structures of the complexes formed by interacting protein pairs in the apoptosis pathway by using the MEGADOCK and to figure out the implications of the newly obtained interactions.

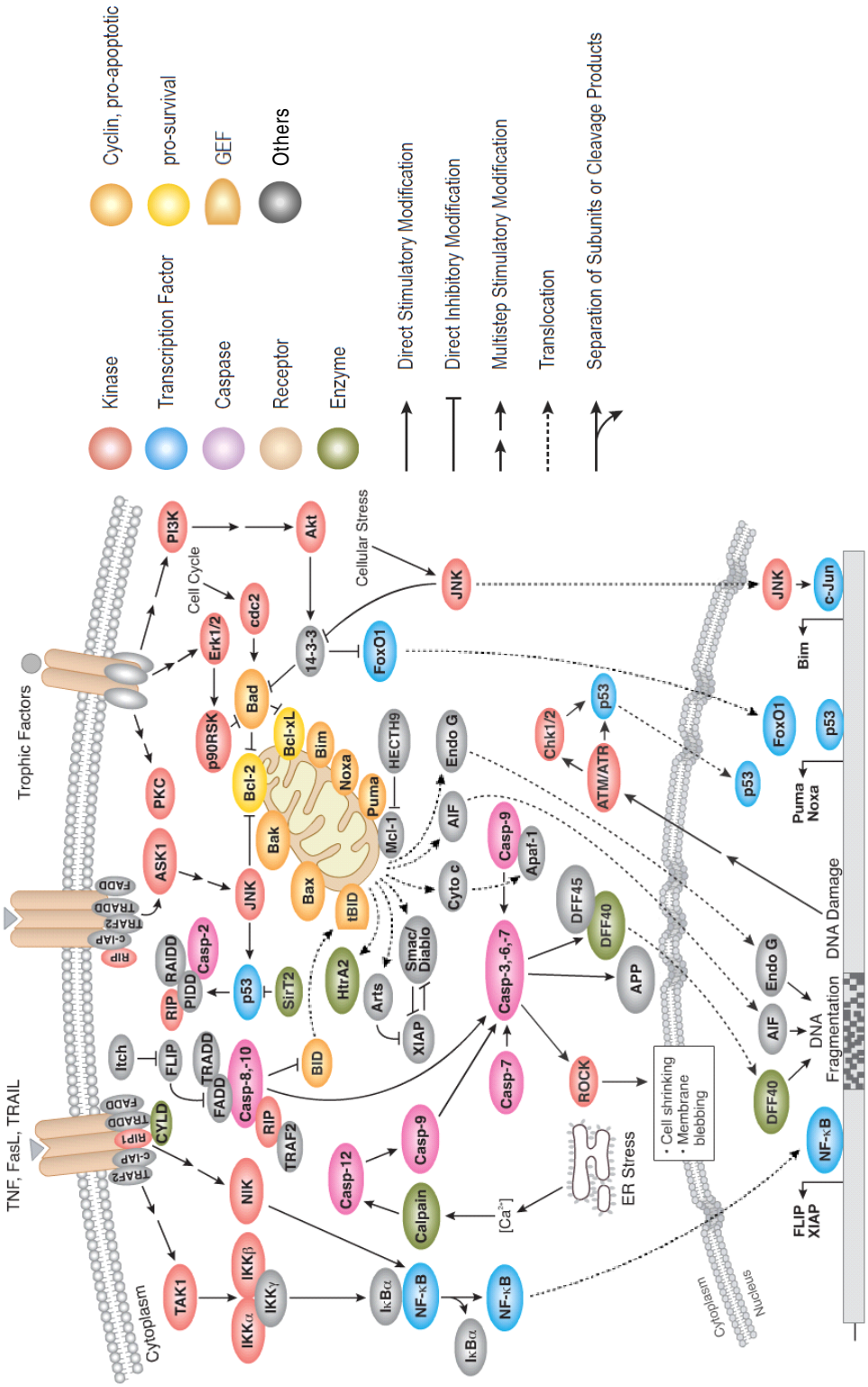


Figure 6.1: The overview of apoptosis pathway. Illustration reproduced courtesy of Cell Signaling Technology, Inc. (`www.cellsignal.com`).

6.1.1 Summary of the human apoptosis

Apoptosis is involved in the pathogenesis of many diseases. If the cells fail to undergo apoptosis, an uncontrolled proliferation rate can cause diseases such as cancer, autoimmune diseases and viral infections. In contrast, accelerated rates of apoptosis may cause diseases that are related to cell loss, such as AIDS (acquired immunodeficiency syndrome), neurodegenerative diseases, ischemic injury and toxin-induced liver disease. It is crucial to know the details of the apoptosis signaling pathway, especially structural details of protein–protein interactions, in order to identify targets and design drugs [111].

Central players in signal transduction in both the extrinsic and intrinsic pathways are the caspases (cysteine-dependent aspartate-directed proteases). Caspases are members of the protease family, which are synthesized as inactive precursors or procaspases. Procaspases are activated by proteolytic cleavage by other members of their family in response to inducing signals. Once they are activated and become caspases, they can activate other procaspases by cleaving them. In this manner, initiator caspases, such as caspase-8, -9 become activated and cleave the inactive effector caspases, such as caspase-3, -6 and -7.

The extrinsic pathway is mediated by the death receptors such as TNF-R, Fas and TRAIL-R. Initiator caspases are activated by the death receptors with death domain-containing adaptor molecules such as FADD. The DISC complex composed TRAIL-R, FADD, caspase-8 and FLIP activates a signaling cascade [112].

On the other hand, the intrinsic pathway is initiated by stress signals, such as UV-irradiation, γ-irradiation, DNA damage, and genotoxic stress, causing cytochrome C (CytC) release from the mitochondria. Released CytC binds to apoptotic protease activating factor 1 (Apaf-1) to form the apoptosome and activate initiator caspase, which activates the executioner caspases [113].

Thus, the different players work in the extrinsic and intrinsic pathways of apoptosis. However, the apoptosis pathways observed to cross-talk via caspase-8, which leads to the initiation of the intrinsic pathway by activating the BID protein and the release of CytC from the mitochondria, in addition to its role of activating caspase-3 and triggering apoptosis in the extrinsic pathway (Fig. 6.1).

Table 6.1: PDB IDs of human apoptosis pathway protein from *hsa04210* KEGG pathway (124).

1A0N	1A1W	1A8M	1AIE	1AUI	1CY5	1CZZ	1D00	1D4V	1DG6
1DU3	1E8Y	1EGJ	1EXT	1F16	1F1J	1F3V	1G73	1H9O	1HE7
1I4O	1I51	1IBX	1ICH	1IKN	1ITB	1IYR	1J3S	1JLI	1JXQ
1KFU	1M6I	1MF8	1MRV	1NFI	1NW9	1O6K	1O6L	1OLG	1P6S
1PBW	1QTN	1RHQ	1SHC	1SVC	1UNQ	1WWW	1XQH	1YC5	1Z6T
1ZCM	2B3G	2B48	2BEC	2BID	2DBF	2DKO	2E30	2ECG	2ENQ
2FOO	2FUN	2G5B	2GF5	2GS0	2IFG	2ILA	2IUG	2IZX	2J32
2JS7	2JVX	2K8F	2KBW	2KNA	2KT1	2NQA	2NRU	2NVH	2POI
2QL9	2R28	2UVL	2V1Y	2VUK	2W3L	2WDP	2X18	2XA0	2XS6
2YGS	3AGM	3BRT	3BRV	3CL3	3CM7	3CQW	3D06	3D9T	3DAB
3EB5	3EB6	3EWT	3EZQ	3FDL	3FX0	3H11	3HHM	3I5R	3IZA
3KNV	3LL8	3LW1	3M0A	3M0D	3M1D	3MOP	3MTT	3MUP	3O4O
3O96	3PK1	3YGS	4TSV						

6.2 Materials and Methods

6.2.1 Dataset

In this study, we focused on the human apoptosis signaling pathway previously analyzed by PRISM [14] because our prediction results can thus be compared directly to the results of the previous study. PRISM and MEGADOCK are based on three-dimensional protein structures and therefore can only be applied to proteins whose tertiary structures are available. Therefore, we searched among proteins involved in the human apoptosis pathway that were present in the Protein Data Bank (PDB) (accessed on July 28, 2012). We selected several proteins that had the highest resolution for the structural group that had high sequence similarity (> 0.9) with the other proteins in the dataset [114]. After filtering according to resolution and sequence similarity, we obtained 158 PDB structures that corresponded to 57 proteins in the human apoptosis pathway described in KEGG (KEGG pathway ID: *hsa04210*) [98]. The PDB IDs in this structure dataset were the same as those used by Ozbabacan *et al.* [114]. Table 6.1 shows the list of PDB IDs of human apoptosis pathway proteins from *hsa04210* KEGG pathway and Table 6.2 shows the list of protein names with PDB chains of this dataset.

Table 6.2: PDB chains of human apoptosis pathway protein from *hsa04210* KEGG pathway (158 chains). The first 4 characters before '_' represent PDB ID and the last 1 character after '_' represents chain name.

Name	PDB ID_chain
AIF	1M6I_A
AKT1	1UNQ_A, 3CQW_A, 3O96_A
AKT2	1MRV_A, 1O6K_A, 1O6L_A, 1P6S_A
AKT3	2X18_A
APAF1	1CY5_A, 1Z6T_A, 2YGS_A, 3IZA_A, 3YGS_C
Bax	1F16_A, 2G5B_I, 2XA0_C, 3PK1_B
BCL-2	2W3L_A, 2XA0_A
BCL-XL	2B48_A, 3FDL_A
BID	2BID_A, 2KBW_B
Calpain1	1ZCM_A
Calpain2	1KFU_L, 2NQA_A
CASP3	1RHQ_A, 1RHQ_B, 2DKO_A, 2DKO_B, 2J32_A
CASP6	2WDP_A
CASP7	1F1J_A, 1I4O_A, 1I51_A, 1I51_B, 2QL9_A, 2QL9_B
CASP8	1QTN_A, 1QTN_B, 2FUN_B, 3H11_B
CASP9	1JXQ_A, 1NW9_B, 3D9T_C, 3YGS_P
Cn(CHP)	2E30_A
Cn(CHP2)	2BEC_A
Cn(PPP3CA)	1AUI_A, 1MF8_A, 2R28_C, 3LL8_A
Cn(PPP3R1)	1AUI_B, 1MF8_B, 3LL8_B
CytC	1J3S_A
DFF40	1IBX_A
DFF45	1IBX_B, 1IYR_A
FADD	1A1W_A, 2GF5_A, 3EZQ_B
Fas	3EWT_E, 3EZQ_A
FLIP	3H11_A
IAP(BIRC2)	3D9T_A, 3M1D_A, 3MUP_A
IAP(BIRC3)	2UVL_A, 3EB5_A, 3EB6_A, 3M0A_D, 3M0D_D
IAP(BIRC4)	1G73_C, 1I4O_C, 1I51_E, 1NW9_A, 2ECG_A, 2KNA_A, 2POI_A, 3CM7_C
IκBα	1IKN_D, 1NFI_E
IKK	2JVX_A, 3BRT_B, 3BRV_B, 3CL3_D, 3FX0_A
IL-1(A)	2ILA_A
IL-1(B)	1ITB_A, 2NVH_A, 3O4O_A
IL-1R(1)	1ITB_B
IL-1R(RAP)	3O4O_B
IL-3	1JLI_A
IL-3R	1EGJ_A
IRAK2	3MOP_K
IRAK4	2NRU_A, 3MOP_G

Table 6.2 (continue)

Name	PDB ID_chain
MyD88	2JS7_A, 3MOP_A
NF-κB(NFKB1)	1IKN_C, 1NFI_B, 1SVC_P, 2DBF_A
NF-κB(RELA)	1IKN_A, 1NFI_A
NGF	1WWW_V, 2IFG_E
PI3K(PIK3CA)	2ENQ_A, 2V1Y_A, 3HHM_A
PI3K(PIK3CG)	1E8Y_A
PI3K(PIK3R1)	1A0N_A, 1H9O_A, 1PBW_A, 2IUG_A, 2V1Y_B, 3HHM_B, 3I5R_A
PI3K(PIK3R2)	2KT1_A, 2XS6_A, 3MTT_A
PRKACA	3AGM_A
PRKAR2A	2IZX_A
TNFα	1A8M_A, 4TSV_A
TNF-R1	1EXT_A, 1ICH_A
TP53	1AIE_A, 1OLG_A, 1XQH_B, 1YC5_B, 2B3G_B, 2FOO_B, 2GS0_B, 2K8F_B, 2VUK_A, 3D06_A, 3DAB_B, 3LW1_P
TRADD	1F3V_A
TRAF2	1CZZ_A, 1D00_A, 1F3V_B, 3KNV_A, 3M0A_A, 3M0D_A
TRAIL	1D4V_B, 1DG6_A, 1DU3_D
TRAIL-R	1D4V_A, 1DU3_A
TrkA	1HE7_A, 1SHC_B, 1WWW_X, 2IFG_A

Note: The abbreviations used are: AIF, apoptosis-inducing factor, mitochondrion-associated, 1 (AIFM1); AKT1, RACalpha serine/threonine-protein kinase; AKT2, RAC-beta serine/threonine-protein kinase; AKT3, RAC-gamma serine/threonine-protein kinase; APAF1, apoptotic peptidase activating factor 1; BCL-2, B-cell lymphoma 2; BCL-XL, BCL extra-large; BID, BH3 interacting domain death agonist; Bax, BCL-2-associated X protein; CASP3/6/7/8/9, caspase-3/6/7/8/9; Cn(CHP), calcineurin B homologous protein 1; Cn(CHP2), calcineurin B homologous protein 2; Cn(PPP3CA), protein phosphatase 3 catalytic subunit alpha isoform; Cn(PPP3R1), protein phosphatase 3 regulatory subunit 1; CytC, cytochrome C; DFF40, DNA fragmentation factor, 40kDa, beta polypeptide; DFF45, DNA fragmentation factor, 45kDa, alpha polypeptide; FADD, Fas-associated via death domain; FLIP, FLICE/CASP8 inhibitory protein (CASP8 and FADD-like apoptosis regulator, CFLAR); Fas, tumor necrosis factor receptor (TNF) superfamily member 6; IAP, inhibitor of apoptosis; BIRC2/3/4, baculoviral IAP repeat-containing protein 2/3/4; IκBα, nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor alpha; IKK, inhibitor of nuclear factor kappa-B kinase; IL-1(A), interleukin-1 alpha; IL-1(B), interleukin-1 beta; IL-1R(1), type 1 interleukin-1 receptor; IL- 1R(RAP), interleukin-1 receptor accessory protein; IL-3, interleukin-3; IL-3R, interleukin-3 receptor; IRAK2/4, interleukin-1 receptor-associated kinase 2/4; MyD88, myeloid differentiation primary response protein MyD88; NF-κB(NFKB1), nuclear factor of kappa light polypeptide gene enhancer in B-cells; NF-κB(RELA), nuclear factor of kappa light polypeptide gene enhancer in B-cells 3; NGF, nerve growth factor (beta polypeptide); PI3K, phosphatidylinositide 3-kinase; PIK3CA, PI3K subunit alpha; PIK3CG, PI3K subunit gamma; PIK3R1, PI3K regulatory subunit alpha; PIK3R2, PI3K regulatory subunit beta; PRKACA, cyclic adenosine monophosphate (cAMP)-dependent protein kinase catalytic subunit alpha; PRKAR2A, cAMP-dependent protein kinase type II-alpha regulatory subunit; TNFα, tumor necrosis factor; TNF-R1, TNF receptor superfamily member 1A; TP53, cellular tumor antigen p53; TRADD, TNF receptor type 1-associated death domain protein; TRAF2, TNF receptor-associated factor 2; TRAIL, TNF receptor superfamily member 10; TRAIL-R, TNF receptor superfamily member 10B; TrkA, neurotrophic tyrosine kinase receptor type 1.

6.2.2 Known PPI information

Known PPIs were collected from the STRING database [106]. We used only experimental data in the literature obtained from STRING with a confidence score > 0.5. The number of known PPIs was 137. Because the database does not contain existing self-interactions, we did not predict self-interactions. Thus, the number of target pairs was ${}_{57}\mathrm{C}_2 = 1,596$.

In addition, we used another database provided by Dr. Vachiranee Limviphuvadh (Agency for Science, Technology and Research; A*STAR) for the possibility of self-interactions and the interactions which is not contained in STRING. The database is called LIM DB in this thesis. LIM DB is integrated several PPI databases based on literature information; BIND [115], BioGRID [116], DIP [117], HPRD [118], IntAct [119], MINT [120], MPact [121] and MPPI [122].

When we evaluate by using LIM DB, a protein pair contained in one of STRING DB and LIM DB as a positive sample and both are not contained as a negative sample. Thus the number of target protein pairs is ${}_{57}\mathrm{C}_2+57 = 1{,}653$ and the number of positive samples is 187.

6.2.3 PPI predictions

We conducted all-to-all PPI prediction by using MEGADOCK with the PDB structures in Table 6.2. For the apoptosis pathway we often obtained more than one structural data element for a protein species. In such a case we calculated affinity scores using all the data we had. When we found at least one positive evaluation between relevant protein pairs, we evaluated the pair as interacting same as the bacterial chemotaxis case (Chapter 5).

6.2.4 Evaluation of prediction performance

Here, we have defined #TP, #FP, #FN, #TN, precision, recall, and the F-measure, which we used to evaluate the prediction results: #TP is the number of predicted PPIs that were also found in database (true-positive), #FP is the number of predicted PPIs that were not in database (false-positive), #FN is the number of PPIs not predicted by the system even though the pair was found to interact in database (false-negative), and #TN is the number of negative predictions that were also not found in database

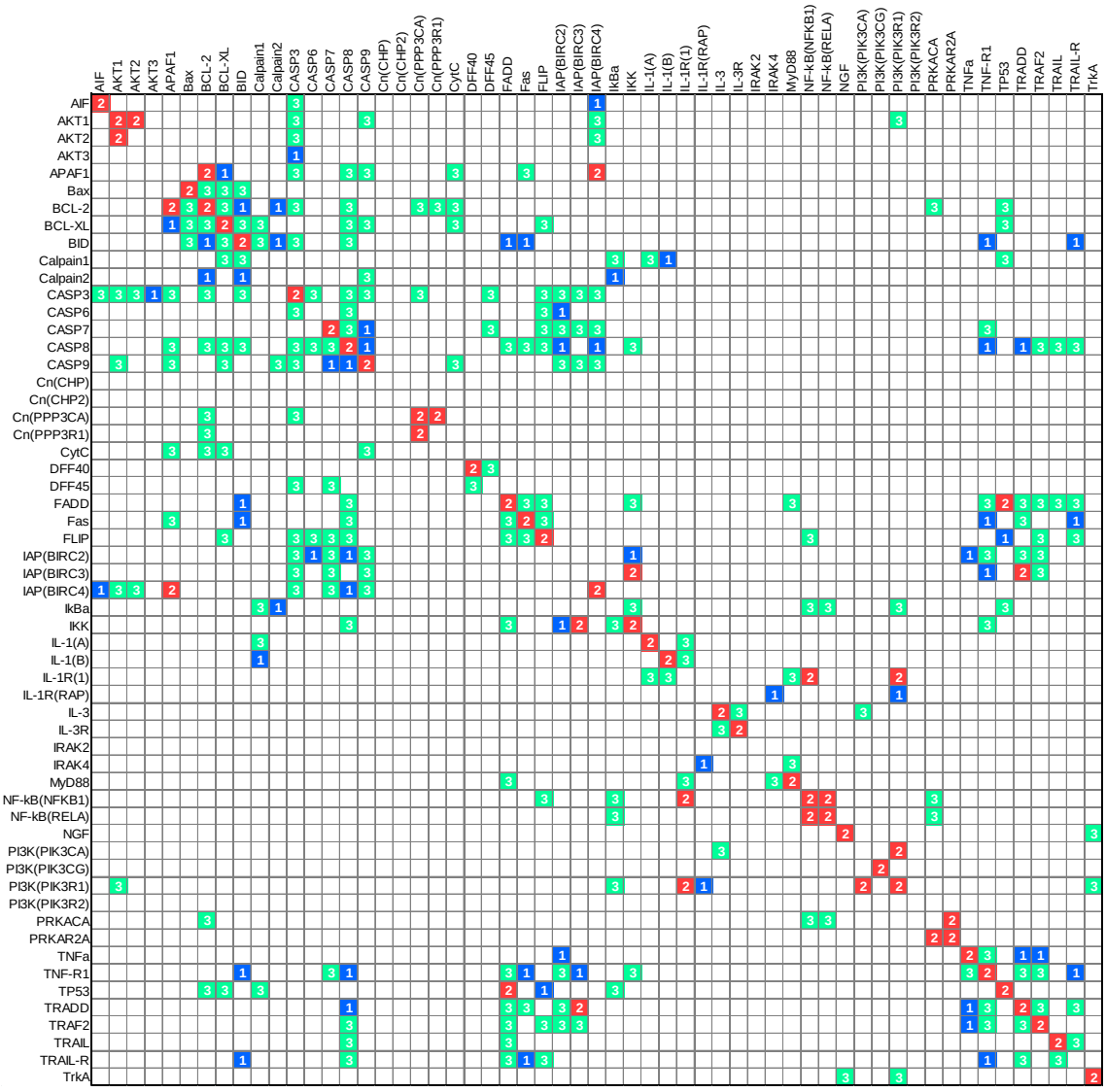


Figure 6.2: The PPIs from STRING DB and LIM DB. Colored cells show interacted protein pairs. '1' (blue) cells are from STRING DB, '2' (red) cells are from LIM DB and '3' (green) cells are both from STRING DB and LIM DB.

(true-negative). Precision, recall, and the F-measure are represented as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F-measure} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

where the F-measure is the harmonic mean of precision and recall. To identify new PPIs in biological experiments after *in silico* screening, precision is more important than recall to reduce the cost of validation.

6.3 Results and Discussion

6.3.1 PPI detection performance

Fig. 6.3 and Fig. 6.4 show the all prediction results of the human apoptosis pathway by MEGADOCK. The accuracies of the results are shown in Table 6.3.

The prediction accuracy of MEGADOCK validated by STRING DB is F-measure = 0.220. This result is a less inferior for PRISM which is a template-base PPI prediction tool developed by Tuncbag, *et al.* [14] and obtained F-measure = 0.296. However, MEGADOCK obtained F-measure value of 0.277 when the results were validated by STRING DB and LIM DB. Although the numbers of target pairs are different, the prediction accuracy is close to the PRISM results.

In addition, the "MEGADOCK$_{\text{TP=56}}$" column in Table 6.3 shows the MEGADOCK results fixed the number of TPs with 56 validated by STRING DB. At this time, the number of FPs of MEGADOCK is 1.8 times larger than PRISM. Practical use of the template structure information by the complex co-crystal structures which PRISM uses is considered to have contributed to reduction of the number of FPs. Fig. 6.5 shows the ROC curve for varying the threshold E^* values.

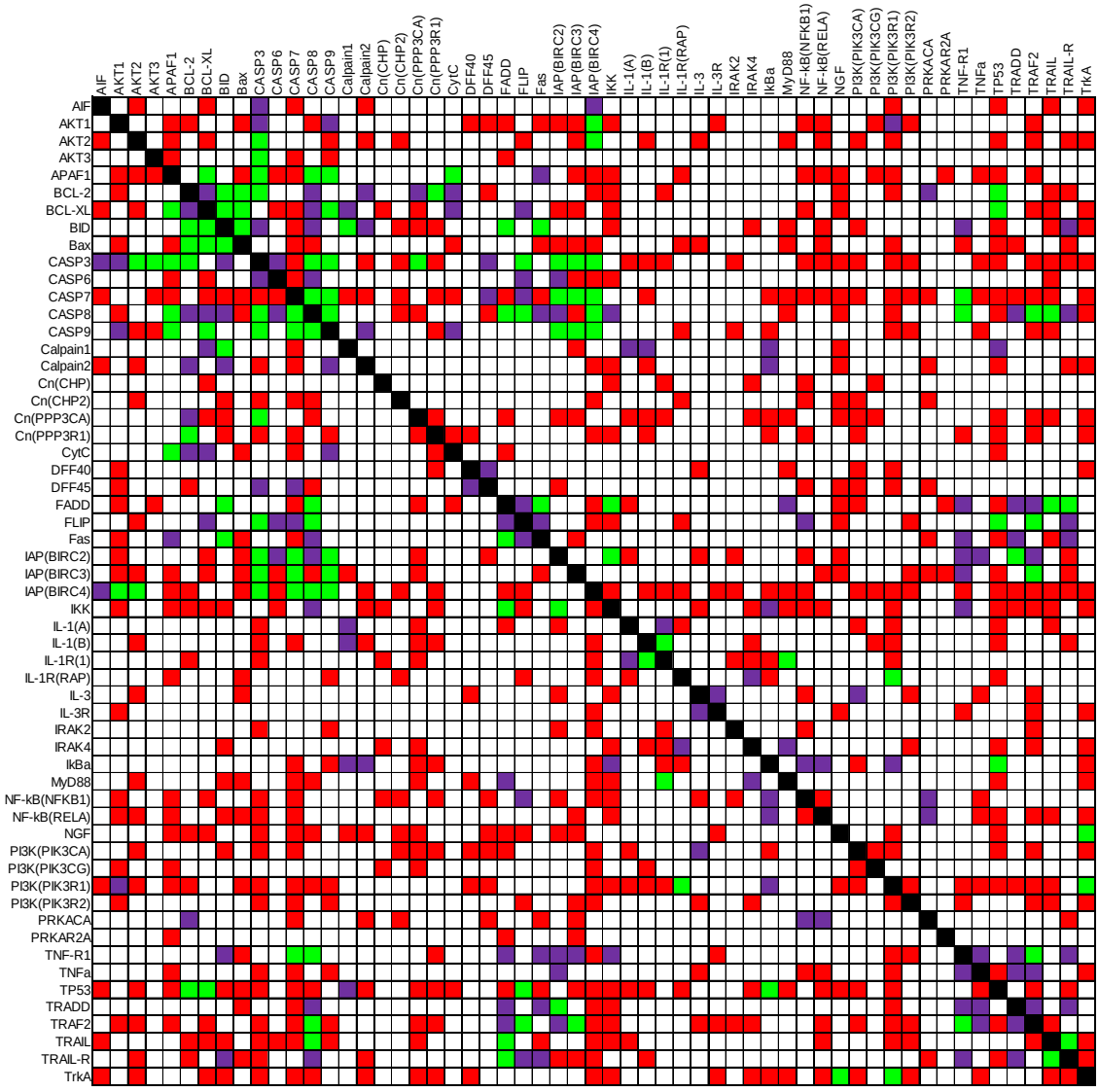


Figure 6.3: Predicted interactions by MEGADOCK. The green colored cells are true positives, the red colored cells are false positives and the purple colored cells are false negatives, validated by STRING database. The diagonal cells (black colored cells) are self-interactions and are not prediction targets, because the STRING database does not contain existing self-interactions.

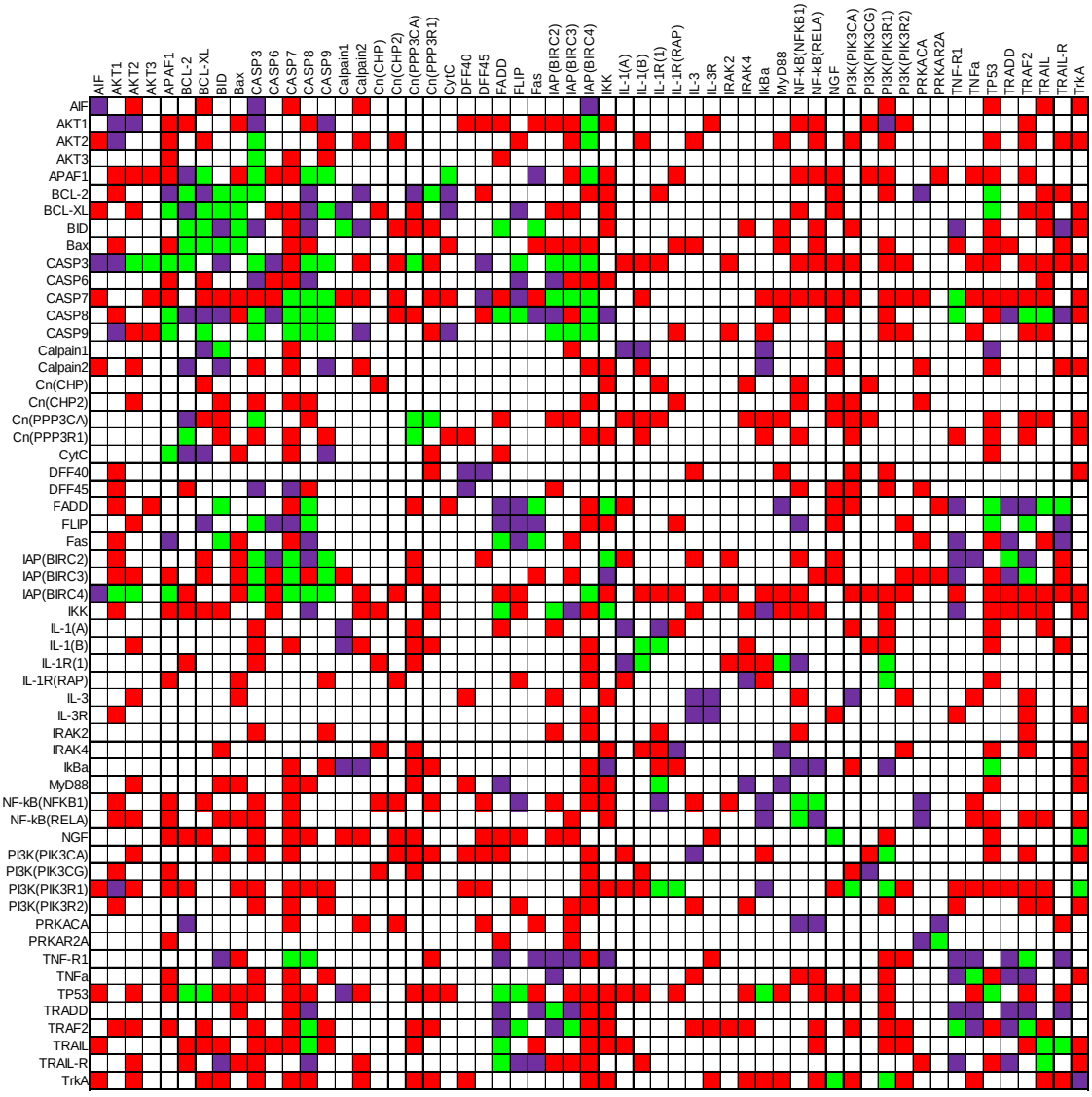


Figure 6.4: Predicted interactions by MEGADOCK. The green colored cells are true positives, the red colored cells are false positives and the purple colored cells are false negatives, validated by LIM database.

Table 6.3: The prediction results of the human apoptosis pathway. The row of "PRISM" shows results of [114].

Method	Database	TP	FP	FN	TN	Precision	Recall	F-measure
MEGADOCK	STRING	62	365	75	1,094	0.145	0.453	0.220
$\text{MEGADOCK}_{\text{TP=56}}$	STRING	56	338	81	1,121	0.142	0.409	0.211
MEGADOCK	STRING+LIM	88	364	96	1,105	0.195	0.478	0.277
PRISM	STRING	56	186	81	1,273	0.231	0.409	0.296

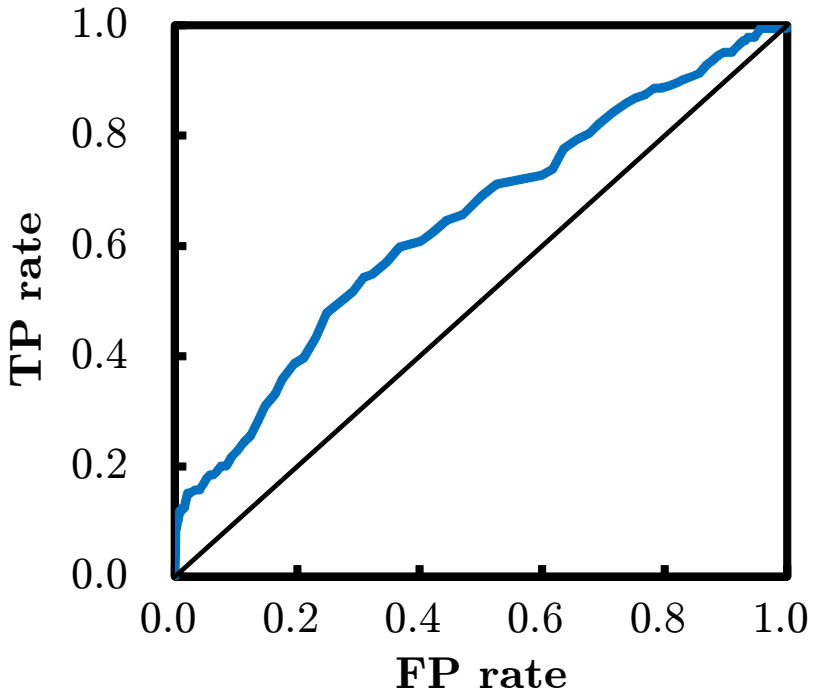


Figure 6.5: Evaluation of the prediction system in apoptosis dataset. The ROC curves for varying the threshold E^* values are shown. x-axis is for the false positive rate ($\frac{\mathrm{TP}}{\mathrm{TP+FN}}$) and y-axis is for the true positive rate ($\frac{\mathrm{FP}}{\mathrm{FP+TN}}$). Random prediction is indicated by the diagonal.

6.3.2 Predicted interactions

In this section, we discuss some predicted protein pairs that are not contained PPI databases (FPs).

(a) CASP3, CASP7

Both CASP3 (caspase-3) and CASP7 (caspase-7) are effector caspases activated by the initiator caspases. However, it is known that some effector caspases activate other effector caspases such as CASP3 and CASP6 that are both effector caspases [123]. Although the initiator and effector caspase cascade is well known, interactions among effector caspases are disputed [124].

CASP3, CASP6, and CASP7 are possibilities of being activated to [125, 126] and mutual although functions differ. Although the functions of CASP3, 6 and 7 are different [125, 126], it is considered enough and the possibility of the interaction of

CASP3 and CASP7 is actually suggested the possibility of interaction by Guerrero, *et al.* [127]. Fig. 6.6 shows the predicted complex structure of CASP3 and CASP7 generated by MEGADOCK. The predicted complex consists of 2DKO chain A (caspase-3, p17 subunit) and 2QL9 chain B (caspase-7, p10 subunit).

Additionally, 2DKO chain B (caspase-3, p12 subunit) and 2QL9 chain B, and 2QL9 chain A (caspase-7, p20 subunit) and 2DKO chain A, respectively, have similar structures. Thus, the predicted complex with each subunit swapped, as shown in Fig. 6.6, is similar to the original heterodimer and possibly predicted to occur with a high score. The interaction among effector caspases, as in this case, has not been examined by biological experiments. In fact, another PPI prediction tool based on template structure and database information, PrePPI [128, 129] (version 1.2.0), predicted the pair of caspase-3 and caspase-7 with a high score (the final probability value was 0.99). This situation is difficult to avoid in large-scale prediction problems. However, efforts such as the Negatome project [130] will help to improve this difficulty in the future.

(b) Akt, Bax

Bax directly induces release of cytochrome C and Akt is said to be Bax activation regulator [131]. Akt can prevent apoptosis upon growth factor withdrawal so Akt is called survival signal transduction factor.

Fig. 6.7 shows the Akt1–Bax complex predicted by MEGADOCK. Akt1 is the one of Akt isoforms.

(c) BID, IKK

BID works in the cross talk of the extrinsic and intrinsic pathways. The direct interaction of BID and IKK (IκK kinase) is not validated but it is known that BID interacts a protein complex containing IKK protein [132]. Fig. 6.8 shows the BID–IKK complex predicted by MEGADOCK.

6.4 Summary

In this chapter, we applied MEGADOCK to human apoptosis pathway related a various diseases and attached great importance as a target of drug discovery.

As compared with the predicted performance of PRISM which is another PPI prediction system based on known complex structure as template information, MEGADOCK brought a slightly low performance.

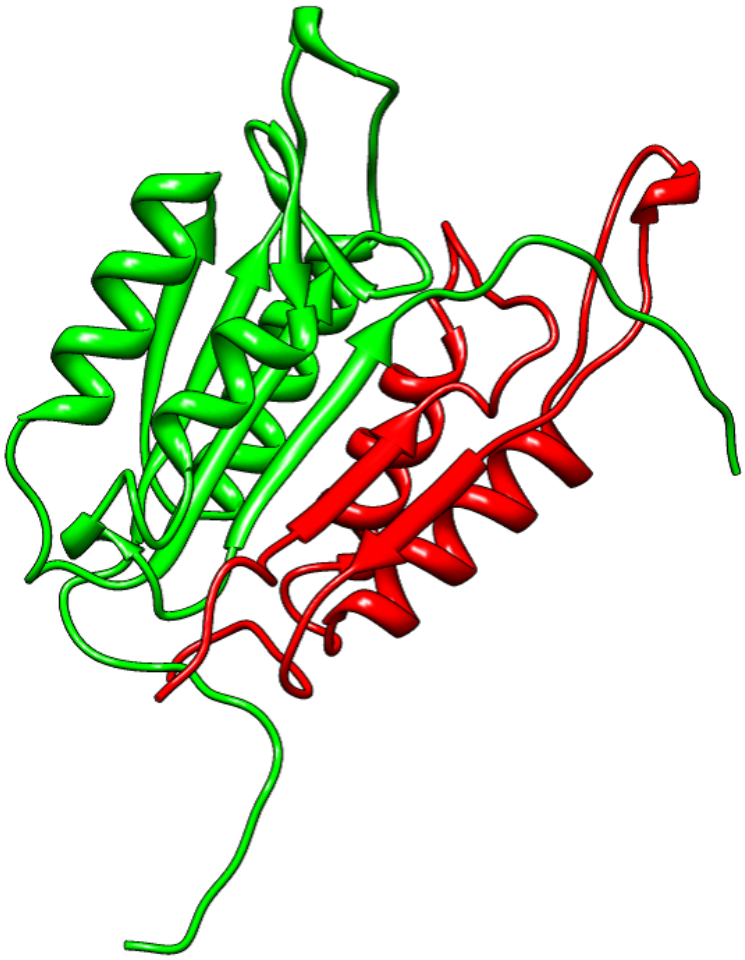


Figure 6.6: The predicted complex structure of CASP3 and CASP7 by MEGADOCK. Green colored protein is CASP3 (PDB: 2DKO_A), red colored protein is CASP7 (P10 subunit, PDB: 2QL9_B).

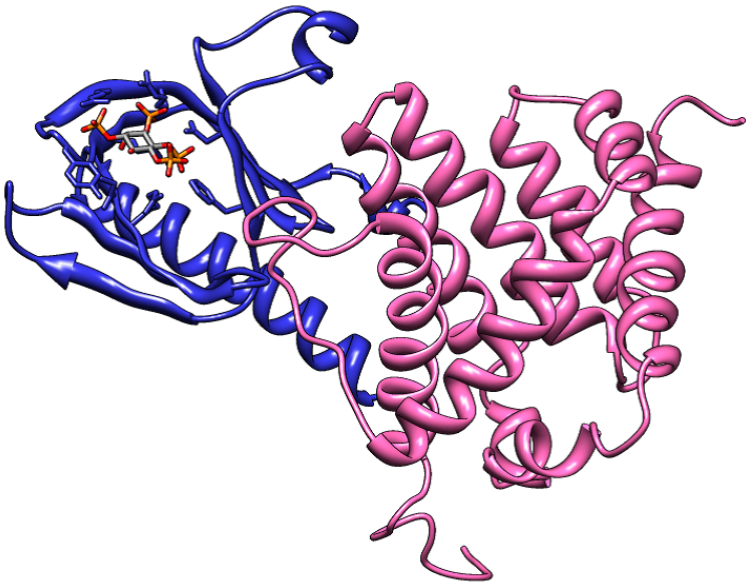


Figure 6.7: The predicted complex structure of Akt1 and Bax by MEGADOCK. Blue colored protein is Akt1 (PDB: 1UNQ_A), pink colored protein is Bax (PDB: 1F16_A).

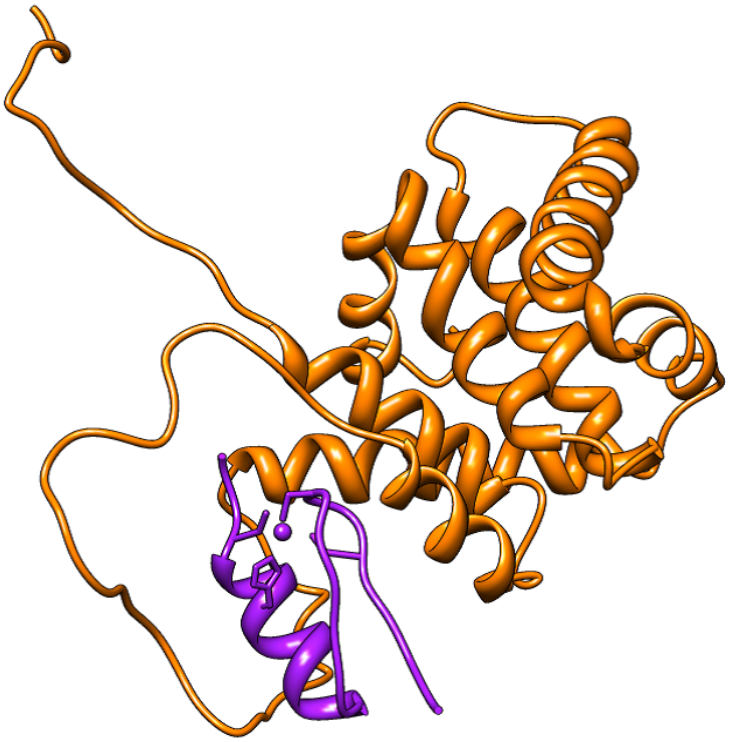


Figure 6.8: The predicted complex structure of BID and IKK by MEGADOCK. Orange colored protein is BID (PDB: 2BID_A), purple colored protein is IKK (PDB: 2JVX_A).

Also, the MEGADOCK results included several new PPI candidates such as CASP3–CASP7, Akt–Bax and BID–IKK.

Chapter 7

Expansion into Protein–RNA Interaction Prediction

7.1 Introduction

Elucidating protein–RNA interactions is important for understanding cellular systems such as protein synthesis, gene expression and regulation. Predicting whether an RNA-binding protein can recognize a given RNA molecule is a great challenge in computational biology. The number of available crystal structures of protein–RNA complexes has recently increased, enabling a protein–RNA interaction prediction based on tertiary structures.

Several statistical studies have been performed on protein–RNA interactions by using known crystal structures, and these studies provided numerous suggestions regarding base-residue interaction propensities [133, 134, 135, 136]. The number of studies on protein–RNA rigid-body docking by using fast Fourier transform (FFT)-based protein-protein docking methods like Molfit [26], FTDock [28] and ZDOCK [31, 32] has gradually increased [137]. Recently, at CAPRI, an international competition on complex structural prediction [47, 84], several research groups studied protein–RNA complexes by using the above mentioned approach. In some studies, for example, the study performed by Peréz-Cano, *et al.* [137], the technique used could be applied to several (but not a large number of) protein–RNA complexes. No studies involving exhaustive interaction prediction for a large number of protein–RNA complexes have yet been performed. However, a large number of proteins and RNA molecules interact in the cell, and it presumed that at least 1,500 RNA-binding proteins are coded for in the human genome [138]. Therefore, it is thought that exhaustive interaction prediction

for proteins and RNA molecules will become increasingly important in the future.

In this study, we developed a new protein–RNA interaction prediction method by extending our protein-protein interaction prediction system MEGADOCK with tertiary structure data.

We evaluated the proposed method by using 78 protein–RNA complex structures from the PDB. We predicted the interactions for pairs in 78×78 combinations. Of these, 78 original complexes were defined as positive pairs, and the other 6,006 complexes were defined as negative pairs.

7.2 Materials and Methods

In this study, a new protein–RNA interaction prediction method by using MEGADOCK has been proposed; this system was developed by enhancing the existing MEGADOCK system for ribonucleotide molecules. This section explains the components of the dataset used and of MEGADOCK.

7.2.1 Dataset

The dataset used for assessing accuracy consisted of 78 X-ray protein–RNA complex structures with a resolution of ≤ 3.0Å and a mutual sequence identity of $\leq 30\%$. We obtained the dataset from the PDB by using the PISCES server [139, 140]. The list of PDB IDs of the protein–RNA complexes used is shown in Table 7.1.

7.2.2 Extend to RNA molecules

MEGADOCK was initially developed for protein-protein interactions. It could not analyze RNA structures because the atomic charge parameters and atomic Van der Waals radius parameters in CHARMM19 are defined only for amino acids. Therefore, we developed a new version of MEGADOCK, which included deoxyribonucleotide and ribonucleotide atomic radius and charges taken from CHARMM27 [141] as done in a protein-DNA docking study [142]. In enhanced MEGADOCK, protein is treated as receptor molecule and RNA is treated as ligand molecule.

7.2.3 Protein–RNA interaction decision

From the results of docking calculations, we predicted whether a protein–RNA pair can interact or not by the same method of protein-protein interaction prediction. How-

Table 7.1: List of the PDB IDs of the 78 protein–RNA complexes used.

RNA type	PDB ID of the complex
tRNA	1ASY, 1B23, 1F7U, 1FFY, 1H3E, 1H4S, 1K8W, 1N78, 1Q2R, 1QF6, 1QTQ, 1R3E, 1SER, 1TFW, 1U0B, 2AZX, 2B3J, 2BTE, 2CT8, 2FK6, 2FMT, 2GJW, 2I82, 2R8S, 2ZZM, 3EPH, 3FOZ
mRNA	1KNZ, 1M8X, 1WPU, 1WSU, 1ZBH, 2ANR, 2F8K, 2HW8, 2IPY, 2J0S, 2PJP, 3K62
rRNA	1FEU, 1MZP, 2ASB, 2BH2, 3AEV
ssRNA	1FXL, 2BX2, 2JLV, 2R7R, 2VNU, 3FHT, 3I5X, 3IEV
dsRNA	1N35, 2AZ0, 2NUG, 2ZKO, 3EQT
siRNA	1SI3, 2BGG, 2F8S, 2ZI0
SRP RNA	1HQ1, 1JID, 1LNG
viral RNA	1F8V, 2E9T, 2QUX, 3BSO
RNA aptamer	1OOA, 3DD2, 3EGZ
others	3IAB (ribozyme), 2GXB (Z-RNA), 1A9N, 2OZB (snRNA), 1SDS, 3HAX (snoRNA)

ever, the reranking tool (ZRANK) is not applicable for RNA molecules. Thus we do not conduct reranking step.

The PRI(i, j) of protein i and RNA j evaluation value E is defined as follows:

$$E = \frac{S_1 - \mu}{\sigma}, \tag{7.1}$$

$$\mu = \frac{1}{D}\sum_{k=1}^{D} S_k, \tag{7.2}$$

$$\sigma^2 = \frac{1}{D}\sum_{k=1}^{D} (S_k - \mu)^2, \tag{7.3}$$

where S_1 is the top-ranked decoy's docking score for a protein–RNA pair, S_k is the k-th ranked decoy's docking score, and D is the number of decoys. In this study, we generated 10,800 decoys ($= D$) by using MEGADOCK. We concluded that a pair interacts if E is larger than threshold E^*:

$$\mathrm{PRI}(i, j) = \begin{cases} \text{True} & \text{if } E > E^* \\ \text{False} & \text{otherwise} \end{cases} \tag{7.4}$$

7.3 Results and Discussion

7.3.1 Performance of protein–RNA docking

To confirm the protein–RNA docking performance of enhanced MEGADOCK, we conducted docking using protein and RNA structures taken from protein–RNA complex structure data (Table 7.1) and confirmed the accuracy of the docking pose prediction (re-docking test). We took 3,600 kinds of ligand molecule (RNA) rotation patterns at intervals of 15 degrees. For each rotation, the translation position with the best docking score was output. The outputs of docking were 3,600 high-ranked decoys by default. We used two values, Best Rank and RMSD_{best} which were shown in Chapter 3 and repeated below, for evaluation:

- Best Rank: Highest rank of the near-native decoys. Here we defined near-native decoys as those that are included in the 3,600 highest scoring decoys and have an RMSD of < 5 Å,
- RMSD_{best}: RMSD of the highest ranked near-native decoy (Å),

where RMSD involves all the atoms between the decoy RNA and original X-ray crystal RNA structure when the receptor (protein) is fixed.

The re-docking test results are shown in Table 7.2, along with the results of MEGADOCK rPSC, rPSC+ES+RDE and ZDOCK 2.1 (PSC). ZDOCK 2.3 (PSC+ES+DE) and ZDOCK 3.0 (PSC+ES+IFACE) uses only amino acid parameters (CHARMM19 atomic charge and Atomic Contact Energy for amino acids) and cannot apply the electrostatics and desolvation score to RNA molecules; therefore, these results represent shape complementarity (PSC score) effects.

The gray cells in Table 7.2 provide the results for a Best Rank of 1. This means that the first ranked decoy has near-native structure. The number of complexes with first ranked decoys that have near-native structure (number of gray cells in Table 7.2) for MEGADOCK rPSC is 47, for MEGADOCK rPSC+ES+RDE is 51, and for ZDOCK is 48. The results for protein–RNA docking were suggestive of the importance of electrostatic interaction and desolvation effect.

Some re-docking complex structures are shown in Fig. 7.1; the structure for the PDB ID 2NUG is shown in (a), 3EPH is shown in (b), and 3FOZ is shown in (c). In each figure, two RNA structures are shown: the green RNA structures are the first ranked decoys generated by MEGADOCK, and the red RNA structures are X-ray crystal structures, that is, native structures. The RMSD for 2NUG is 1.45 Å, 2.25 Å

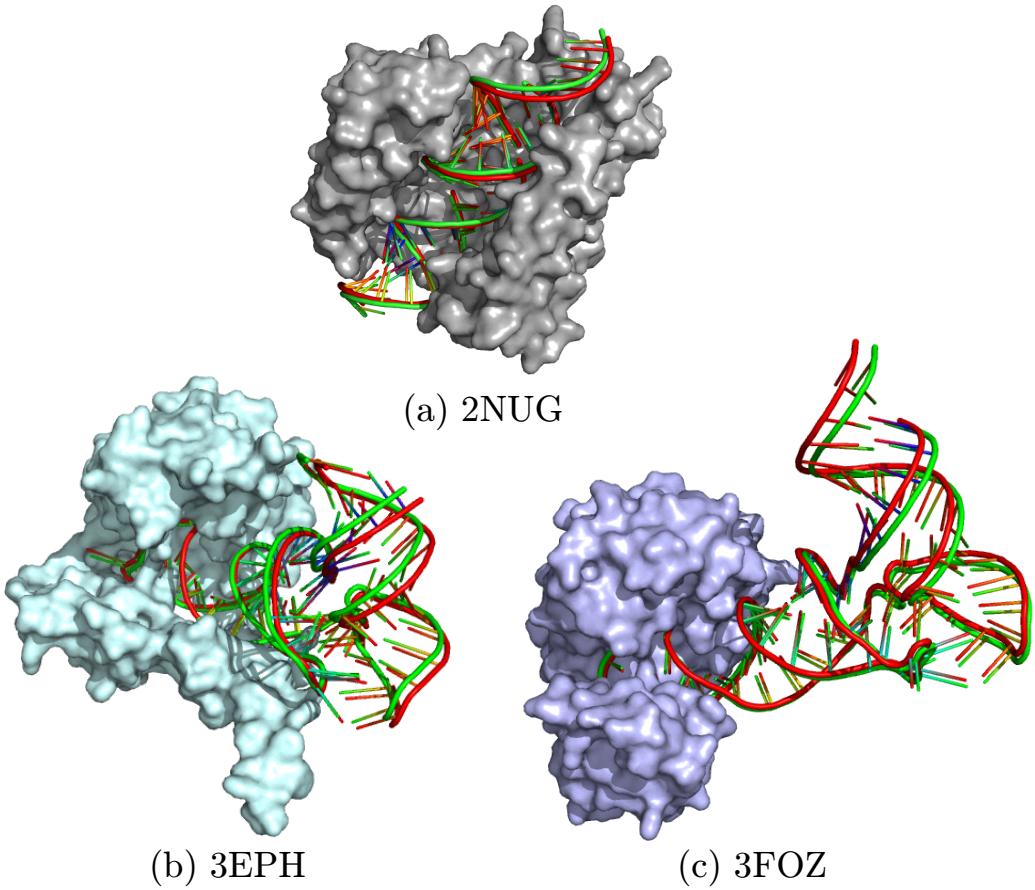


Figure 7.1: The structures after re-docking are shown for (a) 2NUG, (b) 3EPH, and (c) 3FOZ. In each figure, two RNA structures are shown: the green structure is the first ranked decoy generated by MEGADOCK, and the red structure is the original X-ray crystal structure.

for 3EPH, and 3.03 Å for 3FOZ (Table 7.2); the results indicate that MEGADOCK provides good structural predictions.

7.3.2 Performance of protein–RNA interaction prediction

We calculated 78×78 all-to-all docking with enhanced MEGADOCK and predicted possible protein–RNA interactions. The accuracy of the decision regarding whether the given protein–RNA pair could interact was evaluated in terms of sensitivity and

Table 7.2: Results for protein–RNA re-docking test of MEGADOCK and ZDOCK. The gray cells are $RMSD_{best} = 1$. "-" indicates that there was no near-native decoy (RMSD is less than 5 Å) existing in 3,600.

	old MEGADOCK rPSC		enhanced MEGADOCK rPSC+ES+RDE		ZDOCK 2.1 (PSC)	
PDB ID	Best Rank	$RMSD_{best}$	Best Rank	$RMSD_{best}$	Best Rank	$RMSD_{best}$
1A9N	1	1.068	1	1.068	1	1.064
1ASY	1	1.702	1	1.702	1	1.586
1B23	7	3.144	1	3.144	1	3.441
1F7U	1	2.392	1	2.392	1	1.986
1F8V	-	-	-	-	-	-
1FEU	171	1.817	57	1.617	121	2.259
1FFY	1	3.030	1	3.030	1	2.735
1FXL	1	0.899	1	1.055	1	0.714
1H3E	40	1.227	5	1.980	4	1.741
1H4S	3	2.556	2	3.546	1	2.636
1HQ1	4	1.871	1	1.871	1	1.750
1JID	334	1.512	135	1.512	108	1.792
1K8W	1	1.454	1	1.454	1	1.300
1KNZ	1	0.777	1	0.777	1	0.882
1LNG	1	2.683	1	2.683	2	2.325
1M8X	1	2.454	1	1.521	457	2.019
1MZP	1	1.638	1	1.638	1	1.382
1N35	1	0.842	1	1.299	1	1.625
1N78	1	1.980	1	1.980	9	2.774
1OOA	10	1.500	2	1.500	1	2.052
1Q2R	1	1.491	1	1.491	1	3.632
1QF6	1	2.032	1	2.123	1	1.691
1QTQ	1	1.837	1	1.837	1	2.133
1R3E	1	1.318	1	1.318	1	0.723
1SDS	-	-	-	-	1749	4.730
1SER	-	-	642	4.099	9	2.711
1SI3	1	0.878	1	2.714	2	1.078
1TFW	1	1.697	2	1.153	1	1.427
1U0B	1	1.803	1	2.076	1	2.040
1WPU	1	0.661	1	3.483	1	1.842
1WSU	-	-	763	1.468	1866	4.273
1ZBH	7	2.376	1	2.376	34	2.754
2ANR	148	1.915	46	1.915	307	1.967
2ASB	1	0.952	1	0.952	1	1.226
2AZ0	25	1.419	3	2.623	7	1.364
2AZX	-	-	-	-	151	3.468
2B3J	1	1.362	1	1.362	1	1.528

Table 7.2 (continue)

	old MEGADOCK rPSC		enhanced MEGADOCK rPSC+ES+RDE		ZDOCK 2.1 (PSC)	
PDB ID	Best Rank	$RMSD_{best}$	Best Rank	$RMSD_{best}$	Best Rank	$RMSD_{best}$
2BGG	1	0.864	1	0.864	1	1.309
2BH2	1	1.323	1	1.323	1	1.052
2BTE	18	2.115	2	2.207	26	1.430
2BX2	-	-	-	-	-	-
2CT8	-	-	18	2.455	3	2.608
2E9T	1	3.995	1	3.995	2	4.906
2F8K	1931	3.905	4	2.888	-	-
2F8S	-	-	-	-	-	-
2FK6	377	1.292	113	1.292	105	1.474
2FMT	27	2.434	3	2.740	7	4.533
2GJW	2	1.529	1	1.529	2	1.356
2GXB	1	1.222	1	0.680	1	0.763
2HW8	1	1.268	1	1.268	1	1.380
2I82	1	0.927	1	0.927	1	0.902
2IPY	10	1.312	1	1.414	1	1.393
2J0S	5	0.841	7	2.315	1	0.951
2JLV	1	1.924	1	1.924	1	0.860
2NUG	1	1.453	1	1.453	1	1.717
2OZB	7	1.325	1	1.325	1	1.473
2PJP	40	1.471	1	1.471	11	2.016
2QUX	95	0.945	4	1.261	6	1.017
2R7R	2	1.139	7	1.139	10	1.176
2R8S	24	2.706	101	2.706	1	3.109
2VNU	1	1.155	1	1.155	1	0.978
2ZI0	1	1.274	1	1.274	1	1.275
2ZKO	1	1.346	2	1.855	2	1.251
2ZZM	1	2.094	1	2.094	1	2.153
3A6P	1	2.545	1	3.703	1	2.998
3AEV	1	0.898	1	0.898	1	0.922
3BSO	1	0.820	1	0.820	1	1.130
3DD2	93	4.456	59	2.200	6	2.285
3EGZ	1	3.535	7	2.394	1	2.428
3EPH	1	2.249	1	2.249	1	2.135
3EQT	1	0.755	1	0.755	1	1.205
3FHT	1	0.829	1	0.919	1	1.185
3FOZ	1	3.027	1	3.027	1	2.479
3HAX	1	1.926	1	1.452	1	1.814
3I5X	1	0.960	1	0.960	2	1.087
3IAB	1	2.005	1	2.115	1	1.847
3IEV	1	1.004	1	1.004	1	1.040
3K62	1	1.311	1	1.001	1	1.156

specificity:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

where TP is the number of true-positive samples, TN is the number of true-negative samples, FP is the number of false-positive samples and FN is the number of false-negative samples. The overall performance of the protein–RNA interaction prediction was evaluated using the F-measure as follows:

$$F\text{-measure} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}.$$

We predicted the protein–RNA interaction with variation of the parameter E^* (3.0 – 13.0 with increments of 0.1). In the dataset, there were 78 true interactions where each pair had exclusively one interacting partner. The assessment of our method when applied to the dataset is shown in Fig. 7.2. The parameter $E^* = 9.6$ yielded the best F-measure of 0.465 with a sensitivity of 0.385 and a specificity of 0.997. Although proper comparison is not possible because the issues addressed are quite different, a similar prediction method for PPIs that uses MEGADOCK and provides 44×44 all-to-all PPI prediction has been reported; an F-measure value of 0.43 was obtained (see Chapter 6). The accuracy of MEGADOCK on protein–RNA interaction predictions is almost equal to the case of protein-protein interactions.

Moreover, we conducted the same protein–RNA interaction prediction using two subsets of data made by dividing 78 pairs into half, in order to exclude the possibility of the overfitting. The process corresponds to a 2-fold cross validation. The results are shown in Fig. 7.3. Because the two values of E^* that yielded the maximum F-measure value for each subsets were almost equal, it can be said that overfitting did not occur.

The receiver-operator characteristics (ROC) curve [85] for the results of the 78×78 interaction predictions is shown in Fig. 7.4. An ROC curve is a plot of sensitivity (= true-positive fraction) and specificity (or false-positive fraction = 1 − specificity), and shows the trade-off between sensitivity and specificity. A completely random prediction would give a diagonal line from the left bottom to the top right corners in the plot. The points above the diagonal line represent that the prediction is better than random. The ROC curve in Fig. 7.4 clearly shows that our method is better than random prediction. The area under the curve (AUC) of the ROC curve was 0.821,

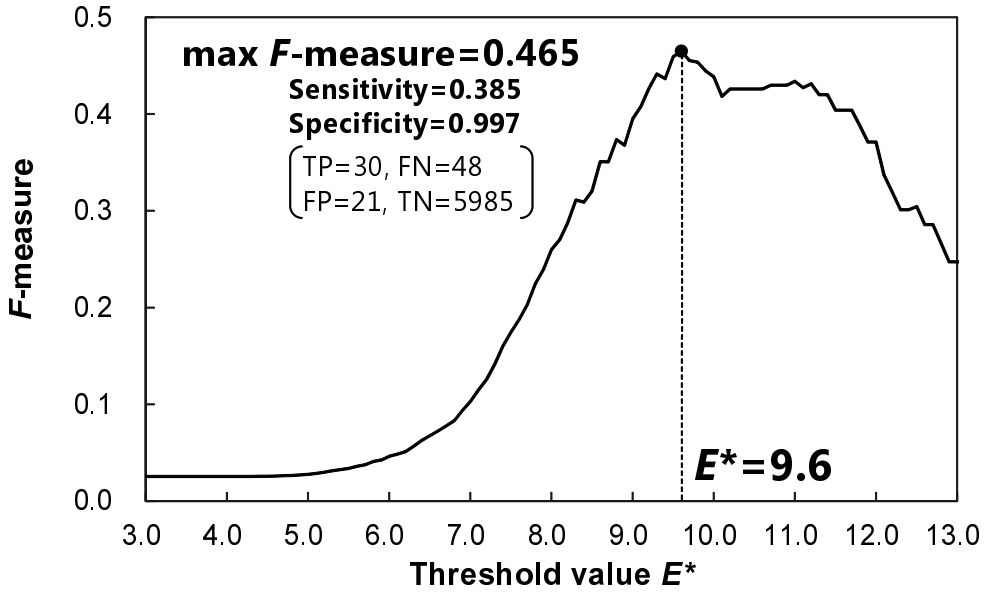


Figure 7.2: Results of the 78×78 predictions. This graph shows the change in the F-measure with respect to the threshold E^*. The maximum F-measure is 0.465 when E^* is 9.6, with a sensitivity of 0.385 and a specificity of 0.997.

showing the better performance of our method over random prediction (diagonal line, its AUC is 0.5).

The protein–RNA interaction map of the all-to-all protein–RNA interaction prediction results is shown in Fig. 7.5. The prediction results for all the pairs from the 78×78 combinations are shown in the form of a heatmap in Fig. 7.5. The red cells are those for which the E-value is larger than $E^*(= 9.6)$. The cell in the diagonal line indicate the combinations that originally exist and are predicted to interact.

7.3.3 False-positive predictions

From among the 21 existing false-positive pairs, we confirmed the structure of RNaseIII from *Aquifex aeolicus* complexed with dsRNA (2NUG) and from tRNA transferase coupled with tRNA (3EPH, 3FOZ). Because the protein taken from 2NUG is RNase, it seems to be natural that our system predicted interactions of this protein with some other dsRNAs in addition to the RNA from the original complex structure. The PDB IDs and the descriptions for the targeted crystallographic structures are shown in Table 7.3, which also shows the PRI value of each pair and the prediction results. The docking structures of the 2NUG protein and RNA of 2GJW, 2ZKO, and

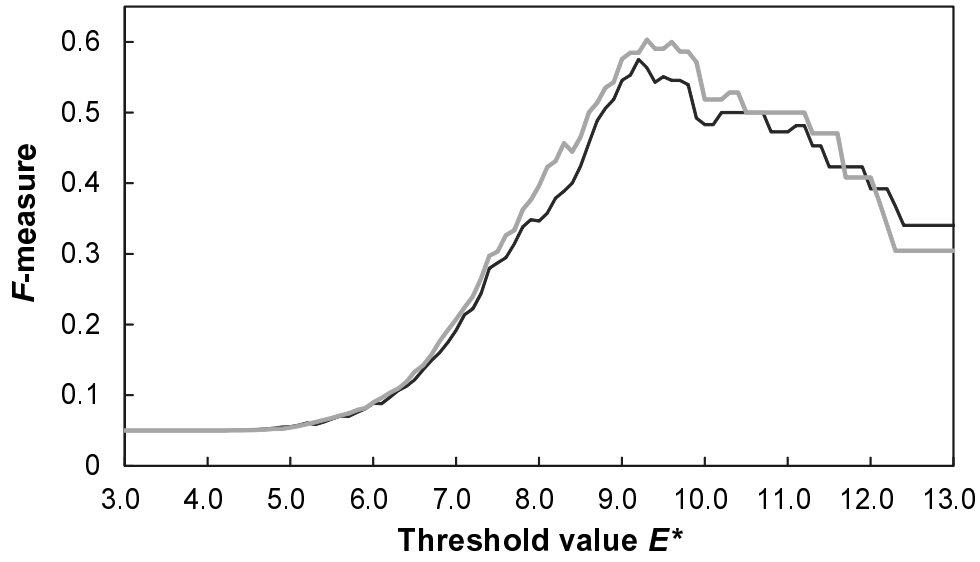


Figure 7.3: Results of 2-fold cross validation prediction performed using the divided 39×39 subset. This graph shows the change of F-measure with respect to the threshold E^*. Because the value of E^* that yielded the maximum F-measure value was almost equal, it can be said that overfitting did not occur.

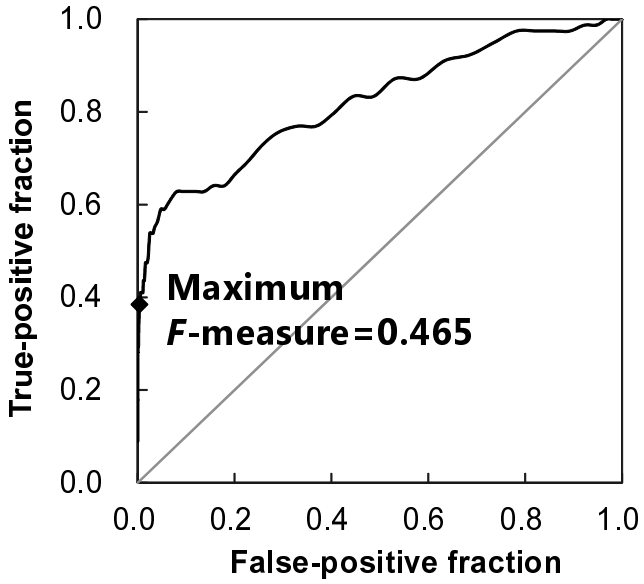


Figure 7.4: ROC curve of 78×78 dataset prediction results. The area under the curve (AUC) is 0.821.

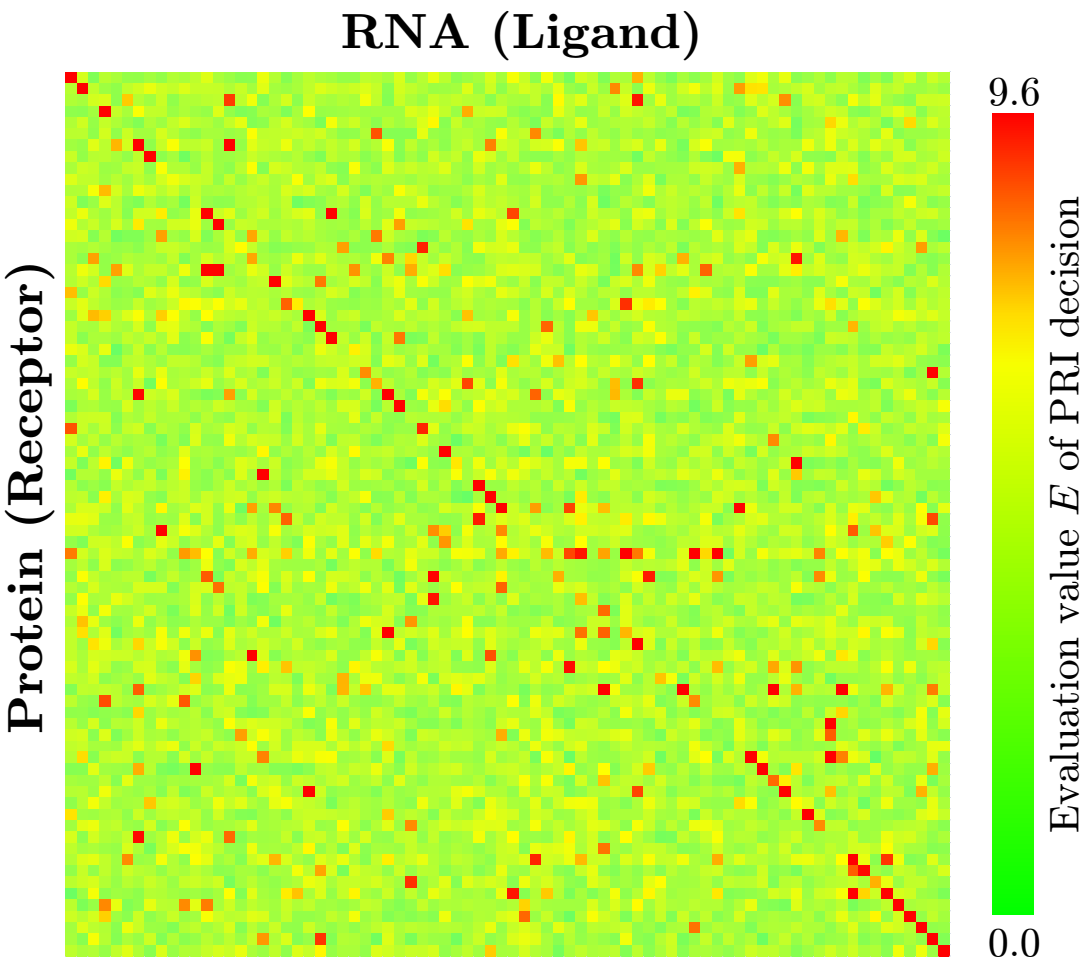


Figure 7.5: 78×78 map of protein–RNA interaction prediction results. The red cells are the cells for which the E-value is more than $E^*(= 9.6)$. The cells have been arranged according to the PDB IDs, which have been arranged in alphabetical order for all axes.

Table 7.3: PDB ID and the description of protein–RNA structures.

PDB ID	description
2NUG	RNase III from *Aquifex aeolicus* and dsRNA
2GJW	*Archaeoglobus fulgidus* tRNA-splicing endonuclease
2ZKO	NS1 protein of human influenza virus A and dsRNA
3EGZ	*Homo sapiens* u1 small nuclear ribonucleo-protein tetracycline aptamer and artificial riboswitch
3EPH	*Saccharomyces cerevisiae* dimethylallyl tRNA transferase and tRNA
3FOZ	*Escherichia coli* isopentenyl tRNA transferase and tRNA

Table 7.4: Interaction prediction results of protein–RNA pairs in Fig. 7.6 and Fig. 7.7.

Pair (Protein–RNA)	PRI value E	Prediction result
2NUG–2NUG	12.5	TP
2NUG–2GJW	9.8	FP
2NUG–2ZKO	9.9	FP
2NUG–3EGZ	11.1	FP
3EPH–3EPH	15.2	TP
3EPH–3FOZ	9.2	TN
3FOZ–3FOZ	17.9	TP
3FOZ–3EPH	12.5	FP

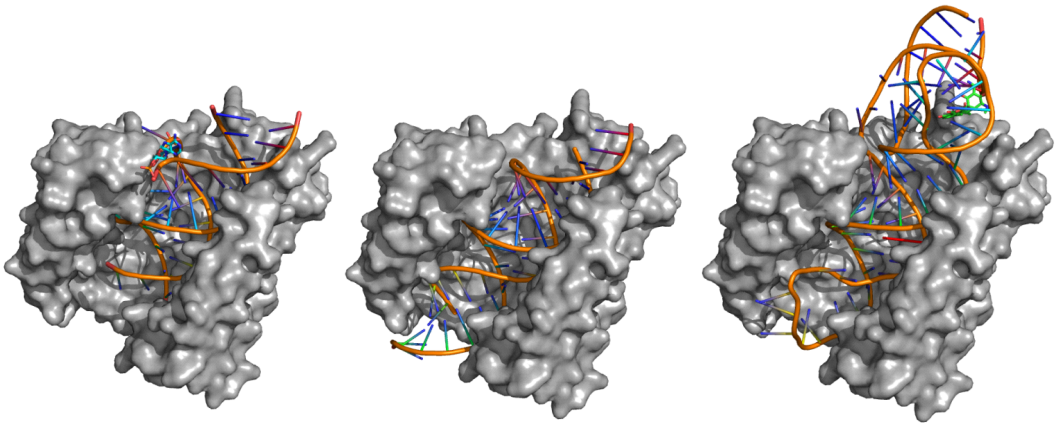
3EGZ are shown in Fig. 7.6. The docking structures of the pair comprising the 3EPH protein and the 3FOZ RNA and of the pair comprising the 3FOZ protein and the 3EPH RNA are shown in Fig. 7.7.

On comparing Fig. 7.1(a) and Fig. 7.6, the predicted structures of protein–RNA complex in Fig. 7.6 are similar to the crystal structure shown in Fig. 7.1. These three structures obtained high evaluation values. Fig. 7.1(b)(c) and Fig. 7.7 also shows protein–RNA pairs with similar predictions of the docking pose. In this result, 3FOZ–3EPH got high PRI value of $E = 9.2$ (Table 7.4).

It is thought that it can reflect a slight difference by the species (*E. coli* vs. *S. cerevisiae*) that only 3EPH-3FOZ's was judged as negative (not-interacting) among four combinations.

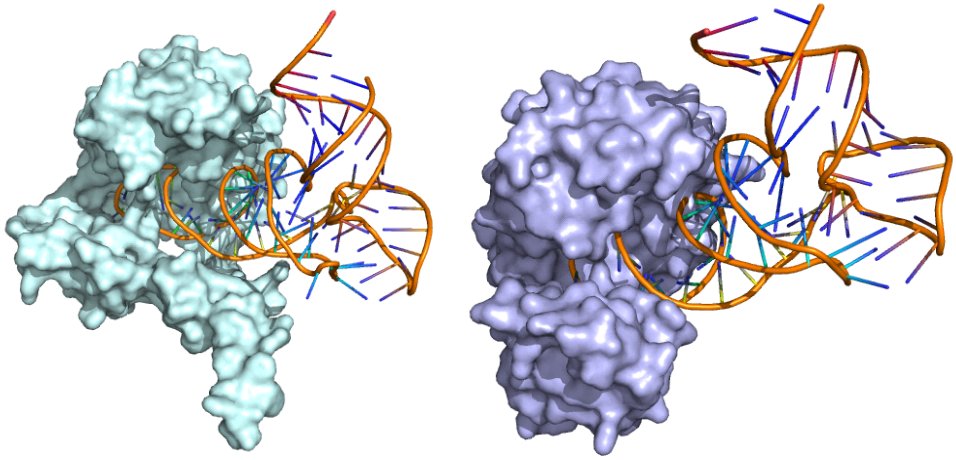
7.3.4 Limitations and challenges

RNA is a very flexible molecule, and therefore, the rigid-body docking approach seems a little inadequate for RNA. However, we think that exhaustive predictions



(a) 2NUG–2GJW (b) 2NUG–2ZKO (c) 2NUG–3EGZ

Figure 7.6: Protein (RNaseIII) of 2NUG and the RNA of the (a) 2GJW, (b) 2ZKO, and (c) 3EGZ docked structures. The structures are first ranked decoys generated by enhanced MEGADOCK



(a) 3EPH–3FOZ (b) 3FOZ–3EPH

Figure 7.7: (a) Protein of 3EPH and RNA for the 3FOZ docking structure and (b) protein of 3FOZ and RNA for the 3EPH docking structure. The structures are first ranked decoys generated by enhanced MEGADOCK

with the rigid-body approach can help elucidate structural similarity between several protein–RNA interaction events in a living cell.

In future work, we will apply our method to ensemble docking in order to account for RNA flexibility in the prediction. In a study involving ensemble docking in protein docking, ensemble docking was executed by ZDOCK by using two or more nuclear magnetic resonance (NMR) conformations [143]. When considering the flexibility of RNA, it is thought that taking the structural ensemble of RNA into account contributes to improving the accuracy of interaction prediction.

We think that further analysis of the protein–RNA interactions from several species that have similar functions is an interesting area for further studies. We have shown that exhaustive PPI prediction can contribute to systems biology research [23], and we will try to combine protein-protein and protein–RNA interaction predictions in future work.

7.4 Summary

In this study, the PPI prediction system MEGADOCK was enhanced for RNA, leading to the development of a protein–RNA interaction prediction system. The enhanced MEGADOCK is a rapid protein–RNA interaction prediction system that uses the rigid-body protein–RNA docking method and has a calculation accuracy almost equal to that of ZDOCK 2.3; the results suggested that electrostatic interaction contributes to a large extent in accuracy of docking pose prediction. Moreover, when it was applied to the exhaustive screening that predicts correct interacting pairs protein and RNA using 78 protein and RNA structure data, we obtained F-measure value of 0.465 and AUC value of 0.821.

In future work, further verification of the possibility of interaction judged false-positive. Possible analysis includes ensemble docking method and verification of method of evaluating mixture of protein-protein and protein–RNA interaction. Additionally, there is a possibility that MEGADOCK find difference of RNA binding proteins that species are different but that have similar functions. We will challenge a further analysis on this aspect.

Part IV

Integration with Other Protein–Protein Docking Methods

Chapter 8

Integration of Two Docking Tools with Different Scoring Models

8.1 Introduction

For improving accuracy of PPI screening, the way of utilizing the information acquired with other prediction tools is one of the answers though which may increase a calculation cost and narrowing an applicable range.

In this chapter, we conducted PPI network prediction by exhaustive docking using two different docking engines: ZDOCK 3.0 [33] and MEGADOCK. ZDOCK uses a scoring function that includes shape complementarity (PSC), electrostatics and a heuristic potential called atomic contact energy (IFACE). MEGADOCK is a similar system to ZDOCK that searches probable docking structures in a grid-based 3D space using FFT. MEGADOCK employs a much simpler score function and thus makes the calculations 8.9 times faster than ZDOCK.

8.2 Material and Methods

We predicted PPIs in bacterial chemotaxis by using MEGADOCK and ZDOCK. In ZDOCK prediction, we used the same procedure by MEGADOCK (see Chapter 4) and only swap docking engine. However, ZDOCK cannot change the parameter of the number of output decoys per rotations. Thus we set parameters of $n_\theta = 3{,}600$ decoys per target pair and $\theta = 15^\circ$ for the ligand rotation step. The target dataset is bacterial chemotaxis dataset shown in Table 5.2.

8.3 Results

8.3.1 Predicted PPIs

Fig. 8.1 shows the predicted PPIs by ZDOCK and MEGADOCK. The best F-measure value was 0.52 (ZDOCK, #TP = 12, #TN = 57, #FP = 11, #FN = 11, recall = 0.52, precision = 0.52, when $E^* = 7.9$) and 0.48 (MEGADOCK, #TP = 14, #TN = 47, #FP = 21, #FN = 9, recall = 0.61, precision = 0.40, when $E^* = 5.5$). For both the ZDOCK and MEGADOCK predictions, parameter values E^* was set as the same values that yielded the best F-measure value when applied to general benchmark data used in a previous study [23]. Previously known PPIs are colored gray in Fig. 8.1.

8.3.2 Considering protein localization

In the real cell, FliG, FliM and FliN proteins are closely associated with the membrane and only CheY is considered capable of interacting with these proteins. When we take into account protein localization, resulting in removing flagellar proteins (Figure 1, proteins circled by the dotted line) in the dataset, the best F-measure value was 0.69 (ZDOCK, #TP = 11, #TN = 34, #FP = 6, #FN = 4, recall = 0.73, precision = 0.65) and 0.54 (MEGADOCK, #TP = 11, #TN = 25, #FP = 15, #FN = 4, recall = 0.73, precision = 0.42). By restricting target proteins using localization information, both ZDOCK and MEGADOCK yielded better F-measure values, with both precision and recall values higher than that of the whole dataset. These results show that more accurate PPI predictions are made if protein localization is taken into consideration.

8.3.3 Comparison of the prediction by using ZDOCK and MEGADOCK

Fig. 8.2 shows a comparison of the results obtained by using ZDOCK and MEGADOCK. In total, 17 out of 23 relevant PPIs were detected when at least one of the docking software programs are used. Among 17 true positives, 9 were predicted by both of the software packages. Among the 28 false positives, 4 were common for both software packages and 7 were specific to ZDOCK, while 17 were specific to MEGADOCK. Thus, a lower precision value was obtained using MEGADOCK when compared to ZDOCK.

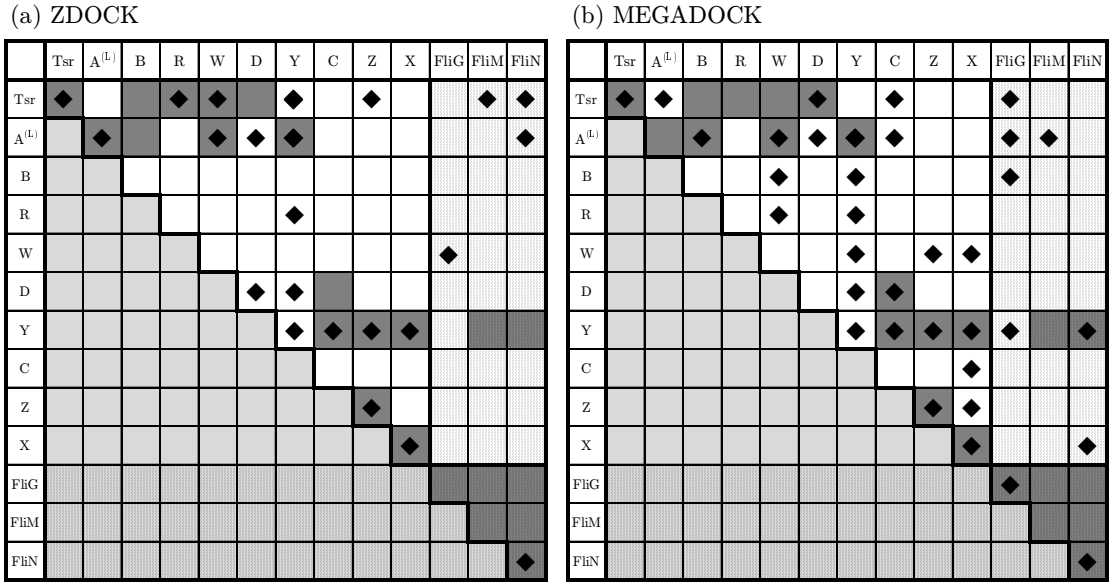


Figure 8.1: Predicted interactions among chemotaxis proteins. Predicted interactions among chemotaxis proteins by using (a) ZDOCK and (b) MEGADOCK as docking engines. The dark grey colored cells indicate known interacting pairs based on conventional studies. Cells with diamond marks indicate predicted interactions. Cells filled with small dots show flagella protein related combinations. Proteins related to the flagellar motor are listed on the right/bottom side. The short form of CheA is known to interact with CheZ [105] but it was not included because the structure was unavailable. A total of seven interactions that are not colored dark grey were found in the STRING database [106] by (i) searching interactions associated with experimental reports or (ii) those annotated in databases (KEGG [98], BioCyc [144]). The interactions are: CheY–FliG, CheY–CheW, CheB–CheW, Tsr–CheZ, Tsr–CheA, CheR–FliN, CheR–CheZ. These interactions were not considered as "correct" in this study because they have not been characterized.

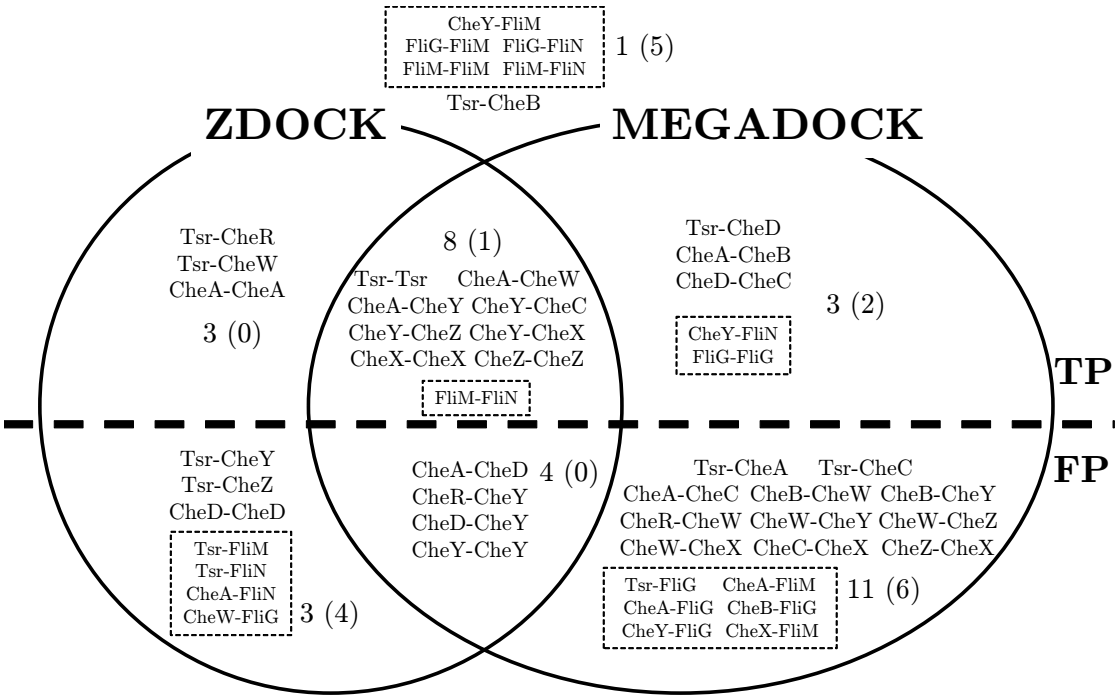


Figure 8.2: Predicted protein–protein interactions. Interactions listed inside the circles and above the dotted line show 'True Positive' pairs, those below the dotted line are 'False Positive' pairs. Pairs that are listed outside both circles are 'False Negative' pairs. Dotted boxes show flagella protein related interactions.

8.4 Discussion

8.4.1 Performance of PPI prediction

On the prediction of binary interactions, both ZDOCK and MEGADOCK yielded F-measure values of more than 0.4. When localizations of proteins were considered, ZDOCK performed better than MEGADOCK. It should be noted that MEGADOCK employs shape complementarity and an electrostatic score function whereas ZDOCK also takes into account heuristic score function based on atomic contact energy (IFACE) [33].

As shown in Fig. 8.2, eight interactions were detected only by one of the docking software programs. Tsr-CheR, Tsr-CheW and CheA-CheA interactions were detected only by ZDOCK. Tsr-CheD, CheA-CheB, CheD-CheC, CheY-FliN and FliG-FliG interactions were detected only by MEGADOCK. Two out of three of the interactions detected by ZDOCK, Tsr-CheW and CheA-CheA, are tight binding interactions that constitute the receptor complex. In the case of MEGADOCK, with the exception of FliG-FliG, all the five detected interactions are transient. These results suggest there are differences in the type of interactions detected when different score functions are applied. It is very interesting to see the difference of the predicted PPIs by using different score functions.

Applying other score functions for docking or conducting re-ranking calculations with more sophisticated score functions to the generated decoys would be useful for analyzing the effects of score function on PPI prediction. To investigate this further we require a more thorough dataset such as that used by Kastritis and Bonvin [145] to evaluate any correlation between score function types and known protein–protein binding affinities.

One such example of applying different PPI prediction procedures is given in Fig. 8.3, which shows PPI prediction results for a chemotaxis dataset using the PRISM protocol [14]. PRISM uses a template dataset of known protein–protein binding interfaces extracted from PDB. The surface of the target monomer protein, for which we want to identify binding partners, is analyzed against all the interface templates by structural alignment. Target protein pairs whose surface structures are aligned to any known interface pairs in the template dataset are then refined and scored by FiberDock [45].

PRISM identified four candidates that potentially interact with most of the proteins in the dataset. Specifically, of the 13 proteins involved in the chemotaxis dataset, CheA and CheZ were predicted to interact with 11 while Tsr and CheY were predicted

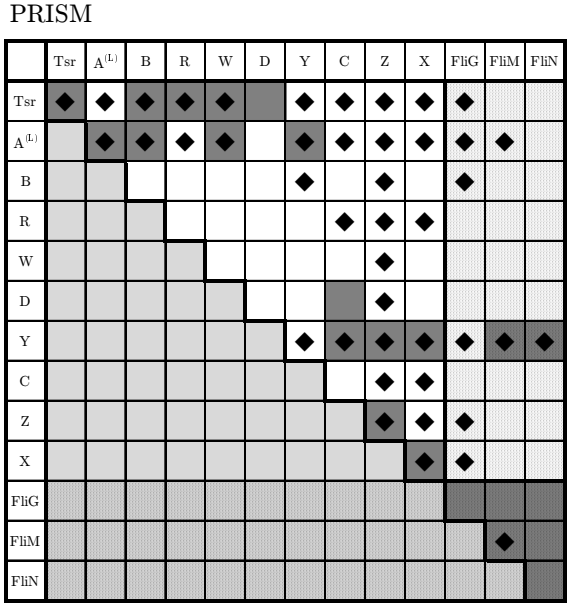


Figure 8.3: Predicted interactions among chemotaxis proteins identified by using PRISM. The cells with a diamond mark indicate the predicted interacting pairs. The prediction was performed by defining an interacting pair of proteins according to the following criteria: (i) if the two potential binding partners have an interaction surface that is aligned to a template dataset constructed from known crystal structures, (ii) the predicted binding event yields less than zero energy by FiberDock calculations. The dark grey coloured cells indicate known interacting pairs based on conventional studies.

to interact with 10. Interestingly, the docking-based procedure applied to CheZ gave fewer predicted binding partners i.e., ZDOCK (three partners) and MEGADOCK (four partners). In this chemotaxis dataset only two interactions are confirmed for CheZ i.e., oligomerization of CheZ and interaction with CheY-p. However, CheZ is also known to interact with the short form of CheA [105] and localize in the cell pole area where receptor complexes are located [146]. CheY, the main target for CheZ, moves between the receptor area and flagellar motor area. The template-based PPI prediction suggests that CheZ may undergo non-specific interactions. Thus, it would be of interest to further analyze the role of CheZ in the receptor complex area.

8.4.2 Protein localization

Both docking tools yielded better performances when flagellar motor related proteins are excluded from the target, while random prediction with a recall value of 0.5 yielded similar F-measure values (0.34 for the whole dataset and 0.35 for the restricted targets). It should be noted that direct binary interactions among flagellar motor proteins are still unclear and the true interacting pairs might be different from the "correct" interactions used here. Combining protein localization prediction methods such as PSORT [147] and SOSUI [148], especially for forecasting whether a given protein is membrane associated or soluble, to our PPI prediction would be useful when applying our method to large numbers of target proteins.

8.4.3 False negative interactions

When using both ZDOCK and MEGADOCK predictions, 7 interactions were not detected; FliG–FliG, FliM–FliM, FliG–FliM, FliG–FliN, FliM–FliN, Tsr–CheB and CheY–FliM (Fig. 8.2).

When we removed interactions among flagellar motors to consider protein localization, only the interaction between Tsr–CheB and CheY–FliM were not detected by both ZDOCK and MEGADOCK. It is known that for these interactions to occur CheB and CheY both need to be phosphorylated [149, 150]. In our dataset there were no protein structures of the phosphorylated forms. However, CheY structures used here include the mimicked activated state by using $BeFe^{3-}$, such as PDB ID: 1ZDM [151]. This activated structure showed only modest differences with the native structure in terms of backbone geometry [151]. Based on these results, we cannot assess whether a rigid-body docking method is capable of distinguishing the phosphorylated from the non-phosphorylated state. Nonetheless, our findings are understandable because we did not use a flexible docking tool that considers phosphorylation mediated conformational changes of CheY and CheB.

One possible mean of obtaining increased sensitivity in our PPI prediction model is to construct likely structural variations of target proteins and then use the ensemble set as a docking target.

8.4.4 False positive interactions

There are four common false positive PPIs (CheD–CheY, CheA–CheD, CheR–CheY, CheY–CheY) predicted both by ZDOCK and MEGADOCK, three of which include

CheY. In total, there are 51 structures of CheY in the dataset and 7 interactions out of 13 target proteins were predicted for CheY by both docking tools. Positive predictions were obtained using various structures of single protein species. This, however, does not mean that only specific protein structure pairs generate positive interactions. CheA has 21 structures in the dataset. Both of the docking tools predicted 5 interactions out of 13. The availability of more structural data for a given protein enriches structural variation and serves to increase sensitivity. In such cases we can consider using higher E^* value to get better precision.

This result is also understandable from the fact that CheY has multiple binding partners. Bacterial chemotaxis is a two-component signal transduction system consisting of a histidine kinase (CheA) and response regulators (CheB, CheY). CheA operates in the form of a complex with receptors and CheW. Phosphorylated CheB works as a modifier of receptor proteins, which accumulate at the cell pole [152]. While CheB operates in the local area around the receptor complex, CheY accepts signals from CheA and transmits them to the flagellar motors, which are evenly distributed around the entire surface of the cell. There are several processes that modify CheY activity during transmission of the signal; CheC, CheX (*T. maritima*) and CheZ (*E. coli*, *S. typhimurium*) have activity that dephosphorylates CheY [104, 107, 153]. Thus, CheY undergoes transient interactions with several different proteins during the signal transduction process. Indeed, our conclusion that CheY undergoes non-specific binding with many types of proteins is in agreement with our findings given in Fig. 8.1. It should also be noted that both docking tools predict CheY undergoes dimerization. Moreover, sequence homology based interlog search using PiSite [5] also suggests that dimerization of CheY is likely.

8.5 Summary

We conducted a reconstruction of the protein–protein interaction network using two distinct physical docking tools. The predicted interactions generated from the two tools were slightly different. However, when the positive predictions from both tools were combined, the vast majority of relevant interactions were represented. Indeed, there were only two exceptions, both of which required phosphorylation to activate the corresponding interaction.

Chapter 9

Integration of Template-based and *De Novo* PPI Prediction

9.1 Introduction

Recently, there are two typical approaches for tertiary-structure-based PPI predictions: a method based on template matching with known protein structures and another method based on *de novo* protein docking like MEGADOCK as discussed. The template-based method is based on the hypothesis that known complex structures or interface architectures can be used to model the complex formed between two target proteins. The hypothesis is logical, and this method provides good prediction performance when complex structural information is available as a template; however, if the template structure information is not available, performance is poor. In addition, because the interface architecture is not always similar for similar interactions, the template-based method has a narrow applicable range. In contrast, the *de novo* protein docking method has a wide applicable range because it uses only tertiary structural information. However, because the advantage provided by existing template information is not utilized, the prediction performance is not so good.

Tuncbag, *et al.* developed a template-based PPI prediction method called PRISM [14], which is based on information regarding the interaction surface of crystalline complex structures. PRISM has been applied for predicting PPIs in a human apoptosis pathway [114] and a p53-protein-related pathway [154], and has contributed to the understanding of the structural mechanisms underlying some types of signal transduction.

We developed a PPI prediction method called MEGADOCK [59] based on protein–

protein docking without interaction surface information. MEGADOCK has been applied for PPI screening for a bacterial chemotaxis pathway (Chapter 5) and a human apoptosis pathway (Chapter 6) and has contributed to the identification of protein pairs that may interact.

However, the prediction results of both template-based and *de novo* protein docking methods in these studies contained many false-positive predictions. PRISM obtained a precision value of 0.231 when applied to a human apoptosis pathway that consisted of 57 proteins, which was higher than the precision obtained with random prediction (precision value of 0.086), and MEGADOCK obtained a precision value of 0.400 when applied to a bacterial chemotaxis pathway that consisted of 13 proteins, which was higher than the precision obtained with random prediction (precision value of 0.253). To identify new PPIs, the prediction results need to be validated using biological experiments. For this purpose, obtaining a low number of predicted interaction candidates with high reliability is more important than obtaining a high number of predictions with low reliability. Thus, this paper aims to improve the reliability of the method used to obtain PPI predictions.

In this study, we combined two different PPI prediction methods to improve the precision of PPI prediction. Because PRISM is a template-based method, its prediction accuracy depends on the template dataset prepared. Only PPIs whose interaction surface structures are conserved are expected to be predicted. In contrast, MEGADOCK is a non-template-based method (also called *de novo* prediction), which has the demerit of generating false-positives for the cases in which no similar structures are seen in known complex structure databases; thus, template-based method would be ruled out from the prediction. However, in situations where template structures are not present in databases, MEGADOCK can still predict PPIs. This qualitative difference between the two methods typically makes their output different. Thus, the combination of both prediction methods may improve prediction accuracy, as the intersection set (AND set) of both results may contain fewer false-positives; this improvement in precision would also contribute to improvement in the prediction reliability provided by the use of just one method.

Such an approach is called a “meta” approach. Meta approaches have already been used in the field of protein tertiary structure prediction [155], and critical experiments have demonstrated improved performance of meta predictors when compared with the individual methods used in the meta predictors. The meta approach has also provided favorable results in protein domain prediction [156] and the prediction of disordered regions in proteins [157]. We have therefore proposed a new PPI prediction method

based on the consensus between template-based and *de novo* docking methods. Generally, a meta prediction method may have low applicability because meta approaches require applicable conditions for every method in the approach. However, if structural information is available, the *de novo* docking method introduced in this study is always applicable with or without template information. Thus, the applicability of the consensus method is not narrower than that of a template-based method.

9.2 Materials and Methods

9.2.1 Template-based PPI prediction

We used PRISM for template-based PPI prediction. PRISM uses two input datasets: the template set and the target set. The template set consists of interfaces extracted from protein pairs that are known to interact. The target set consists of protein chains whose interactions need to be predicted. The two sides of a template interface are compared with the surfaces of two target monomers by structural alignment. If regions of the target surfaces are similar to the complementary sides of the template interface, then these two targets are predicted to interact with each other through the template interface architecture.

The prediction algorithm consists of four steps: (1) interacting surface residues of target chains are extracted using Naccess [158]; (2) complementary chains of template interfaces are separated and structurally compared with each of the target surfaces by using MultiProt [159]; (3) the structural alignment results are filtered according to threshold values, and the resulting set of target surfaces is transformed onto the corresponding template interfaces to form a complex; and (4) the FiberDock [45] algorithm is used to refine the interactions to introduce flexibility, resolve steric clashes of side chains, compute the global energy of the complex, and rank the solutions according to their energies. When the computed energy of a protein pair is less than -10 kcal/mol, the pair is determined to "interact" (personal communication with Ms. Saliha Ece Acuner Ozbabacan, July 12, 2013). This prediction protocol has been described in detail in a previous study [14, 114].

9.2.2 *De novo* PPI prediction

For *de novo* PPI prediction, we used MEGADOCK with rPSC, ES and RDE functions (see Chapters 3 and 4). MEGADOCK does not require template structures for

prediction.

Hear, the PPI prediction scheme used in this study is reproduced below. First, we conducted rigid-body docking calculations based on a simplified energy function considering shape complementarity, electrostatics, and hydrophobic interactions for all possible binary combinations of proteins in the target set. Using this process, we obtained a group of high-scoring docking complexes for each pair of proteins. Next, we applied ZRANK [78] to the predicted complex structures for more advanced binding energy calculation and re-ranked the docking results based on ZRANK energy scores. The deviation of the selected docking scores from the score distribution of high-ranked complexes was determined as a standardized score (Z-score) and was used to assess possible interactions. This prediction protocol has been described in Chapter 4. Potential complexes that had no other high-scoring interactions nearby were rejected using structural differences. Thus, we considered likely binding pairs that had at least one populated area of high-scoring structures, one of which may be the true binding site.

9.2.3 Consensus prediction method

In this study, we proposed a new meta-prediction method by evaluating the consensus between both previously used prediction methods. The proposed method consists of two steps: (1) prediction from the same target set by PRISM and MEGADOCK and (2) consideration that the method provides a prediction regarding target protein pair interaction only when both PRISM and MEGADOCK predict that the target protein pair interacts. Although some true-positives may be dropped by this method, the remaining predicted pairs are expected to have higher reliability because of the consensus between two prediction methods that have different characteristics.

9.2.4 Dataset

In this section, we focused on the human apoptosis signaling pathway previously analyzed by PRISM [114] and MEGADOCK (same as Chapter 6, Table 6.2) because our prediction results can thus be compared directly to the results of the previous study. Table 1 shows the list of PDB IDs and chains of this dataset.

Known PPIs were collected from the STRING database [106]. We used only experimental data in the literature obtained from STRING with a confidence score > 0.5. The number of known PPIs was 137. Because the database does not contain existing self-interactions, we did not predict self-interactions. Thus, the number of target pairs was ${}_{57}\mathrm{C}_2 = 1,596$.

9.2.5 Evaluation of prediction performance

Here, we have defined #TP, #FP, #FN, #TN, precision, recall, and the F-measure, which we used to evaluate the prediction results: #TP is the number of predicted PPIs that were also found in STRING (true-positive), #FP is the number of predicted PPIs that were not in STRING (false-positive), #FN is the number of PPIs not predicted by the system even though the pair was found to interact in STRING (false-negative), and #TN is the number of negative predictions that were also not found in STRING (true-negative). To identify new PPIs in biological experiments after *in silico* screening, precision is more important than recall to reduce the cost of validation.

9.3 Results and Discussion

9.3.1 Comparison of template-based and *de novo* docking methods

Fig. 9.1 (a) and (b) show the prediction results for PRISM and MEGADOCK, respectively, as applied to a human apoptosis pathway. The threshold used for MEGADOCK prediction yielded the best value of the F-measure for this dataset. The diagonal line (black cells) in Fig. 9.1 indicates self-interactions that were not considered as prediction targets. As shown in Fig. 9.1, PRISM was performed with fewer FPs than MEGADOCK. Table 9.1 shows the evaluation of prediction results. With MEGADOCK, we obtained a lower value of precision and a higher value of recall relative to PRISM. When the F-measure was evaluated as a measure of overall performance, MEGADOCK showed lower values than PRISM. Predictions by MEGADOCK contained more FPs because, in contrast to PRISM, MEGADOCK does not restrict interface structures to those found in template structures. In contrast, PRISM obtained lower recall values than MEGADOCK because it only searched interactions whose interface structures could be found in the template set.

9.3.2 Results of the consensus prediction

Fig. 9.2 shows the Venn diagram of the number of TPs and FPs of the results of PRISM and MEGADOCK. A large difference was observed in the results obtained by the two methods. Thus, combining the prediction results of PRISM and MEGADOCK may provide better performance in PPI prediction. All of the predicted pairs of TPs and FPs are shown in Table 9.2.

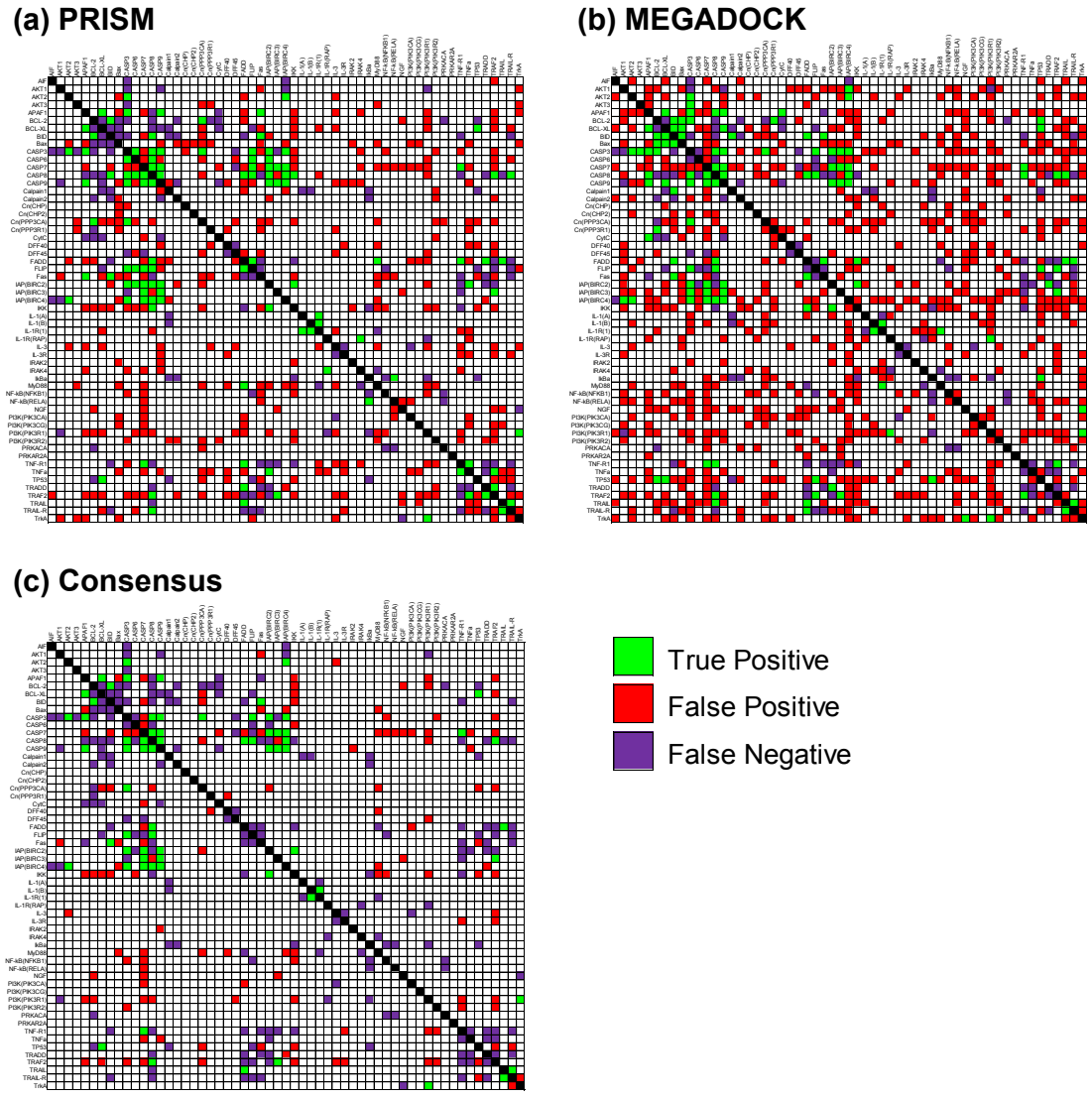


Figure 9.1: Apoptosis prediction by the (a) PRISM, (b) MEGADOCK, and (c) consensus methods. The green cells are true-positives, the red cells are false-positives, and the purple cells are false-negatives. The diagonal cells (black cells) have no PPI information in the STRING database and are excluded from the prediction targets.

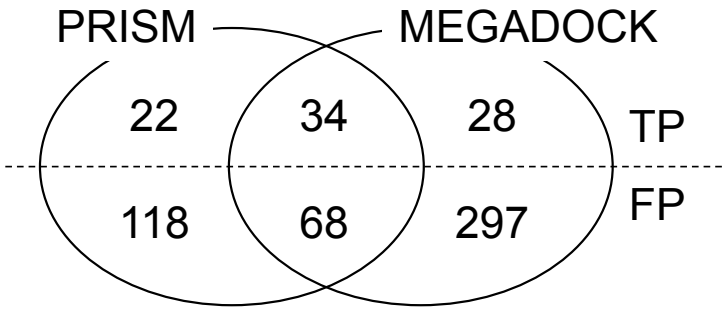


Figure 9.2: Venn diagram of apoptosis pathway prediction results. The common set (#TP=34, #FP=68) is denoted as "Consensus".

Table 9.1: Accuracy of human apoptosis pathway prediction

Method	#TP	#FP	#FN	#TN	Precision	Recall	F-measure
Consensus(AND)	34	68	103	1,391	0.333	0.248	0.285
OR	84	483	53	976	0.148	0.613	0.239
PRISM	56	186	81	1,273	0.231	0.409	0.296
MEGADOCK	62	365	75	1,094	0.145	0.453	0.220

Table 9.2: The list of all true-positive pairs and false-positive pairs predicted by the PRISM, MEGADOCK, and consensus methods; (a) the true-positive list of PRISM predictions, (b) the false-positive list of PRISM predictions, (c) the true-positive list of MEGADOCK predictions, (d) the false-positive list of MEGADOCK predictions, (e) the true-positive list of consensus predictions, and (f) the false-positive list of consensus predictions.

(a) The true-positive list of PRISM predictions

AKT2 – CASP3	CASP6 – FLIP	CASP9 – IAP(BIRC4)
AKT2 – IAP(BIRC4)	CASP6 – IAP(BIRC2)	FADD – FLIP
APAF1 – BCL-XL	CASP7 – CASP8	FADD – MyD88
APAF1 – CASP3	CASP7 – CASP9	FADD – TNF-R1
APAF1 – CASP9	CASP7 – FLIP	FADD – TRADD
APAF1 – Fas	CASP7 – IAP(BIRC2)	FADD – TRAIL
BCL-2 – BID	CASP7 – IAP(BIRC3)	IAP(BIRC2) – IKK
BCL-2 – Cn(PPP3CA)	CASP7 – IAP(BIRC4)	IAP(BIRC2) – TNFα
BCL-XL – TP53	CASP7 – TNF-R1	IAP(BIRC3) – TRAF2
BID – CASP8	CASP8 – CASP9	IL-1(A) – IL-1R(1)
BID – Fas	CASP8 – FADD	IL-1(B) – IL-1R(1)
CASP3 – CASP6	CASP8 – FLIP	IL-1R(RAP) – IRAK4
CASP3 – CASP8	CASP8 – IAP(BIRC2)	IκBα – NF-κB(RELA)
CASP3 – CASP9	CASP8 – IAP(BIRC4)	PI3K(PIK3R1) – TrkA
CASP3 – Cn(PPP3CA)	CASP8 – IKK	TNF-R1 – TNFα
CASP3 – FLIP	CASP8 – TRAF2	TNFα – TRAF2
CASP3 – IAP(BIRC2)	CASP8 – TRAIL-R	TRADD – TRAF2
CASP3 – IAP(BIRC4)	CASP9 – IAP(BIRC2)	TRAIL – TRAIL-R
CASP6 – CASP8	CASP9 – IAP(BIRC3)	

(b) The false-positive list of PRISM predictions

AIF – TRAF2	CASP6 – DFF45	DFF45 – TRAF2
AKT1 – Fas	CASP6 – IKK	FADD – PI3K(PIK3CA)
AKT1 – TrkA	CASP6 – PI3K(PIK3R1)	FADD – PI3K(PIK3CG)
AKT2 – Bax	CASP6 – TRAF2	FADD – PI3K(PIK3R1)
AKT2 – FADD	CASP6 – TRAIL-R	FADD – TNFα
AKT2 – IL-3	CASP7 – CytC	FLIP – TrkA
AKT3 – Cn(PPP3CA)	CASP7 – FADD	Fas – IKK
AKT3 – Cn(PPP3R1)	CASP7 – Fas	Fas – MyD88
AKT3 – MyD88	CASP7 – IKK	Fas – NF-κB(NFKB1)
AKT3 – TrkA	CASP7 – IRAK4	Fas – NF-κB(RELA)
APAF1 – Bax	CASP7 – MyD88	Fas – TP53
APAF1 – CASP7	CASP7 – NF-κB(NFKB1)	Fas – TRAF2
APAF1 – FADD	CASP7 – NF-κB(RELA)	IAP(BIRC2) – MyD88
APAF1 – IKK	CASP7 – NGF	IAP(BIRC2) – PI3K(PIK3R1)
APAF1 – IL-3	CASP7 – PI3K(PIK3CA)	IAP(BIRC3) – NGF
APAF1 – IRAK4	CASP7 – PI3K(PIK3CG)	IAP(BIRC4) – MyD88
APAF1 – PI3K(PIK3R1)	CASP7 – PI3K(PIK3R1)	IAP(BIRC4) – TRADD
APAF1 – TRAF2	CASP7 – TNFα	IKK – IRAK2
APAF1 – TrkA	CASP7 – TRAF2	IKK – MyD88
BCL-2 – FADD	CASP8 – Cn(PPP3CA)	IKK – NF-κB(NFKB1)
BCL-2 – IKK	CASP8 – IAP(BIRC3)	IKK – PI3K(PIK3R1)
BCL-2 – NGF	CASP8 – IL-1R(1)	IKK – PI3K(PIK3R2)
BCL-2 – PI3K(PIK3R1)	CASP8 – PI3K(PIK3R1)	IKK – TP53
BCL-2 – TRAF2	CASP9 – Calpain1	IKK – TRAF2
BCL-XL – Cn(PPP3CA)	CASP9 – DFF40	IL-1R(1) – IL-3
BCL-XL – Fas	CASP9 – FADD	IL-1R(1) – TNF-R1

Table 9.2 (continue)

BCL-XL – IKK	CASP9 – FLIP	IL-1R(1) – TNFα
BCL-XL – IL-1R(1)	CASP9 – IL-3	IL-1R(RAP) – TNFα
BCL-XL – IL-3	CASP9 – IL-3R	IL-3 – MyD88
BCL-XL – PI3K(PIK3R1)	CASP9 – IRAK2	IL-3 – PI3K(PIK3R1)
BCL-XL – TRAF2	CASP9 – IRAK4	IL-3 – TNF-R1
BCL-XL – TRAIL-R	CASP9 – TNFα	IL-3 – TRAF2
BID – CASP7	Calpain1 – FADD	IL-3 – TrkA
BID – Cn(PPP3CA)	Calpain1 – IAP(BIRC4)	IL-3R – TNF-R1
BID – IAP(BIRC2)	Calpain1 – PI3K(PIK3R2)	IL-3R – TNFα
BID – IKK	Calpain2 – IAP(BIRC2)	IL-3R – TRAF2
Bax – CASP9	Calpain2 – PI3K(PIK3R2)	IL-3R – TRAIL-R
Bax – Calpain2	Calpain2 – TRAF2	IRAK4 – TNF-R1
Bax – Cn(CHP)	Cn(PPP3CA) – DFF45	MyD88 – PI3K(PIK3R1)
Bax – Cn(CHP2)	Cn(PPP3CA) – Fas	MyD88 – PI3K(PIK3R2)
Bax – Cn(PPP3CA)	Cn(PPP3CA) – IAP(BIRC2)	NF-κB(NFKB1) – PI3K(PIK3R1)
Bax – Cn(PPP3R1)	Cn(PPP3CA) – IKK	NF-κB(RELA) – NGF
Bax – Fas	Cn(PPP3CA) – MyD88	NF-κB(RELA) – PI3K(PIK3CA)
Bax – IAP(BIRC4)	Cn(PPP3CA) – PI3K(PIK3R2)	NGF – TRAF2
Bax – IRAK2	Cn(PPP3CA) – PRKACA	NGF – TRAIL
Bax – MyD88	Cn(PPP3CA) – TNFα	PI3K(PIK3CA) – TNFα
Bax – PI3K(PIK3CA)	Cn(PPP3CA) – TP53	PI3K(PIK3CG) – TNFα
Bax – PI3K(PIK3CG)	Cn(PPP3CA) – TRAF2	PI3K(PIK3CG) – TRAF2
Bax – TNFα	Cn(PPP3R1) – DFF40	PI3K(PIK3CG) – TRAIL
Bax – TrkA	Cn(PPP3R1) – Fas	PI3K(PIK3R1) – TNF-R1
CASP3 – CASP7	Cn(PPP3R1) – IL-3	PI3K(PIK3R1) – TRAF2
CASP3 – Cn(CHP)	Cn(PPP3R1) – TNFα	PI3K(PIK3R2) – TNF-R1
CASP3 – Fas	CytC – PI3K(PIK3R2)	PI3K(PIK3R2) – TRADD
CASP3 – IKK	DFF40 – IL-3R	PI3K(PIK3R2) – TRAF2
CASP3 – IRAK4	DFF40 – MyD88	TNFα – TRAIL
CASP3 – NF-κB(NFKB1)	DFF40 – TNFα	TNFα – TRAIL-R
CASP3 – PI3K(PIK3R2)	DFF40 – TP53	TP53 – TRAF2
CASP3 – TNF-R1	DFF40 – TRAF2	TP53 – TRAIL
CASP3 – TRAF2	DFF45 – FADD	TP53 – TRAIL-R
CASP6 – CASP7	DFF45 – Fas	TRAF2 – TRAIL-R
CASP6 – CASP9	DFF45 – IL-3	TRAF2 – TrkA
CASP6 – Cn(PPP3CA)	DFF45 – PI3K(PIK3R1)	TRAIL-R – TrkA

(c) The true-positive list of MEGADOCK predictions

AKT1 – IAP(BIRC4)	BID – Fas	CASP9 – IAP(BIRC2)
AKT2 – CASP3	CASP3 – CASP8	CASP9 – IAP(BIRC3)
AKT2 – IAP(BIRC4)	CASP3 – CASP9	CASP9 – IAP(BIRC4)
AKT3 – CASP3	CASP3 – Cn(PPP3CA)	FADD – Fas
APAF1 – BCL-XL	CASP3 – FLIP	FADD – IKK
APAF1 – CASP3	CASP3 – IAP(BIRC2)	FADD – TRAIL
APAF1 – CASP8	CASP3 – IAP(BIRC3)	FADD – TRAIL-R
APAF1 – CASP9	CASP3 – IAP(BIRC4)	FLIP – TP53
APAF1 – CytC	CASP7 – CASP8	FLIP – TRAF2
BCL-2 – BID	CASP7 – CASP9	IAP(BIRC2) – IKK
BCL-2 – Bax	CASP7 – IAP(BIRC2)	IAP(BIRC2) – TRADD
BCL-2 – CASP3	CASP7 – IAP(BIRC3)	IAP(BIRC3) – TRAF2
BCL-2 – Cn(PPP3R1)	CASP7 – IAP(BIRC4)	IL-1(B) – IL-1R(1)
BCL-2 – TP53	CASP7 – TNF-R1	IL-1R(1) – MyD88
BCL-XL – BID	CASP8 – CASP9	IL-1R(RAP) – PI3K(PIK3R1)
BCL-XL – Bax	CASP8 – FADD	IκBα – TP53
BCL-XL – CASP9	CASP8 – FLIP	NGF – TrkA
BCL-XL – TP53	CASP8 – IAP(BIRC4)	PI3K(PIK3R1) – TrkA

Table 9.2 (continue)

BID – Bax	CASP8 – TNF-R1	TNF-R1 – TRAF2
BID – Calpain1	CASP8 – TRAF2	TRAIL – TRAIL-R
BID – FADD	CASP8 – TRAIL	

(d) The false-positive list of MEGADOCK predictions

AIF – AKT2	CASP3 – PI3K(PIK3R1)	FADD – TP53
AIF – BCL-XL	CASP3 – PI3K(PIK3R2)	FLIP – IAP(BIRC4)
AIF – CASP7	CASP3 – TNFα	FLIP – IKK
AIF – Calpain2	CASP3 – TP53	FLIP – IL-1R(RAP)
AIF – PI3K(PIK3R1)	CASP3 – TRAF2	FLIP – NGF
AIF – TP53	CASP3 – TRAIL	FLIP – PI3K(PIK3R2)
AIF – TRAIL	CASP3 – TRAIL-R	Fas – IAP(BIRC3)
AIF – TrkA	CASP3 – TrkA	Fas – PRKACA
AKT1 – APAF1	CASP6 – CASP7	Fas – TP53
AKT1 – BCL-2	CASP6 – IAP(BIRC3)	Fas – TRAIL
AKT1 – Bax	CASP6 – IAP(BIRC4)	IAP(BIRC2) – IL-1(A)
AKT1 – CASP8	CASP6 – IKK	IAP(BIRC2) – IL-3
AKT1 – DFF40	CASP6 – TRAIL	IAP(BIRC2) – IRAK2
AKT1 – DFF45	CASP7 – Calpain1	IAP(BIRC2) – NF-κB(NFKB1)
AKT1 – FADD	CASP7 – Calpain2	IAP(BIRC2) – NGF
AKT1 – Fas	CASP7 – Cn(CHP2)	IAP(BIRC2) – TRAIL-R
AKT1 – IAP(BIRC2)	CASP7 – Cn(PPP3R1)	IAP(BIRC3) – NF-κB(RELA)
AKT1 – IAP(BIRC3)	CASP7 – CytC	IAP(BIRC3) – NGF
AKT1 – IKK	CASP7 – FADD	IAP(BIRC3) – PI3K(PIK3R2)
AKT1 – IL-3R	CASP7 – Fas	IAP(BIRC3) – PRKACA
AKT1 – NF-κB(NFKB1)	CASP7 – IL-1(B)	IAP(BIRC3) – PRKAR2A
AKT1 – NF-κB(RELA)	CASP7 – IκBα	IAP(BIRC3) – TP53
AKT1 – PI3K(PIK3CG)	CASP7 – MyD88	IAP(BIRC3) – TRAIL-R
AKT1 – PI3K(PIK3R2)	CASP7 – NF-κB(NFKB1)	IAP(BIRC4) – IKK
AKT1 – TRAF2	CASP7 – NF-κB(RELA)	IAP(BIRC4) – IL-1(B)
AKT2 – APAF1	CASP7 – NGF	IAP(BIRC4) – IL-1R(1)
AKT2 – BCL-XL	CASP7 – PI3K(PIK3CA)	IAP(BIRC4) – IL-1R(RAP)
AKT2 – CASP9	CASP7 – PI3K(PIK3R1)	IAP(BIRC4) – IL-3R
AKT2 – Calpain2	CASP7 – PI3K(PIK3R2)	IAP(BIRC4) – IRAK2
AKT2 – Cn(CHP2)	CASP7 – PRKACA	IAP(BIRC4) – IκBα
AKT2 – FLIP	CASP7 – TNFα	IAP(BIRC4) – MyD88
AKT2 – IAP(BIRC3)	CASP7 – TP53	IAP(BIRC4) – NF-κB(NFKB1)
AKT2 – IL-1(B)	CASP7 – TRADD	IAP(BIRC4) – PI3K(PIK3CA)
AKT2 – IL-3	CASP7 – TRAF2	IAP(BIRC4) – PI3K(PIK3CG)
AKT2 – MyD88	CASP7 – TRAIL	IAP(BIRC4) – PI3K(PIK3R1)
AKT2 – NF-κB(RELA)	CASP7 – TrkA	IAP(BIRC4) – PI3K(PIK3R2)
AKT2 – PI3K(PIK3CA)	CASP8 – Cn(CHP2)	IAP(BIRC4) – TNF-R1
AKT2 – PI3K(PIK3R1)	CASP8 – Cn(PPP3CA)	IAP(BIRC4) – TP53
AKT2 – TP53	CASP8 – DFF45	IAP(BIRC4) – TRADD
AKT2 – TRAF2	CASP8 – IAP(BIRC3)	IAP(BIRC4) – TRAF2
AKT2 – TRAIL-R	CASP8 – MyD88	IAP(BIRC4) – TRAIL
AKT2 – TrkA	CASP8 – NGF	IAP(BIRC4) – TRAIL-R
AKT3 – APAF1	CASP8 – PI3K(PIK3R1)	IAP(BIRC4) – TrkA
AKT3 – CASP7	CASP8 – TP53	IKK – IL-3
AKT3 – CASP9	CASP8 – TrkA	IKK – IRAK4
AKT3 – FADD	CASP9 – Cn(PPP3R1)	IKK – MyD88
APAF1 – Bax	CASP9 – IL-1R(RAP)	IKK – NF-κB(NFKB1)
APAF1 – CASP6	CASP9 – IRAK2	IKK – NF-κB(RELA)
APAF1 – CASP7	CASP9 – IκBα	IKK – PI3K(PIK3R1)
APAF1 – IAP(BIRC3)	CASP9 – PI3K(PIK3R1)	IKK – TP53
APAF1 – IAP(BIRC4)	CASP9 – PI3K(PIK3R2)	IKK – TRADD

Table 9.2 (continue)

APAF1 – IKK	CASP9 – TNFα	IKK – TRAF2
APAF1 – IL-1R(RAP)	CASP9 – TRAF2	IKK – TRAIL
APAF1 – NF-κB(NFKB1)	CASP9 – TRAIL	IKK – TrkA
APAF1 – NF-κB(RELA)	Calpain1 – IAP(BIRC3)	IL-1(A) – IL-1R(RAP)
APAF1 – NGF	Calpain1 – NGF	IL-1(A) – PI3K(PIK3CA)
APAF1 – PI3K(PIK3CG)	Calpain2 – IAP(BIRC4)	IL-1(A) – PI3K(PIK3R1)
APAF1 – PI3K(PIK3R1)	Calpain2 – IKK	IL-1(A) – TP53
APAF1 – PRKAR2A	Calpain2 – IL-1(B)	IL-1(A) – TRAIL
APAF1 – TNFα	Calpain2 – NGF	IL-1(B) – IRAK4
APAF1 – TP53	Calpain2 – PRKACA	IL-1(B) – PI3K(PIK3CG)
APAF1 – TRAF2	Calpain2 – TP53	IL-1(B) – PI3K(PIK3R1)
BCL-2 – DFF45	Calpain2 – TRAIL-R	IL-1(B) – TP53
BCL-2 – IAP(BIRC4)	Calpain2 – TrkA	IL-1(B) – TRAIL-R
BCL-2 – IKK	Cn(CHP) – IKK	IL-1R(1) – IRAK2
BCL-2 – IL-1R(1)	Cn(CHP) – IL-1R(1)	IL-1R(1) – IRAK4
BCL-2 – NGF	Cn(CHP) – IRAK4	IL-1R(1) – IκBα
BCL-2 – PI3K(PIK3R1)	Cn(CHP) – NF-κB(NFKB1)	IL-1R(1) – PI3K(PIK3R1)
BCL-2 – TRAIL	Cn(CHP) – PI3K(PIK3CG)	IL-1R(RAP) – IκBα
BCL-2 – TRAIL-R	Cn(CHP2) – IAP(BIRC4)	IL-1R(RAP) – TP53
BCL-XL – CASP6	Cn(CHP2) – IL-1R(RAP)	IL-3 – NF-κB(NFKB1)
BCL-XL – CASP7	Cn(CHP2) – NF-κB(NFKB1)	IL-3 – PI3K(PIK3R2)
BCL-XL – Cn(CHP)	Cn(CHP2) – NGF	IL-3 – TNFα
BCL-XL – Cn(PPP3CA)	Cn(CHP2) – PI3K(PIK3CA)	IL-3 – TRAF2
BCL-XL – IAP(BIRC2)	Cn(CHP2) – PRKACA	IL-3R – NGF
BCL-XL – IAP(BIRC3)	Cn(PPP3CA) – Cn(PPP3R1)	IL-3R – TNF-R1
BCL-XL – IKK	Cn(PPP3CA) – FADD	IL-3R – TRAF2
BCL-XL – NF-κB(NFKB1)	Cn(PPP3CA) – IAP(BIRC2)	IL-3R – TrkA
BCL-XL – NGF	Cn(PPP3CA) – IAP(BIRC3)	IRAK2 – NF-κB(NFKB1)
BCL-XL – TRAF2	Cn(PPP3CA) – IL-1(A)	IRAK2 – TRAF2
BCL-XL – TRAIL	Cn(PPP3CA) – IL-1(B)	IRAK4 – PI3K(PIK3R2)
BCL-XL – TrkA	Cn(PPP3CA) – IL-1R(1)	IRAK4 – TP53
BID – CASP7	Cn(PPP3CA) – IRAK4	IRAK4 – TRAF2
BID – Cn(CHP2)	Cn(PPP3CA) – IκBα	IRAK4 – TrkA
BID – Cn(PPP3CA)	Cn(PPP3CA) – MyD88	IκBα – PI3K(PIK3CA)
BID – Cn(PPP3R1)	Cn(PPP3CA) – NGF	IκBα – TrkA
BID – IKK	Cn(PPP3CA) – PI3K(PIK3CA)	MyD88 – TP53
BID – IRAK4	Cn(PPP3CA) – PI3K(PIK3CG)	MyD88 – TrkA
BID – MyD88	Cn(PPP3CA) – TP53	NF-κB(NFKB1) – NF-κB(RELA)
BID – NF-κB(RELA)	Cn(PPP3CA) – TRAF2	NF-κB(NFKB1) – TNFα
BID – PI3K(PIK3CA)	Cn(PPP3CA) – TRAIL	NF-κB(RELA) – TNFα
BID – TP53	Cn(PPP3CA) – TrkA	NF-κB(RELA) – TP53
BID – TRAIL	Cn(PPP3R1) – CytC	NF-κB(RELA) – TRAF2
BID – TrkA	Cn(PPP3R1) – DFF40	NF-κB(RELA) – TRAIL
Bax – CASP7	Cn(PPP3R1) – IAP(BIRC4)	NF-κB(RELA) – TrkA
Bax – CASP8	Cn(PPP3R1) – IKK	NGF – PI3K(PIK3R1)
Bax – CytC	Cn(PPP3R1) – IL-1(B)	NGF – TP53
Bax – Fas	Cn(PPP3R1) – IκBα	PI3K(PIK3CA) – PI3K(PIK3CG)
Bax – IAP(BIRC2)	Cn(PPP3R1) – NF-κB(NFKB1)	PI3K(PIK3CA) – PI3K(PIK3R1)
Bax – IAP(BIRC3)	Cn(PPP3R1) – PI3K(PIK3CA)	PI3K(PIK3CA) – TP53
Bax – IAP(BIRC4)	Cn(PPP3R1) – TNF-R1	PI3K(PIK3CA) – TRAF2
Bax – IL-1R(RAP)	Cn(PPP3R1) – TP53	PI3K(PIK3CA) – TrkA
Bax – IL-3	Cn(PPP3R1) – TRAF2	PI3K(PIK3R1) – PI3K(PIK3R2)
Bax – MyD88	Cn(PPP3R1) – TrkA	PI3K(PIK3R1) – TNF-R1
Bax – NF-κB(RELA)	CytC – FADD	PI3K(PIK3R1) – TNFα
Bax – PI3K(PIK3R1)	CytC – TP53	PI3K(PIK3R1) – TP53
Bax – TNF-R1	DFF40 – IL-3	PI3K(PIK3R1) – TRADD

Table 9.2 (continue)

Bax – TP53	DFF40 – MyD88	PI3K(PIK3R1) – TRAF2
Bax – TRADD	DFF40 – PI3K(PIK3CA)	PI3K(PIK3R1) – TRAIL
Bax – TRAIL-R	DFF40 – PI3K(PIK3R1)	PI3K(PIK3R2) – TNF-R1
CASP3 – CASP7	DFF40 – TrkA	PI3K(PIK3R2) – TRAF2
CASP3 – Calpain2	DFF45 – IAP(BIRC2)	PI3K(PIK3R2) – TRAIL
CASP3 – Cn(CHP2)	DFF45 – NF-κB(NFKB1)	PI3K(PIK3R2) – TrkA
CASP3 – Cn(PPP3R1)	DFF45 – NGF	PRKACA – TRAIL-R
CASP3 – IL-1(A)	DFF45 – PI3K(PIK3CA)	TNFα – TP53
CASP3 – IL-1(B)	DFF45 – PI3K(PIK3R1)	TNFα – TrkA
CASP3 – IL-1R(1)	DFF45 – PRKACA	TP53 – TRAF2
CASP3 – IRAK2	FADD – IAP(BIRC4)	TP53 – TRAIL-R
CASP3 – NF-κB(NFKB1)	FADD – IL-1(A)	TRAF2 – TRAIL
CASP3 – NF-κB(RELA)	FADD – NGF	TRAIL – TrkA
CASP3 – NGF	FADD – PI3K(PIK3CA)	TRAIL-R – TrkA
CASP3 – PI3K(PIK3CA)	FADD – PRKAR2A	

(e) The true-positive list of Consensus predictions

AKT2 – CASP3	CASP7 – TNF-R1	IAP(BIRC2) – CASP9
AKT2 – IAP(BIRC4)	CASP8 – CASP3	IAP(BIRC2) – IKK
APAF1 – BCL-XL	CASP8 – CASP7	IAP(BIRC3) – CASP7
APAF1 – CASP3	CASP8 – CASP9	IAP(BIRC3) – CASP9
APAF1 – CASP9	CASP8 – FADD	IAP(BIRC3) – TRAF2
BCL-2 – BID	CASP8 – FLIP	IAP(BIRC4) – AKT2
BCL-XL – APAF1	CASP8 – IAP(BIRC4)	IAP(BIRC4) – CASP3
BCL-XL – TP53	CASP8 – TRAF2	IAP(BIRC4) – CASP7
BID – BCL-2	CASP9 – APAF1	IAP(BIRC4) – CASP8
BID – Fas	CASP9 – CASP3	IAP(BIRC4) – CASP9
CASP3 – AKT2	CASP9 – CASP7	IKK – IAP(BIRC2)
CASP3 – APAF1	CASP9 – CASP8	IL-1(B) – IL-1R(1)
CASP3 – CASP8	CASP9 – IAP(BIRC2)	IL-1R(1) – IL-1(B)
CASP3 – CASP9	CASP9 – IAP(BIRC3)	PI3K(PIK3R1) – TrkA
CASP3 – Cn(PPP3CA)	CASP9 – IAP(BIRC4)	TNF-R1 – CASP7
CASP3 – FLIP	Cn(PPP3CA) – CASP3	TP53 – BCL-XL
CASP3 – IAP(BIRC2)	FADD – CASP8	TRAF2 – CASP8
CASP3 – IAP(BIRC4)	FADD – TRAIL	TRAF2 – IAP(BIRC3)
CASP7 – CASP8	FLIP – CASP3	TRAIL – FADD
CASP7 – CASP9	FLIP – CASP8	TRAIL – TRAIL-R
CASP7 – IAP(BIRC2)	Fas – BID	TRAIL-R – TRAIL
CASP7 – IAP(BIRC3)	IAP(BIRC2) – CASP3	TrkA – PI3K(PIK3R1)
CASP7 – IAP(BIRC4)	IAP(BIRC2) – CASP7	

(f) The false-positive list of Consensus predictions

AKT1 – Fas	Cn(PPP3CA) – BCL-XL	NF-κB(NFKB1) – CASP7
AKT2 – IL-3	Cn(PPP3CA) – BID	NF-κB(NFKB1) – IKK
APAF1 – Bax	Cn(PPP3CA) – CASP8	NF-κB(RELA) – CASP7
APAF1 – CASP7	Cn(PPP3CA) – IAP(BIRC2)	NGF – BCL-2
APAF1 – IKK	Cn(PPP3CA) – MyD88	NGF – CASP7
APAF1 – PI3K(PIK3R1)	Cn(PPP3CA) – TP53	NGF – IAP(BIRC3)
APAF1 – TRAF2	Cn(PPP3CA) – TRAF2	PI3K(PIK3CA) – CASP7
BCL-2 – IKK	Cn(PPP3R1) – DFF40	PI3K(PIK3CA) – FADD
BCL-2 – NGF	CytC – CASP7	PI3K(PIK3R1) – APAF1
BCL-2 – PI3K(PIK3R1)	DFF40 – Cn(PPP3R1)	PI3K(PIK3R1) – BCL-2
BCL-XL – Cn(PPP3CA)	DFF40 – MyD88	PI3K(PIK3R1) – CASP7
BCL-XL – IKK	DFF45 – PI3K(PIK3R1)	PI3K(PIK3R1) – CASP8
BCL-XL – TRAF2	FADD – CASP7	PI3K(PIK3R1) – DFF45
BID – CASP7	FADD – PI3K(PIK3CA)	PI3K(PIK3R1) – IKK

Table 9.2 (continue)

BID – Cn(PPP3CA)	Fas – AKT1	PI3K(PIK3R1) – TNF-R1
BID – IKK	Fas – Bax	PI3K(PIK3R1) – TRAF2
Bax – APAF1	Fas – CASP7	PI3K(PIK3R2) – CASP3
Bax – Fas	Fas – TP53	PI3K(PIK3R2) – TNF-R1
Bax – IAP(BIRC4)	IAP(BIRC2) – Cn(PPP3CA)	PI3K(PIK3R2) – TRAF2
Bax – MyD88	IAP(BIRC3) – CASP8	TNF-R1 – IL-3R
CASP3 – CASP7	IAP(BIRC3) – NGF	TNF-R1 – PI3K(PIK3R1)
CASP3 – NF-κB(NFKB1)	IAP(BIRC4) – Bax	TNF-R1 – PI3K(PIK3R2)
CASP3 – PI3K(PIK3R2)	IAP(BIRC4) – MyD88	TNFα – CASP7
CASP3 – TRAF2	IAP(BIRC4) – TRADD	TNFα – CASP9
CASP6 – CASP7	IKK – APAF1	TP53 – Cn(PPP3CA)
CASP6 – IKK	IKK – BCL-2	TP53 – Fas
CASP7 – APAF1	IKK – BCL-XL	TP53 – IKK
CASP7 – BID	IKK – BID	TP53 – TRAF2
CASP7 – CASP3	IKK – CASP6	TP53 – TRAIL-R
CASP7 – CASP6	IKK – MyD88	TRADD – IAP(BIRC4)
CASP7 – CytC	IKK – NF-κB(NFKB1)	TRAF2 – APAF1
CASP7 – FADD	IKK – PI3K(PIK3R1)	TRAF2 – BCL-XL
CASP7 – Fas	IKK – TP53	TRAF2 – CASP3
CASP7 – MyD88	IKK – TRAF2	TRAF2 – CASP7
CASP7 – NF-κB(NFKB1)	IL-3 – AKT2	TRAF2 – Cn(PPP3CA)
CASP7 – NF-κB(RELA)	IL-3 – TRAF2	TRAF2 – IKK
CASP7 – NGF	IL-3R – TNF-R1	TRAF2 – IL-3
CASP7 – PI3K(PIK3CA)	IL-3R – TRAF2	TRAF2 – IL-3R
CASP7 – PI3K(PIK3R1)	IRAK2 – CASP9	TRAF2 – PI3K(PIK3R1)
CASP7 – TNFα	MyD88 – Bax	TRAF2 – PI3K(PIK3R2)
CASP7 – TRAF2	MyD88 – CASP7	TRAF2 – TP53
CASP8 – Cn(PPP3CA)	MyD88 – Cn(PPP3CA)	TRAIL-R – TP53
CASP8 – IAP(BIRC3)	MyD88 – DFF40	TRAIL-R – TrkA
CASP8 – PI3K(PIK3R1)	MyD88 – IAP(BIRC4)	TrkA – TRAIL-R
CASP9 – IRAK2	MyD88 – IKK	
CASP9 – TNFα	NF-κB(NFKB1) – CASP3	

Note: The abbreviations used are: AIF, apoptosis-inducing factor, mitochondrion-associated, 1 (AIFM1); AKT1, RACalpha serine/threonine-protein kinase; AKT2, RAC-beta serine/threonine-protein kinase; AKT3, RAC-gamma serine/threonine-protein kinase; APAF1, apoptotic peptidase activating factor 1; BCL-2, B-cell lymphoma 2; BCL-XL, BCL extra-large; BID, BH3 interacting domain death agonist; Bax, BCL-2-associated X protein; CASP3/6/7/8/9, caspase-3/6/7/8/9; Cn(CHP), calcineurin B homologous protein 1; Cn(CHP2), calcineurin B homologous protein 2; Cn(PPP3CA), protein phosphatase 3 catalytic subunit alpha isoform; Cn(PPP3R1), protein phosphatase 3 regulatory subunit 1; CytC, cytochrome C; DFF40, DNA fragmentation factor, 40kDa, beta polypeptide; DFF45, DNA fragmentation factor, 45kDa, alpha polypeptide; FADD, Fas-associated via death domain; FLIP, FLICE/CASP8 inhibitory protein (CASP8 and FADD-like apoptosis regulator, CFLAR); Fas, tumor necrosis factor receptor (TNF) superfamily member 6; IAP, inhibitor of apoptosis; BIRC2/3/4, baculoviral IAP repeat-containing protein 2/3/4; IκBα, nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor alpha; IKK, inhibitor of nuclear factor kappa-B kinase; IL-1(A), interleukin-1 alpha; IL-1(B), interleukin-1 beta; IL-1R(1), type 1 interleukin-1 receptor; IL- 1R(RAP), interleukin-1 receptor accessory protein; IL-3, interleukin-3; IL-3R, interleukin-3 receptor; IRAK2/4, interleukin-1 receptor-associated kinase 2/4; MyD88, myeloid differentiation primary response protein MyD88; NF-κB(NFKB1), nuclear factor of kappa light polypeptide gene enhancer in B-cells; NF-κB(RELA), nuclear factor of kappa light polypeptide gene enhancer in B-cells 3; NGF, nerve growth factor (beta polypeptide); PI3K, phosphatidylinositide 3-kinase; PIK3CA, PI3K subunit alpha; PIK3CG, PI3K subunit gamma; PIK3R1, PI3K regulatory subunit alpha; PIK3R2, PI3K regulatory subunit beta; PRKACA, cyclic adenosine monophosphate (cAMP)-dependent protein kinase catalytic subunit alpha; PRKAR2A, cAMP-dependent protein kinase type II-alpha regulatory subunit; TNFα, tumor necrosis factor; TNF-R1, TNF receptor superfamily member 1A; TP53, cellular tumor antigen p53; TRADD, TNF receptor type 1-associated death domain protein; TRAF2, TNF receptor-associated factor 2; TRAIL, TNF receptor superfamily member 10; TRAIL-R, TNF receptor superfamily member 10B; TrkA, neurotrophic tyrosine kinase receptor type 1.

Fig. 9.1 (c) shows the prediction obtained on consensus between PRISM (a) and MEGADOCK (b); notably, the number of FP samples greatly decreased. The first row of Table 9.1 shows that the consensus method obtained an F-measure value of 0.285, which was comparable to the PRISM result (F-measure = 0.296). The consensus prediction indicated a higher value of precision for the consensus method (0.333) than for PRISM (0.231). The consensus method yielded the highest precision value in the method shown in Table 9.1. This method is useful when validating unknown PPI predictions using biological experiments. In contrast, OR prediction demonstrated high recall (Table 9.1). Thus, the OR method will be useful when prediction with high sensitivity, e.g., in the initial construction of the draft PPI network from the relevant proteins, is required.

9.3.3 Relationship between the number of predicted positives and the number of structures

The structure-based PPI prediction method may generate positives with some bias regarding the type of proteins (rows and columns of Fig. 9.1). From Table 1 and Fig. 9.1, predictions with a large number of protein structures tend to generate more positive pairs. To verify this tendency, the number of PDB chain structures used for PPI prediction and the number of positive predicted pairs containing its protein are plotted in Fig. 9.3. The #TPs are shown in Fig. 9.3 (a) and the #FPs are shown in Fig. 9.3 (b). Pearson's correlation coefficient R and the P-value for the correlation coefficient t-test are shown in Table 9.3.

From the results of the t-tests, the number of chains and the number of positive predictions were clearly correlated with $P < 0.05$ in all cases, which suggests that the structure-based PPI prediction method should address the number of used protein structures without bias. For example, in a template matching-based method such as PRISM, a protein pair with more conformations of structures will have more matches in template complexes and a higher possibility of predicted interaction. In Table 9.3, the correlation coefficient values are particularly high in FP predictions. Therefore, for more precise prediction, we should consider one of the two ways: (i) how to generate the target set without multiple conformations in each protein and (ii) develop a correction method when the target set contains multiple conformations.

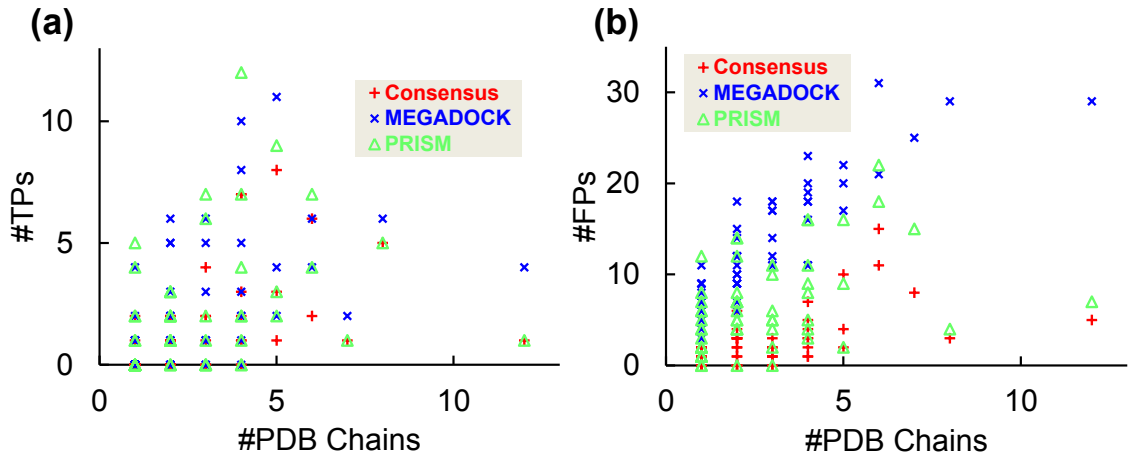


Figure 9.3: Number of PDB chains vs. positive predictions. (a) Shows the number of true-positives and (b) shows the number of false-positives. The horizontal axis is the number of PDB chains used in the interaction prediction, and the vertical axis is the number of positives predicted by using protein structures.

Table 9.3: Pearson's correlation coefficient R and P-value of correlation test on Fig. 9.3

Method	(a) #TPs		(b) #FPs	
	R	P-value	R	P-value
Consensus	0.477	1.784×10^{-4}	0.594	1.121×10^{-6}
PRISM	0.342	9.259×10^{-3}	0.415	1.316×10^{-3}
MEGADOCK	0.488	1.167×10^{-4}	0.864	4.602×10^{-18}

9.3.4 Performance evaluation with various sensitivity parameters

In this study, we used a fixed threshold value for MEGADOCK that provided the best F-measure value for the target dataset. Fig. 9.4 shows a plot of precision vs. F-measure value for prediction results with various threshold values for MEGADOCK. Fig. 9.4 also plots the performance of the consensus method with various threshold values for MEGADOCK prediction while the threshold value for PRISM prediction was fixed. When the threshold value was changed in MEGADOCK, the plotted values remained in the region of low precision (0.0–0.2), and lower F-measure values were observed in the region of higher precision because of the decreased recall value. The consensus prediction method maintained a stable F-measure value when the value of precision was approximately 0.2–0.3, although the performance in the high-precision region (> 0.4) was inferior to that of MEGADOCK. In this region, the consensus prediction provides a better precision value than PRISM while maintaining the same F-measure value. Fig. 9.4 clearly shows that the performance obtained by using the consensus method is better over a wide range of threshold values than the prediction obtained using only MEGADOCK.

The AUC, i.e., the area under the ROC curve [85], is a more general and effective statistical measure. The $\mathrm{ROC}_{0.1}$ curves, which include the ROC curves up to an FP rate of 0.1, are shown in Fig. 9.5. ROC curves were created by plotting the TP rate (#TP/(#TP+#FN)) against the FP rate (#FP/(#FP+#TN)). Regions with high FP rates are not useful for prediction because many FPs are generated, e.g., an FP rate of 0.2 represents #FP = 292. The $\mathrm{ROC}_{0.1}$ curve was thus considered to favor methods that produce a high TP rate at low FP rates, and the associated area under the curve is referred to as $\mathrm{AUC}_{0.1}$. A perfect prediction will produce an $\mathrm{AUC}_{0.1}$ of $(0.1 \times 1 =)$ 0.1, whereas a random prediction will result in an $\mathrm{AUC}_{0.1}$ of $(0.1 \times 0.1/2 =)$ 0.005. Fig. 9.5 shows that the consensus prediction ($\mathrm{AUC}_{0.1} = 0.023$) is better than the MEGADOCK ($\mathrm{AUC}_{0.1} = 0.014$) and random predictions ($\mathrm{AUC}_{0.1} = 0.005$).

9.4 Summary

In this study, we propose a new PPI network prediction method based on the consensus between template-based prediction and non-template-based prediction. The consensus method successfully predicted the PPI network more accurately than the conventional single template/non-template method. Because such precise prediction

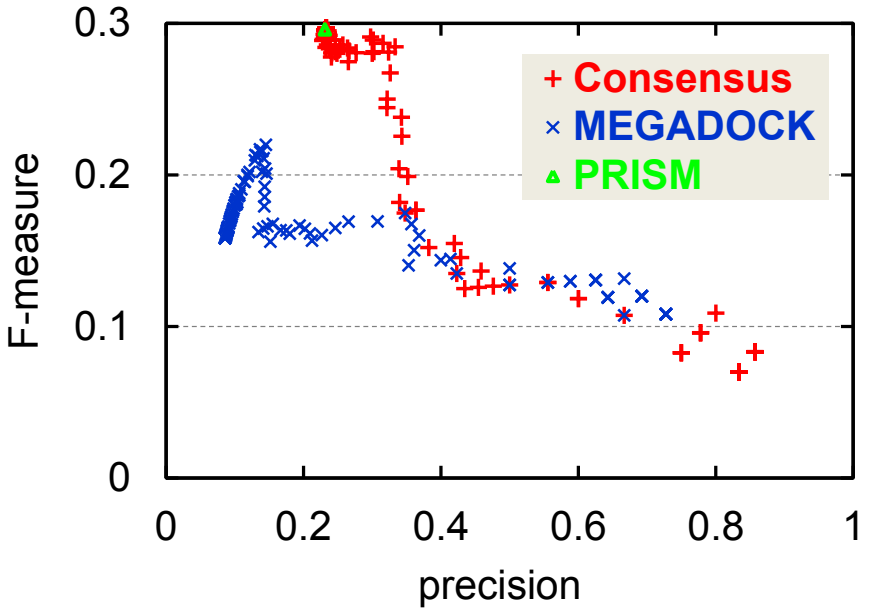


Figure 9.4: F-measure vs. precision for predictions when the MEGADOCK threshold parameter is changed in the apoptosis pathway prediction. The green triangle indicates the results of the PRISM prediction (Table 9.1).

can reduce biological screening costs, it will promote interactome analysis. For further improvement of prediction performance, it is necessary to further improve the combination of the two techniques, e.g., by using a strategy other than taking a simple AND/OR consensus. For example, biological information such as biochemical function and subcellular localization information could be used.

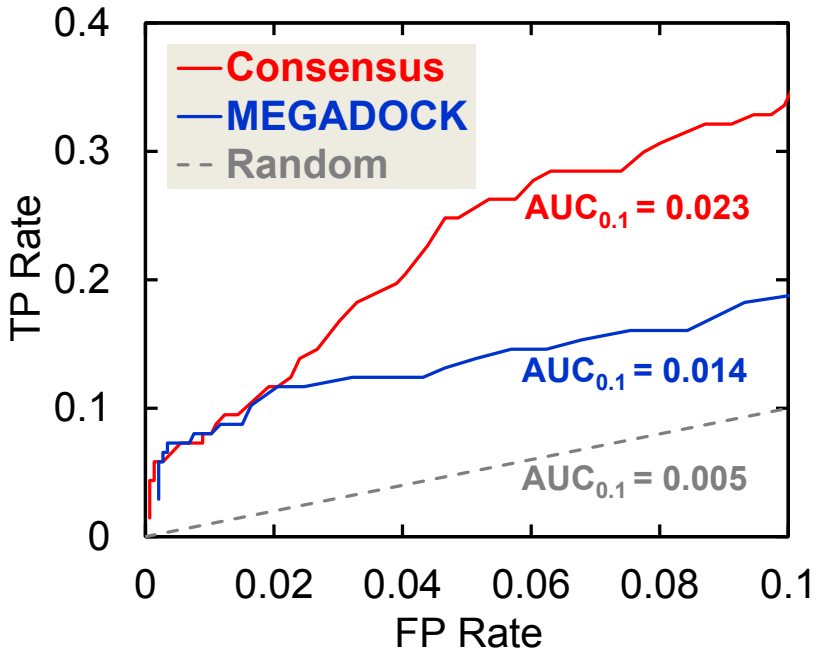


Figure 9.5: $ROC_{0.1}$ curves obtained when the MEGADOCK threshold parameter is changed in the apoptosis pathway prediction. $AUC_{0.1}$ is the area under the $ROC_{0.1}$ curve. For the 0–0.1 FP rate range here, a random prediction produced an $AUC_{0.1}$ of 0.005.

Part V

Concluding Remarks

Chapter 10

Conclusion

10.1 Conclusion

In this thesis, we document a successful PPI network prediction system, MEGADOCK, based on protein tertiary structures. Our scoring function represented three physico-chemical interaction properties with acceptable accuracy and, as the main result of this work, decupled the calculation speed when compared with other methods. In addition, our GPU and parallel implementation achieved 37.0-fold acceleration using one computing node with three GPUs and worked in high-performance computing environments equipped with over ten thousands nodes ($\sim$ 25,000 nodes). Based on the results, MEGADOCK has achieved over 1,000,000-fold acceleration and has made it possible to predict megaorder PPIs. Below, we describe the contributions provided by this work.

10.1.1 Contributions

- In Chapter 3, we proposed a novel shape complementarity score function called real Pairwise Shape Complementarity (rPSC) for FFT-based rigid-body protein–protein docking calculations. The rPSC function that uses only real number representations for shape complementarity was correlated with a conventional score function represented by a complex number. We also proposed a novel desolvation free energy function called Receptor Desolvation Free Energy (RDE). Therefore, it is possible to calculate a total energy score that includes shape complementarity, electrostatic interactions and desolvation effects with only one FFT correlation. As a result, the proposed method was shown to be 9.8 times faster than the conventional tool ZDOCK 3.0 while maintaining acceptable docking

prediction accuracies.

- In Chapter 4, we developed MEGADOCK, an exhaustive PPI screening system, that conducts protein–protein docking and post-analysis with reranking technique on protein tertiary structural data. For the detection of the relevant interacting protein pairs, we obtained an F-measure value of 0.231 and achieved better accuracy than the methods without reranking technique when our method was applied to a subset of a general benchmark dataset.

- In Chapters 5, 6 and 7, we performed real applications in the field of systems biology. In this study, we applied MEGADOCK to (i) a bacterial chemotaxis pathway and (ii) a human apoptosis pathway to reconstruct pathways and determine unknown interactions. In the chemotaxis pathway analysis, all core signaling interactions were correctly predicted with the exception of interactions activated by protein phosphorylation. In the apoptosis pathway analysis, the prediction results included several new PPI candidates that might be suitable targets for drug discovery. In addition, the MEGADOCK was enhanced for RNA, leading to the development of a protein-RNA interaction prediction system.

- In Chapters 8 and 9, we compared MEGADOCK with other structure-based PPI screening tools: (i) ZDOCK [33] that has similar scoring functions to MEGADOCK and (ii) PRISM [14] that is a template-based PPI prediction tool. The predicted interactions generated from MEGADOCK and ZDOCK in chemotaxis pathway analysis were slightly different; however when the positive predictions from both tools were combined, the vast majority of relevant interactions were represented. Indeed, there were only two exceptions, both requiring phosphorylation to activate the corresponding interaction. The consensus between template-based and non-template-based methods successfully predicted the PPI network more accurately than the conventional single template-/non-template-based methods. Because such precise prediction reduces biological screening costs, it should further promote interactome analysis.

- In Appendix A, we implemented our protein–protein docking method to be suitable for running on supercomputers by using hybrid parallelization (MPI/OpenMP), where a number of docking processes are distributed among the nodes by MPI with each docking process that is also calculated in parallel by threads using OpenMP within one node. This implementation has significant advantages that (i) save memory space and (ii) avoid a large overhead because of

handling data communication on numerous core systems such as the K computer running a flat MPI implementation. As a result, we obtained a strong scaling value that is a type of evaluation value for parallel efficiency, of over 0.95 out of a maximum of 1.00 in both K computer and TSUBAME 2.0.

- In Appendix B, we enabled the use of recent computing systems by taking advantage of GPU features. We implemented not only FFT calculations but also generated grid (voxelization) and rotation of protein structures on GPUs to reduce the cost of data transfers. As a result, the system achieved 13.9-fold acceleration using 1 CPU core and 1 GPU, and 37.0-fold acceleration using 12 CPU cores and 3 GPUs by making full use of heterogeneous computing resources.

10.2 Future Work

10.2.1 Improvement of post processing of PPI prediction

Improvement of the accuracy of PPI predictions is an important issue. The following three different points are considered for further improvement.

1. **Optimizing to the threshold parameter E^*.** A parameter of the sensitivity of PPI predictions, E^*, was set up on the basis of optimal F-measure values through this study. However, the size of a dataset, especially the number of protein pairs and ratio of positives and negatives, will change the optimal threshold E^* that attains the optimal F-measure value. In fact, biological networks like the PPI network are conjectured to be scale-free [160, 161]. In addition, biological information such as biochemical function and subcellular localization information may be useful. Development of a method for optimizing parameters based on these points of interest is needed to further improve the prediction accuracy.

2. **Correction of the bias from the number of structures.** As described in Chapter 9, the number of false positive predictions is highly correlated with the number of PDB chain structures used for prediction. We should consider one of two ways to reduce false positives: (i) how to generate the target set without multiple conformations in each protein and (ii) develop a correction method when the target set contains multiple conformations.

3. **Considering the post process.** In the field of protein tertiary structure prediction and docking prediction, post processing, such as clustering and filtering,

is commonly used for improvement of performance. For example, Uchikoga, *et al.* used an interaction fingerprint (IFP) technique for clustering protein decoys and succeeded in improving docking predictions [162, 163]. We did not consider this post process method and adopted only energy reranking for faster calculation and accepted less improvement of PPI predictions in our pre-experiments. To introduce IFP and other post-process techniques should be considered future work for further improvement of docking analysis. In addition, we used ZRANK as a reranking tool through this study, but we need to develop an in-house reranking tool usable on GPU supercomputers for further acceleration of PPI network predictions.

10.2.2 Flexible PPI prediction

Considering the flexibility of protein structures and flexible docking represents an effective solution for improvement of prediction performance, but is computationally intensive. One possible workaround to this problem is to consider ensemble docking. In a pre-existing equilibrium model [164], proteins have a steady-state distribution of tertiary structures, and one such structure corresponds to that of the bound form. This model explains the structural difference between bound and isolated proteins. The equilibrium is disrupted by ligand binding and interaction with other proteins. A possible strategy is to generate structure variations based on the crystallized data by methods such as normal mode analysis [165, 166] and molecular dynamics to construct a hypothetical protein structure ensemble, and then use the possible structures in all-to-all docking. It may also be useful to perform structure sampling if we obtain too much structure data for one type of protein and would like to reduce the number of data elements.

10.2.3 More large-scale pathway analysis

An application to more large-scale pathway analysis is also an area we are currently studying for future work. We are applying our methods to proteins related to the epidermal growth factor receptor (EGFR) signaling pathway. Besides its importance in the proliferation and function of normal cells, when this pathway is altered, inappropriate signaling contributes to the pathogenesis of human cancers [167]. The problem size is approximately $2{,}000 \times 2{,}000$, and the task is well within MEGADOCK's power.

10.2.4 Other hardware acceleration

In the present study, we were able to accelerate the protein–protein docking calculation through GPU implementation (see Appendix B). However, recent supercomputers equip not only GPUs but MIC architecture, such as the Intel Xeon Phi processor. The top ranked supercomputer "Tihanhe-2" at the National Super Computer Center in Guangzhou, China [168], was accelerated by Xeon Phi and achieved 33.9 petaflops. To expand MEGADOCK to the MIC architecture is considered future work to respond to the changes in the field of high-performance computing. Recent progress of hardware accelerators is impressive, so improved acceleration by using several hardware accelerators will open up new doors in the world of science and technology by solving computational complex problems that were previously considered seemingly impossible.

Part VI

Appendix

Appendix A

MPI/OpenMP Hybrid Parallelization of Protein–Protein Docking

A.1 Introduction

The next challenge is to perform interactome level large-scale analysis. In order to address this problem we proposed a rapid protein–protein docking method and post-docking analysis [58, 59, 162, 163]. Using this system, we input protein tertiary structure data to acquire predictions of possible interacting pairs.

For example, when reconstructing the bacterial chemotaxis pathway [23], 101×101 potential combinations of structures were considered. In the human apoptosis pathway [114], 158×158 potential combinations of structures were also considered. In real biology problems, such as searching for the drug induced pathway of EGFR (Epidermal Growth Factor Receptor) signaling, about 200 proteins need to be examined. In our preliminary survey on the EGFR pathway and related proteins data, we identified about 2,000 structures corresponding to these proteins. Therefore, the PPI network prediction system needs to handle about $2{,}000 \times 2{,}000$ combinations of protein structures.

To solve such large-scale problems, a highly efficient computing system is necessary. High performance computers are currently being developed and built [54]. Some top ranked supercomputers have shown a peak performance of 33.4 petaflops (Tianhe-2, National Super Computer Center in Guangzhou, China [168]), 17.6 petaflops (Titan, Oak Ridge National Laboratory, USA [169]) and 10.5 petaflops (K computer, RIKEN,

Advanced Institute of Computer Science (AICS), Japan) in November 2013.

We have implemented a protein–protein docking calculation of MEGADOCK suitable for running on supercomputers by using hybrid parallelization with MPI and OpenMP, where a number of docking processes are distributed among the nodes by MPI with each docking process also calculated in parallel by threads by OpenMP within one node. Data parallelization showed almost linear scaling up to 24,576 nodes on K computer (RIKEN AICS, Japan).

We expect the proposed method can be a useful tool in bioinformatics and systems biology area as a basic tool, assuming we can utilize 10,000 $\sim$ 100,000 CPU cores.

A.2 Implementation

We implemented MEGADOCK by hybrid parallelization (MPI/OpenMP) in order to conduct large numbers of docking jobs for PPI network predictions.

The overall procedure of MEGADOCK is shown in Fig. A.1. On the cluster computers, a master node gets a list of protein structures and distributes the docking jobs to available nodes. Upon docking of receptor and ligand proteins, a ligand protein is rotated to various orientations and translated in the space around the receptor, which is fixed during docking calculation, to search for the best scoring positions. These processes are parallelized by threads.

A.2.1 Hybrid parallelization

Initially, a master node distributes docking jobs to available nodes after obtaining a list of protein pairs. We parallelized the calculation of each docking processes using MPI library (Fig. A.1, red colored loop). After one loop of this MPI, we obtain high scoring poses among all the rotation and translation patterns of assigned protein pairs.

Each docking process in each node is parallelized to threads by OpenMP (Fig. A.1, blue colored loop). Upon docking, the coordinates of the ligand are repeatedly rotated and translated to search for a better complex form with the receptor. The calculations of FFT and inverse FFT for each rotation angle are performed independently. Thus, using OpenMP loop we calculate high scoring poses for various rotation angles in parallel.

The implementation is designed to run efficiently on K computer which has 88,128 nodes with 8 cores per node (i.e., total of 705,024 cores). Each node is equipped with 16 GB of memory. Flat MPI is often used for parallel applications. However, using

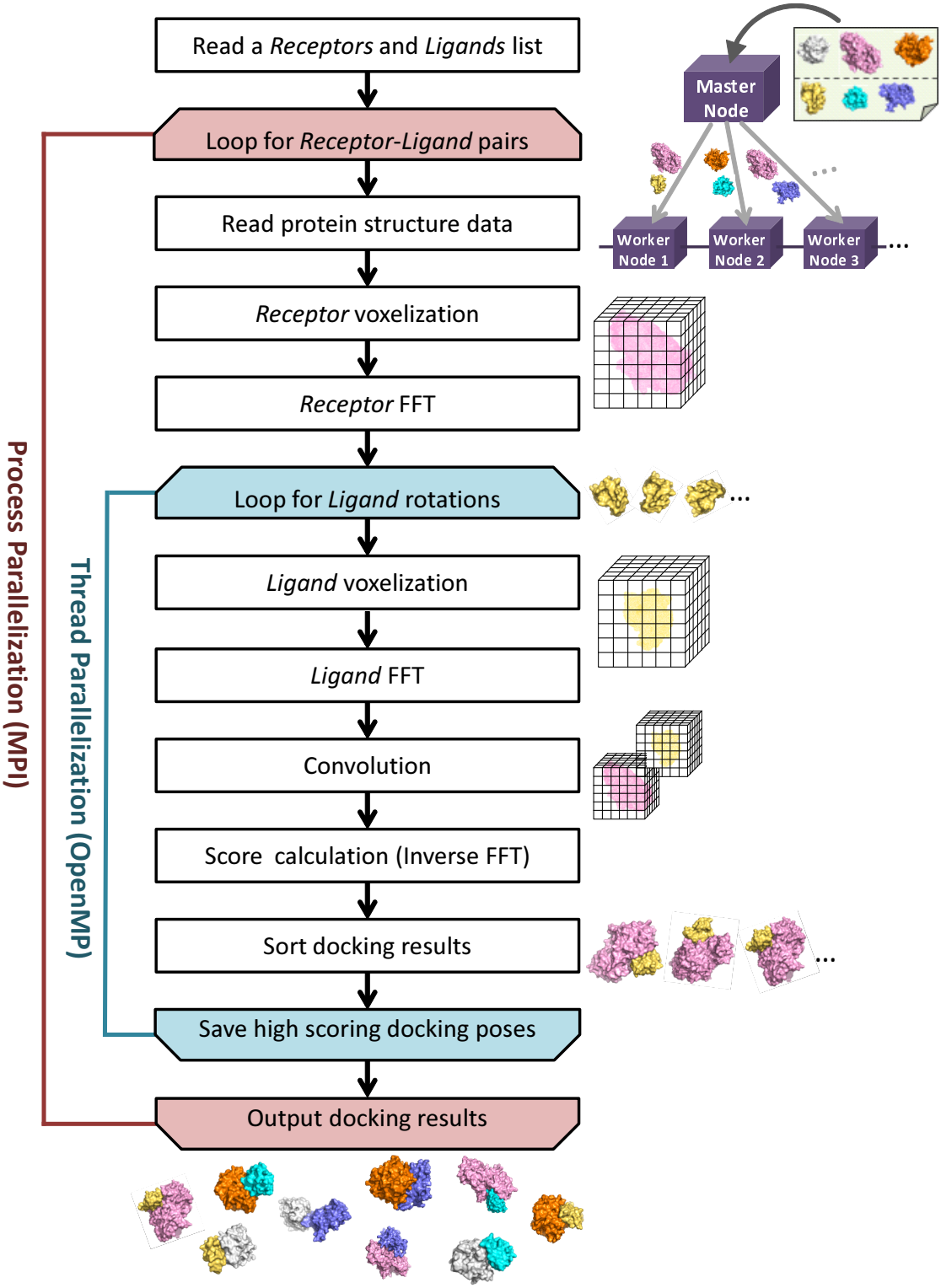


Figure A.1: Flow chart of the MEGADOCK docking process. A master node gets a list of docking targets and distributes each job to the available nodes. Each node calculates one docking job by thread parallelization.

flat MPI on numerous core systems like K computer may result in a large overhead due to handling data communication of ∼700,000 cores. Thus, hybrid parallelization is efficient on such high performance computing systems.

Reducing usage of memory space is important with systems that have many cores per node and relatively small memory size. In flat MPI, the docking job of each protein pair is assigned to each core. Thus, each core requires memory space for input/output data. If a node has n cores, the memory space in the node should be large enough to keep data for n pairs of proteins (in the case of K computer, $n = 8$). In contrast, by implementing hybrid parallelization, we assign one protein pair to each node and then distribute the calculations of ligand rotation by thread parallelization. As such, each node will keep data of one pair of proteins on the memory and threads will share the input/output data on the memory. The memory size needed for docking is dependent on protein size. This implementation is feasible when considering calculations of large proteins. Thus, we implemented MEGADOCK by hybrid parallelization.

A.3 Results and Discussion

A.3.1 Dataset

We used a general benchmark dataset for protein docking (protein–protein docking benchmark 4.0, [82]). For measuring thread parallel scalability, we conducted dockings of a protein complex from PDB, 1ACB (chain E and I). The size of FFT is $N = 108$ in this case. Parallel scalability over nodes using MPI was measured by conducting exhaustive docking of 220 different proteins (220×220 dockings), with an FFT size of $N = 140$.

A.3.2 Test environment

Parallel scalability of MEGADOCK was measured on two supercomputing environments, TSUBAME (Tokyo Institute of Technology Global Scientific Information and Computing Center (GSIC), Japan) and K computer (RIKEN AICS, Japan). The most abundant node type of TSUBAME had an Intel Xeon 5670, 2.93 GHz processor, 12 cores. Each node is capable of up to 24 threads of computation by using the hyper threading technique. K computer has Fujitsu SPARC64 VIIIfx CPUs, 2 GHz, 8 cores.

A.3.3 Calculation speedup

The dataset includes proteins of various size (see Chapter 3, Fig. 3.3). The time required for each docking calculation is dependent on protein size. For example, a protein that requires size 120 FFT calculations (1E96) gave a calculation time of about 547 seconds. Smaller sized protein pairs, such as size 80 FFT (1GCQ) were calculated in about 155 seconds. This variation in calculation time reflects the difference of FFT calculation (size $120 \times 120 \times 120$ and $80 \times 80 \times 80$). The smaller protein pair (size 80 FFT) takes about 0.28 times the elapsed time compared to the larger protein pair (size 120 FFT). This ratio of elapsed time is reasonable. In theory FFT takes the order of $\mathcal{O}(N^3 \log N)$ for calculation. Therefore calculations involving a size of 80 FFT should take $\sim$0.27 times $((80^3 \log 80)/(120^3 \log 120) = 0.271...)$ the elapsed time of a corresponding calculation involving a size of 120 FFT, which is almost the same scale as the calculation time we measured on TSUBAME.

A.3.4 Parallel scalability

Fig. A.2 shows the thread parallel scalability of MEGADOCK by parallelizing ligand rotation and FFT calculation. The calculation time is shown as an average of 10 individual docking events with an FFT size of 108 (1ACB chain E and I) from the benchmark data. We observed a 7.33-fold speedup when using the maximum number of threads, 8 threads, on K computer compared to a single thread calculation. We observed a 9.17-fold speedup for 12 threads of calculation and a 10.42-fold speedup for 24 threads of calculation compared to a single thread calculation. Note that in TSUBAME system we measured time with hyper threading activated, so number of threads more than 12 includes slight speedup including this effect.

Fig. A.3 shows a process level parallel scalability of MEGADOCK. On K computer, where a maximum of 24,576 nodes can be used simultaneously, we measured the time needed to calculate exhaustive dockings of 220 proteins (220×220 dockings), calculated with a size of FFT 140. Calculation time using 24,576 nodes was about 3.76-fold faster than the time needed to solve the same problem on 6,144 (1/4 of 24,576 nodes) nodes. On TSUBAME, we measured the time needed to calculate exhaustive dockings of 44 proteins (44×44 dockings) using up to 400 nodes at a time. Calculation time using 400 nodes was about 3.78-fold faster than that on 100 nodes. MEGADOCK achieved almost linear scalability on both supercomputing environments.

We used a dataset of similar sized proteins (FFT size 140) for the scalability test. It is an unrealistic scenario when calculations conducted by each node are almost

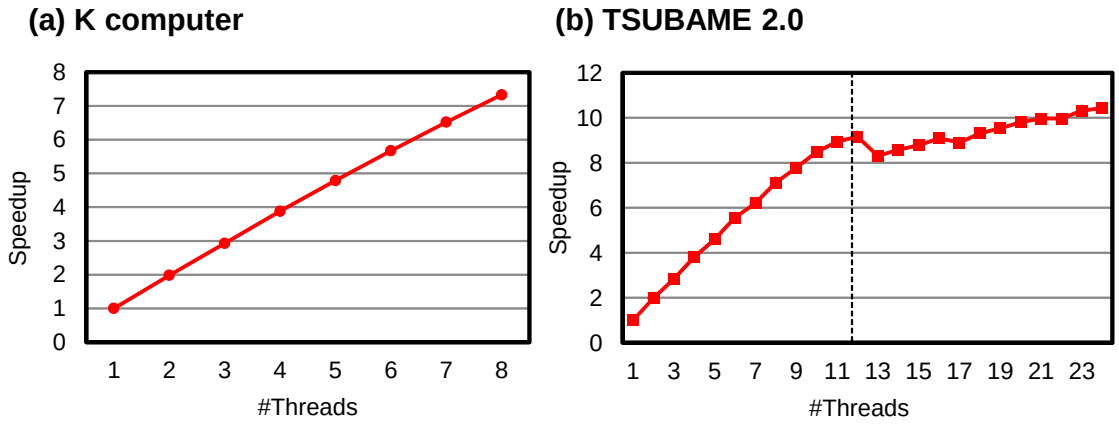


Figure A.2: Scalability of thread parallelization using OpenMP on (a) K computer (8 cores/node) and (b) TSUBAME (12 cores/node, hyper threading enabled). 1ACB chain E and 1ACB chain I was used for docking. Elapsed time was measured from the mean of 30 docking processes. The right area of the dashed line shows speedup by activating hyper threading.

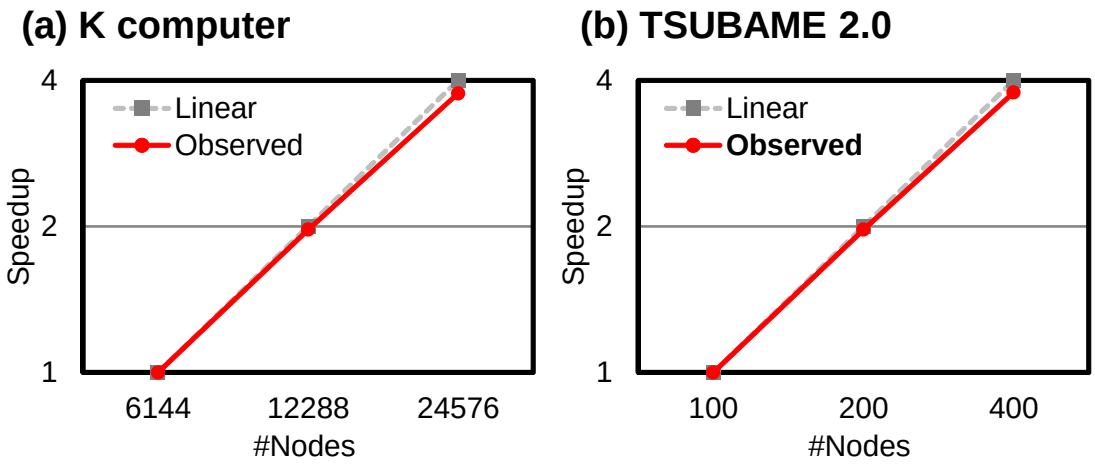


Figure A.3: Scalability of parallelization among nodes by MPI on (a) K computer (6,144 to 24,576 nodes), 220 × 220 dockings of FFT size = 140 protein pairs; (b) TSUBAME (100 to 400 nodes), 44 × 44 dockings of FFT size = 140 protein pairs.

equal. Thus, the only possible overhead by parallelization is the job distribution and checking by the controller nodes. For real problems, which include simulating dockings of proteins with a variety of sizes, a more intelligent controller is needed to efficiently distribute docking tasks according to the protein size.

Another possible improvement to make calculation faster could be on the FFT calculation. A profiler output showed that about 86.4% of the elapsed time was used by FFT and inverse FFT calculations (see Chapter 3, Table 3.9). Users can switch the FFT engine to similar libraries by making small changes to the MEGADOCK source code. We have tried using FFTE [170], FFTW [83] and FFT function (dvcfm1) in CSSL2 (Fujitsu Ltd., Tokyo, Japan). All three implementations yielded equivalent docking outputs. The speed of calculation differs depending on the size of the proteins. For example, CSSL2 was slightly faster than FFTW when applied to FFT size of 128 or other base 2 FFT calculations. By contrast, FFTW outperformed CSSL2 with docking simulations involving other sizes of protein. Thus, FFTW may be the function of choice for applications where the dataset includes proteins of various sizes.

A.4 Summary

In this chapter, we implemented a high-throughput and ultra-fast PPI network prediction system "MEGADOCK" suitable for massively paralleled large-scale analysis of millions of protein combinations. The docking engine of MEGADOCK was implemented by parallelization techniques and shown to be scalable on massively parallel computing environments. MEGADOCK is ideally suited to a large-scale computing system.

Appendix B

Acceleration of Protein-Protein Docking on GPUs

B.1 Introduction

Recently, graphics processing units (GPUs) have been transformed into powerful accelerators for general purpose computing. Current GPUs, such as NVIDIA's Tesla K20, have excellent power efficiency and their computational power supersedes that of CPUs. Also, GPU software development tools, such as NVIDIA's Compute Unified Device Architecture (CUDA) [171], have been developed and they enable us to develop GPU applications much easier. Thus the general-purpose computing on GPUs (GPGPU) techniques have been widely used in various research fields including bioinformatics, such as metagenome sequence mapping [172], molecular dynamics simulation [173] and quantum chemistry calculation [174]. Therefore, the MEGADOCK could be also accelerated using GPU computing techniques.

In this chapter, we mapped the docking calculation of MEGADOCK onto GPUs and developed fast protein-protein docking software named MEGADOCK-GPU. We implemented almost all processes of MEGADOCK including "FFT", "modulation" and "ligand voxelization". We also implemented the system for utilizing all CPU cores and GPUs in a computation node. As results, MEGADOCK-GPU on 12 CPU cores and 3 GPUs achieved a calculation speed that was 37.0 times faster than MEGADOCK CPU version on 1 CPU core.

B.2 Related Work

A GPU implementation of FFT-based protein-protein docking was already done by Sukhwani, *et al.* in 2009 [175]. They mapped a FFT-based protein-protein docking tool PIPER [29] onto GPUs, and demonstrated a calculation speed that was 17.7 times faster with 1 GPU than PIPER with 1 CPU core. Sukhwani, *et al.* used cuFFT library [176] for mapping FFT processes onto GPUs, and also mapped several processes onto GPUs. PIPER is a famous docking tool and has shown its good prediction accuracy through international benchmarks [177], but it uses 22 energy terms and is much slower than ZDOCK and MEGADOCK. Therefore, the performance of PIPER is still insufficient for proteomics-scale studies even with GPUs.

B.3 GPU Acceleration

We developed MEGADOCK-GPU for further acceleration of a protein-protein docking. For the GPGPU implementation, we used CUDA, which is a platform for GPGPU provided by NVIDIA. The system requires CUDA version 5.0 or later because older cuFFT libraries in the previous version of CUDA have problems in the barrier synchronization. We mapped not only the FFT and modulation processes but also the voxelization and finding the best solutions processes onto GPUs. In the previous work by Sukhwani, *et al.*, the targeted system has only a GPU card and their implementation could not utilize multiple GPUs. However, current computing system often has multiple CPU cores and multiple GPU cards in a computing node. It is thus important to make full use of such a computing environment, e.g. 12 CPU cores and 3 GPUs. Therefore, we targeted a computing node with multiple CPU cores and multiple GPU cards.

B.3.1 Profile of MEGADOCK processes

Fig. B.1 shows the flow of the docking processes of MEGADOCK and Table B.1 shows proportion of docking calculation time for each process of MEGADOCK. This profile was obtained from the docking calculation for a protein complex (PDB ID: 1ACB, receptor: chain E, 245 residues, ligand: chain I, 70 residues). The FFT size N of the docking calculation is 108, and it is typical in the current protein structure database. The profile was taken on Intel Xeon 2.93 GHz, 1 CPU core.

FFT processes (P5+P7) occupy majority (86.1%) of total time. On the other hand,

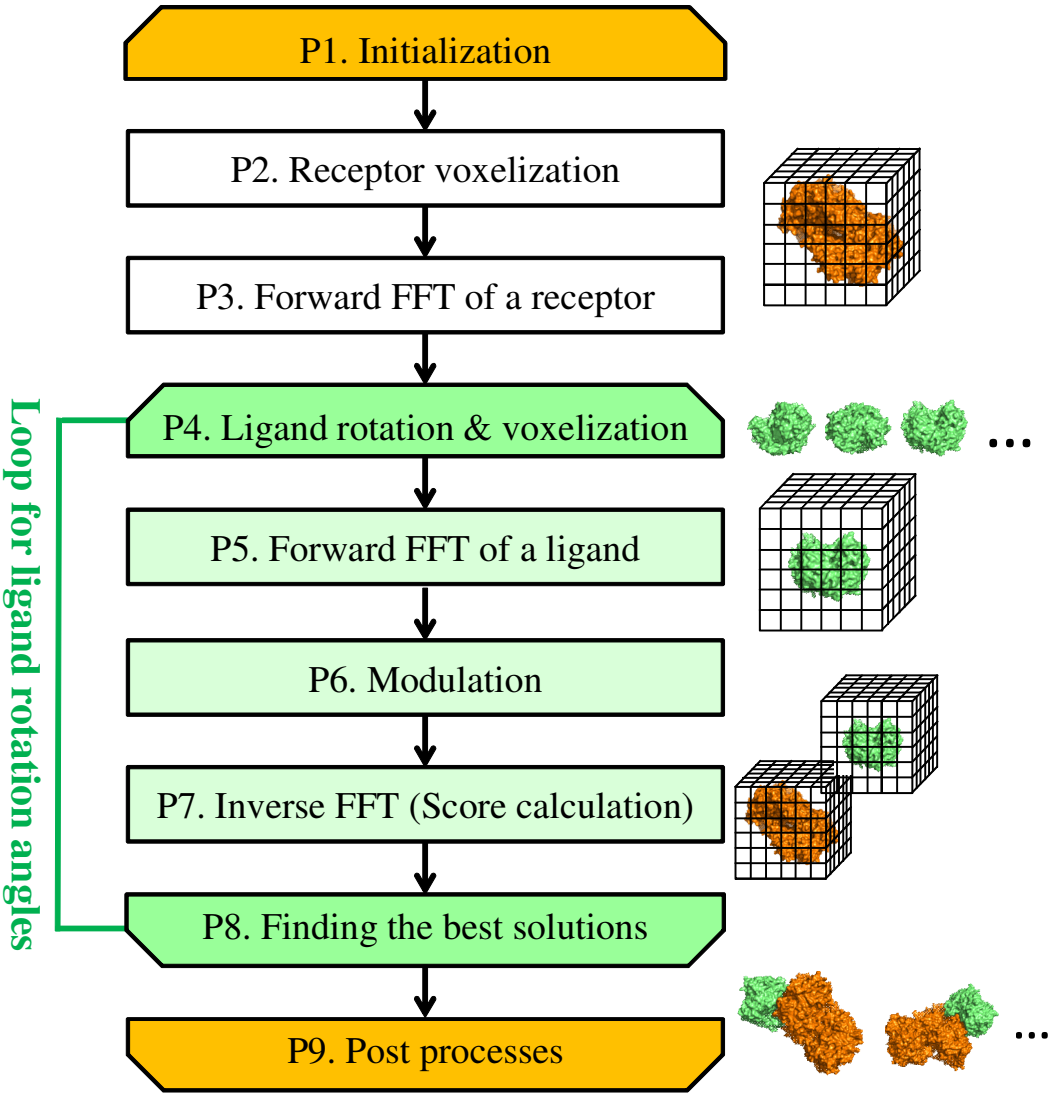


Figure B.1: The process flow of FFT-based docking tools.

Table B.1: The profile of docking calculation on 1 CPU core (PDB ID: 1ACB).

	Time (sec.)	Ratio (%)
P1. Initialization	0.04	0.0
P2+P3. Receptor processes	0.30	0.1
P4. Ligand rotation & voxelization	9.19	2.7
P5. Forward FFT of a ligand	155.99	46.4
P6. Modulation	33.97	10.1
P7. Inverse FFT	133.68	39.7
P8. Finding the best solutions	3.27	1.0
P9. Post processes	0.00	0.0
Total	336.45	100.0

other calculations such as voxelization and finding the best solutions still consume considerable time portions. It is because the time of FFT calculation has already been reduced by employing a simplified scoring function compared with the other docking software. Thus, even processes other than the FFT calculation must be accelerated for significant speedup. Assume that 30-fold acceleration is achieved in the FFT and modulation processes (P5–P7), then estimated time consumption of these processes will be approximately 10 seconds and the total computation time will be 23 seconds. As a result, the computation time of ligand voxelization (P4, currently 9.19 seconds consumed) will occupy about 40% of total time. Therefore, the mapping of almost all processes, which include ligand voxelization and finding the best solutions, onto GPUs is obviously crucial for achieving effective acceleration.

B.3.2 Implementation on GPUs

We have implemented the following processes on GPUs, forward FFT of a receptor (P3), ligand rotation and voxelization (P4), forward FFT of a ligand (P5), modulation (P6), inverse FFT (P7) and finding the best solutions (P8). The details of each implementation are described in the following.

Ligand voxelization

MEGADOCK sets adequate rPSC score, electrostatics values and desolvation free energy score on the ligand voxel model in this process. Ligand voxelization is a process that calculates the distance between the coordinates of an atom and each grid and

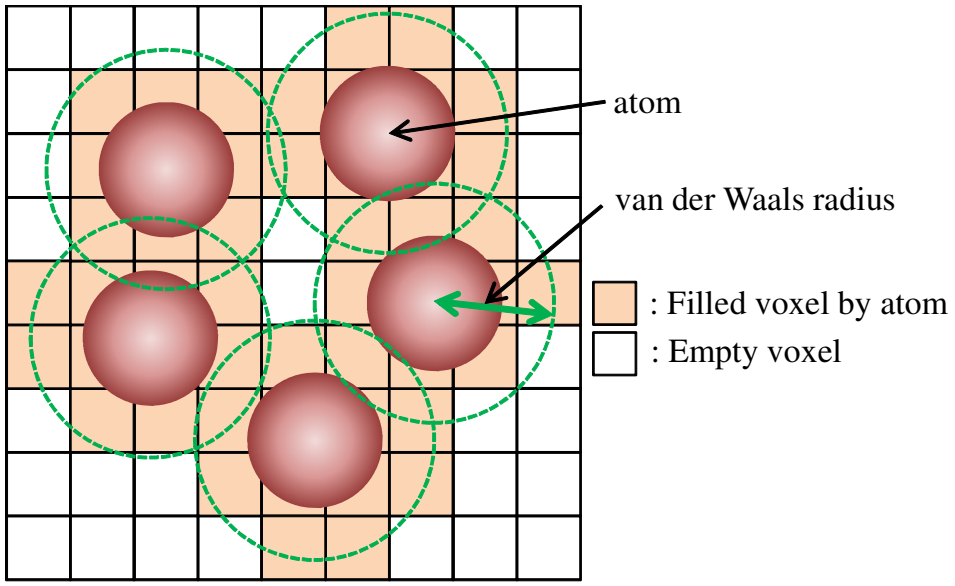


Figure B.2: Assignment of voxels filled by atoms.

assigns a value to each grid within van der Waals radius of the atom (Fig. B.2). The assignment process can be parallelized for each atom. Because the rPSC score and desolvation free energy score of a ligand has only binary states (0 or 1) and the electrostatics value of a grid is calculated as accumulative sum of the values of all adjacent atoms, the calculation order for each atom can be freely exchanged. Therefore, we could process atoms in parallel and mapped them onto GPUs. Thus, multiple atoms are simultaneously processed on different GPU cores in this process.

Forward and inverse FFT

For mapping FFT calculations onto GPUs, we used the NVIDIA cuFFT library [176]. Since cuFFT is optimized for FFT bases $\{2, 3, 5, 7\}$, MEGADOCK-GPU uses FFT size N as a multiple of $\{2, 3, 5, 7\}$. This issue will be discussed again on section 5.6.3.

Finding the best solutions

In this process, the best docking pose is selected according to the docking score. This reduction process is also implemented on GPUs. Current MEGADOCK-GPU only reports a single pose with the best score in a ligand rotation step while original MEGADOCK can report the n-best poses. However, the other protein-protein docking software, such as ZDOCK and PIPER, does not have such option, and it is not so crucial

in a practical case.

Rotation of ligand

In this process, the atom coordinates of a ligand are updated according to a given rotation matrix. The process for each atom is independent and can be fully parallelized. We mapped them onto GPUs.

Modulation

The modulation can be independent for each grid, thus we mapped them onto GPUs.

B.3.3 Data transfer

In the previous work by Sukhwani, *et al.*, they implemented the voxelization process on a CPU because the system needs to perform many FFT calculations and it did not become a bottleneck. Thus, a voxelized protein structure data had to be transferred from a host system to GPUs. Previously, we had also performed the voxelization on a CPU, but the data transfer from a host to GPUs became a bottleneck. Indeed, only one FFT calculation is required in our system and the data transfer then occupies large portion of computation time. Thus, we tried to minimize the data transfer. In our implementation, the transfer of large data from a host to GPUs takes place only once. The data includes the original atom coordinates of a ligand and Fourier transformed receptor grid information are transferred at first. In the loop for each ligand rotation angle, only trivial data transfer is required (12 bytes angular information and 8 bytes calculation result) because all processes are performed on GPUs.

B.3.4 Using multiple CPU cores and multiple GPUs

The latest computing systems tend to have a powerful computing node composed of several multicore CPUs and multiple GPU cards. Thus, we implemented our system to fully use such heterogeneous computing resources. As we mentioned above, the processes for each ligand rotation are parallelized in our system. For utilizing all computing resources, we assign the decomposed jobs to multiple GPUs and CPU cores dynamically using OpenMP. The same number of CPU cores as GPUs is used for controlling GPU processes but the remaining cores perform docking calculation by themselves. Thus, 3 CPU cores are used for controlling 3 GPUs and the remaining 9

Algorithm 1 Parallel algorithm of docking calculation

Require: atom coordinates of receptor R and ligand L, rotational angles of ligand $\{\theta\}$, number of CPU cores C, number of GPUs G
Ensure: List of high-score docking poses $\mathcal{H}$

```
Initialization
Data transfer to GPU (R, L)
R' ← Voxelization(R)
F[R'] ← FFT(R')
for each θ by C threads do
    if thread number < G then
        processed on GPU
    else
        processed on CPU
    end if
    L_θ ← Rotation(L, θ)
    L'_θ ← Voxelization(L_θ)
    F[L'_θ] ← FFT(L'_θ)
    M_θ ← Modulation(F[R'], F[L'_θ])
    Score_θ ← IFFT(M_θ)
    H ← H ∪ Max(Score_θ)
end for
Post processes
```

CPU cores are used for docking calculation, when we use a system with 12 CPU cores and 3 GPUs. Algorithm 1 shows parallel algorithm of docking calculation.

B.4 Evaluation of Performance

B.4.1 Computation environment

All the calculations were conducted on the TSUBAME 2.0 supercomputing system, Tokyo Institute of Technology, Japan. We used its thin nodes in all experiments. The specifications of the node are shown in Table B.2.

B.4.2 Dataset

We used 352 protein complex structures retrieved from a standard protein-protein docking benchmark set (Protein-Protein Docking Benchmark 4.0) [82], which contains protein structures in both bound and unbound forms. The sizes of the proteins in the

Table B.2: Computation environment

CPU	Intel Xeon X5670 2.93 [GHz] (6 cores)×2
Memory	54 [GB]
OS	SUSE Linux Enterprise Server 11 SP1
GPU	NVIDIA Tesla M2050×3 (3 GPUs / 1 node)
Compiler	Intel C++ Compiler 13.0.0
FFTW	FFTW 3.2.2
CUDA	CUDA 5.0
cuFFT	cuFFT 5.0

dataset are distributed widely and it is a fairly-sampled subset of the current known protein structure complexes.

B.4.3 Evaluation method

For evaluating calculation time of each system, we performed same docking calculation for 352 protein pairs three times, and took their average. We used `gettimeofday()` function to measure calculation time. There is no difference between the CPU version of MEGADOCK and the GPU version in the scoring function. Although MEGADOCK-GPU seldom returned different results from those of the CPU version due to the different precision of numeric calculation between a CPU and a GPU, the difference was less than 0.001% and negligibly small.

B.5 Results

B.5.1 Comparison of total docking runtime

Table B.3 shows the results of total and average docking calculation time for 352 protein complexes. MEGADOCK-GPU using 1 CPU core and 1 GPU was 13.9 times faster than MEGADOCK using 1 CPU core. Also, MEGADOCK-GPU using 12 CPU cores and 3 GPUs was 37.0 times faster than MEGADOCK using 1 CPU core. MEGADOCK has been already parallelized using OpenMP and it achieved 8.9-fold speedup using 12 CPU cores. Furthermore, MEGADOCK-GPU using 12 CPU cores and 3 GPUs was approximately 4.2 times faster than MEGADOCK using 12 CPU cores by fully using computer resources of a node. One may expect that MEGADOCK-GPU using 12 CPU cores and 3 GPUs may achieve not 37.0-fold but at least 41.7-fold ($= 13.9 \times 3$)

speedup compared to 1 CPU core, because of using three GPUs. However, the speedup is not fully proportional because the initialization of GPU cannot be parallelized and becomes a bottleneck.

Table B.3: The results of total and averaged docking calculation time for 352 protein complexes.

	1 CPU core	1 CPU core + 1 GPU	12 CPU cores	12 CPU cores + 3 GPUs
Total (hour)	73.5	5.3	8.2	2.0
Average (sec.)	751.4	54.1	84.1	20.3
vs. 1 CPU core	1.0	13.9	8.9	37.0

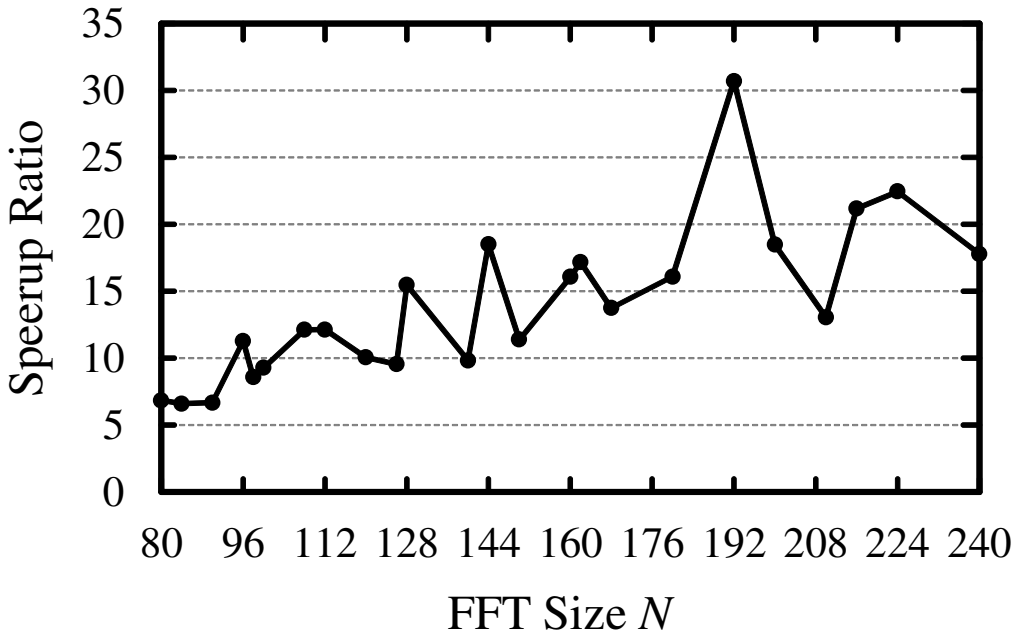


Figure B.3: The distribution of the speedup ratio of MEGADOCK-GPU using 1 CPU core and 1 GPU compared to MEGADOCK using 1 CPU for different FFT size N. Horizontal axis shows FFT size N and vertical axis shows the averaged speedup ratio in protein complexes with same FFT size.

B.5.2 Distribution of computation time for FFT size

Fig. B.3 shows the distribution of the speedup ratio of MEGADOCK-GPU with 1 CPU core and 1 GPU compared to MEGADOCK using 1 CPU core, for each FFT size. In the figure, the horizontal axis is FFT size N and the vertical axis is an averaged speedup ratio in the complexes whose FFT size are same. The ratio generally increases in proportion to the size of FFT. Because, the FFT requires $\mathcal{O}(N^3 \log N)$ calculation but the processes that are hard to be mapped onto GPUs basically take only $\mathcal{O}(N)$. Thus, the speedup is not large for small N. Within this experiments, the best speedup ratio was obtained with FFT $N = 192$ and is 33.7-fold against 1 CPU core.

B.5.3 Speedup on each process

Table B.4 shows the speedup of docking calculation from MEGADOCK using 1 CPU core to MEGADOCK-GPU using 1 CPU core and 1 GPU, for each process. In the FFT and modulation processes, better acceleration is achieved. Speedups on the ligand voxelization and finding the best solutions processes are moderate because they include

reduction processes. However, we succeeded to reduce data transfer drastically by mapping these processes onto GPUs. Thus the GPU implementation of these processes is practically effective. The time of initialization highly increased using a GPU. This is because the startup time is demanded for initializing a GPU and it is difficult to be decreased.

Table B.4: Acceleration ratio for each calculation part (PDB ID: 1ACB).

	1 CPU core (sec.)	1 CPU core & 1 GPU (sec.)	Speedup Ratio
P1. Initialization	0.04	5.34	-
P2+P3. Receptor processes	0.30	0.26	1.2
P4. Ligand rotation & voxelization	9.19	3.72	2.5
P5. Forward FFT of a ligand	155.99	5.92	26.4
P6. Modulation	33.97	1.09	31.3
P7. Inverse FFT	133.68	5.92	22.6
P8. Finding the best solutions	3.27	2.58	1.3
P9. Post processes	0.00	0.00	1.0
Total	336.45	24.83	13.5

B.6 Discussion

B.6.1 Data transfer time

In this work, we mapped almost all processes of a protein-protein docking calculation onto GPUs. Also, we succeeded to reduce the amount of data transfer. As results, the time of data transfer in now only 270 milliseconds is approximately 1.3% of total docking time in the case of a docking for PDB ID: 1ACB. This is a large advantage to calculate all processes in a ligand rotation loop on GPUs.

B.6.2 Initialization of GPU

We achieved to reduce computation time of almost all processes in a protein-protein docking significantly by using GPGPU techniques. As results, the initialization of a GPU has now become one of the bottlenecks. As shown in Table B.4, the initialization of a GPU requires approximately 5 seconds and it occupies more than 20% of whole computation time, because we generate a new process and initialize GPU for each pair now. However, in the practical application like a protein-protein interaction network prediction, we have to perform large number of docking calculations for many protein pairs. Therefore, we have only to initialize GPU once if we modify the system to deal many docking calculations for multiple protein pairs in a single computing process. Assume that the time of GPU initialization can be ignorable; the average calculation time of MEGADOCK-GPU using 12 CPU cores and 3 GPUs will be approximately 15 seconds and it is 50-times faster than that of using 1 CPU core.

B.6.3 Optimization of FFT size

An FFT-based docking tool firstly reads the atom coordinates of a receptor and a ligand, and determines the grid size fitted for the receptor and ligand. The FFT size N is proportional to the grid size, which was automatically calculated from the single grid unit size and the size of proteins. FFTW algorithms [83], which is used in MEGADOCK, are optimized for sizes that represented as a multiple of $\{2, 3, 5, 7, 11, 13\}$. Thus, our algorithm to decide the grid size searches the smallest composite number consisted of those prime factors.

However, cuFFT algorithms [176], which is used in MEGADOCK-GPU, are optimized for sizes that represented as a multiple of $\{2, 3, 5, 7\}$. For the other sizes, slower algorithm is used. Therefore, we should adjust the grid size to optimal one for the

cuFFT library. However, in case of using multiple cores and GPUs, MEGADOCK-GPU uses both the FFTW library for a CPU and the cuFFT library for a GPU. The best set of prime numbers consisting FFT size N is different between them. We conducted an investigation of a set of prime numbers as a previous experiment. For MEGADOCK-GPU, the best results were obtained on a set $\{2, 3, 5, 7\}$ in both using 1 CPU core and 1 GPU, and using 12 CPU cores and 3 GPUs. For evaluating the performances, we used $\{2, 3, 5, 7\}$ for MEGADOCK-GPU and $\{2, 3, 5, 7, 11, 13\}$ for MEGADOCK CPU version.

B.7 Summary

In this chapter, we developed MEGADOCK-GPU and mapped almost all processes of a protein-protein docking onto GPUs. As a result, the system achieved 13.9-fold acceleration using 1 CPU core and 1 GPU, and 37.0-fold acceleration using 12 CPU cores and 3 GPUs by making full use of heterogeneous computing resources.

References

[1] Wass MN, David A, Sternberg MJE. Challenges for the prediction of macromolecular interactions. *Curr Opin Struct Biol*, 21: 382–390, 2011.

[2] Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an *Arabidopsis* interactome map. *Science*, 333: 601–607, 2011.

[3] Fields S, Song O. A novel genetic system to detect protein–protein interactions. *Nature*, 340: 245–246, 1989.

[4] Förster T. Zwischenmolekulare energiewanderung und fluoreszenz. *Ann Physik*, 437: 55–75, 1948.

[5] Higurashi M, Ishida T, Kinoshita K. Identification of transient hub proteins and the possible structural basis for their multiple interactions. *Protein Sci*, 17: 72–78, 2008.

[6] Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein–protein interactions based only on sequences information. *Proc Natl Acad Sci U S A*, 104: 4337–4341, 2007.

[7] Park Y, Marcotte EM. Revisiting the negative example sampling problem for predicting protein–protein interactions. *Bioinformatics*, 27: 3024–3028, 2011.

[8] Valencia A, Pazos F. Prediction of protein–protein interactions from evolutionary information. *Structural Bioinformatics, Second Edition*, 617–634, 2009.

[9] Bowers PM, Cokus SJ, Eisenberg D, Yeates TO. Use of logic relationships to decipher protein network organization. *Science*, 306: 2246–2249, 2004.

[10] Deng M, Mehta S, Sun F, Chen T. Inferring domain-domain interactions from protein–protein interactions. *Genome Res*, 12: 1540–1548, 2012.

[11] Ta HX, Holm L. Evaluation of different domain-based methods in protein interaction prediction. *Biochem Biophys Res Commun*, 390: 357–362, 2009.

[12] Hue M, Riffle M, Vert J-P, Noble WS. Large-scale prediction of protein–protein interactions from structures. *BMC Bioinform* 11:144, 2010.

[13] Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A. PRISM: protein interactions by structural matching. *Nucleic Acids Res*, 33: W331–336, 2005.

[14] Tuncbag N, Gursoy A, Nussinov R, Keskin O. Predicting protein–protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc*, 6: 1341–1354, 2011.

[15] Gromiha MM, Yokota K, Fukui K. Energy based approach for understanding the recognition mechanism in protein–protein complexes. *Mol Biosyst*, 5: 1779–1786, 2009.

[16] La D, Kihara D. A novel method for protein–protein interaction site prediction using phylogenetic substitution models. *Proteins*, 80: 126–141, 2012.

[17] La D, Kong M, Hoffman W, Choi YI, Kihara D. Predicting permanent and transient protein–protein interfaces. *Proteins*, 81: 805–818, 2013.

[18] Selent J, Kaczor AA, Guixà-González R, Carrió P, Pastor M, Obiol-Pardo C. Rational design of the survivin/CDK4 complex by combining protein–protein docking and molecular dynamics simulations. *J Mol Model*, 19: 1507–1514, 2013.

[19] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN. The Protein Data Bank. *Nucleic Acids Res*, 28: 235–242, 2000.

[20] RCSB Protein Data Bank,
`http://www.rcsb.org/pdb/.`

[21] DeLano WL. The PyMOL molecular graphics system. DeLano Scientific 2002. (Available at: `http://www.pymol.org`)

[22] Wass MN, Fuentes G, Pons C, Pazos F, Valencia A. Towards the prediction of protein interaction partners using physical docking. *Mol Syst Biol*, 7: 469, 2011.

[23] Matsuzaki Y, Matsuzaki Y, Sato T, Akiyama Y. *In silico* screening of protein–protein interactions with all-to-all rigid docking and clustering: an application to pathway analysis. *J Bioinform Comput Biol*, 7: 991–1012, 2009.

[24] Yoshikawa T, Tsukamoto K, Hourai Y, Fukui K. Improving the accuracy of an affinity prediction method by using statistics on shape complementarity between proteins. *J Chem Inf Model*, 49: 693–703, 2009.

[25] Grosdidier S, Totrov M, Fernández-Recio J. Computer applications for prediction of protein–protein interactions and rational drug design. *Adv Appl Bioinform Chem*, 2: 101–123, 2009.

[26] Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A*, 89: 2195–2199, 1992.

[27] Eisenstein M, Katchalski-Katzir E. On proteins, grids, correlations, and docking. *C R Biol*, 327: 409–420, 2004.

[28] Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol*, 272: 106–120, 1997.

[29] Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, 65: 392–406, 2006.

[30] Chen R, Weng Z. Docking unbound proteins using shape complementarity desolvation, and electrostatics. *Proteins*, 47: 281–294, 2002.

[31] Chen R, Weng Z. A novel shape complementarirty scoring function for protein–protein docking. *Proteins*, 51: 397–408, 2003.

[32] Chen R, Li L, Weng Z. ZDOCK An initial-stage protein-docking algorithm. *Proteins*, 52: 80–87, 2003.

[33] Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z. Integrating statistical pair potentials into protein complex prediction. *Proteins*, 69: 511–520, 2007.

[34] Pierce BG, Hourai Y, Weng Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS ONE*, 6: e24657, 2011.

[35] Cheng TM, Blundell TL, Fernández-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins*, 68: 503–515, 2007.

[36] Jiménez-García B, Pons C, Fernández-Recio J. pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. *Bioinformatics*, 29: 1698–1699, 2013.

[37] Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res*, 33: W363–367, 2005.

[38] Venkatraman V, Yang YD, Sael L, Kihara D. Protein–protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics*, 10: 407, 2009.

[39] Ritchie DW, Kemp GJ. Protein docking using spherical polar Fourier correlations. *Proteins*, 39: 178–194, 2000.

[40] Ritchie DW, Venkatraman V. Ultra-fast FFT protein docking on graphics processors. *Bioinformatics*, 26: 2398–2405, 2010.

[41] Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*, 331: 281–299, 2003.

[42] Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein–protein docking. *Protein Sci*, 14: 1328–1339, 2005.

[43] Palma PN, Krippahl L, Wampler JE, Moura JG. BiGGER: A new (soft) docking algorithm for predicting protein interactions. *Proteins*, 39: 372–384, 2000.

[44] Andrusier N, Nussinov R, Wolfson HJ. FireDock: fast interaction refinement in molecular docking. *Proteins*, 69: 139–159, 2007.

[45] Mashiach E, Nussinov R, Wolfson HJ. FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins*, 78: 1503–1519, 2010.

[46] Venkatraman V, Ritchie DW. Flexible protein docking refinement using pose-dependent normal mode analysis. *Proteins*, 80: 2262–2274, 2012.

[47] Lensink MF, Wodak SJ. Docking and scoring protein interactions: CAPRI 2009. *Proteins*, 78: 3073–3084, 2010.

[48] Lensink MF, Wodak SJ. Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins*, 78: 3085–3095, 2010.

[49] Lensink MF, Wodak SJ. Docking, scoring and affinity prediction in CAPRI. *Proteins*, 81: 2082–2095, 2013.

[50] Janin J. The Targets of CAPRI Rounds 20–27. *Proteins*, 81: 2075–2081, 2013.

[51] Fleishman SJ, Whitehead TA, Strauch EM, Corn JE, Qin S, Zhou HX, Mitchell JC, Demerdash ON, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko JS, Lee H, Seok C, Bourquard T, Bernauer J, Poupon A, Azé J, Soner S, Ovali SK, Ozbek P, Tal NB, Haliloglu T, Hwang H, Vreven T, Pierce BG, Weng Z, Pérez-Cano L, Pons C, Fernández-Recio J, Jiang F, Yang F, Gong X, Cao L, Xu X, Liu B, Wang P, Li C, Wang C, Robert CH, Guharoy M, Liu S, Huang Y, Li L, Guo D, Chen Y, Xiao Y, London N, Itzhaki Z, Schueler-Furman O, Inbar Y, Potapov V, Cohen M, Schreiber G, Tsuchiya Y, Kanamori E, Standley DM, Nakamura H, Kinoshita K, Driggers CM, Hall RG, Morgan JL, Hsu VL, Zhan J, Yang Y, Zhou Y, Kastritis PL, Bonvin AM, Zhang W, Camacho CJ, Kilambi KP, Sircar A, Gray JJ, Ohue M, Uchikoga N, Matsuzaki Y, Ishida T, Akiyama Y, Khashan R, Bush S, Fouches D, Tropsha A, Esquivel-Rodríguez J, Kihara D, Stranges PB, Jacak R, Kuhlman B, Huang SY, Zou X, Wodak SJ, Janin J, Baker D. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *J Mol Biol*, 414: 289–302, 2011.

[52] K computer (RIKEN Advanced Institute for Computational Science), `http://www.aics.riken.jp/en/kcomputer`.

[53] TSUBAME (Tokyo Institute of Technology, Global Scientific Information and Computing Center), `http://www.gsic.titech.ac.jp/en/tsubame`.

[54] Top500 supercomputer sites, `http://www.top500.org`.

[55] The Green500, `http://www.green500.org/`.

[56] Ohue M, Matsuzaki Y, Matsuzaki Y, Sato T, Akiyama Y. MEGADOCK: an all-to-all protein–protein interaction prediction system using tertiary structure data and its application to systems biology study. *IPSJ TOM*, 3: 91–106, 2010. (in Japanese)

[57] Ohue M, Matsuzaki Y, Akiyama Y. Docking-calculation-based method for predicting protein–RNA interactions. *Genome Informatics*, 25: 25–39, 2011.

[58] Ohue M, Matsuzaki Y, Ishida T, Akiyama Y. Improvement of the protein–protein docking prediction by introducing a simple hydrophobic interaction model: an application to interaction pathway analysis. *Lecture Notes in Bioinformatics*, 7632: 178–187, 2012.

[59] Ohue M, Matsuzaki Y, Uchikoga N, Ishida T, Akiyama Y. MEGADOCK: An all-to-all protein–protein interaction prediction system using tertiary structure data. *Protein Pept Lett.* (in press)

[60] Ohue M, Matsuzaki Y, Shimoda T, Ishida T, Akiyama Y. Highly Precise Protein–protein Interaction Prediction Based on Consensus Between Template-Based and *de Novo* Docking Methods. *BMC Proceedings*, 7(Suppl 7): S6, 2013.

[61] Matsuzaki Y, Ohue M, Uchikoga N, Akiyama Y. Protein–protein interaction network prediction by using rigid-body docking tools: application to bacterial chemotaxis. *Protein Pept Lett.* (in press)

[62] Matsuzaki Y, Uchikoga N, Ohue M, Shimoda T, Sato T, Ishida T, Akiyama Y. MEGADOCK 3.0: A high-performance protein–protein interaction prediction software using hybrid parallel computing for petascale supercomputing environments. *Source Code Biol Med*, 8(1): 18, 2013.

[63] Shimoda T, Ishida T, Suzuki S, Ohue M, Akiyama Y. MEGADOCK-GPU: Acceleration of Protein–protein Docking Calculation on GPUs. In *Procs of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine 2013 (ACM-BCB 2013), 2nd International Workshop on Parallel and Cloud-based Bioinformatics and Biomedicine (ParBio2013)*, 884–890, 2013.

[64] Levinthal C, Wodak SJ, Kahn P, Dadivanian AK. Hemoglobin Interaction in Sickle Cell Fibers. I: Theoretical Approaches to the Molecular Contacts. *Proc Natl Acad Sci USA*, 72: 1330–1334, 1975.

[65] Vajda S, Hall DR, Kozakov D. Sampling and scoring: A marriage made in heaven. *Proteins*, 81: 1874–1884, 2013.

[66] Ritchie DW. Recent progress and future directions in protein–protein docking. *Curr Protein Pept Sci*, 9: 1–15, 2008.

[67] Janin J. Protein–protein docking tested in blind predictions: the CAPRI experiment. *Mol Biosyst*, 6: 2351–2362, 2010.

[68] Vakser IA. Low-resolution structural modeling of protein interactome. *Curr Opin Struct Biol*, 23: 198–205, 2013.

[69] Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF. Protein docking using continuum electrostatics and geometric fit. *Protein Eng*, 14: 105–113, 2001.

[70] Roberts VA, Thompson EE, Pique ME, Perez MS, Ten Eyck LF. DOT2: Macromolecular docking with improved biophysical models. *J Comput Chem*, 34: 1743–1758, 2013.

[71] Vakser IA, Aflalo C. Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins*, 20: 320–329, 1994.

[72] Vakser IA. Protein docking for low-resolution structures. *Protein Eng*, 8: 371-377, 1995.

[73] Vakser IA, Matar OG, Lam CF. A systematic study of low-resolution recognition in protein–protein complexes. *Proc Natl Acad Sci U S A*, 96: 8477–8482, 1999.

[74] Hingerty BE, Ritchie RH, Ferrell TL, Turner JE. Dielectric effects in biopolymers: The theory of ionic saturation revisited. *Biopolymers*, 24: 427–439, 1985.

[75] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem*, 4: 187–217, 1983.

[76] Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins, *J Mol Biol*, 267: 707–726, 1997.

[77] Li L, Chen R, Weng Z. RDOCK: Refinement of rigid-body protein docking predictions. *Proteins*, 53: 693–707, 2003.

[78] Pierce B, Weng Z. ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins*, 67: 1078–1086, 2007.

[79] Chen R, Mintseris J, Janin J, Weng Z. A protein–protein docking benchmark. *Proteins*, 52: 88–91, 2003.

[80] Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. Protein–protein docking benchmark 2.0: an update, *Proteins*, 60: 214–216, 2005.

[81] Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein–protein docking benchmark version 3.0. *Proteins*, 73: 705–709, 2008.

[82] Hwang H, Vreven T, Janin J, Weng Z. Protein–protein docking benchmark version 4.0. *Proteins*, 78: 3111–3114, 2010.

[83] Frigo M, Johnson SG. The design and implementation of FFTW3. In *Procs of IEEE*, 93: 216–231, 2005.

[84] Pons C, Solernou A, Peréz-Cano L, Grosdidier S, Fernández-Recio J. Optimization of pyDock for the new CAPRI challenges: Docking of homology-based models, domain-domain assembly and protein–RNA binding. *Proteins*, 78: 3182–3138, 2010.

[85] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39: 561–577, 1993.

[86] Gao Y, Douguet D, Tovchigrechko A, Vakser IA. DOCKGROUND system of databases for protein recognition studies: unbound structures for docking. *Proteins*, 69: 845–851, 2007.

[87] Chaleil RAG, Tournier AL, Bates PA, Kro M. Implicit flexibility in protein docking: Cross-docking and local refinement. *Proteins* 69: 750–757, 2007.

[88] Andrusier N, Mashiach E, Nussinov R, Wolfson HJ. Principles of flexible protein–protein docking. *Proteins* 73: 271–289, 2008.

[89] Foynes S, Dorrell N, Ward SJ, Stabler RA, McColm AA, Rycroft AN, Wren BW. *Helicobacter pylori* possesses two CheY response regulators and a histidine kinase sensor, CheA, which are essential for chemotaxis and colonization of the gastric mucosa. *Infect Immun*, 68: 2016–2023, 2000.

[90] de Haas CJ, Veldkamp KE, Peschel A, Weerkamp F, Van Wamel WJ, Heezius EC, Poppelier MJ, Van Kessel KP, van Strijp JA. Chemotaxis inhibitory protein of *Staphylococcus aureus*, a bacterial antiinflammatory agent. *J Exp Med*, 199: 687–695, 2004.

[91] Tsuji T, Suzuki M, Takiguchi N, Ohtake H. Biomimetic control based on a model of chemotaxis in *Escherichia coli*. *Artif Life*, 16: 155–177, 2010.

[92] Baker MD, Wolanin PM, Stock JB. Systems biology of bacterial chemotaxis. *Curr Opin Microbiol*, 9: 187–192, 2006.

[93] Wadhams GH, Armitage JP. Making sense of it all: bacterial chemotaxis. *Nat Rev Mol Cell Biol*, 5: 1024–1037, 2004.

[94] Macnab R. The bacterial flagellum: reversible rotary propellor and type III export apparatus. *J Bacteriol*, 181: 7149–7153, 1999.

[95] Kentner D, Sourjik V. Dynamic map of protein interactions in the *Escherichia coli* chemotaxis pathway. *Mol Syst Biol*, 5: 238, 2009.

[96] Matsuzaki Y, Kikuchi S, Tomita M. Robust effects of Tsr-CheBp and CheA-CheYp affinity in bacterial chemotaxis. *Artif Intell Med*, 41: 145–150, 2007.

[97] van Albada SB, Ten Wolde PR. Differential affinity and catalytic activity of CheZ in *E. coli* chemotaxis. *PLoS Comput Biol*, 5: e1000378, 2009.

[98] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28: 27–30, 2000.

[99] Fujibuchi W, Goto S, Migimatsu H, Uchiyama I, Ogiwara A, Akiyama Y, Kanehisa M. DBGET/LinkDB: an integrated database retrieval system. *Pac Symp Biocomput*, 683–694, 1998.

[100] Bren A, Eisenbach M. How signals are heard during bacterial chemotaxis: protein–protein interactions in sensory signal propagation. *J Bacteriol*, 182: 6865–6873, 2000.

[101] Sarkar MK, Paul K, Blair D. Chemotaxis signaling protein CheY binds to the rotor protein FliN to control the direction of flagellar rotation in *Escherichia coli*. *Proc Natl Acad Sci U S A*, 107: 9370–9375, 2010.

[102] Blat Y, Eisenbach M. Oligomerization of the phosphatase CheZ upon interaction with the phosphorylated form of CheY. The signal protein of bacterial chemotaxis. *J Biol Chem*, 271: 1226–1231, 1996.

[103] Zhao R, Collins EJ, Bourret RB, Silversmith RE. Structure and catalytic mechanism of the *E. coli* chemotaxis phosphatase CheZ. *Nat Struct Biol*, 9: 570–575, 2002.

[104] Park SY, Chao X, Gonzalez-Bonet G, Beel BD, Bilwes AM, Crane BR. Structure and function of an unusual family of protein phosphatases: the bacterial chemotaxis proteins CheC and CheX. *Mol Cell*, 16: 563–574, 2004.

[105] Wang H, Matsumura P. Characterization of the CheAS/CheZ complex: a specific interaction resulting in enhanced dephosphorylating activity on CheY-phosphate. *Mol Microbiol*, 19: 695–703, 1996.

[106] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, 39: D561–568, 2011. `http://www.string-db.org`.

[107] Szurmant H, Muff TJ, and Ordal GW. *Bacillus subtilis* CheC and FliY are members of a novel class of CheY-P-hydrolyzing proteins in the chemotactic signal transduction cascade, *Biol Chem*, 279: 21787–21792, 2004.

[108] Rosario MM, Ordal GW. CheC and CheD interact to regulate methylation of Bacillus subtilis methyl-accepting chemotaxis proteins. *Mol Microbiol*, 21: 511–518, 1996.

[109] Chao X, Muff TJ, Park SY, Zhang S, Pollard AM, Ordal GW, Bilwes AM, Crane BR. A receptor-modifying deamidase in complex with a signaling phosphatase reveals reciprocal regulation. *Cell*, 124: 561–571, 2006.

[110] Ghobrial IM, Witzig TE, Adjei AA. Targeting apoptosis pathways in cancer therapy. *CA Cancer J Clin*, 55: 178–194, 2005.

[111] Oltersdorf T, Elmore SW, Shoemaker AR, Armstrong RC, Augeri DJ, Belli BA, Bruncko M, Deckwerth TL, Dinges J, Hajduk PJ, Joseph MK, Kitada S, Korsmeyer SJ, Kunzer AR, Letai A, Li C, Mitten MJ, Nettesheim DG, Ng S, Nimmer PM, O'Connor JM, Oleksijew A, Petros AM, Reed JC, Shen W, Tahir SK, Thompson CB, Tomaselli KJ, Wang B, Wendt MD, Zhang H, Fesik SW, Rosenberg SH. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature*, 435: 677–681, 2005.

[112] Kischkel FC, Hellbardt S, Behrmann I, Germer M, Pawlita M, Krammer PH, Peter ME. Cytotoxicity-dependent APO-1 (Fas/CD95)-associated proteins form a death-inducing signaling complex (DISC) with the receptor. *EMBO J*, 14: 5579–5588, 1995.

[113] Portt L, Norman G, Clapp C, Greenwood M, Greenwood MT. Anti-apoptosis and cell survival: A review. *Biochimica et Biophysica Acta (BBA) - Mol Cell Res*, 1813: 238–259, 2011.

[114] Acuner Ozbabacan SE, Keskin O, Nussinov R, Gursoy A. Enriching the human apoptosis pathway by predicting the structures of protein–protein complexes. *J Struct Biol*, 179: 338–346, 2012.

[115] Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31: 248–250, 2003.

[116] Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M. The BioGRID interaction database: 2013 update. *Nucleic Acids Res*, 41: D816–823, 2013.

[117] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 32: D449–451, 2004.

[118] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human Protein Reference Database–2009 update. *Nucleic Acids Res*, 37: D767–772, 2009.

[119] Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, 40: D841–846, 2012.

[120] Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res*, 38: D532–539, 2010.

[121] Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stümpflen V. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*, 34: D436–441, 2006.

[122] Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, Ruepp A, Frishman D. The MIPS mammalian protein–protein interaction database. *Bioinformatics*, 21(6): 832–834, 2005.

[123] Allsopp TE, McLuckie J, Kerr LE, Macleod M, Sharkey J, Kelly JS. Caspase 6 activity initiates caspase 3 activation in cerebellar granule cell apoptosis. *Cell Death Differ*, 7: 984–993, 2000.

[124] Edgington LE, van Raam BJ, Verdoes M, Wierschem C, Salvesen GS, Bogyo M. An optimized activity-based probe for the study of caspase-6 activation. *Chem Biol*, 19: 340–352, 2012.

[125] Slee EA, Adrain C, Martin SJ. Executioner caspase-3, -6, and -7 perform distinct, non-redundant roles during the demolition phase of apoptosis. *J Biol Chem*, 276: 7320–7326, 2001.

[126] Walsh JG, Cullen SP, Sheridan C, Lüthi AU, Gerner C, Martin SJ. Executioner caspase-3 and caspase-7 are functionally distinct proteases. *Proc Natl Acad Sci U S A*, 105: 12815–12819, 2008.

[127] Guerrero AD, Chen M, Wang J. Delineation of the caspase-9 signaling cascade. *Apoptosis*, 13: 177–186, 2008.

[128] Zhang QC, Petrey D, Garzon JI, Deng L, Honig B. PrePPI: a structure-informed database of protein–protein interactions. *Nucleic Acids Res*, 41: D828–833, 2013.

[129] Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, 490: 556–560, 2012.

[130] Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Rattei T, Frishman D, Ruepp A. The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res*, 38: D540–544, 2010.

[131] Sadidi M, Lentz SI, Feldman EL. Hydrogen peroxide-induced Akt phosphorylation regulates Bax activation. *Biochimie*, 91: 577–585, 2009.

[132] Yeretssian G, Correa RG, Doiron K, Fitzgerald P, Dillon CP, Green DR, Reed JC, Saleh M. Non-apoptotic role of BID in inflammation and innate immunity. *Nature*, 474: 96–99, 2011.

[133] Peréz-Cano L, Fernández-Recio J. Optimal protein–RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins*, 78: 25–35, 2010.

[134] Lejeune D, Delsaux N, Charloteaux B, Thomas A, Brasseur R. Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins*, 61: 258–271, 2005.

[135] Ellis JJ, Broom M, Jones S. Protein–RNA Interactions : Structural Analysis and Functional Classes. *Proteins*, 66: 903–911, 2007.

[136] Jeong E, Kim H, Lee S, Han K. Discovering the interaction propensities of amino acids and nucleotides from protein–RNA complexes. *Molecules and Cells*, 16: 161–167, 2003.

[137] Peréz-Cano L, Solernou A, Pons C, Fernández-Recio J. Structural Prediction of Protein–RNA Interaction by Computational Docking with Propensity-based Statistical Potentials. *Pacific Symp Biocomput*, 293–301, 2010.

[138] Keene JD. Ribonucleoprotein infrastructure regulating the flow of genetic information between the genome and the proteome. *Proc Natl Acad Sci U S A*, 98: 7018–7024, 2001.

[139] Wang G, Dunbrack RL. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res*, 33: W94–98, 2005.

[140] PISCES: A Protein Sequence Culling Server, `http://dunbrack.fccc.edu/PISCES.php`.

[141] MacKerell Jr AD, Banavali N, Foloppe N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, 56: 257–265, 2001.

[142] Fanelli F, Ferrari S. Prediction of MEF2A-DNA interface by rigid body docking: a tool for fast estimation of protein mutational effects on DNA binding. *J Struct Biol*, 153: 278–283, 2006.

[143] Yi H, Qiu S, Cao Z, Wu Y, Li W. Molecular basis of inhibitory peptide maurotoxin recognizing Kv1.2 channel explored by ZDOCK and molecular dynamic simulations. *Proteins*, 70: 844–854, 2010.

[144] Krummenacker M, Paley S, Mueller L, Yan T, Karp PD. Querying and computing with BioCyc databases. *Bioinformatics*, 21: 3454–3455, 2005.

[145] Kastritis PL, Bonvin AM. Are scoring functions in protein–protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J Proteome Res*, 9: 2216–2225, 2010.

[146] Sourjik V, Berg HC. Localization of components of the chemotaxis machinery of Escherichia coli using fluorescent protein fusions. *Mol Microbiol*, 37: 740–751, 2000.

[147] Nakai K, Horton P. PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24: 34–35, 1999.

[148] Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14: 378–379, 1998.

[149] Parkinson JS. Bacterial chemotaxis: a new player in response regulator dephosphorylation. *J Bacteriol*, 185: 1492–1494, 2003.

[150] Welch M, Oosawa K, Aizawa S, Eisenbach M. Phosphorylation-dependent binding of a signal molecule to the flagellar switch of bacteria. *Proc Natl Acad Sci U S A*, 90: 8787–8791, 1993.

[151] Lee SY, Cho HS, Pelton JG, Yan D, Berry EA, Wemmer DE. Crystal structure of activated CheY. Comparison with other activated receiver domains. *J Biol Chem*, 276: 16425–16431, 2001.

[152] Sourjik V. Receptor clustering and signal processing in *E. coli* chemotaxis. *Trends Microbiol*, 12: 569–576, 2004.

[153] Hess JF, Oosawa K, Kaplan N, Simon MI. Phosphorylation of three proteins in the signaling pathway of bacterial chemotaxis. *Cell*, 53: 79–87, 1998.

[154] Tuncbag N, Kar G, Gursoy A, Keskin O, Nussinov R. Towards inferring time dimensionality in protein–protein interaction networks by integrating structures: the p53 example. *Mol Biosyst.* 5: 1770–1778, 2009.

[155] Zhou H, Pandit SB, Skolnick J. Performance of the Pro-sp3-TASSER server in CASP8. *Proteins*, 77: 123–127, 2009.

[156] Saini HK, Fischer D. Meta-DP: domain prediction meta-server. *Bioinformatics*, 21: 2917–2920, 2005.

[157] Ishida T, Kinoshita K. Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, 24: 1344–1348, 2008.

[158] Hubbard SJ, Thornton JM. Naccess. Department of Biochemistry and Molecular Biology, University College London, 1993.

[159] Shatsky M, Nussinov R, Wolfson HJ. A method for simultaneous alignment of multiple protein structures. *Proteins*, 56: 143–156, 2004.

[160] Barabási AL, Albert R. Emergence of scaling in random networks. *Science*, 286: 509–512, 1999.

[161] Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3(4): e59, 2007.

[162] Uchikoga N, Hirokawa T. Analysis of protein–protein docking decoys using interaction fingerprints: application to the reconstruction of CaM-ligand complexes. *BMC Bioinformatics*, 11: 236, 2010.

[163] Uchikoga N, Matsuzaki Y, Ohue M, Hirokawa T, Akiyama Y. Re-docking scheme for generating near-native protein complexes by assembling residue interaction fingerprints. *PLoS ONE*, 8: e69365, 2013.

[164] Tsai CJ, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels, and protein function. *Protein Sci*, 8: 1181–1190, 1999.

[165] Wako H, Kato M, Endo S. ProMode: A database of normal mode analyses on protein molecules with a full-atom model. *Bioinformatics*, 20: 2035–2043, 2004.

[166] Wako H, Endo S. Ligand-induced conformational change of a protein reproduced by a linear combination of displacement vectors obtained from normal mode analysis. *Biophys Chem*, 159: 257–266, 2011.

[167] Selvaggi G, Novello S, Torri V, Leonardo E, De Giuli P, Borasio P, Mossetti C, Ardissone F, Lausi P, Scagliotti GV. Epidermal growth factor receptor overexpression correlates with a poor prognosis in completely resected non-small-cell lung cancer, *Annals of Oncology*, 15: 28–32, 2004.

[168] Tianhe-2 (National Supercomputer Center in Guangzhou), `http://www.nscc-gz.cn`.

[169] Titan (Oak Ridge National Laboratory, Oak Ridge Leadership Computing Facility) `http://www.olcf.ornl.gov/titan`.

[170] FFTE: a fast Fourier transform package, `http://www.ffte.jp`.

[171] NVIDIA CUDA - NVIDIA developer zone, `http://developer.nvidia.com/cuda/`.

[172] Suzuki S, Ishida T, Kurokawa K, Akiyama Y. GHOSTM: a GPU-accelerated homology search tool for metagenomics, *PLoS ONE*, 7: e36060, 2012.

[173] van Meel JA, Arnold A, Frenkel D, Portegies Zwart SF, Belleman RG, Harvesting graphics power for MD simulations. *Mol Simul*, 34: 259–266, 2008.

[174] Ufimtsev IS, Martínez TJ. Quantum chemistry on graphical processor units. 1. Strategies for two-electron integral evaluation. *J Chem Theory Comput*, 4: 222–231, 2008.

[175] Sukhwani B, Herbordt MC. GPU acceleration of a production molecular docking code, In *Procs of 2nd Workshop on General Purpose Processing on Graphics Processing Units (GPGPU-2)*, 19–27, 2009.

[176] CUFFT - NVIDIA developer zone, `http://developer.nvidia.com/cufft/`.

[177] Kozakov D, Hall DR, Beglov D, Brenke R, Comeau SR, Shen Y, Li K, Zheng J, Vakili P, Paschalidis IC, Vajda S. Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13–19, *Proteins*, 78: 3124–3130, 2010.

List of Publications

Journal Papers

1. **Masahito Ohue**†, Yuri Matsuzaki†, Nobuyuki Uchikoga, Takashi Ishida, Yutaka Akiyama. MEGADOCK: An all-to-all protein–protein interaction prediction system using tertiary structure data. *Protein & Peptide Letters*, 2014. (in press) (†: These authors have equally contributed.)

2. Yuri Matsuzaki†, **Masahito Ohue**†, Nobuyuki Uchikoga, Yutaka Akiyama. Protein–protein interaction network prediction by using rigid-body docking tools: application to bacterial chemotaxis. *Protein & Peptide Letters*, 2014. (in press) (†: These authors have equally contributed.)

3. **Masahito Ohue**, Yuri Matsuzaki, Takehiro Shimoda, Takashi Ishida, Yutaka Akiyama. Highly Precise Protein–Protein Interaction Prediction Based on Consensus Between Template-Based and *de Novo* Docking Methods. *BMC Proceedings*, 7(Suppl 7): S6, 2013.

4. Yuri Matsuzaki, Nobuyuki Uchikoga, **Masahito Ohue**, Takehiro Shimoda, Toshiyuki Sato, Takashi Ishida, Yutaka Akiyama. MEGADOCK 3.0: A high-performance protein–protein interaction prediction software using hybrid parallel computing for petascale supercomputing environments. *Source Code for Biology and Medicine*, 8(1): 18, 2013.

5. Nobuyuki Uchikoga, Yuri Matsuzaki, **Masahito Ohue**, Takatsugu Hirokawa, Yutaka Akiyama. Re-docking scheme for generating near-native protein complexes by assembling residue interaction fingerprints. *PLoS ONE*, 8(7): e69365, 2013.

6. Sarel J. Fleishman, Timothy A. Whitehead, Eva-Maria Strauch, Jacob E. Corn, Sanbo Qin, Huan-Xiang Zhou, Julie C. Mitchell, Omar N. A. Demerdash, Mayuko Takeda-Shitaka, Genki Terashi, Iain H. Moal, Xiaofan Li, Paul A. Bates, Martin Zacharias, Hahnbeom Park, Jun-su Ko, Hasup Lee, Chaok Seok, Thomas Bourquard, Julie Bernauer, Anne Poupon, Jerome Aze, Seren Soner, Sefik Kerem Oval., Pemra Ozbek, Nir Ben Tal, Turkan Haliloglu, Howook Hwang, Thom

Vreven, Brian G. Pierce, Zhiping Weng, Laura Perez-Cano, Carles Pons, Juan Fernandez-Recio, Fan Jiang, Feng Yang, Xinqi Gong, Libin Cao, Xianjin Xu, Bin Liu, Panwen Wang, Chunhua Li, Cunxin Wang, Charles H. Robert, Mainak Guharoy, Shiyong Liu, Yangyu Huang, Lin Li, Dachuan Guo, Ying Chen, Yi Xiao, Nir London, Zohar Itzhaki, Ora Schueler-Furman, Yuval Inbar, Vladimir Patapov, Mati Cohen, Gideon Schreiber, Yuko Tsuchiya, Eiji Kanamori, Daron M. Standley, Haruki Nakamura, Kengo Kinoshita, Camden M. Driggers, Robert G. Hall, Jessica L. Morgan, Victor L. Hsu, Jian Zhan, Yuedong Yang, Yaoqi Zhou, Panagiotis L. Kastritis, Alexandre M. J. J. Bonvin, Weiyi Zhang, Carlos J. Camacho, Krishna P. Kilambi, Aroop Sircar, Jeffrey J. Gray, **Masahito Ohue**, Nobuyuki Uchikoga, Yuri Matsuzaki, Takashi Ishida, Yutaka Akiyama, Raed Khashan, Stephen Bush, Denis Fouches, Alexander Tropsha, Juan Esquivel-Rodriguez, Daisuke Kihara, P. Benjamin Stranges, Ron Jacak, Brian Kuhlman, Sheng-You Huang, Xiaoqin Zou, Shoshana J. Wodak, Joel Janin and David Baker. Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology. *Journal of Molecular Biology*, 414(2): 289–302, 2011.

7. **Masahito Ohue**, Yuri Matsuzaki, Yutaka Akiyama. Docking-Calculation-Based Method for Predicting Protein–RNA Interactions. *Genome Informatics*, 25(1): 25–39, 2011.

8. **Masahito Ohue**, Yuri Matsuzaki, Yusuke Matsuzaki, Toshiyuki Sato, Yutaka Akiyama. MEGADOCK: an all-to-all protein–protein interaction prediction system using tertiary structure data and its application to systems biology study. *IPSJ Transactions on Mathematical Modeling and Its Applications*, 3(3): 91–106, 2010. (in Japanese)

International Conference Proceedings

1. Takehiro Shimoda, Shuji Suzuki, **Masahito Ohue**, Takashi Ishida, Yutaka Akiyama. A Comparative Study of GPU and MIC Accelerations in Protein–Protein Docking Calculation. In *Proceedings of European Conference on Parallel Computing 2014 (Euro-Par 2014)*. (submitted)

2. Takehiro Shimoda, Takashi Ishida, Shuji Suzuki, **Masahito Ohue**, Yutaka Akiyama. MEGADOCK-GPU: Acceleration of Protein–Protein Docking Calculation on GPUs. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine 2013 (ACM-BCB 2013), 2nd International Workshop on Parallel and Cloud-based Bioinformatics and Biomedicine (ParBio2013)*, 884–890, 2013.

3. **Masahito Ohue**, Yuri Matsuzaki, Takehiro Shimoda, Takashi Ishida, Yutaka Akiyama. Highly Precise Protein–Protein Interaction Prediction Based on Consensus Between Template-Based and *de Novo* Docking Methods. In *Proceedings of Great Lakes Bioinformatics Conference 2013 (GLBIO2013)*, 100–109, 2013.

4. **Masahito Ohue**, Yuri Matsuzaki, Takashi Ishida, Yutaka Akiyama. Improvement of the Protein–Protein Docking Prediction by Introducing a Simple Hydrophobic Interaction Model: an Application to Interaction Pathway Analysis. In *Proceedings of The 7th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB2012), Lecture Notes in Bioinformatics* , 7632: 178–187, 2012.

Honors and Awards

1. Fourth (FY2013) JSPS* *Ikushi* Prize (2014)
2. JSPS Research Fellow (DC1) (2011–2014)
3. Grant-in-Aid for JSPS Fellows (23·8750) (2011–2014)
4. The 73rd National Convention of IPSJ** Student Award (2011)
5. The 78th IPSJ SIGMPS Presentation Award (2010)
6. 2009 IPSJ SIGBIO Best Student Presentation Award (2010)
7. The Second Forum on Data Engineering and Information Management (DEIM2010) Excellent Student Presentation Award (2010)

*JSPS: Japan Society for the Promotion of Science
**IPSJ: Information Processing Society of Japan