Spring 2014

Prediction Of The Protein Complex Assembly Pathway Using Multiple Docking Algorithm

Yoichiro Togawa
Purdue University

Graduate School ETD Form 9
(Revised 01/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Yoichiro Togawa

Entitled
PREDICTION OF THE PROTEIN COMPLEX ASSEMBLY PATHWAY USING MULTIPLE DOCKING ALGORITHM

For the degree of Master of Science

Is approved by the final examining committee:

Daisuke Kihara

Barry L. Wanner

Cynthia V. Stauffacher

To the best of my knowledge and as understood by the student in the *Thesis/Dissertation Agreement. Publication Delay, and Certification/Disclaimer (Graduate School Form 32)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Daisuke Kihara

Approved by Major Professor(s): _____

_____

Approved by: Richard J. Kuhn 04/15/2014

Head of the Department Graduate Program Date

PREDICTION OF THE PROTEIN COMPLEX ASSEMBLY PATHWAY USING MULTIPLE DOCKING ALGORITHM

A Thesis

Submitted to the Faculty

of

Purdue University

by

Yoichiro Togawa

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Biological Sciences

May 2014

Purdue University

West Lafayette, Indiana

For my parents, wife, and daughter.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Daisuke Kihara, for his support and instruction during my research at Kihara lab. He allowed me to join the lab at the second year of my thesis Master program, despite the lack of my programming and bioinformatics background. Although being extremely busy, Dr. Kihara patiently taught me the ideas that I didn't understand. I greatly appreciate his support, and I also admire his passion to science.

I would like to thank members of my research advisory committee, Professors Barry L. Wanner and Cynthia V. Stauffacher, who game me advice in conducting research, and also supported me to do research at Purdue.

I would like to thank all the people in Kihara lab, who gave me invaluable advice in research, coding, and presentation: Atilla Sit, Bingjie Hu, Hyung Rae Kim, Ishita Khan, Juan Esquivel-Rodoriguez, Kristen Johnson, Lenna Peterson, Lyman Monroe, Shuvra Nath, Xiaolei Zhu, Xin Cheng, Xuejiao Kang, and Yi Xiong. Especially, I would like to thank Juan XXX, who have directly supervised my research, taught me coding, provided me with new idea to my research, and helped me organize and analyze the results. If it were not for his help, I was unable to do a research in this lab.

I would like to thank government of Japan government and my colleagues in Japan, who have allowed me to leave my work and supported my stay at Purdue.

For closing, I'd like to dedicate this thesis to my parents, Toru Togawa and Tomoko Togawa, and my wife and daughter, Yuki Togawa and Haruka Togawa. I would like to thank my parents for bringing me up. I appreciate for their love and effort for bringing me up along with three younger brother and sisters. I would like to thank my wife for taking care of our daughter while I was away. She took good care of our daughter while she was very busy with her work. I appreciate for her love to our daughter.

TABLE OF CONTENTS

Page

Page

ABSTRACT

Togawa, Yoichiro. M.S., Purdue University, May 2014. Prediction of the Protein Complex Assembly Pathway Using Multiple Docking Algorithm. Major Professor: Daisuke Kihara.

Proteins often function as a complex of multiple subunits, and the quaternary structure is important for proper function. An ordered assembly pathway is one of the strategies nature has developed to obtain the correct conformation: studies have shown a relationship between the assembly pathway and evolution of protein complexes. Identification of the assembly pathway and the intermediate structures helps drug development as well. Therefore, elucidation of the assembly pathway of protein complexes is important for understanding biochemical processes central to cellular function. Recent studies have demonstrated the assembly pathway of a protein complex can be predicted from its crystal structure by comparing the buried surface area (BSA) between each subunit. To our knowledge, this is the first and only work that has predicted the assembly pathways of protein complexes from their structure.

In this work, we have developed four methods to predict the assembly pathway from the output of Multi-LZerD, a multiple docking algorithm for asymmetric protein complexes. We found that data from Multi-LZerD predicted

not only the model of the complex but also suggested how the complex is assembled. The four methods were benchmarked, along with the BSA-based method, using a dataset of manually-curated protein complexes. In contrast with the data set used in the BSA-based method, which only contained homomeric and symmetric complexes, our data set includes asymmetric complexes varying in size, topology, and number of subunits. We confirmed that the BSA based-method also worked with asymmetric complexes as they predict the correct pathway in 68% of the cases in our data set. Although the success rate of our methods ranges from 40% to 52%, it improved to as high as 82% for the complexes where Multi-LZerD was successful in modeling near native structures. The results also showed that our method is capable of capturing some of the dimerization events in the assembly pathway, even if the overall pathway prediction was failing. Additionally, there was a case where the BSA-based method failed, but our method was successful, suggesting the limitations in the BSA-based method. These results demonstrate the ability of a multiple docking algorithm to predict the assembly pathway of protein complexes.

CHAPTER 1. BACKGROUND

1.1 Introduction: the importance of protein complexes and its assembly

Proteins carry out various functions that are crucial to life. Proteins are components of cells, a generator of energy, and molecular machines that grow and replicate cells. Importance of proteins was recognized from the dawn of molecular biology, gathering huge interest in elucidating their functions. The one gene - one enzyme hypothesis by Beadle and Tatum [1] led to the birth and development of molecular genetics, which provided scientists various tools to manipulate genome sequences. Scientists were able to elucidate the function of genes or proteins by genetic experiments, for example, by knocking out the gene of interest. After annotating functions, people's interests will shift towards how proteins carry out their function. Because protein structures define its function [2], huge efforts are being made in solving protein structures. The advent of X-ray crystallography was important not only for the discovery of the double helical structure of the DNA [3], but also for determining huge number of protein structures. Solved structures are deposited to the Protein Data Bank (PDB) [4] and they are freely available to the research community. Due to the development of other structure solving methods, such as nuclear magnetic resonance and electron microscopy, and also the results of structural genomics projects [5], the

number of PDB entries has been rapidly growing, having over 97,000 structures as of February, 2014. These structures provide detailed information of how proteins interact with their ligands, the atomic level of interaction, and the mechanism of biochemical reaction. Proteins may interact with each other or with another protein to form a multi-subunit complex, and such complexes constitute significant portion of the proteins in cell. In the case of *Escherichia coli*, monomers only consist one fifth of the protein species in the cell [6]. Protein oligomerization may be an advantage in the evolution of protein by obtaining new features [7], such as allosteric control of oxygen binding in hemoglobin. The intricate function of the large complexes, such as ribosomes and RNA polymerases, would not have been able without the formation of the complex. However, solving the structure of a large protein complex is challenging and these structures were not available until recently [8]. Before the development of structure determination techniques, scientists used biochemical experiments, such as yeast two-hybrid and co-immunoprecipitation (Co-IP), to construct topology of proteins to make estimations of how multi-subunit protein complexes are structured, which is exemplified by the researches done in the past for Arp2/3 complex. Arp2/3 complex, a protein complex consisting of 7 unique protein subunits, has an important role in actin nucleation and branching. Before its first structure was deposited to the PDB in 2001 [9], several biochemical experiments were done to reveal the interaction among the subunits and the role of each subunit in the activity and formation of the complex [10–13]. Head module of the mediator, a transcriptional co-activator, is another such example; researchers did

biochemical experiments to construct topology of the complex [14–16] before it was first crystalized [17]. These experiments allowed researchers to know the stoichiometry of the complex, which protein subunits are interacting, and what functions they have. Combination of the results from the biochemical experiments helps us understand not only how the protein is structured, but also how it is assembled.

Research have shown that numbers of protein complexes assemble via ordered pathway, implicating the importance of the pathway for further understanding of the complexes [18–22]. Therefore, even if the structure of a protein is solved, understanding the mechanism of its assembly itself is still an important scientific question. Teichmann *et al*. have shown that the assembly pathways of a protein could be predicted solely from its crystal structure [20, 21]. As the development of protein structure prediction methods are complementing the limitation of protein structure determination, having computational methods to predict assembly pathway of a given complex should benefit the research community in the same manner. This introduction will cover why some proteins assemble via ordered pathway, how the pathways are determined experimentally, and the motivation of this project.

1.2 Protein have ordered pathway of assembly

Why would a protein assemble in an ordered pathway, rather than assembling randomly? In the field of protein folding, it is now widely believed that protein folding proceeds through energetically favorable pathway [23]; random

search of correct fold will take forever and proteins are unable to fold into correct conformation in biological timescale [24]. Therefore, it is natural for proteins to adopt the same strategy to assemble into the complex both efficiently and correctly.

Assembly mechanisms of membrane proteins are reviewed, and they suggest the benefits of having ordered assembly pathway [18, 19]. First, having ordered pathway of assembly may help proteins to assemble correctly by preventing the aggregation and production of off-pathway subcomplexes. Formation of non-functional complexes is a waste of energy and also a potential threat for the cell survival. These complexes may lead to the misassembly and aggregation that may result in serious consequences [25, 26]. An ordered pathway of assembly is suggested for F_1F_0 ATP synthase [18, 19, 27] (Figure 1.1); F_0, F_1, and stator subunits are assembled independently before forming the functional complex. Because the proton channel of the ATP synthase is formed at the last step of assembly, the ordered pathway is likely to be preventing uncontrolled proton diffusion across membrane [27].

Second, ordered pathway of complex assembly enables cells to perform systematic process. Divisome of Escherichia coli (E. coli), a protein complex that is in charge of cell division, is known to form via an ordered pathway [19]. Cell division must be precisely controlled in order to divide the cell at the right time and location, which involves a series of different reactions. The sequential recruitment may reflect the series of enzymatic reactions that takes place at the site of cell division [19]. The similar phenomenon is observed with protein

complexes that are not assembled on membrane; Kinetochore [28, 29] and RNA spliceosome [30, 31] are other examples of proteins being sequentially recruited at the site where the protein complexes perform their function.

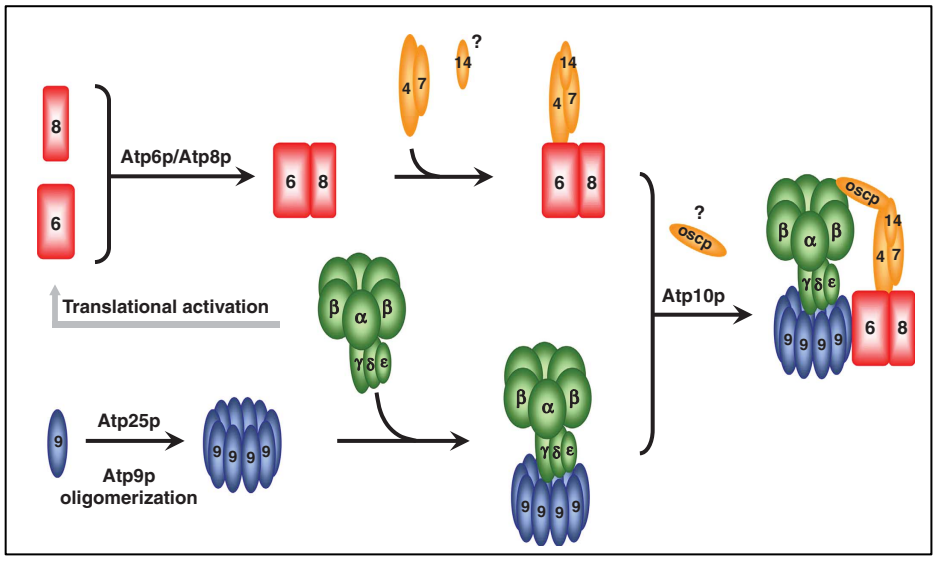


Figure 1.1 Assembly of ATP synthase. Taken from [27]

Having the ordered assembly pathway and having a single assembly pathway is not mutually exclusive. Indeed, promiscuous assembly pathway is observed on maltose transporter. Maltose transporter is a 4-chain complex, consisting of MalF, MalG, and two copies of MalK. Experimental data shows that $MalFGK_2$ can be assembled into correct final complex in multiple pathways [32]. Having multiple pathways could be an advantage in terms of rapid formation of the final complex, because various subcomplexes can be rescued in the pathways, rather than becoming dead-end products. Assembly of huge and complex structure like flagella definitely requires highly ordered pathway and machinery for the assembly [33], but simple complexes may benefit from having multiple pathways. Multiple pathways provide an advantage for the protein to

assemble quickly if the assembly intermediates are having no harmful effects. Although it is tempting to make a statement that whether a protein adopts an ordered pathway or not is dependent on the complexity of the final structure and its function, it is in a realm of mere speculation.

Why would understanding protein's assembly pathway be important? As described at the beginning of this section, elucidating assembly pathway may help us understand the mechanism of how protein complexes function. Also, Teichmann et al have shown that the assembly pathways are conserved in protein evolution [20, 21, 34]. They investigated the relationship between evolution and assembly pathway of proteins by looking at gene fusion. They used the proteins where the two genes encoding a pair of interacting subunit in the complex are known to get fused in the homolog of the protein in other organism. Their result showed that gene fusions occur in a manner that conserves the assembly pathway [21].

The relation between assembly pathway and evolution is also discussed for F_1F_0 ATP synthase [27, 35]. Because F_1 and F_0 subunits are highly likely to have evolved from DNA/RNA helicase and membrane channel respectively [35–37] and they are assembled in the separate pathways before associating with each other, the assembly pathway of the ATP synthase is likely to be recapitulating the evolutional events of the protein [27]. These studies imply that the elucidation of assembly pathway of protein complexes will help us understand the evolutionary path of the protein.

The assembly pathway of protein complexes can also provide invaluable information in the field of drug design. Prevention of the assembly of disease-related protein is one possible cure or prevention of the disease. Therefore, identification of assembly pathway and the assembly intermediates could lead to the discovery of new drug targets. Cholera toxin is such an example. Cholera toxin is a 6-chain complex, consisting of five B subunits and one a subunit. The B subunits form a pentamer ring structure, and funnel-like A subunit is fitted in the ring (Figure 4.19). An experiment have revealed that A subunit is unable to interact with the fully assembled pentamer ring, and that the presence of A subunit promotes the assembly of the pentamer ring [38]. Further research have identified the importance of the hydrophobic interaction between A subunit and B subunit in the holotoxin assembly, and a compound that bind to the hydrophobic region of the pentamer ring pore was found by structural study [39]. Therefore, understanding the assembly pathway of protein complexes has profound importance in understanding biochemical processes central to cellular functions.

1.3 Elucidating the assembly pathway

Various experiments, often in combination, are being used to understand how protein complexes are assembled. Different approach may suggest different assembly pathway, and we see this problem quit often in the case of in vitro experiments. Also, single experiment may suggest not only one, but several assembly pathways for a protein complex. This section briefly introduces the examples of the assembly pathways proposed for some protein complexes.

1.3.1 Elucidating the assembly pathway by experiment

The basic approach to study assembly pathway of a protein complex is the identification of its assembly intermediates. Existence of such subcomplexes may suggest one of the pathway the complex takes for its assembly [22]. Yeast two-hybrid and Co-IP are the commonly used tools for the characterization of subcomplexes formed by the subunits. Yeast two-hybrid assay is a classic, but powerful experiment that allows researchers to detect protein-protein interaction [40]. The assembly pathway of Arp2/3 complex proposed by Zhao et al. is based on the systematic yeast two-hybrid assay [11]. They first identified the pairwise interaction that occur among the subunits, and then the interactions between those dimers and other subunits. The subcomplexes identified by their analysis have good agreement with the Co-IP experiment [10]. Co-IP, sometimes referred to as pull-down assay, is useful in identifying all the subunits interacting with the target subunit both directly and indirectly. By conducting Co-IP assay with various combinations of subunits, one can gather information about interaction beyond dimerization. Assembly pathway of mediator head module is proposed [17], based on the comprehensive Co-IP assay that have revealed the subcomplexes formed by the subunits [14].

Recent development in mass spectrometry (MS) has provided researchers powerful and versatile means of analyzing protein samples. Analysis of protein complex by electrospray ionization mass spectroscopy (ESI-MS) can provide valuable information, such as interaction among subunits, stoichiometry, binding affinity, and conformation of the protein [41, 42]. Teichman et al. used ESI-MS to

identify the subcomplexes that are formed upon dissociation of the complex. Based on the fact that they were able to reassemble the original complex from the dissociated proteins without the formation of off-pathway subcomplexes, they concluded that the assembly pathways are the opposite of the disassembly [20, 21].

A crystal structure can also provide information of how a complex is assembled. The assembly pathway suggested for HypCDE complex is such an example [43]. HypCDE is a hexameric complex that is formed in the process of [NiFe] hydrogenase maturation (Figure 2.4, 4.19). The assembly starts with the formation of three dimers, HypE homodimer and two HypCD heterodimers, followed by the association of the three dimers. The crystal structure shows that each HypE has interface with both HypC and HypD. Their structural analysis revealed that a loop in HypC is interacting with HypE. Because the position of the loop is stabilized by the dimerization of HypC with HypD, they concluded that HypC and HypD dimerization takes place before the association with HypE. This is consistent with their pull-down assay and size exclusion chromatography, which showed that HypE alone is not capable of forming a complex with neither HypE nor HypD [43].

1.3.2 Prediction of the assembly pathway of protein complexes

Bioinformatics have made a significant progress in the field of protein research, such as structure prediction, protein folding, and protein docking. However, to our knowledge, Teichmann et al. are the first and only group that

demonstrated assembly pathway of a protein complex can be predicted from its structure [20, 21]. The basic idea is that the protein-protein interface with a large buried surface area (BSA) is more likely to be formed earlier in the assembly pathway than that with a smaller BSA. In other words, the larger the BSA of an interface is, the earlier its formation is in the assembly pathway. A BSA is defined as the surface area that is not accessible to solvent after binding, and the size of a BSA has a correlation with actual binding affinity of proteins [44]. Figure 1.2 shows how a BSA is calculated [44], where SASA stands for Solvent Accessible Surface Area [45].

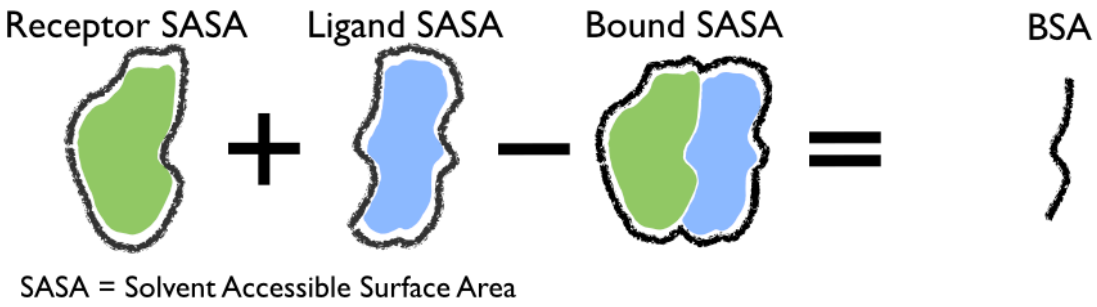


Figure 1.2 Calculation of BSA

They have shown that their prediction of assembly pathways have good agreement with the pathway obtained from their mass spectroscopy experiments [21]. Although their dataset are limited to homomeric [20] and symmetric protein complexes [21], their finding is valuable because research have shown the prevalence of homomeric and symmetric protein complexes in nature [6, 34, 46, 47]. However, asymmetric heteromers play crucial roles in shaping and sustaining life. For example, protein complexes that are found in the transcription

of DNA to RNA, and the translation of RNA to protein, are having asymmetric structures: ribosomes, DNA and RNA polymerases, pre-initiation complex, and RNA spliceosome to name a few. Also, we encounter other such complexes that have important function in cells, such as Arp2/3 complex and ATP synthase described before. Because the data set used by Teichmann et al. contained only homomeric and symmetric complexes, it is unclear if the BSA method is capable for predicting the assembly pathways of asymmetric heteromers. Also, the BSA method requires the structure of fully assembled complex, which is not always available at PDB.

1.4 Using data from Multi-LZerD for assembly pathway prediction

Multi-LZerD, developed by Juan Esquivel Rodriguez at Kihara Lab, is an algorithm for docking multiple proteins [48]. Because pairwise docking itself is a challenging task, not many algorithms are available for multiple docking. Multi-LZerD is capable of performing asymmetric multiple docking without any additional structural information, such as structure symmetry. Multi-LZerD works in two steps: all the possible combination of pairwise docking within the complex is done at the first step, followed by the structure search by the combination of the pairwise models, which we refer to as decoys, generated at the first step. The decoys generated at the first step is evaluated and ranked by the scoring function described in Chapter 2. The generated pairwise models are randomly combined to construct the full complexes, which are then evaluated in another scoring function. We can see which pairwise models were used in terms of the rank.

Although the rank was only used for Multi-LZerD to refer to each pairwise model, this rank seemed to be suggesting how the protein complexes are assembled. The detail of this algorithm is explained in Chapter 2. Literature survey of the Multi-LZerD dataset complexes suggested the ability for Multi-LZerD to predict the assembly pathways of protein complexes, which lead to the project of prediction of the assembly pathway of protein complex using Multi-LZerD. Because BSA method is the only method currently available for the prediction of protein assembly pathways from the crystal structures in PDB, providing another option in this field will benefit the research community.

CHAPTER 2. MATERIALS AND METHODS

2.1 How Multi-LZerD works

As briefly explained in chapter 1, Multi-LZerD works in two steps, pairwise docking and structure search as shown in Figure 2.1. Because this project utilizes the intermediate data of Multi-LZerD, explanation must be given on how the data are obtained and how the data look. The following subsections will briefly explain how Multi-LZerD works, where the data come from, and how the data are processed for the prediction of the assembly pathway. The dataset of protein complexes used in the project is described separately in Chapter 3.

2.1.1 Pairwise docking by LZerD

The first step of Multi-LZerD is pairwise docking by Local 3D Zernike descriptor-based Docking program, named LZerD, which was developed in Kihara Lab [48]. It uses 3D-Zernike descriptor (3DZD) to capture shape of protein surface and to evaluate the complementarity. For the details of the algorithm, please refer to the original article reporting LZerD.

The structure of two proteins, receptor and ligand, are the input of LZerD. First, LZerD will create points on the surface of each protein that are evenly distributed. Surface normal and 3DZD are calculated for each point, which are

used later for evaluating surface shape complementarity. Here, the surface normal is a vector that is orthogonal about the plane at each surface point created earlier. The point patterns from receptor and ligand are matched, and a score is given to each match. The score consists of four elements: angle between the surface point normals, correlation of 3D-Zernike descriptor of the points, size of the interface defined as buried surface area, and the excluded volume, which represents the atoms that are too close to each other. These elements make up the four terms in the scoring function that evaluate docking models. The first two elements, which represent the shape complementarity, are combined and represented in reward and penalty term. BSA represents the extent of surface overlap, which is not considered in the evaluation of shape complementarity, makes the 3rd term. Atoms that are close to each other have repulsive effect, and the effect is measured as excluded volume, which is incorporated in the scoring function as the 4th term. These four terms are linearly combined with weighting factors that are obtained by training the algorithm using a set of proteins obtained from ZDOCK benchmark 0.0 and 1.0, the set of protein structures that are commonly used to test the accuracy of protein docking algorithms [49]. Using the shape-based scoring function described above, LZerD gives score to each of the created docking models or decoy, and rank them with the score. In the field of protein docking, researchers try to get the near native structure ranked top among the other predictions. The score given to each decoy using this scoring function is referred to as shape score hereafter.

At the first step of Multi-LZerD, LZerD performs pairwise docking of all the possible pairwise combinations of subunits in a protein complex. For a four-subunit protein complex A, B, C, D for example, there are six possible combinations of the two subunits: A-B, A-C, A-D, B-C, B-D, and C-D. LZerD is done for all the six pair and the top 54,000 decoys are kept after sorting by the shape score. The top 54,000 decoys are further clustered at 5 Å and 10 Å threshold; all the decoys that are close to each other within each threshold are clustered together, and each cluster is represented by the decoys that have the best shape score. The clustering allows Multi-LZerD to perform the structure search effectively by reducing the numbers of decoys in similar conformation. The clustering completes the first step of Multi-LZerD and the clustered decoys are sorted according to their shape score.

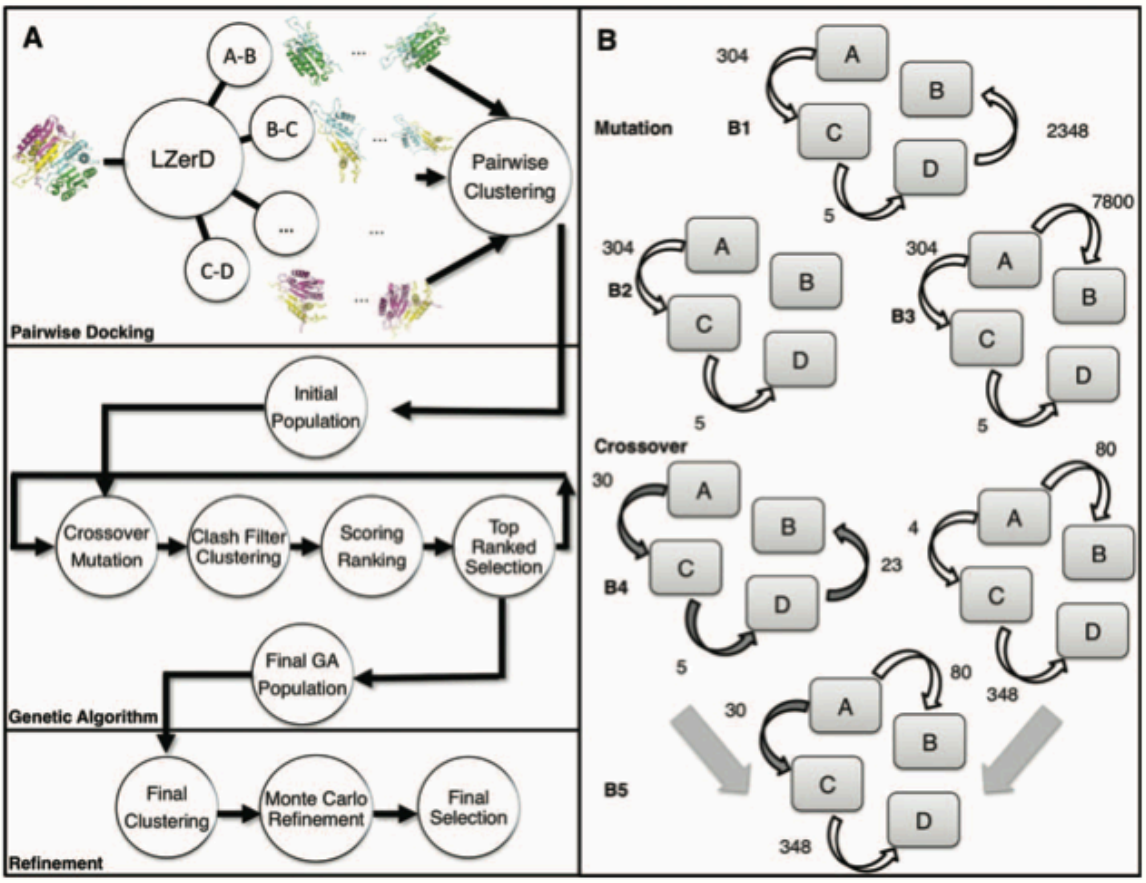


Figure 2.1 Multi-LZerD algorithm taken from [48]

(A) Overview of the algorithm. Upper and middle panel shows the diagram for pairwise docking and structure search respectively. The refinement step described at the bottom was not used in this project. (B) Protein complex in spanning tree representation. Each box denotes subunit of the complex, and the arrows connecting the two boxes define the decoy, i.e., the conformation of the two subunits. Crossover operation was not used in this project.

2.1.2 Structure search by genetic algorithm

The second step of Multi-LZerD is the structure search based on genetic algorithm. The pairwise models generated at the first step are randomly combined to construct the complex structure. Each complex structure is represented in a spanning tree, a graph without any cycle within, where each node and edge representing protein subunit and pairwise models respectively (Figure 2.1B). At the beginning, M numbers of these structures are randomly generated. M is the size of the population, which is set as 200 throughout this project. Each clustered decoy will go through mutation, which is a random replacement of edge in the spanning tree. 2M, namely 400, such operations are done and the resulting population will be clustered with the desired cutoff, which was set to 10 Å. Each clustered structure will be evaluated by the physics-based fitness function, which considers van der Walls, electrostatics potential, hydrogen and disulfide bond, solvation, and knowledge-based atom contact. The score given to each structure by this scoring function is referred to as physics score hereafter. After the scoring of each structure in the population, top M models are selected to proceed to another round of mutation, clustering, and evaluation. If the clustering resulted in a population smaller than M, randomly generated models are added to fulfill the population size. This process is repeated up to 3000 times with 1000 increment. The physics score and RMSD of the models will drop as the iteration of the process (Figure 2.2).

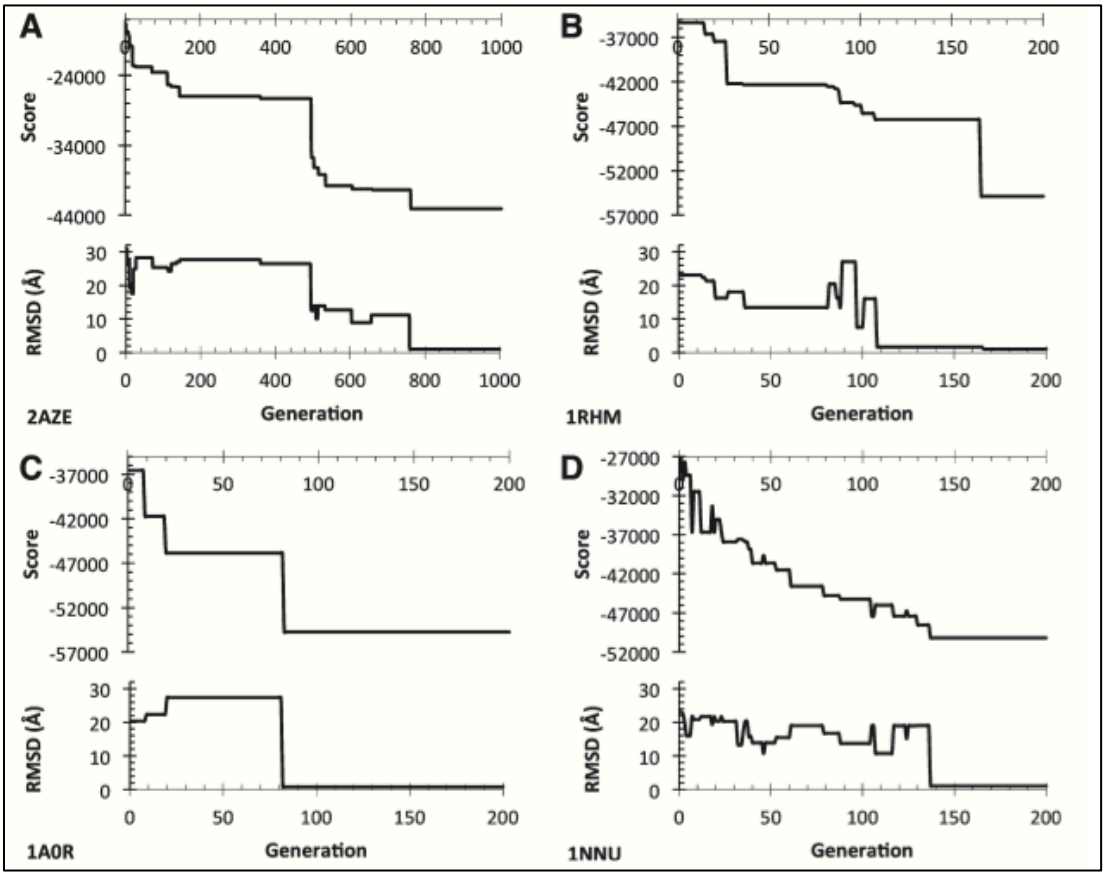


Figure 2.2 Evolution of the score and the RMSD taken from [48] The plot shows the evolution of the physics score and of the model with the best RMSD in each generation. Four panels A, B, C, and D corresponds to the complex 2AZE, 1RHM, 1A0R, and 1NNU respectively. The y-axis shows the physics score of the complex and the RMSD against the native structure. X-axis shows the number of generations. The plot is shown up to the generation where the score and the RMSD converged. The score and RMSD could drop in a stepwise manner (A, B, and C) or in gradient (D). Conversion of the score and the RMSD could be observed early as 100th generation (C), or it may take as long as 1000 generation (A).

2.2 Prediction of assembly pathway using the ranks

As briefly explained in Chapter 1, we are using the rank of the pairwise decoy to predict the assembly pathway of protein complexes. The output of Multi-LZerD allows us to see which pairwise decoys were used to construct a model generated by Multi-LZerD. Figure 2.1 shows the Multi-LZerD output for a heterotrimeric complex and the structure of the complex (PDBID 1A0R [50]). All the figures of protein complexes in this thesis are drawn using Pymol (URL: http://www.pymol.org/). Protein complexes will be referred to by its PDBID hereafter. 1A0R was a very good case, where the model with lowest RMSD was successfully ranked 1st among the others. Figure 2.3 (A) shows that rank 1st and 2nd of pairwise models B-G and B-P respectively, were used to construct the best model. 1A0R is a complex of G-protein βγ subunit dimer bound with phosducin. The βγ heterodimer is often considered as single unit because β and γ subunit does not dissociate unless denatured [50, 51]. The dimer is the functional unit of the protein that is involved in G-protein cycle, which repeats the association and dissociation with α subunit of the heteromeric G-protein. Phosducin regulates this G-protein cycle by binding to the βγ dimer, preventing it from re-associating with alpha subunit [52]. Based on these facts, we assume that assembly pathway of 1A0R in biological context as follows; the assembly of heterotrimer starts with the dimerization of beta and alpha subunit, followed by the association of phosducin with the beta-gamma heterodimer. This pathway of assembly corresponds to the relative relation between the ranks of the pairwise models used to construct the model: 1st ranking B-G decoy and 2nd ranking B-P

decoy were used for the model. We developed four different assembly pathway prediction methods based on this idea, which are separately described in the following subsections.

A

```
PDBID  1A0R
Chains B,G,P
Generations  2001
PopulationSize 200
ClashThreshold 200
ScoreType  physics
DecoyProgram LZerD

0,1,0;0,2,1
0,1,0;1,2,3551
```

B

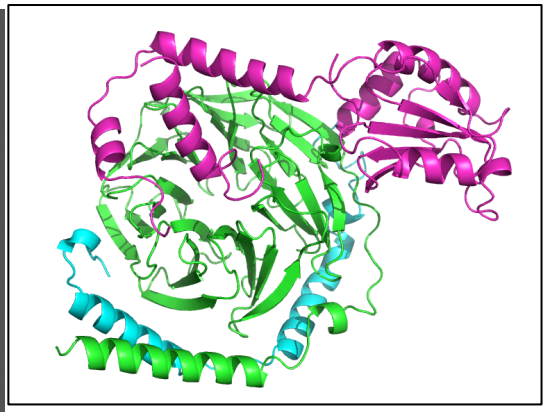


Figure 2.3 Example of Multi-LZerD output (1A0R)
(A) The first seven lines indicate the docked protein complex and the setting of Multi-LZerD parameters. The 1st and 2nd line shows the PDB ID and the subunit chain IDs of the complex that was docked respectively. The 3rd to 7th lines show the settings of Multi-LZerD. The bottom two lines are the actual output of Multi-LZerD. Each line represents a spanning tree, namely a model of the multiple docking. The models are sorted by the physics score, and written in the output according to the rank: the model with the tree "0,1,0;0,2,1" had best physics score thus, coming at the top of the output. Each set of three numbers separated by semicolons represents a pair of subunit and its conformation. The first two integers denote the chain ID; in the example, numbers from 0 to 2 correspond to chain ID alphabets B, G, and P respectively. The third integer denotes the rank of the decoy. The first three comma-separated integers in the 9th line represents 1st B-G decoy. Note that the rank starts from 0, not 1. (B) 1B9X in cartoon representation. Chains B, G and P are shown in green, cyan, and magenta respectively.

2.2.1 Lowest RMSD method

The first method, which we refer to as the lowest RMSD method, is basically the method described above. It uses the Multi-LZerD output model that has the lowest RMSD to the native structure among the final population. Because the scoring function does not necessarily rank the models with low RMSD at the top, we need the native structure to calculate the RMSD with. After the model with lowest RMSD is identified, the model's tree is converted into assembly pathway in the following steps. First, the pairwise models are sorted according to their rank. Then, the pathways are constructed by reading the decoys one by one from the ones with the higher ranks to the ones with lower ranks. The decoys are incorporated to the pathway in three ways: connected to one of the preexisting subcomplexes, connects two of the preexisting subcomplexes, or added to the pathway as a new subcomplex.

Therefore, our predictions always start with the formation of the dimer of the decoy with the highest rank. Whenever there are two or more decoys with the same rank, we use Z-score of the shape score to distinguish the ties. Z-score measures the divergence of an individual value from the mean value of the population. Z-score Z of a raw score χ is calculated by the equation below, where μ and σ are population mean and standard deviation respectively.

$$Z = (\chi - \mu) / \sigma$$

For each pair of subunits, Z-score of shape score is calculated for all the decoys. Because interface size has large contribution to the shape score, it is heavily

affected by the size of protein. Therefore, a large decoy will always be selected over a small one when comparing the raw score.

2.2.2 Low-RMSD decoy combination method

The second method uses the structure generated by the combination of the pairwise decoys with low RMSD. Top five decoys with lowest RMSD are selected, and the decoys are exhaustively combined to generate the models of the fully assembled complex. This method does not involve the structure search by the genetic algorithm, which makes this method faster compared to our methods based on Multi-LZerD. This usually provides some near native models with single digit RMSD. The initial purpose of constructing these models was to check if it is possible for Multi-LZerD to construct the near-native structure, before proceeding to computationally expensive structure search; Multi-LZerD is unable to construct good model unless LZerD is able to generate adequate parts for the near-native structure. The model with lowest RMSD is selected and the pathway is obtained following the same procedure as the lowest RMSD method.

2.2.3 Final generation method

The above two methods require the native structure to calculate the RMSD with. Development of the method that does not rely on the native structure is very important, since not all the protein can have its structure solved. It is also an advantage over the BSA method, because it requires crystal structure of the assembled complex to calculate BSAs. To this end, we have come up with a

method that utilizes not only the model with lowest RMSD, but also all the other models in the last generations of genetic algorithm. The basic idea is the same as the lowest RMSD method, but we are applying it to all the structures in the final generation and taking the consensus of the pathway. First, we get an assembly pathway from each model. Next, occurrences of the identical pathways are counted. Then the pathways are sorted according to its count and the most frequently occurring pathway is taken as the prediction. This method outputs the prediction without referring to the native structure. We call this third method the final population method.

2.2.4 Consensus across generation method

The fourth method is the extension of final generation method: this employs not only the final generation, but also all the population of the output file from 1,000th generation to the final generations. The cut-off was set to 1,000 since the physics score and RMSD of the structure tend to start converging by 1,000th generation [48] (Figure 2.2). As mentioned before, Multi-LZerD was run up to 3,000 generations with the increment of 1,000. Multi-LZerD was stopped if the score was converging. Otherwise, additional 1,000 generation was run up to 3,000 generation. The proteins in the dataset of this project were run for either 2,000 or 3,000 generations.

2.3 Construction of the pathways using ITScore

We also used another scoring function, ITScore for reevaluating our results based on the assumption that ITScore would improve the result. ITScore is a knowledge-based scoring function trained by using a set of 851 dimeric protein complexes with true biological interface [53]. While shape based scoring function only considers the geometric feature of the protein surface, ITScore considers the atomic interaction between true protein dimers. We converted the rank by shape score to the rank by ITScore, by calculating ITScore for all the pairwise decoys and then resorting them. All the four methods were tested after the conversion of the rank.

2.4 BSA method

The BSA method was also benchmarked to all the protein complexes we have analyzed, on view to compare the accuracy of assembly pathway prediction with our results. We also wanted to test if the BSA method works for asymmetric protein complexes. First, we calculated ASA for the fully assembled complex and the subcomplexes with all the possible combination of subunits. The ASA of each subunit and sub complexes were calculated using the program NACCESS v2.1.1 (URL: http://www.bioinf.manchester.ac.uk/naccess/) as described in the introduction. Then, starting from the full complex, a subunit or a subcomplex, which exposes smallest BSA upon separation, is removed from the complex. This process is repeated until a dimer is left. The below is an example of the result of assembly pathway prediction by BSA method on a six chain complex

3VYT, where alphabets A to F denotes each subunit (Figure 2.4). This shows that taking off AB dimer from the full complex exposes the smallest BSA of 2413 $Å^2$. Next, separating the CDEF complex into two dimers, DE and CD, result in the minimum exposure of BSA with 2423 $Å^2$. In this example, the BSA method prediction starts with the dimerization of two dimers, DE and CF, followed by the association of AB dimer to the CDEF tetramer.

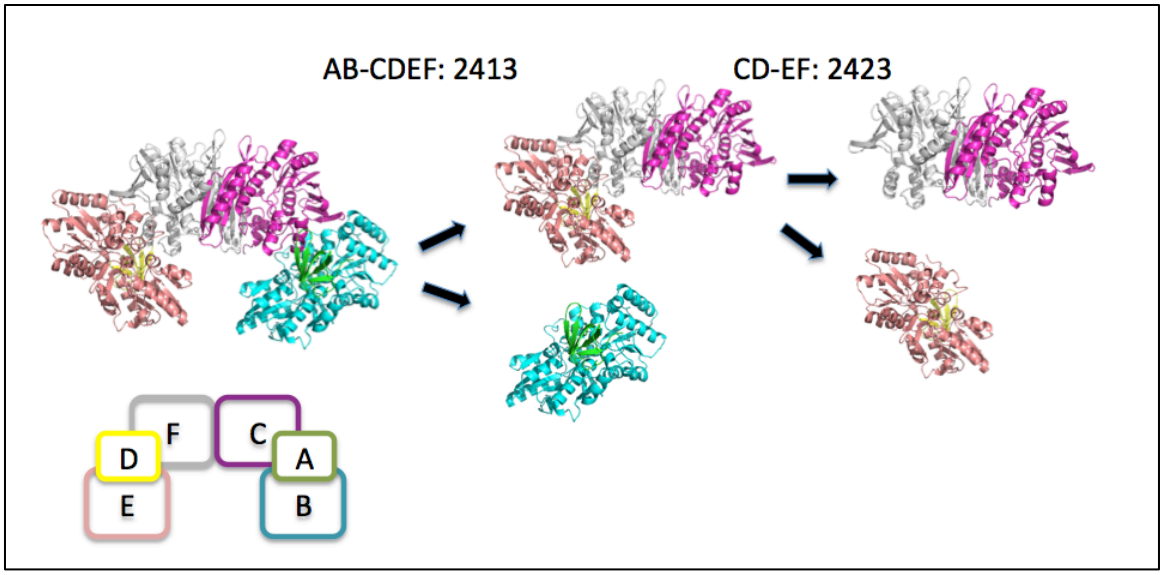


Figure 2.4 Example of the BSA method with a hexamer 3VYT

Summary of the five methods described in this chapter are provided in Table 2.1. BSA method only requires the calculation of BSA, thus, the computation cost is very low. The four methods we have developed may take up to a week depending on the size and number of subunits of the target protein. The low-RMSD decoy combination method only requires the generation of pairwise decoys by LZerD. Number of the subunit determines how many pairwise docking LZerD must perform. The three methods that require the structure search by Multi-LZerD have the highest computational cost.

Table 2.1 Summary of the four prediction methods

Method	Native structure	Rquired process	Computational cost
BSA	Required	BSA calculation by NACCESS	Low
Low-RMSD decoy combination	Required	pairwise docking by LZerD	Middle
Lowest RMSD model	Required	Multi-LZerD (LZerD + structure search)	High
Final population	-	Multi-LZerD (LZerD + structure search)	High
1000/2000 generation	-	Multi-LZerD (LZerD + structure search)	High

BSA method requires native structure to calculate the BSA. Low-RMSD decoy combination method and lowest RMSD model method requires the native structure to calculate the RMSD with. Computational cost is heavily dependent on the size and subunit of the complex. Also, the number of processor used affects the speed. For the heterotrimer 1A0R, BSA can finish the process in less than a minute. Low-RMSD decoy combination will likely to take a day, and rest of the methods would require two days.

CHAPTER 3. DATASET

The dataset of protein complexes used in this project comes from two sources. First source is the Multi-LZerD dataset that were already available at the beginning of this project, most of which are already reported in journal articles [48]. The second source is the result of the search manually done at PDB specifically for this project. It is important to note here the difference between the two sources. The complexes from the first source were run in different Multi-LZerD setting; complexes were ran using 5 Å or 10 Å clustered decoys, or non-clustered decoys, and different clash thresholds were chosen to yield good docking results. In contrast, the same setting was used for the complexes from the second source; using 10 Å clustered decoys and 2,000 as clash threshold. Decoys clustered by 10 Å were used, since it generally has the good performance [48, 54]. Since we are unable to conduct experiments, the assembly pathways were obtained from literature search. The dataset proteins and their details, including the source of assembly pathway, are provided below. Summary of the dataset is provided in Table 4.24.

(1) 1A0R: Transducin β-γ dimer bound with phosducin [50]

Table 4.1 1A0R subunits

Chain ID	Description
B	Transducin β subunit
G	Transducin γ subunit
P	Phosducin

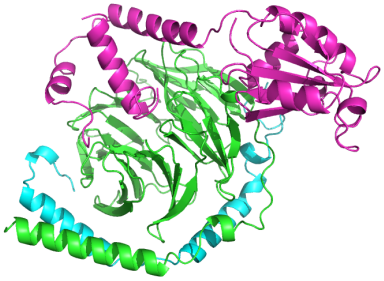


Figure 4.1 Structure of 1A0R

Assembly pathway: BG > BGP

Transducin is a heteromeric GTP-binding protein (G protein) consisting of α, β, and γ subunits. Since β and γ subunits are strongly associated with each other, β-γ dimer is often regarded as single unit [50, 51]. Phosducin regulate the G-protein cycle of transducin by interacting with the β-γ dimer. Based on the strong association between the β-γ dimer and from the biological context, formation of β-γ likely to take place first, followed by the interaction with phosducin.

(2) 1B9X: Transducin β-γ dimer bound with phosducin [52]

Table 4.2 1B9X subunits

Chain ID	Description
A	Transducin β subunit
B	Transducin γ subunit
C	Phosducin

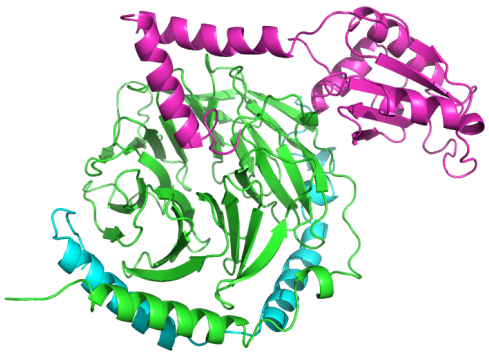


Figure 4.2 Structure of 1B9X

Assembly pathway: AB > ABC

See the description for 1A0R. 1B9X and 1A0R are the structure of same protein submitted by different group. The pairwise sequence identity was calculated using EMBOSS Needle. The sequence identity of β subunit, γ subunit, and phosducin between 1A0R and 1B9X were 98.3 %, 95.6 % and 87.0 % respectively.

EMBOSS Needle URL: URL: https://www.ebi.ac.uk/Tools/psa/emboss_needle/

(3) 1VCB: ElonginBC bound with von Hippel-Lindau (VHL) tumor suppressor [55]

Table 4.3 1VCB subunits

Chain ID	Description
A	Elongin B
B	Elongin B
C	VHL tumor suppressor

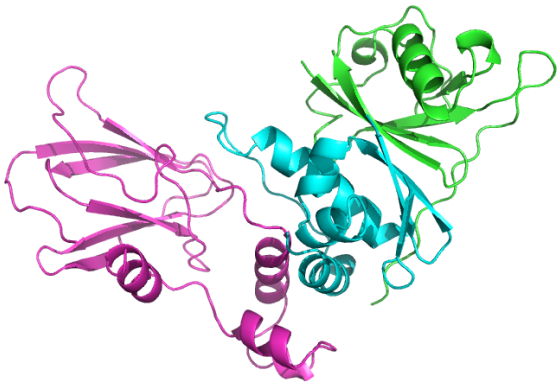


Figure 4.3 Structure of 1VCB

Assembly pathway: AB > ABC

Elongin BC is a component of transcription factor B complex (SIII), which is a ternary complex of Elongins A/A2, B, and C. VHL is a tumor suppressor that binds to Elongin BC and inhibit transcription elongation. Elongin B and C alone have no or very poor interaction with VHL, but the interaction is enhanced when both are present [56]. A model of interaction between Elongin BC and VHL is proposed based the observations above: Elongin BC dimer binds to either VHL or Elongin A, which leads to transcription regulation and elongation respectively.

(4) 2AZE: Structure of the Rb C-terminal domain bound to and E2F1-DP1 dimer [57]

Table 4.4 2AZE subunits

Chain ID	Description
A	Transcription factor DP-1
B	Transcription factor E2F1
C	Retinoblastoma-associated protein (Rb)

Figure 4.4 Structure of 2AZE

Assembly pathway: AB > ABC

DP-1 and E2F1 forms a heterodimer, which functions as a transcription factor regulating cell cycle. Rb is one of the proteins that interact with the transcription factor. The heterodimer is likely to be the functional unit, since E2F1 alone have weak DNA-binding activity and the activity is enhanced under the presence of DP-1 [58]. Also, heterodimerization of DP-1 and E2F-1 results in efficient binding with Rb [58].

(5) 1IKN: I-Kappa-B α/NF-Kappa-B complex [59]

Table 4.5 1IKN subunits

Chain ID	Description
A	NF-Kappa-B p65 subunit (RelA)
B	NF-Kappa-B p50 subunit
C	I-kappa-B α

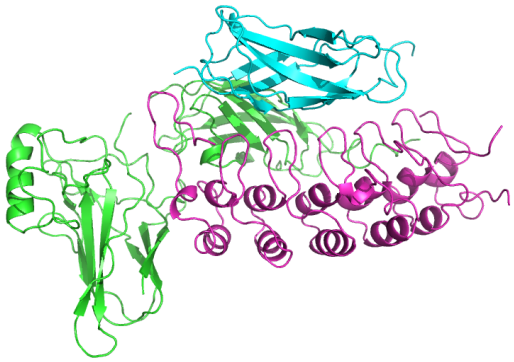


Figure 4.5 Structure of 1IKN

Assembly pathway: AC > ACD

NF-kappa-B is a protein complex involved in transcription. They bind to DNA as homo- or heterodimer, and RelA-p50 heterodimer is one of the major NF-kappa-B dimer in cell [60]. I-kappa-B is an inhibitor of NF-kappa-B, which binds to NF-kappa-B dimer and prevents it from binding to DNA. Proposed signal transduction pathway shows that I-kappa-B α protein bind to NF-kappa-B dimers to inhibit transcription, and its dissociation allow the NF-kappa-dimer to bind to DNA [60, 61].

(6) 1GPQ: IVY complex with its target HEWL [62]

Table 4.6 1GPQ subunits

Chain ID	Description
A,B	Inhibitor of vertebrate lysozyme (IVY)
C,D	Lysozyme C

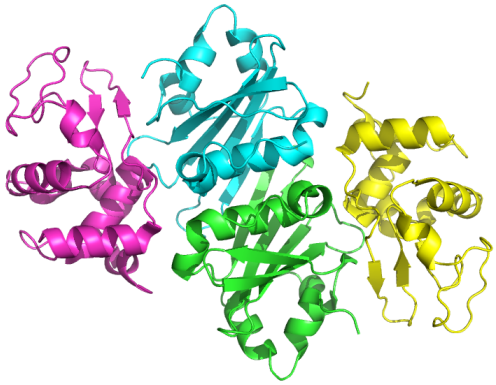


Figure 4.6 Structure of 1GPQ

Assembly pathway: AB > ABC (D) > ABCD

While the functional unit of IVY is homodimer [63], HEWL can be functional in both monomeric and dimeric forms [64, 65]. The crystal structure agree with the 2:2 stoichiometry of IVY and HEWL in other experiment [63]. The two lysozyme subunits do not make contact with each other in the crystal structure. Therefore, the HEWL-IVY tetramer complex is likely to be formed by sequential recruitment of HEWL to IVY homodimer.

(7) 1ES7: Complex between BMP-2 and two BMP receptor IA ectodomains [66]

Table 4.7 1ES7 subunits

Chain ID	Description
A, C	Bone morphogenic protein (BMP)-2
B, D	BMP receptor extracellular domain

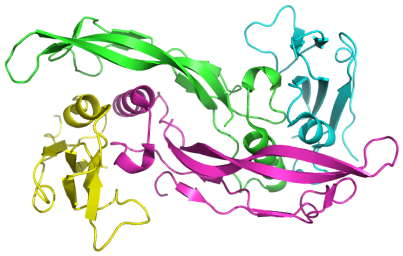


Figure 4.7 Structure of 1ES7

Assembly pathway: AC > ABC(D) > ABCD

BMP-2, a disulfide-linked homodimer, is a growth factor that is involved in bone and cartilage formation, which interacts with two types of serine/threonine receptor kinase, type I and type II [66]. The disulfide linkage gives BMP-2 homodimer high stability, which enabled it to be purified under harsh conditions without inactivation [67]. Chain B and D are the extracellular-ligand binding domain of the BMP receptor type I, and they do not have contact with each other in the crystal structure. Therefore, the assembly of the BMP-2 and the extracellular domain of its receptor is likely to start with homodimerization of BMP-2, followed by the binding of extracellular domain of receptor to the homodimer.

(8) 1REW: Complex between BMP-1 and two BMP receptor IA [68]

Table 4.8 1REW subunits

Chain ID	Description
A, C	Bone morphogenic protein (BMP)-2
B, D	BMP receptor extracellular domain

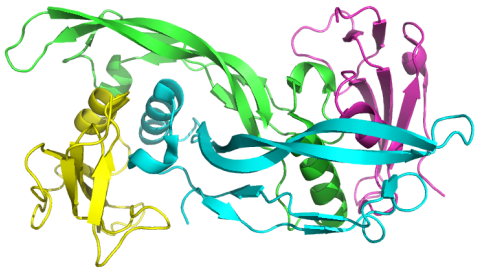


Figure 4.8 Structure of 1REW

Assembly pathway:

AC > ABC(D) > ABCD

See the description for 1ES7. 1REW and 1ES7 are the structure of same protein submitted by a different group. The pairwise sequence identity was calculated using EMBOSS Needle. The pairwise sequence identity of the subunit between 1ES7 and 1REW are 98.3 % for the BMP-2 subunits and 65.9 % for BMP receptor subunits. The relatively low sequence identity between BMP receptor subunits comes from the difference of the sequence length. The Receptor subunits in 1ES7 are missing 49 residues compared to those of 1REW. The missing residues led to the gap penalty when their sequences were aligned.

(9) 2E9X: Human GINS core complex [69]

Table 4.9 1REW subunits

Chain ID	Description
A	DNA replication complex GINS protein Psf1
B	DNA replication complex GINS protein Psf2
C	GINS complex subunit 3 (Psf3)
D	GINS complex subunit 4 (Sld5)

Figure 4.9 Structure of 2E9X

Assembly pathway: DB > ADB > ABCD

Human GINS complex was analyzed by mass spectroscopy under stepwise addition of methanol as disrupting agent, which detected two subcomplexes, Psf2-Sld5 and Psf1- Psf2-Sld5 [70]. Following the hypothesis that disassembly is the opposite of assembly [20, 21], the complex assembly is likely to start with the formation of Psf2-Sld5, then the sequential binding of Psf1 and Psf3.

(10) 2QSP: Bovine hemoglobin at pH 5.7 [71]

Table 4.10 2QSP subunits

Chain ID	Description
A, C	Hemoglobin α subunit
B, D	Hemoglobin β subunit

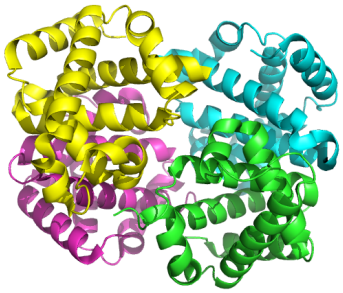


Figure 4.10 Structure of 2QSP

Assembly pathway:

AB, CD > ABCD

Hemoglobin is a dimer of two heterodimers, each of them formed by α and β subunit. Biochemical experiments have shown that hemoglobin assembly starts with the formation of α-β heterodimer, followed by the association of the two heterodimers, and that the formation of the heterodimer is the rate-limiting step of the hemoglobin assembly [72]. The proposed assembly pathway is also supported by the results from ESI-MS [73, 74].

(11) 3FH6: Resting state maltose transporter [75]

Table 4.11 3HF6 subunits

Chain ID	Description
A, B	Maltose/maltodextrin import ATP-binding protein MalK
F	Maltose/maltodextrin import ATP-binding protein MalK
G	Maltose/maltodextrin import ATP-binding protein MalK

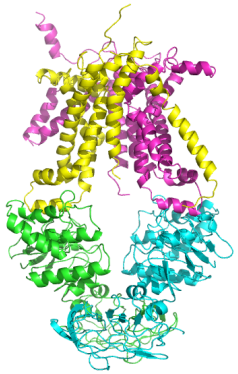


Figure 4.11 Structure of 3FH6

Assembly pathway: AB > ABF > ABFG / AB > ABG > ABFH / AB, FG > ABFG

Maltose transporter is a membrane bound protein where peripheral cytoplasmic protein MalK homodimer is bound to the MalFG heterodimer integrated to the membrane. Co-IP and quantification of interacting subunits have shown that maltose transporter can assemble in multiple pathway [32].

(12) 2BQ1: Ribonucleotide reductase [76]

Table 4.12 2BQ1 subunits

Chain ID	Description
E, F	Ribonucleotide-diphosphate reductase 2 α subunit
I, J	Ribonucleotide-diphosphate reductase 2 β subunit

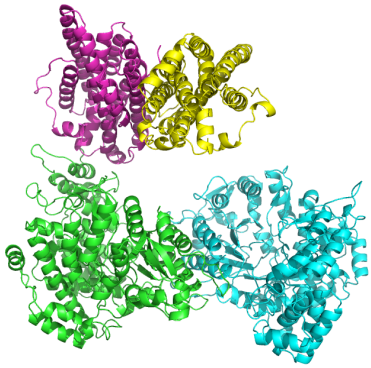


Figure 4.12 Structure of 2BQ1

Assembly pathway: EF, IJ > EFIJ

Ribonucleotide reductase consists of two homodimers, α_2 and β_2. The homodimers exists in the equilibrium of α_2, β_2, $\alpha_2\beta_2$, and $\alpha_4\beta_4$ [77]. Therefore, the $\alpha_2\beta_2$ complex is likely to form by the association of the two homodimers.

(13)1KF6: Fumarate reductase [78]

Table 4.13 1KF6 subunits

Chain ID	Description
A	Fumarate reductase flavoprotein (FrdA)
B	Fumarate reductase iron-sulfur protein (FrdB)
C	Fumarate reductase 15 KDa hydrophobic protein (FrdC)
D	Fumarate reductase 13KDa hydrophobic protein (FrdD)

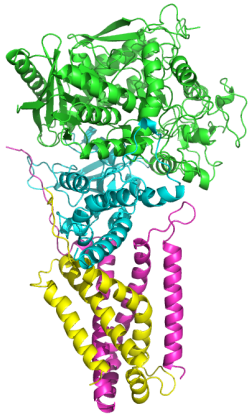


Figure 4.13 Structure of 1KF6

Assembly pathway: CD > BCD > ABCD

Fumarate reductase is a membrane bound protein, FrdA and B bound to FrdCD heterodimer, which functions as an anchor that hold the complex on to the membrane. Assembly of the complex was monitored by pulse-chase experiment [79], and showed that assembly starts with the dimerization of the two hydrophobic protein FrdC and FrdD. FrdCD is quickly inserted to the membrane and gets capped by FrdB. FrdA caps the complex on top of FrdB in a manner that fully enclose FrdB inside the complex [80].

(14) 1HEZ: Antibody-antigen complex [81]

Table 4.14 1HEZ subunits

Chain ID	Description
A, C	Kappa light chain of IG
B, D	Kappa heavy chain of IG
E	Protein L (PpL)

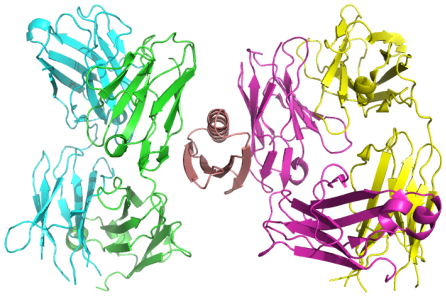


Figure 4.14 Structure of 1HEZ

Assembly pathway: AB, CD > ABE, CD > ABCDE

Protein L, a cell wall-anchored protein from *Peptostreptococcus magnus*, binds to kappa light chain of mammalian IGs [81]. The structure 1HEZ has PpL at its center with two IGs bound to it symmetrically without making contact with each other. The complex formation is likely to start with the formation of two IGs. Then the two IGs will bind to PpL sequentially.

(15) 1W88: Pyruvate dehydrogenase E1 bound to the peripheral subunit binding domain of E2 [82]

Table 4.15 1W88 subunits

Chain ID	Description
A, C	E1 component, α subunit
B, D	E1 component, β subunit
I	Peripheral subunit-binding domain of E2

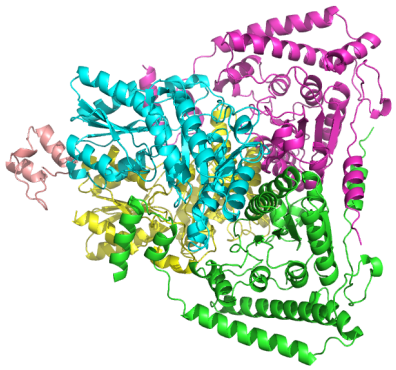


Figure 4.15 Structure of 1W88

Assembly pathway: ABCD > ABCDI

Pyruvate dehydrogenase E1, a component of pyruvate dehydrogenase complex (PDC), is a heterotetramer consisting of two α and two β subunits. A structure of E1 component without E2 component is reported [83]. The structure 1W88 is likely to be assembled with the formation of E1 component, followed by its association with the subunit from E2 component.

(16) 1RLB: Retinol binding protein complexed with transthyretin [84]

Table 4.16 1RLB subunits

Chain ID	Description
A, B, C, D	Transthyretin (TTR)
E, F	Retinol binding protein (RBP)

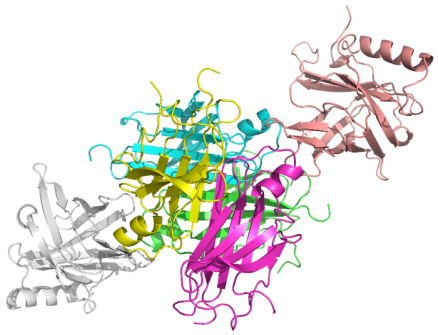


Figure 4.16 Structure of 1RLB

Assembly pathway: ABCD > ABCDE(F) > ABCDEF

TTR, a homotetramer, is a transporter that carries the hormone thyroxine and RBP bound to retinol [84]. The structure is consistent with the 2:1 binding stoichiometry of TTR and RBP [85]. Therefore, it is likely that the complex formation starts with the assembly of TTR, followed by the sequential binding of RBP.

(17) 1DU3: Structure of TRAIL-SDR5 [86]

Table 4.17 1DU3 subunits

Chain ID	Description
A, B, C	Death receptor 5
D, E, F	TNF-related apoptosis inducing ligand (TRAIL)

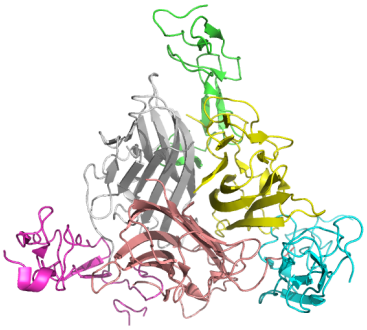


Figure 4.17 Structure of 1DU3

Assembly pathway: DEF > ABCDEF

TRAIL, a homotrimeric protein complex, binds to its receptor and induce apoptosis and the ligand binding induces the homotrimerization of the receptor [86]. The receptor proteins in the structure 1DU3 only contains the ligand binding domain and don't have contact with each other. Therefore, the complex is likely to form starting with the homotrimerization of TRAIL, followed by sequential binding of receptor chains to the trimer.

(18) 1S5B: Cholera holotoxin with an A-subunit [87]

Table 4.18 1S5B subunits

Chain ID	Description
A	A subunit
D, E, F, G, H	B subunit

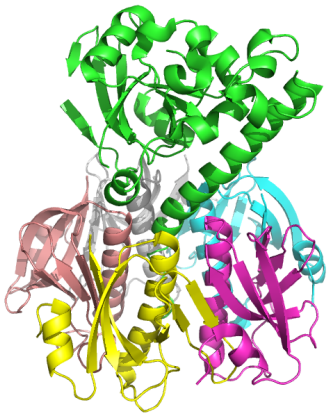


Figure 4.19 Structure of 1S5B

Assembly pathway: $AB_3 + B > AB_4 + B > AB_5$

Cholera toxin has a cone-like structure where CTB pentamer forming a ring at the base and CTA binding on top of the ring. An assembly pathway, $B_n > B_n + A > AB_n + (5\text{-}n)B > AB_5$, was proposed based on the fact that A-subunit is unable to form a complex with fully assembled B-protein pentamer ring, and that the assembly of the CTB ring structure is 3 fold faster under the presence of CTA [38]. Other research has shown that AB_3 or AB_4 assembly intermediates was able to attract additional monomeric B subunits [39]. Structure of cholera toxin have revealed that A subunit is making major contacts with three B subunits, forming a salt-bridge with two of them, indicating that the AB_3 is the stable assembly intermediate of the complex formation [88]. The assemble pathway above was proposed based on the structural analysis [88].

(19) 3VYT: HypCDE complex [43]

Table 4.19 3VYT subunits

Chain ID	Description
A, D	Hydrogenase expression/formation protein HypC
B, E	Hydrogenase expression/formation protein HypD
C, F	Hydrogenase expression/formation protein HypE

3VYT had only chains A, B, and C in the asymmetric unit. The biological assembly file was used to construct the full complex and chain ID was assigned as below.

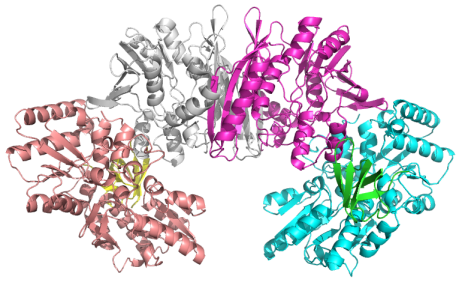


Figure 4.19 Structure of 3VYT

Assembly pathway: AB, DE, CF > ABCDEF

HypCDE has a horseshoe-like structure, where two HypE occupying the toe and two HypD at the tip. HypC is bound at the interface between HypE and HypD. The structure showed that HypE share interface with both HypC and HypD. Because the position of the loop in HypC, which interacts with HypE, is fixed by the dimerization of HypC and HypD, HypCD dimerization is likely to take place before the association with HypE [43]. Pull-down assay and size exclusion chromatography showed that HypE alone is incapable of forming a complex with neither HypE nor HypD [43].

(20) 4HI0: UreF/UreH/UreG complex [89]

Table 4.20 4HI0 subunits

Chain ID	Description
A, C	Urease accessory protein UreF
B, D	Urease accessory protein UreH
E, F	Urease accessory protein UreG

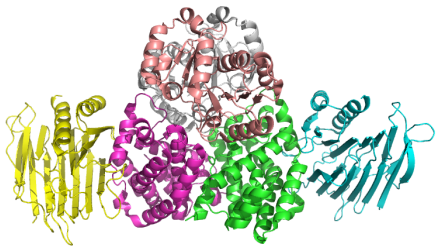


Figure 4.21 Structure of 4HI0

Assembly pathway: AB, CD > ABCD, EF > ABCDEF

The UreF/UreH/UreG complex has structure that is formed by three dimers. Two UreF/UreH heterodimers binds to each other with UreF forming a linear structure, and UreG homodimer is bound to the UreF-UreF interface. The mutation that disrupts the homodimerization of UreF/UreH also leads to the failure to recruit UreG to the complex [89]. Since UreG independently forms a homodimer [89], the UreF/UreH/UreG complex assembly starts with the formation of dimerization of UreF/UreH heterodimer, followed by association of UreG homodimer to the $UreF_2H_2$ heterotetramer.

(21) 4IGC: Bacterial RNA polymerase [90]

Table 4.21 4IGC subunits

Chain ID	Description
A, B	DNA-directed RNA polymerase subunit α
C	DNA-directed RNA polymerase subunit β
D	DNA-directed RNA polymerase subunit β'
E	DNA-directed RNA polymerase subunit ω
X	DNA-directed RNA polymerase subunit σ (RpoD)

Figure 4.21 Structure of 4IGC

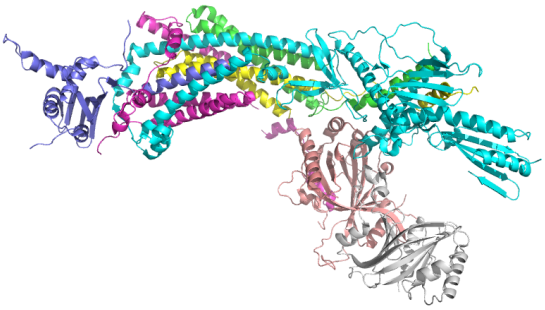
Assembly pathway: AB > ABC, DX > ABCDX > ABCDEX

Assembly of the bacterial RNA polymerase is reviewed as early as 1979 [22]; assembly starts with the formation of $\alpha_2\beta$ hetero trimer, then the sequential recruitment of β′ and σ subunit. α and β subunits are in equilibrium with the $\alpha_2\beta$ subcomplex, which is stabilized by β′ subunit, and the assembly of the σ subunits completes the polymerase assembly. However, the ω subunit was not considered as the subunit of the RNA polymerase at the time when the review was published. ω subunit was later found to be stabilizing β′ and preventing it from aggregation [91]. The identification of ω subunit updates the assembly pathway; $\alpha_2\beta$ and β′-ω subcomplexes are formed separately and assembled, which is then bound with the σ subunit to yield functional RNA polymerase [91].

(22) 4GWP: Mediator head module [92]

Table 4.22 4GWP subunits

Chain ID	Description
A	Mediator subunit 11 (Med11)
B	Mediator subunit 11 (Med17)
C	Mediator subunit 11 (Med8)
D	Mediator subunit 11 (Med22)
E	Mediator subunit 11 (Med18)
F	Mediator subunit 11 (Med20)
G	Mediator subunit 11 (Med6)



Assembly pathway: ABD, CG > ABCDG, EF > ABCDEFG

Mediator is a large protein complex that regulates transcription by RNA polymerase II, and electron microscopy has shown that Mediator has three distinct structures; head, middle, and tail modules. Comprehensive Co-IP assay has revealed the subcomplexes formed by the seven subunits that consists the Mediator head module [14], and an assembly pathway is proposed based on the Co-IP assay [17].

(23) 3UKU: Arp2/3 complex [no publication]

Table 4.23 3UKU subunits

Chain ID	Description
A	Actin like protein 3 (Arp3)
B	Actin like protein 2 (Arp2)
C	C: Actin-related protein 2/3 complex subunit 1B (p41)
D	Actin-related protein 2/3 complex subunit 2 (p34)
E	Actin-related protein 2/3 complex subunit 3 (p21)
F	Actin-related protein 2/3 complex subunit 4 (p20)
G	Actin-related protein 2/3 complex subunit 5 (p16)

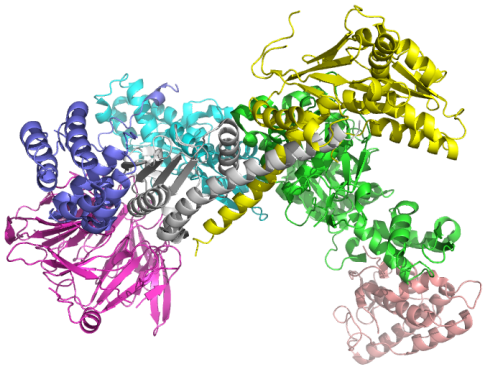


Figure 4.23 Structure of 3UKU

Assembly pathway: DE, CG, AE > ACDEFG > ABCDEFG

Speculative assembly pathways for Arp2/3 complex was proposed based on the systematic pairwise yeast-two hybrid assay [11]. The pathways were incomplete because their assay was not able to detect interactions that involve neither Arp2 nor Arp3, implying their incorporation to the complex at later stage of the assembly [11]. A comprehensive Co-IP assay have revealed that p20-p34 heterodimer, termed core subunits, is critical for the assembly of the complex, and also revealed three sets of peripheral subunits that bind to the core subunits

[10]. We propose the key steps of the Arp2/3 assembly based on the combination of the incomplete assembly pathway and the subcomplexes identified by Co-IP assay.

Table 4.23 Summary of dataset

	# of chains	PDB ID	Structure description	Biological inference	Structural inference	Experimental evidence	Model of assembly
1	3	1A0R	Transducin βγ dimer bound with phosducin	x		x	
2		1B9X	Transducin βγ dimer bound with phosducin	x		x	
3		1VCB	ElonginBC bound with VHL	x		x	x
4		2AZE	Rb C-terminal domain bound to E2F1-DP1	x		x	
5		1IKN	I-Kappa-B alpha/NF-Kappa-B complex	x		x	x
6	4	1GPQ	IVY complex with its target HEWL	x	x		
7		1ES7	Complex between BMP-2 and two BMP receptor		x		
8		1REW	Complex between BMP-1 and two BMP receptor		x		
9		2BQ1	Ribonuclueotide reductase	x	x	x	
11		2QSP	Bovine hemoglobin at pH 5.7			x	x
12		3FH6	Maltose Transporter			x	x
13		2E9X	Human GINS core complex			x	
14		1KF6	Fumarate reductase			x	x
15	5	1HEZ	Antibody-antigen complex	x	x		
16		1W88	Pyruvate dehydrogenase E1 bound to a subunit of E2	x			
17	6	1RLB	Retinol binding protein complexed with transthyretin	x	x		
18		1DU3	Structure of TRAIL-SDR5	x	x		
19		1S5B	Cholera holotixin with an A-subunit			x	x
20		3VYT	HypCDE complex			x	x
21		4HI0	UreF/UreH/UreG complex			x	x
22		4IGC	Bacterial RNA polymerase			x	x
23	7	3UKU	Arp2/3			x	x
24		4GWP	Mediator head module			x	x

Biological inference refers to the assumption based on biological reasoning. For example, if a structure consists of a protein dimer and its inhibitor, we assume that inhibitor binds to the complex after the formation of dimer. Structural inference refers to the assumption based on the structure. For example, if two subunits in the complex are not having contact with each other, we assume that these subunits does not dimerize. Experimental evidence refers to any evidences from experiment that supports certain assembly pathway. For example, subunit interaction network information from experiment and analysis of interaction from crystal structure are in this criterion. Model of assembly shows that assembly pathway is proposed in journal articles.

CHAPTER 4. RESULTS

4.1 Success rate of each method

The result of the pathway prediction is provided in Table 4.1. Overall, BSA method had the best prediction success rate of 67.0%. Using the rank by the shape score, low-RMSD decoy combination method, lowest RMSD method, final generation method, and consensus across generation method all had the success rate of 46.0%. Conversion of the shape score rank to ITScore rank has changed the success rate of lowest RMSD method, final generation method, and consensus across generation method to 39.0 %, 57.0 %, and 57.0 % respectively. The result shows that BSA method is capable in correctly predicting the assembly pathway not only for symmetrical protein, but also for asymmetric proteins. Also, BSA method was the most successful method among the five methods. The subsections below show the results specific for each prediction method.

4.1.1 Low-RMSD decoy combination method

The success rate of the low-RMSD decoy combination method, 46.0 % was lower than our expectation. We expected this method to perform well, since the models generated by this method often have a single digit RMSD to the

native structure. The results show that obtaining the low-RMSD model does not necessarily lead to the prediction of the correct pathway. While the structure search at the genetic algorithm considers physical properties of the complex, this method only considers RMSD to construct the model.

4.1.2 Lowest RMSD method

The lowest RMSD method had the success rate of 46.0 % for the shape score rank, and 38.0 % for the ITScore rank. Table 4.1 shows that this method tends to be successful in the complexes where Multi-LZerD was able to generate a model with low RMSD. Therefore, we checked how the success rate will change if we only considered the complexes with successful multiple docking prediction (Table 4.2). All the complexes with the model lower than 2.0 Å RMSD to the native structure were selected. There were 6 such complexes: 1A0R, 1B9X, 1VCB, 2AZE, 1GPQ, and 1ES7. In multiple docking, shift of single subunit can lead to large RMSD even if other subunits were having near native conformation [48]. Therefore, we checked the RMSD of the partial structure to look for such complex, and we found 2E9X and 1W88. 2E9X is a 4-chain complex consisting of chains A, B, C, and D. Although the overall RMSD was 9.5 Å, trimer subcomplex ABD had the 1.6 Å RMSD to the native structure (Figure 4.1). The 5-chain complex 1W88 had 4.8 Å RMSD to the native structure, but substructure of four subunits had the RMSD of 1.3 Å.

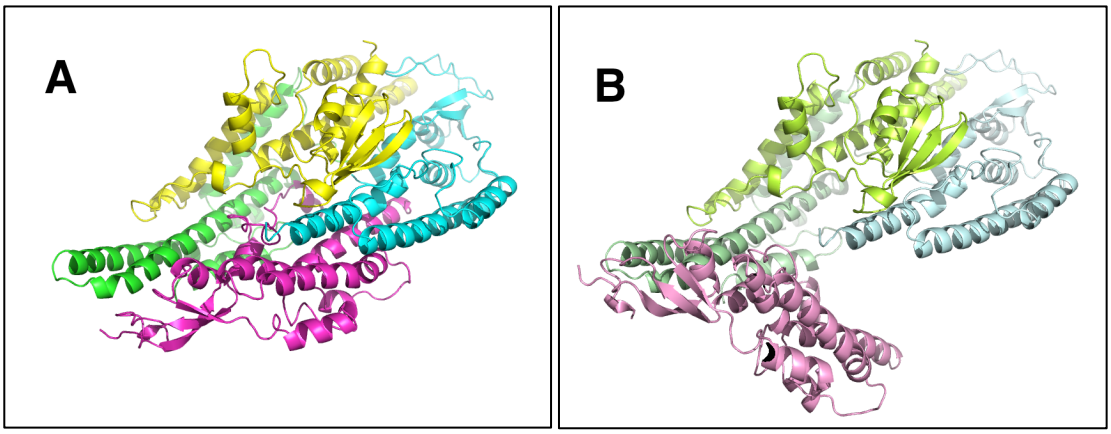


Figure 4.1 Structure of 2E9X
(A) Native structure of 2E9X. Chains A, B, C, and D are shown in green, cyan, magenta, and yellow respectively. (B) Docked model by Multi-LZerD with 9.5 Å RMSD. The same color code, but in opaque color is used. Position of chain C (magenta) is different between the two, but the rest of the three chains are in similar conformation with the 1.6 Å RMSD.

We also manually checked the models with the RMSD lower than 20.0 Å, and selected complexes where the model was having similar topology to the native structure. We follow the definition of topology by 3D complex [46]; 3D complex is a database of protein structures classified by its topology, where protein complexes are represented in graph. Each subunit is represented as node and a pair of nodes are connected with edge if they have an interface. Interface is defined as contact of more than 10 residues between a pair of subunits. Residue-residue interaction is considered as contact if van der Waals radii of any pair of atoms from the two residues are within 0.5 Å. There were three such cases: 1IKN, 1REW, and 1HEZ (Figure 4.2).

We have selected 11 complexes in total, and looked at the success rate specific for this subset. There were improvements in the success rates in all methods (Table 3.2): from 65.0 % to 73.0 %(8 out of 11) for BSA method, from

48.0 % to 64.0 % (7 out of 11) for low-RMSD decoy combination method, from 48.0 % to 82.0 % (9 out of 11) for lowest RMSD method, final generation method, and consensus across generation method. Success rate of the three methods by ITScore ranks also improved from 39.0 % to 73.0 % (8 out of 9) for lowest RMSD method and from 57.0 % to 82.0 % (9 out of 11) for the final generation method and consensus across generation method.

It is not surprising that pathway prediction is affected by the modeling accuracy of the complex. Multi-LZerD may connect a pair of subunits that are not having contact in the crystal structure. If the scoring function favors structure with such wrong topology, they are kept and eventually become prevalent in the population. Since we convert the spanning trees to the pathway, interactions between the wrong subunits directly affects the pathway. In other words, a model with correct topology could give a correct pathway even if it did not have near native RMSD, which is exemplified by the complexes 1IKN, 1HEZ, and 1REW. These results indicate that models that don’t have near-native structure could still be useful for the prediction of the assembly pathway if the model have a good topology.

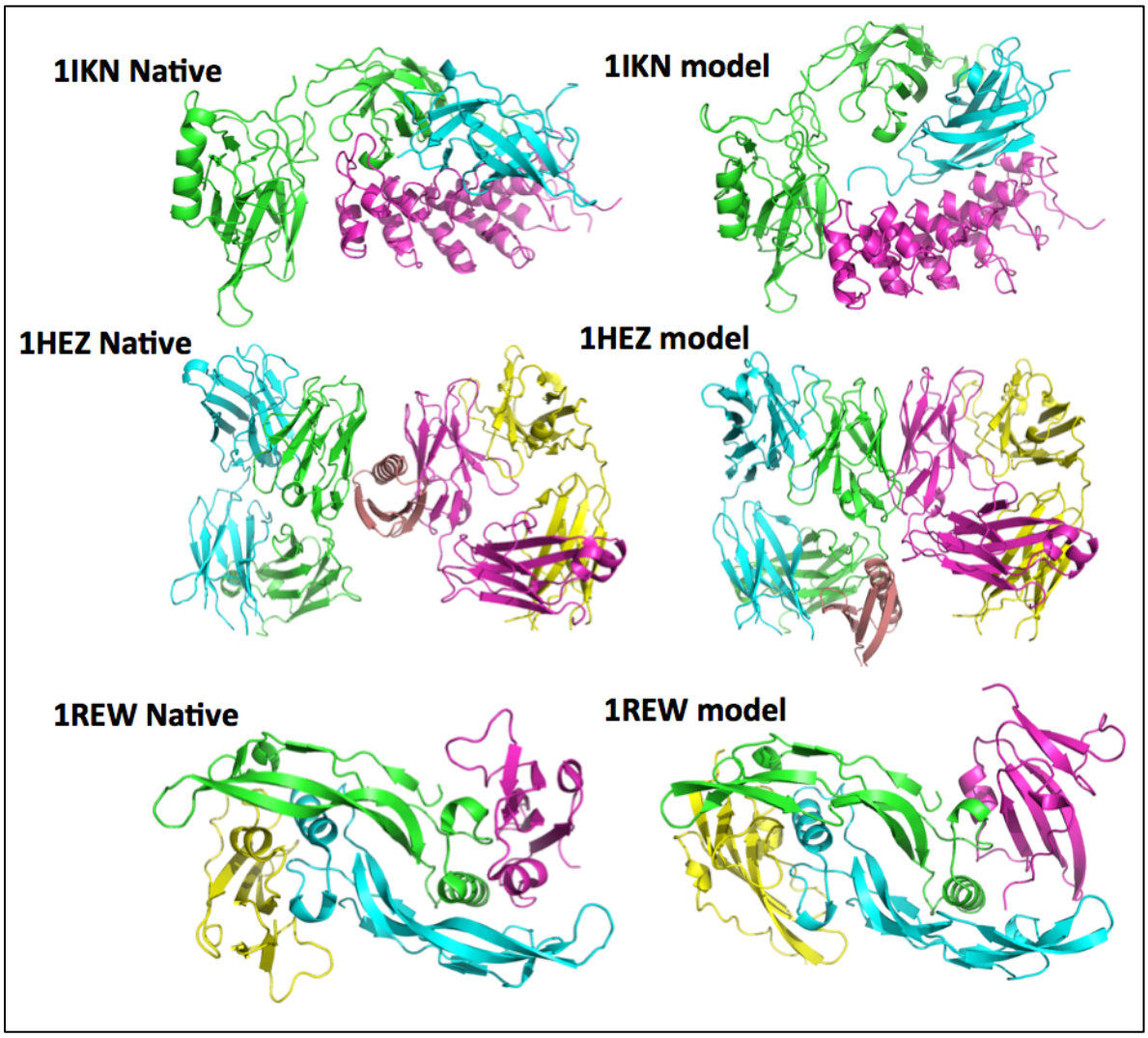


Figure 4.2 Models with similar topology to the native structure
Figures in the left and right columns show the native structure and the models with the lowest RMSD respectively.

4.1.3 The methods utilizing multiple models

The success rate of the final generation method and the consensus across generations method were both 46.0 %. The success rate improved to 54.0 % when the shape score ranks were converted to ITScore ranks. For the final population method, taking top 50 and top 100 structure were also considered, but there were no change in the prediction accuracy. In these two

methods, prediction was considered success when the correct pathway had the highest count among the total population. However, there were cases were the correct pathway was ranked relatively high among the others. Therefore, we also recorded the ranks of all the correct pathways that appeared in the collection of pathways. We allowed exchange of chain ID of identical subunits when looking at the pathways. There are complexes with two or more identical subunits. For example, for a trimer consisting of subunits A, B, and C, where A and B are identical protein, we consider AC dimerization and BC dimerization as equal. For those complexes, we have checked the pairwise RMSD of identical subunits. We regarded such subunits as equal when their pairwise RMSD is below 1.5 Å.

The purpose of developing these two methods was to make prediction of assembly pathway without relying on the final structure. Not all protein structures can be crystallized as the fully assembled complex and BSA method cannot be used for such complexes. Although not using any information from the native structure, these two methods also performed better for the complex with good prediction model by Multi-LZerD (Table 4.2). These results agree with our observation that prediction of the model with good topology plays a significant role in the successful prediction of assembly pathway.

4.2 The effect of conversion to the rank by ITScore

The conversion of the ranks by shape score to the ranks by ITScore did not make significant improvement in the prediction success rate. We expected the ITScore to improve the performance, because, as described in Chapter 2, it is

more accurate than the scoring function of the shape score. However, the ITScore did not improve the success rate for the complexes, especially for the complex with 6 and 7 subunits. This might be due to the way ITScore was constructed; since the scoring function was trained on homodimers and heterodimers, they may not have included the interaction that occur among the subunits in the complexes with larger number of subunits. However, ITScore was capturing some of the dimerization events in the assembly even when the overall assembly prediction had failed. This is explained in the next section.

4.3 Prediction of dimers in the pathway

None of the prediction methods, including the BSA method, was able to predict correct pathway for the asymmetric large complexes having 6 or 7 subunits. However, we have found that our method was predicting some of the key steps even when the prediction of the complete pathway was failing. There were 5 such cases: 3VYT, 4HI0, 4IGC, 3UKU, and 4GWP. Figure 4.1 is the example of the output by final generation method for 3VYT using the rank by ITScore. 3VYT is a hexamer complex HypCDE. There were 21 unique assembly pathways, and 19 of them were predicting the dimerization of chain C and F, which both corresponds to HypE subunit. This homodimerization is consistent with the assembly pathway based on experimental results [43]. 4HI0 is a hexamer of UreF/UreH/UreG. The EF dimer, which was predicted in 18 out of 48 pathway, corresponds to UreG homodimer, agrees with the assembly pathway based on the experiment [89].

Two of dimerization events were well predicted for the hexamer 4IGC, a bacterial RNA polymerase. Frequently observed AB dimer and DE dimer corresponds to alpha-alpha homodimer and beta'-omega dimer respectively (Figure 4.2). Out of 57 unique pathways, 51 pathways were correctly predicting either of the two dimers, and 9 on them were predicting the both. A heterodimer important for the complex formation was well predicted for the heptamer 3UKU. Chains D and F correspond to subunit p34 and p20 of Arp2/3 complex respectively. The experiments have revealed that the heterodimer has a critical role in the formation and stability of the whole complex [10]. Dimerization of chains D and F was correctly predicted for all the pathways for 3UKU.

Another heptamer 4GWP, Mediator head module, was also a successful case in identifying a heterodimer. Chain E and F corresponds to subunits Med18 and Med20, which forms a heterodimer [14]. We later found that 4GWP is missing side chain atoms for 63.9 % of the residues. Chains E and F had only few residues with missing side chains. We used OSCAR [93], a side chain prediction program, to reconstruct side chain and ran Multi-LZerD again. None of the methods was able to predict the correct assembly pathway of the complex. Also, EF pair was no longer observed in the consensus across generation method, probably due to the change in the side chain conformation. The original EF dimer and the one with predicted side chain had the RMSD of 1.19Å, indicating the change in side chain conformation. Also, the side chains assigned to other subunits may have given rise to the dimers that receive a better ITScore.

Table 4.1 Results of assembly pathway prediction

	# of chains	PDB ID	RMSD (Å)	BSA		Rank by shape score								Rank by ltscore					
						Low-RMSD decoy		Lowest RMSD		Final population		Consensus across gen.		Lowest RMSD		Final population		Consensus across gen.	
1	3	1A0R	0.84	x	1/1	x	1/1	x	1/1	x	1/1	x	1/1	-	1/1	x	1/2	x	1/2
2		1B9X	0.63	x	1/1	-	0/1	-	0/1	x	1/2	x	1/3	x	1/1	x	1/2	x	1/3
3		1VCB	1.15	-	0/1	x	1/1	x	1/1	x	1/3	x	1/3	x	1/1	x	1/3	x	1/3
4		2AZE	0.99	x	1/1	-	0/1	x	1/1	x	1/1	x	1/1	x	1/1	x	1/1	x	1/1
5		1IKN	14.51	-	0/1	-	0/1	-	0/1	x	1/3	x	1/3	-	0/1	x	1/3	x	1/3
6	4	1GPQ	1.74	x	2/2	x	2/2	x	2/2	x	1/3	x	1/14	x	2/2	x	1/3	x	1,2/14
7		1ES7	1.86	x	2/2	x	2/2	x	2/2	x	1,3/3	x	1,3/5	x	2/2	x	1/3	x	1,3/6
8		1REW	11.02	x	2/2	x	2/2	x	2/2	x	1,2/12	x	1,2/16	x	2/2	x	1,2/8	x	1,2/11
9		2BQ1	24.27	x	2/2	-	0/2	-	1/2	-	10/10	-	10/13	-	1/2	-	8,10/12	-	10,12/15
10		2QSP	18.4	x	2/2	x	2/2	-	0/2	x	1,9,11/16	x	1,8,11,17/18	-	0/2	x	1,12,13/15	x	1,7,17/17
11		3FH6	35.72	x	2/2	-	0/2	-	1/2	-	4/9	-	7,9/12	-	1/2	-	3/6	-	4/6
12		2E9X	9.5	-	1/2	-	0/2	x	2/2	-	4/9	-	2/10	-	1/2	-	5/10	-	4/17
13		1KF6	22.22	x	1/1	x	1/1	-	0/1	-	0/10	-	2/10	-	1/2	-	0/9	-	5/13
14	5	1HEZ	11.73	x	2/2	x	2/2	x	2/2	x	1,2/3	x	1,2,4/14	x	2/2	x	1,2/3	x	1,2,7/13
15		1W88	4.8	x	1/1	x	2/2	x	1/1	-	6/9	-	10,16,18/42	x	1/1	-	3/6	-	5,15,24/34
16	6	1RLB	22.99	x	2/2	-	0/2	x	2/2	x	1,2/28	x	1,2,3,10,61,66,67/77	x	2/2	x	1,2/22	x	1,2,3,15,55,70/73
17		1DU3	20.86	x	1/1	x	1/1	-	0/1	-	0/11	-	0/23	-	0/1	x	1,2,3/9	x	1,2,3/12
18		1S5B	22.09	-	0/2	-	0/2	x	2/2	-	2,8/10	-	3,9/15	-	1/2	x	1/10	x	1,11/16
19		3VYT	36.81	x	2/2	x	2/2	-	0/2	-	0/24	-	0/38	-	0/2	-	0/18	-	0/21
20		4HI0	40.8	-	0/2	-	1/2	-	0/2	-	0/30	-	0/55	-	1/2	-	4,12,14,21/27	-	6,11,13,1832/48
21		4IGC	53.5	-	0/3	-	1/3	-	1/3	-	0/30	-	0/49	-	1/3	-	4/33	-	4/58
22	7	3UKU	36.6	-	0/3	-	0/3	-	1/3	-	0/84	-	0/210	-	0/3	-	0/91	-	0/227
23		4GWP	34.24	-	0/3	-	0/3	-	0/3	-	0/49	-	0/97	-	0/3	-	0/39	-	0/71
	Count of success			15		11		11		11		11		9		13		13	
	Success rate			0.65		0.48		0.48		0.48		0.48		0.39		0.57		0.57	

4th column shows the RMSD of the best model predicted by Multi-LZerD. 5th column shows the prediction result by the BSA method; cross in the first sub column indicates the success, and the second sub column shows the how many steps were correctly predicted. The columns titled “Low-RMSD decoy” and “Lowest RMSD” shows the results from Low-RMSD decoy combination method and lowest RMSD method respectively. The result format is same as the BSA method. The column titled “Final gen.” and “consensus across gen.” corresponds to final population method and consensus across generation method respectively. The results are presented differently for these two columns; the cross in the first sub column indicates the success, and the second sub column shows the number of occurrence of the correct pathway within the

Table 4.1 continued

generation(s). The number at the right side of the slash shows the number of pathways observed in the generation(s), and the number(s) at the left side shows the rank(s) of the correct pathway. For example, 1,2/4 mean that the correct pathway is ranked 1st and 2nd among 4 different pathway observed.

Table 4.2 Results for the models with good docking predictions

	# of chains	PDB ID	RMSD (Å)	BSA		Rank by shape score								Rank by Itscore					
						Low-RMSD decoy		Lowest RMSD		Final population		Consensus across gen.		Lowest RMSD		Final population		Consensus across gen.	
1	3	1A0R	0.84	x	1/1	x	1/1	x	1/1	x	1/1	x	1/1	-	1/1	x	1/2	x	1/2
2		1B9X	0.63	x	1/1	-	0/1	-	0/1	x	1/2	x	1/3	x	1/1	x	1/2	x	1/3
3		1VCB	1.15	-	0/1	x	1/1	x	1/1	x	1/3	x	1/3	x	1/1	x	1/3	x	1/3
4		2AZE	0.99	x	1/1	-	0/1	x	1/1	x	1/1	x	1/1	x	1/1	x	1/1	x	1/1
5		1IKN	14.51	-	0/1	-	0/1	-	0/1	x	1/3	x	1/3	-	0/1	x	1/3	x	1/3
6	4	1GPQ	1.74	x	2/2	x	2/2	x	2/2	x	1/3	x	1/14	x	2/2	x	1/3	x	1,2/14
7		1ES7	1.86	x	2/2	x	2/2	x	2/2	x	1,3/3	x	1,3/5	x	2/2	x	1/3	x	1,3/6
8		1REW	11.02	x	2/2	x	2/2	x	2/2	x	1,2/12	x	1,2/16	x	2/2	x	1,2/8	x	1,2/11
9		2E9X	9.5	-	1/2	-	0/2	x	2/2	-	4/9	-	2/10	-	1/2	-	5/10	-	4/17
10	5	1HEZ	11.73	x	2/2	x	2/2	x	2/2	x	1,2/3	x	1,2,4/14	x	2/2	x	1,2/3	x	1,2,7/13
11		1W88	4.8	x	1/1	x	2/2	x	1/1	-	6/9	-	10,16,18/42	x	1/1	-	3/6	-	5,15,24/34
	Count of success			8		7		9		9		9		8		9		9	
	Success rate			0.73		0.64		0.82		0.82		0.82		0.73		0.82		0.82	

The table format follows that of Table 4.1.

```
Rank      1, count 151238: CF>CF,BE>CF,BDE>CF,ABDE>ABCDEF
Rank      2, count 140285: CF>CF,BE>CF,BDE>ACF,BDE>ABCDEF
Rank      3, count  26983: CF>CF,BE>CF,BE,AD>ABDE,CF>ABCDEF
Rank      4, count  16109: CF>CF,BD>CF,BDE>CF,ABDE>ABCDEF
Rank      5, count  12332: CF>CF,BE>CDF,BE>ACDF,BE>ABCDEF
Rank      6, count  11016: CF>CF,BE>CF,BE,AD>ACDF,BE>ABCDEF
Rank      7, count  10453: AD>AD,CF>AD,CF,BE>ABDE,CF>ABCDEF
Rank      8, count   5895: CF>CF,BE>ACF,BE>ACF,BDE>ABCDEF
Rank      9, count   5434: AD>AD,CF>AD,CF,BE>ACDF,BE>ABCDEF
Rank     10, count   4701: CF>CF,BE>CF,ABE>CF,ABDE>ABCDEF
Rank     11, count   4568: CF>CF,BD>CF,BDE>ACF,BDE>ABCDEF
Rank     12, count   2880: CF>CF,AD>CF,AD,BE>ACDF,BE>ABCDEF
Rank     13, count   2687: CF>CF,AD>CF,AD,BE>ABDE,CF>ABCDEF
Rank     14, count   1716: AB>AB,CF>ABE,CF>ABDE,CF>ABCDEF
Rank     15, count   1141: CF>CF,BE>ACF,BE>ACDF,BE>ABCDEF
Rank     16, count   1085: AF>ACF>ACF,BE>ACF,BDE>ABCDEF
Rank     17, count   1000: CF>CF,BE>CDF,BE>CDF,ABE>ABCDEF
Rank     18, count    857: CF>CF,BE>CF,ABE>CDF,ABE>ABCDEF
Rank     19, count     17: DF>CDF>CDF,BE>ACDF,BE>ABCDEF
Rank     20, count      2: CF>CF,BE>CF,BDE>BCDEF>ABCDEF
Rank     21, count      1: CF>CF,BD>CF,BDE>BCDEF>ABCDEF
```

Figure 4.1 Pathway prediction for 3VYT by consensus across generation method
The occurrence of each pathway is counted. Then they are sorted according to the count. The prediction of CF dimer is highlighted in bold text with underline.

```
Rank      1, count 136711: AB>AB,DE>ABDE>ABCDE>ABCDEX
Rank      2, count  55200: AB>AB,DE>ABDE>ABDE,CX>ABCDEX
Rank      3, count  45017: AB>AB,DE>ABDE>ABDEX>ABCDEX
Rank      4, count  27634: AB>AB,DE>ABC,DE>ABCDE>ABCDEX
Rank      5, count  21970: AB>AB,DE>AB,CDE>ABCDE>ABCDEX
Rank      6, count  19681: DE>ADE>ACDE>ABCDE>ABCDEX
Rank      7, count  12454: DE>CDE>ACDE>ABCDE>ABCDEX
Rank      8, count  12377: AB>AB,DE>AB,DE,CX>ABDE,CX>ABCDEX
Rank      9, count  10273: DE>BDE>ABDE>ABCDE>ABCDEX
Rank     10, count   7681: DE>DE,AB>ABDE>ABCDE>ABCDEX
Rank     11, count   5112: DE>DE,BC>ADE,BC>ABCDE>ABCDEX
Rank     12, count   4450: AC>ABC>ABC,DE>ABCDE>ABCDEX
Rank     13, count   3933: DE>ADE>ADE,BC>ABCDE>ABCDEX
Rank     14, count   3273: DE>DE,AC>ACDE>ABCDE>ABCDEX
Rank     15, count   3178: DE>DE,BX>ADE,BX>ACDE,BX>ABCDEX
Rank     16, count   2857: DE>ADE>ADE,BX>ACDE,BX>ABCDEX
Rank     17, count   2617: DE>DE,AC>DE,ABC>ABCDE>ABCDEX
Rank     18, count   2490: AB>AB,CD>AB,CDE>ABCDE>ABCDEX
Rank     19, count   2340: DE>ADE>ABDE>ABCDE>ABCDEX
Rank     20, count   2106: DE>ADE>ACDE>ACDE,BX>ABCDEX
Rank     21, count   1827: DE>DE,AB>CDE,AB>ABCDE>ABCDEX
Rank     22, count   1817: BX>BX,DE>BX,CDE>BX,ACDE>ABCDEX
Rank     23, count   1565: DE>DE,BC>BCDE>ABCDE>ABCDEX
Rank     24, count   1375: DE>CDE>ACDE>ACDE,BX>ABCDEX
Rank     25, count   1295: DE>ADE>ABDE>ABDEX>ABCDEX
Rank     26, count   1270: AB>AB,DE>ABC,DE>ABCX,DE>ABCDEX
Rank     27, count   1075: BC>BC,DE>BC,ADE>ABCDE>ABCDEX
Rank     28, count    963: DE>DE,BX>CDE,BX>ACDE,BX>ABCDEX
Rank     29, count    960: DE>CDE>CDE,AB>ABCDE>ABCDEX
Rank     30, count    893: AB>AB,CX>AB,CX,DE>ABDE,CX>ABCDEX
Rank     31, count    816: BC>BC,DE>BCDE>ABCDE>ABCDEX
Rank     32, count    717: BD>BDE>ABDE>ABCDE>ABCDEX
Rank     33, count    665: DE>DE,BC>DE,BCX>DE,ABCX>ABCDEX
Rank     34, count    593: BX>BX,DE>BX,ADE>BX,ACDE>ABCDEX
Rank     35, count    520: BX>BX,DE>BX,DE,AC>ACDE,BX>ABCDEX
Rank     36, count    454: DE>DE,BC>ADE,BC>ADEX,BC>ABCDEX
Rank     37, count    448: DE>ADE>ABDE>ABDE,CX>ABCDEX
Rank     38, count    415: AE>ADE>ADE,BC>ABCDE>ABCDEX
Rank     39, count    409: DE>DE,AB>ABDE>ABDEX>ABCDEX
Rank     40, count    331: DE>CDE>BCDE>ABCDE>ABCDEX
Rank     41, count    152: CX>CX,AB>CX,AB,DE>ABDE,CX>ABCDEX
Rank     42, count    127: CD>CD,AB>CDE,AB>ABCDE>ABCDEX
Rank     43, count    105: DE>ADE>ADE,BC>ADE,BCX>ABCDEX
Rank     44, count    101: BE>BDE>BCDE>ABCDE>ABCDEX
Rank     45, count     79: DE>BDE>BCDE>ABCDE>ABCDEX
Rank     46, count     74: DE>DE,AB>ABDE>ABDE,CX>ABCDEX
Rank     47, count     72: DE>DE,BX>ADE,BX>ABDEX>ABCDEX
Rank     48, count     59: DE>CDE>CDE,BX>ACDE,BX>ABCDEX
Rank     49, count     51: DE>DE,AB>DE,ABC>DE,ABCX>ABCDEX
Rank     50, count      4: DE>DE,BX>DE,BX,AC>ACDE,BX>ABCDEX
Rank     51, count      3: DE>BDE>ABDE>ABDE,CX>ABCDEX
Rank     52, count      2: DE>DE,AC>ACDE>ACDE,BX>ABCDEX
Rank     53, count      2: DE>DE,AC>DE,AC,BX>ACDE,BX>ABCDEX
Rank     54, count      2: DE>DE,AC>DE,ABC>DE,ABCX>ABCDEX
Rank     55, count      2: DE>DE,BC>ADE,BC>ADE,BCX>ABCDEX
Rank     56, count      1: DE>ADE>ADE,CX>ABDE,CX>ABCDEX
Rank     57, count      1: BX>BX,DE>BX,CDE>ABX,CDE>ABCDEX
```

Figure 4.2 Pathway prediction for 4HI0 by consensus across generation method Format is the same as Figure 4.1. The prediction of AB and DE dimers are highlighted.

CHAPTER 5. DISCUSSION

5.1 Why the ranks can be used for the assembly prediction

Our model of pathway prediction method is based on the assumption that the ranks can approximate the Gibbs energy of protein-protein interaction, ΔG_{bind}, which is shown in the equation below. G_{AB}, G_A, and G_B indicates the Gibbs energy of protein A and B in bound state, protein A alone, and protein B alone respectively.

$$\Delta G_{bind} = G_{AB} - (G_A + G_B)$$

In nature, proteins will interact and form a complex if the complex formation is energetically favorable, i.e., if ΔG of the reaction is below. The plot of shape score against the rank shows smooth curve, which makes it reasonable to assume that shape score is continuous (Figure 5.1). Based on the fact that the rank of the decoys used for the model with low RMSD is usually above 1,000, and that the difference of the scores between the ranks are fairly constant for the decoys ranked above 1,000, we assume that the ranks can approximate the ΔG_{bind}. Additionally, we are also making the assumption that the complex formation does not affect the affinity of subunit-subunit interaction. Under these assumptions, our model predicts the assembly pathway of protein complexes based on the idea that assembly starts from the decoys with higher rank.

ITScore has a correlation with actual binding affinity of protein-protein interaction [53] (Figure 5.2). Binding affinity can be directly converted to ΔG_{bind} using the equation $\Delta G = -RT \ln K_d$, where R and T stands for gas constant and absolute temperature respectively. Therefore, the rank by ITScore can also be used for predicting the assembly pathway as the rank by shape score.

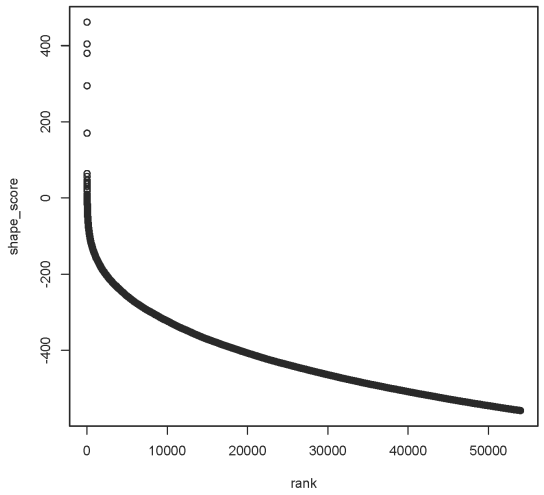


Figure 5.1 Plot of shape score and rank
Shape score of non-clustered 2AZE decoy A-B plotted against the rank of the decoy.

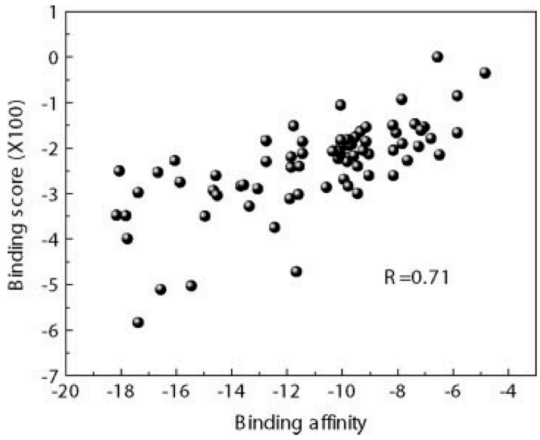


Figure 5.2 Plot of ITScore and binding affinity taken from [53]
ITscore is plotted against the experimentally determined binding affinity of protein-protein interaction.

5.2 Improving the performance of assembly pathway prediction

The results showed that our assembly pathway prediction is hugely affected by the modeling accuracy of Multi-LZerD; obtaining a near-native model or a model with good topology is crucial for the prediction of the correct assembly. Therefore, improving the modeling accuracy of Multi-LZerD is likely to improve the accuracy of our pathway prediction methods. The physics-based scoring function, which is used to evaluate the complex at each generation in Multi-LZerD, is sometimes the bottleneck for the correct structure prediction. Whether a near-native complex is kept in a generation or not depends on the score given by the scoring function. Sometimes, the scoring function fails to give good scores for near native models. There are also cases where the native structure is given a bad physics score. Usually, the range of the physics score of native structures is from -5,000 to -20,000, with negative score being better. However, the physics score of the native structure for 2QSP, bovine hemoglobin, has an unusually high value of -1298. We expected the docking prediction would yield near-native structure for 2QSP; hemoglobin consist of dimer of heterodimer with A_2B_2 stoichiometry, and Multi-LZerD was successful in modeling the complexes with similar topology. The unusual score of the native complex may explain the modeling failure for 2QSP. Increasing the generation will not improve the modeling if the scoring function is not working. We have checked the physics score of the native structure where the modeling resulted in a complex with a RMSD larger than 10Å, and compared it with the physics score of the model that had lowest RMSD among the final population. Out of 18 such complexes, 15 of

them had a better physics score than the native structure, which means that additional generation would not bring the models closer to the native structure in terms of the physics score.

One simple strategy for improving the modeling accuracy is to increase the population size. One of the follow up research of Multi-LZerD revealed that an excessive conformational search is not required for complexes with less than four subunits [54]. However, larger conformational search might improve the modeling accuracy for protein with larger number of subunits, since they could have larger conformation space. We have set the population size as 200 for all the complexes in our dataset. We may need to increase them for the complexes with 6 and 7 chains in order to obtain the models with near-native structure, or the ones with good topology.

5.3 Difference between BSA method and shape-score based methods

Table 4.2 shows that our pathway prediction is performing slightly better than the BSA method for the 11 complexes where Multi-LZerD was able to generate models with near-native RMSD or good topology. Although the difference is not significant due to the size of the dataset, we decided to investigate what made the difference and why the shape-based scores are predicting the assembly pathway in terms of Gibbs energy. Gibbs energy of binding, ΔG_{bind}, can be decomposed as shown in the equation below [94].

$$\Delta G_{bind} = \Delta G_{int} + \Delta G_{solv} + \Delta G_{motion} + \Delta G_{conf}$$

ΔG_{int}, ΔG_{solv}, ΔG_{motion}, and ΔG_{conf} correspond to Gibbs energy of interaction, solvation, motion, and conformation. The last two terms are entropic term, which is not considered in scoring functions. ΔG_{int} and ΔG_{solv} can be further decomposed as shown in Figure 5.3. ΔG_{int} can be decomposed to van der Waals interaction and electrostatic interaction. ΔG_{solv} can be decomposed to electrostatic and nonelectrostatic part [95]. The ΔG_{solv}^{elect} is reflects the energy of interaction between protein and water, and $\Delta G_{solv}^{nonelect}$ reflects the energy of creating the cavity in solution to place the protein, which is proportional to the SASA. The BSA method is basically considering $\Delta G_{solv}^{nonelect}$ part of ΔG_{bind} to predict the assembly pathway. Although shape-based scoring function only considers the geometric feature of the protein-protein interaction, it implicitly considers few elements of ΔG_{bind}. The shape match and clash penalty corresponds to the attractive and repulsive part of van der Waals interaction respectively. Also, shape-based scoring function considers the interface size. Thus, it takes in account the $\Delta G_{solv}^{nonelect}$ as BSA method does. Therefore, we can conclude that the shape-based scoring function better approximates the ΔG_{bind} compared to the BSA method, which may be the reason for the better prediction performance of our prediction methods. Although, we are still working on the investigation of the mechanism behind the prediction using the ranks based on the scores obtained from scoring functions.

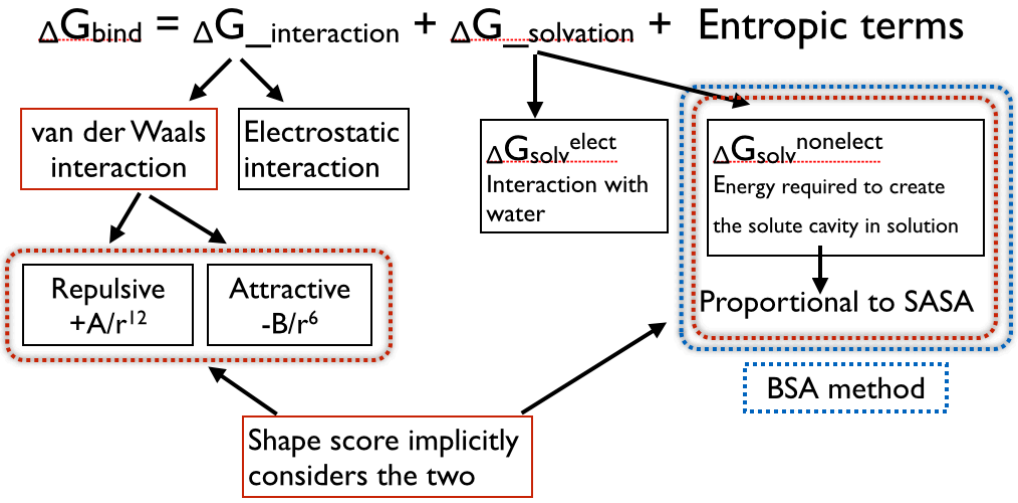


Figure 5.3 Decomposition of ΔG_{bind}
Boxes colored in red is the element implicitly considered in shape score. BSA method only considers the nonelectric portion of ΔG_{solv}, which is proportional to the SASA.

5.4 Future direction

Our result indicates that successful multiple docking also leads to successful prediction of assembly pathway. As discussed in section 5.2, more exhaustive structure search by larger population and more iteration may lead to successful docking. Another thing we could do is to find and optimize the Multi-LZerD parameter suitable for each complex. The complexes in our dataset obtained from previous Multi-LZerD publication was tested with different parameters to obtain near native models. All three types of decoys generated by LZerD, non-clustered decoy and the decoys clustered by 5 and 10 Å cutoff were tested with different atom clash thresholds. On the other hand, Multi-LZerD was run using a single setting for the complexes found for this project. Finding an optimal setting may lead to the improvement in the performance of docking and assembly pathway prediction.

Although not used in this project, Multi-LZerD can be run using interface information. Using the information, Multi-LZerD can restrict the structure search to a conformation that involves the interface residues. Information of interface residues can be obtained not only from the crystal structure, but also from experiments. This may lead to the reduction of incorrect structures and help Multi-LZerD perform the docking efficiently. Rewarding the pair of subunits that are known to interact is another possibility. We have showed that Multi-LZerD was capable in detecting some dimerization events even when the model was not having near native RMSD. By rewarding the decoys that are known to interact, we can save the time that Multi-LZerD required to select the decoy among the others.

5.5 Conclusion

We have developed four methods to predict the assembly pathway of protein complexes using Multi-LZerD, a multiple docking algorithm for asymmetric complexes. Using the manually curated dataset that includes the complexes varying in size, number of subunit, and topology, we have benchmarked our method along with the BSA-based method. We confirmed that the BSA method is able to predict the assembly pathway of both symmetric and asymmetric complexes. While our methods had lower performance compared to the BSA method, our method was successful for the complexes where Multi-LZerD was able to model near-native structures or the structures with good topology. The result indicates the importance of the modeling accuracy in the

success of our assembly pathway prediction. Also, our method was able to capture some dimerization steps, even when the overall pathway prediction failed. Although not complete, our work demonstrates that a multiple docking algorithm can be applied to predict assembly pathway of protein complexes.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Beadle GW, Tatum EL: **Genetic control of biochemical reactions in Neurospora**. *Proc. Natl. Acad. …* 1941, **27**:499–506.

2. Petsko GA, Ringe D: *Protein Stucture and Function (Primers in Biology) (PRIMER IN BIOLOGY)*. Sinauer Associates, Inc.; 2003:180.

3. Watson J, Crick F: **Molecular structure of nucleic acid**. *Nature* 1953.

4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res.* 2000, **28**:235–42.

5. Chandonia J-M, Brenner SE: **The impact of structural genomics: expectations and outcomes.** *Science* 2006, **311**:347–51.

6. Goodsell D, Olson A: **Structural symmetry and protein function**. *Annu. Rev. Biophys. …* 2000.

7. Ali MH, Imperiali B: **Protein oligomerization: how and why.** *Bioorg. Med. Chem.* 2005, **13**:5013–20.

8. Mueller M, Jenni S, Ban N: **Strategies for crystallization and structure determination of very large macromolecular assemblies.** *Curr. Opin. Struct. Biol.* 2007, **17**:572–9.

9. Robinson RC, Turbedsky K, Kaiser D a, Marchand JB, Higgs HN, Choe S, Pollard TD: **Crystal structure of Arp2/3 complex.** *Science* 2001, **294**:1679–84.

10. Gournier H, Goley ED, Niederstrasser H, Trinh T, Welch MD: **Reconstitution of human Arp2/3 complex reveals critical roles of individual subunits in complex structure and activity.** *Mol. Cell* 2001, **8**:1041–52.

11. Zhao X, Yang Z, Qian M, Zhu X: **Interactions among subunits of human Arp2/3 complex: p20-Arc as the hub.** *Biochem. Biophys. Res. Commun.* 2001, **280**:513–7.

12. Machesky LM: **Cell motility: complex dynamics at the leading edge.** *Curr. Biol.* 1997, **7**:R164–7.

13. Mullins RD, Stafford WF, Pollard TD: **Structure, subunit topology, and actin-binding activity of the Arp2/3 complex from Acanthamoeba.** *J. Cell Biol.* 1997, **136**:331–43.

14. Takagi Y, Calero G, Komori H, Brown J a, Ehrensberger AH, Hudmon A, Asturias F, Kornberg RD: **Head module control of mediator interactions.** *Mol. Cell* 2006, **23**:355–64.

15. Guglielmi B, van Berkum NL, Klapholz B, Bijma T, Boube M, Boschiero C, Bourbon H-M, Holstege FCP, Werner M: **A high resolution protein interaction map of the yeast Mediator complex.** *Nucleic Acids Res.* 2004, **32**:5379–91.

16. Kang JS, Kim SH, Hwang MS, Han SJ, Lee YC, Kim YJ: **The structural and functional organization of the yeast mediator complex.** *J. Biol. Chem.* 2001, **276**:42003–10.

17. Imasaki T, Calero G, Cai G, Tsai K-LK, Yamada K, Cardelli F, Erdjument-Bromage H, Tempst P, Berger I, Kornberg GL, Asturias FJ, Kornberg RD, Takagi Y: **Architecture of the Mediator head module.** *Nature* 2011, **475**:240–3.

18. Dalbey RE, Wang P, Kuhn A: **Assembly of bacterial inner membrane proteins.** *Annu. Rev. Biochem.* 2011, **80**:161–87.

19. Daley DO: **The assembly of membrane proteins into complexes.** *Curr. Opin. Struct. Biol.* 2008, **18**:420–4.

20. Levy ED, Boeri Erba E, Robinson C V, Teichmann SA, Erba EB: **Assembly reflects evolution of protein complexes.** *Nature* 2008, **453**:1262–5.

21. Marsh J a, Hernández H, Hall Z, Ahnert SE, Perica T, Robinson C V, Teichmann S a: **Protein complexes are under evolutionary selection to assemble via ordered pathways.** *Cell* 2013, **153**:461–70.

22. Friedman FK, Beychok S: **Probes of subunit assembly and reconstitution pathways in multisubunit proteins.** *Annu. Rev. Biochem.* 1979, **48**:217–50.

23. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE: **How fast-folding proteins fold.** *Science* 2011, **334**:517–20.

24. Levinthal C: **How to fold graciously**. *Mossbauer Spectrosc. …* 1969, **24**:22–24.

25. Sanders CR, Myers JK: **Disease-related misassembly of membrane proteins.** *Annu. Rev. Biophys. Biomol. Struct.* 2004, **33**:25–51.

26. Sanchez de Groot N, Torrent M, Villar-Piqué A, Lang B, Ventura S, Gsponer J, Babu MM: **Evolutionary selection for protein aggregation.** *Biochem. Soc. Trans.* 2012, **40**:1032–7.

27. Rak M, Gokova S, Tzagoloff A: **Modular assembly of yeast mitochondrial ATP synthase.** *EMBO J.* 2011, **30**:920–30.

28. Tan ALC, Rida PCG, Surana U: **Essential tension and constructive destruction: the spindle checkpoint and its regulatory links with mitotic exit.** *Biochem. J.* 2005, **386**:1–13.

29. Roy B, Varshney N, Yadav V, Sanyal K: **The process of kinetochore assembly in yeasts.** *FEMS Microbiol. Lett.* 2013, **338**:107–17.

30. Seraphin B, Rosbash M: **Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing.** *Cell* 1989, **59**:349–58.

31. Jamison SF, Crow A, Garcia-Blanco MA: **The spliceosome assembly pathway in mammalian extracts.** *Mol. Cell. Biol.* 1992, **12**:4279–87.

32. Kennedy K a, Gachelet EG, Traxler B: **Evidence for multiple pathways in the assembly of the Escherichia coli maltose transport complex.** *J. Biol. Chem.* 2004, **279**:33290–7.

33. Aldridge P: **Regulation of flagellar assembly**. *Curr. Opin. Microbiol.* 2002, **5**:160–165.

34. Venkatakrishnan a J, Levy ED, Teichmann S a: **Homomeric protein complexes: evolution and assembly.** *Biochem. Soc. Trans.* 2010, **38**:879–82.

35. Mulkidjanian AY, Makarova KS, Galperin MY, Koonin E V: **Inventing the dynamo machine: the evolution of the F-type and V-type ATPases.** *Nat. Rev. Microbiol.* 2007, **5**:892–9.

36. Gomis-Rüth FX, Moncalián G, Pérez-Luque R, González a, Cabezón E, de la Cruz F, Coll M: **The bacterial conjugation protein TrwB resembles ring helicases and F1-ATPase.** *Nature* 2001, **409**:637–41.

37. Walker J: **ATP Synthesis by Rotary Catalysis (Nobel lecture)****. 1997.

38. Hardy SJ, Holmgren J, Johansson S, Sanchez J, Hirst TR: **Coordinated assembly of multisubunit proteins: oligomerization of bacterial enterotoxins in vivo and in vitro.** *Proc. Natl. Acad. Sci. U. S. A.* 1988, **85**:7109–13.

39. Tinker JK, Erbe JL, Hol WGJ, Holmes RK: **Cholera holotoxin assembly requires a hydrophobic domain at the A-B5 interface: mutational analysis and development of an in vitro assembly system.** *Infect. Immun.* 2003, **71**:4093–101.

40. Young KH: **Yeast two-hybrid: so many interactions, (in) so little time...** *Biol. Reprod.* 1998, **58**:302–11.

41. Hernández H, Robinson C V: **Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry.** *Nat. Protoc.* 2007, **2**:715–26.

42. Heck AJR: **Native mass spectrometry: a bridge between interactomics and structural biology.** *Nat. Methods* 2008, **5**:927–33.

43. Watanabe S, Matsumi R, Atomi H, Imanaka T, Miki K: **Crystal structures of the HypCD complex and the HypCDE ternary complex: transient intermediate complexes during [NiFe] hydrogenase maturation.** *Structure* 2012, **20**:2124–37.

44. Chen J, Sawyer N, Regan L: **Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area.** *Protein Sci.* 2013, **22**:510–5.

45. Lee B, Richards FM: **The interpretation of protein structures: estimation of static accessibility.** *J. Mol. Biol.* 1971, **55**:379–400.

46. Levy E, Pereira-Leal J: **3D complex: a structural classification of protein complexes**. *PLoS Comput. …* 2006, **2**.

47. Levy ED, Teichmann S: *Structural, evolutionary, and assembly principles of protein oligomerization.* 1st edition. Copyright © 2013, Elsevier Inc. All Rights Reserved.; 2013, **117**:25–51.

48. Esquivel-Rodríguez J, Yang YD, Kihara D: **Multi-LZerD: multiple protein docking for asymmetric complexes.** *Proteins* 2012, **80**:1818–33.

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

50. Loew a, Ho YK, Blundell T, Bax B: **Phosducin induces a structural change in transducin beta gamma.** *Structure* 1998, **6**:1007–19.

51. Dingus J, Hildebrandt JD: **GPCR Signalling Complexes – Synthesis, Assembly, Trafficking and Specificity**. 2012, **63**:155–180.

52. Gaudet R, Savage J, McLaughlin J: **A Molecular mechanism for the phosphorylation-dependent regulation of heterotrimeric G proteins by phosducin**. *Mol. Cell* 1999, **3**:649–660.

53. Huang S-Y, Zou X: **An iterative knowledge-based scoring function for protein-protein recognition.** *Proteins* 2008, **72**:557–79.

54. Esquivel-Rodríguez J, Kihara D: **Effect of conformation sampling strategies in genetic algorithm for multiple protein docking.** *BMC Proc.* 2012, **6 Suppl 7**:S4.

55. Stebbins CE, Kaelin WG, Pavletich NP: **Structure of the VHL-ElonginC-ElonginB complex: implications for VHL tumor suppressor function.** *Science* 1999, **284**:455–61.

56. Duan DR, Pause a, Burgess WH, Aso T, Chen DY, Garrett KP, Conaway RC, Conaway JW, Linehan WM, Klausner RD: **Inhibition of transcription elongation by the VHL tumor suppressor protein.** *Science* 1995, **269**:1402–6.

57. Rubin SM, Gall A-L, Zheng N, Pavletich NP: **Structure of the Rb C-terminal domain bound to E2F1-DP1: a mechanism for phosphorylation-induced E2F release.** *Cell* 2005, **123**:1093–106.

58. Bandara LR, Lam EW, Sørensen TS, Zamanian M, Girling R, La Thangue NB: **DP-1: a cell cycle-regulated and phosphorylated component of transcription factor DRTF1/E2F which is functionally important for recognition by pRb and the adenovirus E4 orf 6/7 protein.** *EMBO J.* 1994, **13**:3104–14.

59. Huxford T, Huang D-B, Malek S, Ghosh G: **The Crystal Structure of the IκBα/NF-κB Complex Reveals Mechanisms of NF-κB Inactivation**. *Cell* 1998, **95**:759–770.

60. Gilmore TD: **Introduction to NF-kappaB: players, pathways, perspectives.** *Oncogene* 2006, **25**:6680–4.

61. Perkins ND: **Integrating cell-signalling pathways with NF-kappaB and IKK function.** *Nat. Rev. Mol. Cell Biol.* 2007, **8**:49–62.

62. Abergel C, Monchois V, Byrne D, Chenivesse S, Lembo F, Lazzaroni J-C, Claverie J-M: **Structure and evolution of the Ivy protein family, unexpected lysozyme inhibitors in Gram-negative bacteria.** *Proc. Natl. Acad. Sci. U. S. A.* 2007, **104**:6394–9.

63. Monchois V, Abergel C, Sturgis J, Jeudy S, Claverie JM: **Escherichia coli ykfE ORFan gene encodes a potent inhibitor of C-type lysozyme.** *J. Biol. Chem.* 2001, **276**:18437–41.

64. Cegielska-radziejewska R, Leśnierowski G, Kijowski J: **PROPERTIES AND APPLICATION OF EGG WHITE LYSOZYME AND ITS MODIFIED PREPARATIONS – A REVIEW**. *Polish J. …* 2008, **58**:5–10.

65. Maroufi B, Ranjbar B, Khajeh K, Naderi-Manesh H, Yaghoubi H: **Structural studies of hen egg-white lysozyme dimer: comparison with monomer.** *Biochim. Biophys. Acta* 2008, **1784**:1043–9.

66. Kirsch T, Sebald W, Dreyer MK: **Crystal structure of the BMP-2-BRIA ectodomain complex.** *Nat. Struct. Biol.* 2000, **7**:492–6.

67. Scheufler C, Sebald W, Hülsmeyer M: **Crystal structure of human bone morphogenetic protein-2 at 2.7 A resolution.** *J. Mol. Biol.* 1999, **287**:103–15.

68. Keller S, Nickel J, Zhang J-L, Sebald W, Mueller TD: **Molecular recognition of BMP-2 and BMP receptor IA.** *Nat. Struct. Mol. Biol.* 2004, **11**:481–8.

69. Kamada K, Kubota Y, Arata T, Shindo Y, Hanaoka F: **Structure of the human GINS complex and its assembly and functional interface in replication initiation.** *Nat. Struct. Mol. Biol.* 2007, **14**:388–96.

70. Boskovic J, Coloma J, Aparicio T, Zhou M, Robinson C V, Méndez J, Montoya G: **Molecular architecture of the human GINS complex.** *EMBO Rep.* 2007, **8**:678–84.

71. Aranda R, Cai H, Worley CE, Levin EJ, Li R, Olson JS, Phillips GN, Richards MP: **Structural analysis of fish versus mammalian hemoglobins: effect of the heme pocket environment on autooxidation and hemin loss.** *Proteins* 2009, **75**:217–30.

72. Bunn H: **subunit assembly of hemoglobin: An Important determinant of Hematologic phenotype**. *Blood* 1987, **1**:1–6.

73. Liu J, Konermann L: **assembly of hemoglobin from denatured monomeric subunits: Heme ligation effects and off-pathway intermediates studied by electrospray mass spectrometry**. *Biochemistry* 2013.

74. Boys BL, Konermann L: **Folding and assembly of hemoglobin monitored by electrospray mass spectrometry using an on-line dialysis system.** *J. Am. Soc. Mass Spectrom.* 2007, **18**:8–16.

75. Khare D, Oldham ML, Orelle C, Davidson AL, Chen J: **Alternating access in maltose transporter mediated by rigid-body rotations.** *Mol. Cell* 2009, **33**:528–36.

76. Uppsten M, Färnegårdh M, Domkin V, Uhlin U: **The first holocomplex structure of ribonucleotide reductase gives new insight into its mechanism of action.** *J. Mol. Biol.* 2006, **359**:365–77.

77. Brignole EJ, Ando N, Zimanyi CM, Drennan CL: **The prototypic class Ia ribonucleotide reductase from Escherichia coli: still surprising after all these years.** *Biochem. Soc. Trans.* 2012, **40**:523–30.

78. Iverson TM, Luna-Chavez C, Croal LR, Cecchini G, Rees DC: **Crystallographic studies of the Escherichia coli quinol-fumarate reductase with inhibitors bound to the quinol-binding site.** *J. Biol. Chem.* 2002, **277**:16124–30.

79. Latour DJ, Weiner JH: **Assembly of Escherichia coli fumarate reductase holoenzyme.** *Biochem. Cell Biol.* 1989, **67**:251–9.

80. Van Hellemond JJ, Tielens a G: **Expression and functional properties of fumarate reductase.** *Biochem. J.* 1994, **304 (Pt 2**:321–31.

81. Graille M, Stura E a, Housden NG, Beckingham J a, Bottomley SP, Beale D, Taussig MJ, Sutton BJ, Gore MG, Charbonnier JB: **Complex between Peptostreptococcus magnus protein L and a human antibody reveals structural convergence in the interaction modes of Fab binding proteins.** *Structure* 2001, **9**:679–87.

82. Frank R a W, Titman CM, Pratap JV, Luisi BF, Perham RN: **A molecular switch and proton wire synchronize the active sites in thiamine enzymes.** *Science* 2004, **306**:872–6.

83. Kato M, Wynn RM, Chuang JL, Tso S-C, Machius M, Li J, Chuang DT: **Structural basis for inactivation of the human pyruvate dehydrogenase complex by phosphorylation: role of disordered phosphorylation loops.** *Structure* 2008, **16**:1849–59.

84. Monaco HL, Rizzi M, Coda a: **Structure of a complex of two plasma proteins: transthyretin and retinol-binding protein.** *Science* 1995, **268**:1039–41.

85. Trägårdh L, Anundi H, Rask L, Sege K, Peterson P a: **On the stoichiometry of the interaction between prealbumin and retinol-binding protein.** *J. Biol. Chem.* 1980, **255**:9243–8.

86. Hymowitz SG, Christinger HW, Fuh G, Ultsch M, O'Connell M, Kelley RF, Ashkenazi a, de Vos a M: **Triggering cell death: the crystal structure of Apo2L/TRAIL in a complex with death receptor 5.** *Mol. Cell* 1999, **4**:563–71.

87. O'Neal CJ, Amaya EI, Jobling MG, Holmes RK, Hol WGJ: **Crystal structures of an intrinsically active cholera toxin mutant yield insight into the toxin activation mechanism.** *Biochemistry* 2004, **43**:3772–82.

88. Moss J, Iglewski B, Vaughan M, Tu AT: *Handbook of Natural Toxins: Bacterial Toxins and Virulence Factors in Disease, Volume 8*. CRC Press; 1995:664.

89. Fong YH, Wong HC, Yuen MH, Lau PH, Chen YW, Wong K-B: **Structure of UreG/UreF/UreH complex reveals how urease accessory proteins facilitate maturation of Helicobacter pylori urease.** *PLoS Biol.* 2013, **11**:e1001678.

90. Murakami KS: **X-ray crystal structure of Escherichia coli RNA polymerase σ70 holoenzyme.** *J. Biol. Chem.* 2013, **288**:9126–34.

91. Mathew R, Chatterji D: **The evolving story of the omega subunit of bacterial RNA polymerase.** *Trends Microbiol.* 2006, **14**:450–5.

92. Robinson PJJ, Bushnell D a, Trnka MJ, Burlingame AL, Kornberg RD: **Structure of the mediator head module bound to the carboxy-terminal domain of RNA polymerase II.** *Proc. Natl. Acad. Sci. U. S. A.* 2012, **109**:17931–5.

93. Liang S, Zhou Y, Grishin N, Standley DM: **Protein side chain modeling with orientation-dependent atomic force fields derived by series expansions.** *J. Comput. Chem.* 2011, **32**:1680–6.

94. Raha K, Jr KM: **Calculating binding free energy in protein–ligand interaction**. *Annu. Rep. Comput. Chem.* 2005, **1**.

95. Noskov SY, Lim C: **Free energy decomposition of protein-protein interactions.** *Biophys. J.* 2001, **81**:737–50.