Template Based Modeling and Structural Refinement of Protein-Protein Interactions:

by

Brandon Govindarajoo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2016

Doctoral Committee:

Professor Yang Zhang, Chair
Professor Philip C. Andrews
Professor Brian S. Athey
Professor Daniel M. Burns, Jr.
Assistant Professor Peter L. Freddolino
Assistant Professor Barry Grant

To my mother, family and friends.

ACKNOWLEDGEMENTS

First I would like to thank my advisor, Dr. Yang Zhang, for his support throughout this training process. Through the course of my doctoral training, Dr. Zhang has been an effective advisor; he makes it a high priority to train, interact and work with members in his lab. Through my time here I have seen a large improvement in my ability to analyze and resolve problems efficiently. Through time, consistency and Dr. Zhang's guidance I have developed the skill set to take seemingly complicated problems and reduce them into small problems with perceivable solutions. My experience in Dr. Zhang's lab has been very challenging while still being very engaging. My capabilities as a scientist have improved tremendously and continue to do so, and I greatly appreciate the time Dr. Zhang has given to train and advise me.

Next I would like to thank my committee members for their direction during my doctoral degree. I had a difficult time during the transition from undergraduate to graduate student; I really appreciate the availability and supervision from Dr. Dan Burns, Dr. Brian Athey, and Dr. Phil Andrews. The tremendous amounts of academic and social support has been paramount in the completion of this dissertation. I would also like to thank Dr. Barry Grant; he would regularly come to my lab presentations and provide feedback on the structure, presentation and scope of my projects, and finally Dr. Peter Freddolino for his willingness to participate on my committee given such short notice.

As a member of Dr. Zhang's lab, I have had the chance to interact with many graduate students and research scientist working on protein structure and function predictions. Their suggestions, insights and assistance has been a substantial asset in the development of the methodologies presented in this dissertation in particular the contributions of Dr. Aysam Guerler and Dr. Srayanta Mukherjee have truly been invaluable.

I would like to thank my family and friends for their continued guidance and emotional support during these six years at Michigan. I would like to especially thank my mother, Eleanor, for her years of love and encouragement. She has always been there for me providing support and guidance, and her contributions to my development are more than I can express.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. Introduction to Structural Bioinformatics

Overview

Proteins are a class of biomolecules composed of covalently bonded amino acids; these macromolecules carry out the majority of essential cellular function such as catalysis, cell signaling, immunity responses, cell structure, molecule transport, and signal transduction. Due to their importance determining protein function and how the function is performed is essential to understanding cellular physiology. In order to understand and characterize a proteins function, its three dimension structure is often a prerequisite as a protein is normally only functional when it is folded into its three dimensional structure. Protein structure is divided into four levels. The primary structure consists of the sequence of covalently linked amino acids forming the protein. The secondary structure are local structurally stable motifs in the protein that are created from a specific and periodically occurring hydrogen bond pattern amongst covalently bonded amino acids; the most common types are alpha helices and beta sheets. The tertiary structure consists of singular globular units in a protein chain that are formed from the hydrophobic collapse and the hydrogen bonding between amino acids. The quaternary structure consists of the interface generated by the permanent and transient interactions of the tertiary level folds. Most protein function occurs at the quaternary level, which makes determining the final level of protein structure of upmost importance.

Since the early 1950's it has been possible to experimentally determine the structure of a protein, yet the information regarding protein tertiary and quaternary structure has lagged behind the known protein sequences due to the time and cost associated with experimental determination of protein structures. Currently over 50 million protein sequences are deposited in the UniProt protein sequence database while slightly over 100 thousands structures have known three dimensional structure. The lack of structures has encouraged the development of algorithms that can predict the three dimensional structure of a protein given its sequence. For tertiary structure and function prediction the field of computational biology has seen promising results. Using information form previously resolved protein structures, protein modeling can generate prediction with similar

accuracies as experiments. But for the quaternary structure prediction, the field is still in its infancy. Predicting what proteins interact, the strength of the interactions and the orientation of the individual chains is an open problem.

In this thesis work, I developed methods to predict and model protein quaternary structure, in particular protein-protein interactions. The first algorithm SPRING (Chapter 2, Published in Journal of Chemical Information and Modeling) searches the protein database of known protein interactions to identify possible structural homologs to the query sequences. This algorithm along with the PDB was designed to efficiently predict and model protein interactions for whole genomes. However the number of known interfaces structures in the PDB is incomplete and does not represent all protein interfaces contained in nature. To increase the types of interfaces that can be modelled, we incorporate the interface between domains into the prediction of protein interactions (Chapter 3, Manuscript Completed). Finally a structure prediction pipeline was developed to create full atom quaternary structures (Chapter 4, Manuscript Completed). Using information from known structures such as pair wise residue distance and physical energy potentials incorporated into a folding simulation, medium to high resolution protein structures can be predicted starting from coordinates of identified homologs to the pair of query sequences.

1.1 The Protein Data Bank and Limited Protein Folds

1.1.1 The PDB: Protein Data Bank

The protein data bank was created in 1971 to house experimentally resolved protein structures [1, 2]. At its creation it contained several structures resolved by X-ray crystallography [2], but starting in 1978 the PDB started to see exponential growth [3], and now contains over one hundred thousand structures [4]. There are currently three experimental methods to determine macromolecular structure: X-Rays Crystallography, NMR and Electron Microscopy. X-ray crystallography can handle small to large proteins and association of multiple protein chains with high resolution. NMR (Nuclear Magnetic Resonance) can handle small to medium size proteins. It additionally has the benefit of the structures being solved in solutions closer to physiological conditions [5]. It also generates knowledge regarding different conformations and motions of a protein chain [5]. Finally Electron Microscopy can provide images of large associations of macromolecular complexes. Unfortunately it often generates low to medium resolution images [5].

There are currently 106,827 proteins contained in the PDB [4]. The majority of proteins are resolved by X-ray crystallography, followed by NMR and finally electron microscopy [4]. Regarding protein complexes, there are over 60,000 protein structures in the PDB. SCOPPI clustered the structures based on the interfaces identifying 15,058 unique interface structures [6, 7].

1.1.3 The Classification of Protein Structure and the Fold Recognition Problem

As the structures in the PDB grew the need to track the evolutionary and structural relationships between resolved proteins became apparent. The Structural Classification of Proteins database (SCOP) was developed for this purpose. SCOP is a manually curated hierarchical database that groups protein domains based on similarity; the SCOP database has four structure levels [8]:

I. Classes: This is the top level of the database it consists of four categories based on the proteins overall secondary structure content.
II. Fold: Structures are grouped together if the overall three dimensional topology of the proteins are similar.
III. Superfamily: Groups distantly related proteins
IV. Family: Proteins in this group have high similarity.

As the number of structures increased, several things became apparent: there was a need for automated methods to partition structures into domains and classify them i.e. CATH Database [9], there may be limits to the number of domain folds in nature, and most importantly the limited topologies could be used to predict protein structure [9-11]. This recognition moved the protein structure prediction problem into a fold recognition challenge, which further initiated the structural genomics project [12]. The structural genomics initiative was started to experimentally determine the structure of all domain folds, and since 2009 despite several thousand structures being identified no new domain topologies have been identified, suggesting all the possible folds are currently stored in the PDB [4]. This allows in theory for any protein domain to be modelled using the fold recognition paradigm; the main challenge for single domains prediction is matching it to the correct fold [13, 14].

It has been shown that unknown sequences can be mapped onto known domain folds to guide protein folding and structure prediction. Using known structures, protein models can be generated at very high resolution [10, 13, 15]. Regarding single domain proteins the major challenge is matching the sequence to the correct fold [13, 14]. Additionally it was then checked to see if fold recognition could be extended past single domain proteins.

Determining the limits of homology modeling and its scalability to modeling protein quaternary structure has huge implications on the direction the field of structural biology and structure prediction will take. It has been show that oftentimes similar proteins interact in similar ways. Furthermore structural alignment studies have consistently shown newly deposited structures are similar to preexisting interfaces, and that there is some structural overlap between domain-domain and protein-protein interfaces [7, 16]. It is predicted that there are a finite number of interfaces in nature and that the library will be approaching completion in the next two decades [17, 18]. This shows promise in modeling proteins at the quaternary level, and that over time the coverage of interfaces will continue to increase and soon in theory homology/template based modeling will have a database with enough depth to accurately model all proteins in nature.

1.2 Identifying Similarities amongst Proteins Sequences

Similar sequences often share similar structure and function. Comparing an unknown protein sequence to a database of known sequences with known structure and function can allow inference on the properties of the unknown protein. To compare two sequences two systems are needed: a scoring system for quantifying similarity and an algorithm to maximize the alignment score.

1.2.1 Sequence Similarity Matrix

Many scoring systems have been created to assess the similarity of protein sequences. The field of sequence alignments more or less began with the creation of the Point Accepted Mutation Matrix by Matrix (PAM) in 1978 [19]. The idea was to check the frequency of accepted point mutations in nature among close homologs to create a scoring system for measuring sequence similarity. Regarding PAM, phylogenic trees were constructed from closely related protein sequences. Between nodes in the tree and each position between sequences the occurrence of residue X being matched with residue Y was counted. Next the mutability was evaluated, which is the propensity

of a residue being mutated into another. Mutability is the frequency of not observing a change in residue between nodes in the phylogenic tree. This information was combined and converted into a logs odds matrix forming the 20x20 PAM substitution matrix. This matrix gives a score for determining if it is favorable to match one residue to another in an alignment.

Several attempts were made to improve upon the PAM score matrix, one of the most famous matrices the BLOSUM matrix was derived from similar principles as PAM [20]. The BLOSUM matrix is derived from a block of closely related sequences. The derivation of the scoring matrix considered pairwise point mutation from a larger more diverse set of sequences. For each block within each column all combinations of the residues were taken in pairs of two. This protocol tracked the propensity of one residue capabilities to mutate into another residue. These frequencies are then converted into a logs odd ratio for score sequence alignments.

1.2.2 Needleman Wunsch, Smith-Waterman and BLAST

The aim of global sequence alignment is to obtain the optimal alignment that maximizes the global similarity between two given sequences. Needleman and Wunsch developed a dynamic programing algorithm to identify such an alignment [21]. The algorithm creates an $m \times n$ matrix with m being the length of the first sequence and n being the length of the second chain. Each element is filled by considering the best possible choice at that position, such as an alignment or introducing a gap (insertion/deletion) in the first or second sequence. The maximum values is chosen to fill that position and the direction is recorded: diagonal for an alignment and up or left for the respective gap. Once the matrix has been completed the alignment is generated by the back track procedure where the maximum value of each element is used to generate the alignment [21]. Starting from the last element in the matrix, information is contained on whether an alignment or inserting a gap is the best option for that position. The choice determines which adjacent matrix position to go to next. This procedure is repeated until the algorithm reaches the first node.

Generally local alignment contain important structural and functional information. The Smith-Waterman alignment made a slight alteration to the Needleman-Wunsch alignment to create local matches. A floor was added to prevent values from going negative. This prevents local segments from being connected by long gaps or dissimilar alignment regions [22]. The NW algorithm was

initially designed for using a simple similarity score for deciding alignment values and a constant gap penalty value. However, insertions and deletions need to be properly accounted for when aligning two sequences. The affine gap penalty, developed by Gotoh, efficiently incorporated this procedure into the Needleman-Wunsch alignment algorithm [23]. These three papers form the searching foundations for most modern alignment algorithms.

Although the NW alignment algorithm guaranteed the optimal alignment it was often too slow to compare whole genomes, which led to the creation of faster alignment algorithm based on heuristics, FASTA and BLAST [24, 25] . The BLAST algorithm is a heavily cited algorithm and the foundation for one of the most popular molecular biology tool, PSI-BLAST [26]. The BLAST algorithm splits the sequence into overlapping words i.e. sequential groups of letters from the sequence ranging from 3-5 letters. A residue similarity matrix is then used to create words that are similar to the words created from the initial sequence. The words are then scanned against sequences in the database. If they do not hit the sequence it is then skipped, additionally when a word is hit the words start as initial alignments which are then extended without gaps. If the alignment scores above a threshold, the alignment is considered a match. And if it falls below the threshold its rejected which prevents full alignments of sequence pairs that are very unlikely to result in a strong alignment [25].

1.2.3 Multiple Sequence Alignments and Sequence Profiles

General sequence alignments with a simple *20 x 20* scoring matrix are often limited to finding confident alignments with high homology. Fortunately, iteratively running alignments by incorporating information from previous alignments into the search can significantly increase the depth of alignments by considering conservation of residue positions [27]. Starting from an initial search through the database similar sequences are identified and joined into a multiple sequence alignment, MSA, generally using the neighbor joining algorithm [28, 29]. The MSA is converted into an *L x 20* log-odds scoring matrix referred to as a position specific scoring matrix (PSSM) [27]. L is the length of the query sequence; the PSSM contains conservation information of each residue for each position among evolutionary related sequences. A second alignment uses the PSSM for scoring and creates a new multiple sequence alignments; this procedure is repeated until convergence [26, 30]. Often times the contribution of sequences are weighted based on sequence

similarity. A large number of higher similar sequences are down weighted in order to reduce bias in conservation due to redundancy; likewise low similarity sequences have higher impact on the PSSM [31].

From the success of sequence profile alignments, attempts were soon made to align two profiles. The idea is creating a database where each element in the database contains a PSSM. The query sequence is searched through the database and a query profile is created. Alignments are generated by aligning two profiles [30]. The most popular method for sequence-profile alignment is PSI-BLAST [26] which has been cited over 50,000 times. The algorithm starts with a normal iteration of BLAST, the first round generates a multiple sequence alignment which is used to create a PSSM. This PSSM is used to generate new words based on conservation within the PSSM. The blast algorithm is reran with the new set of information, and the scoring of alignments uses the PSSM information. This process is by default iterated three times [26], and can often identify homologs with at least 30% sequence identity to the query sequence.

1.3 Structure Alignment and Structure Similarity Scores

Structures with similar local and global topologies often share similar functions. Identifying structural features can help identify function of unknown proteins by comparison to known structures with similar topologies. Kabsch developed an algorithm for the optimal rotation given a pair of coordinates to align [32]. Most modern alignment algorithms uses this algorithm for the orientation, and focus on finding the correct pairs to provide to the Kabsch algorithm. The algorithm is optimized to minimize the root mean squared deviation between two sets of vectors. Although better scoring systems have been developed to compare structures, the RMSD and optimal rotation matrix are foundational tools for structural biology.

1.3.1 The Kabsch Algorithm and RMSD

Similar to protein sequences, protein structures need to be aligned in order to evaluate the similarity. Given a superposition, one way is to compare structures is to calculate the root mean squared error between a set of points.

$$RMSD = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

However any rotation of the y coordinates will substantially alter the error without altering the structure of y. So first the y structure needs to be correctly superimposed onto x

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - U * y_i)^2}$$

Kabsch derived the solution for finding the optimal solution of y onto x as to minimize the RMSD. First the equation is squared and mean squared deviation is used for simplicity and the equation is expanded.

$$MSD = \sum_{i=1}^{n}(x_i^2 + y_i^2) - 2 * \sum_{i=1}^{n}(x_i * U * y_i)$$

Notice that the first summation term is not dependent on U and can be removed from consideration. In order to minimize the RMSD we need to maximize.

$$2 * \sum_{i=0}^{n}(x_i * U * y_i)$$

Converting the equation into vector notation leads to the equation below with the goal of maximizing L.

$$L = Trace(X * U * Y^T)$$

Now using the cyclic property of the transpose the equation can be written as follows where R can be calculated and U is unknown.

$$\begin{aligned} L &= Trace(X * U * Y^T) \\ &= Trace(U * Y^T * X) \\ &= Trace(U * R) \end{aligned}$$

R is the correlation matrix and using singular value decomposition it can be rewritten as $R = V * S * W^T$ where V and W are orthogonal matricies and S is a diagonal matrix containing the singular values. Replacing R with the SVD and using the trace cyclic property

$$Trace(U * R) = Trace(U * V * S * W^T) = Trace(S * W^T * U * V)$$

The matrices W^T,U,V are replaced with the matrix T, which is the product of orthogonal matrices, which leads to

$$Trace(S * T) = S_{11} * T_{11} + S_{22} * T_{22} + S_{33} * T_{33}$$

Since T is orthogonal and $T_{ii} <= 1$, the trace is maximized when T equals the Identity Matrix.

$$T = W^T * U * V = I$$

The Kabsch algorithm optimal rotation matrices can either be in the right handed coordinate system or left handed system (determinant of U equals -1). In order to convert a left handed system to right handed the last column of U needs to be multiplied by negative one [33].

Despite being the most widely used metric to evaluate structure similarity, the RMSD measure has three major caveats. First, RMSD is protein length dependent: a pair of long proteins tend to have a larger RMSD than a pair of short proteins. Second, RMSD puts equal weights on all pairwise alignments which makes the score sensitive to local structure variation. Lastly, it is difficult to identify the cutoff for a "good" RMSD score.[34].

1.3.2 Structure Alignment and Similarity Scores

Structural scoring systems that can properly identify global structure similarity are important. Given a pair of aligned protein structure coordinates three scoring systems are still in consistent use. RMSD, the GDT-score [35], and the TM-score [34]. For comparing equivalency between a model and native structure the RMSD is the score of choice, with a value less than 2 Angstroms being considered accurate enough for drug discovery. However when local variation exists in the model or comparing distantly related structures the GDT and TM-score are often the scores of choice.

The TM-score was designed to consider the alignment coverage and pairwise distance proximities in order to calculate a similarity score. Additionally, scores like RMSD, MaxSub [36] and the GDT-score have dependencies based on the length of the alignment which the TM-score tries to circumvent. Here the TM-score, based on MaxSub score, is normalized to the length of the protein and the score is bounded between 0 and 1, which allows simple and easy comparison between large sets of alignments. A similarity threshold of 0.5 was determined confident based on an extreme value distribution model [34]. The d_i are pair wise distance within a cutoff between two

aligned residues and the d_0 is a factor which normalizes the score based on protein length [37]. L_{ali} is the number of aligned residue pairs and L_T is the protein length.

$$TMscore = \frac{1}{L_T}\sum_{i=1}^{L_{Ali}}\frac{1}{1+(\frac{d_i}{d_0})^2}$$

It is often necessary to compare proteins that have distant homology. Due to low sequence identity, algorithms such as Needelman Wunsch paired with the BLOSUM matrix produce poor alignments which in turn produce poor structure superposition. TM-align was developed shortly after the TM-score, with the goal of properly aligning structures with similar folds but highly divergent sequences [38]. The algorithm uses structural features, such as secondary structure, from the two proteins to generate seed alignments for initial structure alignments. The initial pairwise alignments are used to generate the first superposition. All pairwise residue distances after superposition are checked in order to generate the next set of pairwise alignments for superposition. Multiple iterations are performed in order to maximize the TM-score.

1.3.3 Interface Evaluation Scores and Quaternary Structure Alignment

Similar proteins form similar interactions. But often there are many alternative binding modes between a pair of protein folds that have been gained and lost through time [39]. It is not enough to identify and match towards the tertiary level components of the complex; the orientation of the two chains needs to be considered. Similar approaches towards aligning tertiary structure were used in order to best align two quaternary level structures and assess if the interfaces are similar.

Similar to tertiary structure, scoring systems have been developed to recognize orientation similarities between two quaternary structures or between a model and native structure. In the Critical Assessment of Prediction of Interaction (CAPRI) experiments three scores have been developed: FNAT, I-RMSD and Ligand RMSD [40]. FNAT refers to the percentage of shared interface residues between two proteins, I-RMSD is the RMSD of interface residues of the target structure compared to the same residues in the query structure, and finally Ligand RMSD is the measure of the RMSD of the second chain after optimal superposition. FNAT is a measure of local interface structure, IRMSD compares the overall interface structures, and Ligand RMSD evaluates the similarity of the two proteins orientations.

Regarding structural alignment three prominent algorithms have been developed based on the methodologies of TM-align and TM-score to evaluate quaternary structures. MM-align is a direct extension of TM-align with the goal of generating an optimal alignment based on the general tertiary structures and their relative orientations while preventing cross alignments [41]. The algorithms I-align and PCalign focus on the optimal alignment of the interface region. I-align generates a TM-score like score that weights alignments to the interface while PCalign incorporates chemical features into the TM-score [42, 43]. Additionally PCalign allows for non-sequential alignment that comes at an increase in computation time. For comparing a model structure to native, the TM-score or RMSD superposition is used followed by further evaluation by FNAT, I-RMSD and Ligand RMSD.

1.4 Energy Scores and Decoy Recognition

Accurate simulation of protein folding is dependent on the energy/scoring system used to compare to different states. Energy potentials and scores are needed that can drive molecular dynamics or Monte Carlo simulations towards a correctly folded protein. Initially energies were developed completely based on physical properties, unfortunately with limited success in predicting protein structure mainly due to computational expense. Later it was determined that reduced atom simulations with statistical properties of the PDB and homology restraints could be used to fold proteins in a simulation. Often these properties are combined into an energy function as a linear combination where weights are trained on a decoy set, which are a set of structures with various levels of perturbation relative to the native structure.

1.4.1 Physics based Energy Terms

The initial attempts to determine protein structure and model its dynamics consisted of created physical potentials and incorporating them into a molecular dynamics simulation. The Amber and Charm force fields are among the most commonly used [44, 45].

$$V = \sum_{bonds} k_b(b - b_o)^2 + \sum_{angles} k_\theta(\theta - \theta_o)^2 + \sum_{dihedrals} k_\varphi\big(1 + (n\varphi - \delta)\big)$$
$$+ \sum_{VW} \varepsilon\left(\left(\frac{Rmin_{ij}}{r_{ij}}\right)^{12} - \left(\frac{Rmin_{ij}}{r_{ij}}\right)^{12}\right) + \sum_{Elect} \frac{q_i q_j}{er_{ij}}$$

$$+ \sum_{improp} k_w(w - w_0)^2 + \sum_{UB} k_u(u - u_0)^2$$

Amber uses the first five and charm adds the second two. The first two terms refer to proper bond lengths and angles. The dihedrals refer to the phi and psi angles in the protein backbone. The next terms are the Vander Waals and electrostatic potentials. The last terms refer to bond bending. Proteins fold in water and properly tracking all the atoms in the simulation prevents long time scale simulations required for protein folding.

1.4.2 Statistical Energy Scores and Information Based Restraints

Information from the PDB can improve the modeling and prediction of protein structure. These knowledge based simulations incorporate PDB structural information in the form of statistical potentials, folding biases, rotamer libraries and homology derived restraints. The statistical potentials derived from the PDB are based on the theoretical foundations of the Boltzmann distribution. The Boltzmann distribution is an equation that describes the population of different states and the respective energy of that state. With lower the energies corresponding to the more populated states. Features such as pairwise residue distances are extracted from the PDB and distances are binned and the population frequencies are converted into energy potentials [46, 47]. Other statistical features include residue environment and contact propensities [47]. These statistics represent favorable positions for residues that help guide protein folding. Often bioinformatics based predictions of secondary structure, residue position and solvent accessibility are predicted and incorporated into energy functions [48, 49].

Rotamer libraries and folding biases are often included into folding simulation to make more protein like models. Rotamer libraries refer to backbone and sidechain structural elements extracted from experimentally resolved structures. The rotamers can be used to create starting structures for simulations, initial sidechain placement and creating higher resolution structural models [50-52]. Folding bias refer to incorporating intuitive awards and benefits for local and global topologies that are protein like. Often proteins are globular, so radius of gyration is often used as a reward for creating compact structures. Additionally strong rewards are often given for forming and keeping secondary structure elements as well as encouraging hydrogen bond contacts [52, 53].

Coevolution and homology restraints provide global topological information for the proteins shape. Mapping query sequences to homologous proteins provide starting coordinates and distance restraints for modeling proteins [54]. Oftentimes multiple structures are present and distance restraints can be generated by creating statistical energy functions based on the distribution of pairwise residue distances [54, 55]. Another approach involves creating contact restraints from observed coevolution events in multiple sequence alignments. The idea is through evolution if one residue mutates neighboring residues will also mutate to compensate for chemical and physical properties of the new change [56-60]. These contact matrices can be generated and used to guide protein folding simulations.

1.4.3 Interface Energy Terms

Generally the scores generated for folding single domain proteins should be able to recognize and model the correct interface between a pair of proteins. Pairwise residue statistical potentials are often incorporated for identifying interfaces [61-63]. More so two properties have been observed from the PDB that also helps guide selection of the correct interface. First is the geometric surface complementarity. This refers to the shape of the two binding sites fitting together compactly, i.e. two flat surface are often in contact or a bulge in one site is complemented by a grove in the opposite binding site [7, 64-66]. Similarly surface charge patches such as matching hydrophobic or oppositely charged patches are used to identify correct interfaces. Finally it has been observed that homodimers have interfaces that involve symmetry. When predicting homodimer interactions enforcing symmetry can reduce the search space which helps guide accurate prediction of protein complexes [67].

1.4.4 Decoy Sets and Weight Training

It is often necessary to combine statistical or experimental information to guide the simulation, the information/scores are often combined as a linear combination. The simplest approach involves simulating folding using a large set of parameters on a diverse set of proteins and choosing the set that yields the best results. However, the computational expense of doing this properly can prohibit this approach. Another approach revolves around creating sets of deformed model structures from

the native and training scores that rank the native structure high and highly deformed structures low.

Decoy sets are groups of models that have a large range of structural similarities to the correct topology. The goal of the training is developing scores and setting parameters that have high correlation to the structural quality of the models. Many sets have been generated by various structure prediction algorithms i.e. I-TASSER, MODELLER and Rosetta; often consisting of trajectories (coordinates of a structure at a particular time step) that are grouped into bins representing a uniform spread of model quality [68, 69] . Often times the sets are too simple and the native can be recognized by simple looking for irregular local structures and side chain interactions. Unfortunately this is often not enough to guide the simulation to the correct fold of the protein. This encouraged the development if 3DRobot which attempts to make structures with high structural similarity, but difficult to distinguish from the native structure without proper consideration of long distant contacts [69].

Quaternary decoy sets are generated with the goal of determining the correct orientations from sets of false interfaces. Two decoy sets are currently available for evaluating interface scores. The first one is the ZDOCK benchmark set [70]. It contains a set of protein complexes that also have the individual units crystalized in the unbound state. The unbound structures are docked by ZDOCK in order to generate a few thousand decoy interfaces. Potentials can be trained on their ability to distinguish orientations that are close to the native structure. However, when docking, model errors in the tertiary units can disguise the correct orientation from the similarity scores. An additional set of decoys was generated where tertiary models where perturbed and then dock [71]. This allowed potentials to be trained on more complicated sets.

1.5 Fold Recognition and Function Mapping

Most of the success in structure prediction has been using homologous structures from the PDB to guide structure predictions. With evidence showing that the single domain structures are completely covered by the PDB, substantial effort has been used to identify correct folds starting from sequence. Additionally with the realization that quaternary space is limited extensions have

been made for interface recognition. Once folds have been recognized this information can be used to predict the function of the protein based on similarity to proteins with known function.

1.5.1 Fold Recognition and Meta Threading Servers

Fold recognition generally refers to matching a query sequence to an experimental structure with the correct topology regardless of homology. Increasing the depth of sequence alignments began with the creation of profile-profile alignments. This incorporated conservational information of the query and resolved protein structures which improve the recognition of distantly related protein [72, 73]. Another approach is to use multiple sequence alignments to train Hidden Markov Models [74, 75]. Instead of static scoring systems, alignment, insertion and deletion probabilities are trained from individual profiles. Analogous to profile-profile alignment creating a database of HHM profiles and aligning them to query HHM's improved fold recognition. Both of these use sequence conservation for alignment [76]. Another idea was presented to create structure profiles and features to improve alignment accuracy [77]. In addition to sequence conservation, programs like MUSTER incorporate predictions of secondary structure, solvent accessibility, hydrophobicity, and residue depth in the alignment. Predictions are compared to known structures for fold recognition [78]. Due to the importance of fold recognition many algorithms have been developed to address this problem. However amongst the best algorithm none has emerged that consistently and substantially outperforms the other methods. Often times different fold recognition algorithms provide complementary bits of information regarding the query sequences fold. LOMETS, a meta-threading server, incorporates 15 state of the art threading programs for fold recognition. The threading scores are normalized; consensus and ranking are used to identify distantly related folds [14].

1.5.2 Quaternary Threading

A number of significant efforts have been made in recent years to develop bioinformatics based approaches to predict protein interactions [79-81]. Currently there are three classes of bioinformatics based approaches for identifying and modeling protein-protein interactions by sequence alignment: dimer threading, monomeric threading and oligomer mapping, and modeling the constituent chains followed by template interface docking [82].

Dimer threading directly aligns the query sequences to the target complex which allows for interface information to be considered during the alignment, example programs are: MULTIPROSPECTOR [83] by Skolnick's group, HOMBACOP by Kundrotas et. Al [84], the strategy used by Aloy et. al [85]., and COTH [86]. Monomeric threading and oligomer mapping starts with generating query alignments to the monomer library. Complexes are identified using a pre-generated lookup table where every protein chain constituent in an interaction is represented by a homologous structure in the monomer library. The recently developed programs SPRING and PrePPI are example of this protocol [87, 88]. The last approach uses monomer threading to identify homologs for each query sequence. The monomer representatives are compared to an interface library using structural alignment. The best orientation is determined by structural alignment scores and interface energy scores [89-91]. These different protocols are highlighted in the Figure 1.1 below extracted from a recent review on quaternary threading [82].

Each method has subtle differences that often can provide complementary information. The template docking approach uses structural alignments to improve threading depth. Structural alignments are generally more sensitive to similarity than sequence comparisons. Given highly accurate monomer structures, structural alignments are more likely to identify distant homologs. Monomer mapping is similar to the structural alignment approach accept each monomer is mapped to corresponding complex structures. This can reduces the number of structural alignments that are performed by two orders of magnitude. This approach allows for genome wide scale structure prediction, but errors in monomer threading may prevent the correct quaternary template appearing in the set of structures to preform structural alignment comparisons against. Finally, dimer threading benefits include: incorporating interface contact information into the alignment and in conditions where one of the query to template chain alignments are below threading detection levels, the homology of the second alignment can improve the recognition signal. The major drawback to the first two methods is if there is any error in ranking the correct template for the monomers the quaternary structure recognition part is guaranteed to fail. In all three methods, the top ranked monomers are used in the final model. It's incorporated by structural alignment to the complex framework. This can improve coverage and tertiary level similarity, but comes at the cost of generating clashes and nonnative pairwise interface contacts.

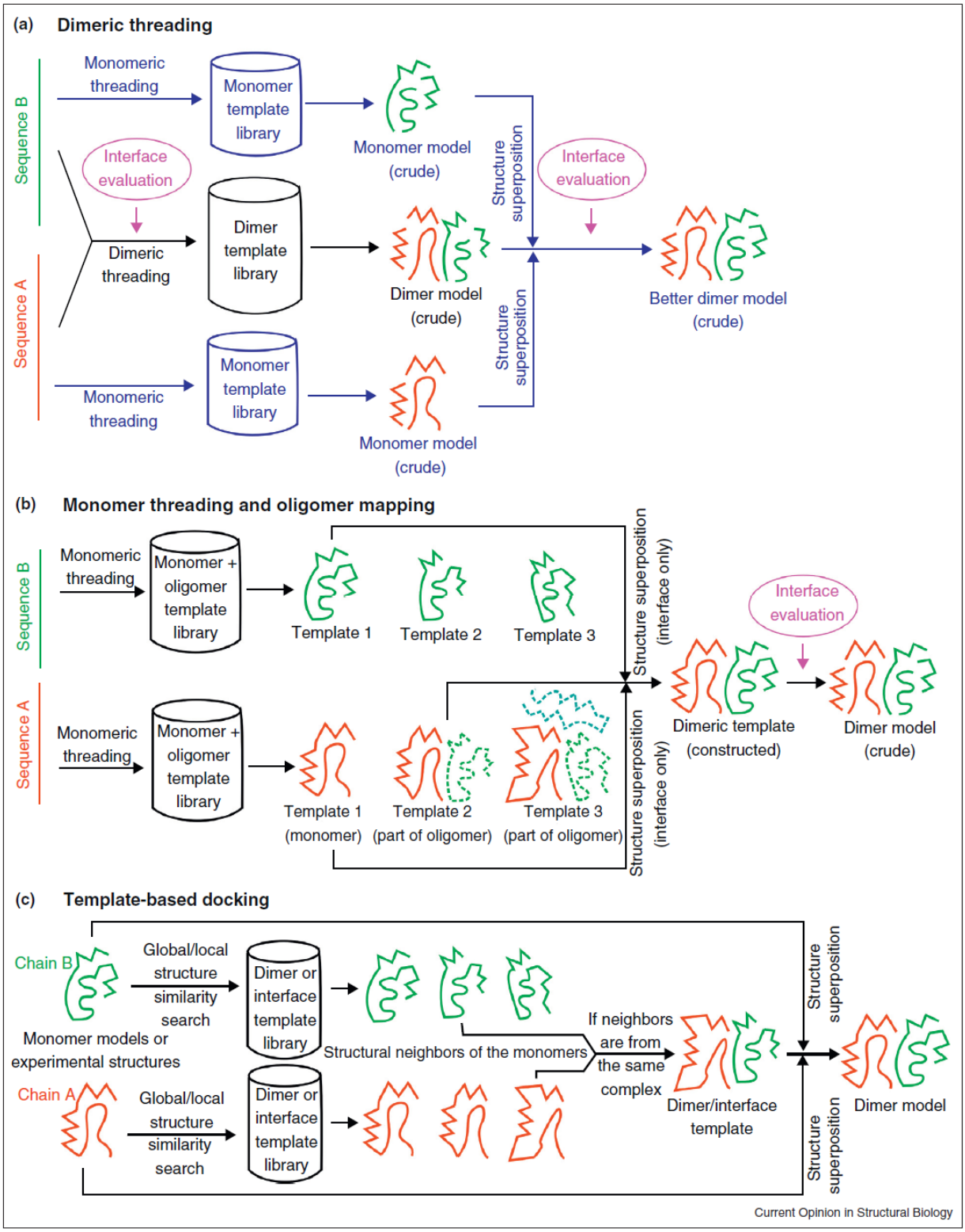


Figure 1.1 Overview of three general approaches for template based identification of protein dimers [82].

1.5.3 Function and Genome Wide Interaction Predictions

Similar protein sequences and structures often share similar functions. This paradigm allows for bioinformatics based approaches to function prediction. Databases have been created that map experimental information regarding biological function. Structures can be screened against a database to identify possible ligand binding sites as well as other biological processes [92, 93]. Similarly this approach can be expanded to the prediction of protein interaction networks. Large scale experimental methods to elucidate these networks are limited to yeast-two hybrid and affinity purification with estimated error rates of up to 90% [94]. Genome wide structure prediction using the current PDB can reduce the error rate of genome interaction predictions [87, 88]. Furthermore databases such as DIP [95], BIND [96] and INTACT [97] contain experimentally validated information regarding protein interactions which can be used to improve prediction accuracy [87, 98]. An additional source of information for prediction is using protein fusion events to predict interactions [94, 99, 100]. With the overall assumption that previously separated genes that function at the quaternary level provide an evolutionary benefit to be permanently connected. This connection may also preserve the original orientation between the two proteins, regardless it has been show it can improve prediction of protein interactions [99].

1.6 Tertiary Structure Prediction

The prediction of structure from sequence is a fundamental problem in structural biology. Initially physical based force fields were incorporated into molecular dynamics simulations to attempt to predict protein structure starting from random coil. There has been a few cases of success, but the most often cited issue is the long simulations times required to complete a folding simulation [101]. More recently a super computer named ATON was built that was specifically designed for protein simulations which significantly improved simulation time [101]. This procedure has shown great promise for folding small domain proteins and uncovering folding paths, but is restricted to small proteins that fold relatively fast. Another approach is folding simulations of a reduced atomic model where only backbone heavy atoms and side chain centers of mass are used to represent the structure. Algorithms such as QUARK and ROSETTA have shown promising results uncovering the overall topology of a protein domain. Often the results are too low resolution for further study however they can distinguish independent domains from sequence allowing for partitioning and

conversion to fold recognition problems [102, 103]. Regarding experimentally useful models, comparative modeling which uses highly homologous structures from the PDB to guide folding using programs such as MODELLER can provide high resolution structures that are useful for drug ligand screening, when the sequence similarity is above 50% [15]. More often though these structures are not present, but structures with the same fold are present that can guide the folding simulation. The starting coordinates usually need to have substantial changes to create a high resolution model [104, 105]. This is where composite techniques come into play where PDB structures along with ab initio potentials and simulations can generate high resolution models [104-108]. Here we present the composite algorithm I-TASSER as it's the foundation of the TACOS Protein Complex Modeling Pipeline.

1.6.1 I-TASSER: Composite Based Structure Prediction

I-TASSER is a composite algorithm that combines secondary prediction, homology detection and threading algorithms, along with physical and statistical scores to drive the initial template model structure closer to the native structure; the algorithm and pipeline were derived from [47, 109, 110]. The protocol for I-TASSER, in Figure 1.2, consists of four steps: threading and secondary structure prediction, replica exchange conformational sampling algorithm starting from the threading templates, clustering and averaging low energy replicas from the simulation, and finally full atom refinement. At the end there is a function annotation search using the modeled structure to predict function [104-106].

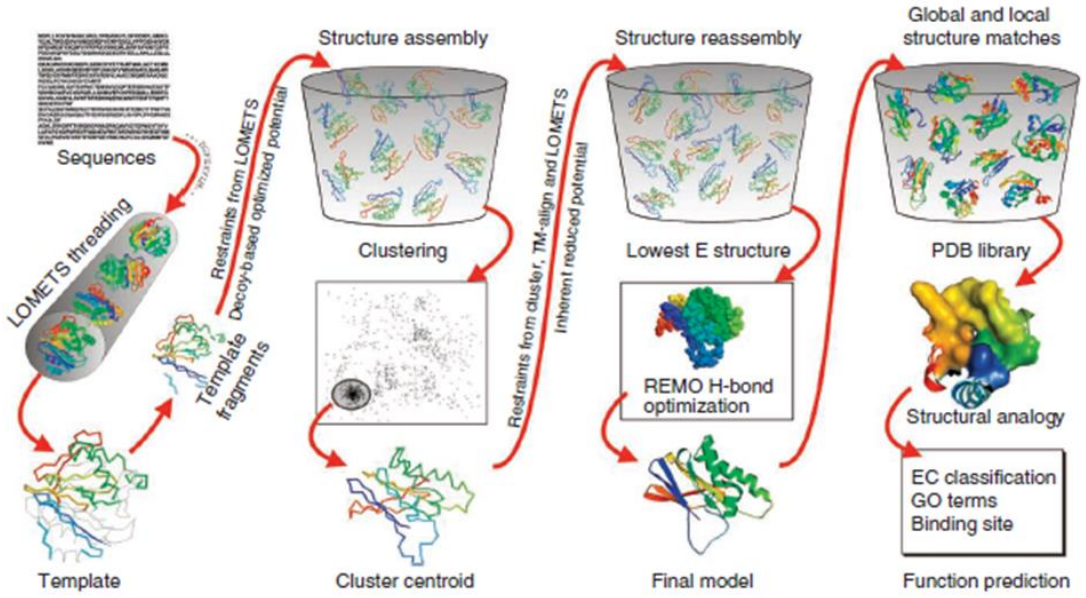


Figure 1.2 I-TASSER pipeline for protein structure prediction [104].

The pipeline starts with a query sequence and uses the meta threading (fold recognition) algorithm LOMETS [14] to identify structures in the PDB that are similar to the structure of the query sequence. Additionally multiple sequence alignments and profiles are created to predict the secondary structure of the query sequence with PSIPRED [111, 112], while also identifying contacts through coevolution as additional restraints for the folding algorithm [113-115]. The template, secondary structure and contact restraints are converted into statistical potentials to guide the replica exchange sampling simulation.

The initial template alignments provide starting positions and restraints for modeling, but need to undergo structural rearrangements to improve its similarity to the native structure. In order to improve protein structure modeling, two things are needed: an accurate scoring system and an efficient conformational search algorithm. The potential is created from a linear combination of physical potentials inherited from Touchstone2 and TASSER [47, 109, 110] and conservation and homology restraints identified by threading. The simulation reduces the nearest neighbor problem by representing each residue by only its Cα position and sidechain center of mass. Different conformations are evaluated by allowing local movements in the structure. If movements are too small the simulation takes too long to generate significant conformational changes, and if the movements are two large the moves are often rejected due to causing clashes or creating unreasonable conformations. I-TASSER circumvents this issue by transforming the simulation

from continuous to discrete space. The residue Cα positions are placed on a discrete grid, and a discrete set of local moves are generated that create protein like local structures that can efficiently sample large conformational changes. The hyperbolic replica exchange simulation [116] samples many conformations and outputs the trajectories at different time steps during the simulation. The next step involves selecting the trajectory that best matches the topology of the native structure.

Monte Carlo simulations sample an ensemble of conformations that follows the Boltzmann Distribution. The folding simulations should converge and pool into the overall correct topology. This convergence is identified by clustering. In the I-TASSER protocols clustering is performed with SPICKER [117]. The algorithm clusters up to 15,000 structures based on RMSD. The largest 5 clusters are further evaluated, for all trajectories within a cluster the structures are superimposed and averaged together in order to generate the consensus structure from the cluster. This average structure removes small perturbations which results in an overall structure that better represents the average. This structure often has an overall better topology, but the local structure is perturbed and often non protein like. A quick simulation is done to remove the irregular local structures. Furthermore the structure is still reduced and the rest of the atoms need to be added. The structures undergo full atom refinement where the full atom protein is built and the structure is refined using full atom energy functions with programs such as REMO and FG-MD [52, 118]. These programs often improve the structural quality and the hydrogen bonding network.

1.7 Quaternary Structure Prediction

Quaternary structure prediction is often referred to as a docking problem. Given two proteins known to interact, quaternary structure prediction tries to model the relative orientation and induced conformational changes of the two proteins. Ideally one would run a simulation that considered all these at once, but this is often beyond the capabilities of most computers. The problem is often broken into smaller steps: first a 6 dimensional rigid body search using the Fast Fourier Transform (FFT) is used, secondly reduced atom backbone models are used with small rigid body rotation and translation for conformations, and finally full atom simulations and with their corresponding potentials are ran. More recently, with the growth of the PDB models can be directly docked based on the orientation of homologous proteins.

1.7.1 Rigid and Soft Body FFT Based Docking

The FFT based approach developed by Katchalsi-Katzir is often the preferred choice as it efficiently searches the 6D (rotational and translational) space by the FFT's property of transforming a convolution problem into multiplication. Shape complementary and geometric fitting are important and often the primary features used in the low resolution search for recognizing correct interface [64, 66, 119]. Initially only shape complementary was the primary score for FFT. This property quickly checks all possible protein-protein orientations and ranks them based on the geometrical fit between the two surfaces. Many docking programs have been developed since then incorporating more detailed scoring systems MolFit [120], Hex [121], GRAMMX [122], FTdock [123], pydock [124], and ZDOCK considers shape complementarity, electrostatics and a pairwise atomic statistical potential [125]. Docking programs such as ZDOCK, improved on this method by incorporating a second more detailed scoring system on the structures generated post Fourier transform [126, 127]. Docking programs are generally successful when there is not much conformational change, less than one angstrom, due to binding [128, 129]. This deficiency prevents high accurate docking using models due to nonnative perturbations in the model that prevent the correct interface from being found [127]. To circumvent this small clashes "soft docking" are allowed during the docking [125, 128]. Moreover state of the art docking programs often occur in three stages: low resolution docking, re-ranking and refinement

.

1.7.2 Flexible Docking and Refinement

The second stage of docking often starts from highly ranked models from FFT docking methods. They allow quick refinement of sidechain positions and backbone conformations [130, 131]. Many of the faster refinements only allow small changes in order to make a physically reasonable model, but rarely improve on the global technology [130, 131]. In order to make substantial improvements on the interface and orientation, algorithms that include backbone and sidechain flexibility are needed [132-136]. Rossetta Dock incorporates rotations and translations in the chains to improve the general orientation [137]. Backbone and sidechain moves are also considered during the Monte Carlo simulation. During the course of the simulation thousands of trajectories are stored and the top 200 are output for further clustering. The cluster centers of the top 200 models are chosen as the final structures.

1.7.3 Structural Alignment to Complex Homolog Frameworks

This set of methods use templates to dock structures. Individual protein models can be built using general modeling prediction algorithms. Next, complex threading algorithms such as COTH or Multiprospector [83, 86] are used to identify homologous protein complexes. Two approaches can be used to dock the monomer models into the complex framework. First, general docking algorithms can accept two models and generate all feasible protein-protein orientations and the homologous pairwise restraint can be used to identify the correct orientation. Secondly, the models can be superposed using TM-align to the constituent proteins in the complex or the interface. More importantly the homolog similarity can provide confidences regarding their potential interaction and the orientation.

1.8 Limitations of Current Methods and Proposed Research

Most quaternary threading algorithm designs are computationally prohibitive for predicting genome wide interactions. Here I developed a computationally practical framework of using the PDB to predict and model protein interactions on the genome scale using the PDB. Additionally, I combined two state of the art dimer threading programs in order to improve template identification. Unlike tertiary structure, the interface library is far from complete which limits the types of interactions that can be predicted or modelled. Next I incorporated multidomain interfaces to bolster the types of quaternary structures that can be modeled. Finally, I use threading templates and domain knowledge from the PDB to accurately predict and model physically reasonable full length protein structures.

CHAPTER 2: Mapping Monomeric Threading to Protein-Protein Structure Prediction

2.1 Introduction

The number of possible protein-protein complexes scales in principle as the square of the number of monomer protein chains in genomes, with estimates of the possible number of distinct protein complexes in the order of millions [137]. Although the currently available high-throughput experimental methods have been employed to identify putative interaction protein pairs on proteome scales, the estimated error rates range from 41% to 90% [94]. These high-throughput methods do not provide structural information, i.e., where and how the proteins interact. Structural determination methods, such as X-ray and NMR techniques, could provide such information but are too costly and labor intensive to be applied on the proteome scale.

To address these issues, many computational approaches have been proposed for predicting the quaternary structures of proteins, which can be categorized as template-based and template-free approaches [138]. In the template-based approaches as applied to dimers [83, 86, 89, 90, 139-141], the quaternary model is constructed by matching a pair of monomer target sequences to a library of related template protein complexes which have the structure experimentally solved. In the template-free approaches [123, 125, 135, 142-146], also known as protein-protein docking, the target protein complex structure is predicted by scoring a large set of protein-protein orientations which are generated by assembling known monomer structure models.

Both methods have advantages and disadvantages. The template-free approaches can in principle treat any protein targets whose monomer structures are known. However, there is no guarantee of a high-quality structural prediction, particularly when bound structures undergo conformational changes from the unbound structures [128]. These usually involve side-chain readjustments and sometimes backbone rearrangements. Furthermore, the docking methods require the information

that the two proteins interact; this restriction is largely due to the limitations of the force fields used for evaluating the interaction energy [147].

Template-based (or homologous modeling) approaches generally have a higher accuracy than docking when homologous templates are available, but the alignment accuracy decreases sharply when the evolutionary relationship between target and template proteins becomes ambiguous, which generally corresponds to the scope of a target-template sequence identity <30%. Recently it was recognized that the structural space of protein-protein interfaces is highly degenerate [17, 18], which implies that the template-based approach can in principle be used to deal with any protein. In practice, the identification of the analogous protein complex pairs is highly challenging because the majority of the neighboring structure pairs have no obvious evolutionary relationship. Thus, development of new approaches to detect distantly homologous protein complex pairs is essential.

Partly toward this goal, we recently developed a method called COTH [86] which first threads both target sequences to a representative complex structure library. The monomer template structures identified by single-chain threading are then shifted to the dimeric framework that was identified by multiple-chain threading. The combination of the tertiary and quaternary libraries demonstrates a significant increase of the alignment coverage from the original complex structure templates, compared with other multiple-chain threading methods. However, the COTH procedure can be laborious since two template libraries (one for monomer and one for dimer) need to be maintained and updated. It is quite often that we found some interactions have been missed in the dimeric library even though we increased the sequence identity cutoff up to 90%. More importantly, the structural superposition can shift the complex structure to a wrong orientation especially when the structural similarity between the monomer structures in the two threading steps is low.

In this work, we address these issues by developing a new single-chain based threading and mapping methods for complex structure prediction, called SPRING (single-chain based prediction of interactions and geometries). Since in most cases one chain structure is taken directly from the original oligomer structure in the PDB, the alignment loss from the monomer-to-dimer

superimposition is kept minimal. Second, the close match of the interface areas from the same oligomers helps improve the coverage and accuracy of interface contact predictions which aims to solve a major issue in previous multiple-threading approaches [83, 84, 86]. Third, since a one-step single-chain threading is conducted, only the monomer structure library is needed in SPRING. It is therefore faster than COTH and other threading approaches, and the library is easier to maintain and update. Meanwhile, a precalculated lookup table is exploited to quickly exclude most of the complex frameworks that have no homologous association to the binding sequences. This is particularly important for speeding up the genome-scale modeling of protein-protein interactions, since only a small subset of interactions need to be pursued after this filtering step. Moreover, the complex template coverage is significantly maximized since there is no sequence cutoff for constructing the library. To examine the efficiency and generality, we will carefully test the method in control with other state-of the-art template-based methods in large-scale benchmarks. The SPRING algorithm is freely through the zhanglabs webserver at http://zhanglab.ccmb.med.umich.edu/spring/.

2.2 Materials and Methods

2.2.1 SPRING Algorithm

SPRING constructs the structure of protein complexes starting from two input chain sequences A and B (Figure 2.1). At first, a list of putative monomer templates (T_A) for sequence A is identified from the monomeric template library using a threading approach, e.g., MUSTER [78], HHsearch [76], or both. The threading provides a template alignment and a Z-score (Z_A) for the input sequence A. Here, the Z-score is defined as the difference between the raw alignment scores and the mean in the unit of variations, which has been widely used to assess the significance of the threading alignments, i.e., a higher Z-score means a higher significance and usually corresponds to a better quality of the alignment. The top template of the highest Z-score (T_{A1}) will be used to construct a monomer model for chain A.

Meanwhile, we thread the sequence of chain B through the monomeric structure library to identify another set of putative templates (T_B) with associated Z-scores (Z_B) (right column of Figure 2.1). Analogous to chain A, we derive a top-ranked monomer model for chain B using the template with

the highest Z-score (T_{B1}).

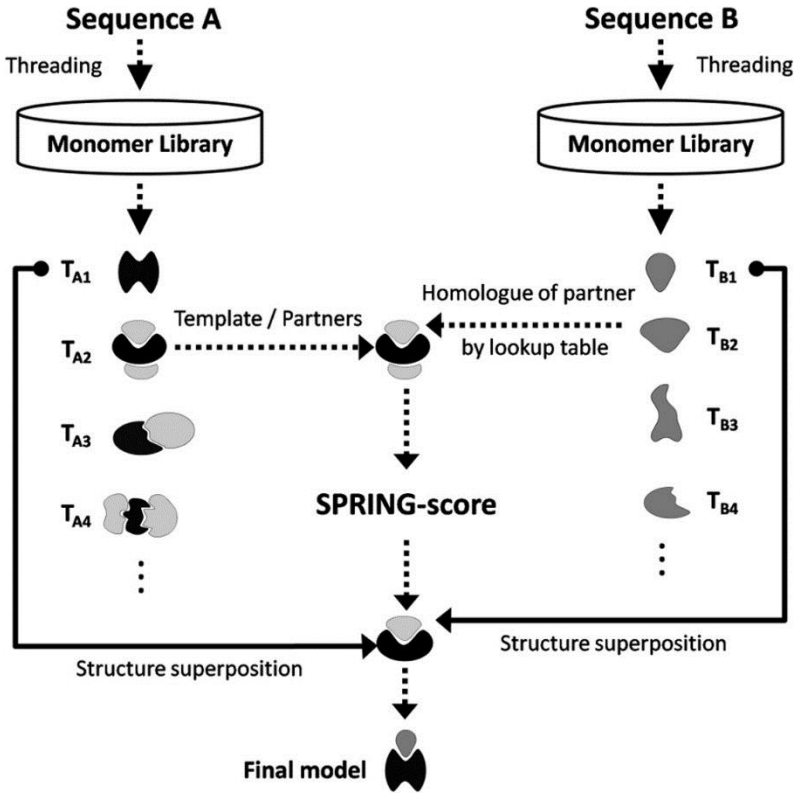


Figure 2.1 Flowchart of SPRING pipeline. Target sequences A and B are first threaded against the monomer template library, which yields two lists of template TA¬ (black) and TB (gray). For chain A, we retrieve all binding partners PA (light gray) from the origin.

To construct structure models of the complex, we now gather a set of template/partner frameworks by using the (top and lower-ranking) monomer templates of chain A (T_A). Therefore we retrieve the corresponding oligomer file of each monomer template T_A from the PDB. Then, all binding partners of the templates T_A are collected from the oligomers (labeled as P_A). These also include binding partners and their respective orientations as deposited by remark "350" of the PDB file. Template/partner frameworks can only be derived from monomer templates T_A with at least one binding partner.

Using the identified template/partner frameworks, we start by structurally aligning the top-ranked monomer model of chain A to all templates (T_A), where the alignment is built on the subset of interface residues. Additionally, we align the to-ranked monomer model of chain B to all binding partner structures of chain A that were retrieved from the PDB oligomers (P_A); the alignment is

based on the subset of interface residues. These two monomer-to-oligomer superimpositions yield a dimeric model, consisting of the reoriented top monomer models for chains A and B based on each of the oligomer frameworks. Here, we note that the tertiary structures of two components are both from the top threading template, although the oligomer frameworks can come from the lower rank threading templates. Based on our training results, using the top-rank monomer templates can generate on average better quality of complex models than using lower-rank monomer templates based on both local and interface scores. This is because the top monomer templates have generally a higher accuracy of alignments than lower-rank ones. Moreover, for reasonable frameworks the structures of component chains and the top monomer templates are often close, and the alignment loss from the superimpositions is minimal. Nevertheless, most of the top complex models by SPRING are built from the top oligomer frameworks. In these cases, the component models of the probe chain are taken directly from the oligomers, and no structural superimposition is needed. To improve the efficiency, we exclude template/partner frameworks if the corresponding binding partner is not homologous to any of the monomer templates (T_B) identified for chain B. The homology can be quickly verified through our precalculated look-up table, which is essentially a one-to-one PDB ID map to associate every binding partner in the oligomers to its closest homologues monomer structure from our tertiary template library (middle column of Figure 2.1). The look-up table was pregenerated by an all-against-all PSI-BLAST search [26] of the PDB library, where the partner/homologue association with the lowest E-value was selected for each binding partner. The look-up table is particularly useful to increase the efficiency for genome-wide all-against-all modeling studies, since only a small subset (~1%) of protein pairs that can find putative template/partner frameworks is needed for the consequent model construction.

The models constructed from monomer-to-oligomer mappings are evaluated by the SPRING-score which is a linear combination of three terms:

$$\text{SPRING-score} = \min(Z_A, Z_B) + w_1 * \text{TM} + w_2 * \text{contact}$$

where the first term is the smaller Z-score of threading of the two target sequences; the second is the TM-score returned by TM-align [38] when aligning the top-ranked monomer model for B to the subset of interface residues of the selected binding partners of chain A (P_A); the third counts

for a distance-specific interface contact potential, which was derived from 3897 non redundant dimer protein structures with a sequence identity <30% to each other [18]. It uses a formula similar as Zhou et al. but with the atomic distances taken from residues in separate chains [46], and w1 =12.4 and w2 = -0.2 are the weights factors balancing the terms. We determine the weighting parameters through a grid search on a separate training set of 200 randomly selected protein complexes by maximizing the number of 'acceptable' models, where an acceptable model refers to the top-ranked models >30% of correctly predicted Cα-atom contacts in the interface.

For heterodimer proteins, this process is repeated using B as probe to identify binding partners for the complex template identification and model construction. The models of the highest SPRING-score in the two processes are finally selected as predicted models. For homodimer proteins, a single threading starting on one chain is sufficient due to the symmetry of the complex structures.

2.2.2 Libraries of Protein Structure Templates.

SPRING is based on monomer threading, and we constructed from the PDB a representative set of 43,571 monomeric protein structures, sharing a pairwise sequence identity of <70%. Obsolete structures and theoretical models were removed. For multiple-domain proteins, both individual domains and whole proteins are included in the library, which has been proven to increase the alignment accuracy of single domain proteins [78].

For benchmarking SPRING with other methods, we also constructed a set of non redundant dimeric complex structures that is needed by COTH and the naïve complex threader using MUSTER, HHsearch and PSI-BLAST. This library was derived from DOCKGROUND [148] with a filter of pairwise sequence identity <70%. In addition, irregular structures, transmembrane complexes, and complexes with alternate binding modes were removed. To rule out crystallization artifacts, complexes with <30 interface residues or with a buried surface area <250 $Å^2$ were not included. It finally contains 7404 dimeric protein structure templates at the same date cutoff of the monomer library.

2.2.3 Test Set of Protein-Protein Complexes.

The evaluation of prediction performance was conducted using a set of 1838 non homologous protein-protein complexes from the PDB, including dimers derived from higher order oligomers,

similar to that used by Lu et al [83]. Each of the 3676 monomer structures from the dimers contain at least 40 interface residues with at least 30 interface residue-residue contacts, where a contact is defined as a pair of residues from different chains with at least one pair of side-chain heavy atoms within 4.5 Å. In addition, the dimers have a sequence identity <35% to each other (i.e., at most one chain in a dimer can have >35% sequence identity to any of the chains in another dimer so that homodimers are included).

2.2.4 Measures of Dimer Model Quality.

The global model qualities are evaluated using TM-score [34], the global complex RMSD, and the sequence-template alignment coverage. Local model qualities are measured using the fraction of native Cα-atom contacts (fnat) in the interface, the interface RMSD (I-RMSD), and the interface alignment coverage, where interface residues are defined as those with a heavy atom distance of <10 Å to any residue of the other atom.

TM-score has been extensively used to assess the quality of monomeric protein structure predictions, because of its attribute to balance alignment accuracy and coverage. In order to calculate TM-score of dimeric models, we convert the dimer into an artificial monomer by connecting the C-terminal of the first chain with the N-terminal of the second and then run TM-score program using the length of the query complex as normalization scale. This definition of complex TM-score is sensitive to the topology of individual chains and their relative orientation. A high complex TM-score indicated the correct modeling of both individual chain structures and their relative orientation [149].

2.3 Results

2.3.1 Control of SPRING with Competing Threading Methods.

SPRING derives complex structure by mapping monomer alignments as identified by single-chain threading algorithms, e.g., MUSTER [78] and HHsearch [76] (Figure 2.1). To examine the gain of the threading and mapping procedure over the traditional dimeric or monomeric threading procedures, we implement SPRING using the monomer alignments from MUSTER (called SPRING-M), in control with COTH (threading and superposition) and a naïve implementation of MUSTER (called NAIVE-M). In NAIVE-M, MUSTER is used to align every chain of the target

complex with that of known proteins in the complex template library. A template model is obtained if both chains from a template are aligned with the target. This procedure is identical to the strategy that was used by several authors in the former studies [83-85].

Figure 2.2 shows a comparison of the three methods on a set of 1838 interacting protein-protein sequence pairs, based on the global TM-score, the fraction of correctly predicted interface contacts (fnat), the interface RMSD, and the global RMSD, respectively. To rule out contamination from close homologous templates which are easy to identify by sequence comparisons, any templates which have a sequence identity >30% to target proteins in the testing set have been excluded from the template libraries. This filter is implemented in all the following threading calculations unless noted specifically.

Overall, the number of successful predictions by SPRING-M is the highest among all methods in each of the TM-score ranges. The same is true for the fraction of interface Cα-atom contacts and the interface and global RMSD results. For instance, if we consider a TM-score threshold of >0.5 SPRING-M, COTH, and NAIVE-M generated valid dimeric models for 1029 (56%), 767 (42%) and 568 (31%) of the 1838 protein targets, respectively. Similarly, if we count for the number of cases which have an I-RMSD < 5 Å and with at least 50% interface residues aligned, the number for SPRING-M, COTH, and NAIVE-M is 638 (35%), 381 (21%), and 359 (20%), respectively.

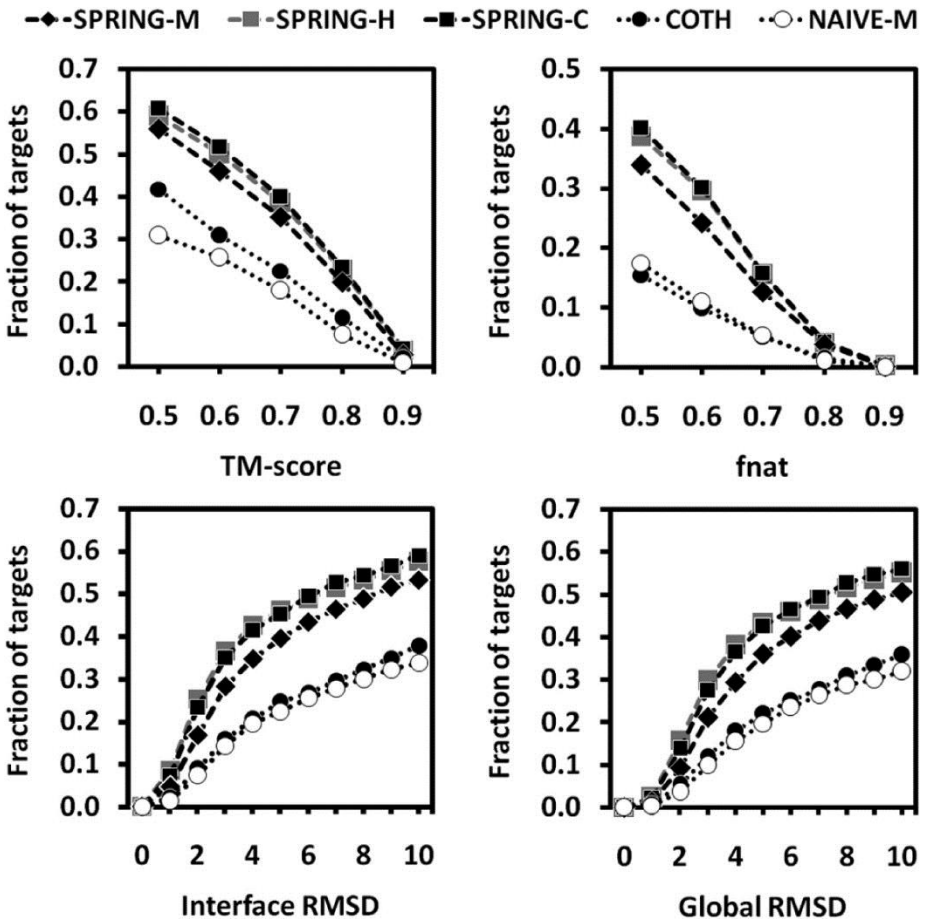


Figure 2.2 Cumulative fraction of TM-score, native contacts, and interface and global RMSD at different threshold cutoffs, for models on 1838 proteins predicted by SPRING-M, SPRING-H, SPRING-C, COTH, and NAÏVE-M, respectively. The shown data are from the best out of five top-ranked models for each protein target.

To further examine the detailed difference between the algorithms, in Figure 2.3A, B we present a head-to-head comparison of dimeric models predicted by SPRING-M and COTH, with regard to the TM-score and contact accuracy (fnat). There are 1023 cases where SPRING-M generates models of a higher TM-score than COTH, where COTH does so in 539 cases. Overall, the average TM-score of the predicted SPRING-M models is 13% higher that than of COTH. For interface structure modeling, SPRING-M models preserve 31% of the native contacts, whereas in COTH it does so in only 17% of cases (see Figure 2.3B). Since both methods used the top-ranked monomeric models to form the dimer models, their global alignment coverage is close (~88%). Thus, this TM-score increase is purely due to the identification of better dimer templates from the SPRING-M threading mapping, which results in more precise chain orientations. This is further manifested by the modeling quality at the interface structures. If we defined a high-quality hit as that with an I-RMSD < 2.5 Å on >90% of interface residues aligned, SPRING-M produced 162

hits compared to 89 by COTH, which corresponds to an increase of 82%.

In Figure 2.3C, D we present a similar head-to-head comparison of SPRING-M with NAIVE-M, where the TM-scores of the dimeric models predicted by SPRING-M are on average 40% higher than that of NAIVE-M. The major reason for the TM-score increase in SPRING-M is due to the boost of template libraries because SPRING-M has the monomer structures built from the tertiary template library (43,571 entries) which is much larger than the quaternary template library (7404 entries), while the latter was the only source used for NAIVE-M for building the complex models. For interface structure, the NAIVE-M alignments conserve 17% of native contacts (see Figure 2.3D), which is similar to that of COTH but 45% lower than that of SPRING-M. Here, although the individual COTH models are on average of higher TM-score, they do not contain more correct interface contacts than NAIVE-M. The poor performance of COTH relative to SPRING-M is mainly due to the alignment strategy that COTH employs to combine the monomers of the identified dimeric template. Since COTH uses a full-length global superposition strategy, it focuses less on the interface conservation, rather than the global topology of the complexes. In contrast, SPRING-M maps the monomer alignment using a subset of interface residues, which guarantee a better match in the interface regions. Meanwhile, many of the top alignments have the partner chain directly coming from the original oligomer entry which helps enhance the shape match of the interface. Third, the quaternary chain orientation of the complexes as identified by SPRING-M mapping has a better quality than that by monomeric or dimeric threading, which further contributes to the interface contacts.

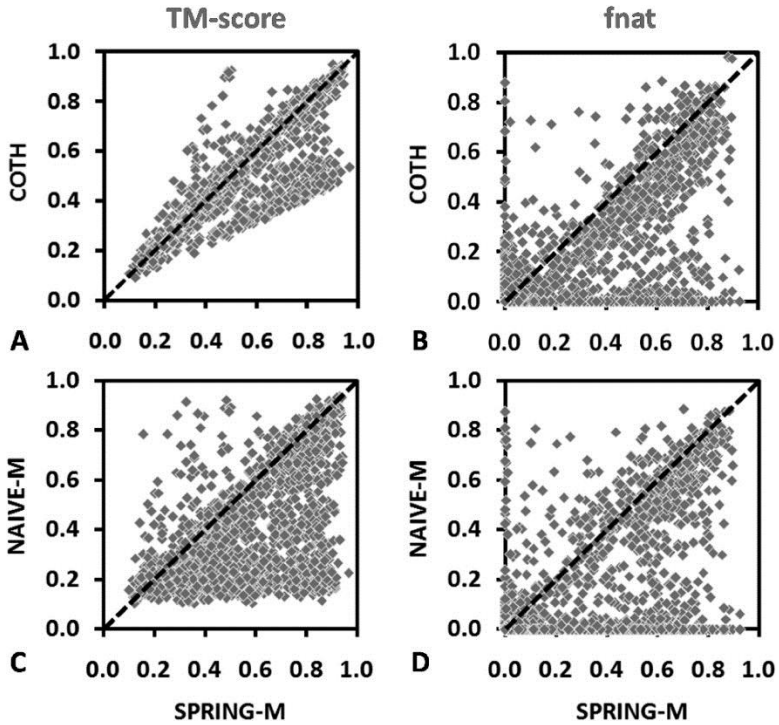


Figure 2.3 Head-to-head comparison of 1838 SPRING-M models with that by the control methods. The left column shows TM-score of the best in top-five complex models, and the right column is the fraction of the correctly predicted interface contacts. (A, B) SPRING-M vs NAÏVE-M.

The observed performance differences of SPRING-M from COTH and NAIVE-M with regard to the TM-scores are statistically significant, which have p-values from the Wilcoxon signed-rank test of 10^{-42} and 10^{-161}, respectively. In Table 2.1, we summarize the overall model qualities for each method, according to the average TM-score, the fraction of native Cα-atom contacts, the number of hits with an I-RMS <2.5 Å, and the global alignment coverage, respectively. The results are shown from both the first model and the best in top five models, where SPRING-M clearly outperforms the control methods in all the criterions.

Methods	TM-score[a]	fnat[b]	hits[c]	Coverage[d]
NAÏVE-P[e]	0.25/0.26	0.10/0.11	42/47	45/47%
NAÏVE-H[f]	0.35/0.37	0.15/0.17	80/93	56/58%
NAÏVE-M[g]	0.38/0.40	0.15/0.17	60/67	87/88%
COTH	0.48/0.50	0.15/0.17	70/89	88/88%
SPRING-M	0.54/0.56	0.26/0.31	133/162	88/88%
SPRING-H	0.55/0.57	0.29/0.33	211/246	81/81%
SPRING-C	0.56/0.58	0.29/0.34	187/219	86/83%

Table 2.1 [a]Average TM-score of predicted complex models. [b]Average fraction of conserved interface native contacts. [c]Number of targets with model of I-RMSD <2.5 Å and >90% interface covered. [d]Average fraction of aligned complex residues. [e]NAÏVE implementation of PSI-BLAST. [f]NAÏVE implementation of HHsearch. [g]NAÏVE implementation of MUSTER.

2.3.2 Illustrative Examples of Dimeric Threading.

To further analyze the strength and weakness of SPRING-M in comparison with the other methods, we dissect in detail several typical examples. Figure 2.4 presents the model predictions for the 1-Cys peroxiredoxin complex (PDB ID: 1XCC), which is a typical homodimer complex. First, NAIVE-M identified a template from the glutathione peroxidase-5 (PDB ID: 2P5Q) with a sequence identity of 11% to the target. The predicted model has a TM-score = 0.37 and an I-RMSD = 12 Å, covering 64% of interface residues. The model predicted by COTH uses the same complex (PDB ID: 2P5Q) as the global template. However, COTH derives both monomer models from the peroxiredoxin-4 protein (PDB ID: 2PN8). The combination of the monomer templates on the dimer framework increases the TM-score from 0.37 to 0.52, which has an I-RMSD of 7.1 A to the native crystal structure complex. In total, it has 376 residues aligned, which are much higher than that in the NAIVE-M alignment.

Finally, SPRING-M derives the orientation based on the single-chain MUSTER threading, which retrieves the dimer template from tryparedoxin (PDB ID: 1UUL). The individual monomers of this dimer template are structurally similar (TM-score = 0.67) to the monomers of the template (PDB ID: 2P5Q) as identified by COTH and NAIVE-M, but the chain orientation in 1UUL is much closer to the native than that in 2P5Q. The closer orientation of the framework allows SPRING-M to generate a dimer model of higher quality after the mapping of monomer structures (see Figure 2.4).

Figure 2.4 Predicted dimer models (dark color) of NAÏVE-M, COTH, and SPRING-M for target protein superposed with native structure of the 1-Cys peroxiredoxin complex (light color, PDB ID: 1XCC). The values below each superposition are TM-score, I-RMSD, and fraction of aligned interface residues.

An interesting question is why 1UUL was only successfully identified by SPRING-M but no other methods since both COTH and NAIVE-M use MUSTER for monomer threading. A closer analysis showed that this template is not included in the non redundant dimer structural library since the complex structure contains a single decamer and thereby multiple alternate binding modes for the homologue chain pairs. Since the other two algorithms did not select particular binding modes from a set of alternatives, none of the putative orientations could be detected. As an essential advantage, however, SPRING-M considers all alternative binding modes from all complexes of the oligomer structure, since it starts from monomer threading with the composite SPRING-score selecting the most suitable pair. In this example, although all monomer templates have a low sequence identity to the target (<30%), the SPRING-score is high (27.5), which gives us a high confidence on the prediction. The overall TM-score of the mapped dimer model is 0.75 with an I-RMSD of 3.0 Å. Again, the fraction of aligned interface residues of both SPRING-M and COTH models is the same (= 83%), where the improvement of SPRING-M is on the choice of the better template framework and the closer mapping of monomer structures in the individual domains.

Figure 2.5 presents another example from the putative kinase complex (PDB ID: 2AN1, chains A and D). In this example, the best template (PDB ID: 1YT5) is included in the dimeric structural library. But the oligomer complex includes 8 biomolecules based on 4 homologues chains; these correspond to 48 dimeric alternative binding modes. The COTH library can chose only one binding mode from the dimeric pair of chains A and D that has the lowest solvent free energy (-183 kcal/mol) as defined in the PDB; this template results in an incorrect orientation for this target (fnat = 0.09 and I-RMD = 14.2 Å), although the individual monomer models are similar to native (TM-scores > 0.77).

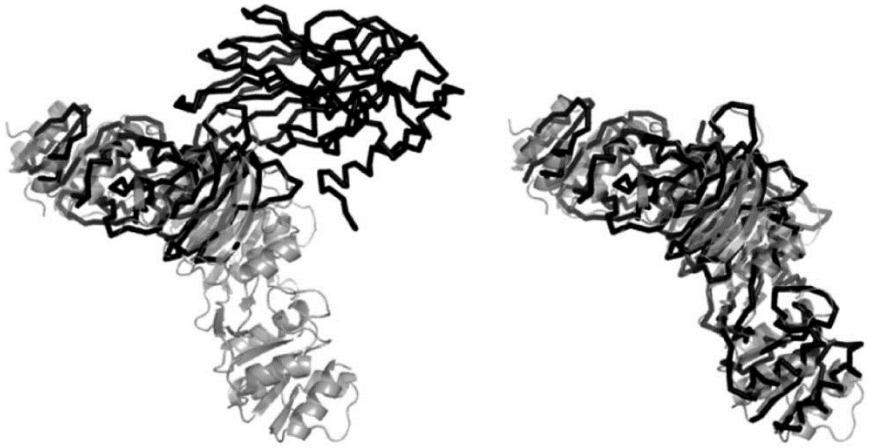


Figure 2.5 Complex models (dark color) are superposed with the native crystal structure of putative kinase complex (light color, PDB ID: 2ANI). (A) COTH and (B) SPRING-M.

In contrast, since SPRING-M retrieves partners from original oligomer complex structure, it naturally considers all 48 putative binding modes in the look-up table. Despite the slightly higher solvent free energy (-139 kcal/mol), the complex of biomolecules 3 and 5 with chains A and B was selected by SPRING-M as the most suitable framework, since the TM-score from TM-align superposition (0.82) and the contact potential (-41) are both better than all other partnerships (TM-score and contact potential values for the A/D pair template are 0.44 and -22, respectively). The choice of this A/B template results in a complex model with much better quality (fnat = 0.70 and I-RMSD = 2.6 Å) than that by COTH (see Figure 2.5). Meanwhile, since only one chain was required for other proteins (instead of both chains in COTH) to be superimposed on the framework, the interface shape match is another contribution to the quality of the interface structures of the SPRING-M models in this example.

2.3.3 Performance of SPRING Using Different Monomeric Threading Algorithms.

In the previous sections, we compared SPRING-M, COTH, and NAIVE-M with all three algorithms based on MUSTER to ensure a fair comparison of different template identification and complex constructing strategies. However, neither SPRING mapping nor the SPRING-score is restricted to specific monomer threading algorithms. An interesting question is whether and how the SPRING pipeline benefits from choosing different target-template alignment algorithms. Here we test the performance of SPRING using another threading program, HHsearch [76] (SPRING-H). While MUSTER generates the target-template alignment based on a composite sequence and

structural profiles, HHsearch uses the hidden Markov models. They can have significantly different results on template selection and target-template alignment for specific cases, although the overall performance in the tertiary template identification was shown comparable in previous benchmark tests [14].

Based on the data of the 1838 protein complexes, we found that SPRING-H identifies on average better quality quaternary templates than that by SPRING-M. For instance, if we consider a TM-score threshold > 0.5, SPRING-H and SPRING-M generated valid dimeric models for 1082 (59%) and 1029 (56%) protein targets, respectively, after excluding homologous templates (see Figure 2.2). Similar conclusion is obtained, regarding the average TM-score, contact accuracy, interface, and global RMSDs. In particular, if we count the number of correct models with an I_RMSD <2.5 Å and >90% interface coverage, SPRING-H has about 1.5 times more hits than SPRING-M (see Table 2.1).

This difference is quite striking since MUSTER and HHsearch alignments have about the same TM-score on the tertiary template recognitions (i.e., <TM-score> = 0.57 for both alignments in our test). A detail analysis showed that the alignment coverage of the MUSTER alignments is ~8% higher than that of HHsearch. These extra residues of alignments have number of acceptable models (TM-score >0.5) increases from 1082 (59%) in SPRING-H to 1115 (61%) in SPRING-C. In Figure 2.6C, D we also present a head-to-head comparison of SPRING-M with SPRING-C, where the TM-scores of dimeric models predicted by SPRING-C are on average 4% higher than that of SPRING-M. Considering the contact accuracy, SPRING-M preserves 31% of native contacts (see Figure 2.6D) compared to 34% by SPRING-C. The overall results of the comparisons are summarized in Table 2.1. These data demonstrate that a complementary alignment from different threading algorithms can further improve the yields of SPRING.

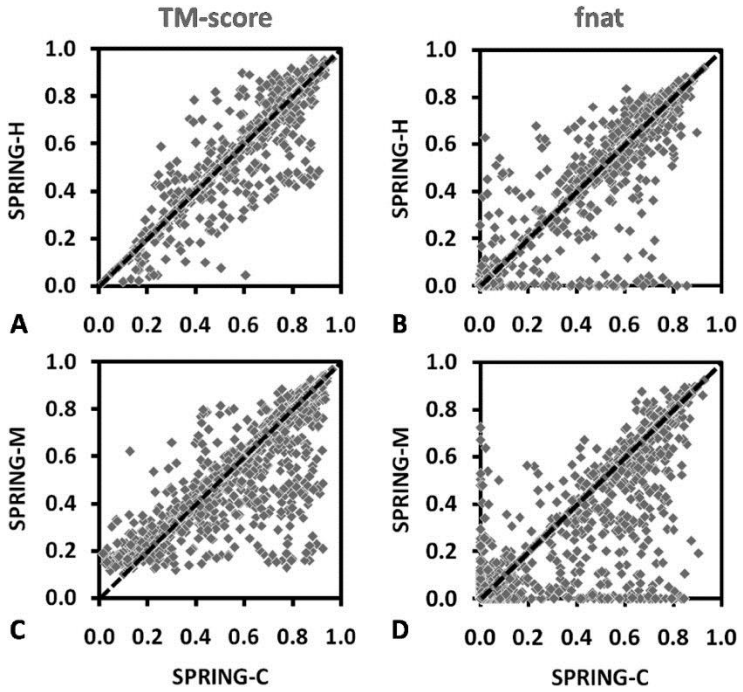


Figure 2.6 Head-to-head comparison of the SPRING models using different monomeric threading methods of 1737 test proteins. The left column shows TM-score of the best in top-five complex models, and the right column is the fraction of the correctly predicted interface contacts. (A,B) SPRING-C vs SPRING-H and (C,D) SPRING-C vs SPRING-M.

2.3.4 Specificity of SPRING Predictions.

The confidence of the SPRING method is assessed by the SPRING-score, which is a combination of threading Z-score, structural mapping TM-score, and the interface contact potential (see eq 1). In this section, we examine whether the SPRING-score is able to distinguish correct from incorrect SPRING predictions, which is important in practical applications since confidence scores of predictions essentially decide how the models should be used by biologist users. We use SPRING-C for the illustration.

Figure 2.7 presents TM-score, fnat, and interface and global RMSDs of the predicted models to the native complexes in different SPRING-score interval. Considering the first models for the 1838 cases, SPRING-C made 987 (54%) predictions with a SPRING-score > 13. In 774 (78%), 579 (59%), 601 (61%), and 519 (53%) cases, the predicted models have a TM-score >0.5, a fnat >0.5, and interface and global RMSDs < 5A (see the dark regions in Figure 2.7). Apparently, when SPRING-score is higher, there is a higher fraction of protein targets that have models with a better quality and vice versa. For example, when considering TM-score >0.5 as a threshold, the fraction of successful modeling is 60, 80, and 88% for the targets in the SPRING-score interval of [13, 15],

[15, 20], and >20, respectively. If we use a threshold of SPRING-score >13 to predict the correct template alignments, the false-positive and false-negative rates for TM-score >0.5 are 0.22 and 0.27, respectively. A similar tendency was also seen when using other criterions (see Figure 2.7.).

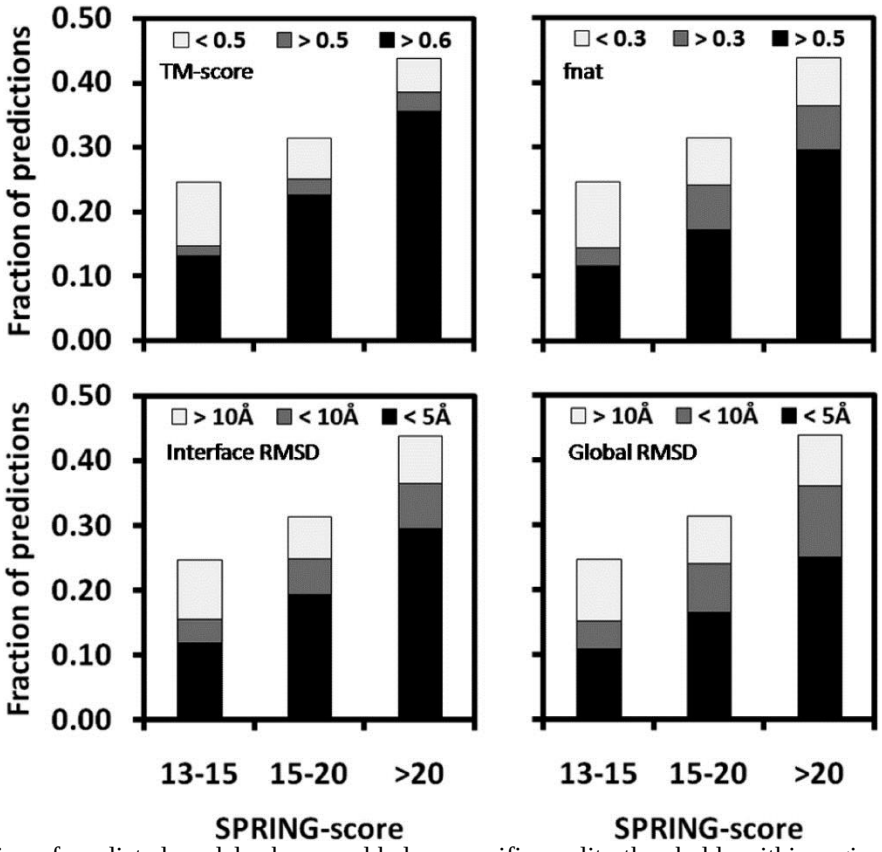


Figure 2.7 Fraction of predicted models above and below specific quality thresholds within a given SPRING-score interval for the top-ranked models. The depicted quality measures are TM-score, fraction of native interface contacts, interface and global RMSD. Models with <50% of aligned residues are included in the RMSD category > 10 Å.

Nevertheless, there is a considerable fraction of proteins which have low specificity, i.e., the proteins that have a high-scoring prediction but with poor model qualities when compared to the native or vice versa. For instance, we identified overall 53 structures which have a SPRING-score >20 but with a TM-score <0.5. In the majority of these cases, we found that SPRING ranks the templates of alternative binding modes as the highest score templates. Incorporation of specific binding affinity energy terms, such as the binding predictions by BSpred [86], can be a possible solution to further improve the specificity of SPRING.

2.3.5 Comparison of SPRING with Other Conventional Threading Strategies.

The majority of above SPRING benchmark data are controlled with other internal algorithms of

COTH [86] and MUSTER [78]. To have a general control with other external threading algorithms, we implement two additional procedures of the naïve extension of PSI-BLAST (NAIVE-P) and HHsearch (NAIVE-H) for complex modeling. Following the traditional homology-based multimeric threading strategy [83-85], these procedures first match the monomer chains through the dimer template library by PSI-BLAST or HHsearch. If the two target chains hit the monomers from the same complex template, the aligned regions isolated form the template constitute the complex models.

As shown in Table 2.1, SPRING significantly outperforms NAIVE-P and NAIVE-H, in terms of global and local quality of the models. For example, the TM-score and the number of native contacts in the first model of SPRING-C is 124 and 190% higher than that of NAIVE-P and 60 and 93% higher than that of NAIVE-H. Among the naïve extensions of the monomer threading algorithms, NAIVE-M has a slightly higher TM-score than NAIVE-H due to the higher alignment coverage but with a lower number of hits considering the I-RMSD cutoffs. Both algorithms have a significantly better model quality than NAIVE-P, which stems from the improved sensitivity of profile-profile alignments by MUSTER and HHsearch on monomer threading over the sequence-profile alignment by PSI-BLAST.

2.3.6 Control of SPRING with Rigid-Body Docking Algorithms.

To have a control of SPRING with the rigid-body docking methods [123, 125, 135, 142-146], we implement SPRING-H on the dimer complexes of the protein docking benchmark set [70] 3.0, which have both complex and unbound monomer structures solved in the PDB. Since SPRING has often partial structure aligned, we implement another version of SPRING-UB which superimposes the unbound monomer structures to the SPRING-H models after threading.

In Figure 2.8, we present the modeling results of SPRING-H and SPRING-UB in terms of the number of targets with an I-RMSD < 5 Å in the top-five models. As expected, the SPRING algorithm strongly depends on the level of filters for excluding homologous templates. At the sequence identity cutoffs of 30, 50 and 70%, SPRING-H generated models with I-RMSD <5 Å for 16, 23, and 28 targets, respectively. SPRING-UB has a slightly better result (with 17, 24, and 31 targets, respectively) due to the better model of the monomer structures.

As a control, two ZDOCK programs (V2.32 and V3.02) are implemented, which represents one of the best rigid-body algorithms according to the CAPRI experiments [150]. While both ZDOCK programs use the fast Fourier transformation technique to sample the conformation space of docking, ZDOCK V3.02 incorporates a new statistical pairwise potential to improve modeling selections [61]. ZDOCK V2.32 generates models of I-RMSD <5 Å for 11 targets, which is lower than both SPRING-H and SPRING-UB. However, the new pairwise potential significantly improves the ZDOCK V3.02 program with models of I_RMSD <5 Å for 26 targets, where SPRING could produce a similar number of correct models only if the homologue filter cutoff increases up to 50-70%.

In the right column of Figure 2.8, we also show the results of a hybrid modeling which has two models selected from SPRING-UB and three from ZDOCK V3.02. This combined approach outperformed all the four individual methods at different sequence identity thresholds (30, 50, and 70%) with correct models in the top five for 32, 36, and 40 targets, respectively. The results illustrate that the approaches of SPRING and ZDOCK are complementary to each other and a combination can lead to improved prediction accuracy.

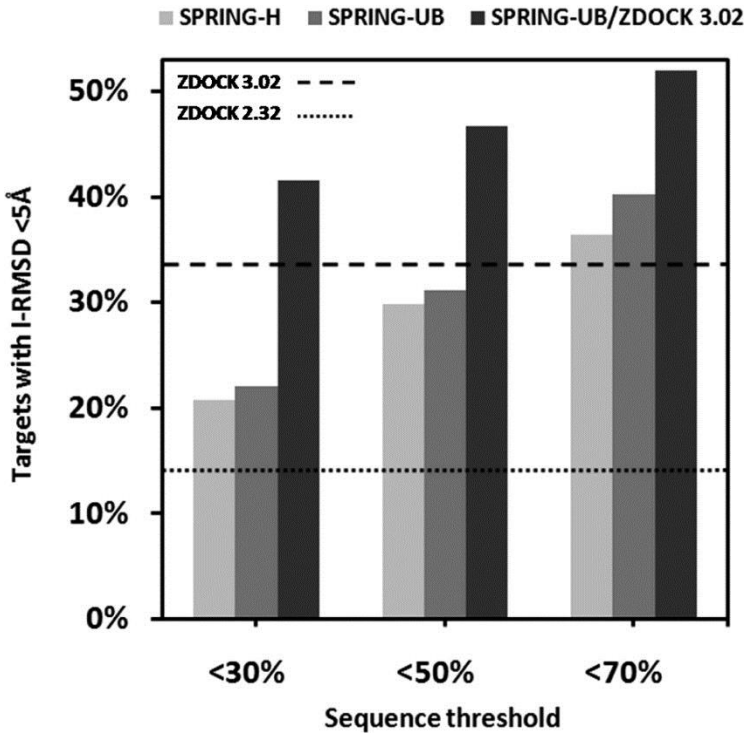


Figure 2.8 Comparison of SPRING and ZDOCK models at different target-template sequence similarity thresholds (30, 50, and 70%) on 77 heterodimeric protein complexes. The number of correct prediction, i.e., with I-RMSD < 5 Å, is shown for SPRING-H, SPRING-UB, ZDOCK (V2.32 and V3.02) and a combination of SPRING-UB and ZDOCK V3.02.

2.4 Conclusion

We presented SPRING, a new method to identify protein complex structural templates by mapping single-chain-based threading alignments with complex frameworks. Large-scale benchmark testing was performed in control with a recently developed cothreading method COTH [86] and the naïve extension of three monomer threading algorithms (MUSTER, HHsearch and PSI-BLAST), where the latter strategy is identical to that used by other authors in former template identification studies [84-86].

Based on a large test set of 1838 non homologous protein complexes, we showed that SPRING can produce models in the top five for 1115 (61%) targets with a TM-score >0.5, after all homologous template with a sequence identity >30% are excluded. The average TM-score for all targets is 0.58 with 34% of native interface contacts correctly predicted. In our recent studies, we have demonstrated that a TM-score >0.5 is statistically significant, which corresponds to a model of the correct fold in tertiary structure prediction [37] and in quaternary structure comparisons [18]. These data demonstrate that SPRING has the ability to generate reasonable correct complex

models for more than half of non homologous targets.

On the same benchmark protein set with the same homology filter, the TM-score of the SPRING models is 16, 45, 57, and 123% higher than that by COTH, MUSTER, HHsearch, and PSI-BLAST, respectively. The differences are statistically and all with p-values $<10^{-42}$ in the Wilcoxon signed-rank test. Considering the fraction of correctly predicted interface contacts, the SPRING models preserve at least twice as many native contacts compared to the competing methods. The corresponding p-values of the Wilcoxon signed-rank test are below 10^{-06} in all the comparisons. The number of targets with high quality models (i.e., with an I-RMSD <2.5 Å and >90% of interface residues aligned) was 219 in SPRING, compared to 89/67/93/47 in the competing methods, respectively.

Compare to COTH, a method that is conceptually closest to SPRING among the control methods, the major advantage of SPRING is the employment of the monomer-to-oligomer mapping which allows the use of the entire PDB library for complex frame derivation, while COTH exploits only a subset of complex structures at certain sequence identity cutoff which renders a loss of template frameworks; in particular the different binding modes from same monomer sequences (see the example in Figure 2.5). Such loss cannot be recovered by improving the scoring function of ranking.

We also control SPRING with the rigid-body docking algorithms on the docking benchmark databases [70]. As expected, the relative performance of algorithms strongly relies on the thresholds that are used to filter out homologous templates. However, a combination of the two approaches outperforms individual ones at all homologous cutoffs, which demonstrates the complementarities of the algorithms. Thus, a combination of both threading and rigid-body docking methods should represent a promising and reliable approach to the genome-wide structure modeling, where various targets with different levels of homology and difficulty need to be modeled.

For the evaluation of model qualities, we illustrated that there is a strong correlation between SPRING-score and the quality of the predicted models. If we consider a cutoff of good quality

models of TM-score >0.5, the false-positive and false-negative rates for a SPRING-score >13 are 0.22 and 0.27, respectively. These data not only underline the high specificity of the SPRING predictions but also highlight the limitation of current threading-based approaches, since SPRING does not have high confidence predictions in nearly 50% of testing cases. This is partly due to the limited availability of analogous template structures since all homologous templates with a sequence identity >30% have been excluded in the test. Nevertheless, considering the large number of possible protein-protein interactions in genomes, high accuracy predictions for even less than half of all interactions would yield highly valuable new insights, not saying that a higher successful rate should not be possible if homologous templates are included.

As a threading-based modeling approach, SPRING only provides partial structures on the target sequences, with C-alpha structural models derived from complex templates. The full length atomic structural models need to be generated using separate assembly and refinement procedures, such as TACOS (http://zhanglab.ccmb.med.umich.edu/TACOS/). Moreover, in the presented version, SPRING only considers pairwise protein sequences known to interact. The extension of the method to the high-order complex prediction is straightforward since no additional template library and monomer complex lookup table are needed. We are working on addressing these issues and plan to apply the SPRING mapping technique to the construction of genome-wide structural networks.

CHAPTER 3. Improving Quaternary Homology Based Structure Prediction by Inclusion of Intramolecular and Intermolecular Domain-Domain Interfaces

3.1 Introduction

Protein structure prediction has shifted from a physics based problem to a fold recognition problem where new sequences are matched to structures in the Protein Data Bank (PDB) in order to guide modeling [13, 14, 75]. For single domain proteins, the library is nearly complete and theoretically any domain can be folded using homology based approaches [2]. Modeling quaternary structures is analogous to using known homologs to model the structure of a single protein chain; identifying homologous structures at the quaternary level, interlogs, can be used to model and predict protein quaternary structure [151, 152]. Quaternary structures can be confidently modeled when a template is provided, but unlike the single domain library the quaternary library is not complete.

Though the number of interactions stored in the PDB is growing, and a confident template can often be identified for all queries [16]; the library of protein interactions is far from complete [18]. Due to the fact that proteins often associate in multiple orientations [153, 154] a confident template does not always align with finding the correct orientation. It has been hypothesized that there are a limited number of protein interface types and that it will take another twenty five years before the protein interface library is near completion [18]. Fortunately, it may be possible to extract more structural information from the PDB than is currently being utilized, since the interface between domains within a protein have some similarity between protein interfaces [6, 99, 100].

Its hypothesized that multidomain protein chains mainly arose from a series of gene fusions [155]. There is evidence that proteins that interact in one genome may be fused together in others; this fusion can provide an evolutionary benefit if the function occurs at the quaternary level, and this information can be used to confidently predict protein interactions [99]. Here domain-domain interactions are searched to observe for preservation of quaternary interface. Figure 3.1 provides

an overview of the methodology. Using the combined information of domain and protein interfaces can further extend the prediction capabilities of template based modeling of protein quaternary structure. Furthermore, the effectiveness of using the protein interaction library to orient the domains within a protein chain is also investigated. In this study we compare monomer and complex threading along with their respective libraries to evaluate their independent and combined potential for quaternary assembly and function prediction.

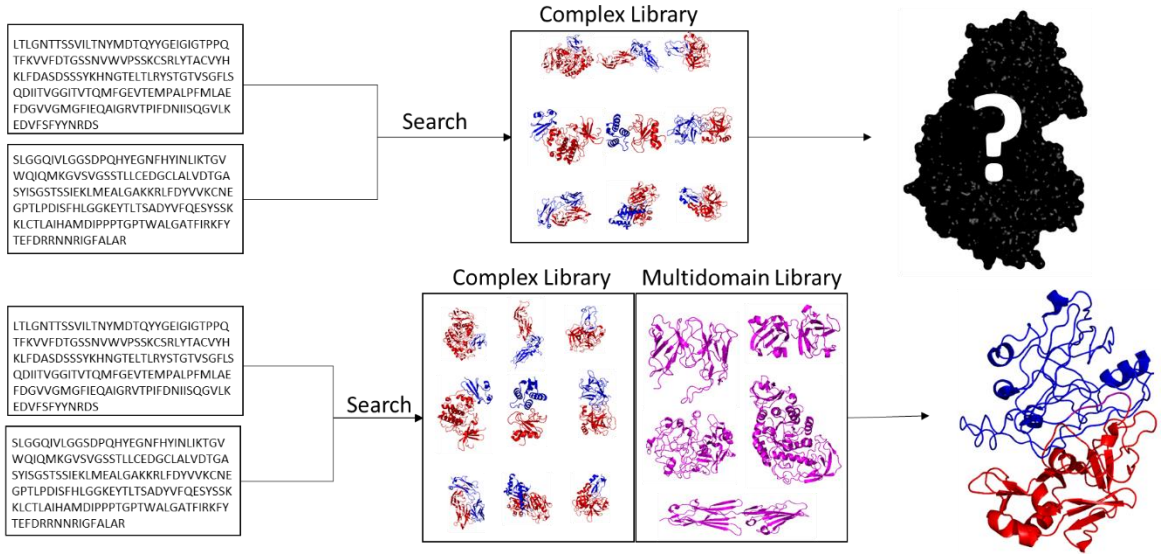


Figure 3.1 Incorporation of domain-domain template structures into the protein dimer library. Protein dimers homologous to the query sequence pair are searched for in the dimer library. The library is incomplete and often no information is present. By additionally searching through the monomeric multidomain library the domain orientations can be used as homologous templates for the query sequences.

3.2 Methods

3.2.1 Evaluation

Evaluating the global similarity of a protein model to the native structure is normally sufficient, but for protein interactions the quality of the interface is also important. The critical assessment of protein interactions (CAPRI) is a blind competition to assess the state of the art methods for predicting the structures of protein-protein interactions [40]. One of the CAPRI scores used for identifying the correct orientation between model and native is the fraction of native contacts (FNAT). The FNAT are the percentage of correctly predicted Cα-Cα interface contacts contained in the model within 8 Angstroms. An interface is considered similar if the model to native interfaces share at least 30% FNAT. The FNAT threshold is used here to consider if a template model correctly identifies the query interface. For monomer similarity the TM-score is used. TM-

score is used to compare global similarity between a model and native structure; the similarity score ranges from 0 to 1, with a value of 0.5 or greater being considered significant [34, 37].

3.2.2 Databases and Datasets

The domain interface benchmark was generated by creating a non-redundant set from over one hundred thousand monomer chains contained in the PDB. The structures were ordered by length and then filtered by a 70% sequence identity threshold. The protein chains were checked for multiple domains using the domain parser algorithm [11]. Each pair of domains were required to have at least 20 Cα interface contacts within 8 Å. Additionally, no discontinuous domains were allowed in the benchmark set. This resulted in 8942 chains containing 11838 domain pairs.

The complex library was created by extracting all experimentally determined complexes from the PDB [1]. All alternative binding modes of complexes were obtained from the PDB, and complexes with more than two chains were split into all possible dimeric combinations of protein chains. Then they were filtered by interface and monomer similarity. A new dimer is excluded from Dimer Library if it has at least 70% sequence identity and a TM-score similarity greater than 0.7 to a structure already in the library. This resulted in 29,454 dimer complexes. For identifying useful multidomain proteins as templates, the full monomeric PDB Library of over one hundred thousand chains is used.

3.2.3 Monomer and Protein Complex Fold Recognition

Monomer threading was preformed using the HHsearch algorithm [76] which is currently a preliminary step for quaternary fold recognition using SPIRNG; SPRING maps monomer alignment to protein complexes using a pre-calculated lookup table [88]. The lookup table contains listings of monomer structures that are similar to protein constituents in protein complexes. When the first chain is matched to a protein complex, a list of binding partners in the PDB file is obtained. The homologs identified from the second chain are quickly matched to the binding partners using the lookup table. A complex template that is found to be similar to both groups of homologs representing query chain A and B respectively, pass the first filter. The top ranked monomers for both sequences are structurally aligned to the complex frameworks and the template model is further evaluated by a statistical energy potential. The ranking of structures is evaluated by the

SPRING score, which is a linear combination of the HHsearch E-value, structural similarity from the docking of monomers to the complex framework and the interface statistical energy.

3.2.4 Identifying Protein Interaction Templates from Multidomain Protein Chains

Given two sequences, each are independently run through the monomer library using HHsearch. The protocol identifies single chains that are identified by both query sequences independently. When the query sequences align to the same set of residues in the template, preference is given to the query template alignment with the lower E-value. Multidomain templates identified with overlapping residues covering more than twenty percent of both alignments are realigned. The alignment with a higher similarity score blocks those residues from being considered from the alignment of the second chain. The realignment consists of structural alignments of the initial template model from the multidomain structure and the top ranked monomer template. These two structures are docked to the remaining region of the multidomain protein chain and the highest scoring alignment based on TM-score is retained. If neither has a TM-score above 0.5 the structure is removed from the confident hit list.

3.3 Results

3.3.1 Assessment and Improvement of Protein Dimer Library by Inclusion of Multidomain Protein Chains

Two approaches are generally used to evaluate the completeness of the PDB library. The first involves observing trends in the PDB and looking at rates of new information being added to the library; and the latter removes a set of structures from the library and looks for homologs using structural alignment below a sequence identity threshold of 30%. The density of the Dimer Library was evaluated using the latter approach, SPRING [88] and structural alignment with TM-align [38] was run on all protein complexes to evaluate the current completeness and the ability of threading to properly capture the information in the PDB. Table 3.1, shows the percentage of correctly identified templates at different sequence identity thresholds: 70%, 50% and 30%. An FNAT of at least 30% is generally considered to indicate a matching interfaces between two complex structures [156, 157]. Benchmark settings consist of excluding matching structures above a certain sequence identity threshold. At the 30% threshold SPRING identified confident templates for 72% of the complexes. There are often several binding modes that can be confidently predicted

given a pair of sequences [158]; however in this test only 39.4% of the target hits contained information regarding the orientation of the target complex. The current information identification gap between identifying the correct template with threading and its presence in the library was 8.7% under benchmark settings. The sparsity of the heterodimers is highlighted in Table 3.1. Regardless of orientation less than half are confidently identified at a high homology threshold of 70% sequence identity. Using the benchmark setting only 15.5% have confident hits and 6.7% of the targets have homologs with correct orientation. The difficulty in identifying heterodimers is due to the structural distribution of information in the PDB.

Identification Method	70% Seq ID Cutoff	50% Seq ID Cutoff	30% Seq ID Cutoff
Structure Alignment Hit	**64.9%**	**61.2%**	**48.7%**
SPRING Hit	**58.5%**	**53.7%**	**39.4%**
Multidomain Hit	**6.03%**	**5.9%**	**5.6%**
SPRING and Multidomain Hit	**59.3%**	**54.5%**	**41.1%**
SPRING Predicted Dimer Hit	**85.4%**	**82.5%**	**72.6%**
SPRING Predicted Heterodimer Hit	**46.3%**	**36.6%**	**15.5%**

Table 3.1 Percentage of Dimer Library that has homologs identified by structure and sequence alignment matches below sequence identity thresholds of 70%, 50% and 30%. Structure Alignment uses TM-align to search the PDB. SPRING and Multidomain Hit uses threading through the Dimer Library and Multidomain Library respectively. The predicted hits have confident sequence alignment matches that may also preserve orientation to the target.

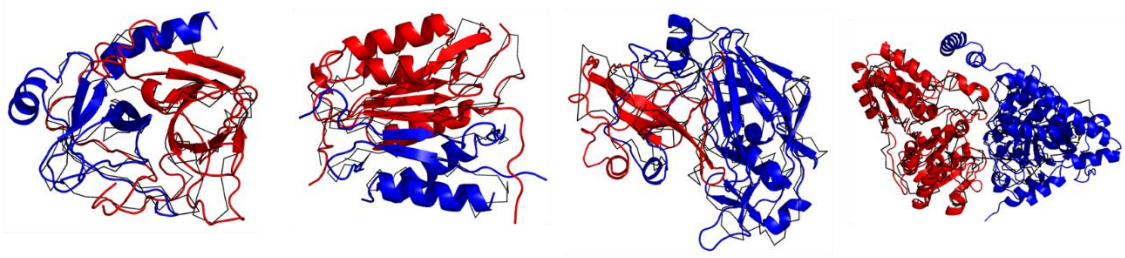
The current PDB is dominated by homodimers and structures that have diverged from them. Protein dimeric complexes are often grouped into two structural classes homodimers and heterodimers based on sequence and/or structural similarity of the interacting chains. Here the structural classes are determined by TM-align [38], if there is structural similarity between units they are classified as structural homodimers. Homodimer are interacting chains sharing high sequence and structural similarity. A complex whose components have dissimilar sequences but highly similar topologies are often classified as heterodimers, but they can often be modelled by homodimers due to their divergence from an ancestral homodimer. The term heterodimer here refers to structures whose interacting constituents are structurally distinct; they are often the most interesting and difficult cases to model. Only 9.5% of the PDB falls into this class. 11,838 pairs of interacting domains within 8,942 single chain multidomain proteins were also classified into homodimers and heterodimers. Unlike the dimer library, intramolecular domain-domain interactions are dominated by heterodimers which make up 73% of domain-domain interactions. The structural similarity between domain-domain interfaces and protein-protein interfaces were

then combined to test for improvement in predicting the orientations of protein-protein interactions.

Protein complexes were mapped to multidomain structures using HHsearch. Each alignment search requires each constituent protein to confidently align to the multidomain protein structure at different locations with a strict E-value threshold of 0.01. Under benchmark settings, 1658 out of 29454 complex structures were matched to multidomain structures that preserved the orientation and structure of the constituent chains. SPRING threading properly identified 11,603 complexes. The combination of the two libraries resulted in the identification of 12,098 structures. More prominent though was the improvement in identified heterodimers, Table 3.2. 111 heterodimers were identified within the Multidomain Library compared to 190 from the Complex Library. The procedures provided complementary results recognizing different types of heterodimers; combining the methods resulted in 266 identified complexes which provided a 40% improvement in heterodimer recognition. Figure 3.2 illustrates four examples of heterodimeric protein targets that now have suitable templates due to the inclusion of the multidomain library. Using benchmark setting of 30% sequence identity, SPRING using the dimer library was unable to identify structures with the correct quaternary structure resulting in all orientation having a FNAT of zero. The four example targets were matched to multidomain proteins that have highly significant global structural and interface similarity as defined by the TM-score and FNAT. The multidomain templates identified would allow for medium to high resolution structures to be modelled where it would not have been possible before.

Identification Method	**70% Seq. ID Cutoff**	**50% Seq. ID Cutoff**	**30% Seq. ID Cutoff**
SPRING	**882**	**627**	**190**
Mulditomain Hit	**160**	**152**	**111**
SPRING and Multidomain Hit	**935**	**680**	**266**
Improvement	**53**	**53**	**76**

Table 3.2 Percentage of the 2823 heterodimer structures where threading using dimer and multidomain libraries can identify templates below sequence identity thresholds of 70%, 50% and 30%.



Target	1ab9BC	1f9eAB	1lyaAB	1w85BC
Template	2vidA	4i1rA	2b42A	2o1xA
TMscore	0.719	0.726	0.510	0.829
FNAT	0.922	0.873	0.822	0.806
SEQID	0.173	0.169	0.187	0.201

Figure 3.2 Four examples of heterodimer targets that are not identified by SPRING using the dimer library at the 30% sequence identity threshold but are contained in the multidomain library. The red and blue colored chains are the target heterodimer structures; the black line is the C-alpha trace of the multidomain template model.

3.3.2 Orientation of Domains Using the Complex Library

Incorporating interfaces in the dimer library in order to orient the domains in a multidomain protein was also investigated. The general approach to modeling multidomain protein chains is to use threading algorithms to identify full coverage templates that give structural information about the topology of the individual domains and their respective orientations [104, 159]. Often full coverage templates are not available and the sequence needs to be partitioned into individual domains modelled separately and then reoriented [159]. However the sequence partition and reorientations of the domains are both difficult and unsolved problems [159-161].

Multidomain protein are treated as artificial protein complexes and searched for through the dimer library using SPRING. Multidomain protein sequences are given ideal partitions into domains based on the proteins structure using the domain parser algorithm [11]. Domain pairs were submitted to the SPRING algorithm as artificial complexes. Additionally as a control, each pair of interacting continuous domains were run through HHsearch as an artificial monomer. The results are tabulated in Table 3.3. At benchmark settings, 553 of the domain pairs are identified by protein interactions. The majority of the domain pairs, 8701 out of 11838, were identified by general monomeric threading using HHsearch. There was substantial overlap between the targets identified by both methods. Combining the two methods identified 8797 domain interface pairs. Although multidomain proteins are often hard targets to model, the current library shows that interacting continuous domains stored in the monomer library are well represented.

Identification Method	70% Seq. ID Cutoff	50% Seq. ID Cutoff	30% Seq. ID Cutoff
HHSEARCH	**10200**	**9856**	**8701**
SPRING	**714**	**677**	**553**
SPRING and HHSEARCH	**10263**	**9923**	**8797**
Improvement	**63**	**67**	**96**

Table 3.3 A tabulation of the number of successes of correctly orienting domain-domain interactions with monomer threading (HHSEARCH) and dimer threading (SPRING) along with their corresponding libraries. 11838 domain pairs were oriented using HHSEARCH and SPRING templates below three sequence identity threshold.

3.3.3 Diversity of Multidomain Templates for Dimeric Modeling

There are a multitude of ways two proteins can interact and form distinct interfaces amongst themselves; it is important to investigate the range of unique dimer interfaces that can be modeled using multidomain structures. Overall 1,936 protein dimer structures were mapped to 5,895 single chain proteins. The dimer library was clustered into families and interfaces in order to evaluate what class of dimeric interactions are similar with domain-domain interaction. A family consist of all interlogs where an interlog is a protein-protein interaction which is preserved among pairs of homologs. An interface cluster is a group of interlogs where the interaction and interface are preserved. Two protein complexes belong to the same family if two quaternary structures have interacting chains that are structurally similar. If the interlog pair A and B are similar to another pair, they belonged to the same family. Structures are considered to have similar tertiary structure if they shared a TM-score greater than or equal to the 0.5 threshold after structure alignment with TM-align. SPRING was used to quickly rank similar structures in the dimer library and check for tertiary and interface similarity. If two pairs contained at least 30% of similar interface contacts, they are grouped into the same interface cluster.

Single-linkage clustering is the criteria used for including a new structure or merging two clusters together. Under single linkage clustering if two groups of structures share any pairwise similarity the two groups are merged into one cluster. SPRING threading along with structural alignment were used to compare each protein complex to all other protein complexes. The first protein complex SPRING results are obtained and all homologs within the similarity threshold are considered neighbors. Then for each neighbor the SPRING results are checked again for new structures that meet the similarity threshold. This process is repeated until no new structures are added to the cluster. The next cluster is generated starting with a protein complex not currently in

a cluster. Hierarchical clustering is completed when all complexes belong to a cluster. The clustering generated 4,234 family clusters and 13,046 interface clusters. The domain-domain interactions that matched protein interactions were contained in 211/4,234 family clusters and 311/13,046 interfaces.

3.3.4 Functional Conservation and Alternative Binding Modes

Confident alignments between multidomain proteins and query pair interacting proteins can be used as templates for predicting structure and function; here it's investigated how often confident dimer alignments to multidomain structures correlate to robust structure and function information. Gene ontologies (GO terms) were used to investigate functional conservations amongst protein fusions with both conserved and new interfaces identified compared to the complex constituents. Gene ontologies provide a standard definition of biochemical properties of macromolecules [162, 163]. Here the conservation of molecular functions is looked at. Not all proteins in the PDB are covered by the GO ontologies. Of the 302 heterodimer targets only 137 had GO terms for both target chains and the multidomain templates. On average, after removal of generic functions, each remaining protein chain had 2.42 GO terms. Two proteins are classified as having the same conserved functionality if the share any go terms.

Multidomain protein chains that were identified to be homologous to both chains in a dimeric structure often shared similar go terms to both individual components of the complex. Table 3.4 highlights conservation of molecular function and protein orientations amongst domain-domain and protein-protein interactions. Each dimer target can map to many individual protein chain structures with confidence. Each row shows confident matches within a ranking threshold that share functional or structural interface similarity to the target dimer. The columns are counts of targets for which multidomain structures can be used to predict the dimer structure or function with increasing confidence due to consensus information from multiple protein fusion matches. The functions in our testing set are shown to be 1.7 times as likely to be preserved as the specific orientation. The vast majority, 92%, of highly ranked homologs that preserve the orientation also preserve the function.

Top Ranked Templates	Go Term Hit	Go and Interface Hit	Interface Hit
Top 5	**100**	**53**	**57**
Top 10	**102**	**55**	**58**
Top 50	**104**	**56**	**61**
All Confident Templates	**104**	**56**	**61**

Table 3.4 Structural and functional recognition from multidomain templates. The table provides a count of the number of heterodimers out of 137 that are matched to homologs that share GO terms and interface structure.

The ideal match for protein structure and function prediction refers to recognizing homologous structure that has a high structural and functional similarity. Figure 3.3 is an example of an ideal match. The protein 1a5k in *Enterobacter aerogenes* is permanently associated in a higher level organism *Cajanus cajan*. Three separate chains of protein 1a5k have significant alignments to the protein chain 4g7eA, and both structures are ureases. Although this structure is easily identified using the dimer library the structure represents a protein containing three separate chains and highlights the potential of orienting more than two proteins at a time using protein fusions.

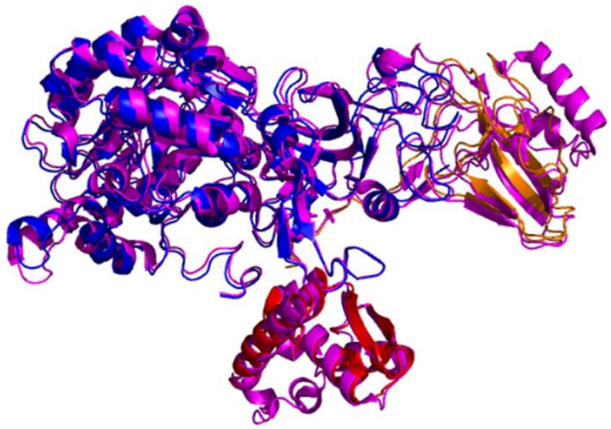


Figure 3.3 1a5k trimeric complex mapped to protein fusion 4g7eA.

Surprisingly many of the highly confident alignments preserved function but not orientation to the target structure. Changes in orientation did not have much of an effect on the specific molecular function of the protein complex or multidomain structure. We present two interesting cases where the orientation in the two images where expected to be preserved but were not. It is anticipated that when two structures that interact are found within a single chain with a long enough linker regions the original interface should be preserved. Figure 3.4 shows an example of a confident hit where the domains A and B of the single chain are highly similar to chains A′ and B′ respectively. The linker is long enough to preserve the orientation of complex but it is not contained in the

image. It appears that the error is due to the simplification of an assumption regarding how proteins interact. Here we assume that each binary interface within a large multimeric complex can form pairwise interaction independently. However, this is not necessarily true. The case study 3qqc has two chains that map to 4ddiA. Yet the 3qqc structure is actually a tetramer where there may be an order of assembly. The first two proteins may undergo an interaction that causes a conformational change in the second chain that allows the third chain to bind. In the tetramer the third chain is floating in the cytoplasm and the formation of the complex is restricted by the third chain finding the complex. However in the multidomain structure the second and third chains are fused together. This can increase the formation speed of the tetramer due to the permanent localization of the third chain. In Figure 3.5, the protein complex 3vonAC matched to a protein chain 4ddiA that has similar tertiary components, but the orientation is not preserved. However the 4ddi protein is homo-mer in the PDB and the interface in the complex 3vonAC is used to construct the quaternary structure of the homomeric protein 4ddi.

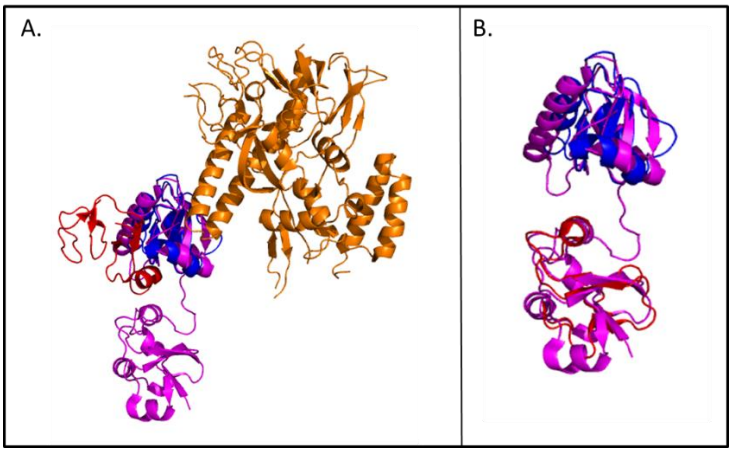


Figure 3.4 Case study of interaction loss between domains. The target dimer 3qqcDE has tertiary structural similarity to the protein chain 2exuA. A.) Shows the superposition of the two domain structure to the trimer complex 3qqcCDE. The red and blue chains are used to search for multidomain structure; the orientation is not preserved although there is a high similarity between the protein fusion and dimer and the linker region appears to be long enough to not disrupt the assembly of the dimer interface. B.) The D chain colored red shares a TM-score of 0.83 to the first domain while the E chain colored blue has a TM-score of 0.77. The sequence identities are 0.232 and 0.328 respectively.

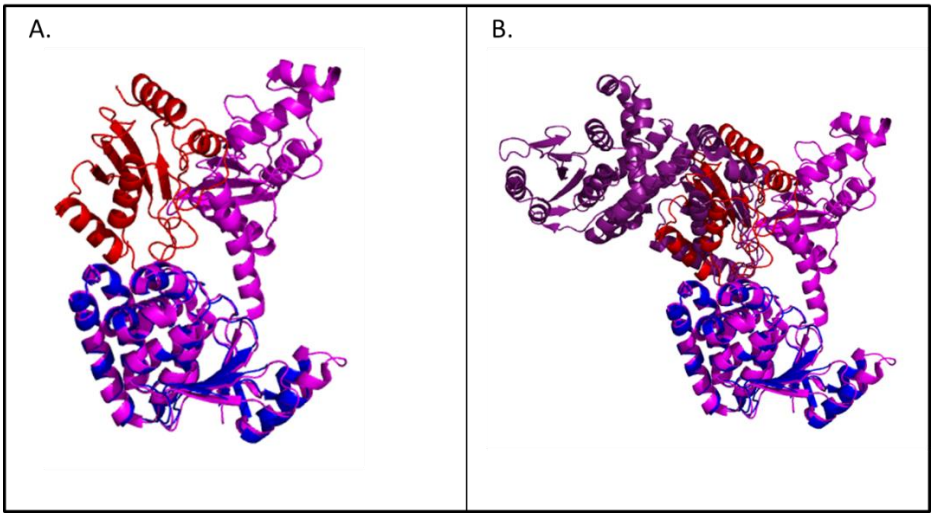


Figure 3.5 Dimer 3vonAC (red and blue) aligns to both domains in multidomain chain 4ddiA (purple), but interaction is preserved between homodimers contained in 4ddi protein. A.) Shows the superposition of 3vonAC to 4ddiA. Despite the high tertiary similarity of 3vonAC having TM-scores of 0.97 and 0.922 to the constituent domains the interface is not preserved. B.) The3vonAC orientation is observed between two separate protein chains in the 4ddi PDB file.

3.3.5 Multidomain Structures Potential for Predicting and Modeling Protein Interactions.

The goal of incorporating multidomain structures is increasing the current capabilities of bioinformatics based approaches towards predicting and modeling novel dimeric structure. The previous sections were used to validate the inclusion of the library by observing how many structurally resolved dimers could be confidently identified using the single chain library. 162 heterodimeric complexes with similar orientations and topologies to single chain proteins were identified. 302 complex structures had high confident matches, where the tertiary constituents had higher similarity to a multidomain protein but the orientations were different. These interfaces may still represent actual protein interfaces that cannot be verified by the current PDB. Furthermore it is worth seeing the potential for interactions to be modelled by proteins fusions. If two proteins align to a single chain, it is worth considering that these two structures or homologs of them interact. Here we look at all possible combinations of proteins that may interact by mapping them to multidomain protein chains. 22,136 non redundant by sequence protein chains were obtain. Every possible pairwise combinations was compared to the monomer library. 7,085,057 proteins pairs independently mapped to 10,057 protein chains. This is 1.7 times as many protein structure as were identified using the dimer library, and suggest there are many more interactions that multidomain proteins chains can model that the dimer library cannot.

3.6 Discussion

There are two similar yet separate hypothesis that guided this project: there are limited number of interface topologies and domain-domain interfaces are ancestral fused protein interactions. Similar to the single domain case there may be a structural limit to how two domains can interact which is irrespective of the domains interacting within a single polypeptide chain or between two polypeptide chains. Due to the sparsity of the interfaces in the PDB the domain and protein interaction sets appear to be separate with minor overlap, but with time the structural overlap will be significant. The alternate hypothesis is if two separate protein chains are only functional at the quaternary level there is an evolutionary benefit to permanently associating the structures together; a protein interaction in one genome may be permanently fused in another one [99]. Regardless the merging of the two sets provides complementary information that improves the breadth of information regarding multidomain and protein-protein interaction orientations. The augmented orientation libraries were compared to the independent constituent monomer and dimer libraries to assess similarities, limitations and enhancements of template based structure and function prediction of quaternary structure.

The depth of the dimer library was examined using structure alignments and the ability of the SPRING dimer alignment algorithm to identify dimer templates was assessed. Structural alignments reveal extraction of information under a perfect scenario, when correct tertiary templates can be identified and a template is available, the quaternary structure homolog can be identified 48.7% times using the current dimer library under benchmark exclusion settings whereas a similar experiment using single domain proteins always found a suitable match [10]. A proper template is identified 39.4% of the time when performing sequence alignments using SPRING when excluding homologous matches above 30% sequence identity. The test reveals that the dimer library is still sparse and several years of unique additions to the PDB are required to complete the dimer library which is especially true for heterodimers. Most of the sequentially defined heterodimers still have highly similar tertiary level topologies; these are often believed to be the results of divergence and gene duplications from a homodimer ancestor [164]. Those heterodimers formed from structurally distinct proteins represent only 9.5% and are often the sole representative of a pairwise interaction. Intramolecular domain-domain interfaces have some overlap with protein

interfaces, and furthermore the structural classes of domain-domain interactions are mostly in the category lacking in the dimer library, the heterodimers.

Combining the two libraries resulted in significant improvement in the heterodimer set and a minor overall improvement for overall protein interactions. Minimal improvement was found in orienting domain-domain interactions using combined library. Although identifying domain-domain orientations remains an open and difficult problem; the region where the protein interactions overlapped with multidomain proteins was abundant in the monomer library. Whereas the intersection of the sets is contained in a sparse data region of protein dimer library. Additionally integration of the two data sets can provide new information such as conservation of residue position in a sequence/structure profile which is an important feature for bioinformatics based hotspot detection and binding energy predictions [165]. Still even with confident sequence alignments caution has to be taken due to the many alternate binding modes two proteins that interact can have. Two cases Figure 3.4 and 3.5 present confident matches with high sequence identity of the dimer target to a protein fusion chain; however neither contains the target interface. The molecular function appears to have higher conservation then the set of biological interaction networks or the assembly order of the complex.

The matching of a pair of dimer sequences to a single chain protein provides a threshold of confidence that the protein preserves the tertiary and quaternary structure of the protein dimer. PDB structural information has been shown to be useful for predicting genome wide protein networks [87, 88, 166]. The primary test set revolved around verifying the validity of this argument. However the interesting cases are predicting and modeling new interactions not already stored in the dimer library. Here a sequentially non-redundant set of protein chains were obtained and every pair of chains was checked to see if they both map to a protein fusion chain. HHsearch confidence scores above its threshold has low false positive rates; given two protein chains mapping to a single protein fusion suggest the tertiary structure of both independent chains is represented by the protein fusion chain. The dimer library confidently mapped to 5,895 protein fusion chains whereas the all against all confidently mapped to 10,057 protein fusion chains. This suggest that there are many structural pairs of proteins that map to a single chain which in turn may provide interface information for a distantly related interlog.

3.7 Conclusion

The dimer library is far from complete, even with a perfect fold recognition algorithm there is a significant gap between the dimer interactions in nature and those that can be modeled and predicted with the current PDB. Excluding homologous structures at 70, 50 and 30% sequence identity thresholds only 58.5, 53.7 and 39.4% of the Dimer Library have a suitable template identified by SPRING, whereas structural alignment identifies 64.9, 61.2 and 48.7% of the library respectively. This demonstrated that the current limitations in dimeric threading is not the searching algorithms, but the limited number of structures in the library. To circumvent this problem the interface between domains with a protein chain can be used to boost the dimer library. Excluding homologous structures above 30% sequence identity, the new comprehensive library had an overall improvement of 1.7% for recognizing dimers, and a 40% improvement for identification of heterodimeric templates. A similar attempt was made using the comprehensive library to model multidomain structures that led to a 1% improvement in identifying a correct template. The two libraries were found to share a structural overlap of 5.6% under benchmark settings, the major improvement revolves around heterodimers being sparse in the dimer library which represents 9.5% of the database whereas structurally distinct domains interacting within a single protein chain constitutes 73% of the pairwise interactions. The correct orientation and function of heterodimer targets could be extracted from multidomain templates 76% and 41% of the time respectively. Overall the information regarding orientation of domains and proteins is still rather sparse within the PDB. Template based modeling is currently the most accurate approach for predicting protein structure and function. Integrating structures from domain-domain and protein-protein interactions into a comprehensive library can further extend the capabilities template based structure and function prediction on the quaternary level.

CHAPTER 4. Full-length Structure Prediction of Protein Complexes from Sequence by Template Identification and Atomic-level Structural Refinement

4.1 Introduction

Proteins are large macromolecules that carry out numerous essential cellular functions. The majority of protein function occurs at the quaternary structure level, which is composed of multiple interacting protein chains, some of which are permanent while others are transient in nature. To obtain system level understanding of living cells, it is essential to obtain structural and functional level information of the protein interactome. While a number of high-throughput methodologies exist that are capable of elucidating functional level information of the interactome [167, 168] (such as the constituents of a protein complex) there is a dearth of information in the three dimensional structural space. Accurate structural determination methods such as X-ray and NMR techniques could in principle provide this information, but the cost and labor intensiveness of these methods have caused structural genomics to lag behind the number of validated protein-protein interactions [169].

Regarding protein structure prediction, there are two classes, modeling *ab initio* and template based methods. *ab initio* methods use first principle based potentials to fold a protein from a random state. There have been some success but the computational expense of these methods limit them to small proteins and peptides. Template based modeling identifies structural analogous to the target sequences and uses their constraints to model the protein structure. Template based modeling of tertiary level structures is capable of producing high resolution structures when a suitable template is identified [13]. A limiting factor in template based modeling is the requirement of a structural analog being present in the PDB; for single domain proteins there is strong evidence that all the protein folds are already contained in the PDB [10]. Fortunately for protein complexes, there is evidence that the diversity of structural interfaces is finite and the PDB library for

complexes is approaching completion [16, 17]. Given the encouraging progress observed in the field of template-based structure prediction of monomeric proteins [13, 170, 171] along with the current set of representative interfaces within the PDB, a similar level of success is expected for these methods extensions towards quaternary structures.

Two of the major existing problems in template-based protein structure prediction are the detection of remote homologous templates and the refinement of the template closer to the target. A number of significant efforts have been made in recent years to develop bioinformatics based approaches to predict protein interactions [40, 79-81]. Currently there are three classes of bioinformatics based approaches for identifying and modeling protein-protein interactions by sequence alignment: dimer threading, monomeric threading and oligomer mapping and modeling the constituent chains followed by docking [82]. Dimer threading directly aligns the query sequences to the target complex which allows for interface information to be considered during the alignment, example programs are: MULTIPROSPECTOR [83]by Skolnick's group, HOMBACOP by Kundrotas et. Al [84], the strategy used by Aloy et. al. [85] and the work by Sinha et al [89], and COTH [86]. Monomeric threading and oligomer mapping starts with generating query alignments to the monomer library. Complexes are identified using a pre-generated lookup table where every protein chain constituent in an interaction is represented by a homologous structure in the monomer library. The recently developed programs SPRING and PrePPI are examples of this protocol [87, 88]. The last approach for identifying possible binding orientations uses monomer models which are docked using a physics potential [125], statistical based potential, and in some cases template coordinates [87, 88, 166]. Physics based docking can be applied to the structural constituents of a protein complex using programs such as ZDOCK, but successful cases are limited to special classes of protein-protein interactions [125]. Additionally, ZDOCK has a reduced performance when using homology models [127]. Proper template identification is an essential prerequisite for homology modeling, combining multiple methods that produce complementary results can improve modeling. The three approaches independently have inherent strengths and weaknesses for identifying protein complexes, but by combining the results together in a manner analogous to the monomer meta threading servers [14] may improve the coverage of complex interaction can be improved. Here we investigate the use of this methodology to quaternary level orientation

prediction using the three classes of bioinformatics based interface prediction algorithms represented by COTH, SPRING, and docking using I-TASSER models docked by ZDOCK.

Threading templates do not generate full length models, and docking often needs refinement regarding clashes, interface contacts and flexible regions. Here, we describe a new algorithm, TACOS, a hybrid approach geared towards generating full-length protein dimer structures from sequences alone similar to the approach used by M-TASSER which is a modelling program using templates identified by MULTIPROSPECTOR as starting structures [140]. TACOS starts from threading alignments, and construct full-length structure models by modelling threading gapped regions *ab initio*, and reassembling the continuous aligned template fragments in a course-grained schematic similar to the monomer structure prediction algorithm I-TASSER [47, 110, 172]. Importantly, in addition to TACOS retaining the portion of the I-TASSER energy and protein folding methodology, it introduces a set of statistical-based interface potentials to capture the unique idiosyncrasies of protein-protein interactions, along with a movement for improving the relative orientation of the chains.

4.2 Methods

TACOS is a sequence to structure algorithm for protein complexes, Figure 4.1. There are four critical steps for template based modeling; (1) template identification; (2) a robust funnel shaped energy function; (3) an efficient protocol to search the conformational space; (4) identifying the best decoy from a set of ten thousands possible structures; (5) and generating full atom models. Each of the steps is described in detail in the following sections.

4.2.1 Template selection

In the first step, TACOS attempts to identify homologs/structural analogs of the given query sequence of the complex by threading it across representative libraries of structures for monomers and protein complexes. The complex library is obtained from the PDB biounit files and screened for structural templates at 70% sequence identity and TM-score of 0.8. Complex templates are searched for by COTH/SPRING, while the individual chains of the query complex are threaded across the monomer structure library, which is six times larger, by LOMETS. LOMETS is a threading alignment approach that combines multiple complementary threading alignment algorithms for fold recognition [14]. The currently included threading program in LOMETS are:

FFAS (sequence profile-profile match) [173], HHsearch (hidden Markov model to hidden Markov model alignment) [76], Muster (multiple structural profile-profile alignments) [78], PRC (hidden Markov model match) [174], PROSPECT2 (contact-assisted profile-profile alignment) [175], dPPAS (depth profile profile alignment) [176], SAM-T02 (sequence to hidden Markov alignment) [75], SPARKS (profile alignment assisted with single-body potential) [112], SP3 (profile alignment assisted with fragment depth) [177]. The individual chain templates thus identified by LOMETS are superimposed on the dimer template framework identified by COTH/SPRING threading.

4.2.2 Mapping of the dimer onto an artificial monomer on a CAS lattice

The initial template is course-grained and represented by the C-Alpha Side Chain Based (CAS) model consisting of only the Cα atom for each residue and the side chain center of mass (SG). The portion of the template without alignment to the query sequence, gapped regions, are initially missing from the model. To generate the initial full-length CAS model, a Cα framework for the full-length structure is first constructed by using a Cα random walk on a cubic lattice to build the gapped regions between aligned fragments. If any of the unaligned regions between two aligned fragments cannot be connected completely by a series of 3.8 Å Cα-Cα bonds, then an external spring-like harmonic force is applied to bring the fragments together until reasonable bond lengths are achieved [104]. The structure of the complex is then segregated into "template-aligned" and "template-unaligned" regions and placed on the CAS on- and off-lattice model used by I-TASSER. Here, the Cα atoms of the template unaligned (gapped) regions are placed on the lattice for computational efficiency and are built *de novo*, while the template aligned fragments are placed off-lattice for maximum accuracy and subjected only to rigid body adjustments. The side-chain SG atoms are always placed off-lattice. The dimer is represented on the lattice as a single artificial monomer with a long "psuedobond" connecting the C-terminal of the first chain with the N-terminal of the second chain. The pseudobond is kept completely flexible during the assembly and refinement simulations and can have any length. This convenient trick allows the well-established simulation protocol of I-TASSER to essentially treat the complex as a monomer prediction problem and ensures the direct adoption of the many energy potentials and movement schemes in I-TASSER to TACOS.

4.2.3 Structure assembly

The initial dimer structure developed in the previous step is then placed on the CAS lattice and subjected to parallel hyperbolic Monte Carlo sampling [47, 104, 116]. The movement scheme for the Monte-Carlo sampling can be divided into 2 distinct types; 1) local intra-chain moves for packing the individual chains of the complex and 2) large rigid-body inter-chain moves for identifying the correct orientation of the chains with respect to each other. In general for each replica, each large inter-chain move is followed by multiple intra-chain moves to stabilize the local structure of the individual chains and the process is repeated. A flow-chart of the movement scheme of TACOS is shown in Figure 4.1.

Local intra-chain movements. The local intra-chain moves adapted from I-TASSER can be classified into two types: 1) on-lattice bond rebuilding of the unaligned regions and 2) off-lattice rigid body moves for the template aligned fragments. On-lattice movement includes extensive bond rebuilding moves for the *ab initio* generation of the template unaligned regions. For the on-lattice moves, 312 bond vectors restricted to the cubic-lattice points are pre-computed with a bond length varying between 3.26-4.25 Å (the variability of the bond length allows for larger conformational flexibility). During each attempted movement, 2 or 3 continuous bond vectors picked randomly from any of the on-lattice regions are replaced by the pre-computed bond vectors. Larger movements are achieved by a combination of multiple 2 or 3 bond replacements. Each move is accepted based on the standard Metropolis Monte Carlo criteria.

Inter-chain movement: Inter-chain movement was incorporated on the premise that the initial orientation of the two chains may be incorrect or require readjustments. Here, one of the chains of the dimer (the smaller one for heterodimers and either chain for the homodimers) is first randomly moved to a new position and then drawn closer by a short independent Monte Carlo simulation. To draw the chain closer, the vector between the center of mass (COM) of both chains is defined and the chain is subjected to small randomly selected rigid body rotation and translation motions with each move accepted or rejected based on the standard metropolis Monte Carlo criteria. During these moves, the center of mass is kept fixed along the original COM vector. Newly defined inter-chain specific potential terms were defined to guide the movement. At the end of each cycle of the

inter-chain move, the final position is rejected or accepted once again based on the Metropolis criteria.

4.2.4 Energy Function

A statistical, knowledge based energy function was designed and optimized to drive the TACOS simulation. Since TACOS seeks to simultaneously build both the individual chains of the dimer as well as modeling their orientation and interface match, the potential terms belong to two distinct classes: 1) local terms aimed at mimicking the monomeric conformational energy landscape and 2) inter-chain terms to maximize the complementarity of the dimer interface required to stabilize the interaction. Due to the use of the pseudobond, the simulation protocol can essentially treat the problem as a monomer prediction problem, thus allowing all the inherent I-TASSER potential terms [47, 110] to be directly carried over and used to guide the local conformation search.

The new inter-chain energy terms contain a mixture of template based restraints and knowledge-based potentials derived from the structure library of protein complexes. The new energy terms are discussed in more detail in the following. The coefficient *w* indicates the weight of the energy term (all terms are combined linearly), which was carefully optimized by large-scale benchmarking on the training set proteins.

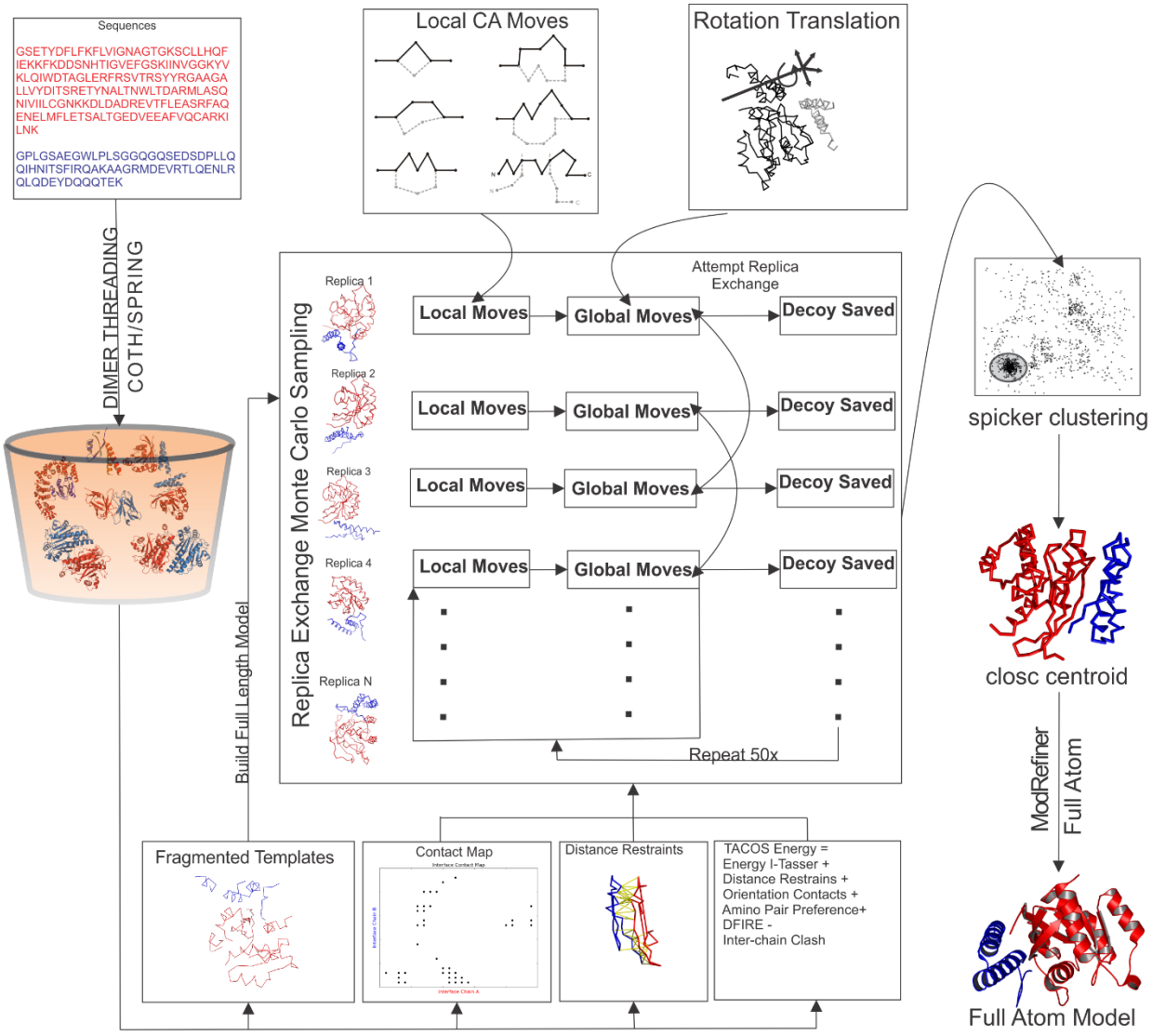


Figure 4.1 Flowchart of the TACOS, Template-Based Assembly of Complex Structures, protocol. Given two protein sequences known to be involved in a protein-protein interaction, TACOS first searches a curated structure library of dimeric protein complexes using COTH/SPRING. The TACOS energy, predicted interface contacts and interface distance restraints are derived from the Dimer PDB library. The identified templates coordinates are used as starting positions and an initial full length model is built from them. This initial structure is placed on the C-Alpha Side-chain (CAS) based on-off lattice system similar to that used by the I-TASSER. The templates are then reassembled and refined using the TACOS replica-exchange Monte Carlo simulation. The decoys (native-like protein conformations) thus generated are then clustered by SPICKER and the cluster centroid is refined further by the ModRefiner program to generate full atomic models.

i) E_{COM}: This term based on the distance between the Center of Mass (COM) of the two dimer chains is required to prevent the two chains from drifting too far away during the simulation procedure. The equation is given by

$$E_{COM} = w \times d_{COM}^2 \qquad (1)$$

where d_{COM} is the distance between the two centers of masses. On the other hand, this potential can dictate one chain into collapsing onto the other and therefore needs to be balanced with a large clash penalty to ensure a roughly accurate placement of the chains with respect to each other.

ii) E_{clash}: A large clash penalty is assessed if any atom (Cα or SG) of one chain has a distance < 3.8 Å from any atom in the opposite chain.

iii) $E_{Ncontact}$: To be stable, a number of inter-chain contacts are required to stabilize the dimer interface. Accordingly, based on the hypothesis that at least 30 inter-chain contacts are required for a stable complex formation, a large penalty was assessed for decoys with no inter-chain contacts. This penalty is gradually decreased as more inter-chain contacts are formed, eventually becoming a constant for more than 30 inter-chain contacts. The equation for this energy term is given by

$$E_{Ncontact} = w\begin{cases} 15 - N \text{ if } N < 30 \\ -15 \text{ if } N \geq 30 \end{cases} \qquad (2)$$

where N is the number of inter-chain contacts. The energy is kept constant after 30 inter-chain contacts are formed to prevent the structures being compressed into being flat sheets where all residues are forming contacts.

iv) $E_{oricontact}$: For any residue i and j in opposite chains which are in contact, the orientation of the unit bisector vectors of i and j can be in three different orientations as defined by their dot product: parallel, anti-parallel or perpendicular. This energy term is described in the form of a general exclusion volume potential for the SG atoms and is given by

$$E_{oricontact} = -w\sum_{i=1}^{Lch1}\sum_{j=1}^{Lch2} E_{i,j}(s_{i,j}) \qquad (3)$$

where

$$E_{i,j}(s_{i,j}) = \begin{cases} -6 & \text{when } s_{i,j} \le R_{\min}(A_i, A_j, \gamma_{i,j}) \text{ and } c_{i,j} \le 6 \\ e(A_i, A_j, \gamma_{i,j}) & \text{when } R_{\min}(A_i, A_j, \gamma_{i,j}) \le s_{i,j} \le R_{\max}(A_i, A_j, \gamma_{i,j}) \text{ and } c_{i,j} \le 6 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Here, *Lch1* and *Lch2* are the lengths of chain 1 and 2 respectively, $c_{i,j}$ ($s_{i,j}$) is the distance between the Cα (SG) atoms of residue *i* and *j*, A_i (A_j) is the amino acid type for residues *i* (*j*), $\gamma_{i,j}$ is the orientation of the bisector vectors of *i* and *j*, $R_{min}(A_i,A_j,\gamma_{i,j})$ ($R_{max}(A_i,A_j,\gamma_{i,j})$) is the minimum (maximum) distance observed between amino acids A_i and A_j for either of the three $\gamma_{i,j}$ types in the complex structure library and $e(A_i, A_j, \gamma_{i,j})$ is the probability of an amino acid pair to be in the orientation $\gamma_{i,j}$ (equal to the total number of times any particular amino acid pair is observed in the orientation γ divided by the total number of times that particular amino acid pairing is observed in the protein complex structure library).

v) $E_{respref}$: This is defined as the preference of the Cα atom of an amino acid A_i to be present in one chain when the Cα of another amino acid A_j is present at a distance less the 6.0 Å on the opposite chain and is given by the equation:

$$E_{respref} = -w \sum_{i=1}^{Lch1} \sum_{j=1}^{Lch2} P(A_i, A_j) \quad (5)$$

where

$$P(A_i, A_j) = \begin{cases} \dfrac{f(A_i, A_j)}{\sum_{j=1}^{20} f(A_i, A_j)} \times \dfrac{t(A_i)}{\sum_{i=1}^{20} t(A_i)} & \text{if } c_{i,j} \le 6 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Here, $f(A_i,A_j)$ is the total number of times the pairing of amino acids A_i and A_j is observed at a Cα distance less than 6.0 Å among the complex structures in the library, while $t(A_i)$ is the total number of times the amino acid A_i is observed among the structures in the library.

vi) $E_{resdistpref}$: This potential terms seeks to account for the preferred distance between the Cα atoms of any two pair of amino acids A_i and A_j. Since we are only interested in the interface residues in this case, the range of distance considered is from 4.0 Å to 12.0 Å which was divided into 8 distance bins $\lambda_{i,j}$ of 1.0 Å each. Thus the final potential is given by the equation

$$E_{resdistpref} = -w \sum_{i=1}^{Lch1} \sum_{j=1}^{Lch2} D(A_i, A_j) \quad (7)$$

where

$$D(A_i, A_j) = \begin{cases} q(A_i, A_j, \lambda_{i,j}) & \text{if } 4.0 \le c_{i,j} \le 12.0 \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Here, $q(A_i,A_j,\lambda_{i,j})$ is derived from the protein complex structure library and is given by the total number of times the Cα atoms of the amino acids A_i and A_j belonging to different chains of a complex are present in the distance bin $\lambda_{i,j}$ divided by the total number of time the Cα atoms of A_i and A_j are present within 4.0 Å to 12.0 Å of each other.

vii) $E_{distmap}$: This energy function is a template-based restraint which penalizes the deviation observed between the distance of residue *i* in chain 1 and residue *j* in chain 2 in the generated decoys with respect to the template and is given by the equation

$$E_{distmap} = w\sum_{i=1}^{Lch1}\sum_{j=1}^{Lch2}\left|r_{ij} - d_{ij}\right| - \delta_{ij} \quad (10)$$

where r_{ij} is the distance between the residue *i* and *j* in the decoy, d_{ij} is the average distance between residue *i* and *j* in the top templates while δ_{ij} is the standard deviation.

viii) $E_{tcontact}$: A penalty of 1 is assessed when residue *i* and *j* belonging to opposite chains of the complex are found to be in contact ($d_{ij} \le 4.5$ Å) in multiple templates but are not in contact in a given decoy.

ix) E_{dfire}: This potential terms seeks to account for the preferred distance between the Cα atoms of any two pair of amino acids A_i and A_j. Since we are only interested in the interface residues in this case, the range of distance considered is from 0.0 Å to 10.0 Å which was divided into 20 distance bins $r_{i,j}$ of 0.5 Å each. Thus the final potential is given by the equation

$$E_{dfire} = -w\sum_{i=1}^{Lch1}\sum_{j=1}^{Lch2} D(A_i, A_j, r_{ij}) \quad (11)$$

where

$$D(i,j,r) = \begin{cases} -wln\dfrac{N_{obs}(i,j,r)}{\left(\frac{r}{r_{cut}}\right)^{\alpha}\left(\frac{\Delta r}{\Delta r_{cut}}\right)N_{obs}(i,j,r_{cut})}, r < r_{cut} \\ 0 \qquad r < 4 \; or \; r > 12 \end{cases} \quad (12)$$

Here, $D(A_i,A_j,\lambda_{i,j})$ is derived from the protein complex structure library using the DFIRE equation [46]. Nobs is given by the total number of times the Cα atoms of the amino acids A_i and A_j

belonging to different chains of a complex are present in the distance bin $r_{i,j}$ divided by the total number of time the Cα atoms of A_i and A_j are present within 0.0 Å to 10.0 Å of each other. $\Delta r(\Delta r_{cut})$ is the bin width and it is set to 0.5 Angstrom, rcut is 10 Angstroms, and α equals 1.61.

4.2.5 Ranking and refinement for generation of full atomic models

The TACOS simulation is implemented in replica exchange cycles, and the decoy created at the end of each cycle is stored for the five lowest temperature replicas. At the end of the simulations, these decoys are clustered by SPICKER [117] using a global RMSD matrix. The cluster centroids of the ten largest clusters are then selected and the full atomic structure including side-chain atoms is generated using ModRefiner [53]. ModRefiner attempts to optimize the hydrogen bonding network, remove clashes, and impart a general protein like conformation on the final models.

4.2.6 Evaluation

Evaluating the global similarity of a protein model to the native structure is normally sufficient, but for protein interactions the quality of the interface is also important. The critical assessment of protein interactions (CAPRI) is a blind competition to assess the state of the art methods for predicting the structures of protein-protein interactions [40]. CAPRI uses the interface RMSD, ligand RMSD, fraction of native contacts (FNAT), global RMSD, FNAT, and accuracy (ACC) for evaluation purposes. Where RMSD is the root mean squared deviation, FNAT are the percentage of correctly predicted ca-ca interface contacts contained in the model within 10 Angstroms, and ACC is the accuracy of the predicted contacts [40]. The ligand RMSD is calculated after optimal superposition of the native structure onto the model. The CAPRI criteria for an acceptable interface hit requires the model has at least 20% of the correctly predicted interface contacts while having a ligand RMSD less than 10 Angstroms.

A common metric for evaluating protein models is the TM-score. The TM-score is a score that measures the structural similarity between two proteins, and it is used here to compare the quality of the model to native [34]. The TM-score ranges between 0 and 1 with a score greater than 0.5 being highly significant [37]. Additionally a new score, the reciprocal TM-score (rTM-score), is introduced for assessing protein complexes. It's a score that considers the individual model quality of both chains while assessing the correct orientation. Similar to the TM-score it ranges from 0 to 1.

$$rTMscore = \frac{2}{\left(\frac{1}{TMscore_{chain1}}\right) + \left(\frac{1}{TMscore_{chain2}}\right)}$$

4.3 Results and Discussion

4.3.1 Benchmark Set.

The TACOS data set contained 500 non-redundant protein complex with medium and long sequence lengths which were used for the training and testing of the TACOS pipeline. The complexes in this set have a sequence similarity cutoff of 30% for each chain. Over half of the structures are enzymes; 37 are antibody-antigen complexes, 45 are enzyme-inhibitor complexes and the remaining fall into a large assortment of biological classes. The set was split into 150 training and 350 testing sets. The training set contained 92 homodimers and 58 heterodimers while the test set contained 227 homodimers and 123 heterodimers. The average complex in the TACOS benchmark contains 424 residues, with the smallest complex containing 112 residues and the largest containing 712 residues.

4.3.2 Benchmark Target Classification

During benchmarking of monomeric threading target structures are normally classified into three groups: easy, medium and hard. The classification is based on confidence of identified templates determined by various threading programs. Easy targets have multiple high confidence templates identified by threading which allows for consensus based restraints for modeling. Medium targets have one confident threading result, and hard targets have no confident templates for modeling. The individual template confidence levels along with the target classification can be used to assess the quality of the final model.

The classification for protein interactions is more complicated. The templates can still be classified as easy, medium and hard but the confidence scores are not as robust as the monomeric analogues due to possible alternative binding modes. The tertiary sequence to structure similarity is a major component for template recognition for complexes, however groups of protein interactions with similar tertiary components may share little to no similarity regarding the orientation of their respective interactions. A target is classified as ‘easy’ if multiple templates have a COTH alignment score above 2.5 or a SPRING score above 20, as ‘medium’ if only one target has an alignment score above the COTH or SPRING thresholds and hard otherwise. Regarding the

TACOS benchmark set 290 are classified as easy, 25 medium and 35 hard. However, only 162 of the 290 easy targets have templates that provide some level of correct information regarding the correct orientation of protein chains. The training set consisted of 70 easy, 23 medium and 57 hard targets.

4.3.3 Template Selection and Interface Restraints:

Currently there are three classes of bioinformatics based approaches for identifying and modeling protein-protein interactions: dimer threading (i.e. COTH), monomeric threading and oligomer mapping (i.e. SPRING) and docking with homology models (i.e. ZDOCK) [82]. In an approach analogous to meta threading, which is a consensus based approach where multiple complementary threading methods are used for monomeric structure prediction, we evaluated the potential for combining methods for quaternary structure prediction. Using updated protocols and databases for SPRING and COTH, we ran them on a benchmark set of 1830 structures at 30% sequence identity threshold to evaluate their performances, Figure 4.2. COTH identified 948 and SPRING identified 953 hits within the top 10 structures. Combining the top 5 structures from SPRING and COTH, threading were able to identify 1046 hits with an overall improvement of 5.3 percent compared to COTH and 5.1 percent compared to SPRING.

The consensus threading was then evaluated against ZDOCK using the ZDOCK benchmark set of 99 complexes. ZDOCK was ran with two sets of starting structures: ZDOCK was given models generated by I-TASSER and ZDOCK NATIVE used the native unbound monomeric constituents. Using the native structure, ZDOCK at most sequence identity thresholds was able to outperform COTH/SPRING, but threading was still able to identify interfaces not predicted by ZDOCK. Given monomer models threading outperformed ZDOCK at all sequence thresholds. ZDOCK using models and native structures identified 6 and 25 targets respectively, whereas threading identified 7, 21 and 27 when homologous templates are excluded at 30, 50, and 70 percent sequence identity thresholds. The combination of docking with models and threading identified 12, 25 and 31 of the targets at the respective sequence thresholds. This showed promise in combining docking with threading for the modeling protein complex.

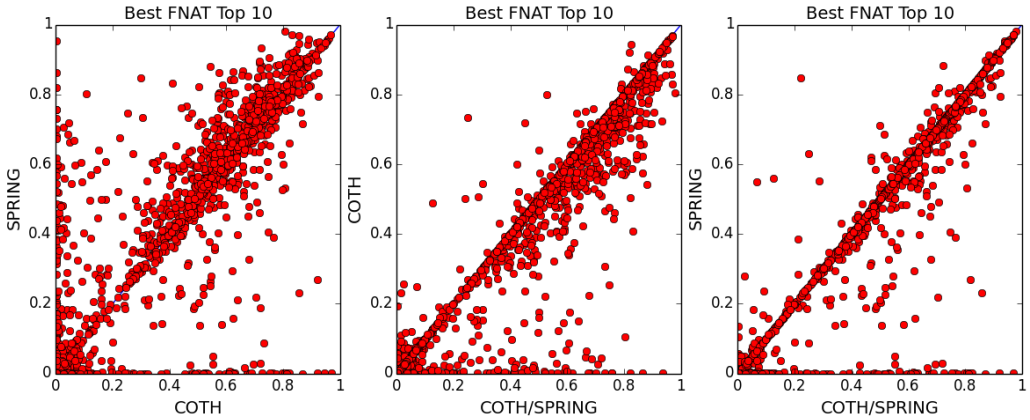


Figure 4.2 COTH and SPRING consensus threading and comparison by FNAT. The left plot is a comparisons of the top ten templates generated by COTH and SPRING. The middle plot is the top ten templates by COTH compared to the top 5 from COTH and SPRING. The plot on the right is the top ten generated by SPRING compared to the top 5 in COTH and SPRING.

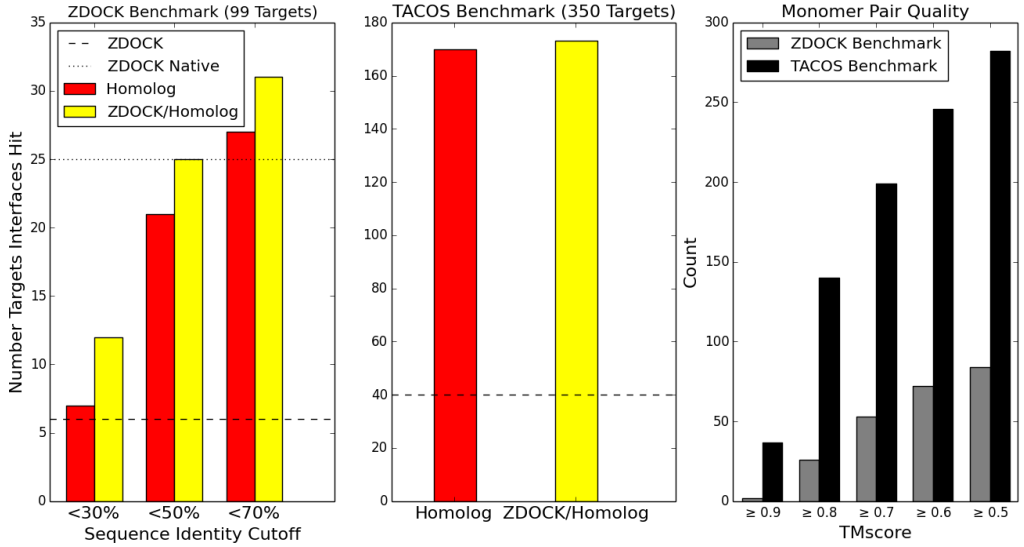


Figure 4.3 Comparison of threading compared to docking using ZDOCK. The ZDOCK benchmark has 99 target structure where ZDOCK NATIVE starts with the native constituents for docking. ZDOCK, using I-TASSER models, is compared to threading at three different sequence identity thresholds. The second plot is the TACOS benchmark containing 350 target structures. ZDOCK is given the best I-TASSER model in the top ten determined by TM-score, and threading excludes all templates with a sequence identity greater than or equal to 30%. The CAPRI criteria is used to designate a hit. The third plot contains the complex targets constituent quality generated by I-TASSER. The Y-axis is a count of the targets where both monomer models for a complex have TM-score's above the thresholds on the X-axis.

The comparison of protein docking with threading was repeated using I-TASSER models of the individual units from the 350 TACOS benchmark set, Figure 4.3. ZDOCK was given the best monomer, by TM-score, modeled by I-TASSER at a 30% sequence identity threshold within the top 10. ZDOCK identified 43 of the targets complexes having a FNAT of at least 20% and only

34 of the complex model meet the CAPRI criteria. Whereas COTH/SPRING identified 170 targets with an FNAT of at least 20%. Combining the two increased coverage to 173 targets which was less than a percent increase.

4.3.4 Energy Function Correlation with Native Structure

To partly assess the quality of our force field, the Pearson's correlation coefficient was calculated between the energy and the TM-score of decoys to the structure of the native complex. The correlation can have variation depending on the quality of the complex and monomeric templates. When the correct monomeric and complex templates are correctly identified with a single binding orientation the correlation average is 0.791. For easy cases where there are multiple binding orientations predicted by threading, the correlation average drops to 0.748. Cases where the complex templates are wrong but the monomeric templates are correct the correlation average is 0.751. Finally when both monomeric and complex threading fails to identify the correct templates the correlation drops to 0.505.

In Figure 4.4, we show 3 representative examples of each modeling category (easy, medium, and hard) showing the correlation between TM-score and energy. In general, the decoy set for the easy cases spanned a larger TM-score range compared to the medium and hard cases and showed increased sampling in the higher TM-score ranges. The increased specificity of the TACOS force-field towards native-like structures in the easy and medium cases can therefore be attributed to the agreement in restraints between the monomer and complex templates.

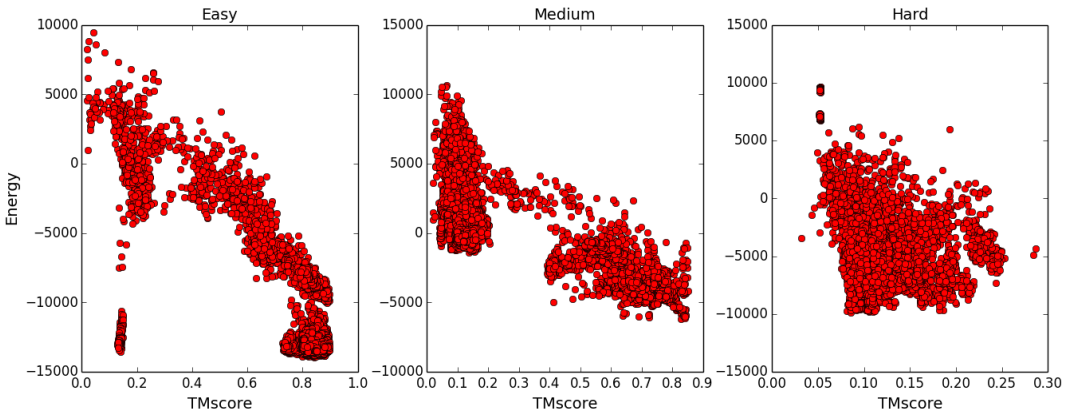


Figure 4.4 Correlation of TACOS energy with TM-score. Three representative examples, one each for easy (left), medium (middle) and hard (right) modeling targets, are shown, which illustrate the correlation between energy and TM-score for each category.

4.3.5 Modelling Protein Complexes with TACOS.

The primary goal of homology modeling is to generate full length full atom models that are closer to native than the starting template. It is important to check if the pipeline is able to generate models that are of higher quality than the starting template trajectory identified by COTH and SPRING. Additionally, it is necessary to compare it to other methods that can generate protein complex models.

Overall TACOS made improvements to the global quality of the protein models. When considering the global similarity using the TM-score, and then new rTM-score, which is more sensitive to orientational differences, TACOS was consistently able to make improvements compared to the starting templates. For the top ranked structure 300/350 structures were improved when being evaluated by the TM-score. Regarding the rTM-score 240/350 models showed improvement. Note that the rTM-score will be close to or at zero for orientations that are far away from native. Out of the 350 structure 149 had rTM-scores equal to zero for the highest ranked template. When comparing the best structures in the top 10, 298/350 TM-score and 273/350 rTM-score showed improvement compared to the best template. The interface I-RMSD for templates is much lower than the I-RMSD of the TACOS model. Part of this is due to the templates being fragmented, and TACOS attempting to rebuild those regions *ab initio*. That being said, the other measurement and assessment scores values can be affected by the varying lengths between the template and full length model. COTH and SPRING templates often contain gap regions that are unable to contribute to the overall TM-score and rTM-score. Thus for comparison purposes, full length models were also built by the homology modeling program MODELLER [54] which can accept quaternary level templates.

MODELLER builds full length models from the template with only slight deviations from the starting CA positions. The overall global similarity to native is usually improved when using MODELLER. Comparing the first ranked structures of the modeling programs, TACOS energy potential and clustering is able to correctly rank 108 model orientations compared to threading/MODELLERs 77. However, in Figure 4.5, there are three PDB target structures where MODELLER appears to generate rank 1 models with significantly better interface structure than TACOS: 1a22, 1jag and 1osg. Regarding 1a22, dimer threading results are dominated by high

confidence templates with very similar orientations that are not close to the native orientation. The second model built by TACOS contains the correct orientation. The second protein 1jag is actually correct. TACOS ranks another correct alternative binding mode that was not in the 1jag file; the pdb file 1jag was superseded by 2ocp which contains the rank 1 TACOS orientation. Finally 1osg, has the correct orientation but the structure of the interface is poor. In general, TACOS has improved global and interface structure when compared to MODELLER. For the highest rank models, in terms of the TM-score TACOS outperforms MODELLER 260/350 cases and has a lower interface RMSD 213/350 cases. Figure 4.5 shows a head to head comparison of the two programs. The series of data points on the X-axis for the FNAT, ACC and the rTM-score plots highlights TACOS energy function and clustering capabilities regarding reranking the models when the first ranked template contains the wrong orientation, but the correct orientation is contained in the template set.

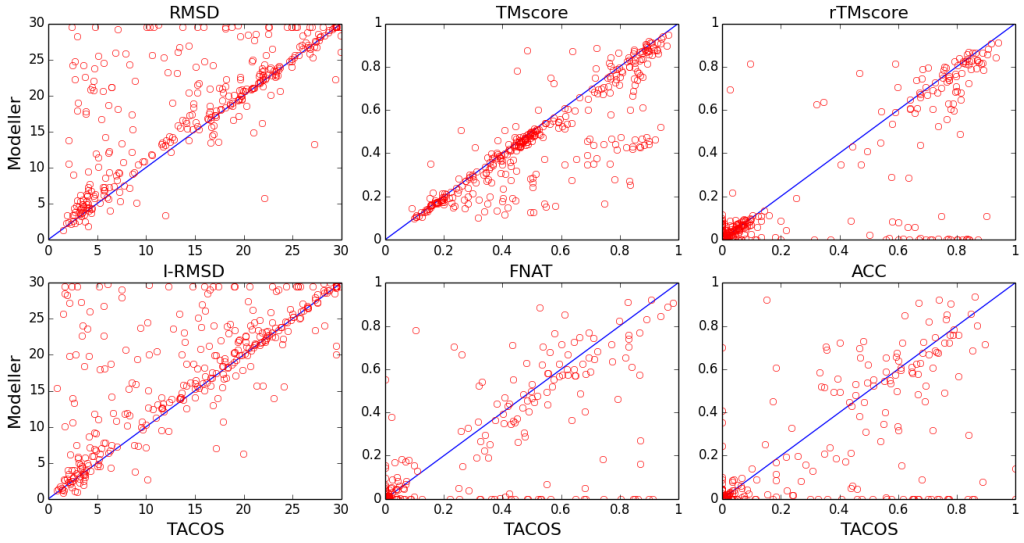


Figure 4.5 Set of six scatter plots showing benchmark results of TACOS on a test set of 350 proteins compared to MODELLER for the 1st ranked structure. The six scores are Root Mean Squared Deviation (RMSD), TM-score, rTM-score, Interface-RMSD (I-RMSD), Fraction of Correctly Interface Contacts (FNAT), and the Accuracy of the predicted interface contacts (ACC). For TM-score, rTM-score, FNAT and ACC, points below the diagonal show better performance by TACOS. For RMSD and I-RMSD, points above the diagonal show better performance by TACOS.

On average TACOS has better assessment scores than MODELLER regarding the top ranked structure and the best in top 10. Table 4.1 and Table 4.2 contain summary information of the two categories respectively. Regarding the best in top 10, TACOS outperforms the other methods in terms of the global and interface structure at a p-value threshold of 10^{-3} using the Wilcoxon signed-

rank test. For many of the large and medium sized complexes, increasing the FNAT is not an issue of the orientation but the quality of the monomeric constituents. Although the two dimensional contact maps of the best TACOS model is only marginally better than the best MODELLER model the three dimensional structure of the interface is generally better than MODELLER's. Additionally it was noticed that TACOS tends to maintain alpha helix structure at the interface compared to beta sheet. Prediction of alpha helices tends to be higher than for beta sheets, also helices often form individually whereas beta sheets occur in groups, which requires accurate long distance contacts.

Method	Median I-RMSD	FNAT	ACC	RMSD	TM-score	rTM-score	Hits
ZDOCK I-TASSER	18.139	0.034	0.053	18.306	0.469	0.072	11
Dimer Threading	16.490	0.153	0.180	17.241	0.452	0.201	---
MODELLER	17.730	0.161	0.166	18.471	0.464	0.207	71
TACOS	14.732	0.210	0.227	15.343	0.530	0.274	108

Table 4.1 Comparison of TACOS against controls for the rank 1 structures.

Method	Median I-RMSD	FNAT	ACC	RMSD	TM-score	rTM-score	Hits
ZDOCK I-TASSER	13.602	0.090	0.137	14.093	0.515	0.173	34
Dimer Threading	8.032	0.311	0.350	10.920	0.551	0.380	---
MODELLER	10.838	0.331	0.322	12.941	0.587	0.399	137
TACOS	9.325	0.353	0.323	11.032	0.622	0.414	154

Table 4.2 Comparison of TACOS against controls for the top ten structures.

In Figure 4.6, three high resolution structures with different topologies are presented and compared to the best ranked template model with the correct orientation. Human glutathionine transferases, 16gs, is the first example, it contains a majority α-helix homodimeric complex. The template had a global RMSD of 4.13 Angstroms of which TACOS was able to refine to 2.35 Angstroms. The 1c3h complex is a serum protein with beta sheet topology involved in metabolism. TACOS refined the template from 2.95 to 2.4 Angstroms. Finally the hetero complex 1go3 is a part of the eukaryotic RNA polymerase and it contains recognition sites for RNA motifs. TACOS built a 2.06 Angstroms compared to 3.77 Angstroms from the template.

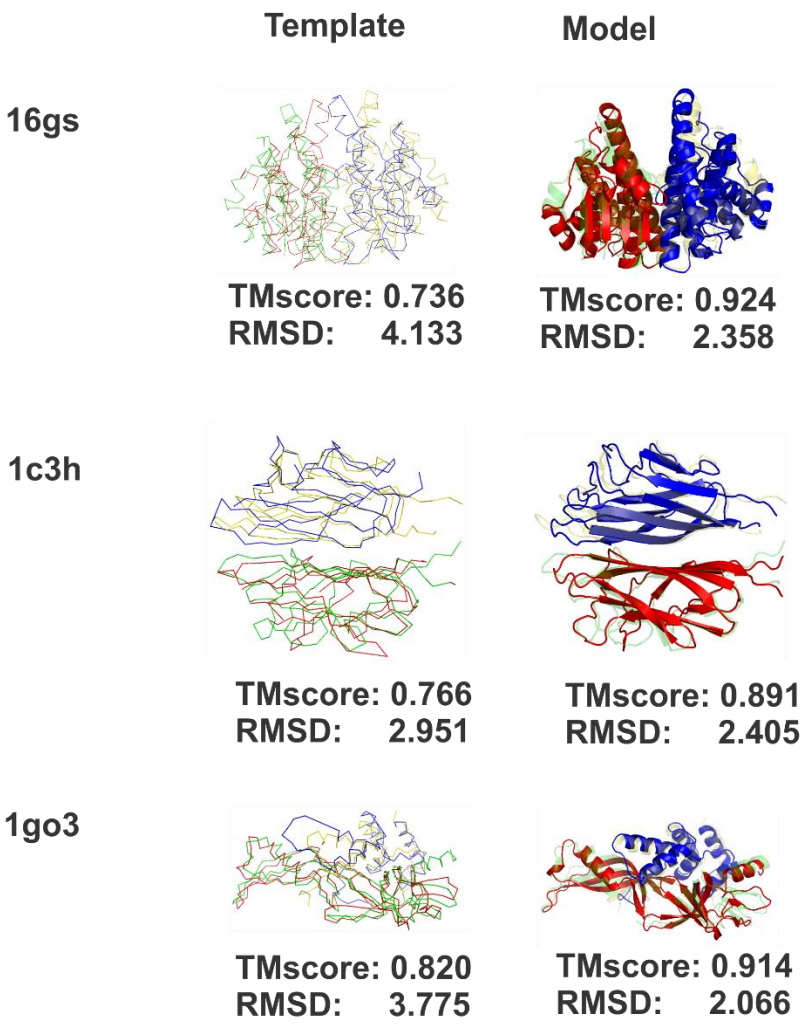


Figure 4.6 Near-native models built by TACOS. Plot showing examples of TACOS modeling for both homo- and heterodimers. The predicted models are shown in red and slate for chain A and B; the native structure shown in transparent green and yellow is superimposed onto the model structure.

4.3.6 Information quality required for TACOS models against different classes.

TACOS uses information to guide protein structure prediction. The quantity and quality of the information has a major impact on the quality of the final models. Quality is determined from threading and quantity is dependent on the structural diversity and size of the protein library which is limited for quaternary structure. Consensus of predicted contacts from the template alignments are considered as important conserved regions for binding and are more heavily restrained in the simulation than those without.

Unlike monomer structure prediction, there can be multiple alternate quaternary structures. Most cases often contain only one template per orientation, or identify multiply confident templates with the incorrect orientation which occurs for 128 of the 290 targets with confident templates. This makes it difficult to generate statistics on interface contacts. The majority of medium to high resolution structures generated by TACOS occurs when there are at least five identified templates with similar orientations that have at least thirty percent of the correct interface contacts.

Protein-protein interactions are often grouped into homodimers and heterodimers. Sixty percent of the complexes in the PDB are homodimers. Eighty five percent consists of complexes that are homodimers or complexes that diverged from homodimers, i.e. the interacting chains are structurally similar. The last fifteen percent consists of heterodimers where the interacting chains do not share similar topologies. This distribution results in template(s) for modeling being more readily available for modeling homodimers than heterodimers. In our benchmark set, 117 homodimers and 37 heterodimers meet the CAPRI criteria.

4.4 Conclusion

Learning from the experiences gathered in the field of protein structure prediction, we developed a new algorithm, TACOS, to predict the structure of protein-protein complex structures from sequence alone. TACOS uses a hybrid comparative modeling-*ab initio* approach which first identifies putative templates from a non-redundant protein complex structure library by COTH/SPRING threading. Simultaneously, TACOS uses LOMETS singe chain threading to generate intra-chain restraints for the individual subunits of the protein complex. In the second step, TACOS uses a lattice-based replica exchange Monte-Carlo simulation to build *ab initio* the template un-aligned regions and further refine the template aligned regions through rigid body moves. TACOS also seeks to search the ideal orientation for the component chains of the complex with respect to each other by using a newly designed inter-chain movement which implements a predefined discrete set of rotation and translation moves to produce the best fit at the interface. The TACOS simulation is driven by an energy function composed of intra-chain template based restraints from LOMETS, inter-chain distance and contact restraints from templates identified by SPRING and COTH threading, and knowledge based terms. While some of the knowledge based potential terms were adapted from the well-known monomeric structure prediction algorithm I-

TASSER, other newly derived inter-chain potential terms were added to recreate the uniqueness of the protein-protein interface in a course-grained fashion.

The TACOS simulation parameters were trained on a non-redundant set of 150 dimeric protein structures and tested on an independent 350 protein dataset. No homologous templates with ≥ 30% sequence identity to the query were used for either training or testing. Despite this, TACOS performs well and can predict full length structures with the same basic fold as the target when a confident template is identified. In 44% of the cases the final TACOS models meets the CAPRI requirements for an acceptable solution. TACOS also tends to make improvements in the starting aligned regions.

Another important observation noted was that TACOS performed better overall for certain structural classes of protein complexes. Generally complexes with the majority of secondary structure as alpha helixes were the easiest to model. Targets that had substantial interface structure consisting of beta sheets tended to be more difficult to generate higher resolution structures. The performance of TACOS is greater for homodimers than for heterodimers due to the depth of homodimer templates in the PDB.

TACOS thus represents one of the first algorithms designed to predict the structure of dimeric protein complexes given the sequence alone. Importantly, TACOS capabilities will improve with time as the depth of the protein complex structure library will continue to grow in the years to come. Also, since TACOS models both chains simultaneously while taking into account their relative orientation, it can potentially model the conformational changes brought about by complex formation, a task that is difficult to do currently with rigid body docking algorithms. The TACOS algorithms can be used freely by the academic community through the web-server made available at http://zhanglab.ccmb.med.umich.edu/TACOS/.

CHAPTER 5. Conclusions

Predicting the structure and function of proteins from sequence is a forefront problem in computational biology. Over the last few years there have been many advances for predicting tertiary structures using primarily two approaches: ab initio (free modeling) and template based modelling with the latter showing promise for generating high quality structural models [171, 178]. While the first approach attempts to model a protein solely using an energy potential, the second approach incorporates spatial restraints from identified homologous proteins. Although the majority of recent protocols have focused on predicting protein tertiary folds, the function of proteins vastly occurs at the quaternary level where multiple protein chains assemble into one large complex. The most common approach to assembling complexes is docking, where a 6-dimensional search is performed to identify the interface, but in many cases the correct structure is not identified [71, 127]. Fortunately in recent years the number of available quaternary structures has grown substantially allowing template based modeling to be incorporated into quaternary structure prediction [16]. Here the state of the art template based modeling pipeline I-TASSER is extended to handle quaternary structure. In this dissertation, I developed methods for the three main challenges regarding the extension of template based modeling to quaternary structure prediction: identifying homologs in the PDB, increasing the number of interfaces in the dimer library, and refining the initial model towards the correct structure.

The first and most essential component of template based modeling is correctly identifying a homolog to the target in the PDB Library. I have developed a threading method, SPRING, to address the first requirement. SPRING uses monomer threading to identify protein complexes. The monomer protein structures are connected to its binding partners obtained from the PDB. A precalculated look up table assesses whether a pair of monomeric templates are structurally similar to an interacting pair of proteins in the Dimer Library. After a set of dimer templates are identified, the dimer template models are ranked by the tertiary similarity to the top ranked monomer model,

the threading Z-score and an interface potential calculated with DFIRE [88]. Similar to the monomer threading case for deep distant homology searches, no single algorithm can recognize all of the homologs. Hence, several algorithms are run in a meta threading approach to identify protein homologs. Thus, I combined SPRING with a previously created threading program, COTH, in order to recognize lower homology quaternary templates. COTH, an extension of the MUSTER algorithm with an incorporated interface profile, treats dimers as artificial monomers during the alignment. Using updated protocols and databases (incorporating all biomolecules) for SPRING and COTH, I ran them on a benchmark set of 1830 structures at 30% sequence identity threshold to evaluate their performances. COTH and SPRING identified 948 and 953 hits respectively, within the top 10 structures. Combining the identified structures from SPRING and COTH, threading were able to identify 1046 hits with an overall improvement of 5.3% and 5.1% compared to COTH and SPRING, respectively. Despite the improvement in template identification, the consensus program misses 42.8% of the target set due to the lack of representation in the PDB.

The major limitation of the template based modeling for quaternary structures is the scarcity of suitable templates in the Dimer Library. Even with a perfect quaternary threading alignment program the limited size of the dimeric information in the PDB would prevent most targets from being modelled using template based methods. Here I assessed the current limitations of the dimer library and threading algorithms as well as incorporating the interface information between domains within single polypeptide chains in order to boost the dimer library. Structural alignment under benchmark exclusion settings was only able to identify a suitable template 48.7% of the time due to the sparsity of the dimer library. The limited number of dimer interactions restricts template based prediction and modeling of dimer interactions. However, it has been shown that the interface between domains within and between protein chains have structural similarity. Incorporating domain interfaces into the modeling of protein complexes improved template recognition overall by 1.7% and by 40% for heterodimers. The incorporation of domain orientations within the dimer library allows for a larger set of protein interactions to be confidently modelled.

Although threading can provide structural frameworks of dimeric interactions they often only produce partial Cα trace structures. To this effect, I developed a sequence to full atom structure

pipeline, TACOS. Starting from fragmented templates identified from SPRING/COTH initial full length models are built by connecting fragments using a random walk. The TACOS conformational search uses a hybrid comparative modeling-*ab initio* approach to generate a model that is closer to the native like structure. A combination of monomeric spatial restraints from LOMETS and interface restraints from COTH/SPRING are merged with physical potentials inherited from I-TASSER along with its replica exchange Monte Carlo search engine to refine the initial model towards a more native topology. Given a confident set of templates, the energy score on average has a correlation of 0.79 with the TM-score. The TACOS pipeline was tested against two independent methods: homology models built by MODELLER and I-TASSER models of the constituent chains docked by ZDOCK. TACOS outperformed both methods in terms of model ranking, global and interface evaluation scores.

Most high order biological processes occur at the quaternary level, which necessitates determining the final assembly of protein chains fundamental in order to understand cell physiology. Structural information around a proteins interface can provide significant insight into which residues are essential for binding and which residues are most susceptible to disease related mutations. Furthermore query alignments to structures in the PDB can provide further validation of interactions determined by large scale low confident experiments. Due to the importance of determining and modeling protein quaternary structure, the developed methods were integrated into a webserver to provide easy access to the general scientific community. The methods presented in this dissertation are available as three separate webservers hosted at http://zhanglab.ccmb.med.umich.edu/.

5.1 Future Directions

5.1.1 Improving the TACOS Pipeline

As shown in chapter two there is a gap between the templates that are readily available in the PDB library and those identified by SPRING. Currently, SPRING only uses HHsearch for its monomeric search. However, previous research has shown [14], meta-threading programs consistently outperform the constituent threading algorithms. Here, LOMETS can simply be exchanged for HHsearch with regard to identifying monomeric templates for SPRING. Additionally the precalculated lookup table was created using a simple blast alignment which can

easily be replaced with structural alignment using TM-align. This should further allow for mapping monomers to dimeric templates with low homology.

Often the dimer template frameworks represent a small portion of the whole dimer complex causing extensive loop regions. An alternative is to try to structurally dock the top monomer into the image using structural alignment, but this may lead to heavy clashes or loss of interface structure. Another alternative is to try to dock it into the framework, but this reverts the assembly back into the fundamental problem that TACOS is trying to circumvent. A quick alternative is to use I-TASSER models as restraints in the creation of an initial model. A weighted RMSD, to emphasize restraints on consensus regions, with a simple interface potential can potentially quickly construct strong starting conformations before the template refinement by the full TACOS simulation. Additionally, the other forms of docking models into dimer frameworks can be incorporated and ranked by the TACOS energy to recognize strong starting conformations before the standard simulation takes place.

5.1.2 Genome Wide Modeling of *E. coli*

The entire *E. coli* genome was modelled by the structure prediction pipelines I-TASSER and QUARK. Using SPRING 46,033 dimer complexes were predicted in the *E. coli*. Using the pregenerated data from I-TASSER and QUARK [179], TACOS can be used to refine and assemble the predicted protein dimeric interactions. The models from I-TASSER and QUARK can be used in the initial stage of dimer threading to identify distant homology complexes and improve recognition of conserved binding residues at the interface.

5.1.3 Extending TACOS to Higher Order Quaternary Assemblies

Extending the threading and modeling protocols to higher order structures is technically a straight forward process. SPRING would only require threading additional sequences and searching for a single PDB files that contain all of the constituent templates. The virtual bond trick used in TACOS to treat the dimer as an artificial monomer can be extended to as many protein chains as needed. The main limitations would be the size of the system being modelled and the limited number of large quaternary structures in the PDB.

5.1.4 Structural Models as a Feature for Prediction of Interface Mutation Stability

Many disease mutations in the human genome have been found to be present at the interface between proteins, suggesting many disease disrupt protein interaction networks [180]. A recent method in the Zhang Lab was developed to predict the effect of point mutations on the stability of protein interfaces [165]. The prediction accuracy should be improved by incorporating structural models as features into the prediction. Incorporating high quality structural models into function prediction generally results in better prediction algorithms.

References

1. Bank, P.D., *Protein Data Bank.* Nature New Biol, 1971. **233**: p. 223.
2. Berman, H.M., et al., *The protein data bank.* Nucleic acids research, 2000. **28**(1): p. 235-242.
3. Abad-Zapatero, C., *Notes of a protein crystallographer: on the high-resolution structure of the PDB growth rate.* Acta Crystallographica Section D: Biological Crystallography, 2012. **68**(5): p. 613-617.
4. Research Collaboratory for Structural Bioinformatics. *RCSB Protein Data Bank*. January 2016; Available from: www.rcsb.org.
5. Wüthrich, K., *Protein structure determination in solution by NMR spectroscopy.* Journal of Biological Chemistry, 1990. **265**(36): p. 22059-22062.
6. Kim, W.K., et al., *The many faces of protein-protein interactions: A compendium of interface geometry.* PLoS Comput Biol, 2006. **2**(9): p. e124.
7. Christof Winter. *Structural Classification of Protein-Protein Interfaces*. January 2016; Available from: www.scoppi.org.
8. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* Journal of molecular biology, 1995. **247**(4): p. 536-540.
9. Orengo, C.A., et al., *CATH–a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-1109.
10. Zhang, Y., et al., *On the origin and highly likely completeness of single-domain protein structures.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(8): p. 2605-2610.
11. Xu, Y., D. Xu, and H.N. Gabow, *Protein domain decomposition using a graph-theoretic approach.* Bioinformatics, 2000. **16**(12): p. 1091-1104.
12. Burley, S.K., et al., *Structural genomics: beyond the human genome project.* Nature genetics, 1999. **23**(2): p. 151-157.
13. Zhang, Y., *Progress and challenges in protein structure prediction.* Current opinion in structural biology, 2008. **18**(3): p. 342-348.
14. Wu, S. and Y. Zhang, *LOMETS: a local meta-threading-server for protein structure prediction.* Nucleic acids research, 2007. **35**(10): p. 3375-3382.
15. Zhang, Y., *Protein structure prediction: when is it useful?* Current opinion in structural biology, 2009. **19**(2): p. 145-155.
16. Kundrotas, P.J., et al., *Templates are available to model nearly all complexes of structurally characterized proteins.* Proceedings of the National Academy of Sciences, 2012. **109**(24): p. 9438-9441.
17. Gao, M. and J. Skolnick, *Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected.* Proceedings of the National Academy of Sciences, 2010. **107**(52): p. 22517-22522.

18. Garma, L., et al., *How many protein-protein interactions types exist in nature.* PloS one, 2012. **7**(6): p. 13.
19. Dayhoff, M.O. and R.M. Schwartz. *A model of evolutionary change in proteins*. in *In Atlas of protein sequence and structure*. 1978. Citeseer.
20. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proceedings of the National Academy of Sciences, 1992. **89**(22): p. 10915-10919.
21. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* Journal of molecular biology, 1970. **48**(3): p. 443-453.
22. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* Journal of molecular biology, 1981. **147**(1): p. 195-197.
23. Gotoh, O., *An improved algorithm for matching biological sequences.* Journal of molecular biology, 1982. **162**(3): p. 705-708.
24. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison.* Proceedings of the National Academy of Sciences, 1988. **85**(8): p. 2444-2448.
25. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-410.
26. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic acids research, 1997. **25**(17): p. 3389-3402.
27. Gribskov, M., A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins.* Proceedings of the National Academy of Sciences, 1987. **84**(13): p. 4355-4358.
28. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Molecular biology and evolution, 1987. **4**(4): p. 406-425.
29. Clustal, W., *improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice Thompson, Julie D.; Higgins, Desmond G.; Gibson, Toby J.* Nucleic Acids Research, 1994. **22**(22): p. 4673-80.
30. Panchenko, A.R., *Finding weak similarities between proteins by sequence profile comparison.* Nucleic acids research, 2003. **31**(2): p. 683-689.
31. Henikoff, S. and J.G. Henikoff, *Position-based sequence weights.* Journal of molecular biology, 1994. **243**(4): p. 574-578.
32. Kabsch, W., *A solution for the best rotation to relate two sets of vectors.* Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, 1976. **32**(5): p. 922-923.
33. Kabsch, W., *A discussion of the solution for the best rotation to relate two sets of vectors.* Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, 1978. **34**(5): p. 827-828.
34. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins: Structure, Function, and Bioinformatics, 2004. **57**(4): p. 702-710.
35. Moult, J., et al., *Critical assessment of methods of protein structure prediction (CASP): round III.* Proteins: Structure, Function, and Bioinformatics, 1999. **37**(S3): p. 2-6.
36. Siew, N., et al., *MaxSub: an automated measure for the assessment of protein structure prediction quality.* Bioinformatics, 2000. **16**(9): p. 776-785.

37. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score= 0.5?* Bioinformatics, 2010. **26**(7): p. 889-895.
38. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic acids research, 2005. **33**(7): p. 2302-2309.
39. Levy, E.D., et al., *Assembly reflects evolution of protein complexes.* Nature, 2008. **453**(7199): p. 1262-1265.
40. Lensink, M.F. and S.J. Wodak, *Docking and scoring protein interactions: CAPRI 2009.* Proteins: Structure, Function, and Bioinformatics, 2010. **78**(15): p. 3073-3084.
41. Mukherjee, S. and Y. Zhang, *MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming.* Nucleic acids research, 2009. **37**(11): p. e83-e83.
42. Gao, M. and J. Skolnick, *iAlign: a method for the structural comparison of protein–protein interfaces.* Bioinformatics, 2010. **26**(18): p. 2259-2265.
43. Cheng, S., Y. Zhang, and C.L. Brooks, *PCalign: a method to quantify physicochemical similarity of protein-protein interfaces.* BMC bioinformatics, 2015. **16**(1): p. 33.
44. Cornell, W.D., et al., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.* Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.
45. Karplus, M., *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.* J Comput Chem, 1983. **4**: p. 187217.
46. Zhou, H. and Y. Zhou, *Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction.* Protein science, 2002. **11**(11): p. 2714-2726.
47. Zhang, Y., A. Kolinski, and J. Skolnick, *TOUCHSTONE II: a new approach to ab initio protein structure prediction.* Biophysical journal, 2003. **85**(2): p. 1145-1164.
48. Heffernan, R., et al., *Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.* Scientific reports, 2015. **5**.
49. Wu, S., A. Szilagyi, and Y. Zhang, *Improving protein structure prediction using multiple sequence-based contact predictions.* Structure, 2011. **19**(8): p. 1182-1191.
50. Xu, D. and Y. Zhang, *Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field.* Proteins: Structure, Function, and Bioinformatics, 2012. **80**(7): p. 1715-1735.
51. Simons, K.T., et al., *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* Journal of molecular biology, 1997. **268**(1): p. 209-225.
52. Zhang, J., Y. Liang, and Y. Zhang, *Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling.* Structure, 2011. **19**(12): p. 1784-1795.
53. Xu, D. and Y. Zhang, *Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization.* Biophysical journal, 2011. **101**(10): p. 2525-2534.
54. Šali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* Journal of molecular biology, 1993. **234**(3): p. 779-815.
55. Meier, A. and J. Söding, *Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling.* PLoS Comput Biol, 2015. **11**(10): p. e1004343.

56. Jones, D.T., et al., *PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.* Bioinformatics, 2012. **28**(2): p. 184-190.
57. Jones, D.T., et al., *MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins.* Bioinformatics, 2015. **31**(7): p. 999-1006.
58. Savojardo, C., et al., *BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming.* Bioinformatics, 2013: p. btt555.
59. Skwark, M.J., A. Abdel-Rehim, and A. Elofsson, *PconsC: combination of direct information methods and alignments improves contact prediction.* Bioinformatics, 2013. **29**(14): p. 1815-1816.
60. Kaján, L., et al., *FreeContact: fast and free software for protein contact prediction from residue co-evolution.* BMC bioinformatics, 2014. **15**(1): p. 85.
61. Mintseris, J., et al., *Integrating statistical pair potentials into protein complex prediction.* Proteins: Structure, Function, and Bioinformatics, 2007. **69**(3): p. 511-520.
62. Lu, H., L. Lu, and J. Skolnick, *Development of unified statistical potentials describing protein-protein interactions.* Biophysical journal, 2003. **84**(3): p. 1895-1901.
63. Liu, S., et al., *A physical reference state unifies the structure-derived potential of mean force for protein folding and binding.* Proteins: Structure, Function, and Bioinformatics, 2004. **56**(1): p. 93-101.
64. Zielenkiewicz, P. and A. Rabczenko, *Protein-protein recognition: Method of finding complementary surfaces of interacting proteins.* Journal of theoretical biology, 1984. **111**(1): p. 17-30.
65. Walls, P.H. and M.J. Sternberg, *New algorithm to model protein-protein recognition based on surface complementarity: Applications to antibody-antigen docking.* Journal of molecular biology, 1992. **228**(1): p. 277-297.
66. Greer, J. and B.L. Bush, *Macromolecular shape and surface maps by solvent exclusion.* Proceedings of the National Academy of Sciences, 1978. **75**(1): p. 303-307.
67. Pierce, B., W. Tong, and Z. Weng, *M-ZDOCK: a grid-based approach for Cn symmetric multimer docking.* Bioinformatics, 2005. **21**(8): p. 1472-1478.
68. Tsai, J., et al., *An improved protein decoy set for testing energy functions for protein structure prediction.* Proteins: Structure, Function, and Bioinformatics, 2003. **53**(1): p. 76-87.
69. Deng, H., Y. Jia, and Y. Zhang, *3DRobot: automated generation of diverse and well-packed protein structure decoys.* Bioinformatics, 2015: p. btv601.
70. Hwang, H., et al., *Protein–protein docking benchmark version 4.0.* Proteins: Structure, Function, and Bioinformatics, 2010. **78**(15): p. 3111-3114.
71. Anishchenko, I., et al., *Protein models: The Grand Challenge of protein docking.* Proteins: Structure, Function, and Bioinformatics, 2014. **82**(2): p. 278-287.
72. Edgar, R.C. and K. Sjölander, *A comparison of scoring functions for protein sequence profile alignment.* Bioinformatics, 2004. **20**(8): p. 1301-1308.
73. Wang, G. and R.L. Dunbrack, *Scoring profile-to-profile sequence alignments.* Protein Science, 2004. **13**(6): p. 1612-1626.
74. Haussler, D., A. Krogh, and K. Sjölander. *Protein modeling using hidden Markov models: Analysis of globins*. in *System Sciences, 1993, Proceeding of the Twenty-Sixth Hawaii International Conference on*. 1993. IEEE.

75. Karplus, K., C. Barrett, and R. Hughey, *Hidden Markov models for detecting remote protein homologies.* Bioinformatics, 1998. **14**(10): p. 846-856.
76. Söding, J., *Protein homology detection by HMM–HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-960.
77. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-170.
78. Wu, S. and Y. Zhang, *MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information.* Proteins: Structure, Function, and Bioinformatics, 2008. **72**(2): p. 547-556.
79. Vajda, S. and C.J. Camacho, *Protein–protein docking: is the glass half-full or half-empty?* Trends in biotechnology, 2004. **22**(3): p. 110-116.
80. Aloy, P., M. Pichaud, and R.B. Russell, *Protein complexes: structure prediction challenges for the 21 st century.* Current opinion in structural biology, 2005. **15**(1): p. 15-22.
81. Russell, R.B., et al., *A structural perspective on protein–protein interactions.* Current opinion in structural biology, 2004. **14**(3): p. 313-324.
82. Szilagyi, A. and Y. Zhang, *Template-based structure modeling of protein–protein interactions.* Current opinion in structural biology, 2014. **24**: p. 10-23.
83. Lu, L., H. Lu, and J. Skolnick, *MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading.* Proteins: Structure, Function, and Bioinformatics, 2002. **49**(3): p. 350-364.
84. Kundrotas, P.J., M.F. Lensink, and E. Alexov, *Homology-based modeling of 3D structures of protein–protein complexes using alignments of modified sequence profiles.* International journal of biological macromolecules, 2008. **43**(2): p. 198-208.
85. Aloy, P., et al., *Structure-based assembly of protein complexes in yeast.* Science, 2004. **303**(5666): p. 2026-2029.
86. Mukherjee, S. and Y. Zhang, *Protein-protein complex structure predictions by multimeric threading and template recombination.* Structure, 2011. **19**(7): p. 955-966.
87. Zhang, Q.C., et al., *PrePPI: a structure-informed database of protein–protein interactions.* Nucleic acids research, 2012: p. gks1231.
88. Guerler, A., B. Govindarajoo, and Y. Zhang, *Mapping monomeric threading to protein–protein structure prediction.* Journal of chemical information and modeling, 2013. **53**(3): p. 717-725.
89. Sinha, R., P.J. Kundrotas, and I.A. Vakser, *Docking by structural similarity at protein-protein interfaces.* Proteins: Structure, Function, and Bioinformatics, 2010. **78**(15): p. 3235-3241.
90. Keskin, O., R. Nussinov, and A. Gursoy, *PRISM: protein-protein interaction prediction by structural matching*, in *Functional Proteomics*. 2008, Springer. p. 505-521.
91. Ogmen, U., et al., *PRISM: protein interactions by structural matching.* Nucleic acids research, 2005. **33**(suppl 2): p. W331-W336.
92. Yang, J., A. Roy, and Y. Zhang, *Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment.* Bioinformatics, 2013. **29**(20): p. 2588-2595.
93. Roy, A., J. Yang, and Y. Zhang, *COFACTOR: an accurate comparative algorithm for structure-based protein function annotation.* Nucleic acids research, 2012: p. gks372.

94. Von Mering, C., et al., *Comparative assessment of large-scale data sets of protein–protein interactions.* Nature, 2002. **417**(6887): p. 399-403.
95. Xenarios, I., et al., *DIP: the database of interacting proteins.* Nucleic acids research, 2000. **28**(1): p. 289-291.
96. Bader, G.D., D. Betel, and C.W. Hogue, *BIND: the biomolecular interaction network database.* Nucleic acids research, 2003. **31**(1): p. 248-250.
97. Kerrien, S., et al., *The IntAct molecular interaction database in 2012.* Nucleic acids research, 2011: p. gkr1088.
98. Li, H.-D., et al., *Functional networks of highest-connected splice isoforms: from the Chromosome 17 Human Proteome Project.* Journal of proteome research, 2015. **14**(9): p. 3484-3491.
99. Enright, A.J., et al., *Protein interaction maps for complete genomes based on gene fusion events.* Nature, 1999. **402**(6757): p. 86-90.
100. Marsh, J.A., et al., *Protein complexes are under evolutionary selection to assemble via ordered pathways.* Cell, 2013. **153**(2): p. 461-470.
101. Shaw, D.E., et al., *Anton, a special-purpose machine for molecular dynamics simulation.* Communications of the ACM, 2008. **51**(7): p. 91-97.
102. George, R.A. and J. Heringa, *SnapDRAGON: a method to delineate protein structural domains from sequence data.* Journal of molecular biology, 2002. **316**(3): p. 839-851.
103. Zhang, W., et al., *Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11.* Proteins: Structure, Function, and Bioinformatics, 2015.
104. Roy, A., A. Kucukural, and Y. Zhang, *I-TASSER: a unified platform for automated protein structure and function prediction.* Nature protocols, 2010. **5**(4): p. 725-738.
105. Zhang, Y., *I-TASSER server for protein 3D structure prediction.* BMC bioinformatics, 2008. **9**(1): p. 40.
106. Yang, J., et al., *The I-TASSER Suite: protein structure and function prediction.* Nature methods, 2015. **12**(1): p. 7-8.
107. Källberg, M., et al., *Template-based protein structure modeling using the RaptorX web server.* Nature protocols, 2012. **7**(8): p. 1511-1522.
108. Källberg, M., et al., *RaptorX server: a resource for template-based protein structure modeling*, in *Protein Structure Prediction*. 2014, Springer. p. 17-27.
109. Zhou, H. and J. Skolnick, *Ab initio protein structure prediction using chunk-TASSER.* Biophysical journal, 2007. **93**(5): p. 1510-1518.
110. Wu, S., J. Skolnick, and Y. Zhang, *Ab initio modeling of small proteins by iterative TASSER simulations.* BMC biology, 2007. **5**(1): p. 17.
111. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* Journal of molecular biology, 1999. **292**(2): p. 195-202.
112. Zhou, H. and Y. Zhou, *Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition.* Proteins: Structure, Function, and Bioinformatics, 2004. **55**(4): p. 1005-1013.
113. Cheng, J. and P. Baldi, *Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms.* Bioinformatics, 2005. **21**(suppl 1): p. i75-i84.
114. Cheng, J. and P. Baldi, *Improved residue contact prediction using support vector machines and a large feature set.* BMC bioinformatics, 2007. **8**(1): p. 113.

115. Wu, S. and Y. Zhang, *A comprehensive assessment of sequence-based and template-based methods for protein contact prediction.* Bioinformatics, 2008. **24**(7): p. 924-931.
116. Zhang, Y., D. Kihara, and J. Skolnick, *Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding.* Proteins: Structure, Function, and Bioinformatics, 2002. **48**(2): p. 192-201.
117. Zhang, Y. and J. Skolnick, *SPICKER: A clustering approach to identify near-native protein folds.* Journal of computational chemistry, 2004. **25**(6): p. 865-871.
118. Li, Y. and Y. Zhang, *REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks.* Proteins: Structure, Function, and Bioinformatics, 2009. **76**(3): p. 665-676.
119. Kuntz, I.D., et al., *A geometric approach to macromolecule-ligand interactions.* Journal of molecular biology, 1982. **161**(2): p. 269-288.
120. Redington, P.K., *Molfit: A computer program for molecular superposition.* Computers & chemistry, 1992. **16**(3): p. 217-222.
121. Macindoe, G., et al., *HexServer: an FFT-based protein docking server powered by graphics processors.* Nucleic Acids Research, 2010: p. gkq311.
122. Tovchigrechko, A. and I.A. Vakser, *GRAMM-X public web server for protein–protein docking.* Nucleic acids research, 2006. **34**(suppl 2): p. W310-W314.
123. Gabb, H.A., R.M. Jackson, and M.J. Sternberg, *Modelling protein docking using shape complementarity, electrostatics and biochemical information.* Journal of molecular biology, 1997. **272**(1): p. 106-120.
124. Cheng, T.M.K., T.L. Blundell, and J. Fernandez-Recio, *pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking.* Proteins: Structure, Function, and Bioinformatics, 2007. **68**(2): p. 503-515.
125. Chen, R., L. Li, and Z. Weng, *ZDOCK: An initial-stage protein-docking algorithm.* Proteins: Structure, Function, and Bioinformatics, 2003. **52**(1): p. 80-87.
126. Pierce, B. and Z. Weng, *ZRANK: reranking protein docking predictions with an optimized energy function.* Proteins: Structure, Function, and Bioinformatics, 2007. **67**(4): p. 1078-1086.
127. Hwang, H., et al., *Performance of ZDOCK and ZRANK in CAPRI rounds 13–19.* Proteins: Structure, Function, and Bioinformatics, 2010. **78**(15): p. 3104-3110.
128. Bonvin, A.M., *Flexible protein–protein docking.* Current opinion in structural biology, 2006. **16**(2): p. 194-200.
129. Zacharias, M., *Accounting for conformational changes during protein–protein docking.* Current opinion in structural biology, 2010. **20**(2): p. 180-186.
130. Andrusier, N., R. Nussinov, and H.J. Wolfson, *FireDock: fast interaction refinement in molecular docking.* Proteins: Structure, Function, and Bioinformatics, 2007. **69**(1): p. 139-159.
131. Mashiach, E., et al., *FireDock: a web server for fast interaction refinement in molecular docking.* Nucleic acids research, 2008. **36**(suppl 2): p. W229-W232.
132. Schueler-Furman, O., C. Wang, and D. Baker, *Progress in protein–protein docking: Atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility.* Proteins: Structure, Function, and Bioinformatics, 2005. **60**(2): p. 187-194.

133. Venkatraman, V. and D.W. Ritchie, *Flexible protein docking refinement using pose-dependent normal mode analysis.* Proteins: Structure, Function, and Bioinformatics, 2012. **80**(9): p. 2262-2274.
134. Schindler, C.E., S.J. de Vries, and M. Zacharias, *iATTRACT: Simultaneous global and local interface optimization for protein–protein docking refinement.* Proteins: Structure, Function, and Bioinformatics, 2015. **83**(2): p. 248-258.
135. Wang, C., P. Bradley, and D. Baker, *Protein–protein docking with backbone flexibility.* Journal of molecular biology, 2007. **373**(2): p. 503-519.
136. Wang, T. and R.C. Wade, *Implicit solvent models for flexible protein–protein docking by molecular dynamics simulation.* Proteins: Structure, Function, and Bioinformatics, 2003. **50**(1): p. 158-169.
137. Lu, L., et al., *Multimeric threading-based prediction of protein–protein interactions on a genomic scale: Application to the Saccharomyces cerevisiae proteome.* Genome Research, 2003. **13**(6a): p. 1146-1154.
138. Gao, M. and J. Skolnick, *New benchmark metrics for protein-protein docking methods.* Proteins: Structure, Function, and Bioinformatics, 2011. **79**(5): p. 1623-1634.
139. Aloy, P. and R.B. Russell, *Interrogating protein interaction networks through structural biology.* Proceedings of the National Academy of Sciences, 2002. **99**(9): p. 5896-5901.
140. Chen, H. and J. Skolnick, *M-TASSER: an algorithm for protein quaternary structure prediction.* Biophysical journal, 2008. **94**(3): p. 918-928.
141. Günther, S., et al., *Docking without docking: ISEARCH—prediction of interactions using known interfaces.* Proteins: Structure, Function, and Bioinformatics, 2007. **69**(4): p. 839-844.
142. Dominguez, C., R. Boelens, and A.M. Bonvin, *HADDOCK: a protein-protein docking approach based on biochemical or biophysical information.* Journal of the American Chemical Society, 2003. **125**(7): p. 1731-1737.
143. Fernández-Recio, J., M. Totrov, and R. Abagyan, *Soft protein–protein docking in internal coordinates.* Protein Science, 2002. **11**(2): p. 280-291.
144. Kozakov, D., et al., *PIPER: an FFT-based protein docking program with pairwise potentials.* Proteins: Structure, Function, and Bioinformatics, 2006. **65**(2): p. 392-406.
145. Sandak, B., H.J. Wolfson, and R. Nussinov, *Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers.* Proteins Structure Function and Genetics, 1998. **32**(2): p. 159-174.
146. Vakser, I.A., *Protein docking for low-resolution structures.* Protein Engineering, 1995. **8**(4): p. 371-378.
147. Kastritis, P.L. and A.M. Bonvin, *Are scoring functions in protein− protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark.* Journal of proteome research, 2010. **9**(5): p. 2216-2225.
148. Douguet, D., et al., *Dockground resource for studying protein–protein interfaces.* Bioinformatics, 2006. **22**(21): p. 2612-2618.
149. Lorenzen, S. and Y. Zhang, *Identification of near-native structures by clustering protein docking conformations.* PROTEINS: Structure, Function, and Bioinformatics, 2007. **68**(1): p. 187-194.
150. Janin, J., et al., *CAPRI: a critical assessment of predicted interactions.* Proteins: Structure, Function, and Bioinformatics, 2003. **52**(1): p. 2-9.

151. Walhout, A.J., et al., *Protein interaction mapping in C. elegans using proteins involved in vulval development.* Science, 2000. **287**(5450): p. 116-122.
152. Aloy, P., et al., *The relationship between sequence and interaction divergence in proteins.* Journal of molecular biology, 2003. **332**(5): p. 989-998.
153. Prabu, M.M., K. Suguna, and M. Vijayan, *Variability in quaternary association of proteins with the same tertiary fold: a case study and rationalization involving legume lectins.* Proteins: Structure, Function, and Bioinformatics, 1999. **35**(1): p. 58-69.
154. Park, S.-Y., et al., *In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(32): p. 11646-11651.
155. Pasek, S., J.-L. Risler, and P. Brézellec, *Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins.* Bioinformatics, 2006. **22**(12): p. 1418-1423.
156. Janin, J., *Protein–protein docking tested in blind predictions: the CAPRI experiment.* Molecular BioSystems, 2010. **6**(12): p. 2351-2362.
157. Gray, J.J., et al., *Protein–protein docking predictions for the CAPRI experiment.* Proteins: Structure, Function, and Bioinformatics, 2003. **52**(1): p. 118-122.
158. Andreani, J., G. Faure, and R. Guerois, *Versatility and invariance in the evolution of homologous heteromeric interfaces.* 2012.
159. Xue, Z., et al., *ThreaDom: extracting protein domain boundary information from multiple threading alignments.* Bioinformatics, 2013. **29**(13): p. i247-i256.
160. Xue, Z., et al., *Extending Protein Domain Boundary Predictors to Detect Discontinuous Domains.* PloS one, 2015. **10**(10): p. e0141541.
161. Cheng, T.M., T.L. Blundell, and J. Fernandez-Recio, *Structural assembly of two-domain proteins by rigid-body docking.* BMC bioinformatics, 2008. **9**(1): p. 1.
162. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.
163. Consortium, G.O., *The Gene Ontology (GO) database and informatics resource.* Nucleic acids research, 2004. **32**(suppl 1): p. D258-D261.
164. Ispolatov, I., et al., *Binding properties and evolution of homodimers in protein–protein interaction networks.* Nucleic acids research, 2005. **33**(11): p. 3629-3635.
165. Brender, J.R. and Y. Zhang, *Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles.* PLoS Comput Biol, 2015. **11**(10): p. e1004494.
166. Zhang, Q.C., et al., *Structure-based prediction of protein-protein interactions on a genome-wide scale.* Nature, 2012. **490**(7421): p. 556-560.
167. Rain, J.-C., et al., *The protein–protein interaction map of Helicobacter pylori.* Nature, 2001. **409**(6817): p. 211-215.
168. Uetz, P., et al., *A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-627.
169. Montelione, G.T., *The Protein Structure Initiative: achievements and visions for the future.* F1000 biology reports, 2012. **4**.
170. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-96.
171. Moult, J., *A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.* Current opinion in structural biology, 2005. **15**(3): p. 285-289.

172. Zhang, Y. and J. Skolnick, *Automated structure prediction of weakly homologous proteins on a genomic scale.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(20): p. 7594-7599.
173. Rychlewski, L., et al., *Comparison of sequence profiles. Strategies for structural predictions using sequence information.* Protein Science, 2000. **9**(2): p. 232-241.
174. Madera, M., *Profile Comparer: a program for scoring and aligning profile hidden Markov models.* Bioinformatics, 2008. **24**(22): p. 2630-2631.
175. Xu, Y., et al., *Protein threading by PROSPECT: a prediction experiment in CASP3.* Protein engineering, 1999. **12**(11): p. 899-907.
176. Yan, R., et al., *A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction.* Scientific reports, 2013. **3**.
177. Zhou, H. and Y. Zhou, *Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments.* Proteins: Structure, Function, and Bioinformatics, 2005. **58**(2): p. 321-328.
178. Yang, J., et al., *Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade.* Proteins: Structure, Function, and Bioinformatics, 2015.
179. Xu, D. and Y. Zhang, *Ab Initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment.* Scientific reports, 2013. **3**.
180. Wang, X., et al., *Three-dimensional reconstruction of protein networks provides insight into human genetic disease.* Nature biotechnology, 2012. **30**(2): p. 159-164.