STRUCTURAL MODELING OF PROTEIN-PROTEIN INTERACTIONS USING MULTIPLE-CHAIN THREADING AND FRAGMENT ASSEMBLY

By

SRAYANTA MUKHERJEE

Submitted to the graduate degree program in Bioinformatics
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctorate of Philosophy

Chairperson Ilya A. Vakser, PhD

Co-Chairperson Yang Zhang, PhD

John Karanicolas, PhD

Eric Deeds, PhD

Mark Richter, PhD

Date Defended:

The Dissertation Committee for Srayanta Mukherjee
certifies that this is the approved version of the following dissertation:

STRUCTURAL MODELING OF PROTEIN-PROTEIN INTERACTIONS USING MULTIPLE-CHAIN THREADING AND FRAGMENT ASSEMBLY

Chairperson Ilya Vaker, PhD

Co-Chairperson Yang Zhang, PhD

Date approved:

Abstract

Since its birth, the study of protein structures has made progress with leaps and bounds. However, owing to the expenses and difficulties involved, the number of known protein structures has not been able to catch up with the number of protein sequences and in fact has steadily lost ground. This necessitated the development of high-throughput, but accurate, computational algorithms capable of predicting the three dimensional structure of proteins from its amino acid sequence. While progress has been made in the realm of protein tertiary structure prediction, the advancement in protein quaternary structure prediction has been limited by the fact that the degree of freedom for protein complexes is even larger and fewer number of protein complex structures are present in the PDB library. In fact, protein complex structure prediction has largely remained a docking problem where automated algorithms aim to predict the complex structures starting from the unbound crystal structure of its component subunits and has remained largely limited in scope. Secondly, since docking essentially treats the unbound subunits as "rigid-bodies" it has limited accuracy when conformational change accompanies protein-protein interaction.

In one of the first of its kind effort, this study aims for the development of protein complex structure prediction algorithms which require only the amino acid sequence of the interacting subunits as input. The study aimed to adapt the best features of protein tertiary structure prediction including template detection and *ab initio* loop modeling and extend it for protein-protein complexes. The algorithm thus performs simultaneous modeling of the three dimensional structure of the component subunits while attempting to ensure the correct orientation of the chains at the protein-protein interface. Essentially, the algorithms are dependent on knowledge-based statistical potentials for both fold recognition and structure modeling.

First, as a way to compare known structure of protein-protein complexes, a complex structure alignment program MM-align was developed. MM-align joins the chains of the complex structures to be aligned to form artificial monomers in every possible order. It then aligns them using a heuristic dynamic programming based approach using TM-score as the objective function. However, the traditional NW dynamic programming was redesigned to prevent the cross alignment of chains during the structure alignment process.

Driven by the knowledge obtained from MM-align that protein complex structures share evolutionary relationships and the current protein complex structure library already contains homologous/structurally analogous protein quaternary structure families, a dimeric threading approach, COTH was designed. The new threading-recombination approach boosts the protein complex structure library by combining tertiary structure templates with complex alignments. The query sequences are first aligned to complex templates using the modified dynamic programming algorithm, guided by a number of predicted structural features including *ab initio* binding-site predictions. Finally, a template-based complex structure prediction approach, TACOS, was designed to build full-length protein complex structures starting from the initial templates identified by COTH. TACOS, fragments the template-aligned regions of the templates and reassembles them while building the structure of the threading unaligned regions *ab inito* using a replica-exchange monte-carlo simulation procedure. Simultaneously, TACOS also searches for the best orientation match of the component structures driven by a number of knowledge-based potential terms. Overall, TACOS presents one of the first approaches capable of predicting full length protein complex structures from sequence alone and introduces a new paradigm in the field of protein complex structure modeling.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mentor, Dr Yang Zhang, without whose mentorship, guidance and unstinted support, nothing was possible. He inspired me to seek the limits of my potential which I probably never fully realized but not due to lack of effort from his side. As a world-renown expert in the field, his insights and vision were invaluable and it was a great learning experience. It was a pleasure, privilege and honor to work under his apprenticeship.

I would like to thank all past and current members of the Zhanglab; it was a pleasure to work with all of them. Specially, I would like to thank Dr. Sitao Wu and Dr. Dong Xu for all the stimulating discussions we shared. I would like to specially thank Dr. Ambrish Roy who has been a friend, philosopher and often my “private tutor” from the day I started with my studies.

I would like to acknowledge The Center for Bioinformatics at the University of Kansas and Center for Computational Medicine and Bioinformatics at the University of Michigan which provided me with support and unparalled environments to work in.

I would like to thank all my collaborators, and the professors who taught me at KU. I would especially like to thanks all my committee members for providing their valuable time and advice.

Last but not the least a special vote of thanks to my family members and my in-laws Sumit and Shipra Chaudhury who stood like a rock behind me at all times. None of this would have been possible without the constant love, encouragement and support of my parents Susanta and Srabani Mukherjee and I am proud to have lived upto their expectations. Sukanya, in these PhD years, went from being my fiancé to my wife to the mother of my son. The one constant factor is that she was and has remained my strength and catharsis.

TABLE OF CONTENTS

TABLE OF CONTENTS ... vi

LIST OF TABLES ... ix

LIST OF FIGURES ... x

LIST OF ABBREVIATIONS ... xii

CHAPTER 1. Introduction ... 15

1.1 PROTEIN STRUCTURE: A JOURNEY THROUGH THE YEARS ... 16

1.1.1 Early breakthroughs ... 16

1.1.2 The first structures ... 17

1.1.3 The birth of the PDB and exponential growth ... 18

1.2 COMPARING SEQUENCE, STRUCTURE AND PROFILES OF PROTEINS. ... 19

1.2.1 Substitution matrices ... 19

1.2.2 Sequence alignments ... 19

1.2.4. Sequence profiles, PSI-BLAST and profile-profile alignment ... 21

1.2.2 Comparing protein structures ... 23

1.3 PROTEIN STRUCTURE PREDICTION ... 25

1.3.1 I-TASSER Methods ... 27

1.4 PROTEIN COMPLEX STRUCTURES: IMPORTANCE AND SHORTFALLS. ... 31

1.5 PREDICTION OF PROTEIN COMPLEX STRUCTURES ... 34

1.5.1 FFT-based docking by ZDOCK and RDOCK refinement ... 36

1.5.2 Monte-carlo based docking by Rosetta ... 37

1.6 OVERVIEW OF RESEARCH ... 37

1.7 REFERENCES ... 38

CHAPTER 2. MM-Align: A Quick Algorithm for Aligning Multiple-Chain Protein Complex Structures Using Iterative Dynamic Programming ... 51

2.1. RESULTS ... 52

2.1.1 Benchmark Sets ... 52

2.1.2 Prevention of Cross-chain Alignment ... 53

2.1.3 Option for Interface Enhanced Alignments ... 56

2.1.4 Functional Relevance of Structure Alignments 57

2.1.5 Alignment of Large Oligomers 60

2.2 MATERIALS AND METHODS 63

2.2.1 TM-score for complexes 64

2.2.2 Chain Selection and Order of Chain Joining 64

2.2.3 Initial Alignments 65

2.2.4 Heuristic Iterations 68

2.2.5 Preferential Alignment on Interfaces 69

2.3 DISCUSSION 71

2.4 REFERENCES 73

CHAPTER 3. Protein-protein complex structure predictions by multimeric threading and template recombination 77

3.1 RESULTS 78

3.1.1 Overall results of COTH on testing proteins 78

3.1.2 Comparison of different alignment algorithms 84

3.1.3. Structure combination of threading templates 88

3.1.4 Comparison of COTH with docking algorithms 92

3.2 MATERIALS AND METHODS 99

3.2.1 Template libraries 100

3.2.2 Single-chain monomeric threading 100

3.2.3 Protein-Protein Interface Prediction by BSpred 101

3.2.4 COTH threading 103

3.2.5 Template Selection and Target Classification 104

3.2.6 Structural superposition and combination of monomer templates 105

3.3 DISCUSSIONS 107

3.4 REFERENCES 109

CHAPTER 4. Reassembly and Refinement of Multimeric Templates Using Replica-Exchange Monte-Carlo Simulations 115

4.1 RESULTS 118

4.1.1 Benchmark Sets 118

4.1.2 Improvement over initial templates 119

4.1.3 Comparisons with docking 125

4.1.4 Performance of TACOS across different protein complex classes........................ 129

4.2 MATERIALS AND METHODS.. 130

4.2.1 Template selection 131

4.2.2 Mapping of dimer onto an artificial monomer on CAS lattice 132

4.2.3 Structure Assembly: Movements.. 133

4.2.4 Structure Assembly: Inter-chain energy terms 135

4.2.5 Evaluation 139

4.3 DISCUSSION... 142

4.4 REFERENCES 144

CHAPTER 5. Conclusions.. 148

5.1 REFERENCES 152

Appendix I 155

Appendix II 175

Appendix III.. 176

Appendix IV.. 178

LIST OF TABLES

Table 2.1. Summary of results from TM-align and MM-align on complex structure alignments. 54

Table 3.1. Naming conventions of methods described. 79

Table 3.2. Template identification by different methods on 500 testing proteins. 87

Table 3.3. Summary of the best in top 10 models on 77 ZDOCK benchmark proteins. 99

Table 4.1. Comparison of performance of TACOS with respect to controls. 128

LIST OF FIGURES

Figure 1.1. A timeline showing key events which helped shape modern protein structural biology 16

Figure 1.2. A schematic diagram of the I-TASSER[42, 63-64, 73] structure and function prediction protocol 31

Figure 1.3: Mapping of the number of complex structures per year 33

Figure 2.1. TM-score histogram of 205 protein complexes and their best-matching structures identified by MM-align in a non-redundant set of 3,897 protein complexes 53

Figure 2.2.Example of cross-aligned chains by TM-align........ 55

Figure 2.3. Three examples of protein dimeric complex alignments identified by MM-align, from three different protein classes (alpha-, alpha/beta-, and beta-proteins). 59

Figure 2.4. The structural alignment of casein kinase (1cki) with its best-matching structures in a non-redundant protein complex library. 60

Figure 2.5. Examples of MM-align on big oligomers 62

Figure 2.6. An illustration of the chain-joining procedure in MM-align........ 63

Figure 2.7. An illustration of the modified dynamic programming algorithm with cross-chain alignment prevented......... 68

Figure 2.8. A modified dynamic programming scheme with the alignment of interface residue pairs reinforced 70

Figure 3.1: Flowchart of the COTH algorithm for protein complex template identification 79

Figure 3.2: Complex threading results by COTH on 500 non-redundant test proteins 81

Figure 3.3: Comparison of TM-score of the complex templates as identified by different threading methods 86

Figure 3.4: Examples of improvement of COTH over controls. 88

Figure 3.5: Structure superposition improves template quality 91

Figure 3.6: Modeling result of ZDOCK and COTH on the Ran-Importin beta complex 97

Figure 3.7: TM-score versus Z-score of the first COTH templates for 180 training proteins ... 105

Figure 4.1: Flowchart of the TACOS, template-based assembly of complex structures, protocol 117

Figure 4.2: Correlation of TACOS energy with TM-score 119

Figure 4.3: Plot showing benchmark results of TACOS on 350 test set proteins 121

Figure 4.4. Comparison of TACOS with control methods 122

Figure 4.5. Near-native models built by TACOS 123

Figure 4.6: Success rate of ZDOCK on I-TASSER models versus that on the unbound x-ray structures 126

Figure 4.7: Plot showing CAS lattice based representation of dimer structure 133

Figure 4.8. Flowchart showing the replica-exchange monte-carlo movement schematic of the TACOS simulation 135

Figure 5.1 Graphical network representation of the similarity of protein-protein complex structures by Cytoscape. 150

LIST OF ABBREVIATIONS

BLAST:	Basic Local Alignment Search Tool
BLOSUM:	Block Substitution Matrix
BSpred:	Binding site predictor
CAS:	C-alpha and side chain center
CATH:	Class, Architecture, Topology, Homologous superfamily
CE:	Combinatorial Extension
CHARMM:	Chemistry at Harvard Molecular Mechanics
CM:	Comparative Modeling
C-MUSTER:	Complex-Multi Source Threader
COM:	Center of Mass
COTH:	Co-threading
C-PPA:	Complex-Profile Profile Alignment
DP:	Dynamic Programming
FFT:	Fast Fourier Transform
FG-MD:	Fragment guided Molecular Dynamics
GDT-TS:	Global Distance Test-total score
GO:	Gene Ontology
HMM:	Hidden Markov Model
HOMBACOP:	Homology Based Complex Prediction
I-RMSD:	Interface-Root Mean Square Deviation
I-TASSER:	Iterative Threading Assembly and Refinement
LOMETS:	Local Meta Server
MM-align:	Multimer Align
M-TASSER:	Multimer-Threading Assembly and Refinement

MUSTER:	Multi Source Threader
NMR:	Nuclear Magnetic Resonance
NOE:	Nuclear Overhauser Effect
NW:	Needleman Wunsch
PAM:	Point Accepted Mutation
PDB:	Protein Data Bank
PPA:	Profile Profile Alignment
PSI:	Protein Structure Initiative
PSI-BLAST:	Position Specific Iterative Basic Local Alignment Search Tool
PSSM:	Position Specific Substitution Matrix
RMSD:	Root Mean Square Deviation
rTM-score:	reciprocal Template Modeling Score
SCOP:	Structural Classification of Proteins
TACOS:	Template-based Complex Structures
TASSER:	Threading Assembly and Refinement
TM-score:	Template Modeling score

This thesis is dedicated to my grandparents, Kalidas and Shanti Mukherjee and Nandalal and Deepa Sanyal and my granduncle Gurudas Mukherjee.

My PhD research is dedicated to my parents Susanta and Srabani Mukherjee and to the two individuals who complete me; my wife Sukanya Chaudhury and son Shraman Mukherjee.

CHAPTER 1. Introduction

Protein structure is commonly divided into four levels; i) primary structure or amino acid sequence where each amino acid is linked together by covalent peptide bonds ii) secondary structure or the local conformations (helices and beta strands) held together by main chain hydrogen bonding iii) tertiary structure or the global monomeric fold primarily driven by the need for a hydrophobic core and hydrophilic surface and iv) quaternary structure or the oligomeric state, the driving forces of which can vary depending on the nature of the complex. The rise of X-ray crystallography in the late 1950's and protein nuclear magnetic resonance (NMR) in the 1980's made the study of protein structures possible. In Figure 1, a timeline is presented detailing some key contributions down the years which made study of protein structures a reality.

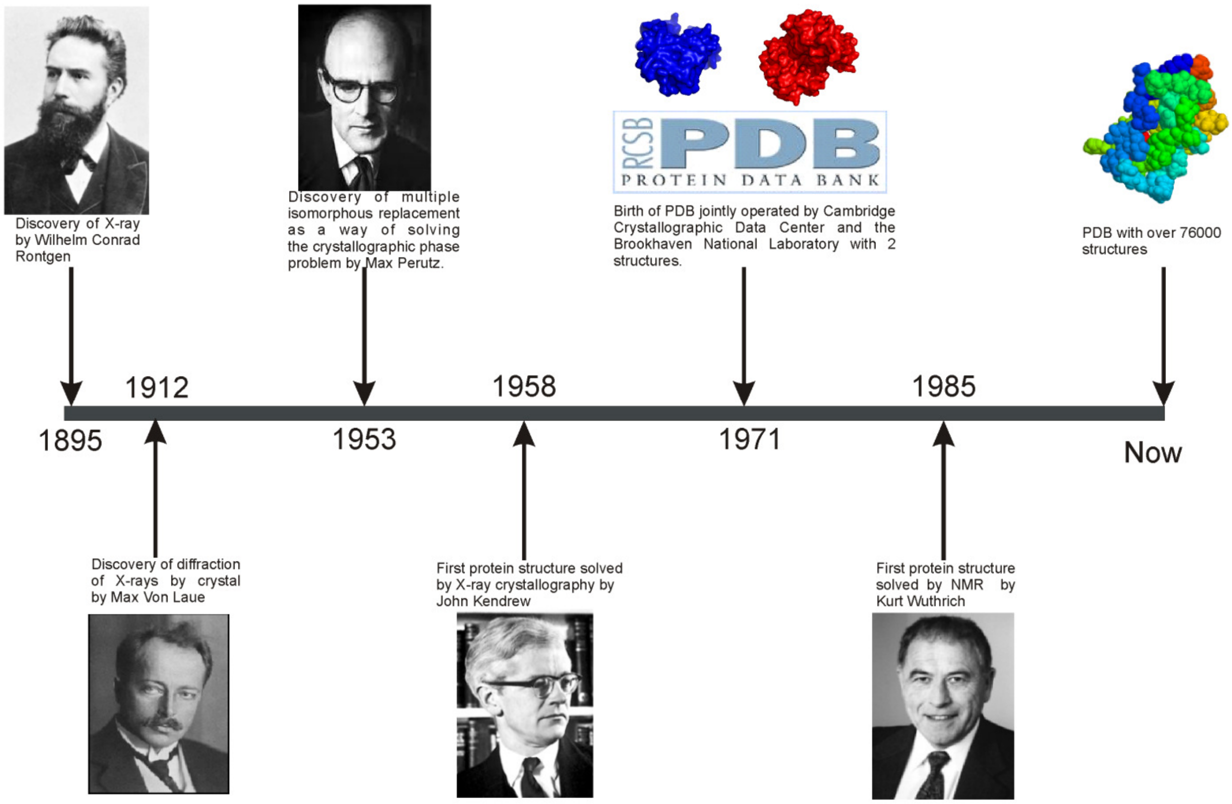
Discovery of X-ray by Wilhelm Conrad Rontgen
Discovery of multiple isomorphous replacement as a way of solving the crystallographic phase problem by Max Perutz.
RCSB PDB PROTEIN DATA BANK
Birth of PDB jointly operated by Cambridge Crystallographic Data Center and the Brookhaven National Laboratory with 2 structures.
PDB with over 76000 structures
1912
1958
1985
1895
1953
1971
Now
Discovery of diffraction of X-rays by crystal by Max Von Laue
First protein structure solved by X-ray crystallography by John Kendrew
First protein structure solved by NMR by Kurt Wuthrich

Figure 1.1. A timeline showing key events which helped shape modern protein structural biology

The field has now broken new grounds with the rise of computational algorithms capable of predicting, from primary amino acid sequence alone, the three dimensional structure of proteins with increasingly high accuracy. However, among the four levels of protein structure, least amount of information is available regarding the last level or quaternary structure. This fact is a by-product of a combination of factors including complexities involved, expenditure incurred for a thorough investigation and other factors. Even computationally, handling protein complexes are tougher due to their size as well as due to an added parameter, the orientation of the individual chains with respect to each other. This investigation therefore makes an attempt to try and bridge the gap using a theoretical approach via computational modeling of protein dimers. In what follows, groundwork is first laid to aid in the comprehension of later chapters. First, a review of protein structural biology from a historical perspective is presented followed by tracing the rise of bioinformatics and protein structure and protein complex structure prediction while reviewing basic concepts. Selected state of the art methodologies in protein structure and complex prediction are also reviewed.

1.1 PROTEIN STRUCTURE: A JOURNEY THROUGH THE YEARS

1.1.1 Early breakthroughs

After the discovery of an electromagnetic radiation with a wavelength range now called X-rays by William Rontgen in 1895, Max Von Laue demonstrated the phenomenon of diffraction of X-ray by crystals in 1912. William Braggs was the first to report the structure of a crystal, that of NaCl, in 1913 using X-ray diffraction patterns and when James Sumner, in 1926, demonstrated that enzymes can be isolated and crystallized, the field of protein X-ray crystallography was born. One of the earliest significant contributions to the field was made by

William Astbury who not only obtained some of the best X-ray photographs of protein crystals in the early 1930s but also proposed that mainchain-mainchain hydrogen bonding was a significant contributor towards the stabilization of protein structures in 1931. This culminated in Linus Pauling building upon Astbury's findings to correctly propose, in 1951, that the primary structural motifs in protein structures were alpha helices and beta sheets [1] held together by hydrogen bonding. In 1953 and 1962, two significant breakthroughs were made to solve the phasing problem of X-ray diffraction. The first event, in 1953, was the demonstration by Max Perutz that the phase problem of protein crystals could be solved by multiple isomorphous replacement [2] i.e. by comparing the diffraction patterns of the native protein itself and that of the protein after it had been soaked in a solution containing heavy metal ions under different conditions. In 1962, a second approach to solving the phase problem was formalized by Michael Rossmann and colleagues using a technique known as "molecular replacement".

1.1.2 The first structures

The first structure to be solved was that of the sperm whale myoglobin in 1958 by John Kendrew [3] followed soon after by the first complex structure, that of human hemoglobin in 1963 by Max Perutz [4-5]. The lysozyme structure [6] and the structure of a lysozyme-inhibitor complex [7] (the first structure of an enzyme-inhibitor complex) followed in 1965. More structures followed including ribonuclease in 1967 [8-9], papain in 1968 [10] and a small but complicated molecule, insulin, by Dorothy Hodgkins and co-workers in 1971 [11-12]. Interestingly, Tom Blundell who worked on the structure of insulin would also go on to do pioneering work much later in the field of computational protein structure biology. In the early 1980's, an alternative to X-ray crystallography emerged, led by the pioneering work of Kurt Wuthric and Richard Ernst [13] on Nuclear magnetic resonance (NMR) and the two dimensional nuclear overhauser effect (2D NOE). This culminated with the first solution NMR structure, that of proteinase inhibitor IIA in 1985.

1.1.3 The birth of the PDB and exponential growth

In between, in 1971, a collaborative effort between the Brookhaven National Laboratory and Cambridge Crystallographic Data Center resulted in the birth of a computerized central repository for storing the atomic co-ordinates of solved structures in a common universal format, called the Protein Data Bank (PDB) under the direction of Walter Hamilton [14]. The PDB began with two structures, myoglobin (1MBN) and hemoglobin (1DHB, superseded by 2DHB) and within 5 years grew to 23 structures [15]. Since then, the PDB has seen exponential growth and currently contains, close to 77000 structures. In 2010 alone, 7929 structures were deposited or at an average rate of 22 structures per day. However, despite this mind-boggling rate the number of sequences in sequence databases far exceeds the number of structures, more than 200 times as of 2011. Due to the costs, complexity and time required to solve protein structures it is clear that it is impossible for structures to keep pace with sequences. Additionally, many proteins are not amenable to crystallization or NMR based studies. These factors, coupled with rapid rise of computational power and availability, led to the development of computational algorithms to predict theoretically, the structure of proteins from their amino acid sequence alone. Simultaneously, attempts were also made to develop algorithms capable of predicting the association of two different protein molecules and to predict the structure of the protein complex when the structures of the individual subunits are known. The concept of profile-profile alignment for fold-recognition by David Eisenberg and colleagues in 1991 [16] had profound effects on revolutionizing the field of protein structure prediction. Similarly the use of a grid based Fast Fourier Transform correlation search technique by Katchalski-Katzir, Ilya Vakser and colleagues in 1992 [17] had profound impacts on protein complex structure prediction by docking.

1.2 COMPARING SEQUENCE, STRUCTURE AND PROFILES OF PROTEINS.

1.2.1 Substitution matrices

An amino acid substitution matrix is a 20×20 matrix (representing the 20 amino acid types) which computes the rate at which a particular amino acid mutates to the other 19 amino acids and is accepted by the evolutionary natural selection process. It can therefore be interpreted as a mathematical formulation which quantifies the intuitive hypothesis that a random mutation of an amino acid is more likely to be accepted by nature if it mutates to an amino acid with similar physiochemical properties. The first such mutation matrix, the Point Accepted Mutation (PAM) matrix, was derived by Margaret Dayhoff in 1978 [18]. The PAM was computed by aligning the sequence of 71 closely related families and each cell of the matrix represents the probability of each particular mutation being observed in nature.

The BLOSUM [19] or Block Substitution Matrix, on the other hand, was constructed using gapless local alignment of conserved zones or "blocks" of distantly related protein families. The frequency of each possible mutation was calculated and normalized to compute the mutational probability. The log-odds of each value were thereby calculated to generate the final 20×20 matrix. The numerical suffix of the BLOSUM matrices, for example BLOSUM62, BLOSUM90, represents the sequence identity cutoff for the sequences used to compute the matrix. In this study, the BLOSUM62 has been used throughout.

1.2.2 Sequence alignments

Sequence alignment is defined as the method used to measure the similarity or dissimilarity of two sequences. The Needleman-Wunsch (NW) dynamic programming method [20] developed in 1970 and its variations [21-22] have, for long, remained the method of choice for performing pairwise alignment of two sequences. In this study, the NW algorithm and a variation of it (discussed in Chapter 2) have been used multiple times and hence merit a thorough discussion.

The most widely used implementation of the NW algorithm is described in the following. It has two distinct steps; 1) matrix filling and 2) path traceback. A step-by-step procedure to implement the NW dynamic programming is as follows:

1. During the first step, a $N1 \times N2$ matrix is first created where *N1* and *N2* are the length of two sequences being compared.
2. The value of each cell of the matrix is denoted as $M_{i,j}$ where *i* and *j* are the positions on sequence 1 and sequence 2 respectively.
3. The values of the $M_{i,j}$ are then progressively calculated in a top-down fashion, beginning at the first position of both sequence 1 and sequence 2, and are given by the equation:

   $$M_{i,j} = \max\left\lfloor M_{i-1,j-1} + S(A_i, B_j), M_{i,j-1} + GP, M_{i-1,j} + GP \right\rfloor \qquad (1)$$

   where $S(A_i,\ B_j)$ is the score (generally obtained from a substitution matrix like BLOSUM62) of aligning amino acid type *A* with amino acid type *B* which are present at the *i*th and *j*th position of sequence 1 and 2 respectively.
4. The gap penalty, *GP* is given by the equation $GP = u + kv$ where *u* is the gap-opening penalty, *v* is the gap extension penalty and *k* is the length of the gap. It is imperative to represent the gap penalty in this way because it allows (by recursively calculating the total gap penalty for any previous path already traversed) the algorithm to be completed in $N1 \times N2$ steps.
5. Importantly, the NW algorithm not only computes the values of each cell but also stores the direction, diagonal, horizontal or vertical, from which the value was generated. For

example if $M_{i-1,j-1} + S(A_i, B_j) > M_{i,j-1} + GP$ and $M_{i-1,j-1} + S(A_i, B_j) > M_{i-1,j} + GP$ then the direction from which the values of $M_{i,j}$ is derived is recorded to be from the diagonal cell.

6. The algorithm then proceeds to fill up the rest of the matrix till all cells have been filled up.
7. In the second step, path traceback, the algorithm proceeds in the opposite direction that is from the cell $M_{i,j}$ where $i=N1$ and $j=N2$.
8. At each step, it looks for the parent cell or the direction from which the value of $M_{i,j}$ was derived.
9. If the value of a particular cell was derived from the diagonal direction, then the two amino acids in sequence 1 and sequence 2 are aligned with each other. If the value of $M_{i,j}$ was derived from the horizontal or vertical direction then a gap is introduced in the alignment.
10. The algorithm then moves to the parent cell of $M_{i,j}$ and continues to traceback the optimal path till it reaches the cell where $i=1$ and $j=1$.

The NW algorithm is based on the principle of divide and conquer where the best globally optimal solution is obtained by breaking the problem into overlapping sub-problems and finding the best locally optimal solution for each sub-problem. However, the algorithm is computationally expensive and cannot realistically be used to perform a very large number of sequences, like comparing two genomes. This necessitated the birth of many orders faster, albeit sub-optimal, heuristic alignment procedures like FASTA [23] and BLAST [24].

1.2.4. Sequence profiles, PSI-BLAST and profile-profile alignment

When multiple sequences need to aligned, repeated pairwise alignments can be used to generate a multiple sequence alignment. Extending further, for any given query sequence a multiple sequence alignment can be generated against entire sequence databases. The homologous sequences identified by the multiple sequence alignment can then be used to generate a “sequence profile” or a position specific score matrix (PSSM). The sequence profile, introduced by David Eisenberg and co-workers [25], is the log-odds frequency of each of the 20 amino acids at any given position of the multiple sequence alignment and is represented by a 20×N matrix where N is the length of the query sequence. This profile matrix can then be used in place of substitution matrix as the scoring function when aligning the query sequence to other proteins or protein profiles to detect evolutionary similarity. When aligning two profiles i.e. performing profile-profile alignment, the score function is the product of the frequency matrix of the query sequence and the log-odds profile of the template sequence.

PSI-BLAST [26] uses iterative sequence-profile alignments to detect homologous sequence through sequence databases. In the first step, the query sequence is aligned to the sequences in the databases to create a multiple sequence alignment. The homologues identifyied in the first round of search are then used to create a PSSM for the query sequence and this PSSM is re-aligned to the database sequences to identify a new set of homologues. The new homologue set is used to recalculate the sequence profile and the search step is repeated. The entire process is then repeated till the search converges and no new homologues are identified. PSI-BLAST is an extremely fast yet a powerful tool for detection of homologues. It is capable of searching through large databases containing millions of sequences and has established itself as the tool of choice across the community. It has also been instrumental in establishing sequence profiles as one of the most powerful properties to establish evolutionary relatedness.

1.2.2 Comparing protein structures

Like protein sequences, protein structures can also be compared for similarity/dissimilarity and in fact may be more meaningful since protein structures are generally more conserved in nature than sequence [27]. Hence, two proteins which have similar structures often share similar functions. It is also important, when comparing two objects, to represent the similarity in the form of a quantifiable scoring function. A number of such scoring functions exist which include RMSD [28], GDT-TS score [29], MaxSub [30] score and TM-score [31] among others. These scoring functions can be used to not only compare the structures of two different proteins but can also be used to assess the quality of models and templates predicted by structure prediction and docking algorithms. In this investigation, RMSD and TM-score have been used as the scoring functions of choice (some other specific ones have been used and are described in more detail in the later chapters) both for comparison of different proteins as well as for model assessment and hence merits more detailed discussion.

Root mean square deviation (RMSD) is a widely used scoring function where the two structures are first superposed onto each other by rotating and translating one structure onto the other to minimize the average distance between the atoms of the two structures. After superposition the RMSD between the two structures is calculated using the equation

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \qquad (2)$$

where n is the total number of atoms and x_1, y_1 ,z_1 and x_2, y_2, z_2 are the x, y, and z coordinates of structure 1 and structure 2 respectively. For proteins, the unit of RMSD is generally angstroms (Å), and lower the RMSD more similar the two structures are. One disadvantage of the RMSD is that it does not provide the complete information about the similarity of the two structures. For example, it is not necessary or obvious that two structures with a RMSD of 1.0 Å but with only

30% of the residues aligned is more similar than a different pair of structures with an RMSD of 3.0 Å but with 70% of the residues aligned.

The TM-score was defined as a measure to assess the structural similarity of protein monomer chains and is given by the equation

$$\mathrm{TM-score} = \max\left[\frac{1}{L}\sum_{i=1}^{L_{ali}}\frac{1}{1+d_{ij}^2/d_0^2}\right] \tag{3}$$

where L is the total length of all chains in the target structure and L_{ali} is the number of the aligned residue pairs in the two structures. d_{ij} is the distance between the Cα atoms of the aligned residues i and j after superposition of the structures, and d_0 is given by $d_0 = 1.24\sqrt[3]{L-15}-1.8$. To calculate the TM-score, the aligned residues are broken into fragments of equal length for both structures. Then, beginning from the N-terminal of both structures, the first fragment of one structure is superposed onto the first fragment of the second structure according to the RMSD rotation matrix. The rotation matrix returning the optimal superposition for the fragment pair is then used to rotate and translate the rest of the atoms of the first structure. The TM-score, according to Eq (3), is then calculated for the full length of the two structures where two residues are considered aligned if the distance d_{ij} between the two atoms is less than the cutoff d_o. The first fragment of the first structure is then superposed on to the second fragment of the second structure and once again the TM-score is calculated. This procedure is repeated till all possible fragment pairs, scanning from the N- to the C-terminal end, has been superposed. In the next round, the length of the fragment is increased and the whole procedure is repeated till all possible superpositions with increasing fragment lengths are completed. Finally, the superposition which yields the highest TM-score is returned as the optimal superposition and the TM-score for that is returned as the final TM-score between the two structures.

One major advantage of TM-score over the often-used RMSD in assessing structural alignments is that TM-score accounts for both the similarity of the aligned regions and the alignment coverage in a single parameter. Second, even when alignments with the same coverage are evaluated, TM-score is more sensitive to the global topology of the structures because it down-weights the larger distances between aligned Cα pairs compared to the smaller ones. In RMSD all distances are taken into account with equal weights, and therefore a local error (e.g. a mis-oriented tail) will result in a big RMSD value even though the global topology of the two structures may be similar. As reported by Xu and Zhang [32], a TM-score of 1 means that the two structures are identical, a TM-score>0.5 indicates that two structures have a similar topology and share similar folds, and a TM-score<0.17 indicates that the structural similarity is close to random.

1.3 PROTEIN STRUCTURE PREDICTION

Rapid strides have been taken in the field of protein structure prediction from amino acid sequence using computational methods [33]. The obvious advantage of computational methods is their speed and low cost, making genome-scale structure prediction and functional annotations a reality. Protein structure prediction methods can be divided into three main categories based on the approach that is adopted [33]: 1) comparative or homology modeling[34-36] 2) threading or fold recognition [16, 37-40] and 3) *ab initio* or *de novo* methods [41-49].

In comparative modeling (CM), the protein structure is constructed by matching the sequence of the protein of interest (query protein) to an evolutionarily related protein with a known structure (template protein) in the PDB. Thus, a prerequisite for comparative modeling technique is the presence of a homologous protein in the PDB [50] library. For proteins with >50% sequence identity to their templates, models built by CM techniques can have up to a 1.0 Å

RMSD from the native structure for the backbone atoms. For proteins which have a 30 to 50% sequence identity with their template, the models often have ~85% of their core regions within an RMSD of 3.5 Å from the native structure, with errors mainly in loop regions. When the sequence identity drops below 30% (in the twilight zone [51]), modeling accuracy sharply decreases because of substantial alignment errors and lack of significant template hits. Also, by definition, models built by CM usually have a strong bias towards the template structure rather than being closer to the native structure of the target protein [52-53].

Threading or fold recognition is similar to CM modeling in the sense that it also searches a structure library to identify a known structure which would "best fit" a given query sequence. However, an evolutionary relationship (homology) between the query and the template is not a prerequisite in this case. These "sequence to structure" alignment approaches usually employ a wide range of scoring functions to find the best alignment, and may rely on profile-profile alignment [16], distance dependent potentials [54], predicted secondary structure [55], solvent accessibility [56-57], and other predicted structural features. Most of the successful threading approaches use scores combining sequence features and predicted structural information [39, 58-59], with a search engine of either NW dynamic programming [20, 22] or Hidden Markov model (HMM) [60-61] for remote homology detection and fold recognition.

Ab initio or *de novo* methods originally referred to approaches purely based on physicochemical properties; however, some of the contemporary algorithms in this category do use evolutionary and knowledge-based information to collect spatial restraints or to detect structural fragments to assist structure assembly. Still, by definition, *ab initio* methods are not dependent on the presence of known structures which are sequentially or structurally similar to a given query sequence. The guiding principle of this approach is the Anfinsen hypothesis [62], which states that the native structure of the protein lies at the global energy minimum of the conformational energy landscape. Therefore, *ab initio* approaches try to fold a given protein based on various force fields via conformational search. Though some notable developments have been made in this field [41-49], predicting the three-dimensional structure of proteins longer than 150 amino acids is still an unsolved problem due to the inaccuracy of the available force fields and the bottlenecks arising out of insufficient conformational search.

Significant progress has been achieved in developing composite structure predictions which combine various approaches of comparative modeling, threading and *ab initio* folding. The Threading ASSEmbly Refinment (TASSER) [31] and Iterative Threading ASSEmbly Refinement (I-TASSER) [42, 63-64] methods are notable examples in this category.

1.3.1 I-TASSER Methods

I-TASSER is a composite structure prediction method and is an extension of TASSER (developed in the Skolnick lab) [31, 65] involving a hierarchical combination of template search by threading, followed by assembly and rearrangement of continuous fragments excised from the templates. The protein conformation is specified by an on-and-off-lattice system with an energy function integrating a number of structural restraints which are predicted from the threading templates. The on-and-off-lattice-based conformational search is used to generate thousands of conformations or “decoys” which are then subjected to iterative structural clustering for the selection of the final models [66].

The I-TASSER predictions begin by taking the amino acid sequence as input, which is then subjected to “sequence-structure alignment” or threading by LOMETS [40] against a comprehensive threading library. The threading process utilizes close and distant sequence profiles and predicted secondary structure information from PSIPRED [67] and other predicted structural features to find the best match. The alignment is performed using the NW dynamic programming algorithm [20] or HMM based alignments, and the raw alignment score and the alignment length are used to obtain the statistical significance (Z-score) of the alignment. The alignments on different templates are ranked by the Z-score, which is also used to classify the query protein into an "easy", a "medium" or a "hard" target. The "hard" category basically means that no good threading template is detectable in the library, and the structure will have to be largely predicted by an "*ab initio*" method. I-TASSER also includes (a) sequence-based contact predictions from SVMSEQ to guide the *ab initio* simulations [68-69]; (b) REMO, to refine the hydrogen-bonding network of reduced models [70]; (c) iterative TASSER reassembly [42]; (d) integration of structure-based functional annotations.

The templates found by the threading process are divided into continuously aligned (>5 residues) and gapped regions, and placed onto the CAS (C-Alpha and Side-chain center of mass) on-and-off-lattice model. The local structure of the aligned regions remains unchanged during the simulation; their Cα atoms are excised from the template and placed off-lattice in order to keep the fidelity of the structures. In the gapped or *ab initio* regions, Cα atoms are placed on the lattice points with a grid spacing of 0.87 Å. The side-chain centers of mass are off-lattice for all regions. The gapped regions are first filled up using a random walk of Cα-Cα bond vectors to generate a full-length model which is subsequently subjected to parallel hyperbolic Monte Carlo sampling [71]. Once again the CAS model differentiates between the on- and off-lattice atoms with regard to the movements they are subjected to. The off-lattice atoms are subjected to rigid-body translation and rotation. Care is taken to ensure that the acceptance probability of a movement is approximately the same for different fragment lengths, implemented by normalizing the amplitude of movement by the length of the fragment. On the other hand, on-lattice atoms are subjected to two- to six-bond movements and sequence shifts of multiple bonds.

The I-TASSER energy function integrates three different classes of energy terms. The first term consists of a number of knowledge-based statistical potentials derived from the PDB [50], including long-range side-chain pair interactions, hydrogen-bond potential terms, hydrophobic interaction and local Cα correlations. The second class includes the propensity of an amino acid to assume a particular secondary structure as predicted by PSIPRED[67] in order to impose a general "protein-like bias" to the decoys generated, while the third class includes protein specific tertiary structure contact restraints and a distance map calculated by LOMETS from the generated threading templates. New potential terms that have been incorporated in I-TASSER include the predicted accessible surface area (ASA) [42, 57] and sequence-based contact predictions [68]. Both energy terms have been derived and optimized using machine learning methods. The overall correlation between the actual exposed area as calculated by STRIDE [72] and that predicted by a neural network is 0.71, based on a test on 2,234 non-homologous proteins. In the latest version of I-TASSER [73], the sequence-based pairwise residue contact information from SVMSEQ [68], SVMCON [74] and BETACON [75] are used to constrain the simulation search and improve the funnel around the global minimum of the energy landscape.

The trajectories of the low-temperature replicas from the first-round of I-TASSER simulations are clustered by SPICKER [66]. The cluster centroids are obtained by averaging all the clustered structures after superposition and are ranked based on the structure density of the cluster. Cluster centroids generally have a number of non-physical steric clashes between Cα atoms and can be over-compressed. Starting from the selected SPICKER cluster centroids, the I-TASSER Monte Carlo simulation is performed again (see Figure 1.2). While the inherent I-

TASSER potential remains unchanged in the second run, external constraints are added. These are derived by pooling the initial high-confidence restraints from threading alignments, the distance and contact restraints from the combination of the centroid structures, and the PDB structures identified by the structure alignment program TM-align [76] using the cluster centroids as query structures. The conformation with the lowest energy in the second round is selected as the final model. The main purpose of this iterative strategy is to remove the steric clashes of the cluster centroids. To increase the biological usefulness of protein models, all-atom models are generated by REMO [70] simulations, which include three general steps: (1) removal of steric clashes by moving each of the Cα atoms that clash with other residues; (2) backbone reconstruction by scanning a backbone isomer library collected from the solved high-resolution structures in the PDB library; (3) hydrogen bonding network optimization based on predicted secondary structure from PSIPRED. Finally, Scwrl3.0 [77] is used to add the side-chain rotamers.

Recently, I-TASSER was extended by an additional component to predict the biological function of the query proteins. The procedure involves matching the I-TASSER-generated structural models against representative libraries of proteins with known function using both global and local structure alignment based methods in order to find the best functional homologs in the PDB library. Based on a large-scale benchmark test set of more than 300 non-homologous proteins, it was found that even when the structures are predicted after removing all the homologous templates from the template library, the correct function (EC number and GO terms) and binding site could be identified with high confidence in more than 80% of the cases.

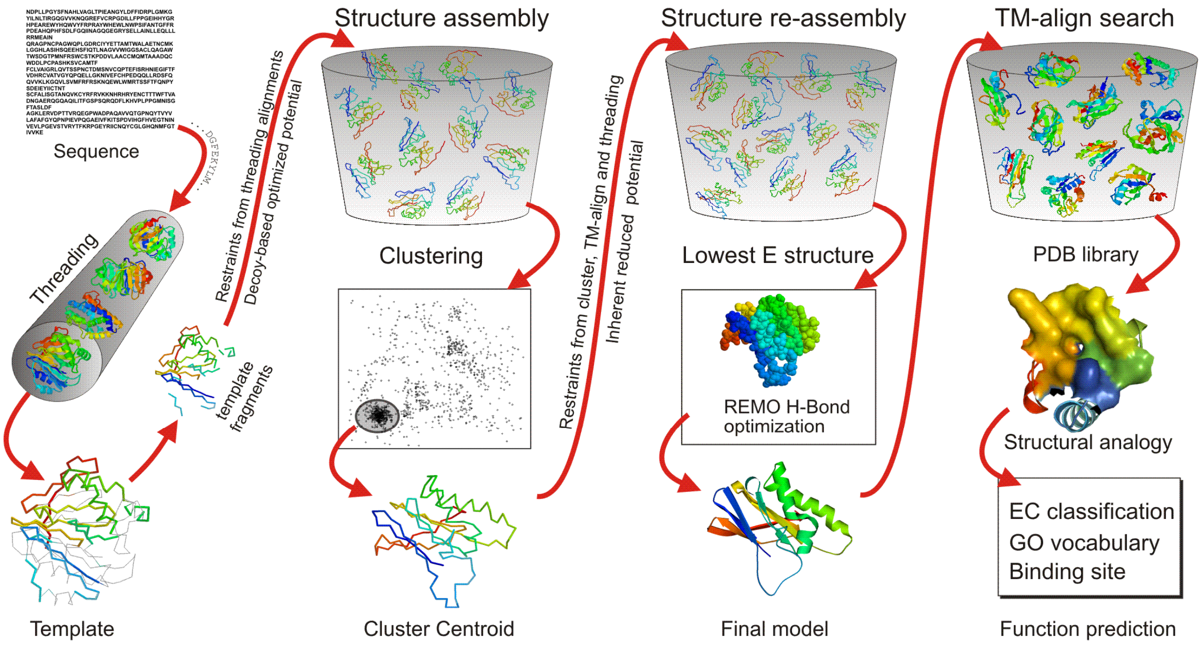

Figure 1.2. A schematic diagram of the I-TASSER[42, 63-64, 73] structure and function prediction protocol. Templates for the query protein are identified by LOMETS[40], which provides template fragments and spatial restraints. The template fragments are then assembled by parallel hyperbolic Monte Carlo simulations[71]. The conformations generated during the simulation are clustered using SPICKER[66], in order to find the structure with the lowest free energy. As an iterative strategy, the cluster centroids are then subjected to second round of simulation with the purpose of refining the structure and removing clashes. The final all-atom models are generated by REMO[70] through the optimization of hydrogen-bonding networks. Functional homologs (PDB structures that have an associated EC number/GO term/known binding site) of the final models are identified using both global structural search[76] and local structure alignment programs which aim at finding matches between the binding/active sites of the predicted structure and templates with known function.

1.4 PROTEIN COMPLEX STRUCTURES: IMPORTANCE AND SHORTFALLS.

Protein-protein interactions play a central role in almost all cellular processes and pathways including metabolism, signaling, DNA replication, transcription, translation, splicing and apoptosis [78-80]. High throughput proteomics methods like yeast two-hybrid assays [81] and affinity purification followed by mass spectrometry [82] have provided a wealth of information regarding putative protein complexes spanning whole genomes [83-89]. The information from such studies reveals that most proteins interact with other protein partners, even though the interaction may only be transient. On average, each protein interacts with at least 9 other proteins

[90-91]. Due to the myriad of important roles that protein-protein interactions play in the cellular machinery they are attractive targets for novel drug design [92-96].

A prerequisite (in most cases) to obtaining a substantiative understanding of the mechanism of protein-protein interactions, is to have knowledge of the 3D structure of the complex [97]. Information about the orientation of atoms in three dimensional coordinate space not only provides more detailed information about the possible biochemical and biophysical parameters involved in the process but can also prove invaluable in rational structure-based drug design. While it can be argued that explicit knowledge of structure is not a pre-requisite for a successful case study of a pair of interacting proteins, it can be very difficult or nearly impossible to explain the observed phenomenon without knowledge of the atomistic details. Structure largely dictates function and therefore to obtain a complete mechanistic understanding of function, knowledge of structural details is an indispensable asset.

While it was earlier believed that protein-protein interfaces are not attractive candidates for development of small molecule inhibitors [27] primarily due to the large surface area involved and the relatively flat interfaces, a paradigm shift was brought about by the identification of “hot spot” residues at protein interfaces by Wells and co-workers [98]. Since then a number of success stories exist in the field of novel drug design targeted towards inhibition of protein-protein interactions [99-101].

Unfortunately, complex structures are more difficult to solve experimentally due to problems associated with co-crystallization and the cumbersome process of solving complex structures by NMR. While 102,308 pairwise complexes are present in the PDB as of September, 2011, a large number of them are crystallization artifacts and/or redundant. If the PDB, is screened to eliminate complexes with buried surface area less than 250Å^2, at least 30 interface residues and a sequence identity cutoff of 70%, only 7701 structures remain according to the protein complex structure database, DOCKGROUND [102-103].

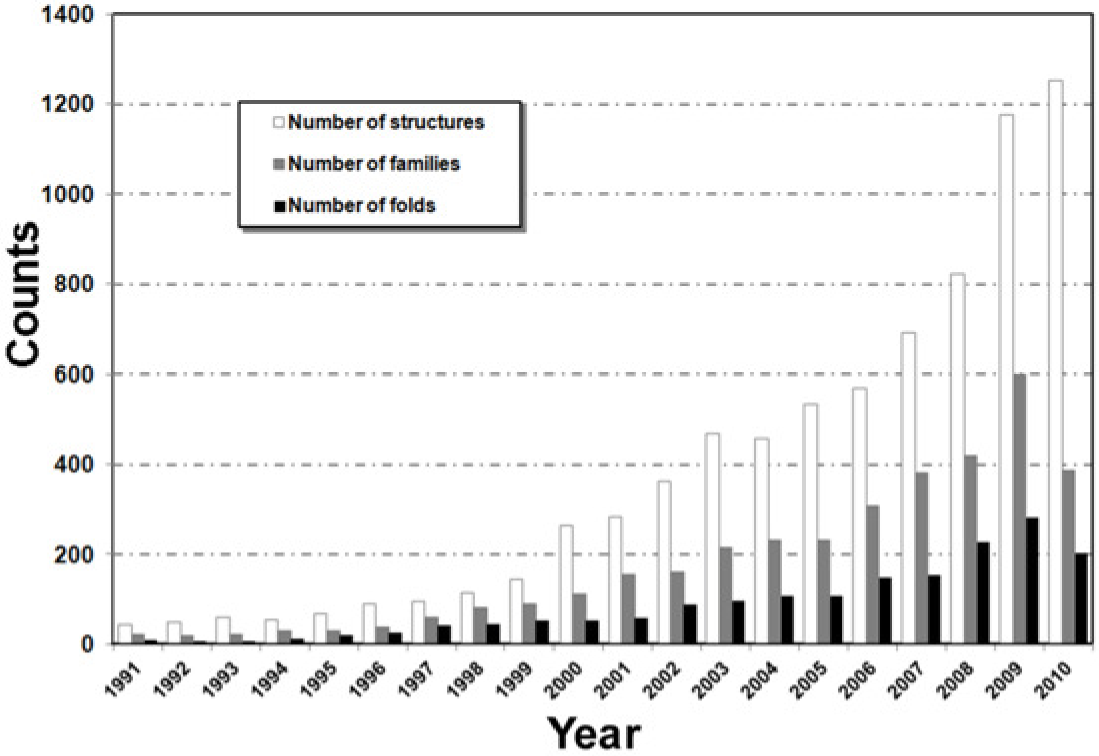

Figure 1.3: Mapping of the number of complex structures per year. A plot showing the number of new quaternary families and number of new quaternary folds that were deposited in the PDB for the last 20 years till April, 2011.

In Figure 1.3, a mapping of the number of complex structures, the number of new quaternary families and number of new quaternary folds that were deposited in the PDB for the last 20 years till April, 2011 is shown. After showing a trend of steady increase of all three categories, a jump was observed in the year 2000 in the actual number of structures deposited though that did not correspond to a simultaneous jump in the number of new families and folds that were deposited. One of the reasons for that is in the year 2000 a number of Antibody-Antigen complexes were deposited which did not necessarily differ in the quaternary family or fold. Another sharp rise in the actual number of structures deposited was observed in 2003 which also resulted in an increase in the corresponding number of new families and folds. This also corresponds to the Phase I cycle of the Protein Structure Initiative (PSI) as well as the SPINE2 and 3D Reperotire projects. A final jump in the number of structures was observed in 2009, though the number of families and folds decreased from 2009 to 2010. This can either be attributed to the completion of the 3D repertoire and the Phase II cycle of PSI or may be an indication of a slow down due to higher completeness of the library. If we were to assume that in the following years the growth curve would be a mirror image of the curve in the last 20 years (with technological advancements being offset by more of the fold space being covered up) then it will take roughly 25 years from now to reach approximately 4000 unique quaternary folds or a complete set of possible quaternary folds in nature. This necessitates the need for developing efficient computational algorithms for predicting the structure of protein complexes.

1.5 PREDICTION OF PROTEIN COMPLEX STRUCTURES

Some of the earliest efforts which defined the field of protein docking were made in the late 1970's and early 1980's [104-107] and established the role of shape complementarity as a central paradigm on which many advancements have been made over the years. Thus, when the atomic resolution structures of two proteins are available and the two proteins are known to interact, docking methods aim to predict the structure of the complex mainly by trying to maximize the shape and physiochemical complementarity between the two structures. The docking process involves two distinct stages i) global search for generation of decoys ii) scoring or ranking to identify near-native decoys. For the first "search" stage, emphasis is placed on geometric shape complementarity and is generally done on lower-resolution "smoothened" structures. There are two schools of thought for the implementation of the search stage; one which involves a more exhaustive search using grid based Fast-Fourier Transforms (FFT) to maximize surface correlation and the second which uses Monte-carlo methods to sample the rugged energy landscape. Methods like GRAMM-X [108] and ZDOCK [109] among others use FFT-based correlation techniques for sampling while methods like ICM-DISCO [110] and ROSETTA [111] use monte-carlo methods for decoy generation. Increasingly, docking methods are also shifting to a two-stage search process, an initial low resolution search followed by high resolution refinement [112-114], to account for side-chain flexibility and conformational change during the protein complex formation process. Moreover, the initial search stage docking decoys are generally always subjected to a re-ranking procedure which includes higher resolution physics based potentials like electrostatics, hydrogen bonding, desolvation and hydrophobicity [115-116].

One disadvantage of current docking methods is that it essentially treats the two 2 unbound subunits of the complex as a "rigid-body". Therefore, while current methods are quite successful when the RMSD between the bound and unbound forms of the subunits are low (typically < 1.0Å) [112-114], accounting for large conformational changes poses a significant challenge. To counter this problem, some strategies have evolved over the years which can treat the problem implicitly by using "soft-docking" and "ensemble-docking" or explicitly by implementing a two-stage search step to allow for side-chain and some backbone flexibility in the second-stage high resolution refinement procedure. Below, we present a more detailed discussion of two representative examples of both approaches, ZDOCK for FFT based docking and RosettaDock as a model of monte-carlo based docking methods.

A third approach which has recently emerged and is in nascent stages of development is homology modeling of complex structures [117-119]. These approaches are built on the premise that if the individual subunits of the query complex are homologous to the individual subunits of a complex of known structure then the query complex is expected to share the same orientation as the template complex. Four significant efforts have been made in this direction in recent years which includes the development of MULTIPROSPECTOR [118] by Skolnick's group, HOMBACOP by Kundrotas et. al [119], the strategy used by Aloy et. al. [117] and the work by Sinha et al [120].

1.5.1 FFT-based docking by ZDOCK and RDOCK refinement

ZDOCK [109] is a grid based initial stage docking algorithm where the receptor and ligand structures are treated as "rigid bodies" and are subjected to 6 rotational and translation degree of freedom. ZDOCK uses the FFT based search technique pioneered by Katchalski-Katzir et. al [17]. The novelty of the approach was the use of three distinct scoring functions to represent the protein on the grid; protein shape complementarity, desolvation [121] and electrostatics. The proteins, receptor and ligand, are placed in a cubic grid lattice and discrete functions are used to represent the three different scoring functions for the receptor and the ligand separately. The final energy is a correlation of the discrete functions used.

The initial stage generation of decoys is performed on a "smoothened" grid to allow for some conformational flexibility. Thereafter, the decoys generated by ZDOCK are subjected to a high resolution, three step refinement process. The three step minimization includes a) removal of clashes b) optimization of the summation of van der waals energy and coulombic interactions c) optimization of "charge-charge" ionic interactions. The molecular mechanics package CHARMM [122] is used to implement the refinement procedure. Finally, the refined decoys are re-ranked by re-evaluation of the desolvation potential and electrostatic energy component of the CHARMM force-field.

1.5.2 Monte-carlo based docking by Rosetta

Rosetta also employs a two stage (low and high) docking procedure where a monte-carlo sampling technique is used as the search engine. In the initial low resolution stage, the structures are represented by the backbone atoms and side chain centroids only and a “glancing” contact is enforced by rigid body rotation and translation moves of one protein along the surface of the other (sliding). The force field includes a) residue-residue interaction b) residue environment potential c) rewarding of inter-chain contacts and d) penalty for clash. After multiple low resolution search steps, the lowest ranking decoys (according to energy) is subjected to a high resolution refinement procedure.

During the refinement procedure the side chains atoms are represented explicitly and is introduced using a backbone-based rotamer packing algorithm [123]. The decoys are then extensively minimized to search for the local minima. The full atomic force-field used in this step includes full-atomic Van der Waal’s interactions, a Gaussian solvation potential, hydrogen bonding, rotamer probability, knowledge-based pairwise interaction of residues, electrostatic potential and surface area of solvation. The search procedure (low resolution search followed by high resolution refinement) is repeated through many cycles and finally the 200 lowest energy decoys are retained. These decoys are then clustered and the members of the top cluster, where the clusters are ranked according to cluster density, are used as the final predictions [124].

1.6 OVERVIEW OF RESEARCH

The basic premise of this study is that similar to protein tertiary structure, protein quaternary structure is also evolutionarily conserved. Hence, utilizing the knowledge available from already existing protein complex structures in the PDB, it is possible to identify structure templates and use them to predict the structure of a given query complex sequence. First, a structural alignment tool which could be used to compare two protein complex structures was designed by extending the protein structure alignment method TM-align. The method used a modified NW dynamic programming protocol to align multiple chains simultaneously while preventing cross-alignment of chains.

The number of protein-protein complex structures is nearly 6-times smaller than that of tertiary structures in PDB which limits the power of homology-based approaches to complex structure modeling. Therefore, a new threading-recombination approach, COTH, was developed to boost the protein complex structure library by combining tertiary structure templates with complex alignments. The query sequences are first aligned to complex templates using a modified dynamic programming algorithm, guided by *ab initio* binding-site predictions. The monomer alignments are then shifted to the multimeric template framework by structural alignments.

Finally, the threading templates identified by COTH were subjected to a rigorous reassembly and refinement process to generate full-length structures. Replica-exchange monte-carlo simulation was implemented as the sampling technique of choice guided by a knowledge-based energy function. The TACOS force-field is composed of general statistically derived potential terms as well as inter-chain and intra-chain distance and contact restraints. The decoys generated were clustered and cluster centroid was refined using fragment-guided molecular dynamics (FG-MD) [125] to produce full-atomic models of protein complex structures.

1.7 REFERENCES

1. Pauling, L. and R.B. Corey, *The polypeptide-chain configuration in hemoglobin and other globular proteins.* Proc Natl Acad Sci U S A, 1951. **37**(5): p. 282-5.

2. Perutz, M.F., *An optical method for finding the molecular orientation in different forms of crystalline haemoglobin; changes in dichroism accompanying oxygenation and reduction.* Proc R Soc Lond B Biol Sci, 1953. **141**(902): p. 69-71.

3. Kendrew, J.C., et al., *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.* Nature, 1958. **181**(4610): p. 662-6.

4. Perutz, M.F. and L. Mazzarella, *A Preliminary X-Ray Analysis of Haemoglobin H.* Nature, 1963. **199**: p. 639.

5. Muirhead, H. and M.F. Perutz, *Structure of Haemoglobin. A Three-Dimensional Fourier Synthesis of Reduced Human Haemoglobin at 5-5 a Resolution.* Nature, 1963. **199**: p. 633-8.

6. Blake, C.C., et al., *Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution.* Nature, 1965. **206**(986): p. 757-61.

7. Johnson, L.N. and D.C. Phillips, *Structure of some crystalline lysozyme-inhibitor complexes determined by X-ray analysis at 6 Angstrom resolution.* Nature, 1965. **206**(986): p. 761-3.

8. Wyckoff, H.W., et al., *The structure of ribonuclease-S at 3.5 A resolution.* J Biol Chem, 1967. **242**(17): p. 3984-8.

9. Wyckoff, H.W., et al., *The structure of ribonuclease-S at 6 A resolution.* J Biol Chem, 1967. **242**(16): p. 3749-53.

10. Drenth, J., et al., *Structure of papain.* Nature, 1968. **218**(5145): p. 929-32.

11. Blundell, T.L., et al., *Atomic positions in rhombohedral 2-zinc insulin crystals.* Nature, 1971. **231**(5304): p. 506-11.

12. Blundell, T.L., et al., *X-ray analysis and the structure of insulin.* Recent Prog Horm Res, 1971. **27**: p. 1-40.

13. Kumar, A., R.R. Ernst, and K. Wuthrich, *A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules.* Biochem Biophys Res Commun, 1980. **95**(1): p. 1-6.

14. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J Mol Biol, 1977. **112**(3): p. 535-42.

15. Berman, H.M., *The Protein Data Bank: a historical perspective.* Acta Crystallogr A, 2008. **64**(Pt 1): p. 88-95.

16. 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. 164-70.

17. Katchalski-Katzir, E., et al., *Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.* Proc Natl Acad Sci U S A, 1992. **89**(6): p. 2195-9.

18. Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt, eds. *A model of evolutionary change in proteins.* An atlas of protein sequence and structure, ed. M.O. Dayhoff. Vol. 5. 1978, National Biomedical Research Foundation: Washington D.C.

19. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.

20. Needleman, S. and C. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* Journal of Molecular Biology, 1970. **48**: p. 443-453.

21. Gotoh, O., *An improved algorithm for matching biological sequences.* J Mol Biol, 1982. **162**(3): p. 705-8.

22. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.

23. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison.* Proc Natl Acad Sci U S A, 1988. **85**(8): p. 2444-8.

24. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

25. Gribskov, M., A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins.* Proc Natl Acad Sci U S A, 1987. **84**(13): p. 4355-8.

26. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

27. Bogan, A.A. and K.S. Thorn, *Anatomy of hot spots in protein interfaces.* J Mol Biol, 1998. **280**(1): p. 1-9.

28. Kabsch, W., *A solution for the best rotation to relate two sets of vectors.* Acta Crystallogr A, 1976. **32**: p. 922-923.

29. Zemla, A., et al., *Processing and analysis of CASP3 protein structure predictions.* Proteins, 1999. **Suppl 3**: p. 22-9.

30. Siew, N., et al., *MaxSub: an automated measure for the assessment of protein structure prediction quality.* Bioinformatics, 2000. **16**(9): p. 776-85.

31. Zhang, Y. and J. Skolnick, *Automated Structure Prediction of Weekly Homologous Proteins on a Genomic Scale* Proceedings of The National Academy of Science, 2004. **101**: p. 7594-7599.

32. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score = 0.5?* Bioinformatics, 2010. **26**(7): p. 889-95.

33. Zhang, Y., *Progress and challenges in protein structure prediction.* Curr Opin Struct Biol, 2008. **18**(3): p. 342-8.

34. Marti-Renom, M., et al., *Comparative protein structure modeling of genes and genomes* Annual Review of Biophysics and Biomolecular Structure, 2000. **29**: p. 291-325.

35. Sali, A. and T. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* Journal of Molecular Biology, 1993. **234**: p. 779-815.

36. Fiser, A., R.K. Do, and A. Sali, *Modeling of loops in protein structures.* Protein Sci, 2000. **9**(9): p. 1753-73.

37. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A new approach to protein fold recognition.* Nature, 1992. **358**(6381): p. 86-9.

38. Xu, Y. and D. Xu, *Protein threading using PROSPECT: design and evaluation.* Proteins, 2000. **40**(3): p. 343-54.

39. Skolnick, J., D. Kihara, and Y. Zhang, *Development and large scale benchmrk testing of the Prospector_3 threading algorithm.* Proteins, 2004. **56**(3): p. 502-518.

40. Wu, S. and Y. Zhang, *LOMETS: a local meta-threading-server for protein structure prediction.* Nucleic Acids Res, 2007. **35**(10): p. 3375-82.

41. Kolinski, A. and J. Skolnick, *Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme.* Proteins, 1994. **18**(4): p. 338-52.

42. Wu, S., J. Skolnick, and Y. Zhang, *Ab initio modelling of small proteins by iterative TASSER simulations.* BMC Biol, 2007. **5**: p. 17.

43. Zhang, Y., A. Kolinski, and J. Skolnick, *TOUCHSTONE II: a new approach to ab initio protein structure prediction* Biophysical Journal, 2003. **85**: p. 1145-1164.

44. Liwo, A., et al., *Protein structure prediction by global optimization of a potential energy function.* Proc Natl Acad Sci U S A, 1999. **96**(10): p. 5482-5.

45. Simons, K.T., et al., *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* J Mol Biol, 1997. **268**(1): p. 209-25.

46. Kihara, D., et al., *TOUCHSTONE: An ab initio protein structure prediction method that uses threading based tertiary restraints* Proceedings of The National Academy of Science, 2001. **98**: p. 10125-10130.

47. Bradley, P., K. Misuara, and D. Baker, *Towards high-resolution de novo structure prediction for small proteins.* Science, 2005. **309**: p. 1868-1871.

48. Oldziej, S., et al., *Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests.* Proc Natl Acad Sci U S A, 2005. **102**(21): p. 7547-52.

49. Klepeis, J.L., et al., *Ab initio prediction of the three-dimensional structure of a de novo designed protein: a double-blind case study.* Proteins, 2005. **58**(3): p. 560-70.

50. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

51. Rost, B., *Twilight zone of protein sequence alignments.* Protein Eng, 1999. **12**(2): p. 85-94.

52. Tramontano, A. and V. Morea, *Assesment of homology based predictions in CASP 5.* Proteins, 2003. **53**(Suppl 6): p. 352-368.

53. Read, R.J. and G. Chavali, *Assessment of CASP7 predictions in the high accuracy template-based modeling category.* Proteins, 2007. **69 Suppl 8**: p. 27-37.

54. Sippl, M. and S. Weitckus, *Detection of native like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations* Proteins, 1992. **13**: p. 258-271.

55. McGuffin, L. and D. Jones, *Improvement of GenTHREADER method for genomic fold recognition.* Bioinformatics, 2003. **19**(7): p. 874-881.

56. Zhang, B., et al., *Similarities and differences between non-homologous proteins with similar folds: evaluation of threading strategies.* Folding and Design, 1997. **2**(5): p. 307-317.

57. Chen, H. and H.X. Zhou, *Prediction of solvent accessibility and sites of deleterious mutations from protein sequence.* Nucleic Acids Res, 2005. **33**(10): p. 3193-9.

58. Wu, S. and Y. Zhang, *MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information.* Proteins, 2008. **72**: p. 547-556.

59. Zhou, H. and Y. Zhou, *Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments.* Proteins, 2005. **58**(2): p. 321-8.

60. Karplus, K., C. Barrett, and R. Hughey, *Hidden markov models for detecting remote protein homologies.* Bioinformatics, 1998. **14**(10): p. 846-856.

61. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**: p. 951-960.

62. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96): p. 223-30.

63. Zhang, Y., *Template-based modeling and free modeling by I-TASSER in CASP7.* Proteins, 2007. **69 Suppl 8**: p. 108-17.

64. Zhang, Y., *I-TASSER server for protein 3D structure prediction.* BMC Bioinformatics, 2008. **9**: p. 40.

65. Zhang, Y. and J. Skolnick, *Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins.* Biophysical Journal, 2004. **87**: p. 2647-2655.

66. Zhang, Y. and J. Skolnick, *Spicker: Approach to clustering protein structures for near native model selection.* J . of Comp. Chem., 2004. **25**: p. 865-871.

67. Jones, D., *Protein secondary structure prediction based on position-specific scoring matrices.* Journal of Molecular Biology, 1999. **292**: p. 195-202.

68. Wu, S. and Y. Zhang, *A comprehensive assessment of sequence-based and template-based methods for protein contact prediction.* Bioinformatics, 2008. **24**(7): p. 924-31.

69. Wu, S. and Y. Zhang, *Improving protein tertiary structure assembly by sequence based contact predictions.* Submitted, 2009.

70. 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, 2009.

71. Zhang, Y., D. Kihara, and J. Skolnick, *Local energy landscape flattening: Parallel hyperbolic monte-carlo sampling of protein folding.* Proteins, 2002. **48**: p. 192-201.

72. Frishman, D. and P. Argos, *Knowledge-based protein secondary structure assignment.* Proteins, 1995. **23**(4): p. 566-79.

73. Zhang, Y., *I-TASSER: Fully automated protein structure prediction in CASP8.* Proteins, 2009: p. In press.

74. Cheng, J. and P. Baldi, *Improved residue contact prediction using support vector machines and a large feature set.* BMC Bioinformatics, 2007. **8**: p. 113.

75. Cheng, J. and P. Baldi, *Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms.* Bioinformatics, 2005. **21 Suppl 1**: p. i75-84.

76. 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.

77. Canutescu, A.A., A.A. Shelenkov, and R.L. Dunbrack, Jr., *A graph-theory algorithm for rapid protein side-chain prediction.* Protein Sci, 2003. **12**(9): p. 2001-14.

78. Berggard, T., S. Linse, and P. James, *Methods for the detection and analysis of protein-protein interactions.* Proteomics, 2007. **7**(16): p. 2833-42.

79. Kuroda, K., et al., *Systems for the detection and analysis of protein-protein interactions.* Appl Microbiol Biotechnol, 2006. **71**(2): p. 127-36.

80. Phizicky, E.M. and S. Fields, *Protein-protein interactions: methods for detection and analysis.* Microbiol Rev, 1995. **59**(1): p. 94-123.

81. Ito, T., T. Chiba, and M. Yoshida, *Exploring the protein interactome using comprehensive two-hybrid projects.* Trends Biotechnol, 2001. **19**(10 Suppl): p. S23-7.

82. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.

83. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-1736.

84. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proceedings of The National Academy of Science, 2001. **98**(8): p. 4569-4574.

85. Gavin, A., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141-147.

86. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-627.

87. Ho, Y., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectroscopy.* Nature, 2002. **415**(6868): p. 180-183.

88. Rain, J., et al., *The protein-protein interaction map of Helibacter pylori.* Nature, 2001. **409**(6817): p. 211-215.

89. Li, S., et al., *A map of the interactome network of the metazoan C.elegans.* Science, 2004. **303**(5657): p. 540-543.

90. Merring, C.v., et al., *Comparative assesment of large scale datasets of protein-protein interactions.* Nature, 2002. **417**(6887): p. 399-403.

91. Schumacher, M.A., et al., *Structure of the gating domain of a Ca2+-activated K+ channel complexed with Ca2+/calmodulin.* Nature, 2001. **410**(6832): p. 1120-4.

92. Arkin, M.R. and J.A. Wells, *Small-molecule inhibitors of protein-protein interactions: progressing towards the dream.* Nat Rev Drug Discov, 2004. **3**(4): p. 301-17.

93. Arkin, M., *Protein-protein interactions and cancer: small molecules going in for the kill.* Curr Opin Chem Biol, 2005. **9**(3): p. 317-24.

94. Berg, T., *Modulation of protein-protein interactions with small organic molecules.* Angew Chem Int Ed Engl, 2003. **42**(22): p. 2462-81.

95. Che, Y., B.R. Brooks, and G.R. Marshall, *Development of small molecules designed to modulate protein-protein interactions.* J Comput Aided Mol Des, 2006. **20**(2): p. 109-30.

96. Drews, J., *Drug discovery: a historical perspective.* Science, 2000. **287**(5460): p. 1960-4.

97. Vakser, I., *PSI has to live and become PCI: Protein Complex initiative.* Structure, 2008. **16**(1): p. 1-3.

98. Bass, S.H., M.G. Mulkerrin, and J.A. Wells, *A systematic mutational analysis of hormone-binding determinants in the human growth hormone receptor.* Proc Natl Acad Sci U S A, 1991. **88**(10): p. 4498-502.

99. Arkin, M.R., et al., *Binding of small molecules to an adaptive protein-protein interface.* Proc Natl Acad Sci U S A, 2003. **100**(4): p. 1603-8.

100. Clackson, T. and J.A. Wells, *A hot spot of binding energy in a hormone-receptor interface.* Science, 1995. **267**(5196): p. 383-6.

101. Wang, J.L., et al., *Structure-based discovery of an organic compound that binds Bcl-2 protein and induces apoptosis of tumor cells.* Proc Natl Acad Sci U S A, 2000. **97**(13): p. 7124-9.

102. Douguet, D., et al., *DOCKGROUND resource for studying protein-protein interfaces.* Bioinformatics, 2006. **22**(21): p. 2612-8.

103. Gao, Y., et al., *DOCKGROUND system of databases for protein recognition studies: unbound structures for docking.* Proteins, 2007. **69**(4): p. 845-51.

104. Zielenkiewicz, P. and A. Rabczenko, *Protein-protein recognition: method for finding complementary surfaces of interacting proteins.* J Theor Biol, 1984. **111**(1): p. 17-30.

105. Greer, J. and B.L. Bush, *Macromolecular shape and surface maps by solvent exclusion.* Proc Natl Acad Sci U S A, 1978. **75**(1): p. 303-7.

106. Wodak, S.J. and J. Janin, *Computer analysis of protein-protein interaction.* J Mol Biol, 1978. **124**(2): p. 323-42.

107. Kuntz, I.D., et al., *A geometric approach to macromolecule-ligand interactions.* J Mol Biol, 1982. **161**(2): p. 269-88.

108. Tovchigrechko, A. and I.A. Vakser, *GRAMM-X public web server for protein-protein docking.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W310-4.

109. Chen, R., L. Li, and Z. Weng, *ZDOCK: an initial-stage protein-docking algorithm.* Proteins, 2003. **52**(1): p. 80-7.

110. Fernandez-Recio, J., M. Totrov, and R. Abagyan, *ICM-DISCO docking by global energy optimization with fully flexible side-chains.* Proteins, 2003. **52**(1): p. 113-7.

111. Gray, J.J., et al., *Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.* J Mol Biol, 2003. **331**(1): p. 281-99.

112. Zacharias, M., *Accounting for conformational changes during protein-protein docking.* Curr Opin Struct Biol, 2010. **20**(2): p. 180-6.

113. Russell, R.B., et al., *A structural perspective on protein-protein interactions.* Curr Opin Struct Biol, 2004. **14**(3): p. 313-24.

114. Bonvin, A.M., *Flexible protein-protein docking.* Curr Opin Struct Biol, 2006. **16**(2): p. 194-200.

115. Pierce, B. and Z. Weng, *ZRANK: reranking protein docking predictions with an optimized energy function.* Proteins, 2007. **67**(4): p. 1078-86.

116. Cheng, T.M., T.L. Blundell, and J. Fernandez-Recio, *pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking.* Proteins, 2007. **68**(2): p. 503-15.

117. Aloy, P., et al., *Structure-based assembly of protein complexes in yeast.* Science, 2004. **303**(5666): p. 2026-9.

118. Lu, L., H. Lu, and J. Skolnick, *MULTIPROSPECTOR: An algotihm for the predictio of protein-protein interactions by multimeric threading.* Proteins:Structure,Function and Genetics, 2002. **49**: p. 350-364.

119. 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.* Int J Biol Macromol, 2008. **43**(2): p. 198-208.

120. Sinha, R., P.J. Kundrotas, and I.A. Vakser, *Docking by structural similarity at protein-protein interfaces.* Proteins, 2010. **78**(15): p. 3235-41.

121. Zhang, C., et al., *Determination of atomic desolvation energies from the structures of crystallized proteins.* J Mol Biol, 1997. **267**(3): p. 707-26.

122. Brooks, B.R., et al., *CHARMM: the biomolecular simulation program.* J Comput Chem, 2009. **30**(10): p. 1545-614.

123. Kuhlman, B. and D. Baker, *Native protein sequences are close to optimal for their structures.* Proc Natl Acad Sci U S A, 2000. **97**(19): p. 10383-8.

124. Li, X., I.H. Moal, and P.A. Bates, *Detection and refinement of encounter complexes for protein-protein docking: taking account of macromolecular crowding.* Proteins, 2010. **78**(15): p. 3189-96.

125. Zhang, J., Y. Liang, and Y. Zhang, *Atomic-level protein structure refinement using fragment guided molecular dynamics conformation sampling.* Structure, 2011. **In Press**.

CHAPTER 2. MM-Align: A Quick Algorithm for Aligning Multiple-Chain Protein Complex Structures Using Iterative Dynamic Programming

Protein-protein complex structures have rapidly accumulated in various protein quaternary structure libraries [1-3]. As a consequence, large-scale automated structure comparisons of multiple-chain protein complexes have become routine in most contemporary structure biology studies, ranging from structure-based functional annotation [4-6] to protein quaternary structure modeling [7-8]. While extensive efforts have been focused on the development of protein *tertiary* structure comparisons [9-11], there is no efficient structure alignment algorithm for comparing protein *quaternary* structures.

Tertiary structure alignment algorithms, which were developed for structurally aligning two monomer structures, cannot be directly exploited for multimeric proteins. A simple treatment might be to join the multiple chains into an artificial monomer and then align the two "monomers" using existing programs such as Dali [9], CE [10], or TM-align [11]. However, non-physical cross-chain alignments, i.e. the alignment of one chain in the first complex to several chains in the second complex, will arise because the programs do not differentiate residues of different chains. Also, if the two protein complexes include more than two chains then a combinatorial problem arises which the available methods are not designed to handle. An alternative approach is to align the monomer chains of the two complexes separately. However, this alignment cannot account for the differences in chain orientations within the complexes. Moreover, the structure of interface regions is usually of special importance in both biological function annotation and structural modeling. Neither one of these approaches take the special characteristics of the interface structure into account. Alternatively, some sequence independent approaches like Galinter [12], I2I SiteEngine [13], MAPPIS [14] can compare protein-protein interfaces but does not help in analyzing the global structural similarity of complexes.

This chapter describes the development of a new algorithm, MultiMer-align (or MM-align), dedicated to multimeric protein structure alignment, as an extension of the monomeric alignment program TM-align [11]. TM-align, developed by Zhang and Skolnick, uses a heuristic dynamic programming alignment procedure. Because the objective function and the rotation matrix in TM-align are consistent with each other, and are both based on TM-score [15], the dynamic programming iteration converges faster than that in many other heuristic algorithms. On average, TM-align is about 20 times faster than Dali and 4 times faster than CE; and yet the alignments of monomer structures have higher TM-scores on average. Nevertheless, for monomer alignments, there are still some cases where TM-align does not identify the best alignment because of the limited number of initial alignments. The purpose of this work was, first to improve the efficiency of TM-align by exploring more extensive search and then to extend the algorithm to deal with the problem of unphysical cross-chain alignments and the variance of chain orientations in protein complex structures. The alignment of interface residues is also reinforced in MM-align.

2.1. RESULTS

2.1.1 Benchmark Sets

Dimers constitute by far the largest subgroup of multimeric protein complexes and therefore it is on dimers that MM-align was mostly tested. However, MM-align also has the capability of accurately aligning larger multimers and is tested on a number of higher-order multimeric cases. For testing MM-align on dimeric complexes, two sets of protein complex structures were constructed. The first set consists of 205 non-redundant dimers with various sizes and a pair-wise sequence identity of < 30%. The second set consists of 3,897 dimers collected from Dockground [2], with a pair-wise sequence identity < 70%. The pair-wise sequence identity between the first and the second complex sets are < 98%. A complete list of the two benchmark sets is available at http://zhanglab.ccmb.med.umich.edu/MM-align/benchmark.

2.1.2 Prevention of Cross-chain Alignment

At first, the ability of MM-align to exclude the unphysical cross-chain alignments was tested. Using MM-align, all dimer structures in the first benchmark set were aligned against all dimers in the second set. For each of the 205 complexes in the first set, the complex from the 3,897 complexes in the second set that has the best match based on TM-score was selected. A summary of the results on the 205 pairs is presented in Table 2.1. For most dimers in the first set, MM-align identified similar dimer structures in the second set. As shown in Figure 2.1, 83% of protein complex pairs have a TM-score > 0.5, indicating similar topology of the two complexes [15].

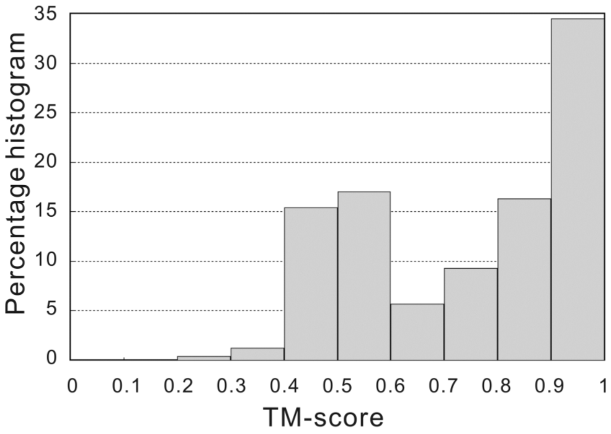

Figure 2.1. TM-score histogram of 205 protein complexes and their best-matching structures identified by MM-align in a non-redundant set of 3,897 protein complexes.

The average sequence identity between these best complex pairs is 44%. For the protein pairs having a sequence identity < 30%, the average TM-score is 0.59, indicating that MM-align can identify structures with similar topology even when the sequence identity is very low. A complete list of the alignments for the 205 best complex pairs is available in Appendix I.

As a comparison, TM-align was run on the complexes, directly aligning them with chains joined and treating them as "artificial monomers" (the results are shown as TM-align-I in Table 2.1). As expected, because TM-align does not distinguish between the different chains, a substantial portion of residue pairs gets non-physically aligned across chains. In the MM-align alignment, however, due to the exclusion of the cross-chain alignment paths in the DP matrix, there is no cross-chain alignment in any of the 205×3,897 alignments.

Method	<TM-score>	<RMSD>	<Cov[c]>	<N_{cross}[d]>
MM-align	0.759	2.65 Å	60.4%	0
TM-align-I[a]	0.750	2.70 Å	60.2%	12.6
TM-align-II[b]	0.710	3.00 Å	58.5%	0

[a] Using TM-align to align joined-chain complex structures.
[b] Same as TM-align-I but after removing the cross-chain aligned residues.
[c] Average fraction of the aligned residue pairs divided by the length of the target complex.
[d] Average number of the non-physical cross-chain alignments.

Table 2.1. Summary of results from TM-align and MM-align on complex structure alignments.

One example is presented in Figure 2.2, where both chains of the C. AhdI protein complex (PDB ID: 1y7y) are aligned by TM-align on the A chain of the *Xenopus laevis* nudix hydrolase nuclear SnoRNA decapping protein (PDB ID: 1u20). But when MM-align is used on the same structure pair, there is no cross-chain alignment and the interfaces of the two complexes are correctly aligned. Remarkably, despite the fact that MM-align searches far fewer possible alignment paths than TM-align (i.e. neglecting all the paths of cross alignments, see Figure 2.7), the average TM-score and RMSD of the best alignments by MM-align are better than those produced by TM-align-I. This improvement is mainly attributed to the newly added initial alignments in MM-align and the improved DP search in the existing paths of TM-align. On the other hand, the fact that much fewer paths find equivalent structure matches with similar or even better TM-scores reflects that the protein quaternary structures have inherent structural similarities of separate domain/chains. If we remove the cross-chain parts of the alignment from TM-align (shown as "TM-align-II" in Table 2.1), the alignment score and coverage are much lower than that of MM-align, i.e. TM-score/RMSD/coverage by TM-align-II are 0.71/3.0Å/58.5% versus 0.759/2.65Å/60.4% for MM-align (Table 2.1).

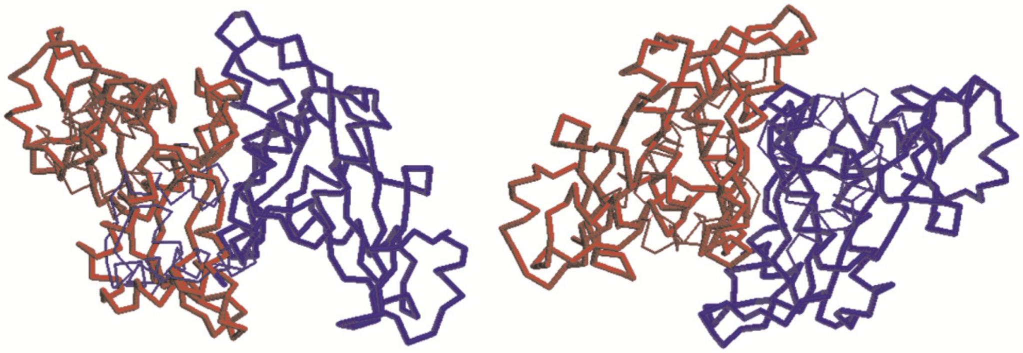

Figure 2.2.Example of cross-aligned chains by TM-align. A typical example structures aligned by TM-align, containing cross-chain alignments (left panel), and the same structures aligned without cross-chain alignment by MM-align (right panel). The two complexes are from PDB files 1u20 (thick trace) and 1y7y (thin trace), with the two chains represented in blue and red, respectively.

Although no-chain-crossing rule is requested in most multimeric complex structure comparisons, there are also occasions where it may not be the case, e.g. aligning protein complexes which involve domain swapping [16]. For dealing with this issue, MM-align has a special option which allows cross-chain alignment between chains when users suspect that domain swapping may be involved (or for any other reason where cross-chain alignment prevention is not required). There are also cases where no one-to-one correspondence is specified between subunits (e.g. gene-fusions [17] or aligning proteolytically cleaved chains to an uncleaved chain).Therefore, another special option of MM-align is setup for aligning one chain to multiple chains. Similarly, the no-chain-crossing rule is taken off by using the normal DP for alignment instead of the modified DP illustrated in Figure 2.2.

2.1.3 Option for Interface Enhanced Alignments

Interface residues are usually related to biological activity, and evolutionarily more conserved than other regions of the protein [18-21]. Matching subunit interfaces is of special importance when complex structures are compared. For protein complexes with obvious structural similarity and consistent interfaces, the normal version of MM-align can align both global structures and interfaces correctly. But when structural similarity is weak, the procedure may place the interfaces arbitrarily along the alignment path. For users interested only in aligning the interfaces of such complexes, MM-align provides an option to optimize the interface match in addition to optimizing the TM-score.

To reinforce the alignment of the interfaces, MM-align assigns a higher weight to the alignment scores and a higher gap penalty if the alignment involves the interface residues as described in Equation 3 and Figure 2.8. For testing this option, we randomly selected 2,000 complex pairs from the 205×3,897 pairs which have a TM-score < 0.4 and an interface coverage $< 10\%$ by the normal MM-align alignment. This set of protein complexes is different from the training protein pairs used to train the parameters as described in Methods. The average fraction of aligned interface residues versus all interface residues is 3.3% in the normal MM-align alignments. After applying the interface-enhancement option, the average fraction of aligned interface residues increases to 14.3%, but the overall TM-score is similar to that without using the interface option (though the global structural match in this TM-score range is not very meaningful).

2.1.4 Functional Relevance of Structure Alignments

The biological function of protein complexes depends on their 3D structures [6, 22]. An important goal of protein structural alignment algorithms is to assist in identifying function-related structural similarities between complexes.

Out of the 205 non-redundant protein complex pairs identified by MM-align, 153 (75%) pairs have related functions as judged by the annotations in the original PDB files and Gene Ontology (GO) [23] annotations. The function of the complexes has been manually assessed by the following procedure: If the “molecular function” GO term of the query and template complexes were the same, they were considered to have the same function. In a few cases, no "molecular function" was associated with a complex; we then looked at the “biological process” GO term. If the “biological process” term was also missing, which occurs quite rarely, we further referred to the “Classification” record in the PDB file. The function of all the 205 complexes could be obtained by this procedure.

Among the 135 protein complex pairs having a TM-score > 0.7, 133 (98.5%) have the same function. 21% of these protein pairs have a sequence identity below 30%. Out of the two complex pairs with different functions, one has a TM-score of 0.959 because both complexes are coiled-coils with very little deviation in structure. The sequence identity of this pair is very low (10%). In the other case, the TM-score is 0.709, but the compared structures are only fragments of their respective proteins rather than the complete complex structures. A complete list of TM-scores, RMSDs, and functional assignments of all 205 complex pairs is presented in Appendix I. These data demonstrate the ability of MM-align to identify structural similarities related to biological function.

In Figure 2.3, three illustrative examples of protein pairs from different protein classes (alpha-, beta- and alpha/beta-proteins); each having a high structural similarity but low sequence identity are presented. The first target complex is from the protein allophycocyanin (PDB ID: 1all), a light harvesting protein [24] found in the cyanobacterium *Spirulina platensis*. Both its chains belong to the "mainly alpha" class in CATH [25], and have an orthogonal bundle architecture and a globin-like topology. The complex selected by MM-align based on TM-score is alpha-phycoerythrocyanin (PDB ID: 2j96), which is also involved in photosynthesis [26] in the thermophilic alga *Mastigocladus laminosus*. According to CATH and SCOP, its both chains have the same architecture and topology as allophycocyanin [27]. The sequence identity of the complex pair is 27% and the TM-score from MM-align is 0.895 (Figure 2.3a).

In the second example, the protein alcohol dehydrogenase from *Drosophila lebanonensis* (PDB ID: 1a4u) has an oxidoreductase activity [28] and its both chains are classified by CATH as alpha-beta proteins having a Rossman fold. The structurally closest complex chosen by MM-align is sorbitol dehydrogenase (PDB ID: 1k2w) from the bacterium *Rhodobacter sphaeroides,* which has the same activity [29] and belongs to the same class and fold according to CATH. The sequence identity of the complex pair is 24% and the TM-score from MM-align is 0.818 (Figure 2.3b).

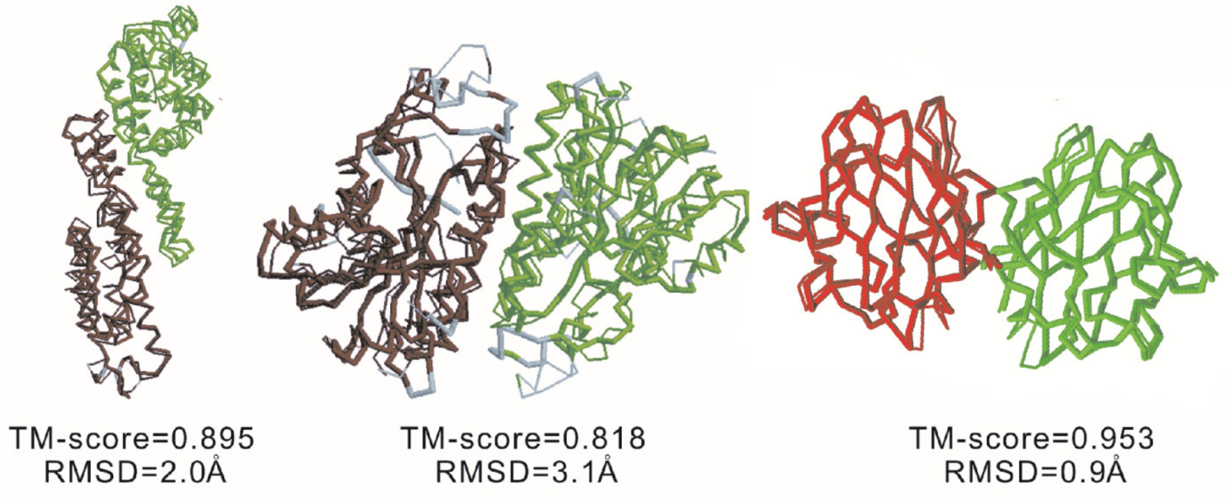

Figure 2.3. Three examples of protein dimeric complex alignments identified by MM-align, from three different protein classes (alpha-, alpha/beta-, and beta-proteins). Thick and thin lines represent the Cα traces of different complexes, and red and green indicate different chains. The grey regions are those with a distance >5Å in the superposition.

The complexes in the third example are two "mainly beta" proteins as classified by SCOP and CATH. The query protein is human copper superoxide dismutase (PDB ID: 1do5), and the best match found by MM-align is copper-zinc superoxide dismutase from *Xenopus laevis* (PDB ID: 1xso). The two proteins share a low sequence identity around 50% but have extremely similar structures with a TM-score of 0.953 (Figure 2.3c). Both have a similar topology and architecture of an immunoglobin-like sandwich according to CATH.

When the structural similarity is very high, functionally related protein pairs may also be identified by the naïve application of TM-align. However, cross-chain alignments may occur, and may lead to incorrect assignment of the protein family. One such example is casein kinase from *Rattus norvegicus* (PDB ID: 1cki). The closest complex identified by TM-align is the calcium binding protein S100P (PDB ID: 1j55) (see Figure 2.4). When we search Set 2 by MM-align, the closest protein complex found is a tyrosine kinase from human (PDB ID: 1fgk). In this example, the aligned complex structures derived from the naïve version of TM-align have a higher TM-score (0.409) than that from MM-align (0.396), but with 26 residue pairs aligned to the wrong chain, which results in an incorrect function assignment. By preventing the cross-chain alignment, MM-align aligns the complex structure correctly and assigns a similar function to it by the structure comparison. Only one chain is aligned by MM-align because of the different chain orientations.

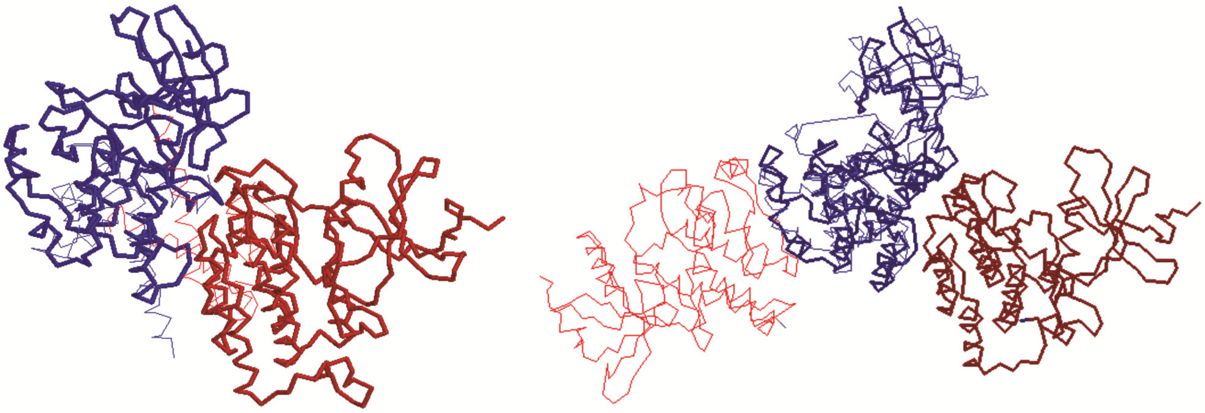

Figure 2.4. The structural alignment of casein kinase (1cki) with its best-matching structures in a non-redundant protein complex library. TM-align picks up human S100P (1j55) with 26 residues aligned across chains (left panel); MM-align picks up the tyrosine kinase domain of fibroblast growth factor (1fgk), without cross-aligned residues.

2.1.5 Alignment of Large Oligomers

One of the important purposes of MM-align is to align large oligomeric proteins. Because the number of solved higher-order complexes in the PDB is much smaller than that of dimers, in Figure 2.5 four examples are shown of MM-align alignments with structures randomly selected from four families of big complexes. These include two with unequal number of chains and two with equal number of chains. The size of the complexes varies from 3 to 20 chains.

Figure 2.5a is an alignment of the photosynthetic reaction center of *Rhodobacter sphaeroides* (PDB ID: 2jiy) with that of *Rhodopseudomonas viridis* (PDB ID: 1dxr), which are randomly selected from the same family of bacteria. 2jiy has three subunits while 1dxr has four (the cytochrome C subunit is extra). Their alignment by MM-align yields a TM-score of 0.669

with the three chains of 2jiy being aligned to the second, third and fourth subunit of 1dxr, respectively. The first chain of 1dxr, which is cytochrome C, remains unaligned.

Figure 2.5b is another example of big complexes with unequal number of chains. The Cytochrome bc1 complex from chicken (PDB id: 1bcc) has 10 chains while the bovine mitochondrial cytochrome bc1 complex (PDB id: 1qcr) has 11 chains. The automated MM-align procedure identified the correct chain combination and generated a structural match of TM-score=0.907 and RMSD=2.7 Å.

Figure 2.5c is an example of complexes with equal chain numbers, which come from phycocyanins in the *Gleobacter violaceus* (PDB id: 2vml) and the red algae *Gracilaria chilensis* (PDB id: 2bv8). Both complexes include 12 protein chains. MM-align correctly selects the chain combination and generates an alignment of TM-score=0.657 and RMSD=2.13 Å.

Figure 2.5d is an alignment of complexes of maximum size by MM-align. The structures come from the bacterial ribosome in *E.coli* (PDB id: 2qbd) and the ribosome of the bacterial species *Thermus thermophilus* (PDB id: 1fjg); both have 20 protein chains. MM-align generate a structure match of TM-score=0.517 and RMSD=4.16 Å. Owing to the large number of possible chain combinations, it takes MM-align nearly 1 hour at a 2.6 GHz AMD processor to generate the best alignment in this example.

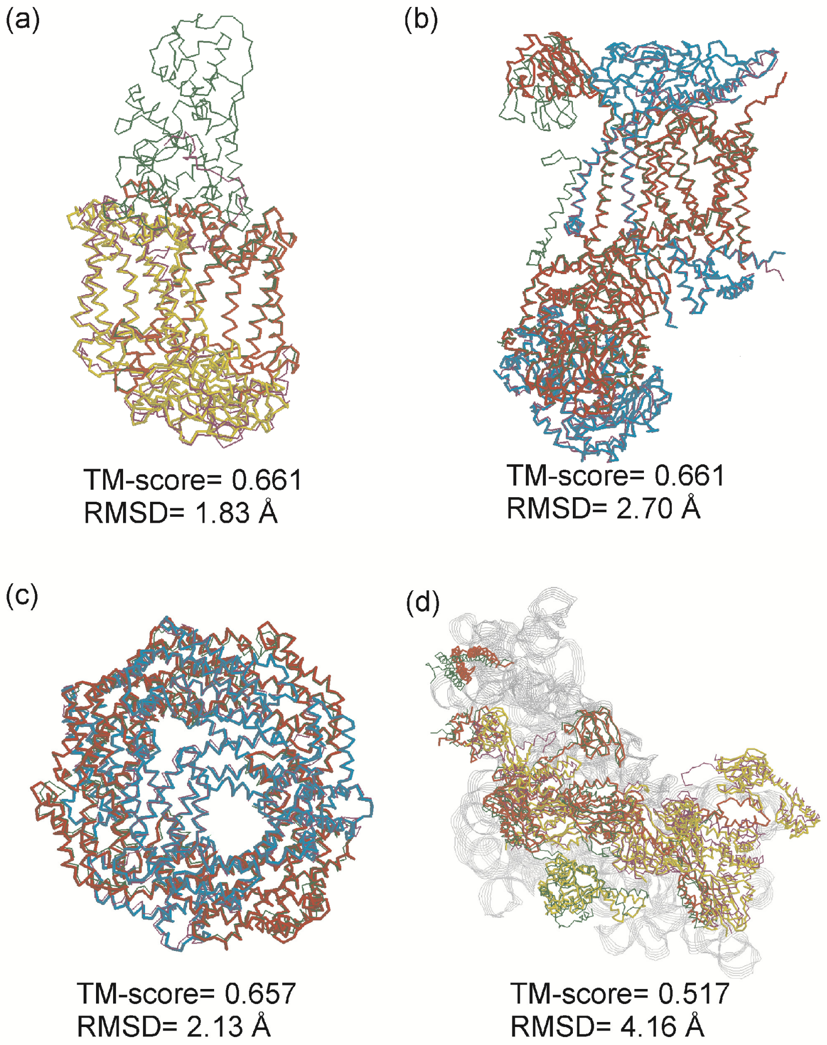

Figure 2.5. Examples of MM-align on big oligomers. (a) Alignment of the photosynthetic reaction center from *Rhodobacter sphaeroides* (PDB id: 2jiy, 3 chains, thick backbone) with that from *Rhodopseudomonas viridis* (PDB id: 1dxr, 4 chains, thin backbone). Yellow, cyan and yellow are for the first, second, and third chains of 2jiy; dark green, magenta, dark green and magents are for the first, second, third and fourth chains of 1dxr. (b) Alignment of Cytochrome bc1 complex from chicken (PDB id: 1bcc, 10 chains, thick backbone) with bovine mitochondrial cytochrome bc1 complex (PDB id: 1qcr, 11 chains, thin backbone). The chains are colored red and cyan alternatively for 1bcc and green and magenta for 1qcr. (c) Alignment of phycocyanin from the *Gleobacter violaceus* (PDB id: 2vml, 12 chains, thick backbone) with phycocyanin from the red algae *Gracilaria chilensis* (PDB id: 2bv8, 12 chains, thin backbone). The chains are colored in red and cyan alternatively for 2vml and green and magenta for 2bv8. (d) Alignment of bacterial ribosome from *E.coli* (PDB id: 2qbd, 20 chains, thick backbone) with ribosome of the bacterial species *Thermus thermophilus* (PDB id: 1fjg, 20 chains, thin backbone). The chains are colored red and yellow alternatively for 2qbd and green and magenta for 1fjg. The grey strands in background are RNA from 2qbd superimposed onto the aligned complexes.

2.2 MATERIALS AND METHODS

For two given protein complex structures containing n and m chains ($n \geq m$), respectively, MM-align starts by generating all possible $P(n,m)=n!/(n-m)!$ permutations for selecting m chains in the first complex. MM-align then proceeds to join the C-terminus of one protein chain with the N-terminus of another chain, in the order generated by the permutation step, and treats the combined artificial chains as rigid-body alignment units (An example of dimeric complexes shown in Figure 2.6).

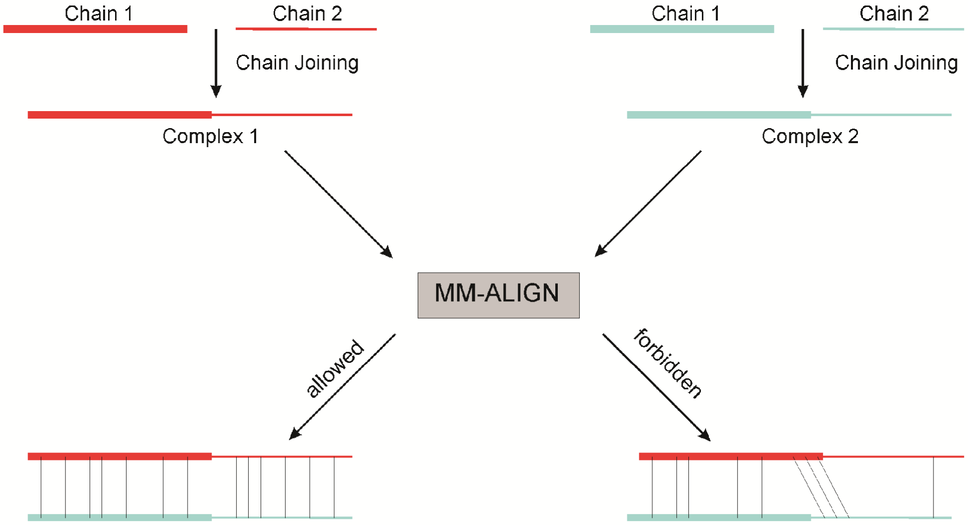

Figure 2.6. An illustration of the chain-joining procedure in MM-align. Both chains of the compared dimers are merged into single artificial chains and then aligned with cross-alignments forbidden. The chains corresponding to each other are presented by the same type of lines (thick and thin). Complex 1 is in red and Complex 2 is blue.

The structural alignment procedure is subdivided into three phases: (1) Selection of chains and chain order for chain-joining; (2) constructing initial alignments; (3) performing the heuristic iteration of the superposition to optimize the TM-score. In general, several alignments are initially constructed, and the inter-complex distance matrix between the superimposed structures is used to guide a heuristic iteration to refine the alignment. The chains are joined in every possible order and the alignment obtained from the order with the highest TM-score is finally returned. For the purpose of saving time in comparing big complexes of more than 3 chains, MM-align first sum the TM-scores obtained from a quick alignment of individual chain pairs and then proceeds with those combinations which have a sum of individual TM-score higher than 90% of the maximum sum of the individual TM-scores.

2.2.1 TM-score for complexes

The TM-score, as mentioned in Chapter 1, was originally defined as a measure to assess the structural similarity of protein monomer chains [15]. Here, the definition was extended to multiple-chain protein complexes, i.e.

$$\mathrm{TM-score} = \max\left[\frac{1}{L}\sum_{i=1}^{L_{ali}}\frac{1}{1+d_{ij}^{2}\big/d_{0}^{2}(L)}\right] \tag{1}$$

where L is the total length of all chains in the target complex and L_{ali} is the number of the aligned residue pairs in the complexes. d_{ij} is the distance between the C□ atoms of the aligned residues i and j after superposition of the complexes, and $d_0(L)$ is given by $d_0 = 1.24\sqrt[3]{L-15} - 1.8$.

2.2.2 Chain Selection and Order of Chain Joining

For a pair of protein complexes with multiple chains, a combinatorial problem arises if the two proteins need to be aligned without cross-chain alignments. For example, consider two proteins containing *n* and *m* chains (*m*<*n*). Then *m* chains need to be selected from the *n* chains of the larger complex, which can be done in *C*(*n*,*m*) = *n*!/(*m*!(*n* —*m*)!) ways. These *m* chains can be joined in *m*! ways, giving rise to a total of *P*(*n*,*m*) = *n*!/(*n*—*m*)! ways of comparison. If the numbers of chains in both proteins are equal, the number of comparisons will become *n*!. When the number of chains is large the number of possible chain orders becomes prohibitably large due to both memory and time constraints (e.g. 10 chains mean more than 3 million possible chain joining). Therefore to limit the number of total comparisons to a treatable range but without missing the meaningful matches, MM-align quickly calculates the monomer TM-score for each chain in the first complex to match with the chains of the second complex based on a modified version of the TM-align program which exploits only the initial alignment from gapless threading. For each chain order, MM-align sums the TM-scores of the monomer chains which have been prescribed to be aligned. If the sum of the TM-scores of the monomer chains is > 90% of the maximum sum of the monomer TM-scores obtained so far from previous steps, it then proceeds further to align the complex as a whole. Otherwise, MM-align discards the particular chain order and moves on to the next order of chain joining. We find that the omission of these low-TM-score joining does not decrease the average performance of MM-align in our testing results.

2.2.3 Initial Alignments

MM-align uses five quickly constructed initial alignments, which are detailed below. (1) An alignment of secondary structure (SS) elements using Needleman-Wunsch (NW) dynamic programming [30], using a score of 1 (0) for matching (non-matching) SS types (helix, strand, or coil) of two aligned residues, and a gap penalty of -1. (2) Gapless alignment of the two structures (i.e. generating all possible gapless alignments by sliding one sequence along the other one with each step jumping 5 residues; the best alignment is selected on the basis of TM-score). Moreover, if the TM-score of any of the gapless alignments is greater than a cutoff (i.e. > 95% of the maximum TM-score obtained so far), the alignment is further optimized by dynamic programming, and the alignment with the highest TM-score is selected. It is observed that the implementation of DP helps in generating much better starting alignments. But since only high-scoring gapless alignments are selected to do DP, this procedure does not increase the overall CPU time of the MM-align algorithm. (3) An alignment from dynamic programming where the score matrix is a half/half combination of the SS score matrix and the distance score matrix extracted from the second initial alignment. The gap-opening penalty is set to -1. (4) The fourth initial alignment is also gapless threading but the superposition of the structures is restricted to the longest continuous segments in each complex. This initial alignment is added because the second initial alignment could miss the best superposition when the joined chains have gaps (chain breaks) in the structure. This is especially the case when the algorithm is used to align interface structures that consist of chain fragments. (5) A fragment of 5 continuous residues starting from the N-terminus of one protein is superimposed onto a similar fragment of 5 residues starting from the N-terminus of the second protein. The global TM-score is quickly calculated based on the rotation matrix of the 5-residue fragments. If the TM-score is higher than 12% of the best TM-score obtained from the previous superimpositions, a DP alignment is performed to refine the initial alignment using the inter-residue distances from the initial superposition. The procedure is repeated for all 5-residue fragments of either protein and the best alignment based on TM-score is finally selected. For saving CPU time, however, we skip those 5-residue fragment pairs which do not have similar secondary structure content.

Compared with TM-align, the last two initial alignments are new and the initial alignment (2) is improved by the additional DP iteration. These changes result in considerable improvement of the search engine of TM-align. In a benchmark test of aligning 4000 monomer pairs, TM-score increased in 1337 cases, while the total CPU cost is kept essentially unchanged.

To prevent cross-chain alignment in the initial alignments, the conventional NW algorithm [30] was altered so that regions in the DP matrix corresponding to cross-chain alignment are ignored as shown in Figure 2.7. For example, if chains 1 and 2 of Complex 1 are to be aligned to chains 1 and 2 of Complex 2 respectively, the DP matrix regions corresponding to aligning chain 2 of Complex 1 with chain 1 of Complex 2 are omitted when filling up the alignment paths during DP (an example of aligning a three-chain complex pair shown in Figure 2.7). The filling up of the DP matrix can be considered as a three-step process: 1) The region corresponding to the first chain (by the order prescribed by the chain joining step) of both complexes is filled up. 2) A pseudo-layer uniformly assumes the value of the last cell of the preceding block; by doing this the gap extension penalty will be ignored at the respective first residues of the second chains. 3) The region corresponding to the second chains (as per the order of chain joining) of both complexes is now filled up starting from the pseudo-layer values (instead of 0 which is used as the initial value for the first block). The process is repeated when aligning complexes with more than two chains.

While tracing back the pathway, the reverse order is followed and the traceback is started in the region corresponding to the last chain of both complexes, crossing the junction of the diagonal blocks, and then continuing the traceback in the area corresponding to the "next to last" chains of both complexes. Traceback continues until the first residue of the first chain of both complexes is reached. MM-align thus avoids the cross-alignment zones completely, and forces the alignment to traverse a path which does not lead to alignment of any residue of chains not prescribed to be aligned for that particular iteration. An illustration of the modified DP for a trimer is presented in Figure 2.7. An alternative treatment would be to employ a large penalty for cross-aligned regions, which is, however, more CPU-expensive because of the filling and backtracing procedures in the forbidden areas.

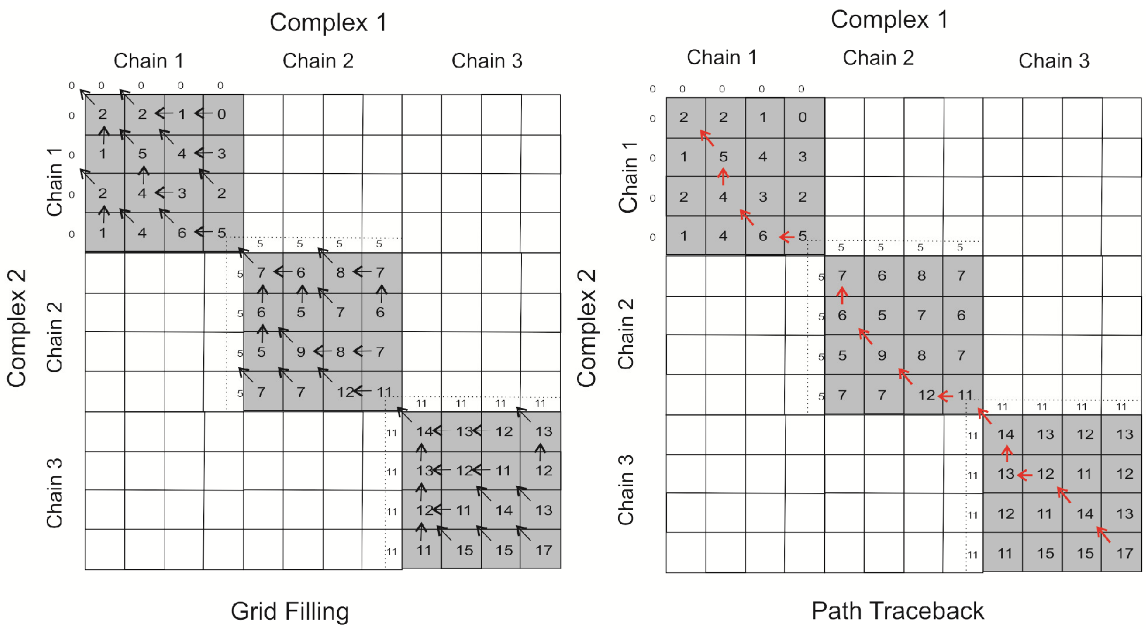

Figure 2.7. An illustration of the modified dynamic programming algorithm with cross-chain alignment prevented. The picture on the left panel illustrates the process of filling up the grid, with the cross-alignment zones (empty grids) ignored. The dashed lines represent a pseudo-layer which assumes the value in the last cell of the preceding block. The values of the pseudo-layer (5 and 11 in this example) are used as starting score of the next block corresponding to the next chain of both complexes. The picture on the right panel shows the traceback path (indicated by red arrows).

The five initial alignments thus derived are passed on to the heuristic iteration phase for further refinement.

2.2.4 Heuristic Iterations

Once an alignment is obtained, the structures of the two complexes can be spatially superimposed by the TM-score rotation matrix [15]. Based on the superimposed structures, a similarity scoring matrix is defined as

$$S_{ij} = \begin{cases} \dfrac{1}{1 + d_{ij}^2 / d_0^2(L_{\min})}, & \text{if } i \text{ and } j \text{ are aligned without cross} \\ \text{ignored,} & \text{if } i \text{ and } j \text{ are aligned with cross} \end{cases} \quad (2)$$

where d_{ij} is the same as that defined in Eq. (1). $d_0 = 1.24\sqrt[3]{L_{\min} - 15} - 1.8$, and $L_{\min}$ is the total length of the smaller complex. The purpose of using $L_{\min}$ instead of the target length (L) here is to avoid the asymmetry resulting when aligning Complex 1 to Complex 2 versus Complex 2 to Complex 1. Like in Figure 2.7, we omit the residue pair when i and j are from cross-aligned chains.

A new alignment can be generated based on the score matrix of Eq. (2) by the modified NW dynamic programming as explained in Figure 2.7, with an optimal gap-opening penalty of -0.6. Based on the new alignment, MM-align superimposes the complex structures by the TM-score rotation matrix again, which gives rise to a new similarity scoring matrix and can again be used for the modified NW dynamic programming. The procedure is repeated a number of times until the alignment between two protein complexes becomes stable. The alignment with the maximum TM-score encountered during the iterations starting from the five initial alignments is returned as the final alignment.

Because the score matrix of Eq. (2) is consistent with the target function of TM-score of Eq. (1), the iteration converges very fast, and usually 2-3 iterations are enough to find the best alignment. As we are mainly interested in the topological match between the compared complexes, no gap extension penalty is applied.

2.2.5 Preferential Alignment on Interfaces

The structures of protein-protein interfaces are usually more conserved than other regions, and generally have special importance in the inference of biological function [18]. The MM-align program has a special option for preferentially aligning the interface residues of dimers, which constitutes the largest subgroup of multimeric protein complexes.

For the given dimer structures, the interface residues are defined using a default Cα distance cutoff of 8 Å (a different value can optionally be specified by the user), i.e. any residue whose Cα atom is at a distance <8 Å from any Cα atom in the other chain of the complex is considered to be an interface residue. The alignment of the interface residues can be enhanced by a modified dynamic programming scheme where the alignment path is defined by

$$P(i,j) = \max\{P(i-1,j-1) + wS_{ij}, \max_{k\geq 1}\{P(i-k,j) + xGP_k\}, \max_{l\geq 1}\{P(i,j-l) + xGP_l\}\} \tag{3}$$

where $P(i, j)$ is the maximum score of an alignment path ending at (i, j) and $GP_k<0$ is the normal gap penalty. Because we have no gap extension penalty, GP_k actually does not depend on k. For non-interface residue pairs, w=x=1. If both i and j are from interfaces, w>1 is used to encourage the alignment of the interfaces and x>1 to discourage gaps at the interfaces (see Figure 2.8). The gap penalty is always neglected at the boundary of two chains.

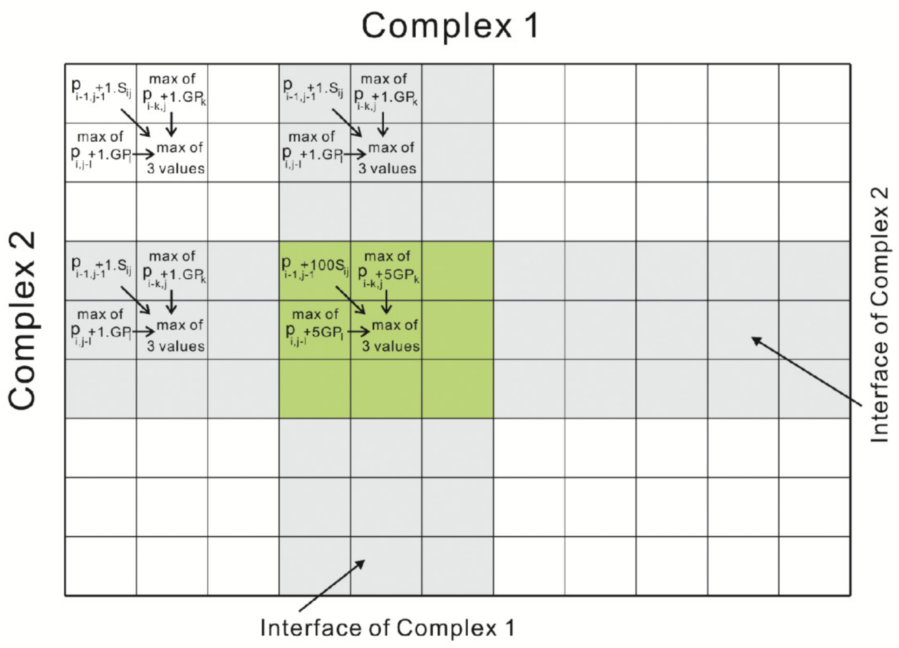

Figure 2.8. A modified dynamic programming scheme with the alignment of interface residue pairs reinforced. The interface areas are highlighted in color. If the residue pairs are both from an interface (the area in green), the score is increased by a factor w and the gap penalty is increased by a factor x.

Since interface alignment is most important for complex pairs with weak structural similarity (see above), we optimized the parameters of *w* and *x* based on 2,000 complex pairs with TM-scores <0.3 and interface alignment coverage below 10%. In general, higher values of *w* and *x* will increase the number of aligned interface residues but too large values will reduce the TM-score of the overall alignment. After a comprehensive grid search of the parameter space, we found that *w=100* and *x=5* work the best for generating the highest number of aligned interface residue pairs while still maintaining a reasonable TM-score of global alignments.

2.3 DISCUSSION

A new algorithm, MM-align, was developed for quickly aligning and comparing the structures of multiple-chain protein complexes. Bearing in mind the importance of protein-protein interactions in structural biology studies, and the lack of computer algorithms dedicated to multimeric structure alignments, the MM-align method is expected to be of immense use across many aspects of the field. The algorithm performs simultaneous alignment of all chains of protein complexes with both the monomer similarity and the relative chain-orientations accounted for by a single TM-score. The biologically irrelevant cross-chain alignments are efficiently prevented by the implementation of a modified dynamic programming algorithm which ignores the cross-alignment blocks of the DP matrix while filling up the cells and tracing back the pathway. This results in halving the necessary CPU time. Because of the consistency of the rotation matrix and the objective function, the convergence of the heuristic iteration stage is fast. For aligning a pair of protein dimers of 400 residues each, the average CPU cost is 0.35s on a 2.6 GHz AMD processor.

The algorithm also includes a user-specified option to reinforce the structural alignment in the interface regions. The default weight for aligned interface residues has been carefully optimized using a benchmark set, balancing the overall topology match and the accuracy of interface alignment. Higher weights would result in aligning a higher number of interface residues but would, on average, deteriorate the overall structure match. This option is especially useful when the global structural match is inconsistent with the interface similarities but the user is interested in the interface match. In cases where there is reason to believe that prevention of cross-chain alignment is not desirable (e.g. complexes involving domain swapping or gene fusion), MM-align has a special option to utilize normal DP which does not prevent cross-chain alignment. It also allows alignment between one chain with multiple chains.

Noting the fact that proteins often function as complexes, a functional annotation study based on the conserved complex structures is relevant. In a test on 205 non-homologous proteins, MM-align was able to detect functionally similar proteins within a non-complete benchmark dataset of 3,897 complexes. It often prevents false positives that may arise when dimer structures are aligned with tools dedicated to single-chain alignments only, like TM-align. MM-align also has the capability of aligning large multimeric complexes up to 20 chains and correctly identifying the corresponding subunits and the structure match. These data show that MM-align may serve as an effective function annotation tool if used for querying a complete library such as all complexes in the PDB. Because MM-align provides a single TM-score describing the global similarity of the complexes, it can also be conveniently used for automated and quantitative classification of protein complex structures which forms the focus of Chapter 3. An online MM-align server and the source code of the program are freely available at: http://zhang.bioinformatics.ku.edu/MM-align.

2.4 REFERENCES

1. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

2. Douguet, D., et al., *DOCKGROUND resource for studying protein-protein interfaces.* Bioinformatics, 2006. **22**(21): p. 2612-8.

3. Henrick, K. and J.M. Thornton, *PQS: a protein quaternary structure file server.* Trends Biochem Sci, 1998. **23**(9): p. 358-61.

4. Arakaki, A.K., Y. Zhang, and J. Skolnick, *Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment.* Bioinformatics, 2004. **20**(7): p. 1087-96.

5. Graille, M., et al., *Structure-based functional annotation: yeast ymr099c codes for a D-hexose-6-phosphate mutarotase.* J Biol Chem, 2006. **281**(40): p. 30175-85.

6. Zhang, Y., *Protein structure prediction: When is it useful?* Corr Opin Struct Biol, 2009: p. doi:10.1016/j.sbi.2009.02.005

7. Janin, J., et al., *CAPRI: a Critical Assessment of PRedicted Interactions.* Proteins, 2003. **52**(1): p. 2-9.

8. Vajda, S. and C.J. Camacho, *Protein-protein docking: is the glass half-full or half-empty?* Trends Biotechnol, 2004. **22**(3): p. 110-6.

9. Holm, L. and C. Sander, *Dali: a network tool for protein structure comparison.* Trends Biochem Sci, 1995. **20**(11): p. 478-80.

10. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Eng, 1998. **11**(9): p. 739-47.

11. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res, 2005. **33**(7): p. 2302-9.

12. Zhu, H., et al., *Alignment of non-covalent interactions at protein-protein interfaces.* PLoS ONE, 2008. **3**(4): p. e1926.

13. Mintz, S., et al., *Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions.* Proteins, 2005. **61**(1): p. 6-20.

14. Shulman-Peleg, A., et al., *MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W260-4.

15. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**(4): p. 702-10.

16. Bennett, M.J., S. Choe, and D. Eisenberg, *Domain swapping: entangling alliances between proteins.* Proc Natl Acad Sci U S A, 1994. **91**(8): p. 3127-31.

17. Mitelman, F., B. Johansson, and F. Mertens, *The impact of translocations and gene fusions on cancer causation.* Nat Rev Cancer, 2007. **7**(4): p. 233-45.

18. Bogan, A.A. and K.S. Thorn, *Anatomy of hot spots in protein interfaces.* J Mol Biol, 1998. **280**(1): p. 1-9.

19. Pawson, T. and P. Nash, *Protein-protein interactions define specificity in signal transduction.* Genes Dev, 2000. **14**(9): p. 1027-47.

20. Phizicky, E.M. and S. Fields, *Protein-protein interactions: methods for detection and analysis.* Microbiol Rev, 1995. **59**(1): p. 94-123.

21. Valencia, A. and F. Pazos, *Computational methods for the prediction of protein interactions.* Curr Opin Struct Biol, 2002. **12**(3): p. 368-73.

22. Wilson, C.A., J. Kreychman, and M. Gerstein, *Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.* J Mol Biol, 2000. **297**(1): p. 233-49.

23. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

24. Brejc, K., et al., *Isolation, crystallization, crystal structure analysis and refinement of allophycocyanin from the cyanobacterium Spirulina platensis at 2.3 A resolution.* J Mol Biol, 1995. **249**(2): p. 424-40.

25. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-108.

26. Schmidt, M., et al., *Structural basis for the photochemistry of alpha-phycoerythrocyanin.* Biochemistry, 2007. **46**(2): p. 416-23.

27. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.

28. Benach, J., et al., *The refined crystal structure of Drosophila lebanonensis alcohol dehydrogenase at 1.9 A resolution.* J Mol Biol, 1998. **282**(2): p. 383-99.

29. Philippsen, A., et al., *Structure of zinc-independent sorbitol dehydrogenase from Rhodobacter sphaeroides at 2.4 A resolution.* Acta Crystallogr D Biol Crystallogr, 2005. **61**(Pt 4): p. 374-9.

30. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.

CHAPTER 3. Protein-protein complex structure predictions by multimeric threading and template recombination

While rapid progress has been made in protein tertiary structure prediction [1-3], the challenges in generating atomic level protein quaternary structures from amino acid sequence has remained relatively unexplored [4-7]. The effort in complex structure modeling has been mainly focused on rigid-body docking of monomer structures [8-12], with success often depending on size and shape complementarity of the interface area, and the hydrophobicity of interface residues [4]. One of the major challenges in protein-protein docking is the modeling of binding-induced conformational changes [7, 13-14] in which some progress has recently been made with the development of new docking methods, e.g. SnugDock [15], MdockPP [16], ATTRACT [17] and others. Progress in this area was also observed in the recent community-wide docking experiments, CAPRI [7, 18-20]. However, as an inherent limit, protein-protein docking can be performed only when the structures of the component monomers are known. The second way of constructing protein-protein complex structures is through homology modeling which has attracted considerable attention in recent years [21-23] as reviewed in Chapter 1.

Here, a new method, COTH, is presented for protein-protein complex structure prediction, based on co-threading the sequences of both chains simultaneously through the protein quaternary structure library. To boost the capacity of the protein complex library, a monomer-based threading was performed in parallel through the tertiary structure library with the resultant alignments shifted to the complex framework by structure alignments. A new *ab initio* interface predictor, BSpred, was developed to adjust the complex alignment. The algorithm has been tested on two large-scale bound and unbound benchmarks to examine the strength and weakness in comparison with the conventional rigid-body docking and homology modeling methods, which demonstrated promising new avenues for protein complex structural predictions.

3.1 RESULTS

3.1.1 Overall results of COTH on testing proteins

The COTH protocol consists of three consecutive steps: 1) Dimeric threading through a multiple-chain complex structure library for chain orientation prediction (called “COTH threading” throughout the article); 2) single-chain threading through tertiary structure library; 3) recombination of tertiary templates and model selection of complex structures (Figure 3.1).

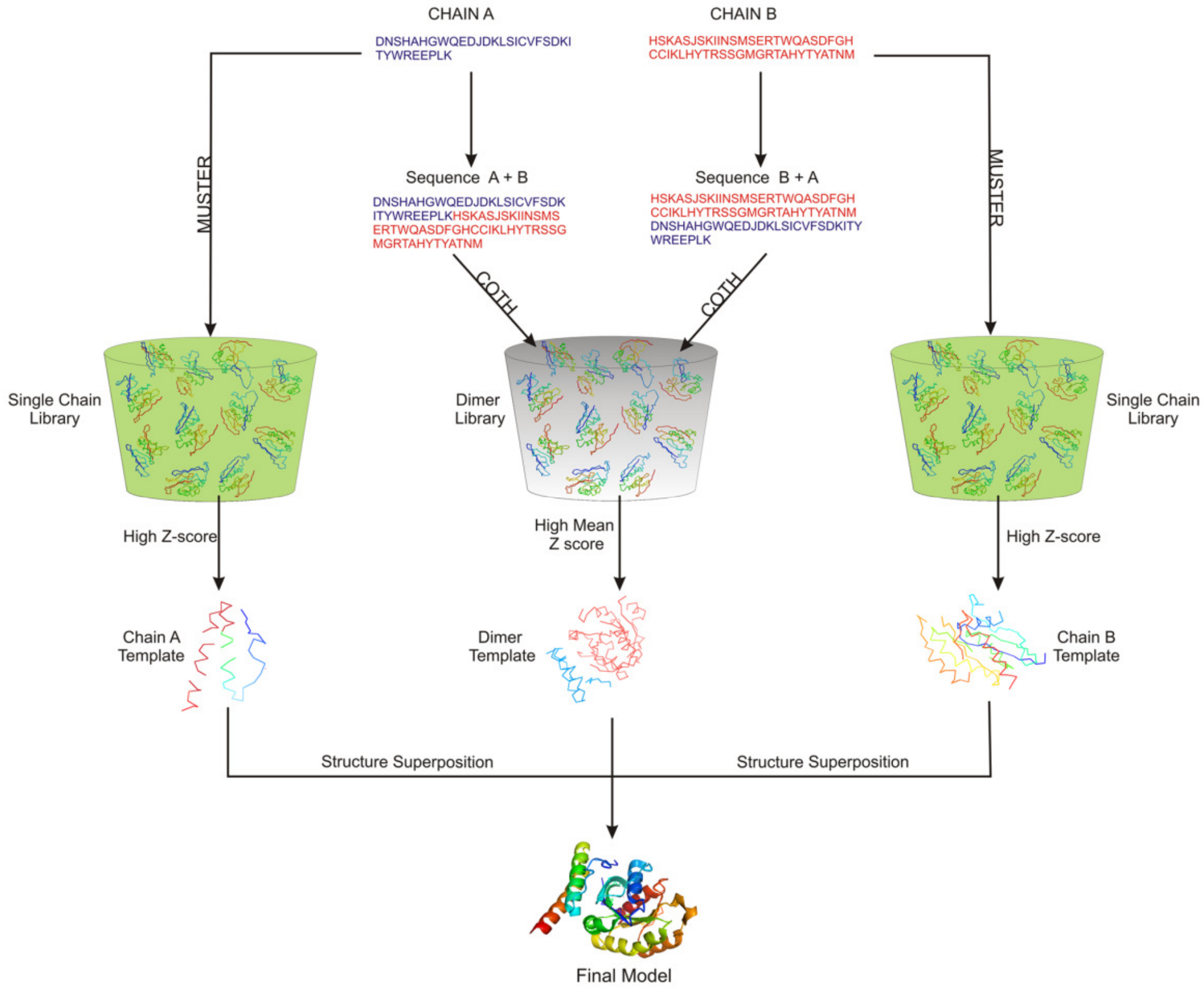

Figure 3.1: Flowchart of the COTH algorithm for protein complex template identification. The sequences are first joined in both permutation and threaded against a complex structure library to identify complex templates. Both monomer chains are individually threaded by MUSTER against the tertiary structure library to obtain tertiary structures. The monomer templates are then structurally superposed to the dimer template to generate the final template models. To avoid confusion, a list of the programs described in this article is presented in Table 3.1.

Name	Description
C-PPA	A multiple-chain threading algorithm with scoring function including profile-profile and secondary structure matches. It is an extension of the PPA algorithm for monomer threading [24].
C-MUSTER	A multiple-chain threading algorithms with scoring function including similar terms to C-PPA, plus multiple structure-based terms derived for torsion-angle and structural profile matches. It is an extension of the MUSTER algorithm for monomer threading [25].
COTH threading	A multiple-chain threading algorithm with scoring function including similar terms to C-MUSTER, plus the binding site match. The binding sites for targets are predicted by BSpred.
COTH	Models are generated by combining the tertiary templates from MUSTER with the quaternary templates from COTH-threading through structure superposition.
COTH-exp	Models are generated by superimposing the experimental unbound monomer structures onto the templates from COTH-threading.
COTH-model	Models are generated by superimposing the full-length monomer models onto the templates from COTH-threading. The monomer models were predicted by MUSTER with loops filled by MODELLER.
ZDOCK-exp	Models are generated by ZDOCK which docks the experimental unbound monomer structures followed by RDOCK refinement.
ZDOCK-model	Models are generated by ZDOCK which docks the full-length monomer models predicted by MUSTER and MODELLER, followed by RDOCK refinement.

Table 3.1. Naming conventions of methods described.

To test COTH, a non-redundant set of 500 dimeric proteins from the PDB was constructed, which is also non-redundant to (below 30% in sequence identity with) the 180 training proteins used in algorithm optimization (Materials and Methods). A list of the testing and training proteins is shown at http://zhanglab.ccmb.med.umich.edu/COTH/proteinlist.html. When COTH is executed, all homologous templates, which have a sequence identity > 30% or are detectable by PSI-BLAST with an E-value < 0.5 to the query, are excluded from both dimer and monomer template libraries. These criterions are widely used in protein structure predictions for excluding homologous templates [26-27].

Evaluation of the global template quality is mainly carried out by TM-score [28], complex RMSD, and the alignment coverage. TM-score has been extensively used for quality assessment of protein structure predictions because of its ability in combining alignment accuracy and coverage. TM-score was originally developed for comparing monomers. To calculate the TM-score of dimer models, we first convert the dimer structure into an artificial monomer by connecting the C-terminal of the 1st chain and the N-terminal of the 2nd chain, and then run the TM-score program with the length of the query complex sequence as the normalization scale (as described in Chapter 2). This definition of complex TM-score has the value beween [0, 1] and is sensitive to both the topology of individual chain structures and the relative orientation of two components. In general, either the incorrect component structure or the wrong orientation of the components will result in low TM-score. In other words, a high complex TM-score means the correct modeling of both individual chain structures and the relative orientation [29].

In Figure 3.2a, RMSD versus alignment coverage of the first COTH models is shown. Here, RMSD means the root-mean-squared-deviation of the threading model and the native structure in the threading aligned region (unless specified, RMSD indicates the global complex RMSD throughout the Chapter). Even though all homologous templates are excluded, COTH identified notable templates from non-homologous proteins. For example, there are 269 cases (or 293 in the top 10 models) which have the first template with a TM-score > 0.4. The average sequence identity between template and query is only 21.2% for the 269 proteins. Despite the low sequence identity, the average alignment coverage is 85.1% and the average RMSD to the native is 5.9 Å in the aligned regions. This demonstrates the ability of COTH to identify non-homologous templates. Alternatively, if templates with an RMSD < 6.5Å and alignment coverage > 70% are considered to be reliable, 272 out of 500 targets have reliable templates in the best in top 10 predictions. In Figure 3.2b the distribution of TM-score of the first templates is shown. The majority of targets have templates with a TM-score >0.3, which is significantly higher than the random template selection (TM-score <0.17) [28]. In cases where TM-score is in the 0.3-0.4 range, targets often have only the chain orientation correctly predicted but with substantial regions of monomer structures missing or wrongly aligned. This provides opportunities for improvement by further structure refinement based on monomer structure recombination as explored in Chapter 4.

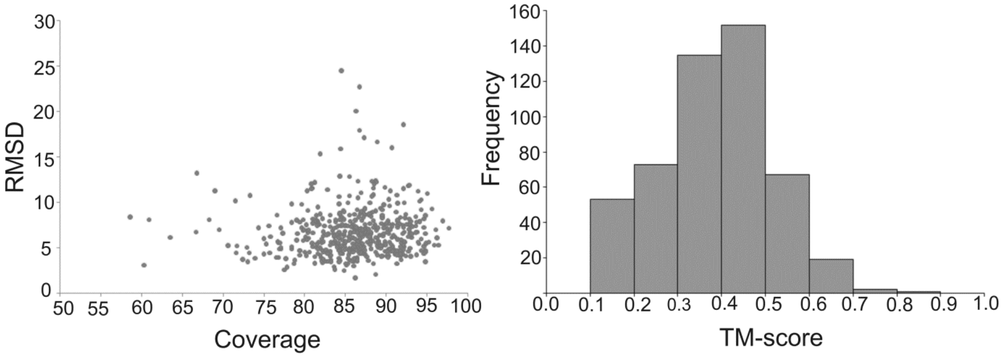

Figure 3.2: Complex threading results by COTH on 500 non-redundant test proteins. a) RMSD versus alignment coverage for the best in Top 10 models. b) Histogram of TM-score for the first model.

There are 39 cases, however, which are all hard cases (i.e. with a significance score of alignments relative to the random, Z-score < 2.5) (see Figure 3.7), where the TM-score of individual ligand and receptor templates are > 0.5 but the complex TM-score is < 0.4. In these cases, though the quality of the individual chains is good, their predicted orientation is incorrect. The average accuracy for the interface prediction by BSpred is, as expected, poor at only 42.3% with coverage of 14.9%. It should be noted that among the 39 targets, 21 cases do have templates of correct orientations with a TM-score > 0.4 as identified from the complex library by our complex structural alignment algorithm MM-align [30] when using the native structure as the probe. Thus, improvement of the accuracy of BSpred in binding-site predictions is essential to recognizing the correct chain orientations for these cases.

Other than the TM-score of the global complex structure, an assessment is made about the modeling quality of protein-protein interface structures, the quality of which is of key importance for the functional annotation of protein complexes. Here, a residue is defined to be at the interface if the distance of the Cα atom to any Cα atoms in the opposite chain is below 10 Å. The interface RMSD, I-RMSD, is the root-mean-squared-deviation of the model and the native structure in the aligned region of the interfaces. The interface coverage, I-cov, is the ratio of the threading aligned interface residues divided by the total number of interface residues in the target. For the 500 targets, the average I-RMSD and I-cov is 12.9 Å and 61.1%, respectively, for the best in the top-5 models (Table 3.2). This high I-RMSD value is partly due to a few hard cases, which have a very high I-RMSD (> 25 Å) because of completely wrong alignments. If a successful threading "hit" is defined as the model which has an I-RMSD ≤ 5.0 Å with at least 50% of the interface residues aligned, there are 186 cases in which COTH generates at least one hit in the top-5 models, despite the exclusion of homologous templates, which represents 37% of the overall sample.

Enzyme-ligand and antigen-antibody are two major complex classes found predominantly in nature. In the testing set there are 236 enzyme-ligand complexes and 169 antigen-antibody complexes. The first COTH templates for enzyme-ligand have an average TM-score of 0.441, and an average RMSD 4.1 Å with alignment coverage 86.2%. For the antigen-antibody complexes, the COTH models have an average TM-score of 0.410, and an average RMSD of 4.6 Å across 86.3% residues. There is a tendency that COTH performs better on enzyme-ligand complexes than antigen-antibody complexes, which somewhat surprisingly coincides with that of rigid-body docking methods which also performs better on average at docking the enzyme-ligand structures because of the inherent shape complementarity in the complex structures while antigen-antibody interactions have usually larger backbone and side-chain variations at the interfaces [4, 9, 11, 13-14, 31-33]. For COTH however, the higher TM-score is mainly due to higher conservation of the enzyme-ligand sequences while antigen-antibody complexes can vary greatly in the sequence space. In the test set proteins for example, the average number of sequence homologies as identified by PSI-BLAST from non-redundant sequence databases is 3.12 for enzyme-ligand complexes, which is about two-fold higher than that of antibody-antigen (1.67). This therefore allows on average a better construction of sequence profiles for COTH. It should be noted however, that for both our test set proteins and the proteins in the template library, the complexes are represented as dimers although more often than not the antigen-antibody complexes are trimers (the heavy chain and light chain of the antibody and the antigen chain). So, by antigen-antibody complexes here only one chain of the antibody (the light chain or the heavy chain) and the antigen chain is modeled each time, and the result shown is the average of all antibody chains with the antigen.

3.1.2 Comparison of different alignment algorithms

To have an objective control of the COTH performance, it is compared with other template alignment algorithms which are implemented on the same template library and with the same sequence identity cutoffs. Despite a number of published template detection algorithms, due to the lack of publicly available web-servers or downloadable programs which are capable of predicting protein complex structures based on homology modeling, here we focus our comparison mainly on PSI-BLAST and several in-house developed programs.

First, PSI-BLAST, a widely-used tool to identify evolutionarily related proteins through iterative sequence-profile alignments [34] was used as control. PSI-BLAST was run on our complex structure library after joining the two chains together using the BLASTPGP program. The templates were ranked according to the PSI-BLAST E-value. Figure 3.3A shows a comparison of the templates detected by PSI-BLAST and C-PPA, where the latter is a profile-profile alignment method assisted by secondary structure predictions from PSI-PRED [35]. In 71% of cases, the C-PPA templates have a higher TM-score than that by PSI-BLAST. The major difference between these two methods is that PSI-BLAST only uses the template sequence while C-PPA uses sequence profile from multiple sequence alignments to represent the templates in the profile-profile alignments, which often contain additional motif conversation signals that aids in the detection of weak evolutionary relationships. Another reason is that C-PPA uses predicted secondary structures (with an accuracy >80%) to assist in adjusting local secondary structure alignments.

To test the usefulness of additional structure information in complex template identification, C-MUSTER which is a dimeric threading algorithm extended from the monomer threading MUSTER program [25] was developed and tested. In addition to the profile-profile and secondary structure matches as implemented in C-PPA, C-MUSTER contains multiple structural features predicted from sequences. Figure 3.3B shows a head-to-head comparison of C-MUSTER and C-PPA. There are obviously more cases (389 versus 92) which are above the diagonal line. The reason for the improvement is that even though sometimes no obvious sequence similarity exists between two proteins, they may share a similar structural framework. Thus, the use of solvent accessibility, torsion angles, structural profile, and hydrophobicity predictions provides insight into the structure of two proteins.

The major difference between C-MUSTER and COTH threading is that COTH threading contains binding site matches from a neural network based prediction algorithm, BSpred. For the 500 testing proteins, the average accuracy of the binding site prediction is 66.8% with coverage of 14.2%. This accuracy is significantly higher than that of random predictions (34.2%) with a p-value $<10^{-5}$. Figure 3.3C shows the comparison of C-MUSTER versus COTH threading. Overall there are 311 cases which have a higher TM-score in the COTH threading alignment than that in C-MUSTER, demonstrating the usefulness of adding the binding-site predictions.

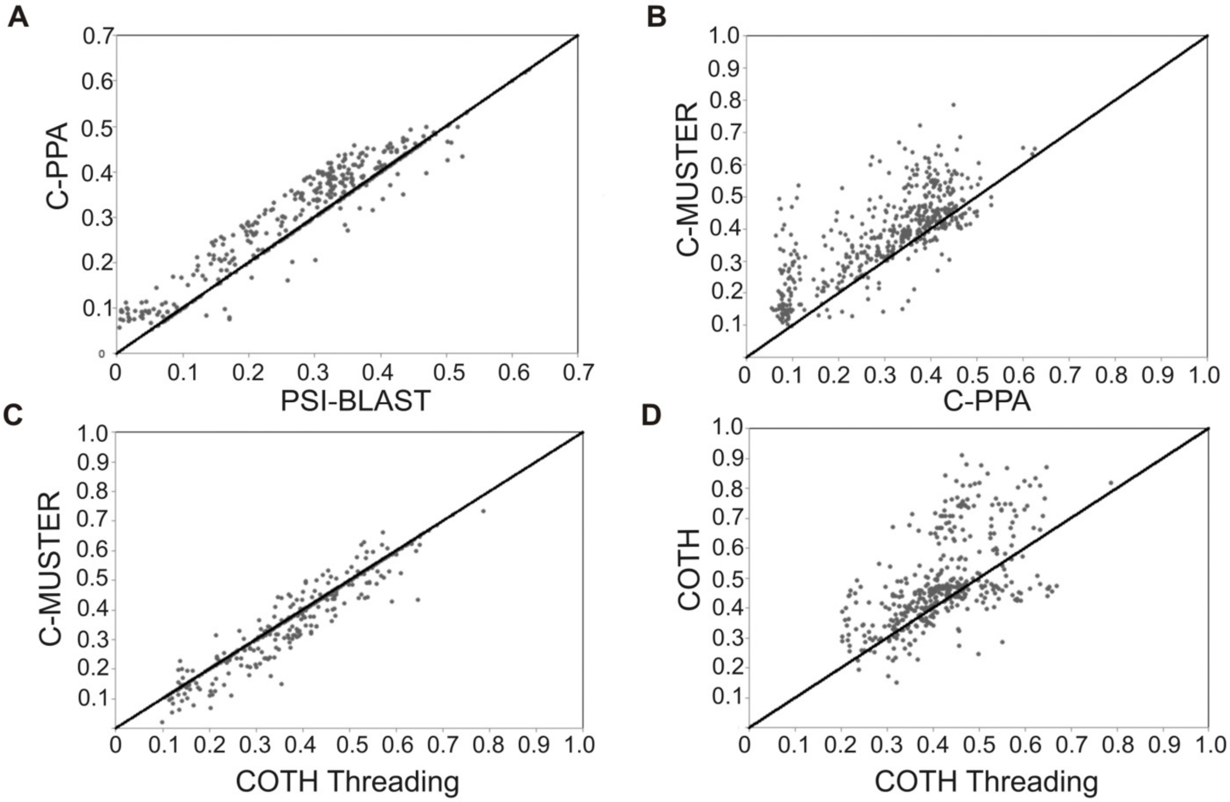

Figure 3.3: Comparison of TM-score of the complex templates as identified by different threading methods. a) C-PPA versus PSI-BLAST; b) C-MUSTER versus C-PPA; c) C-MUSTER versus COTH threading; d) COTH threading versus COTH.

Table 3.2 summarizes the average TM-score, RMSD, alignment coverage, I-RMSD, I-cov and the number of hits of the template models identified by different methods (PSI-BLAST, C-PPA, C-MUSTER, COTH threading). Compared with PSI-BLAST, C-PPA identifies templates of higher coverage (64.8% versus 63.3%) but with significantly lower RMSD (5.43 Å versus 8.19 Å) which results in a 20% increase in TM-score for the first model. Correspondingly, COTH threading identifies better templates than C-MUSTER and C-PPA in both accuracy and coverage. Overall, the TM-score of COTH threading (0.394) is 46% higher than that by PSI-BLAST (0.269) and there are dominantly more cases with higher TM-score in COTH (427) than in PSI-BLAST. The interface accuracy of the COTH threading is also much higher than PSI-BLAST as indicated by the I-RMSD and I-cov (12.6 Å/55.8% vs. 13.7 Å/42.9%). Again, if a hit is defined as a model with I-RMSD < 5 Å with I-cov > 50%, the number of hits in the COTH threading models is 168 which is 35% higher than that by PSI-BLAST (124).

Methods	**TM-score (first/best in top 5)**	**RMSD (coverage)**[1]	**N_{Hit}**[1]	**I-RMSD/I-cov**[1]
PSI-BLAST	0.269/0.293	8.19 Å (63.3%)	124	13.7 Å (42.9%)
C-PPA	0.321/0.334	5.43 Å (64.8%)	145	13.1 Å (49.3%)
C-MUSTER	0.381/0.412	4.51 Å (69.8%)	161	12.8 Å (54.6%)
COTH threading	0.394/0.421	4.45 Å (71.0%)	168	12.6 Å (55.8%)
COTH	0.438/0.477	4.30 Å (77.6 %)	186	12.9 Å (61.1%)

[1] Data are shown as the best in top 5 models.

Table 3.2. Template identification by different methods on 500 testing proteins.

Figure 3.4 is a typical example of a dimer structure (PDB ID: 16gsA0-16gsB0), which reflects the difference of alignments identified by the different methods. First, both PSI-BLAST and C-PPA identify 2c8uA0-2c8uB0 as the best template but C-PPA produces a more accurate alignment and an increased coverage (57.2% for PSI-BLAST and 65.4% for C-PPA) which accounts for the rise in TM-score from 0.523 to 0.602. C-MUSTER identifies 1k3oA0-1k3oB0 as the top template with a sequence identity 25% to the query sequence 16gsA0-16gsB0, which leads to an overall higher coverage 89.9% and a much improved TM-score 0.786. COTH threading, on the other hand, chooses a different protein 1gtaA1-1gtaA2 as the highest scoring template with alignment coverage of 94%; the resulting template has a maximum TM-score of 0.818. This better template selection is mainly due to the BSpred binding-site prediction which has an accuracy of 79.4%. The orientation of 1gtaA1-gtaA2 is more similar to the query protein than 1k3oA0-1k3oB0 as identified by the BSpred prediction, which predicts 31 interface residues of the query 16gsA0-16gsB0 and leads to a better alignment reflecting the orientation of the chains correctly.

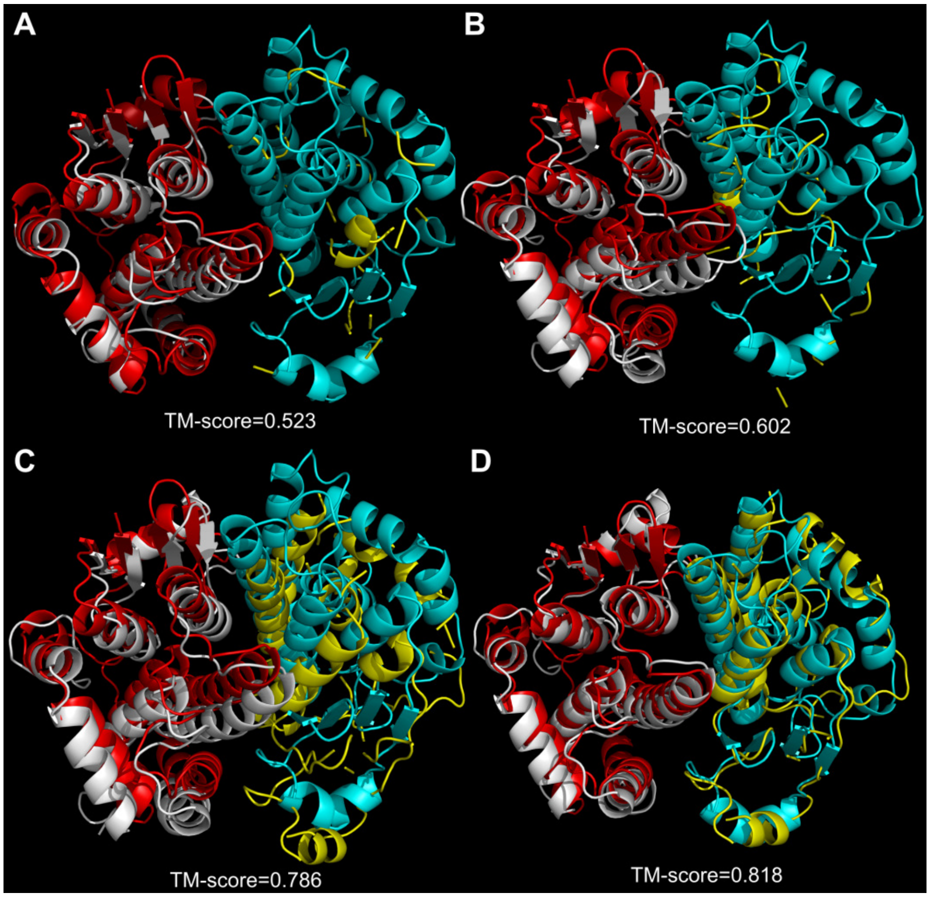

Figure 3.4: Examples of improvement of COTH over controls. Template structures produced for the human pi class glutathione transferase by four different threading methods from (a) PSI-BLAST (b) C-PPA (c) C-MUSTER (d) COTH-threading, which have been superimposed onto the experimental structure of the query protein. The experimental structure of 16gsA0-16gsB0 is shown in red for chain 1 and cyan for chain 2 while the models from the threading algorithms are represented in silver for chain1 and yellow for chain 2.

3.1.3. Structure combination of threading templates

Template complexes with similar structures are essential for COTH threading. However, the algorithm can be constrained due to the limited number of available structures in the complex structure library (6,118 structures at 70% sequence identify cutoff in the PDB. Please refer to Methods section for details). The tertiary structure library, on the other hand, is much larger (38,884 structures at the same cutoff) and hence monomer threading has much greater scope to identify homologous or analogous structures. In fact, Zhang and Skolnick [36] demonstrated that the current PDB library is sufficiently complete to solve in principle the protein structure prediction problem for single-domain proteins. This means that for any single-domain protein there is at least one protein in the PDB which is close enough to the target protein such that a full-length model of correct topology can be constructed by template-based modeling methods. Thus, it was hypothesized that the tertiary structure of the component chains may be predicted with a better quality by monomeric threading through a tertiary structure library and the quaternary structure prediction should benefit if tertiary templates are combined with the COTH threading frames.

In Figure 3.3D, we present a head-to-head comparison of the templates by COTH threading versus that by COTH threading followed by monomer structure recombination (called "*COTH*" instead of "*COTH threading*" throughout, see naming convention in Table 3.1). In the latter case, COTH first identifies monomeric templates by MUSTER [25] using monomer sequence as the query, and identify dimer templates by COTH threading using dimer sequences as the query. In the second step, the monomer templates are superposed on the COTH threading templates by the TM-score program [28] to obtain the final complex models by combining the monomer and dimer alignments. All structures in the chain with longer alignment which has a steric clash with another chain during structure combination are excluded. For the 1,000 (500×2) test set monomers, the MUSTER templates have a higher TM-score than that from the COTH threading in 893 cases. When combining the MUSTER templates with the COTH threading, in almost all the cases, this structure recombination results in an increase in alignment coverage, while in 399 out of 500 cases, the global RMSD of the complexes decreases despite the increase in alignment coverage. Overall, the TM-score of the final COTH model is higher than the original COTH threading template in 443 cases. The average TM-score of the first COTH model is 0.438, 11% higher than that of the COTH threading templates (Table 3.2).

In Figure 3.5, two typical examples are cited to illustrate the improvement of structure recombination, one is a heterodimer and another is a homodimer. Figure 3.5A is an example of a near-native heterodimeric structure identified by threading for 1z0kA-1z0kB. The figure on the left shows the first template identified by COTH threading superimposed on the native structure which has a TM-score 0.786 and a RMSD/coverage of 2.16Å/86.9%. Despite the correct chain orientation of the template, the alignments of some loops in Chain A and considerable portion of Chain B are missed. The figure on the right is the final template model predicted by COTH. The majority of missed regions in original COTH threading alignment are recuperated through MUSTER alignments with the coverage increased from 86.9% to 94.7%; the alignment accuracy is also slightly improved and the RMSD decreased from 2.16 Å to 2.01 Å. This results in an overall TM-score increase from 0.786 to 0.906.

The second example is from the homodimer 1f2dA0-1f2dB0 shown in Figure 3.5B. The dimeric template identified by the COTH threading is extracted from the homodimer 1wdwB0-1wdwD0 which shares a sequence identity of 14.5%. The TM-score of this template to native is 0.696 and the RMSD/coverage is 4.02Å/90.7%. MUSTER, on the other hand, identifies 1j0aA from the tertiary structure library as template for both component chains. After the superposition and combination of the MUSTER templates, the TM-score of the complex model increases to 0.884. Again, the MUSTER templates improve both the alignment coverage and the alignment accuracy of COTH, with RMSD/coverage changed to 2.42Å/93.5%.

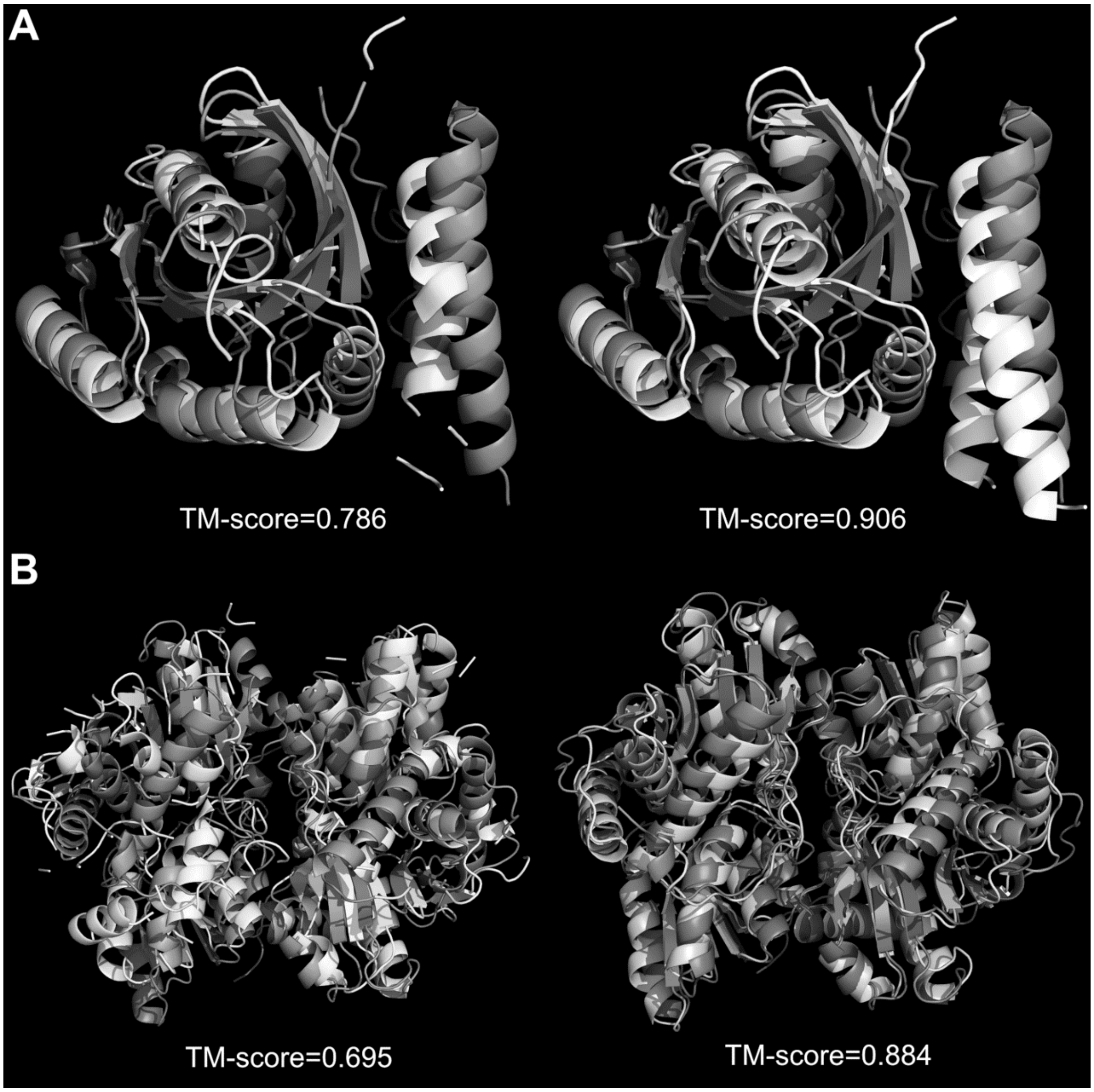

Figure 3.5: Structure superposition improves template quality. Superposition of the native structure (darker shade) with the template structures generated by COTH threading (lighter shade, left) and COTH threading plus recombination (lighter shade, right). a) GTP-Bound Rab4Q67L GTPase (PDB ID: 1z0kA0-1z0kB0). b) 1-aminocyclopropane-1-carboxylate deaminase (PDB ID: 1f2dA0-1f2dB0).

Here, although COTH uses monomer threading from MUSTER, it is essentially different from the separate monomer-based alignments in many of the former methods [21-23]. In these former methods, the single-chain threading is on the monomers extracted from the complex structure library and both monomer and dimer structures are dictated by the dimer structure library. But in COTH, the single-chain threading of MUSTER is through the independent tertiary structure library, which are then recombined with the dimer alignments. Overall, the chain orientation is eventually decided by the dimer threading while the MUSTER single-chain threading serves to improve the quality of monomers and the alignment coverage of the complexes by the use of a nearly 6-fold more complete tertiary structure library.

3.1.4 Comparison of COTH with docking algorithms

Docking and threading-recombination are different approaches towards modeling of protein-protein complex structures. While the goal of the docking algorithms is to find the correct orientation and binding sites of the components given the bound/unbound monomer structures, COTH is designed to generate complex structures from sequences with the aid of template identification. Nevertheless, it is of interest to examine the overall modeling results of COTH and the well-established rigid-body docking algorithms with the purpose for understanding where the two methods stand in a head to head comparison.

For this study, ZDOCK [31, 37-38] was selected as a representative example of the rigid-body docking algorithms partly due to its consistently good performance in the CAPRI experiments. The ZDOCK package is also publically downloadable at http://zdock.bu.edu. Because the threading-based methods have only part of the chain with the structure predicted while docking is usually performed on full-length structures, to have fair comparisons, 4 additional controls were designed which are all on full-length structures. First, ZDOCK was run on the unbound experimental structures, i.e. running the first step rigid body docking using ZDOCK followed by refinement with RDOCK, which is called “ZDOCK-exp” in Tables 3.1 and 3.3. In the second method, full-length models were constructed for each individual chain by MUSTER [25] and MODELLER [39] and then ZDOCK was used to dock the full-length models (called “ZDOCK-model” in Table 3.1 and 3.3). In the third method, full-length complex structures were generated by superposing the unbound experimental structures of individual chains to the template frame from COTH-threading, called “COTH-exp”. In the fourth control, the full-length models of the individual chains modeled by MUSTER and MODELLER were superposed onto the COTH-threading template frame, called “COTH-model” in Tables 3.1 and 3.3. There were no further refinements conducted in the latter two COTH-based modeling.

It should be mentioned that the models generated by COTH (and all other threading methods) are Cα only which were copied from the template proteins. But for COTH-exp and COTH-model, since the monomer structures are full-atomic, the final combined models are full-atomic as well (similar to the ZDOCK models).

Table 3.3 summarizes the results (the best in top ten models) of the five methods on 77 dimeric complexes in the ZDOCK Benchmark Set 3.0 [40] (the rest of complexes are higher order oligomers and were thus omitted from this study). Since the unbound monomer structures in docking studies are usually similar to the native, instead of examining TM-score and RMSD of the global structure, here we assess the model quality mainly by the interface structure predictions, in a similar way as the CAPRI experiments [7, 13-14].

Interface residue prediction. For the assessment of the interface residue predictions, we define the Accuracy and Coverage of interface residues as

$$Accuracy = \frac{\text{No. of residues correctly predicted to be interface residues}}{\text{No. of residues predicted to be interface residues}} \qquad (1)$$

$$Coverage = \frac{\text{No. of residues correctly predicted to be interface residues}}{\text{No. of actual interface residues in native complex}} \qquad (2)$$

where an “interface residue” is defined as the residue whose Cα atom lies within 10Å of any Cα atoms of any residues in the opposite chain. Since models constructed from threading are Cα only, we do not use the full-atom definition of interface residue as used in CAPRI [41].

However, since our definition is consistent for all the methods compared here, it should allow for an objective assessment of our method. It is found that COTH-based approaches generally have higher binding-site prediction accuracy, but with lower coverage, than the models by ZDOCK irrespective of whether the experimental unbound structures (70.2% vs. 67.7% accuracy and 39.8% and 64.5% coverage) are used or the MODELLER models (63.3% vs. 56.4% accuracy and 38.7% and 49.7% coverage) for docking. For the 12 "hard" targets as classified in the ZDOCK benchmark dataset (most are antigen-antibody complexes), for example, the average accuracy of the predicted interface residues is 44.8% with coverage of 42.6% in the ZDOCK models, while the models constructed by superposition of unbound structures to the COTH templates have an average interface accuracy of 60.3% with coverage of 30.3%. Of the 12 cases, the ZDOCK models have an accuracy higher than 50% in 4 cases while 7 of the COTH models have the accuracy over 50%.

Interface contact prediction. Since the binding-site prediction accuracy only counts for the total number of the correctly predicted residues in the interface area which nevertheless may interact with incorrect residues of the opposite chain in the model, in Column 3 of Table 3.3 we list the accuracy of the interface contacts predicted for the best in the top 10 models. Similarly, the accuracy of interface contact predictions is defined as the number of correctly predicted contacts across two chains divided by the total number of cross-chain contacts in the model; the coverage is the number of correctly predicted interface contacts divided by the observed inter-chain contacts in the native structure.

Since threading alignments provide only C_α traces, we defined the inter-chain residue contacts based on amino acid specific 20×20 C_α distance and standard deviation matrices, which were calculated from 6,118 non-redundant dimer structures in our library (Appendix II).

In the calculations, since the experimental complex structures are full-atomic, we defined the inter-chain residue pairs as contact if the distance of any heavy atoms is below 5 Å. Interestingly, the mean distance of C_α atoms is generally smaller between the same amino acids than that between different amino acid types (Appendix II), which indicates that the similar amino acids tend to be packed tighter than different amino acid pairs. Two residues are predicted to be in contact if the distance between their C-alpha atoms is ≤ $(d_{i,j} + sd_{i,j})$ where $d_{i,j}$ is the mean C-alpha distance between residue i and residues j and $sd_{i,j}$ is the standard deviation.

In general ZDOCK, generates models of comparable contact accuracy and coverage as COTH when experimental unbound structures are used for docking and for structure superposition, i.e. 0.466 vs 0.474 for accuracy and 48.8% vs 42.3% for coverage, by ZDOCK and COTH respectively. When the predicted full-length models (by MUSTER + MODELLER) are used, the contact accuracy by COTH-model (0.405) is higher by 35% than ZDOCK-model (0.301), whereas the coverage of the contact predictions by the two methods is similar (40.3% vs. 40.4%). Interestingly, the accuracy of COTH-model, which combines full-length models to the COTH templates, is also better than COTH itself that combines MUSTER threading templates (34.2%). This is mainly due to around 1/3 test cases where the MUSTER threading has substantial gaps in the interface area which reduce the accuracy and coverage of the contact predictions. When the full-length models are constructed, the gapped regions were filled and the overall accuracy and coverage of contacts are increased.

Even using the experimental unbound structures, COTH slightly outperforms ZDOCK in the hard cases when conformational changes are involved in protein-ligand binding [40]. In the 12 hard cases, for example, the ZDOCK models have a contact accuracy >50% in 4 cases

(2nz8A:B, 2ot3A:B, 1r8sA:E, 2c0lA:B) while the COTH models have an accuracy higher than 50% in 5 cases (1iraY:X, 2ot3A:B, 2c0lA:B, 1ibrA:B, 1pxvA:C). Of the 5 COTH "winning" cases, only two (2ot3A:B and 2c0lA:B) has ZDOCK models with a contact accuracy >50%; for the other 2 cases where ZDOCK has an accuracy >50% both the COTH models have a contact accuracy below 50%, which demonstrates that the two methods are essentially complementary to each other in terms of predicting the structure of protein complexes. Again, in all the contact predictions, ZDOCK generally has higher coverage than COTH.

In Figure 3.6, we show one example of the hard targets from the Ran-Importin beta complex (PDB ID 1ibrA:B). ZDOCK (the best in top 10 models, ranked 5 in this case) puts the Ran chain on the convex site of the crescent structure of the Importin beta chain but in the native structure Ran actually binds on the concave site, which resulted in a high I-RMSD (9 Å) with the interface contact accuracy and coverage being 0% (Figure 3.6A). On the other hand, COTH-threading (the best in top 10 models, ranked 2 in this case) detected the template of mDIA1-RhoC complex (PDB ID: 1z2c) with a sequence identity of 12.4% to the target which has 79.4% of residues aligned. Despite the wrong topology of the C-terminal of the template on the Importin beta chain, the Ran chain was aligned at an approximately correct location at the concave site, which has an I-RMSD=4.7 Å with an interface contact accuracy of 68.6% and coverage of 57.5% (Figure 3.6B). When we superposed the experimental unbound structures to the template, we got a complex model with an I-RMSD=4.8 Å, with an interface contact accuracy of 70.1% and coverage of 74.2%. Because the unbound experimental structures have a closer topology to the target than the COTH-threading template, after the COTH superposition, the global topology of the complex structure is also markedly improved with the overall TM-score increasing from 0.435 to 0.692 and the RMSD decreasing from 5.4 Å to 3.85 Å (Figure 3.6C).

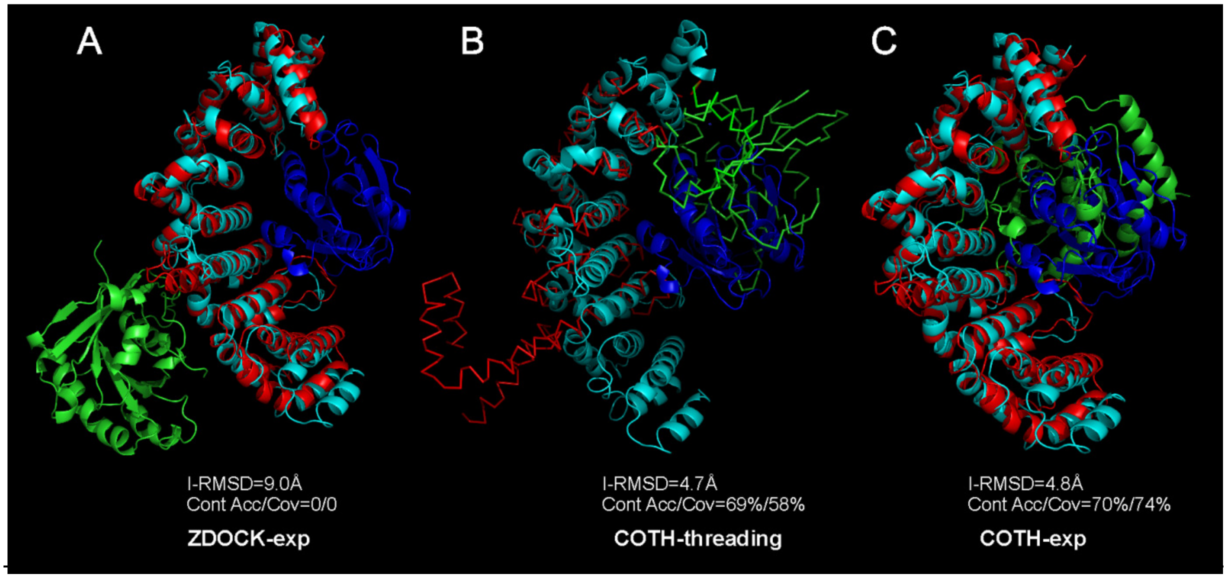

Figure 3.6: Modeling result of ZDOCK and COTH on the Ran-Importin beta complex. The native complex is represented in Cyan (larger chain) and Blue (smaller chain) while the predicted models represented as Red (larger chain) and Green (smaller chain) respectively. (A) ZDOCK-exp; (B) COTH-threading; (C) COTH-exp with unbound experimental structures superimposed on the COTH-threading template.

In general, the ZDOCK model has a higher coverage in the interface and contact predictions. One reason for the difference is that ZDOCK tries to geometrically match the ligand and receptor structures and the contact area of two chains in ZDOCK is usually maximized. In COTH on the other hand, the threading alignment is designed to identify the best global structure and chain-orientation match. When the unbound experimental structures or predicted single-chain models are combined with the threading templates, they were simply shifted through superposition to the complex frame without any attempt to maximize the geometric contact area of the interface. Therefore, even though the orientation of the monomer chains is correctly modeled in COTH, the coverage of interface contact predictions is usually lower. Further docking refinement simulations, e.g. by backbone displacement and side-chain optimization as done in ROTAFIT [42], may be used to fine-tune the complex structure and improve the interface coverage and contact accuracy. Another factor for the reduction in coverage is the alignment gaps in COTH threading which may appear in the interface regions and reduce the residue coverage. This has been partly amended in COTH-exp and COTH-model when full-length structures were used.

Accuracy of interface structure. The accuracy of the interface structure is assessed by the interface RMSD, I-RMSD. A full list of the I-RMSD values by the five methods, COTH, COTH-exp, COTH-model, ZDOCK-exp, ZDOCK-model, is given in Appendix III. For all such analysis reported here, the best in top 10 (according to rank) models for each method has been used. The average I-RMSD by different methods is almost randomly distributed due to the large fluctuations of a few high I-RMSD targets. In Column 4 of Table 3.3, the number of hits in the 77 targets was counted. For COTH, since gaps may be present in the interface area, it is requested that a hit should have at least 50% of the interface residues aligned. Overall, the number of hits by the four methods with full-length models is similar, ranging from 20 to 26, where ZDOCK is slightly better on experimental unbound structures and COTH has only one more hit on predicted models. The COTH models have the highest number of hits (28) which is partly due to the lower alignment coverage.

Methods	**Interface-Accuracy (Coverage)**[1]	**Contacts-Accuracy (Coverage)**[2]	**N_{Hit}**[3]	**Median I-RMSD**
COTH	59.8% (31.7%)	34.2% (33.4%)	28	6.37 Å
COTH-exp	70.2% (39.8%)	47.4% (42.3%)	23	7.76 Å
COTH-model	63.6% (38.7%)	40.5% (40.3%)	21	7.92 Å
ZDOCK-exp	67.7% (64.5%)	46.6% (48.8%)	26	8.29 Å
ZDOCK-model	56.4% (49.7%)	30.1% (40.4%)	20	9.78 Å

[1] Accuracy (coverage) of the predicted interface residues.
[2] Accuracy (coverage) of the predicted inter-chain contacts.
[3] Number of hits which have an I-RMSD ≤ 5Å to the native.

Table 3.3. Summary of the best in top 10 models on 77 ZDOCK benchmark proteins.

Again, the COTH-based methods are highly complementary to the docking-based methods. For example, there are only 12 targets commonly hit by both COTH-exp and ZDOCK-exp methods. If we take the top 5 models (according to rank) from each of the methods, the number of hits in the top 10 models will increase from 26 to 33. Meanwhile, there are only 9 targets commonly hit by both COTH-model and ZDOCK-model methods. If we take the top 5 models from each of these two methods, the number of hits in the top 10 models will increases from 21 to 28. In Column 5 of Table 3.3, the median I-RMSD of the models by different methods are reported, where the COTH based models have generally a lower median I-RMSD than the ZDOCK models.

3.2 MATERIALS AND METHODS

COTH is a hierarchical threading approach to fold-recognition and structural recombination of protein-protein complexes. For a given complex protein, COTH takes only the amino acid sequences of both chains (i.e. Chain A and B) as the input. It proceeds by joining the chains in both order, i.e. ChainA-ChainB and ChainB-ChainA, to represent the dimer sequence for template identification. The joined dimeric sequences are then threaded through a representative complex library of the PDB by a process called "COTH threading", to identify complex templates of similar quaternary structure to the target. Meanwhile, the individual chains of the complex are threaded separately through a representative tertiary structure library by the monomer threading algorithm MUSTER, to identify the monomer templates of similar tertiary structure to the individual target chains. Finally, the top monomer template structures from MUSTER are superimposed onto the top complex templates from

COTH-threading, to generate complex structure models which are the output of the COTH pipeline (Figure 3.1).

3.2.1 Template libraries

Two libraries were created for COTH. The first is a representative monomer structure library collected from the PDB at a pair-wise sequence identity <70%. Obsolete structures and theoretical models are removed. For multiple domain proteins, both individual domains and the whole proteins are used as the template entries. The second is a non-redundant *dimeric* structure library screened from DOCKGROUND [44] with the pair-wise sequence identity cutoff at 70% after an initial filtering to remove irregular structures, transmembrane complexes and the complexes with alternate binding modes. Complexes with less than 30 interface residues or with a buried surface area ≤ 250 Å^2 are ignored to rule out possible crystallization artifacts. However, if a new structure has an overall sequence identity >70% to an old structure existing in the library but has one chain sharing less than 70% sequence identity to the corresponding chain of the old structure, the new structure is also included in the library. This helps account for the targets which have big common receptor structures but with different small ligand proteins (often with different orientation). Higher-order complexes are split into dimers by taking all possible dimeric combinations. As of February, 2010, the libraries consist of 38,884 monomer and 6,118 dimer structures.

3.2.2 Single-chain monomeric threading

The single-chain threading is carried out by an execution of the MUSTER algorithm [25] through the tertiary structure library. The scoring function of MUSTER is based on the close and remote sequence profile-profile alignments, assisted by the secondary structure predictions, structural profiles accounting for residue depth in the structure, solvent accessibility, torsion angle prediction, and hydrophobic scale.

3.2.3 Protein-Protein Interface Prediction by BSpred

To better identify the orientation of one protein chain relative to another in protein-protein complexes, prior knowledge about the interface residues of both chains is helpful. Accordingly, a new machine-learning method called BSpred is developed, which is capable of predicting the binding status of each residue from the amino acid sequence alone.

The input features of BSpred are the following 1) The Position Specific Scoring Matrix (PSSM) generated by PSI-BLAST search using an E-value cutoff =0.001. 2) The secondary structure (SS) of the query sequence, predicted by PSI-PRED, which is to detect the SS preference at the interface residues. The SS is represented by a 3-element vector ([0 0 1] for random coil, [0 1 0] for alpha helix, [1 0 0] for beta strands). 3) The solvent accessibility (SA) predicted from an independent neural network predictor[45]. The predicted solvent accessibility (whether buried or exposed) is a 2-element vector ([0 1] for buried, [1 0] for exposed). 4) The distinctive hydrophobicity of amino acids in protein-protein interfaces. Each amino acid is assigned by a hydrophobicity score, taken from the Eisenberg hydrophobicity scale [46] which lies between 0 and 1 for all the amino acids. The NN software used in BSpred is from Fast Neural Network (FNN) [47]. By trial and error, 3 layers with 50 hidden neurons for NN are chosen which gives the best performance on training data. The training algorithm of NN is the standard Back-Propagation (BP) algorithm.

For prediction of interface residues, the neighboring residues around a central residue also contribute to the formation of the interface [48]. A window size of 21 is used to specify the *i*th residue, which includes residue indices from *i*-10 to *i*+10. Since there are 26 (=20+3+2+1) feature values for a residue, the number of features for a window around the trained residue is 546 (=21× 26). At the N and C terminals, the input values for the neighboring residues which are not present are represented by 0. The NN output value is between -1 and 1 for each residue where larger values indicate higher confidence for that particular residue to be at the interface. Accordingly, a carefully optimized cutoff value (to obtain a balance between accuracy and coverage of prediction) is selected based on the performance on a set of training proteins which is non-redundant to the testing proteins of this work. Any residue with an output value higher than the cutoff is considered as an interface residue. It is found that the NN output cutoff =-0.1 have the best balance of accuracy and coverage.

Based on the observation that interface residues are often sequentially clustered together, we introduce a second-step post-processing for smooth filtering of raw neural network predictions, i.e. a residue with NN output score >-0.1 is finally considered as an interface residue only if at least 6 other residues in its direct sequence neighborhood (from i-3 to i+3) are also predicted to be interface residues (NN output score>-0.1). For the N-terminal and C-terminal residues, at least 3 neighboring residues should be at the interface. Also, since an interface residue must be solvent exposed at the monomer structure, any predicted interface residues which were not predicted to be solvent exposed are eliminated from our final interface predictions.

The method has been tested on a set of 150 single-chain proteins which are non-redundant to the training proteins and are known to participate in dimer formation. For assessment of the interface and chain orientation predictions, the *Accuracy* and *Coverage* of interface residues is defined as in Eq (1) and Eq (2).

The final accuracy of the interface prediction by BSpred is 65.6% with coverage of 13.7%. The BSpred program and the on-line server are freely available at http://zhanglab.ccmb.med.umich.edu/BSpred.

3.2.4 COTH threading

The alignment of the query and template complexes is generated by a modified dynamic programming algorithm that is designed to avoid unphysical cross alignments and was also implemented in MM-align (Chapter 2). The scoring function for aligning the *i*th residue of the query and the *j*th residue of the template is given by

$$\begin{aligned}\text{Score}(i,j) \; &= \sum_{k=1}^{20}\left(Pc_q(i,k)+Pd_q(i,k)\right)L_t(j,k)/2+c_1\delta(s_q(i),s_t(j))+c_2\sum_{k=1}^{20}Ps_t(j,k)L_q(i,k)\\ &+c_3(1-2\,|\,SA_q(i)-SA_t(j))+\mathrm{c}_4(1-2\,|\,\phi_q(i)-\phi_t(j)+c_5(1-2\,|\,\varphi_q(i)-\varphi_t(j)\,|)\\ &+\mathrm{c}_6M(AA_q(i),AA_t(j))+c_7\delta\left(I_q(i),I_t(j)\right)+c_8\end{aligned} \qquad (3)$$

where ‘*q*’ stands for the query and ‘*t*’ for the template. The first term in Eq. 3 represents the sequence-derived profiles where $Pc_q(i, k)$ is the frequency of the *k*th amino acid at the *i*th position of the multiple sequence alignment by PSI-BLAST at an E-value cutoff of 0.001; $Pd_q(i, k)$ is the “remote homology” frequency matrix by PSI-BLAST with E-value<1.0; $L_t(j, k)$ is the PSSM log-odds profile of the template. The second term denotes the secondary structure match and $\delta(s_q(i), s_t(j))$ equals 1 when the secondary structures of *i* and *j* are the same and -1 when the secondary structures are different. The third term counts the depth of the aligned residues where $Ps_t(j, k)$ is the depth dependent structure profile and $L_q(i, k)$ is the PSSM profile of the query. The fourth, fifth and sixth terms compute the match between the solvent accessibility, phi angle and psi angle of the query and the template, respectively. The seventh term counts the hydrophobicity match of the residues based on the hydrophobicity scoring matrix. The eighth term computes the match between the predicted interface residues of the query by BSpred and the interface residues of the template, where $I_q(i)$ is the interface index of *i*th query residue (0 or 1) and $I_t(j)$ is that for *j*th residue on the template. The last parameter of c_8 is introduced to avoid the alignment of unrelated residues in the local regions.

Thus, COTH threading has 10 free parameters (8 weights in Eq. 1, a gap opening (G_o) and a gap extension penalty (G_e)). To determine the parameters, we constructed a 10-dimensional parameter space and ran COTH on 180 randomly selected non-homologous proteins from DOCKGROUND that are also non-homologous to the test proteins, with parameters taken from each of the grid lattice in the 10-dimension system. The optimal parameters are selected when the highest average TM-score for the 180 training proteins is achieved. As a result, the optimized parameters are: c_1=0.80, c_2=0.34, c_3=1.7, c_4=0.29, c_5=0.29, c_6=0.37, c_7=0.20, c_8=-4.90, G_o=10.11, $G_{e=}$0.95.

3.2.5 Template Selection and Target Classification

The significance of a threading alignment in COTH is assessed by Z-score:

$$\mathrm{Z-score} = \frac{R_{\mathrm{score}} - \langle R_{\mathrm{score}} \rangle}{\sqrt{\langle R_{\mathrm{score}}^2 \rangle - \langle R_{\mathrm{score}} \rangle^2}} \tag{4}$$

where R_{score} is the raw alignment score R'_{score} from the dynamic programming normalized by the length of the query dimer sequence (L_{query}) i.e. $R_{\mathrm{score}}=R'_{\mathrm{score}}/L_{\mathrm{query}}$. Because the dynamic programming of COTH uses an unique path for both chains, the overall raw alignment score and the Z-score has a bias towards the larger of the two chains, especially when the receptor is significantly larger than the ligand; this may lead to artificially high Z-scores even though the ligand is poorly aligned. To balance this bias, we rank the COTH models based on the mean Z-score of the ligand, the receptor and the complex:

$$\mathrm{Mean\ \ Z\text{-}score} = (\mathrm{Z\text{-}score}_{\mathrm{complex}} + \mathrm{Z\text{-}score}_{\mathrm{receptor}} + \mathrm{Z\text{-}score}_{\mathrm{ligand}})/3 \tag{5}$$

In Figure 3.7, we show the mean Z-score versus the TM-score of the first COTH models of the 180 training proteins. There is a positive correlation between Z-score and TM-score with correlation coefficient=0.77. Accordingly, the query proteins are categorized into "easy" or "hard" targets based on the Z-score, i.e. when a query has at least one template alignment with an average Z-score >2.5 it is defined as an "easy" target; otherwise it is labeled as a "hard" target. If templates with TM-score above 0.4 are considered to be reliable, the false positive and false negative rates of Z-score=2.5 are 8.2% and 5.1%, respectively, for the training proteins. When applying this definition of Z-score to the 500 test proteins, 296 cases are considered as "easy" targets and 204 that are "hard" targets. The average TM-score for "easy" and "hard" proteins are 0.478 and 0.245, respectively. These data demonstrate that the Z-score can be used as a reliable indicator of the template quality.

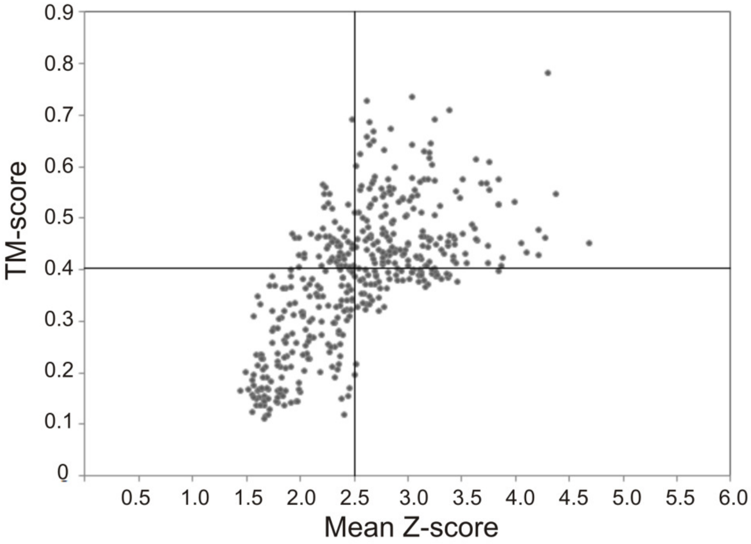

Figure 3.7: TM-score versus Z-score of the first COTH templates for 180 training proteins.

3.2.6 Structural superposition and combination of monomer templates

Most proteins in the PDB library have been solved in monomer form [49]. As a result, the number of available structures in monomer structure library (38,884) is much higher than that in dimer structure library (6,118). The structure space is far more complete in the tertiary structure library than in the complex structure library. It is therefore expected that by combining both the tertiary and quaternary structure libraries the COTH threading alignments can be further improved. To achieve this, first the normal COTH threading procedure as described above is used to identify the template frames of complex structures. Meanwhile, COTH threads the monomer sequence (the individual chains of the dimer) to the tertiary structure library by the MUSTER algorithm. Finally, the MUSTER monomer templates, which usually have a better tertiary structure quality than that of the COTH-threading templates, are superimposed by the TM-score rotation matrix on the COTH complex templates, based on the commonly aligned residues. It should be noted here that only the aligned regions in the MUSTER and COTH threading alignments was used for superposition (but not the original PDB structures of the templates structurally aligned together). If no commonly aligned residues are present between the MUSTER and COTH threading template (which actually never happened in any of our training or testing proteins), the program simply discards the superposition step and retains the original COTH threading template alignments.

The final complex model consists of the re-oriented structures of the MUSTER templates. If there are regions which are aligned by COTH threading but not aligned by MUSTER, the structural coordinates are not copied to the final models because these regions may have steric clashes with the MUSTER templates though it increases the coverage. The advantage of the superimposition step is that the resultant template retains the information regarding the relative orientation of the chains extracted by the COTH threading alignment while the tertiary structure qualities of the individual chains are significantly improved since the MUSTER templates have been generated from a much larger structure library. However, it is possible that in some rare cases, the combination step may result in unphysical inter-chain clashes (the inter-chain Cα-Cα distance <3.8Å). To rule out the clashes, the COTH program automatically discards the residues from the chain of higher alignment coverage which has a distance <3.8 Å to any residues in another chain.

When superimposing and combining monomer and dimer templates, COTH takes the top-five templates from MUSTER for each chain; each of the monomer templates is then superimposed on the top-ten dimer templates from COTH, which results in 250 dimeric structures (=5×5×10). To rank the 250 structures, we structurally aligned each of the structure to the other 249 structures by the multimeric structure alignment program MM-align and calculate the average TM-score of the structure compared with others. The structure of the highest TM-score to other template, which means a consensus, is selected as the final COTH model.

3.3 DISCUSSIONS

A new algorithm for protein complex structure modeling by threading-based template identification and the monomer-dimer alignment combination, COTH, was developed. The algorithm takes the advantage of the well-established threading alignment methods in protein structure prediction and the complement of tertiary and quaternary structure libraries. The *ab initio* binding site prediction is further exploited to assist the chain orientation selections.

The COTH method has been tested on two independent sets of protein-protein complexes. In the first test on 500 non-homologous complexes, COTH produces predictions with a TM-score >0.4 (or RMSD<6.5 Å with alignment coverage >70%) for nearly half of the cases when all homologous templates with a sequence identity >30% or detectable by PSI-BLAST with E-value <0.5 are excluded. Detailed comparisons of four different alignment methods show

COTH threading with *ab initio* binding site predictions outperforms C-MUSTER, a direct extension of the tertiary threading algorithm combining multiple structural information; C-MUSTER in turn performs better than the profile-profile based alignments methods, which outperforms the sequence-profile alignment by PSI-BLAST. Overall, COTH threading (combining the advantages of the profile-profile alignment and multiple-resource structure information) outperforms PSI-BLAST by 46% in TM-score. When combining the tertiary threading alignments, the improvement over PSI-BLAST increase to 63%. Another observed trend in COTH is that the threading-based methods tend to be more reliable for enzyme-ligand complexes as compared to antibody-antigen complexes due to the conservation in sequence profiles in the former.

In the second test of 77 protein complexes from ZDOCK benchmark 3.0, COTH was compared with ZDOCK, which constructs complex structures by docking unbound experimental structures (or predicted full-length monomer models). It is found that COTH performs favorably with a higher accuracy than ZDOCK in predicting the binding-site interface residues; however, the number of interface residues in the COTH prediction is lower. For the interface contact prediction and the accuracy of interface structure represented by interface RMSD, COTH shows complementarity in performance with respect to ZDOCK, especially for the hard cases when binding-induced conformational changes are involved.

Since COTH has benefited from recombination of monomer threading templates from MUSTER, the algorithm can be further improved by exploiting the meta-server threading approaches. A recent experiment showed that combining templates from multiple threading programs results in at least 7% TM-score increase compared to the best single threading methods [43]. The COTH method currently takes, on average, 30 minutes for a medium sized dimer protein of about 400 amino acids on 2.6GHz AMD processors. This efficiency in CPU cost ensures the feasibility of accommodating increasingly larger structure libraries as well as including more single-chain based meta-server threading approaches. It also represents a favorable performance in terms of speed of calculation as compared to the docking methods which usually costs several hours for docking one pair structures. Thus, COTH represents one of the first, fast and reliable methods for predicting template structures of protein complexes from the sequence information. The COTH on-line server is publicly accessible at http://zhanglab.ccmb.med.umich.edu/COTH.

3.4 REFERENCES

1. Moult, J., et al., *Critical assessment of methods of protein structure prediction-Round VIII.* Proteins-Structure Function and Bioinformatics, 2009. **77**: p. 1-4.
2. Kryshtafovych, A., K. Fidelis, and J. Moult, *CASP8 results in context of previous experiments.* Proteins, 2009. **77 Suppl 9**: p. 217-28.
3. Zhang, Y., *Progress and challenges in protein structure prediction.* Curr. Opin. Struct. Biol., 2008. **18**(3): p. 342-8.
4. Vajda, S. and C.J. Camacho, *Protein-protein docking: is the glass half-full or half-empty?* Trends Biotechnol, 2004. **22**(3): p. 110-6.
5. Aloy, P., M. Pichaud, and R.B. Russell, *Protein complexes: structure prediction challenges for the 21st century.* Curr Opin Struct Biol, 2005. **15**(1): p. 15-22.
6. Russell, R.B., et al., *A structural perspective on protein-protein interactions.* Curr Opin Struct Biol, 2004. **14**(3): p. 313-24.
7. Lensink, M. and S. Wodak, *Docking and scoring protein interactions: CAPRI 2009.* Proteins, 2010: p. DOI: 10.1002/prot.22818.

8. Kozakov, D., et al., *Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19.* Proteins, 2010. **78**(15): p. 3124-30.

9. Gray, J.J., et al., *Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.* J Mol Biol, 2003. **331**(1): p. 281-99.

10. Tovchigrechko, A. and I.A. Vakser, *Development and testing of an automated approach to protein docking.* Proteins, 2005. **60**(2): p. 296-301.

11. Katchalski-Katzir, E., et al., *Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.* Proc Natl Acad Sci U S A, 1992. **89**(6): p. 2195-9.

12. Hwang, H., et al., *Performance of ZDOCK and ZRANK in CAPRI rounds 13-19.* Proteins, 2010. **78**(15): p. 3104-10.

13. Mendez, R., et al., *Assessment of blind predictions of protein-protein interactions: current status of docking methods.* Proteins, 2003. **52**(1): p. 51-67.

14. Mendez, R., et al., *Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures.* Proteins, 2005. **60**(2): p. 150-69.

15. Sircar, A. and J.J. Gray, *SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models.* PLoS Comput Biol, 2010. **6**(1): p. e1000644.

16. Huang, S.Y. and X. Zou, *MDockPP: A hierarchical approach for protein-protein docking and its application to CAPRI rounds 15-19.* Proteins, 2010. **78**(15): p. 3096-103.

17. Zacharias, M., *ATTRACT: protein-protein docking in CAPRI using a reduced protein model.* Proteins, 2005. **60**(2): p. 252-6.

18. Sircar, A., et al., *A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13-19.* Proteins, 2010. **78**(15): p. 3115-23.

19. Fiorucci, S. and M. Zacharias, *Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT.* Proteins, 2010. **78**(15): p. 3131-9.

20. Janin, J., *The targets of CAPRI Rounds 13-19.* Proteins, 2010. **78**(15): p. 3067-72.

21. Aloy, P., et al., *Structure-based assembly of protein complexes in yeast.* Science, 2004. **303**(5666): p. 2026-9.

22. Lu, L., H. Lu, and J. Skolnick, *MULTIPROSPECTOR: An algotihm for the predictio of protein-protein interactions by multimeric threading.* Proteins:Structure,Function and Genetics, 2002. **49**: p. 350-364.

23. Kundrotas, P., M. Lensink, and E. Alexov, *Homology based modelling of 3D structures of protein complexes using alignments of modified sequence profiles.* International Journal of Biological Macromolecules, 2008. **43**(2): p. 198-208.

24. Wu, S., J. Skolnick, and Y. Zhang, *Ab initio modeling of small proteins by iterative TASSER simulations.* BMC Biol, 2007. **5**: p. 17.

25. Wu, S. and Y. Zhang, *MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information.* Proteins, 2008. **72**(2): p. 547-56.

26. Zhang, Y. and J. Skolnick, *Automated structure prediction of weakly homologous proteins on a genomic scale.* Proc. Natl. Acad. Sci. USA, 2004. **101**: p. 7594-7599.

27. Simons, K.T., C. Strauss, and D. Baker, *Prospects for ab initio protein structural genomics.* J. Mol. Biol., 2001. **306**: p. 1191-1199.

28. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**: p. 702-710.

29. Lorenzen, S. and Y. Zhang, *Identification of near-native structures by clustering protein docking conformations.* Proteins, 2007. **68**: p. 187-194

30. Mukherjee, S. and Y. Zhang, *MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming.* Nucleic Acids Res, 2009. **37**(11): p. e83.

31. Chen, R., L. Li, and Z. Weng, *ZDOCK: an initial-stage protein-docking algorithm.* Proteins, 2003. **52**(1): p. 80-7.

32. Comeau, S.R., et al., *ClusPro: a fully automated algorithm for protein-protein docking.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W96-9.

33. Tovchigrechko, A. and I.A. Vakser, *GRAMM-X public web server for protein-protein docking.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W310-4.

34. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

35. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* J. Mol. Biol., 1999. **292**: p. 195-202.

36. Zhang, Y. and J. Skolnick, *The protein structure prediction problem could be solved using the current PDB library.* Proc. Natl. Acad. Sci. USA, 2005. **102**: p. 1029-1034.

37. Li, L., R. Chen, and Z. Weng, *RDOCK: refinement of rigid-body protein docking predictions.* Proteins, 2003. **53**(3): p. 693-707.

38. Wiehe, K., et al., *The performance of ZDOCK and ZRANK in rounds 6-11 of CAPRI.* Proteins, 2007. **69**(4): p. 719-25.

39. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J. Mol. Biol., 1993. **234**(3): p. 779-815.

40. Hwang, H., et al., *Protein-protein docking benchmark version 3.0.* Proteins, 2008. **73**(3): p. 705-9.

41. Lensink, M.F. and S.J. Wodak, *Blind predictions of protein interfaces by docking calculations in CAPRI.* Proteins, 2010. **78**(15): p. 3085-95.

42. Lorenzen, S. and Y. Zhang, *Monte Carlo refinement of rigid-body protein docking structures with backbone displacement and side-chain optimization.* Protein Sci, 2007. **16**(12): p. 2716-25.

43. Wu, S.T. and Y. Zhang, *LOMETS: A local meta-threading-server for protein structure prediction.* Nucl. Acids. Res., 2007. **35**: p. 3375-3382.

44. Douguet, D., et al., *DOCKGROUND resource for studying protein-protein interfaces.* Bioinformatics, 2006. **22**(21): p. 2612-8.

45. Chen, H. and H.X. Zhou, *Prediction of solvent accessibility and sites of deleterious mutations from protein sequence.* Nucleic Acids Res, 2005. **33**(10): p. 3193-9.

46. Eisenberg, D., R.M. Weiss, and T.C. Terwilliger, *The hydrophobic moment detects periodicity in protein hydrophobicity.* Proc Natl Acad Sci U S A, 1984. **81**(1): p. 140-4.

47. Nissen, S. and E. Nemerson, *Fast artificial neural network. Available at http://fann.sourceforge.net.*

48. Ofran, Y. and B. Rost, *ISIS: interaction sites identified from sequence.* Bioinformatics, 2007. **23**(2): p. e13-6.

49. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

CHAPTER 4. Reassembly and Refinement of Multimeric Templates Using Replica-Exchange Monte-Carlo Simulations

Subsequent to the steady success observed in the field of automated structure prediction of monomeric proteins[1-4], the time is ripe for investment of the considerable knowledge gathered from the experience, to attempt the prediction of protein-protein complex structures from the primary amino acid sequence alone. Admittedly, the three major existing problems in protein structure prediction; 1) detection of remote homologs/structural analogs for use as templates 2) efficient search of the conformational phase space and 3) design of accurate funnel-shaped force fields[2, 5] is multiplied when trying to predict not one but two protein chains in association with each other. However, carefully adopting the strengths of existing algorithms designed for accurate structure prediction of monomers in conjunction with newly designed force fields and conformational search techniques to reflect the conformational space of interacting protein partners can lead to low to medium resolutions models in a number of cases[5].

A major bottleneck in the prediction of protein quaternary structures is the high degree of incompleteness of the protein quaternary structure library. Without the availability of reliable templates, a large number of query sequences become “free modeling” or “*ab initio*” targets for which success is very limited. Though in a stage of relative infancy[6-9], a number of significant efforts have been made in recent years to try and circumvent this problem[5, 10-12] as discussed in Chapter 3. M-TASSER[5] went one step further, used the templates identified by MULTIPROSPECTOR and subjected it to multiple reassembly and refinement steps to generate full length complex structures.

One drawback of the above monomer based threading algorithms is that the cooperativity of multiple-chain alignments, for example, binding specificity and burial interactions, cannot be correctly accounted for during the course of threading alignments because the alignment result of one chain is independent from that of another chain. Similarly, though ambitious in scope M-TASSER falls short because it essentially treats the problem as one of simultaneously modeling two protein chains without searching the orientational space between the two subunits of the complex.

COTH, the multimeric threading algorithm discussed in Chapter 3, aligns both chains of a given target complex simultaneously to both chains of putative templates present in a protein complex library. COTH was shown to outperform traditional multimeric threading algorithms based on PSI-BLAST and was also shown to perform comparably with rigid body docking algorithm ZDOCK[13-14] in terms of I-RMSD (Interface-RMSD), accuracy and coverage of predicted interface residues. This was possible even after removing all homologous templates (sequence identity > 30%) while performing threading. However, one major problem observed was the gaps in the alignment which meant that the predicted structures were not full length complexes. Secondly, in approximately 50% of the cases, COTH was unable to find remote homologs/structural analogs which led to prediction of non-native folds. The deficiencies of COTH necessitated the development of the next logical step; to generate full length complexes by modeling *ab initio* the alignments gaps and reassembling of the aligned template fragments. It was also important to refine the orientation between the two chains of a dimer by conformational search.

Here, we describe a new algorithm, TACOS, a hybrid approach geared towards generating full length protein dimer structures from sequences. It starts from identified templates from

COTH multimeric threading. From there, it fragments them and subsequently reassembles and refines the templates (Figure 4.1) using a course-grained schematic similar to the successful monomer structure prediction algorithm I-TASSER[15-16]. Notably, I-TASSER, which was ranked as the top automated structure prediction methodology in the recently concluded CASP7[4], CASP8[4] and CASP9[17] experiments, also uses a similar lattice based strategy to predict protein monomer structures from sequence. Importantly, in addition to TACOS retaining the distinct flavor of the salient features of the I-TASSER methodology, it introduces a number of novel strategies and knowledge-based potentials to capture the unique idiosyncrasies of protein-protein interactions.

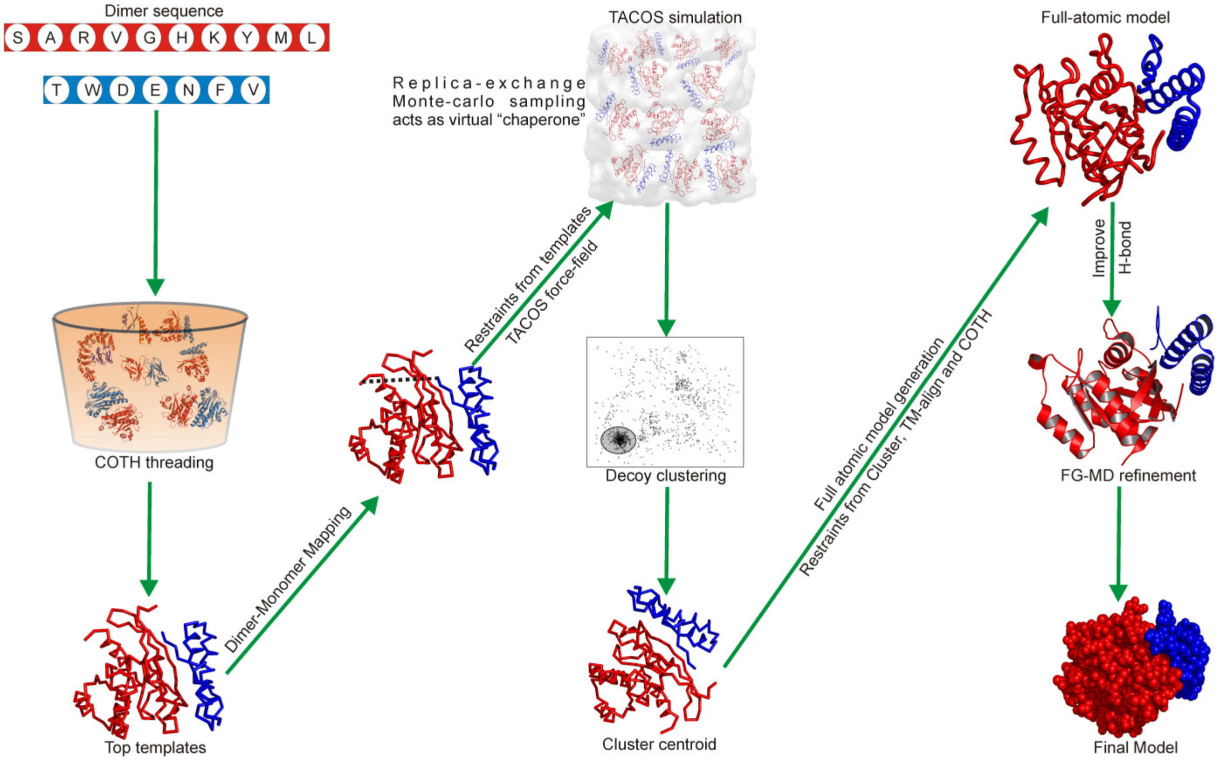

Figure 4.1: Flowchart of the TACOS, template-based assembly of complex structures, protocol.

4.1 RESULTS

4.1.1 Benchmark Sets

500 non-redundant small to medium length proteins were selected from the protein complex structure library at a sequence identity cutoff of 30%. The 500 proteins were divided into two sets; 1) Training set: 150 complexes consisting of 92 homodimers and 58 heterodimers. The heterodimers in the training set also contained a mixture of different biological classes of complexes, 33 enzyme-inhibitor complexes, 19 antigen-antibody complexes and 16 other complexes. 2) Testing set: 350 complexes made up of 213 homodimers and 137 heterodimers. The training set was exclusively used for optimization of the TACOS simulation scheme and the TACOS energy function. The method was then extensively tested on the remaining test set complexes. During training and testing, all templates with a sequence identity ≥ 30% to the target protein complex were removed.

The complexes in the training and testing set can be classified into "easy", "medium" or "hard" modeling targets. A target complex is classified as an "easy" target if multiple templates are identified by COTH with a Z-score ≥ 2.5, as a "medium" target if atleast 1 template with a Z-score ≥ 2.5 is detected and as "hard" if no good templates exist. The training set consisted of 70 easy, 23 medium and 57 hard targets while the testing set consists of 181 easy, 69 medium and 110 hard targets. The average number of decoys generated for the 350 test set proteins is 8313 with more decoys being generated for smaller proteins in general. Better overall result was also obtained when homodimers and heterodimers were trained separately as opposed to when the two types of complexes were trained together.

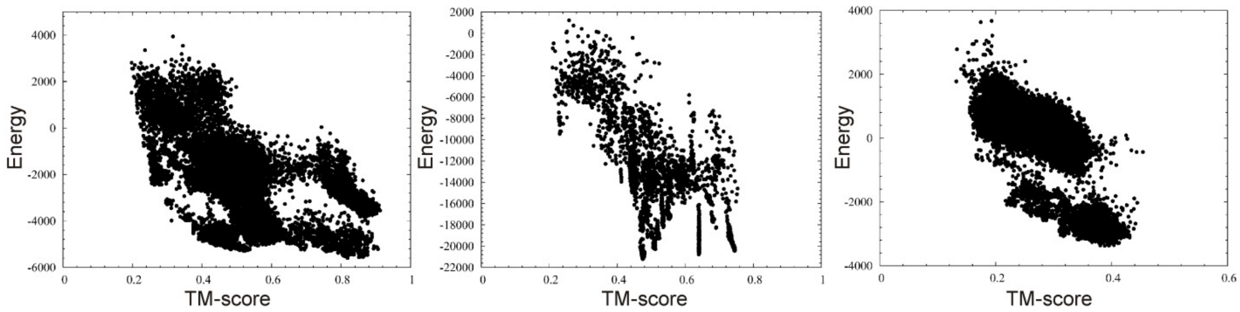

Figure 4.2: Correlation of TACOS energy with TM-score. Three representative examples, one each for easy (left), medium (middle) and hard (right) modeling targets, showing the correlation between energy and TM-score.

Overall, the correlation between energy of the final decoys generated and TM-score of the decoys to native was quite high for the test set proteins. The mean Pearson's correlation coefficient for the 350 test case proteins is 0.768. In Figure 4.2, we show 3 representative examples of each modeling category (easy, medium and hard) showing the correlation of TM-score and energy. The general trend observed was that the decoys for easy cases spanned a larger TM-score range and the sampling was also better in the higher TM-score ranges which can be attributed to multiple reliable templates. However, even though the medium and hard case proteins have less number of decoys in the higher TM-score bins, the correlation between the energy and the TM-score was still very high with the mean correlation for the medium cases only being 0.771 while that for the hard cases only was 0.756.

4.1.2 Improvement over initial templates

Overall improvement: One of the primary goals of TACOS was to try and draw the initial templates closer to the native structure. Hence, it is imperative to check whether that was indeed the case by comparing the initial templates from COTH threading with the final models generated by TACOS. In Figure 4.3A, the TM-score of the initial top 1 COTH threading templates is plotted against the top 1 cluster centroid generated after the clustering of decoys. Figure 4.3B shows the rTM-score of the best in top 10 templates plotted against the best in top

5 cluster centroids. Overall, TACOS simulation improves the TM-score of the top 1 model to the native structure over the COTH threading top 1 template in 285 out of 350 cases or in 81.5% of cases. When best in top 5 TACOS models is compared with the best in top 10 COTH threading templates, the TM-score is improved in an overwhelming 324 out of 350 cases (92.5%). In terms of rTM-score, 264 out of 350 (75.4%) cases are improved for the top 1 model against the top 1 template and 308 out of 350 (88%) for best in top 5 model versus best in top 10 template. The average rTM-score/TM-score of the top 1 COTH threading template was 0.302/0.371 while that of top 1 TACOS model was 0.336/0.430, an improvement of 11% in terms of rTM-score and 16% in terms of TM-score. The improvement is more stark when the best in top 10 templates are compared with best in top 5 models; the average rTM-score/TM-score for COTH is 0.317/0.391 while that of TACOS is 0.377/0.475 indicating a 19% improvement in terms of rTM-score and 21% in terms of TM-score. A histogram showing the distribution of TM-score across various TM-score bins is shown in Figure 4.3D.

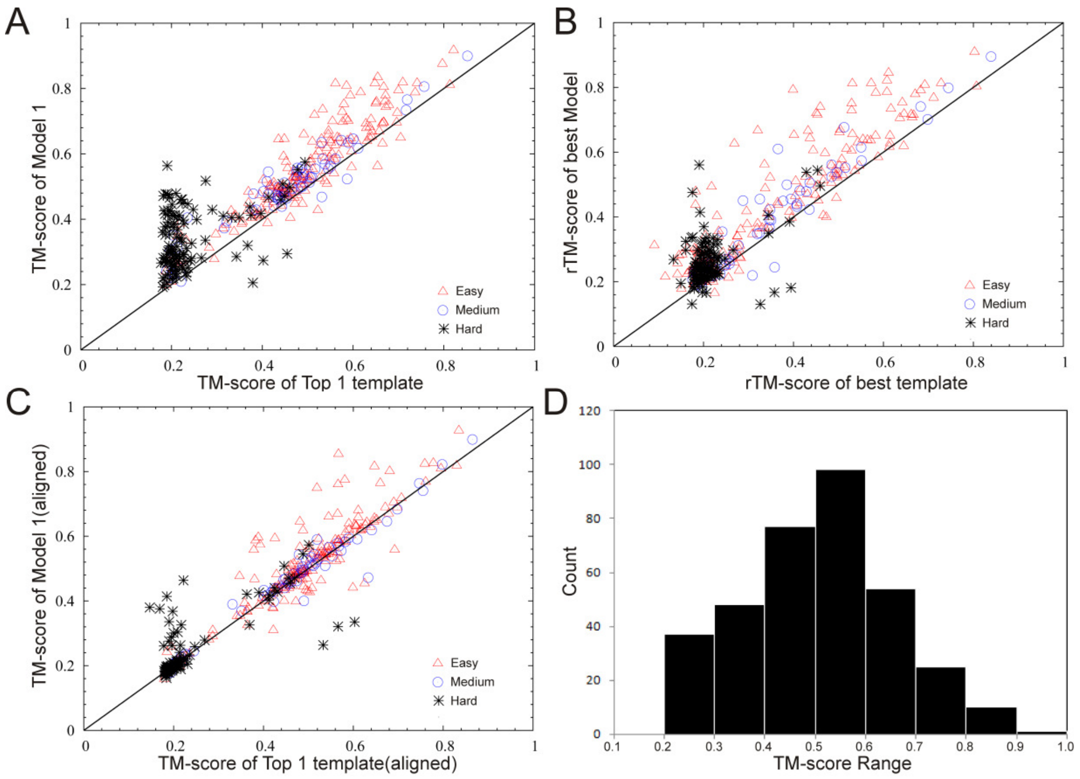

Figure 4.3: Plot showing benchmark results of TACOS on 350 test set proteins. A) Figure showing comparison of TM-score of top 1 COTH threading template versus top 1 model generated by TACOS. The points above the diagonal line indicate targets showing improvement in TM-score for the model generated by TACOS. B) The same plot shown for rTM-score the best in top 10 COTH threading template against the best in top 5 TACOS model. C) The same comparison for Top 1 template and top 1 model but in the threading aligned region only. In this case, the length of the template alignment was used for normalization during the calculation of the TM-score. D) Histogram showing the distribution of TM-score of top 1 model for all 350 targets.

Improvement in template aligned regions: Since, when calculating TM-score/rTM-score normalization is done by the length of the native structures and the COTH threading templates contain gaps in the unaligned regions, it is possible that the TM-score/rTM-score improvement observed overall is simply because of the increased coverage in the final models. To verify whether that was indeed the case or whether the TACOS simulation was successful in refining the template aligned regions as well, we compared the TM-score to native in the template aligned regions only for the Top 1 COTH threading template and the top 1 cluster centroid of TACOS. In this case the TM-score was normalized by the length of the threading aligned regions. As shown in Figure 4.3C, TACOS successfully refines the templates in the aligned regions with an increased TM-score being observed for 244 cases (70%).

Performance for easy, medium and hard targets: To examine further the effect of the quality of the initial COTH templates on the final models generated by TACOS, we analyzed the performance of TACOS for the easy, medium and hard category targets separately. In all three categories, as shown in Figure 4.4, the average TM-score/rTM-score of the final TACOS models is significantly higher than the initial COTH threading templates. Among the 181 easy cases, a TM-score/rTM-score increase (for the top 1 model against top 1 template) was noted in 150/141 cases. For the 69 medium targets TACOS showed an improvement in 56/52 targets and for the hard targets, 79/70 targets showed an overall increase in TM-score/rTM-score.

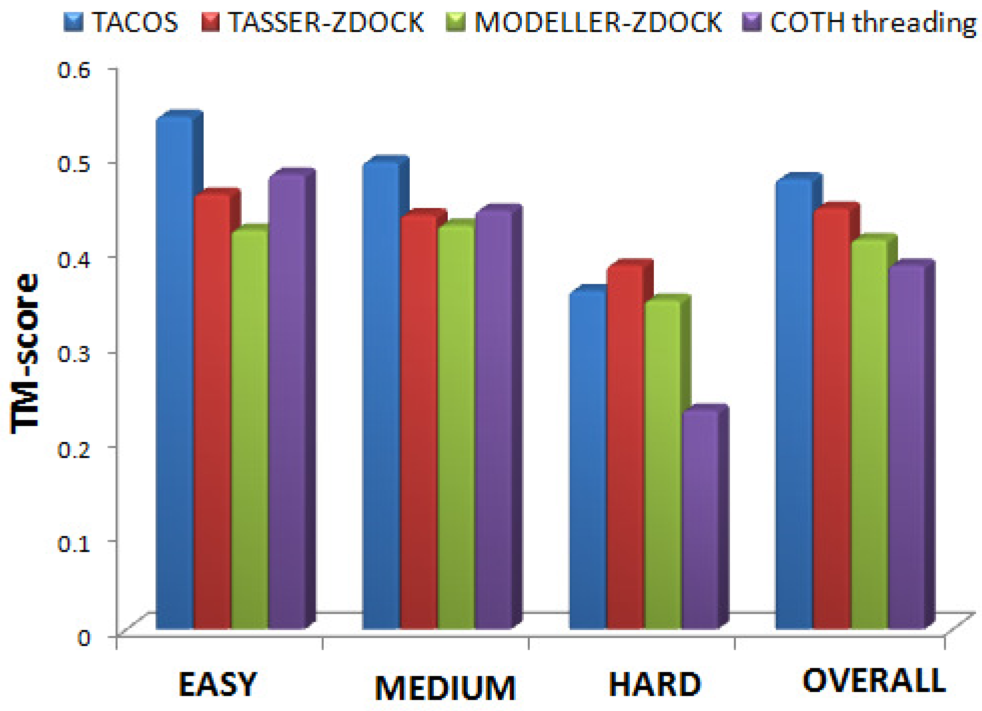

Figure 4.4. Comparison of TACOS with control methods. Plot showing comparison of the best in top 5 models of TACOS as compared to ITASSER-ZDOCK, MODELLER-ZDOCK and COTH threading in terms of TM-score. While TACOS is observed to be the best performing method overall and for easy and medium cases, ITASSER-ZDOCK has a slightly higher average TM-score for the 110 hard targets as compared to TACOS.

What went right: Out of the 324 cases where the final model showed an improvement over the template, 136 had an improvement in TM-score of atleast 20%. The reason for the improvement can be attributed to an increase in coverage as well as refinement of the template aligned regions. Importantly for 39 of these 136 cases, a increase of > 20% was observed in terms of inter-chain contact prediction accuracy and coverage which basically implies that the orientation of the chains with respect to each other were significantly refined. Again, in terms of high-accuracy predictions, there are 36 cases with a TM-score and rTM-score ≥ 0.7 and a global RMSD ≤ 5.0 Å.

In Figure 4.5, we show four examples of successful near-native models built by TACOS, two from homodimers and two from heterodimers. Figure 4.5A presents an example of an all-alpha homodimeric complex from *B. subtilis* which is a putative HTH-type transcriptional regulator (PDBID: 1sgmA0-1sgmB0). The best initial template identified by COTH is from the putative tetR-type transcriptional regulator form *S. coelicolor* (2hyjA1-2hyjA2) which has a sequence identity of 17.1% to the target. The RMSD of the template to native is 3.21 Å in the aligned region. TACOS refined the backbone fragments and the top-ranked model has a RMSD of 2.06 Å to the native. In the same threading aligned region, $\mathrm{RMSD}_{\mathrm{ali}}$ is reduced from 3.21 to 2.02 Å. Accordingly, the TM-score of template is increased from 0.763 to 0.891. Figure 4.5B is another homodimeric example but with an all-beta topology. The TM-score/RMSD of the best template was improved by TACOS from 0.654/4.10 Å to 0.833/3.14 Å.

Figure 4.5C is an example of alpha-beta type of heterodimeric complex between Ulp1 protease and ubiquitin like protein SMT3 (PDB ID: 1euvA0-1euvB0). The best COTH template, the sentrin-specific protease 8 and Neddylin (2brkA0-2brkB0) which has a sequence identity of 21.5% to the target, had an RMSD/TM-score 0.726/4.33Å to the native. It was refined by TACOS with the final model having a TM-score/RMSD of 0.884/2.76 Å. Similarly, the template of the heterodimer 1d9kA0-1d9kB0 in Figure 4.5D was refined from 0.754/3.18 Å to 0.934/1.78 Å in terms of TM-score/RMSD.

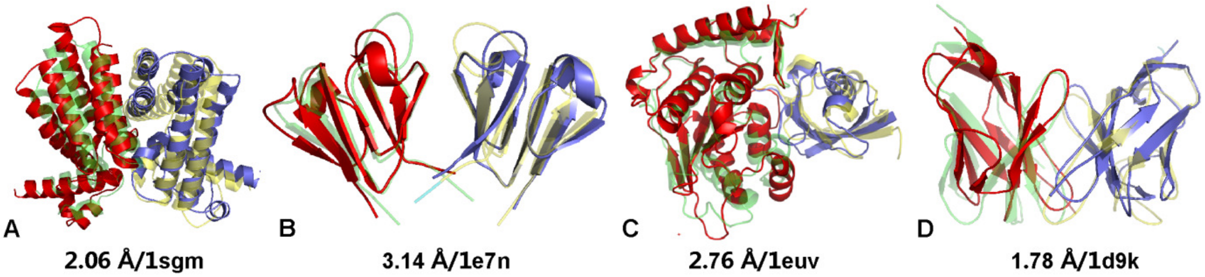

Figure 4.5. Near-native models built by TACOS. Plot showing examples of TACOS modeling for both homo and heterowhere a glycine linker was used to connect both chains into an artificial chain. The first predicted models, shown in red and slate cartoon, are superimposed on the native structure, shown in green and yellow transparent cartoon for chain A and B, respectively.

In all the four cases, both the interface region and the global topology have been improved compared to the threading templates. For 1sgm/1e7n/1euv/1d9k, the interface RMSD of the COTH templates were 2.8/4.6/3.2/2.1 Å while that of the final models were 2.6/3.6/2.9/1.8 Å. In all the cases, the interface backbone atoms were modified by 2.8-4.2 Å by TACOS modeling, where the interface residue packing was driven in a correct direction as demonstrated by reduction of I-RMSD.

What went wrong: 110 of the 350 targets are identified as "hard" targets by COTH which indicates that no reliable templates exist for these cases and are essentially treated as to be built *ab intio.* While TACOS does manage to improve the TM-score in a large majority of these hard targets three notable examples stand out (1ym3A1-1ym3A2, 1q06A-1q06B, 1lq9A-1lq9B) where there is a large regression in TM-score in the final TACOS model as compared to the template. All these three cases (seen quite clearly placed well below the diagonal line in Figure 4.3A, B and C) were classified as "hard" targets by COTH because the alignment coverage for COTH threading was very low but the orientation of the template was in fact correct. Unfortunately, TACOS (because of the low coverage which lead to a low Z-score for the templates) recognized it as a hard target and altered the orientation of the chains since the weight of the template-based restraints are kept low for hard cases. This resulted in the models being driven away from the native orientation leading to low TM-scores. However, the silver lining is that the numbers of such cases are very few. Using a combination of *ab initio* interface prediction methods like BSpred when seen to share a consensus with template-based interface predictions may offer a solution to modeling such targets.

4.1.3 Comparisons with docking

Performance of model-model docking: The current alternative to modeling the dimer as a whole when the structures of the component subunits are unknown is to model the dimers separately and then dock them together using rigid-body docking algorithms[13, 18-23]. Although, docking of protein models has been attempted before[24-25], a systematic study with advanced modeling methods like I-TASSER has not been attempted. Accordingly, we first evaluated the feasibility of performing docking experiments with protein models instead of unbound experimental structures. To perform this study, 77 dimeric proteins of the ZDOCK Benchmark 3.0[26] was selected and the unbound sequences of all the subunits were modeled using I-TASSER (154 total chains). The models were then docked together using ZDOCK (RDOCK refinement was not performed in this case). Simultaneously, we also ran ZDOCK with experimental unbound structures and the results were compared in terms of number of "hits" in the decoy pool where a hit is defined as a target with atleast one model among the top *N* decoys with Interface-RMSD (I-RMSD) less than a cutoff (5.0 Å and 2.5 Å cutoffs were used for this study). Overall the performance of ZDOCK when using modeled protein structures were found to be comparable to docking starting from the unbound experimental structures when using the cutoff of 5.0 Å to define a hit. Figure 4.6 shows the comparison of model-model and unbound native-unbound native docking when the top 1000 decoys are considered. Interestingly, the quality of the decoy pool remains very similar and the performance of ZDOCK when using I-TASSER modeled structures or experimental unbound structure proceeds neck-and-neck till the top 1000 decoys.

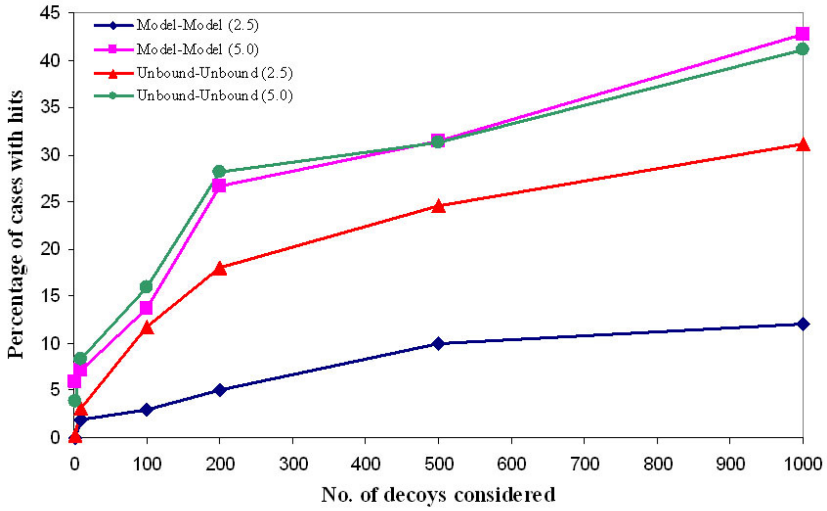

Figure 4.6: Success rate of ZDOCK on I-TASSER models versus that on the unbound x-ray structures.

Playing devil's advocate, it can be argued that the performance of ZDOCK when using experimental unbound structures was significantly better when a cutoff of 2.5 Å is used to define a hit. However, this is very much expected since, no matter how accurate current state-of-the-art structure prediction is, it will contain small local errors and since model-model docking involves two models per case (meaning double the error) a 5.0 Å cutoff is still reasonable. Further, even though the criterion for a hit is slightly lenient, as is the case when using a 5.0 Å cutoff, the fact that both methods were compared using the same criteria ensures that the analysis does not lose out on objectivity. Hence, it can be stated that model-model docking can be considered a viable alternative to experimental unbound structure docking and therefore can be reliably used when the unbound experimental structures of the subunits of a complex are not known.

Comparison of TACOS performance against model-model docking: To obtain an objective analysis of the performance of TACOS as compared to rigid-body docking algorithms, both chains of the 350 complex test set were modeled using I-TASSER. Since I-TASSER is an in-house algorithm, we also modeled the constituent chains of the dimers using MODELLER[27] with MUSTER templates and alignments being provided as input. The models of the individual models were subsequently docked using ZDOCK followed by refinement using RDOCK[14] (the methods are referred to as ITASSER-ZDOCK and MODELLER-ZDOCK). The final models (best in Top 5 models) were evaluated according to the following criterion i) TM-score ii) rTM-score iii) Median I-RMSD iv) Number of hits v) Inter-chain contact accuracy and coverage. Modeling by both I-TASSER and MODELLER was carried out using the "benchmark" setting i.e. all templates with sequence identity ≥ 30% to the target sequence were excluded.

Out of the 700 individual chains (350×2) 521 were classified as "easy" targets, 123 were classified as "medium" targets and 106 were classified as "hard" targets by LOMETS. The average TM-score of the 700 models generated by I-TASSER was 0.663 while that by MODELLER was 0.612. Overall, among the 350 dimers of the test set, 223 cases had both chains modeled by I-TASSER with a TM-score ≥ 0.5 which according to Xu and Zhang[28] indicates the model share the same basic topology as the target.

Method	**rTM-score**	**TM-score**	**Median I-RMSD**[a]	**C_{Acc}/C_{Cov}**[b]	**Hits**[c]
TACOS	0.377	0.475	8.34 Å	46.8%/45.3%	107
ITASSER-ZDOCK	0.354	0.443	8.97 Å	45.6%/44.3%	102
MODELLER-ZDOCK	0.326	0.411	9.83 Å	32.4%/41.4%	88
COTH threading	0.313	0.384	6.66 Å	33.9%/34.8%	-

[a] I-RMSD stands for Interface Root Mean Square deviation. Since the range of I-RMSD can vary greatly when the orientations of the chains are wrong we calculate the median value instead of the mean.

[b] C_{Acc}/C_{Cov} stands for the average accuracy/coverage of prediction of native inter-chain contacts in the best in top 5 models (according to TM-score) as compared to native (SI 4).

[c] "Hits" is defined as the number of cases in the 350 complex test set which has atleast model in top 5 predictions with an I-RMSD ≤ 5.0 Å to native.

Table 4.1. Comparison of performance of TACOS with respect to controls.

As shown in Table 4.1, TACOS outperforms both ITASSER-ZDOCK and MODELLER-ZDOCK in terms of all four evaluation criteria. Interestingly, even COTH threading alone is comparable to MODELLER-ZDOCK in-term of average TM-score and rTM-score but as reportedly previously, COTH threading acts in complementarity with rigid-body docking. Overall, TACOS outperforms ITASSER-ZDOCK 0.377/0.475 to 0.354/0.443 in terms of mean TM-score/rTM-score of the best in top 5 models, an improvement of 6.5%/7.2%. When compared to MODELLER-ZDOCK the scores are 0.326/0.411 for TACOS and MODELLER-ZDOCK respectively or an improvement of 15.6%/15.5%. Importantly, TACOS beats ITASSER-ZDOCK in 226 out of 350 cases while it beats MODELLER-ZDOCK in 259 out of 350 cases. In terms of median I-RMSD as well, TACOS shows an improvement of 6.9%/15.2% over ITASSER-ZDOCK/MODELLER-ZDOCK. If the number of hits are considered, TACOS generates atleast 1 hit among the top 5 models in 107 out of 350 cases while ITASSER-ZDOCK and MODELLER-ZDOCK produces a hit in 102 and 88 cases respectively.

Among the 110 targets which are classified as TACOS “hard”, the average TM-score/rTM-score of ITASSER-ZDOCK is slightly higher than that of TACOS while that of MODELLER-ZDOCK is comparable. However, among the 110 hard targets, 69 have atleast 1 chain which are hard targets for I-TASSER as well (as classified by LOMETS in this case). If only these 69 targets are considered, then the average TM-score/rTM-score of TACOS is higher than that of ITASSER-ZDOCK and MODELLER-ZDOCK. Therefore, according to our evaluation, TACOS is a more robust and versatile option to model-model docking, in order predict protein-protein complex structures when the experimentally determined constituents of the dimer are not known.

Predicting native contacts: One of the most important objectives of complex structure modeling is to predict the native inter-chain contacts correctly. Therefore, we found it imperative to assess the predicted complex structures based on accuracy and coverage of native contact prediction. As a control, we also assessed the performance of ITASSER-ZDOCK in correctly predicting contacts at the interface. Even though TACOS lays more emphasis on predicting the complex as a whole, it still predicts the native contacts with higher reliability than ITASSER-ZDOCK. Overall the prediction accuracy of inter-chain contacts by TACOS is 46.8% with coverage of 45.3% while that for I-TASSER-ZDOCK is 45.6% with coverage of 44.3%. Moreover, TACOS has 43 targets with both C_{Acc} and C_{Cov} above 75% while ITASSER-ZDOCK has 37 cases in the corresponding category. Based on performance therefore, it can be concluded that TACOS is successful not only in predicting the overall structure of complexes but also successful in recuperating reliably, the native inter-chain contacts starting from sequence alone.

4.1.4 Performance of TACOS across different protein complex classes

Protein complexes can be categorized into various classes based on different criteria like sequence identity of constituent chains to each other (homo and heterodimers), lifetime of association (permanent and transient) and biological process involved (enzyme-inhibitor, antigen-antibody and others) to name a few. The different complex classes can have few idiosyncratic characteristics which makes them unique. We therefore considered it prudent to analyze the performance of COTH across different categorizes of complexes in order to check its consistency. However, we restricted our analysis to homo- and heterodimers and according to the biological process involved. Unfortunately, though a very important categorization, it is often not possible to automatically judge (based just on its structure or sequence) the lifetime of a protein complex. Hence, it was not possible to judge the performance of TACOS on permanent and transient complexes.

Overall, if the performance of TACOS for homodimers is better than that for heterodimers. For the 213 homodimers in the test dataset, the average TM-score/rTM-score for the best in top 10 templates is 0.513/0.402 while that for the heterodimers the mean scores are 0.415/0.338. First, the number of heterodimers in our library is lesser and in general heterodimers have more sequence diversity which is why COTH threading finds it more difficult to identify good templates from which full length structures can be modeled. Secondly, unlike homodimers, many types of heterodimeric interactions are side chain dependent (especially antigen-antibody interaction) which our course-grained approach is less successful in reproducing. However, in some case TACOS is still able to predict near-native models of heterodimeric complexes. In Figure 4.5 we show examples of near-native structure (one homodimers, one heterodimer) predicted by TACOS. Among the 137 heterodimers, there are 76 enzyme-inhibitor complexes, 39 antigen-antibody complexes and 22 other complexes. The performance of TACOS for enzyme-inhibitor complexes was significantly better as compared to antigen-antibody complex in terms of rTM-score, a trend similar to that of COTH threading. The reason for this can be attributed to the fact that even though TACOS predicted the antibody chain with high TM-score in most cases, the orientation of the antigen chain was often incorrect leading to low rTM-scores.

4.2 MATERIALS AND METHODS

The TACOS methodology has four distinct steps; 1) template selection 2) mapping of dimer onto an artificial monomer placed on lattice 3) structure assembly and 4) model selection and refinement. Each of the steps is described in detail in the following.

4.2.1 Template selection

In the first step, TACOS attempts to identify homologs/structural analogs of the given complex query sequence by threading it across representative monomer and complex structure libraries screened at 70% sequence identity by using the COTH complex threading algorithm [29]. The search for templates by COTH requires two distinct steps. First, the complete dimeric sequence is threaded across a complex structure library by "COTH threading". The scoring function of COTH threading includes multiple sources of predicted structure information like secondary structure, solvent accessibility, torsion angles, hydrophobicity as well as putative interface residues predicted using a *ab initio* neural network based approach. COTH threading also implements a novel modification of the Needleman-Wunsch dynamic programming where the alignment of chains between the query and the template are forbidden.

In the second step, the individual chains of the query complex are threaded across the six times larger monomer structure library by LOMETS. The individual chain templates thus identified by LOMETS are thereby superimposed on the dimer template framework identified by COTH threading. The use of LOMETS to identify templates for the individual chains of the complex seeks to circumvent the problem of the incompleteness of the complex structure library. The use of the much larger monomer library ensures more accurate templates for the individual chains while superposing them on the COTH threading framework ensures the orientation information indentified by dimeric threading is retained.

Based on the hypothesis that the COTH threading templates and the LOMETS templates provides valuable information regarding the structure of the complex, restraints are added to the TACOS simulation which ensures that the decoys retain the information of the templates and are not completely driven away from it. Thus, intra-chain distance and contact restrains are generated from the LOMETS templates while inter-chain restraints containing orientation information of 2 chains are collected from the COTH threading templates and used as energy terms to drive the TACOS simulation.

4.2.2 Mapping of dimer onto an artificial monomer on CAS lattice

The entire complex structure is course-grained and represented only by the Cα atom and the side chain center of mass (SG). First, the template unaligned or gapped regions are built using a Cα random walk to connect the continuous template aligned fragments and build an initial full length structure. If any of the unaligned regions between two aligned fragments cannot be connected completely by 3.8Å Cα-Cα bonds then a large bond remains and an external spring like force is applied until a reasonable bond length is achieved. The complex structure is then segregated into "template-aligned" and "template-unaligned" regions and placed on the CAS on and off lattice model also used by I-TASSER. Here, the Cα atoms of the template un-aligned (gapped) regions are placed on-lattice and are treated as "to be built *de novo*" while the template aligned fragments are placed off-lattice and subjected only to rigid body adjustments. The SG atoms are always placed off-lattice. Finally, the dimer is represented on lattice as an 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 to treat the complex as essentially a monomer prediction problem and ensures the direct adoption of the many I-TASSER energy potentials and movement schemes to TACOS. The representation of the complex structure on the CAS lattice joined by the pseudobond has been shown in Figure 4.7.

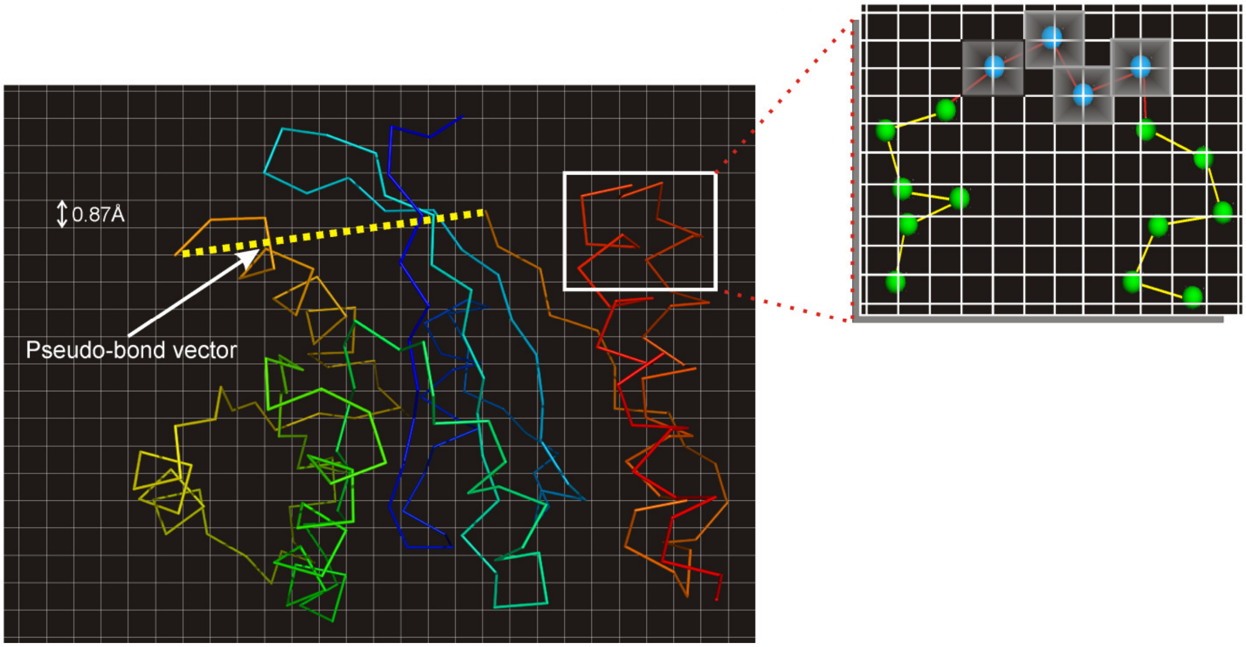

Figure 4.7: Plot showing CAS lattice based representation of dimer structure. A completely flexible pseudo-bond is used to connect the C-terminal of the first chain with the N-terminal of the second chain. The dimer is represented in an on- and off- lattice system as shown in the blowout (top-right) with the template aligned Cα residues being placed off-lattice (green spheres) and the unaligned Cα residues (blue spheres) are placed on-lattice.

4.2.3 Structure Assembly: Movements

Local intra-chain movements. The local intra-chain moves, adapted from I-TASSER, can be classified into two types 1) On-lattice bond rebuilding and 2) Off-lattice rigid body moves for continuous template aligned fragments. On-lattice movement includes extensive bond rebuilding moves for *ab inito* generation of the template unaligned regions. 312 basis vectors or bond vectors, restricted to the cubic-lattice points, are pre-computed with bond lengths varying between 3.26 Å and 4.25 Å (this allows for larger conformational flexibility). The average bond length is 3.8 Å which corresponds to the physiological Cα-Cα bond length. The rebuilding steps can be further divided into 3 types i) Basic 2-bond and 3-bond movements: here 2 to 3 continuous bonds in the present decoy is replaced by the pre-fabricated bond vectors of the same length. ii) Higher 4-, 5- and 6- bond movements: this is effected by replacement of 4-,5- or 6- continuous bonds in the present decoy by a combination of the 2- and 3-bond moves. iii) Terminal random walk: here the entire region between any random point *m* to the N- or C-terminal is rebuilt using a Cα random walk. For the off-lattice moves, a continuous template aligned fragments is subjected to a rigid body translation, rotation and deformation (2- or 3- bond movement within the fragment).

Inter-chain movement: This movement was built on the premise that initial orientation of the two chains may be incorrect or require readjustments. Accordingly, a new inter-chain movement was designed to explore the orientational search space. Here, one of the chains of the dimer (smaller one for heterodimers and either one for homodimers) is first randomly moved to any new position. Following this, the vector between the center of the mass (COM) of both chains is defined and the chain is subjected to randomly selected small rigid body rotation and translation motions (while keeping the center of mass fixed along the COM vector) with each move being selected or rejected based on the standard Metropolis monte-carlo criteria. Newly defined inter-chain specific potential terms were defined to guide the movement. It should be noted here that the large inter-chain move disrupts the lattice representation of the mobile chain and hence the CAS lattice is redefined after every inter-chain move. At the end of one cycle of the inter-chain move, the final energy (inter-chain energy plus local conformational energy) is calculated and the new position is rejected or accepted once again based on the standard Metropolis criteria. The overall simulation schema, balancing large inter-chain moves with local refinements is shown in Figure 4.8.

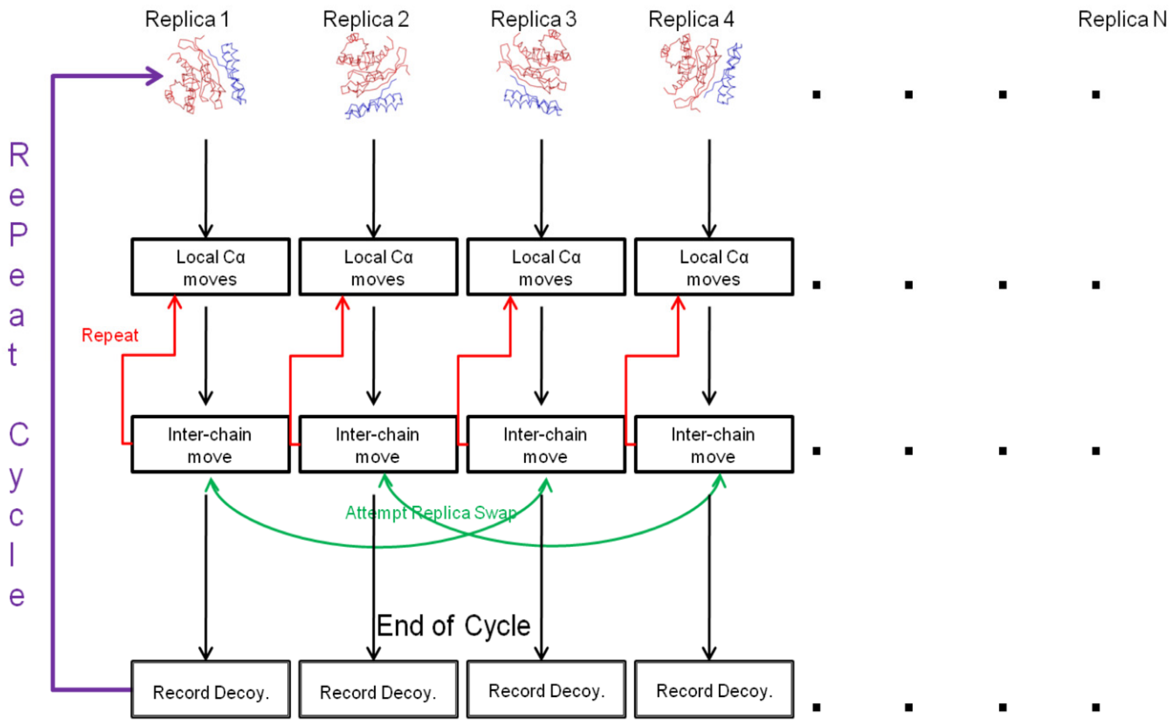

Figure 4.8. Flowchart showing the replica-exchange monte-carlo movement schematic of the TACOS simulation.

4.2.4 Structure Assembly: Inter-chain energy terms

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[15, 30] to be directly carried over and are 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 complex structure library. The new energy terms are discussed in more detail in the following. The *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.

i) E_{COM}: Computes the distance between the Center of Mass (COM) of the two dimer chains and is required to prevent the two chains from drifting too far away during the simulation procedure where the equation is given by

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

where d_{COM} is the distance between the two center of masses. On the other hand, this potential can dictate one chain into collapsing onto the other and hence 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 on 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 atleast 30 inter-chain contacts are required for a stable complex formation, a large penalty was assessed for decoys with no inter-chain contacts and then gradually decreased as more inter-chain contacts were 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

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

where

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

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 a amino acid pair to be in orientation $\gamma_{i,j}$ (total number of times any particular amino acid pair is observed in orientation γ divided by the total number of times the that particular amino acid pairing is observed).

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:

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

where

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

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

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

where

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

Here, $q(A_i,A_j,\lambda_{i,j})$ is derived from the 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

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

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 ($d_{ij} \leq 4.5$ Å) in multiple templates but are not in contact in a given decoy.

4.2.5 Evaluation

The similarity of protein tertiary structures is often evaluated by TM-score[31], which can be simply extended to the comparison of complex structures:

$$\text{TM} - \text{score} = \max\left[\frac{1}{L_c}\sum_{i=1}^{L_{ali}}\frac{1}{1+\left(\frac{d_i}{d_0(L_c)}\right)^2}\right] \qquad (11)$$

where L_c is the total length of all chains in the target complex and L_ali is the number of the aligned residue pairs in the two complexes. d_i is the distance of *i*th pair of Cα atoms after the superposition of the complex structures. $d_0(L_\text{c}) = \sqrt[3]{L_\text{c} - 15} - 1.8$ is a length-dependent scale to normalize the distance so that the overall TM-score of random complex structures is independent of the protein size. max[…] indicates the optimal superposition to maximize the overall TM-score value.

For complexes, TM-score in Eq. 3 can be factorized as two additive parts from two chains:

$$\mathrm{TM-score} = \frac{L_r}{L_c}\mathrm{TM-score}_r + \frac{L_l}{L_c}\mathrm{TM-score}_l \qquad (12)$$

where L_r and L_l are lengths of the receptor and ligand, respectively; TM-score$_\mathrm{r}$ and TM-score$_\mathrm{l}$ are their TM-scores calculated based on the same rotation matrix of the complex superposition. Therefore, one disadvantage of TM-score, when used to compare complex structures, is that it becomes more sensitive to the tertiary structure of the monomers, due to the linear dependence of the monomer TM-scores. For example, for a pair of homodimers, if the structure of one chain is identical, the TM-score is at least 0.5 even if the orientation of the other chain is completely different. For heterodimeric complexes, if one chain is much bigger than the other, the TM-score can be dominated by the structural similarity of the bigger chain regardless of the structure and orientation of the smaller chain because the weighting factor for the small chain (L_l/L_c) is too small in Eq. 12. To overcome this drawback, we define a new score called reciprocal TM-score, or rTM-score, given by

$$\mathrm{rTM\text{-}score} = \frac{2}{\dfrac{1}{\mathrm{TM\text{-}score_r}} + \dfrac{1}{\mathrm{TM\text{-}score_l}}}. \qquad (13)$$

Here, the factor 2 in the numerator is used to normalize the range of rTM-score within [0, 1].

The definition of rTM-score in Eq. 13 makes the score more sensitive to the overall structure similarity of the complex, i.e. the relative orientation of the component chains, rather than the individual monomer structures. For instance, if the structure or orientation of one chain is very different (i.e. TM-score$_\mathrm{l}$~0), the rTM-score of the complex structure will be close to 0 even if the structure of another chain is identical (TM-score$_\mathrm{r}$~1). In other words, two complexes have a high rTM-score only when both the monomer tertiary structure and the relative orientation are similar.

Quantitatively, for tertiary protein structures, it has been shown[28] that the posterior probability of TM-score of random protein structure pairs has a rapid phase transition at TM-score=0.5 and the structures of TM-score >0.5 approximately corresponds to the same protein folds as defined by the SCOP[32] and CATH[33] databases. Similarly, we define rTM-score >0.5 as the complexes of the same interactions. Mathematically, this corresponds to two protein complexes which have two chains with the similar relative orientation and the similar folds (i.e. TM-score$_{r,l}$ > 0.5) according to Eq. 13.

Additional to rTM-score/TM-score of the global complex structure, we also assess the modeling quality of protein-protein interface structures, the quality of which is of key importance for the functional annotation of protein complexes. Here, a residue is defined to be at the interface if the distance of the Cα atom to any Cα atoms in the counterpart chain is below 10 Å. The interface RMSD, I-RMSD, is the root-mean-squared-deviation of the model and the native structure in the aligned region of the interfaces. Accordingly a "hit" is defined as a target where at least one of the top 5 five models has an I-RMSD ≤ 5Å. For the 350 targets of the test set, since the I-RMSD values can be very large if the orientation of the chains is incorrect, the average I-RMSD can be skewed due to these extremely high values. Hence, we use median I-RMSD instead as an additional measure of the performance of the TACOS method compared to control methods.

Finally, it is of great importance in complex structural biology to correctly identify inter-chain residue-residue contacts at the protein-protein interface. Therefore we define two additional score accuracy and coverage (C_{Acc} and $C_{cov)}$ to assess the performance of the various methods in correctly predicting these all important inter-chain contacts. The accuracy of interface contact predictions is defined as the number of the correctly predicted contacts across two chains divided by the total number of cross-chain contacts in the model; the coverage is the number of correctly predicted interface contacts divided by the observed cross-chain contacts in the native structure. They are given by the equations

$$C_{Acc} = \frac{\text{No. of correctly predicted contacts}}{\text{No. of predicted contacts}} \tag{14}$$

$$C_{Cov} = \frac{\text{No. of correctly predicted contacts}}{\text{No. of actual inter - chain contacts in native complex}} \tag{15}$$

4.3 DISCUSSION

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 and therefore first identifies putative templates from a non-redundant protein complex structure library by COTH 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 refines 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 random move of one chain followed by a short independent metropolis monte-carlo simulation 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 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 admirably and can predict full length structures with the same basic fold in ~60% cases. Importantly, the TACOS simulation is highly successful in refining the initial threading templates with an increase in TM-score noted for > 80% cases. It was also noted that the increase in TM-score was not simply a product of increased coverage as TACOS is also successful in refining the templates aligned regions in 70% cases. However there are a few cases which show an overall regression of the model as compared to the initial top template which provides direction for attempting further improvements.

Modeling of the 2 subunits of a dimer separately followed by docking them together using rigid-body protein docking algorithms can serve as alternative to TACOS when prediction of the complex structure starting from sequence alone is desired. Hence, we compared the performance of TACOS to ITASSER-ZDOCK and MODELLER-ZDOCK which model the component chains separately using I-TASSER and MODELLER respectively and then dock the modeled structures together using ZDOCK. Overall, TACOS is found to perform better than either approach in terms of a number of different evaluation criteria. There exist some cases which are hard modeling targets for TACOS due to the lack of reliable dimeric templates but the component chains are easy modeling targets individually. These cases can be modeled better overall with ITASSER-ZDOCK than is possible using TACOS. Another important observation that was noted was that TACOS performed better overall for homodimers than for heterodimers and among the heterodimers, enzyme-inhibitor complexes are comparatively easier for TACOS than antigen-antibody complexes.

TACOS thus represents one of the first algorithms designed to predict the structure of dimeric protein complexes given the sequence alone. Importantly, the high incompleteness of the protein complex structure library implies that the performance of TACOS will improve in the years to come as more and more complex structures are available in the PDB. 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 version of TACOS which can refine rigid-body docking decoys to model interaction induced backbone conformational changes is currently under preparation. 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/.

4.4 REFERENCES

1. Moult, J., et al., *Critical assessment of methods of protein structure prediction - Round VIII.* Proteins, 2009. **77 Suppl 9**: p. 1-4.
2. Zhang, Y., *Progress and challenges in protein structure prediction.* Curr Opin Struct Biol, 2008. **18**(3): p. 342-8.
3. Zhang, Y., *Protein structure prediction: when is it useful?* Curr Opin Struct Biol, 2009. **19**(2): p. 145-55.
4. Moult, J., et al., *Cristical assessment of methods of protein structure prediction (CASP) - round VII* Proteins, 2007. **69**(Suppl 8): p. 3-9.
5. Chen, H. and J. Skolnick, *M-TASSER: an algorithm for protein quaternary structure prediction.* Biophys J, 2008. **94**(3): p. 918-28.

6. Vajda, S. and C.J. Camacho, *Protein-protein docking: is the glass half-full or half-empty?* Trends Biotechnol, 2004. **22**(3): p. 110-6.

7. Aloy, P., M. Pichaud, and R.B. Russell, *Protein complexes: structure prediction challenges for the 21st century.* Curr Opin Struct Biol, 2005. **15**(1): p. 15-22.

8. Russell, R.B., et al., *A structural perspective on protein-protein interactions.* Curr Opin Struct Biol, 2004. **14**(3): p. 313-24.

9. Lensink, M. and S. Wodak, *Docking and scoring protein interactions: CAPRI 2009.* Proteins, 2010: p. DOI: 10.1002/prot.22818.

10. Aloy, P., et al., *Structure-based assembly of protein complexes in yeast.* Science, 2004. **303**(5666): p. 2026-9.

11. Kundrotas, P., M. Lensink, and E. Alexov, *Homology based modelling of 3D structures of protein complexes using alignments of modified sequence profiles.* International Journal of Biological Macromolecules, 2008. **43**(2): p. 198-208.

12. Lu, L., H. Lu, and J. Skolnick, *MULTIPROSPECTOR: An algotihm for the predictio of protein-protein interactions by multimeric threading.* Proteins:Structure,Function and Genetics, 2002. **49**: p. 350-364.

13. Chen, R., L. Li, and Z. Weng, *ZDOCK: an initial-stage protein-docking algorithm.* Proteins, 2003. **52**(1): p. 80-7.

14. Li, L., R. Chen, and Z. Weng, *RDOCK: refinement of rigid-body protein docking predictions.* Proteins, 2003. **53**(3): p. 693-707.

15. Wu, S., J. Skolnick, and Y. Zhang, *Ab initio modelling of small proteins by iterative TASSER simulations.* BMC Biol, 2007. **5**: p. 17.

16. Zhang, Y. and J. Skolnick, *Automated Structure Prediction of Weekly Homologous Proteins on a Genomic Scale* Proceedings of The National Academy of Science, 2004. **101**: p. 7594-7599.

17. Mariani, V., et al., *Assessment of template based protein structure predictions in CASP9.* Proteins: Structure, Function, and Bioinformatics, 2011. **79**(S10): p. 37-58.

18. Gray, J.J., et al., *Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.* J Mol Biol, 2003. **331**(1): p. 281-99.

19. Gray, J.J., et al., *Protein-protein docking predictions for the CAPRI experiment.* Proteins, 2003. **52**(1): p. 118-22.

20. Katchalski-Katzir, E., et al., *Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.* Proc Natl Acad Sci U S A, 1992. **89**(6): p. 2195-9.

21. Tovchigrechko, A. and I.A. Vakser, *GRAMM-X public web server for protein-protein docking.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W310-4.

22. Vakser, I.A., *Protein docking for low-resolution structures.* Protein Eng, 1995. **8**(4): p. 371-7.

23. Comeau, S.R., et al., *ClusPro: a fully automated algorithm for protein-protein docking.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W96-9.

24. Tovchigrechko, A., C.A. Wells, and I.A. Vakser, *Docking of protein models.* Protein Sci, 2002. **11**(8): p. 1888-96.

25. Kundrotas, P.J. and I.A. Vakser, *Accuracy of protein-protein binding sites in high-throughput template-based modeling.* PLoS Comput Biol, 2010. **6**(4): p. e1000727.

26. Hwang, H., et al., *Protein-protein docking benchmark version 3.0.* Proteins, 2008. **73**(3): p. 705-9.

27. Sali, A. and T. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* Journal of Molecular Biology, 1993. **234**: p. 779-815.

28. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score = 0.5?* Bioinformatics, 2010. **26**(7): p. 889-95.

29. Mukherjee, S. and Y. Zhang, *Protein-protein complex structure predictions by multimeric threading and template recombination.* Structure, 2011. **19**(7): p. 955-66.

30. Zhang, Y., A. Kolinski, and J. Skolnick, *TOUCHSTONE II: a new approach to ab initio protein structure prediction* Biophysical Journal, 2003. **85**: p. 1145-1164.

31. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**: p. 702-710.

32. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J. Mol. Biol., 1995. **247**(4): p. 536-40.

33. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-1108.

CHAPTER 5. Conclusions

Even though comparative modeling of protein tertiary structures have made rapid progress in the last two decades[1-3], homology-based modeling approaches for protein-protein complex structures have remained few and far between[4-9]. The problem of predicting the structure of a protein complex has thus largely been restricted to a "docking" problem where the unbound subunits of the complex are fitted or docked together if the native structures of the unbound subunits are known [10-14]. This work therefore presents one of the first comprehensive efforts to utilize the vast information that can be obtained from evolutionary relationships towards the development of methods capable of predicting the structure of a protein-protein complex from its primary amino acid sequence alone. As more and more protein complex structures are experimentally solved and the protein complex structure library nears completeness, the efficiency and accuracy of homology assisted methods like COTH and TACOS are expected to increase manifolds. Traditionally *ab initio* methods like docking have both limited accuracy and scope. It is therefore hoped that the development of methods discussed here and new ones developed in the future will help bring about a paradigm shift in the way the scientific community approaches the problem of predicting protein complex structures computationally.

Initially, it was imperative to have a tool which could be used to compare the "similarity" of two protein complex structures in a quantitative fashion. To this end, MM-align was developed to structurally align 2 complex structures and return a RMSD and a TM-score of the complexes as a whole. These scores can be used to quantitatively assess structural homology/analogy of two structures. In a related work, MM-align was used to structurally align all known non-redundant protein complex structure families (defined according to Pfam classification of two chains). Clustering of related structures thereby revealed that 62% of the known structures are “orphans” i.e. no other complex structures exists in the current protein complex structure library shares the same quaternary fold as these structures (shown in Figure 6.1). Importantly, a strict logarithmic dependence was observed between the number of current known quaternary structure families and the number of quaternary structure folds currently known which could be then used to estimate that approximately 4000 unique quaternary folds are expected to exist in nature.

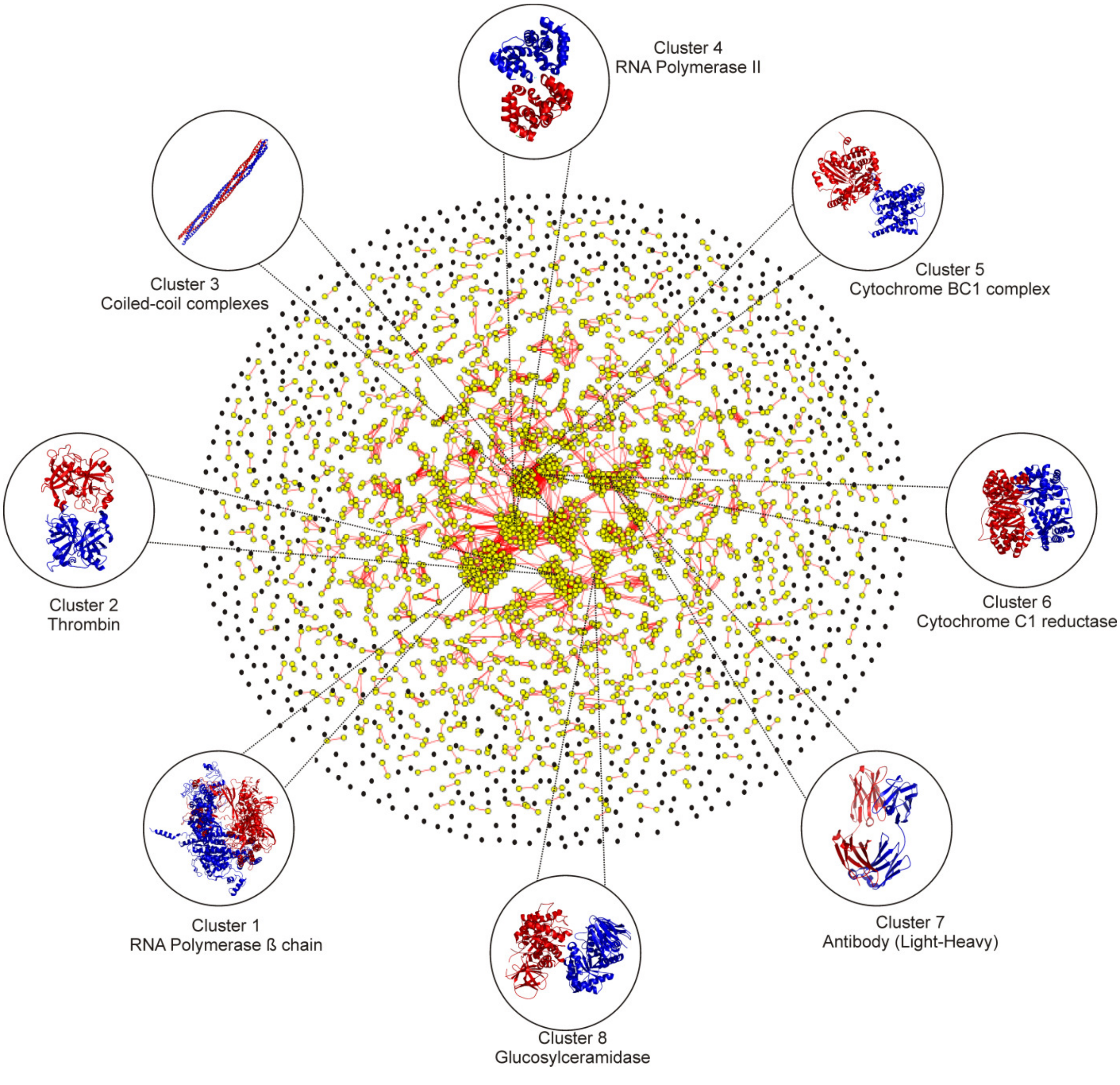

Figure 5.1 Graphical network representation of the similarity of protein-protein complex structures by Cytoscape. Each node represents a known complex structure and two nodes are connected by an edge if the rTM-score between the two structures is >0.5. The orphan nodes are shown in black while nodes which are connected by at least one edge are shown in yellow. Representative examples from the eight largest clusters are listed together with the protein name.

Also, the fact that functionally homologous protein complexes could be identified from the library of protein complex structures (Appendix I) lent credence to the possibility of using threading based methods to predict the structure of protein-protein complexes given the primary amino acid sequence as input.

Buoyed by the information gathered from MM-align that protein complex structures are evolutionarily conserved, COTH was developed to recognize the possible fold adopted by a query dimer sequence based on sequence-structure alignment. COTH (which was tested on 500 non-redundant protein complex structures) was able to correctly predict a structure sharing the same global fold for > 50% cases. This highlighted the fact that prediction of protein complex structures from sequence is indeed a realistic possibility even though the protein complex structure library is largely incomplete. Furthermore, COTH was found to be complementary to *ab initio* methods like rigid body docking. It can therefore be argued that until more unique folds are deposited in the PDB, a combination of COTH like threading based methods and rigid body docking methods maximizes the possibility of generation of acceptable quality models.

Finally, a template fragment based reassembly and refinement protocol, TACOS, was developed to build upon the initial templates detected by COTH to generate full-length protein complex structures. TACOS was designed to perform as a hybrid methodology which incorporated features of homology modeling (rigid-body motion of template aligned fragments), *ab initio* structure prediction (*de novo* generation of template unaligned regions) and protein-protein docking (inter-chain rigid body movement). A new knowledge-based energy function was developed to drive the replica-exchange monte-carlo sampling scheme which contained intra-chain and inter-chain restraints coupled with statistically derived potential terms both at the tertiary structure level as well as at the protein-protein interface level. TACOS was extensively optimized and a large-scale benchmarking proved that on average TACOS was able to outperform the more traditional approach of modeling the subunits of a complex separately and then docking them together.

Since TACOS attempts to tackle the problem of protein structure prediction and protein-protein interaction simultaneously, it is computationally expensive. Secondly, when the structures of unbound subunits of a complex are known, it is not desirable to predict the individual chains from the sequence. On the other hand, if the unbound subunit structures are known, rigid-body docking fails to account for backbone conformational changes, especially at the interface. TACOS treats the backbone as flexible and can hence be realistically used to refine initial complex structures generated by rigid body docking. The use of TACOS to model the binding induced conformational change starting from initial docked models would thus present a more direct approach as compared to the protocols currently adopted [15-16].

Conveniently, TACOS does not depend on the presence of known structures of the unbound subunits of protein complex and hence it can be used to conduct modeling experiments of protein complexes on a genomic scale. High-throughput experimental methods have made available the protein interaction networks of a number of genomes [17-23] along with their sequences. Databases like Database of Interacting Proteins (DIP) [24], MIPS [25] among others store genomic scale protein interaction data in a user-friendly, easily downloadable fashion. Therefore, the possibility of modeling all protein complexes in a genome no longer remains a far-fetched pipe dream.

Finally, even though TACOS currently is concentrated on modeling protein dimers only, it can technically be extended to model higher order oligomers. Admittedly, predicting the structure of higher order multimers is a very challenging problem since the degrees of freedom involved is prohibitably large. The major reason why TACOS currently does not attempt to model larger oligomers is the lack of sufficient number of templates in the PDB library. However, when more number of larger oligomers are deposited in the PDB or given some experimental restraints to guide the simulation process, modeling these larger macromolecular structures will be a distinct possibility.

5.1 REFERENCES

1. Zhang, Y., *Progress and challenges in protein structure prediction.* Curr Opin Struct Biol, 2008. **18**(3): p. 342-8.
2. Moult, J., et al., *Cristical assessment of methods of protein structure prediction (CASP) - round VII* Proteins, 2007. **69**(Suppl 8): p. 3-9.
3. Moult, J., et al., *Critical assessment of methods of protein structure prediction - Round VIII.* Proteins, 2009. **77 Suppl 9**: p. 1-4.
4. 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.* Int J Biol Macromol, 2008. **43**(2): p. 198-208.
5. Aloy, P., et al., *Structure-based assembly of protein complexes in yeast.* Science, 2004. **303**(5666): p. 2026-9.
6. Aloy, P. and R.B. Russell, *Ten thousand interactions for the molecular biologist.* Nat Biotechnol, 2004. **22**(10): p. 1317-21.

7. Russell, R.B., et al., *A structural perspective on protein-protein interactions.* Curr Opin Struct Biol, 2004. **14**(3): p. 313-24.

8. Chen, H. and J. Skolnick, *M-TASSER: an algorithm for protein quaternary structure prediction.* Biophys J, 2008. **94**(3): p. 918-28.

9. Lu, L., H. Lu, and J. Skolnick, *MULTIPROSPECTOR: An algotihm for the predictio of protein-protein interactions by multimeric threading.* Proteins:Structure,Function and Genetics, 2002. **49**: p. 350-364.

10. Kozakov, D., et al., *Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19.* Proteins, 2010. **78**(15): p. 3124-30.

11. Gray, J.J., et al., *Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.* J Mol Biol, 2003. **331**(1): p. 281-99.

12. Tovchigrechko, A. and I.A. Vakser, *Development and testing of an automated approach to protein docking.* Proteins, 2005. **60**(2): p. 296-301.

13. Katchalski-Katzir, E., et al., *Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.* Proc Natl Acad Sci U S A, 1992. **89**(6): p. 2195-9.

14. Hwang, H., et al., *Performance of ZDOCK and ZRANK in CAPRI rounds 13-19.* Proteins, 2010. **78**(15): p. 3104-10.

15. Bonvin, A.M., *Flexible protein-protein docking.* Curr Opin Struct Biol, 2006. **16**(2): p. 194-200.

16. Sircar, A., et al., *A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13-19.* Proteins, 2010. **78**(15): p. 3115-23.

17. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-1736.

18. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proceedings of The National Academy of Science, 2001. **98**(8): p. 4569-4574.

19. Gavin, A., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141-147.

20. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-627.

21. Ho, Y., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectroscopy.* Nature, 2002. **415**(6868): p. 180-183.

22. Rain, J., et al., *The protein-protein interaction map of Helibacter pylori.* Nature, 2001. **409**(6817): p. 211-215.

23. Li, S., et al., *A map of the interactome network of the metazoan C.elegans.* Science, 2004. **303**(5657): p. 540-543.

24. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.

25. Mewes, H.W., et al., *MIPS: a database for genomes and protein sequences.* Nucleic Acids Res, 2002. **30**(1): p. 31-4.

Appendix I

Query PDB ID[a] (Chain1,Chain2)	PDB Class[b]: GO term[c]	Temp PDB ID[d] (Chain1,Chain2)	PDB Class[b]: GO term[c]	TM-score[e]	R(Å)[f]
1djt (A,B)	Toxin: Ion channel inhibitor activity	1aap (A,B)	Protease/Inhibitor complex: serine type endopeptidase activity and inhibitor	0.379	4.7
1dkf (A,B)	Hormone/Growth Factor Receptor: DNA binding and transciption factor activity	1xb7 (A1,A2)	DNA binding and Transcription factor activity	0.849	3.1
1dl5 (A,B)	Transferase: Methyltransferase activity	1utx (A,B)	DNA binding protein: Sequence specific DNA binding	0.442	3.9
1dlf (L,H)	Immunoglobin: Antibody	1j05 (L,H)	Immunoglobin: Antigen binding	0.935	1.5
1do5 (A,B)	Chaperone: Copper ion binding activity	1xso (A,B)	Superoxide Acceptor: Copper ion binding	0.960	1.0
1dos (A,B)	lyase: fructose-biphosphate aldolase	1rvg (A,B)	lyase: fructose-biphosphate aldolase	0.766	3.8
1dp4 (A,C)	Hormone/Growth Factor Receptor: protein kinase activity	1dz3 (A1,A2)	Response Regulator: DNA binding and transcription factor activity	0.462	3.3
1dqp (A,B)	Transferase: transferring glycosyl groups	1grv (A,B)	Transferase: transferring glycosyl groups	0.857	3.0
1dqw (A,B)	Lyase: orotidin-5-phosphate decarboxylase	2jgy (A,B)	Transferase: orotidin-5-phosphate decarboxylase	0.952	1.8
1ds6 (A,B)	Signalling	1cc0 (A,E)	Signalling	0.899	2.2

	Protein: GTPase activity		Protein: GTPase activity		
1dx5 (M,I)	Trypsin like Serine Protease	1h9h (E,I)	Trypsin like Serine Protease	0.842	1.6
1dz3 (A1,A2)	Response Regulator: 2-component response regulator activity	1srr (A1,A2)	Response Regulator: 2-component response regulator activity	0.586	4.3
1ega (A,B)	Hydrolase: GTP binding	1j2j (A,B)	Transport protein: GTP binding	0.565	3.4
1e05 (L,I)	Serpin: Endopeptidase inhibitor activity involved in blood coagulation	1oc0 (A,B)	Serine Protease inhibitor: Endopeptidase inhibitor activity involved in blood coagulation	0.835	2.1
1e0b (A,B)	Chromatin binding: Transcription regulator	1igq (A,C)	DNA binding: Transcription regulator	0.458	3.8
1e0o (C,D)	Growth factor: Heparin binding protein with kinase activity	1ev2 (A,E)	Growth factor: Heparin binding protein with kinase activity	0.724	2.0
1e19 (A,B)	Transferase: Carbamate kinase activity	1b7b (A,C)	Transferase: Carbamate kinase activity	0.911	2.6
1e2k (A,B)	Transferase: Thymidine Kinase activity	1p6x (A,B)	Transferase: Thymidine Kinase activity	0.886	2.4
1e6u (A1,A2)	Epimerase: GDP-L-fucose synthase activity	1udb (A1,A2)	Isomerase: UDP-glucose-4 epimerase activity	0.784	3.6
1e7w (A,B)	Dihydrofolase Reductase: oxidoreductase activity (pteridin)	1vl8 (A,B)	Dihydrofolase Reductase: oxidoreductase activity (torpin)	0.893	2.4
1e8i (A,B)	Hematopoeitic Cell Receptor: Receptor activity	1yxk (A,B)	Lipid Binding protein: Receptor activity	0.727	2.8
1e9g (A,B)	Hydrolase:	1ygz (A,B)	Hydrolase:	0.640	4.9

	Inorganic phosphatase activity		Inorganic phosphatase activity		
1edz (A1,A2)	Oxidoreductase: methylenetetrahydrofolate reductase activity	2be9 (A2,B2)	Transferase: Aspartate carbamoyltransferase activity	0.557	4.9
1eeq (A,B)	Immunoglobin: Antigen binding activity	1bww (A,B)	Immunoglobin: Antigen binding activity	0.932	1.4
1ehi (A,B)	Ligase:D-alanine D-alanine ligase activity	2fb9 (A1,A2)	Ligase:D-alanine D-alanine ligase activity	0.909	2.3
1c3i (A,B)	Hydrolase: Metaloendopeptidase activity	2j0t (A,D)	Hydrolase: collagenease and metalloprotease inhibitor activity	0.556	0.9
1c40 (A,B)	Hemoglobin: Oxygen transport	1ird (A,B)	Hemoglobin: Oxygen transport	0.978	1.0
1c47 (A,B)	Transferase: Metal ion binding with transferase activity	1srr (A1,A2)	Regulatory protein: Metal ion binding with tranferase activity	0.469	1.5
1c4z (A,D)	Ligase: ubiquitin ligase activity	2c2v (B,S)	Heat shock protein complex	0.582	1.5
1c6v (A,X)	DNA binding protein: RNA dependent DNA replication	1exq (A,B)	DNA binding protein: RNA dependent DNA replication	0.472	1.9
1c8k (A1,A2)	Transferase:Transferase activity	1et1 (A,B)	Hormone : parathyroid hormone receptor binding activity	0.434	8.3
1cc0 (A,E)	Signalling Protein: GTPase activity	1ds6 (A,B)	Signalling Protein: GTPase activity	0.916	2.2
1cgi (E,I)	Serine Protease inhibitor: Endopeptidase inhibitor activity	2f3c (E,I)	Serine Protease inhibitor: Endopeptidase inhibitor activity	0.940	1.5
1ci6 (A,B)	Transcription factor: Transcription activator activity	1zik (A,B)	Leucine Zipper: Transcription activator activity	0.950	0.6

1cjb (C,D)	Transferase: Metal ion binding with transferase activity	1izl (E,F)	Photosynthesis: Metal ion binding electron carrier activity	0.483	5.0
1cje (A,B)	Electron Transport: Metal ion binding electron carrier activity	1aar (A,B)	Ubiquitin: transcription regulator activity	0.482	5.0
1cki (A,B)	Phosphotransferase: Casein kinase activity	1fgk (A,B)	Phosphotransferase: Threonine kinase kinase activity	0.396	3.6
1clv (A,I)	Hydrolase: Alpha-amylase activity	1bvn (P,T)	Hydrolase: Alpha-amylase activity	0.846	2.3
1clx (A,B)	Xylanase: endo-1,4-xylanase activity	1ta3 (A,B)	Xylanase/Xylanase inhibitor: endo-1,4-xylanase activity	0.638	5.3
1cm5 (A,B)	Transferase:acyl transferase activity	1et1 (A,B)	Hormone : parathyroid hormone receptor binding activity	0.515	2.7
1d2g (A,B)	Transferase:methyl transferase activity	2ov2 (A,I)	Transferase: Toxin and serine threonine kinase	0.450	4.9
1d3y (A,B)	Isomerase: DNA topoisomerase activity	1l0l (A,K)	Oxidoreductase: Metal ion binding involved in transport	0.470	4.6
1d4x (A,G)	Contractile Protein: Actin bound to actin binding protein	2btf (A,P)	Contractile Protein: Actin bound to actin binding protein	0.723	2.5
1d5z (A,C)	Immunoglobulin: MHC class II receptor activity	1klu (A,D)	Immunoglobulin: MHC class II receptor activity	0.966	0.9
1d6f (A1,A2)	Transferase: Naringenin-chalcone synthase activity	1u0u (A,B)	Transferase: Dihydropinosylvin synthase activity	0.989	0.8
1d7f (A,B)	Transferase: Calcium Ion	1clv (A,I)	Hydrolase: Alpha-amylase	0.715	3.5

	binding		activity with calcium binding		
1d7m (A,B)	Contractile Protein: Actin filament binding coiled coil	1t6f (A,B)	Cell cyle protein: Protein binding coiled coil	0.959	0.7
1d9k (A,B)	Immunoglobulin: T-cell receptor	1ac6 (A,B)	Immunoglobulin: T-cell receptor	0.875	2.0
1db2 (A,B)	Hydrolase inhibitor: Endopeptidase inhibitor activity involved in blood coagulation	1oc0 (A,B)	Serine Protease inhibitor: Endopeptidase inhibitor activity involved in blood coagulation	0.888	1.2
1dba (L,H)	Immunoglobulin : Fab Light and Heavy chains	1d5i (L,H)	Immunoglobulin : Fab Light and Heavy chains	0.965	1.6
1dbq (A,B)	DNA binding protein: transcription repressor activity	2fep (A1,A2)	DNA binding protein: Transcription regulator activity	0.831	3.3
1dc6 (A,B)	Oxidoreductase: Glyceraldehyde-3-phosphate dehydrogenase	1b7g (O,Q)	Oxidoreductase: Glyceraldehyde-3-phosphate dehydrogenase	0.748	3.7
1dcf (A1,A2)	Response Regulator: 2-component response regulator activity	1eay (A,C)	Response Regulator: 2-component response regulator activity	0.543	3.4
1ddz (A,B)	Lyase: Carbonic dehydratase activity	1ym3 (A1,A2)	Lyase: Carbonic dehydratase activity	0.496	5.6
1dfj (E,I)	Complex of ribonuclease with ribonuclease inhibitor	1z7x (Z,Y)	Complex of ribonuclease with ribonuclease inhibitor	0.953	1.7
1dfk (A,Z)	Contractile Protein: Myosin head with motor activity	1w7j (A1,B1)	Contractile Protein: Myosin head with motor activity	0.840	3.2
1dhf (A,B)	Oxidoreductase: Dehydrofolate reductase activity	2pln (A1,A2)	Signalling protein: Two component	0.470	4.5

	(human)		regulator activity		
1dk4 (A,B)	Hydrolase: inositol phosphatase activity	1vdw (A,B)	Hydrolase: inositol phosphatase activity	0.916	2.3
1dle (A,B)	Hydrolase: Serine type endopeptidase activity	1h9h (E,I)	Hydrolase/Hydrolase inhibitor : Trypsin like serine protease activity	0.757	3.1
1dml (A,B)	DNA binding protein: DNA dependent DNA polymerase activity	1t6l (A1,A2)	DNA binding protein: DNA dependent DNA polymerase processivity factor activity	0.421	3.2
1dok (A,B)	Chemokine: Signal transducer activity	1eqt (A,B)	Chemokine: Signal transducer activity	0.811	2.3
1dov (A1,A2)	Cell Adhesion: Cadherin binding	2bf9 (A1,A2)	Pancreatic hormone	0.742	2.2
1dpg (A,B)	Oxidoreductase: Glucose-6-phosphate dehydrogenase	1dz3 (A1,A2)	Response Regulator: 2-component response regulator activity	0.459	4.9
1dqz (A,B)	Immunoprotein: Acyltransferase activity	2fe1 (A1,A2)	Hypothetical protein from Pyrobaculum aerophilum	0.494	5.3
1dsu (A,B)	Serine Protease: Serine-type endopeptidase activity	1h9h (E,I)	Trypsin like Serine Protease	0.826	1.5
1dxg (A,B)	Non heme iron protein: Iron ion binding activity	1dfn (A,B)	Defensin: Defense response to bacteria	0.312	4.3
1e50 (C,D)	Transcription factor: Transcription activator activity	1h9d (A,B)	Transcription factor: Transcription activator activity	0.999	0.2
1e51 (A0,B0)	Dehydratase:Porphyrin biosynthetic	2dqw (A,B)	Transferase: Folic acid and derivative	0.579	4.7

	process		biosynthetic process		
1e8u (A,B)	Sialidase: Host cell surface receptor binding	1ofz (A,B)	Lectin: Sugar Binding	0.613	6.4
1ecj (A,B)	Transferase: amidophosphoribosyl transferase activity	2dy0 (A,B)	Transferase: phosphoribosyl transferase activity	0.620	4.8
1edh (A,B)	Cell Adhesioin: Calcium ion binding	1ncg (A1,A2)	Cadherin: Calcium ion binding	0.601	3.9
1epa (A,B)	Retinoic acid binding protein: Transporter activity	2akq (A,B)	Transport protein: Retinol binding	0.671	4.5
1eqw (A,C)	Oxidoreductase: Superoxide dismutase activity	2aqp (A,B)	Oxidoreductase: Superoxide dismutase activity	0.866	1.5
1es0 (A,B)	Immunpprotein: MHC class II antigen processing activity	1jl4 (A,B)	Immunpprotein: MHC class II antigen processing activity	0.955	1.0
1ete (A,B)	Cytokine: Positive regulation of cell proliferation	2o27 (A,B)	Cytokine: Stem cell factor receptor binding	0.712	3.6
1eui (A,C)	Hydrolase/Hydrolase inhibitor: Uracil DNA-N glycosylase activity	1uug (A,B)	Hydrolase/Hydrolase inhibitor: Uracil DNA-N glycosylase activity	0.942	0.8
1euv (A,B)	Hydrolase: SUMO specific peptidase activity	2ckh (A,B)	Hydrolase: SUMO specific peptidase activity	0.902	1.7
1a22 (A,B)	Hormone/Hormone Receptor: Growth hormone bound to growth hormone receptor	1bp3 (A,B)	Hormone/Hormone Receptor: Growth hormone bound to growth hormone receptor	0.862	2.5
1a2d (A,B)	Fatty Acid Binding Protein:	1ftp (A,B)	Fatty Acid Binding Protein:	0.462	2.1

	Transporter activity		Transporter activity		
1a4y (A,B)	Inhibitor/Nuclease: Ribonuclease activity protein bound to inhibotor	2bex (A,C)	Inhibitor/Nuclease: Ribonuclease activity protein bound to inhibotor	0.913	2.5
1a50 (A2,B2)	Lyase: tryptophan synthase activity	1k8y (A1,B1)	Lyase: tryptophan synthase activity	0.994	0.7
1a6d (A1,A5)	Chaperonin: Unfolded protein binding	1we3 (G,U)	Chaperone: Unfolded protein binding	0.627	3.7
1aih (A,B)	DNA integration: DNA recombination integrationa and transposition	1utx (A,B)	DNA binding protein: Sequence specific DNA binding	0.435	5.2
1aor (A,B)	Oxidoreductase: aldehyde-ferrodoxin reductase activity and electron carrier activity	1sph (A,B)	Phosphotransferase: Kinase activity	0.443	5.0
1aoz (A,B)	Oxidoreductase: L-ascorbate oxidase activity	1a25 (A,B)	Calcium Binding Protein: Protein kinase C activity	0.451	5.3
1az3 (A,B)	Endonuclease: Magnesium ion binding with endonuclease activity	1k0z (A,B)	Hydrolase: Magnesium ion binding with endonuclease activity	0.468	5.8
1b0n (A,B)	DNA binding protein: Sequence specific DNA binding with DNA sporulation activity	1y7y (A,B)	DNA binding protein: Sequence specific DNA binding	0.425	1.8
1b34 (A,B)	RNA binding protein:splicesomal snRNP biogenesis and assembly	1igq (A,C)	Transcription factor: Transcription activator activity	0.428	3.8
1b4f (G2,H2)	Signal protein:	1b4f (A1,B1)	Signal protein:	0.972	0.4

	Transmenbrane ephrin receptor with protein tyrosine kinase activity		Transmenbrane ephrin receptor with protein tyrosine kinase activity		
1b73 (A1,A2)	Isomerase: Glutamate racemase activity	1r4a (B,F)	Transport protein: GTP binding involved in vescicle mediated transport	0.436	5.9
1b98 (A,M)	Hormone/Growth factor: growth factor involved in regulation of synaptic plasticiy	1bnd (A2,B2)	Hormone/Growth factor: growth factor involved in positive regulation of glial cell differentiation	0.877	1.8
1b9x (A,B)	Signalling protein: signal transducer activity	1tbg (B,F)	Signalling protein: signal transducer activity	0.973	1.3
1bd2 (A,B)	MHC Class I protein Complex: Antigen processing	1qvo (A1,B1)	MHC Class I protein Complex: Antigen processing	0.986	0.8
1bi7 (A1,A2)	Kinase: Cyclin dependent protein kinase activity	2a1a (A1,B1)	Translation initiation factor bound to protein serine/threonine kinase	0.689	1.6
1bi8 (C,D)	Kinase: Cyclin dependent protein kinase bound to kinase inhibitor	1bi7 (A1,B1)	Kinase: Cyclin dependent protein kinase bound to multiple tumor supressor	0.957	1.3
1bjf (A,B)	Calcium Binding Protein: Clathrin, tubulin and actin binding	2bn1 (B1,B2)	Radiation Damage protein: Insulin receptor binding	0.442	5.0
1bml (A,C)	Blood Clotting: peptidase activity in blood coagulatio	1gl1 (C,K)	Peptidase/Inhibitor complex: Peptidase activity in digestion and	0.842	2.3

			proteolysis		
1bog (A1,B1)	Antibody-peptide complex: Antigen binding	1ggb (L,H)	Immunogloulin: Light and Heavy chains	0.962	1.4
1bp6 (A1,A2)	Transferase: thymidilate synthase activity	1sqv (A2,K2)	Oxidoreductase: Metalloendopeptidase activity	0.491	1.2
1bqu (A,B)	Signalling protein: Interleukin-6 receptor activity	1pvh (A,B)	Signalling protein/cytokine complex: Interleukin-6 receptor activity	0.621	3.8
1btg (B,C)	Growth factor: Growth factor activity	1bnd (A2,B2)	Hormone/Growth factor: growth factor involved in positive regulation of glial cell differentiation	0.937	1.4
1bth (H,P)	Serine Protease/inhibitor : Endopeptidase activity involved in blood coagulation with inhibitor	1eaw (C,D)	Serine Protease/inhibitor : Endopeptidase activity involved in blood coagulation with inhibitor	0.957	1.3
1buo (A1,A2)	Gene Regulation: transcription repressor activity involved in myeloid cell regulation	2if5 (A1,A2)	Gene Regulation: transcription repressor activity involved in myeloid cell regulation	0.957	1.4
1bww (A,B)	Immunoglobulin: Antigen binding	1eeq (A,B)	Immunoglobulin: Antigen binding	0.949	1.4
1byl (A1,A2)	Antibiotic resistant protein: drug binding	1qto (A1,A2)	Antibiotic resistant protein: drug binding	0.960	1.1
1byr (A1,A2)	Endonuclease: endonuclease activity	2ppx (A2,A4)	Strutural Protein: Sequence specific DNA binding	0.489	3.4
1u0s (Y,A)	Response Regulator: 2-component response	1eay (A,C)	Response Regulator: 2-component response	0.598	3.6

	regulator activity		regulator activity		
1u3h (C,D)	Immunoprotein : MHC class I protein binding	1jl4 (A,B)	Immunoprotein : MHC class I protein binding	0.993	0.6
1u5t (A,B)	Transport Protein: G-protein signalling regulator and telomere maintainenance	1u5t (A,C)	Transport Protein: telomere maintainenance	0.548	1.1
1u7u (A1,A2)	Ligase:phosphopantothenate--cysteine ligase activity	3rap (R,S)	Signalling Protein: GTP binding	0.573	5.1
1uad (A,C)	Endocytosis/exocytosis: Ras protein signal transduction and in exocytosis	1k8r (A,B)	Transport Protein: Ras related protein involved in signal tranduction and tranport	0.780	3.4
1ub9 (A1,A2)	Transcription factor: Transcription activator activity	2bf9 (A1,A2)	DNA binding Protein: Termination of DNA replication	0.545	3.0
1uc4 (A,G)	Lyase: Propanediol dehydratase activity	1tyg (C2,G2)	Biosynthetic Protein: Thymine biosynthesis activity	0.515	4.7
1ugh (E,I)	Glycosylase: uracil DNA-N glycosulase activity involved in bsae excision repair	1uug (A,B)	Glycosylase: uracil DNA-N glycosulase activity involved in bsae excision repair	0.971	1.0
1ul1 (X,A)	Hydrolase/DNA binding protein: DNA repair and DNA polymerase processivity factor activity	1ul1 (Z,C)	Hydrolase/DNA binding protein: DNA repair and DNA polymerase processivity factor activity	0.598	4.9
1ulz (A1,A2)	Ligase: Biotin binding with carbamoyl synthase activity	1dv1 (A,B)	Ligase: Biotin binding with carbamoyl synthase activity	0.964	1.9

1uuf (A1,A2)	Oxidoreductase: Metal ion binding oxidoreductase activity	1q1n (A1,A2)	Oxidoreductase: Metal ion binding oxidoreductase activity	0.838	3.3
1uw4 (A,B)	RNA binding protein: nonsense mediated decay	1a7g (E1,E2)	Transcription factor: Transcription activator activity	0.507	4.5
1uyt (A1,A2)	Transeferase: biotin carboxylase activity	1zm7 (A,B)	Transcription factor: phosphotransferase activity	0.477	5.5
1a3a (A,C)	Phosphotransferase: Sugar-hydrogen symporter activity	1aar (A,B)	Ubiquitin: transcription regulator activity	0.491	4.1
1a4r (A,B)	Hydrolase: Establishment and maintainence of cell polarity via GTP dependent protein binding	2ov2 (A,I)	Transferase: GTP binding with involvement in actin cytoskeleton assembly	0.816	1.4
1a4u (A,B)	Oxidoreductase: Alcohol dehydrogenase activity	1k2w (A,B)	Oxidoreductase: L-iditol 2 dehydrogenase activity	0.820	3.2
1a6z (A,B)	MHC Class I protein Complex: Antigen processing	1qo3 (A,B)	MHC Class I protein Complex: Antigen processing	0.920	1.9
1ac6 (A,B)	Receptor: T-cell receptor	1d9k (A,B)	Receptor: T-cell receptor	0.879	2.0
1acb (E,I)	Hydrolase: Serine type endopeptidase activity	1gl1 (C,K)	Hydrolase: Serine type endopeptidase activity	0.932	1.6
1ad1 (A,B)	Transferase: Dihydopteorate synthase activity	2dqw (A,B)	Transferase: Dihydopteorate synthase activity	0.864	2.3
1ad3 (A,B)	Oxidoreductase: aldehyde dehydrogenase	1ez0 (A,D)	Oxidoreductase: aldehyde dehydrogenase	0.751	4.0

	activity		activity		
1ade (A,B)	Ligase: Adenylosuccinate synthase activity	1loo (A1,A2)	Ligase: Adenylosuccinate synthase activity	0.944	2.0
1adj (A,B)	Histidyl tRNA synthase:aminoacyl tRNA synthase activity	1htt (A,B)	Histidyl tRNA synthase:aminoacyl tRNA synthase activity	0.933	2.4
1ae1 (A,B)	Oxidoreductase: tropine dehydrogenase activity	2ae2 (A,B)	Oxidoreductase: tropine dehydrogenase activity	0.960	0.8
1afs (A,B)	Oxidoreducatse: 3-alpha-hydroxysteroid dehydrogenase activity	1exb (A1,E1)	Oxidoreducatse: 3-alpha-hydroxysteroid dehydrogenase activity	0.611	2.8
1agr (A,E)	Signal Transduction : GTP bound signal transduction activity	2ihb (A,B)	Signal Transduction : GTP bound signal transduction activity	0.991	0.8
1aiz (A,B)	Electron transport: Cadmium binding protein	1nwp (A,B)	Electron transport: Cadmium binding protein	0.704	4.1
1all (A,B)	Light harvesting protein: Protein chromophore linkage	2j96 (A,B)	Light harvesting protein: Protein chromophore linkage	0.901	2.0
1aoh (A,B)	Cellulosome subunit: Hydrolyzing O-glycosyl compounds	1ohz (A1,B1)	Cellulosome Scaffolding proteinwith dockerin complex having O-glycosyl hydrolyzing activity	0.776	2.6
1ap2 (A,B)	Immunoglobulin: Antibody variable domain like	1j05 (L,H)	Immunoglobulin: Antibody variable domain like	0.964	1.1
1aqu (A,B)	Transferase:	2f1r (A,B)	Biosynthetic	0.492	5.1

	estrone sulfotransferase activity		Protein: Mo-Molybdopterine biosynthesis activity		
1ati (A,B)	Protein biosynthesis: glycine tRNA ligase activity	2g4c (A,C)	Transferase: glycine tRNA ligase activity	0.855	3.1
1aui (A,B)	Hydrolase: serine/threonine phpsphatase activity	2o8g (A,I)	Hydrolase: serine/threonine phpsphatase activity	0.783	1.8
1avg (H,I)	Blood Clotting: peptidase activity in blood coagulation	1h9h (E,I)	Hydrolase/Hydrolase inhibitor : Trypsin like serine protease activity	0.841	1.5
1aw2 (A,B)	Isomerase: triose phosphate isomerase activity	1tre (A,B)	Isomerase: triose phosphate isomerase activity	0.972	1.2
1ay1 (L,H)	Immunoglobulin: Antibody variable domain like	1mf2 (M,N)	Immunoglobulin: Antibody variable domain like	0.943	1.9
1ay7 (A,B)	Enzyme/Inhibitor complex: Endoribonuclease activity with Barstar	1b2s (A,D)	Enzyme/Inhibitor complex: Endoribonuclease activity with Barstar	0.765	2.0
1azt (A,B)	Hydrolase: GTP binding with signal transducer activity	1fqj (A,C)	Hydrolase: GTP binding with signal transducer activity	0.843	1.7
1azy (A,B)	Glycosyltransferase: thymidine phosphorylase activity	2dsj (A,B)	Glycosyltransferase: thymidine phosphorylase activity	0.943	2.6
1b43 (A,B)	Transferase: magnesium binding protein involved in DNA repair	2izo (A,C)	Hydrolase: magnesium binding protein involved in DNA repair	0.496	1.7
1b49 (A,C)	Methyltransferas	2ftn (A1,A2)	Methyltransferas	0.746	3.0

	e: Thymidylate synthase activity		e: Thymidylate synthase activity		
1b5q (A,B)	Oxidoreductase: Polyamine oxidase activity	2v1d (A,B)	Oxidoreductase: Electron carrier activity	0.519	3.8
1b67 (A,B)	DNA binding protein: Sequence specific DNA binding	1ku5 (A,B)	DNA binding protein: Sequence specific DNA binding	0.945	1.0
1b6d (A,B)	Immunoglobulin: Antibody variable domain like	1bww (A,B)	Immunoglobulin: Antibody variable domain like	0.966	0.7
1b6s (A,B)	Lyase: phosphoribosylaminoimidazole carboxylase activity	2dwc (A,B)	Lyase: phosphoribosylglycinamide formyltransferase 2 activity	0.760	3.0
1b78 (A,B)	Nucleoside triphosphatase activity	2car (A,B)	Nucleoside triphosphatase activity	0.820	3.0
1b7b (A,C)	Transferase: Carbamate kinase activity	1e19 (A,B)	Transferase: Carbamate kinase activity	0.895	2.6
1b8a (A,B)	Liagase: Carbamate kinase activity	1wyd (A,B)	Liagase: Carbamate kinase activity	0.966	1.9
1b8m (A,B)	Growth factor/neurotrophin: cytokine activity	1bnd (A2,B2)	Growth factor/neurotrophin: cytokine activity	0.934	1.5
1b8z (A,B)	DNA binding protein: Mitotic chromosome condensation	2o97 (A1,B1)	DNA binding protein: Mitotic chromosome condensation	0.913	0.9
1bc2 (A,B)	Hydrolase: beta lactamase activity	1vgn (A,B)	Hydrolase: beta lactamase activity	0.513	4.1
1bcc (A,B)	Oxidoreductase: Metalloendopeptidase activity	1ezv (A1,B1)	Oxidoreductase: Metalloendopeptidase activity	0.929	2.5
1bdm (A,B)	Oxidoreductase: Malate dehydrogenase activity	1b8p (A1,A2)	Oxidoreductase: Malate dehydrogenase activity	0.966	0.9

1bdy (A,B)	Calcium Binding Protein: Protein kinase C activity	1edm (B,C)	Coagulation factor: Calcium binding activity	0.458	4.1
1bfo (A,B)	Antibody: Constant domain like	1dvf (A,B)	Antibody: Constant domain like	0.960	1.2
1bh5 (A,B)	Lyase: Lactoglutathione lyase activity	1f9z (A,B)	Lyase: Lactoglutathione lyase activity	0.866	2.6
1bht (A,B)	Heparin binding protein: serine type endopeptidase activity involved in epithelial to mesonchymal activity	1gmo (G,H)	Heparin binding protein: serine type endopeptidase activity involved in epithelial to mesonchymal activity	0.995	0.5
1bk5 (A,B)	Transpor protein: protein import into nucleus	2c1t (A,C)	Transpor protein: protein import into nucleus	0.924	1.6
1bkn (A,B)	DNA repair: Mismatched DNA binding	1hss (A,B)	Cereal Inhibitor: Serine type endopeptidase inhibitor	0.453	4.9
1blx (A,B)	Kinase: Cyclin dependent protein kinase bound to kinase inhibitor	1bi8 (A,B)	Kinase: Cyclin dependent protein kinase bound to kinase inhibitor	0.946	1.3
1bo1 (A,B)	Transferase:phosphatidylinositol phosphate kinase activity	2gk9 (A,D)	Transferase:phosphatidylinositol phosphate kinase activity	0.959	1.7
1bp3 (A,B)	Hormone/growth factor: involved in ematopoietin/interferon-class (D200-domain) cytokine receptor activity	1a22 (A,B)	Hormone/growth factor: involved in ematopoietin/interferon-class (D200-domain) cytokine receptor activity	0.885	2.5
1bqq (M,T)	Hydrolase/Inhibitor: Metalloendopeptidase activity	2e2d (A,C)	Hydrolase/Inhibitor: Metalloendopeptidase activity	0.904	2.2

1br1 (A,B)	Muscle protein: Motor activity	1w7j (A1,B1)	Muscle protein: Motor activity	0.688	3.2
1brc (E,I)	Protease/Inhibitor complex: serine type endopeptidase activity and inhibitor	1taw (A1,B1)	Protease/Inhibitor complex: serine type endopeptidase activity and inhibitor	0.981	0.9
1bsl (A,B)	Flavoprotein: monooxygenase activity in biolumination	1luc (A,B)	Flavoprotein: monooxygenase activity in biolumination	0.956	1.8
1bt6 (A,B)	Ligand (S100)/annexin complex: Calcium ion binding protein involved signal tranduction	1k96 (A1,A2)	S100 protein: Calcium ion binding protein involved xenobiotic metabolic process	0.891	1.6
1btk (A,B)	Transferase: protein tyrosine kinase activity	1r4a (B,F)	Transport Protein: GTP binding	0.513	3.5
1bvn (P,T)	Hydrolase/inhibitor: alpha amylase activity and its inhibitor	1clv (A,I)	Hydrolase/inhibitor: alpha amylase activity and its inhibitor	0.953	2.3
1bvr (A,B)	Oxidoreductase: enoyl reductase activity	1eny (A1,A2)	Oxidoreductase: enoyl reductase activity	0.992	0.8
1u0u (A,B)	Transferase: Acyl transferase activity	1d6f (A1,A2)	Transferase: Acyl transferase activity	0.989	0.8
1u20 (A,B)	Hydrolase:snoRNA binding protein	2bn1 (B1,B2)	Hydrolase:hydroloase activity	0.522	3.4
1u2w (A,B)	DNA binding protein: DNA dependent regulation of transcription	1r1u (A,B)	DNA binding protein: DNA dependent regulation of transcription	0.871	2.0
1u41 (A,B)	DNA binding protein (NF kappa B mutant): DNA dependent regulation of	1my7 (A,B)	DNA binding protein (NF kappa B mutant): DNA dependent regulation of	0.928	0.7

	transcription		transcription		
1u5k (A,B)	Response to DNA damage stimulus	2v1c (A1,C1)	Response to DNA damage stimulus	0.597	3.3
1u5w (A,B)	Hypthetical protein	1msc (A1,A2)	Virial protein: RNA binding	0.470	5.8
1u60 (A,B)	Hydrolase: glutaminase activity	1hss (A,B)	Cereal Inhibitor: Serine type endopeptidase inhibitor	0.486	5.5
1u6e (A,B)	Transferase: Acyl transferase activity	1hnj (A1,A2)	Transferase: Acyl transferase activity	0.968	1.8
1u73 (A,B)	Hydrolase: Phospholipase A2 activity	2bf9 (A1,A2)	Pancreatic hormone	0.484	3.9
1u75 (A,C)	Oxidoreductase: Cytochrome C - peroxidase activity	2vnz (X1,X2)	Oxidoreductase: L-ascorbate peroxidase activity	0.488	2.7
1u8s (A,B)	Transcription factor: glycine cleavage system transcriptional repressor	1usm (A1,A2)	Transcriptional stimulator: lyase activity	0.576	3.4
1uc8 (A,B)	Biosynthetic protein: lysine biosynthesis	1i7n (A,B)	Neuropeptide:Neurotranmitter secretion	0.681	3.4
1udi (E,I)	Hydrolase/inhibitor: uracil DNA-N gycosylase activity	1ugh (E,I)	Hydrolase/inhibitor: uracil DNA-N gycosylase activity	0.967	1.0
1udv (A,B)	DNA binding protein: Double stranded DNA binding	1h0x (A,B)	DNA binding protein: Double stranded DNA binding	0.873	1.4
1ueh (A,B)	Transferase: di-trans,poly-cis-decaprenylcistransferase activity involved in peptidoglycaan biosynthesis process	1f75 (A,B)	Transferase: di-trans,poly-cis-decaprenylcistransferase activity involved in peptidoglycaan biosynthesis process	0.930	2.1

1un8 (A,B)	Kinase: Glycerone kinase activity	1oi2 (A,B)	Kinase: Glycerone kinase activity	0.839	2.8
1unk (A,C)	Immunoprotein: Toxin binding	2guz (A,B)	Protein Transport: protein transporter activity	0.473	4.1
1unl (A,D)	Cyclin Dependent Kinase 5:protein serine/threonine kinase activator activity	1unh (B,E)	Cyclin Dependent Kinase 5:protein serine/threonine kinase activator activity	0.986	0.9
1usl (A,B)	Isomerase: ribose-5-phosphate isomerase activity	1nn4 (A,D)	Isomerase: ribose-5-phosphate isomerase activity	0.883	1.5
1uth (A,B)	Transcription regulator: DNA dependent regulation of trnascription	2fyi (A,B)	Transcription regulator: DNA dependent regulation of trnascription	0.769	3.9
1utx (A,B)	DNA binding protein: Sequence specific DNA binding	1y7y (A,B)	DNA binding protein: Sequence specific DNA binding	0.705	2.1
1uty (A,B)	Viral Protein: RNA binding	1moy (A1,A3)	Biotin binding protein	0.393	4.9
1uu0 (A,B)	Transferase:histidinol-phosphate transaminase activity	1lc5 (A1,A2)	Transferase: N-succinyldiaminopimelate aminotransferase activity	0.805	3.0
1uuz (A,D)	Lyozyme/Inhibitor complex	1gpq (B,C)	Lyozyme/Inhibitor complex	0.876	2.1
1uz6 (E,F)	Antigen-Antibody: Antbody variable domain like	1rvf (B,C)	Antigen-Antibody: Antbody variable domain like	0.914	2.1
1amh (A,B)	Hydrolase: serine-type endopeptidase	1h9h (E,I)	Hydrolase: serine-type endopeptidase	0.919	2.2

	activity		activity		
1avw (A,B)	Trypsin/Trypsin inhibitor	1h9h (E,I)	Trypsin/Trypsin inhibitor	0.933	1.8
1uer (A,B)	Oxidoreductase: Superoxide dismutase activity	2nyb (A,B)	Oxidoreductase: Superoxide dismutase activity	0.967	1.1
1ugs (A,B)	Lyase:nitrile hydratase activity	1ahj (A,B)	Lyase:nitrile hydratase activity	0.900	2.2

[a] PDB and Chain ID of the proteins in Benchmark 1.

[b] Classfication of the complexes as annotated in the PDB library (1).

[c] Gene Ontology terms (2).

[d] PDB and Chain ID of the proteins in Benchmark 2 which is the best match to the complexes in the first column as identified by DM-align.

[e] TM-score of the complex structures in Column 1 and Column 3.

[f] RMSD of complex structures in the aligned region.

Appendix II

Table 1. Mean distance (Å) between C_α atoms of amino acids that are in inter-chain contact.

	G	**A**	**V**	**L**	**I**	**S**	**T**	**C**	**M**	**P**	**D**	**N**	**E**	**Q**	**K**	**R**	**H**	**F**	**Y**	**W**
G	3.3	5.5	5.6	6.1	6.0	5.6	5.6	5.3	5.7	5.6	5.8	5.6	5.6	5.9	6.5	6.5	5.8	6.2	6.4	7.0
A	5.5	3.8	6.1	6.5	6.6	5.3	5.7	5.4	5.9	6.1	5.8	6.1	6.5	5.7	6.3	6.8	5.8	6.6	6.3	6.8
V	5.5	5.4	4.1	7.0	6.3	5.9	6.0	5.5	6.3	5.8	5.9	6.2	6.3	6.0	6.5	6.5	5.9	7.1	6.7	7.1
L	5.6	5.9	6.8	4.3	6.9	5.5	6.3	5.8	6.9	6.5	5.9	6.6	6.1	6.2	6.5	7.3	6.1	7.4	7.1	7.7
I	5.5	5.9	6.4	6.9	4.3	5.6	6.0	5.9	6.6	6.1	5.7	6.0	6.2	6.0	6.2	7.0	6.4	7.9	7.0	7.3
S	5.3	5.5	6.3	5.8	6.0	4.4	7.3	5.5	6.6	6.5	6.0	7.2	6.0	6.0	6.5	7.5	6.0	6.6	6.8	6.7
T	5.5	5.5	5.7	6.5	5.9	6.8	4.4	5.4	6.1	6.0	5.6	7.7	5.8	6.6	6.5	7.6	6.5	6.5	6.9	7.2
C	5.4	5.7	6.3	6.7	6.1	5.4	5.8	4.6	6.2	5.8	5.3	5.7	5.7	5.8	5.8	6.4	5.8	6.8	7.1	6.7
M	6.3	6.3	6.7	7.5	7.0	7.1	6.7	6.4	4.5	6.4	6.5	6.9	7.2	7.3	8.8	7.9	6.7	7.5	7.2	9.0
P	5.2	5.4	5.9	6.7	5.9	6.3	5.9	5.4	6.2	4.9	5.5	6.1	6.0	6.2	6.1	8.1	5.9	6.3	7.2	7.1
D	5.6	5.9	6.5	6.5	6.8	6.0	6.3	5.5	7.1	5.8	4.9	6.9	7.1	6.6	7.7	8.0	6.7	6.4	7.5	7.1
N	5.5	5.8	6.3	7.1	6.1	6.8	7.8	5.5	6.2	5.9	6.3	4.3	6.5	6.4	7.6	8.3	6.4	6.6	7.3	6.7
E	6.1	6.0	6.6	6.7	6.9	6.6	6.2	5.1	6.5	6.2	6.6	6.7	4.5	6.5	7.7	8.0	6.9	7.1	7.4	8.6
Q	6.3	5.9	6.8	7.4	6.8	6.5	6.6	6.2	7.2	6.6	6.9	6.8	6.9	4.6	6.9	7.5	7.2	7.4	8.0	7.5
K	5.8	6.0	6.1	6.6	6.4	6.0	6.2	5.4	7.1	6.0	7.1	6.4	6.8	6.4	4.2	7.2	6.9	7.0	7.2	6.9
R	6.5	6.8	7.2	7.1	7.2	7.5	7.7	6.0	7.9	7.6	7.8	7.8	8.0	7.2	7.3	4.8	7.3	7.9	8.0	8.5
H	5.9	6.3	6.3	7.1	7.0	6.2	6.9	6.5	6.8	6.5	7.4	6.7	7.3	7.0	8.3	8.2	4.1	7.1	7.6	7.6
F	6.0	6.2	7.6	7.6	7.5	6.3	7.2	6.4	7.3	6.3	6.3	6.8	6.7	6.9	7.9	7.9	7.0	5.8	7.6	8.8
Y	6.3	6.7	7.0	7.2	7.4	6.8	7.0	6.3	7.4	7.1	7.2	7.2	7.7	7.3	7.6	7.8	7.3	7.7	5.0	8.3
W	7.1	7.9	8.6	9.0	7.9	7.8	9.2	6.4	8.5	7.5	8.6	7.6	8.7	6.8	9.4	9.7	7.9	8.9	7.9	5.8

Table 2. Standard deviation (Å) of C_α distance for amino acids that are in inter-chain contact.

	G	**A**	**V**	**L**	**I**	**S**	**T**	**C**	**M**	**P**	**D**	**N**	**E**	**Q**	**K**	**R**	**H**	**F**	**Y**	**W**
G	0.8	2.1	2.3	2.6	2.5	2.0	2.1	1.1	1.2	2.3	2.5	1.5	1.7	2.1	2.7	2.0	1.6	2.4	2.1	2.4
A	2.3	1.3	2.9	2.8	2.8	1.7	1.9	1.1	1.7	2.5	2.4	2.2	3.2	2.0	3.2	3.2	1.8	2.8	2.0	2.4
V	2.3	2.0	1.3	2.9	2.3	2.4	2.7	1.2	1.7	2.1	2.5	2.6	2.9	2.2	2.9	2.5	1.7	2.8	2.5	2.1
L	2.2	2.3	2.8	1.5	2.3	1.8	2.7	1.3	2.3	2.8	2.4	2.7	2.4	2.3	3.2	3.1	1.8	2.6	2.2	2.3
I	2.5	2.6	2.7	2.6	1.4	2.1	2.2	2.1	1.9	2.6	2.5	2.5	3.0	2.1	2.8	3.0	2.3	3.2	2.6	2.2
S	2.1	2.3	2.9	1.9	0.6	1.4	3.5	1.2	3.0	2.9	3.0	3.3	2.4	1.8	3.1	3.1	1.8	2.7	2.2	2.2
T	2.3	2.2	2.2	2.7	1.8	3.3	1.5	1.2	1.8	2.7	1.6	3.5	1.6	2.7	2.9	2.9	2.3	2.0	2.4	2.7
C	2.2	2.5	2.8	2.8	2.3	1.7	2.5	1.8	1.3	2.4	1.9	1.9	2.1	1.7	2.0	2.5	1.3	2.6	2.7	2.4
M	2.5	2.6	2.6	2.8	2.8	2.7	2.7	2.2	1.6	2.5	2.5	2.8	3.3	3.2	3.4	3.3	2.3	2.4	2.0	2.9
P	1.8	2.1	2.2	2.9	2.1	2.7	2.3	1.3	2.0	1.9	1.6	2.4	2.7	1.6	2.6	3.6	1.4	2.2	2.2	2.4
D	2.6	2.9	3.2	2.9	3.5	2.6	2.9	2.1	3.7	2.3	1.3	3.1	3.6	2.8	3.4	2.8	1.9	2.9	2.7	2.7
N	2.1	2.3	2.8	3.2	2.5	3.1	3.6	1.3	2.1	2.0	2.8	1.9	3.0	2.2	4.2	3.2	2.1	2.6	2.7	1.5
E	2.6	2.6	3.1	2.9	3.1	2.7	2.4	1.6	2.5	2.4	2.7	2.6	1.4	2.4	2.8	2.9	2.0	3.1	2.3	3.1
Q	2.9	2.6	3.1	3.2	2.9	2.7	2.8	2.3	3.1	2.9	3.0	2.7	3.1	1.2	3.1	2.5	2.7	2.9	2.9	2.9
K	2.2	2.7	2.7	2.8	3.0	2.1	2.3	1.4	3.3	2.4	3.1	2.5	2.6	2.0	1.6	3.1	2.8	3.0	2.5	2.1
R	2.4	2.9	3.1	2.7	2.8	2.8	3.1	1.7	3.4	2.9	2.9	2.9	2.9	2.6	3.2	1.9	2.6	3.2	2.4	3.0
H	2.3	2.7	2.1	2.8	2.6	2.3	2.8	2.2	2.3	2.6	2.9	2.4	2.6	2.7	4.1	3.1	1.1	2.5	2.2	1.8
F	2.1	2.2	3.1	2.6	2.6	2.1	3.0	1.4	2.0	2.0	2.4	2.3	2.7	2.1	3.7	3.0	2.2	1.6	2.3	2.4
Y	2.0	2.7	2.7	2.0	2.6	2.2	2.6	1.8	2.2	2.3	2.0	2.5	2.9	2.2	2.9	3.1	2.3	2.6	1.3	2.7
W	3.1	3.3	3.6	3.4	2.7	3.0	3.7	2.2	3.1	2.9	3.6	3.0	3.8	2.4	4.1	4.1	2.9	3.2	2.7	1.9

Appendix III

Comparison of I-RMSD (Å) of predicted models*

Name	ZDOCK-exp	ZDOCK-model	COTH	COTH-exp	COTH-model
1avxA-1avxB	2.16 (5)	4.43 (2)	4.44 (1)	5.00	4.81
1ay7A-1ay7B	11.77 (6)	13.87 (4)	8.96 (3)	11.64	12.53
1bvnP-1bvnT	1.97 (10)	3.64 (7)	4.53 (1)	6.11	6.64
1cgiE-1cgiI	9.63 (1)	12.57 (2)	3.86 (8)	4.30	4.78
1d6rA-1d6rI	8.04 (4)	11.09 (1)	13.54 (10)	14.86	15.30
1dfjE-1dfjI	2.09 (10)	2.22 (1)	7.48 (1)	7.87	7.74
1e6eA-1e6eB	2.33 (5)	3.83 (5)	1.52 (3)	2.42	3.14
1eawA-1eawB	3.21 (1)	3.87 (4)	7.86 (6)	8.89	7.18
1ewyA-1ewyC	2.26 (7)	4.51 (8)	6.37 (4)	8.72	9.50
1f34A-1f34B	13.62 (10)	16.74 (10)	11.25 (1)	13.75	14.58
1mahA-1mahF	1.35 (8)	2.34 (1)	2.43 (2)	3.55	3.54
1ophA-1ophB	6.20 (2)	6.62 (3)	7.57 (9)	9.40	10.01
1ppeE-1ppeI	1.29 (1)	3.40 (4)	5.08 (1)	6.66	6.85
1tmqA-1tmqB	12.33 (7)	13.67 (7)	14.29 (1)	14.13	14.18
1udiE-1udiI	3.86 (2)	7.86 (6)	2.36 (3)	3.68	3.92
2b42B-2b42A	1.85 (1)	4.60 (5)	1.04 (4)	3.07	3.68
2o8vA-2o8vB	11.05 (8)	14.89 (5)	4.09 (5)	4.74	4.86
2pccA-2pccB	9.44 (3)	12.85 (2)	7.31 (5)	8.89	9.42
2sicE-2sicI	1.69 (4)	2.94 (1)	3.42 (6)	2.48	3.50
2sniE-2sniI	6.42 (8)	7.08 (1)	3.08 (4)	4.35	4.91
2uuyA-2uuyB	2.49 (4)	4.19 (9)	3.48 (1)	3.40	3.70
7ceiA-7ceiB	2.22 (6)	5.74 (10)	5.93 (1)	7.76	8.38
1ak4A-1ak4D	6.77 (9)	9.52 (2)	10.60 (7)	11.55	9.97
1b6cA-1b6cB	2.77 (1)	4.35 (3)	2.72 (10)	3.79	4.15
1buhA-1buhB	13.97 (5)	14.44 (1)	7.06 (8)	8.30	8.72
1e96A-1e96B	2.72 (7)	5.71 (8)	9.45 (9)	11.90	12.72
1efnB-1efnA	8.29 (8)	9.61 (7)	2.17 (1)	1.94	2.32
1fc2C-1fc2D	5.36 (7)	8.94 (3)	3.50 (7)	5.87	6.67
1fqjA-1fqjB	14.71 (3)	17.26 (2)	19.40 (4)	18.52	18.82
1gcqB-1gcqC	9.68 (9)	12.44 (5)	5.02 (5)	6.10	6.31
1ghqA-1ghqB	11.5 (7)	13.25 (4)	6.90 (6)	6.35	6.53
1glaG-1glaF	2.50 (1)	6.38 (2)	8.68 (2)	8.64	8.65
1gpwA-1gpwB	2.03 (8)	2.56 (3)	4.46 (9)	6.42	7.08
1he1C-1he1A	8.03 (8)	10.89 (1)	4.35 (2)	7.21	8.17
1j2jA-1j2jB	3.93 (3)	5.41 (8)	8.83 (10)	11.40	12.26
1kacA-1kacB	2.18 (10)	2.51 (10)	3.43 (1)	4.00	4.81
1ktzA-1ktzB	9.46 (1)	13.19 (6)	6.58 (3)	8.92	9.69
1kxpA-1kxpD	3.27 (1)	4.69 (2)	3.08 (4)	4.98	5.29
1qa9A-1qa9B	12.65 (2)	14.72 (10)	8.18 (6)	10.72	11.57
1s1qA-1s1qB	13.32 (1)	16.62 (1)	19.03 (5)	20.58	21.09
1sbbA-1sbbB	9.73 (4)	9.78 (5)	3.53 (4)	2.70	2.98
1t6bX-1t6bY	6.48 (6)	8.84 (6)	4.68 (1)	5.86	6.89
1xd3A-1xd3B	4.06 (6)	4.78 (8)	4.81 (8)	6.38	6.90
1z0kA-1z0kB	2.06 (3)	2.37 (8)	8.39 (3)	9.13	9.38
1z5yD-1z5yE	8.31 (1)	10.98 (7)	1.36 (3)	2.41	3.67
1zhiA-1zhiB	9.56 (9)	12.31 (1)	13.22 (2)	15.75	16.59

2ajfA-2ajfE	11.59 (10)	12.36 (9)	14.89 (5)	17.82	18.79
2btfA-2btfP	13.03 (8)	13.65 (7)	5.76 (7)	7.92	8.65
2hleA-2hleB	2.34 (1)	3.25 (4)	4.91 (1)	6.99	7.35
2hqsA-2hqsH	12.22 (10)	13.65 (8)	3.93 (4)	4.65	4.92
2oobA-2oobB	5.25 (5)	5.35 (1)	7.94 (7)	10.19	10.95
2i25N-2i25L	8.57 (5)	11.62 (5)	3.94 (1)	5.44	5.95
1kxqH-1kxqA	2.02 (1)	4.96 (6)	2.97 (6)	2.60	3.20
1acbE-1acbI	4.65 (3)	4.85 (10)	4.26 (2)	5.00	5.75
1m10A-1m10B	17.47 (4)	18.09 (3)	10.47 (1)	12.07	12.61
1nw9B-1nw9A	8.43 (6)	9.53 (1)	3.82 (1)	3.56	4.47
1grnA-1grnB	16.78 (9)	17.05 (8)	17.59 (7)	18.95	19.41
1he8B-1he8A	32.26 (6)	35.75 (7)	26.22 (4)	27.65	28.12
1i2mA-1i2mB	13.50 (2)	15.32 (2)	9.75 (7)	11.99	12.74
1wq1R-1wq1G	8.12 (9)	10.55 (3)	13.32 (1)	15.04	15.61
1xqsA-1xqsC	9.16 (5)	9.42 (3)	11.27 (6)	12.47	10.47
2cfhA-2cfhC	8.79 (1)	12.08 (7)	2.22 (1)	1.56	2.78
2h7vA-2h7vC	12.04 (1)	14.83 (1)	5.18 (6)	6.66	7.16
2hrkA-2hrkB	10.27 (5)	12.46 (10)	13.86 (8)	16.38	17.21
2nz8A-2nz8B	5.57 (6)	7.03 (7)	14.57 (6)	15.04	16.16
1fq1A-1fq1B	13.56 (8)	13.89 (5)	14.22 (3)	15.75	16.26
1pxvA-1pxvC	13.87 (3)	17.42 (6)	5.59 (5)	5.75	6.81
1atnA-1atnD	17.54 (9)	17.60 (1)	17.81 (1)	20.51	21.40
1bkdR-1bkdS	15.58 (2)	16.06 (3)	10.73 (4)	13.00	13.75
1h1vA-1h1vG	19.06 (7)	20.12 (10)	23.10 (10)	23.08	23.08
1ibrA-1ibrB	9.00 (5)	9.29 (5)	4.72 (2)	4.81	5.00
1iraY-1iraX	21.93 (1)	25.07 (2)	17.91 (6)	17.60	17.70
1r8sA-1r8sE	7.57 (7)	10.99 (10)	14.43 (7)	15.95	16.11
1y64A-1y64B	19.62 (4)	21.33 (1)	14.09 (5)	16.84	17.76
2c0lA-2c0lB	9.81 (3)	9.83 (2)	5.70 (3)	7.37	7.92
2ot3B-2ot3A	4.67 (7)	5.55 (8)	4.03 (9)	4.50	4.97
1r0rE-1r0rI	7.68 (7)	9.59 (1)	12.63 (4)	13.83	14.23
Average	8.47	10.30	8.07	9.31	9.78

* The table shows a comparison of the DOCKING methods (ZDOCK-exp and ZDOCK-model) and COTH based methods (COTH, COTH-exp and COTH-model) in terms of I-RMSD. The values in the table indicate the I-RMSD of the best in top 10 (as ranked by the independent programs) models while the values in parentheses indicate the rank of the models. The ranks for COTH-exp and COTH-model are not indicated since they are the same as that for COTH.

Appendix IV

List of Publications:

- Emanuele Bultrini, Kevin Brick, Srayanta Mukherjee, Yang Zhang, Francesco Silvestrini, Pietro Alano, Elisabetta Pizzi. Revisiting the Plasmodium falciparum RIFIN family: from comparative genomics to 3D-model prediction. **BMC Genomics**, vol 10, 445 (2009).
- Srayanta Mukherjee, Yang Zhang. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. **Nucleic Acids Research**, vol 37, e83 (2009).
- Srayanta Mukherjee, Andras Szilagyi, Ambrish Roy, Yang Zhang. Genome-wide protein structure prediction. **Multiscale approaches to protein modeling: structure prediction, dynamics, thermodynamics and macromolecular assemblies**, Chapter 11, Edited by Andrzej Kolinski, (Springer-London, 2010), P. 255-280.
- Srayanta Mukherjee, Yang Zhang. Protein-protein complex structure prediction by multimeric threading and template recombination. **Structure**, vol 19, P. 955-966 (2011).
- Rajasree Menon, Ambrish Roy, Srayanta Mukherjee, Saveily Belkin, Yang Zhang and Gil Omenn. Functional implications of structural predictions for alternatively spliced proteins expressed in Her2/neu induced breast cancers. **Journal of Proteomics Research**. In press (2011). (Co-first author)
- Iris Dror, Shula Shazman, Srayanta Mukherjee, Yang Zhang, Fabian Glaser and Yael Mandel-Gutfreund. Predicting nucleic acid binding interfaces from structural models of proteins. **Proteins: Structure, Function and Bioinformatics.** In press (2011).
- Leonardo Garma, Srayanta Mukherjee and Yang Zhang. How many protein-protein interactions exist in nature? Submitted (2011). (Co-first author)
- Srayanta Mukherjee and Yang Zhang. TACOS: Automated structure prediction of protein-protein complex structures. Manuscript under preparation (2011).
- Srayanta Mukherjee and Yang Zhang. Refinement of protein complex structures predicted using docking by re-optimization of backbone structure using TACOS. Manuscript under preparation (2011).
- Ambrish Roy, Srayanta Mukherjee, P.S. Hefty and Yang Zhang. Remote homolog detection and function prediction using global and local structure similarity approach. Manuscript under preparation (2011).