BIOINFORMATICS TOOLS & ANALYSIS OF PROTEIN STRUCTURE AND FUNCTION

by

FEI JI

(Under the Direction of Ying Xu)

ABSTRACT

This dissertation mainly focuses on protein structure and functional studies from the viewpoint of Bioinformatics. My dissertation consists of three bioinformatics projects, which all utilized bioinformatics tools to understand the structure and function of proteins. The first project introduced an optimal strategy to identify optimal mutation sites in NMR experiments, and to predict trans-membrane proteins topology with minimum number of PRE sites. The second project is to develop a novel template based structure prediction tool using segmental structure instead of whole chain structure. The structural segments could help to identify the proteins of novel structures without proper templates. In the last project, I applied multiple bioinformatics tools to model the structure of a protein complex, ferredoxin hydrogenase in *Thermotoga maritima*. The model structure gives a new perspective on our understanding of the redox proteins and mechanism of H_2 production in anaerobic bacteria.

INDEX WORDS: Bioinformatics, Machine Learning, Protein Structure, Threading, Homology Modeling, Nuclear Magnetic Resonance, Hydrogenase

BIOINFORMATICS TOOLS & ANALYSIS OF PROTEIN STRUCTURE AND FUNCTION

by

FEI JI

BS, Fudan University, China, 2009

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOGOPHY

ATHENS, GEORGIA

2015

BIOINFORMATICS TOOLS & ANALYSIS OF PROTEIN STRUCTURE AND FUNCTION

by

FEI JI

Major Professor:	Ying Xu
Committee:	James Prestegard
	Liming Cai
	Natarajan Kannan

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2015

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

Proteins are the molecular devices where the biological functions are performed. The dynamic processes of life cycle of reproduction, metabolism and defense are all carried out by proteins. All protein functions are dependent on their structure, which, in turn, depends on physical and chemical parameters. This is other important fact on studying these molecules; classical biological, physical, chemical, mathematical and informatics sciences have been working together in a new area known as bioinformatics to allow a new level of knowledge about life organization [1].

Proteins have traditionally being studied individually. A protein of interest had its coding sequence identified and cloned in a proper expression vector. Hence, provided that cloning, expression and purification were successful, enough quantities of pure proteins could be employed in biochemical experiments or used to prepare solutions for NMR spectroscopy or to grow crystals for structure determination by X-ray crystallography. With the sequencing technology advances, protein genomic sequences and functional data are produced in high-throughput manner.

At present, over six million unique protein sequences have been deposited in the public databases, and this number is growing rapidly. Meanwhile, despite the progress of high-throughput structural genomics initiatives, just over 50,000 protein structures have so far been experimentally determined. This enormous disparity between the number of sequences and structures has driven research toward computational methods for

predicting protein structure from sequence. My dissertation consists of three bioinformatics projects, which all utilized computational tools to study the protein structures and related functions.

In Chapter 2, a strategy for predicting trans-membrane (TM) proteins with nuclear magnetic resonance paramagnetic relaxation enhancement (PRE) labels is presented. PRE measures long-range distance to isotopically labeled residues, providing useful distance constraints information in NMR for protein structure prediction. I focused on developing a computational strategy to determine TM proteins packing topology with minimizing the number of PRE labels on multiple positions. Tests on four helices DsbB experimental data using just one label correctly predicted the topology. Benchmark results using simulation data show that the correct topology for five and six helices can be predicted using minimum two labels, with an average success rate of 72%.

In Chapter 3, a new template based protein structure prediction method, SPRED, is introduced. Unlike the traditional method using full chain structure as structural templates, SPRED aligned to the segment structures that span several secondary structure units. SPRED has been tested on 317 non-homologous proteins from Protein Data Bank (PDB). The overall TM-scores by the PSRED alignments increase by 11.4% compared with those by the best whole-chain threading methods.

In Chapter 4, a novel mechanism of bacteria ferredoxin hydrogenase complex is proposed using various bioinformatics structure analysis tools. The trimeric ferredoxin hydrogenase is found to oxidize NADH and ferredoxin synergistically to produce hydrogen in *Thermotoga maritime*, which the molecular mechanism remains unknown.

This challenge is solved by protein complex structure modeling using state of art tertiary structure and protein docking tools. Modeled structure revealed an alternative interaction of trimeric hydrogenase in microorganisms, gives a new perspective on our understanding of the redox proteins and mechanism of H_2 production in anaerobic bacteria.

.

CHAPTER 2

MEMBRANE PROTEIN STRUCTURE AND OPTIMAL MUTATION SITES PREDICTION FOR PRE DATA[1]

[1] Huiling Chen, Fei Ji, Victor Olman, Charles K Mobley, Yizhou Liu, Yunpeng Zhou, John H Bushweller, James H Prestegard, Ying Xu. 2011. *Structure*. 19(4):484-495. Reprint here with permission of the publisher.

INTRODUCTION

Transmembrane (TM) proteins play central roles in cellular transport processes. They comprise ~60% of all drug targets [2-6]. In humans, ~27% of all proteins are TM helical proteins [7] but only 2.6% of the determined structures (2240 of 113331) in the Protein Data Bank (PDB) [8] are TM helical proteins up to date. The scarcity of the TM helical structures reflects the difficulty in determining such protein structures using techniques such as X-ray crystallography or nuclear magnetic resonance (NMR) [9-12]. The methods presented in this chapter attempt to facilitate solution NMR structure determination of membrane proteins by combining efficiently chosen small numbers of experimental constraints with rigorous computational structure prediction.

Solution NMR has only recently been used to determine the structures of polytopic helical membrane proteins. Successful examples of application, such as the structure determination of *Escherichia coli* proteins DsbB [13] and DAGK [14] determined using both solution and solid-state NMR methods[15-17]. Because of the nature of membrane proteins, a combination of multiple sources of data is generally required to solve TM helical protein structures. In the cases of both DsbB and DAGK, extensive paramagnetic relaxation enhancement distance constraints, residual dipolar coupling data, and long-range Nuclear Overhauser Effects (NOEs) were collected and used for solving the structures.

In the past few years, computational structure-prediction methods have improved to a point where predicted structures based on limited experimental data, possibly of low-resolution, become increasingly useful for studying protein functions and associated mechanisms [18-20]. Often, a low-resolution structure is useful enough as it can serve as

a starting point for more accurate structure determination using additional computational techniques. Barth et al. [21] showed that, when coarse-grained decoy structures with near-native topologies (<4 Å) were generated, de novo methods can predict high-resolution structures (<2.5 Å) for TM helical proteins with up to 145 residues. The major challenge in applying this approach for larger systems is in developing effective sampling procedures to consistently generate near-native topologies at a coarse-grained level. My focus in this chapter is to develop a computational strategy that identifies a minimal set of NMR data that will be adequate to determine the correct packing topology of TM helical proteins.

In this chapter, I present a computational method for selecting a minimal set of mutation sites in a given protein sequence for PRE data collection. The method is based on a theoretical analysis and it is validated through a computational study using a distance geometry-based algorithm. DsbB, a membrane protein with four TM helices is chosen as a test system; both a crystal structure and PRE data from nine cysteine sites are available for this protein [13, 22]. We demonstrate that it is possible to determine the correct packing topology by using PRE data collected on one specific cysteine-mutation site or any two cysteine-mutation sites within the protein if they are at the ends of helices and on the same side of the membrane. Using simulated PRE data, we extend the study to 10 proteins ranging from four to seven TM helices and with diverse topologies. The correct topology can be determined reliably for proteins with up to seven helices using PRE data collected on two or three sites, predicted by our program. These results show promises in predicting a minimal set of mutational sites needed for PRE data collection;

this in turn can guide experimental design and improve efficiency of membrane protein structure determination.

METHODS

PRE constraints

Paramagnetic relaxation enhancement (PRE) can provide long range distance constraints (15-25 Å) between a paramagnetic center and an NMR active nucleus such as a proton attached to a ^{15}N or ^{13}C enriched site [23-25]. Application of these constraints began with proteins that have native paramagnetic metal centers, but application has recently expanded with the use of cysteine mutagenesis and site-directed spin-labeling (SDSL) of cysteine sites with nitroxide labels [26]. Whereas direct interaction between NMR active nuclei (NOEs) provides distance information that rarely goes beyond 5-6 Å, the much larger interaction energy between an electron and a nucleus makes PRE effective at significantly longer distances. For example, perturbation of proton spin relaxation rates by a nitroxide spin label can yield distance constraints of 15-25 Å with accuracy approaching ±15%. Thus PRE can be particularly helpful in determining the global fold of perdeuterated polypotic TM helical proteins.

Distance matrix

Our system consists of m amino acids and n PRE labels, a total of $m+n$ points, hence the distance matrix is a $(m+n) \times (m+n)$ matrix containing distance constraints, upper and lower bound, for each pair of residues in the system. Since the PREs measure the distance from a label to the HN atom of a residue, the amino acid residues were represented by their HN atoms in the distance matrix. The distances between the spin-

label (OAB atom) and any HN atom in the structure are categorized into three ranges: (0, 15 Å), (15, 25 Å), and (25 Å, 150 Å). Only within the range of 15–25 Å, a spin label can yield distance constraints of 15-25 Å with accuracy approaching ±15%. Besides the PRE constraints derived from experiments, we also used distance constraints within each transmembrane helix predicted by TMHMM2 [27]. In addition to distance constraints determined by PRE labels, the distance constraints between pairs of HN atoms in the same TM helix are calculated from an ideal helical structure, with the error set to 10% of the HN-HN distance. Additional constraints are used to assure that the helices are roughly parallel to each other and perpendicular to the membrane surface. Specifically, for all pairs of helices, the end-to-end distances on the same side of membrane are set to be equal, and the end-to-end distances across the membrane are set to be the hypotenuse of a right-angled triangle, where the sides are the length of the ideal helix and the end-to-end distance on the same side of the membrane. The errors were set to 15% of the distances. If two helices have an unequal number of residues, the end residues on the longer helix are adjusted accordingly to match the length of the shorter helix.

Structure prediction from distance constraints

We developed an algorithm based on the stochastic proximity embedding (SPE) procedure to implement a distance geometry search. The procedure starts from a random initial conformation by assigning random coordinates to m+n residues in the system. Next it calculate the distance matrix D' and distance discrepancy matrix V = |D-D'|, in which D is the desired distance matrix descripted before. The system randomly selects a

pair of residues *i* and *j* with the probability proportional to V_{ij}. Then it updates the coordinates by moving them to satisfy the distance constraint D_{ij}. The process is repeated until all constraints are satisfied. In case the algorithm does not converge to a structure satisfying all the constraints in a prescribed number of iterations (N=1,000,000), the algorithm is stopped and restarted with a different set of initial random coordinates. This algorithm generates structures that satisfy all the distance constraints both faster and with fewer inconsistencies than the traditional matrix embedding [28] of conventional distance geometry algorithms [29].

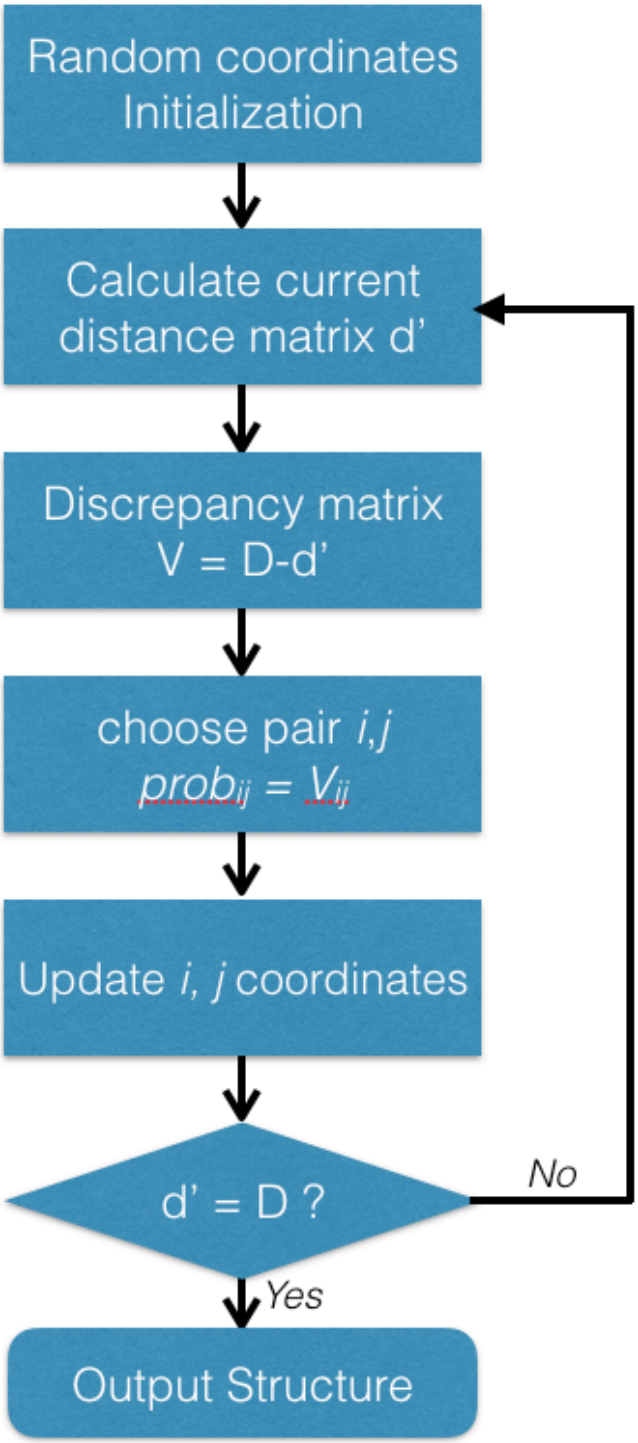

Figure 1.1 The Schematic of distance geometry based structure prediction

Structure Clustering selection

1,000 structures that satisfy all the constraints for each protein were generated using the algorithm described before. To select the best model, the structures were clustered based on their pair-wise RMSDs by structural alignment [30], and then the model clusters are ranked by the size. For each cluster, a centroid structure was generated using the SPIKER program [31]: first, the structure which has the smallest RMSD to all the other members was identified and designated as the cluster center; second, all the member structures were superimposed to the cluster center model and their new coordinates were averaged to create the centroid structure. The centroid structure may have distorted helical structures. The theoretical helices were then structurally aligned to the centroid structure to create the final model. When the cluster centroid structure has steric clashes, we use the closest-to-centroid structure (i.e., the single structure with the best RMSD to the centroid structure) to superimpose the ideal helices. For benchmarking results, the best centroid (or closest-to-centroid) models of the top 10 clusters are used as prediction.

TM Helix Packing Topology 2D Analog

In this study, we focus on strategizing positions of PRE mutation sites. Statistical analyses of helix-packing motifs in membrane proteins indicate that interacting helix pairs are in general approximately vertical to the membrane surface and are nearly parallel to one another [32]. For an approximate model, we assume that the helices are parallel and nearly perpendicular to the membrane surface. Therefore, the problem of finding the relative positions of *n* helices can be reduced to identifying the geometry of the *n* termini on a plane. The lengths of the loops are also assumed to be long enough not

to be a determinant of helix packing. The PRE spin labels will be attached to cysteines at the ends of helices as the introduction of cysteine mutations and nitroxide labels in the middle of helices are more likely to disrupt the structure. In a non-channel forming helical bundle, helices will generally maximize interactions with other helices forming pairwise interactions with two to six other helices [33]. For four helices this generally implies a rhombus packing topology, where every pair of neighboring helices on the sides of a rhombus interact and the helices on the opposite sides of the short rhombus diagonal also interact with each other. To see how well the model superimposed on real structures, we constructed a 3D model using ideal helices, which are parallel to each other and perpendicular to the membrane surface, assuming the correct helical arrangement. Structural alignment of the 3D rhombus model to five unique four-helix bundles in our benchmark set, namely DsbB (2hi7B), ligand gated ion channel (2vl0A), leukotriene C4 synthase (2uuhA), V-type sodium ATPase (2bl2A), and particulate methane monooxygenase (1yewC), shows the model has a root-mean-square deviation (RMSD) over the Cα atoms to the natives at 4.1 Å, 3.9 Å, 5.0 Å, 3.3 Å, and 4.4 Å respectively.

Determine Packing Topology using One to Three Labels

For the four points forming a rhombus, there are 12 possible helix-packing topologies consisting of six pairs of mirror structures (Figure 1.2). We now examine how to use PRE data to distinguish among the six pairs. The long diagonal of each rhombus has a unique distance different from all the other distances within the rhombus, i.e., 21 Å in this case. For example, in Figure 1.2A this distance is the distance between the helix termini 2 and 4 in the first model. Placement of a PRE spin-label on either of these two sites would provide the unique distance of 21 Å to an NMR observable nucleus at the

other site, thus allowing for the identification of the topology. The other two sites, 1 and 3, can only provide the non-unique distances of 12 Å. Because only two of the four sites for PRE spin labeling have a 21Å distance, one has 50% of chance to pick a site that will allow for the unique determination of the correct topology.

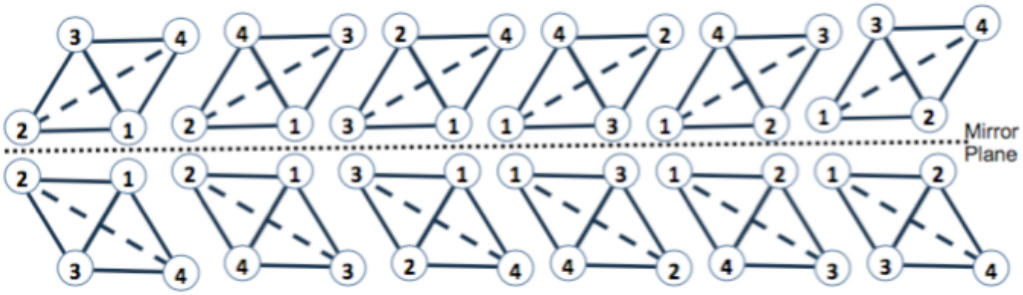

Figure 1.2 Total 12 helix packing topologies for the rhombus model of a four-helix bundle

In the event when the first PRE label is not placed on a helix at one end of the long diagonal, the placement of a second label at the end of any helix on the same side of the membrane will allow for identification of the correct topology, regardless whether the second label provides the 21 Å distance measure (hence unique) or just a set of 12 Å distance measures. In the latter case, the two involved helices are on the opposite sides of the short diagonal and therefore the other two helices must be on the opposite sides of the long diagonal.

We conclude that we can uniquely determine the correct topology out of the six possible topologies with a 50% probability of success based on a placement of one PRE label and with a 100% probability of success based on placement of two PRE labels, as

long as the labels are placed on the same side of the membrane, assuming no non-helical linker can cross the membrane. Restriction to the same side is easily done knowing the connectivity of TM helices in the protein sequence.

The analysis was extended to up to seven helix bundles because there is so much interest in seven helix GPCRs. We observed in solved structures that helix bundles with more than four helices have rhombus-shape substructures, and most helices interact with at least two other helices from either the same protein monomer or other protein subunits. This motivated us to build the topologies for proteins with a higher-number of helices by adding one helix at a time to the rhombus-based models, assuming each new helix interacts with at least two existing helices. Figure 1.3 shows all possible geometric models for the layouts of five to seven helical bundles, as viewed from either side of the membrane. There is one exception to this rule, a six-helix channel, 6-1, which can be generated by removing a central helix from 7-1 of the seven helix models. Each model has a large number of permutations of helix order.

We now examine the minimal number of sites needed to distinguish the correct topology for each model. Specifically, we examine all combinations consisting of a fixed number of sites to check which of them gives rise to the PRE data that can determine the correct topology. Assuming PRE data can distinguished the short and long pairs in the rhombus shape, for example 1-3 and 2-4 of first model Figure 1.2, Table 1.1 lists all the correct combinations with the minimal number of sites needed for each model. From the table, we can see that the minimal number of sites for five to seven helical bundles is two, and the probabilities for selecting the correct two sites for five, six, and seven helical bundles are on average 60%, 58%, and 29%, respectively.

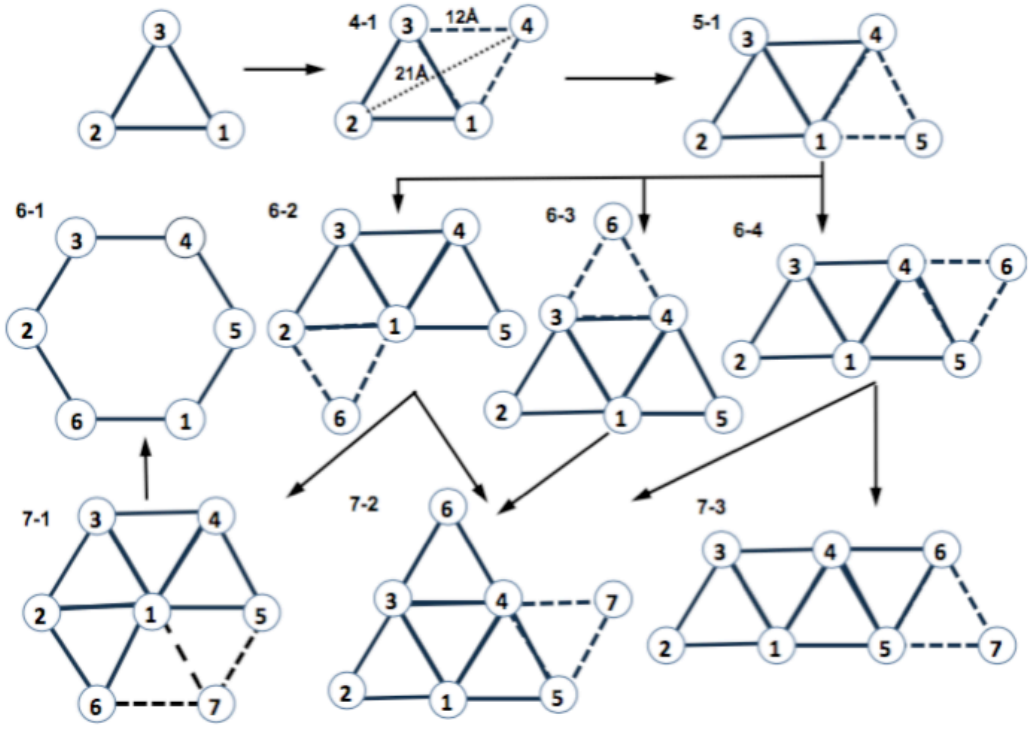

Figure 1.3 Geometric Models for the Layouts of Four to Seven Helix Bundles. The models for four to seven helix bundles, derived by adding one helix at a time to the models of the previous set.

Table 1.1 Theoretical Analysis of the Minimal Number of Mutation Sites Needed to Determine the Correct Packing Topology. Sites number and model are corresponding to models shown in Figure 1.3

Model	Pairs of mirror topologies	Minimal sites needed	% combination
4-1	6	1: 2, 4	50%
5-1	60	2: 3+4, 2+5, 2+3, 4+5, 1+5, 1+2	60%
6-1	60	2: 6 pairs like 2+4, 6 adjacent pairs like 2+3	80%

6-2	360	2: 3+4, 2+3, 5+6, 4+5, 2+6	33%
6-3	120	2: 3 pairs like 2+5, 6 adjacent pairs like 1+2	60%
6-4	180	2: 3+6, 2+5, 2+6, 6 adjacent pairs like 3+4	60%
7-1	420	2: 6 adjacent pairs	29%
7-2	2520	2: 2+5, 2+7, 2+6, 1+5, 3+6, 2+3	29%
7-3	2520	2: 3+6, 3+4, 4+6, 1+5, 1+2, 5+7	29%

RESULTS

Structure Prediction of DsbB Constrained by PRE DATA

The utility of the above prediction capability can be examined by using both experimental and simulated PRE data and comparing predicted with observed structural topologies. Firstly we show an application of our prediction capability to protein DsbB, which has a crystal structure, an NMR solution structure, and some PRE data available [13]. The protein is 176 residues long and has four TM helices. The predicted TM residues using TMHMM2 [27] are: TM1 (A14–V35), TM2 (I45–A64), TM3 (Y71–Y89), and TM4 (W145–I162). PRE data were collected from nine mutational sites, six of which are located at helix termini, i.e., A14, V72, and V161 on the intracellular side of the membrane, and L30, L87, and Y89 on the extracellular side. Three other sites (Q122, F137, and G139) are located in loops. Since our model was designed for determination of topology of helix topology and orientation, the loop mutation sites were excluded and only the first six sites were used.

Table 1.2 lists the number of PRE data from each site to the helices grouped into three ranges: (0, 15 Å), (15, 25 Å), and (25 Å, 150 Å), and the associated distances to the

ends of helices on the same side of membrane as label-to-end distances. A distance only within the range of 15–25 Å can be measured with an error of 2–4 Å whereas for the other two ranges, we can only say that the distance is <15 Å or >25 Å, respectively. We refer the first type of PREs as specific and the other two types as loose constraints. The crystal structure of the TM regions of DsbB (PDB: 2hi7B) and the predicted spin-label locations on the structure are shown in Figure 1. The labels are not included in the crystal structure hence they are computational predicted. In the shown model, the spin label is on average 5-7Å away from its attached helix ends.

Table 1.2. Experimental PRE data for DsbB. Number of constraints to helices list the total number and each number of constrains in the three ranges of distance of <15Å, 15-25Å, >25Å. Distance between label site and other helix ends on the same side of membrane.

Side	Label Site	Location on helix	Number of constraints to helices				Distance to other helix ends (Å)			
			1	2	3	4	1	2	3	4
Intracellular	A14	1	13 (5,1,7)	18 (0, 10, 8)	14 (0, 5, 9)	16 (4, 8, 4)	-	16.6	19.7	16.6
	V72	3	15(0, 2, 13)	17 (4, 8, 5)	7 (2, 2, 3)	13 (0, 5, 8)	21.6	16.3	-	18.6
	V161	4	18 (0, 3, 15)	18 (0, 9, 9)	5 (1, 3, 1)	7 (0, 1, 6)	19.8	16.9	<15	-
Extracellular	L30	1	10 (3, 3, 4)	11 (3, 6, 2)	11 (0, 0, 11)	11 (0, 0, 11)	-	20.1	>25	>25
	L87	3	14 (0, 0, 14)	9 (1, 3, 5)	4 (0, 2, 2)	13 (0, 9, 4)	>25	<15	-	18.4
	Y89	3	17 (0, 0, 17)	18 (0, 8, 10)	13 (3, 4, 6)	14 (5, 6, 3)	>25	19.1	-	<15

We computationally folded the structure using constraints by the PRE data. The folding results are listed in Table 1.3. The models were compared with the crystal structure on the TM helical region using the TM-score program[34]. TM-score is a structure similarity score that ranges in [0, 1] with a higher value indicating a stronger

structural similarity. TM-score > 0.4 means statistically significant structural similarity [34]. Visual examination indicates TM-score > 0.4 generally give correct helical arrangement. Thus, TM-score = 0.4 was used as cutoff for correct topology. The best model predicted by the algorithm has correct topology by using any single label placed at the ends of the long diagonal and with specific constraints to all helices (i.e., label14, label72), whereas the algorithm using any label without a specific constraint (i.e., label30, label87, label89) or placed on the end of the short diagonal (i.e., label161) did not lead to correct topology. The DsbB structure was then folded using PRE data associated with any two labels on the same side of the membrane (except for label87 and label89 that are on the same helix terminus). The results are listed in Table 1.3 and the models are shown in Figure 3B. The models for all possible site combinations have the correct topology with an average rmsd 4.8 Å to the crystal structure.

Table 1.3 Results of Structure Prediction for DsbB using Experimental PRE data

Label Site	Rank of best cluster	Size of best cluster (%)	Best cluster centroid structure RMSD (Å) / TM-scrore
14	4	16.1	5.37 / 0.40
72	4	13.0	5.69 / 0.43
161	4	12.9	6.31 / 0.36
30	6	10.9	9.36 / 0.31
87	2	17.0	6.74 / 0.30
89	1	20.3	6.97 / 0.32
Average		15.0 ± 3.4	6.74 ± 1.42 / 0.35 ± 0.05
14, 72	1	32.0	4.33 / 0.47
14, 161	1	30.1	4.38 / 0.45

72, 161	2	24.1	5.10 / 0.42
30, 87	1	37.0	4.85 / 0.41
30, 89	1	30.5	5.47 / 0.40
Average		30.7 ± 4.6	4.83 ± 0.48 / 0.43 ± 0.03

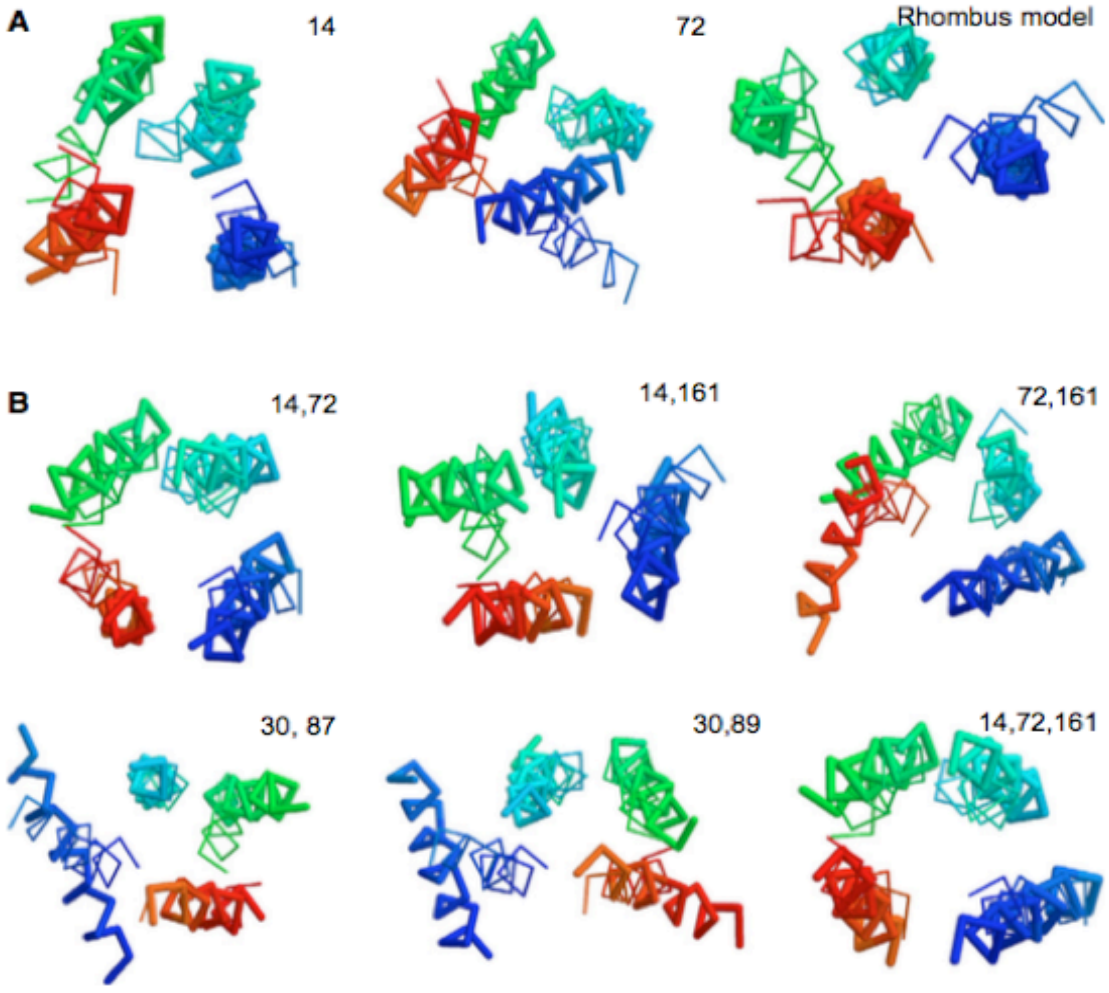

Figure 1.4 DsbB predicted structure from experimental PRE data. (A) Structures based on a single PRE label. (B) Structures based on two PRE labels. The predicted structure (thick line) is aligned to the crystal structure (thin line) and colored from blue (N terminal) to red (C terminal)

Higher Order Helix Bundles using simulated PRE data

To extend the test to proteins with diverse topologies, simulated PRE data derived from crystal structures of a set of unique proteins with four to seven TM helices. A set of TM protein structures were collected from the OPM database [35] as follows : (1) all the polytopic chains were collected and culled at the PISCES server [36] by the criteria that the structure was determined by X-ray crystallography with resolution <3.5 Å, and that the sequence identity was <35%, resulting in 57 chains with ≥4 TM helices; (2) Pair-wise structure alignment using TM-align [34] was conducted on the TM helical region. A TM-score > 0.5 means two structures share the same structural fold [37]. For structures belonging to the same fold, the one with the longest TM sequence or the highest resolution was retained, which resulted in 25 unique structures; (3) The structures were manually examined and those chains that form a single bundle (i.e., each helix interacts with at least two other helices) with other chains but have extended monomeric structure were excluded (these are mainly multi-subunit proteins involved in photosystems or photosynthetic reaction centers). At this stage, we only consider monomeric bundle structures without considering complex structures involving multiple proteins, although the intra-protein and inter-protein helices association may be the same. Ten proteins with 4-7 TM helices were obtained from such a filtering scheme (Table 1.4). The X-ray structure of DsbB (2hi7B), initially excluded because its resolution was lower than the 3.5 Å cutoff, was also added to the set. The resulting 11 proteins include 5 four-helix bundles, 5 six-helix bundles, and 1 seven-helix bundle. Since there is no individual test case for a five-helix bundle, we used the sub-bundle structures formed by five helices from the six or seven helical proteins, providing they were not structurally similar to one

another. Such sub-bundle structures were also included in the four-helix and six-helix groups, resulting in a total of 18 test cases (Table 1.4).

To simulate the PRE data from crystal structures, we mutated *in silico* the amino acids at the selected sites to a cysteine residue carrying a PRE spin-label using LEaP package in AMBER [29]. The spin-labeled sites are those residues predicted by the TMHMM2 program [27] to be at the ends of TM helices. An energy minimization step is carried out on each mutated residue to remove steric clashes and minimize the van der Waals energy in AMBER. The distance between the spin-label (OAB atom) and any HN atom in the structure is calculated and grouped into three ranges: (0, 15 Å), (15, 25 Å), and (25 Å, 150 Å). Only within the range of 15–25 Å, is a distance specifically constrained (with an error of ±3 Å).

Based on our observation on optimal two label selection, most exposed helix should be selected first for PRE label. The strategy for optimal site prediction is as follows: The lipid accessible surface area (ASA) for each residue was predicted from sequence using the ASAP server [38]. For 4-6 helical bundles, the most exposed helix is selected as the first helix to label. For seven-helix bundles, if the 7-3 topology is identified by the lipid accessibility prediction (i.e., the top two lipid accessibility's are significantly higher than the others), the third most exposed helix will be selected as the first site; otherwise, the most exposed helix will be selected as the first site. The second and subsequent sites (if needed) are selected iteratively based on the next most exposed helix and the results confirm the theoretical prediction that it is possible to use PRE data from a minimal two sites to predict the correct topology for up to seven-helix bundles if they are properly selected (Table 1.4)

Table 1.4 Results of structure predictions for benchmark proteins using simulated PRE data

Protein Name	PDB Chain	Model	Averge RMSD (Å)	Models from Predicted Sites(Å/TM)
DsbB	2hi6B	4-1	5.0 ± 1.1	4.6 / 0.49
Ligand Gated ion channel	2vl0A	4-1	6.5 ± 1.0	4.9 / 0.47
Leukotriene C4 synthase	2uuhA	4-1	4.8 ± 0.9	3.9 / 0.50
Particulate methane monooxygenase	1yewC(1-4)	4-1	6.5 ± 1.0	5.8 / 0.44
Particulate methane monooxygenase	1yewB(2-5)	4-1	5.0 ± 0.6	5.2 / 0.39
Calcium ATPase	1wgpA(5-8)	4-1	4.4 ± 0.8	3.5 / 0.54
Average		4	5.4	4.6 / 0.47
Calcium ATPase	1wpgA(5-8,10)	5-1	5.3 ± 1.0	5.8 / 0.46
Protease gIpG	2ic8A(1-5)	5-1	6.4 ± 0.9	6.2 / 0.36
Particulate methane monooxygenase	1yewB(2-5,7)	5-1	5.4 ± 0.9	5.2 / 0.43
Bacteriorhodopsin	1m0lA(2-4,6-7)	5-1	6.2 ± 1.1	4.9 / 0.47
Average		5	5.8	5.5 / 0.43
Calcium ATPase	1wpgA(5-10)	6-2	6.1 ± 0.9	4.5 / 0.54
Aqaporin Aqpm	2f2bA(1-2,4-6,8)	6-1	6.4 ± 1.1	6.2 / 0.43
Protease gIpG	2ic8A	6-3	6.4 ± 0.9	6.5 / 0.44
Particulate methane monooxygenase	2yewB(1-5,7)	6-2	6.4 ± 1.3	6.1 / 0.38
Bacteriorhodopsin	1m0lA(2-7)	6-4	7.4 ± 1.0	6.2 / 0.44
Average		6	6.5	5.9 / 0.45
Bacteriohodopsin	1m0lA	7-3	8.2 ± 1.2	7.9 / 0.43
V-type sodium ATPase	2bl2A	4-1	7.4 ± 0.6	7.0 / 0.42

DISCUSSION

We have theoretically analyzed and computationally verified that one to two PRE sites should be sufficient to constrain solution NMR structure prediction for four to seven helical bundles. Our approach for the optimal site prediction successfully predicts the minimal sites for up to six element helical bundles. Improving the lipid accessibility prediction will likely improve the prediction results for seven-helix bundles.

Because only a few structures of membrane protein families are currently available, template-based structure prediction methods do not work in general for membrane proteins [39]. At the same time, *ab initio* approaches suffer from a major hurdle in that a significant portion of conformation space must be sampled to derive a final structure. This often makes the approach computationally unfeasible. By determining the correct helix-packing topology of a membrane protein and producing a starting point having a native fold, the computational space can be significantly reduced and an accurate structure may be determined using additional prediction methods. The study presented here provides a useful approach to deriving starting models for membrane proteins having a correct topology using a small number of experimental data and a simple structure prediction method.

CHAPTER 3

IDENTIFICATION OF STRUCTURAL MOTIFS BY SEGMENT THREADING[2]

[2] Fei Ji, Huiling Chen, Victor Olman, Sitao Wu, Yang Zhang and Ying Xu. Submitted to PLOS One, 10/25/2015.

INTRODUCTION

The rapid advancement in *omic* techniques has made it possible to produce genomic sequences and functional data such as transcriptomic and metabolic data in a high-throughput manner. Due to the challenging nature of the problem, protein structure solution has not been amenable to high-throughput approaches. This has made structural data collection a bottleneck when connecting genomic sequence data to the low-resolution functional data such as gene-expression data for functional mechanism studies. Fortunately computational techniques can offer useful structural data, albeit not at the highest possible resolution, based on a widely-believed hypothesis: the number of structural folds is relatively small for all protein structures in nature [40, 41] and the already solved protein structures in the Protein Data Bank (PDB) may have covered the majority of all the possible structural folds [42].

Among different classes of protein (tertiary) structure prediction techniques, threading represents the most generally applicable. A number of protein structure libraries, a key component of threading-based prediction methods, have been developed based on PDB structures to facilitate protein structure prediction using threading methods, such SCOP [43] and CATH [44]. A variety of algorithmic methods have been developed to execute the ideas of threading-based structure prediction, such as sequence profile-profile alignment [45, 46], structural profile alignment [47, 48], hidden Markov model [49-51], machine learning [52] or pairwise optimal scoring search [53-56].

In this chapter, we present a new segment-identification and assembly based threading algorithm and software package, SPRED, for protein structure prediction. The segments structure units identified from PDB protein structures, are used as structure template library for protein structure prediction through (1) substructure-based threading and (2) assembly of identified substructures. SPRED has been tested on 317 non-homologous proteins from Protein Data Bank (PDB) [57].

The overall TM-scores by the SPRED alignments increase by 11.4% compared with those by the best whole-chain threading methods. When tested on 82 CASP10 targets, SPRED improves the TM-score by 13.0% compared to the best predictions of whole-chain threading methods. The significant improvement is achieved predominantly due to the accurate identification of native-like substructures of target proteins from different structures of possibly different structural folds. The segment based threading in theory could predict proteins with novel folds, and concurrently requires substantially less computational resource compared to the traditional fragment-based *de novo* methods.

METHODS

Structure Segment Library

To generate our substructure-level template library, we divide each full-chain structure in PDB to a collection of structural segments based on the secondary structures. For each PDB structure, a *secondary structure unit* (SSU) is defined as a complete α-helix or β-strand determined using DSSP [26] and then refined using the following criterion: 1) only SSUs with at least four residues are considered and the shorter ones are

defined as non-SSU; and 2) adjacent SSUs are joined together if they are sequentially separated by at most two amino acids. A *segment* is defined as substructure consisting two or three consecutive SSUs having at least 30 residues. We have extracted all the segments from the PISCES database of non-homologous protein structures. This database consists of 15,605 full-chain proteins with lengths ranging from 100 to 500 amino acids and the highest pairwise sequence identity within the set being 25%. A total of 203,569 segments are generated from these protein structures, with the average length of 43 residues per segment, which serve as the Segment Structural Library in SPRED. Figure 3.1 shows an example of three segments of a protein structures with four SSUs.

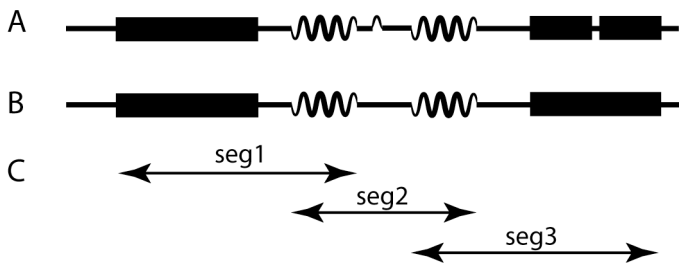

Figure 3.1: An illustration of segment definition. (A) A protein having multiple SSUs with each wave line representing a helix, a band denoting a strand and a thin line for a coil. (B) A short secondary structure unit in the middle is converted to being a part of a coil and two nearby strands are merged into one. (C) Three segments represent all the segments generated from the protein.

Alignment

In SPRED, substructure segments are used as the fundamental structural templates for a threading alignment. We have employed two state-of-the-art threading programs, MUSTER [27] and HHpred [13], to generate the initial segment alignments, separately, denoted as M-align and H-align. Considering that these two methods were designed for full-chain sequence alignments, a number of adjustments are made for each initial alignment by the two programs since segment-based alignments are much shorter and do not necessarily have compact local structures.

M-align uses a scoring function similar to that of MUSTER, which consists of terms related to the sequence-based profile-profile alignment, structure profile alignment, secondary structure, solvent accessibility, torsion angle and hydrophobic residue matches [27]. Therefore, the score for matching the *i*th residue of a query sequence to *j*th residue of a segment template is given by

$$\text{Score}(i,j) = E_{\text{seq-prof}} + E_{\text{sec}} + E_{\text{struc\_prof}} + E_{\text{sa}} + E_{\text{phi}} + E_{\text{psi}} + E_{\text{hydro}}$$

The parameters for each term and the scaling factors across all terms are refined based on our segment alignment training set. The optimal alignment on a query sequence for each segment template was identified using a dynamic programing approach as in MUSTER, and the raw score for each alignment was normalized to a Z-score. The final candidate alignments of M-align are selected based on a Z-score threshold determined based on our training data.

The H-align is derived based on HHpred, which detects homologous proteins based on an optimal alignment between a hidden Markov model (HMM) developed for each template structure and an HMM for the query sequence. To use the program, we have downloaded the full-chain template structure library from the HHpred server, and divided the HMM profile for each full-length protein at the segment boundaries defined in SPRED to generate a segment-based HMM profile library. The candidate alignments are generated using HHpred against our segmental HMM profile library, and ranked based on the p-values determined based on our training data.

Assembly

For a query protein sequence, numerous segment alignments would be generated from both the M-align and the H-align, respectively. Note that the lengths of the segments in the template library range from 30 to 98 amino acids, therefore our segment-based alignments cover only short range pairwise interactions between residues within each segment. To deal with longer range interactions that are potentially significant for the determination of the overall structural fold, SPRED merges each pair of aligned segments to a *coupled-alignment* if the two segments are from the same protein and have (the closest) inter-segment Cα atoms closer than 8Å in the original structure (Figure 3.2). The so defined coupled-alignments, regardless of discontinuous or not in the original sequence, provide longer-range pairwise interaction information; and hence segment alignments against them will have no gap penalty for the non-consecutiveness between the two relevant substructures.

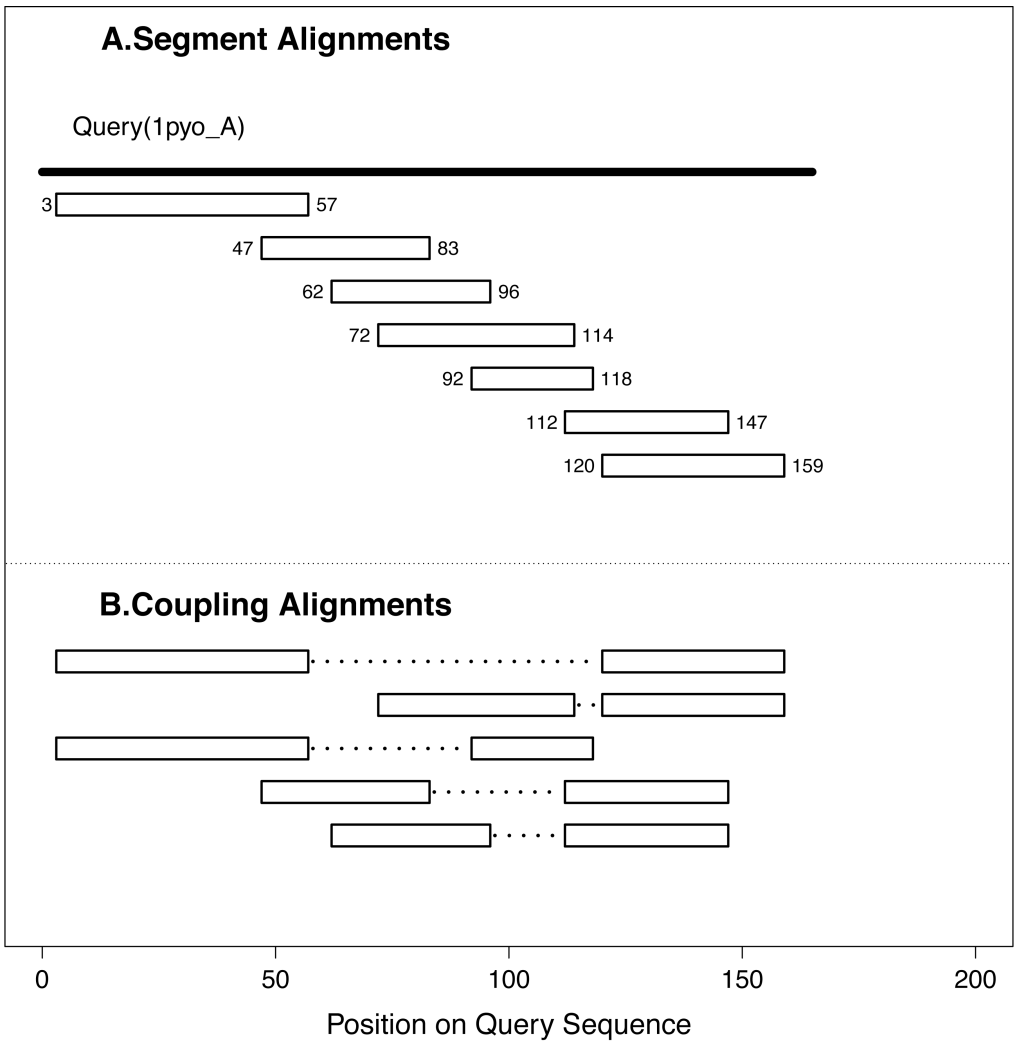

Figure 3.2: An illustration for segment alignment and coupled segment-alignments. (A) 1nw9_B has a total of 7 segments aligned with different regions of a query sequence, with the horizontal position of each box denoting the aligned position of each template segment on the query sequence. The numbers marked on each box are the positions on the template protein. (B) If a pair of aligned non-overlapping segments is spatially close in the same template structure, they form a coupled-alignment without a gap penalty in the middle (dashed line).

Numerous segment alignments and coupled-alignments will be generated from the previous steps, each covering a part of the query protein and some segment alignments possibly overlapping. To generate a full structure for the query protein from these aligned regions, SPRED randomly selects a minimal set of overlapping segments and coupled-

segments that together maximally cover the query sequence to generate a full-chain model by the multiple alignment module of MODELLER [28]. Specifically, the program extracts distance and dihedral angle restraints from aligned substructures and then optimizes the all-atom full structure using the CHARMM22 force-field [29]. The conformations generated by different sets of aligned substructures are then clustered using our in-house clustering algorithm [30], and for each cluster, a centroid structure is selected using SPICKER [31]. The centroid structures from the largest clusters are selected as final predictions.

Protein Structure Dataset

We retrieved all the non-homologous single-chain proteins from the PISCES server [32] that satisfy the following criteria: structural resolution cutoff at 1.6Å, R factor cutoff at 0.25 and pairwise sequence identity cutoff at 25%, which gives a total of 953 proteins. These proteins are divided into two sets, randomly assigning 67% and 33% to the training and the test sets, respectively. The test set contains 317 proteins, 216 of which are considered as "easy" targets and 101 as "hard" targets based on the Z-score of full-chain alignment program MUSTER. Specifically, if the alignment Z-score > 7.5, the topology and folding prediction is usually correct, and hence considered as "easy" targets; and if Z-score < 7.5, the targets are defined as "hard". The training set has 636 proteins, with 473 easy and 163 hard targets.

RESULTS

Predictions are evaluated using the TM-score [33] along with root mean square distance (RMSD). The TM-score was designed to assess the alignment quality in terms of both accuracy and coverage, which overcomes the length effect issue of RMSD: longer

proteins tend to have higher RMSD [34]. A TM-score ranges between 0 and 1.0 with TM-score = 1.0 indicating identical structures, and the average TM-score between two randomly selected protein structures is 0.17.

We have tested the performance of SPRED on 317 proteins against our segment library described before. The average TM-score over all the alignments between each of the 317 proteins and its native structure is 0.589, with an average RMSD at 4.6 Å; the "best in top 5" prediction has a higher average TM-score at 0.601. The performance of SPRED on different categories of target proteins is summarized in Table 1. We also listed RMSD result in each category by SPRED and by MUSTER and HHpred in Table 3.1, and SPRED achieved significantly lower RMSD than both full-chain methods.

Table 3.1: The average TM-score of the structures predicted by different programs on the test set and CASP10 targets. Boldface numbers show the best result in each category

	SHRED	**HHpred**	**MUSTER**
All(317)	0.589	0.561	0.557
Hard(101)	0.399	0.358	0.364
Easy(216)	0.678	0.655	0.646
CASP10	0.538	0.476	0.497

Despite longer proteins tend to have more alignment segments, no strong correlation has been observed between TM-scores and query sequence lengths across all prediction targets. This suggests that MODELLER has integrated structural constraints from multiple template segments without any artifacts during the full-chain structure modeling.

Table 3.2: The average RMSD the structures of first alignments and best in top 5 alignments predicted by different programs on the test set and CASP10 targets. Boldface numbers show the best result in each category.

	Top1			Best in Top 5		
	MUSTER	HHpred	SHRED	MUSTER	HHpred	SHRED
Hard(101)	8.14Å	7.68Å	6.23Å	7.92Å	7.57Å	6.08Å
Easy(216)	4.10Å	4.23Å	3.67Å	4.02Å	4.11Å	3.52Å
CASP10	5.96Å	5.03Å	4.44Å	5.89Å	4.86Å	4.13Å

To evaluate the performance of SPRED, we compared the prediction results with HHpred and MUSTER, based on which SPRED is developed. SPRED shows consistent improvement on average TM-scores over both HHpred and MUSTER across all categories of prediction targets (Table 3.1). All 317 test proteins have an average TM-score of 0.589, increasing 5.1% and 5.9% over that by HHpred and MUSTER, respectively. The TM-score improvement on the 101 hard targets is 9.6% and 11.4% over MUSTER and HHpred, respectively. A detailed TM-score comparison between SPRED and MUSTER and between SPRED and HHpred is shown in **Figure 3.3** across all the 317 proteins. Each point in the figure represents a protein and points above the diagonal line are the proteins on which SPRED outperforms MUSTER or HHpred.

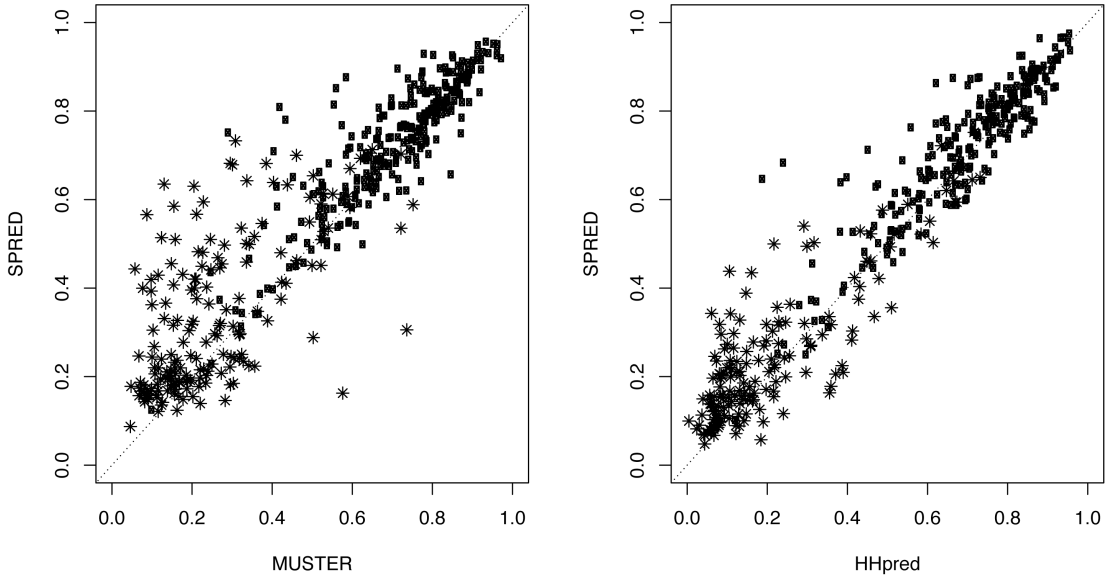

Figure 3: TM-scores of the best threading alignments for 317 test targets with substructure identified by SPRED *versus* those by full-chain threading programs MUSTER and HHpred, respectively. Dots are for easy targets and asterisks are for hard targets.

For the proteins whose SPRED TM-scores are at least 0.2 better than those of MUSTER, MUSTER tends to miss their optimal structural templates in its template library, but interestingly HHpred tends to identify them correctly. This is also true for cases on which HHpred did substantially poorer than SPRED but MUSTER tends to do well. These strongly suggest that SPRED has captured the strengths of each program and is capable of selecting the better templates selected by the two methods. Hence overall the combination of the two data sources has helped to increase the final prediction accuracy. Overall, SPRED has better TM-scores on 144 out of the 317 protein targets than both MUSTER and HHpred.

To assess the contribution by substructure identification and assembly in SPRED, apart from template selection, we compared to a modified LOMET server as a control. LOMET [35] is a meta-sever for automated prediction of protein structures through combining prediction results by nine state-of-the-art threading programs, including MUSTER and HHpred. We compared LOMET selections based only on predictions of MUSTER and HHpred (denoted as LOMET_MH) with SPRED predictions on the test set. Table 3.3 shows the comparison results. Clearly on this test set, the SPRED predictions are consistently better than the selections of LOMET_MH, which indicates that substructure threading and assembly method employed by SPRED indeed improves the quality of a threading approach as the two programs under comparison use essentially the same threading scoring functions and algorithms except that SPRED uses segment-based threading while the other two use full-chain based threading.

Table 3.3: The average TM-scores of the structures predicted by LOMET selection from MUSTER and HHpred predictions (LOMET_MH) and of SPRED on the test set and CASP10 targets.

	All(317)	**Hard(101)**	**Easy(216)**	**CASP10**
SPRED	0.590	0.399	0.678	0.538
LOMET_MH	0.572	0.382	0.661	0.516

On the 317 test proteins, a total of 32,236 aligned segments are above cutoff scores and used as candidate segments for structure assembly. The average number of candidate segments used per protein is 102, with the average aligned segment length of 40.7 residues, which is close to the average segment length (42 residues) in the template library. This suggests no bias towards either long or short segments. For the 32,236 candidate segments, 42% are from templates in structural folds different from that of the query protein according to SCOP [7], which are referred to as cross-fold alignments. We noted that 63% of the cross-fold alignments have TM-scores > 0.5 or RMSD < 2, which are generally considered as "accurate" alignments and are crucial for the overall accuracy of the final structure assembly. For segments aligned from templates in the same folds, this ratio rises to 76% having "accurate" alignments. It is worth mentioning that the cross-fold alignments account for nearly half of the segment alignments in SPRED (42% for all test targets and 64% for hard targets). The cross-fold structural information was clearly not considered by full-chain threading methods, but captured by our segment-based method, which has made a significant contribution here for the improved structure prediction.

Note that segments used in our program are substantially longer than the fragments (~10 residues) typically used in fragment-based methods [36]. Clearly the longer segment sizes used in SPRED have greatly reduced the computational cost when compared to fragment-based *de novo* methods such as ROSSETA [36]. For example, the average computational time for SPRED for each of the 317 targets is under one hour on our computers with Intel® Xeon® CPU @2.67GHz, while ROSETTA takes on average 20

hours on the same computer. Yet, SPRED clearly has the capability in predicting protein structures with novel folds.

Figure 3.4 shows an example of assembled structure by SPRED using multiple segments of different folds. For this target (PDB_ID:2VZC, calponin homology domain of alpha parvin), the TM-scores of the best alignments by MUSTER and HHpred are 0.74 and 0.66, respectively. In comparison, the SPRED structure has a TM-score at 0.87. Significant improvements on various segments can be observed in the predicted structures by SPRED *versus* those by the other two programs. Furthermore, the best aligned single structure in our library has a TM-score of 0.76, lower than the SPRED final structure. The improvement is clearly due to the combination of alignments from multiple templates, including those from different structural folds.

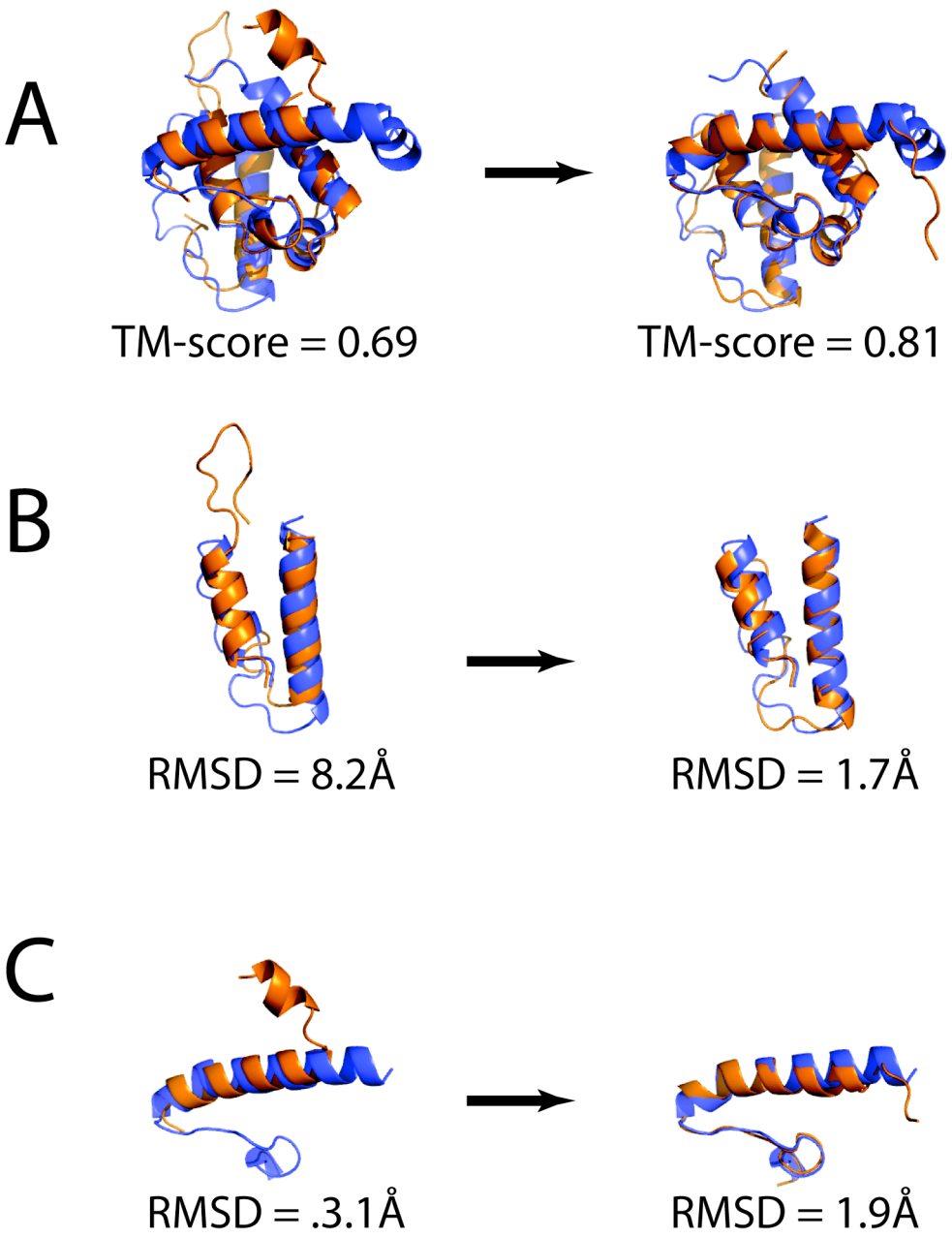

Figure 3.4: Segment-based structure prediction of 2VZC. (A) The left side is the best alignment identified by full-chain threading tools MUSTER and HHpred (LOMET_MH) (orange) superimposed on the native structure (blue). The right side is the SPRED prediction model (orange). (B) and (C) are representative segment alignments compared between LOMET_MH (left) and SPRED (right).

While the major improvement by SPRED is on hard targets with novel folds, it is of interest to compare the segment-based and full-chain threading methods on easy targets, most of which have full-chain templates existing in the library. 216 out of the 317 target proteins fall into the category of easy prediction targets, most of which have the correct structural templates identified by both MUSTER and HHpred. A consistent TM-score improvement is observed by SPRED over MUSTER and HHpred on these prediction targets, as shown in Table 1, specifically achieving 4.9% and 3.5% improvement in the average TM-score over the two programs, respectively. The improvement in SPRED is mainly due to the more accurate local alignments which will be discussed in the following section. Overall, SPRED has better TM-scores than both MUSTER and HHpred on 96 of the 216 prediction targets. It is worth noting that for the targets with well-aligned templates (TM-score > 0.8) identified by a full-chain model, MUSTER and HHpred tend to perform more accurately than SPRED. Our analysis indicates that this is mainly due to the artifacts introduced by structural modeling by MODELLER in the structure-assembly step. Specifically, MODELLER used average structural constraints taken from multiple templates instead of the optimal template among the top threaded structures. Hence full-chain threading methods like MUSTER and HHpred remain to be the method to use for targets with a high level of homology with known PDB structures, while segment-based threading like SPRED is a better choice for hard targets or targets with novel folds. We need to mention here that the "homology" indicates structure similarity instead of sequence similarity, since all templates with a sequence identity >25% have been excluded.

DISCUSSION

We have developed a segment-based threading method SPRED, which predicts a protein structure through combining substructure threading and structure assembly, hence enabling structure prediction on proteins without native-like structural folds among the solved structures as templates. The assessment results clearly indicate that the method provides a highly effective tool for protein structure prediction complementary to the existing threading-based techniques as well as short-fragment based *de novo* prediction method in terms of applicability and practical usefulness.

The idea of using substructure identification and their assembly has been applied in several fragment-based methods for protein structure prediction. For example, ROSSETA [36] used fragments of fixed lengths, i.e., nine residues, to predict local structures and then assemble them into global structures. Chunk-TASSER [41] and I-TASSER [40] identify aligned structural fragments from a template structure library and assemble them into global structures by using Monte Carlo simulation for structural refinement. All these fragment-based methods are computationally expensive [22] and not practical for large-scale structure prediction. To the best of our knowledge, only one segment-based structure prediction method, SEGMER, has been published [42]. The key difference between SPRED and SEGMER is that SEGMER splits a query sequence into secondary structures and aligns each partitioned segments onto full-chain structure templates in SEGMER, while SPRED, in comparison, aligns a query sequence onto a set of pre-defined segment structures (or substructures). Since SEGMER only predicts the segment structures defined on secondary structure prediction of query protein without the whole-

chain assembly, it is hard to compare the accuracy of either the segment or whole-chain structure prediction with that of SPRED. Still, the advantage of SPRED is three-fold: (i) the program avoids the uncertainty of secondary structure prediction; (ii) it contains more alignment scoring functions (from MUSTER and HHpred) than SEGMER and (iii) it excludes the discontinuous segments used in SEGMER from consideration, which helps to substantially reduce the computational cost from an average of 5,799 segments per protein in SEGMER to 104 segments in SPRED.

It is worth mentioning that the segment alignment will not replace full-chain threading since they each have their strengths and weakness as discussed earlier. The full-chain threading remains an effective approach for identification of global folds for targets with well-aligned global structural templates as shown before. One possible weakness of a segment-based approach is that it may introduce structural variations when using multiple template structures; in addition, the assembly process is computationally more expensive than the threading alignment alone. Nevertheless, segment-based threading could prove to be a key to accurate detection of substructure motifs, as needed in functional annotation and identification of active sites. These advantages on segment-based threading suggested a novel method for ligand binding site prediction. Unlike the traditional threading based functional site prediction tools [43-46], one can use structural segments adjacent to the functional sites as the templates instead of full-chain structural templates for threading prediction. This will not only reduce the computational cost, but also improve the local alignment accuracy as shown in Table 3.3.

It is also worth noting that the alignment step in SPRED can be accomplished by different threading programs, and only minimal modification is needed when combined with a different threading program. Hence, SPRED could take advantage of the combination of different threading coring function and yield better results. In addition, the program can be easily integrated into other threading programs.

CHAPTER 4

A STRUCTURAL PERSPECTIVE ON IRON-HYDROGENASE UTILIZES BOTH FERREDOXIN AND NADH ON HYDROGEN PRODUCTION[3]

[3] Fei Ji, Xizeng Mao, Minseok Cha, Janet Westpheling, Ying Xu. To be submitted to PLOS One.

INTRODUCTION

In the last chapter, I presented a segment based structure prediction method to identify sub-structure similarity. In this chapter, application of such sub-structure identification helped us revealing a novel molecular model of trimeric hydrogenase in *Termotoga maritime*. The oxidation of NADH and reduced ferredoxin is coupled to H_2 production by trimeric [FeFe] hydrogenase in *Thermotoga maritima*, but the molecular mechanism remains unknown. The challenge has been to solve the 3-D complex structure by application of state of art tertiary structure and protein docking tools. Complex structure suggests that [FeFe] hydrogenase utilizes an electron from NADH by interaction with a homologous NADH catalytic subunit from NADH oxidoreductase. This finding revealed an alternative interaction of trimeric hydrogenase in microorganisms under different conditions. Comparative genomic analysis shows such mechanism is retained in multiple anaerobic species with a conserved regulatory transcription factor. The discovery gives a new perspective on our understanding of the redox proteins and mechanism of H_2 production in anaerobic bacteria.

Molecular hydrogen is a key intermediate in the metabolic interactions of a wide range of microorganisms. The main routes for hydrogen production are photoproduction and dark fermentation with the latter providing higher rates of gas evolution without external energy requirements and the possibility of converting a wide range of biomass-based substrates into hydrogen. [FeFe]-hydrogenases are key enzymes present in these microorganisms and are responsible for major bioproduction of molecular hydrogen. They are found in diverse organisms, including bacteria, anaerobic archaea, rhizobia,

protozoa, fungi and some green algae. Several efforts are currently underway to understand how their active sites are assembled, and to improve the development of hydrogenase analogs in renewable energy applications [58-61].

[FeFe]-hydrogenases are usually found in monomeric form in most bacteria, in which they oxidize reduced-ferredoxin for hydrogen production through multiple Fe-S clusters. Fe-S clusters are known for their role as electron transfer chains in the oxidation-reduction reactions. However, an alternative pathway was recently proposed for *Thermotoga maritima*, in which a trimeric bifurcating hydrogenase simultaneously oxidizes reduced ferredoxin and NADH under low partial hydrogen pressure [62].

EC 1.12.1.4: $5H^+$ + NADH + 2 Fd-reduced = $2H_2$+ NAD^+ +2 Fd-Oxidized

The cytoplasmic [FeFe] hydrogenase from *T.maritima* does not use either Fd or NADH as the sole electron donor. In this pathway, the oxidation of feredoxin and NADH is coupled *in vivo* to H_2 production by hydrogenase. Previous genome sequence analysis suggested the catalytic subunit TM1424 is part of the operon that encodes heterotrimetric hydrogenase (TM1424-TM1426), while roles of other subunits and synergistic mechanism of hydrogen reduction and ferredoxin oxidation remained unclear. In this study, we applied multiple state of the art structural modeling tools to establish a molecular model of the trimeric complex. The structural analysis identified a catalytic site in the subunit TM1424 and hydrogen production site in TM1426. Our trimeric complex model illustrates a novel type of hydrogenase electron transfer chain which

provides efficient catalysis of H_2 production in which both NADH and Ferredoxin serve as electron donors.

METHODS

Structure Modeling

The tertiary structures were modeled using I-TASSER and SPRED [63]. I-TASSER is a popular computer program for protein tertiary structure prediction from protein sequence. It first generates structural models for various sequence fragments of a given protein sequence using a threading approach against a library of 3D structures of short peptides generated from experimentally solved 3D structures. Then the structural models for the sequence fragments are used to assemble full-length models preserving the sequential order of the sequence fragments through energy minimization by using replica-exchange Monte Carlo simulation. The final model is selected among models with the lowest energy and then further refined by using atomic level refinement.

Complex Structure Docking

Two computer programs are used to dock the three component structures into one trimeric complex, namely a semi-rigid body docking program ZDOCK [64] and a template-based docking program SPRING [65]. ZDock predicts a docked structural model between two protein structures through optimizing a combination of three energy terms, namely desolvation energy, grid-based shape complementarity (GSC) and electrostatics energy by using Fast Fourier Transform (FFT). Component proteins are treated as rigid bodies, and rotational and translational transformation between the two

structures are fully examined and scored using an energy function. The models with the lowest energy are selected as the final prediction. In comparison, SPRING is a template-based method for protein-protein complex structure prediction. It predicts a complex structure of two protein sequences through identifying the optimal threading of each of the two sequences onto the each of the complexed structures found in the PDB database, like a protein threading prediction but against a complex structure with two sequences. In general, ZDOCK exhaustively scans through the conformational space defined by all possible translations and rotation between two rigid structures to form a complex structure for energy minimum models, while SPRING searches through all known oligomer structures in PDB to find the optimal complex structural templates for the two input protein sequences.

RESULTS

A Structure Model of Trimeric [Fd-only] Hydrogenase Complex

The tertiary structure of each subunit of the trimeric hydrogenase complex (TM1424, TM1425 and TM1426) was modeled *in silico* by I-TASSER [66]. I-TASSER applies template-based threading followed by fragment assembly in an iterative manner for protein structure prediction. The aligned threading templates of each of the three subunits in the [Fe-Fe] hydrogenase complex have the sequence identity > 40%, coverage > 0.7 and alignment Z-score > 7 (Table S1), where Z-score is used for normalization of the alignment threading score in I-TASSER. It has been shown that alignments with – scores > 3 represent good threading alignments [63]. High Z-scores for all threading alignments against different template structures typically suggest that identified

homologous templates tend to share similar structural folds with query protein. The predicted models with high Z-scores were then refined through energy minimization by using replica-exchange Monte Carlo simulation. Table 4.1 summarizes the prediction accuracies of the medium resolution models (2-5Å RMSD) among all the predicted structures.

Table 4.1. Estimated reliability of predicted models by I-TASSER for each subunit in [Fe-Fe] hydrogenase complex.

	Alignment Coverage	**Threading Z-score**	**Est RMSD**
TM1424	0.99	7.62	1.7 ± 1.5 Å
TM1425	0.79	9.18	2.6 ± 1.9 Å
TM1426	0.85	9.28	2.9 ± 2.1 Å

TM1426 – alpha subunit

The alpha subunit (TM1426) is the largest subunit in the trimeric complex (645 amino acids), and represents the [Fe-Fe] hydrogenase subunit in the complex, based on our finding that it is a homologue of the [Fe-Fe] hydrogenase in *Clostridium pasteurianum* (PDB ID: 3C8Y). The Z-score of the predicted model by I-TASSER is 8.96, and its structural accuracy level is predicted at RMSD 2.9±2.1Å after refinement. Similar to all known structures of the other [Fe-Fe] hydrogenases, the overall structure of the alpha subunit can be divided into two non-overlapping structural domains, a catalytic domain (residue 1 to 208) and a [Fe-S]-cluster domain (residue 209 to 645).

The catalytic domain contains a unique active site at the C-terminus, known as the H-cluster in the [Fe-only] hydrogenase. The H-cluster domain contains conserved cysteine residues involved in the coordination of active site in all known [Fe-Fe] hydrogenases. This is the only hydrogenase active site found in all three units (TM1424-TM1426) of the trimeric complex and it is predicted to be responsible for catalyzing the hydrogen production in the EC 1.12.1.4 enzymatic function. The structure of the catalytic domain consists of two twisted beta sheets, each with four strands and flanked by a number of alpha helices forming two nearly identical lobe-like structures, with one beta sheet and associated helices contained in one lobe. The active-site, H-cluster, is located at the interface between the two lobes near the interaction site with the adjacent domain. The H-cluster consists of a [2Fe] center bridged to a [4Fe-4S] cubane [67], which is coordinated by four conserved cysteines Cys^{294}, Cys^{295}, Cys^{350} and Cys^{486} (Fig 1).

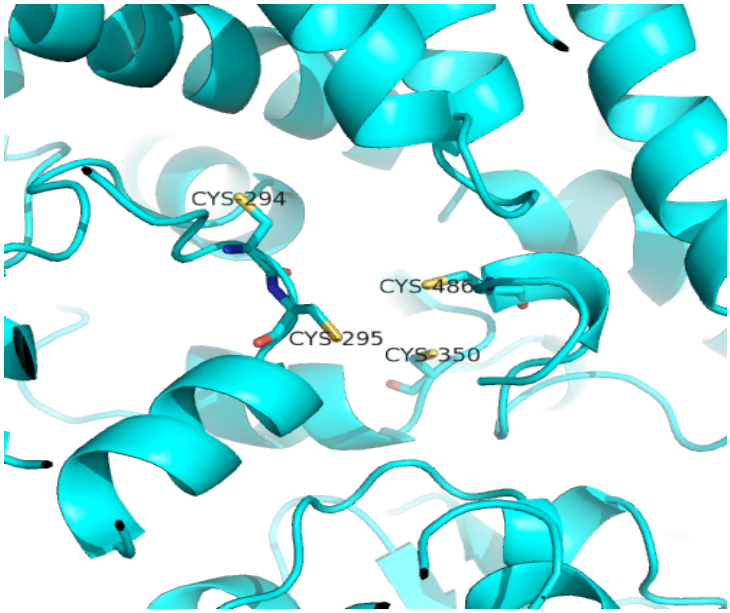

Figure 4.1. The active site H-cluster coordinated by four conserved cysteines in TM1426.

The [Fe-S] cluster domain consists of three [4Fe-4S] clusters and one [2Fe-2S] cluster, each coordinated by four conserved cysteine residues putatively responsible for the binding with a [Fe-S] cluster (Table 1); and each such domain is immediately adjacent to the catalytic domain. Site-directed mutagenesis analysis in a monomeric [Fe-Fe] hydrogenase revealed that all these conserved cysteines are essential to the maturation and activation of the enzyme [68]. Generally, [Fe-S] clusters are known for their role in electron transfer in the redox metabolism [67], which requires two consecutive [Fe-S] clusters at most 10 Å apart since otherwise electron transfer will not take place. A total of six [Fe-S] clusters were found within the trimeric [Fe-Fe]-hydrogenase complex, which may represent a novel pathway for electron transfer leading to the oxidation of NADH and production of hydrogen synergistically. It will be elaborated in the following sections.

Table 4.2. Alignment of Fe-S clusters.

	Type	Conserved Residues	ID
TM1424	2Fe-2S	Cys81, Cys86, Cys122, Cys126	FS1
TM1425	4Fe-4S	Cys485, Cys488, Cys491, Cys531	FS2
TM1426	2Fe-2S	Cys34, Cys45, Cys48, Cys60	FS3
	4Fe-4S	Cys143, Cys146, Cys149, Cys196	FS4
	4Fe-4S	His92, Cys96, Cys99, Cys105	FS5
	4Fe-4S	Cys153, Cys186, Cys189, Cys192	FS6

TM1425 – beta subunit

The beta unit, TM1425, of the complex is found to be homologous to the NADH binding subunit in NADH:ubiquinone oxidoreductase in *Thermus thermophilus* (PDB ID: 3I9V_B), with sequence identity 45%, threading alignment Z-score 9.18 and the resolution of the I-TASSER predicted structure at RMSD 2.6Å.

This subunit contains a conserved NADH-binding site and a [4Fe-4S] cluster FS2, which is predicted to be responsible for transferring an electron from NADH via the catalytic reaction (EC 1.12.1.4). The NADH binding site contains conserved residues for a flavin mononucleotide (FMN) bound NADH binding site. Functional analyses of the trimeric hydrogenase in a previous study [62] found that the synergistic production of hydrogen was only observed with the presence of FMN, indicating the essential role of FMN in the catalytic reaction. We noted that the cavity structure and all the essential residues for the NADH binding site in the NADH:ubiquinone oxidoreductase are conserved in TM1425. Within this solvent-exposed NADH binding cavity, the Glu^{315}/Glu^{316} locations make hydrogen bonds to the ribose of the adenosine moiety; residues 196 to 201, forming a glycine-rich loop, can bind to the phosphate groups of the substrate while the aromatic rings of Phe^{337} and $Phe^{210,}$ near the entrance to the cavity, are so positioned (8.5Å apart) to surround an adenine ring by side chains through aromatic stacking interactions. All these indicate that the cavity can accommodate one NADH/FMN molecule, which is validated by the minimum energy model using ligand-docking tool AutoDock [69]. The conserved residues that are putatively interacting with NADH/FMN are shown in Figure 4.2.

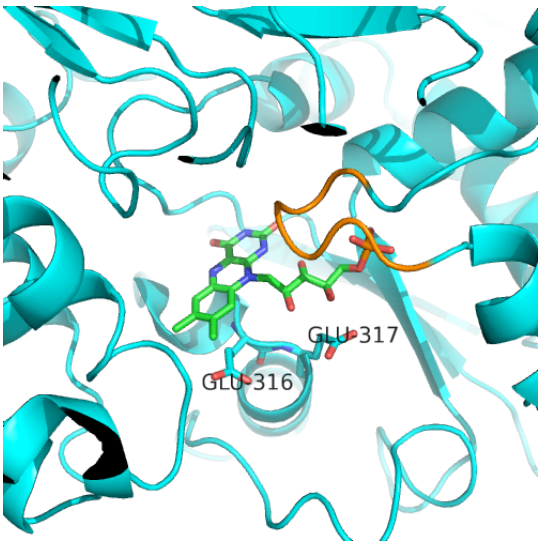

Figure 4.2. NADH/FMN cavity structure in TM1425 with key residues labels. Glycine-rich loop is marked in orange.

In addition, a [4Fe-4S] cluster is also found in TM1425, around 10Å away from the NADH binding site. The [4Fe-4S] cluster is coordinated by four cysteines consistent with the [4Fe-4S] motif $CX_2CX_2CX_{48\text{-}42}C$. The [4Fe-4S] cluster is coordinated in the cubane geometry by conserved cysteines Cys^{485}, Cys^{488}, Cys^{491} and Cys^{531} (Table 1). The first three cysteines are on the loop between the helices of the N-terminal helix bundle and the last one on the loop between adjacent helices.

TM1424 – gamma subunit

The gamma subunit (TM1424) is the smallest subunit (164 amino acids) in the trimeric complex, having the highest estimated accuracy (RMSD ~1.7Å) from structure modeling. The top structure template used for homology modeling is the [2Fe-2S] ferredoxin subunit in NADH:ubiquinone oxidoreductase from *Theremus thermophilus* [PDB ID: 3I9V_A], with sequence identity of 39% and threading Z-score 7.62. The C-

terminal domain consists of a mixed beta sheet flanked by two alpha helices and contains a conserved binding site with four conserved cysteines (cysteine^{81}, cysteine^{86}, cysteine^{122} and cysteine^{126}) that coordinate the [2Fe-2S] cluster.

Trimeric Complex Structure

It has been reported [62] that all three genes, TM1424, TM1425 and TM1426, are required for the enzymatic function (EC 1.12.1.4). However, the molecular mechanism by which the trimeric [Fe-Fe]-hydrogenase catalyzes the hydrogen production has not been fully elucidated. Hence, we have built a complex structure consisting of the three proteins to propose a model the electron flow during the oxidation of NADH and production of hydrogen based on the detailed structural features of the model.

Both semi-rigid body docking (ZDOCK[64]) and template-based docking (SPRING [65]) methods were used to predict the tertiary structure of the complex model, which has a conformation similar to that of the NADH:ubiquinone oxidoreductase complex as predicted earlier, where TM1424 and TM1425 correspond to chains A and B in the NADH:ubiquinone oxidoreductase complex (Figure 4.3). Since TM1425 and TM1424 are homologous to oxidoreductase subunits as described in previous sections, it is no surprise that the optimal binding interface retained for these two subunits. Interestingly, TM1426 replaced chain C in NADH:ubiquinone oxidoreductase in the complex conformation even though these two proteins share low sequence identity (27%) and low structural similarity (RMSD 15.47Å). To understand how the NADH:ubiquinone subunits might interact with Fd hydroganse, we applied sub-structure alignment using SPRED. Surprisingly, it revealed that both proteins contain a [Fe-S] cluster domain on

the binding interface (light green domain in Figure 4.3), which have sequence identity at 49% and RMSD 1.56Å between their sequences and structures, respectively. It suggests that through the interaction with the partial complex of oxidoreductase, [Fe-Fe] hydrogenase not only utilizes electron from ferredoxin in its monomeric form, but also acquired the ability to transfer electron from NADH for H_2 production as well.

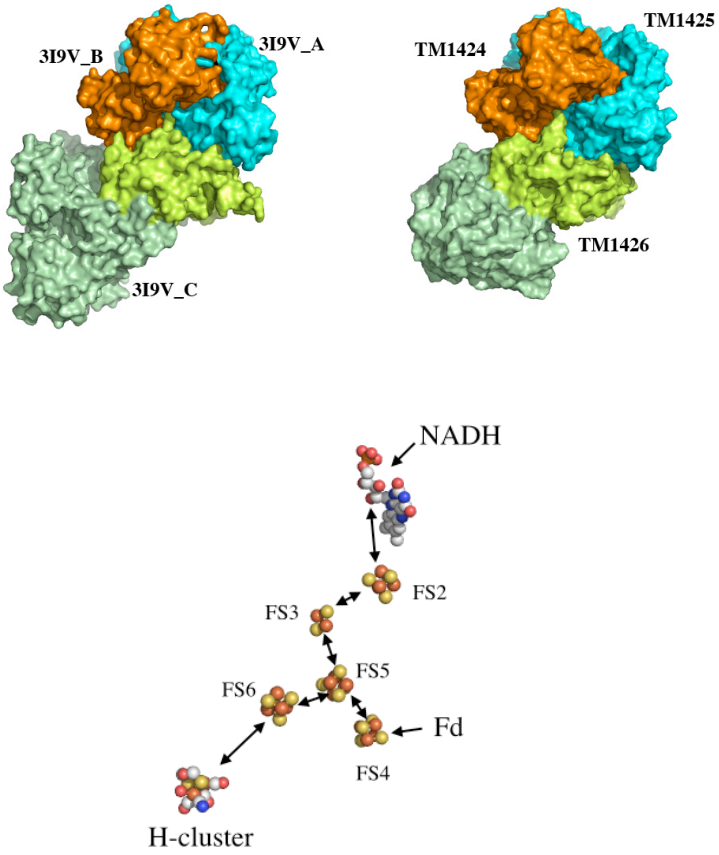

Figure 4.3. Comparison between complex structures of NADH:ubiquinone oxidoreductase (left) and the [Fe-Fe] hydrogenase (right). The light green domains in TM1426/3I9V_C are the conserved [Fe-S] cluster domains at the protein interaction interface. The bottom of the figure shows the proposed electron transfer chain.

It is well known that [Fe-S] clusters play significant role in oxidative-reductive reactions by acting as media for electron transfer. Six [Fe-S] clusters (Table 2) have been identified in the trimeric [Fe-Fe] hydrogenase complex with within 10Å between two consecutive such clusters (Figure 4.3). These [Fe-S] clusters combining with catalytic sites represent a novel electron transfer chain for enzymatic reaction EC 1.12.1.4, in which [Fe-Fe] hydrogenase utilizes electrons from both NADH and ferredoxin to produce hydrogen synergistically.

DISCUSSION

[Fe-Fe]-hydrogenases are key enzymes in anaerobic bacteria, responsible for biohydrogen production. We established a structural model to investigate the [Fe-Fe] hydrogenase complex, which has a novel enzymatic function but its molecular mechanism is unclear. From our structural model, a trimeric [Fe-Fe] hydrogenase complex in *T. maratima* oxidizes NADH and ferredoxin by interaction between the monomeric [Fe-Fe] hydrogenase and the NADH binding subunit as in NADH:ubiquinone oxidoreductase. A high number of [Fe-S] clusters and conserved cysteine patterns were found within the catalytic and accessory domains, which enables efficient interactions with numerous redox partners. The proposed interaction reveals an alternative mechanism to recycle NADH and reduced ferredoxin in these microorganisms.

To determine if such a mechanism also exists in other bacteria, extensive genomic analyses were performed to find homologous trimeric [Fe-Fe] hydrogenase in all sequenced bacteria genomes in NCBI genome database [70]. Specifically, a hidden Markov model (HMM) is built for the three genes and used to search against all the

bacterial genomes through HMM profile alignment [71], which has led to the identification of 23 bacterial species containing all three homologs within the same operon based on the DOOR operon database [72]. Since genes in the same operon are likely transcribed in a single mRNA molecule, the stoichiometry among the three component proteins should be guaranteed when forming the trimeric complex. In addition, examination of the upstream promoter sequences of these operons showed that 11 organisms contain a highly conserved palindromic motif in their promoter regions (Table 4.3, Figure 4.4). This suggests that not only the novel hydrogen-production function is conserved, but the regulatory mechanism is also conserved across this group of anaerobic organisms. Unfortunately, no transcription factor was found in the existing database that has the similar motif.

Table 4.3. Predicted operons that contain genes encoding the NADH/Fd [Fe-Fe] hydrogenase in bacteria along with a predicted cis regulatory motif.

Organism	Genes	Motif Location	Upstream Motif
Caldicellulosiruptor bescii	Athe_1295-Athe_1299	NC_012034: 1385662	TGTTAAATTTCTAACA
Caldicellulosiruptor hydrothermalis	Calhy_1427-Calhy_1431	NC_014652: 1462959	TGTTAAATTTCTAACA
Caldicellulosiruptor kristjanssonii	Calkr_1280-Calkr_1283	NC_014721: 1337008	TGTTAAATTTCTAACA
Caldicellulosiruptor owensensis OL	Calow_1086-Calow_1088	NC_014657: 1172486	TGTTAAATTTCTAACA
Clostridium clariflavum DSM 19732	Clocl_1309-Clocl_1311	NC_016627: 1514902	TGTTAATTTGTTAACA
Caldicellulosiruptor obsidiansis	COB47_1258-COB47_1261	NC_014392: 1382154	TGTTAAATTTCTAACA
Clostridium phytofermentans	Cphy_3801-Cphy_3804	NC_010001: 4665022	TGATAGTTTTTTAACA
Caldicellulosiruptor saccharolyticus	Csac_1860-Csac_1864	NC_009437: 2003859	TGTTAAATTTCTAACA
Clostridium thermocellum	Cthe_0338-Cthe_0342	NC_009012: 426900	TGTTAAATTGTTAACA
Mahella australiensis 50-1 BON	Mahau_1250-Mahau_1252	NC_015520: 1305260	TGTTAAAATCATAACG
Thermotoga maritima	TM1424-TM1426	NC_000853: 1435446	TGTGAAGTAGGTAACA

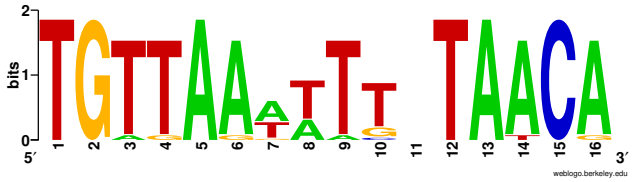

Figure 4.4. Conserved palindromic motif for NADH/Fd hydrogenase.

From the modeled complex structure, we were able to predict and characterize the binding pockets and the active sites in the trimeric [Fe-Fe] hydrogenase for electron transfer. These results provide a deeper and new knowledge of the molecular mechanism driving the [Fe-Fe] hydrogenase and a novel perspective of interactions between the [Fe-S] cluster-containing proteins in these bacteria under different conditions. Future experiments are needed to evaluate the roles of conversed residues in the process of [Fe-Fe] hydrogenase maturation for both the monomeric and trimeric forms, and their relationship to application on industrial hydrogen bioproduction.

CHAPTER 5

CONCLUSION

A Bioinformatics approach in conjunction with experimental evidence provides a powerful combination to carry out research. We have applied different techniques and approaches to confront some of challenging issues in protein structure research. A wide range of questions were addressed here.

Optimal PRE mutation sites can be selected by lipid accessibility area prediction of residues on trans-membrane helix. Using a stochastic distance matrix algorithm, we can correctly predict the topology of four helices in DsbB using minimum one PRE label with experimental data. In benchmark simulation, minimum two labels are sufficient with an average success rate of 76%. It provides a useful approach to deriving starting models for membrane proteins using a small number of experimental data.

SPRED, a new segment identification and assembly based threading algorithm for protein structure prediction, improves the accuracy of alignments compared to the best predictions of whole chain threading methods. In addition, it provides a flexible framework to incorporate threading tools as a meta-server and requires substantially less computational resource compared to the traditional fragment-based *de novo* methods.

Novel complex of ferredoxin hydrogenase in *Thermotoga maritima* suggests the [FeFe] hydrogenase utilizes electrons from NADH by interaction with homologous NADH catalytic subunit from NADH oxidoreducates. Comparative genomic analysis

shows such mechanism is retained in multiple anaerobic species with conserved regulatory transcription factor. The discovery gives a new perspective on our understanding of the redox proteins and mechanism of H_2 production in anaerobic bacteria.

As presented here, the bioinformatics approaches can be used to study various types of biological problem, and it will lead to a better understanding of protein structures and functions.

REFERENCES

1. Luscombe, N.M., D. Greenbaum, and M. Gerstein, *What is bioinformatics? A proposed definition and overview of the field.* Methods Inf Med, 2001. **40**(4): p. 346-58.
2. Yildirim, M.A., et al., *Drug-target network.* Nat Biotechnol, 2007. **25**(10): p. 1119-26.
3. Bespalov, M.M. and M. Saarma, *GDNF family receptor complexes are emerging drug targets.* Trends Pharmacol Sci, 2007. **28**(2): p. 68-74.
4. Tang, X.L., et al., *Orphan G protein-coupled receptors (GPCRs): biological functions and potential drug targets.* Acta Pharmacol Sin, 2012. **33**(3): p. 363-71.
5. Pieper, U., et al., *Coordinating the impact of structural genomics on the human alpha-helical transmembrane proteome.* Nat Struct Mol Biol, 2013. **20**(2): p. 135-8.
6. Shoichet, B.K. and B.K. Kobilka, *Structure-based drug screening for G-protein-coupled receptors.* Trends Pharmacol Sci, 2012. **33**(5): p. 268-72.
7. Almen, M.S., et al., *Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin.* BMC Biol, 2009. **7**: p. 50.
8. Sussman, J.L., et al., *Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules.* Acta Crystallogr D Biol Crystallogr, 1998. **54**(Pt 6 Pt 1): p. 1078-84.
9. Wiener, M.C., *A pedestrian guide to membrane protein crystallization.* Methods, 2004. **34**(3): p. 364-72.
10. Tronin, A.Y., et al., *Direct evidence of conformational changes associated with voltage gating in a voltage sensor protein by time-resolved X-ray/neutron interferometry.* Langmuir, 2014. **30**(16): p. 4784-96.
11. Morales-Rios, E., et al., *Structure of ATP synthase from Paracoccus denitrificans determined by X-ray crystallography at 4.0 A resolution.* Proc Natl Acad Sci U S A, 2015. **112**(43): p. 13231-6.
12. Wang, S., et al., *Solid-state NMR spectroscopy structure determination of a lipid-embedded heptahelical membrane protein.* Nat Methods, 2013. **10**(10): p. 1007-12.
13. Zhou, Y., et al., *NMR solution structure of the integral membrane enzyme DsbB: functional insights into DsbB-catalyzed disulfide bond formation.* Mol Cell, 2008. **31**(6): p. 896-908.
14. Van Horn, W.D., et al., *Solution nuclear magnetic resonance structure of membrane-integral diacylglycerol kinase.* Science, 2009. **324**(5935): p. 1726-9.
15. Hong, M., Y. Zhang, and F. Hu, *Membrane protein structure and dynamics from NMR spectroscopy.* Annu Rev Phys Chem, 2012. **63**: p. 1-24.

16. Goldbourt, A., *Biomolecular magic-angle spinning solid-state NMR: recent methods and applications.* Curr Opin Biotechnol, 2013. **24**(4): p. 705-15.
17. Gayen, S., Q. Li, and C. Kang, *Solution NMR study of the transmembrane domain of single-span membrane proteins: opportunities and strategies.* Curr Protein Pept Sci, 2012. **13**(6): p. 585-600.
18. Barth, P., B. Wallner, and D. Baker, *Prediction of membrane protein structures with complex topologies using limited constraints.* Proc Natl Acad Sci U S A, 2009. **106**(5): p. 1409-14.
19. Toukach, F.V. and V.P. Ananikov, *Recent advances in computational predictions of NMR parameters for the structure elucidation of carbohydrates: methods and limitations.* Chem Soc Rev, 2013. **42**(21): p. 8376-415.
20. Bahrami, A., et al., *Robust, integrated computational control of NMR experiments to achieve optimal assignment by ADAPT-NMR.* PLoS One, 2012. **7**(3): p. e33173.
21. Barth, P., J. Schonbrun, and D. Baker, *Toward high-resolution prediction and design of transmembrane helical protein structures.* Proc Natl Acad Sci U S A, 2007. **104**(40): p. 15682-7.
22. Inaba, K., et al., *Crystal structure of the DsbB-DsbA complex reveals a mechanism of disulfide bond generation.* Cell, 2006. **127**(4): p. 789-801.
23. Bertini, I., et al., *NMR spectroscopy of paramagnetic metalloproteins.* Chembiochem, 2005. **6**(9): p. 1536-49.
24. Wang, S., et al., *Paramagnetic relaxation enhancement reveals oligomerization interface of a membrane protein.* J Am Chem Soc, 2012. **134**(41): p. 16995-8.
25. Gottstein, D., et al., *Requirements on paramagnetic relaxation enhancement data for membrane protein structure determination by NMR.* Structure, 2012. **20**(6): p. 1019-27.
26. Clore, G.M., C. Tang, and J. Iwahara, *Elucidating transient macromolecular interactions using paramagnetic relaxation enhancement.* Curr Opin Struct Biol, 2007. **17**(5): p. 603-16.
27. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.* J Mol Biol, 2001. **305**(3): p. 567-80.
28. Agrafiotis, D.K., *Stochastic proximity embedding.* J Comput Chem, 2003. **24**(10): p. 1215-21.
29. Case, D.A., et al., *The Amber biomolecular simulation programs.* J Comput Chem, 2005. **26**(16): p. 1668-88.
30. Zhou, F., V. Olman, and Y. Xu, *Barcodes for genomes and applications.* BMC Bioinformatics, 2008. **9**: p. 546.
31. Zhang, Y. and J. Skolnick, *Automated structure prediction of weakly homologous proteins on a genomic scale.* Proc Natl Acad Sci U S A, 2004. **101**(20): p. 7594-9.
32. Walters, R.F. and W.F. DeGrado, *Helix-packing motifs in membrane proteins.* Proc Natl Acad Sci U S A, 2006. **103**(37): p. 13658-63.
33. Harris, N.L., S.R. Presnell, and F.E. Cohen, *Four helix bundle diversity in globular proteins.* J Mol Biol, 1994. **236**(5): p. 1356-68.
34. 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.

35. Lomize, M.A., et al., *OPM: orientations of proteins in membranes database.* Bioinformatics, 2006. **22**(5): p. 623-5.
36. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.
37. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**(4): p. 702-10.
38. Yuan, Z., et al., *Predicting the solvent accessibility of transmembrane residues from protein sequence.* J Proteome Res, 2006. **5**(5): p. 1063-70.
39. Fleishman, S.J. and N. Ben-Tal, *Progress in structure prediction of alpha-helical membrane proteins.* Curr Opin Struct Biol, 2006. **16**(4): p. 496-504.
40. Kolodny, R., et al., *On the universe of protein folds.* Annu Rev Biophys, 2013. **42**: p. 559-82.
41. Levitt, M., *Nature of the protein universe.* Proc Natl Acad Sci U S A, 2009. **106**(27): p. 11079-84.
42. Zhang, Y., et al., *On the origin and highly likely completeness of single-domain protein structures.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(8): p. 2605-10.
43. Fox, N.K., S.E. Brenner, and J.M. Chandonia, *SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures.* Nucleic Acids Res, 2014. **42**(1): p. D304-9.
44. Sillitoe, I., et al., *New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures.* Nucleic Acids Res, 2013. **41**(Database issue): p. D490-8.
45. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nat Methods, 2012. **9**(2): p. 173-5.
46. Hildebrand, A., et al., *Fast and accurate automatic structure prediction with HHpred.* Proteins, 2009. **77 Suppl 9**: p. 128-32.
47. Shi, J., T.L. Blundell, and K. Mizuguchi, *FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties.* J Mol Biol, 2001. **310**(1): p. 243-57.
48. Kelley, L.A., R.M. MacCallum, and M.J. Sternberg, *Enhanced genome annotation using structural profiles in the program 3D-PSSM.* J Mol Biol, 2000. **299**(2): p. 499-520.
49. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.
50. Karplus, K., C. Barrett, and R. Hughey, *Hidden Markov models for detecting remote protein homologies.* Bioinformatics, 1998. **14**(10): p. 846-56.
51. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-63.
52. Cheng, J. and P. Baldi, *A machine learning information retrieval approach to protein fold recognition.* Bioinformatics, 2006. **22**(12): p. 1456-63.
53. Xu, J., et al., *RAPTOR: optimal protein threading by linear programming.* J Bioinform Comput Biol, 2003. **1**(1): p. 95-117.
54. McGuffin, L.J. and D.T. Jones, *Improvement of the GenTHREADER method for genomic fold recognition.* Bioinformatics, 2003. **19**(7): p. 874-81.

55. Xu, Y. and D. Xu, *Protein threading using PROSPECT: design and evaluation.* Proteins, 2000. **40**(3): p. 343-54.
56. Jones, D.T., *GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences.* J Mol Biol, 1999. **287**(4): p. 797-815.
57. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
58. Nicolet, Y., et al., *A novel FeS cluster in Fe-only hydrogenases.* Trends Biochem Sci, 2000. **25**(3): p. 138-43.
59. Drescher, T., et al., *Sulfide catalysis without coordinatively unsaturated sites: hydrogenation, cis-trans isomerization, and H2/D2 scrambling over MoS2 and WS2.* J Am Chem Soc, 2012. **134**(46): p. 18896-9.
60. Bui, E.T. and P.J. Johnson, *Identification and characterization of [Fe]-hydrogenases in the hydrogenosome of Trichomonas vaginalis.* Mol Biochem Parasitol, 1996. **76**(1-2): p. 305-10.
61. Beer, L.L., et al., *Engineering algae for biohydrogen and biofuel production.* Curr Opin Biotechnol, 2009. **20**(3): p. 264-71.
62. Schut, G.J. and M.W. Adams, *The iron-hydrogenase of Thermotoga maritima utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production.* J Bacteriol, 2009. **191**(13): p. 4451-7.
63. Yang, J., et al., *The I-TASSER Suite: protein structure and function prediction.* Nat Methods, 2015. **12**(1): p. 7-8.
64. Chen, R., L. Li, and Z. Weng, *ZDOCK: an initial-stage protein-docking algorithm.* Proteins, 2003. **52**(1): p. 80-7.
65. Guerler, A., B. Govindarajoo, and Y. Zhang, *Mapping monomeric threading to protein-protein structure prediction.* J Chem Inf Model, 2013. **53**(3): p. 717-25.
66. Zhang, Y., *I-TASSER server for protein 3D structure prediction.* BMC Bioinformatics, 2008. **9**: p. 40.
67. Peters, J.W., et al., *X-ray crystal structure of the Fe-only hydrogenase (CpI) from Clostridium pasteurianum to 1.8 angstrom resolution.* Science, 1998. **282**(5395): p. 1853-8.
68. King, P.W., et al., *Functional studies of [FeFe] hydrogenase maturation in an Escherichia coli biosynthetic system.* J Bacteriol, 2006. **188**(6): p. 2163-72.
69. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.
70. Tatusova, T., et al., *RefSeq microbial genomes database: new representation and annotation strategy.* Nucleic Acids Res, 2015. **43**(7): p. 3872.
71. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.
72. Mao, F., et al., *DOOR: a database for prokaryotic operons.* Nucleic Acids Res, 2009. **37**(Database issue): p. D459-63.