

Assessment of PRISM with previous CAPRI rounds

by

Efe ELBEYLİ

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of
Masters of Science
in
Biomedical Sciences and Engineering



November 16, 2017

Assessment of PRISM with previous CAPRI rounds

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's dissertation by

Efe ELBEYLİ

and have found that it is complete and satisfactory in all respects,

and that any and all revisions required by the final

examining committee have been made.

Committee Members:

Prof. Özlem Keskin (Advisor)

Prof. Attila Gürsoy (Co-Advisor)

Prof. Burak Erman

Prof. İ. Halil Kavaklı

Assist. Prof. S. Ece Özbabacan

Date _____



That's one giant step for [a] man, one small leap for mankind.

ABSTRACT

Proteins are key elements of a cell to perform the wide range of molecular and cellular activity. Proteins perform their function through binding to other proteins, DNA, RNA, and small molecules. Therefore, predicting how a protein interacts with its binding partners is one of the most important objectives of structural biology. As a result of the improvements in experimental structure determination methods, the number of individual protein structures in PDB has increased vastly. However, the number of complex structures does not increase as fast as the individual proteins which creates demand for new approaches to predict complexes from the individual protein structures. Currently, there are a lot of computational approaches to predict the complex structures. Critical Assessment of Prediction of Interactions (CAPRI) is a well-known community-wide experiment with the purpose of establishing a routine which allows testing the performance of several different docking algorithms created. The success rate of predicted structures is verified by the several evaluation criteria determined by the CAPRI association. Those criteria are interface and ligand rmsds (I-rmsd, L-rmsd), native residue contacts and number of clashes. In this study, I assessed the performance of PRISM (Protein Interaction by Structural Matching) using CAPRI evaluation criteria. The main objectives were to determine how much PRISM is successful in predicting the complex structures of available CAPRI targets and to force the limits of PRISM by stretching the parameters of PRISM. PRISM could not predict the correct complex structures for 33% of targets majority of which correspond to homodimers and enzyme/inhibitor complexes. The results also indicate that considering just the structures with the negative energy score results in the loss of 36% of successful predictions implying a problem in the scoring function. For further testing, RosettaDock was used as an alternative scoring function. Both scoring approaches yielded a correlation. To increase the success rate, some parameters of PRISM were changed however, no significant improvement has been achieved.

ÖZETÇE

Proteinler hücre içerisindeki çeşitli moleküler ve hücresel aktivitelerin gerçekleşmesinde rol alan kilit elemanlardan birisidir. Proteinler görevlerini gerçekleştirmek için diğer proteinlerle, DNA, RNA ve diğer küçük moleküllere bağlanırlar. Bu sebeple, bir proteinin partnerine nasıl bağlandığı konusunu incelemek yapısal biyolojinin en temel görevlerinden biri olmuştur. Deneysel yöntemlerin protein yapısını elde etmekte gösterdiği gelişme sonucu PDB’de bulunan tekil protein yapılarının sayısı yüksek bir artış göstermiştir. Ancak, karmaşık protein yapılarının sayısı tekil protein yapılarının sayısı kadar hızlı bir artış göstermemektedir ki bu durum tekil proteinlerden yola çıkarak karmaşık protein yapılarını tahmin eden yöntemlere yönelik bir talep oluşturmuştur. Günümüzde karmaşık protein yapılarını tahmin eden çok sayıda hesaplamalı yöntem bulunmaktadır. Protein etkileşimlerinin tahmin eden algoritmaların performansını eleştirel olarak inceleyen CAPRI herkese açık gerçekleştirilen rutin bir deneydir. Tahmin edilmiş yapılar, CAPRI grubunun belirlediği birtakım değerlendirme koşullarına tabi tutularak doğrulanır. Bu koşullar arayüz rmsd (I-rmsd), ligand rmsd (L-rmsd), doğal aminoasit temaslarının sayısı ve çarpışan aminoasit sayılarından oluşmaktadır. Bu çalışmada ben, PRISM performansını önceki CAPRI turlarını kullanarak değerlendirdim. PRISM’in CAPRI turlarında göstereceği performansı değerlendirmek ve PRISM’in parametrelerini esneterek performansının sınırlarını görmek bu çalışmanın asıl amaçlarıdır. PRISM kullanılan hedeflerin %33’ü için doğru yapı tahmini yapamamıştır ki bu hedeflerin çoğunluğunu eş-ikililer ve enzim/inhibitor çiftleri oluşturmaktadır. Sonuçlar göstermektedir ki, sadece negatif enerjili yapıları değerlendirmeye almak doğru sonuçların %36’sının kaybolmasına sebep olmaktadır ve yapıların puanlamasında önemli sorunlara işaret etmektedir. Değerlendirmenin devamında RosettaDock alternatif puanlama için kullanılmıştır. Her iki puanlama yönteminin birbiriyle denk sonuçlar verdiği görülmektedir. PRISM’in başarı oranını artırmak amacıyla bazı parametreler değiştirilmiş, ancak kayda değer bir performans artışı gözlemlenmemiştir.

ACKNOWLEDGEMENT

I would like to thank all the people who contributed in some way to the work described in this thesis. First and foremost, I would like to express my sincere gratitude to my advisors Prof. Özlem Keskin and Prof. Attila Gürsoy for their continuous support, patience, and motivation. Their guidance helped me in all the time of research and writing of this thesis. One simply could not wish for friendlier supervisors.

I appreciate the feedback offered by my thesis jury members Prof. Burak Erman, Prof. İ. Halil Kavaklı and Assist. Prof. S. Ece Özbabacan. Advice and comments given by them have been a great help in the progress of my studies.

Every result described in this thesis was accomplished with the help and support of fellow officemates. Thus, I would like to thank to former and current COSBI lab members: Emine Güven-Maiorov, Güray Kuzu, Engin Çukuroğlu, Nilay Karahan, Serena Muratcioğlu, Farideh Halakou, Doğuş Doğru, A. Derya Cavga, E. Sıla Özdemir, Hande Güneş, Deniz Demircioğlu, Alper Başpınar, Emel Şen Kılıç, Cemal Yamak, Bilgesu Erdoğan, Emre Büyüktuna, Hazal Karadağ, Mahan Rajaei Golsefidi, A. Tuğrul Balcı, and Ayşe S. Yazgılı who have been great friends to me.

We were so close, so tight and powerful enough to declare our own republic on the second floor for the past two years. I would like to thank to all members of Solidarity Between Offices (OFAD): Bakış Doğru, Houman Bahmani Jalali, Rustamzhon Melikov, Nazlı Ataç, Ömer Kırnap, Aysima Hacısüleyman, Anıl Fettahoğlu and Özgün Can Önder for being perfect comrades on our action.

Some people touches your life more than others. Bilgesu Erdoğan, Serena Muratcioğlu, Bakış and Doğuş Doğru and of course Laviş (the cat) were being there whenever I need. I can say that I have an academic family from now.

I could not forget my true F.R.I.E.N.D.S Selena Kazancıoğlu, Erinç Bulça, Gürkan Ağyüz, Ece Bayram, Burcu Kara, Selin Akan, Samet Şentürk and Alican Koluman. My life would not be good enough without those people in it.

Last but not least, I owe my deepest gratitude to my family especially my parents, my brother, my grandmother and my uncle. Their love, understanding and emotional support keep me focused on the path I want to go through. Their endless encouragement made me able to pursue this research. Thank you for making me feel competent and confident. Thank you for always being there whenever I need. I have always felt so lucky and proud that I have such wonderful and loving family.

I also want to thank my family for funding my research and for their contribution to computational science.

TABLE OF CONTENTS

LIST OF FIGURES.....	X
LIST OF TABLES	XIV
NOMENCLATURE.....	XVI
CHAPTER 1.....	1
INTRODUCTION.....	1
CHAPTER 2	4
LITERATURE REVIEW	4
2.1 STRUCTURE BASED PROTEIN DOCKING.....	4
2.1.1 <i>Current Docking Algorithms</i>	7
2.1.1 <i>Current Scoring Methods</i>	10
2.2 BACKGROUND INFORMATION ABOUT PRISM	11
2.2.1 <i>Insights of Interface</i>	12
2.2.2 <i>Creation of Non-redundant Template Data Set</i>	14
2.2.3 <i>PRISM Prediction Algorithm</i>	15
2.3 CRITICAL ASSESSMENT OF PREDICTION OF INTERACTIONS (CAPRI)	18
2.3.1 <i>CAPRI Evaluation Protocol</i>	18
CHAPTER 3.....	22
MATERIALS AND METHODS	22
3.1 TARGET DATASET	23
3.2 PRISM AS DOCKING ALGORITHM.....	27
3.3 ROSETTA AS ALTERNATIVE SCORING ALGORITHM	28
3.3.1 <i>Rosetta Prepack Protocol (for local docking)</i>	28
3.3.2 <i>Rosetta Docking Protocol (for local docking)</i>	28
3.4 LIGAND RMSD (L-RMSD) CALCULATION	29
3.5 INTERFACE RMSD (I-RMSD) CALCULATION	30
3.6 NATIVE AND NON-NATIVE RESIDUE FREQUENCY (F_{NAT} & $F_{\text{NON-NAT}}$) CALCULATION.....	31
CHAPTER 4.....	33
RESULTS AND DISCUSSION	33

4.1 DETAILED RESULTS AND DISCUSSION FOR EACH TARGET	37
4.1.1. Target 01: Hpr – HprK/P	37
4.1.2. Target 02: Rotavirus VP6 – Fab	38
4.1.3. Target 03: Influenza Hemagglutinin – Fab HC63	39
4.1.4. Target 04, 05, and 06: α -Amylase – Camelide Antibody VH Domains 1,2,3	40
4.1.5. Target 07: T-cell receptor β -chain - Pyrogenic Exotoxin A.....	43
4.1.6. Target 08: Laminin – Nidogen	44
4.1.7. Target 09: LicT Homodimer.....	45
4.1.8. Target 10: Trimeric form of TBEV Envelope	47
4.1.9. Target 11 and 12: Cohesin – Dockerin	49
4.1.10. Target 13: Major Surface Antigen P30 - SAG1 Fab.....	51
4.1.11. Target 14: Protein Phosphatase 1 – Myosin Phosphatase Targeting Subunit 1	53
4.1.12. Target 15: Colicin D - ImmD	54
4.1.13. Target 16: Family 10 Xylanase – XIP-1 Xylanase Inhibitor.....	55
4.1.14. Target 17: Family 11 Xylanase – XIP-1 Xylanase Inhibitor.....	57
4.1.15. Target 18: Family 11 Xylanase – TAXI-I.....	58
4.1.16. Target 19: Ovine Prion – Fab.....	60
4.1.17. Target 20: HemK – RF1	61
4.1.18. Target 21: Orc1 – Sir1	63
4.1.19. Target 22: Spliceosomal Protein (U5-15K) - Cytoplasmic Domain Binding Protein (U5-52K) (CD2BP2).....	65
4.1.20. Target 23: hGBP1 homodimer	66
4.1.21. Target 24 and 25: Arf1 – ARHGap10	68
4.1.22. Target 26: TolB - Pal.....	70
4.1.23. Target 27: Hip2 - Ubc9	71
4.1.24. Target 28: NEDD4 Catalytic Domain Homodimer.....	72
4.1.25. Target 29: tRNA Methyltransferase 8 - tRNA Methyltransferase 82	74
4.1.26. Target 30: Rho Family GTPase Rnd1 – Plexin-B1	75
4.1.27. Target 32: Barley- α -Amylase Subtilis Inhibitor – Subtilis Savinase.....	76
4.1.28. Target 35 and 36: CBM22-1 of Xylanase – GH10 of Xylanase	78
4.1.29. Target 37: Leucine Zipper 2 – ADP-ribosylation Factor (ARF6).....	80
4.1.30. Target 38 – 39: Centurin- α 1 – FHA domain of KIF13B.....	81
4.1.31. Target 40: Bovine Trypsin - APIA.....	83
4.1.32. Target 41: Colicin E9 – Immunity Protein 2.....	85
4.1.33. Target 42: Tetrcopeptide Repeat (TPR) homodimer	86
4.1.34. Target 46: Methyltransferase MtrA – Trm112	88

4.1.35. Target 47: Colicin E2 Immunity Protein – Colicin E9.....	89
4.1.36. Target 50: Hemagglutinin Receptor Binding Domain and Membrane Fusion Domain – Hypothetical Protein.....	90
4.2. EXAMINING THE EFFECT OF MONTE CARLO MINIMIZATION ON FIBERDOCK SCORING	92
4.3. COMPARISON OF RESULTS FROM PRISM SCORING WITH FIBERDOCK AND SCORING WITH ROSETTADOCK	94
4.3.1 FiberDock Scoring Function.....	94
4.3.2 RosettaDock Scoring Function.....	95
CHAPTER 5 – NEW PARAMETERS	97
CHAPTER 6 – CONCLUSION.....	102
BIBLIOGRAPHY	104

List of Figures

Figure 2.1. Basic representation of bound/bound docking.	5
Figure 2.2. The basic representation of rationale of PRISM.	12
Figure 2.3. Basic representation of interface (green) between two proteins (blue and red).	13
Figure 2.4: Formulation of root-mean square deviation.	19
Figure 2.5: Basic representations of CAPRI evaluation criteria.....	21
Figure 3.0.1. Overall route and individual steps followed for this study.....	23
Figure 3.1.2. CAPRI .brk file format and information on a starting file.	27
Figure 4.1.1. Best model of docking prediction for Target 01, HPr - HPrK/P complex. Blue: HprK/P; red: crystal HPr; green: predicted HPr.....	38
Figure 4.1.2. Best model of docking prediction for Target 02, Rotavirus VP6 / Fab complex. Blue: Rotavirus VP6; red: native Fab; green: predicted Fab.	39
Figure 4.1.3. Best model of docking prediction for Target 03, Hemagglutinin / Fab HC63 complex. Blue: Hemagglutinin; red: native Fab; green: predicted Fab.	40
Figure 4.1.4. Best models of docking prediction for Target 04, 05 and 06 on the same α -amylase. α -Amylase / Camelide Antibody VH Domains 1,2,3 complex. Blue: α -Amylase; red: native VH Domains; green: predicted VH domains.....	42
Figure 4.1.5. Best model of docking prediction for Target 07, T-cell receptor β -chain / Pyrogenic Exotoxin A complex. Blue: T-cell receptor; red: native exotoxin; green: predicted exotoxin.....	44
Figure 4.1.6. Best model of docking prediction for Target 08, Laminin / Nidogen complex. Blue: Nidogen; red: native laminin; green: predicted laminin.....	45
Figure 4.1.7. Best model of docking prediction for Target 09, LicT homodimer complex. Blue: LicT monomer; red: native second LicT monomer; green: predicted second LicT monomer.	47

Figure 4.1.8. Best model of docking prediction for Target 10, TBEV homodimer complex. Blue: two TBEV monomers in trimer; red: native TBEV monomer; green: predicted TBEV monomer.....	49
Figure 4.1.9. Best model of docking prediction for Target 11 (left) and Target 12 (right), Cohesin / Dockerin complex. Blue: Cohesin; red: native dockerin; green: predicted dockerin.....	51
Figure 4.1.10. Best model of docking prediction for Target 13, Major Surface Antigen P30 / SAG1 Fab complex. Blue: SAG1 Fab; red: native MSA P30; green: predicted MSA P30.....	52
Figure 4.1.11. Best model of docking prediction for Target 14, Protein Phosphatase 1 / Myosin Phosphatase Targeting Subunit 1 complex. Blue: MPTS1; red: native Phosphatase 1; green: predicted Phosphatase 1.	54
Figure 4.1.12. Best model of docking prediction for Target 15, Colicin D - ImmD complex. Blue: Colicin D; red: native ImmD; green: predicted ImmD.	55
Figure 4.1.13. Best model of docking prediction for Target 16, Family 10 Xylanase – XIP-1 Xylanase Inhibitor complex. Blue: XIP-1; red: native family 10 xylanase; green: predicted family 10 xylanase.	56
Figure 4.1.14. Best model of docking prediction for Target 17, Family 11 Xylanase – XIP-1 Xylanase Inhibitor complex. Blue: XIP-1; red: native family 11 xylanase; green: predicted family 11 xylanase.	58
Figure 4.1.15. Best model of docking prediction for Target 18, Family 11 Xylanase / TAXI-1 complex. Blue: TAXI-1; red: native family 11 xylanase; green: predicted family 11 xylanase.	60
Figure 4.1.16. Best model of docking prediction for Target 19, Ovine Prion / Fab complex. Blue Fab; red: native Ovine Prion; green: predicted Ovine Prion.	61
Figure 4.1.17. Best model of docking prediction for Target 20, HemK / RF1 complex. Blue RF1; red: native HemK; green: predicted HemK.....	63
Figure 4.1.18. Best model of docking prediction for Target 21, Orc1 / Sir1 complex. Blue Orc1; red: native Sir1; green: predicted Sir1.....	64

Figure 4.1.19. Best model of docking prediction for Target 22, Spliceosomal Protein / Cytoplasmic Domain Binding Protein complex. Blue Spliceosomal Protein; red: native Cytoplasmic Domain Binding Protein; green: predicted Cytoplasmic Domain Binding Protein.	66
Figure 4.1.20. Best model of docking prediction for Target 23, hGBP1 homodimer complex. Blue: hGBP1 monomer; red: native hGBP1 monomer from complex; green: predicted hGBP1 monomer.....	67
Figure 4.1.21 Best model of docking prediction for Target 24 (left) and Target 25 (right), Arf1 – ARHGap10 complex. Blue: Arf1; red: native ARHGap10; green: predicted ARHGap10.....	69
Figure 4.1.22. Best model of docking prediction for Target 26, TolB / Pal complex. Blue: TolB; red: native Pal; green: predicted Pal.	71
Figure 4.1.23. Best model of docking prediction for Target 27, Hip2 / Ubc9 complex. Blue: Hip2; red: native Ubc9; green: predicted Ubc9.	72
Figure 4.1.24. Best model of docking prediction for and Target 28, NEDD4L Catalytic Domain Homodimer complex. Blue: NEDD4L monomer; red: native NEDD4L monomer in complex; green: predicted NEDD4L monomer.	73
Figure 4.1.25. Best model of docking prediction for Target 29, tRNA Methyltransferase 8 / tRNA Methyltransferase 82 complex. Blue: tRNA Methyltransferase 82; red: native tRNA Methyltransferase 8; green: predicted tRNA Methyltransferase 8.....	75
Figure 4.1.26. Best model of docking prediction for Target 30, Rho Family GTPase Rnd1 / Plexin-B1 complex. Blue: Rnd1; red: native Plexin-B1; green: predicted Plexin-B1.....	76
Figure 4.1.27. Best model of docking prediction for Target 32, Barley- α -Amylase Subtilis Inhibitor / Subtilis Savinase complex. Blue: Subtilis Savinase; red: native Subtilis Inhibitor; green: predicted Subtilis Inhibitor.	77
Figure 4.1.28. Best model of docking prediction for Target 35 (left) and Target 36 (right), CBM22-1 of Xylanase – GH10 of Xylanase complex. Blue: GH10; red: native CBM22; green: predicted CBM22.....	79

Figure 4.1.29. Best model of docking prediction for Target 37, Leucine Zipper 2 / ADP-ribonylation Factor (ARF6) complex. Blue: ARF6; red: native Leucine Zipper 2; green: predicted Leucine Zipper 2.	81
Figure 4.1.30. Best model of docking prediction for Target 38 (left) and Target 39 (right), Centurin- α 1 / FHA domain of KIF13B complex. Blue: Centurin- α 1; red: native KIF13B; green: predicted KIF13B.	83
Figure 4.1.31. Best model of docking prediction for Target 40, Bovine Trypsin / APIA complex. Blue: Bovine Trypsin; red: native APIA; green: predicted APIA.	84
Figure 4.1.32. Best model of docking prediction for Target 41, Colicin E9 – Immunity Protein 2 complex. Blue: Colicin E9; red: native Immunity Protein 2; green: predicted Immunity Protein 2.	86
Figure 4.1.33 Best model of docking prediction for Target 42, Tetrcopeptide Repeat (TPR) homodimer complex. Blue: TPR; red: second dimer of native TPR; green: second dimer of predicted TPR.....	87
Figure 4.1.34. Best model of docking prediction for Target 46, Mtq2 / Trm112 complex. Blue: Trm112; red: native Mtq2; green: predicted Mtq2.....	89
Figure 4.1.35. Best model of docking prediction for Target 47, Colicin E2 / Colicin E9 complex. Blue: Colicin E9; red: native Colicin E2; green: predicted Colicin E2.	90
Figure 4.1.36. Best model of docking prediction for Target 50, HA Receptor Binding Domain and HA Membrane Fusion Domain / Hypothetical Protein complex. Blue: HA domains; red: native Hypothetical Protein; green: predicted Hypothetical Protein	92
Figure 4.2.1. The graph shows energy value comparison between MC minimization equals 0 and MC minimization equals 50. Black dots are responsible for the negatively affected models with increased energy scores after MC minimization. Red dots are responsible for the models with lower energy scores after MC minimization.....	93
Figure 4.3.1: FiberDock scoring function formulation.....	94
Figure 4.3.2. The graph shows energy value comparison between FiberDock Energy score and Rosetta I _{sc} score.....	95

List of Tables

Table 2.1. Summary of CAPRI criteria for ranking predicted complexes [61].	19
Table 3.1. All CAPRI targets with protein names and final PDBs.	24
Table 4.1. Summary of PRISM Performance in CAPRI. Results for best model on each target are given. Classification column gives the number for each category in the order of high/medium/acceptable/incorrect.	33
Table 4.2. Targets, used interfaces, search results on PIFACE and model quality.	35
Table 4.1.1. Evaluation Statistics of all predicted models for CAPRI Target 01.	37
Table 4.1.2. Evaluation Statistics of all predicted models for CAPRI Target 02.	39
Table 4.1.3. Evaluation Statistics of all predicted models for CAPRI Target 03.	40
Table 4.1.4. Evaluation Statistics of all predicted models for CAPRI Targets 4, 5 and 6.	42
Table 4.1.5. Evaluation Statistics of all predicted models for CAPRI Target 7.	43
Table 4.1.6. Evaluation Statistics of all predicted models for CAPRI Target 8.	45
Table 4.1.7. Evaluation Statistics of all predicted models for CAPRI Target 9.	46
Table 4.1.8. Evaluation Statistics of all predicted models for CAPRI Target 10.	48
Table 4.1.9. Evaluation Statistics of all predicted models for CAPRI Targets 11 and 12.	50
Table 4.1.10. Evaluation Statistics of all predicted models for CAPRI Target 13.	52
Table 4.1.11. Evaluation Statistics of all predicted models for CAPRI Target 14.	53
Table 4.1.12. Evaluation Statistics of all predicted models for CAPRI Target 15.	55
Table 4.1.13. Evaluation Statistics of all predicted models for CAPRI Target 16.	56
Table 4.1.14. Evaluation Statistics of all predicted models for CAPRI Target 17.	57
Table 4.1.15. Evaluation Statistics of all predicted models for CAPRI Target 18.	59
Table 4.1.16. Evaluation Statistics of all predicted models for CAPRI Target 19.	61
Table 4.1.17. Evaluation Statistics of all predicted models for CAPRI Target 20.	62
Table 4.1.18. Evaluation Statistics of all predicted models for CAPRI Target 21.	64

Table 4.1.19. Evaluation Statistics of all predicted models for CAPRI Target 22.	65
Table 4.1.20. Evaluation Statistics of all predicted models for CAPRI Target 23.	67
Table 4.1.21.1. Evaluation Statistics of all predicted models for CAPRI Target 24.	69
Table 4.1.21.2. Evaluation Statistics of all predicted models for CAPRI Target 25.	69
Table 4.1.22. Evaluation Statistics of all predicted models for CAPRI Target 26.	70
Table 4.1.23. Evaluation Statistics of all predicted models for CAPRI Target 27.	72
Table 4.1.24. Evaluation Statistics of all predicted models for CAPRI Target 28.	73
Table 4.1.25. Evaluation Statistics of all predicted models for CAPRI Target 29.	74
Table 4.26. Evaluation Statistics of all predicted models for CAPRI Target 30.	76
Table 4.1.27. Evaluation Statistics of all predicted models for CAPRI Target 32.	77
Table 4.1.28. Evaluation Statistics of all predicted models for CAPRI Targets 35 and 36.	79
Table 4.1.29. Evaluation Statistics of all predicted models for CAPRI Target 37.	80
Table 4.1.30. Evaluation Statistics of all predicted models for CAPRI Targets 38 and 39.	82
Table 4.1.31. Evaluation Statistics of all predicted models for CAPRI Target 40.	84
Table 4.1.32. Evaluation Statistics of all predicted models for CAPRI Target 41.	85
Table 4.1.33. Evaluation Statistics of all predicted models for CAPRI Target 42.	87
Table 4.1.34. Evaluation Statistics of all predicted models for CAPRI Target 46.	88
Table 4.1.35. Evaluation Statistics of all predicted models for CAPRI Target 47.	90
Table 4.1.36. Evaluation Statistics of all predicted models for CAPRI Target 50.	91
Table 5.1. The comparison of same models predicted by different parameters for targets 40, 41, 46 and 47. (E/I : Enzyme/Inhibitor complex, A/A: Antibody/Antigen complex).	99
Table 5.2. The comparison of the best models predicted by different parameters for targets 09, 10, 23 and 28. (E/I: Enzyme/Inhibitor complex, A/A: Antibody/Antigen complex).	101

Nomenclature

APIAs	Arrowhead Protease Inhibitors
Arf1	ADP-Ribosylation Factor 1
ASP	Astex Statistical Potential
ACE	Atomic Contact Energy
BAH	Bromo-Adjacent Homology
BASI	Barley- α -Amylase Subtilis Inhibitor
CAPRI	Critical Assessment of Prediction of Interactions
CASP	Critical Assessment of Methods of Protein Structure Prediction
CD2BP2	CD2 Receptor Binding Protein
FFT	Fast Fourier Transform
GA	Genetic Algorithm
GAP	GTPase-Activating Protein
GH10	Family 10 of Glycosyl Hydrolases
GH11	Family 11 of Glycosyl Hydrolases
HprK/P	HPr kinase / phosphorylase
HA	Hemagglutinin
hGBP1	Human Guanylated Binding Protein 1
Im	Immunity proteins
I-rmsd	Interface Root-mean Square Deviation
L-rmsd	Ligand Root-mean Square Deviation
MC	Monte Carlo

MD	Molecular Dynamics
MHC	Major Histocompatibility Complex
MM/GBSA	Molecular Mechanics / Generalized-Born Surface Area
MM/PBSA	Molecular Mechanics / Poisson-Boltzmann Surface Area
MYPT1	Myosin Phosphatase Targeting Subunit 1
NMA	Normal Mode Analysis
NMR	Nuclear Magnetic Resonance
Orc	Origin Recognition Complex
PAL	Peptidoglycan-Associated Lipoprotein)
PDB	Protein Data Bank
PP1	Protein Phosphatase 1
PPI	Protein-Protein Interaction
PRISM	Protein Interaction by Structural Matching
RF1	Release Factor 1
RNP	Ribonucleoprotein
RSA	Relative Surface Accessibilities
RMSD	Root-mean Square Deviation
SAGs	Superantigens
Sir1	Silent Information Regulator 1
TAXI-I	Triticum Aestivum Xylanase Inhibitor-I
TBEV	Tick-borne Encephalitis Virus Envelope
TCR	T-cell Receptor
TPR	Tetratricopeptide repeat

Trm112	tRNA methyltransferase 112
Ubc9	Ubiquitin-Conjugating Enzyme 9
XIP-I	Xylanase Inhibitor Protein I



Chapter 1

Introduction

Proteins are key elements of a cell to perform the wide range of molecular and cellular activity. A protein is barely functional by itself. Interaction of a protein with other proteins, DNA, RNAs, and other small molecules are very common for cellular metabolic and developmental activities [1]. Therefore predicting the protein's correct interaction is one of the most important objectives of structural biology. We focused mostly on interactions between two proteins. Current docking methods perform the protein-protein interaction predictions in two ways: template-free and template-based [2].

PRISM (Protein Interaction by Structural Matching) [3] is a template-based docking tool to predict the correct complex form of two proteins by using the interface templates constructed from the interfaces of well-known complex structures found in literature. PRISM has unique characteristics of performing docking according to both geometric complementarity and flexible refinement. In this study, an extensive assessment of PRISM is done. Large dataset of individual proteins and their complex forms are used for the assessment to test the accuracy of algorithm and to identify the strongest and weakest features of PRISM. The dataset is taken from the previous rounds of well-known organization called Critical Assessment of Prediction of Interactions (CAPRI) [4]. Thus, the performance of PRISM could be compared with the performances of other prediction algorithms. Other than docking, scoring is another important step for PRISM to demonstrate the best results it predicts. In this study, the scoring performance of PRISM is tested and compared with different versions of itself and different scoring functions currently used by other prediction algorithms.

Critical Assessment of Prediction of Interactions (CAPRI) is a well-known community-wide experiment with the purpose of establishing a routine which allows

testing the performance of several different docking algorithms created by the structural biologists. CAPRI has been organizing yearly rounds since 2001. The targets for every round are selected from unpublished complex proteins with available three-dimensional structures of component proteins. Difficulties of rounds may vary to test the accuracy of algorithms in several cases. The success rate of predicted structures is verified by the several evaluation criteria determined by the CAPRI association. Those criteria are interface and ligand rmsds (I-rmsd, L-rmsd), native residue contacts and number of clashes which are globally accepted and necessary to be calculated by the researchers who want to participate in that experiment.

PRISM is a complex algorithm which invokes sub-algorithms for different steps. Each sub-algorithm has its own parameters which could be modified for special uses. The prediction depends on every single parameter which are defined for general uses. In this study, the assessment covers testing the effect of change in parameters over the prediction accuracy of PRISM. Three versions of PRISM is created and the results are compared with the results coming from original PRISM.

The outline of the thesis is arranged as follows:

In Chapter 2, an extensive literature review is included starting with the history of structure-based protein docking. This section starts with explaining the logic behind the protein docking algorithms. Afterward, the section continues by the current successful algorithms and explaining their capabilities. Then, it provides detailed background information about the template-based docking methods especially about PRISM. Finally, it explains the CAPRI organization, its purpose and the evaluation criteria determined by this association which is used for determination of the success of predicted protein complexes.

In Chapter 3 the methodology is presented in detail. The path followed in this thesis is given in a scheme. All target dataset is presented as a table with protein names and PDB IDs. The way targets are selected, the details of starting files are explained in

the first part. Afterward, small modifications on the PRISM are explained. FiberDock [5] scoring function is explained. As an alternative scoring, Rosetta Local Refinement [6] is explained in detailed manner. Then, the types of criteria selected for the evaluation of a predicted complex structure are explained. Each evaluation criteria is described with their rules and restrictions. The codes developed for the calculation of CAPRI evaluation criteria are expressed in a detailed manner.

Chapter 4 presents the detailed analysis of results. The accuracies of predictions are compared. The amounts of successful and failed predictions are expressed. Detailed background information about every individual protein in each target is explained. Statistical analysis of PRISM is performed to see the types of protein pairs which algorithm predicts successfully or unsuccessfully. Then the comparison of scoring functions, FiberDock (in two versions) and Rosetta, are expressed. Results are lined up with those two scoring functions. Thus, positive and negative points of scoring functions are compared in this section.

Chapter 5 presents the results and discussion of models coming from additional trials of PRISM with modifications over the parameters belonging to the algorithms PRISM invokes. The effects of new parameters over the performance of PRISM will be examined in detail. Also the performance of MULTIPROT [7] and FiberDock are indirectly tested with this work.

As a final chapter, Chapter 6 concludes the thesis study with the summary of results, final discussion and future works.

Chapter 2

Literature Review

Proteins are essential elements in a cell which involve almost any biological pathways varying from signaling to enzymatic reactions or from immune response to forming molecular machines like ribosomes. These biological pathways require interactions of proteins with other biological molecules to maintain regulatory and metabolic pathways that together constitute the process of life [1]. It has been known that protein structure mostly determines the protein function. Specific folding patterns give specific roles to a protein or to a domain of more complex proteins [8]. Most proteins consist of multiple domains which give a protein the ability to interact a wide range of many other molecules including other proteins, RNA, DNA, small-organic and metal-organic compounds [9][10, 11]. The binding mode between the protein and other biomolecule is of fundamental importance in modern structure-based studies such as drug design, functional proteomics, computational biology and structural biology.

2.1 Structure Based Protein Docking

Protein docking is a computational method to predict the formation of intermolecular complexes of proteins from the structures of individual proteins in three-dimensional space which basically is a process of ligand molecule placed into the special site of a protein [2][12]. The first docking methods were developed for protein interactions with other ligands at predetermined binding sites rather than protein-protein complexes. In the early 90s, implementing the technique called Fast Fourier Transform (known as FFT docking) from computer science provides a significant uptake in the development of protein docking techniques because FFT allows an exhaustive search of the full six-dimensional docking space in a computationally feasible manner [2].

On the beginnings of protein – protein docking, the steric complementarity was the only criterion. Most approaches were based on the simplistic assumption that two proteins are two rigid bodies. These approaches were useful for “bound docking problem” (Figure 2.1), which is re-docking two protein structures derived from the separation of their co-crystallized complex structures [2, 10]. Protein Data Bank (PDB) [13] has always been the major source of three-dimensional structures of proteins for structural biologists. Due to the improvements on X-ray crystallography [14] and nuclear magnetic resonance (NMR) [15] techniques, the number of individual proteins structures on PDB has increased vastly. Unfortunately, the number of complex protein structures remains lower, since it is experimentally unlikely to isolate all interactions of a single protein with many others. That shortage created demand for new approaches to be able to predict complexes from the separately crystallized protein structures [2, 13].

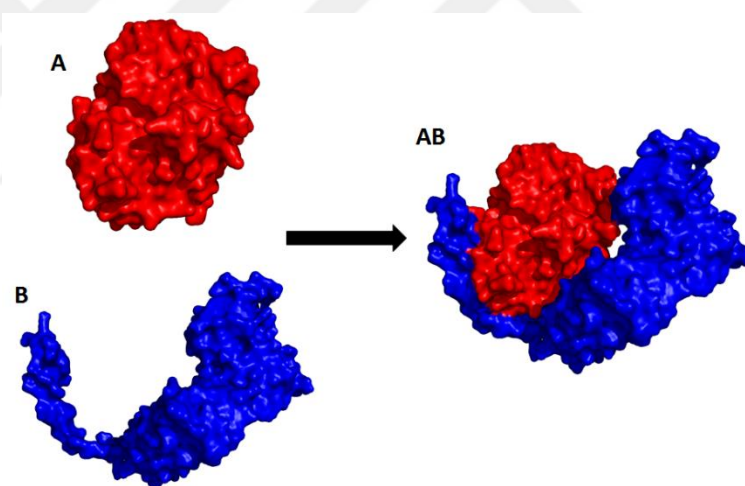


Figure 2.1. Basic representation of bound/bound docking.

The major challenges for protein-protein interaction prediction are dealing with flexibility and conformational changes [16]. Rigid-body assumption ignored one of the most common phenomena called induced fit. Induced fit is the property of a protein structure experiencing small conformational changes during the complex formation, especially on the contacting surface (interface) [10]. Addition of protein flexibility, for

both backbone and side-chain atoms, to the docking algorithm requires the calculation of high number of degrees of freedom by using normal mode analysis (NMA). Side-chain flexibility could be included by using rotamer libraries constructed with Molecular Dynamics (MD) simulations or statistical studies. Both of these calculations highly depend on the size of the proteins which significantly increases the running time and gives higher rate of false-positive results [17]. At present, there are dozens of docking tools approaching from different perspective to the protein flexibility issue. Many of the current programs generate ensembles of conformations allow ligand to be fully flexible while keeping the other component fixed. Some recent programs can perform “induce fit” with partial or complete conformational flexibility of the residues in protein binding site. Andrusier et al. concludes in the review that the current methods to combine side-chain and backbone optimization is still not to be fully optimal, and further work in this direction is required [17].

To eliminate false-positive predictions, scoring the predictions becomes important for a docking algorithm. Like all other force fields, scoring functions are based on a summation of independent terms that together they create binding affinity. For that reason, the general problem is suffering from size dependence of score. Summation favors larger molecules with higher scores [18]. A useful scoring function should distinguish the complex structures that most closely resemble to the experimental structures by calculating their binding free energies which are generally the sum of van der Waals and electrostatic interactions. Different scoring approaches could weight structural parameters differently to find the best fit with experimental constraints. Although they need to be improved, overall, Ferrara et al. concludes that the success rate of almost all current scoring functions can reach 80 - 90% [19].

Currently, docking approaches can be divided into two main classes of algorithms: template-free (docking) approaches and template-based approaches. Each approach has its own advantages or disadvantages [20]. Template-free methods are based on the physical properties of the component proteins for the prediction. Geometric

complementarity and conservation of matching hotspots creates the computational core of docking algorithms. These methods provide fast search algorithms with strong scoring functions which make these methods sensitive to conformational changes upon binding [20, 21]. In contrast, template-based methods use similarities with known complex structures for generating the prediction. By using high-quality non-redundant template dataset, known protein complexes are compared with target proteins in terms of structural or sequential similarities to reduce the solution space and help docking approaches by limiting the possible orientations. Main limitation for template-based approaches is availability of similar template to the target proteins to form the complex. If the template data set does not have the similar template, the complex structure cannot be predicted [21]. With the help of increasing rate of availability of new protein structures and complexes on PDB and constant updating of the template dataset, the effect of that limitation could be minimized.

2.1.1 Current Docking Algorithms

Glide [22, 23], Grid-Based Ligand Docking With Energetics, is an algorithm based on geometric search of ligand position, orientation and conformation on protein complex. Before the docking, Glide uses hierarchical filters with precomputed grids. Grids help to represent the shape and properties of the receptor in tandem with several different sets of fields. Binding site of the receptor is enclosed by the grid box centered according to the mass center of the ligand. The minimization of ligands is achieved by OPLS-AA force field and possible ligand conformations are subjected to MC to generate ligands with different torsional minima. Glide uses GlideScore function, a version of ChemScore [24] function, to re-score the models based on calculating solvation and repulsive interactions [25-27].

Surflex [28] is a docking algorithm starting with creation of idealized ligand structure called “promotol”. The ligand is fragmented into several small pieces with different rotatable bonds. Those individual fragments are incrementally placed into the active site of protein with a purpose of maximizing similarity to the promotol. Docking ends when

the ligand is fully reconstructed. For scoring, Surflex uses an empirical scoring function called Hammerhead [25-27].

GOLD [29], Genetic Optimization for Ligand Docking, is a genetic algorithm (GA) starting with the search for conformational and rotational flexibility of the ligand. GA adds some fixing points to the hydrogen binding group both on protein and ligand to optimize the ligand geometry and dihedrals of ligand and protein. For scoring, GOLD uses three different scoring functions ChemScore, GoldScore and Astex Statistical Potential (ASP) which are based on the molecular mechanics[26, 27, 30].

AutoDock [31] is also another tool uses genetic algorithm same as Gold. Differently, the algorithm generates poses of the ligand with Lamarckian version of GA inside a protein active site. The pose of ligand is gradually extended with the changes in conformations adopted by molecules after in situ optimization. Same as GOLD, the active side of protein is selected based on the native ligand position [26].

FRED divides the prediction process into two main steps: shape fitting and optimization. The algorithm uses 0.5-Å-resolution grid box to place the ligand and uses Gaussian potential to encompass all active site atoms including hydrogens. On the second step, FRED uses three different filters for optimization of predicted complex. The filters are for refining the hydrogen atom positions on the ligand, rigid body optimization and ligand pose optimization by changing the dihedral angles. FRED uses 4 different scoring functions but the ChemScore is the primary choice [25].

FlexX [32] is another fragment-based prediction tool. The first and most critical step is deciding the base fragment of ligand which creates the the principle interaction with the target protein. FlexX uses the active site of the protein for prediction. Torsion angle database is used to generate different poses of a fragment that was considered as rigid in further steps. FlexX uses its own scoring function which considers the formal charges during the pose scoring [26].

PatchDock [33] is a geometry-based algorithm aimed to find the transformations lead to the best shape complementarity between docking partners. The algorithm divides the surfaces of two components into the patches and superimposes those patches using shape matching algorithms. Finally, complementary patches are matched and candidate transformation is generated with avoiding steric clashes. PatchDock uses its own scoring function that considers atomic desolvation and geometric fit together [33].

ZDOCK [34] is FFT based method which performs global search in the rotational and translational space of ligand with respect to receptor. As a unique feature of ZDOCK, Atomic Contact Energy (ACE) is used to measure desolvation free energy which is basically the energy change of a protein-water contact to protein-protein contact. ZDOCK also allows the conformational changes on side chains of the surface residues. For scoring the complexes, ZDOCK considers electrostatic energy which is generated by the receptor and the charges of the ligand atoms. Partial charges are derived from CHARMM19 [35] potential. Shape complementarity, desolvation and electrostatic energy scores are considered together for every predicted complex [34].

HADDOCK [36], High Ambiguity Driven DOCKing, has three major steps on its docking protocol. The protocol starts with random rotations of two individual partner proteins located about 150Å away from each other to calculate minimum intermolecular energy. Then the docking is performed by translocations and rotations with rigid body energy minimization. The best models are selected and exposed to three simulated annealing refinements. At first, two proteins are considered as rigid-bodies and only the orientation is optimized. Then, only the side-chains at interface residues are allowed to move. And last, both side-chains and backbone atoms at the interface are allowed to move for conformational rearrangements. HADDOCK uses its own energy function which calculates the average buried surface area and average interaction energies including electrostatic energies and van der Waals energies [36].

RosettaDock [37] has an algorithm that mimics the physical process of protein interactions. The algorithm performs the combination of low-resolution stage involving recognition process with Monte Carlo search of translations and rotations and high-resolution (atomic-scale) stage involving simultaneous optimization of side-chain positions and rigid-body displacement. RosettaDock has a comprehensive scoring function which considers physical, chemical and statistical potentials. The score is dominated by packing interactions, solvation, rotamer probability, hydrogen bonding, and repulsive van der Waals interactions and includes contributions for electrostatics [37].

2.1.1 Current Scoring Methods

Commonly, prediction algorithms are designed to produce complexes as much as possible to avoid losing a chance to predict near-native complex form of protein and the ligand. Thus, it becomes necessary to use fast and reliable scoring function that can estimate the strength of complex. Basically, a scoring function tries to mimic the experimental binding affinities of complexes [38]. The binding affinity depends both on geometric matching of the two interacting molecules and strength of the interaction based on physicochemical parameters. The major problem of scoring functions is the estimation of overall binding energy as the sum of individual physicochemical parameters which creates large dependency between size of ligand and its score. Currently, the scoring functions could be grouped under three categories: (I) force-field based methods, (II) empirical methods, and (III) statistical (knowledge-based) methods [26].

Solvation-based methods are considered more reliable especially used together with current scoring functions. Under these approaches, the receptor flexibility and the presence of water molecules contribute to a more realistic description of the complex, which could have an influence in binding energy calculations. Methods such as MM/PBSA (Molecular Mechanics / Poisson-Boltzmann Surface Area) and MM/GBSA (Molecular Mechanics / Generalized-Born Surface Area) [39, 40], mimics the forces same as an implicit solvent in order to improve docking accuracy and correct scoring

function values. Also, Molecular Dynamics (MD) simulations are used to get several different conformation of the complex obtained by docking algorithms. Subsequent calculation of the binding energy by averaging the score values for different poses extracted from the trajectory could present more reliable value as a score [38].

In this study, we focused on one of the most powerful template-based docking (strictly speaking “interface-based docking”) algorithms called PRISM and its assessment. Also, because of its consideration of wide range of physicochemical parameters on scoring calculation, we select the scoring function of RosettaDock to compare the performance with PRISM’s scoring function.

2.2 Background Information about PRISM

PRISM is a hybrid protocol generating protein complexes from template interfaces considering geometric complementarity of proteins and evolutionary conservation of matching hotspots. The rationale of PRISM is based on a simple assumption: two protein structures “possibly interact” with each other through a particular region on their surfaces, if that region resembles the complementary partners of a known interface. In other words, If A and B is known to interact with each other, A' has similar region on its surface with the binding site of A , B' has similar region on its surface with the binding site of B , then we predict that A' interacts with B' [3, 41]. Basically, PRISM depends on structural similarity of template interfaces (known interfaces) to target surfaces [21]. Since the interface dataset is one of the most important elements on PRISM, the protein interfaces should be explained in more detailed way.

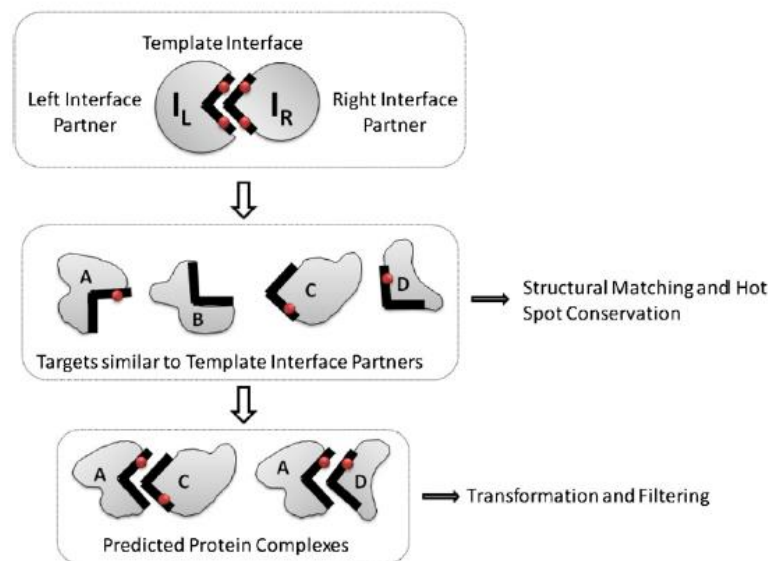


Figure 2.2. The basic representation of rationale of PRISM [21].

2.2.1 Insights of Interface

Interfaces are specific regions on the protein surface having much higher binding affinities than the rest of the surface. For a functional protein, the structure and the shape of interface region is crucial. Evolution has repeatedly utilized favorable, stable folds adapting them to a broad range of regulatory, enzymatic and packaging/structural roles [42]. That is why the interfaces were proven as evolutionary more conserved regions on proteins' surface. Choi et al. analyzed a large dataset of proteins and showed that among a total of 2646 non-antibody protein interfaces, 76.83% (2033 interface) were more conserved than the rest of the protein surface [43]. Interface-based approaches take advantage of this evolutionary conservation to predict the correct complex structure [44]. In fact, Zhao et al. showed that less than a half of sequence identity (45%) between template surface and modeled protein is enough for detecting more than 48% of the important atoms on the surface and for generating the interaction interface successfully [45]. Sufficiency of less than half of the interface region corresponds much less percentage of necessity for whole protein structure (depends on the size of protein). Therefore, the focuses of docking community have slide more on local alignments instead

of global alignment. Vakser et al. implied the necessity of template-based docking with partial structural alignments rather than full structure alignments [46].

Keskin et al. (2004) defines an interface to be the region between two polypeptide chains that are not covalently linked. The interface residues on a chain are selected according to the closeness to the residues on opposite chain. An interface consists of “interacting” residues and “nearby” residues [42]. Tsai et al (1996) creates two schemes to define interacting residues and nearby residues on an interface: (i) two residues, one from each chain, may be defined as interacting across the interface, if the distance between C^α atoms is less than 9 Å (ii) if the distance between any two atoms of two residues from different chains is less than the sum of their corresponding van der Waals radii plus 0.5 Å, these residues are interacting [47]. In PRISM, we use the second definition. “Nearby” residues are defined as a residue which has the distance between its C^α atom and the C^α atom of an interacting residue less than 5Å. Nearby residues are important to shape distinguishable architecture of interface and help aligning one interface scaffold against another [21].

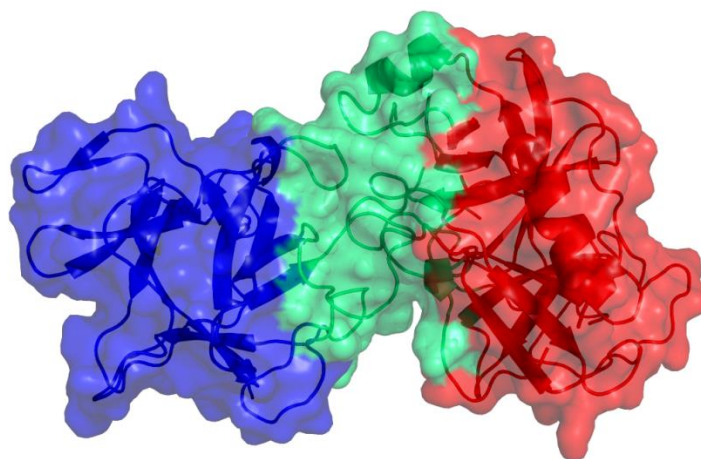


Figure 2.3. Basic representation of interface (green) between two proteins (blue and red).

The evolutionary conservation of certain residues at protein interfaces is another characteristic of binding sites [3]. Experimental researches over interface reveal that

interface residues of a protein do not contribute equally in binding. Clackson and Wells stated in 1995 that some residual contacts were more important than others. They have identified those residues, hot spots, by performing alanine scanning mutagenesis technique which is mutating a single one of interface residues to Alanine in each time and calculating the free energy change ($\Delta\Delta G$) [48]. More detailed researches revealed that hot spots are not randomly located residues on the interface; instead, they tend to cluster together near the center of interface [49]. Furthermore, hot spots are evolutionary conserved not only sequentially but also structurally [50, 51]. Due to experimentally identified hot spot information are limited, computational methods are needed to be developed for that service [50]. Current computational tools can predict hot spots with the success rate of 88% [52]. Lately, hot spots are also found to be buried and tightly packed with other residues resulting in densely packed clusters called “hot regions” [51]. These characteristics of interface should be considered on the creation of non-redundant template dataset.

2.2.2 Creation of Non-redundant Template Data Set

Interface consists of interacting residues and nearby residues that belong to two different chains. Thus, interfaces consist of pieces of each of the chain, and some isolated residues. Since there is no sequential order for interface residues, it is essential to carry out structural comparisons of the interfaces independent of their amino acid sequence order to generate a non-redundant dataset [42]. Geometric Hashing Paradigm is explained by Nussinov et al. as an efficient way of truly three-dimensional sequence-order-independent structural comparison [53]. Protein interfaces are represented by six-letter nomenclature. First four letters correspond to the PDB name and the following two letters correspond to chain identifiers. In other words, the interface between chain A and B of the PDB structure 1c1y is named 1c1yAB [54]. Creation of non-redundant template dataset has many steps and iterations. First, the similarity between any two structures is calculated. Next, two most similar structures are clustered to generate a set of clusters with selecting one of them as representative to that cluster. This procedure is iterated,

until the extent of similarity between the un-clustered structures and the cluster representative is below specific threshold. A single clustering cycle starts with forming a new cluster for the first entry in the initial list of interfaces. The next entry in the interface list is compared to the first. If the similarity between those two is above a predefined threshold, the second interface is added to the cluster of the first, or else it forms its own new cluster. Next, the third interface is compared with the cluster already formed. This procedure is repeated, until all structures are assigned to clusters. Six clustering cycles were performed to create more condensed dataset. At each cycle, the similarity threshold is gradually relaxed. This yields a hierarchy among group of clusters with different similarity levels. At the end of the final (sixth) iteration, the number of interface is reduced to 22,604 distinct clusters with at least 0.5 connectivity score [42]. To eliminate redundancy among interfaces in each cluster, 50% similarity threshold was imposed. Thus, one of the two sequences in a cluster is removed from cluster if they share sequence similarity of more than 50%. This yields a data set of interfaces with structurally similar but sequentially dissimilar members [42]. The members of the 3799 interface clusters were put through a filtering process which eliminated the redundant sequences from the clusters. A cluster was defined as non-redundant if it contained at least five non-homologous sequences [3].

2.2.3 PRISM Prediction Algorithm

PRISM constitutes four consecutive stages:

- (i) Surface extraction of target proteins,
- (ii) Structural alignment between target surface and template interface,
- (iii) Transformation and filtering,
- (iv) Flexible refinement and energy scoring.

2.2.3.1 Surface Extraction

As an interface-based approach, PRISM structurally compares the partners of the template interfaces with surface of target proteins. Considering only the surface of the

protein prevents inaccurate matching of templates to the core of the proteins [55]. The surface residues are found by calculating Relative Surface Accessibilities (RSA) of residues using the NACCESS program [56]. RSA is defined as percentage of accessibility of a residue of interest compared with the accessibility of that residue type in an extended ALA-X-ALA tripeptide [41]. NACCESS calculates the atomic accessible area with a probe rolling around the Van der Waal's surface of three-dimensional structure of the protein. The program uses the Lee & Richards' method [57], whereby a probe, generally picked as a water molecule, with 1.4Å radius is rolled around the surface of the molecule, and the path traced out by its center is called solvent accessible surface. Jones and Thornton [58] argued that residues with Relative Surface Accessibilities (RSA) more than 15% can be considered to be on the surface [41].

2.2.3.2 Rigid – Body Alignment

Each representative interface picked from the template dataset is split into its constituent partners. PRISM invokes MULTIPROT [7] for structural alignments of each partner with the target surface. MULTIPROT derives partial multiple alignments from simultaneous superposition of input molecules to detect common geometrical cores between given protein structures in a sequence-order-independent way [7, 41]. Since the protein interfaces have sequential discontinuity, sequence-order-independent alignment makes MULTIPROT a favorable selection for the task [41].

MULTIPROT returns 10 best sub-structural matches for every constituent partner of all interfaces in dataset. PRISM uses the best 3 matches. Best matches are ranked according to similarity scoring function which is $\alpha f_{\text{evolution}} + (1 - \alpha) f_{\text{structure}}$. $f_{\text{evolution}}$ represents evolutionary similarity scoring function which is the number of identically matched hot spots. $f_{\text{structure}}$ is structural similarity scoring functions which is the size and quality of the alignment along the target–template alignment. The coefficient α creates the relative importance between evolutionary similarity and structural similarity. PRISM gives hotspots greater importance in defining an interface than geometrical complementarity [41].

2.2.3.3 Transformation and Filtering

PRISM checks all structural matching results in that part. For passing to the next step, 30% of the residues of template chains should geometrically match the target surfaces. This threshold is 50% for template chains containing less than 50 residues. Such relatively small interfaces are likely to align perfectly with target surfaces and yield high similarity scores, causing biased and unselective results [54, 55]. In both cases, number of matching residues should be at least 15 to get rid of interfaces with small contact areas that could cause spurious matchings [41]. In addition to structural similarity, hot spot filtering incorporates evolutionary similarity between the target surface and template interface. If the template interface has computational hot spots, at least one hot spot in each template partner should correctly match with the target surface. The maximal RMSD allowed in the structural alignment is 2 Å. Thus, while the alignment is rigid, it can handle movements within 2 Å [55]. Finally, if two partners have (more than five) spatially colliding residues, the match is eliminated. Side-chain clashes are not considered at this stage [54].

2.2.3.4 Flexible Refinement

PRISM invokes FiberDock [5] for docking which considers both backbone and side-chain flexibility. Currently, the refinement algorithm models backbone flexibility in one of the proteins (the receptor) and side-chain flexibility in both of them. FiberDock uses iterative process to minimize the structural conformations of the interacting proteins and optimizes their rigid-body orientation. It starts with side-chain optimization modeled by a rotamer library to find optimum combinations of rotamers with lowest total energy. Then, it continues with backbone refinement based on normal mode analysis. The relevance of a mode is calculated according to its correlation with the chemical forces applied on each atom. Rigid-body Monte Carlo (MC) minimization is the last step of iteration cycle before scoring the solutions [5]. To score the solutions, FiberDock calculates energies by using CHARMM force field [54, 59].

2.3 Critical Assessment of Prediction of Interactions (CAPRI)

CAPRI [4] is a community - wide experiment to assess the capacity of protein docking methods to predict the protein - protein interactions. The first round of CAPRI is established in 2001. CAPRI was evolved from CASP (Critical Assessment of Methods of Protein Structure Prediction) which focuses on the prediction of 3D protein structure. Since the protein docking and structure prediction overlaps in many ways, the need for another CASP-like experiment, CAPRI, rises. CAPRI has already been a powerful drive for the community of computational biologists who development docking algorithms [60].

The blind docking is based on predicting the complex form of proteins without knowing the correct result. The three-dimensional structures of each component must be known independently but the complex structure should be soon-to-be-released which means there is no published three-dimensional structure for the complex form of components. Any chemical or biological background information is allowed to use, if it helps algorithm for prediction [2].

Targets are selected as protein-protein complexes or protein-DNA complexes. If only one component has available three-dimensional structure online, the second component may be carved from the complex, or could be homology modelled from the sequence. Taking both component structures from the complex biases docking too much towards the correct solution. In many cases, the protein structures with randomly re-oriented backbone coordinates are considered [4].

2.3.1 CAPRI Evaluation Protocol

In first CAPRI evaluation meeting in 2002, a protocol was established to assess geometric and biological properties of predicted models. The committee decided to evaluate the models based on the combination of three measures: the fractions of native residue – residue contact, interface RMSD and ligand RMSD. Then, the prediction models are ranked as high accuracy, medium accuracy, acceptable and incorrect. Detailed expressions for each rank can be seen on the Table 2.1 [61].

Table 2.1. Summary of CAPRI criteria for ranking predicted complexes [61].

Rank	F _{nat}	L _{rms}	or I _{rms}
High	≥ 0.5	≤ 1.0	or ≤ 1.0
Medium	≥ 0.3	1.0 < x ≤ 5.0	or 1.0 < x ≤ 2.0
Acceptable	≥ 0.1	5.0 < x ≤ 10.0	or 2.0 < x ≤ 4.0
Incorrect	< 0.1		

Root-mean Square Deviation (RMSD) represents the sample standard deviation of the differences between two proteins or protein complexes. The RMSD is calculated according to the following equation [5]:

$$\text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^n \|v_i - u_i\|^2}$$

Figure 2.4: Formulation of root-mean square deviation [5].

2.3.1.1 Ligand RMSD (L-RMSD)

To identify the ligand, CAPRI committee chooses to call the larger component of a complex as the receptor (R), and the other component as the ligand (L). The L-RMSD is a measure of distance of atoms of two ligand proteins calculated after the superposition of receptor proteins. Geometric distances and superposition are both computed on backbone atoms (N, C_α, C, and O). For further superimposition of the ligand after superimposed the receptors, residual rigid body rotation angle (θ_L) and residual translation vector of the geometric center (d_L) is required. One should keep in mind that, L-RMSD is a global measure that depends on the size of the ligand [61].

2.3.1.2 Interface RMSD (I-RMSD)

The global characteristics of L-RMSD may not always provide a good measure for the fit of the interface. If the ligand protein is very large, the orientation on interface becomes slightly different in the predicted model than in the target. Therefore I-RMSD

is chosen to be calculated. In that case, interface is re-defined as residues having at least one atom within 10 Å of an atom on the other residue. The backbone atoms of these selected residues are superimposed on their equivalents in the predicted complex to calculate I-RMSD [61].

2.3.1.3 Native and Nonnative Residue Frequencies (f_{nat} & $f_{\text{non-nat}}$)

Pairwise contacts between receptor and ligand are another important aspect to be considered. A contact is defined as two residues on different sides of the interface with any atoms within the distance of 5 Å. Native contact frequency (f_{nat}) is basically the ratio of correct number of contacted residues on the predicted complex divided by the total number of contacted residues on the target complex. In other words, if the receptor interface comprises N_R residues in the target and n_R of those residues are in the model, then $f_R = n_R / N_R$ measures how well the interface of receptor is predicted. Similarly, the native contact is the ratio of residue pairs in contact presented in the model over the residue pairs in contact in the target ($f_{\text{NC}} = n_C / N_C$). Because n_C could be manipulated artificially by pushing the ligand into the receptor, fraction number of nonnative contacts ($f_{\text{non-nat}}$) becomes important. It is defined as the ratio between the number of non-native residue-residue contacts in the predicted complex and total number of contacts in the prediction [61].

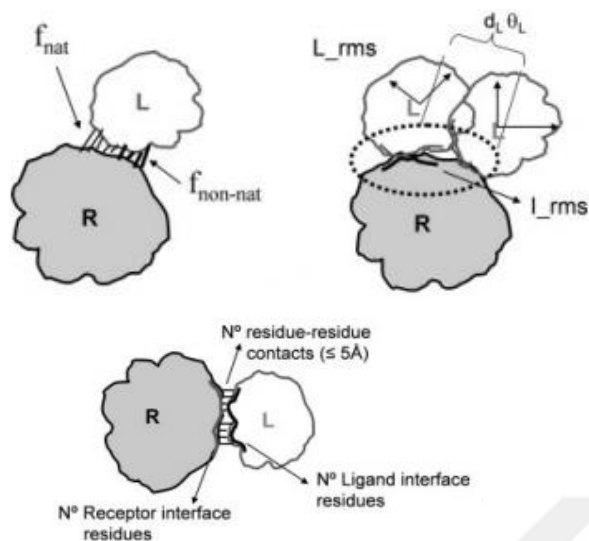


Figure 2.5: Basic representations of CAPRI evaluation criteria.

All three parameters (L-rmsd, I-rmsd and f_{nc}) are considered together for ranking the model. Fraction of native contacts is important criteria for evaluation. A complex should have at least 10% of native residue contacts. Otherwise no further evaluation is needed. However, the usefulness of I-RMSD and L-RMSD depends on the size of the protein. To eliminate the bias, it is enough for a complex to ensure one of the RMSD criteria (or statement).

Chapter 3

Materials and Methods

The route followed in this study is expressed roughly on Figure 3.0.1 and explained in detail on further following sub-chapters. The scheme shows three major routes. On first approach, the current version of PRISM is used for previous CAPRI targets to see the actual performance. The only parameter changed in the first approach is the threshold value for FiberDock scores. On default, PRISM takes models only with negative scores. To evaluate models in wider energy range, the threshold is changed to much larger value (<500) in order to see the positively scored models. In the second approach, the purpose is to see the effect of Monte Carlo (MC) minimization over the energy scores. MC minimization is basically assigning random numbers to the coordinates of atoms on backbone and side-chains of proteins in order to reduce the energy. On default, PRISM has final 50 steps of MC minimization at the end of FiberDock part. In that approach, the 50 steps reduced to zero and the results are compared with the results coming from the first approach. On the third approach, the major purpose is to compare the performance of two scoring functions. RosettaDock Local Refinement is used to compare the scoring performance with FiberDock scores. The models coming from second approach is subjected to the third approach. All results are presented on the following chapters.

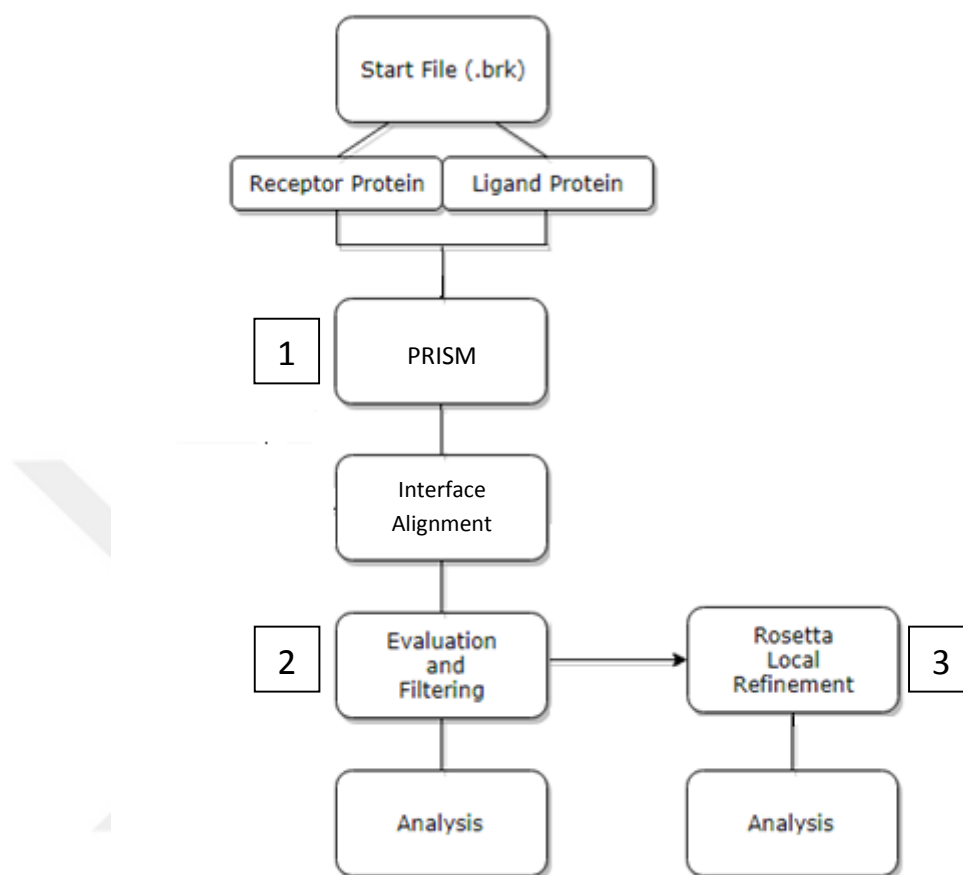


Figure 3.0.1. Overall route and individual steps followed for this study.

3.1 Target Dataset

For this study, targets are selected from the previous rounds of CAPRI (http://www.ebi.ac.uk/msd-srv/capri/pdb_ids.html). Currently, CAPRI has 40 rounds and 125 different targets in total. Targets could be selected to predict the correct interactions between enzyme and inhibitors, antibody and antigens or proteins and proteins. Currently the website is not up to date and some parts of the website need authorization (username and password) to be reached. Detailed information about all rounds, target proteins and the result PDB id's could be found in the Table 3.1. The targets without final PDB id's were compared with the final X-ray structures provided by the CAPRI website.

Table 3.1. All CAPRI targets with protein names and final PDBs.

Round	Target	Protein Name 1	Protein Name 2	PDB final
Round 1	T01	HPr	Phosphovcarrier	1KKL
	T02	Bovine Rotavirus VP6	Fab	-
	T03	Flu Hemagglutinin	Fab HC63	1KEN
Round 2	T04	Alpha-amylase complex	Camelide antibody VH domain 1 (AMYD10)	1KXT
	T05	Alpha-amylase complex	Camelide antibody VH domain 2 (AMY07)	1KXV
	T06	Alpha-amylase complex	Camelide antibody VH domain 3 (AMY09)	1KXQ
	T07	T cell receptor beta chain	Pyrogenic exotoxin A (SpeA)	1L0X
Round 3	T08	Nidogen-G3 Domain	Laminin	1NPE
	T09	Wild type LicT (Homodimer)		-
	T10	TBEV Envelope Protein (Trimeric form)		1URZ
Round 4	T11	Cellulosome cohesin (unbound)	Endoglucanase SS Dockerin complex (homology model)	1OHZ
	T12	Cellulosome cohesin (unbound)	Endoglucanase SS Dockerin complex (bound)	1OHZ
	T13	SAG1 Fab chain H (unbound) SAG1 Fab chain L (unbound)	Major Surface Antigen P30 (Antibody complex) (bound)	1YNT
	T14	Protein Serine/Threonine Phosphatase-1	Myosin Phosphatase Targeting Subunit 1	1S70
Round 5	T15	ImmD Immunity Protein	CoLD Catalytic Domain	1V74
	T16	Xylanase Inhibitor Protein (XIP-1) (1OM0)	Endo-1,4-Beta-Xylanase (1BG4)	1TA3
	T17	Xylanase Inhibitor Protein (XIP-1) (1OM0)	Endo-1,4-Beta-Xylanase (1UKR)	1TE1
	T18	Xylanase Inhibitor TAXI	Endo-1,4-Beta-Xylanase (1UKR)	1T6G
	T19	Ovine prion	FAB complex (Heavy – Light chain)	-
Round 6	T20	Protein Methyltransferase HemK	Release Factor 1 (RF1)	2B3T
Round 7	T21	Origin Recognirion Complex Subunit 1 (Orc1)	Saccharomyces Cerevisiae (Sir1)	1ZHI
Round 8	T22	Spliceosomal Protein U5-15K (1QGV) U5-52 (1GYF)	Cytoplasmic Domain Binding Protein (CD2BP2)	1SYX
	T23	Interferon-Induced Guanylate-Binding Protein 1 (GBP1) (dimer)		2B8W
Round 9	T24	ADP-Ribosylation Factor 1 (Arf1) (unbound) (1O3Y)	ArfBD (1BTN) (model-built)	2J59
	T25	ADP-Ribosylation Factor 1 (Arf1) (unbound) (1O3Y)	ArfBD (bound)	2J59
Round 10	T26	TolB	Peptidoglycan Associated Lipoprotein (Pal)	2HQS
Round 11	T27	Ubiquitin-Conjugating Enzyme (E2-25K)	Ubc9	2O25
Round 12	T28	E3 Uniquitin Protein Ligase NEDD4-like Protein	Monomer may be built from (1ZVD) or (1ND7)	2ONI
Round 13	T29	Trm82 form complex Trm8	Trm82 tRNA Guanine-N(7)-Methyltransferase	2VDU, 2VDV
Round 14	T30	Crystal structure of Rnd1-GTP bound to the RBD dimer		2R20
	T31	NMR structure of Rac1-GTP (1MH1)	Plexin-B1	2JPH
Round 15	T32	Protease savinase (1SVN)	Bifunctional Inhibitor BASI (1AVA_C)	3BX1
	T33	Protein / RNA complex		-
	T34	Protein / RNA complex		-
	T35	BM22-1- GH10 modules of Clostridium (1DYO) – (1N82)	Xyn10B xylanase	2W5F

	T36	BM22-1- GH10 modules of Clostridium (1DYO) – (1N82)	Xyn10B xylanase	2W5F
Round 16	T37	ADP-Ribosylation Factor 6	JNK-Interacting Protein 4 Leucine Zipper (JIP4-LZ2)	2W83
Round 17	T38	Centaurin-alpha-1	Kinase-Like Protein KIF1C (FHA domain)	3FM8
	T39	Centaurin-alpha-1	Kinase-Like Protein KIF1C (FHA domain)	3FM8
Round 18	T40	Beta-Trypsin	Protease Inhibitor (API-A)	3E8L
Round 19	T41	Colicin E9 DNase	Im2 immunity protein	2WPT
	T42	Oligomeric assembly of a designed tetratricopeptide CTPR2		2WQH
Round 20	T43	Discriminating correct models		-
	T44	Discriminating correct models		-
Round 21	T45	-	-	-
Round 22	T46	MTQ2	TRM112	3Q87
Round 23	T47	Colicine-E2 Immunity Protein	Colicin-E9	3U43
	T48	Hydroxylase	Ferredoxin	4I57
	T49	Hydroxylase (Different hexamer conformation)	Ferredoxin	4I57
Round 24	T50	Hemagglutinin Receptor Binding Domain HA1	Hemagglutinin Membrane Fusion Domain HA2	3R2X
	T51	CBM13	Fn3	4BXG
		GH5-CBM6	CBM13	
Round 25	T52	-	-	-
Round 26	T53	Alpharep-4	Alpharep-2	4JW2
	T54	Neocarzinostatin	Rep16	4JW3
	T55	Hemagglutinin HA1 chain	Hemagglutinin HA2 chain	4EEF
	T56	Hemagglutinin HA1 chain	Hemagglutinin HA2 chain	4EEF
Round 27	T57	BT4661	Heparin	4AK2
	T58	Lysozyme G	Inhibitor of G-type Lysozyme	4G9S
Round 28	T60	Peptide	Importin-alpha NLS binding domain (residues 70-529, PDB entry 1EJL)	3ZIN
	T61	Peptide	Importin-alpha NLS binding domain (residues 70-529, PDB entry 1EJL)	3ZIO
	T62	Peptide	Importin-alpha NLS binding domain (residues 70-529, PDB entry 1EJL)	3ZIP
	T63	Peptide	Importin-alpha NLS binding domain (residues 70-529, PDB entry 1EJL)	3ZIQ
	T64	Peptide	Importin-alpha NLS binding domain (residues 70-529, PDB entry 1EJL)	3ZIR
Round 29	T65	Peptide of C-terminus of the single-stranded DNA binding protein (SSB)	RNASE H	4Z0U
	T66	Peptide of C-terminus of the single-stranded DNA binding protein (SSB)	Pria DNA Helicase	4NL8
	T67	Peptide of C-terminus of the single-stranded DNA binding protein (SSB)	WW3 domain from NEDD4	4N7H
Round 30	T68-94			
Round 31	T95	Chromatin-Conjugating Enzyme	Histone H3.2 Histone H4 Histone H2A Histone H2B 1.1 DNA	4R8P
	T96	Artificial Alpha-repeat Protein (ALPHAREP-4) (EGF-A)	Green Fluorescent Protein.	4XL5
Round 32	T98	UCH- L5	RPN13	4UEM
	T99	UCH- L5	RPN13	4UEL
	T100	UCH- L5	INO80G (NFRKB) Ubiquitin-Propargyl	4UF6

	T101	UCH- L5	INO80G (NFRKB)	4UF5
Round 33	T103	UBE2Z	FAT10	5A4P
Round 34	T104	Pyocin AP41 DNASE Domain	AP41 Immunity Protein	4UHP
	T105	Pyocin S2 DNASE Domain	S2 Immunity Protein	4QKO

The website of CAPRI provides the starting files of all previous rounds in “.brk” format (Link: http://www.ebi.ac.uk/msd-srv/capri/pdb_ids.html). This format has the same fundamental constructions with a standard PDB file format. The same format should be used when a predictor submits models. The format starts with the keyword HEADER which should have the CAPRI-target number. Keyword COMPND contains the descriptions of chain assignments used for different proteins. Keyword AUTHOR should include the submitter’s information. Keyword SEQRES has the residue sequence of each protein. Any other annotation on methods may be included in keyword REMARK. Since one predictor may send up to 10 models for every target, each model should start with keyword MODEL and should end with keyword ENDMDL similar to the standard format of NMR ensembles. The chains of a model should be ended with keyword TER. Finally, the keywords ATOM and HETATM have the coordinates in standard PDB format. In Figure 3.1.2, the sample of the file format could be found. Depending on the target proteins on the starting file, the receptor and ligand proteins are selected and created as separate PDB files with necessary format arrangements included for each of them. Then the target dataset becomes ready for following steps.

```

HEADER      CAPRI 02
COMPND      MOL ID: 1;
COMPND      2 MOLECULE: VIRAL CAPSID VP6;
COMPND      3 CHAIN: A, B, C;
COMPND      4 MOL ID: 2;
COMPND      5 MOLECULE: FAB, KAPPA LIGHT CHAIN;
COMPND      6 CHAIN: L;
COMPND      7 MOL ID: 3;
COMPND      8 CHAIN H;
COMPND      9 MOLECULE: FAB, IG HEAVY CHAIN
EXPDTA      THEORETICAL MODEL,1 STRUCTURES
AUTHOR      submitter
REMARK      Predictor remarks

CRYST1      1.000      1.000      1.000  90.00  90.00  90.00 P 1
SCALE1      1.000000      0.000000      0.000000      0.000000
SCALE2      0.000000      1.000000      0.000000      0.000000
SCALE3      0.000000      0.000000      1.000000      0.000000
MODEL       1
ATOM        4  N   MET A   1  23.616 -6.024   3.029  1.00  20.67
ATOM        5  CA  MET A   1  23.057 -4.914   3.794  1.00  20.80
(...)
ATOM       3175  NZ  LYS A 397  13.121   4.639  20.328  1.00  50.11
ATOM       3176  OXT LYS A 397  17.870   0.132  22.866  1.00  50.72
TER        3177      A 397
ATOM       3181  N   MET B   1   3.029  23.616 -6.024  1.00  20.67
(...)
ATOM       6352  NZ  LYS B 397  20.328  13.121   4.639  1.00  50.11
ATOM       6353  OXT LYS B 397  22.866  17.870   0.132  1.00  50.72
TER        6354      LYS B 397
ATOM       6358  N   MET C   1  -6.024   3.029  23.616  1.00  20.67
ATOM       6359  CA  MET C   1  -4.914   3.794  23.057  1.00  20.80
(...)
ATOM       9529  NZ  LYS C 397   4.639  20.328  13.121  1.00  50.11
ATOM       9530  OXT LYS C 397   0.132  22.866  17.870  1.00  50.72
TER        9531      LYS C 397
ATOM         1  CB  ASP L   1  -5.859  28.852  76.956  1.00  55.84
ATOM         2  CG  ASP L   1  -6.496  29.295  75.685  1.00  55.84
(...)
ATOM       1637  O   ARG L 211  -40.305 -17.023  48.625  1.00  66.48
ATOM       1638  OT  ARG L 211  -41.130 -14.944  48.552  1.00  60.71
TER        1639      ARG L 211
ATOM       1639  CB  GLU H   0   3.480  -2.102  95.482  1.00  95.00
ATOM       3236  O   GLU H 208  -30.889 -26.738  60.340  1.00  61.67
ATOM       3237  OT  GLU H 208  -31.330 -25.575  58.484  1.00  67.04
TER        3238      GLU H 208
ENDMDL
MODEL       2
(...)
ENDMDL
END

```

Figure 3.1.2. CAPRI .brk file format and information on a starting file.

3.2 PRISM as Docking Algorithm

In this part of the research, PRISM is subjected for some small modifications on its algorithm. The working principle of PRISM is explained in detail in Chapter 2.2.3 before. FiberDock ranks all prediction models according to its scoring function. PRISM, then, selects the predicted models only with the negative scores. This threshold eliminates all of predictions with scores larger than zero after FiberDock scoring. To analyze the results in wider perspective, that threshold is changed to 500 in order to eliminate the complexes with much larger energy value. That parameter could be reached by the

following directory in PRISM standalone version: run_files > flexibleRefinement.py > line 141.

Rigid body Monte Carlo (MC) energy minimization is performed by FiberDock after back-bone refinement is successfully done for the proteins. As default, FiberDock has fifty cycles of MC minimization at the end. To see the true effects of minimization, last fifty minimization cycles is reduced to zero. Each target is performed parallel with both fifty cycles of minimization and zero cycle of minimization. That parameter could be reached by the following directory in PRISM standalone version: run_files > flexibleRefinement.py > line 124.

3.3 Rosetta as Alternative Scoring Algorithm

RosettaDock has the ability to find high-resolution binding modes by local searching and refine proposed docking positions. With that service, RosettaDock can be combined with other docking servers for refinement and ranking the predictions [6].

3.3.1 Rosetta Prepack Protocol (for local docking)

Rosetta prepack protocol is highly useful to prepare a protein-protein complex for docking via standard Rosetta docking protocol. Basically, prepack protocol changes the conformations of side chains of all residues other than the interface residues to ensure that the outside of interface have a low energy conformation which is necessary for scoring.

The prepack algorithm consists of three steps: First (i), the docking partners are moved to a distance which makes partners out of contact. Then (ii), side chain optimization of the unbound complex is performed by packing. Third (iii), unbound docking partners are brought back to the original orientation. Detailed information could be reached on www.rosettacommons.org.

3.3.2 Rosetta Docking Protocol (for local docking)

RosettaDock is a Monte Carlo (MC) based multi-scale docking algorithm with both rigid-body orientation and side-chain conformation for both low-resolution (local)

and high-resolution (global) stages. Typically, the difference between local and global docking is the information that the algorithm gets from the user. Global Docking is used to randomizing initial orientation of the two partners if there is no structural or biochemical information known about the protein. On the other hand, if a user wants to define starting pose of proteins, local perturbation takes place with initial orientations. Detailed information could be reached on www.rosettacommons.org.

3.4 Ligand RMSD (L-RMSD) Calculation

In the chapter 2.3.1.1, ligand rmsd (L-RMSD) is explained in detail. PyMOL is used to calculate L-RMSD values for all results in all rounds.

PyMOL is an open-sourced molecular visualization tool built on Python language. Users can easily develop their own functions and use with PyMOL in addition to current large script library [62].

To calculate L-RMSD in PyMOL, receptor proteins and ligand proteins should be selected as individual selections for both target complex and predicted complex. PyMOL selections could be handled in many ways. The single line command that is used in this purpose is following:

```
- select [selection name], / [protein name] / [segid] / [chain name] / [residue id] / [atom name]
```

After creating four selections for two complexes, receptor proteins of both target complex and predicted complex should be superimposed. There are two functions in PyMOL for superimposition of proteins: fit and align. The description for “align” command in PyMOL user guide says that “align” performs a sequence alignment followed by a structural alignment, and then carries out zero or more cycles of refinement in order to reject structural outliers found during the fit. Since the “align” function performs sequence alignment, it is more useful for our purpose. The output of “align”

command is the RMSD values of every cycle and the visual superposition of two proteins. The following command should be typed:

- align [*target receptor*], [*prediction receptor*]

After superimposition of receptor proteins, the atomic distance between the current positions of ligands should be calculated. For that purpose PyMOL has a function called “rms_cur”. The description for “rms_cur” function in PyMOL user guide says that “rms_cur” computes the RMS difference between two atom selections without performing any fitting. The following command should be typed:

- rms_cur [*target ligand*], [*prediction ligand*]

Possible Errors and Solutions:

- PyMOL has very strict rules on comparing two protein structures. Two proteins should have matching atoms which have identical in chain, residue name, residue number, atom name, segid and alternate position identifier if present. PDB files must be modified properly before using these functions.
- Creating objects instead of selections may cut the relationship between receptors and ligands. Superimposition of receptors without moving ligands has no meaning for L-RMSD calculation.

3.5 Interface RMSD (I-RMSD) Calculation

In the chapter 2.3.1.2, I-RMSD is explained in detail. A python script from literature is used to calculate I-RMSD values for all results in all rounds.

Trac is an enhanced wiki and issue tracking system for software development projects.” irmsd.py” open-source code derived from Trac specifically developed by Author Shruthi Viswanath to calculate interface RMSD between a model structure and reference structure. This code is used to calculate I-RMSD values for all results in all rounds.

The source code could be found on the website under the following directory:

<https://wiki.ices.utexas.edu/clsb/browser/branches/capriScripts/compareStructures/irmsd/irmsd.py>

The code could be run with the following configurations with exact order:

*<model1> <model1's receptor chain> <model1's ligand chain> <reference structure>
<reference receptor chain> <reference ligand chain>*

Example: m0.pdb A B 1ABC.pdb A B

3.6 Native and Non-native Residue Frequency (f_{nat} & $f_{\text{non-nat}}$) Calculation

In the chapter 2.3.1.3, the methods for calculating fraction of native and non-native contacts are explained in detail. UCSF Chimera is used to calculate native and non-native residue contact fractions for all results in all rounds.

UCSF Chimera [63] is a computer program for interactive visualization and analysis of molecular structures. Wide variety of molecular analysis includes density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles. UCSF Chimera is also used for generating high-quality images and animations. Chimera includes complete documentation and several tutorials which guide new users.

To find contacts between the receptor and ligand proteins, UCSF Chimera has function under the “Tools > Structure Comparison > Find Clash/Contact” tab. Two chains of interest are designated as the first and second set of atoms. Van der Waals overlap radius is selected as very low value (such as -5) to not miss any contacts between receptor and ligand proteins. Chimera has an option to create a text file and write the contacted residues with information including residue names, residue numbers, atom name and the distance between atoms. A simple python script code is developed for parsing the file and eliminating the residue contacts with the distance lower than 3 Å and higher than 5 Å.

Then, another python script can compare each contact in a predicted model with the contacts on target complex to count the number of matched (native) and unmatched (non-native) contacts. Final values of f_{nat} and $f_{\text{non-nat}}$ are written into a new text file.



Chapter 4

Results and Discussion

CAPRI target involves mostly enzyme – inhibitor and antibody – antigen complexes. However, not only the protein – protein interactions are selected for targets but also protein – RNA interactions and some scoring rounds may be the topic for a CAPRI target. Table 4a gives the summary of PRISM performance in previous CAPRI rounds. One should keep in mind that, this study purely tests the performance of PRISM. There are no pre-modifications over PDB structures.

Table 4.1. Summary of PRISM Performance in CAPRI. Results for best model on each target are given. Classification column gives the number for each category in the order of high/medium/acceptable/incorrect.

Target	Interface	Energy score	Total # of models	Fnat	I-rmsd	L-rmsd	Classification
T01	1kkmBI	25.67	16	0.724	1.738	4.979	0/5/0/11
T02	3n6yAB	-8.17	5	-	18.423	87.385	0/0/0/5
T03	1kenCU	-0.24	11	0.839	1.421	4.219	0/4/0/7
T04	1kxvBD	12.00	7	0.897	0.582	1.197	4/0/0/3
T05	1kxtAB	-29.39	3	0.844	0.573	0.876	2/0/0/1
T06	1kxqBG	-76.88	8	0.887	0.677	1.836	6/0/0/2
T07	2aq3AB	11.84	14	0.839	1.572	3.837	0/4/2/8
T08	1npeAB	1.13	14	0.864	1.526	6.478	0/6/0/8
T09	2go7AB	-53.48	124	-	20.357	45.487	0/0/0/124
T10	1oanAB	-98.11	84	-	47.078	86.554	0/0/0/84
T11	3p0dAC	-9.28	4	0.370	3.302	6.194	0/0/1/3
T12	3p0dAC	-71.53	7	0.963	0.542	1.097	3/0/0/4
T13	1yntDG	-44.98	27	0.833	1.139	19.019	0/4/0/23
T14	1s70AB	-48.80	21	0.327	0.852	1.675	0/5/1/15
T15	1tfkAB	-8.46	7	0.074	2.158	5.248	0/0/2/5
T16	1ta3AB	-41.39	20	0.879	0.998	1.172	2/0/1/17
T17	1te1AB	-31.24	9	0.405	2.233	5.753	0/0/6/3
T18	3hd8AB	-22.57	18	0.594	2.236	5.489	0/9/6/3
T19	1tqcAC	2.26	16	0.716	1.165	33.942	0/6/0/10

T20	2b3tAB	39.96	32	0.149	3.987	31.03	0/0/2/30
T21	1zbxAB	6.33	11	0.707	1.375	5.899	0/3/3/5
T22	1syxEF	-3.60	2	0.375	2.793	4.439	0/0/2/0
T23	2b8wAB	-51.84	23	-	5.510	18.420	0/0/0/23
T24	1ej6CE	11.60	4	-	12.362	37.936	0/0/0/4
T25	2j59BN	-51.87	13	0.769	1.523	3.609	0/6/0/7
T26	2hqsAH	50.75	9	0.559	1.323	5.918	0/6/0/3
T27	2o25AD	-7.58	49	0.72	2.125	17.802	0/0/4/45
T28	2yc1AD	-8.75	40	-	23.160	54.681	0/0/0/40
T29	2vduBE	-7.92	8	0.574	1.785	5.032	0/6/0/2
T30	2rexAB	-54.6	13	0.727	1.496	4.187	0/2/0/11
T32	3bx1AC	-58.49	10	0.802	1.072	2.811	0/4/0/6
T35	3fxdBd	-17.99	6	-	9.753	29.4038	0/0/0/6
T36	2a24AB	-23.21	16	-	17.334	38.439	0/0/0/16
T37	2kwiAB	16.42	21	0.041	3.080	30.406	0/0/0/21
T38	3mdbBD	7.14	8	0.255	2.989	5.315	0/0/1/7
T39	3mdbBD	-29.33	10	0.804	0.757	1.612	2/4/0/4
T40	2qyiCD	-16.55	7	-	12.651	27.476	0/0/0/7
T41	3zu0AB	22.9	2	-	14.759	36.542	0/0/0/2
T42	2q00AB	-22.65	52	-	12.994	39.008	0/0/0/52
T46	3do8AB	-26.85	8	-	18.152	48.8167	0/0/0/8
T47	2pmyAB	8.05	4	-	10.52	28.239	0/0/0/4
T50	2y43AB	-41.68	29	-	9.649	17.934	0/0/0/29

Totally, 42 previous CAPRI targets are used in this study. Among those 42 targets, %64 of targets (27 of them) gives the results ranked at least acceptable. Since PRISM is a template based approach, the reference PDB structures should not be selected as templates for its own targets. Therefore the models predicted with these problematic interfaces coming from the reference PDB structures are eliminated. Among those 27 at-least-acceptable targets, 41% of the targets (11 of them) are derived from the interfaces which are not from the reference PDB complex. For those models predicted with reference interfaces, alternative interface search is done among the database. The interface database could be reached under an online tool which is called PIFACE [64], constructed by Cukuroglu et al.

These problematic interfaces are representatives for their clusters and PIFACE gives no other alternative interface to use except for one. For target 26, interface 2hqsAB is the representative of that cluster. In the same cluster there is another member, 2w8bAH, which could be used for the predictions. Unfortunately, PRISM could not give any results with this interface which carry us where we started. Table 4b gives detailed summary for the target interfaces and PIFACE search for alternative interfaces.

Table 4.2. Targets, used interfaces, search results on PIFACE and model quality.

Targets	Reference PDB	Interface	PIFACE Status	Other Interfaces	Model Quality
Target 03	1KEN	1kenCU	Representative	None	Medium
Target 04	1KXV	1kxvBD	Representative	1 from same PDB	High
Target 05	1KXT	1kxtAB	Representative	2 from same PDB	High
Target 06	1KXQ	1kxqBG	Representative	3 from same PDB	High
Target 08	1NPE	1npeAB	Representative	None	Medium
Target 13	1YNT	1yntDG	Representative	1 from same PDB	Medium
Target 14	1S70	1s70AB	Representative	None	Medium
Target 16	1TA3	1ta3AB	Representative	None	High
Target 17	1TE1	1te1AB	Representative	None	Acceptable
Target 20	2B3T	2b3tAB	Representative	None	Acceptable
Target 22	1SYX	1syxEF	Representative	3 from same PDB	Acceptable
Target 25	2J59	2j59BN	Representative	5 from same PDB	Medium
Target 26	2HQS	2hqsAH	Representative	1 different interface	Medium
Target 27	2O25	2o25AD	Representative	1 from same PDB	Acceptable
Target 29	2VDU	2vduBE	Representative	1 from same PDB	Medium
Target 32	3BX1	3bx1AC	Representative	1 from same PDB	Medium

For 42 targets, total 782 models were constructed by PRISM. 126 models are predicted with a quality enough for ranking at least acceptable. In this study, PRISM is modified by changing parameters towards giving more predictions to analyze even those predictions have positive energy scores. An important fact has appeared here that there are many successful predictions with positive energy scores going to be lost on the results because of their energies. Among those qualified 126 models, 36% (45 models) has positive energy scores which cover 28.7% of total targets. Scoring part gives problematic results which significantly reduces the success rate of PRISM.

PRISM could not find any successful predictions for 36% of all 42 targets which makes 14 unsuccessful targets. The majority, 36%, of unsuccessful targets are the targets with homodimer interactions (Targets 09, 10, 23, 28 and 42). All of the targets for homodimer interactions ended with unsuccessful results. The second majority (with 28%) of unsuccessful targets are enzyme/inhibitor complexes. Targets 40, 41, 46 and 47 are the ones PRISM could not find any correct models. Targets 24, 35 and 36 have the proteins homology modeled from sequential knowledge which creates the 21,4% of PRISM's unsuccessful results. Target 50 uses a hypothetical protein for prediction which also belongs to the unsuccessful 14 targets. The other one, target 22, is an antibody/antigen prediction which has low number of structural predictions. The last unsuccessful one, Target 37, has large sequential difference between the structure of Leucine Zipper 2 protein given on starting file and the reference PDB structure. That difference creates large frequency decrease on native residue contacts which is the fundamental reason of failure for the target 37 models.

4.1 Detailed Results and Discussion for Each Target

4.1.1. Target 01: *Hpr* – *HprK/P*

Target 01 aims to predict the correct structure of complex between Hpr protein and HPr Kinase / Phosphorylase protein which is 1KKL in PDB. HPr kinase / phosphorylase (HprK/P) arranges the carbon source usage on Gram-positive bacteria by determining the phosphorylation state of the phosphocarrier protein HPr. HprK/P acts bidirectional and catalyzes both ATP-dependent phosphorylation of HPr and its dephosphorylation [65]. HPrK/P protein is in a hexamer form. Each subunit of hexamer contains an ATP-binding domain similar to other kinases. The ATP-dependent phosphorylation occurs by binding inorganic phosphate from ATP to typical P-loop structure formed by the Walker motif A present in many nucleotide-binding proteins [66].

PRISM returns total of 16 models constructed with 8 different interfaces. Among those 16 models, 6 of them are classified other than incorrect. Table 4.1.1 shows the details of best 6 models predicted by PRISM. Important thing is that none of them has negative energy score, which means with default parameters, PRISM would not show those models at all. The best model is selected as *hprk_0_phcr_0* with interface 1kkmbI which has I-rmsd value as 1.738 Å with the 72% of native residue contacts. The model is classified as medium quality. Figure 4.1.1 show the comparison of our best predicted model with the reference PDB structure 1KKL.

Table 4.1.1. Evaluation Statistics of all predicted models for CAPRI Target 01.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1kkmbI	hprk_0_phcr_1	16.88	1.987	0.621	0.486	6	5.971	Medium
1kkmbI	hprk_1_phcr_1	16.88	1.987	0.621	0.486	6	5.971	Medium
1kkmbI	hprk_0_phcr_0	25.67	1.738	0.724	0.523	6	4.979	Medium
1kkmbI	hprk_1_phcr_0	25.67	1.738	0.724	0.523	6	4.979	Medium
1kklCJ	hprk_1_phcr_0	33.55	1.903	0.621	0.609	6	7.051	Medium
1kklCJ	hprk_0_phcr_0	33.65	1.844	0.655	0.596	6	6.598	Medium

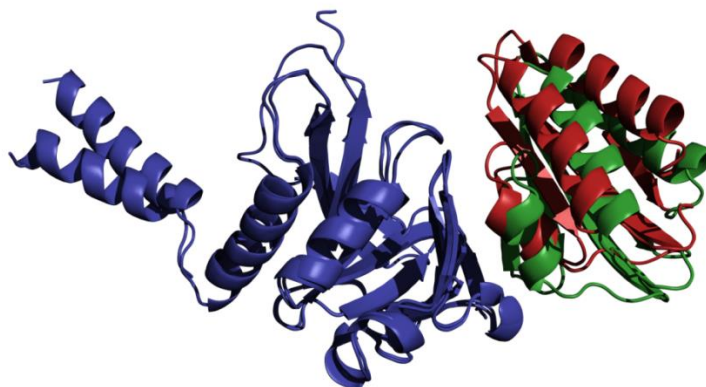


Figure 4.1.1. Best model of docking prediction for Target 01, HPr - HPrK/P complex. Blue: HPrK/P; red: crystal HPr; green: predicted HPr.

4.1.2. Target 02: Rotavirus VP6 – Fab

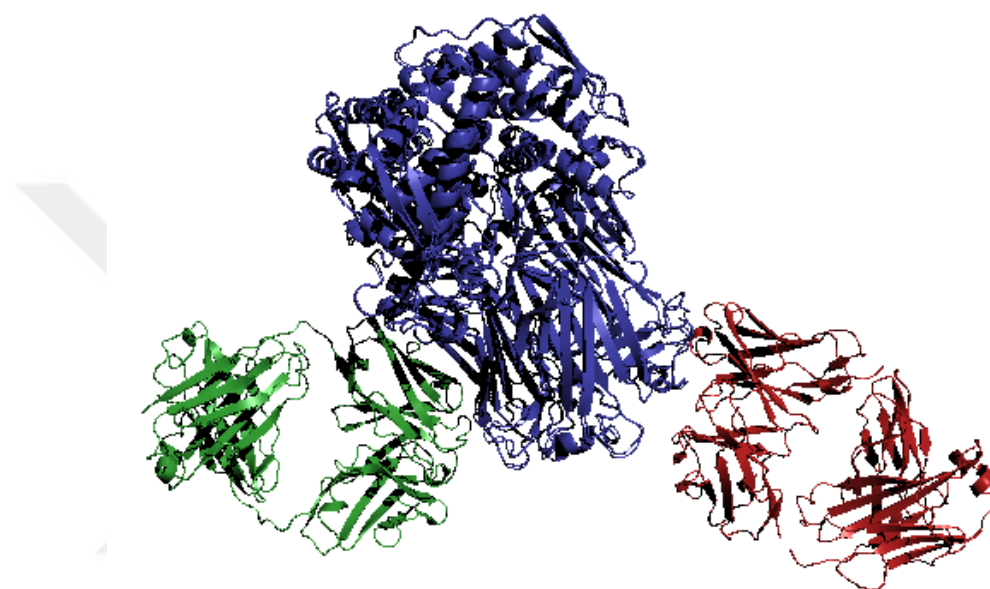
Target 02 aims to predict the correct structure of complex between Bovine Rotavirus VP6 protein and Fab protein. VP6 is a trimeric protein which contacts with both inner protein shells and outer protein shells. Outer layer interaction is necessary for a virus to enter the cell and could be performed through a distal beta-barrel domain. Inner layer interaction is necessary for packaging the genomic RNA and could be performed through proximal alpha-helical domain. So, VP6 acts as a connector protein between those two distinct biological functions of a virus [67].

PRISM returns total of 5 models constructed with 2 different interfaces. Unfortunately, all models are classified as incorrect. Table 4.1.2 shows the details of all models predicted by PRISM. No model has I-rmsd value less than 18 Å. Since both I-rmsd and L-rmsd values are too high, no further evaluation would be necessary. It can be concluded that PRISM could not find the correct complex form for Target 02.

Figure 4.1.2 shows the comparison of our best prediction (lowest energy), *fabh_0_vcvp_0* with interface 3n6yAB, with the reference complex structure. It can be seen that, Fab protein interacts almost opposite side of VP6 proteins.

Table 4.1.2. Evaluation Statistics of all predicted models for CAPRI Target 02.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
3n6yAB	fabh_0_vcvp_0	-8.17	18.423	87.385	Incorrect
3n6yAB	fabh_0_vcvp_1	-3.73	23.181	76.625	Incorrect
3n6yAB	vcvp_0_fabh_0	9.30	23.034	75.908	Incorrect
3n6yAB	vcvp_1_fabh_0	10.31	18.310	87.824	Incorrect
1v1pAB	vcvp_0_fabh_2	31.73	34.236	99.161	Incorrect

**Figure 4.1.2.** Best model of docking prediction for Target 02, Rotavirus VP6 / Fab complex. Blue: Rotavirus VP6; red: native Fab; green: predicted Fab.

4.1.3. Target 03: Influenza Hemagglutinin – Fab HC63

Target 03 aims to predict the correct structure of complex between Flu Hemagglutinin protein and Fab HC63 protein. Hemagglutinin (HA) is a trimeric receptor-binding protein of influenza virus with long peptides which allows HA act as a fusion membrane glycoprotein. In immune system, HA is a target for infectivity-neutralizing antibodies [68].

PRISM returns total of 11 models constructed with 3 different interfaces. Among those 11 models, 4 of them classified other than incorrect. Table 4.1.3 shows the details of best 4 models predicted by PRISM. The best model is selected as *hema_0_fabb_0* with

interface 1kenCU which has I-rmsd value as 1.421 Å with the 84% of native residue contacts. The model is classified as medium quality. Unfortunately, second best two predictions Figure 4.1.3 show the comparison of our best predicted model with the reference PDB structure 1KEN.

Table 4.1.3. Evaluation Statistics of all predicted models for CAPRI Target 03.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1kenCU	hema_0_fabb_0	-0.24	1.421	0.839	0.373	6	4.219	Medium
1kenCU	hema_1_fabb_0	-0.24	1.421	0.839	0.373	6	4.219	Medium
1kenCU	hema_0_fabb_1	3.07	1.672	0.774	0.400	6	4.142	Medium
1kenCU	hema_1_fabb_1	3.07	1.672	0.774	0.400	6	4.142	Medium

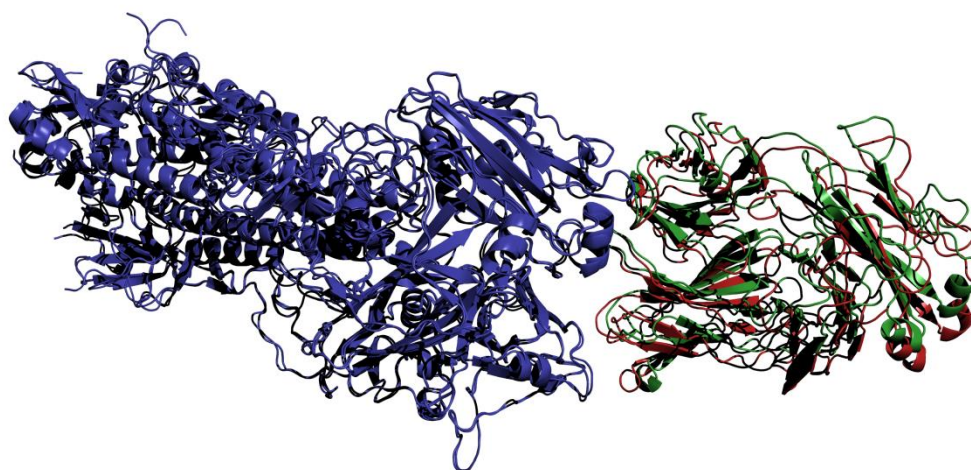


Figure 4.1.3. Best model of docking prediction for Target 03, Hemagglutinin / Fab HC63 complex. Blue: Hemagglutinin; red: native Fab; green: predicted Fab.

4.1.4. Target 04, 05, and 06: α -Amylase – Camelide Antibody VH Domains 1,2,3

Targets 04, 05 and 06 aim to predict the correct structures of complexes between α -amylase protein and different VH domains of Camelide antibody protein. All immunoglobulins consist of heterodimeric heavy (H) and light (L) chains where each of

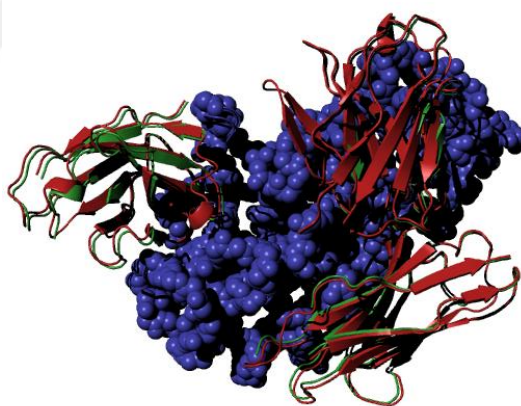
four or five domains is linked by disulfide bridges. Different from other antibodies, Camelid antibodies are functional without light chains and CH1 domains. Therefore, one single domain of heavy chain VH is enough for antigen-binding. Here, three VHH structures are selected for targets: AMB7, AMD9 and AMD10. Although, there are no important alterations in their structure, two VHHs bound outside the catalytic site. Only AMD9 establish an interaction with the active site which makes it a powerful amylase inhibitor [69].

Table 4.1.4 returns the cumulative evaluation statistics for targets 04, 05 and 06. PRISM returns total of 7 models constructed with 4 different interfaces for target 04. The best model, *paam_0_ca10_0* with interface 1kxvBD has I-rmsd value of 0.582 Å with 90% of native residue contacts. Unfortunately, the scoring function fails the correct predictions by giving them positive values which normally could not be seen with default PRISM.

For targets 5 and 6, best models have the lowest energy scores. For target 05, the best model is selected as *paam_1_ca07_0* with interface 1kxtAB which has 84 % of native residue contacts with 0.566 Å I-rmsd value. In that case, again, the best model does not have the lowest energy score. For target 06, PRISM returns 8 models with 3 different interfaces. *paam_0_ca09_0* with interface 1kxqBG is selected as one of the best models. It has 89 % of native residue contacts with 0.677 Å I-rmsd value. For three targets, total of 12 models (out of 18 models) are classified as high quality. Figure 4.1.4 shows three VHH proteins together with one α -Amylase.

Table 4.1.4. Evaluation Statistics of all predicted models for CAPRI Targets 4, 5 and 6.

	Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnnat	Clash	L-rmsd	Classification
Target 04	1kxvBD	paam_1_ca10_0	0.59	0.809	0.879	0.480	5	1.710	High
	1kxvBD	paam_1_ca10_1	0.59	0.809	0.879	0.480	5	1.710	High
	1kxvBD	paam_0_ca10_0	12.00	0.582	0.897	0.395	5	1.197	High
	1kxvBD	paam_0_ca10_1	12.00	0.582	0.897	0.395	5	1.197	High
Target 05	1kxtAB	paam_0_ca07_0	-29.39	0.573	0.844	0.365	3	0.876	High
	1kxtAB	paam_1_ca07_0	-27.44	0.566	0.844	0.372	3	0.898	High
Target 06	1kxqBG	paam_0_ca09_0	-76.88	0.677	0.887	0.292	13	1.836	High
	1kxqBG	paam_1_ca09_0	-76.88	0.677	0.887	0.292	13	1.836	High
	1kxqBG	paam_2_ca09_0	-76.88	0.677	0.887	0.292	13	1.836	High
	1kxqBG	paam_0_ca09_1	-75.66	0.744	0.887	0.284	13	2.171	High
	1kxqBG	paam_1_ca09_1	-75.66	0.744	0.887	0.284	13	2.171	High
	1kxqBG	paam_2_ca09_1	-75.66	0.744	0.887	0.284	13	2.171	High

**Figure 4.1.4.** Best models of docking prediction for Target 04, 05 and 06 on the same α -amylase. α -Amylase / Camelid Antibody VH Domains 1,2,3 complex. Blue: α -Amylase; red: native VH Domains; green: predicted VH domains.

4.1.5. Target 07: T-cell receptor β -chain - Pyrogenic Exotoxin A

Superantigens (SAGs) are disease-associated substances that stimulate immune system originated from bacterial or viral organisms. SAGs bind both to class II major histocompatibility complex (MHC) and T cell receptor (TCR) molecules at the same time which are located on the surfaces of antigen-presenting cells and T lymphocytes. Pyrogenic Exotoxin A is classified as SAG secreted by a Gram-positive bacteria *Streptococcus pyogenes* [70]. Target 7 aims to predict the correct structure of complex between β chain of T cell receptor protein and Pyrogenic Exotoxin A protein.

PRISM returns 14 models constructed with 8 different interfaces. Among those 14 models, 6 of them classified other than incorrect. Table 4.1.5 shows the details of best 6 models predicted by PRISM. The best model is selected as *tcar_0_spa1_0* with interface 2aq3AB which has I-rmsd value as 1.572 Å with the 84% of native residue contacts. The model is classified as medium quality. Unfortunately; all of our best models have negative energy scores, which means PRISM would not show those models at all. Figure 4.1.5 show the comparison of our best predicted model with the reference PDB structure 1L0X.

Table 4.1.5. Evaluation Statistics of all predicted models for CAPRI Target 7.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
2aq3AB	tcar_0_spa1_1	8.28	1.723	0.865	0.529	7	4.535	Medium
2aq3AB	tcar_1_spa1_1	9.58	1.642	0.838	0.537	7	4.394	Medium
2aq3AB	tcar_2_spa1_1	9.58	1.642	0.838	0.537	7	4.394	Medium
2aq3AB	tcar_0_spa1_0	11.84	1.572	0.838	0.483	7	3.837	Medium
2aq3AB	tcar_1_spa1_0	13.66	2.258	0.730	0.518	7	6.165	Acceptable
2aq3AB	tcar_2_spa1_0	13.66	2.258	0.730	0.518	7	6.165	Acceptable

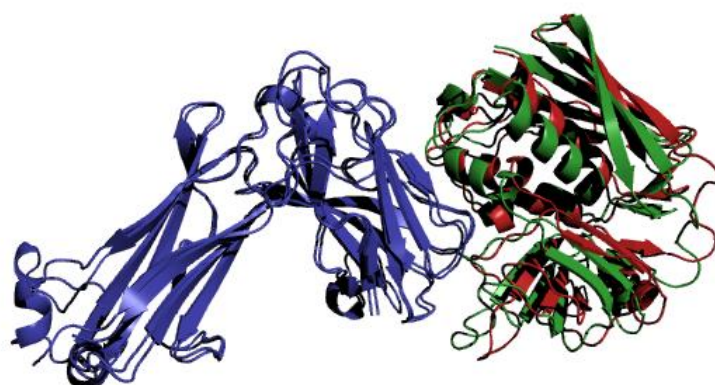


Figure 4.1.5. Best model of docking prediction for Target 07, T-cell receptor β -chain / Pyrogenic Exotoxin A complex. Blue: T-cell receptor; red: native exotoxin; green: predicted exotoxin.

4.1.6. Target 08: Laminin – Nidogen

Laminins are cross-shaped heterotrimeric glycoproteins composed of α , β , and γ chains with different variants. Nidogen is a protein to connect and stabilize the Laminin with Collagen IV proteins. Nidogen is also helps to other membrane proteins, proteoglycans and perlecans for integration. Nidogen – Laminins integration with high-affinity involves a single binding site on the Laminin γ 1 chain. Almost all Laminin isoforms uses the same binding site [71]. Target 8 aims to predict the correct structure of complex between Nidogen-G3 Domain and Laminin protein.

PRISM returns 14 models constructed with 4 different interfaces. Among those 14 models, 6 of them classified other than incorrect. Table 4.1.6 shows the details of best 6 models predicted by PRISM. The best model is selected as *ndg3_0_lmnn_0* with interface 1npeAB which has I-rmsd value as 1.526 Å with the 86% of native residue contacts. The model is classified as medium quality. Unfortunately; all of our best models have positive energy scores, which mean PRISM would not show those models at all.

Figure 4.1.6 show the comparison of our best predicted model with the reference PDB structure 1NPE.

Table 4.1.6. Evaluation Statistics of all predicted models for CAPRI Target 8.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1npeAB	ndg3_0_lmnn_0	1.13	1.526	0.864	0.260	9	6.478	Medium
1npeAB	ndg3_1_lmnn_0	1.13	1.526	0.864	0.260	9	6.478	Medium
1npeAB	ndg3_2_lmnn_0	1.13	1.526	0.864	0.260	9	6.478	Medium
1npeAB	ndg3_0_lmnn_1	6.97	1.962	0.515	0.393	9	5.842	Medium
1npeAB	ndg3_1_lmnn_1	6.97	1.962	0.515	0.393	9	5.842	Medium
1npeAB	ndg3_2_lmnn_1	6.97	1.962	0.515	0.393	9	5.842	Medium

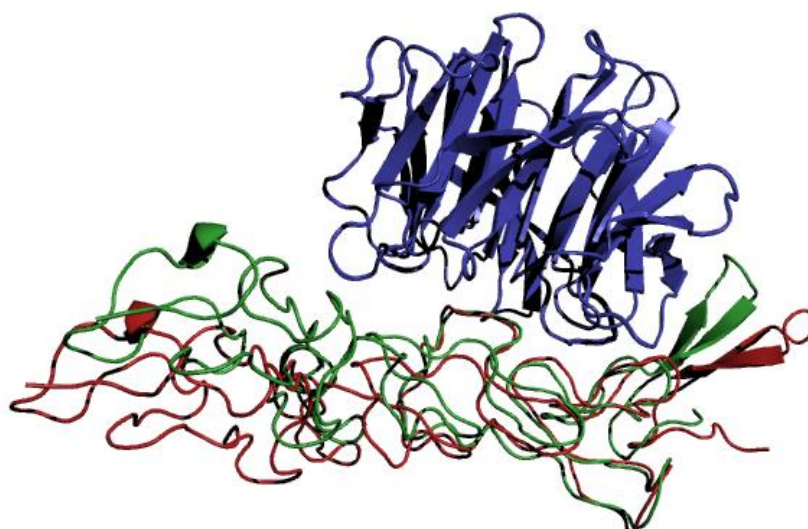


Figure 4.1.6. Best model of docking prediction for Target 08, Laminin / Nidogen complex. Blue: Nidogen; red: native laminin; green: predicted laminin.

4.1.7. Target 09: LicT Homodimer

LicT is a transcriptional anti-terminator protein which regulates the expression of operons involved in β -glucoside metabolism in *Bacillus subtilis*. As an anti-terminator, LicT binds to the RNA on a region partially overlaps to the terminator region, broke the

termination process and allows the transcription through the operon. LicT protein has analogous α -helical domains which form homodimer structure. Homodimer complex formation of LicT requires many tertiary and quaternary rearrangements [72]. Target 9 aims to predict the correct structure of complex of homodimer Wild type LicT.

PRISM returns 124 models constructed with 56 different interfaces. Among those 124 models, none of them shows the correct complex formation. No model has I-rmsd value less than 10 Å. Since both I-rmsd and L-rmsd values are too high, no further evaluation would be necessary. Table 4.1.7 shows the details of best 11 models predicted by PRISM. According to the energy score, the best model is selected as *lita_0_litb_0* with interface 2go7AB which has I-rmsd value as 20.357 Å which is classified as incorrect. It can be concluded that PRISM could not find the correct complex form for Target 09.

Figure 4.1.7 shows the comparison of our best prediction (lowest energy) with the reference complex structure. As it can be seen that, the predicted LicT monomer is placed almost perpendicular to the native form of LicT monomer.

Table 4.1.7. Evaluation Statistics of all predicted models for CAPRI Target 9.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
2go7AB	<i>lita_0_litb_0</i>	-53.48	20.357	45.487	Incorrect
2wp2AB	<i>lita_1_litb_0</i>	-44.93	20.567	39.600	Incorrect
2wp2AB	<i>litb_1_lita_0</i>	-36.42	20.584	45.755	Incorrect
3dhxAB	<i>litb_2_lita_0</i>	-34.72	13.151	18.596	Incorrect
2wp2AB	<i>lita_0_litb_2</i>	-31.63	20.873	46.617	Incorrect
2go7AB	<i>litb_0_lita_0</i>	-25.76	20.635	44.411	Incorrect
2gmyBC	<i>lita_1_litb_1</i>	-24.42	25.595	55.186	Incorrect
2wp2AB	<i>litb_0_lita_1</i>	-22.71	20.636	39.724	Incorrect
3bdzAB	<i>lita_0_litb_1</i>	-20.89	16.431	23.355	Incorrect
1k44AD	<i>lita_1_litb_1</i>	-18.69	24.521	56.732	Incorrect
3k7eCD	<i>lita_1_litb_1</i>	-14.98	21.208	43.198	Incorrect

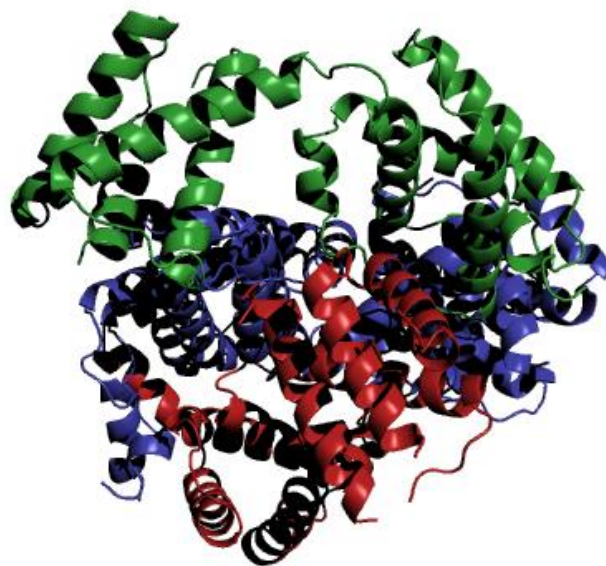


Figure 4.1.7. Best model of docking prediction for Target 09, LicT homodimer complex. Blue: LicT monomer; red: native second LicT monomer; green: predicted second LicT monomer.

4.1.8. Target 10: Trimeric form of TBEV Envelope

Tick-borne encephalitis virus envelope (TBEV) is a fusion glycoprotein which is effective in trimeric form. It is used for membrane fusion by enveloped viruses during the cell entry. Complex formation of TBEV requires subunit rearrangements in the form of clustering peptide loops at one end. The conformational changes of subunits are highly depending on the pH of environment. Trimeric form of TBEV could be induced by lowering the pH [73]. Target 10 aims to predict the correct structure of complex of homodimer TBEV Envelope Protein.

PRISM returns 84 models constructed with 32 different interfaces. Among those 84 models, none of them shows the correct complex formation. No model has I-rmsd value less than 10 Å. Since both I-rmsd and L-rmsd values are too high, no further evaluation would be necessary. Table 4.1.8 shows the details of best 11 models predicted by PRISM. According to the energy score, the best model is selected as *tbcb_1_tbca_0*

with interface 1oanAB which has I-rmsd value as 47.078 Å which is classified as incorrect. It can be concluded that PRISM could not find the correct complex form for Target 10.

Figure 4.1.8 shows the comparison of our best prediction (lowest energy) with the reference complex structure. As it can be seen that, the predicted TBEV monomer is placed almost 180° opposite side from the native form of TBEV monomer.

Table 4.1.8. Evaluation Statistics of all predicted models for CAPRI Target 10.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
1oanAB	tbc_b_1_tbc_a_0	-98.11	47.078	86.554	Incorrect
1k4rBC	tbc_a_0_tbc_b_0	-90.94	46.684	86.410	Incorrect
3uajAB	tbc_b_0_tbc_a_0	-69.03	45.844	85.505	Incorrect
1oanAB	tbc_a_0_tbc_b_0	-62.7	46.381	86.556	Incorrect
3uajAB	tbc_a_0_tbc_b_1	-37.19	46.059	85.983	Incorrect
3uajAB	tbc_a_0_tbc_b_0	-37.08	45.536	84.930	Incorrect
1w9aAB	tbc_a_0_tbc_b_1	-31.24	33.161	58.884	Incorrect
2ffjAB	tbc_b_0_tbc_a_2	-22.45	24.348	60.833	Incorrect
3uajAB	tbc_b_0_tbc_a_1	-22.22	41.939	80.862	Incorrect
1gulFG	tbc_b_0_tbc_a_0	-19.62	17.780	44.495	Incorrect
1y07AB	tbc_a_1_tbc_b_0	-15.25	31.870	59.892	Incorrect

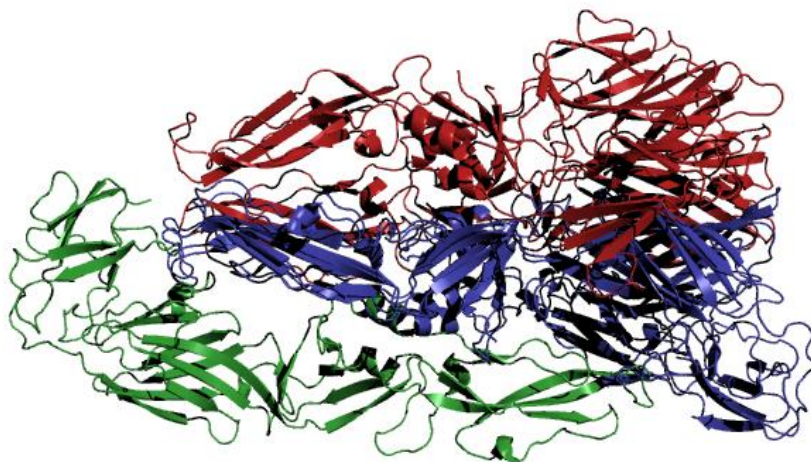


Figure 4.1.8. Best model of docking prediction for Target 10, TBEV homodimer complex. Blue: two TBEV monomers in trimer; red: native TBEV monomer; green: predicted TBEV monomer.

4.1.9. Target 11 and 12: Cohesin – Dockerin

Cohesin and Dockerin are two names of enzymatic domains interacting with each other found on the megadalton-wide catalytic multiprotein machinery called cellulosome. Anaerobic bacteria use cellulosome for degradation of cellulose and other plant wall polysaccharides [74]. Dockerin domain has 4 characteristic residues located on duplicated sequence which are highly involved on Cohesin interaction [75]. Target 11 and 12 aim to predict the correct structure of complex between Cellulosome Cohesin protein and Endonuclease SS Dockerin protein. The difference between target 11 and 12 is that, Target 11 uses homology modeled Dockerin protein. On the other hand, Target 12 uses Dockerin in bound form.

PRISM returns total 4 models for Target 11 constructed with 2 different interfaces. For Target 11, only one model has I-RMSD value lower than 4 Å. That model has 37% of native residue contacts with I-RMSD of 3.61 Å which is classified as acceptable. Table 4.1.9 gives detailed information about both Target 11 and Target 12.

On the other hand, Target 12 has total 7 models constructed with 2 different interfaces. Among those 12 models, 3 models have 97% percent of native residue contacts. The best model is selected as *docn_1_chs2_1* with interface 3p0dAC which has I-rmsd value of 0.542 Å and classified as high quality. The results show that instead of protein with homology modeling, bound form of the same protein leads the algorithm towards the correct complex structure more efficiently. Figure 4.1.9 shows the difference between two targets, by comparing the best two predictions with the reference PDB structure 1OHZ.

Table 4.1.9. Evaluation Statistics of all predicted models for CAPRI Targets 11 and 12.

	Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
Target 11	3p0dAC	doc1_0_chs1_0	-9.28	3.302	0.370	0.683	6	6.194	Acceptable
	3p0dAC	docn_1_chs2_1	-71.53	0.542	0.963	0.350	6	1.097	High
Target 12	3p0dAC	docn_1_chs2_2	-71.53	0.542	0.963	0.350	6	1.097	High
	3p0dAC	docn_1_chs2_0	-70.22	0.609	0.963	0.381	6	1.354	High



Figure 4.1.9. Best model of docking prediction for Target 11 (left) and Target 12 (right), Cohesin / Dockerin complex. Blue: Cohesin; red: native dockerin; green: predicted dockerin.

4.1.10. Target 13: Major Surface Antigen P30 - SAG1 Fab

SAG1 is a surface protein which is an important immune target with providing strong response against invasive *Toxoplasma gondii*. The forms of parasite devoid of SAG1 on their surface could be avoid from getting attacked and multiply. The structure of SAG1 consists of 3 major fragments F1, F2 and F3 which are undetectable by them individually but easily recognized together. The three fragments are able to induce the proliferation of anti-SAG1 T cells. During an immune response, monoclonal antibody produces SAG1 complexed with Fab rise against tachyzoite particles [76],[77]. Target 13

aims to predict the correct structure of complex between Major Surface Antigen P30 protein and both heavy and light chains of SAG1 Fab protein.

PRISM returns 27 models constructed with 19 different interfaces. Among those 27 models, 4 of them classified other than incorrect. Table 4.1.10 shows the details of best 4 models predicted by PRISM. The best model is selected as *sgfb_0_ap30_0* with interface 1yntDG which has I-rmsd value as 1.139 Å with the 84% of native residue contacts. The model is classified as medium quality. Figure 4.1.10 shows the comparison of our best predicted model with the reference PDB structure 1YNT.

Table 4.1.10. Evaluation Statistics of all predicted models for CAPRI Target 13.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1yntDG	sgfb_0_ap30_1	-44.98	1.139	0.833	0.375	7	19.019	Medium
1yntDG	sgfb_1_ap30_1	-42.91	1.163	0.847	0.378	7	18.885	Medium
1yntDG	sgfb_1_ap30_0	-42.41	1.215	0.833	0.400	7	18.899	Medium
1yntDG	sgfb_0_ap30_0	-42.17	1.224	0.847	0.408	7	18.947	Medium

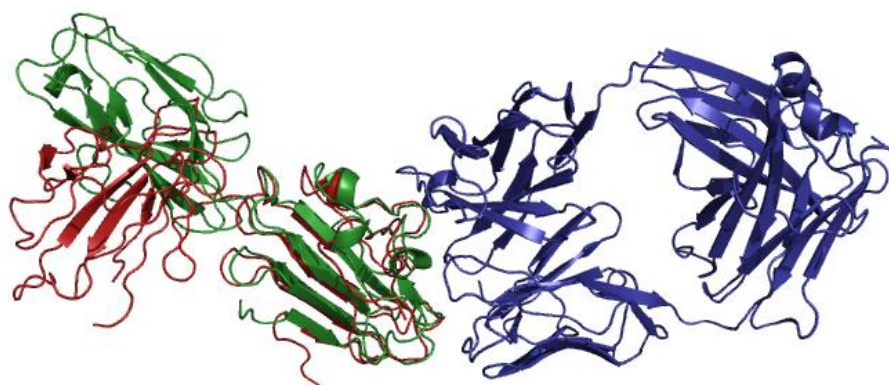


Figure 4.1.10. Best model of docking prediction for Target 13, Major Surface Antigen P30 / SAG1 Fab complex. Blue: SAG1 Fab; red: native MSA P30; green: predicted MSA P30.

4.1.11. Target 14: Protein Phosphatase 1-Myosin Phosphatase Targeting Subunit 1

Protein Phosphatase 1 (PP1) is a protein responsible for the regulation of broad range of cellular functions including cell-cycle, glycogen metabolism or muscle relaxation. For that reason, PP1 has a broad range number of regulatory subunits. Myosin Phosphatase Targeting Subunit 1 (MYPT1) is a smooth muscle protein which regulates the PP1 function on muscle relaxation. Complex form of these two proteins needs reshaping of the catalytic cleft of PP1 [78]. Target 14 aims to predict the correct structure of complex between Protein Serine/Threonine Phosphatase-1 protein and Myosin Phosphatase Targeting Subunit 1 protein.

PRISM returns 21 models constructed with 13 different interfaces. Among those 21 models, 6 of them classified other than incorrect. Table 4.1.11 shows the details of best 6 models predicted by PRISM. 5 out of 6 models are classified as medium and only the model *pstp_2_mpts_0* is acceptable quality. The best model is selected as *pstp_0_mpts_0* with interface 1s70AB which has I-rmsd value as 0.852 Å with the 33% of native residue contacts. The model is classified as medium quality. One important detail is that, the model with lowest energy score is not the best model. Figure 4.1.11 shows the comparison of our best predicted model with the reference PDB structure 1S70.

Table 4.1.11. Evaluation Statistics of all predicted models for CAPRI Target 14.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1s70AB	pstp_2_mpts_1	-50.54	0.893	0.300	0.365	20	1.863	Medium
1s70AB	pstp_0_mpts_0	-48.80	0.852	0.327	0.400	20	1.675	Medium
1s70AB	pstp_1_mpts_0	-48.80	0.852	0.327	0.400	20	1.675	Medium
1s70AB	pstp_0_mpts_1	-45.39	0.902	0.300	0.333	20	1.963	Medium
1s70AB	pstp_1_mpts_1	-45.39	0.902	0.300	0.333	20	1.963	Medium
1s70AB	pstp_2_mpts_0	-19.53	2.774	0.155	0.493	20	6.877	Acceptable

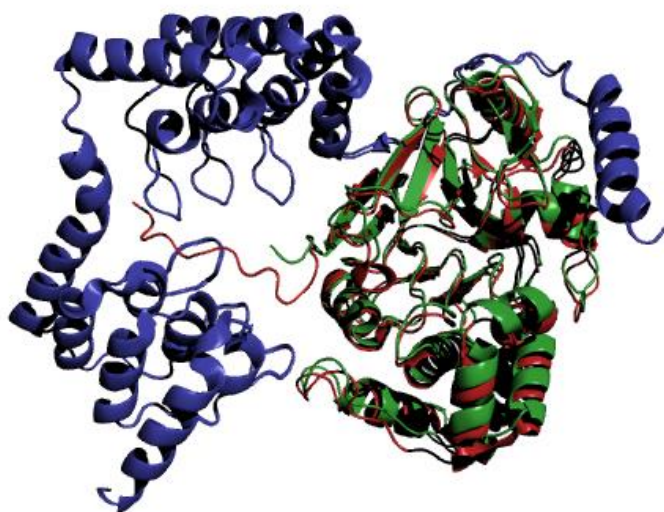


Figure 4.1.11. Best model of docking prediction for Target 14, Protein Phosphatase 1 / Myosin Phosphatase Targeting Subunit 1 complex. Blue: MPTS1; red: native Phosphatase 1; green: predicted Phosphatase 1.

4.1.12. Target 15: Colicin D - ImmD

Escherichia coli secrete toxins called Colicins to kill the competitors. Colicin D is one of those toxins which consist of three significant domains: translocation domain, receptor-binding domain and cytotoxic domain. As a toxin, Colicin D cleaves the anticodon loop of tRNA^{Arg} to block the protein synthesis thereby leading to cell death. Colicin-infected cells becomes having reduced activity only for accepting amino acid arginine [79]. ImmD is an immunity protein which forms complex with that toxin. ImmD mimics the tRNA substrate backbone and completely block the active site pocket of Colicin D [80]. Target 15 aims to predict the correct structure of complex between Colicin D protein and ImmD protein.

PRISM returns total 7 models constructed with 5 different interfaces. Among those 7 models, 2 of them classified other than incorrect. Table 4.1.12 shows the details of best 2 models predicted by PRISM. The best model is selected as *cdcd_0_imip_0* with interface 1tfkAB which has I-rmsd value as 2.158 Å with the 29% of native residue

contacts. The model is classified as acceptable quality. Figure 4.1.12 shows the comparison of our best predicted model with the reference PDB structure 1V74.

Table 4.1.12. Evaluation Statistics of all predicted models for CAPRI Target 15.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1tfkAB	cdcd_0_imip_0	-8.46	2.158	0.286	0.660	10	5.248	Acceptable
1tfkAB	cdcd_1_imip_0	-8.40	2.183	0.339	0.604	10	5.543	Acceptable

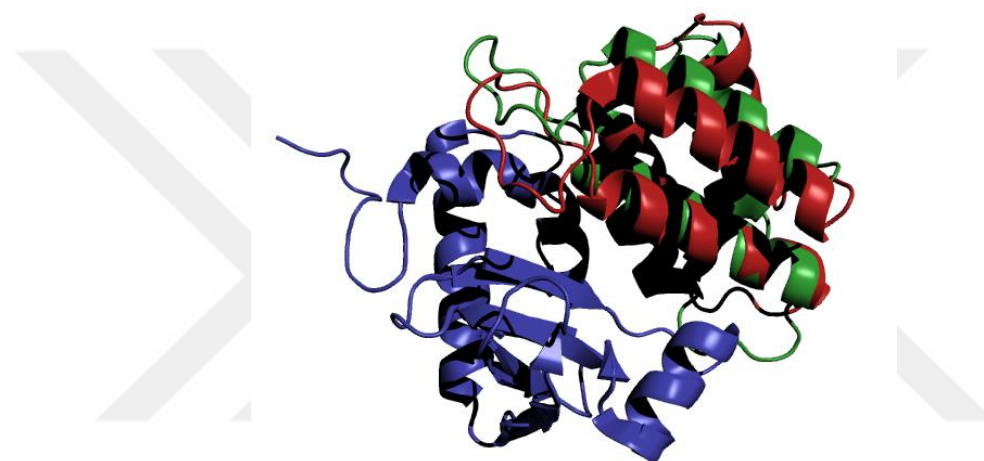


Figure 4.1.12. Best model of docking prediction for Target 15, Colicin D - ImmD complex. Blue: Colicin D; red: native ImmD; green: predicted ImmD.

4.1.13. Target 16: Family 10 Xylanase – XIP-1 Xylanase Inhibitor

Xylanase is an enzyme degraded xylan in plant cell walls into the cellulose without damaging it. Xylanase derived from *Penicillium simplicissimum* belongs to family 10 of glycosyl hydrolases (GH10) [81]. Xylanase Inhibitor Protein I (XIP-I) is protein inhibitor derived from *Triticum aestivum* which can be used to inhibit xylanases belongs to GH10 and GH11 families. XIP-1 has two independent active-site for enzyme binding which makes it useful for enzymes with different conformations [82]. Target 16 aims to predict the correct structure of complex between Xylanase Inhibitor Protein (XIP-1) and Endo-1,4-Beta-Xylanase protein from glycoside hydrolase family 10.

PRISM returns total 20 models constructed with 13 different interfaces. Among those 20 models, 3 of them classified other than incorrect. Table 4.1.13 shows the details of best 3 models predicted by PRISM. The best model is selected as *xyip_0_bxyl_0* with interface 1ta3AB which has I-rmsd value as 0.998 Å with the 88% of native residue contacts. The model is classified as high quality. Another model, *xyip_0_bxyl_2*, has the same interface but the I-rmsd value almost touches the limit. For that model, the energy score is positive 9.02 which is normally not considered as result. Figure 4.1.13 shows the comparison of our best predicted model with the reference PDB structure 1TA3.

Table 4.1.13 Evaluation Statistics of all predicted models for CAPRI Target 16.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1ta3AB	xyip_0_bxyl_0	-41.39	0.998	0.879	0.402	7	1.172	High
1ta3AB	xyip_0_bxyl_1	-41.39	0.998	0.879	0.402	7	1.172	High
1ta3AB	xyip_0_bxyl_2	9.02	3.970	0.167	0.843	7	11.586	Acceptable

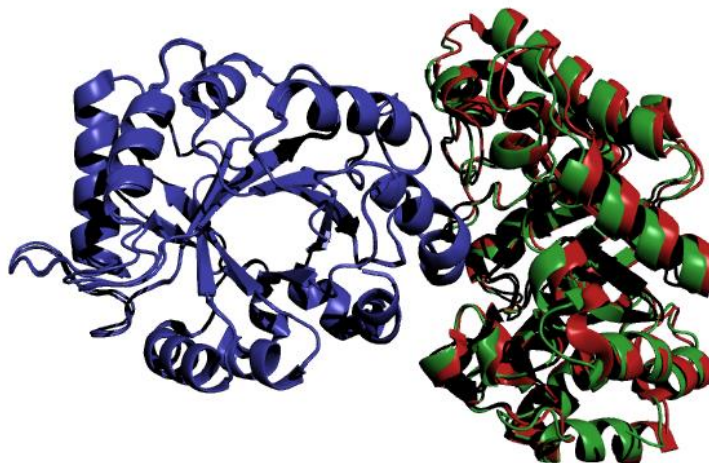


Figure 4.1.13. Best model of docking prediction for Target 16, Family 10 Xylanase – XIP-1 Xylanase Inhibitor complex. Blue: XIP-1; red: native family 10 xylanase; green: predicted family 10 xylanase.

4.1.14. Target 17: Family 11 Xylanase – XIP-1 Xylanase Inhibitor

The role and importance of xylanase enzyme was explained before on Target 16. In target 17, Xylanase enzyme is derived from *Aspergillus niger* belongs to family 11 of glycosyl hydrolases (GH11) [83]. Different from family 10 Xylanase, family 11 Xylanase mostly consists of β -sheets packed against each other which gives a unique structure. The cleft on one side of strongly twisted β -sheets becomes active site [84]. As it was mentioned before in Target 16, Xylanase Inhibitor Protein I (XIP-I) is protein inhibitor derived from *Triticum aestivum* which can be used to inhibit xylanases belongs to GH10 and GH11 families. Here, target 17 aims to predict the correct structure of complex between Xylanase Inhibitor Protein (XIP-1) and Endo-1,4-Beta-Xylanase protein from glycoside hydrolase family 11.

PRISM returns 9 models constructed with 4 different interfaces. Among those 9 models, 6 of them classified other than incorrect. Table 4.1.14 shows the details of best 6 models predicted by PRISM. The best model is selected as *xip1_0_bxyn_1* with interface 1te1AB which has I-rmsd value as 2.292 Å with the 41% of native residue contacts. The model is classified as acceptable quality. One important detail is that, the model with lowest energy score is not the best model. Furthermore, the model *xip1_1_bxyn_2* has the lowest L-rmsd but the highest energy score among the best 6 models. Figure 4.1.14 shows the comparison of our best predicted model with the reference PDB structure 1TE1.

Table 4.1.14. Evaluation Statistics of all predicted models for CAPRI Target 17.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1te1AB	xip1_1_bxyn_1	-32.92	2.415	0.419	0.551	7	5.052	Acceptable
1te1AB	xip1_0_bxyn_2	-31.24	2.233	0.405	0.565	7	5.753	Acceptable
1te1AB	xip1_0_bxyn_1	-27.09	2.292	0.405	0.571	7	4.443	Acceptable
1te1AB	xip1_1_bxyn_0	-21.12	3.686	0.216	0.652	7	11.428	Acceptable
1te1AB	xip1_0_bxyn_0	-19.27	2.432	0.473	0.563	7	4.647	Acceptable
1te1AB	xip1_1_bxyn_2	-8.95	2.495	0.405	0.630	7	3.761	Acceptable

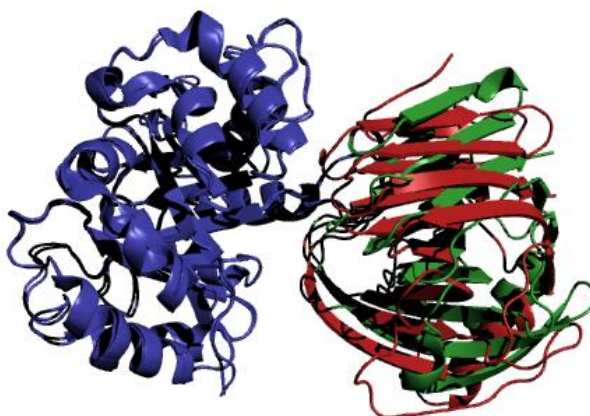


Figure 4.1.14. Best model of docking prediction for Target 17, Family 11 Xylanase – XIP-1 Xylanase Inhibitor complex. Blue: XIP-1; red: native family 11 xylanase; green: predicted family 11 xylanase.

4.1.15. Target 18: Family 11 Xylanase – TAXI-I

As it was mentioned before on target 17, Xylanase enzyme is derived from *Aspergillus niger* belongs to family 11 of glycosyl hydrolases (GH11) [83]. The role and importance of family 11 Xylanase was also mentioned before in that chapter. The other protein called *Triticum aestivum* xylanase inhibitor-I (TAXI-I) is developed by the organism itself as an arsenal for inhibiting the cell wall-degrading enzymes. Structurally, TAXI-I is homologous with the pepsin-like family of aspartic proteases [85]. Target 18 aims to predict the correct structure of complex between Xylanase Inhibitor TAXI protein and Endo-1,4-Beta-Xylanase protein from glycoside hydrolase family 11.

PRISM returns 18 models constructed with 5 different interfaces. There are two prominent interfaces for target 18: 1t6gBD and 3hd8AB. The best model is selected as *taxi_0_edbx_2* with interface 1t6gBD which has I-rmsd value as 1.016 Å with the 80% of native residue contacts. That model is classified as medium quality. On the other hand, with interface 3hd8AB, the best model is selected as *taxi_0_edbx_0* which has I-rmsd value as 2.236 Å with the 60% of native residue contacts. That model is classified as

acceptable quality. Figure 4.1.15 shows the comparison of our best predicted model, *taxi_0_edbx_0* with interface 3hd8AB, with the reference PDB structure 1T6G.

Table 4.1.15. Evaluation Statistics of all predicted models for CAPRI Target 18.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1t6gBD	taxi_0_edbx_2	-27.30	1.016	0.797	0.522	10	1.846	Medium
1t6gBD	taxi_1_edbx_2	-27.30	1.016	0.797	0.522	10	1.846	Medium
1t6gBD	taxi_2_edbx_2	-27.30	1.016	0.797	0.522	10	1.846	Medium
3hd8AB	taxi_0_edbx_0	-22.57	2.236	0.594	0.594	10	5.489	Acceptable
3hd8AB	taxi_0_edbx_1	-22.57	2.236	0.594	0.594	10	5.489	Acceptable
3hd8AB	taxi_1_edbx_0	-22.57	2.236	0.594	0.594	10	5.489	Acceptable
3hd8AB	taxi_1_edbx_1	-22.57	2.236	0.594	0.594	10	5.489	Acceptable
3hd8AB	taxi_2_edbx_0	-22.57	2.236	0.594	0.594	10	5.489	Acceptable
3hd8AB	taxi_2_edbx_1	-22.57	2.236	0.594	0.594	10	5.489	Acceptable
1t6gBD	taxi_0_edbx_0	-22.52	1.515	0.667	0.535	10	3.118	Medium
1t6gBD	taxi_1_edbx_0	-22.52	1.515	0.667	0.535	10	3.118	Medium
1t6gBD	taxi_2_edbx_0	-22.52	1.515	0.667	0.535	10	3.118	Medium
1t6gBD	taxi_0_edbx_1	-7.45	1.314	0.652	0.511	10	2.276	Medium
1t6gBD	taxi_1_edbx_1.	-7.45	1.314	0.652	0.511	10	2.276	Medium
1t6gBD	taxi_2_edbx_1	-7.45	1.314	0.652	0.511	10	2.276	Medium

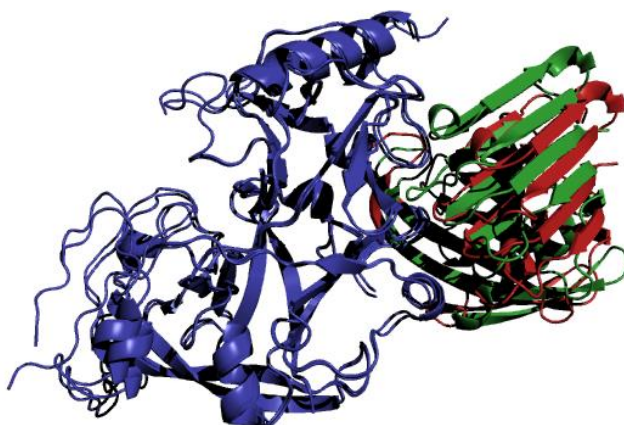


Figure 4.1.15. Best model of docking prediction for Target 18, Family 11 Xylanase / TAXI-1 complex. Blue: TAXI-1; red: native family 11 xylanase; green: predicted family 11 xylanase.

4.1.16. Target 19: Ovine Prion – Fab

Prion is a disease which is known as “mad cow disease” caused by the mutation on ovine prion protein. Normally, Ovine Prion protein has globular domain consists of 3 α -helices and a short anti-parallel β -sheet [86]. Studies show that, the conformational change in α -helix rich prion protein into the β -structure-rich prion causes the infection. Fab is an antibody protein forming a complex with prion protein during the infection which makes it the best coworker to study the Prion disease [87]. Target 19 aims to predict the correct structure of complex between Ovine Prion protein and Fab protein.

PRISM returns 16 models constructed with 9 different interfaces. Among those 16 models, 6 of them classified other than incorrect. Table 4.1.16 shows the details of best 6 models predicted by PRISM. The best model is selected as *ovip_0_ovhl_0* with interface 1tqcAC which has I-rmsd value as 1.165 Å with the 72% of native residue contacts. The model is classified as medium quality. One important detail is that, the model with lowest energy score is not the best model. Furthermore, none of the best models has negative energy score which means PRISM would not show those models at

all. Figure 4.1.16 shows the comparison of our best predicted model with the reference PDB structure.

Table 4.1.16. Evaluation Statistics of all predicted models for CAPRI Target 19.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1tqcAC	ovip_0_ovhl_2	1.63	1.175	0.741	0.294	12	33.652	Medium
1tqcAC	ovip_1_ovhl_2	1.63	1.175	0.741	0.294	12	33.652	Medium
1tqcAC	ovip_0_ovhl_0	2.26	1.165	0.716	0.301	12	33.942	Medium
1tqcAC	ovip_0_ovhl_1	2.26	1.165	0.716	0.301	12	33.942	Medium
1tqcAC	ovip_1_ovhl_0	2.26	1.165	0.716	0.301	12	33.942	Medium
1tqcAC	ovip_1_ovhl_1	2.26	1.165	0.716	0.301	12	33.942	Medium

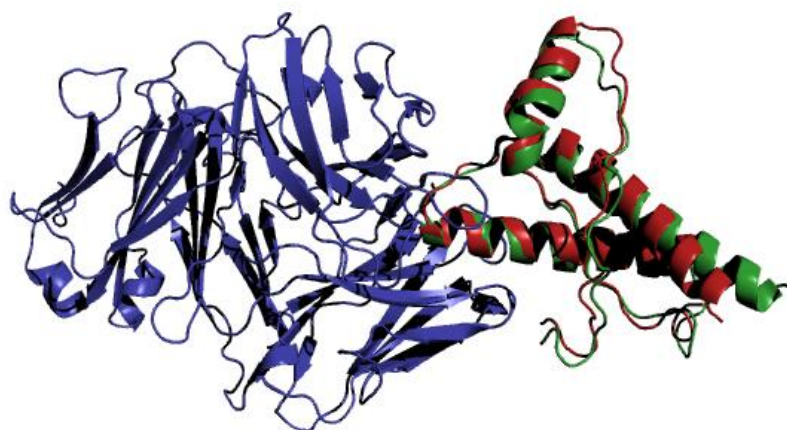


Figure 4.1.16. Best model of docking prediction for Target 19, Ovine Prion / Fab complex. Blue Fab; red: native Ovine Prion; green: predicted Ovine Prion.

4.1.17. Target 20: HemK – RF1

Release Factor 1 (RF1) is a protein in charge with translation termination in a cell. Ribosome needs to read a stop signal on mRNA in order to terminate the translation. Class I release factors bind to the ribosome and trigger the hydrolysis of peptidyl-tRNA with the trigger of stop codon [88]. In bacteria, RF1 proteins recognize the stop codons UAG and UAA, and RF2 recognize UAA and UGA. However, eukarya and archea have

one release factor (eRF1/aRF1) that recognizes all stop codons [89]. Both RFs share one motif called GGQ which interacts with ribosome. Hemk is a protein from *Escherichia coli* responsible for modulating the release factors to control the termination activity of ribosomal protein synthesis [90]. Target 20 aims to predict the correct structure of complex between Protein Methyltransferase HemK protein and Release Factor 1 (RF1) protein.

PRISM returns 32 models constructed with 21 different interfaces. Among those 32 models, 2 of them classified other than incorrect. Table 4.1.17 shows the details of best 2 models predicted by PRISM. The best model is selected as *hemk_1_ref1_0* with interface 2b3tAB which has I-rmsd value as 3.987 Å with the 15% of native residue contacts. The model is classified as acceptable quality. One important detail is that, both 2 best models have positive energy score, which means PRISM would not show those models at all. Figure 4.1.17 shows the comparison of our best predicted model with the reference PDB structure 2B3T.

Table 4.1.17. Evaluation Statistics of all predicted models for CAPRI Target 20.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
2b3tAB	hemk_1_ref1_0	39.96	3.987	0.149	0.868	19	31.03	Acceptable
2b3tAB	hemk_2_ref1_0	39.96	3.987	0.149	0.868	19	31.03	Acceptable

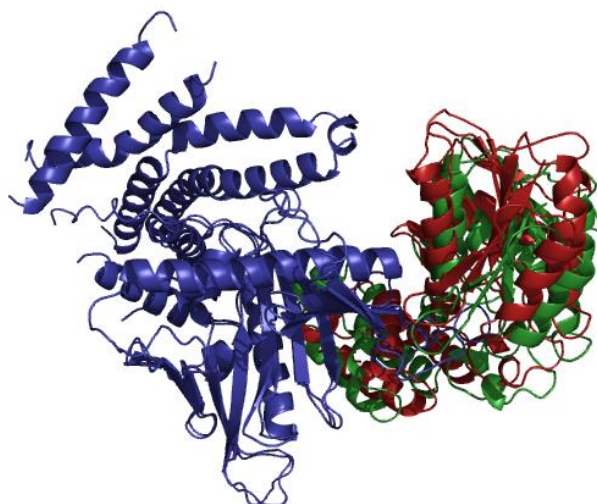


Figure 4.1.17. Best model of docking prediction for Target 20, HemK / RF1 complex. Blue RF1; red: native HemK; green: predicted HemK.

4.1.18. Target 21: Orc1 – Sir1

Silent Information Regulator 1 (Sir1) is a protein with an important role in chromatin silencing which is found in yeast *Saccharomyces cerevisiae*. Silencer proteins are bound by DNA-binding proteins specific for certain sequence and recruit Sir Proteins which cause transcriptionally repressive modification of nucleosome array [91]. On the other hand, origin recognition complex (Orc) proteins, again from *Saccharomyces cerevisiae*, also have a role in transcriptional silencing. Orc1 protein has two important domains: bromo-adjacent homology (BAH) domain also found on Sir1 and non-conserved helical sub-domain which is necessary for Orc1-Sir1 interaction [92]. Target 21 aims to predict the correct structure of complex between Origin Recognition Complex Subunit 1 (Orc1) protein and *Saccharomyces Cerevisiae* (Sir1) proteins.

PRISM returns 11 models constructed with 6 different interfaces. Among those 11 models, 6 of them are classified other than incorrect. Table 4.1.18 shows the details of best 6 models predicted by PRISM. The best model is selected as *orcs_0_sir1_1* with interface 1zbxAB which has I-rmsd value as 1.375 Å with the 71% of native residue contacts. The model is classified as medium quality. One important detail is that, the

model with lowest energy score is not the best model. Furthermore, the model *orcs_1_sir1_0* has the lowest energy score among the best 6 models but the classification of that model is acceptable. Another important thing is that, our best model *orcs_0_sir1_1* has positive energy score. Normally PRISM would not show that model at all. Figure 4.1.18 shows the comparison of our best predicted model with the reference PDB structure 1ZHI.

Table 4.1.18. Evaluation Statistics of all predicted models for CAPRI Target 21.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1zbxAB	orcs_1_sir1_0	-15.16	2.421	0.634	0.552	6	8.178	Acceptable
1zbxAB	orcs_1_sir1_2	-15.16	2.421	0.634	0.552	6	8.178	Acceptable
1zbxAB	orcs_0_sir1_1	6.33	1.375	0.707	0.442	6	5.899	Medium
1zbxAB	orcs_1_sir1_1	11.11	3.998	0.293	0.745	6	16.873	Acceptable
1zbxAB	orcs_0_sir1_0	13.25	1.413	0.756	0.415	6	6.088	Medium
1zbxAB	orcs_0_sir1_2	13.25	1.413	0.756	0.415	6	6.088	Medium



Figure 4.1.18. Best model of docking prediction for Target 21, Orc1 / Sir1 complex. Blue Orc1; red: native Sir1; green: predicted Sir1.

4.1.19. Target 22: Spliceosomal Protein (U5-15K) - Cytoplasmic Domain Binding Protein (U5-52K) (CD2BP2)

Spliceosome is a large ribonucleoprotein (RNP) responsible for splicing the pre-mRNA in the cell nucleus of eukaryotic cells. This molecular machine consists of five major snRNP proteins (U1, U2, U4, U5 and U6) and several non-snRNP proteins. U5 snRNP proteins have particle specific versions, like 15K, 40K, 52K, which are important for spliceosome assembly, maturation and dissociation. U5-52K, also identical with CD2 Receptor Binding Protein (CD2BP2), is associated with spliceosome and plays a role in immune response. U5-15K has 5 β -strands surrounded by 4 α -helices and GYF-domain which mediates the interaction of U5-52K and U5-15K has an anti-parallel β -sheet flanked by 3 α -helices. The interaction between U5-52K and U5-15K is necessary for the association of U5-52K with U5 snRNP [93]. Target 22 aims to predict the correct structure of complex between Spliceosomal protein (U5-15K) and Cytoplasmic Domain Binding protein (U5-52K).

PRISM returns 2 models constructed with the same interface. Both models are classified as acceptable. Table 4.1.19 shows the details of 2 models predicted by PRISM. The best model is selected as *spu5_0_cdbp_0* with interface 1syxEF which has I-rmsd value as 2.793 Å with the 38% of native residue contacts. One important detail is that, the model with lowest energy score is not the best model. The model *spu5_0_cdbp_1* has the lowest energy score but higher I-rmsd value. Figure 4.1.19 shows the comparison of our best predicted model with the reference PDB structure 1SYX.

Table 4.1.19. Evaluation Statistics of all predicted models for CAPRI Target 22.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
1syxEF	spu5_0_cdbp_1	-8.11	3.154	0.425	0.500	4	7.54	Acceptable
1syxEF	spu5_0_cdbp_0	-3.60	2.793	0.375	0.464	4	4.439	Acceptable

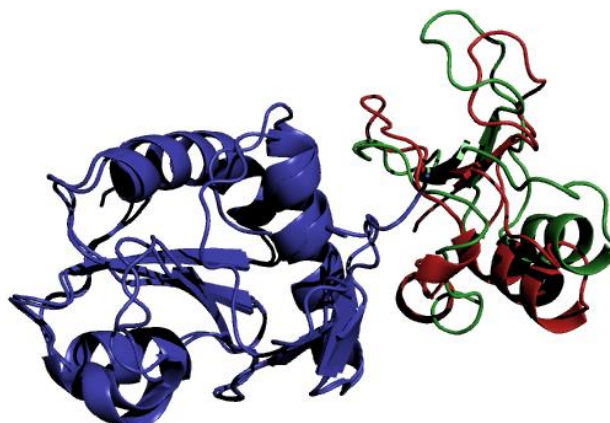


Figure 4.1.19. Best model of docking prediction for Target 22, Spliceosomal Protein / Cytoplasmic Domain Binding Protein complex. Blue Spliceosomal Protein; red: native Cytoplasmic Domain Binding Protein; green: predicted Cytoplasmic Domain Binding Protein.

4.1.20. Target 23: hGBP1 homodimer

Human guanylated binding protein 1 (hGBP1) is a kind of large GTP binding proteins. It also has the ability to hydrolyze GTP to GMP in consecutive cleavage reactions. The hGBP1 has several roles in the cell. On inflammation, hGBP1 antagonizes the proliferation and angiogenic activity of endothelial cells. It also provides resistance against the proliferation of two specific viruses: stomatitis and encephalomyocarditis [94]. Target 23 aims to predict the correct complex structure of homodimer hGBP1.

PRISM returns 23 models constructed with 9 different interfaces. Among those 23 models, all of them are classified as incorrect. Table 4.1.20 shows the details of best 11 models predicted by PRISM. The best model is selected as *igbp_1_igbp_1* with interface 2b8wAB which has I-rmsd value as 5.510 Å. Since none of the models has I-rmsd value less than 4 Å (or L-rmsd less than 10 Å), there is no further evaluation necessary. It can be concluded that PRISM could not find the correct complex form for

Target 23. Figure 4.1.20 shows the comparison of our best predicted model with the reference PDB structure 2B8W.

Table 4.1.20. Evaluation Statistics of all predicted models for CAPRI Target 23.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
2b8wAB	igbp_1_igbp_1	-51.84	5.510	18.420	Incorrect
1zvpAB	igbp_0_igbp_0	-27.24	33.200	77.502	Incorrect
3mcwAB	igbp_1_igbp_2	-23.36	32.430	72.519	Incorrect
2ddzAB	igbp_0_igbp_0	-22.46	34.631	77.701	Incorrect
2y1hAB	igbp_0_igbp_0	-10.92	34.311	80.281	Incorrect
2xguAB	igbp_0_igbp_0	-9.48	34.339	77.918	Incorrect
2b8wAB	igbp_2_igbp_0	-0.35	4.720	11.119	Incorrect
2g37AB	igbp_0_igbp_2	0.41	33.454	74.822	Incorrect
2fpwAB	igbp_0_igbp_0	1.97	32.257	70.255	Incorrect
2rhmcD	igbp_0_igbp_0	4.28	33.513	74.279	Incorrect
2fw7AB	igbp_0_igbp_0	5.59	34.226	80.515	Incorrect

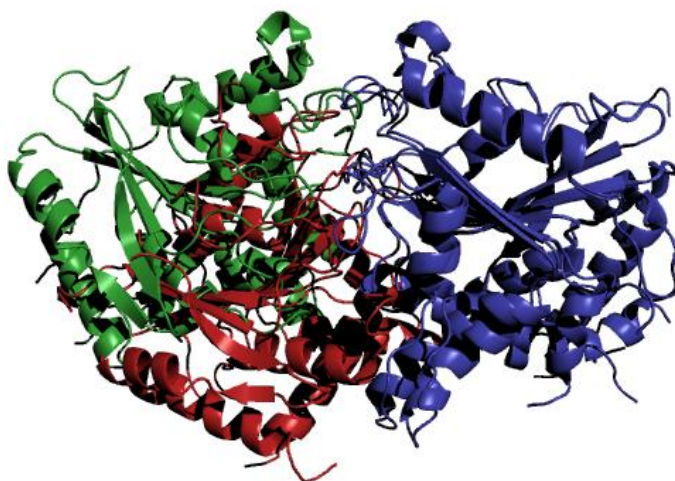


Figure 4.1.20. Best model of docking prediction for Target 23, hGBP1 homodimer complex. Blue: hGBP1 monomer; red: native hGBP1 monomer from complex; green: predicted hGBP1 monomer.

4.1.21. Target 24 and 25: Arf1 – ARHGap10

ARHGap10 is a novel human gene responsible for the coding a Rho family GTPase-activating protein (RhoGAP). RhoGAPs, on Golgi complex, regulate the activity of small GTPase proteins. ARHGap10 expression is detected high on brain and muscle which claims that it is closely related with cell differentiation. ARF1 is another small GTPase protein forms complex structure with ARHGap10 [95]. Target 24 and 25 aim to predict the correct complex structure between ADP-Ribosylation Factor 1 (Arf1) protein and ARHGap10 (ArfBD) protein.

For target 24 ARHGap10 is modelled from PDB 1BTN. PRISM returns 4 models constructed with 4 different interfaces. Table 4.1.21.1 shows the details of 4 models predicted by PRISM. The best model is selected as *afb1_0_adr1_1* with interface 1ej6CE which has I-rmsd value as 12.362 Å. All models for target 24 are classified as incorrect quality. One important detail is that, none of the models for target 24 has negative energy score which means PRISM would not show those models at all. It can be concluded that PRISM could not find the correct complex form for Target 24.

For target 25, ARHGap10 protein is given in its bound form. PRISM returns 13 models constructed with 8 different interfaces. Among those 13 models, 6 of them are classified other than incorrect. The best model is selected as *adr2_1_afb2_0* with interface 2j59BN which has I-rmsd value as 1.523 Å with 77% of native residue contacts. The best model is classified as medium quality.

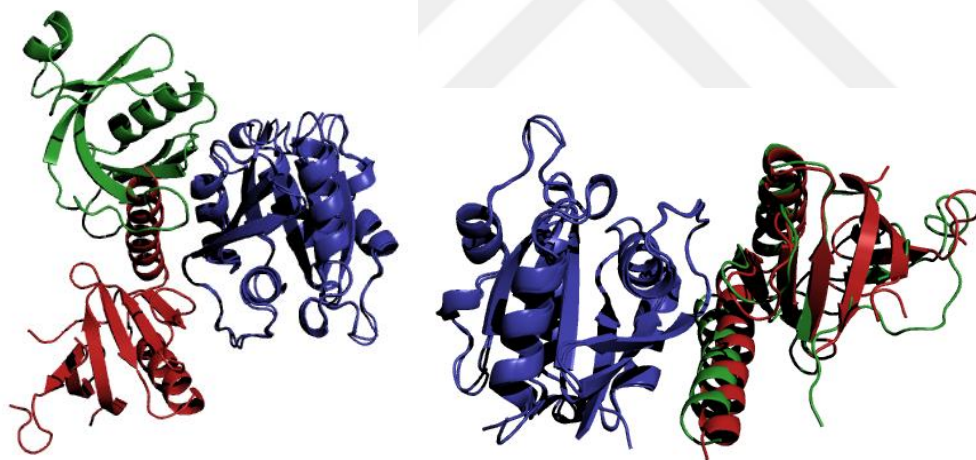
The results show that instead of protein with homology modeling, bound form of the same protein leads the algorithm towards the correct complex structure. Figure 4.1.21 shows the comparison of our best predicted models for both targets with the reference PDB structure 2J59.

Table 4.1.21.1. Evaluation Statistics of all predicted models for CAPRI Target 24.

	Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
Target 24	1ej6CE	afb1_0_adr1_1	11.60	12.362	37.936	Incorrect
	3f9sAB	adr1_1_afb1_1	33.77	14.423	54.688	Incorrect
	1yafBC	adr1_0_afb1_0	54.65	15.679	41.531	Incorrect
	3do8AB	adr1_2_afb1_1	67.39	16.152	51.628	Incorrect

Table 4.1.21.2. Evaluation Statistics of all predicted models for CAPRI Target 25.

	Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
Target 25	2j59BN	adr2_1_afb2_0	-51.87	1.523	0.769	0.459	3	3.609	Medium
	2j59BN	adr2_1_afb2_1	-51.87	1.523	0.769	0.459	3	3.609	Medium
	2j59BN	adr2_0_afb2_0	-49.64	1.618	0.769	0.512	3	3.921	Medium
	2j59BN	adr2_0_afb2_1	-49.64	1.618	0.769	0.512	3	3.921	Medium
	2j59BN	adr2_2_afb2_0	-49.64	1.618	0.769	0.512	3	3.921	Medium
	2j59BN	adr2_2_afb2_1	-49.64	1.618	0.769	0.512	3	3.921	Medium

**Figure 4.1.21** Best model of docking prediction for Target 24 (left) and Target 25 (right), Arf1 – ARHGap10 complex. Blue: Arf1; red: native ARHGap10; green: predicted ARHGap10.

4.1.22. Target 26: TolB - Pal

Tol system is a multiprotein complex found on the cell envelope of Gram-negative bacteria *Escherichia coli*. Tol system consists of 4 major proteins: TolA, TolB, TolR and TolQ [96]. TolB outshines from the other Tol proteins because of its role on outer cell membrane integrity in complex with PAL (peptidoglycan-associated lipoprotein) [97]. Pal has a role of providing a stable, non-covalent bridge between the outer membrane and the peptidoglycan layer by inserting itself into the outer membrane. On the other hand, TolB protein acts as a network hub, providing several other interactions between Pal, TolA and other outer membrane proteins like Lpp, OmpA [98]. Target 26 aims to predict the correct structure of complex between TolB protein and Peptidoglycan Associated Lipoprotein (Pal) protein.

PRISM returns 9 models constructed with 3 different interfaces. Among those 9 models, 6 of them classified other than incorrect. Table 4.1.22 shows the details of best 6 models predicted by PRISM. The best model is selected as *tolb_0_pplp_0* with interface 2hqsAH which has I-rmsd value as 1.323 Å with the 56% of native residue contacts. The model is classified as medium quality. One important detail is that, none those models has negative energy score, which means PRISM would not show those models at all. Figure 4.1.22 shows the comparison of our best predicted model with the reference PDB structure 2HQS.

Table 4.1.22. Evaluation Statistics of all predicted models for CAPRI Target 26.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
2hqsAH	tolb_0_pplp_0	50.75	1.323	0.559	0.297	13	5.918	Medium
2hqsAH	tolb_0_pplp_1	50.75	1.323	0.559	0.297	18	5.918	Medium
2hqsAH	tolb_0_pplp_2	50.75	1.323	0.559	0.297	18	5.918	Medium
2hqsAH	tolb_1_pplp_0	50.75	1.323	0.559	0.297	18	5.918	Medium
2hqsAH	tolb_1_pplp_1	50.75	1.323	0.559	0.297	18	5.918	Medium
2hqsAH	tolb_1_pplp_2	50.75	1.323	0.559	0.297	18	5.918	Medium

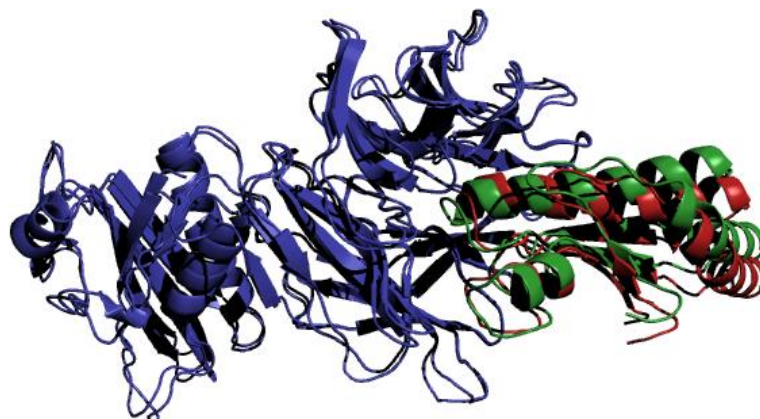


Figure 4.1.22. Best model of docking prediction for Target 26, TolB / Pal complex.

Blue: TolB; red: native Pal; green: predicted Pal.

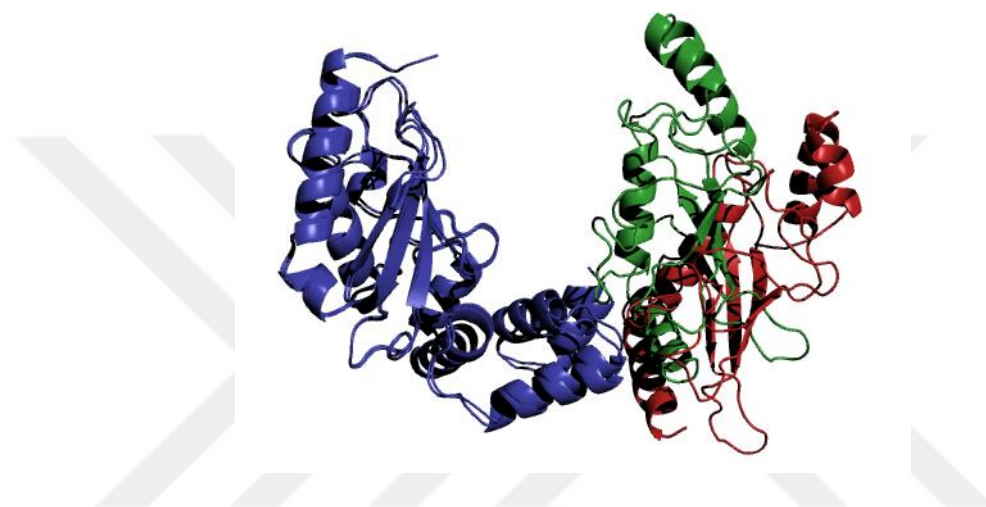
4.1.23. Target 27: Hip2 - Ubc9

Accumulation of abnormal proteins is a hallmark of neurodegenerative diseases. E2-25K/Hip2 is an ubiquitin-conjugating enzyme which is involved in regulation of amyloid- β neurotoxicity [99]. Ubiquitin-conjugating enzyme 9 (Ubc9) is an important enzyme necessary for the yeast cell *Saccharomyces cerevisiae* for the G₂-M transition of the cell cycle [100]. Target 27 aims to predict the correct structure of complex between Ubiquitin-Conjugating Enzyme (E2-25K) protein and Ubc9 protein.

PRISM returns 49 models constructed with 22 different interfaces. Among those 49 models, 4 of them classified other than incorrect. Table 4.1.23 shows the details of best 4 models predicted by PRISM. The best model is selected as *hip2_0_ubc9_0* with interface 2o25AD which has I-rmsd value as 2.125 Å with the 72% of native residue contacts. The model is classified as acceptable quality. One important detail is that, the model with lowest energy score is not the best model. The model *hip2_0_ubc9_1* has the lowest energy score among the best 6 models but does not have the lowest I-rmsd. Figure 4.1.23 shows the comparison of our best predicted model with the reference PDB structure 2O25.

Table 4.1.23. Evaluation Statistics of all predicted models for CAPRI Target 27.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
2o25AD	hip2_0_ubc9_1	-11.79	2.170	0.760	0.525	2	17.167	Acceptable
2o25AD	hip2_1_ubc9_1	-11.79	2.170	0.760	0.525	2	17.167	Acceptable
2o25AD	hip2_0_ubc9_0	-7.58	2.125	0.720	0.514	2	17.802	Acceptable
2o25AD	hip2_1_ubc9_0	-7.58	2.125	0.720	0.514	2	17.802	Acceptable

**Figure 4.1.23.** Best model of docking prediction for Target 27, Hip2 / Ubc9 complex. Blue: Hip2; red: native Ubc9; green: predicted Ubc9.

4.1.24. Target 28: NEDD4 Catalytic Domain Homodimer

NEDD4 is an enzyme also known as E3 ubiquitin-protein ligase which targets the proteins for ubiquitination. Ubiquitination is a very common way for protein regulation in the cell. NEDD4 is recruited to regulate many membrane proteins, membrane receptors and ion channels [101]. Target 28 aims to predict the correct structure of homodimer complex of NEDD4 protein.

PRISM returns 40 models constructed with 19 different interfaces. Among those 40 models, all of them are classified as incorrect. Table 4.1.24 shows the details of best 11 models predicted by PRISM. The best model is selected as *sur2_1_surf_0* with interface 2yc1AD which has I-rmsd value as 23.160 Å. No model has I-rmsd value less

than 4 Å. Since both I-rmsd and L-rmsd values are too high, no further evaluation would be necessary. It can be concluded that PRISM could not find the correct complex form for Target 28. Figure 4.1.24 shows the comparison of our best predicted model with the reference PDB structure.

Table 4.1.24. Evaluation Statistics of all predicted models for CAPRI Target 28.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
2yc1AD	sur2_1_surf_0	-8.75	23.160	54.681	Incorrect
1pv2GH	surf_1_sur2_1	0.35	32.699	99.799	Incorrect
2dwyAB	sur2_0_surf_2	7.63	24.159	88.942	Incorrect
1pv2GH	sur2_1_surf_1	9.79	32.580	68.975	Incorrect
2c9jEG	sur2_0_surf_1	10.95	21.088	49.121	Incorrect
3chiAB	surf_0_sur2_0	12.97	30.482	88.282	Incorrect
2c9jEG	surf_0_sur2_1	18.35	20.835	43.367	Incorrect
1ux9AB	surf_1_sur2_0	19.74	32.863	87.022	Incorrect
1g5uAB	sur2_0_surf_0	20.2	34.561	83.058	Incorrect
3chiAB	sur2_0_surf_0	20.52	31.096	86.893	Incorrect
1g5uAB	surf_0_sur2_0	21.02	35.030	84.833	Incorrect

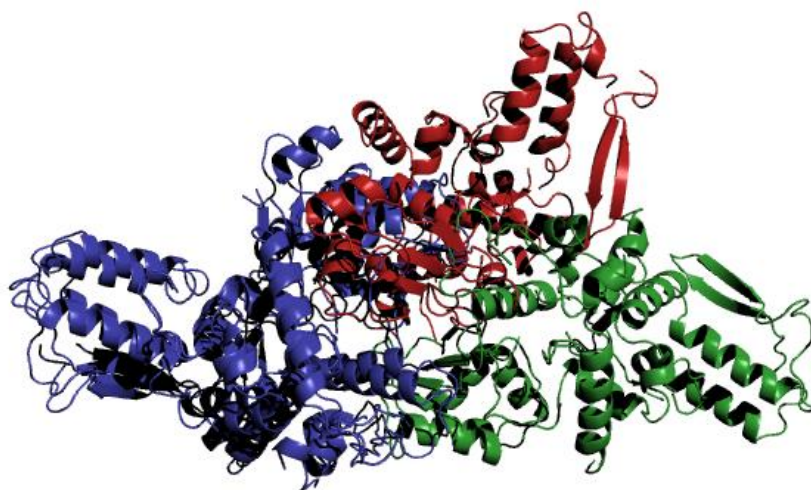


Figure 4.1.24. Best model of docking prediction for and Target 28, NEDD4L Catalytic Domain Homodimer complex. Blue: NEDD4L monomer; red: native NEDD4L monomer in complex; green: predicted NEDD4L monomer.

4.1.25. Target 29: tRNA Methyltransferase 8 - tRNA Methyltransferase 82

Modifications on nucleosides are common way for the regulation of many RNA related cell activity. Methylation is one of the extensive modifications for tRNA activity [102]. Trm8 and Trm82 are two proteins forms heterodimeric complex to catalyze the loss of N7-methylguanosine (m7G) modification which leads tRNA degradation pathway. Trm8 binds from the catalytic subunit and needs subtle conformational change [103]. Target 29 aims to predict the correct structure of complex between Trm8 protein and Trm82 protein.

PRISM returns 8 models constructed with 3 different interfaces. Among those 8 models, 6 of them classified other than incorrect. Table 4.1.25 shows the details of best 6 models predicted by PRISM. The best model is selected as *tr82_0_trm8_0* with interface 2vduBE which has I-rmsd value as 1.785 Å with the 57% of native residue contacts. The model is classified as medium quality. Figure 4.1.25 shows the comparison of our best predicted model with the reference PDB structure 2VDU.

Table 4.1.25. Evaluation Statistics of all predicted models for CAPRI Target 29.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
2vduBE	tr82_0_trm8_0	-7.92	1.785	0.574	0.446	2	5.032	Medium
2vduBE	tr82_1_trm8_0	-7.92	1.785	0.574	0.446	2	5.032	Medium
2vduBE	tr82_2_trm8_0	-7.92	1.785	0.574	0.446	2	5.032	Medium
2vduBE	tr82_0_trm8_1	-3.77	1.855	0.611	0.389	2	2.842	Medium
2vduBE	tr82_1_trm8_1	-3.77	1.855	0.611	0.389	2	2.842	Medium
2vduBE	tr82_2_trm8_1	-3.77	1.855	0.611	0.389	2	2.842	Medium

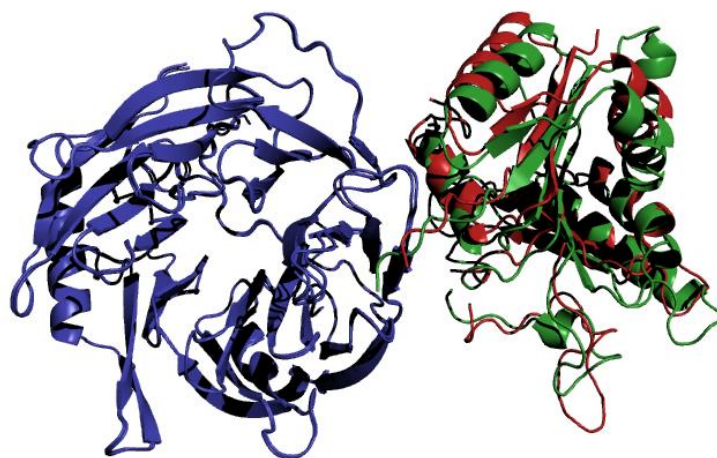


Figure 4.1.25. Best model of docking prediction for Target 29, tRNA Methyltransferase 8 / tRNA Methyltransferase 82 complex. Blue: tRNA Methyltransferase 82; red: native tRNA Methyltransferase 8; green: predicted tRNA Methyltransferase 8.

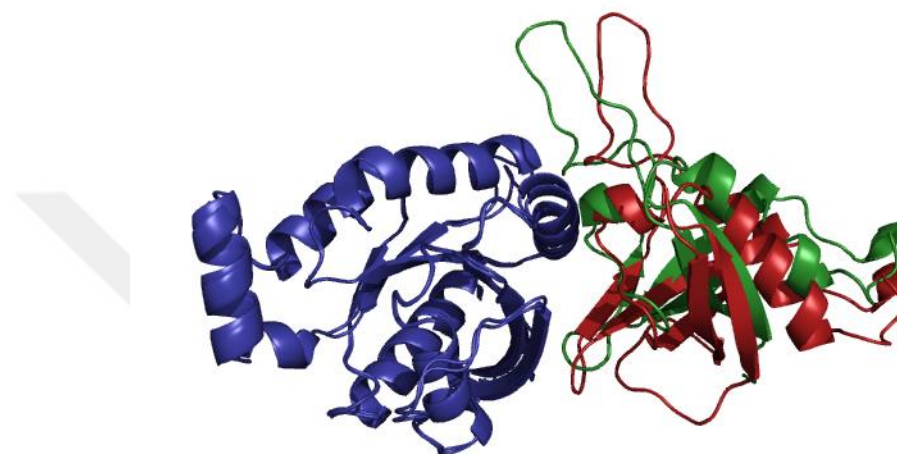
4.1.26. Target 30: Rho Family GTPase Rnd1 – Plexin-B1

Plexins are transmembrane receptor proteins located on cell surface specialized for semaphoring molecules [104, 105]. Plexin-B1 is a receptor for Semaphorin 4D (sema4D) which directly stimulates the guanosine triphosphatase (GTPase) activity of one of the members of the GTP-binding proteins Ras superfamily, R-Ras, to promote cell adhesion and cell growth. Rnd1 is a small GTP-binding protein from the Rho family. The interaction between Plexin-B1 and Rnd1 is needed for R-Ras activation in hippocampal neurons [106]. Target 30 aims to predict the correct structure of the complex between Rnd1 protein and Plexin-B1 protein.

PRISM returns 13 models constructed with 10 different interfaces. Among those 13 models, 2 of them are classified other than incorrect. Table 4.1.26 shows the details of the best 2 models predicted by PRISM. The best model is selected as *plb1_1_rhop_0* with interface 2rexAB which has I-rmsd value as 1.496 Å with the 73% of native residue contacts. The model is classified as medium quality. Figure 4.1.26 shows the comparison of our best predicted model with the reference PDB structure 2HQS.

Table 4.26. Evaluation Statistics of all predicted models for CAPRI Target 30.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
2rexAB	plb1_1_rhop_0	-54.6	1.496	0.727	0.396	3	4.1871	Medium
2rexAB	plb1_0_rhop_0	-29.27	1.893	0.568	0.286	3	6.4413	Medium

**Figure 4.1.26.** Best model of docking prediction for Target 30, Rho Family GTPase Rnd1 / Plexin-B1 complex. Blue: Rnd1; red: native Plexin-B1; green: predicted Plexin-B1.**4.1.27. Target 32: Barley- α -Amylase Subtilis Inhibitor – Subtilis Savinase**

Subtilisins are endopeptidases belong to large class of microbial serine proteases secreted by the *Bacillus* species. Active site of that proteins located on the shallow groove on the surface [107]. Barley- α -Amylase enzyme has a role of starch degradation during seed germination. Barley- α -Amylase Subtilis Inhibitor (BASI) inhibits amylase and prevents the degradation of endosperm starch to protect the seed from attack by pathogens [108]. Target 32 aims to predict the correct structure of complex between Protease savinase protein and Bifunctional Inhibitor BASI protein.

PRISM returns 10 models constructed with 6 different interfaces. Among those 10 models, 4 of them classified other than incorrect. Table 4.1.27 shows the details of

best 4 models predicted by PRISM. The best model is selected as *savi_0_baip_0* with interface 3bx1AC which has I-rmsd value as 1.072 Å with the 80% of native residue contacts. The model is classified as medium quality. Figure 4.1.27 shows the comparison of our best predicted model with the reference PDB structure 3BX1.

Table 4.1.27. Evaluation Statistics of all predicted models for CAPRI Target 32.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
3bx1AC	savi_0_baip_0	-58.49	1.072	0.802	0.216	7	2.811	Medium
3bx1AC	savi_1_baip_0	-58.49	1.072	0.802	0.216	7	2.811	Medium
3bx1AC	savi_0_baip_1	-53.26	1.487	0.779	0.368	7	3.670	Medium
3bx1AC	savi_1_baip_1	-53.26	1.487	0.779	0.368	7	3.670	Medium

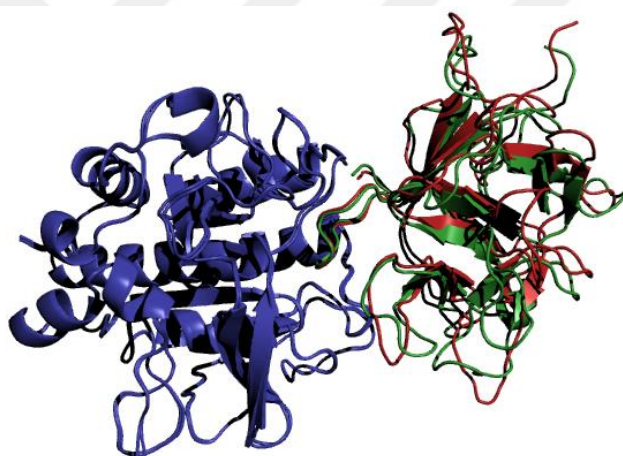


Figure 4.1.27. Best model of docking prediction for Target 32, Barley- α -Amylase Subtilis Inhibitor / Subtilis Savinase complex. Blue: Subtilis Savinase; red: native Subtilis Inhibitor; green: predicted Subtilis Inhibitor.

4.1.28. Target 35 and 36: CBM22-1 of Xylanase – GH10 of Xylanase

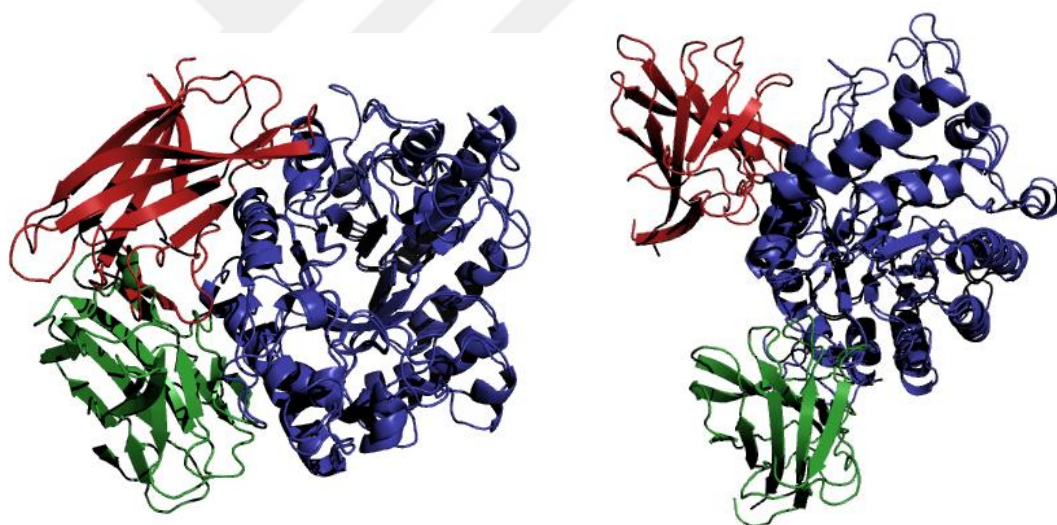
Xyn10B is a multifunctional protein derived from the organism *Clostridium thermocellum* which performs the xylanase and esterase activities. It has multiple domains with different functions. CBM22-1 and GH10 are the subdomains of xylanase protein. CBM22-1 has a special fold called β -jelly-roll and GH10 has typical β/α barrel fold [109]. Target 35 and Target 36 aim to predict the correct structure of complex between CBM22-1 of xylanase protein and GH10 of xylanase protein. Homology modelled GH10 is used for both targets but bound CBM22-1 module is replaced by homology modeled CBM22-1 for Target 36.

For target 35, homology modeled GH10 and CBM22 are used as initial structures. PRISM returns 6 models constructed with 6 different interfaces. None of those 6 models is classified other than incorrect. Table 4.1.28 shows the details of all models predicted by PRISM. The best model is selected as *xyla_0_xber_0* with interface 3fxdBD which has I-rmsd value as 9.753 Å.

For target 36, CBM22 is replaced by its bound conformation. PRISM returns 16 models constructed with 15 different interfaces. None of those 16 models is classified other than incorrect. Table 4.1.28 shows the details of all models predicted by PRISM. The best model is selected as *xyae_0_xnbd_0* with interface 2a24AB which has I-rmsd value as 17.334 Å. One important detail is that, it is expected to be more successful with target 36 because of the bound form of CBM22. However, all models on both targets are classified as incorrect. It can be concluded that PRISM could not find the correct form of complex for both targets. Figure 4.1.28 shows the comparison of our best predicted models with the reference PDB structure 2W5F.

Table 4.1.28. Evaluation Statistics of all predicted models for CAPRI Targets 35 and 36.

	Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
Target 35	3fxdBD	xyla_0_xbet_0	-17.99	9.753	29.404	Incorrect
	3jygAB	xbet_0_xyla_1	17.84	15.703	31.414	Incorrect
	2ywbCD	xyla_0_xbet_1	22.22	11.456	19.802	Incorrect
	2bnlBC	xbet_0_xyla_0	25.97	22.172	66.776	Incorrect
	1p27AC	xbet_1_xyla_2	38.76	13.248	20.160	Incorrect
	1zdnAB	xbet_0_xyla_0	39.22	17.263	35.768	Incorrect
Target 36	2a24AB	xyae_2_xnbd_0	-23.21	17.334	38.439	Incorrect
	1u2eAC	xyae_0_xnbd_1	-13.32	10.385	25.224	Incorrect
	2pnoGI	xnbd_0_xyae_1	-11.92	19.657	63.352	Incorrect
	2a24AB	xyae_2_xnbd_1	1.22	15.003	27.193	Incorrect
	2iabAB	xnbd_2_xyae_2	8.62	14.314	35.077	Incorrect
	2pcrBD	xyae_2_xnbd_0	10.17	16.253	37.164	Incorrect

**Figure 4.1.28.** Best model of docking prediction for Target 35 (left) and Target 36 (right), CBM22-1 of Xylanase – GH10 of Xylanase complex. Blue: GH10; red: native CBM22; green: predicted CBM22.

4.1.29. Target 37: Leucine Zipper 2 – ADP-ribosylation Factor (ARF6)

ADP-ribosylation factor (ARF) is small, GTP-binding protein which has a role in regulation membrane traffic and organization of actin cytoskeleton. ARF6 is located in the plasma membrane regulating receptors and lipids [110]. Leucine zipper is dimerized two alpha-helix monomers with repeated Leucine residue which promotes the dimerization / oligomerization of proteins. Target 37 aims to predict the correct structure of complex between Leucine Zipper 2 protein and ARF6 protein.

PRISM returns 21 models constructed with 8 different interfaces. Among those 21 models, 1 of them has I-rmsd value less than 4 Å. Table 4.1.29 shows the details that model predicted by PRISM. The name of that model is *arf6_0_lzp2_0* with interface 2kwiAB which has I-rmsd value as 3.080 Å. The problem is the sequence identity of Leucine Zipper 2 proteins. The common residues are too rare to compare the native contacts of our predicted models. Since the Fnat is calculated as 4.1% the model classified as incorrect. It can be concluded that PRISM could not find the correct complex form for Target 37. Figure 4.1.29 shows the comparison of our best predicted model with the reference PDB structure 2W83.

Table 4.1.29. Evaluation Statistics of all predicted models for CAPRI Target 37.

Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
2kwiAB	arf6_0_lzp2_0	16.42	3.080	0.041	0.939	7	30.406	Incorrect

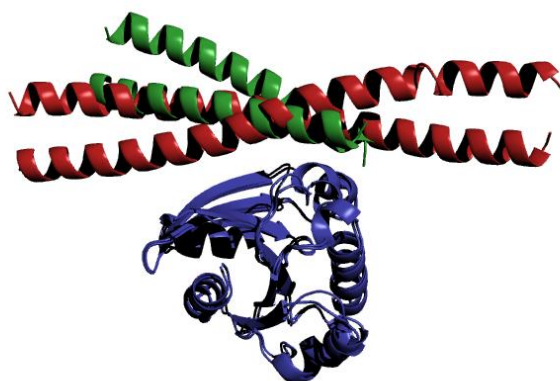


Figure 4.1.29. Best model of docking prediction for Target 37, Leucine Zipper 2 / ADP-ribonylation Factor (ARF6) complex. Blue: ARF6; red: native Leucine Zipper 2; green: predicted Leucine Zipper 2.

4.1.30. Target 38 – 39: Centurin- α 1 – FHA domain of KIF13B

Centurin- α 1 protein is a GTPase activating protein for ARF family proteins. Motor protein KIF13B's stalk domain and Centurin- α 1's protein GAP domain interaction is highly expressed in brain and kidney. This interaction is essential for Centurin- α 1 localization to the leading edges of the cell [111]. Target 38 and 39 aim to predict the correct structure of complex between Centurin- α 1 protein and motor protein KIF13B.

For target 38, PRISM returns 8 models constructed with 5 different interfaces. Among those 8 models, one is classified other than incorrect. Table 4.1.30 shows the details of that model predicted by PRISM. The model is named as *kif1_1_cea1_1* with interface 3mdbBD which has I-rmsd value as 2.989 Å with 26% of native residue contacts. That model is classified as acceptable. Unfortunately, that model has positive energy score, which means PRISM would not show that model at all.

For target 39, KIF13B is replaced by its bound conformation. PRISM returns 10 models constructed with 5 different interfaces. Among those 10 models, 6 models are classified other than incorrect. Table 4.1.30 shows the details of all models predicted by

PRISM. The best model is selected as *kif2_0_cea2_1* with interface 3mdbBD which has I-rmsd value as 0.757 Å with 80% native residue contacts. That model is classified as high. One important detail is that, the models with lower energy scores are qualified as medium, even though they have almost 90% of native residue contacts. Figure 4.1.30 shows the comparison of our best predicted models with the reference PDB structure 3FM8.

Table 4.1.30. Evaluation Statistics of all predicted models for CAPRI Targets 38 and 39.

	Interface Name	Model Name	Energy Score	I-rmsd	Fnat	Fnonnat	Clash	L-rmsd	Classification
Target 38	3mdbBD	kif1_1_cea1_1	7.14	2.989	0.25 5	0.606	6	5.315	Acceptable
	3mdbBD	kif2_1_cea2_0	-29.92	1.202	0.90 2	0.233	6	2.238	Medium
	3mdbBD	kif2_0_cea2_0	-29.82	1.235	0.88 2	0.237	6	2.334	Medium
Target 39	3mdbBD	kif2_2_cea2_0	-29.82	1.235	0.88 2	0.237	6	2.334	Medium
	3mdbBD	kif2_1_cea2_1	-29.35	1.108	0.84 3	0.246	6	2.122	Medium
	3mdbBD	kif2_0_cea2_1	-29.33	0.757	0.80 4	0.268	6	1.612	High
	3mdbBD	kif2_2_cea2_1	-29.33	0.757	0.80 4	0.268	6	1.612	High

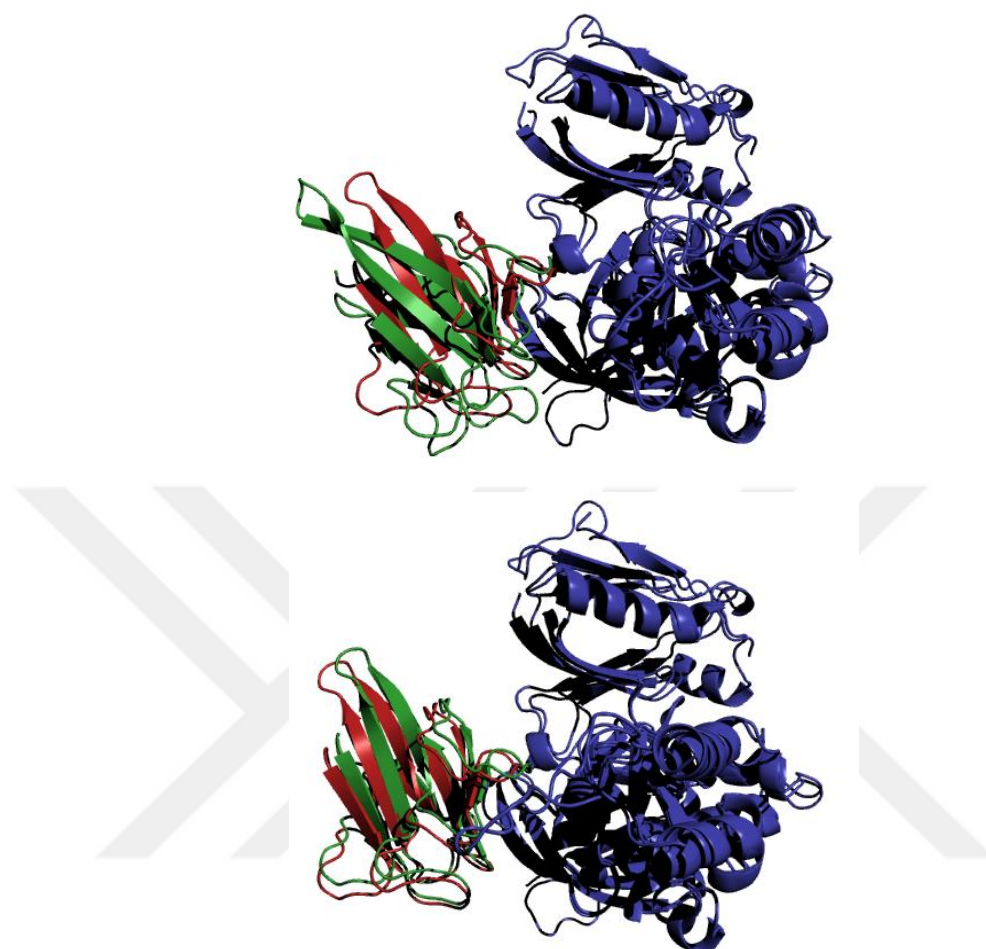


Figure 4.1.30. Best model of docking prediction for Target 38 (left) and Target 39 (right), Centurin- α 1 / FHA domain of KIF13B complex. Blue: Centurin- α 1; red: native KIF13B; green: predicted KIF13B.

4.1.31. Target 40: Bovine Trypsin - APIA

Bovine Trypsin is a member of setine proteases which degrades proteins by cleaving the peptide bonds. Double-headed arrowhead protease inhibitors (APIAs) derived from *Sagittaria sagittifolia* has two distinct reactive sites that can inhibit a variety of serine proteases including trypsin [112]. Target 40 aims to predict the correct structure of complex between Bovine Trypsin protein and APIA protein.

PRISM returns 7 models constructed with 4 different interfaces. Among those 7 models, none of them classified other than incorrect. Table 4.1.31 shows the details of all 7 models predicted by PRISM. The best model is selected as *btyp_1_apia_0* with interface 2qyiCD which has I-rmsd value as 12.651 Å. No model has I-rmsd value less than 4 Å. Since both I-rmsd and L-rmsd values are too high, no further evaluation would be necessary. It can be concluded that PRISM could not find the correct complex form for Target 40. Figure 4.1.31 shows the comparison of our best predicted model with the reference PDB structure 3E8L.

Table 4.1.31. Evaluation Statistics of all predicted models for CAPRI Target 40.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
2qyiCD	btyp_1_apia_0	-16.55	12.651	27.476	Incorrect
2qyiCD	btyp_0_apia_0	-5.35	11.861	24.885	Incorrect
2qyiCD	btyp_2_apia_0	-5.35	11.861	24.885	Incorrect
3nr1AB	btyp_2_apia_2	-5.03	18.501	64.042	Incorrect
3s28BC	apia_1_btyp_1	18.76	17.857	58.914	Incorrect
3nr1AB	apia_2_btyp_0	19.04	19.627	62.959	Incorrect
3mqqAB	btyp_0_apia_1	35.21	19.900	60.097	Incorrect

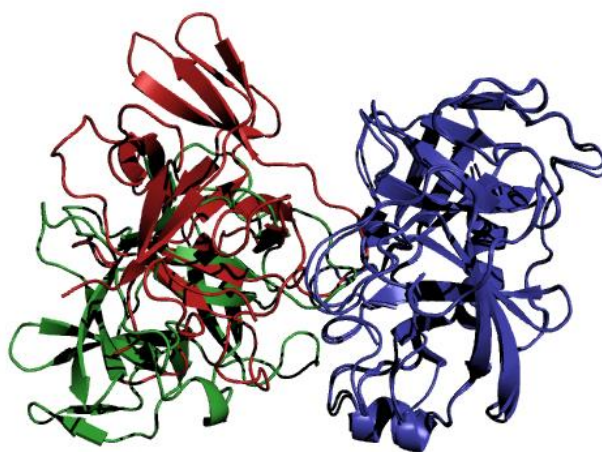


Figure 4.1.31. Best model of docking prediction for Target 40, Bovine Trypsin / APIA complex. Blue: Bovine Trypsin; red: native APIA; green: predicted APIA.

4.1.32. Target 41: Colicin E9 – Immunity Protein 2

Colicins are bacteria producing endonucleases produced at the times of nutrient shortage or environmental stress on the purpose of reducing competition from other microbial populations [113]. Colicins could be inactivated by small immunity proteins binding high affinity and specificity to protect cells from cytotoxic activity of endonucleases. The immunity protein Im2 can inhibit the endonuclease domain of toxin colicin E9 [114]. Target 41 aims to predict the correct structure of complex between Colicin E9 protein and Immunity Protein 2 protein.

PRISM returns 2 models constructed with 2 different interfaces. Both models are classified as incorrect. Table 4.1.32 shows the details of these 2 models predicted by PRISM. The best model is selected as *cln2_0_cln9_1* with interface 3zu0AB which has I-rmsd value as 14.759 Å. . No model has I-rmsd value less than 4 Å. Since both I-rmsd and L-rmsd values are too high, no further evaluation would be necessary. It can be concluded that PRISM could not find the correct complex form for Target 41. Figure 4.1.32 shows the comparison of our best predicted model with the reference PDB structure 2WPT.

Table 4.1.32. Evaluation Statistics of all predicted models for CAPRI Target 41.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
3zu0AB	cln2_0_cln9_1	22.9	14.759	36.542	Incorrect
3putAB	cln9_0_cln2_0	30.81	18.899	55.364	Incorrect

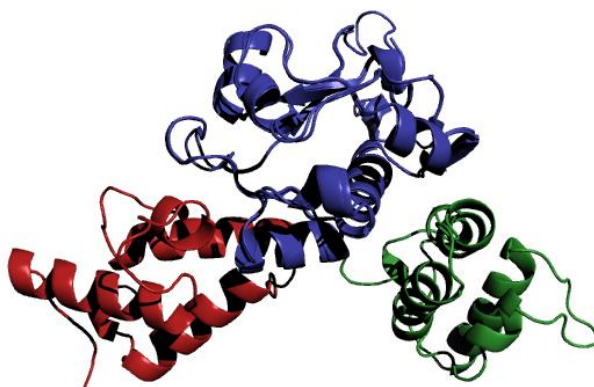


Figure 4.1.32. Best model of docking prediction for Target 41, Colicin E9 – Immunity Protein 2 complex. Blue: Colicin E9; red: native Immunity Protein 2; green: predicted Immunity Protein 2.

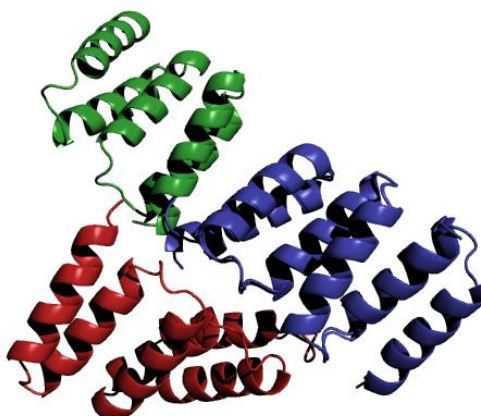
4.1.33. Target 42: Tetratricopeptide Repeat (TPR) homodimer

Tetratricopeptide repeat (TPR) is a special motif which organizes the protein-protein interactions. TPR has a role in transcriptional regulation, mRNA processing, protein folding and translocation. It consists of asymmetric trimer of helix-turn-helix structure. It has been discovered that the C-terminal domain with hydrophilic capping helix found in most TPRs is key for oligomerization [115]. Target 42 aims to predict the correct complex structure of homodimer TPR protein.

PRISM returns 52 models constructed with 14 different interfaces. Among those 52 models, none of them shows the correct complex formation. No model has I-rmsd value less than 4 Å. Since both I-rmsd and L-rmsd values are too high, no further evaluation would be necessary. Table 4.1.33 shows the details of best 11 models predicted by PRISM. According to the energy score, the best model is selected as *ctpa_0_ctpb_0* with interface 2q00AB which has I-rmsd value as 12.994 Å which is classified as incorrect. It can be concluded that PRISM could not find the correct complex form for Target 42. Figure 4.1.33 shows the comparison of our best prediction (lowest energy) with the reference complex structure 2WQH.

Table 4.1.33. Evaluation Statistics of all predicted models for CAPRI Target 42.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
2q00AB	ctpa_0_ctpb_0	-22.65	12.994	39.008	Incorrect
2q00AB	ctpb_0_ctpa_0	-20.52	13.042	36.853	Incorrect
2w2uAB	ctpb_0_ctpa_0	7.62	14.345	44.308	Incorrect
2fi7AB	ctpa_0_ctpb_0	9	12.514	33.398	Incorrect
2fi7AB	ctpb_0_ctpa_0	9.16	12.572	33.465	Incorrect
2w2uAB	ctpb_1_ctpa_0	14.56	14.693	41.868	Incorrect
2w2uAB	ctpa_0_ctpb_0	16.6	14.385	40.894	Incorrect
3ks2AB	ctpb_0_ctpa_0	17.13	14.257	35.849	Incorrect
3sidAB	ctpb_0_ctpa_0	21.34	13.668	41.343	Incorrect
3ks2AB	ctpa_0_ctpb_0	21.63	14.726	36.595	Incorrect
1em1AB	ctpa_0_ctpb_0	21.89	13.400	39.629	Incorrect

**Figure 4.1.33** Best model of docking prediction for Target 42, Tetrapeptide Repeat (TPR) homodimer complex. Blue: TPR; red: second dimer of native TPR; green: second dimer of predicted TPR.

4.1.34. Target 46: Methyltransferase Mtq2 – Trm112

Methylation is one of the most important types of modification applied on DNA, RNA and many proteins. It plays a key role in gene expression, mRNA translation and protein function. In eukaryotes, this modification could be performed by the complex enzyme called Mtq2-Trm112 holoenzyme. tRNA methyltransferase 112 (Trm112) activates the catalytic subunit of Mtq2 and other Trm proteins [116]. Target 46 aims to predict the correct structure of complex between Trm112 protein and Mtq2 protein.

PRISM returns 8 models constructed with 8 different interfaces. Among those 8 models, none of them shows the correct complex formation. No model has I-rmsd value less than 4 Å. Since both I-rmsd and L-rmsd values are too high, no further evaluation would be necessary. Table 4.1.34 shows the details of all 8 models predicted by PRISM. According to the energy score, the best model is selected as *mtq2_0_trm1_2* with interface 3do8AB which has I-rmsd value as 18.152 Å which is classified as incorrect. It can be concluded that PRISM could not find the correct complex form for Target 46. Figure 4.1.34 shows the comparison of our best prediction (lowest energy) with the reference complex structure 3Q87.

Table 4.1.34. Evaluation Statistics of all predicted models for CAPRI Target 46.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
3do8AB	mtq2_0_trm1_2	-26.85	18.152	48.817	Incorrect
3eo6AB	trm1_0_mtq2_0	-19.03	16.514	47.218	Incorrect
1tilEF	mtq2_0_trm1_1	-7.96	17.131	27.043	Incorrect
2f66AB	trm1_0_mtq2_1	16.28	19.827	47.101	Incorrect
3chiAB	mtq2_1_trm1_0	20.75	18.449	49.062	Incorrect
1sg9AC	mtq2_0_trm1_0	20.78	22.040	48.500	Incorrect
1ylmAB	trm1_0_mtq2_1	27.17	19.057	37.353	Incorrect
2gmyBC	trm1_2_mtq2_0	43.03	20.910	57.824	Incorrect

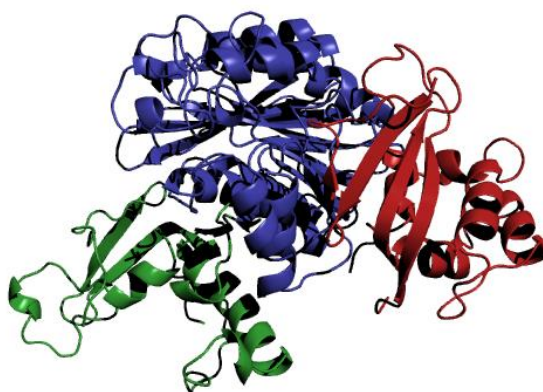


Figure 4.1.34. Best model of docking prediction for Target 46, Mtq2 / Trm112 complex. Blue: Trm112; red: native Mtq2; green: predicted Mtq2.

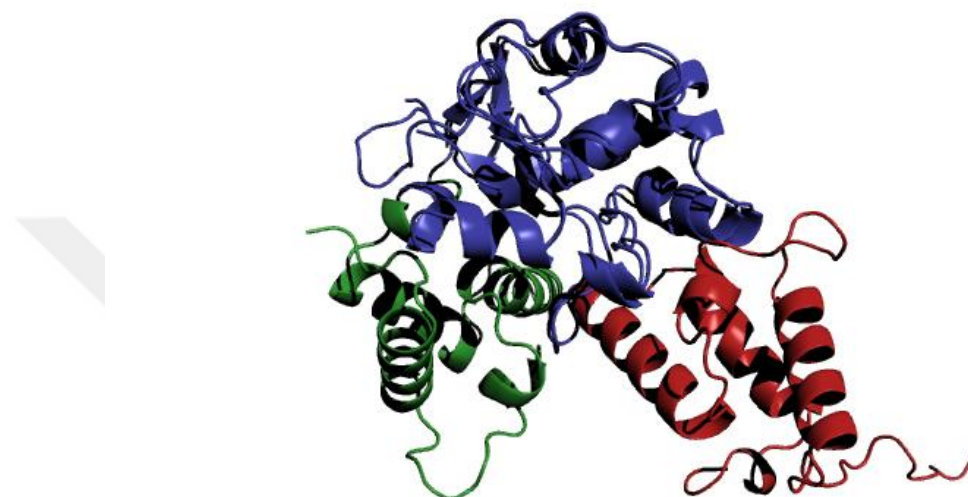
4.1.35. Target 47: Colicin E2 Immunity Protein – Colicin E9

Colicin proteins are inducible antibiotics encoded by plasmids specifically to attack the cells of *Escherichia coli*. Colicin group E proteins (E2, E7, E8, and E9) degrade bacterial genome through the C-terminal DNase domain [117]. Immunity proteins (Im) generally perform high affinity interactions with Colicins. They have conserved and variable helical structures that binds hypervariable IPE domain on DNase surface [118]. Target 47 aims to predict the correct structure of complex between Im2 protein and Colicin E9 protein.

PRISM returns 4 models constructed with 4 different interfaces. Among those 4 models, none of them shows the correct complex formation. No model has I-rmsd value less than 4 Å. Since both I-rmsd and L-rmsd values are too high, no further evaluation would be necessary. Table 4.1.35 shows the details of all models predicted by PRISM. According to the energy score, the best model is selected as *che9_0_che2_1* with interface 2pmyAB which has I-rmsd value as 10.52 Å which is classified as incorrect. It can be concluded that PRISM could not find the correct complex form for Target 47. Figure 4.1.35 shows the comparison of our best prediction (lowest energy) with the reference complex structure.

Table 4.1.35. Evaluation Statistics of all predicted models for CAPRI Target 47.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
2pmyAB	che9_0_che2_1	8.05	10.52	28.239	Incorrect
2i8bAB	che2_1_che9_0	35.34	9.182	26.442	Incorrect
2wymDF	che9_1_che2_0	53.91	15.012	50.625	Incorrect
3f6wCD	che9_1_che2_1	60.26	13.867	43.670	Incorrect

**Figure 4.1.35.** Best model of docking prediction for Target 47, Colicin E2 / Colicin E9 complex. Blue: Colicin E9; red: native Colicin E2; green: predicted Colicin E2.

4.1.36. Target 50: Hemagglutinin Receptor Binding Domain and Membrane Fusion Domain – Hypothetical Protein

Hemagglutinin (HA) is a special protein secreted by the influenza viruses. During a virus invasion, HA has a role as an antigen and binds to human antibody Fab proteins to neutralize them. That makes HA a prime candidate for drug development. Specially designed proteins generally successful to bind the head of HA which is a region that can rapidly vary. HA stem is more conserved region among all group 1 influenza stains which could be targeted recently by two antibodies CR6261 and F10 [119, 120]. Target 50 aims to predict the correct structure of complex between Hemagglutinin Receptor Binding Domain HA1 with Hemagglutinin Membrane Fusion Domain HA2 protein and Hypothetical protein.

PRISM returns 29 models constructed with 20 different interfaces. Among those 29 models, none of them shows the correct complex formation. No model has I-rmsd value less than 4 Å. Since both I-rmsd and L-rmsd values are too high, no further evaluation would be necessary. Table 4.1.36 shows the details of best 11 models predicted by PRISM. According to the energy score, the best model is selected as *hypp_0_hmab_2* with interface 2y43AB which has I-rmsd value as 9.649 Å which is classified as incorrect. It can be concluded that PRISM could not find the correct complex form for Target 50. Figure 4.1.36 shows the comparison of our best prediction (lowest energy) with the reference complex structure.

Table 4.1.36. Evaluation Statistics of all predicted models for CAPRI Target 50.

Interface Name	Model Name	Energy Score	I-rmsd	L-rmsd	Classification
2y43AB	hypp_0_hmab_2	-41.68	9.649	17.934	Incorrect
1hssAB	hmab_0_hypp_1	-24.88	16.959	39.938	Incorrect
3l49CD	hypp_1_hmab_2	-13.39	16.008	47.298	Incorrect
3oq2AB	hmab_1_hypp_0	-12.68	12.154	29.840	Incorrect
1xg5CD	hmab_1_hypp_0	-8.38	21.138	58.156	Incorrect
1gylAB	hypp_0_hmab_0	-5.67	14.128	35.798	Incorrect
3o0nAB	hmab_2_hypp_0	-3.06	19.982	49.961	Incorrect
1gylAB	hypp_1_hmab_2	-2.47	16.498	39.956	Incorrect
1gylAB	hypp_1_hmab_1	0.35	16.877	38.408	Incorrect
2vrnAB	hypp_0_hmab_2	7.09	11.76	30.687	Incorrect
2owrBD	hmab_2_hypp_1	9.26	13.424	38.409	Incorrect

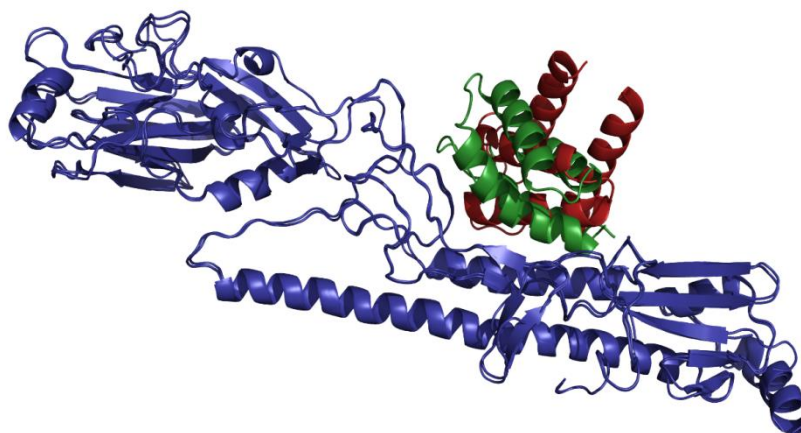


Figure 4.1.36. Best model of docking prediction for Target 50, HA Receptor Binding Domain and HA Membrane Fusion Domain / Hypothetical Protein complex. Blue: HA domains; red: native Hypothetical Protein; green: predicted Hypothetical Protein

4.2. Examining the Effect of Monte Carlo Minimization on FiberDock Scoring

The principle of FiberDock algorithm is explained before on Chapter 2.2.3.4. Before assigning the final score for a prediction, FiberDock iterates 50 cycles of Monte Carlo (MC) minimization steps. Basically, minimization aims to reduce the final energy score by slightly changing the conformation of complex until finding the best conformation with lowest energy.

In this study, total of 812 models were predicted from 42 targets. The same predictions were performed with a change on the 50 steps of MC minimizations which is reduced to zero. Then the same models were compared according to their two energy scores. It is expected to see that, the energy score would be reduced when MC minimization is enabled. On one hand, minimization steps recover 91 models with positive energy scores and make their energies negative. For default PRISM algorithm, those models would be predicted complex structures appeared as results. Moreover, minimization steps reduce the energy of 610 models and make those models much more

stable on their chemical conformation. For the 75% of the data, FiberDock has done its job accurately as expected.

However, 202 models (out of 812 models) have smaller value when the MC minimization is shut. It means that, for some cases minimization steps work on the contrary. 25% of the data effected negatively because of MC minimization. Figure 4.2.1 shows the point comparison graph for the same models both with MC minimization cycle equals zero and equals fifty. In the graph, the black points show the negatively affected models and red points show the models without any problem with minimization.

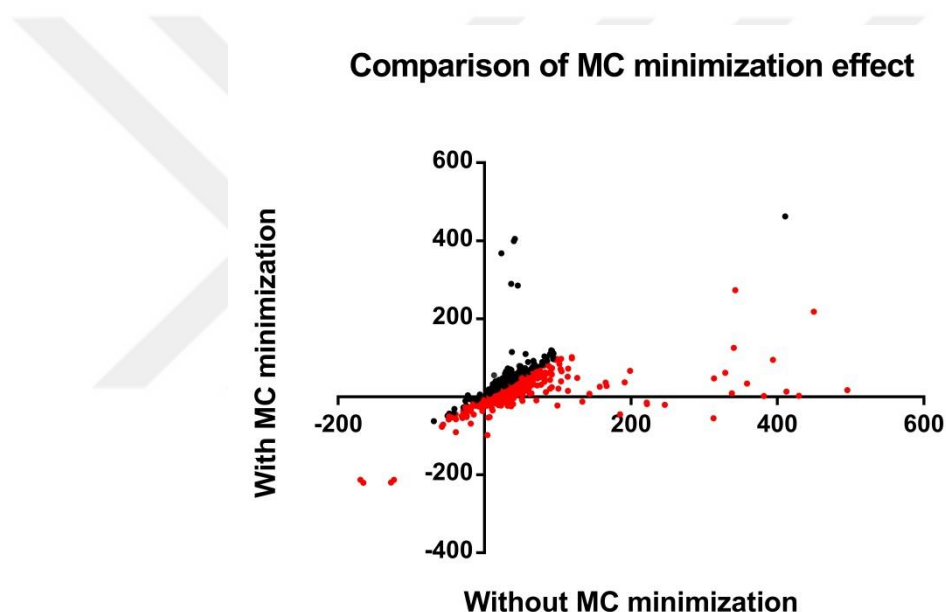


Figure 4.2.1. The graph shows energy value comparison between MC minimization equals 0 and MC minimization equals 50. Black dots are responsible for the negatively affected models with increased energy scores after MC minimization. Red dots are responsible for the models with lower energy scores after MC minimization.

4.3. Comparison of results from PRISM scoring with FiberDock and scoring with RosettaDock

Scoring functions are essential elements for a docking procedure to analyze the decoy predicted form algorithm. A good scoring function should be able to identify the predicted complexes with closest structure to the native complex structure. Here, we try to compare the two effective scoring functions. One belongs to FiberDock which is used by PRISM and the other function belongs to RosettaDock which is another widely used docking algorithm. The basic principles for both scoring function are explained on the following parts. The end of this chapter shows the comparison results of these two scoring functions.

4.3.1 FiberDock Scoring Function

The energy function of the structure minimization procedure is composed of the attractive and repulsive van der Waals energy and penalty deformation energy term that prevents the minimization from over-distorting the structure. The following equation is used to calculate binding energy:

$$E = KE_{\text{attrVdW}} + E_{\text{repVdW}} + \lambda \sum_{i=1}^M (V_i^{\text{freq}})^2 |V_i^{\text{amp}}|, \quad (7)$$

Figure 4.3.1: FiberDock scoring function formulation [5].

In this equation, K represents weight of the attractive van der Waals term which yields the best performance on the value of 5, M represents the number of normal modes, λ is a scaling factor which was set to 0.05. V_i^{freq} denotes the frequency of the i^{th} normal mode and V_i^{amp} denotes its amplitude. This energy function is also used for scoring the model.

4.3.2 RosettaDock Scoring Function

RosettaDock gives two types of scores: total score and I_sc. I_sc is the "interface score," which is the total score of the complex minus the sum of the scores of the individual docking partners. Values of about -5 to -10 typically indicate a good decoy. It is important to also look at the total score, as a clash or other problem in one of the partners can distort the I_sc significantly. These problems are most common when incorporating flexible backbones into docking.

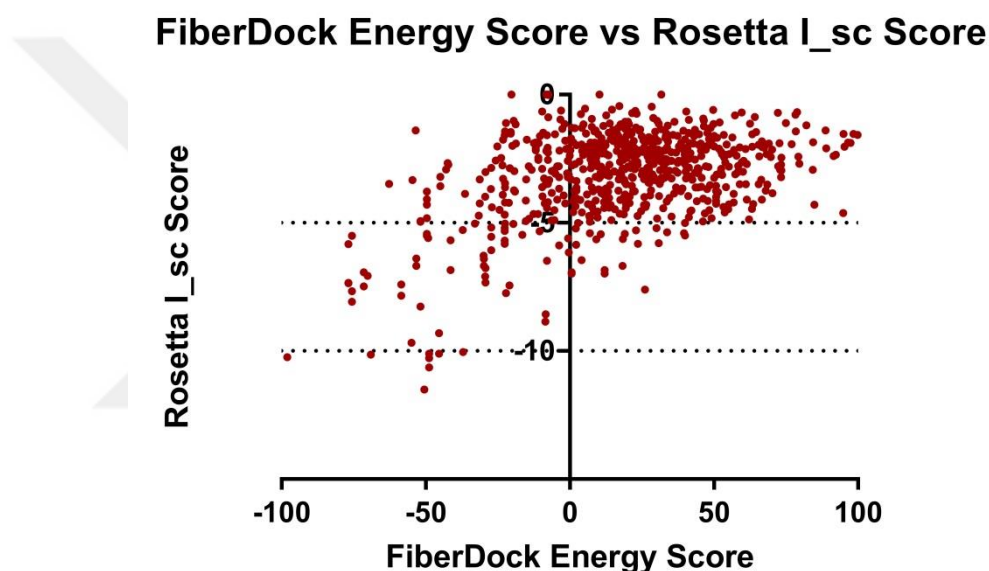


Figure 4.3.2. The graph shows energy value comparison between FiberDock Energy score and Rosetta I_sc score.

In this part of the study, the PRISM is modified with zero MC minimization cycle to avoid the possible modifications on complex conformation. Then the models are used as input for Rosetta Local Docking to reduce the energy by rotating the backbone and side chains of the complex structures. Then, for each model, the I_sc scores are recorded and compared the same model's FiberDock score which were previously calculated by the default PRISM (with 50 cycles of MC minimization on the

last step of FiberDock algorithm). Figure 4.3.2 shows the comparison of FiberDock energy scores with RosettaDock I_{sc} scores.

In this part, total 750 models are used for comparison. The majority of the models, 70% of all, are located on the top right corner of the graph which is the area that models have positive FiberDock energy score and Rosetta I_{sc} score larger than -5. That is consistent with the expectations that the models with positive energy scores are also considered as incorrect with Rosetta scoring. 72 models are plotted between -5 and -10 according to their I_{sc} score. On the other hand, according to FiberDock energy score, 208 models have negative energy score which shows almost 3 times more possible correct model is identified by FiberDock. Detailed examination shows that, Rosetta misses almost 45% of the models FiberDock identifies. Important point in here is that for targets 4, 8 and 20, Rosetta could identify the correct models successfully even though they have positive energy score from FiberDock. However, for target 26, both scoring functions have failed to identify the models with lowest I-rmsd scores.

Overall, those two scoring functions have their own advantages and disadvantages from each other. Their performance may vary from target to target. Rosetta uses a strategy to compress any models between -5 and -10 in terms of I_{sc} score. In that perspective FiberDock could cover much more models and increases its change to guess the correct ones. The scoring accuracy of both models is controversial. However, at the end of the study, we can conclude that FiberDock could find more models than RosettaDock.

Chapter 5 – New Parameters

PRISM is a compound algorithm uses three other tools: NACCESS, MULTIPROT and FiberDock. Each tool contains several modifiable parameters which indirectly affects the performance of PRISM. In this chapter, the effects of changes in the algorithms over the PRISM are tested and the results are compared with default PRISM.

For this chapter, targets are selected among the targets with no correct predictions in previous chapter which are T02, T09, T10, T23, T24, T28, T35, T36, T40, T41, T46 and T47. The purpose is testing whether changing some of the parameters could increase the performance of PRISM or not. As it was mentioned before, MULTIPROT aligns the template interfaces on the each protein's surface. The alignment of two fragments is checked by the maximal RMSD threshold. MULTIPROT defines this threshold as 2.0 Å. The purpose is to see the performances of template interfaces just above the threshold. To increase the model variety and possible interface number, that threshold is stretched a little more to 3.0 Å. Another change on MULTIPROT parameters is on residue classification. On default, hydrophilic residues are classified (classification number 3) together as polar hydrophilic, positively charged hydrophilic and negatively charged hydrophilic residues. There is another option (classification number 2) which considers those three groups of hydrophilic residues separately. To see the effects of parameters over the success of MULTIPROT, the models are predicted with those changes. In PRISM standalone, the folder called "run_files" has the "params.txt" file that leads to the parameters mentioned above.

FiberDock is responsible for the flexible refinement, docking and scoring the complex protein structures. The detailed information about the steps FiberDock follows has explained on chapter 2.2.3.4 Flexible Refinement. FiberDock allows user to change almost all parameters to optimize the program for current proteins. The parameters for FiberDock could be managed by the file called "buildFiberDockParams.pl" which could

be found in the following directory in PRISM standalone version: `external_tools > fiberdock > buildFiberDockParams.pl`

The algorithm provides four options for complex type: Enzyme/Inhibitor, Antibody/Antigen, Protein/RNA and Default for other kinds. PRISM uses default option for complex type. Protein/RNA option is not considered for PRISM because the template interfaces are extracted only from protein-protein complexes. However, to improve the accuracy of PRISM, other complex types could be selected specific for each target. FiberDock has another parameter called RISCO/FISCO. RISCO means restricted side-chain optimization and FISCO means full side-chain optimization. Side-chain optimization is an important factor for energy calculation of the complex model and it could be a reason giving no models well enough for the CAPRI classifications. As default, FiberDock uses restricted side-chain optimization. For this study, it is converted to use full side-chain optimization to see the effect of side-chains over the energy of complex structure.

Targets 40, 41, 46, and 47 are enzyme/inhibitor complexes and target 02 is an antibody/antigen complex. For those targets, the effects of different complex type options and effect of side-chain optimization are examined by the analysis of predictions. Table 5.1 gives the detailed comparison of different parameters for the same complex models. For target 40, 41 and 46 there is no significant change in terms of I-rmsd. All models on each parameter have slight changes on the decimals or around maximum 1Å. However, in terms of energy, A/A always gives the lowest energy score, even though our models are classified as enzyme/inhibitor complexes. Especially for target 46 and 47, default PRISM gives the maximum number of predictions rather than two other versions. So it could be concluded that changes on MULTIPROT gave no new interface for predictions, but changes on FiberDock affect the scoring the same model and increase the availability of the best predictions as results.

Table 5.1. The comparison of same models predicted by different parameters for targets 40, 41, 46 and 47. (E/I : Enzyme/Inhibitor complex, A/A: Antibody/Antigen complex)

	Interface Name	Model Name	Default		E/I		A/A	
			Energy	I-rmsd	Energy	I-rmsd	Energy	I-rmsd
Target 40	2qyiCD	btyp_1_apia_0	-16.55	12.651	14.2	12.327	-22.19	12.327
	2qyiCD	btyp_0_apia_0	-5.35	11.861	0.95	12.812	-43.4	12.812
	2qyiCD	btyp_2_apia_0	-5.35	11.861	14.2	12.327	-22.19	12.327
	Interface Name	Model Name	Default		E/I		A/A	
			Energy	I-rmsd	Energy	I-rmsd	Energy	I-rmsd
Target 41	3zu0AB	cln9_0_cln2_0	22.9	14.759	30.95	14.092	-3.71	14.092
	3zu0AB	cln2_0_cln9_1	-	-	62.81	18.863	29.35	18.863
	3putAB	cln9_0_cln2_0	30.81	18.899	-	-	-	-
	Interface Name	Model Name	Default		E/I		A/A	
			Energy	I-rmsd	Energy	I-rmsd	Energy	I-rmsd
Target 46	3do8AB	mtq2_0_trm1_2	-26.85	18.152	10.4	19.163	-	-
	3eo6AB	trm1_0_mtq2_0	-19.03	16.514	20.37	16.102	-33.07	16.102
	1tileF	mtq2_0_trm1_1	-7.96	17.131	36.95	17.207	-27.05	17.207
	2f66AB	trm1_0_mtq2_1	16.28	19.827	43.75	19.245	-1.94	19.245
	3chiAB	mtq2_1_trm1_0	20.75	18.449	22.4	19.213	-0.8	19.213
	1sg9AC	mtq2_0_trm1_0	20.78	22.040	44.14	21.894	-11.52	21.894
	1ylmAB	trm1_0_mtq2_1	27.17	19.057	23.47	18.919	-	-
	2gmyBC	trm1_2_mtq2_0	43.03	20.910	42.86	21.123	-	-
	Interface Name	Model Name	Default		E/I		A/A	
			Energy	I-rmsd	Energy	I-rmsd	Energy	I-rmsd
Target 47	2pmyAB	che9_0_che2_1	8.05	10.52	-	-	-	-
	2i8bAB	che2_1_che9_0	35.34	9.182	28.6	8.303	-40.12	8.303
	2wymDF	che9_1_che2_0	53.91	15.012	58.82	15.3	-	-
	3f6wCD	che9_1_che2_1	60.26	13.867	88.67	11.531	-	-

Targets 09, 10, 23, and 28 are homodimer complexes which are the weakest spot of PRISM on this study. The same test is applied for those 4 targets. Table 5.2 gives the detailed comparison of different parameters for the best complex models. First of all, PRISM versions with new parameters could not find any promising models too. There is no model with I-rmsd less than 4 Å. Homodimer complex formation is still the weakest spot for PRISM. Obviously, it is not the fact that the same model has to be appeared on

all three versions. For target 09, the best complex structure predictions from all three versions are different. Among those three models, the best model is constructed by default parameters. For target 10, the best models are the same for default and E/I versions which could not be found with A/A. I-rmsd of that model is really close in both versions but E/I construct the model slightly better than default version. According to energy scores, default version assigns lower energy score (even it is positive) which is preferable. For target 23, the best model by far is coming from default version with greatest I-rmsd value of 4.720 Å which could not be constructed by other two versions.

Last, for target 28, changes on MULTIPROT give a new interface which has never used before on other versions, especially on default. The interface is named as 2pwmEF. A model constructed with this interface by using E/I version of FiberDock gives the smallest I-rmsd value. Even though there is still no prediction enough to be considered as true complex structure, that model reflects the fundamental purpose of this chapter. New parameters provides new interface and new model could not be predicted by other two versions and get the closest complex formation to the reference structure. One other important detail could be observed with a model of target 28. A model *surf_0_sur2_0* with interface 3l49CD is predicted with all three versions which allow us to compare them together. I-rmsd values are almost the same (different on second decimals) but energy value of the model derived from A/A is much smaller than the other two.

Table 5.2. The comparison of the best models predicted by different parameters for targets 09, 10, 23 and 28. (E/I: Enzyme/Inhibitor complex, A/A: Antibody/Antigen complex)

	Interface Name	Model Name	Default		E/I		A/A	
			Energy	I-rmsd	Energy	I-rmsd	Energy	I-rmsd
Target 09	3dhxAB	litb_2_lita_0	-34.72	13.151	-	-	-	-
	3beyDE	lita_2_litb_1	-	-	24.78	20.442	-	-
	3bdzAB	lita_2_litb_0	-	-	-	-	-14.12	16.149
Target 10	1urzDE	tbca_1_tcb_0	36.6	12.672	47.13	12.545	-	-
	1gulFG	tbca_0_tcb_0	-	-	-	-	-62.81	16.821
Target 23	2o6iAB	igbb_0_igbp_0	-	-	40.26	31.758	-5.64	31.758
	2b8wAB	igbp_2_igbp_0	-0.35	4.720	-	-	-	-
Target 28	2c9jEG	surf_0_sur2_1	18.35	20.835	40.86	21.588	-	-
	2pwmEF	surf_0_surf_0	-	-	74.7	20.04	-	-
	3l49CD	surf_0_sur2_0	44.43	23.477	46.34	23.464	1.99	23.464

Overall, changes in parameters of MULTIPROT and FiberDock are not enough for PRISM successfully predict the complex formation of targets. Even though there are examples of models reflect the main intend of the chapter, no model could reach the classification level other than incorrect. Recruiting new template interfaces could be helpful as it was thought. However, the complex type parameter on FiberDock has no major impact on complex model accuracy besides the impact over energy scores. A/A version always tend to decrease the energy scores of models even the I-rmsd remains the same and the targets are not antibodies and antigens.

Chapter 6 – Conclusion

The main purpose of this research is the performance assessment of PRISM using previous CAPRI targets. It also explains the weakness and strength of the PRISM on types of targets and scoring performances.

Protein-protein interaction (PPI) prediction tools help guide experimentalists in complex structure prediction and enrich the available complex structure data. Thus, it is of capital importance to evaluate the success of these tools in finding the correct structures which is an indication of their credibility. PRISM, as a template based PPI algorithm, uses all available interfaces in literature to predict the interaction among two proteins. Utilizing the fact that interfaces are evolutionary conserved regions, PRISM models new protein complexes. This study aims to clarify how successful the PRISM algorithm is in finding correct complex structures presented in CAPRI.

CAPRI is an organization that annually performs an experiment to test and compare the performances of participated PPI prediction algorithms. It is called blind docking since the predictors have no information about the correct complex form. The aim is to predict the possible closest complex structure. CAPRI committee designates the evaluation criteria to determine the quality of predicted structures and rank these structures as high, medium, low or incorrect. In this study, the predictions are evaluated and ranked according to the same criteria.

Scoring a protein complex prediction reflects the topological and functional accuracy of a PPI. Most scoring functions are based on the energy calculations of the predicted model. Since the prediction tools are designed to create largest possible decoys for a prediction, researchers should trust the scoring functions for selecting the best model and eliminating the rest. That increases the importance of scoring and questions the reliability of scoring functions. PRISM uses FiberDock for scoring the predicted models which is based on the energy calculation of the complex under the CHARMM force field.

Here, I focused on the effect of scoring function over the performance of PRISM by comparing it with other scoring system.

PRISM utilizes three other algorithms to predict the perfect protein complex structures. Each algorithm uses their own standard parameters to optimize the accuracy of their outcome. Standardization is not always the best choice because each case carries its own special conditions. Here I test the several combinations of parameters to test their effect over the success of PRISM. This study shows the detailed results for each targets under the different PRISM versions.



BIBLIOGRAPHY

1. Alberts, B., Wilson, J., & Hunt, T. (2008). *Molecular biology of the cell*. New York: Garland Science.
2. Vakser, I.A., *Protein-protein docking: from interaction to interactome*. Biophys J, 2014. **107**(8): p. 1785-93.
3. Ogmen, U., et al., *PRISM: protein interactions by structural matching*. Nucleic Acids Research, 2005. **33**(Web Server): p. W331-W336.
4. Janin, J., *Welcome to CAPRI: A Critical Assessment of PRedicted Interactions*. Proteins: Structure, Function, and Genetics, 2002. **47**(3): p. 257-257.
5. Mashiaich, E., R. Nussinov, and H.J. Wolfson, *FiberDock: Flexible induced-fit backbone refinement in molecular docking*. Proteins, 2010. **78**(6): p. 1503-19.
6. Lyskov, S. and J.J. Gray, *The RosettaDock server for local protein-protein docking*. Nucleic Acids Res, 2008. **36**(Web Server issue): p. W233-8.
7. Shatsky, M., R. Nussinov, and H.J. Wolfson, *A method for simultaneous alignment of multiple protein structures*. Proteins, 2004. **56**(1): p. 143-56.
8. Hegyi, H. and M. Gerstein, *The relationship between protein structure and function: a comprehensive survey with application to the yeast genome*. J Mol Biol, 1999. **288**(1): p. 147-64.
9. Apic G, Gough J, Teichmann SA. An insight into domain combinations. Bioinformatics (Oxford, England) 2001. pp. S83–89.
10. Lengauer, T. and M. Rarey, *Computational methods for biomolecular docking*. Curr Opin Struct Biol, 1996. **6**(3): p. 402-6.
11. Rentzsch, R. and C.A. Orengo, *Protein function prediction using domain families*. BMC Bioinformatics, 2013. **14 Suppl 3**: p. S5.
12. Wandzik, Ilona . “Current Molecular Docking Tools and Comparisons Thereof.” MATCH Communications in Mathematical and in Computer Chemistry, 17 Nov. 2005, pp. 271–278.
13. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000. **28**(1): p. 235-42.
14. Smyth, M.S. and J.H. Martin, *x ray crystallography*. Mol Pathol, 2000. **53**(1): p. 8-14.
15. James, T. L. (1975) *Nuclear magnetic resonance in biochemistry*, Academic Press, New York.
16. Bonvin, A.M., *Flexible protein-protein docking*. Curr Opin Struct Biol, 2006. **16**(2): p. 194-200.
17. Andrusier, N., et al., *Principles of flexible protein-protein docking*. Proteins, 2008. **73**(2): p. 271-89.
18. T. Schulz-Gasch, M. Stahl, “Scoring Functions for Protein-ligand Interactions: A Critical Perspective”, *Drug Discovery Today: Technologies*, 1, 231-239 (2004).

19. Ferrara, P., et al., *Assessing scoring functions for protein-ligand interactions*. *J Med Chem*, 2004. 47(12): p. 3032-47.
20. Vreven, T., et al., *Evaluating template-based and template-free protein-protein complex structure prediction*. *Brief Bioinform*, 2014. 15(2): p. 169-76.
21. Tuncbag, N., A. Gurses, and O. Keskin, *Prediction of protein-protein interactions: unifying evolution and structure at protein interfaces*. *Phys Biol*, 2011. 8(3): p. 035006.
22. Friesner, R.A., et al., *Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy*. *J Med Chem*, 2004. 47(7): p. 1739-49.
23. Halgren, T.A., et al., *Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening*. *J Med Chem*, 2004. 47(7): p. 1750-9.
24. Eldridge, M.D., et al., *Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes*. *J Comput Aided Mol Des*, 1997. 11(5): p. 425-45.
25. Kellenberger, E., et al., *Comparative evaluation of eight docking tools for docking and virtual screening accuracy*. *Proteins*, 2004. 57(2): p. 225-42.
26. Plewczynski, D., et al., *Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database*. *J Comput Chem*, 2011. 32(4): p. 742-55.
27. Perola, E., W.P. Walters, and P.S. Charifson, *A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance*. *Proteins*, 2004. 56(2): p. 235-49.
28. Jain, A.N., *Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine*. *J Med Chem*, 2003. 46(4): p. 499-511.
29. Jones, G., et al., *Development and validation of a genetic algorithm for flexible docking*. *J Mol Biol*, 1997. 267(3): p. 727-48.
30. Lape, M., C. Elam, and S. Paula, *Comparison of current docking tools for the simulation of inhibitor binding by the transmembrane domain of the sarco/endoplasmic reticulum calcium ATPase*. *Biophys Chem*, 2010. 150(1-3): p. 88-97.
31. Goodsell, D.S. and A.J. Olson, *Automated docking of substrates to proteins by simulated annealing*. *Proteins*, 1990. 8(3): p. 195-202.
32. Rarey, M., et al., *A fast flexible docking method using an incremental construction algorithm*. *J Mol Biol*, 1996. 261(3): p. 470-89.
33. Schneidman-Duhovny, D., et al., *PatchDock and SymmDock: servers for rigid and symmetric docking*. *Nucleic Acids Res*, 2005. 33(Web Server issue): p. W363-7.
34. Chen, R. and Z. Weng, *Docking unbound proteins using shape complementarity, desolvation, and electrostatics*. *Proteins*, 2002. 47(3): p. 281-94.

35. Brooks, B.R., et al., CHARMM: the biomolecular simulation program. *J Comput Chem*, 2009. 30(10): p. 1545-614.
36. Dominguez, C., R. Boelens, and A.M. Bonvin, HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 2003. 125(7): p. 1731-7.
37. Gray, J.J., et al., Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*, 2003. 331(1): p. 281-99.
38. Ramirez, D. and J. Caballero, Is It Reliable to Use Common Molecular Docking Methods for Comparing the Binding Affinities of Enantiomer Pairs for Their Protein Target? *Int J Mol Sci*, 2016. 17(4).
39. Adasme-Carreno, F., et al., Performance of the MM/GBSA scoring using a binding site hydrogen bond network-based frame selection: the protein kinase case. *Phys Chem Chem Phys*, 2014. 16(27): p. 14047-58.
40. Lee, M.R. and Y. Sun, Improving Docking Accuracy through Molecular Mechanics Generalized Born Optimization and Scoring. *J Chem Theory Comput*, 2007. 3(3): p. 1106-19.
41. Aytuna, A.S., A. Gursoy, and O. Keskin, Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 2005. 21(12): p. 2850-5.
42. Keskin, O., et al., A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci*, 2004. 13(4): p. 1043-55.
43. Choi, Y.S., et al., Evolutionary conservation in multiple faces of protein interaction. *Proteins*, 2009. 77(1): p. 14-25.
44. Naveed, H. and J.J. Han, Structure-based protein-protein interaction networks and drug design. *Quantitative Biology*, 2013. 1(3): p. 183-191.
45. Zhao, J., et al., Accuracy of functional surfaces on comparatively modeled protein structures. *J Struct Funct Genomics*, 2011. 12(2): p. 97-107.
46. Sinha, R., P.J. Kundrotas, and I.A. Vakser, Docking by structural similarity at protein-protein interfaces. *Proteins*, 2010. 78(15): p. 3235-41.
47. Tsai, C.J., et al., A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol*, 1996. 260(4): p. 604-20.
48. Clackson, T. and J.A. Wells, A hot spot of binding energy in a hormone-receptor interface. *Science*, 1995. 267(5196): p. 383-6.
49. Bogan, A.A. and K.S. Thorn, Anatomy of hot spots in protein interfaces. *J Mol Biol*, 1998. 280(1): p. 1-9.
50. Tuncbag, N., A. Gursoy, and O. Keskin, Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, 2009. 25(12): p. 1513-20.
51. Keskin, O., B. Ma, and R. Nussinov, Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol*, 2005. 345(5): p. 1281-94.

52. Gao, Y., R. Wang, and L. Lai, *Structure-based method for analyzing protein-protein interfaces*. *J Mol Model*, 2004. 10(1): p. 44-54.
53. Nussinov, R. and H.J. Wolfson, *Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques*. *Proc Natl Acad Sci U S A*, 1991. 88(23): p. 10495-9.
54. Tuncbag, N., et al., *Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM*. *Nat Protoc*, 2011. 6(9): p. 1341-54.
55. Tuncbag, N., et al., *Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement*. *Proteins*, 2012. 80(4): p. 1239-49.
56. Hubbard, S.J. and Thornton, J.M. (1993) NACCESS Computer Program. Department of Biochemistry and Molecular Biology, University College London
57. Lee, B. and F.M. Richards, *The interpretation of protein structures: estimation of static accessibility*. *J Mol Biol*, 1971. 55(3): p. 379-400.
58. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions*. *Proc Natl Acad Sci U S A*, 1996. 93(1): p. 13-20.
59. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins*. *J Phys Chem B*, 1998. 102(18): p. 3586-616.
60. Janin, J., et al., *CAPRI: a Critical Assessment of PRedicted Interactions*. *Proteins*, 2003. 52(1): p. 2-9.
61. Mendez, R., et al., *Assessment of blind predictions of protein-protein interactions: current status of docking methods*. *Proteins*, 2003. 52(1): p. 51-67.
62. The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.
63. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis*. *J Comput Chem*, 2004. 25(13): p. 1605-12.
64. Cukuroglu, E., et al., *Non-redundant unique interface structures as templates for modeling protein interactions*. *PLoS One*, 2014. 9(1): p. e86738.
65. Fieulaine, S., et al., *X-ray structure of a bifunctional protein kinase in complex with its protein substrate HPr*. *Proc Natl Acad Sci U S A*, 2002. 99(21): p. 13437-41.
66. Fieulaine, S., et al., *X-ray structure of HPr kinase: a bacterial protein kinase with a P-loop nucleotide-binding domain*. *EMBO J*, 2001. 20(15): p. 3917-27.
67. Mathieu, M., et al., *Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion*. *EMBO J*, 2001. 20(7): p. 1485-97.
68. Skehel, J.J. and D.C. Wiley, *Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin*. *Annu Rev Biochem*, 2000. 69: p. 531-69.
69. Desmyter, A., et al., *Three camelid VHH domains in complex with porcine pancreatic alpha-amylase. Inhibition and versatility of binding topology*. *J Biol Chem*, 2002. 277(26): p. 23645-50.

70. Sundberg, E.J., et al., Structures of two streptococcal superantigens bound to TCR beta chains reveal diversity in the architecture of T cell signaling complexes. *Structure*, 2002. 10(5): p. 687-99.
71. Mayer, U., E. Kohfeldt, and R. Timpl, Structural and genetic analysis of laminin-nidogen interaction. *Ann N Y Acad Sci*, 1998. 857: p. 130-42.
72. van Tilbeurgh, H., D. Le Coq, and N. Declerck, Crystal structure of an activated form of the PTS regulation domain from the LicT transcriptional antiterminator. *EMBO J*, 2001. 20(14): p. 3789-99.
73. Bressanelli, S., et al., Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. *EMBO J*, 2004. 23(4): p. 728-38.
74. Carvalho, A.L., et al., Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *Proc Natl Acad Sci U S A*, 2003. 100(24): p. 13809-14.
75. Mechaly, A., et al., Cohesin-dockerin recognition in cellulosome assembly: experiment versus hypothesis. *Proteins*, 2000. 39(2): p. 170-7.
76. Marc Graille, Enrico A. Stura, Marc Bossus, Bruno H. Muller, Odile Letourneur, Nicole Battail-Poirot, Geneviève Sibai, Marie Gauthier, Dominique Rolland, Marie-Hélène Le Du, Frédéric Ducancel, Crystal Structure of the Complex between the Monomeric Form of Surface Antigen 1 (SAG1) and a Monoclonal Antibody that Mimics the Human Immune Response, *Journal of Molecular Biology*, Volume 354, Issue 2, 2005, Pages 447-458, ISSN 0022-2836
77. F. Velge-Roussel, T. Chardès, P. Mévélec, M. Brillard, J. Hoebeke, D. Bout, Epitopic analysis of the major surface antigen SAG1, *Molecular and Biochemical Parasitology*, Volume 66, Issue 1, 1994, Pages 31-38, ISSN 0166-6851
78. Terrak, M., et al., Structural basis of protein phosphatase 1 regulation. *Nature*, 2004. 429(6993): p. 780-4.
79. Tomita, K., et al., A cytotoxic ribonuclease which specifically cleaves four isoaccepting arginine tRNAs at their anticodon loops. *Proc Natl Acad Sci U S A*, 2000. 97(15): p. 8278-83.
80. Graille, M., et al., Structural inhibition of the colicin D tRNase by the tRNA-mimicking immunity protein. *EMBO J*, 2004. 23(7): p. 1474-82.
81. Schmidt, A., et al., Structure of the xylanase from *Penicillium simplicissimum*. *Protein Sci*, 1998. 7(10): p. 2081-8.
82. Payan, F., et al., The dual nature of the wheat xylanase protein inhibitor XIP-I: structural basis for the inhibition of family 10 and family 11 xylanases. *J Biol Chem*, 2004. 279(34): p. 36029-37.
83. Tahir, T.A., et al., Specific characterization of substrate and inhibitor binding sites of a glycosyl hydrolase family 11 xylanase from *Aspergillus niger*. *J Biol Chem*, 2002. 277(46): p. 44035-43.
84. Torronen, A. and J. Rouvinen, Structural comparison of two major endo-1,4-xylanases from *Trichoderma reesei*. *Biochemistry*, 1995. 34(3): p. 847-56.

85. Sansen, S., et al., *Structural basis for inhibition of Aspergillus niger xylanase by triticum aestivum xylanase inhibitor-I*. *J Biol Chem*, 2004. 279(34): p. 36022-8.
86. Lopez Garcia, F., et al., *NMR structure of the bovine prion protein*. *Proc Natl Acad Sci U S A*, 2000. 97(15): p. 8334-9.
87. Eghiaian, F., et al., *Insight into the PrPC-->PrPSc conversion from the structures of antibody-bound ovine prion scrapie-susceptibility variants*. *Proc Natl Acad Sci U S A*, 2004. 101(28): p. 10254-9.
88. Graille, M., et al., *Molecular basis for bacterial class I release factor methylation by PrmC*. *Mol Cell*, 2005. 20(6): p. 917-27.
89. Heurgue-Hamard, V., et al., *The hemK gene in Escherichia coli encodes the N(5)-glutamine methyltransferase that modifies peptide release factors*. *EMBO J*, 2002. 21(4): p. 769-78.
90. Yang, Z., et al., *Structural characterization and comparative phylogenetic analysis of Escherichia coli HemK, a protein (N5)-glutamine methyltransferase*. *J Mol Biol*, 2004. 340(4): p. 695-706.
91. Hou, Z., et al., *Structural basis of the Sir1-origin recognition complex interaction in transcriptional silencing*. *Proc Natl Acad Sci U S A*, 2005. 102(24): p. 8489-94.
92. Zhang, Z., et al., *Structure and function of the BAH-containing domain of Orc1p in epigenetic silencing*. *EMBO J*, 2002. 21(17): p. 4600-11.
93. Nielsen, T.K., et al., *Structural basis for the bifunctionality of the U5 snRNP 52K protein (CD2BP2)*. *J Mol Biol*, 2007. 369(4): p. 902-8.
94. Ghosh, A., et al., *How guanylate-binding proteins achieve assembly-stimulated processive cleavage of GTP to GMP*. *Nature*, 2006. 440(7080): p. 101-4.
95. Basseres, D.S., et al., *ARHGAP10, a novel human gene coding for a potentially cytoskeletal Rho-GTPase activating protein*. *Biochem Biophys Res Commun*, 2002. 294(3): p. 579-85.
96. Carr, S., et al., *The structure of TolB, an essential component of the tol-dependent translocation system, and its protein-protein interaction with the translocation domain of colicin E9*. *Structure*, 2000. 8(1): p. 57-66.
97. Abergel, C., et al., *Structure of the Escherichia coli TolB protein determined by MAD methods at 1.95 Å resolution*. *Structure*, 1999. 7(10): p. 1291-300.
98. Bonsor, D.A., et al., *Molecular mimicry enables competitive recruitment by a natively disordered protein*. *J Am Chem Soc*, 2007. 129(15): p. 4800-7.
99. Ko, S., et al., *Structural basis of E2-25K/UBB+1 interaction leading to proteasome inhibition and neurotoxicity*. *J Biol Chem*, 2010. 285(46): p. 36070-80.
100. Tong, H., et al., *Crystal structure of murine/human Ubc9 provides insight into the variability of the ubiquitin-conjugating system*. *J Biol Chem*, 1997. 272(34): p. 21381-7.
101. Fouladkou, F., et al., *A naturally occurring human Nedd4-2 variant displays impaired ENaC regulation in Xenopus laevis oocytes*. *Am J Physiol Renal Physiol*, 2004. 287(3): p. F550-61.

102. Purta, E., et al., *Sequence-structure-function relationships of a tRNA (m7G46) methyltransferase studied by homology modeling and site-directed mutagenesis*. *Proteins*, 2005. 59(3): p. 482-8.
103. Leulliot, N., et al., *Structure of the yeast tRNA m7G methylation complex*. *Structure*, 2008. 16(1): p. 52-61.
104. Tong, Y., et al., *Binding of Rac1, Rnd1, and RhoD to a novel Rho GTPase interaction motif destabilizes dimerization of the plexin-B1 effector domain*. *J Biol Chem*, 2007. 282(51): p. 37215-24.
105. Oinuma, I., et al., *Direct interaction of Rnd1 with Plexin-B1 regulates PDZ-RhoGEF-mediated Rho activation by Plexin-B1 and induces cell contraction in COS-7 cells*. *J Biol Chem*, 2003. 278(28): p. 25671-7.
106. Oinuma, I., et al., *The Semaphorin 4D receptor Plexin-B1 is a GTPase activating protein for R-Ras*. *Science*, 2004. 305(5685): p. 862-5.
107. Micheelsen, P.O., et al., *Structural and mutational analyses of the interaction between the barley alpha-amylase/subtilisin inhibitor and the subtilisin savinase reveal a novel mode of inhibition*. *J Mol Biol*, 2008. 380(4): p. 681-90.
108. Vallee, F., et al., *Barley alpha-amylase bound to its endogenous protein inhibitor BASI: crystal structure of the complex at 1.9 Å resolution*. *Structure*, 1998. 6(5): p. 649-59.
109. Najmudin, S., et al., *Putting an N-terminal end to the Clostridium thermocellum xylanase Xyn10B story: crystal structure of the CBM22-1-GH10 modules complexed with xylohexaose*. *J Struct Biol*, 2010. 172(3): p. 353-62.
110. Isabet, T., et al., *The structural basis of Arf effector specificity: the crystal structure of ARF6 in a complex with JIP4*. *EMBO J*, 2009. 28(18): p. 2835-45.
111. Venkateswarlu, K., T. Hanada, and A.H. Chishti, *Centaurin-alpha1 interacts directly with kinesin motor protein KIF13B*. *J Cell Sci*, 2005. 118(Pt 11): p. 2471-84.
112. Bao, R., et al., *The ternary structure of the double-headed arrowhead protease inhibitor API-A complexed with two trypsins reveals a novel reactive site conformation*. *J Biol Chem*, 2009. 284(39): p. 26676-84.
113. Li, W., et al., *Dual recognition and the role of specificity-determining residues in colicin E9 DNase-immunity protein interactions*. *Biochemistry*, 1998. 37(34): p. 11771-9.
114. Kuhlmann, U.C., et al., *Specificity in protein-protein interactions: the structural basis for dual recognition in endonuclease colicin-immunity protein complexes*. *J Mol Biol*, 2000. 301(5): p. 1163-78.
115. Krachler, A.M., A. Sharma, and C. Kleanthous, *Self-association of TPR domains: Lessons learned from a designed, consensus-based TPR oligomer*. *Proteins*, 2010. 78(9): p. 2131-43.
116. Liger, D., et al., *Mechanism of activation of methyltransferases involved in translation by the Trm112 'hub' protein*. *Nucleic Acids Res*, 2011. 39(14): p. 6249-59.

117. Meenan, N.A., et al., *The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. Proc Natl Acad Sci U S A*, 2010. 107(22): p. 10080-5.
118. Wojdyla, J.A., et al., *Structure of the ultra-high-affinity colicin E2 DNase--Im2 complex. J Mol Biol*, 2012. 417(1-2): p. 79-94.
119. Fleishman, S.J., et al., *Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science*, 2011. 332(6031): p. 816-21.
120. Ekiert, D.C., et al., *Antibody recognition of a highly conserved influenza virus epitope. Science*, 2009. 324(5924): p. 246-51.

