UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Iyanar VETRIVEL

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : *Biologie Santé*
Discipline : *Aspects moléculaires et cellulaires de la biologie*
Spécialité : *Bioinformatique structurale*
Unité de recherche : *UFIP UMR CNRS 6286*
Soutenue le *30/10/2017*

Exploring new applications of a structural alphabet in protein structure analysis

JURY

Président du jury	**Catherine ETCHEBEST, Professeur d'Université, Université Paris Diderot**
Rapporteurs :	**Anne-Claude CAMPROUX**, Professeur d'université, Université Paris Diderot
	Juliette MARTIN, Chargé de recherche, CNRS – Université de Lyon
Examinateur :	**Srinivasan NARAYANASWAMY,** Professeur d'Université, Indian Institute of Science
Directeur de Thèse :	**Bernard OFFMANN**, Professeur d'université, Université de Nantes
Co-directeur de Thèse :	**Frédéric CADET**, Professeur d'université, Université de La Réunion

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Iyanar VETRIVEL

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : *Biologie Santé*
Discipline : *Aspects moléculaires et cellulaires de la biologie*
Spécialité : *Bioinformatique structurale*
Unité de recherche : *UFIP UMR CNRS 6286*
Soutenue le *30/10/2017*

Exploring new applications of a structural alphabet in protein structure analysis

JURY

Président du jury	**Catherine ETCHEBEST, Professeur d'Université, Université Paris Diderot**
Rapporteurs :	**Anne-Claude CAMPROUX**, Professeur d'université, Université Paris Diderot
	Juliette MARTIN, Chargé de recherche, CNRS – Université de Lyon
Examinateur :	**Srinivasan NARAYANASWAMY,** Professeur d'Université, Indian Institute of Science
Directeur de Thèse :	**Bernard OFFMANN**, Professeur d'université, Université de Nantes
Co-directeur de Thèse :	**Frédéric CADET**, Professeur d'université, Université de La Réunion

Table of contents

Synopsis.....16

Synopsis in English.....16

Synopsis en français.....20

Chapter 1. Introduction.....24

1.1. Primary structure: string of amino acids & peptide bond.....24

1.1.1. Amino acids.....24

1.1.2. Peptide bond.....28

1.2. Secondary structure: strands, helices and loops.....31

1.2.1. Helices.....31

1.2.2. β- Sheets.....33

1.2.3. Loops and turns.....33

1.3. Tertiary and quaternary structure.....34

1.4. Fragment libraries and structural alphabets.....34

1.5. Protein structure determination.....42

1.6. Protein structure prediction.....45

1.7. Comparing and classifying proteins.....46

1.8. Objectives.....51

Chapter 2:Analysis of structural variations in proteins using a structural alphabet.....53

2.1. Introduction.....53

2.2. Material and Methods.....55

2.2.1. Datasets.....55

2.2.2. Protein Blocks.....56

2.2.3. Molecular dynamics.....56

2.2.4. Generating the variation matrices.....57

2.3. Results and discussion....58

2.3.1. Inferences from qualitative analysis of structural plasticity....58

2.3.2. Inferences from quantitative analysis of structural plasticity....60

2.3.3. Case studies of structural variations for alternate conformations of the same protein....62

2.3.4. Case studies of structural variations between two homologous proteins having alternate conformations....65

2.3.5. Further analyses of molecular dynamics trajectories using PBs....67

Chapter 3:Knowledge based prediction of protein backbone conformation using a structural alphabet....73

3.1. Abstract....73

3.2. Author summary....74

3.3. Introduction....75

3.4. Results....77

3.4.1. PENTAdb, a database of pentapeptides from protein structures....77

3.4.2. Not all possible tri-PB combinations are observed in known protein structures....77

3.4.3. Completeness of PENTAdb for knowledge-based prediction....78

3.4.4. Prediction accuracies....79

3.4.5. PB predictions....84

3.4.6. Measure of accuracy....86

3.4.7. Case studies....88

3.4.8. Implementation of the PB-kPRED methods as a web-tool....91

3.5. Discussion....91

3.6. Methods....96

3.6.1. Dataset....96

3.6.2. Protein Blocks....96

3.6.3. Database of pentapeptide conformations from protein structures....97

3.6.4. Prediction scheme....98

3.6.5. Evaluating PB-kPRED using different subsets of PENTAdb....101

3.6.6. A scoring function to estimate the accuracy of the predictions.....103

3.7. Supporting information.....105

3.8. Acknowledgements.....112

3.9. kPRED predictions for additional datasets.....112

3.10. Accuracy gain due to hybrid method with noise filtering.....119

3.11. Geometric scoring function and distribution of pentaPBs in non redundant datasets 120

3.12. Connecting the dots: correlation between the scoring function and prediction accuracy.....124

3.13. References.....127

Chapter 4: PENTAdb- A database of pentapeptides and its implications in protein local structure prediction.....132

4.1. Introduction.....132

4.2. Expanding the scope of PENTAdb.....133

4.2.1. Auto updating pipeline.....133

4.2.2. Challenges in expanding the scope of PENTAdb to be comprehensive.....134

4.2.3. Statistics on the contents of PENTAdb.....135

4.3. Web-interface for the database.....137

4.4. Structural variability of identical pentapeptides.....138

4.5. A quantitative measure of the structural variability of pentapeptides.....147

4.6. Effect of the contents of PENTAdb on PB prediction using the kPRED algorithm…148

Chapter 5: Comprehensive networks of protein structures.....151

5.1. Abstract.....151

5.2. Introduction.....151

5.3. Material and methods.....155

5.3.1. Dataset.....155

5.3.2. Protein Blocks.....156

5.3.3. PB-ALIGN....157

5.3.4. Clusters and Networks....159

5.3.5. PB-ALIGN score vs RMSD....160

5.3.6. Cross referencing with SCOP....160

5.4. Results....161

5.4.1. All vs all alignments....161

5.4.2. Alignment based functional annotation....164

5.4.3. Global analysis....166

5.4.4. Attempts to assign domains using alignment information....167

5.5. Discussion....172

Chapter 6: Conclusions and Perspectives....177

References....182

Acknowledgements

I would like to start by thanking my family, especially my parents and my wife. Their love, moral support and patience has fueled this thesis. I owe everything I have achieved in life to my parents and no words can express my affection and gratitude towards them.

I would like to gratefully acknowledge the enthusiastic supervision of Prof. Bernard Offmann. He has always been a friend first and an advisor next. His subtle balance of providing me the autonomy to develop as an independent researcher and yet supporting me with his guidance and advice when I needed it, has been the key to me successfully completing this thesis. Apart from research, I have imbibed various great qualities from him- work-life balance, perseverance towards perfection, never ending optimism, efficient multitasking etc.. His passion for science and technical prowess will always continue to inspire me. I would also like to thank my co-director Prof. Frederic Cadet for his guidance and believing in my skills to take up this research project.

My sincere thanks to Prof. N. Srinivasan and Dr. Alexandre de Brevern for their stimulating suggestions and invaluable guidance during the thesis committee discussions. Their immense knowledge and insightful advice has been important to set the direction for this thesis. I thank Prof. N. Srinivasan for welcoming me in his lab in IISc, Bangalore during the initial stages of my thesis work. I thank Dr. Yves-Henri Sanejouand and Dr. Stephane Teletchea for their motivation and discussions on specific aspects of this thesis work. I would like to express my gratitude to Prof. Sowdhamini Ramanathan for her support and introducing me to Prof. Bernard in the first place. I feel truly lucky to have such great scientific mentors, from whom I have learnt much more than science.

I would like to take this opportunity to thank all my predecessors who have worked on the PB project. Though I have had the privilege of only working directly with Dr. Swapnil Mahajan, I equally thank Dr. Manoj Tyagi, Dr. Agnel Praven Joseph and many others. Dr. Swapnil has been like an elder brother to me and has helped me both personally and professionally at numerous instances. I have been greatly inspired and benefited by their work. The following quote efficiently summarizes my experience of working on this project:

"Standing on the shoulders of giants"

-Bernard of Chartres

(French philosopher)

Special thanks go to Lionel Hoffmann for taking up multiple roles like synchronizer, database manager, C++ geek etc.. to continuously help me throughout the duration of my thesis. I also thank my *camarade* Johann Hendrickx for the countless technical and personal helps. I also thank the internship student who have worked with me-Thomas Coudoux, Tristan Rialland, Sara Bachiri, Venceslas Douillard and Caroline Roze. Each one of them have added a new dimension and perspective to my work. I greatly acknowledge all the current and former UFIP lab members. UFIP has been a stimulating and fun place to work under the able directorship of Prof. Charles Tellier and Prof. Bernard Offmann. In fact I have enjoyed working in UFIP and under the guidance of Prof Bernard so much that many a times I have wanted my doctoral days to never end !

Heartfelt thanks to all my friends and colleagues from Nantes. The memories of the countless parties, trips, outings, intellectually stimulating coffee time talks, pizza + movie nights etc...that I have enjoyed with them will always stay fresh in my mind. They have helped me experience and enjoy French culture and cuisine and make Nantes a home away from home.

This work was supported by the Région Réunion and the Fond Social Européen [grant no. 20131528] to Iyanar Vetrivel. This work was in part supported by Conseil Régional des Pays de la Loire in the framework of GRIOTE project.





To my Parents & Wife,

Vetrivel, Shanthy and Dharani

List of Tables

Table 1: Chronology of milestones in the development of protein structural biology.....25

Table 2: A few amino acids with their unique features.....28

Table 3: Regular secondary structure elements (SSEs) and their corresponding f and y values.....32

Table 4: Example of features used to define fragment libraries or structural alphabets.....35

Table 5: Correlation among the three PB variation matrices (NMRMAT, altconfMAT, dynamicsMAT) and the PB transition matrix (homologueMAT).....59

Table 6: Raw counts of PB transitions observed during a 10 ns MD simulation of an endoglucanase (PDBID 1A3H).....68

Table 7: Average time between PB transitions in pico seconds observed during the 10 ns MD simulation of an endoglucanase (PDBID 1A3H).....69

Table 8: Content of the different subsets of PDB in terms of pentapeptides accessible to PB-kPRED. 78

Table 9: Assessment of the richness of PENTAdb towards knowledge-based prediction of protein backbone in terms of protein blocks.....79

Table 10: Evaluation of performance of PB-kPRED knowledge-based approach to predict local conformations of protein backbone in terms of protein blocks.....81

Table 11: Assessment of the performance of PB-kPRED using the hybrid method with noise filtering scheme for querying the database.....82

Table 12. Assessment of the performance of PB-kPRED and comparison with other previously reported methods.....87

Table 13. Impact of the availability of known homologues on the accuracy of PB-kPRED.....90

Table 14: Evaluation of the performance of kPRED for the PDB100 dataset.....116

Table 15: Quantile distribution of pentapeptide counts in different subsets of PENTAdb.....137

Table 16. Number of pentapeptides that map to 1,2,3,>=10 or all 16 PBs from PENTAdb filtered at different levels of redundancies.....139

Table 17. Distribution of counts and average RMSDA when compared to the standard torsion angles for the pentapeptide “LGRRE”.....140

Table 18. Summary of 19 out of the 441 instances the penatpeptide “LGRRE” occurs in PENTAdb.141

Table 19: Distribution of number of alignments at different score cut-offs for the comprehensive and non-sequence redundant datasets.....164

Table 20: Distribution of unannotated alignments at different score cut-offs at the superfamily level from the SCOPe database (Chandonia, Fox, and Brenner 2017) for the PDB100 dataset 165

List of Figures

Figure 1.The twenty standard amino acids.....27

Figure 2. The peptide bond.....29

Figure 3. Ramachandran plot for all the residues of insulin.....30

Figure 4. Examples of pentapeptides representing each of the 16 structural prototypes that constitute the protein blocks (PBs) structural alphabet.....42

Figure 5. Growth in the number of structures deposited in the PDB.....43

Figure 6. Rigid vs flexible superposition of L-2-haloacid dehalogenase (grey) and β-phosphoglucomutase (yellow) by the FATCAT program.....49

Figure 7. Tri colored heat maps of the four substitution matrices.....60

Figure 8. Individual PB wise correlation between the four matrices.....62

Figure 9. Comparison between 3 structures of human tyrosine kinase C-SRC by rigid body & flexible superimposition.....63

Figure 10. PB alignment between two structures of human tyrosine kinase C-SRC bound to ANP ligand (2SRC) and MPZ ligand (1Y57) captured by altconfMAT matrix.....64

Figure 11. PB alignment between two structures of human tyrosine kinase C-SRC in the unbound form (1FMK) and bound with ANP ligand (2SRC).....64

Figure 12. Example of the structural variation captured by dynamicsMAT matrix after 5 and 10 ns MD simulations.....65

Figure 13. Network showing structural (dis)similarity between mouse and human sulfotransferases each having multiple alternate conformations.....66

Figure 14. Distribution of PB-ALIGN normalized alignment scores.....67

Figure 15. Root mean squared fluctuation (RMSF) and root mean squared deviation (RMSD) calculated for backbone atoms of PDB ID 1A3H during a 10 ns MD simulation.....70

Figure 16. Structural variability of pentapeptides along the backbone of 1A3H during a 10 ns MD simulation.....71

Figure 17. PB map for 1A3H during a MD simulation.....71

Figure 18. PB logos showing variation of PBs along the backbone of 1A3H during a 10 ns MD simulation.....72

Figure 19. Histogram depicting the distribution of the observed prediction accuracies for 15,544 query proteins by hybrid method using noise filtering scheme.....83

Figure 20. Comparison of the majority rule method without noise filtering scheme and the hybrid method with noise filtering scheme.....84

Figure 21. Assessment of the ability of the scoring function to estimate the prediction accuracy of PB-kPRED.....88

Figure 22. The knowledge-based methodology behind PB-kPRED.....99

Figure 23. Details of the scoring schemes underlying the majority rule method and the hybrid method. 100

Figure 24. Diagrammatic representations of the schemes used by PB-kPRED for querying PENTAdb. 102

Figure 25. Distribution of prediction accuracy by hybrid method for the subset of the PDB100 dataset that have a crystallographic resolution >1.8Å.....113

Figure 26. Distribution of prediction accuracy by hybrid method for the subset of the PDB100 dataset that have been solved by NMR.....114

Figure 27. Distribution of prediction accuracy by hybrid method for the subset of the PDB100 dataset that have been solved by electron microscopy.....115

Figure 28. Distribution of prediction accuracy for the CASP11 dataset.....117

Figure 29. Distribution of prediction accuracy for the CASP10 dataset.....117

Figure 30. Comparison of prediction accuracy along with standard deviation for all possible PBs method, majority method and hybrid method without noise filtering.....118

Figure 31. Histograms of distribution of prediction accuracy for the CASP10 dataset.....119

Figure 32. Exercise followed to estimate the accuracy gain at each sequence identity cutoff level for the PDB30 dataset.....120

Figure 33. Plots for the predicted vs actual PB sequence scoring for 58 CASP 11 targets with penalties ranging from 0 to -11.....123

Figure 34. Evaluation of how the performance of the scoring functions I and II that test the geometric compatibility between consecutive predicted PBs.....126

Figure 35. Contents of PENTAdb as a function of sequence identity cut-off.....136

Figure 36. Intra PB class RMSDA distribution for all the 16 PBs from the entire PENTAdb. In all cases the mean is indicated by the red vertical line.....142

Figure 37. Global distribution of RMSDA for the 61.6 million pentapeptides in PENTAdb.....143

Figure 38. Solvent accessibility for the pentapeptide “LGRRE”.....144

Figure 39. Superposition and polar contacts corresponding to the alternate conformations adopted by the pentapeptide “LGRRE”.....146

Figure 40. Relationship between protein local structure flexibility and predictability by kPRED algorithm.....150

Figure 41. Illustration of the asymmetric nature of using PB-ALIGN in the recursive mode.....162

Figure 42. Illustration of a sample PB-ALIGN alignment and scores.....163

Figure 43. Distribution of normalized alignment scores for the PDB100 dataset.....164

Figure 44. Clusters for Acyl phosphatase (PDB ID 1APS_A) at score cut-offs of 0.5, 0.8, 1.0 and 1.2. 168

Figure 45. Clusters for Rho GTPase (PDB ID 2MIO_A) at score cut-offs of 0.5, 0.8, 1.0 and 1.2.....169

Figure 46. Global analysis of structural similarities derived from PB alignments for the PDB100 dataset.....170

Figure 47. Domain assignment after PB alignment.....171

Figure 48. Analysis of the effect of alignment score cut-off on the number and size of clusters and plot of pairwise alignment score vs corresponding rmsd.....173

Figure 49. Increasing the score-cut-off from (a) 1.2 to (b) 1.6 weeds out the potential false positives for alignment based functional annotation.....175

List of Abbreviations

1D	one dimension, one dimensional
3D	three dimension, three dimensional
ANN	**A**rtificial **N**eural **N**etwork
BLAST	**B**asic **L**ocal **A**lignment **S**earch
CASP	**C**ritical **A**ssessment of protein **S**tructure **P**rediction
CATH	**C**lass, **A**rchitecture, **T**opology and **H**omology
DSSP	**D**ictionary of **S**econdary **S**tructure of **P**roteins
ECOD	**E**volutionary **C**lassification of **P**rotein **D**omains
GA	**G**enetic **A**lgorithm
GROMACS	**GRO**ningen **MA**chine for **C**hemical **S**imulations
HMM	**H**idden **M**arkov **M**odel
kPRED	**k**nowledge based **PRED**iction of protein backbone
MD	**M**olecular **D**ynamics
NMR	**N**uclear **M**agnetic **R**esonance
PALI	**P**hylogeny and **ALI**gnment of homologous protein structure
PBs	**P**rotein **B**locks
PDB	**P**rotein **D**ata **B**ank
PENTAdb	**PENTA**peptide **d**ata**b**ase
Pfam	**P**rotein **fam**ily
RMSDA	**R**oot **M**ean **S**quare **D**eviation on **A**ngular values
SA	**S**tructural **A**lphabet
SCOP	**S**tructural **C**lassification **O**f **P**roteins
SCOPe	**S**tructural **C**lassification **O**f **P**roteins extended
SOM	**S**elf **O**rganizing **M**ap
SSEs	**S**econdary **S**tructure **El**ements
SVI	**S**tructural **V**ariability **I**ndex
SVM	**S**upport **V**ector **M**achines

Synopsis

Synopsis in English

Chapter 1 introduces the reader to the fundamentals of protein structure and the current state of art in protein structure analysis. The organization of proteins as primary, secondary, tertiary and quaternary structure are discussed. A brief review of the fragment libraries and structural alphabets developed till date is included before elaborating on Protein Blocks (PB) in particular. Next the different methods used for determining and predicting protein structure are discussed. Finally, a section is dedicated to the advancements in classifying and comparing proteins. The chapter concludes by setting the context for the work described in the subsequent chapters of this thesis.

Chapter 2 explores the utility of PBs as a tool to analyze structural variations in proteins. Here the collection of multiple X-ray crystallographic structures deposited in the PDB for the same protein are used as the dataset. The structural variations observed in these 23,631 proteins are summarized as a substitution matrix of their corresponding PBs. A similar substitution matrix is constructed based on the structural variations observed during a 10ns molecular dynamics (MD) on a collection of 130 unrelated proteins. The structural variations were analyzed both quantitatively and qualitatively. These two substitution matrices are compared to two previously established substitution matrices- one for homologous proteins (Tyagi et al. 2006) and other for NMR ensembles (Mahajan et al. 2014). Interestingly NMR ensembles and MD simulations seem to capture different aspects of protein structure plasticity, indicated by the least correlation between the two respective substitution matrices. In both the alternate crystallographic structures substitution matrix and MD substitution

matrix, the PBs corresponding to the loops and capping regions of regular secondary structure elements account for almost 80% of the structural variations.

Chapter 3 deals with the knowledge based prediction of protein local structure in terms of PBs. This chapter is closely interlinked to Chapter 4 that deals with PENTAdb, the database of pentapeptides. We describe the evolution of the kPRED algorithm from a purely probabilistic approach (majority rule method) to a context specific approach (hybrid method). The gain in prediction accuracy due to taking into consideration the structure of the preceding and succeeding pentapeptide is illustrated in different prediction exercises. Further a noise filtering mechanism is implemented in which the pentapeptides from close homologues to the query are privileged in being made accessible to the kPRED algorithm. This provides on an average 12% increase in accuracy. The methodology is rigorously tested for 8 different hypothetical test cases when different levels of homologous information is made available to the prediction algorithm. We are able to achieve an average prediction accuracy of 66.3% for a collection of 15,544 unrelated proteins. The strengths and weaknesses of the kPRED methodology are discussed with illustrative examples. Our methodology was compared to previously developed PB prediction methods. An scoring function was developed to auto-evaluate our predictions based on the probabilistic distribution of penta-PBs in the predicted PB sequence. This scoring function correlates well with the actual predictions with an R^2 of 0.82. This opens up an application of the scoring function to act as a guideline towards improving the accuracies of PB prediction using other methods.

Chapter 4 focuses on analyzing the sequence-structure relationship at the pentapeptide level. The challenges and rewards of making the PENTAdb a comprehensive database are described here. The structural plasticity of pentapeptides are analyzed using a statistical approach.

Efforts are made to understand why a particular pentapeptide would prefer a particular conformation in a particular protein. To achieve this objective, pentapeptides are analyzed based on their solvent accessibility and polar contacts at the local level and via hydrogen bonding pattern and contact maps at the global level. The pentapeptide "LGRRE" is taken as an illustrative case study because it can correspond to all the 16 different PBs in PENTAdb. The PB wise root mean squared deviation on angular values (RMSDA) are illustrated graphically. This represents how much the 61.6 million pentapeptides currently in PENTAdb are deviant from the standard PB torsion angles. The relationship between the flexibility of the pentapeptide and its predictability by the kPRED methodology is studied using the structural variability index (SVI). The web-interface to access the PENTAdb database and use its advanced search features are described.

In **Chapter 5** we describe an attempt to understand the entire protein structure space known till date in the form of an interconnected network. The analysis is performed on the PDB100 dataset to not miss out on any of the structural diversity. An all-vs-all exhaustive comparison of the proteins in this dataset is performed using the PB-ALIGN program (M Tyagi et al. 2008). PB-ALIGN is a structure comparison tool that takes advantage of representing protein structures as PB sequences. The 282 million structural alignments made for this work greatly benefit from the speed and accuracy of PB-ALIGN. The alignments are filtered at a length cut-off of 60 residues and are performed recursively (unaligned regions are trimmed and re-aligned) so as to capture the maximum number of biologically meaningful relationships between proteins. The results are filtered at different alignment score cut-offs. Networks are constructed using the Cytoscape software (Shannon et al. 2003) to visually represent the clustering of proteins based on structural similarity. The clustering is based on the alignment score such that two proteins aligned with a higher alignment score come closer on the network

and vice-versa. The alignments were cross referenced with the domain annotation from the SCOPe database to explore the potential to contribute to novel annotations. The confidence of making these domain annotations increase with more stringent alignment score cut-offs. At a length cut-off of 60 and score cut-off of 2.0, there are more than 13, 000 unannotated alignments.

In conclusions and perspectives chapter we discuss the chief findings from each of the previous chapters and their implications on our previous understanding about protein structures.

Synopsis en français

Le **chapitre 1** introduit le lecteur sur les fondamentaux de la structure des protéines et l'état actuel de l'art dans l'analyse de la structure des protéines. L'organisation des protéines comme structure primaire, secondaire, tertiaire et quaternaire est discutée. Une brève description des bibliothèques de fragments et des alphabets structuraux développés à ce jour est incluse avant d'élaborer plus en détail sur les blocs protéiques (BP) en particulier. Ensuite, les différentes méthodes utilisées pour déterminer et prédire la structure des protéines sont discutées. Enfin, une section est consacrée aux avancées dans la classification et la comparaison des protéines. Le chapitre se termine en définissant le contexte du travail décrit dans les chapitres suivants de cette thèse.

Le **chapitre 2** explore l'utilité des BP en tant qu'outil d'analyse des variations structurelles des protéines. Dans ce chapitre, les protéines présents en plusieurs exemplaires dans la PDB et résolus par cristallographie aux rayons X ont été collectées pour constituer une base de données. Les variations structurelles observées dans ces 23 631 protéines sont résumées sous forme d'une matrice de substitution de leurs PB correspondant. Une matrice de substitution similaire a également été construite à partir des variations structurelles observées au cours de simulations de dynamique moléculaire de 10ns exécutées sur une collection de 130 protéines non apparentées. Les variations structurelles ont été analysées quantitativement et qualitativement. Ces deux matrices de substitution sont comparées à deux autres matrices de substitution établies précédemment, une pour les protéines homologues (Tyagi et al., 2006) et une autre pour les ensembles RMN (Mahajan et al., 2014). Il est intéressant de noter que les ensembles RMN et les simulations de dynamique moléculaire semblent capturer différents aspects de la plasticité de la structure protéique, indiqués par une moindre corrélation entre les

deux matrices de substitution respectives. Il est intéressant de noter que dans les matrices de substitution établies à partir des données cristallographiques et à partir des simulations, les PB correspondant aux boucles et aux régions de coiffant les éléments de structure secondaires réguliers représentent près de 80% des variations observées.

Le **chapitre 3** traite du développement d'une méthode de prédiction de la structure locale des protéines en terme de BP en s'appuyant sur les informations provenant d'une base de données de pentapeptides (PENTAdb). Ce chapitre est étroitement lié au chapitre 4 qui traite spécifiquement de cette base de données. Nous faisons évoluer l'algorithme kPRED d'une approche purement probabiliste (méthode de la règle de la majorité) à une approche contextuelle (méthode hybride). Nous démontrons que la précision de l'algorithme est amélioré quand l'information structurale des pentapeptides flanquant sont pris en compte. En outre, nous montrons comment la mise en place d'un mécanisme permettant de filtrer l'excédant d'informations provenant de la base de données permet aussi d'améliorer les performances de l'algorithme en moyenne de 12 %. Ce mécanisme consiste à privilégier d'abord l'information structurale de pentapeptides provenant d'homologues lorsque disponible. La méthodologie est rigoureusement évaluée dans 8 situations différentes où l'algorithme a accès à 8 sous ensembles de la base de données, d'homologie décroissante avec la protéine à prédire. Avec ces améliorations, nous sommes en mesure d'obtenir une précision moyenne de prédiction de 66,3% pour le jeu de 15 544 protéines non apparentées. Les points forts et les faiblesses de la méthodologie KPRED sont discutés à l'aide d'exemples. Notre méthodologie a été comparée à des méthodes de prédiction PB précédemment développées. Une fonction de score probabiliste a été développée pour évaluer automatiquement nos prédictions *a priori*. Elle s'appuie sur la probabilité d'observer la succession des penta-PB prédits. Nous montrons que ce score est corrélée aux taux de prédictions avec un R2 de 0,82. Dans l'annexe à ce

chapitre, nous décrivons comment ce score peut servir de guide pour améliorer la précision de la prédiction en partant du résultant fourni par kPRED. Nous illustrons cela à travers l'exemple du développement d'une méthode basée sur un algorithme génétique.

Le **chapitre 4** est consacré à l'analyse de la relation séquence-structure au niveau du pentapeptide. Les défis et les avantages de faire de PENTAdb une base de données couvrant l'ensemble de la PDB sont décrits ici. La diversité structurale des pentapeptides est analysée à l'aide d'une approche statistique. Nous essayons de comprendre pourquoi un pentapeptide particulier préfère une conformation particulière dans une protéine particulière. Pour atteindre cet objectif, les pentapeptides sont analysés en fonction de leur accessibilité au solvant, des contacts polaires au niveau local, des liaisons hydrogènes et des cartes de contact au niveau global. Le pentapeptide "LGRRE" est pris comme exemple car il a été assigné aux 16 PB différents. Une analyse des racines des moyennes des carrés des écarts sur les valeurs angulaires (RMSDA) par BP a été effectuée sur les 61,6 millions de pentapeptides présents actuellement dans PENTAdb. Cela permet de mesurer comment les angles de torsions observés dans les structure s'écartent des valeurs standards définis pour chacun des BP. La relation entre la variabilité structurale d'un pentapeptide et sa prévisibilité par la méthodologie KPRED a été étudiée en utilisant un indice de variabilité structurale (SVI). L'interface Web pour accéder à la base de données PENTAdb et ses fonctions de recherche avancée sont décrites. Enfin, le chapitre se termine par une brève remarque concernant la relation entre le pentapeptide autour du site catalytique des enzymes et sa variabilité structurale en termes de SVI.

Au **chapitre 5**, nous décrivons une tentative de d'analyse de l'espace des repliements connus ce jour sous forme d'un graphe. L'analyse est effectuée sur l'ensemble de la PDB filtré à

100 % d'identité de séquence pour ne pas manquer à la diversité structurale. Une comparaison exhaustive de toutes les protéines dans cet ensemble a été réalisée à l'aide du programme PB-ALIGN (M Tyagi et al., 2008). PB-ALIGN est un outil de comparaison de structures qui utilise la représentation des structures protéiques en tant que séquences BP. Plus de 282 millions d'alignements structuraux ont été réalisés grâce à la rapidité et de la précision de PB-ALIGN. Les alignements de longueurs inférieurs à 60 résidus n'ont pas été retenus. Les alignements ont été effectués de manière récursive (les régions non alignées sont extraites et ré-alignées) afin de saisir un nombre maximal de similarité structurale entre les protéines. Différents seuils de score d'alignement ont été utilisés pour filtrer les résultats. Les graphes ont été construits à l'aide du logiciel Cytoscape (Shannon et al., 2003) pour représenter visuellement le regroupement des protéines en fonction de la similarité structurale. Le regroupement est basé sur le score d'alignement de sorte que deux protéines alignées avec un score d'alignement plus élevé sont plus proches dans le graphe. Les alignements ont été liés aux annotations provenant de la base de données SCOPe. Nous montrons qu'il est possible d'effectuer de nouvelles annotations structurales et fonctionnelles pour plus de 13000 protéines non annotés dans la PDB. La fiabilité dans ces annotations augmente avec le seuil du score d'alignement choisi.

Dans le **sixième et dernier chapitre**, les conclusions et les perspectives, nous récapitulons les principaux résultats de chacun des chapitres précédents et leurs implications quant à notre compréhension des structures protéiques.

Chapter 1. Introduction

Proteins are ubiquitous and indispensable in living systems. The number of proteins involved in mediating any given biological process is mind boggling. Let's try enlisting the proteins involved in the visual perception of say, the text in this thesis. The light reflected from the text enters the eye through the cornea. The cornea is rich in the protein albumin. The lens of the eye is made of type IV collagen protein. The focusing of the lens involves the ciliary muscles of the eye, composed of actin and myosin proteins. Next the light strikes the photoreceptor proteins- rhodopsin and photopsin that transduce the visual signal to the optic nerve. We already have six proteins in our list and the optical signal is yet to leave the eye, furthermore this is not an exhaustive list. At any given instance the human body requires approximately 20,000 to 30,000 different types of proteins for its normal functioning (Venter et al. 2001; Lander et al. 2001). No living organism on earth is exception to the fact that life relies immensely on the vital bio-molecule protein.

This thesis represents yet another attempt to understand a bit more about the structure and functioning of proteins. With reference to the main development of protein structural biology (Table 1), this chapter aims to provide an overview of the body of knowledge that we have accumulated by studying proteins for the past several decades.

1.1. Primary structure: string of amino acids & peptide bond

1.1.1. Amino acids

Amino acid represents the monomeric units that when combined in a specific order forms the polymer, which is primary structure of the protein. The specific order in which these

amino acids must be linked to form proteins is encoded in the DNA. The synthesis of the protein is carried out by the transcription and translation mechanisms of the cell. Nature has 20 amino acids in its repertoire that are mixed and matched in different combinations to form all the different proteins.

Table 1: Chronology of milestones in the development of protein structural biology.

Year	Key discoveries and developments in the field of structural biology
1951	Frederick Sanger determines the amino acid sequence of Insulin.
1962	John Kendrew and Max Perutz share the Nobel prize in chemistry for determining the atomic structure of myoglobin using X-ray crystallography.
1963	G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan develop the Ramachandran plot
1982	Nobel prize awarded to A. Klug for developing electron microscopy and discovering the structure of nucleic acid-protein complexes.
1988	Nobel prize awarded to J. Deisenhofer, R. Huber, H. Michel for determining the 3D structure of the photosynthetic reaction center.
1989	First NMR structure deposited in the PDB by Driscoll et. al.
1991	First EM structure deposited in the PDB by Henderson et. al.
1993	PDB has 1000 structures in its database.
1994	SCOP v1.0 is released by Alexey G. Murzin and colleagues.
1997	Nobel prize awarded to P.D. Boyer, J.E. Walker, J.C. Skou for elucidating the mechanism underlying the synthesis of ATP.
2002	Nobel prize awarded to K. Wuthrich for the development of NMR spectroscopy to determine the structure of biological macromolecules in solution.
2003	Nobel prize awarded to R. MacKinnon for potassium channels.
2009	Nobel prize awarded to V. Ramakrishnan, T.A.Steitz, A.E. Yonath for studies of the structure an function of the ribosome.
2014	PDB reaches a milestone of 100,000 structures in its database.

Louis Pasteur discovered chirality in molecules way back in 1848. All the amino acids have a carbon atom whose four valencies are satisfied by four different groups, making it chiral. Due the presence of a chiral center, amino acids can be found in two enantiomeric forms, L-amino acids and the D-amino acids. Proteins predominantly have L-amino acids in

them and the D- enantiomer is restricted to bacterial cell walls (Lam et al. 2009) and mammalian brain (Kim et al. 2010)

A hydrogen atom, a carboxyl group and an amino group satisfy the first three valencies. Since there are both an acidic and a basic group, amino acids are amphoteric in nature. The differentiating factor between the 20 amino acids is the chemical group that satisfies the fourth valency of the amino acid (see Figure 1). This chemical group is commonly called as the side chain of the amino acid. It is this side chain or R-group that lends unique characteristics to the proteins they occur in. Amino acids can be classified based on their polarity. Polar amino acids are hydrophilic, tend to interact with water and are generally found on the outer surface of proteins. The non-polar amino acids on the other hand tend to interact with each other to form hydrophobic interactions and are generally found in the hydrophobic core of the protein. The classification of the 20 amino acids are as follows:

- Polar (Hydrophilic)
- Uncharged side chain- Ser, Thr, Cys, Asn, Gln.
- Positively charged side chain- Lys, His, Arg.
- Negatively charged side chain- Asp, Glu.
- Non polar (Hydrophobic)
- Aliphatic side chain- Gly, Ala, Pro, Val, Leu, Ile, Met.
- Aromatic side chain- Phe, Tyr, Trp.

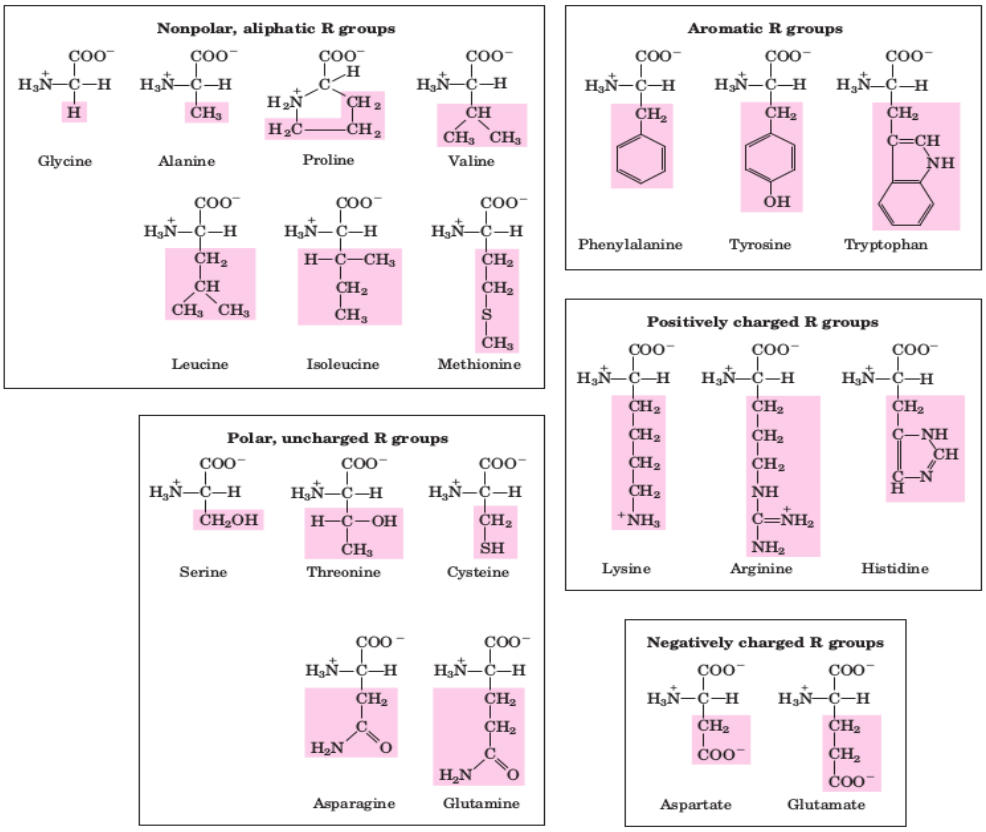


Figure 1.The twenty standard amino acids.
The classification is based on the physiological characteristics of their R-groups. Figure taken from Lehninger's Principles of Biochemistry 4th ed, 2004.

Other unique features of amino acids are also mentioned in Table 2.

Despite their fundamental role in the functioning of the body, the human body cannot synthesis all the 20 amino acids. We can synthesize 11 amino acids *de novo* and the remaining 9 amino acids need to be supplied from the diet (Young 1994). The classification of amino acids as essential and non essential is as follows:

- Non essential amino acids (can be synthesized *de novo*)- Ala, Arg, Asp, Cys, Glu, Gly, Pro, Ser, Tyr, Asp.

- Essential amino acids (needs to be incorporated from diet)- His, Ile, leu, Lys, Met, Phe, Thr, Trp, Val.

Table 2: A few amino acids with their unique features.

Amino acid	Unique feature
Proline	Presence of a cyclic imino group side chain disrupts secondary structure formation.
Cysteine	Ability to form disulphide bonds.
Tryptophan, Tyrosine and Phenylalanine	Absorb UV light at 280nm, helps in the experimental characterization of proteins.
Histidine	Has a side chain with pK_a close to physiological pH, plays a key role in enzymatic reactions.
Glycine	Has a hydrogen atom as the R-group making it the only non-chiral amino acid.

Apart from the 20 standard amino acids encoded in the genetic code, there are a few non-standard amino acids too. Most notably *N*-formylmethionine, pyrrolysine, 4-hydroxyproline, 5-hydroxylysine and selenocysteine. Further there are several amino acids that do not form a part of proteins and they are called non-proteinogenic amino acids (Wagner and Musso 1983).

1.1.2. Peptide bond

The amino acids in a protein chain are frequently referred to as "residues". This is because of the chemical reaction that connects them together. The α-carboxyl group of the nth amino acid undergoes a condensation reaction with the α-amino group of the $n+1^{th}$ amino acid to form the peptide bond. As a by-product of the formation of the peptide bond, a molecule of water is produced. What remains on the growing polypeptide chain after the elimination of the water is hence called the "residue".

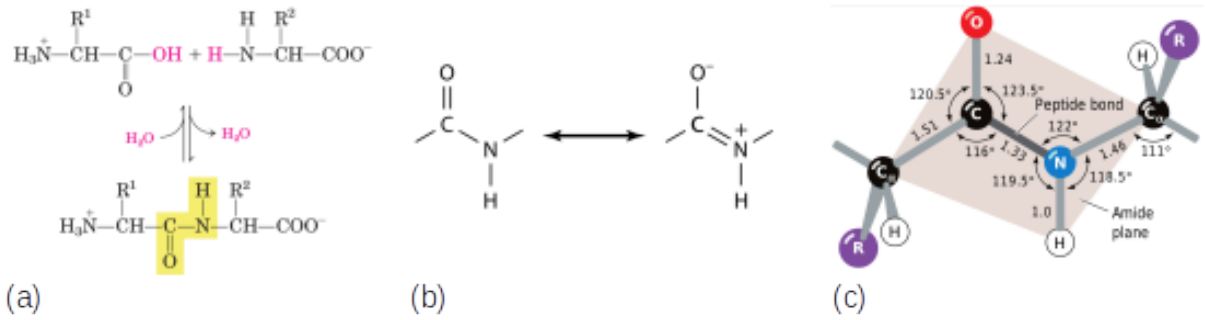


Figure 2. The peptide bond.
(a) A condensation reaction results in the formation of the peptide bond and elimination of a water molecule. (b) The partial double bond character of the peptide bond is due to the partial charges on the oxygen and nitrogen atoms. (c) Bond angles and bond lengths around the peptide bond, the peptide bond has a length intermediate (1.33 Å) to the C=N and C-N lengths. Figures taken from (Nelson and Cox 2004; Murray et al. 2003; Voet and Voet 2011) respectively.

There are n peptide bonds in a peptide containing n+1 residues. The key features of the peptide bond (Figure 2) are as follows:

- The rotation around the N-C bond is restricted due to its partial double bond character. The carbonyl oxygen has a partial negative charge (δ-) and the amide hydrogen has a partial positive charge (δ+). This makes the peptide bond planar.

- The carbonyl oxygen is in trans position to the amide hydrogen.

- While the peptide bond is not rotatable, the single bonds between $C_{\alpha 1}$-C (ψ) and N-$C_{\alpha 2}$ (ϕ) are rotatable.

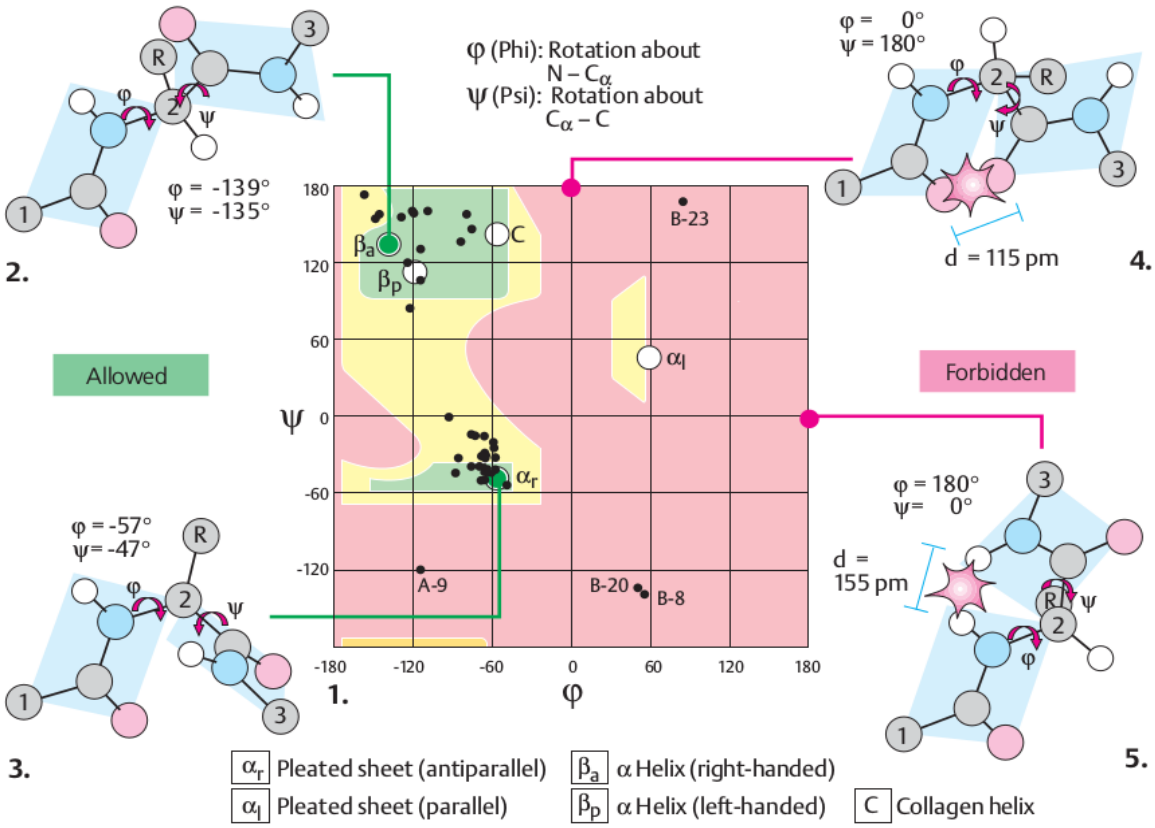


Figure 3. Ramachandran plot for all the residues of insulin.
The 8th, 20th and 23rd residues from the B chain (all glycine) are exception to the forbidden region (pink) of the ramachandran plot. Figure taken from (Koolman and Roehm 2005).

In 1968 GN Ramachandran and V Sasisekharan charted out the conformational space that amino acids could explore without having a steric clash between the R-group and the protein back bone (Ramachandran and Sasisekharan 1968). The resulting phi vs psi plots are famously called the Ramachandran plots. A sample ramachandran plot for insulin is depicted in Figure 3. Except glycine and proline, the remaining 18 amino acids have phi vs psi values that lie within the region indicated in yellow in the figure. Glycine can sample a wide range of phi vs psi values because of its very small R-group (Hydrogen atom) that has minimal steric hindrance. Proline is the other end of the spectrum with maximum hindrance.

1.2. Secondary structure: strands, helices and loops

When segments of the string of amino acids form regular structural patterns, they are called secondary structure elements. α-helices and β-strands are the two main classes of secondary structure elements (SSEs). Regions of the protein structure that cannot be classified into any one of these two classes are generalized as loops. Pauling, Corey and Branson had identified these SSEs in 1951 (Pauling and Corey 1951; Pauling, Corey, and Branson 1951) itself, much before the first 3D structure was determined by Kendrew in 1958 (Kendrew et al. 1958). SSEs can be characterized mainly based on three factors, the ϕ and ψ angles of the residues, hydrogen bonding pattern between the amine hydrogens and carbonyl oxygens and simply the interatomic distances of the protein backbone atoms (Levitt and Greer 1977; Colloc'h et al. 1993; Frishman and Argos 1995). Standard ϕ and ψ angles of the residues are featured in Table 3.

There are many programs that assign SSEs for proteins whose atomic co-ordinates are known based on one or more of the above characteristics. Some of them are- DSSP (Kabsch and Sander 1983), STRIDE (Heinig and Frishman 2004), KAKSI (J. Martin et al. 2005), DEFINE (Richards and Kundrot 1988), SST (Arun S Konagurthu, Lesk, and Allison 2012) etc. These different methods do not agree with each other on their SSE assignments, especially on the boundaries between the regular SSE and loops (Manoj Tyagi et al. 2009). Among the methods currently available, DSSP is the most widely used. The variations in the SSEs are described in the next subsections.

1.2.1. Helices

A spiral arrangement of the protein backbone around an imaginary axis forms an helix. The α-Helix is the most common form of the helix and also SSEs in general because of its

energetically favorable arrangement of hydrogen bonds. It has 3.6 residues per helical turn, pitch of the α-helix is 5.4 Å and is always right handed. The α-helix has hydrogen bonds between the $(n+4)^{th} \rightarrow n^{th}$ residue (Pauling, Corey, and Branson 1951; Kendrew et al. 1960). The 3_{10} and π helices are the less frequently observed variants of the helices. The 3_{10} helix derives its name due to the fact that it has 3 residues per turn and 10 atoms are involved in hydrogen bonding (Donohue 1953). Going by the same nomenclature, the α-helix is 3.6_{13}. The 3_{10} helices are generally found at the boundaries of large α-helices and have hydrogen bonding between $(n+3)^{th} \rightarrow n^{th}$ residue (Armen, Alonso, and Daggett 2003). π helices are found in the functional sites of proteins (Weaver 2000), have hydrogen bonding between $(n+5)^{th} \rightarrow n^{th}$ residue and 4.1 residues per turn (Low and Baybutt 1952).

Table 3: Regular secondary structure elements (SSEs) and their corresponding f and y values.

Secondary Structure	ϕ (deg)	ψ (deg)
Right-handed α helix (α)	−57	−47
Parallel β pleated sheet (↑↑)	−119	113
Antiparallel β pleated sheet (↑↓)	−139	135
Right-handed 3_{10} helix (3)	−49	−26
Right-handed π helix (π)	−57	−70
2.2_7 ribbon (2)	−78	59
Left-handed polyglycine II and poly-L-proline II helices (II)	−79	150
Collagen (C)	−51	153
Left-handed α helix (α_L)	57	47

[After Flory, P.J., *Statistical Mechanics of Chain Molecules*, *p*. 253, Interscience (1969); *and* IUPAC-IUB Commission on Biochemical Nomenclature, *Biochemistry* **9**, 3475 (1970).]

The ϕ and ψ angles of the helices tend to lie at the interface between the 2nd and 3rd quadrant of the Ramachandran map (Figure 3) (Ramachandran and Sasisekharan 1968). The amino acids Alanine, Leucine, Arginine, Methionine, Lysine and Glutamine have high propensity to form helices whereas Proline and Glycine tend disrupt the helix (Chou and

Fasman 1974; Nick Pace and Martin Scholtz 1998). At a sub secondary structure level, helices show irregularities in the form of kinks and bends, extensive work has been done in characterizing these structural features and their impact on the proteins in which they occur (Bansal, Kumar, and Velavan 2000; Kneissl et al. 2011; Law et al. 2016).

1.2.2. β- Sheets

β-sheets are the other regular SSEs that are found in protein structures. They are formed by multiple β-strands lying next to one another connected by hydrogen bonds. β-sheets occur in three forms- parallel, anti-parallel and mixed. As the name suggests, the constituent β-strands in these structures run parallel, anti-parallel or in a combination of these. The β-strands that lie next to each other in a β-sheet can be from distal parts in the protein chain or from different poly peptide chains altogether. In parallel β-sheets the units repeat every 6.5 Å and have a disordered hydrogen bonding pattern. Hence they are less favorable than the anti parallel β-sheets, that have linear hydrogen bonding with units repeating every 7.0 Å. Larger β-sheet arrangements are rarely flat and occur frequently with a right handed twist.

1.2.3. Loops and turns

Loops are the SSEs that are irregular in structure and cannot be classified as either α-helix or a β-sheet. Loops and turns connect the other two regular secondary structure elements and correspond to approximately 50% of the protein residues. Despite their irregular nature, loops serve several important functions like nucleotide binding as in the helix-loop-helix motiff, epitopes of antibodies, structures that form ligand controlled switches etc. Turns on the other hand serve a more specific function of reversing the direction of a polypeptide chain by 180°. They are stabilized by hydrogen bonds between the 1st and 4th residues and frequently

connect adjacent strands of an anti-parallel β-sheet. Proline and glycine are highly represented in the turns.

1.3. Tertiary and quaternary structure

Tertiary structure refers to the arrangement of SSEs in 3D to form domains and also the inter-relation between these domains. Super secondary structures or motiffs are regular structural patterns that do not qualify as domains. Domains in addition to having a well defined structural signature, perform a specific function and can fold and evolve independent of the rest of the protein. Proteins can have a single or multiple domains depending on their function. For example protein kinases have two domains, one ATP binding domain and the other to bind to the substrate. Tertiary structures are stabilized by different weak interactions with the exception of the disulfide bond, a covalent bond. The disulfide bond contributes to stability by being strong whereas ionic interactions, hydrophobic interactions and hydrogen bonds contribute by being in large numbers. Also the interaction of the residues with the solvent, usually water establishes a hydrophobic core and a hydrophilic exterior for the protein.

Some proteins are composed of multiple polypeptide chains. The quaternary structure represents the spatial orientation of these polypeptide chains. These chains can be identical or different giving raise to homopolymeric or heteropolymeric proteins. The distinct chains are represented by the Greek alphabets (α, β, γ...) and their counts as subscripts. For example haemoglobin is $\alpha_2\beta_2$.

1.4. Fragment libraries and structural alphabets

By the late 1980s it had become evident that the three state i.e. helix, β-sheet and loop classification of protein structure could not efficiently capture all the idiosyncrasies of protein

local structure. The description of sub secondary structure features and the sub classification of loops required a more assorted library of repeating structural elements. Thus several groups started developing fragment libraries that could serve this purpose. The objective of these fragment libraries is to capture the structural redundancy underlying in protein structures. The fragment libraries differed in the criteria used for the classification, size of the fragments and the number of fragments.

The first fragment library to be developed was a collection of *building blocks* by Unger and colleagues in 1989 (Unger et al. 1989). They observed that the structures corresponding to hexamers in protein structures could be clustered into discrete groups that could faithfully recreate any protein structure with a minimal RMSD. The library consisted of a collection of 103 *building blocks* obtained from a dataset of 82 non homologous protein structures. The clustering algorithm used was K-nearest neighbor method (Lu and Fu 1978) and the criteria to become a member of a cluster was that the hexamer had to be structurally similar to any other member of the cluster with an C_α RMSD of =<1 Å. Further to be chosen as a *building block* from a cluster, the hexamer had to satisfy a more stringent criteria that it had to have at the maximum 1 Å RMSD compared to all members of the particular sub-cluster. This fragment library could approximate all parts of the 82 proteins with a mean RMSD of 0.74 Å.

Table 4: Example of features used to define fragment libraries or structural alphabets.

Criteria used for clustering	**Clustering method**
C_α RMSD, C_α distances, Dihedral angles, α angle	k-means, Kohonen maps, hierarchical clustering, HMM, ANN
Fragment size	**Library size**
4,5,6,7,8,9 and combinations thereof	Above 4 up to a few hundreds

Several other libraries were developed by considering the C_α RMSD as a criteria for clustering. Rooman et. al. (Rooman, Rodriguez, and Wodak 1990) used hierarchical clustering to identify fragments ranging in length from 4 to 7 amino acids from a collection of 75 well refined and low sequence identity structures. Despite defining only 4 recurrent local structural motifs, they could differentiate between secondary structure elements like extended and curved coils and also β-bulges from β-strands. Similarly Micheletti et. al. (Micheletti, Seno, and Maritan 2000) and Kolodny et. al. (Kolodny et al. 2002) also use fragment lengths varying between 4 and 7 residues but with significantly larger library sizes. The library size also varies for different fragment lengths. Micheletti et. al. use an iterative clustering algorithm similar to the one used by Lacey and Cole (C. Lacey and S. Cole 1993) to identify the structural patterns they call *oligons*. They reported that the top 15 ranking pentamers and hexamers contribute to 75% and 47% of the entire database of fragments respectively. Kolodny et. al. use the K-means simulated annealing technique on a training set of 200 proteins. They tested the ability of their fragment libraries to reproduce the original protein structure on a test set of 145 structures. The criteria tested were the local and global C_α coordinate RMSD for fragment of length 4,5,6 and 7. The relationship between complexity i.e. number of states per amino acid and RMSD are also discussed. An important factor to remark here is that their library was created by clustering on non-overlapping fragments.

The distance between the C_α atoms in the protein backbone has been considered as a criteria for clustering short peptide fragments. Fetrow et. al. (Fetrow, Palumbo, and Berg 1997) used an auto associative artificial neural network (autoANN) based clustering method on a collection of 116 non-redundant proteins from the PDB SELECT database (Griep and Hobohm 2010). The fragment size was selected as 7. Apart from the 15 non consecutive C_α

distances, five virtual bond angles and four virtual dihedral angles were also used as input to the autoANN. All of these values were normalized to vectors having lengths between 0 and 1 as this is a prerequisite for ANN. A total of 6 prototypes were identified, represented by α, β, ζ, η, τ & ι and called structural building blocks (SSBs). They reported the utility of SSBs in identifying regular secondary structure elements, their capping regions and regular patterns in "random coils". Amino acid propensities in SSBs were also studied.

Similarly Camproux and colleagues (Camproux et al. 1999) used non-consecutive C_α distances from 100 proteins of the PDB SELECT database to define short structural building blocks (SSBBs). Atomic co-ordinates and a consensus secondary structure prediction from multiple methods were used as input to a hidden Markov model (HMM). The input is in the form of N four dimensional vectors, where N is the number of fragments of length four which is the chosen fragment length. Each vector of length four corresponds to three non-consecutive C_α distances and an angle formed between the 4th C_α and the plane containing the other three C_α atoms. A total of 12 SSBBs were identified and were classified into 5 clusters $A(\alpha_1\ \alpha_2)$, $A'(\alpha',\ \alpha'_-,\ \alpha'_+)$, $C(\gamma_1, \gamma_2, \gamma_3)$, $B'(\beta'_+, \beta'_-)$, $B(\beta_2, \beta_1)$. The SSBBs are assigned to the third residue of the tetra peptide. These 12 fragments are used to study the connections between regular secondary structures and also repeating patterns in the original dataset of 100 proteins (A. C. Camproux et al. 1999). Later in 2004 they expanded the fragment library to a size of 27 fragments (A.C Camproux, Gautier, and Tufféry 2004).

Sander et. al. (Sander, Sommer, and Lengauer 2006) also used a collection of proteins that shared 20% or less sequence homology from the PDB SELECT dataset (Griep and Hobohm 2010). They used the HSSP database (Schneider, de Daruvar, and Sander 1997) to build sequence profiles but also iteratively searched the SWISS-PROT to identify the profiles

for proteins not covered in HSSP. A total of 27 structural prototypes were identified from a dataset of 1999 proteins. The fragments were of length 7 residues and were clustered using the Leader algorithm (J. Hartigan. 1975) and k-means clustering was used for iterative refinement based on Cα-Cα distance matrix comparisons.

A set of 27 vectors called *protein folding shape codes* (PFSC) were used to describe 3D folding shapes of 5 successive C_α atoms by Yang (Yang 2008). The PFSC were represented by the 26 alphabets A-Z and the $ sign. These PFSCs were identified from 268 structures from the SALIGN benchmark (Marti-Renom, Madhusudhan, and Sali 2004). The vector space is equally divided based on the two torsion angles and the pitch distance (distance between $C_{a(n-2)}$ and $C_{a(n+2)}$) into three parts in each case. Even though 9 of the 27 PFSCs cover 97% of the local structures and PFSC M is never observed in the SALIGN dataset, they are included for the completeness of the vector space. The 20 different conformations of *E.coli* glutaredoxin obtained by solution NMR were analyzed using PFSC. The performance of PFSC was compared to SSB (Camproux et al. 1999), PB (Brevern, Etchebest, and Hazout 2000) and other popular secondary structure assignment methods.

Tung et. al. (Tung, Huang, and Yang 2007) studied pentamers from 1348 protein structures with the main objective of developing a tool called 3D- BLAST. The structural alphabets were defined by a κ–α angle plot. κ angle is the virtual bond angle between any three consecutive C_α atoms and α angle is the virtual torsion angle between four consecutive C_α atoms. 23 structural letters were identified by nearest neighbor clustering. Of the 23 structural letters, 8 described helices and 5 described β-sheets. They built a 23 X 23 structural alphabet substitution matrix (SASM) similar to the BLOSSUM 62 matrix. This SASM is used in conjugation with the BLAST algorithm (Altschul et al. 1997) to develop the 3D-BLAST

tool. Further the structural alphabet database (SADB) and E-value thresholds were also established. The performance of 3D-BLAST is compared with PSI-BLAST (Altschul et al. 1997), CE (I N Shindyalov and Bourne 1998) and MAMMOTH (Ortiz, Strauss, and Olmea 2009).

Prestrelski et. al. (Steven J. Prestrelski, Williams, and Liebman 1992) identified fragments of length 8 residues from a collection of 14 proteins that could be classified into 113 classes. The criteria that they used for classification were linear distance (LD) and backbone dihedral angle (BDA) or α torsion angle. Fragments that were observed repeatedly in their dataset of proteins were called *substructures*. They used cost function as a measure of the similarity (or dissimilarity) between the *substructures* that were being compared. They reported that 67% of the residues from the dataset of 14 proteins could be represented by the *substructures.* When dataset size is increased from 14 to 20, the coverage increases to 77% on introducing redundancy into the dataset this further increases to 96%. In a separate paper they discuss the correlation between these *substructures* and spectra obtained from Fourier transform infrared spectroscopy (FTIR) (S J Prestrelski, Byler, and Liebman 1992).

In 1996, Schuchhardt and colleagues (Schuchhardt et al. 1996) used the Kohonen unsupervised learning algorithm on a dataset of 136 non-redundant proteins (Griep and Hobohm 2010). The 100 local structural motifs are organized in the form of a 10 X 10 2D Kohonen feature map. They report sequence conservation in members belonging to a specific motif cluster. The feature map is color coded to represent motifs corresponding to α-helices in red and β-sheets in yellow. Any protein can be represented as a collection of the coordinate of this feature map. This form of representation was also used to compare full length protein structures. The four proteins that were chosen for this comparative analysis were cytochrome

b_{562}, T4 lysozyme, cytochrome b_5 and γ IV crystallin. Representation in the form of histograms and trajectories for these proteins helps in visually understanding the difference in proportion of α helices and β-sheets.

Ku and Hu (Ku and Hu 2008) used a combination of self organizing maps (SOM), minimum spanning trees and k-means clustering to identify 18 structural alphabets from a non redundant subset of the PDB. Similar to *protein blocks (Brevern, Etchebest, and Hazout 2000)*, they used 5 amino acid long fragments defined by eight torsion angles. A substitution matrix called TRISUM 169 (TRained Iteratively for SUbstitution matrix SCOP 1.69) was created by self training on proteins from SCOP v1.69. This substitution matrix was used with the FASTA algorithm to create a structural alphabet based alignment tool called SA-FAST.

HSSP is a database of multiple sequence alignment of proteins belonging to the same family (Schneider, de Daruvar, and Sander 1997). Bystroff and Baker (Bystroff and Baker 1998) studied 471 such families that have a representative as a member of the PDB SELECT-25 list (Griep and Hobohm 2010). They clustered protein segments of 3-15 residues in length using the K-means algorithm (Everitt et al. 2011). The clustering criteria was sequence profiles. Two other factors used to evaluate how a cluster was organized were RMS distance matrix error (*dme*) and maximum deviation in backbone torsion angles (*mda*). A library called the *I-sites library* was created with 82 clusters grouped into 13 classes. The ability to predict protein local structure using fragments of length 8 was evaluated using *confidence values*. The predictions by the *I-sites library* was compared with the conventional secondary structure predictions by the PHD server (Burkhard Rost and Sander 1994). The predictions by the *I-sites library* had an Q_3 of 64%, but when both methods were used in conjugation they consistently performed better than PHD alone.

Protein Blocks (PBs) are a 16 letter structural alphabet developed by de Brevern and colleagues in 2000 (Brevern, Etchebest, and Hazout 2000). They showed that the orientation of the backbone atoms of any five consecutive residues in protein structures can be defined by 16 sets of 8 torsion angles (ϕ,ψ). PBs are defined by these 16 structural prototypes. An example of representative pentapeptides for each of these sets is given in Figure 4. These standard torsion angles were obtained by performing an unsupervised clustering (Kohonen 1982) of the backbone torsion angles of pentapeptides from a non-redundant collection of proteins. The PBs are represented by the alphabets *a* to *p.* In 2005 Etchebest et. al. (Etchebest et al. 2005) demonstrated the robustness of PBs in approximating protein structures even after a three fold increase in the size of the PDB.

PBs have been used for the development of several structure analysis tools and many research groups have predicted protein local structure in terms of PBs. As the theme of this thesis revolves around the utility of PBs for structure comparison and prediction, a more detailed description of PBs have been included in the context of the appropriate chapters.

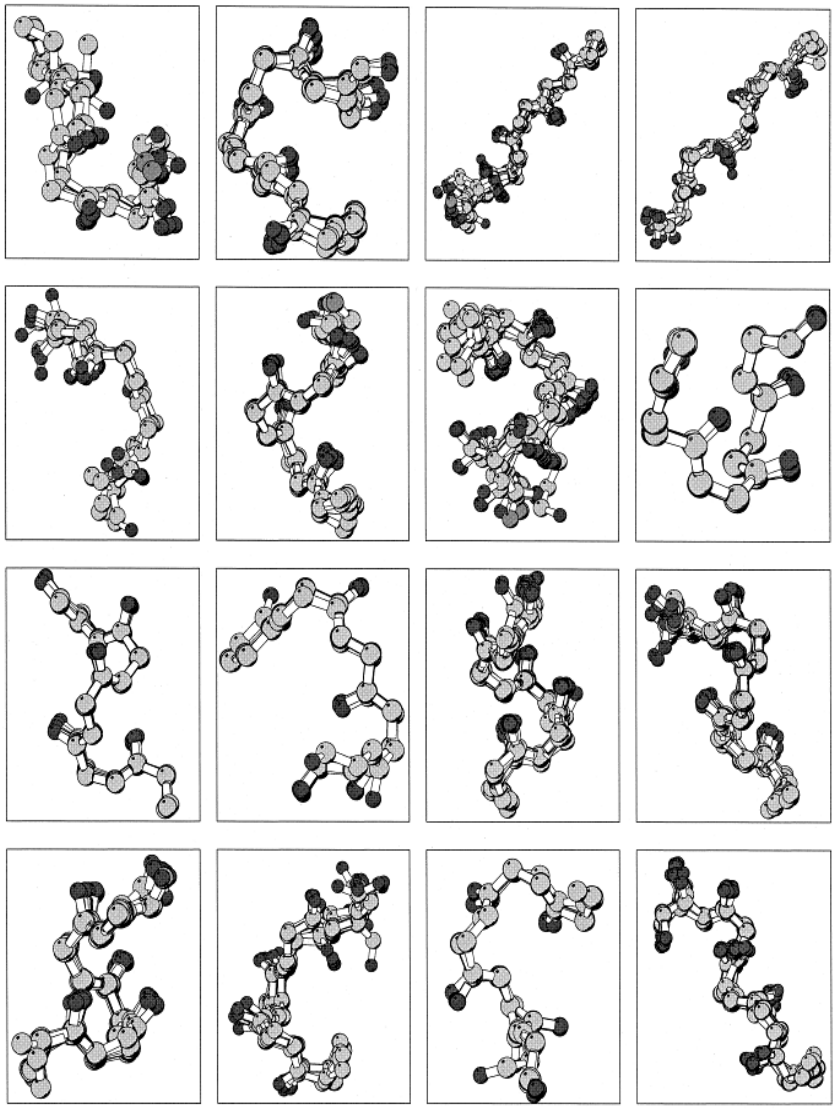


Figure 4. Examples of pentapeptides representing each of the 16 structural prototypes that constitute the protein blocks (PBs) structural alphabet.
Figure was taken from de Brevern et. al. 2000 (de Brevern, Etchebest, and Hazout 2000).

1.5. Protein structure determination

Determining the structure of proteins at atomic resolutions is mainly possible due to x-ray crystallography (90% of structures) but also by NMR spectroscopy (9% of structures) and electron microscopy (1% of structures). The 125,000+ protein structures solved until today can be accessed through the websites of any one of the members of the wwPDB (PDBe,

RCSB PDB and PDBj). While the growth of the PDB has been exponential in past 20 years, the growth of the sequence database has been even more impressive (Figure 5). The Uniprot sequence database is approximately 600 times larger than the PDB (Pundir et al. 2015). It has been a continuous quest for structural biologists to bridge this gap between the sequence and structure databases.

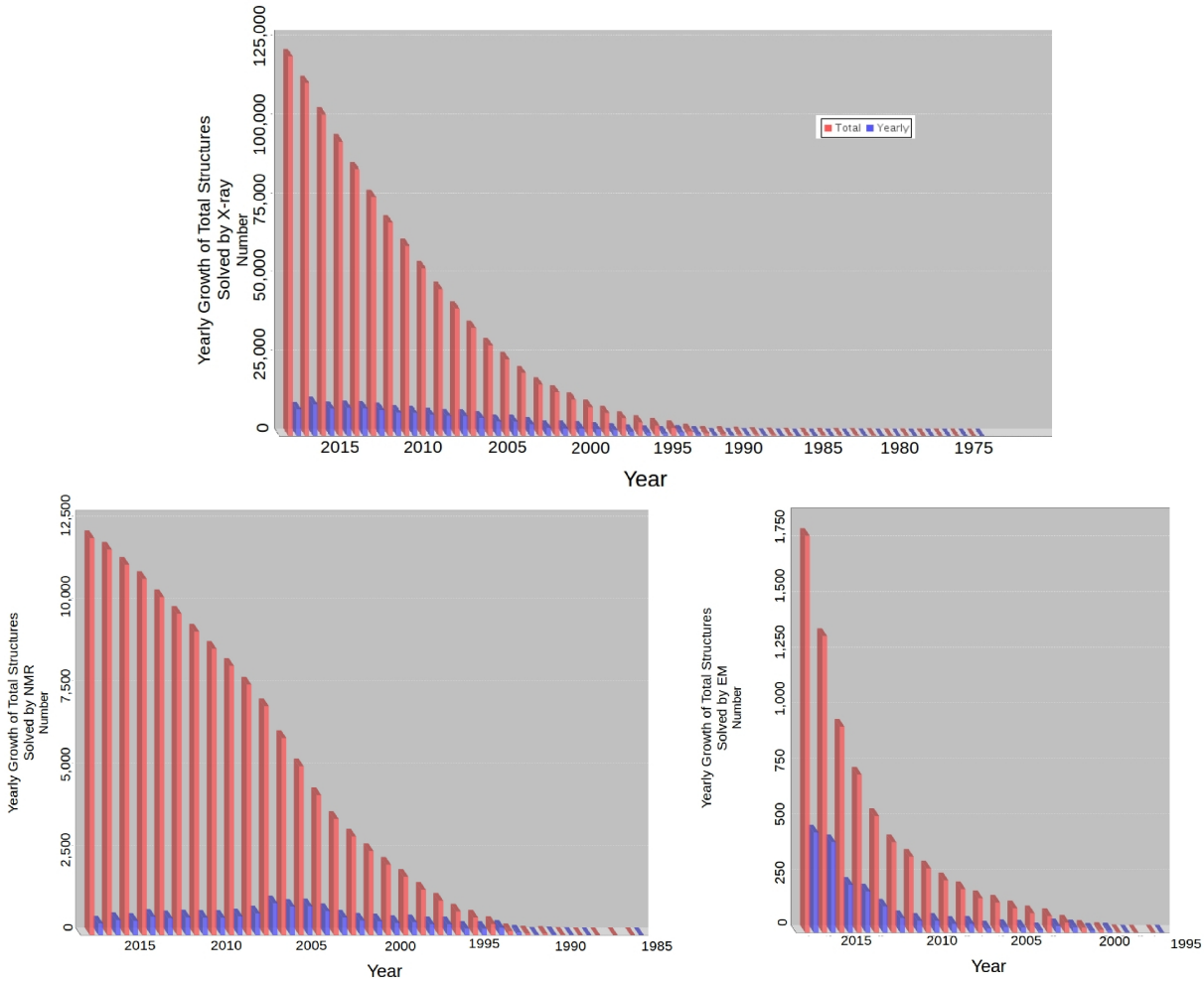


Figure 5. Growth in the number of structures deposited in the PDB.
Shown are number of deposited structures after X-ray crystallography, NMR spectroscopy and electron microscopy methods from 1972-2016.

X-ray crystallography was used initially to determine the structure of small molecules, the first structure to be determined was that of table salt in 1914. Use of crystallography to solve the structure of biological macromolecules like proteins began in the late 1950s. John Kendrew and Max Perutz shared the Nobel prize in chemistry in 1962 for their work on

solving the structure of sperm whale myoglobin (Kendrew et al. 1958). The prerequisite for x-ray crystallography is to obtain the protein in a pure crystalline form. Crystallization of the protein is achieved by trying out different combinations of pH, solvents and precipitating agents. Highly monochromatic x-ray beams are used to obtain diffraction patterns of the protein crystals and computer algorithms are used to decipher these diffraction patterns. While there have been large developments in computer tools to analyze diffraction patterns, the limiting factor for x-ray crystallography has been to obtain sizable protein crystals.

Nuclear magnetic resonance (NMR) spectroscopy of proteins acts as an complimentary method to x-ray crystallography as it does not limit the study to proteins that are crystallizable. The first structure to be determined by NMR spectroscopy was that of a proteinase inhibitor by Kurt Wüthrich and colleagues in 1985 (Williamson, Havel, and Wüthrich 1985). It is based on the principle that the absorption of radio frequency electromagnetic waves by the nuclei of certain atoms is a characteristic of that atom. The shift in the resonance frequency of the nuclei after absorption of the radio frequency waves is called chemical shift. Analyzing chemical shift data gives insights into the local geometry and inter atomic distances. Two dimensional NMR spectroscopy permits us to study the 3D structure of proteins and more importantly also their dynamics. This technique is limited by the size of proteins that it can analyze. Larger proteins result in overlapping peaks that are difficult to resolve.

Electron microscopy is the third type of structure determination methodology wherein the protein is *fixed* by cross-linking or freezing in vitreous ice, then electron beams are used to *visualize* the protein as electron beams have wavelengths 0.1 million times shorter than visible light hence a better resolving power. The latter method of sample (protein) preparation is

called cryo-electron microscopy and is used in preference to the conventional electron microscopy to avoid vacuum damage to the sample. Cryo-electron microscopy is typically well suited for large proteins or protein assemblies and produces low resolution structures. But this is changing rapidly with recent developments in the field.

1.6. Protein structure prediction

The short comings of the structure determination methods namely X-ray crystallography, NMR spectroscopy and electron microscopy in keeping up with the growth of the sequence databases has resulted in the proliferation of the structure prediction methods. Secondary structure prediction methods stop at the level of predicting the stretches of the protein sequence that corresponding to α-helices, β-strands, loops and turns. This information itself can significantly boost fold recognition, protein classification and also be used as input for ab-initio structure prediction. The use of information theory based techniques, machine learning algorithms and hidden markov models (HMM) has resulted in secondary structure prediction accuracies to reach 80% and above. Some of the popular secondary structure prediction tools are- GOR-IV (Garnier, Gibrat, and Robson 1996), YASSPP (Karypis 2006), PSI-PRED (Jones 1999), PHD (B Rost and Sander 1993), SSpro/ACCpro 5 (Magnan and Baldi 2014).

Tertiary and quaternary structure prediction methods fall under two categories, namely the ab-initio and homology modeling methods. The ab-initio approach predicts protein structure based on first principles, the physical interactions between all the atoms that constitute the protein. The objective is to attain a global energy minimum conformation by using an energy function as a guideline. The popular Folding@home (http://folding.stanford.edu) project is an ab-initio approach that utilizes distributed computing

to overcome the high computational requirement of ab-initio approaches. The homology modeling approach relies on the fact that protein structure is significantly more conserved than protein sequence, hence taking cues from homologous proteins. The performance of such methods depends on the ability to efficiently detect distant homologues.

1.7. Comparing and classifying proteins

Since the development of the PAM (point accepted mutation) matrices (Dayhoff, Margaret O., Schwartz R.M. 1978) by Margaret Dayhoff in 1978 for the purpose of aligning protein primary sequences, there have been a large number of advances for comparing proteins. Several programs have been developed to perform pairwise comparisons, multiple comparisons (multiple sequence alignment) and also build phylogenetic trees. Clustal Ω (Sievers and Higgins 2014), T-Coffee (Notredame, Higgins, and Heringa 2000), MUSCLE (Edgar 2004) and MAFFT (Katoh et al. 2005) are a few such commonly used programs. The Clustal family of programs are the most widely used programs for aligning protein sequences with each version improving on the previous. It derives its name from the fact that it works based on the UPGMA clustering approach (Higgins and Sharp 1988). It supports multiple input and output formats and provides the user an option to supply the guide tree based on which the multiple sequence alignment would be generated. MUSCLE and MAFFT also derive their names from the methods they are based on, **L**og-**E**xpectation and **F**ast **F**ourier **T**ransform respectively. T-Coffee is similar to Clustal but with additional features to analyze, combine and convert alignments between different formats. All the methods work for aligning nucleotide sequences as well. The popular program BLAST (Altschul et al. 1990) helps in comparing a given protein sequence with all the other proteins present in a protein database and the PSI-BLAST performs the same process iteratively to fish out distant homologues.

The similarity between the compared protein primary sequences are reported in terms of percentage sequence identity. It is easy to infer evolutionary relatedness among proteins sharing 25% or more sequence similarity but when the similarity drops below this threshold, structural alignment has to be performed to confirm relatedness. This ambiguity in determining the relationship between proteins sharing sequence identity in the vicinity of the twilight zone (25% identity) has led to the development of the protein structure alignment methods (William R. Taylor and Orengo 1989). There are several examples where the proteins have completely unrelated functions despite having high sequence identity and also the vice-versa. Hence the presence of sequence similarity between proteins only implies a evolutionary relatedness and the presence of a common function cannot be guaranteed.

A number of 3D structure comparison programs have been developed to ascertain functional and evolutionary relationships, to perform database scans, to help in automation of protein classification etc.. 3D comparison is difficult not only due to the addition of another dimension of complexity but also due to the need to differentiate between similarities arising due to constraints in folding versus true evolutionary relatedness (Murzin 1994). The diversity in structure comparison methods arise due to both the "criteria" used for comparison and the "method". The most commonly used criteria is the occurrence of similar short fragments between the two proteins to be aligned and commonly used method is the dynamic programming approach. A non-exhaustive list of comparison methods grouped based on the criteria is given below, methods that use dynamic programming are highlighted in red.

- **Inter residue distances:** SSAP (W R Taylor and Orengo 1989), COMPARER (Šali and Blundell 1990) , DALI (Holm and Sander 1993)

- **Fragment similarity (fragment size):** DALI(6), CE(8) (Ilya N Shindyalov and Bourne 1998), MUSTANG(>6) (A S Konagurthu et al. 2006), MAMMOTH(7)(Ortiz, Strauss, and Olmea 2002), MULTIPROT(>3) (M Shatsky, Nussinov, and Wolfson 2004), FATCAT(variable) (Y. Ye and Godzik 2003), FlexProt(variable) (Maxim Shatsky, Nussinov, and Wolfson 2002), YAKUSA(4) (Carpentier, Brouillet, and Pothier 2005), SALAMI(6) (Margraf, Schenk, and Torda 2009), 3D-BLAST(5) (Tung, Huang, and Yang 2007), SA-FAST(variable) (Ku and Hu 2008), PBALIGN(5), iPBA(5), mulPBA (5)

- **Sequence alignment:** PROFIT (Martin, A. C. R. and Porter, C. T.), STAMP (Russell and Barton 1992)

- **Secondary structure patterns:** SSM (Krissinel and Henrick 2004), GANGSTA (Kolbeck et al. 2006), VAST (Gibrat, Madej, and Bryant 1996), TMALIGN (Zhang and Skolnick 2005), MASS (Dror et al. 2003)

The famous program DALI splits the protein into hexapeptides and then computes the inter-residue distance matrix, hence it is featured in both of the categories. Apart from dynamic programming, some of the other commonly used methods are Monte-carlo approach, least square fitting, geometric hashing, graph matching etc.. Some programs take into consideration additional features like residue environments, solvent accessibility etc..e.g. COMPARER, CLICK (Nguyen and Madhusudhan 2011). A subset (e.g. FATCAT, FlexProt) of these methods take into consideration the flexible nature of proteins and introduce twists to improve the alignment. This greatly improves the alignment results for proteins that undergo rigid body motions (see Figure 6 for an example). Apart from being able to align two protein structures, some programs can align multiple proteins as well (e.g. MUSTANG, MAMMOTH, MULTIPROT).

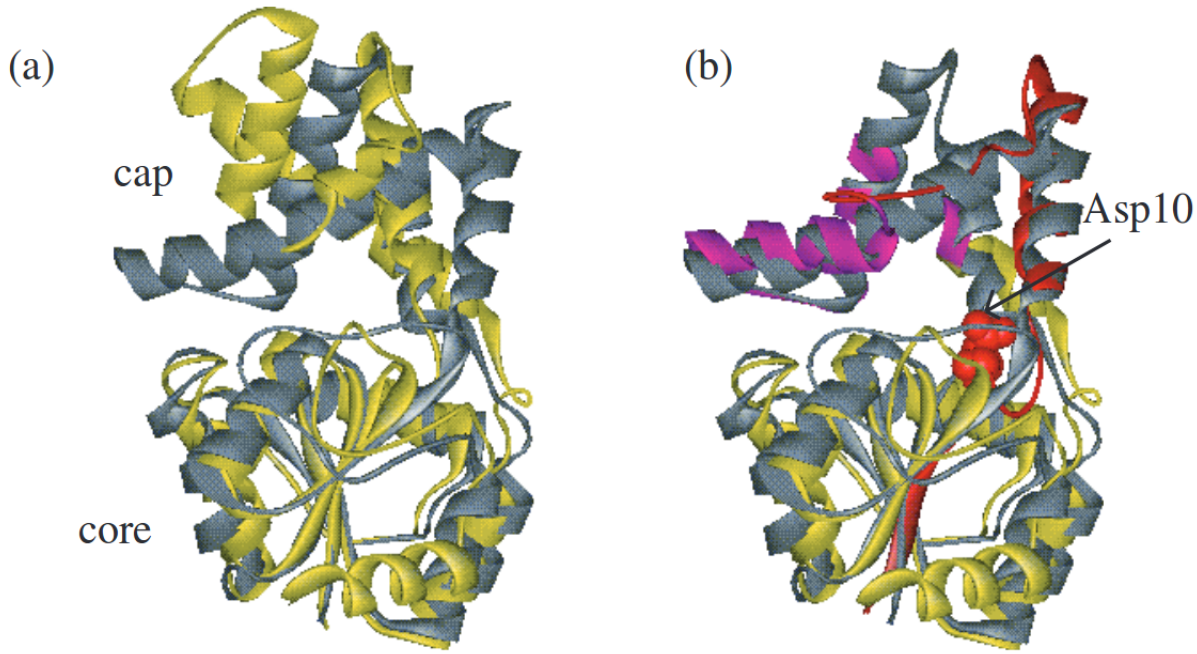


Figure 6. Rigid vs flexible superposition of L-2-haloacid dehalogenase (grey) and β-phosphoglucomutase (yellow) by the FATCAT program.
(a) In rigid superposition the core domains are well aligned but not the cap domains. (b) Flexible superposition greatly improves the alignment of the cap domain, indicated in red and pink. The nucleophile Asp10 of β-phosphoglucomutase is shown in CPK. Figure taken from Ye and Godzik, 2003 (Ye and Godzik 2003).

PB-ALIGN is a PB based alignment tool that relies on a dedicated PB substitution matrix (M Tyagi et al. 2008). All the alignments for the project "Comprehensive networks of protein structures" described in Chapter 5 of this thesis have been performed using PB-ALIGN. A detailed description of this method has been included in Chapter 5. Later both the underlying substitution matrix and the alignment methodology were improved to create the iPBA program (Joseph, Srinivasan, and de Brevern 2011). A variant of the iPBA program dedicated to aligning multiple protein structures called mulPBA was also developed (Joseph, Srinivasan, and de Brevern 2012). PB based alignment tools are not affected by the rigid body motions because the hinge region where the motion occurs are indicated by a few PB changes and the rest of the protein is well aligned e.g. like the cap domain in the above example in Figure 5.

After pairwise and multiple structure alignments, the natural progression was to organize all the known structural information hierarchically. This effort was spearheaded by Alexey G. Murzin and colleagues in the form of the SCOP database (Murzin et al. 1995). The SCOP database uses the hierarchy of Class, Fold, Superfamily, Family, Protein and Species to organize proteins. SCOP is unique in a way that all of its contents are manually curated and every entry in the database is backed by human expertise. But this limited the ability of SCOP to be comprehensive, especially during the 21st century that saw an exponential growth in the protein structural knowledge. SCOPe builds upon the SCOP database by incorporating automation methods and hence drastically increasing coverage (Fox, Brenner, and Chandonia 2014). SCOP in its original hierarchical format was developed till 2009 (v1.75) and more recently SCOP 2 has shifted to a network representation to better represent the relationships between proteins (Andreeva et al. 2014).

CATH is another scheme for the hierarchical classification of protein structures (Orengo et al. 1997). The hierarchy in CATH is represented as Class, Architecture, Topology and homologous superfamily. The CATH database is updated annually and currently holds more than 300,000 protein domain entries. A more recent evolutionary relationship based classification system called ECOD has been developed in the Grishin lab (Cheng et al. 2014) . ECOD gives importance to homology rather than topology for classification and is maintained up-to-date corresponding to the weekly update cycle of the PDB. All these classification schemes are based on the premise that proteins are composed of functionally distinct subunits called domains. Pfam (Finn et al. 2014) is a database of protein families generated using hidden Markov models and in contrast to the other methods is purely sequence based.

1.8. Objectives

All the work contained in this thesis revolves around using the structural alphabet, Protein Blocks for comparing and predicting protein structure and also understanding the relationship between sequence and structure at the pentapeptide level. The central objective has been to demonstrate the utility of a structural alphabet in attaining a satisfying compromise between the detailed atomistic representation of a protein versus a simplistic three state representation as a combination of secondary structure elements.

The first objective was to get a complete picture of the local structural variations in proteins in terms of PB substitutions. Towards this objective, we continued on the foot steps of previous works on observing the PB substitutions in homologous proteins (Manoj Tyagi et al. 2006; Gelly et al. 2011) and NMR ensembles (Mahajan et al. 2014). We wanted to focus on the PB variations from alternate crystallographic structures of the same protein and also during the course of a molecular dynamics (MD) simulation. We were interested to compare the results we obtained to the previously established PB substitution matrices. Further we wanted to go into the details of how much each of the 16 different PB were contributing to the variations. Finally we also wanted to explore the potential of PBs as a tool to analyze MD simulation trajectories.

Secondly, we wanted to incorporate the key findings of Kabsch and Sanders (W Kabsch and Sander 1984), i.e. protein local structure is highly dependent on sequence context into PB sequence prediction. We wanted to test the performance of the knowledge based protein local structure prediction algorithm (kPRED) in different prediction scenarios and datasets. Our intention to understand the performance of the kPRED algorithm, led us to develop a scoring function that could help us in auto-evaluation. Equipped with this scoring

function as a guideline we became curious to see if we could further improve the PB prediction accuracy. Hence our target became the development of a genetic algorithm based PB prediction improvement scheme.

Next, we were motivated to make PENTAdb, the database of pentapeptide conformations comprehensive and up-to-date. We wanted to quantitatively and qualitatively analyze the local structural diversity contained in it. Exploring the possible causes of structural plasticity/rigidity in certain pentapeptides became our next objective. We also explored the relationship between structural variability and predictability. Our objective was also to bring to the notice of the reader specific challenges we faced while parsing PDB files to incorporate into PENTAdb. Hence these examples are included in the appendix to Chapter 4.

Lastly, we wanted to explore the ability of the PB-ALIGN program (M Tyagi et al. 2008) to conduct large scale alignments to map the protein structure space. Our aim was to build a global network representing the structural similarities in an encyclopedic manner. We were interested in leveraging on speed and accuracy of the PB-ALIGN program to uncover novel relationships between proteins. Based on these alignment data we wanted to add functional annotations and hypothesize evolutionary links.

Chapter 2: Analysis of structural variations in proteins using a structural alphabet

2.1. Introduction

It is a well established fact that proteins exhibit a higher degree of structural conservation when compared to sequence conservation (Chothia and Lesk 1986; Bashford et. al. 1987; Bowie et al. 1990; Hark Gan et al. 2002). This observation stems from the fact that proteins require to maintain certain key residues in a specific orientation to fulfill their functional constraints (Nagl, Freeman, and Smith 1999; Lecomte, Vuletich, and Lesk 2005; Dessailly, Lensink, and Wodak 2007; Najmanovich, Kurbatova, and Thornton 2008; Saldaño et al. 2016). At the same time we also know that proteins are intrinsically dynamic entities. Intrinsic dynamic nature is probably encoded in the amino acid sequence itself as suggested by the high dynamics associated with certain intriscically disordered proteins with characteristic amino acid sequences with composition bias (Hansen et al. 2006). These conformational changes occur also due to intrinsic factors such as binding of a ligand, association with other proteins or nucleic acids and solvent content. Extrinsic factors like molecular packing in the crystal lattice, pH, temperature, pressure and crystallization conditions are known to superfluously affect the protein conformation in the process of determining the protein structure. There are several studies that document the dynamic nature of proteins (Gerstein, Lesk, and Chothia 1994; Gerstein and Krebs 1998; Qi, Lee, and Hayward 2005; Monzon et al. 2013; Varadi and Tompa 2015).

In this work we quantify the extent of changes in the backbone conformation of a protein by comparatively analyzing multiple crystal structures available for the same protein. The local structural changes observed over a large collection of these pairs of crystal

structures of same protein have been summarized in the form of a substitution matrix of short structural motifs called Protein Blocks (PBs). We call this PB substitution matrix the *altconfMAT*. Protein blocks are a 16 state classification of pentapeptides that can accurately abstract almost any part of a protein backbone structure (de Brevern, Etchebest, and Hazout 2000; de Brevern 2005).

Protein blocks have been shown to be very sensitive in identifying local structural variations in proteins. It is also effective in differentiating between regions of similarity and regions that show variations in protein local structure (Tyagi et al. 2008; Joseph, Srinivasan, and de Brevern 2011). Further we have also used PBs to monitor how the backbone structure changes during a molecular dynamics (MD) simulation for a collection of unrelated protein structures. The cumulation of the PB changes occurring during the MD simulation have been summarized in a PB substitution matrix called *dynamicsMAT*. Our methodology can be compared to using a flip-book to simulate motion. Where each of the static snapshot of protein structure in the PDB can be thought of as being analogous to the individual images in the flip-book.

Similar matrices have been constructed by observing the PB changes among homologous proteins (Tyagi et al. 2006; Gelly et al. 2011) and NMR ensembles (Mahajan et al. 2014).The first PB substitution matrix (Tyagi et al. 2006) was constructed by cumulating the PB changes observed in homologous proteins from the PALI (Phylogeny and Alignment of homologous protein structures) database (Balaji et al. 2001; Gowri et al. 2003). This substitution matrix was also use for the development of PB-ALIGN (Tyagi et al. 2006; Tyagi et al. 2008) a program based on dynamic programming for aligning PB sequences. Later both the alignment methodology and the underlying substitution matrix were improved upon

(Gelly et al. 2011). This new substitution matrix is henceforth referred to as the *homologueMAT*. Structure of a protein determined by NMR represents the structure as an ensemble rather than a single structure. The structural differences between the different conformers represented in the ensemble were summarized in a PB variation matrix (Mahajan et al. 2014). Henceforth this matrix is referred to as the *NMRMAT*. Further the correlation between the inherent structural variations of a protein and the structural variations among homologues were also explored in the study by Mahajan et. al. (Mahajan et al. 2014).

2.2. Material and Methods

2.2.1. Datasets

The section of the PDB that is considered redundant and excluded from several protein structural studies forms the starting point for our analysis. We started by clustering the entire collection of protein structures available in the PDB (Rose et al. 2015) at 100% sequence identity. This means that all the deposited structures of a given protein become members of a given cluster. This process was accomplished using the BLASTClust algorithm (Altschul et al. 1990). This resulted in 56,829 clusters with the largest cluster being the one with the capsid protein of HIV-1 with 2,580 chains. Many of these are viral capsids in multimeric forms and also there are several groups who have solved the structure of this protein. Out of these 56,829 clusters we filtered out clusters that are singletons because we needed a minimum of two structures in a cluster in order to study their structural variability. Further we restricted our study to structures solved by X-ray crystallography with a resolution of 2.5 Å or better. After this filtering process we were left with a total of 23,631 clusters.

There are 254,979 protein chains in these 23,631 clusters and from this dataset we make 7.8 million pairwise comparisons. For example in a cluster with 30 protein chains, we make 435 pairwise comparisons ((30*(30-1))/2).

2.2.2. Protein Blocks

Protein Blocks are a 16 letter structural alphabet denoted by the letters *a* to *p*. The orientation of the backbone atoms of any five consecutive residues in a protein structure can be defined by a series of 8 torsion angles (ϕ,ψ). PBs are defined by 16 sets of these 8 torsion angles. This standard library of torsion angles were determined by unsupervised clustering of the torsion angles from a non-redundant set of proteins (de Brevern, Etchebest, and Hazout 2000). Parsing a protein structure through a sliding window of 5 residues enables abstraction of the backbone 3D structure into 1D string of PBs. The PBE-T (Protein Block Transform) tool available on the PBE (Protein Blocks Expert) web-server (http://www.bo-protscience.fr/pbe/?page_id=10) can be used to encode protein structures into PB sequences (Tyagi et al. 2006). There are many advantages to approximating protein backbone conformation using PBs and a number of applications have been developed based on PBs (Joseph et al. 2010).

2.2.3. Molecular dynamics

A collection of 130 structures of unrelated proteins were subjected to molecular dynamics simulation using the GROMACS (Abraham et al. 2015) program for a duration of 10ns each. These proteins share less than 30% sequence identity with one another, are single chain proteins with no chain breaks and no ligands bound. Further structural criteria imposed were a X-ray crystallographic resolution better than 1.75 Å, R-factor <0.18 and free R-factor <0.24. The simulations were carried out using the Amber99SB-ILDN force field with TIP3P

water model at 300K temperature and 1 atm pressure. Global charge was neutralized to 0.15 mol/L with Na^+ and Cl^- ions.

2.2.4. Generating the variation matrices

Once we had this dataset of 254,979 protein chains, they were transformed into their respective PB sequences by comparing their torsion angles to the 16 standard PB torsion angles. This was done using an in-house perl script. All the PB sequences within a cluster were compared in an all-vs-all fashion. To ensure only structurally equivalent regions of both the proteins were being compared, the proteins with chain breaks were removed. Also the missing residues at the N-terminal and C-terminal were adjusted by introducing appropriate number of gaps. The PB variations from these 7.8 million pairwise comparisons were cumulated and used to populate a 16 X 16 variation matrix. Henceforth this matrix is referred to as the *altconfMAT* because it summarizes the alternate conformations of crystallographic structures from the PDB.

A similar 16 x 16 matrix was constructed that summaries the PB changes between snapshots during the MD simulation. Out of the 10ns simulation, the first 5ns were not taken into consideration for the analysis to allow for the system to stabilize. Coordinates of the protein structure were written out after every 250ps. Hence we get 19 sets of coordinates for each of the 130 simulations. These coordinates were converted into PB sequences exactly in the same manner as described before. The PB variations among these PB sequences were used to populate the second PB variation matrix, henceforth called the *dynamicsMAT*.

Since all the 16 PBs do not occur in the same frequency in protein structures, the matrices so obtained needs to be normalized. The normalization was carried out as per the procedure described by Johnson and Overington (Johnson and Overington 1993) (Eqn. 1). $S_{x,y}$

is the normalized score for variation of PB x to y, $N_{x,y}$ is the raw count of PB x varying to PB y and M is the total number of PBs *i.e.* 16:

$$S_{x,y} = \log_e \left[\frac{N_{x,y} \times \sum_{x=1}^{M} \sum_{y=1}^{M} N_{x,y}}{\sum_{x=1}^{M} N_{x,y} \times \sum_{y=1}^{M} N_{x,y}} \right]$$

In the *dynamicsMAT* dataset there were no PB $n \leftrightarrow$ PB d changes observed, hence in order to avoid undefined values during the normalization process, the value was changed from 0 to a nominal value of 0.0001. This value was chosen to be lower than the lowest observed value in the matrix.

2.3. Results and discussion

2.3.1. Inferences from qualitative analysis of structural plasticity

It has been previously established that the structural variations observed among models in an NMR ensemble collected over a large number of NMR ensembles correlate well with the structural differences observed in topologically equivalent regions of homologous proteins with a Pearson correlation coefficient of 0.94 (Mahajan et al. 2014). We were interested to see how the newly established PB variation matrices related to the previously established *NMRMAT* and the *homologueMAT* matrices. The correlations between each of these matrices are summarized, in Table 5. The best correlation is still between *NMRMAT* and *homologueMAT* matrices. All the other matrices show similar correlation patterns except for the *NMRMAT* and *dynamicsMAT* matrices, they have a Pearsons correlation coefficient of 0.87 only. Both of the currently established PB variation matrices- *altconfMAT* and *dynamicsMAT* are symmetric across the diagonal. The values in the diagonal correspond to retention of PBs and the off diagonal values correspond to PBs that get substituted.

Table 5: Correlation among the three PB variation matrices (NMRMAT, altconfMAT, dynamicsMAT) and the PB transition matrix (homologueMAT).

	homologueMAT	NMRMAT	altconfMAT	dynamicsMAT
homologueMAT	1			
NMRMAT	0.945	1		
altconfMAT	0.923	0.919	1	
dynamicsMAT	0.910	0.872	0.922	1

The torsion angles of the PBs are organized in such a way that specific PBs have a higher predisposition to be present in certain secondary structure elements (SSE). PBs *a-c* correspond to the N-terminal capping regions of β-strands, PB *d* corresponds to the central regions of β-strands, PBs *e* and *f* correspond to C-terminal caps of β-strands. The counterparts for the helices are PBs *k* and *l* for the N-terminal caps, PB *m* for the central region and PBs *n-p* for the C-terminal caps. PBs *g-j* represent the loop regions. Therefore, as expected, we see that the PBs representing the ends of the spectrum namely PBs *m* and *d* show the least structural variation. This is represented by rows (or columns) of PB m and PB d having the highest shades of red in the heat map (Figure 7). This pattern is consistent across all the four matrices. Apart from the PBs *m* and *d,* the PB *f* and PB *k* are highly represented in the top five invariant PBs. They represent the C-terminal of β-strands and N-terminal of helices respectively. The PBs *g, j, b, h, i* are highly represented in the top 5 varying PBs across all the four matrices. PB *b* is the exception in this list as it typically corresponds to the N-terminal cap of β̃-strands whereas the other PBs correspond to the loop regions.

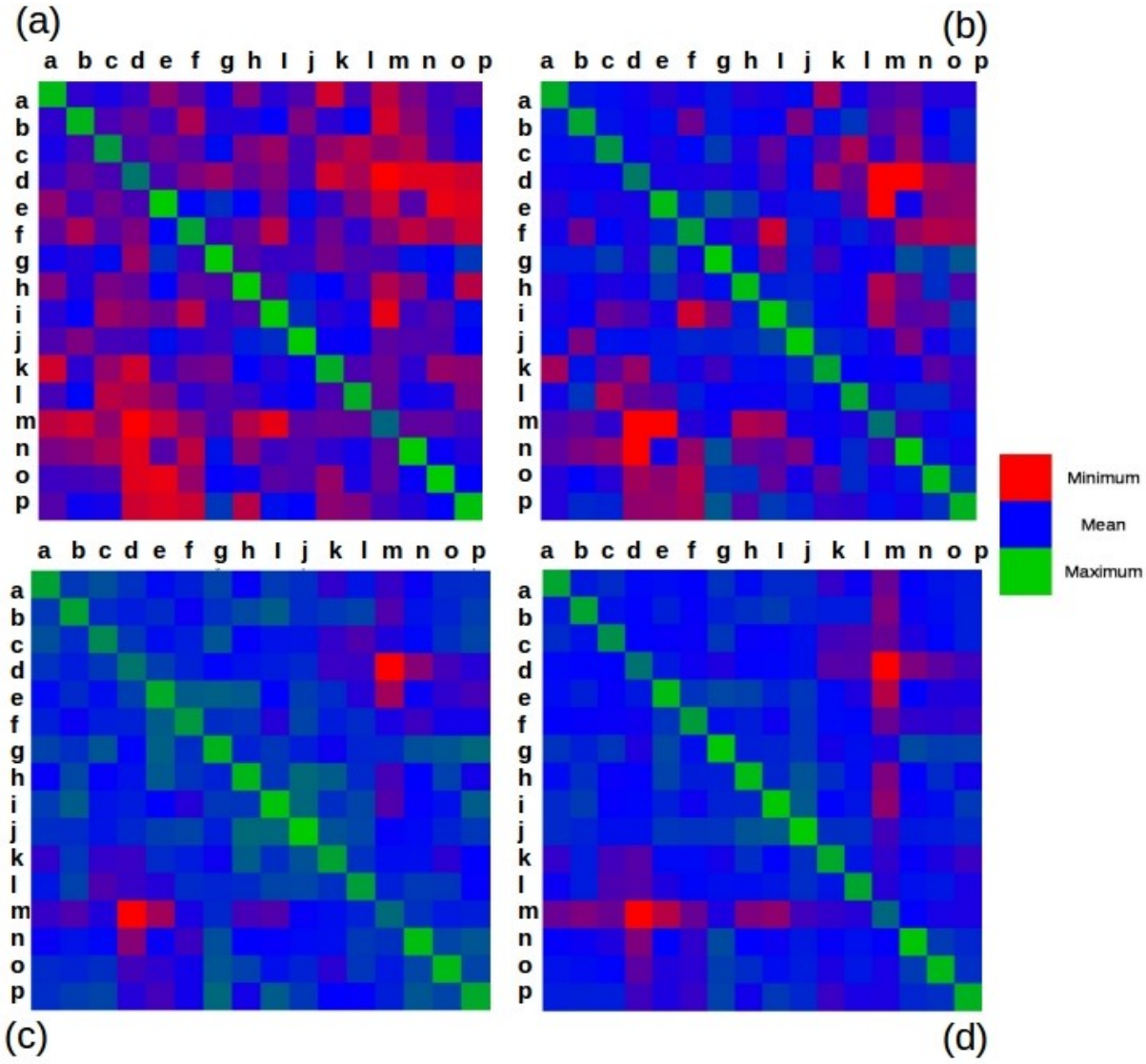


Figure 7. Tri colored heat maps of the four substitution matrices.
(a) *altconfMAT.* (b) *dynamicsMAT.* (c) *homologueMAT.* (d) *NMRMAT.*

2.3.2. Inferences from quantitative analysis of structural plasticity

The proportion of the sum of the raw counts of off-diagonal elements to the sum of all the raw counts of the matrix reflects the amount of structural variability captured by that particular matrix. For the *altconfMAT* this value is 4.88% and it is 7.04% for the *dynamicsMAT* matrix. The sum of the values of a row in the normalized matrix is a measure of how favorable or unfavorable the particular PB is to structural variations. The more negative this value is, the more resistant that particular PB is to structural variations (data not

shown). The degree to which PBs *d* and *m* are refractory to structural variation are very comparable in the *altconfMAT* and *dynamicsMAT.* Whereas this is not the case for the previously established matrices- the *NMRMAT* and the *homologueMAT.* In both these cases, PB *m* is significantly better conserved than PB *d.*

Once it is established that the central regions of the regular SSE are structurally invariant, we now focus on the loop regions and the interface of loop and SSE. In the *altconfMAT* matrix, this amounts to 78.2% of the structural variations out of which 64.3% are from the capping regions of the SSEs. In the case of *dynamicsMAT,* 65.9% of the variations are from the capping regions and this value increases to 79.4% if the loop regions are included. It is important to consider here that the distribution of PBs in the central regions of SSEs, capping regions of SSEs and the loop regions are not absolute. For example, although the PBs *g-j* represent the loop regions we can occasionally find other PBs in the loop regions as well.

The correlation coefficients represented in Table 5 are overall values for all the 16 PBs. They can hide differences in correlation at the individual PB level. Hence we decided to calculate the individual PB wise correlation between these four matrices. Figure 8 is a plot of these PB-wise correlations. In each case the red line corresponds to the overall correlation, the points above contribute positively to the correlation and the ones below negatively. The *NMRMAT vs dynamicsMAT* PB wise correlation shows the maximum dispersion.

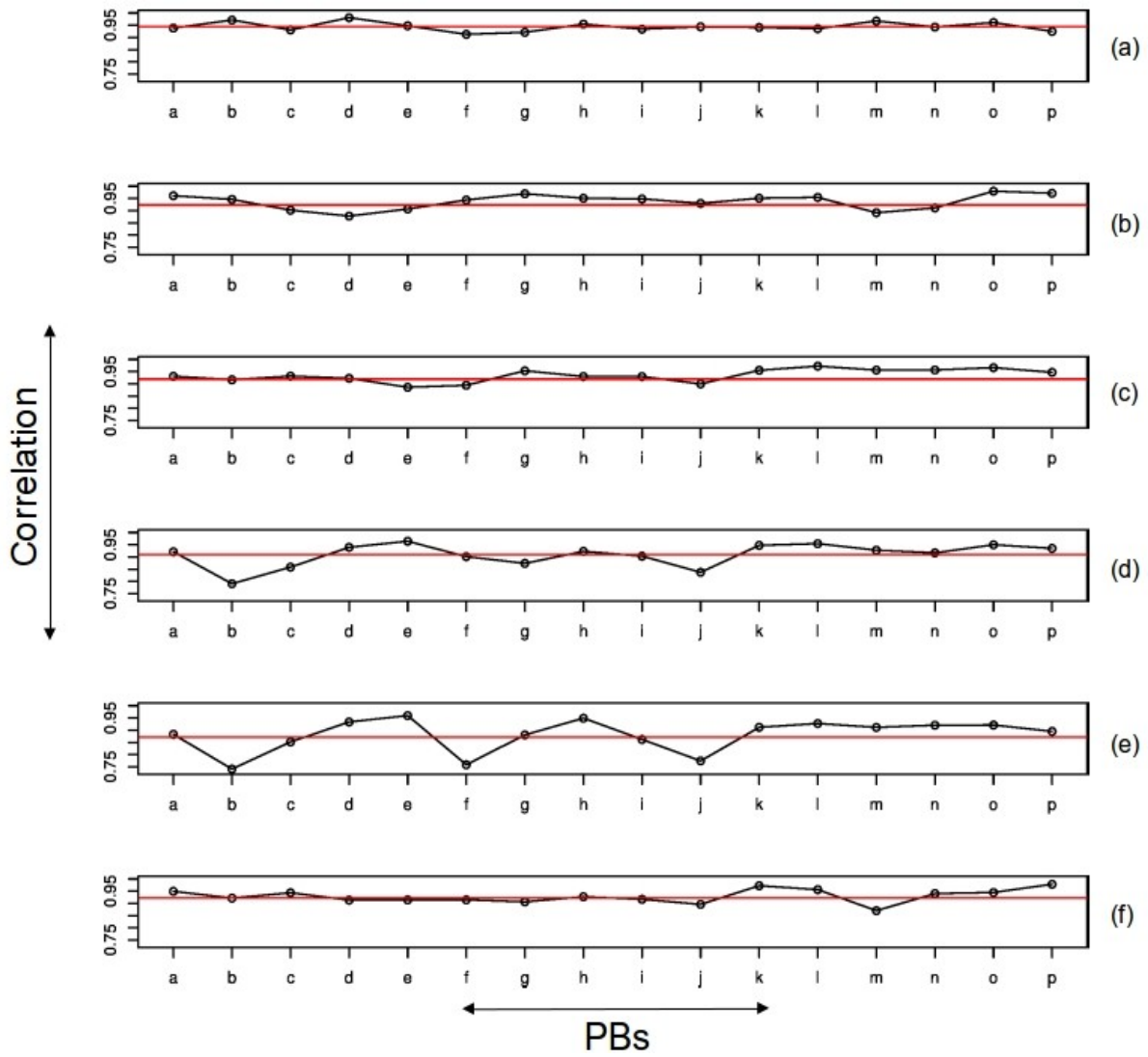


Figure 8. Individual PB wise correlation between the four matrices.

(a)*homologueMAT vs homologueMAT (b)homologueMAT vs altconfMAT, (c)NMRMAT vs altconfMAT, (d)homologueMAT vs dynamicsMAT, (e)NMRMAT vs dynamicsMAT, (f)altconfMAT vs dynamicsMAT.* The global correlation values in Table 1 are represented by the horizontal red line. Points above this line contribute positively to the correlation and the ones below negatively.

2.3.3. Case studies of structural variations for alternate conformations of the same protein

The cluster we get from BLASTClust filtered at 100% sequence identity for human tyrosine kinase consists of three structures (PDBID: 1Y57, 1FMK and 2SRC). Their superimposition featured in Figure 9 clearly shows that there are large structural variations that are induced upon binding to MPZ ligand.

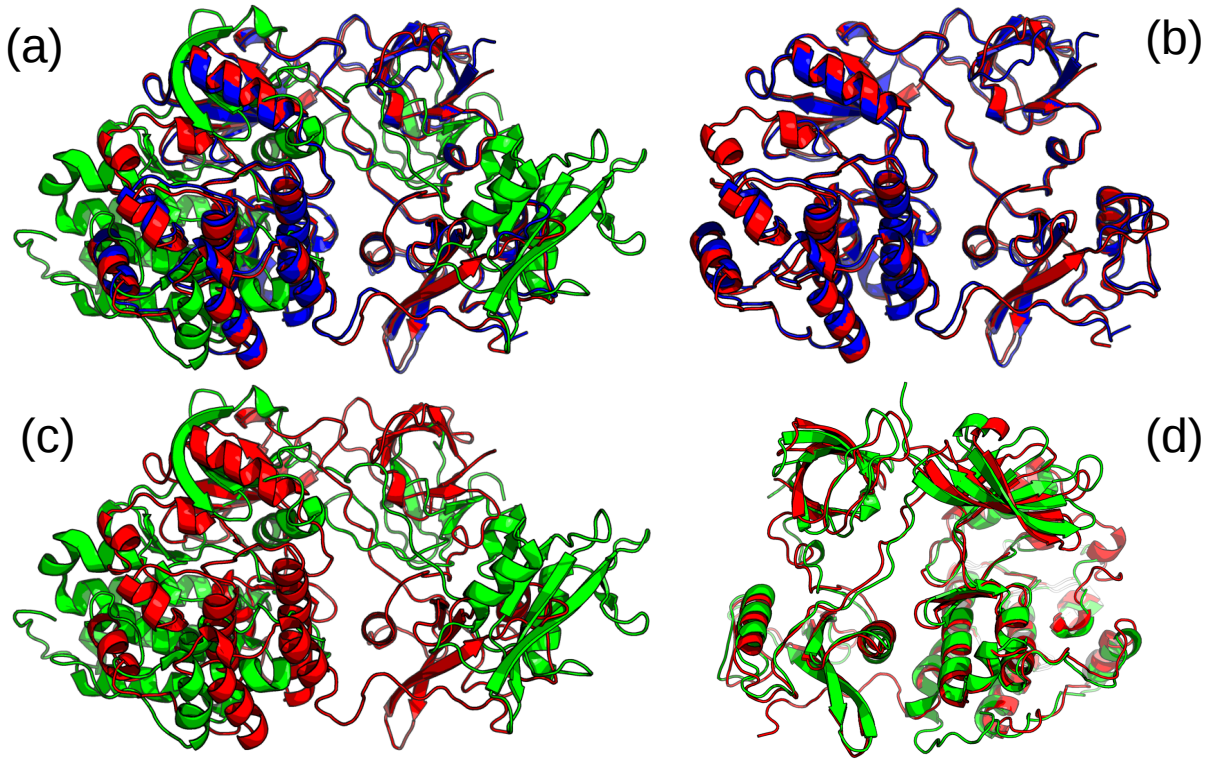


Figure 9. Comparison between 3 structures of human tyrosine kinase C-SRC by rigid body & flexible superimposition.
In blue is the unbound form (PDBID 1FMK), in red is the conformation (PDBID 2SRC) when bound to phosphoaminophosphonic acid adenylate ester (ANP) (Xu et al. 1999) and in green is bound form to MPZ ligand (PDBID 1Y57) (Cowan-Jacob et al. 2005). Ligands are not shown for clarity. (a) Shown is the rigid body superimposition of the 3 structures. (b) Shown is the rigid body pairwise superimposition between 2SRC and 1FMK with a calculated RMSD of 0.806 Å. (c) Shown is the rigid body pairwise superimposition between 2SRC and 1Y57 with a calculated RMSD of 21.03 Å. (d) Shown is the pairwise flexible superimposition between 2SRC and 1Y57 after introducing a twist at position 259 using FATCAT that resulted into a calculated RMSD of 2.72 Å.

These are further compared using protein blocks below. Figure 10 shows the PB alignment between the structures of human tyrosine kinase C-SRC bound to ANP ligand (2SRC) and to MPZ ligand (1Y57). The regions marked in red are the regions that show structural changes based on PBs. A total of 75 out of the 452 PBs variations are observed. Rigid body superposition of the two structures using the ProFit software[1] results in a high RMSD of 21.3 Å (Figure 9c). When the same two structures are aligned flexibly using the

1 - Martin, A.C.R., http://www.bioinf.org.uk/software/profit/

FATCAT program (Ye and Godzik 2004) the RMSD reduces to 2.7 Å (Figure 9d). This was achieved by the FATCAT algorithm (Ye and Godzik 2004) by introducing a twist located at residue 259. Indeed, in this region, PB sequence between the two structures showed a major difference with a twist at the level of the residue leucine 255: *BEHIA* is aligned with *ADDFK* (Figure 10). When the same structure is compared with the ligand unbound form (PDB ID 1FMK) 68 PBs show variation and most importantly the region of the twist shows only one PB change as opposed to 5 PB changes in the previous case(Figure 11).

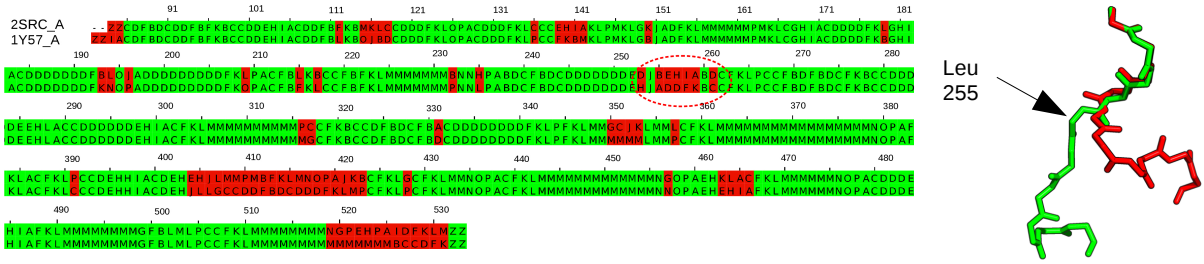


Figure 10. PB alignment between two structures of human tyrosine kinase C-SRC bound to ANP ligand (2SRC) and MPZ ligand (1Y57) captured by *altconf*MAT matrix.
Shown is alignment of chain A for both structures. The region responsible for the high RMSD value calculated after rigid superimposition is circled in red. This region display drastic PB variations that reflects a major twist in the backbone at the level of leucine 255.

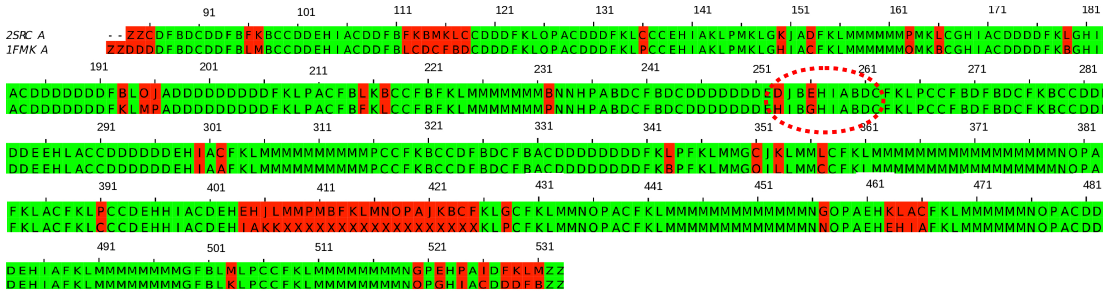


Figure 11. PB alignment between two structures of human tyrosine kinase C-SRC in the unbound form (1FMK) and bound with ANP ligand (2SRC).
This is an example of structural variation captured by *altconf*MAT matrix. Shown is alignment of chain A for both structures. The region featured in Figure 10 is again circled here. This region display minor PB variations that reflects very little change in the backbone conformation in this region.

Figure 12 shows the structure of a mutant amidase from *Geobacillus pallidus* (PDBID 4LF0) at the end of 5 and 10 ns of a molecular dynamics simulation. The regions marked in red are the regions that show structural changes between these two states. In this case 25 out of 340 PBs showed variations. In both the cases, the PBs corresponding to the capping regions of secondary structural elements and loop regions are highly represented. The logical succession to this work would be to characterize the structural changes during a molecular dynamics simulation at different time steps.

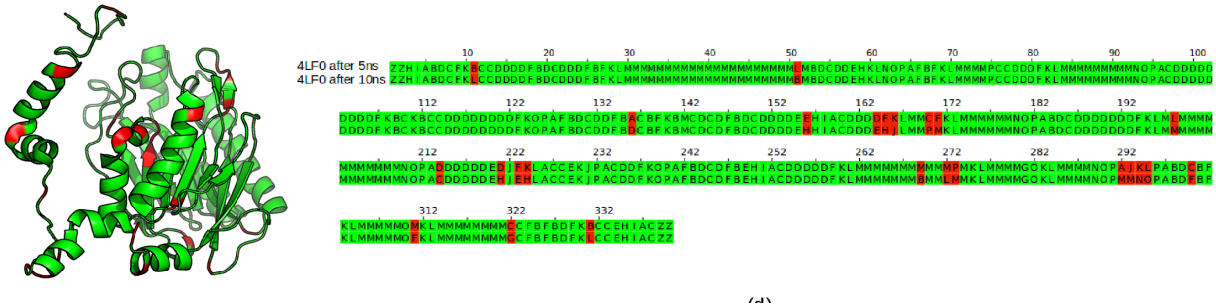


Figure 12. Example of the structural variation captured by *dynamicsMAT* matrix after 5 and 10 ns MD simulations.
Shown is the PB variations observed in an amidase (4LF0) after 5 and 10 ns MD simulations. On the left is the cartoon representation of 4LF0 and on the right are its corresponding PB sequences after 5 and 10 ns MD simulation. The regions marked in red represent structurally varying regions between the two structures extracted at these simulation times.

2.3.4. Case studies of structural variations between two homologous proteins having alternate conformations

As an extension to the previous case study (see section 2.3.3 above), we further used PB align to compare two homologous proteins each having multiple alternate conformations. The main results are featured in the form of network (Figure 13). The cluster in red corresponds to alternate structures of human sulfotransferase and the ones in green are their homologues from mouse. All the members of a single group share 100% sequence identity but inter group members share 59% sequence identity. The edges in red and green represent intra

group alignments and the blue edges are inter group alignments. The alignments were carried out using the PB-ALIGN program (Tyagi et al. 2008) and the alignment scores were normalized by length. In all cases the edge length corresponds to the normalized alignment scores, higher alignment scores correspond to shorter edges. Global distribution of the normalized alignment scores are featured in (Figure 14).

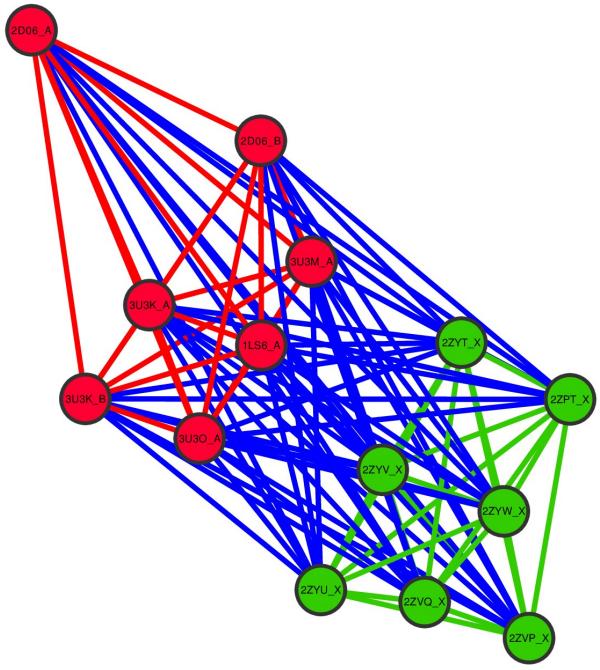


Figure 13. Network showing structural (dis)similarity between mouse and human sulfotransferases each having multiple alternate conformations.

The red nodes are the group of human sulfotransferase structures while the green nodes represent the group of mouse structures. Intragroup alignments are coloured in red and green respectively. Intergroup alignments are coloured in blue. Distance between nodes are inversely proportional to the normalized alignment scores. Intragroup distances for the mouse sulfotransferases are lower than any intergroup distance. Some of the intragroup distances for the human sulfotransferases are greater than some its intergroup distances.

The closest inter group structures (1LS6_A and 2ZYV_X) have a normalized alignment score of 1.829 and the farthest pair (2D06_A and 2ZVP_X) have 1.652. Their corresponding all atom superposition RMSDs are 0.401 Å and 0.459 Å respectively. The difference in normalized alignment score (0.177) and RMSD (0.058 Å) seem to be minor but

it is noteworthy that this variability is seen in homologues sharing a relatively high sequence identity of almost 60%. We can expect this variability effect to be more pronounced in homologues sharing sequence identity in the range of 25%-30% and also when the size of the groups are larger. **More importantly, as shown in Figure 13, taking one conformation of the human sulfotransferase as a reference (1LS6_A), we see a situation where mouse sulfotransferase (2ZYV_X) conformation is strikingly more structurally similar to it than another conformation of the same human sulfotransferase (2D06_A).**

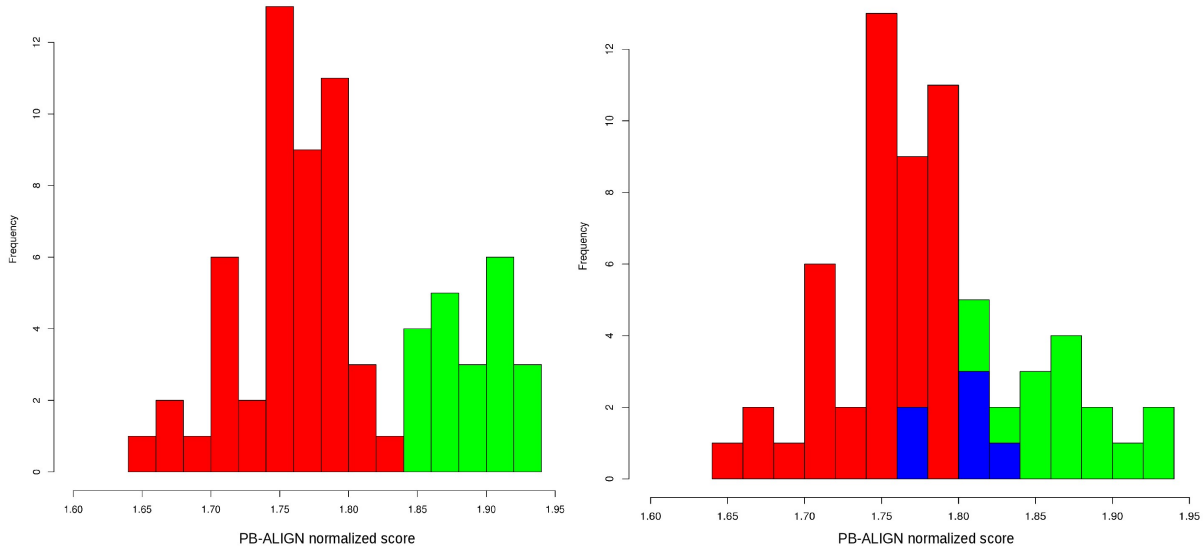


Figure 14. Distribution of PB-ALIGN normalized alignment scores.
On the left is the distribution of the normalized alignment scores for the mouse sulfotransferase structures group. On the right is the distribution of the normalized alignment scores for the human sulfotransferase structures group. In red are the inter group normalized alignment scores. In green are intra group alignment scores. Intragroup scores for the mouse sulfotransferases are higher than any of its intergroup scores. Intragroup scores for the human sulfotransferases show some overlap with some of its intergroup scores. This overlap is indicated in blue in the histogram on the left.

2.3.5. Further analyses of molecular dynamics trajectories using PBs

Two additional analyses were performed on the same 130 proteins that were subjected to molecular dynamic simulations as described above. The first consisted in the global analysis of PB transitions that took place during the 10 ns simulation between two

consecutive time steps of 2 ps. All these PB transitions were cumulated into a PB transition matrix. Such matrices were generated for each of the 130 proteins. A sample of this matrix for an endoglucanase (PDB ID 1A3H_A) is shown in Table 6. For a time step of 2 ps we see an almost symmetric matrix, indicating several flip-flop motions. We also see that PBs *m*↔*k, m*↔*o, d*↔*c, d*↔*f* and *l*↔*b* have high number of transitions (>2000 transitions). This observation is expected due to the high proportion of occurrence of PBs *m* and *d* but still *m*↔*n, d*↔*e* transitions are still low (<200 transitions). It is also surprising to see high *l*↔*b* transitions.

Table 6: Raw counts of PB transitions observed during a 10 ns MD simulation of an endoglucanase (PDBID 1A3H).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	0	175	900	358	0	16	740	1	547	595	6	871	10	87	192	165
B	178	0	810	989	1034	0	94	0	73	0	989	2516	75	0	22	1437
C	897	820	0	2742	51	853	1443	2	27	0	28	0	1107	0	21	877
D	356	991	2742	0	71	2382	685	501	0	0	0	4	0	0	0	0
E	0	1034	45	71	0	0	1025	1486	0	0	732	0	0	8	0	0
F	13	0	857	2386	0	0	0	5	0	24	179	1723	849	0	0	1
G	739	89	1445	685	1023	0	0	32	0	0	0	98	242	77	29	385
H	1	0	1	501	1482	5	35	0	339	271	0	1	0	0	367	1
I	533	72	32	0	0	0	0	338	0	100	0	243	0	0	5	526
J	594	0	0	0	0	25	0	273	99	0	3	0	1	0	703	12
K	7	987	30	0	730	178	0	1	1	2	0	109	2994	99	0	1
L	871	2520	0	3	1	1722	96	1	225	0	110	0	1976	0	233	88
M	8	76	1099	0	0	855	236	0	1	1	2996	1974	0	169	2835	1292
N	85	0	0	0	9	0	71	0	0	1	96	0	183	0	6	76
O	198	14	28	0	0	1	23	366	5	704	0	234	2833	6	0	566
P	182	1440	871	0	0	0	398	0	533	13	1	71	1268	81	567	0

Another observation we can make is that PB *d* corresponding to the central region of the β-strands does not show any transitions towards the other end of the spectrum of PBs (PBs *i* to *p*) but PB *m* corresponding to the central region of α-helices has significant number of transitions to PBs *c, f* and *g* respectively. Nonetheless these observations are highly dependent on the PB distribution in the particular protein and also the time step chosen for the analysis. For example, an all-β protein could have considerable number of transitions that have not been observed in the example here (1A3H).

Correspondingly we were interested to see the average time (in ps) of these transitions. This is represented by the matrix in Table 7. Interestingly, the transition times are not symmetrical. For example, four *d* to *l* events were observed during the simulation and this occurred on average every 434 ps. On the other hand, two *l* to *d* events occurred during the same period with an average PB transition time of only 2 ps. Short delays in PB transitions could probably account for strained local conformations coming back to a more energetically favoured state.

Table 7: Average time between PB transitions in pico seconds observed during the 10 ns MD simulation of an endoglucanase (PDBID 1A3H).

A zero value is indicative of absence of events.

PB	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	0	8.07	24.48	14.85	0	485.88	6.4	14	5	8.05	6.67	20.08	184.2	6.02	7.54	3.66
B	15.06	0	14.11	17.94	6.21	0	2.81	0	2.05	0	7.29	11.66	2.37	0	2	4.41
C	37.68	12.6	0	14.64	2.16	2.65	26.84	2	2.15	0	2.07	0	2.49	0	39.05	10.16
D	60.21	26.32	10.59	0	340.56	14.62	11.66	33.63	0	0	0	434	0	0	0	0
E	0	3.46	16.22	2.11	0	0	9.61	11.64	0	0	2.62	0	0	3	0	0
F	4	0	47.2	8.51	0	0	0	18.8	0	233.17	160.37	13.35	13.63	0	0	528
G	3.31	52.67	6.99	2.67	3.67	0	0	3.06	0	0	0	75.53	2.46	2.31	3.24	2.5
H	2	0	2	2.59	3.51	2.8	2.06	0	30.6	2.27	0	2	0	0	25.09	2
I	2.65	37.69	31	0	0	0	0	27.5	0	2.1	0	2.8	0	0	30	27.99
J	5.28	0	0	0	0	359.6	0	7.81	6.46	0	3.33	0	4	0	6.9	5.5
K	2	9.93	6	0	33.15	3.7	0	3786	2	613	0	191.93	10.41	30.91	0	472
L	3.15	10.43	0	2	2	3.48	2.19	2	2.88	0	63.89	0	23.19	0	2.27	2.95
M	2	2.79	23.43	0	0	6.46	14.41	0	2	2	9.36	17.74	0	96.07	11.62	4.3
N	3.08	0	0	0	2.89	0	2.65	0	0	2	2.31	0	33.65	0	4	2.11
O	2.6	2.29	63.07	0	0	2	7.3	5.96	2	2.77	0	76.22	3.34	2.33	0	3.43
P	2.2	6.77	6.33	0	0	0	27.7	0	2.68	2.15	2	2.23	5.26	9.26	4.49	0

Apart from the classical tools to study MD trajectories like RMSD and RMSF fluctuations during the length of the simulation (Figure 15), we propose to view the simulation in the light of PBs as valuable alternate visual representations of the dynamics. Figures 13, 14 and 15 are examples of such PB-based representations. Figure 16 is the PB wise variability during the simulation calculated as per the Structural Variability Index (SVI)

described in Chapter 4 of this thesis. While SVI is a quantitative measure indicating the variability of every pentapeptide found along the backbone, the qualitative nature of the changes can be seen in Figure 17 and 18. In Figure 17, the 16 PBs are color coded and the PB at each residue position during the length of the molecular dynamics is plotted. The PB changes can be identified by color changes. For example the change in PB *l* to PB *k* at position 204 can be seen as a change from blue to green at approximately 7400 ps. Specific events can be "zoomed" after having this global view by re-plotting for specific range of residues and between certain time intervals.

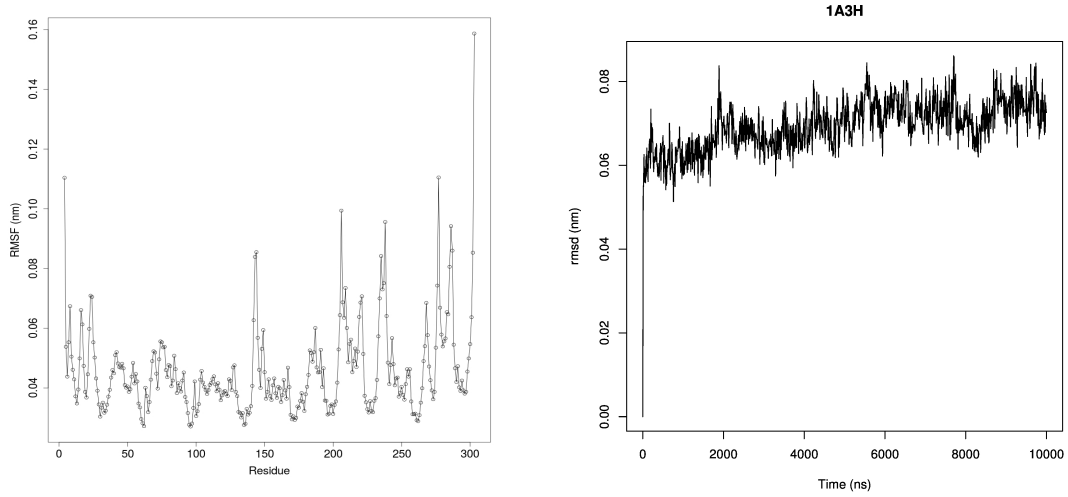


Figure 15. Root mean squared fluctuation (RMSF) and root mean squared deviation (RMSD) calculated for backbone atoms of PDB ID 1A3H during a 10 ns MD simulation.

Weblogo is an application for generating sequence logos (Crooks et al. 2004). Sequence logos were originally developed to represent multiple alignments of nucleic acid and protein sequences (Schneider and Stephens 1990). Here we have adapted the stand alone version of the weblogo program for the purpose of representing PB changes during a MD simulation. The alternate PBs observed at a particular position are represented in a vertical column. The height of the PB letter is proportional to the occurance frequency of that particular PB. Correspondingly we can see PBs *l* and *k* at position 204.

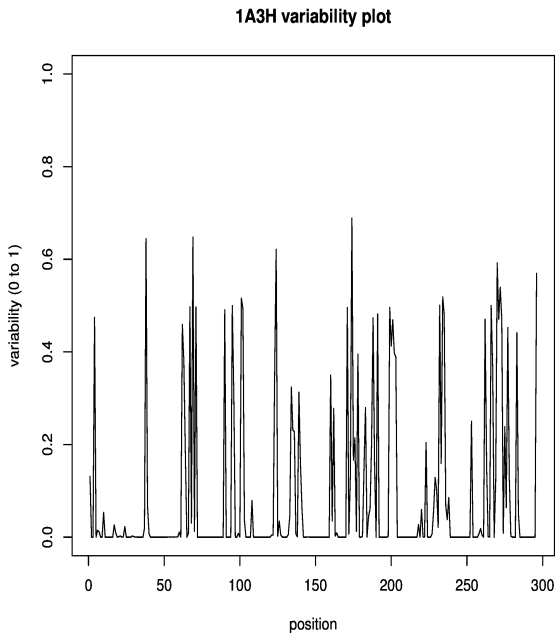


Figure 16. Structural variability of pentapeptides along the backbone of 1A3H during a 10 ns MD simulation.

This is based on the calculation of a structural variability index (SVI) as defined in chapter 4. A value of zero indicates no variability and an index of 1 is indicative of high variability.

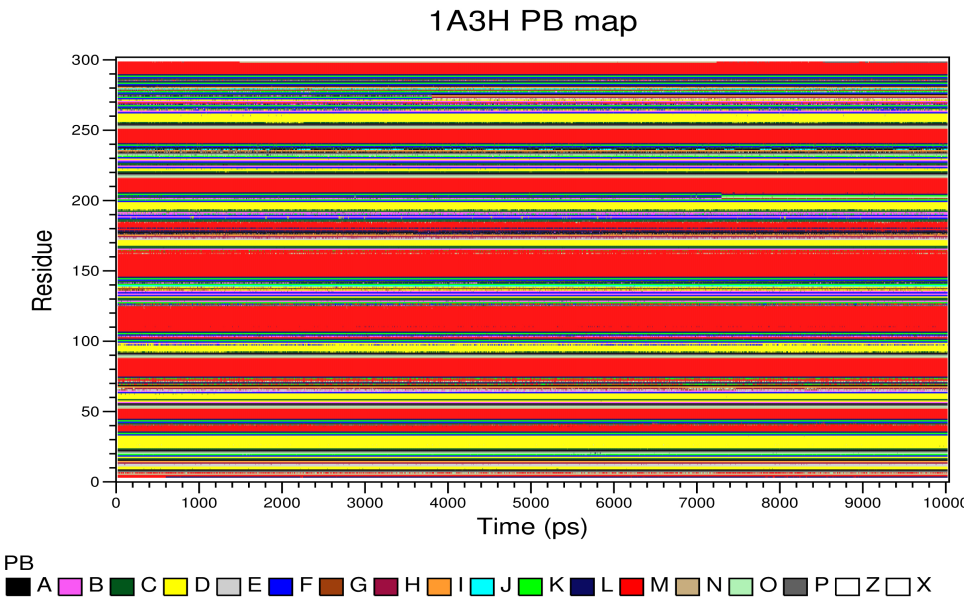


Figure 17. PB map for 1A3H during a MD simulation.

Shown is the variation of PB state for each position along a 10 ns simulation.

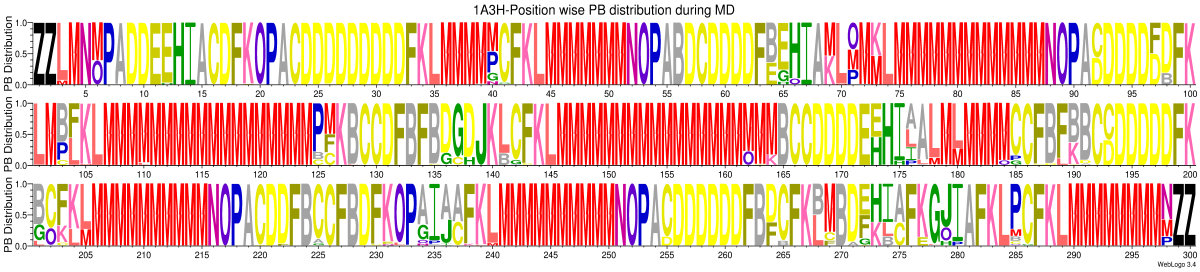


Figure 18. PB logos showing variation of PBs along the backbone of 1A3H during a 10 ns MD simulation.

Chapter 3: Knowledge based prediction of protein backbone conformation using a structural alphabet

The first part of this chapter (sections 3.1 to 3.8) corresponds to a manuscript that has been submitted to the PLOS ONE journal and is hence formatted as per the journal requirements including the citation styling. It has also been deposited in the preprint server bioRxiv. The citation of our preprint manuscript is:

Vetrivel I, Mahajan S, Tyagi M, Hoffmann L, Sanejouand Y-H, Srinivasan N, de Brevern A.G, Cadet F, Offmann B. Knowledge based prediction of protein backbone conformation using a structural alphabet. BioRxiv 127423 doi:10.1101/238423

The second part of the chapter (sections 3.9 to 3.12) corresponds to additional analysis related to prediction of protein local structures in terms of PBs. The list of references in section 3.13 is specific for this chapter.

3.1. Abstract

Libraries of structural prototypes that abstract protein local structures are known as structural alphabets and have proven to be very useful in various aspects of protein structure analyses and predictions. One such library, Protein Blocks, is composed of 16 standard 5-residues long structural prototypes. This form of analyzing proteins involves drafting its structure as a string of Protein Blocks. Predicting the local structure of a protein in terms of protein blocks is the general objective of this work. A new approach, PB-kPRED is proposed towards this aim. It involves (i) organizing the structural knowledge in the form of a database of pentapeptide fragments extracted from all protein structures in the PDB and (ii) applying a

knowledge-based algorithm that does not rely on any secondary structure predictions and/or sequence alignment profiles, to scan this database and predict most probable backbone conformations for the protein local structures. Though PB-kPRED uses the structural information from homologues in preference, if available.

The predictions were evaluated rigorously on 15,544 query proteins representing a non-redundant subset of the PDB filtered at 30% sequence identity cut-off. We have shown that the kPRED method was able to achieve mean accuracies ranging from 40.8% to 66.3% depending on the availability of homologues. The impact of the different strategies for scanning the database on the prediction was evaluated and is discussed. Our results highlights the usefulness of the method in the context of proteins without any known structural homologues. A scoring function that gives a good estimate of the accuracy of prediction was further developed. This score estimates very well the accuracy of the algorithm (R^2 of 0.82). An online version of the tool is provided freely for non-commercial usage at http://www.bo-protscience.fr/kpred/.

3.2. Author summary

Proteins mediate vital bio-processes in all living organisms. Understanding their structure at the atomistic level is of major interest for understanding the functions they play. To date the proportion of the proteins for which we know the structure at the atomistic resolution is still a small fraction of all the proteins known to exist. Nonetheless predictions of protein structures have played an important role in their functional annotation and classification. These predictions are frequently a secondary structure description of the components that build up protein structures, namely helices, strands, loops and turns. In this current work we describe and also rigorously test a more informative and hence potentially more useful 16 state

prediction of protein structure. The method that we propose is a knowledge based approach that relies only on the structural knowledge available in the protein structure repository. We have tested this method on a collection of more than 15000 unrelated proteins and are able to achieve an average accuracy of 66.3%. Different test cases involved predicting when varying levels of pertinent data are permitted in the structural knowledge base. Further we have developed and benchmarked a scoring function that gives us a good estimate of the accuracy of prediction.

3.3. Introduction

Knowledge of protein structure considerably helps towards understanding protein function. The Protein Data Bank (PDB) that serves as the central repository of knowledge for the protein structural biology community contains more than 125,000 protein structures and its growth has been considerable in the past decade [1]. This number is however still far below the ~70 million protein sequences referenced in UniProt database [2]. Hence it is at stake to find methods to bridge this considerable gap. Computational methods for predicting protein secondary and tertiary structure have persistently tried to fill it. In this paper, we explore the ability of a structural alphabet based prediction method to fulfill in part this role.

Since the seminal works by Kabsch and Sander in 1984 [3], one of the most popular and rewarding computational method to predict and analyze protein structures is by breaking them down to their constituent parts in the so-called fragment-based approach. Multiple fragment libraries have been developed so far and they differ in the number of fragments, the length of the fragments, the methods used for clustering and the criteria used for clustering. The first fragment library was developed by Unger and co-workers [4]. There are reviews that give a good overview of the different fragment libraries developed since then [5,6]. Also referred to as structural alphabets (SAs), these have shed some light on the sub-secondary structure level

intricacies in proteins [7]. By identifying redundant structural fragments found in proteins, structural alphabets help in abstracting protein structures accurately. Such collections of fragments have been used in information theory based methods that attempt to predict secondary structure (see for example [8] and to reconstitute protein structures [9-11].

In that respect, a SA called *protein blocks* (PBs) was developed for the purpose of describing and predicting the local backbone structure of proteins [12,13]. This SA accounts for all local backbone conformations in protein structures available in the Protein Data Bank (PDB). Since then, PBs have been used in various applications [12]: for structural motif identification [14–16], structural alignments [17,18] and fold recognition [19,20]. There have also been various efforts to use PBs to predict protein local structure. These approaches are based on the Bayes theorem [10,13] support vector machines [21–23] and neural networks [24]. Some of these methods have used prior predictions of classical three state secondary structures (svmPRAT [23] uses YASSPP [25], SVM-PB-Pred [22] uses GOR [26] and Etchebest et. al. [27] explored the utility of PSI-PRED [28]) and sequence alignment profiles like position specific scoring matrices (PSSMs) are used by LOCUSTRA [21], SVM-PB-Pred [22] and Dong and coworkers methodology [24]. The currently available web-based tools that can predict local structure in terms of protein blocks are LocPred [29] and SVM-PB-Pred [22]. The former implements a Bayesian methodology and the latter is SVM-based.

In this work we describe PB-kPRED, a fragment and knowledge-based approach to predict local backbone structure of proteins in terms of protein blocks and a web-based tool that implements the method. In essence, it takes no other inputs than the amino acid sequence of a query and interrogates a database of pentapeptides extracted from protein structures, relying on evolutionary information only when available, hence useful in the context of proteins without any known structural homologues. It returns the predicted local structures of the

polypeptide chain in the form of a sequence of protein blocks. Very importantly, PB-kPRED also implements a scoring function that efficiently auto-evaluates the quality of the prediction.

3.4. Results

3.4.1. PENTAdb, a database of pentapeptides from protein structures

A total of 68.84 million pentapeptides obtained from the 0.26 million protein chains and their corresponding local structure represented as one of the 16 PBs were obtained and stored in a dedicated database named PENTAdb. Out of the total number of theoretically possible 3.2 million (20^5) pentapeptides, 2.26 million (70.9%) are found in PENTAdb. These 2.26 million pentapeptides occur in different counts amounting to a total of 68.84 million pentapeptides. PENTAdb had 8 different threshold of sequence identity, and can be directly used by PB-kPRED (see Table 8). A 32-fold decrease (from 68.62 to 5.13 million) in the number of pentapeptides in PENTAdb is observed when PDB chains not sharing more than 30% sequence identity are considered. Nonetheless, the decrease in the number of unique pentapeptides present in PENTAdb is only of 1.3 fold at this threshold.

3.4.2. Not all possible tri-PB combinations are observed in known protein structures

Out of all the theoretically possible 4,096 (16^3) tri-PBs, a total of 1,375 (i.e 33.5%) were never observed in a non-redundant subset of PDB filtered at 30% sequence identity, subsequently termed as PDB30 dataset (see Methods section). Likewise, out of all the 1.04 million (16^5) theoretically possible penta-PB motifs, only 40,130 (3.8%) were observed in the PDB30 dataset. These results are indicative of the possibility that many combinations of three or five consecutive PBs are stereochemically unfavorable. The distributions of penta-PB motifs at other sequence identity cut-offs i.e. 40%, 50%, 70%, 90%, 95%, 100% and the entire PDB were also computed (see supporting information S1 Table). Towards higher

sequence identity cut-offs, a continuous increase in the penta-PB coverage is observed, but it comes at the price of the addition of redundant data. Still the entire PDB covered less than 10% of the total penta-PB space. The penta-PB frequency table derived from the PDB30 dataset was used to develop a scoring function (see Methods section), further termed as accuracy score, even if it contained only 40,130 penta-PB motifs. This is a small fraction but was sufficient to efficiently score PB sequences (see below).

Table 8: Content of the different subsets of PDB in terms of pentapeptides accessible to PB-kPRED.

Seq. identity cut-off values (%)	Total number of chains	Total number of unique pentapeptides	Total number of pentapeptides
<30	24,564	1,742,890	5,126,423
<40	28,590	1,881,813	6,153,846
<50	32,588	1,985,203	7,074,571
<70	37,741	2,095,663	8,336,277
<90	42,594	2,148,064	9,351,888
<95	44,714	2,157,239	9,768,748
<100	64,129	2,189,924	14,791,285
Full PDB	**274,920**	**2,268,307**	**68,621,454**

3.4.3. Completeness of PENTAdb for knowledge-based prediction

A quantitative assessment of how often the correct PB can be found in the list of all possible PBs reported for every query protein was performed. This assessment gives the highest accuracy attainable for a query protein using the proposed knowledge-based approach. To this end, for every query protein sequence, different portions of the pentapeptide database were made accessible to the prediction algorithm. This was possible thanks to the hierarchical clustering at different sequence identity levels by the BLASTClust algorithm [30]. PDB30 dataset was used to benchmark the approach. For each of the 15,544 unrelated query protein sequences of this dataset, only pentapeptides coming from a subset of the PDB that shared

sequence identities below an indicated cut-off values (from 30% to 100%) and excluding the query itself were used by PB-kPRED (see Table 8 for the size of the database for each subset). The results of the predictions are detailed in Table 9. At 30% sequence identity cut-off, the correct PB was found in the list of all possible PBs for 71.4% of the case and the success rate increased to 77.3% when only "homologues" sharing 100% sequence identity to the queries were filtered out. When full PDB was used (but excluding the query) as a database, the percentage times the correct PB is in the list of all possible PBs topped to 99.93% (see supporting information S2 for the PB-wise breakdown of these values).

Table 9: Assessment of the richness of PENTAdb towards knowledge-based prediction of protein backbone in terms of protein blocks.

Shown is the percentage of pentapeptide queries for which the correct local conformation was found in the list of all possible PBs reported by the PB-kPRED algorithm after querying PENTAdb. A total number of 15,544 query proteins not sharing more that 30% sequence identity (PDB30 dataset) was used in this assessment.

Sequence identity cut-off values (%)	Correct PB found in the list of all possible PBs (%)
<30	71.4%
<40	71.5%
<50	71.6%
<70	71.9%
<90	72.5%
<95	72.8%
<100	77.3%
FULL PDB (excluding query)	99.93%

3.4.4. Prediction accuracies

Two methods, further termed as *majority rule method* and *hybrid method*, were explored to predict the optimal PB sequence within the list of all the possible PBs obtained after querying

the database (see Methods section for details). The prediction accuracy is calculated as the percentage of PBs correctly predicted for each of the queries. This value is then averaged across the entire prediction dataset to obtain the average prediction accuracy. The average prediction accuracies for the PDB30 query proteins using the *majority rule method* and the *hybrid method* using the classic scheme for querying the database are given in Table 10. When homologues sharing ≥30% sequence identity with each of the queries were removed from the database, PB-kPRED performed with an average accuracy of 39.2% and 40.8% for the *majority rule method* and *hybrid method* respectively (Table 10). Surprisingly, the effect of enlarging the database to include closer homologues sharing <95% sequence identity with the queries improved only marginally the prediction accuracies reaching on average 40.4% and 42.4% for *majority rule method* and *hybrid method* respectively. Accuracy topped to 58.0% and 54.6% respectively when full PDB (excluding the query itself) was used as database for prediction. This overall gain in accuracy is due to an incremental increase of accuracy across all the 16 PBs.

Table 10: Evaluation of performance of PB-kPRED knowledge-based approach to predict local conformations of protein backbone in terms of protein blocks.

Shown are the average prediction accuracies for the PDB30 dataset using both the *majority rule method* and *hybrid method*. For each of the 15,544 query protein sequences, the portion of PENTAdb accessible for prediction was dynamically determined using MySQL queries: only pentapeptides coming from protein chains in PDB that shared sequence identities below the indicated cut-off values were accessible to PB-kPRED for prediction of local structures in terms of PBs.

Experiment number	Sequence identity cut-off values (%)	Average prediction accuracies by Majority rule method (%)	Average prediction accuracies by Hybrid method without noise filtering (%)
A1	FULL PDB (excluding query)	58.0%±12.9	54.6%±22.0
A2	<100	44.0%±14.7	48.0%±18.7
A3	<95	40.4%±12.7	42.4%±16.1
A4	<90	40.1%±12.5	42.0%±15.9
A5	<70	39.7%±12.5	41.4%±15.8
A6	<50	39.4%±12.5	41.0%±15.9
A7	<40	39.3%±12.5	40.9%±15.9
A8	<30	39.2%±12.5	40.8%±15.9

To improve the prediction rates, the *hybrid method* was tested using the *noise filtering scheme* whereby, for each query pentapeptide, data in PENTAdb coming from closest homologues was queried in preference to data from more distant homologues (see also supporting information S7). Results are detailed in Table 11. When compared to the *without noise filtering scheme* (Table 10), the prediction rates improved to reach a maximum of 66.3%. Interestingly, for experiments B2 to B7 where closest homologues to be queried first are in the range of <40% to <95% sequence identities, the prediction accuracies remained high at a level of about 61.6%. Only in experiment B8 the prediction accuracy rate dropped to 40.8%. It must be noticed that this experiment is identical to the one featured for <30%

threshold (see Table 10) using the *hybrid method* and the *without noise filtering scheme* for querying the database.

Table 11: Assessment of the performance of PB-kPRED using the hybrid method with noise filtering scheme for querying the database.

For each query protein, the portion of the database accessible to the algorithm is first restricted to the closest homologues and if no hits were found, only then the more distant homologues are made accessible progressively. Eight results shown here correspond to the eight experiments represented schematically in Figure 6. Shown are the prediction accuracies averaged over 15,544 query proteins from the PDB30 dataset that was used in this assessment.

Experiment	Closest homologues to be queried first	Average prediction accuracies by Hybrid method with noise filtering (%)
B1	100 %	66.31±27.62
B2	<100 %	61.61±24.50
B3	<95 %	61.60±24.49
B4	<90 %	61.59±24.48
B5	<70 %	61.59±24.48
B6	<50 %	61.59±24.47
B7	<40 %	61.58±24.47
B8	<30 %	40.79±15.90

All results further detailed hereafter are concerned with data obtained in experiment B1 where *hybrid method* was applied using the *noise filtering scheme* for querying PENTAdb and where best predictions were obtained.

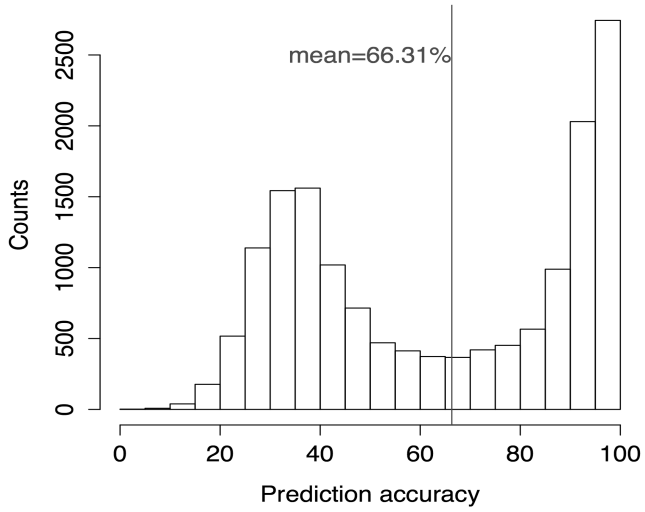


Figure 19. Histogram depicting the distribution of the observed prediction accuracies for 15,544 query proteins by hybrid method using noise filtering scheme.

The distribution of the prediction accuracies for experiment B1 (Table 11) is bimodal (Figure 19). A spike in frequency is observed at the >80% range representing the set of queries which have closely related proteins of known structure available in the PDB and for which the method is able to perform extremely well. Many queries in this set indeed had prediction accuracy values above 90%. At the other end of the spectrum, there is an almost normal distribution with an average around the 35%-40% accuracy range. Hence, the mean falls in between these two at 66.31% accuracy. This distribution did not substantially vary when homologues sharing less than 100% to 40% sequence identity to the query corresponding to experiments B2 to B7 respectively were queried first (data not shown). However, once the twilight zone of 30% sequence identity is crossed, the accuracy distribution drastically changes to that of a unimodal distribution with a very sharp peak at the 40% range and gradually tapering tail towards the higher accuracies (see supporting information S3).

The accuracy by the *hybrid method* using the *noise filtering scheme* was compared to the *majority rule method* (Figure 20). As shown by the data points below the diagonal, the *hybrid method* performed significantly better than the *majority rule method* for a total of 8,195 cases (52.7%) out of the 15,544 protein queries. For remaining 7,245 cases, the *majority rule method* performed slightly better than the *hybrid method*.

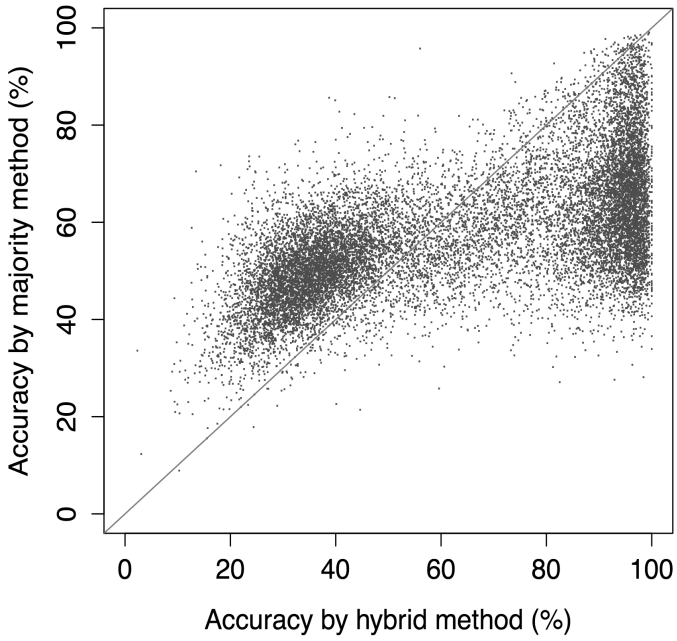


Figure 20. Comparison of the majority rule method without noise filtering scheme and the hybrid method with noise filtering scheme.
Shown are the predictions accuracies for the 15,544 query proteins from PDB30 dataset. The diagonal line separates the points where the majority method performs better and the points where the *hybrid method* performs better. Points lying along the diagonal (bisector) represents the situation where both the methods perform equally.

3.4.5. PB predictions

Results from the best performing condition (experiment B1 featured in Table 4) were further analyzed for the PB-wise prediction rates and compared with published rates from other methods (Table 5). The rates are heterogeneous across the 16 PBs. Top two best-predicted PBs by PB-kPRED were PB *m* and PB *a*, with accuracies of 75.9% and 67.2%,

respectively. On the other hand, the two most badly predicted PBs by PB-kPRED were PB *j* and PB *g* with prediction rates of 49.9% and 43.5% respectively. Analysis of the corresponding confusion matrix (supporting information S4) shows that, the prediction algorithm frequently gets confused between the PBs *c* and *d.* PB *c* is wrongly predicted as PB *d* almost 31,000 times (22.4%). The vice-versa, PB *d* being predicted as PB *c* is more than 42,000 times (16.9%). These PBs are in fact highly related (i) as seen from a pure structural point of view, low Root Mean Squared Deviation on Angular values (RMSDA*)* and similar transitions [28] and (ii) as they have been seen to be highly interchangeable thanks to PB substitution matrix [31,32]. As some PBs are highly similar, it is possible to relax the assessment, *i.e*. considering two PB series as equivalent. With such relaxed criteria, the accuracy increases from 66.31% to 68.87% (a 2.56% gain on average). Interestingly, significant increases in accuracies were observed for PB *g* (from 43.5% to 67.4%) and for PB *j* (from 49.98% to 67.2%).

PB-kPRED globally outperformed two other PB prediction methods (Table 12). Its predictions were better for all the 16 PBs when compared to the Bayes method and better than almost all PBs when compared to LOCUSTRA. Only PBs *d* and *m* were better predicted by this latter method[20]. kPRED predictions were also evaluated against secondary structure assignments from DSSP. This is not a straightforward activity as the 16 PBs are not a categorization of the secondary structure elements. Nonetheless, the results of the comparison are presented in supporting information table S8 and the results are favorable when compared to popular secondary structure prediction algorithms, at 80.94%. It is important to note that, secondary structure is useful mainly for a good description of repetitive structures, i.e. alpha-helix and beta-sheet and not for the remaining approximately 50% of the protein structure that are not defined, namely the coil.

A Mathews correlation coefficient (MCC) close to +1 indicates a good agreement between the observed and the predicted outcomes and a MCC of close to -1 otherwise. For our analysis all the 16 PBs had MCCs between 0.5 and 0.7. PBs *a* and *m* were close to 0.7, PB *g* at 0.51 and the remaining fluctuated around the 0.6 mark. The sensitivity and specificity ranges were 0.4-0.7 and 0.9-1.0 respectively. A common pattern is observed in the case of PBs corresponding to the regular secondary structure elements (PBs *d* and *m*): in both these cases, the sensitivity values peak while the specificity values plummet. Although the sensitivity values varied between 0.4 and 0.7, the specificity values were consistently above 0.9 indicating that the method was able to achieve a very high true negative rate. High specificity may be explained by the binary nature of the test while sensitivity is a sixteen state test.

3.4.6. Measure of accuracy

A probabilistic scoring function was developed for the *a posteriori* analysis of the predicted PB sequences so as to provide a measure of how accurate the *hybrid method* using the *noise filtering scheme* was performing. An assessment of the scoring function is provided in Figure 21. It shows that the score is correlated with the accuracy of the prediction with a Pearson's correlation coefficient of 0.82. The two distinct clusters of data points correspond to those featured in the histogram in Figure 19.

Table 12. Assessment of the performance of PB-kPRED and comparison with other previously reported methods.

Shown are the PB-wise prediction accuracies for experiment A8 of the majority method and four different experiments of *hybrid method* with noise filtering. These are compared with PB-wise results from LOCUSTRA [20] and Bayesian approach [26]. Experiment B1 PB-wise accuracies were compared with the other two methods and corresponding cell values in bold represent the best accuracy achieved between the experiment B1 of *hybrid method*, LOCUSTRA and Bayes method.

PBs	PB frequency	Majority method	Hybrid method						Other methods	
		Expt A8	Expt B8	Expt B4	Expt B2	Expt B1			LOCUSTRA	Bayes method
		Accuracies					Specificity	MCC	Accuracies	
a	3.68%	34.40%	45.93%	64.60%	64.60%	**67.20%**	98.15%	0.69	58.16%	56.60%
b	4.29%	15.19%	18.99%	44.52%	44.58%	**52.15%**	97.72%	0.56	26.14%	20.90%
c	8.31%	23.45%	28.76%	51.62%	51.66%	**58.53%**	95.95%	0.58	44.81%	32.90%
d	18.68%	39.24%	40.09%	61.83%	61.85%	67.00%	94.12%	0.63	**71.58%**	54.00%
e	2.18%	21.78%	28.59%	52.30%	52.38%	**57.45%**	98.96%	0.62	44.74%	38.60%
f	6.45%	24.87%	29.49%	54.64%	54.64%	**60.30%**	97.21%	0.61	41.45%	30.90%
g	1.14%	10.33%	14.94%	36.79%	36.87%	**43.45%**	99.19%	0.51	26.84%	30.10%
h	2.10%	24.25%	33.02%	56.90%	56.89%	**61.05%**	98.82%	0.64	38.45%	40.90%
i	1.49%	20.69%	30.45%	55.36%	55.35%	**59.17%**	99.18%	0.63	36.87%	38.10%
j	0.83%	15.48%	20.81%	42.57%	42.68%	**49.98%**	99.40%	0.56	48.19%	49.70%
k	5.25%	30.46%	35.59%	59.30%	59.32%	**63.93%**	97.67%	0.65	46.46%	33.40%
l	5.20%	26.56%	31.88%	54.87%	54.93%	**59.99%**	97.69%	0.62	42.71%	35.50%
m	32.89%	60.96%	55.68%	72.38%	72.40%	75.89%	91.03%	0.67	**83.76%**	70.60%
n	1.78%	26.48%	35.40%	58.16%	58.21%	**62.15%**	99.05%	0.65	52.08%	50.00%
o	2.44%	29.62%	38.40%	59.49%	59.52%	**63.19%**	98.70%	0.66	55.10%	48.10%
p	3.29%	23.22%	31.72%	54.27%	54.28%	**59.24%**	98.25%	0.62	40.80%	29.20%

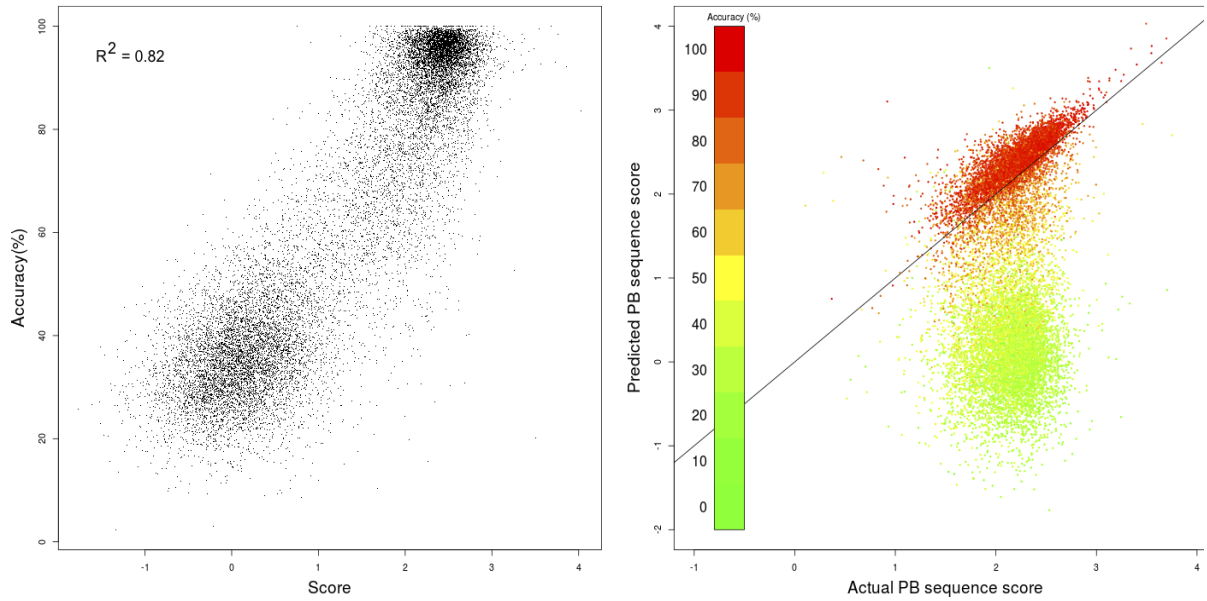


Figure 21. Assessment of the ability of the scoring function to estimate the prediction accuracy of PB-kPRED.
On the left is shown the plot of score versus accuracy and on the right the plot of the scores for predicted PB versus real PB sequences for the 15,544 query proteins of PDB30 dataset. Data points are coloured based on level of accuracy of predictions.

As a further assessment of the scoring function, the scores for the predicted PB sequences were compared with the scores for the real PB sequences (Figure 21). It shows that in case of more accurate predictions (rates above 60%), the two scores correlated very well (red points along the diagonal in Figure 21) with both score values mostly ranging between +1 and +3. In the case of less accurate predictions (rates below 60%), the two scores were no more correlated (green dots below the diagonal in Figure 21) and scores for predicted PB sequences ranged mostly between -2 and +1.

3.4.7. Case studies

Here were considered the predictions for 5 specific cases to look at the strengths and limitations of the PB-kPRED algorithm namely in presence or absence of homologues of known structure (see Table 13). These case studies correspond to counter intuitive prediction instances where (i) prediction accuracy is high despite not having any close homologues and

(ii) prediction accuracy is low despite having sequences of PBs of closely-related proteins in PENTAdb.

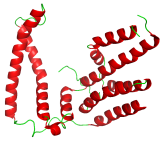
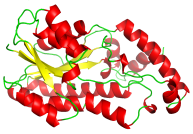
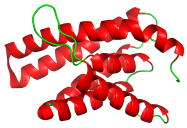
Regarding prediction in employing information from homologues of known structure, three contrasting cases were studied. The first case relates to chain A of a hypothetical DNA binding protein from *Salmonella cholera* (PDB id 2HX0_A) which has a homologue from *Salmonella typhimurium* (PDB id 2NMU) that is 100% identical, 100% accuracy was achieved as shown by the high accuracy score of 2.81. Both structures aligned very well with a Root Mean Squared Deviation (RMSD) of 0.14 Å (supporting information S5). The second case relates to an energy-coupling factor transporter transmembrane protein EcfT from *Lactobacillus brevis* (PDB id 4HUQ_T) which has two other "homologues" (PDB id 4RFS_T and 4HZU_T) that are 100% identical to the query. Here, the prediction accuracy is 73.4% only with an accuracy score of 1.31. 3-D structural alignment with these two "homologues" resulted in RMSDs 1.41 Å and 1.90 Å respectively displaying some structural variations (supporting information S5) despite being 100% identical at the amino acid sequence level. These structural variations were due to rigid body movement. The third case is chain A of a pyrimidine deaminase / uracil reductase from *Thermotoga maritima* (PDB id 2HXV_A) which had only ten very distantly related proteins that shared less than 30% sequence identity in the PDB. Prediction rate is even lower here with accuracy reaching a value of 39.1% as shown by the low accuracy score of 0.30.

As for predictions in absence of homologues of known structure, two contrasting cases were studied. The first case is about a human hydroxysteroid dehydrogenase and the second case is a membrane protein associated with Ecf transporter from *Lactobacillus brevis*. The prediction performed quite well in the first case with an accuracy of 75.4% as shown by the

high accuracy score of 1.99 while in the second case, the prediction almost completely failed with the accuracy of only 9.37% and also shown by the unfavorable accuracy score of -0.28.

Table 13. Impact of the availability of known homologues on the accuracy of PB-kPRED.

Query PDB chains with known homologues and with no known homologues are featured. The *hybrid method with noise filtering* scheme for querying the database was applied for the prediction whereby the conditions were identical to experiment B1 (see Methods section) as featured in Table 6. The queries themselves were excluded from the database prior to prediction. Shown here are the accuracies of the predictions and the numbers of known homologues for different sequence identity thresholds.

	Queries with known homologues in PDB			**Queries with no known homologues in PDB**	
Sequence identity thresholds (%)	**2HX0_A** (hypothetical DNA binding protein) 	**4HUQ_T** (energy-coupling factor transporter EcfT) 	**2HXV_A** (deminase/ reductase) 	**1A27_A** (hydroxysteroid dehydrogenase) 	**4HZU_S** (transmembrane protein associated with Ecf transporter) 
100	2	3	1	1	1
95	2	3	1	1	1
90	2	3	1	1	1
70	2	3	1	1	1
50	2	3	1	1	1
40	2	5	1	1	1
30	2	5	10	1	1
Accuracy (%)	100%	73.41%	39.13%	75.44%	9.37%
Accuracy score	2.81	1.31	0.30	1.99	-0.28

3.4.8. Implementation of the PB-kPRED methods as a web-tool

The PB-kPRED methodology has been implemented as a web-tool that is freely available to the community at http://www.bo-protscience.fr/kpred/. Both *majority rule* and *hybrid method without the noise filtering scheme* for querying the database have been implemented. The tool provides a predicted PB sequence for each query amino acid sequence and also provides the *accuracy score* that serves as an *a posteriori* estimation of the prediction accuracy. In case the prediction score value is below -1, the prediction accuracy cannot be estimated and the user is notified. Users can provide multiple query protein sequences. All results are downloadable as FASTA formatted flat files. Optionally when submitting numerous query sequences, the user can provide an email address to which a notification will be sent when the job is completed.

3.5. Discussion

The exponential growth in the structural knowledge of proteins has warranted the necessity of competent knowledge-based prediction algorithms for local structure prediction. At the level of short protein segments like pentapeptides, this increase in structural knowledge invariably brings with it an unprecedented signal to noise ratio for deciding on the most probable local conformations. Indeed, it is well established that similar pentapeptides can adopt different local conformations [3,33]. This is verified when the content of PENTAdb is inspected.

Hazout's team along with defining the protein blocks also predicted the local structure in terms of PBs using a Bayesian approach [13]. They achieved an accuracy of 34.4% using a 15-residue window and this increased to 40.7% upon supplementing the Bayesian predictor with sequence profiles in the form of *sequence families*. In 2005, Etchebest and colleagues

[27] used a combination of statistical optimization procedure and improved sequence family data to bump up the accuracy to 48.7%. Incorporating secondary structure predictions from PSI-PRED into the PB prediction process did not contribute much to improve the accuracy *i.e* only 1% gain resulting in 49.9%. Machine learning techniques have also been used to predict protein local structure in terms of PBs. Support vector machine based methods like LOCUSTRA [21], svmPRAT [23] and SVM-PB-Pred [22] achieve mean accuracies of 61.0%, 67.0% and 53.0% respectively. A dual layer neural network based prediction method achieved 58.5% accuracy [25]. The most refined version of the PB-kPRED method proposed here, i.e *hybrid method with noise filtering scheme*, outperformed most of the previously developed methods for PB prediction except for svmPRAT where it performed equivalently. Although all the methods evaluated their accuracies on non-redundant sets of proteins, an even comparison is hindered by difference in datasets, varying training regimes for the machine learning methods and different levels of sequence identity used as input in the prediction process. This motivated us to perform a battery of tests on the algorithm to estimate the prediction accuracy when incremental levels of sequence identities are made available in PENTAdb for the prediction (see Table 8). Importantly, to our knowledge, this is the first report of a querying scheme that dynamically filters out, on a per query basis, homologues at different cut-off values so that the portion of the PENTAdb that is made accessible for prediction is calculated on the fly. For each of the 15,544 query sequences of PDB30, 16 experiments were performed, hence amounting to a total dataset size of 248,704 proteins. Using the *noise filtering* strategy, PB-kPRED was able to efficiently weed out the noise present in the database due to redundancy and hence to narrow down the search in the database to find the most appropriate local structure for a given pentapeptide. Therefore, filtering out from the database the pentapeptides from proteins that shared less than 30% sequence identity with the

query indeed improves the prediction efficiency. In general, in those cases where pentapeptides from very close homologues were indeed present in the database, kPRED performed extremely well.

Remarkably, the reported average accuracy (66.31%) is very high considering the fact that for every pentapeptide, there are 16 possibilities. Indeed, if random PB predictions were made taking into account respective PB frequencies, prediction rate (Q16 value) would be very poor with an average value of about 7 to 8%. Similarly, when predicting secondary structure, there are only 3 possibilities and random Q3 prediction accuracies would be of about 30% .

When the *majority method* and the *hybrid method* (Figure 20) were compared, two distinct clusters were noticed. Upon further investigating the reason for this distinct clustering, we note that, irrespective of the sequence identity cut-off, the points below the diagonal were found in more populated clusters while the points above the diagonal were found in least populated clusters. Hence the *hybrid method* using the *noise filtering scheme* will perform better when there are some closely-related protein structures to look-up to in PENTAdb. In a real-life scenario, this will not be always the case. Indeed, proteins for which we want to predict the structure and which do not have any homologues even at 30% sequence identity are not so uncommon. This brings us to the conclusion that even though overall the *hybrid method* performs better, we cannot ignore the *majority rule method* all together.

Nonetheless, this method still has room for improvement as it can be seen from the values in Table 9. The list of all possible PBs reported by the PB-kPRED algorithm after querying PENTAdb database indeed shows that the good PB was present in more than 70% of

the cases. However, owing to the scoring functions S1 and S2 (see Methods for more details), both *majority rule* and *hybrid methods* failed to pick up these good PBs as predictions in several instances.

Interestingly, once the local backbone of a protein was predicted in the form of a PB sequence, we were able to provide an *a posteriori* assessment of how accurate was the prediction. The method used here to achieve this relied on the simple idea that successions of PBs should follow the rule that not all combinations of PBs would be allowed. This intuition turned out to be correct since there was a remarkable correlation between the score and the accuracy of the predictions. Noteworthy, the *accuracy scores* for actual (native) PB sequences are overwhelmingly distributed between +1 and +3, while poorly predicted PB sequences have scores below +1. This scoring of PB sequences could also serve as an indicator towards improving predictions. Because the calculation of the score of a PB sequence is very fast, one could imagine implementing a score-guided optimization procedure to climb the prediction accuracy gradient using Monte-Carlo or genetic algorithms for example.

The case studies documented in this work (Table 13) indicate that the relationship between local structure predictability and the number of homologues of the query available in PDB are not very straightforward. Optimistically, in spite of not having any homologues, the PB-kPRED algorithm can perform a good prediction if the pentapeptides constituting the query adopt consensus local structures for the respective pentapeptides. Two such examples were provided but with contrasting outcomes, one achieving good accuracies and the other failing to predict correctly the PB sequence. Interestingly, the *accuracy scores* provided by our scoring function helped to reliably differentiate one prediction from the other. On the other hand, even if a query has multiple homologues in the PDB, its prediction accuracy will

take a hit if the homologues are contrasting structural analogues of the query. For example, the activation of human pancreatic lipase involves considerable conformational transition in the form of a 'lid movement'. The hypothetical prediction case when the query is the 'lid open form' and PENTAdb has pentapeptides from the 'lid closed form' would confuse the prediction algorithm despite both the forms of lipase being identical in amino acid sequences. Hence these case studies establish two take home messages: (i) there are exceptions to the general observation that the presence of homologues improves the prediction accuracy of PB-kPRED and (ii) the *accuracy score* used to evaluate the predictions is a reliable gauge for estimating the accuracy of the method as illustrated in Figure 21.

PB-kPRED web-server could form a vital link in the pipeline of PB based structure analysis tools. Namely, it can be bridged with PB-based fast structure comparison tools like iPBA [18] and PB-ALIGN [17] and help to mine for similar structures and map the fold space. It can also be used to predict the occurrence of structural motifs in protein sequences. Indeed, the alpha version of the server which was made available on-line earlier, has already been used by some research groups for the structural characterization of RNA binding sites in protein structures and predicting proteins sequences that contain RNA binding sites [34,35] and also in predicting β-turns and their types [36].

The web-based tool currently does not feature the *hybrid method with noise filtering scheme* because it would require running an instance of BLASTClust on every query. We plan to implement this functionality in a future version of the tool.

Our objective in this work was to predict protein backbone represented as protein blocks. We believe that the assembly of protein blocks fragments can ultimately lead to 3D structure prediction using sophisticated algorithms like the one implemented in Rosetta.

Indeed, Rosetta achieves this by leveraging on the I-Sites fragment libraries developed by Bystroff and Baker [10]. These fragment libraries, including PB-based fragment libraries, can be of great interest in the research of protein fragments in the twilight zone when doing 3D-modeling. We are currently developing a method and tool towards this end that takes advantage of existing algorithms implemented in Modeller [37] and Rosetta [10].

3.6. Methods

3.6.1. Dataset

All the protein chains from PDB [1] were segregated into clusters culled at 30% sequence identity using the BLASTClust algorithm [30] resulting in a collection of 15,544 clusters. The dataset comprises of 15,554 protein chains each corresponding to the best representative structure available from each of these clusters. It is hereafter termed as "PDB30 dataset". Preference was given to crystallographic structures over NMR and electron microscopy structures and also preferring better resolution and lowest R-value structures. Out of these 15,544 structures, 14,207 are crystallographic structures, 1,128 are from NMR experiments and 209 are solved by electron microscopy. Further, chains smaller than 100 residues were filtered out. We preferred to keep the NMR and EM structures, as we wanted to investigate if the experimental method impact on the quality of the predictions. For each of these 15,544 proteins the subsets of PDB that were homologous at 30%, 40%, 50%, 70%, 90%, 95% and 100% as reported by BLASTClust were also calculated in order to implement the "*Hybrid method with noise filtering*" scheme.

3.6.2. Protein Blocks

The set of protein blocks (PBs) is a structural alphabet composed of 16 structural prototypes each representing backbone conformation of a fragment of 5 contiguous residues

[12,13]. The 16 PBs are represented by the letters *a* to *p* and were identified from a collection of 228 non-redundant proteins. Clustering these pentapeptides was based on the 8 dihedral angles (ψ_{i-2}, φ_{i-1}, ψ_{i-1}, φ_i, ψ_i, φ_{i+1}, ψ_{i+1}, φ_{i+2}) that define their local backbone conformation. An unsupervised learning algorithm (Kohonen algorithm [38]) was used to arrive to an unbiased classification of the dihedral vectors and to the definition of standard dihedral angles for each PB. Protein blocks are assigned on the basis of the dissimilarity measure called RMSDA between observed dihedral angles and the standard dihedral angles for the 16 PBs. The PB with lowest RMSDA is assigned to the central residue of the pentapeptide region. The choice of fragment size as 5 and library size as 16 for the PBs was because 5 consecutive residues capture well the local contacts in regular secondary structures (α-helices and β-strands) and 16-library size is a good balance between the specificity and sensitivity of predictions [13].

All the 15,544 protein chains from PDB30 dataset were encoded into their corresponding protein blocks sequences (PB sequences) after comparing their backbone φ and ψ torsion angles with the corresponding standard torsion angles for the 16 PBs [12] using an in-house Perl script. Sequence of PBs as observed in crystal and NMR structures were later used as a reference to assess the accuracy of predicted PB sequences.

3.6.3. Database of pentapeptide conformations from protein structures

A database of pentapeptide conformations (PENTAdb) was developed using known 3D structures of proteins. PENTAdb is essentially the entire structural information contained in the PDB, broken down into chunks of pentapeptides. A sliding window of 5 residues was used to extract structural features for every overlapping pentapeptide of a polypeptide chain. The dihedral vector associated with the five consecutive residues that is required to assign PBs as described in the previous section was obtained from the DSSP [39] program. All the

information was stored as a MySQL relational database. PENTAdb is maintained up-to-date; the update frequency corresponds to the weekly updates of PDB. The protein chains from which the pentapeptides are extracted are filterable at 30%, 40%, 50%, 70%, 90%, 95% and 100% sequence identity thresholds.

3.6.4. Prediction scheme

The overall scheme for predicting the local structure in terms of PBs is based on querying the PENTAdb database for every constitutive pentapeptides of a query protein sequence using a sliding window of 5 residues (Figure 22a). Hits from the database are reported as predicted protein blocks (PBs). Predicted PBs are assigned to the central residue of each query pentapeptide. The prediction results are presented at different levels of refinement. The prediction in the coarsest form consists of the list of all the possible PBs for a particular pentapeptide of the query protein sequence. This is the case when multiple hits from PENTAdb database are obtained for a particular query pentapeptide (Figure 22b). The multiple hits correspond to the different conformations, which the pentapeptide has been seen to adopt in protein structures (Figure 22b). When the query pentapeptide is not found in PENTAdb, the information available for the tetrapeptides covering the first four residues with a wildcard for the fifth position was used (Figure 22b) to identify the list of possible PBs with first 4 amino acid residues matching this query. The position of wildcard did not influence the outcome of the results (data not shown). The list of hits thus obtained is referred as *all possible PBs*. This list serves as a framework from which the most probable PB sequence is predicted.

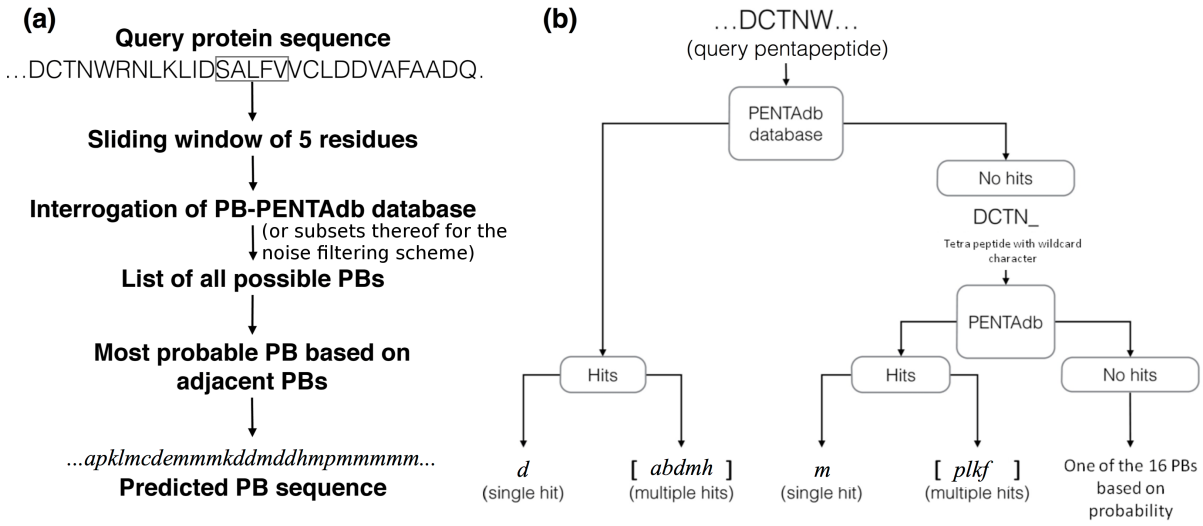


Figure 22. The knowledge-based methodology behind PB-kPRED.
(a) Overview of the scheme followed by PB-kPRED for the prediction process. (b) The different outcomes possible when PENTAdb database is queried for a pentapeptide sequence: hits are reported as a single PB or multiple PBs.

Two methods were explored to predict the optimal PB sequence within the list of all the possible PBs obtained after querying the database (Figure 23). The first method, termed as *majority rule method*, is purely probabilistic and consists of simply picking up the most frequently observed PB for each query pentapeptide. It corresponds to the PB that has highest S1 score, where S1 scores are simply the raw counts of all possible PBs reported by PENTAdb database for the query pentapeptide. In cases when there is no decisive majority (two or more equi-probable PB), both of them are reported as predictions.

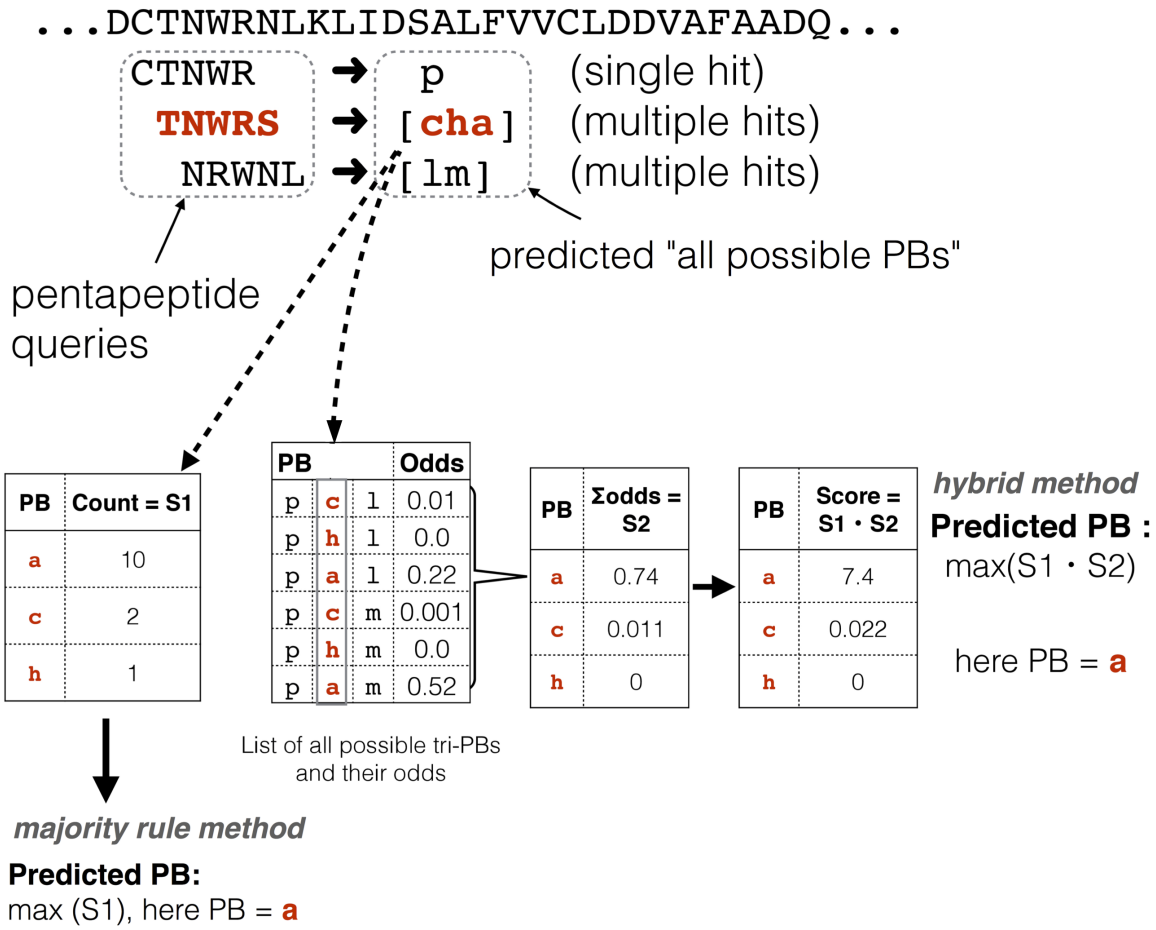


Figure 23. Details of the scoring schemes underlying the majority rule method and the hybrid method.
S1 scores are simply the raw counts of all possible PBs reported by PENTAdb database for a given query pentapeptide. S2 scores are calculated through the summation of the odds of tri-PBs that have a common PB in the central position. For the majority rule method, predictions are based only on the ranking of the S1 scores. For the *hybrid method*, predictions are based on the ranking of the product of scores S1 and S2.

However, it is known that the structure adopted by a short peptide can be highly dependent on its local environment[3]. A second method that integrates contextual information was hence developed and is hereafter termed as *hybrid method*. Here, to predict the local structure of a pentapeptide, the information about the structural status (in terms of PBs) of the two immediately adjacent and overlapping pentapeptides (preceding and succeeding) is also taken into account (Figure 23). It requires a normalized frequency look-up table for observed

motifs of 3 consecutive PBs also termed as tri-PBs (see supporting information S6 for more details). For each query pentapeptide, in complement to the calculated S1 score, an additional S2 score is calculated as follows. A list of all possible combinations of three successive PBs (tri-PB motifs) is built. This is derived from the list of *all possible PBs* for the query pentapeptide and for its two adjacent pentapeptides (Figure 23). For each possible tri-PB motif, their normalized frequencies ("odds") are looked up in the tri-PB normalized frequency table. S2 scores are calculated through the summation of the odds of tri-PB motifs that have a common PB in the central position. The predicted PB for the query pentapeptide is determined after multiplying S1 scores by their corresponding S2 scores and taking the highest value among these products (Figure 23). This approach is called the *hybrid method* because it combines the *majority rule method* with contextual information in the prediction process.

3.6.5. Evaluating PB-kPRED using different subsets of PENTAdb

Two evaluation schemes were developed to benchmark the PB-kPRED methodology. As mentioned above, the query dataset used here constituted of the 15,544 proteins from the PDB30 dataset. The schemes relied on the ability to control which subsection of PENTAdb will be accessible to the prediction algorithm for every query. For example, allowing only pentapeptides in PENTAdb from non-homologues to be accessible by the prediction algorithm emulates a scenario of attempting to predict the local structure of a protein with no known homologues. On the other hand, as in the case of other local structure prediction methods [13,21–23,27], it can be advantageous to have the ability to privilege information from homologous structures when these are available to predict the local structure of a query protein. Such a scheme can be emulated by allowing only pentapeptides in PENTAdb from closest detectable homologues to be accessible by the search algorithm (Figure 24).

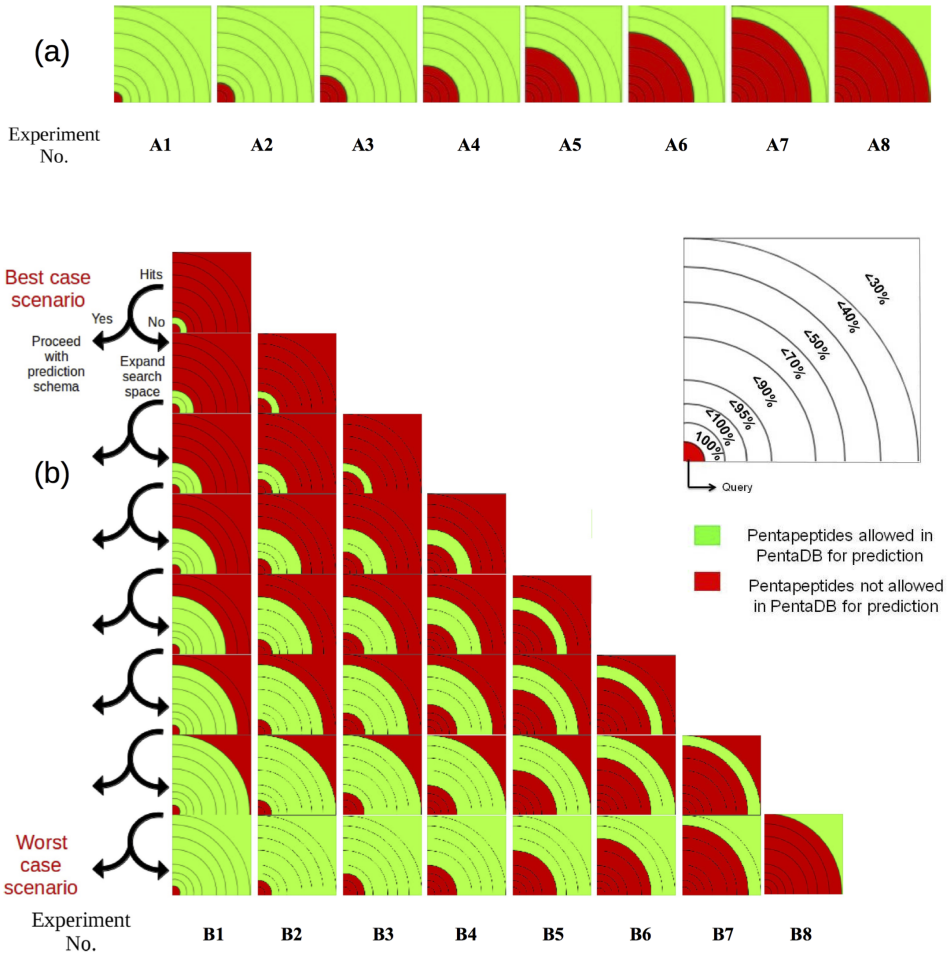


Figure 24. Diagrammatic representations of the schemes used by PB-kPRED for querying PENTAdb.
Shown are (a) the so-called "classic" or "without *noise filtering scheme*" and (b) the "with *noise filtering scheme*". Sections of the database accessible are indicated in green and those not accessible in red. The sections are delimited by sequence identity thresholds. In both schemes, eight different experiments (A1 to A8 and B1 to B8) were performed. See supporting information S7 for detailed legend of this figure.

In first instance, the prediction methodology was assessed with increasing sequence identity cut-offs ranging from 30%, 40%, 50%, 70%, 90%, 95% to 100%, named experiments A1-A8 (Figure 24a). This scheme is subsequently termed as "*without noise filtering scheme*". In second instance, an alternative assessment scheme hereby called the "*with noise filtering scheme*" was applied to further assess the PB-kPRED methodology (experiments B1-B8, see Figure 24b). It aimed at evaluating how privileging information from close homologues, when

available, contributed to improve the quality of the predictions. In brief, the algorithm initially searches for a pentapeptide among the closest homologues first. If the search finds a hit, then the hit is used for the prediction; otherwise the search space is increased to include the immediately next level of more distantly related homologues. This process is repeated until a hit is obtained. Due to this process of introducing more distant homologues in a conditional fashion, wrong pentapeptides (noise) from PENTAdb were potentially filtered out, hence the name *with noise filtering scheme.* In all the cases, care was taken to exclude the pentapeptides from the query proteins themselves.

Reducing the PB predictions into a binary outcome permits the use of classical Mathews correlation coefficient (MCC) to compare our predictions to a random choice. MCCs for the 16 PBs were evaluated based on a confusion matrix similar. For each PB, MCC was calculated according to Equation 1.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (1)$$

3.6.6. A scoring function to estimate the accuracy of the predictions

A probabilistic scoring function was developed for the *a posteriori* analysis of the predicted PB sequence through namely the analysis of its content in penta-PB motifs, with the objective of providing a measure of how accurate PB-kPRED was performing. The principle of the analysis relies on the fact that not all penta-PBs are found in proteins at the same frequency. Indeed, many successions of 5 consecutive PBs are highly improbable because they are geometrically not allowed as explicated by the Ramachandran rules. The probabilistic function is hence based on the look-up table of normalized frequencies of successive penta-PB motifs observed in a non-redundant set of protein structures (see S6 Additional method in supporting information). In brief, using a sliding window of 5 consecutive PBs (penta-PB

motif) along the predicted PB sequence, the normalized frequencies of all penta-PB motifs were looked-up in the penta-PB frequency table. The logarithm of these normalized frequencies were then summed and divided by the length of the predicted PB sequence to generate an *accuracy score* (*A*) as shown here:

$$A = \frac{\sum_{i=1}^{l-4} \log\left(N_i\right)}{l} \quad (2)$$

where *A* is the *accuracy score* for a predicted PB sequence, *l* is the length of the PB sequence, *N* is the normalized frequency of the penta-PB motif observed at window position *i* in the PB sequence. Since an overlapping sliding window of five consecutive PBs is used, the total number of penta-PB motifs (*i.e* the number of windows) is *l*-4. In the case a particular penta-PB motif has a null value in the frequency table (*i.e* it is never observed), a penalty of -5 was instead added to the score.

3.7. Supporting information

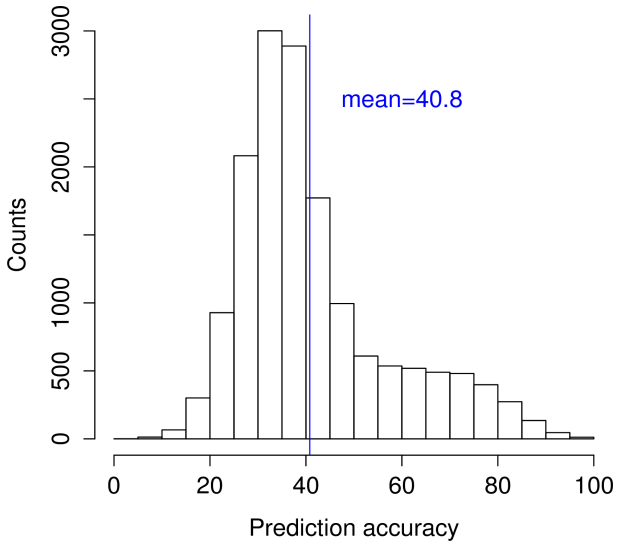
S1 Table. Distribution of total penta-PBs and unique penta-PBs at different sequence identity cut-offs in the PDB.

Seq. identity cut-off (%)	Total number of penta-PBs	Total number of unique penta-PBs
30	4,366,040	40,130
40	5,239,216	42,779
50	6,025,703	45,182
70	7,053,751	48,186
90	7,895,276	50,597
95	8,229,825	51,514
100	12,508,483	57,889
Full PDB	**61,208,019**	**94,263**

S2 Table. Assessment of the list of PBs reported after querying PB-PENTAdb for the 15,544 query proteins from PDB30 datatset. Shown is the number of times the correct PB were in the list of all possible PBs.

Protein Block	Total	Correct PB in list	Percentage
a	145,102	96,501	66,5%
b	172,135	86,256	50,1%
c	332,295	207,031	62,3%
d	753,725	573,645	76,1%
e	84,282	48,114	57,1%
f	257,369	156,029	60,6%
g	45,250	18,251	40,3%
h	83,152	48,627	58,5%
i	58,657	31,993	54,5%
j	32,598	17,356	53,2%
k	209,416	131,400	62,7%
l	207,772	127,305	61,3%
m	1,294,063	1,119,179	86,5%
n	69,954	41,171	58,9%
o	95,682	59,422	62,1%
p	130,506	74,971	57,4%
Total	**3,826,856**	**2,740,750**	**71,6%**

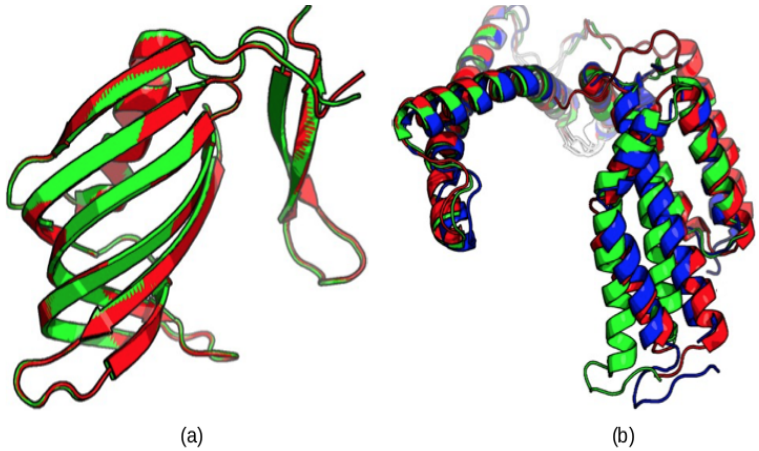
S3 Figure. Histogram depicting the distribution of prediction accuracies for 15,544 query proteins by hybrid method using noise filtering scheme when only distant homologues (<30%) are allowed in PENTAdb. The distribution is unimodal with sharp peak at the 40% accuracy range and gradually tapering towards the higher accuracies.



S4 Table. Confusion matrix for the predictions by Hybrid method using *noise filtering scheme* for querying PENTAdb. The PBs in the column are the actual PBs and the ones in the rows are the predictions. The diagonals are the raw counts for the correctly predicted PBs and these served for calculating the PB-wise identities and the overall accuracy of PB-kPRED (see also Table 4). The off-diagonal values highlighted in grey are the confused predictions that are still acceptable for they correspond to PB substitutions frequently observed among homologous preoteins (Tyagi et al. 2006; Joseph, Srinivasan, and De Brevern 2011). These served for calculating the PB-wise similarities and the overall Q16 similarity. The last column (Δ) represents the PB-wise difference between similarity and identity. It highlights those PB-wise predictions which gained upon considering similarity as a criteria for evaluating the predictions.

	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***	***l***	***m***	***n***	***o***	***p***	Identity	Similarity	Δ
a	97,530	3,235	8,333	8,735	856	3,307	1,125	951	1,108	640	3,010	3,013	9,290	608	1,331	2,063	67.20%	72.94%	5.74%
b	4,388	89,835	9,249	13,708	1,797	5,395	1,196	4,246	2,175	794	7,254	9,615	14,759	1,122	3,195	3,535	52.15%	55.88%	3.73%
c	11,022	8,894	194,601	30,899	2,640	11,788	3,639	3,113	1,849	1,592	7,000	6,465	34,343	1,804	4,605	8,226	58.53%	62.94%	4.41%
d	14,121	19,601	42,215	505,429	10,584	26,320	3,788	9,320	5,167	3,805	14,682	13,552	67,308	3,622	5,949	8,887	67.00%	67.00%	0.00%
e	1,167	1,631	2,201	6,703	48,449	4,745	2,928	1,861	661	664	3,128	1,273	4,765	2,913	551	689	57.45%	68.76%	11.31%
f	4,518	6,082	11,838	19,794	5,775	155,294	2,192	3,496	1,737	1,782	8,398	7,433	22,369	1,897	1,920	2,990	60.30%	63.24%	2.93%
g	1,251	942	2,841	2,139	2,499	1,602	19,673	492	348	277	1,177	1,152	5,401	2,387	1,037	2,061	43.45%	67.36%	23.91%
h	1,282	2,687	2,126	4,031	1,437	2,122	559	50,780	874	727	3,424	2,377	4,375	626	5,003	751	61.05%	71.00%	9.95%
i	1,362	1,685	1,575	2,648	407	1,279	340	760	34,715	2,090	1,224	1,830	3,076	378	623	4,674	59.17%	73.58%	14.40%
j	792	501	1,016	1,700	460	1,220	256	842	2,703	16,307	846	899	1,663	279	365	2,777	49.98%	67.18%	17.20%
k	3,927	5,781	6,710	9,723	2,864	6,944	1,571	4,461	2,129	1,752	133,967	6,028	16,908	2,183	2,161	2,429	63.93%	66.90%	2.97%
l	4,767	8,735	7,073	9,661	1,301	7,174	1,390	2,711	2,668	1,641	6,150	124,727	21,404	1,982	3,569	2,969	59.99%	60.78%	0.79%
m	17,605	20,513	41,891	69,175	6,385	26,869	8,090	7,242	4,910	4,379	25,529	26,605	982,610	14,685	17,029	21,327	75.89%	75.89%	0.00%
n	843	906	1,456	1,993	2,394	1,286	2,024	541	364	352	1,821	1,668	8,402	43,511	1,107	1,343	62.15%	68.54%	6.39%
o	1,588	2,265	2,939	2,904	411	1,344	790	4,830	622	365	1,573	2,788	9,575	893	60,484	2,346	63.19%	64.95%	1.76%
p	2,365	2,982	5,814	5,445	563	2,274	1,627	740	4,406	2,531	2,137	2,261	16,615	1,575	1,880	77,334	59.24%	65.07%	5.83%
																Average Q16	66.31%	68.85%	

S5 Figure. Structure alignment of two query cases with their respective homologues. (a) Shown is the alignment of hypothetical DNA protein binding protein with its identical homologue. RMSD is 0.14Å. (b) Shown is alignment of energy-coupling factor transporter EcfT with its 2 other 100% homologues. RMSD values with query are 1.4Å and 1.9Å.



S6 Additional method

Tri-PB and penta-PB normalized frequency tables

The approach that was used in this work to predict the structure of pentapeptides and to evaluate the predictions relied on two frequency tables built from the PDB30 dataset. The first look-up table constituted of all theoretically possible 4,096 (16^3) tri-PB motifs and was used for predicting the structure of a pentapeptide by taking into account the structures of its immediate preceding and succeeding pentapeptides (see *Prediction scheme* section). The second look-up table is composed of all 1.04 million (16^5) possible penta-PB motifs and was used for scoring the predictions fo PB-kPRED.. The frequency of occurrence of each tri-PB and penta-PBs was normalized following equation (1) below:

$$N_{j,k} = \frac{n_{j,k}}{\prod_{i=1}^{k} p\left(PB_{i,j}\right) \cdot \sum_{j=1}^{16^k} n_{j,k}} \quad \text{(Eqn. 1)}$$

where j is an integer associated with a PB motif, k is the length of the PB motif j (*i.e* the number of successive pentapeptides) with k=3 for tri-PB motifs and k=5 for penta-PB motifs, $N_{j,k}$ is the normalized frequency of the PB motif j, $n_{j,k}$ is the occurrence (raw count) of the PB motif j in PDB30 dataset and $p(\mathrm{PB}_{i,j})$ is the frequency of the observed PB at position i in the PB motif j.

S7 Additional legend to Figure 24. Schemes used by PB-kPRED for querying PENTAdb

Sections of the database accessible are indicated in green and those not accessible in red. The sections are delimited by sequence identity thresholds. In the classic scheme (a), eight different experiments (A1 to A8) were performed. They differ by the sections of PENTAdb that were made accessible for every pentapeptide query. For each of the 15,544 query protein sequences, the portion of PENTAdb accessible for prediction was dynamically determined using MySQL queries: only pentapeptides coming from protein chains in PDB that shared sequence identities below the indicated cut-off values were accessible to PB-kPRED. For example, in experiment A1, only the query was left out, the full PDB was accessible while in experiment A8, only proteins not sharing more than 30% sequence identity were made accessible for predictions. In the *noise filtering scheme*, the subset of PENTAdb that is accessible for query is dynamically expanded. Each individual vertical column represents an experiment (B1 to B8). For each experiment, a maximum homology level was used to define the subset of PENTAdb accessible by the PB-kPRED algorithm for predicting the local structure of each individual query pentapeptide. The defined maximum homology levels were ≤100%, <100%, <95%, <90%, <70%, <50%, <40% and <30% . For each query pentapeptide, data in PENTAdb only coming from closest homologues was queried first. For example, in experiments B1, B2 and B3, the highest allowed homology ranges were respectively 100%, ≥95 to <100% and ≥90 to <95% . If hits were found, these were used to constitute the list of all possible PBs (see text and Figure 22) and the algorithm proceeded with the next query pentapeptide. If no hits were found, subsequent subsets of PENTAdb comprising of more distant homologues were made available successively until hits were found. This is a procedure followed at an individual pentapeptide level and not at the level of the query protein sequence. Hence, in the best case scenario, this would represent a total of (l-4) queries to PENTAdb, l being the length of the sequence and in the worst case scenario, it would be 8 x (l-4) considering the 8 sequence identity cut-off levels. For the PDB30 dataset, this amounts to a total of 3.97 million queries in the best case scenario and up to 31.76 million queries in the worst case scenario.

S8 Comparison of kPRED to secondary structure assignment from dictionary of protein secondary structure (DSSP)

Sequence identity cut-off values	α-helix/3_{10} helix/π-helix mapped to PBs *k/l/m/n/o* and extended *β*-strand mapped to PBs *a/b/c/d/e/f*	α-helix/3_{10} helix/π-helix mapped to PBs *l/m/n* and extended *β*-strand mapped to PBs *c/d/e*
100 %	80.94 %	70.30 %
<100 %	78.81 %	67.96 %
<95 %	78.81 %	67.96 %
<90 %	78.81 %	67.96 %
<70 %	78.81 %	67.96 %
<50 %	78.81 %	67.96 %
<40 %	78.80 %	67.96 %
<30 %	65.90 %	55.20 %

Shown above are the results of comparing the 16 state PB prediction from kPRED to the reduced 3 state secondary structure assignment by DSSP (Kabsch and Sander, 1983). *PBs k, l, m, n, o* were related to helical repetitive structures (3_10, alpha and pi helices) while PBs *a, b, c, d, e, f* were related to extended structures (beta strands). These PBs correspond to central part of these repetitive structures and to their capping regions as shown by de Brevern (de Brevern, 2000). The kPRED algorithm could be predicting sub-secondary structure details for example, kinks and bends in helices by PB other than *k,l,m,n,o* but the three state accuracy (Q3) will favor us negatively in these cases. Hence measuring the performance of kPRED by Q3 accuracy is not totally justified because kpred algorithm was not designed with the objective of secondary structure prediction in mind. Interestingly, this mapping of kPRED prediction to SS status shows that kPRED accuracy (80.94%) is largely better than the prediction accuracy of (i) 66.4% of the method by Zhang et. al., which does not use evolutionary information (J. Mol. Biol. 225, 1049-1063) and (ii) 76.5% of the PSI-PRED method which uses evolutionary information (Jones, D.T., 1999, J. Mol. Biol. 292, 195-202).This simple evaluation highlight the quality of kPRED, here for something not designed for it.

3.8. Acknowledgements

Authors thank Tristan Rialland and Sara Bachiri for providing technical support in the development of the PB-kPRED web server.

3.9. kPRED predictions for additional datasets

The kPRED predictions by the all possible PBs method, majority rule method and the hybrid method were performed on a more comprehensive but redundant dataset of proteins. This dataset is the PDB100 dataset, i.e. the entire PDB filtered at 100% sequence identity by using the BLASTClust algorithm. The procedure adopted for preparing this dataset is similar to the one for preparing the PDB30 dataset described in the manuscript. Also in this case protein chains shorter than 100 residues were filtered out. This dataset comprises of 43,964 protein chains, of these 41, 110 were crystallographic structures, 2241 NMR structures and 613 were resolved by electron microscopy.

The absence of bias of kPRED predictions towards the method by which the protein structure was resolved was analyzed. In the case of crystallographic structures, we further checked if better resolved structure were favored by kPRED. This exercise was performed at different x-ray crystallographic resolution cut-off levels. In all the cases kPRED predictions showed no significant change in the distribution of accuracies. The histograms in Figures 25-27, show the distribution of prediction accuracies at 7 different sequence identity cut-off thresholds for crystallographic structures with resolution better than 1.8Å, NMR structures and electron microscopy structures (see also Table 14). These predictions were performed using the hybrid method without noise filtering and we see similar unbiased results for the all possible PBs and majority rule methods.

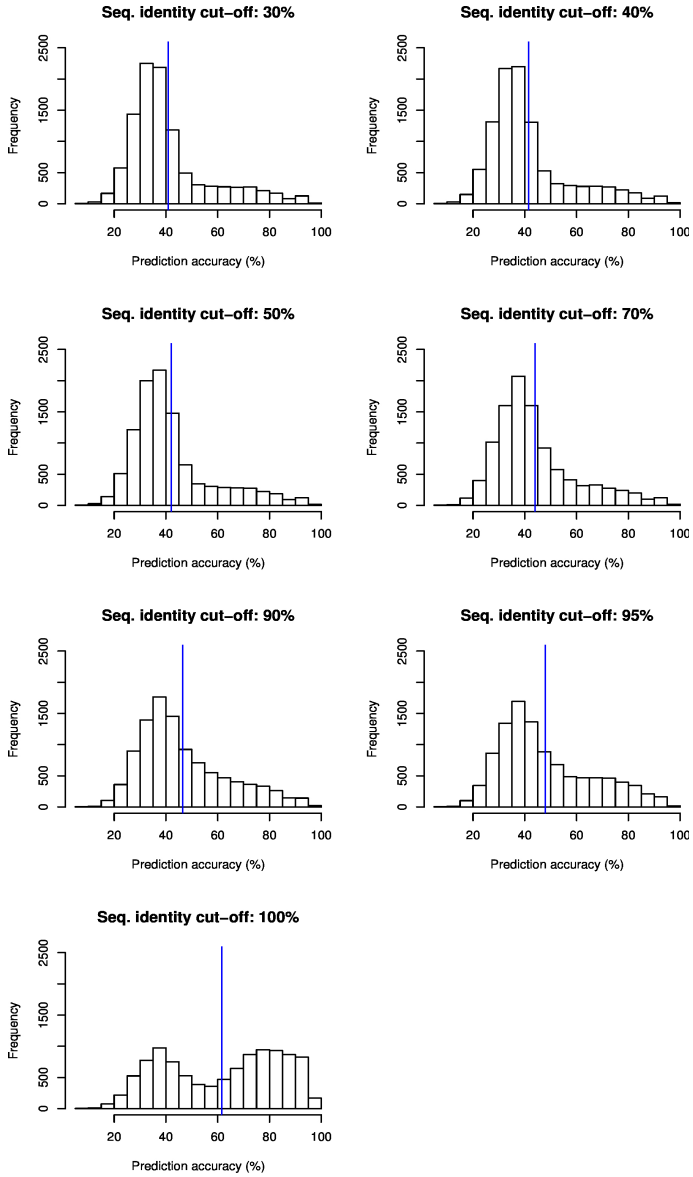


Figure 25. Distribution of prediction accuracy by hybrid method for the subset of the PDB100 dataset that have a crystallographic resolution >1.8Å

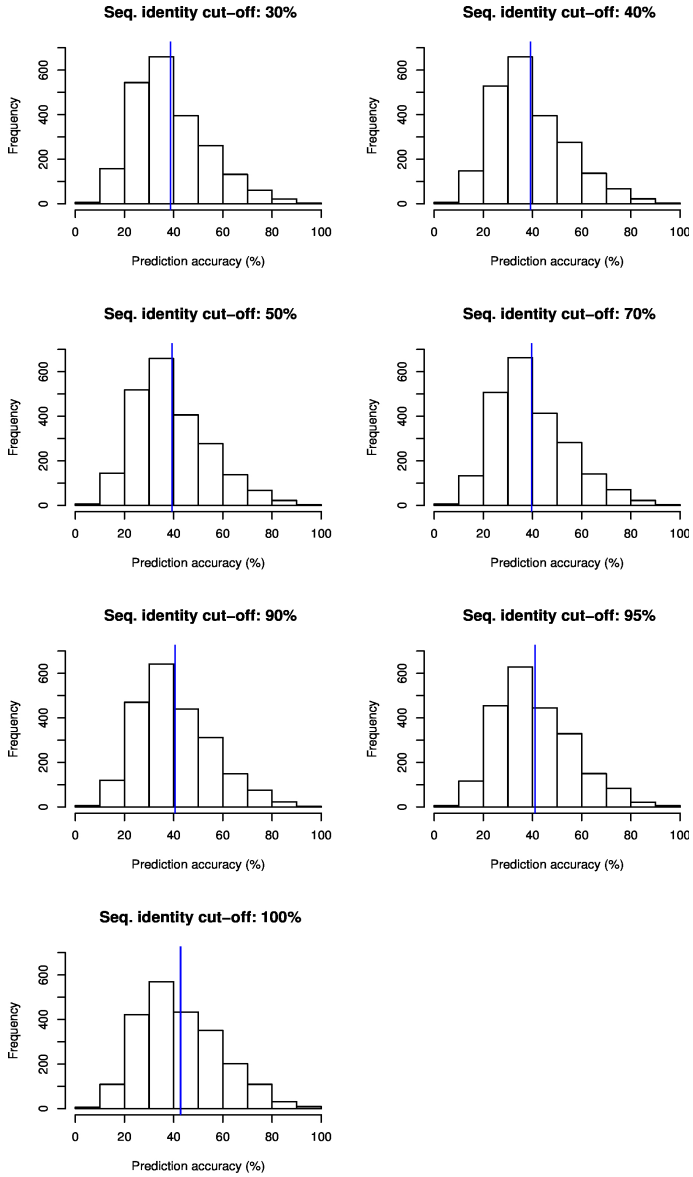


Figure 26. Distribution of prediction accuracy by hybrid method for the subset of the PDB100 dataset that have been solved by NMR.

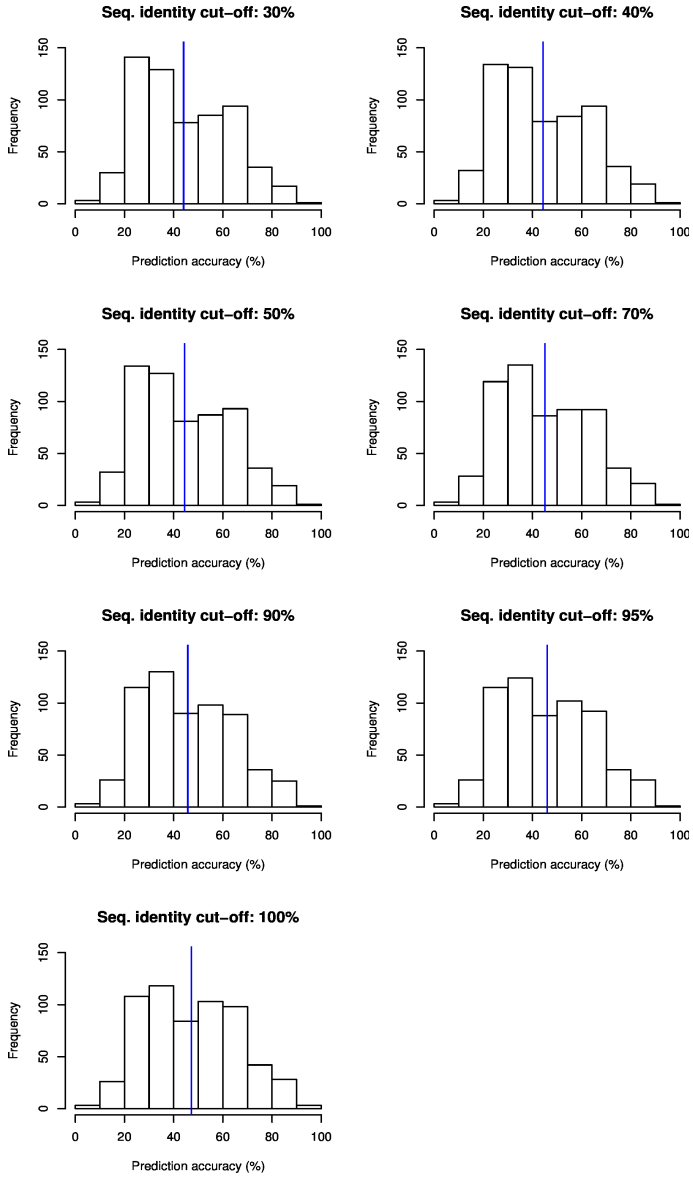


Figure 27. Distribution of prediction accuracy by hybrid method for the subset of the PDB100 dataset that have been solved by electron microscopy.

Table 14: Evaluation of the performance of kPRED for the PDB100 dataset.

Seq. identity (%)		30% seq. id.			100% seq. id.		
Total number of queries		12,223,015					
Method		**Correct**	**Wrong**	**no hits**	**Correct**	**Wrong**	**no hits**
Majority rule	Number of pentapeptides	4,863,025	7,346,688	13,302	6,832,476	5,382,113	8,426
	Prediction accuracy	39.79%	60.11%	0.11%	44.03%	55.90%	0.07%
Hybrid method (without noise filtering)	Number of pentapeptides	5,048,732	7,174,283	0	7,349,073	4,873,942	0
	Prediction accuracy	41.31%	58.69%	0.0%	60.12%	39.88%	0.0%
All possible PBs	Number of pentapeptides	8,743,039	3,466,674	13,302	10,418,743	1,795,846	8,426
	Prediction accuracy	71.53%	28.36%	0.11%	85.24%	14.69%	0.07%

kPRED predictions using the three methods were also performed for two Critical Assessment of Techniques for Protein Structure Prediction (CASP) datasets: CASP10 and CASP11. The CASP targets were downloaded from the CASP website[2]. Only the targets for which the actual structures were released publicly were considered for this analysis, so that we could later compare our predictions with the actual PB sequence. The CASP10 dataset comprised of 83 protein chains and the CASP11 had 58 protein chains. The true measure of the ability of a prediction method is its accuracy of predicting proteins that have very distant homologues. The CASP targets are ideal candidates for such analysis. The predictions performed for the CASP11 targets yields mean accuracies of 60.03%, 34.20% and 34.11% for the *all possible*, *majority rule* and *hybrid methods* respectively (Figure 28). The corresponding results for the CASP10 targets are summarized in the box plots of Figure 29. In all the boxplots, the bottom and top of the box represent the 1st and the 3rd quartiles respectively. The middle band represents the median, the lower whisker corresponds to *max(min(accuracy%), Q1-1.5*IQR)*, the upper whisker corresponds to *min(max(accuracy%), Q3+1.5*IQR)* where Q1 is the 1st quartile, Q3 is the 3rd quartile and IQR is the interquartile

2 - http://www.predictioncenter.org

range (Q3-Q1). What is remarkable for both the cases of the CASP datasets and the PDB100 dataset is the substantial gap between the *all possible PBs* accuracy and the *hybrid method* accuracy. This gap is the head room available for improvement of predictions by the kPRED methodology.

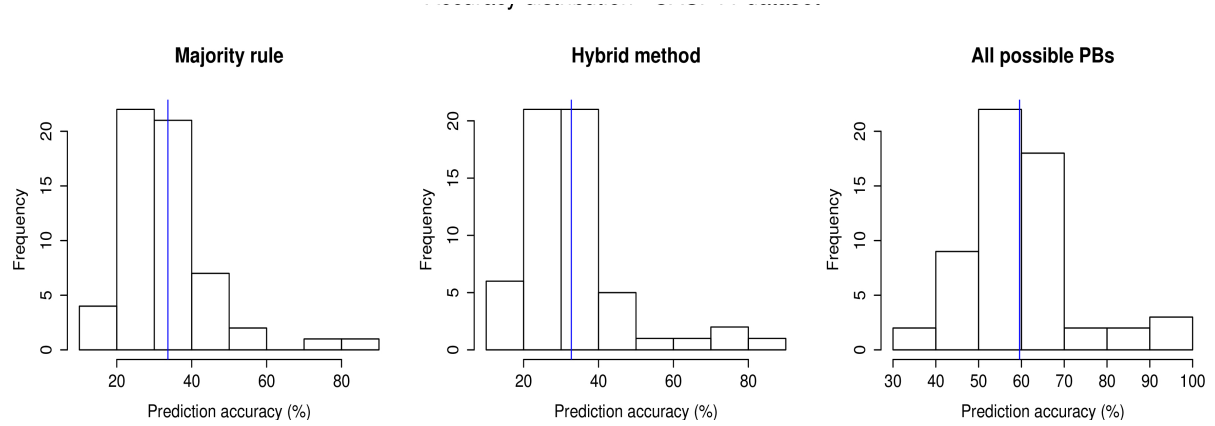


Figure 28. Distribution of prediction accuracy for the CASP11 dataset.

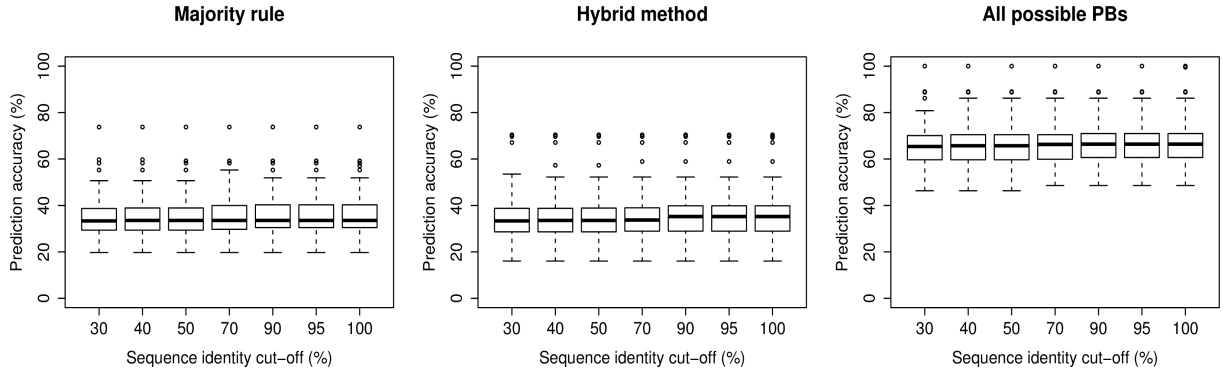


Figure 29. Distribution of prediction accuracy for the CASP10 dataset.

The above analysis were performed only with the variant of the *hybrid method without noise filtering* and not with *noise filtering* because the noise filtering methodology was developed at a later stage after the above analysis were performed. Moreover newer CASP11 targets did not appear in the BLASTClust dataset and the BLASTClust algorithm would have to be implemented within the kPRED prediction methodology in order to use noise filtering. This is one of the perspectives ahead for the kPRED methodology. The noise filtering would

have partially bridged the gap between the green line and the red line in Figure 30. But this effect would be less apparent for the CASP datasets because of the lack of homologues. In the histograms in Figures 25-27 we see a shift in the distribution towards higher accuracies when more sequence homologues are made accessible to the prediction algorithm but this is not the case for the histograms in Figure 31 for CASP10 dataset.

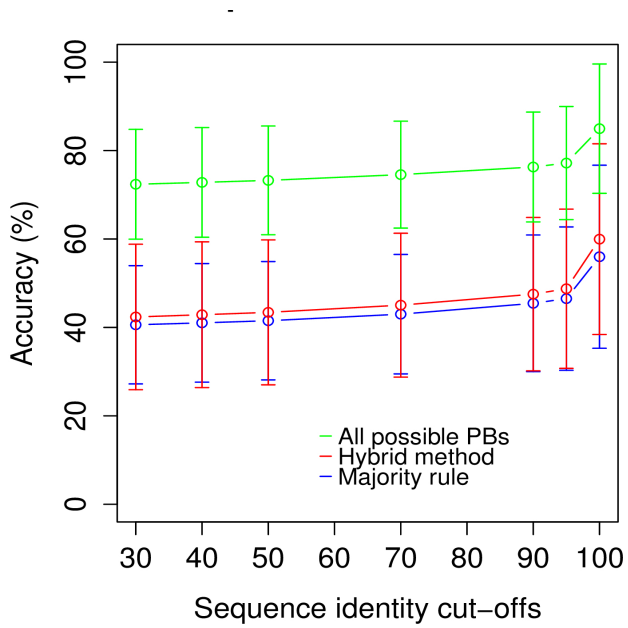


Figure 30. Comparison of prediction accuracy along with standard deviation for all possible PBs method, majority method and hybrid method without noise filtering.

Baker and colleagues produced 15 proteins purely using the principles of de novo protein design [37]. These proteins had no sequence homologues and we were interested to see how kPRED would perform on this dataset. Surprisingly the performance was very good with the minimum, maximum and average accuracies being 71.17%, 97.74% and 88.78% respectively for 20 chains from these 15 proteins. **This high accuracy despite the lack of even distant homologues is because the de novo designed proteins were helical repeats and kPRED was able to pick up the similarities at the local level despite the lack of it at the global level.**

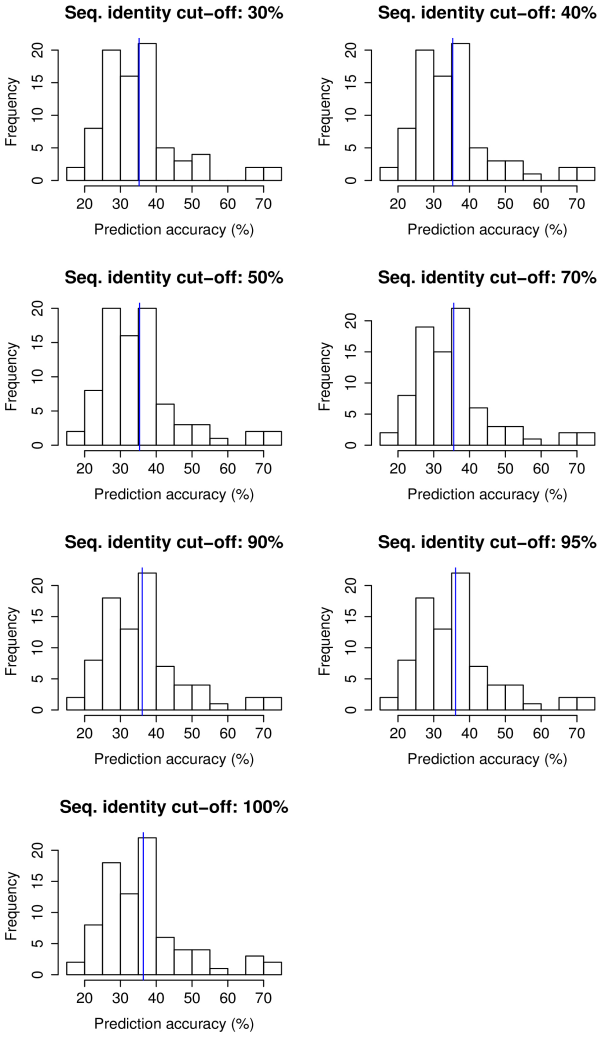


Figure 31. Histograms of distribution of prediction accuracy for the CASP10 dataset.

3.10. Accuracy gain due to hybrid method with noise filtering

In order to estimate the gain in accuracy due to the noise filtering process, additionally the schema illustrated in Figure 32 was followed. It is important to note the difference between this analysis and the one described in Figure 24b of the manuscript. The sectors marked in green are the ones available to PB-kPRED for prediction and the ones marked in red are not. The difference between consecutive values shown in Figure 32 is the gain in

accuracy achieved due to the noise filtering process at each sequence identity cut-off level. Corroborating our intuition, it can bee seen that the highest accuracy gain that was achieved by filtering out the pentapeptides from proteins that shared less than 30% sequence homology with the query. This amounts to 11.6%, which is substantially higher in comparison to less than 0.02 on average for the other cases.

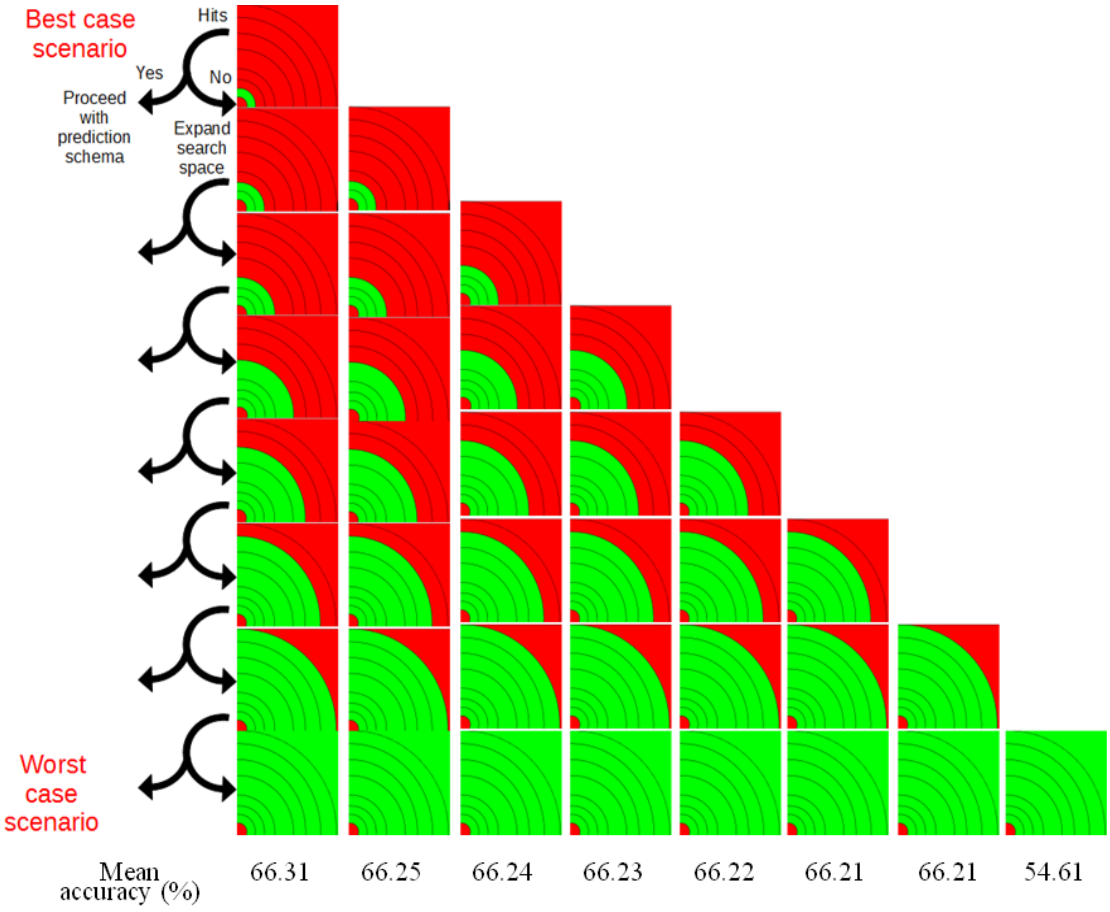


Figure 32. Exercise followed to estimate the accuracy gain at each sequence identity cutoff level for the PDB30 dataset.

3.11. Geometric scoring function and distribution of pentaPBs in non redundant datasets

In the section 3.6.6 of the manuscript, a probabilistic scoring function was described. This scoring function assigns a score to a sliding window of penta PBs in the PB sequence

based on the distribution of penta PBs in a non-redundant dataset. Here we describe an alternate scoring function that is based on the geometric compatibility of two successive PBs. Like in the previous case, we measure the ability of the scoring function to correlate with the prediction accuracy. A native unfolded protein tries to attain the lowest possible energy state while still respecting the structural constraints to attain its final 3D structure. The geometric scoring function tries to mimic these conditions by favoring geometrically compatible PB arrangements and penalizing less compatible PB arrangements while assigning a score.

The probabilistic scoring function requires a look-up table of penta PB distribution in a non-redundant dataset. The dataset that we chose to study the distribution of penta PBs is the PDB30 dataset this is prepared in a similar manner as the dataset described in the PB prediction schema. Higher quality X-ray crystallographic structures are preferred over NMR structures for choosing representatives of a BLASTClust performed on the entire PDB at 30% sequence identity. These representatives were obtained from RCSB PDB.

The PDB30 dataset as of July 2014 amounts to 22,653 protein chains. All these protein chains were converted into PB sequences. These 22,653 PB sequences contained a total of 4,366,040 penta PBs of which 40,130 were unique (see supporting Table S1, page 105). Theoretically the total number of penta PBs possible are 1.04 million (16^5). Out of this the 40,130 (3.8%) observed in the PDB30 dataset might seem to be a small fraction. But surprisingly, this is sufficient to effectively score PB sequences. The score correlates to the accuracy of prediction with a R^2 of 0.79 while tested on the PDB100 dataset. The distribution of pentaPBs in other non-redundant datasets i.e. PDB40, 50, 70, 90, 95, 100 and the entire PDB were also computed. Supporting Table S1 summaries these results.

Although there is a steady increase in the number of unique penta PBs, using higher sequence identity cut-off datasets introduces redundancy. Hence the PDB30 dataset with 40,130 unique penta PBs is used for the scoring. Each of these 40,130 geometrically favoured penta PBs are assigned a normalized odds score as per the formula: *Normalized pentaPB count = Observed penta PB count / Expected pentaPB count* (see Eqn. 1)

When the scoring function encounters a pentaPB that is not in the look-up table, then it penalizes it by introducing a penalty of -5. The choice of the value for the penalty was empirical. 58 CASP 11 targets were scored with penalties ranging from 0 to -11 in increments of 1. The pattern observed was that the higher the penalty, more the points disperse away from the diagonal, implying a better differentiation between the predicted and actual PB sequences.

It is important to note here that a particular pentaPB not being observed in the look-up table does not mean that it is geometrically disallowed, just that it is geometrically less favorable. The choice of a different representative for the clusters from BLASTClust could yield a slightly different penta PB distribution. But within this framework of variations, the overall distribution will be conserved. We have to settle for a trade-off between introducing redundancy into the dataset and including more unique pentaPBs. **The inability of having a comprehensive distribution of pentaPBs without introducing redundancy into the dataset is a limitation of this method of scoring PB sequences.** The total number of pentaPBs observed in the entire PDB i.e. 94,263 is still only 9% of the total number of theoretically possible pentaPBs, 1,048,576 (16^5). The extent of this limitation will diminish with the increase in the number of protein structures in the PDB.

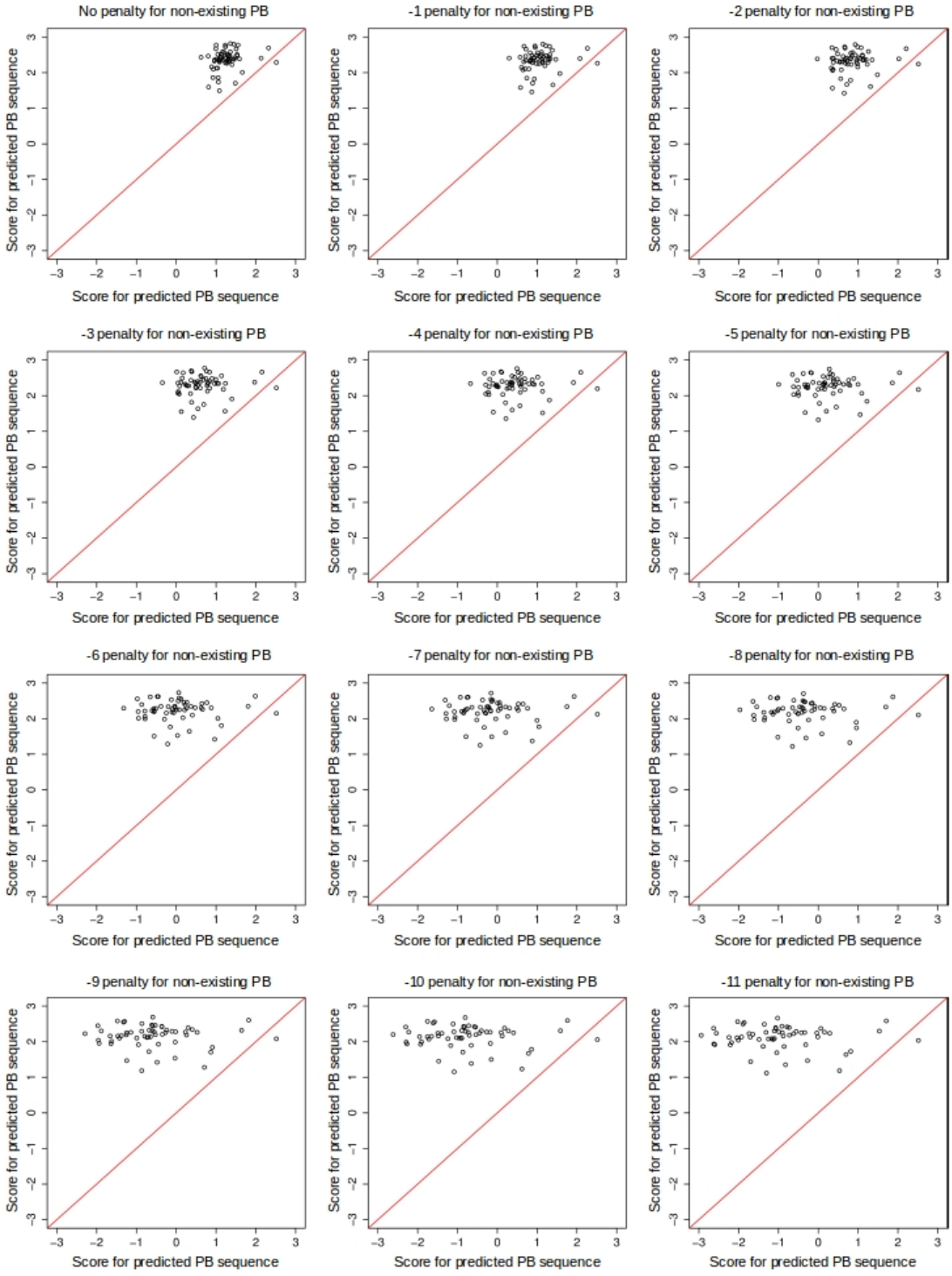


Figure 33. Plots for the predicted vs actual PB sequence scoring for 58 CASP 11 targets with penalties ranging from 0 to -11.

This inherent limitation of the probability based scoring function motivated us to develop a scoring function that was independent of the dataset used to prepare it. A PB is defined by a set of 8 phi-psi angles, the torsion angles. Two adjacent PBs share 6 of these torsion angles. The first 6 torsion angles are shared with the preceding PB and the last 6 are shared with the succeeding PB. The lower the value of the angular root mean square deviation (RMSDa) for these six torsion angles, the more these two PBs prefer to be next to each other. A cumulation of these RMSDa values over the length of a PB sequence is good estimate of its geometric compatibility.

Here we further implemented two different modes for calculating the geometric compatibility score. The first mode (geometric score I) consisted in averaging the RMSDA of the overlapping segments between 5 consecutive PBs followed by their summation over the length of the PB sequence and their subsequent length normalization. The second mode (geometric score II) consisted in a simple summation of the RMSDA of all overlapping segments between 5 consecutive PBs and their length normalization. The lower the score, the better the compatibility between the predicted local structures.

3.12. Connecting the dots: correlation between the scoring function and prediction accuracy

The above two modes for scoring a PB sequence were evaluated based on their ability to differentiate between a predicted PB sequence and an actual PB sequence, similar to the analysis described earlier (see section 3.6.6). In real case scenarios, the predicted PB sequence have a range of accuracies associated with them. So the better the correlation of the score with the accuracy, the better that scoring function is. This opens up an application of the scoring function to act as a guideline towards improving the accuracies of PB prediction. Here, the correlations (R^2) between the geometric scores I and II and the accuracy were of 0.43 (Figure

34a) and 0.64 (Figure 34c) respectively. Compared to the probabilistic scoring function that featured a R^2 of 0.79, they are found to be less performant. These values are surprising because the probabilistic scoring function is able to capture the correlation based on <5% of the theoretically possible pentaPBs but the geometric scoring function was less efficient to do so. Nonetheless, the geometric scoring function has the possibility of being more useful for the purpose of improving the prediction accuracy. In all the cases, we observe that the PB sequences predicted with high accuracy cluster around the diagonal (Figure 34b and 34d). This indicates that the predicted and actual score are approximately equal for the well predicted PB sequences.

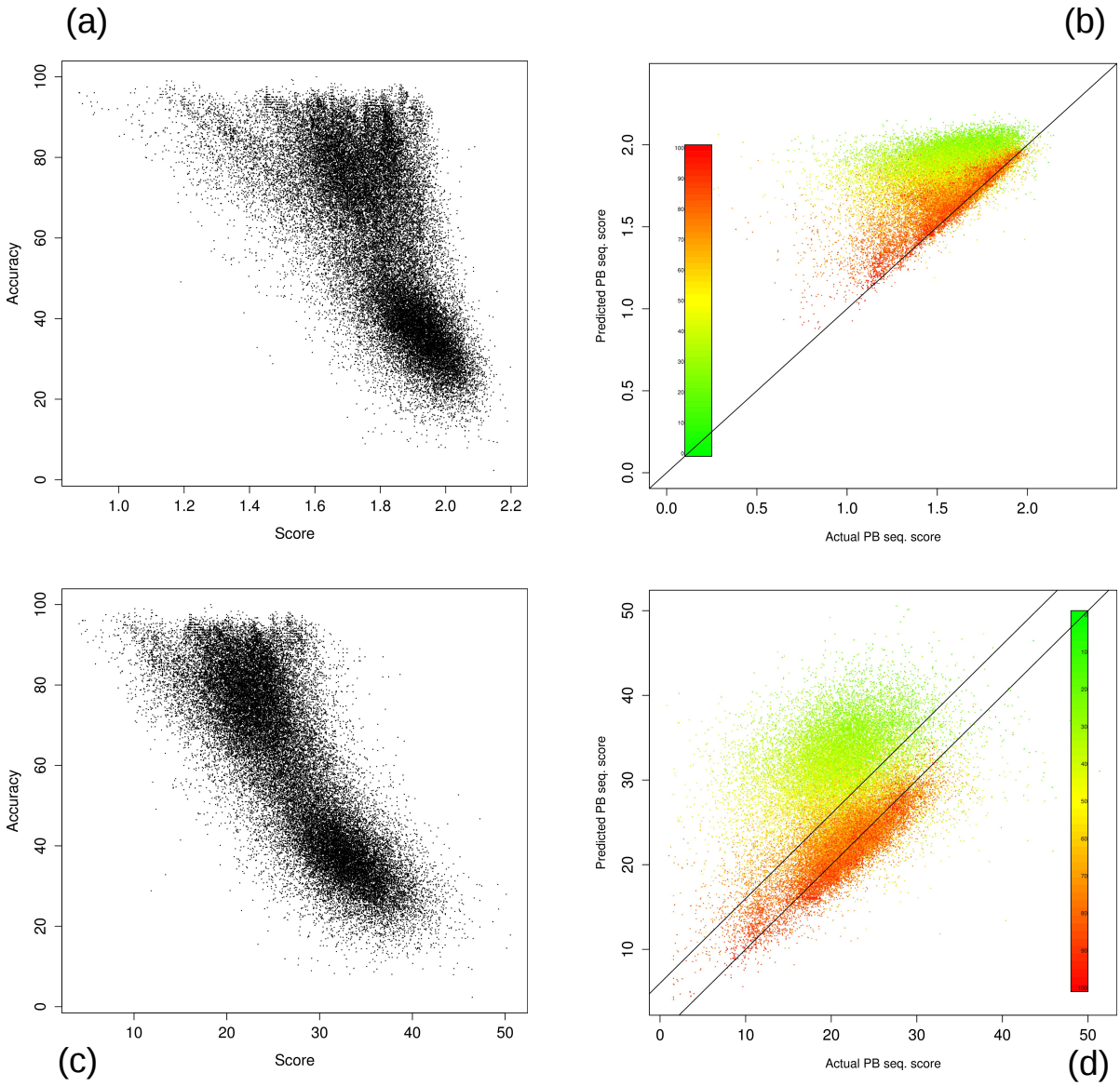


Figure 34. Evaluation of how the performance of the scoring functions I and II that test the geometric compatibility between consecutive predicted PBs.
(a) and (c) are accuracy vs score plots for the two geometric scoring functions I and II respectively. (b) and (d) are the plots of actual vs predicted scores for the two scoring functions respectively. Shown are the data for each of the 15,541 predicted PB sequences of the PDB30 dataset.

3.13. References

(specific for this chapter)

1. Rose PW, Prlić A, Bi C, Bluhm WF, Christie CH, Dutta S, et al. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res. 2015; 43:D345–D356.
2. Uniprot Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43:D204–D212.
3. Kabsch W, Sander C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. Proc Natl Acad Sci USA. 1984; 81:1075–1078.
4. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins.1989; 5:355–373.
5. Karchin R, Cline M, Karplus K. Evaluation of local structure alphabets based on residue burial. Proteins. 2004; 55:508–518.
6. Offmann B, Tyagi M, de Brevern AG. Local Protein Structures. Current Bioinformatics. 2007; 33:165–202.
7. Tyagi M, Bornot A, Offmann B, de Brevern AG. Protein short loop prediction in terms of a structural alphabet. Comput Biol Chem. 2009; 33:329–333.
8. Cheng H, Sen TZ, Jernigan RL, Kloczkowski A. Consensus data mining (CDM) protein secondary structure prediction server: Combining GOR V and fragment database mining (FDM). Bioinformatics. 2007; *23*(19): 2628–2630.
9. Bystroff C, Simons KT, Han KF, Baker D. Local sequence-structure correlations in proteins. Curr Opin Biotechnol. 1996; 7:417–421.

10. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol. 1997; 268:209–225.

11. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. J Mol Biol. 2002; 323:297–307.

12. Joseph AP, Agarwal G, Mahajan S, Gelly J-C, Swapna LS, Offmann B, et al. A short survey on protein blocks. Biophys Rev. 2010; 2:137–147.

13. de Brevern AG, Etchebest C, Hazout S. Bayesian Probabilistic Approach for Predicting Backbone. Proteins. 2000; 287:271–287.

14. Dudev M, Lim C. Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. BMC Bioinformatics. 2007; 8:106.

15. Wu CY, Chen YC, Lim C. A structural-alphabet-based strategy for finding structural motifs across protein families. Nucleic Acids Res. 2010; 38:e150.

16. Schneider B, Cerný J, Svozil D, Cech P, Gelly J-C, de Brevern AG. Bioinformatic analysis of the protein/DNA interface. Nucleic Acids Res. 2014; 42:3381–3394.

17. Tyagi M, de Brevern AG, Srinivasan N, Offmann B. Protein structure mining using a structural alphabet. Proteins. 2008; 71:920–937.

18. Joseph AP, Srinivasan N, de Brevern AG. Improvement of protein structure comparison using a structural alphabet. Biochimie. 2011; 93:1434–1445.

19. Mahajan S, de Brevern AG, Sanejouand Y-H, Srinivasan N, Offmann B. Use of a structural alphabet to find compatible folds for amino acid sequences. Protein Sci. 2015; 24:145–153.

20. Ghouzam Y, Postic G, de Brevern AG, Gelly J-C. Improving protein fold recognition with hybrid profiles combining sequence and structure evolution. Bioinformatics. 2015; 31:3782-9.

21. Zimmermann O, Ulrich HEH. LOCUSTRA: Accurate Prediction of Local Protein Structure Using a Two-Layer Support Vector Machine Approach. J Chem Inf Model. 2008; 48:1903–1908.

22. Suresh V, Parthasarathy S. SVM-PB-Pred: SVM based protein block prediction method using sequence profiles and secondary structures. Protein Pept Lett. 2014; 21(8):736–742.

23. Rangwala H, Kauffman C, Karypis G. svmPRAT: SVM-based protein residue annotation toolkit. BMC Bioinformatics. 2009; 10:439.

24. Dong Q, Wang X, Lin L, Wang Y. Analysis and prediction of protein local structure based on structure alphabets. Proteins. 2008; 72:163–172.

25. Karypis G. YASSPP: Better kernels and coding schemes lead to improvements in protein secondary structure prediction. Proteins. 2006; 64(3):575–586.

26. Garnier J, Gibrat JF, Robson B. GOR secondary structure prediction method version IV. Methods Enzym. 1996; 266:540–553.

27. Etchebest C, Benros C, Hazout S, de Brevern AG. A structural alphabet for local protein structures: Improved prediction methods. Proteins. 2005; 59:810–827.

28. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999; 292:195–202.

29. de Brevern AG, Benros C, Gautier R, Valadié H, Hazout S, Etchebest C. Local backbone structure prediction of proteins. In Silico Biol. 2004; 4:381–386.

30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410.

31. Tyagi M, Gowri VS, Srinivasan N, de Brevern AG De, Offmann B. A Substitution Matrix for Structural Alphabet Based on Structural Alignment of Homologous Proteins and its Applications. Proteins. 2006; 39:32–39.

32. Joseph AP, Srinivasan N, de Brevern AG. Improvement of protein structure comparison using a structural alphabet. Biochimie. 2011; 93:1434–1445.

33. Argos P. Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. Strategies for protein folding and a guide for site-directed mutagenesis. J Mol Biol. 1987; 197:331-48.

34. Suresh V, Liu L, Adjeroh D, Zhou X. RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Res. 2015; 43:1370–1379.

35. Luo J, Liu L, Venkateswaran S, Song Q, Zhou X. RPI-Bind: a structure-based method for accurate identification of RNA-protein binding sites. Sci Rep. 2017; 7:614.

36. Nguyen L, Dang X, Le T, Saethang T, Tran V, Ngo D, et al. Predicting Beta-Turns and Beta-Turn Types Using a Novel Over-Sampling Approach. Journal of Biomedical Science and Engineering. 2014; 7:927-940.

37. Webb B, Sali A. Comparative protein structure modeling using MODELLER. Curr Protoc Bioinforma 2014; 5.6.1-32.

38. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern 1982;43(1):59–69.

39. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22(12):2577–2637.

Chapter 4: PENTAdb- A database of pentapeptides and its implications in protein local structure prediction

4.1. Introduction

The structure adopted by pentapeptides in protein structures forms the basis for classifying them as Protein Blocks (PBs). PENTAdb is a database of pentapeptides enriched with other associated data like solvent accessibility, corresponding secondary structure etc. It was initially set up in 2006 by Manoj Tyagi (Tyagi, 2006) and updated in 2013 by Swapnil Mahajan and mainly comprised of data from SCOP domains (Murzin, Brenner, Hubbard, & Chothia, 1995). No update of the database was done since then. Hence this chapter is built on the strong foundations laid by my predecessors. It deals with the further development of PENTAdb with the objective of expand its coverage to include data from the entire PDB and also keeping it up-to-date in an automated manner. The relationship between PENTAdb and kPRED is also explored.

There have been several studies that document the structural diversity and flexibility of short peptide fragments from protein structures. Starting from the pioneering works by Kabsch and Sanders in 1984 who reported their observation that identical pentapeptides from unrelated proteins can adopt significantly contrasting conformations (W Kabsch & Sander, 1984). They further stated that co-operativity between 6 or more residues was necessary to efficiently predict protein local structure. This was the motivation behind us to develop the *hybrid method* of the kPRED algorithm. The *hybrid method* takes into consideration the structure of the preceding and succeeding pentapeptide while trying to predict the structure of a particular pentapeptide, (see Figure 23 from chapter 3). This provided us a modest gain in prediction accuracy, see Figure 30 from Chapter 3. Argos expanded the study of the structural

diversity of pentapeptides to a larger dataset and also did not limit the study to identical pentapeptides but to similar pentapeptides (singly substituted) as well (Argos, 1987). By studying the substitution rates in pentapeptides he gave certain guidelines for site-directed mutagenesis. Cohen et. al. conducted a similar study on hexapeptides (Cohen, Presnell, & Cohen, 1993). They noted that hexapeptides mostly adopted similar conformations when they were from proteins belonging to the same structural class (all-α, all-β, α/β or α+β). Also the relationship between the local environment and the conformation was studied in-terms of solvent accessibility.

The global take home message conveyed by these studies is that the local structure adopted by two identical short peptides is highly dependent on the sequence similarity between the proteins they form a part of. This inspired us to develop the *noise filtering scheme* for the kPRED algorithm, (see Figure 24 from chapter 3). The adoption of the noise filtering scheme brought about substantial increase in prediction accuracy, almost 12% on average, (see Table 11 from Chapter 3).

4.2. Expanding the scope of PENTAdb

4.2.1. Auto updating pipeline

We decided that the section of PENTAdb that corresponded to SCOP v1.75a would be retained in its static from. On the contrary, we worked out so that the PDB counterpart is maintained up to date in an automated fashion. The update cycle corresponds to the weekly updates of the RCSB PDB. The approximately 150-200 proteins newly added to the PDB every week and their corresponding DSSP secondary structure assignment files are downloaded from the FTP sites of RCSB PDB and DSSP respectively. The scripts to perform

the above parsing and processing tasks are executed consecutively using CRON job scheduling.

4.2.2. Challenges in expanding the scope of PENTAdb to be comprehensive.

The process of making PENTAdb comprehensive involved parsing all the 2,34,523 protein chains from more than 0.12 million PDB files. Although the PDB files follow a well documented format set by the wwPDB (wwPDB, 2008), the PDB file parser designed to populate the PENTAdb faced a fair share of idiosyncrasies. The general format is that the coordinates corresponding to the amino acids appear in the "ATOM" records section. The modified residues and non-standard amino acids coordinates appear in the "HETATM" section. In both cases the alternate location of atoms or residues are indicated before the residue name. The insertions are mentioned after the residue number. In the header section of the PDB file, the missing residues are listed under "REMARK 465" and the missing atoms under "REMARK 470". Specifically in our case if the missing atom is one of the protein backbone atoms, then the entire residue is not taken into consideration as PBs cannot be assigned if the backbone torsions are not known.

The one letter code for non standard amino acids line hydroxyproline, selenocysteine and pyrrolysine are mapped to their corresponding amino acids i.e. proline, cystein and lysine but there are several other non standard amino acids for which this mapping cannot be done. Next coming to a few specific cases, there are reverse insertions in the PDB entry 1HAG. PDB ID 1I3G_L has residues 27 and 28 missing but not mentioned in "REMARK 465". PDB ID 1IAO_B has random insertions and random numbering, residue numbering goes 1, 2, 323... PDB ID 1XMN has insertions in the missing residues. PDB ID 2Q50 is a structure solved by X-ray crystallography but it has 8 models. PDB ID 2V93_A has multiple atoms

with fractional occupancies (0.05, 0.95, 0.16). In some cases the summation of the occupancy exceeds 1.0 as for *e.g*. the ATOM SG in CYS211. There are also several cases of very large structures that exceed 62 chains in the structure, in these cases the corresponding mmCIF files have to be parsed or a collection of minimal PDB format files are available e.g. the ribosome structure with PDB ID 3J36. PDB ID 1H4X has 0.5 occupancy residues with no alternate locations. PDB ID 7PCK has 99 insertions in each chain and also insertions in missing residues. This is by no means a comprehensive list of complications encountered while parsing PDB files. Proteins are complex biomolecules, this complexity cascades into the PDB files and ultimately into parsing the PDB files.

4.2.3. Statistics on the contents of PENTAdb

The effect of expanding PENTAdb from just covering the pentapeptides from SCOP domains to a comprehensive coverage of all the protein chains in the PDB is detailed hereafter. A drastic increase in both the total count of pentapeptides and the number of distinct pentapeptides covered was observed. Currently, as of December 2016, there are 61.6 million pentapeptides in the PENTAdb database that come from 234,523 protein chains. This represents a 72% increase in the number of distinct pentapeptides in the database which contained data only from the ASTRAL100 dataset before the update. We see a 950% increase in the total count of pentapeptides in PENTAdb increasing from 5.9 million to 62.2 million. More importantly these values are continuously increasing with the inclusion of new pentapeptides during the updates of PENTAdb corresponding to the weekly updates of PDB. This maintains the structural information of PENTAdb as comprehensive as possible. Similar to filtering the pentapeptides from the SCOP domains based on the sequence identity (Mahajan, 2013), the pentapeptides from PDB are also filtered at 7 different sequence identity cut-offs (Figure 35a). The sequence filtering of the domains was based on the procedure by

the ASTRAL compendium but the BLASTClust algorithm was used in the case of entire protein chains.

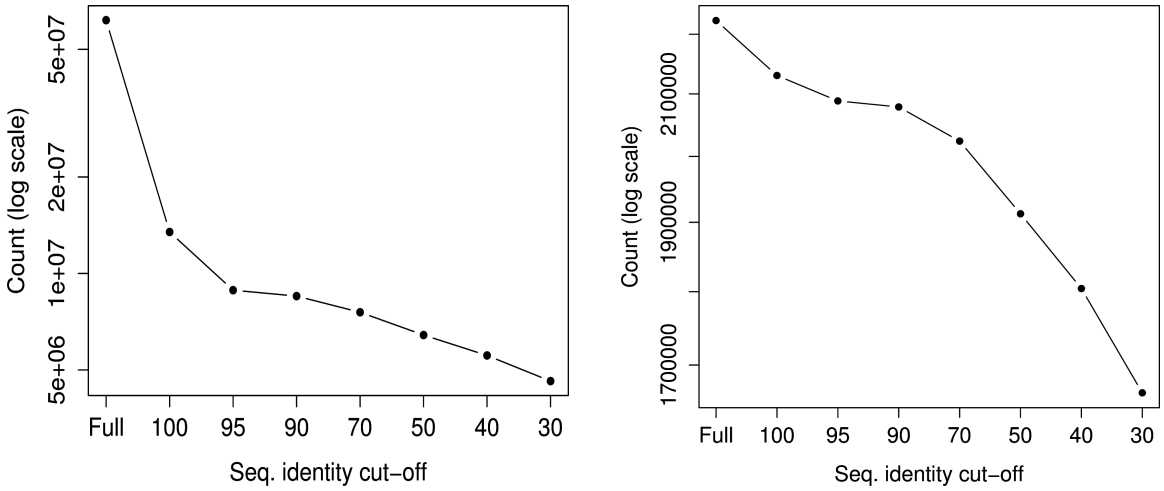


Figure 35. Contents of PENTAdb as a function of sequence identity cut-off.
(a) Distribution of total number of pentapeptides at different sequence identity cut-offs. This plot is indicative of the redundancy in the PDB at the pentapeptide level and (b) Distribution of distinct pentapeptides at different sequence identity cut-offs. The plateauing of the curve at higher sequence identity cut-offs indicates that closer homologues do not have as many new pentapeptides as unrelated proteins.

In terms of the coverage of the total theoretically possible 3.2 million (20^5) pentapeptides, this increased from 40% to 70%. PENTAdb now covers 2.12M (66.3%) pentapeptides (Figure 35b). Apart from the 1.07M (33.7%) pentapeptides that are never observed in the PDB, the remaining pentats are also not recruited equally in protein structures. For example, the pentapeptide AAFVW is observed only once in the entire PDB, it is from an actin related protein called Arp8 that has no sequence homologues in the PDB. On the other end of the spectrum, we have pentapeptides like NLLTQ that occurs 1115 times in 1079 different proteins. The quantile distribution of the counts of all the pentapeptides in PENTAdb is presented in Table 15. We can see that 50% of PENTAdb has a count of 12 or less and it would be safe to say that pentapeptides with counts 30 or more are frequently recruited in proteins. These pentapeptides form the top 25% of frequently recruited pentats.

Table 15: Quantile distribution of pentapeptide counts in different subsets of PENTAdb. Almost 50% of the pentapeptides in the non-redundant dataset are unique whereas this represents only 10% in the version of the PENTAdb built from the entire PDB.

Quantile	Count			
	Full	100	70	30
0%	1	1	1	1
10%	2	1	1	1
20%	3	1	1	1
30%	5	2	1	1
40%	8	3	2	1
50%	12	3	2	2
60%	18	5	3	2
70%	26	6	4	3
80%	39	9	5	4
90%	69	14	8	6
100%	1627	1447	1207	798

There are 396 pentapeptides that occur 1000 times or more in PENTAdb. But, it is important to note that pentats that form a part of protein tags like the poly His tag, S-tag, c-myc tag, FLAG tag etc. that are present in proteins as cloning artifacts (Rosano & Ceccarelli, 2014) have to be discounted from this list of frequently occurring pentapeptides. Some of these pentats include HHHHH, GPSSG, GSSGS, LEHHH etc.

4.3. Web-interface for the database

PENTAdb is a database for analyzing pentapeptides in protein structures. The database can be queried so as to obtain pentapeptides from protein chains from the entire PDB or just the SCOP domains (Murzin et al., 1995) . Further, this data can be filtered through 7 sequence identity cut-offs ranging from 30%-100%. PENTAdb accepts a pentapeptide as a query and provides the user with a range of information pertaining to the query. The different local conformations the pentapeptide can adopt is represented by either of the 16 PBs that it can correspond to and also their counts. The secondary structure assignment and solvent

accessibility values are displayed as computed by the DSSP program (Wolfgang Kabsch & Sander, 1983). In addition the eight torsion angles that define the structure of the pentapeptide, the angular root mean square deviation (RMSDA) of these values from the standard torsions of the PB and the B-factor information are also provided. All the outputs from PENTAdb are down-loadable as text files. The outputs from PENTAdb are cross linked to the RCSB PDB and SCOP databases. The location of the pentapeptide in the protein chain or SCOP domain can be visualized in a Jmol applet. PENTAdb is essentially a MySQL server in the back end and a combination of PERL CGI and HTML pages in the front end.

An advanced search option is included that allows the user to directly query the MySQL server at the backend of PENTAdb to obtain specific results. For every interrogation made by a user to the PENTAdb database, the corresponding MySQL query generated to fetch the results are also displayed. If the user wishes, they could further build upon this query to get customized results directly from the MySQL database from the advanced search option. The ability to use the regular expressions directly in the query makes this a powerful tool to study local structures in proteins.

4.4. Structural variability of identical pentapeptides

It is well established now that identical peptides can adopt significantly different structural conformations based on their occurrence in unrelated proteins. The natural continuation of this understanding is to characterize the structural diversity of these short peptides. Here we take advantage of the 16 state classification of local conformation of pentapeptides in terms of PBs. A similar study was previously undertaken on two non-redundant subsets of proteins namely *pdbselect* and *pisces25* datasets (Tyagi, 2006). Extending the study to an exhaustive dataset gives us a complete picture of the structural

diversity at the least and has the potential to reveal new insights into the sequence-structure relationship. Both the number of different PBs that a pentapeptide corresponds to and their counts are used as quantifiers. The structural diversity exhibited by pentapeptides from redundant and non-redundant subsets of PENTAdb are summarized in Table 16.

Table 16. Number of pentapeptides that map to 1,2,3,>=10 or all 16 PBs from PENTAdb filtered at different levels of redundancies.

The second row indicates the pentapeptides that map to a single PB and the values in the brackets are the pentapeptides that occur only once in the database.

	Full PDB	**PDB 100**	**PDB 70**	**PDB 30**
Null	1,078,479	1,122,836	1,217,033	1,566,156
1 (unique)	806,164 (161,226)	884,896 (496,454)	914,951 (642,336)	915,240 (675,507)
2	518,146	550,274	535,903	431,468
3	328,192	318,339	287,294	183,425
10 or more	9,080	1,363	521	105
All 16	73	18	10	5
Total	2,121,521	2,077,164	1,982,967	1,633,844

The extent of local structural flexibility is such that more than 9,000 pentapeptides can correspond to 10 or more PBs and 73 pentapeptides can adopt structure corresponding to all of the 16 different PBs, for an example see Table 17. It is intuitive to think that in order to have such high flexibility, the smaller amino acids will be over represented in these pentapeptides to avoid steric clashes with the side chain. Out of 365 amino acids from the 73 pentapeptides, glycine (76), serine (55) and alanine (32) constitute almost 45% of the residues. From this collection of 73 pentapeptides we take the example of “LGRRE” for further analyzing the structural diversity, see Table 18. Here we see a case where an identical pentapeptide adopts:

- Contrasting local conformations, PB *m* (typically central region of helices) and PB *d* (typically central region of β-strands) in unrelated proteins.

- Similar local conformation, PB *d* in unrelated proteins.

- Similar local conformation, PB *m* in unrelated proteins and importantly in this case these proteins belong to different structural classes.

- Dissimilar local conformations, PBs *m* and *k* in homologous proteins.

- Similar local conformation, PB *d* in homologous proteins (as expected).

- Similar local conformation, PB *m* in homologous proteins (as expected).

Table 17. Distribution of counts and average RMSDA when compared to the standard torsion angles for the pentapeptide "LGRRE".

This pentapeptide can adopt a local conformation corresponding to all the 16 PBs.

PB	LGRRE counts (%)	Average RMSDA
a	302 (68.9)	73.11
b	7 (1.5)	78.90
c	15 (3.4)	43.08
d	7 (1.5)	33.81
e	8 (1.8)	86.36
f	2 (0.4)	53.07
g	3 (0.6)	82.98
h	9 (2.0)	49.29
i	11 (2.5)	78.74
j	5 (1.1)	67.72
k	17 (3.8)	71.38
l	29 (6.6)	78.45
m	13 (2.9)	70.08
n	6 (1.3)	73.70
o	2 (0.4)	72.68
p	2 (0.4)	63.51

Table 18. Summary of 19 out of the 441 instances the penatpeptide "LGRRE" occurs in PENTAdb.

AA seq	PB	PDB ID	Pos.	Sec. str.	Relative solvent accessibility						Class	Protein	
					Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5	Average			
LGRRE	M	3ql9_A	84	TCHHH	50	18	154	174	45	88.20	α/β	Chromatin remodeling protein implicated in Thalassemia	Group 1
LGRRE	M	3qla_A	88	HHHHH	40	22	180	209	33	96.80	α/β		
LGRRE	M	3qlc_A	88	HHHHH	30	0	14	120	45	41.80	α/β		
LGRRE	M	3qlc_B	88	HCHHH	42	22	183	212	49	101.60	α/β		
LGRRE	M	3qln_A	82	HCHHH	49	25	209	206	41	106.00	α/β		
LGRRE	M	4w5a_A	88	HCHHH	39	26	175	215	45	100.00	α/β		
LGRRE	M	4w5a_E	81	HHHHH	15	9	147	74	49	58.80	α/β		
LGRRE	K	3qla_D	80	HCHHH	44	25	169	212	52	100.4	α/β		
LGRRE	K	3qln_B	81	HCHHH	49	26	188	206	52	104.2	α/β		Group 2
LGRRE	K	4w5a_B	88	TCHHH	25	0	17	104	43.00	37.8	α/β		
LGRRE	M	4huk_A	308	HHSCC	28	12	121	92	73	65.20	All-α membrane protein	Multi drug transport protein from *Neisseria gonorrhoeae*	Group 3
LGRRE	M	4hul_A	308	HHSSC	22	13	109	102	68	62.80	All-α membrane protein		
LGRRE	M	4hum_A	308	HHSCC	25	20	121	84	71	64.20	All-α membrane protein		
LGRRE	M	4hun_A	308	HHSCC	24	12	144	92	54	65.20	All-α membrane protein		
LGRRE	D	1al7_A	161	CCCCH	111	15	116	18	60.00	64.00	α/β	Glycolate oxidase	Group 4
LGRRE	D	1gyl_A	161	CCCBS	86	10	147	32	74.00	69.80	α/β		
LGRRE	D	1gyl_B	161	CCCCT	91	14	112	30	110.00	71.40	α/β		
LGRRE	D	5e7q_A	392	EEEEC	0	0	14	21	22.00	11.40	α/β	Acyl-CoA synthetase	Group 5
LGRRE	D	5e7q_B	392	EEEEC	0	0	15	22	24.00	12.20	α/β		

the local conformation corresponding to the PB *a* but its occurrences as the other 15 PBs are not negligible. In many of the cases, the average RMSDA is significantly higher than the average RMSDA for corresponding PBs in the entire PENTAdb (Figure 36). This implies that the pentapeptide "LGRRE" adopts conformations more deviant from the standard torsion than rest of the pentapeptide in PENTAdb. Figure 36 summarizes the structural plasticity observed within a PB class. Similar analysis have been performed previously for non-redundant datasets (Etchebest, Benros, Hazout, & De Brevern, 2005; Tyagi, 2006) but never for a comprehensive dataset. Intra PB class RMSDA distribution shows different patterns for different PBs, bimodal distributions for PBs *l, o, p* and roughly gaussian distribution for PBs *c, g, j*. The PBs corresponding to the regular secondary structures, PBs *m* and *d* though occurring in maximum counts show minimum intra PB class conformational diversity. The mean RMSDA across all the PBs in PENTAdb is 28.42 as shown in Figure 37 but is only 20.65 for PBs *m* and *d* alone.

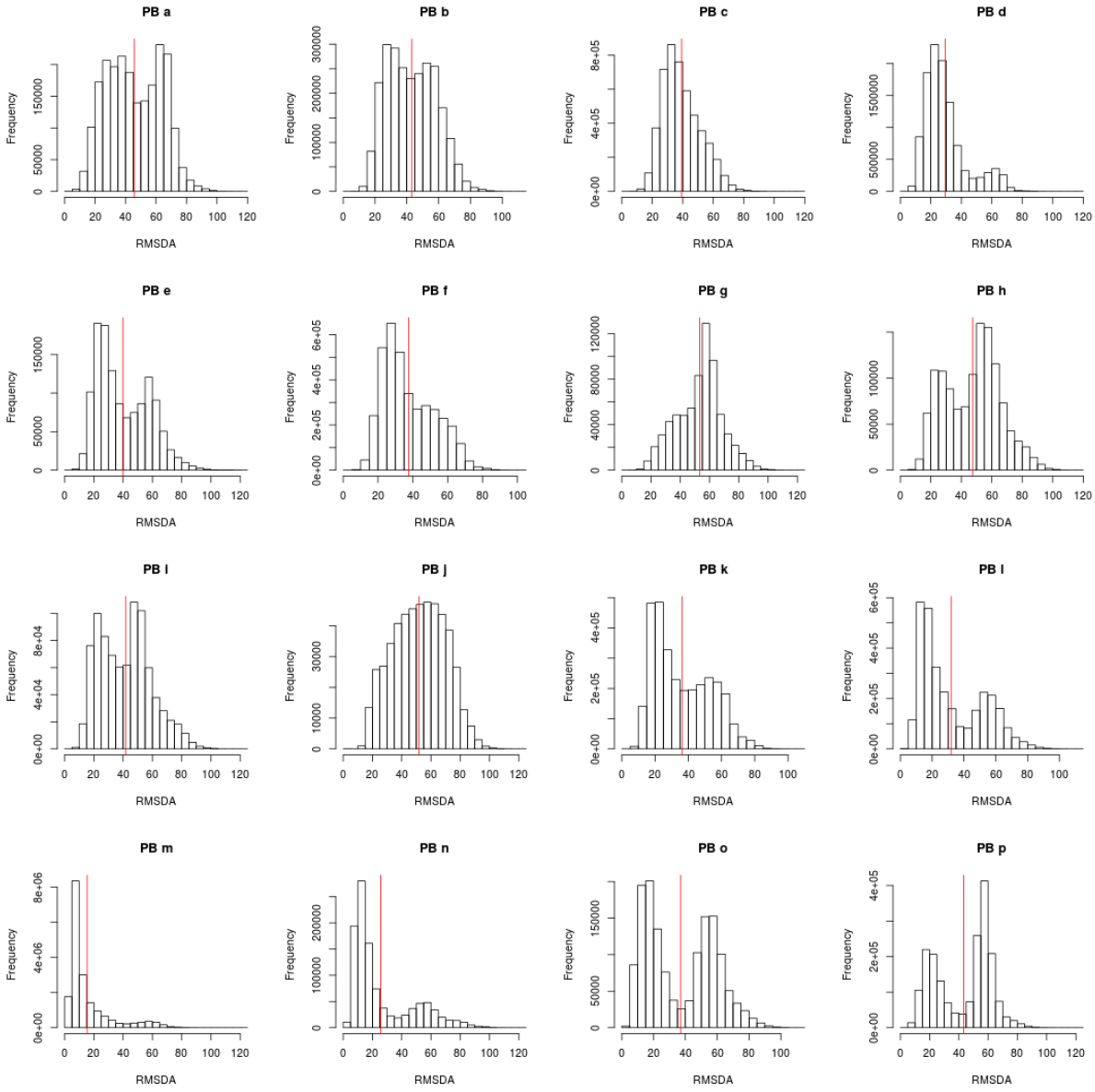


Figure 36. Intra PB class RMSDA distribution for all the 16 PBs from the entire PENTAdb. In all cases the mean is indicated by the red vertical line.

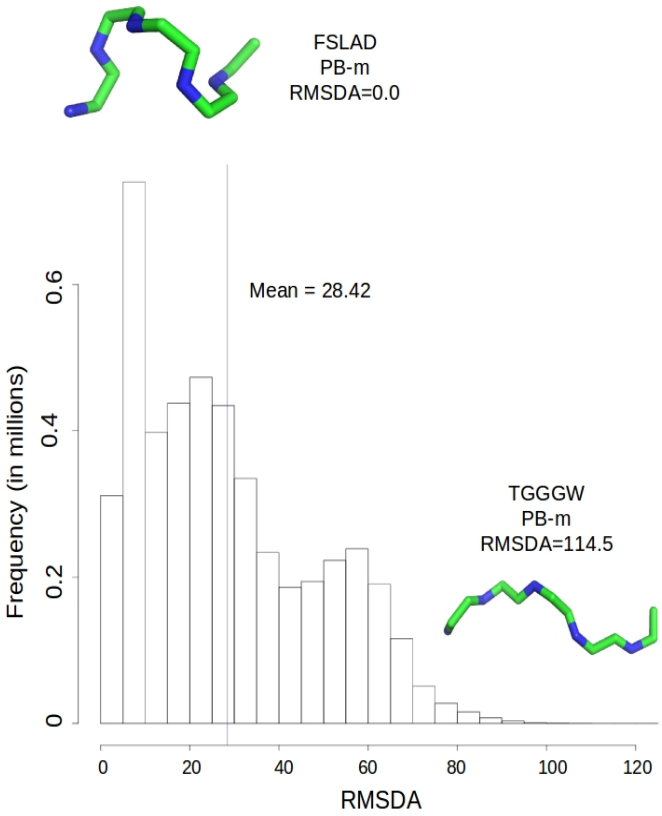


Figure 37. Global distribution of RMSDA for the 61.6 million pentapeptides in PENTAdb.

Coming back to the example of the pentapeptide "LGRRE", the local environment was studied in-terms of solvent accessibility (S. Hubbard and J. Thornton, (c) 1992-6) and polar contacts. The corresponding global picture was studied in-terms of HB plots (Bikadi, Demko, & Hazai, 2007) and contact maps (Vehlow et al., 2011). The 19 instances of the pentapeptide are studied as 5 groups represented in Table 18. Cohen et. al. have used solvent accessibility to study hexapeptide local environment (Cohen et al., 1993), here we use the solvent accessibility of pentapeptides as calculated by the NACCESS program (S. Hubbard and J. Thornton, (c) 1992-6). NACCESS is implemented within the DSSP program. The solvent accessibility of groups 1 and 3 are in Figure 38a, of group 2 in Figure 38b and groups

4 and 5 in Figure 38c. Each of the group has its characteristic solvent accessibility except groups 1 and 2.

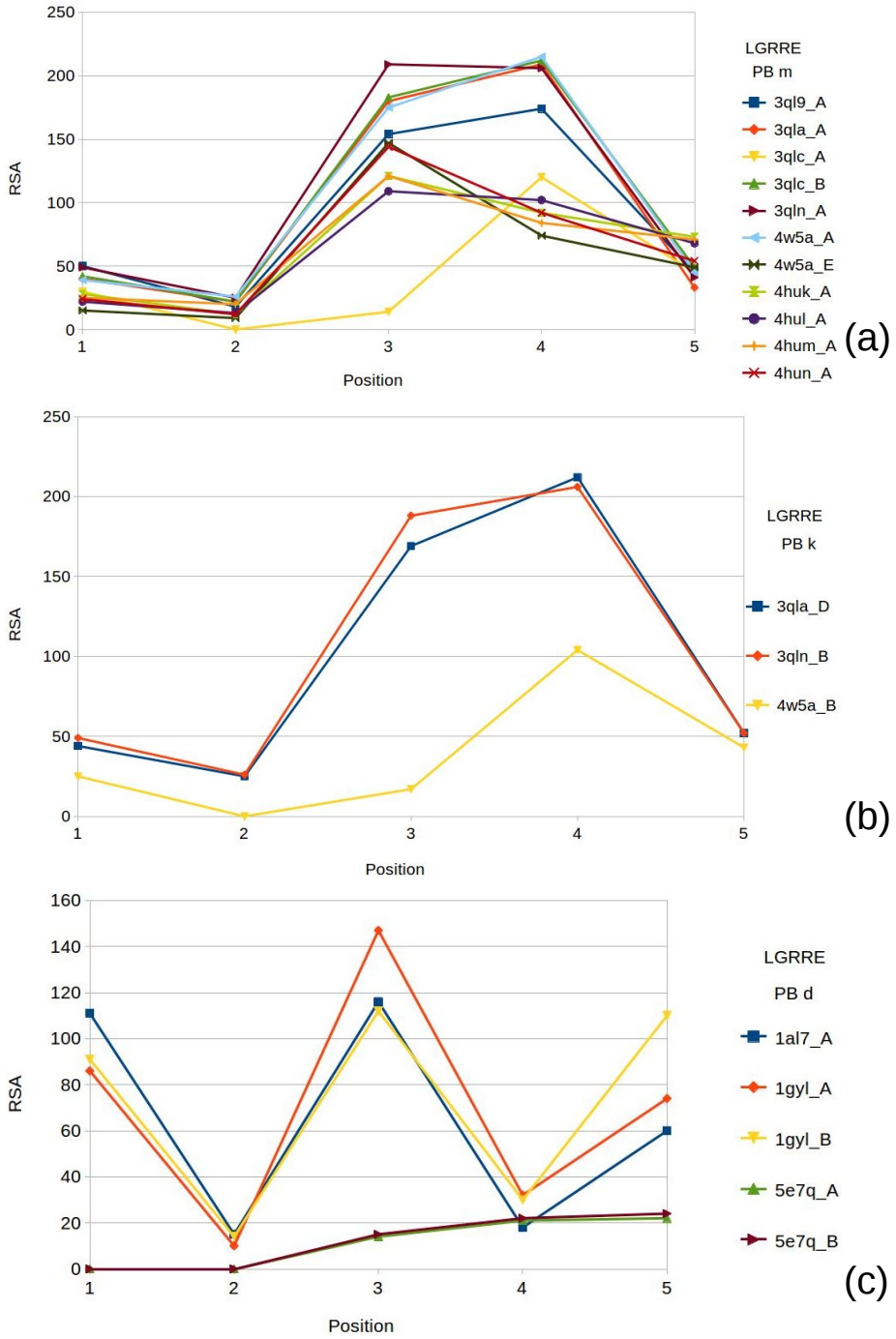


Figure 38. Solvent accessibility for the pentapeptide "LGRRE".
(a) Solvent accessibility for proteins belonging to groups 1 & 3 from Table 18. (b) Solvent accessibility for proteins belonging to group 2 from Table 18. (c) Solvent accessibility for proteins belonging to groups 4 & 5 from Table 18.

Despite corresponding to different PBs (*m* and *k*), they have similar solvent accessibility patterns because they belong to close homologues and also the structural difference between PB m and PB k is not very pronounced in this particular case (Figure 39a and 39b). Also in Figure 38c very different solvent accessibility patterns correspond to the same PB (PB *d*). Hence there seems to be many-to-one kind of relationship between solvent accessibility patterns and local structure for identical pentapeptides. It is a comparable case for polar contacts, similar polar contact patterns lead to similar local structures but this is not exclusive because the same local conformation can be achieved with different polar contacts. In Figure 39i and 39j, despite different polar contact patterns, in both cases the local conformation corresponds to PB *d*.

In order to visualize the non-local interactions that might be influencing the local structure of a pentapeptide, we utilize the HB plot program[3] to make a plot of the hydrogen bonding network in the protein. The HB plot program is in turn based on the HBPlus program (McDonald & Thornton, 1994). The HB plot differentiates between main chain-main chain, main chain-side chain and side chain-side chain hydrogen bonds. They are also classified as short, intermediate and long based on distance. A contact map is a more generalized representation of the physical proximity of the atoms in 3D space. It is a *N x N* matrix where *N* is the number of residues in the protein. Two residues are said to be in "contact" if they lie within the user defined distance cut-off. Contact maps can be made for a specific set of atom types *e.g*. Cα, Cβ, all atoms etc. The open source CMView software (Vehlow et al., 2011) was used for generating contact maps. Further the CMView software allows for comparison of contact maps of two proteins by differentiating the common contact points and the ones unique to each of the protein.

3 - http://virtuadrug.com/products/hb-plot/index.html

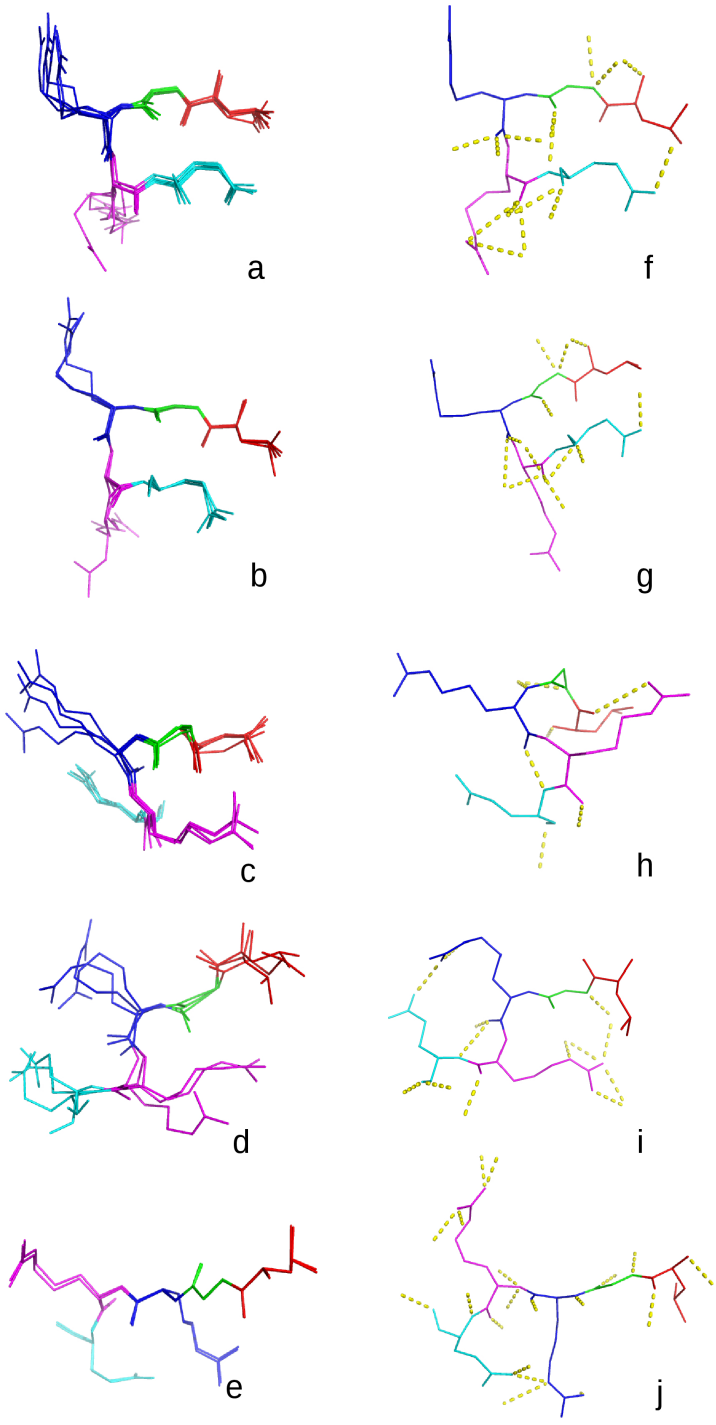


Figure 39. Superposition and polar contacts corresponding to the alternate conformations adopted by the pentapeptide "LGRRE".
(a-e) Superposition of the structures corresponding to the pentapeptide "LGRRE" from groups 1-5 in Table 18. (f-j) Polar contacts for each of the first members of groups 1-5 from Table 18.

Thus the local structure of a pentapeptide can be thought of as the result of a tug-of-war between the local and non-local interactions acting on the pentapeptide. Pentapeptides from unrelated proteins will correspond to the same PB if the effect of local interactions overpower non-local interactions (assuming dissimilar non-local interactions in unrelated proteins) and vice-versa. Complimenting the local structure information contained in PENTAdb with information about potential non-local interactions can help in decreasing the sequence-structure ambiguity at the pentapeptide level. Servers that are dedicated for contact prediction like PSICOV (Jones, Buchan, Cozzetto, & Pontil, 2012) and CCMpred (Seemayer, Gruber, & Soding, 2014) can improve PB prediction accuracies if used in conjuncture.

4.5. A quantitative measure of the structural variability of pentapeptides

As shown above, PBs have been used to describe the conformation adopted by a particular pentapeptide in a particular protein. But to summarize the different conformations it can adopt across multitudes of proteins, a metric called the structural variability index (SVI) was previously developed (Mahajan, 2013). This index is an approximation of the Shannon's entropic index (Shannon and Weaver 1949) or Simpson's diversity index (D, see equation below) used in population genetics (Simpson 1949):

$$D = \sum_{i=1}^{S} p_i^2$$

where, S is the total number of species and p_i is the fraction of all organisms which belong to the ith species.

The formula for calculating SVI is recalled here:

$$SVI = 1 - \sum_{i=1}^{k} \frac{n_i(n_i - 1)}{n(n-1)}$$

where, k is the number of different PBs i.e. 16, n_i is its occurrence in PB class i and n is the total number of times this pentapeptide is found in the database.

SVI summaries the structural variability of the pentapeptide as a single value between 0 (least variable) and 1 (maximally variable). The calculation of SVI strongly depends on the dataset of pentapeptides used. We calculate SVI on the database of pentapeptides derived from the entire PDB without any redundancy filtering to take advantage of all the structural knowledge contained in the PDB. Unique pentapeptides (single occurrence in the entire database) will result in an undefined SVI value, these SVI values have been approximated to 0.

Apart from displaying the SVI values for every query to PENTAdb, the SVI values for all the consecutive pentapeptides of a protein sequence are made available to the user. Provisions are made to visualize these values in the form of a graph and also make bulk queries of up to 1000 sequences.

4.6. Effect of the contents of PENTAdb on PB prediction using the kPRED algorithm

As shown in the previous chapter, we have developed an information or knowledge based method for protein local structure prediction. The method proposes a putative local structure as one of the 16 PBs for each of the pentapeptide represented in a protein sequence. kPRED makes this prediction based on the entire collection of pentapeptides in PENTAdb and also takes into consideration the local environment in terms of the preceding and succeeding

pentapeptide. As shown previously, the effect of hiding information from homologous proteins to the kPRED algorithm can substantially reduce the prediction accuracy (Table 11).

We were also interested to see how the SVI of a pentapeptide related to the ability of kPRED to correctly predict the PB. The PB prediction accuracy for a collection of 15,547 non-redundant proteins was plotted against their mean SVI values. We expected a trend of inverse proportionality, higher the structural variability lower the accuracy of prediction. Even for a subset of this dataset for which we are able to predict with an accuracy of 60% or higher, we see no explicit correlation between the mean SVI and prediction accuracy. The correlation coefficient is -0.503. Since the mean SVI value of the entire protein is too approximate to correctly gauge predictability, we also plot the percentage of pentapeptides with SVI greater than 0.8 against the PB prediction accuracy. From Figure 40 we can clearly see the trend that more higher is the percentage of pentapeptides with SVI >=0.8, lower is the prediction accuracy.

Further the dataset was split into two subsets, one with mean SVI values above the median SVI value of 0.45 (average accuracy 37.60%) and the subset having lower than 0.45 SVI (average accuracy 45.53%). By performing a Mann-Whitney-Wilcoxon test on them, we can reject the null hypothesis (H_0) that both the subsets follow the same distribution with a p-value < 2.2e-16 at a confidence interval of 0.99.

Out of the 3.06 million pentapeptides for which the predictions were made, the mean SVI for the 1.25 million correctly predicted pentapeptides is 0.328 and 0.492 for the 1.8 million wrong predictions. Again with the help of the Mann-Whitney-Wilcoxon test we can say with 99% accuracy that a pentapeptide with SVI of 0.328 or less has an above average chance of being correctly predicted.

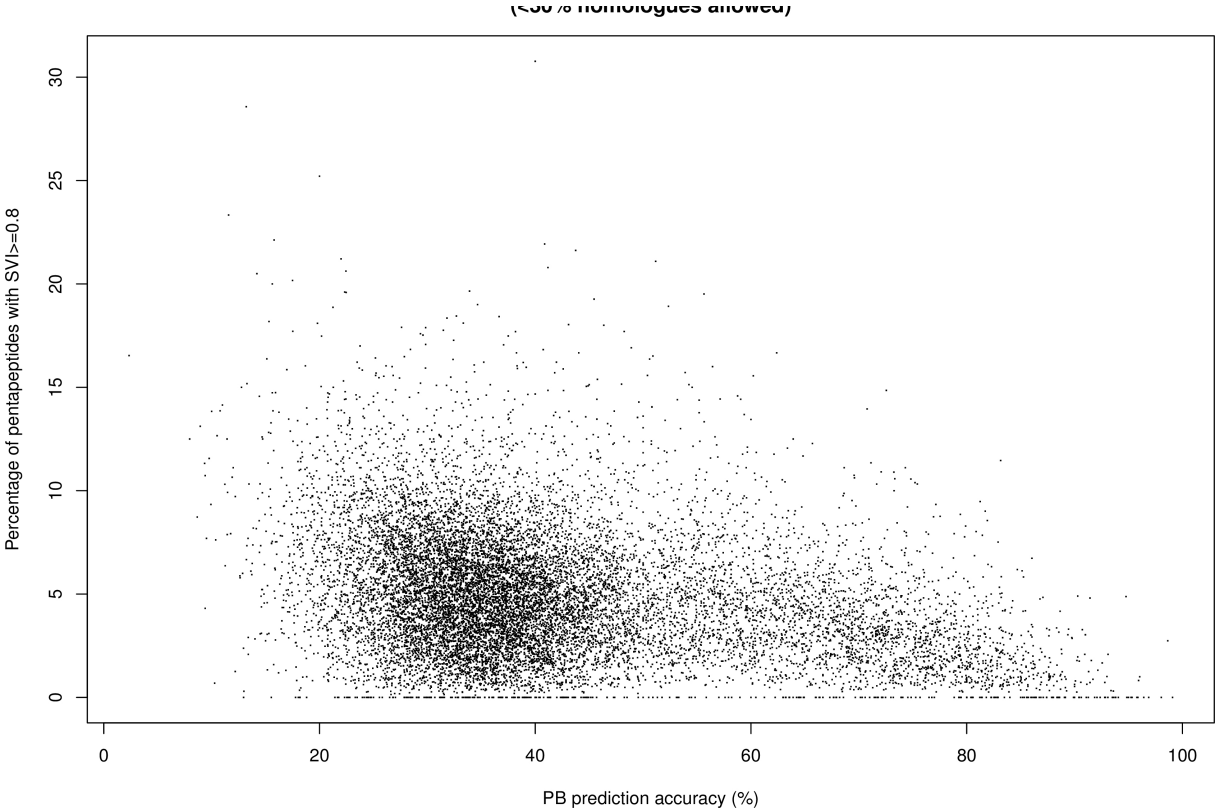


Figure 40. Relationship between protein local structure flexibility and predictability by kPRED algorithm.

Chapter 5: Comprehensive networks of protein structures

Lessons from large scale protein alignments using a structural alphabet

5.1. Abstract

Knowledge of the protein structure space is an invaluable resource for predicting the structure of novel proteins, understanding correlations between protein structure and function and for protein design. Many research groups have studied the interrelation between protein structural patterns, resulting in maps of the protein fold space. Most studies considering protein domains as nodes to build networks but some compare protein chains in their entirety as well. Nonetheless all studies filter their datasets at some level of sequence redundancy. We hypothesise that there could be important structural diversity lurking behind this sequence redundancy. Our quest with the current work was to chart out a comprehensive protein structure space and consider proteins to be redundant if and only if they shared the exact primary structure. We use a structural alphabet based structure comparison tool called PB-ALIGN previously developed by us to efficiently handle the large comprehensive dataset. The utility of this alignment data for functional annotation and identifying possible evolutionary relationships are discussed with examples. To aid the scientific community in discovering previously unidentified structural relationships, we plan to made available all the alignment data as a web based tool in the future.

5.2. Introduction

The diversity in the function and properties of proteins we see today is a direct consequence of the diversity of the living organisms they form a part of. Efforts to organise

and classify the rapidly growing wealth of protein structural information began with the introduction of the Structural Classification Of Proteins (SCOP) database (Murzin et al. 1995). SCOP is a manually curated hierarchical classification scheme, the hierarchy is represented by class, fold, superfamily, family and domain. SCOP remains an exception to the other automated and semi-automated methods of classification. The Class, Architecture, Topology, Homologous superfamily (CATH) (Orengo et al. 1997) and the more recently developed Evolutionary Classification of protein Domains (ECOD) (Cheng et al. 2014) are examples of such methods.

Decomposing proteins structures into their constituent domains has been the most fruitful method to classify proteins. Hence extensive efforts and a variety of programs have been dedicated towards this objective (Crippen 1978; Lesk and Rose 1981; L Holm and Sander 1994; Sowdhamini and Blundell 1995; Islam, Luo, and Sternberg 1995; Wernisch, Hunting, and Wodak 1999; Xu, Xu, and Gabow 2000; Alexandrov and Shindyalov 2003; Vinayagam et al. 2003; Redfern et al. 2007; Zhou, Xue, and Zhou 2007; Siddiqui and Barton 2008; Reddy et al. 2008). Majority of these methods identify domains based on the criteria that there are more inter-residue interactions than intra-residue interactions among the residues that constitute the domain. Few consensus based methods have also been developed (Jones et al. 1998; Alden, Veretnik, and Bourne 2010) but achieving a consensus level greater than the 80% mark between the methods has not been possible. Major points of disagreement between the methods include defining domain boundaries, extent of domain decomposition and identifying discontinuous domains. Multiple definitions as to what constitutes a domain and the disagreement between the methods of domain assignments have direct implications on protein classification and how we study proteins.

There have been several studies focusing on mapping the protein fold space (Liisa Holm and Sander 1998; Dokholyan, Shakhnovich, and Shakhnovich 2002; Hou et al. 2005; Pascual-García et al. 2009; Alva et al. 2010a; Nepomnyachiy, Ben-Tal, and Kolodny 2014; Edwards and Deane 2015). All these studies rely on non-redundant subsets of the SCOP and CATH domains databases except for the studies by Holm et. al., Hou et. al. and Dokholyan et. al. Both Holm et. al. and Hou et. al. utilize the subset of PDB filtered at 25% sequence identity whereas Dokholyan et. al. use the domains from the FSSP database. These analyses also vary in the choice of the tool for structure comparison, they include- Dali, MAMMOTH, HHsearch, SSM, jFatcat, TM-align and ESA. In fact Edwards et. al use four different structural alignment methods to build structural bridges across the fold space and discuss the effect of the choice of alignment method on the outcome. These studies explore the continuous vs discontinuous nature of the fold space at stringent and lax thresholds of structural similarity. At stringent thresholds the domains corresponding to the SCOP classes all-α, all-β and α+β generally separate out as smaller clusters and the α/β domains remain connected, but at sufficiently lax thresholds all the domains can form one giant connected component. The objective of these studies have been to establish a structural similarity threshold that represents the maximum evolutionary relatedness between the proteins. But since different proteins evolve at different rates, these methods in turn establish a range of thresholds, that can be varied to suit the requirements. Alva et. al and Edwards et. al also use this alignment data to correlate between the connectedness of the domains and the age of the domains (Alva et al. 2010a; Edwards and Deane 2015).

The view of the protein universe that we obtain mainly depends on two criteria- the dataset used and the alignment tool used for making the comparisons. In the current work we have explored the protein structure space without any prior notion of domains. Moreover even

the developers of the SCOP database have made a drastic shift in the way domains are organised by moving from a hierarchical model (all versions upto 1.75) to a network based model (version 2.0). The network model better represents the interrelations between protein domains (Andreeva et al. 2014). Hence we have used the structures available in the PDB that have been separated into unique chains. Unlike the studies by Holm et. al. and Hou et. al., we do not restrict ourselves to only non-redundant subsets of the PDB based on sequence similarity because it has been shown that sequence related proteins can still have contrasting structures (Kosloff and Kolodny 2008). We did not want to miss out on these cases. This is the reason we chose to perform our analysis on two datasets, one which is non-redundant and another that is comprehensive in terms of sequence space coverage. The number of comparisons to be made for the comprehensive dataset is extremely large, hence we use an efficient and accurate structural alphabet based alignment program called PB-ALIGN, that has been developed by our group previously (M Tyagi et al. 2008).

Since the identification of domains from less than 20 proteins by Wetlaufer in 1973 (Wetlaufer 1973), the number of known protein structures has grown exponentially hence the identification of domains is not a straight forward activity. The ambiguity involved in defining a domain has been recently explored (Postic et al. 2017) this also a reason why domain assignment is a non-trivial activity (Holland et al. 2006). This was another motivation behind us choosing to build our structural similarity networks on entire protein chains rather than protein domains. The alignment program PB-ALIGN helps us perform large number of structural alignments at the speed and accuracy of sequence alignment because it combines abstracting protein structures as PBs with dynamic programming (M Tyagi et al. 2008). Further using RMSD as a measure of structural similarity could be potentially misleading in cases of rigid body motions, hence using the alignment scores from PB-ALIGN helped us fish

out true structural similarities beyond the hurdles of rigid body motions. Thus we hope this combination of a comprehensive dataset and uniquely suited alignment program will help the protein structural biology community have an exhaustive view of the structural similarity network. We wish to include a word of caution that the networks described in this work are mere structural similarity networks and not protein fold networks as described in similar works by other groups (Alva et al. 2010b; Nepomnyachiy, Ben-Tal, and Kolodny 2014; Edwards and Deane 2015).

5.3. Material and methods

5.3.1. Dataset

Two subsets of the PDB (Rose et al. 2015) filtered at 20% and 100% sequence identity cut-offs were used as the datasets, henceforth they are referred to as PDB20 and PDB100 respectively. The BLASTClust algorithm (Altschul et al. 1990) was used to obtain the PDB100 dataset. BLASTClust (Altschul et al. 1990) clusters the protein sequences based on a user given sequence identity threshold, 100% identity in our case. Then the best structures (better resolution and R-factor) available from each of these clusters are chosen as representatives and they form a part of the final dataset. The PDB20 dataset was obtained from the PISCES sequence culling server (Wang and Dunbrack 2003) with the culling criteria of a maximum resolution of 1.6 Å and maximum R-factor of 0.25. We have deliberately chosen two datasets with a stringent and a lax threshold in order to study the alignments in a strictly non-redundant collection of proteins and also to be exhaustive in terms of coverage and not miss out on any of the structural diversity observed in the PDB. A few structures such as the ones with just the Cα trace in their PDB files are removed from both the datasets as they cannot be used in our current analysis. After this filtering process we get a total of 2,352 proteins for the PDB20 dataset and 61,433 protein for the PDB100 dataset. All these proteins

are transformed into their respective PB sequences (detailed in the subsequent section) by comparing their torsion angles to the standard PB torsion angles.

5.3.2. Protein Blocks

Protein Blocks are a 16 letter structural alphabet that help us represent the three dimensional structure of a protein as a combination of repetitive structural patterns. These 16 repetitive structural patterns are represented by the alphabets *a* to *p*. They were identified by Pr. Serge Hazout's team by performing an unsupervised clustering (Kohonen 1982) of the backbone torsion angles of pentapeptides from a non-redundant collection of proteins (A G De Brevern, Etchebest, and Hazout 2000). Since the PBs are assigned to the central residue of the pentapeptide, the entire protein (except the first two and last two residues) can be approximated as a string of PBs by using a sliding window of 5 amino acids. PBs make this approximation with a minimal loss of structural features. This makes PBs an ideal tool to identify and describe structural variations in loops and also at the sub-secondary structure level.

Thus each of the 16 PBs are the backbone orientations adopted by pentapeptides which in turn is defined by 8 standard torsion angles. To approximate the protein structure as a string of PBs, every consecutive set of 8 torsion angles are compared with the standard torsion angles to find a match with least Root Mean Square Deviation on Angular values (RMSDA) (Schuchhardt et al. 1996), hence assigning a PB. PB assignment for the dataset described above was performed using an in-house developed PERL script, but the same can be done using a web-based tool called Protein Blocks Expert-Transform (PBE-T) (M. Tyagi et al. 2006). The robustness of PBs to approximate protein structures even after a three fold increase in the size of PDB was demonstrated by an insignificant change in the mean and

median RMSDA values for PB assignment by Etchebest et. al. (Catherine Etchebest et al. 2005). Number of advances have been made in the field of predicting PBs from the amino acid sequence and developing applications based on PBs (Joseph et al. 2010). A non-exhaustive list to illustrate the diversity of applications developed is: fold recognition (Mahajan et al. 2015; Ghouzam et al. 2015; Suresh, Ganesan, and Parthasarathy 2012), structure comparison (M Tyagi et al. 2008; Joseph, Srinivasan, and de Brevern 2011), predicting RNA-protein interaction (Suresh et al. 2015), effect of mutations on proteins (C. Etchebest et al. 2007), studying protein flexibility (Alexandre G de Brevern et al. 2012; Craveur et al. 2015), Magnesium binding motif discovery (Dudev and Lim 2007). PBs are one of the widely used structural alphabets, Karchin et. al. have evaluated the utility of different structural alphabets for fold recognition and protein alignment and find PBs to be highly informative (Karchin et al. 2003; Karchin, Cline, and Karplus 2004).

The 16 PBs have different affectations to Secondary Structure Elements (SSEs). PBs *d* and *m* typically represent the central regions of β-strands and helices respectively. PBs *a-c* and *e-f* correspond to N-terminal and C-terminal regions of β-strands and PBs *k-l* and *n-p* correspond to N-terminal and C-terminal regions of helices respectively. The remaining PBs *g-j* mainly form loops and turns. While these are general observations, PBs *m* and *d* can occasionally be found in loop regions and vice- versa loop region PBs can be found at the boundaries of regular SSEs.

5.3.3. PB-ALIGN

Once two protein structures are converted into their respective PB sequences, the exercise of structure comparison essentially reduces to a PB sequence alignment problem. In the current method, we leverage on the fact that we are able to perform structural alignments

at the speed and computational efficiency of sequence alignment. This gives us an opportunity to perform an all vs all structural alignment for large datasets that would otherwise be very computationally demanding.

The PB-ALIGN program (M Tyagi et al. 2008) is used to perform the alignments of the PB sequences. PB-ALIGN is an implementation of the dynamic programming algorithm specifically for PB sequences, it can perform both local as well as global alignments with a user defined value for gap penalty. An in- depth analysis of the application of the program in structure mining and the effect of gap penalty on the quality of the alignment was also done by Tyagi et. al. (M Tyagi et al. 2008). The PB substitution matrix required for dynamic programming was developed by observing the PB substitutions in aligned homologous proteins from the Phylogeny and ALIgnment of Homologous protein structures (PALI) (Balaji et al. 2001) database (Manoj Tyagi et al. 2006). Later the substitution matrix was improved by performing a normalization prior to incorporating the substitutions from the PALI database (Joseph, Srinivasan, and de Brevern 2011), this improved substitution matrix has been used in the current work. Also in the current work we have used PB-ALIGN only in the local alignment mode with a gap penalty of -5. This gap penalty achieves a good balance between identifying the correct fold and also maintaining the quality of the alignment (M Tyagi et al. 2008). We chose the minimum alignment length to be 60 PBs in length and alignments shorter than this were ignored as they tend to be biologically less relevant.

Moreover the PB-ALIGN program is run in a recursive manner. For example when protein A is aligned to protein B, we might get a region of protein A that is aligned to protein B. The rest of the non aligned regions of protein A are trimmed and re-aligned with protein B. This exercise helps us to capture all the information such as multiple domains, domain

duplication, discontinuous domains, and overlapping domains etc. that are observed in proteins due to molecular evolution. This trim and re-align process is continued iteratively till the post-trimming fragments are greater than 60 PBs in length. The PB-ALIGN program is run n^2 times and not $n(n-1)/2$ times because aligning sequence A to sequence B is not the same as aligning sequence B to sequence A when it is run recursively.

The total number of all vs all alignment for the 61,433 proteins dataset amounts to 3.8 billion recursive alignments to be made. This took approximately 2700 hours of computation, when parallelized across 30 CPU cores clocked at 2.3 GHz. An one vs all alignment takes approximately 3 minutes, but this time varies and depends on the length of the PB sequence.

5.3.4. Clusters and Networks

The normalized alignment score has been used as a clustering criteria to generate clusters of structurally related proteins. Clustering was performed at six different normalized alignment score cut-offs ranging from 1.0 to 2.0 in increments of 0.2. The *igraph* (http://igraph.org) R package was used to perform the clustering. If any two proteins share an alignment with an alignment score greater than or equal to the chosen cut-off, they are eligible to become members of the particular cluster. The results of clustering was analysed based on three criteria- (i) Number of clusters, (ii) Total number of proteins in clusters i.e. non-singletons and (iii) Size of the largest cluster. Figure 3 (a-c) are plots illustrating these results.

To visualize these clusters, networks were constructed using cytoscape (Shannon et al. 2003) at the same six score cut-off ranges. These networks depict the global inter relation between protein chains. In the network, protein chains are represented as nodes and two nodes are connected by an edge if they share regions of similarity greater than or equal to 60 PBs in length. The alignment length and score cut-offs can be adjusted to more stringent values to

obtain more sparsely populated networks. The length of the edges are inversely proportional to the alignment score, i.e. higher the alignment score between two protein chains, closer they are placed in the network.

5.3.5. PB-ALIGN score vs RMSD

It has been shown previously (M Tyagi et al. 2008) that performing structural superpositions that are guided by the PB-ALIGN output produce good results that are comparable to other structural alignment methods like YAKUSA (Carpentier, Brouillet, and Pothier 2005) and CE (Shindyalov and Bourne 1998). Here we wanted to establish a relationship between the distribution of alignment scores for alignments of non-related proteins vs their corresponding RMSD. A random collection of 30,000 alignments from the PDB20 dataset were chosen for this experiment. The amino acid sequence corresponding to these PB alignments were used as guideline for performing least squares fitting using the ProFit v3.1 software (McLachlan 1982). The co-ordinates of the aligned regions of the protein are extracted out from the entire PDB files to perform the structural superimposition.

5.3.6. Cross referencing with SCOP

The alignments of the PDB100 dataset were cross referenced with the domain assignments from SCOPe v2.05 (Fox, Brenner, and Chandonia 2014; Chandonia, Fox, and Brenner 2017). The domain descriptions were obtained from the parseable files downloaded from the SCOPe website. All the domains common between the reference and query protein at the fold level are listed out. The domains are searched only on the basis of presence of the alignment and without considering the overlap between the reference and query. This approximation is done to reduce the complexity of deciding whether the aligned region is in a domain or not and moreover it would involve setting an arbitrary length limit for the overlap.

Thus all the domains common between the two aligned proteins are listed irrespective of the aligned regions. There can arise two scenarios, case one- no domain annotation for the query despite the presence of alignment and case two, no domain annotation for the reference itself. Both these cases are marked differently while cross referencing with SCOPe.

5.4. Results

5.4.1. All vs all alignments

We get a total of 412.7 million aligned fragments from the 3.8 billion comparisons made for the PDB100 dataset. But this contains many redundant alignments due to recursively running PB-ALIGN. After filtering out these cases, we are left with 282.8 million non-redundant alignments. An example of the asymmetric nature of aligning using PB-ALIGN is provided in Figure 41. Thus the recursive version of the PB-ALIGN program picks up an additional 76.4 million alignments when compared to the non-recursive version. Correspondingly for the PDB20 dataset we get 0.6 million and 0.41 million redundant and non-redundant alignments respectively after applying a length cut-off of 60 PBs.

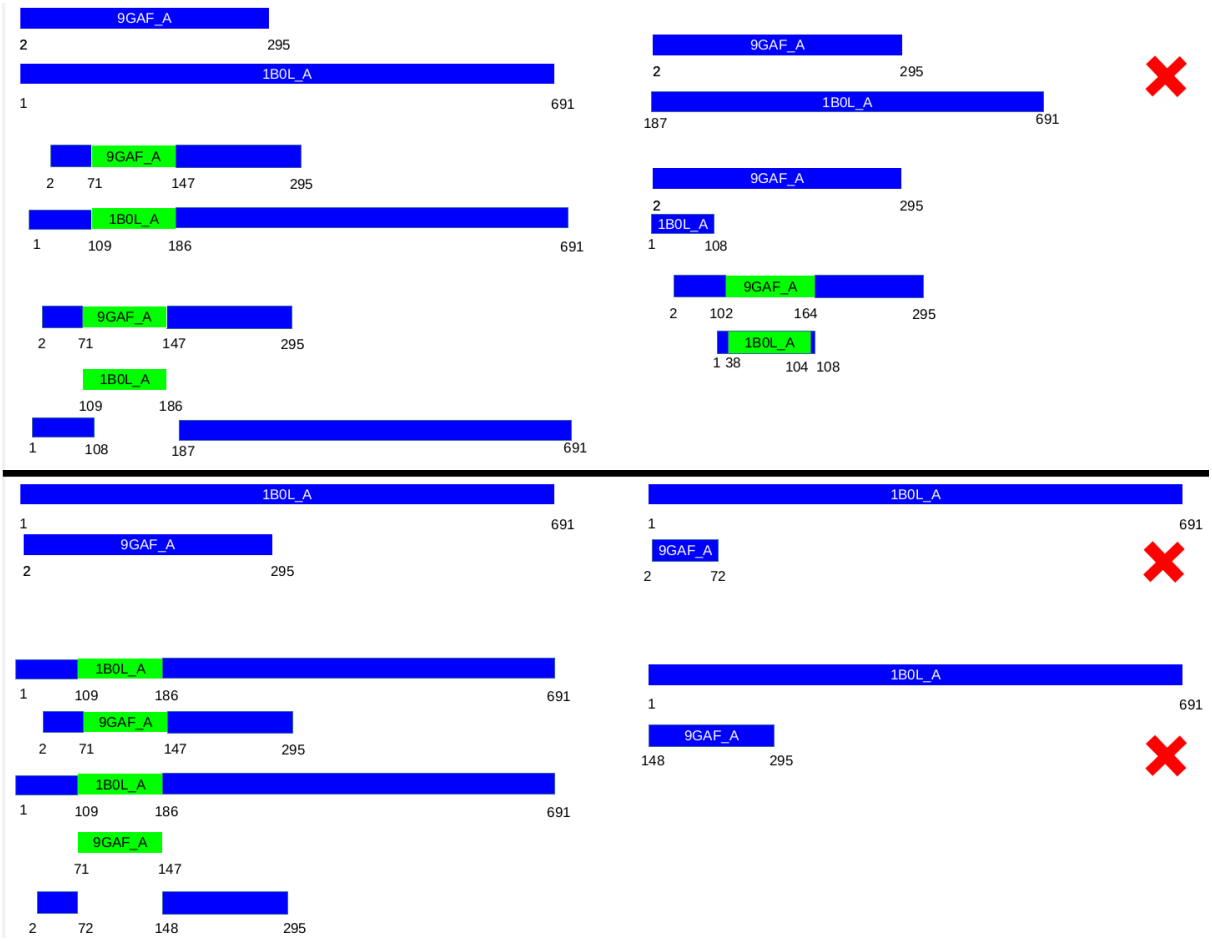


Figure 41. Illustration of the asymmetric nature of using PB-ALIGN in the recursive mode.

Aligning protein A to protein B is not always the same as aligning protein B to protein A. The entire protein sequence is in blue and the aligned regions are in green. In the first case we get two alignments but only one in the second case.

For each of these millions of alignments, we get the aligned PB sequences, the aligned region corresponding to each protein chain, a raw alignment score, the alignment length and a normalized alignment score. The raw alignment score is a summation of the substitution scores of the aligned PBs from the PB substitution matrix. The normalized alignment score is the raw score divided by the length of the alignment. Apart from this, a break up of the alignment length into identical PBs, similar PBs and gaps is also calculated. PBs with a positive substitution score in the substitution matrix are considered as similar PBs. A sample alignment is illustrated in Figure 42.

```
2VK8_A_vs_1Q6Z_A
4-539 vs 2-522

KLMMMMMMMMNOPABDCDDDDFKLMMMMMMMMMPGHIACEKPABFKLMMMMMMMMMMNOPACDDDDEHJ-LMMMMMMMMMMMMNOPACDDDDDDDDFK
KLMMMMMMMMNOPABDCDDDDFKL--MMMMGOPAFKBCCDDDFBFKLMMMMMMMMMMMNOPACDDDDEHJKLMMMMMMMMMMMMNOPACDDDDDDDDFK

LMMNOPACEHIAHLOPAKLMMMMMMMMCFBDCDDFBFKLMMMMMMMMMMMMMMNO-PACDDDDDFKLMMPCCDFKL---MMPCCDFBDCDDDFKLMMMMM
LMMNOP--GHIA-GH--KL-MMMP-CKBFBDCDDFBFKLMMMMMMMMMMMMMMPCCDDDDDDDDDFKLMMPCCDFKLMMMMMOCCDFBDC-DDFKL----M

MMMMMMMMMPMBDCDDDDFKLMMNOPMMMMMMMMMMNOPACDDDFKLGOJA-CFKBFKLGBJCEHJMPCFKLMMMMMM-PCBDCDEHIACDFKLNOGOIA
MMMMMMMMMPMBDCDDDDFKLMMNOPMMMMMMMMMMNOPACDDDEHIABDCBDFKBFKLGBJAD-F-B-FKLMMMMMGOIABDCDEHIACDDDFBDCEDD

CCFBFBDCDDDFKBCCE-HJBCCEHIAFKLMMMMMMMMMMMMMMPCCDDDDDDDDDDDDDDFKLCCDFKLMMMMMMMMPCGHIACDDDEDJK-LMMMMM
FBDGHIACDDDDFBFKLMMMPCGDCDDDDD-FKLMMMMMMMMPCCDFBDCDDDDDDDDDDDDDFBDCDFKLMMMMMMMMMCFKBCCDDDDFKLMMMMMMM

PCCD-EHIACDDDFKNOPGHJLMMMMMMMMMMMMMMOMMBCCDDDDDFKLMMMMMMMMMMMMMNOPACDCDDDDEHIAFKL---MMMNOP-LBFKLPCCD
PCCFBGHIBC-CDFK-OPGHKL---MMMMMMMMMM-MMBCCDDDDDFKLMMMMMMMMMMMMMNOPABDCDDDDEHIAFKLMMMMMMNOPGHIEHKACGH

DDFKLMMMMNOPABDCDDDDFBFKLMMMMMMPMKLMPCFBCCDDDDDDD
IAFKLMMMMNOPA--C-DDDFBFKLMMMMMMM-MMPC-BDCDDDDDDD

Raw score:404.343
Normalized alignment score:0.737
Alignment length:549
Identity:390
Similarity:24
Gaps:43
```

Figure 42. Illustration of a sample PB-ALIGN alignment and scores.
This is an exceptional example where the alignment score is relatively low despite high global similarity (RMSD 2.8Å). Local alignment acts as global alignment.

The distribution of the alignment scores for these alignments shows a slightly right tailed distribution, especially after the 1.0 score cut-off. There are 0.13 million alignments having a score greater than 2.0. The distribution of number of alignments at other score cut-offs are listed in Table 19. For the case of alignment lengths, only 17.5% of the 282.8 million alignments are greater than 100 PBs in length. We see much higher proportions of longer alignments at higher score cut-offs. It is 82.17%, 81.9% and 80.0% of alignments that are longer than 100 PBs at 1.4, 1.6 and 1.8 score cut-offs respectively. While these are much higher values than for the entire dataset, there is a gradual decrease after the 1.4 score cut-off. For the entire dataset there are 0.44 million alignments greater than 300 PBs in length.

Table 19: Distribution of number of alignments at different score cut-offs for the comprehensive and non-sequence redundant datasets.

Normalized alignment score cut-off	No. of alignments	
	PDB100 (million)	PDB20
Total	282.8	0.41 million
≥ 1.0	6.9	4573
≥ 1.2	2.44	433
≥ 1.4	1.44	62
≥ 1.6	0.76	11
≥ 1.8	0.31	3
≥ 2.0	0.068	1

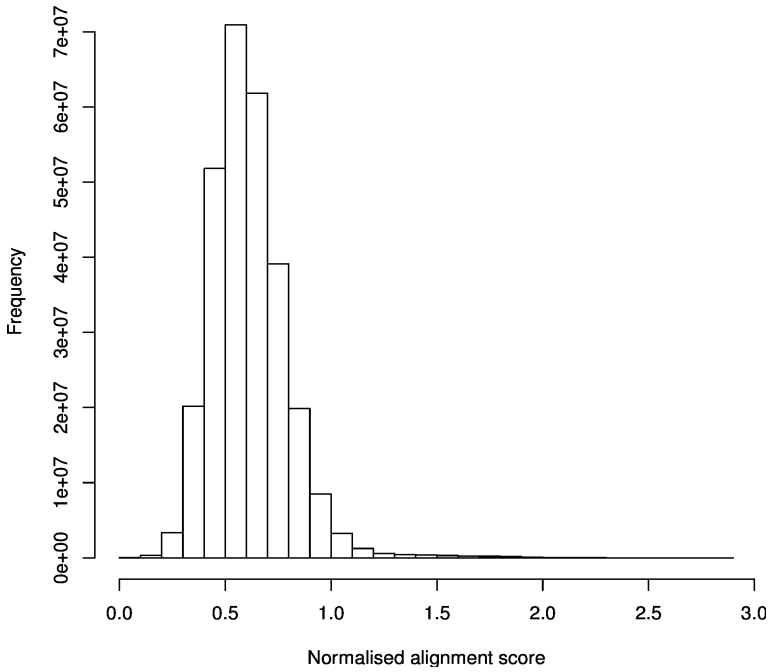


Figure 43. Distribution of normalized alignment scores for the PDB100 dataset.

5.4.2. Alignment based functional annotation

From the results of cross referencing the alignments with SCOPe a large number of unannotated alignments are revealed. The distribution of unannotated alignments for the PDB100 dataset at the six different alignments score cut-offs ranging from 1.0 to 2.0 are listed

in Table 20. It is important to note here that at each of the specified alignments score cut-offs the protein chains share regions of similarity greater than or equal to 60 residues. This does not necessarily imply any evolutionary relationship between the protein segments. We can improve the confidence of extrapolating the functional annotation to an unannotated protein by setting a more stringent alignment score cut-off. While it is true that the general trend of more stringent scores picks-up higher proportions of evolutionary related proteins, a globally applicable score threshold for all protein families cannot be set. This is mainly due to the fact that different protein families are under different evolutionary pressures and hence evolve at different rates. The solution is to set a dynamic threshold on a case by case basis.

Table 20: Distribution of unannotated alignments at different score cut-offs at the superfamily level from the SCOPe database (Chandonia, Fox, and Brenner 2017) for the PDB100 dataset .

Normalized alignment score cut-off	Unannotated alignments ≥ 60 residues in length
≥ 1.0	4,599,359
≥ 1.2	869,225
≥ 1.4	403,900
≥ 1.6	201,307
≥ 1.8	68,457
≥ 2.0	13,217

To illustrate the “one size does not fit all” nature of fixing a single score cut-off to identify all the members of a superfamily we have a comparative case study. The one vs all alignment clusters for acyl phosphatase (PDB ID 1APS_A) and SH3 domain of Rho GTPase activating protein (PDB ID 2MIO_A) are compared at cut-offs of 0.5, 0.8, 1.0 and 1.2. In both the cases the nodes are coloured based on the SCOPe annotation (see methods section). The nodes in white mean that the aligned pair of proteins do not share any common domains at the superfamily level or they have not been annotated for it yet. 1APS_A belongs to the ferredoxin like fold (59 superfamilies) and 2MIO_A belongs to the SH3 barrel like fold (21

superfamilies). At a score cut-off of 0.5 in both Figures 44 and 45, we see many of the superfamilies are represented in the clusters. But at a more stringent threshold of 0.8, in Figure 44 we have only the acyl phosphatase superfamily whereas in Figure 45 there are still 4 different superfamilies. Even at higher cut-offs, two superfamilies still remain. Hence structural similarity based identification of superfamilies is not possible. In both cases at a score cut-off of 1.0, we can analyse the unannotated nodes. Two of the five unannotated nodes in the Rho GTPase cluster have SH3 domains as per pfam (Finn et al. 2014) annotation. The other three have trypsin domain, molybdopterin binding domain and a nucleic acid binding domains respectively. Thus we can hypothesise about the evolutionary relatedness between these domains and the SH3 domain due to the high level of structural similarity between them. Similarly for the cluster in Figure 45, one of the unannotated has an acyl phosphatase domain and the other has a glutamine synthetase domain.

5.4.3. Global analysis

At the default length cut-off of 60 and a score cut-off of 1.2 there are 926 clusters with the largest cluster having 44,795 nodes. At the other end of the spectrum we see many bi-nodal clusters, indicating that they have only one structural homologue in the PDB at this given cut-off. Figure 46(a) is a global view of the network that is generated at a score cut-off of 1.2 and Figure 46(b) focuses on the cluster around cytochrome C3. We can see the disintegration of the cluster as the score cut-off is made more stringent from 1.0-2.2. These structural similarity networks help in visualizing the structural redundancy of the entire PDB in its current state. At a lax threshold of 1.0 normalized alignment score, the 84 clusters that maintain their identity and do not form a part of the single large cluster are interesting cases. These proteins have a structurally unique fold that is unlike any other proteins that constitute the bulk of the PDB. One such cluster is that of the cytochrome C3 describe above. As

scientific interest involves focusing on clusters around specific proteins like the cytochrome, in the future we plan to make available all the alignments via a web based tool.

5.4.4. Attempts to assign domains using alignment information

We tried several strategies to identify domains based on one vs all alignments from the comprehensive PDB100 dataset. The alignments were sorted based on the alignment length from shortest to longest and the domain boundaries were progressively expanded corresponding to the starting and ending position of the alignment. We attempted this exercise with alignments filtered at different score cut-offs. Representations like the one in Figure 7(a) were also generated to visualise where the alignments were located along the entire length of the protein. Full length alignments were excluded in the first round to prevent fusing all the domains together in multiple domain proteins and we chose different alignment lengths to qualify as "full length alignments" eg. 80% of the entire length of the protein. Other approach we tried were to sort the alignments by the starting position of the alignments and raw alignment score.

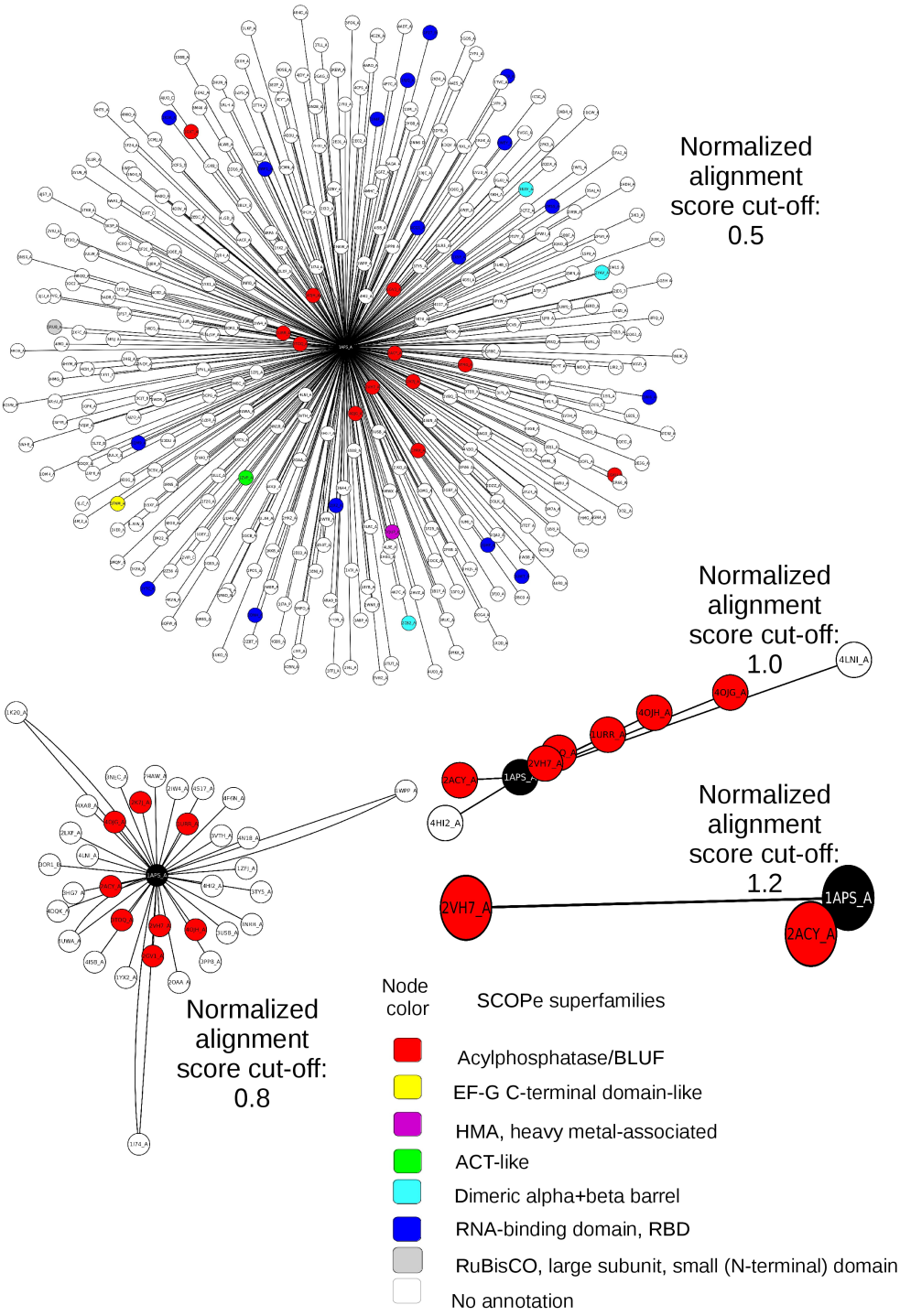


Figure 44. Clusters for Acyl phosphatase (PDB ID 1APS_A) at score cut-offs of 0.5, 0.8, 1.0 and 1.2.

This is to illustrate the fact that a single score cut-off cannot identify all the members of a superfamily. Different cut-offs are required for different proteins because each superfamily is under different evolutionary pressure hence exhibiting varied levels of structural diversity. In both cases, members belonging to the same superfamily are colored identically. The edge lengths are proportional to the alignment scores.

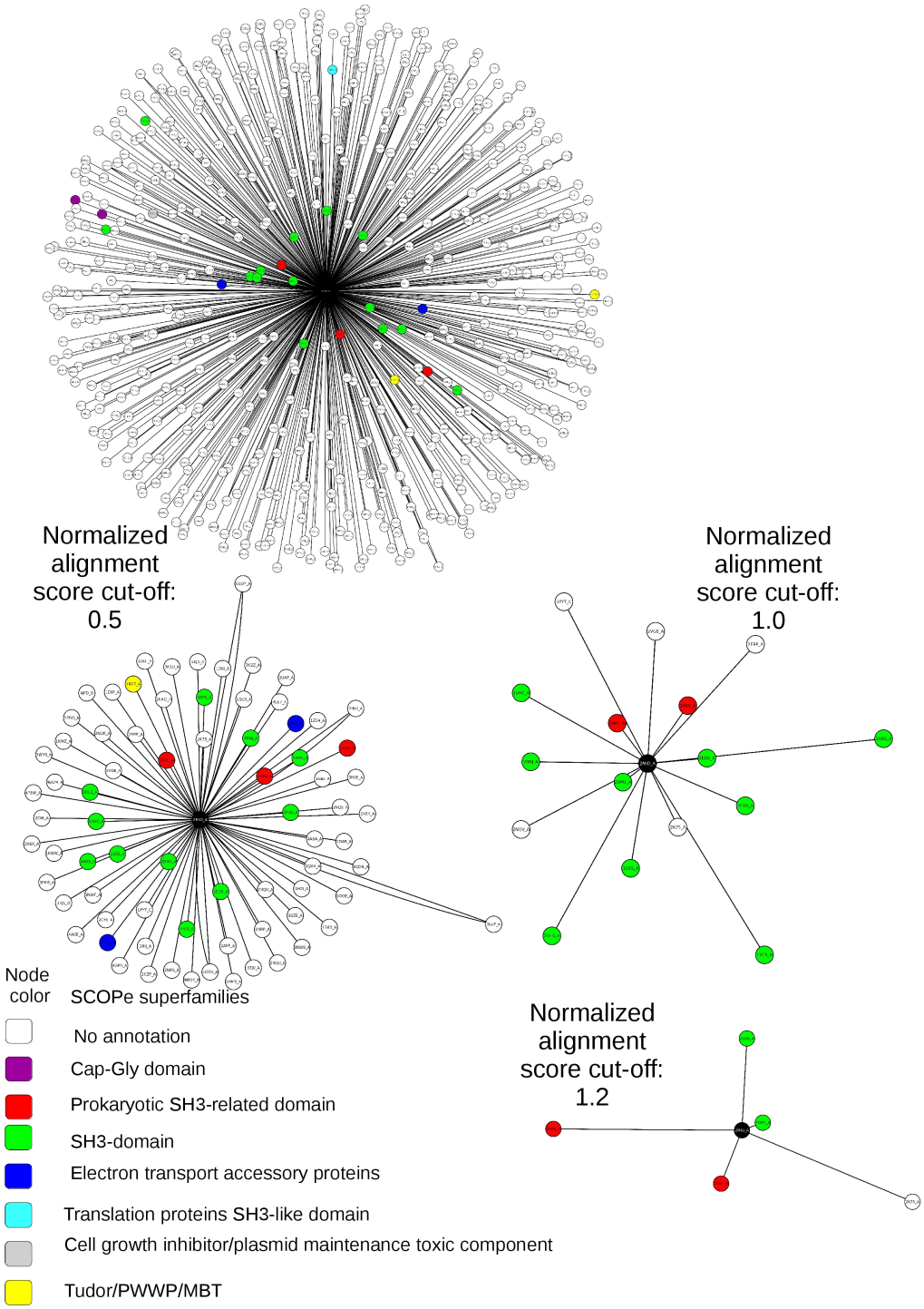


Figure 45. Clusters for Rho GTPase (PDB ID 2MIO_A) at score cut-offs of 0.5, 0.8, 1.0 and 1.2.

This is to illustrate the fact that a single score cut-off cannot identify all the members of a superfamily. Different cut-offs are required for different proteins because each superfamily is under different evolutionary pressure hence exhibiting varied levels of structural diversity. In both cases, members belonging to the same superfamily are coloured identically. The edge lengths are proportional to the alignment scores.

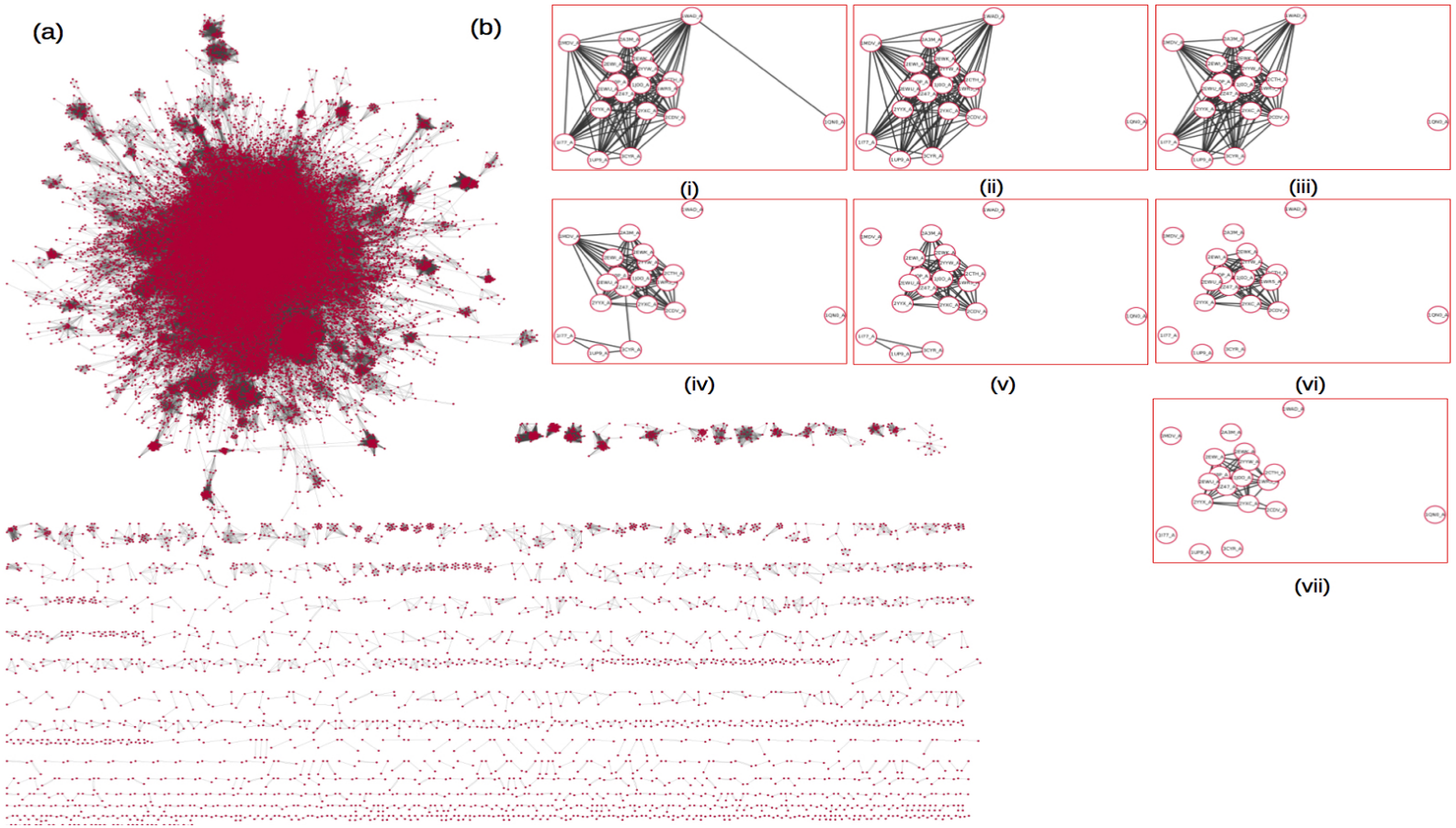


Figure 46. Global analysis of structural similarities derived from PB alignments for the PDB100 dataset.
(a) Structural similarity network at a normalized alignment score cut-off of 1.2. It shows the 926 clusters with the largest cluster having 44,794 members. (b) Evolution of the cytochrome C3 cluster at normalized alignment scores 1.0 to 2.2 (i to vii)

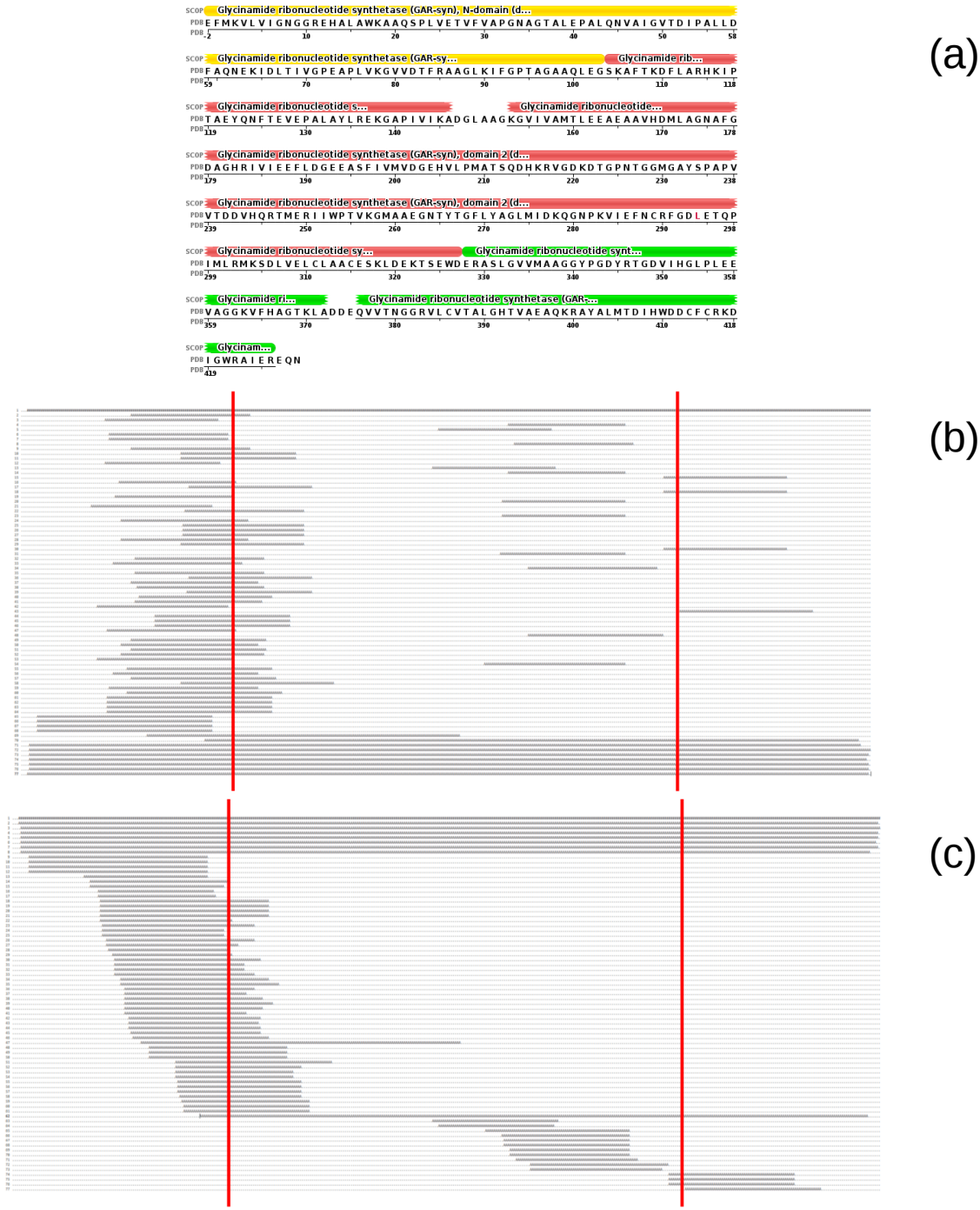


Figure 47. Domain assignment after PB alignment.
(a) Three domains in GAR-SYN from *E.coli* (PDB ID 1GSO_A) as annotated by SCOP. (b) One vs all alignments for 1GSO_A sorted by length. (c) One vs all alignments for 1GSO_A sorted by starting position of alignment. In both cases the red line denotes the domain demarcations

As an example here we discuss our attempt to assign domains for Glycinamide Ribonucleotide Synthetase (GAR-SYN) (pdbID 1GSO_A), a ligase from *E. coli.* According

to SCOP there are three domains as shown in Figure 47(a). There were 77 alignments greater than 1.0 normalized alignment score. These alignments sorted based on the length and starting position of alignment are represented in Figure 47(b) and 47(c) respectively. In Figure 47(c) the first seven alignments are ligases from other organisms but the subsequent alignments are a diverse collection of hydrolases, isomerases, lyases, oxidoreductases etc. But again alignments 47 and 62 are ligases. We see that domain assignment is not possible based on criteria like alignment length, normalized alignment score or raw alignment score. The structure space is sufficiently continuous that even at stringent alignment cut-offs, domain boundaries cannot be demarcated.

5.5. Discussion

The plot of number of proteins in clusters vs normalised alignment score (Figure 48(a)) indicates a value of 52,075 at a score cut-off of 1.0. This implies that out of the 61,433 proteins in the PDB100 dataset, 52,075 proteins have at least one structural neighbour and the remaining 9,358 proteins are singletons. As the score cut-off increases, the number of proteins in cluster gradually decreases. This trend does not apply for the plot in Figure 48(b). The size of the largest cluster has a sharp drop between the 1.2 and 1.6 score cut-offs and flattens later. The 52,075 clustered proteins are distributed as 51,803 proteins in a single large cluster and the remaining 272 proteins distributed over the remaining 84 clusters (data not shown). The number of clusters plot peaks at a score cut-off of 1.6 and tapers on both extremities, the tapering at higher cut-offs is due to proteins not having structural neighbours but at lower cut-offs, the proteins form members of other larger clusters (Figure 48(c)). Thus 1.6 is the inflection point at which we are able to identify the most number of structurally distinct proteins for the current composition of the PDB.

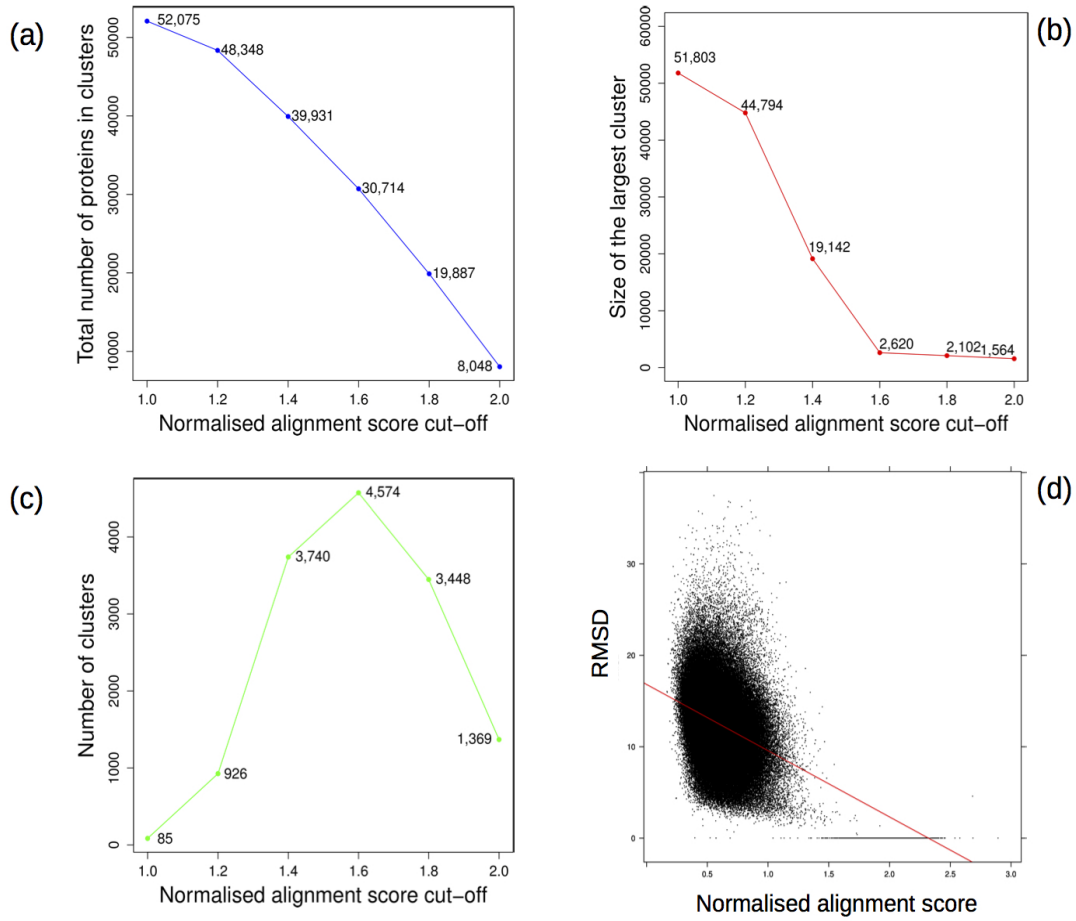


Figure 48. Analysis of the effect of alignment score cut-off on the number and size of clusters and plot of pairwise alignment score vs corresponding rmsd.
(a) Total number of proteins in cluster, (b) size of the largest cluster, (c) number of clusters at 6 score cut-offs and (d) plot of rmsd vs normalized alignment score calculated on PDB20 dataset. The red line denotes the linear trend line.

While cross referencing the alignments of PDB100 dataset with the annotations from SCOPe, two scenarios can arise (see methods section). The case when the query does not share any common domain annotation with the reference protein at the family level, this can be a potential case where we can contribute to a novel annotation. However it is important to note here that it is only a potential case scenario. We can improve the confidence of this

potential annotation by imposing a more stringent thresholds for normalized alignment score and the alignment length. In Figure 49 the lesser probable novel targets for annotation are weeded out by increasing the alignment score cut-off from 1.2 to 1.6. The two non-annotated queries identified when human lactoferrin (pdbID 1B0L.A) is the reference protein are just the N-terminal lobe of human lactoferrin (pdbID 2PMS.A) and the C-terminal lobe of bovine lactoferrin (pdbID 4N6P.A). They align with scores of 1.899 and 1.892 respectively and they also superimpose well with low all atom RMSD values of 0.475 and 0.497 respectively. Thus we can safely extrapolate the annotations to the pdbIDs 2PMS.A and 4N6P.A.

There could be potentially more number of proteins (white) with the lactoferrin domain in 1.2 score cut-off representation of Figure 49(a). It is difficult to identify a single global cut-off score across different domains as the structural diversity among them varies greatly. The safe bet would be to keep a stringent threshold, the number of unannotated domains identified at different score thresholds are listed in Table 2. Nonetheless, we see an expected trend of a higher concentration of same family domains in the core and non-annotated and neighboring family domains in the periphery.

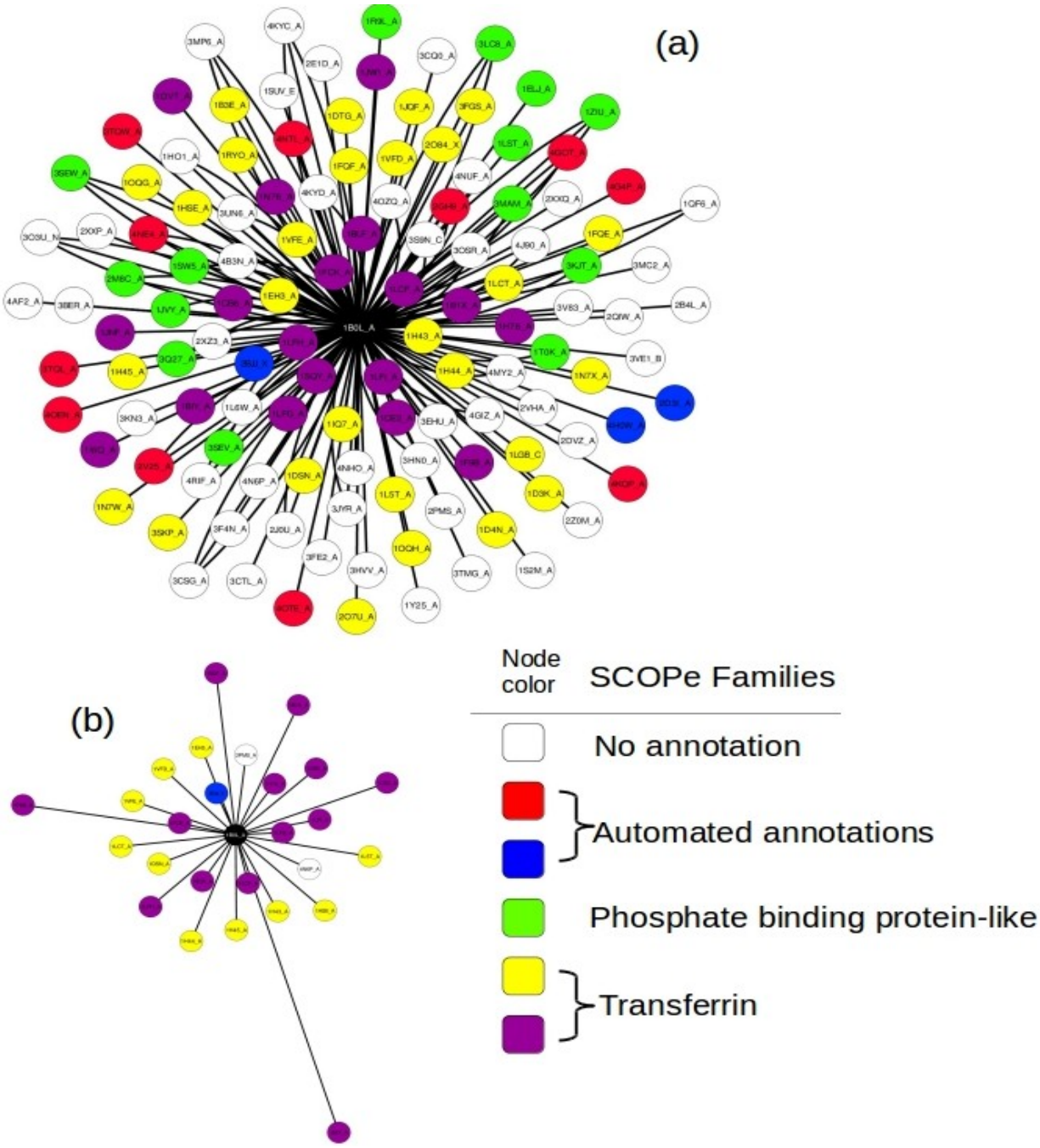


Figure 49. Increasing the score-cut-off from (a) 1.2 to (b) 1.6 weeds out the potential false positives for alignment based functional annotation.

The structural variations among the members of a family are so extensive that some members from the neighboring family align better with the query than with the members of its own family. In the case of human lactoferrin, the members belonging to the same family are

indicated in purple (two domain occurrences) and yellow (one domain occurrence). If we were to attempt to separate out the members of the transferrin family from the *phosphate binding protein-like* family (green) and other automated matches (red and blue), it would be impossible with the structural alignment scores alone. We could set the score threshold to a very stringent value (*e.g* 1.6) and miss out on a substantial number of members of the transferrin family. On the other hand if we set a very lax threshold, the collection of transferrin proteins will be contaminated by members belonging to the *phosphate binding protein-like* family and other automated matches.

Chapter 6: Conclusions and Perspectives

The principle theme of this thesis had been to capitalize on the advantages of representing a protein as a series of repetitive structural patterns, Protein Blocks. The ability of PBs to abstract protein structure with minimal loss of structural information has been used to carry out certain analysis that would be very difficult or even impossible to perform without this approximation. In this section the results, key findings and future prospects of the studies performed in this thesis are discussed.

The high sensitivity of PBs to identify local structural changes in proteins without being mislead by rigid body motions have proven it to be an ideal tool to summarize structural variations across large collections of proteins or during an MD simulation. This is evident in the case study from section 2.3.3 of Chapter 2 that captures a rigid body motion as a few drastic PB changes, but the rest of the protein is well aligned. Summarizing the structural variations observed in redundant crystallographic structures as a PB substitution matrix has helped us to compare these results with previously established PB substitution matrices. The important message from this analysis is that the structural variations inherently observed in proteins seem to be similar to the ones observed between homologous proteins. We also see the interesting case where the dynamics of the human sulfotransferase are so extensive that some conformations are structurally more resemblant of the mouse sulfotransferase than other conformations of human sulfotransferase.

We further constructed a similar substitution matrix for the variations observed between different snapshots during a MD simulation. Interestingly the conformations sampled by a protein during MD simulations seem to be contrasting when compared to the

conformations observed in NMR ensembles. Further a few examples are illustrated to show the potential of utilizing PBs as a tool for analyzing MD trajectories e.g. PB logo, PB map etc. In this project, MD simulations for a duration of 10 ns were analyzed in time steps of 2 ps. As a future perspective, it could be interesting to analyze MD trajectories of longer duration simulations and also in varying time steps. It will be interesting to view the dynamics of proteins that exhibit significant structural movement in the light of PB as for e.g the lid movement in lipases.

The prediction of protein local structure by a knowledge-based method, kPRED is able to reach average prediction accuracies of up to 66.3% for a dataset of more than 15,000 unrelated proteins. Interestingly, for some proteins, prediction accuracies could reach very high values close to 100%. Taking into factor the co-operativity between the pentapeptide under consideration and its neighboring pentapeptides while performing the predictions provides a marginal increase in accuracy. On the other hand, incorporating sequence context into the prediction process provides on an average 12% boost to the prediction accuracy. The evaluation of the performance of the kPRED methodology while making different levels of homologues available to the algorithm helped us to estimate its performance in real life prediction scenarios. kPRED predictions fair well when compared to previously established machine learning based PB prediction methods. Specific case studies are used for demonstrating the strengths and weaknesses of the kPRED methodology. The scoring function used to auto-evaluate our predictions performs well with a high correlation (R^2 =0.82) between the prediction accuracy and the score. Further, a plot of actual vs predicted scores shows a good separation of the well predicted and badly predicted PB sequences. For the well predicted PB sequences (accuracy >60%) there is a good correlation between the actual PB sequence score and the predicted PB sequence score. This suggests that the distribution of

pentaPBs in protein structures shows a distinctive pattern that can be exploited to separate them from spurious PB sequences. This improved kPRED methodology, including the scoring function, is implemented as a web based tool.

We also demonstrated that progressively introducing erroneous PBs into the correct PB sequence will show a corresponding gradual decrease in scores. This was the motivation for us to explore a genetic algorithm based method for improving the PB prediction using the scoring function as a guideline towards higher accuracies. We are not providing details of the results obtained because the method is still in the preliminary stages of development. It requires standardization of the parameters like mutation rate, number of generations and number of off-springs per generation. Also incorporating complimentary information like solvent accessibility, contact map etc. can be explored as future means to improve prediction accuracy.

PENTAdb, the database of pentapeptide was updated to have a comprehensive coverage of all the pentapeptides in PDB. Automation of the update mechanism and synchronizing it to correspond to the weekly update cycle of PDB greatly benefited us to have an exhaustive view of the pentapeptide space. This all-inclusive version of the PENTAdb gave us new insights into the structural plasticity observed in pentapeptides and 73 pentapeptides were identified as highly structurally variable as they could correspond to all 16 PBs. Extensive structural plasticity within PB classes was observed as summarized in the form of RMSDA plots for the 61.6 million pentapeptides. This reflects that categorization of ϕ an ψ torsion angles at the level of pentapeptides into 16 classes may not be optimal. Analyzing specific pentapeptides in the light of solvent accessibility, contact maps and polar contacts helped us understand the sequence-structure correlation in them. Increasing the coverage of PENTAdb has shown to have a profound effect on the PB prediction by the kPRED

methodology. It provided better coverage of the pentapeptide space which helps improve the prediction. However it also resulted in the addition of redundant information from homologues which introduced noise and biased the predictions. We also showed that the structural variability index (SVI) proved to be a useful metric to relate pentapeptide flexibility and predictability.

Large scale structural alignments carried out using the PB-ALIGN program were used to construct protein structure networks at different thresholds of structural similarity. The collection of proteins of the PDB100 dataset that were used for this analysis were clustered by using the normalized alignment scores as a clustering criteria. The alignment data was cross referenced to annotations from the SCOPe database to explore the possibility of contributing to novel annotations. A total of 13,217 alignments were identified that could benefit from putative annotation at a relatively high alignment score cut-off of 2.0. In cases where annotation could not be extrapolated, evolutionary relationships could be hypothesized. The Cytoscape software proved to be an ideal tool to construct and visualize networks. From the study of the alignment results at 6 different score cut-offs of 1.0, 1.2, 1.4,...2.0, the score cut-off of 1.6 was identified as an inflection point when maximum number of clusters were formed. Another important conclusion we arrived at from this study is that, since different protein families are at different evolutionary pressures, a global normalized alignment score cut-off cannot be used across different protein families. The examples of Acyl phosphatase and GTPase were useful to illustrate this observation. The network generated at a lax score cutoff of 1.0 enabled us to identify the protein families that had the most distinctive folds. There were 84 such distinct clusters that did not associate with the largest cluster, hence implying that they were most structurally distinct e.g the cytochrome P450 cluster. Since the score cut-off for identifying all members of a family has to be decided on a case by case basis,

the future perspective of this project would be to make all this alignment data available to the scientific community as a web based tool. This would allow the users to dynamically filter out the relevant data from the 282.8 million alignments performed.

Now that we have shown that it is possible in some cases to predict with high accuracy the PB sequence of a protein sequence, one can think of developing methods to assemble protein blocks and generate protein structures at atomic level. This is not a trivial task. Indeed, the combinatorial behind the assembling of fragments to generate 3D models of proteins is huge. However, some algorithms, like the one implemented in early versions of Rosetta, have been successful in this endeavor. Rosetta is one of the so-called *ab-initio* methods. In absence of templates, these methods are valuable for predicting protein structures. The collection of pentapeptides grouped into 16 Protein Blocks offers an opportunity to develop methods for *ab-initio* protein structure prediction via the assembly of fragments. To reduce the complexity of the task, one could think at using PENTAdb as a collection of pentapeptides to derive useful constraints that can feed sophisticated backbone reconstruction algorithms.

References

Abraham, Mark James, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. 2015. “GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers.” *SoftwareX* 1: 19–25.

Alden, Kieran, Stella Veretnik, and Philip E Bourne. 2010. “dConsensus: A Tool for Displaying Domain Assignments by Multiple Structure-Based Algorithms and for Construction of a Consensus Assignment.” *BMC Bioinformatics* 11 (1): 310. doi:10.1186/1471-2105-11-310.

Alexandrov, Nickolai, and Ilya Shindyalov. 2003. “PDP: Protein Domain Parser.” *Bioinformatics (Oxford, England)* 19 (3): 429–30.

Altschul, S F, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research* 25 (17): 3389–3402. http://www.ncbi.nlm.nih.gov/pubmed/9254694.

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Alva, Vikram, Michael Remmert, Andreas Biegert, Andrei N. Lupas, and Johannes Söding. 2010. “A Galaxy of Folds.” *Protein Science* 19: 124–30. doi:10.1002/pro.297.

Andreeva, Antonina, Dave Howorth, Cyrus Chothia, Eugene Kulesha, and Alexey G Murzin. 2014. “SCOP2 Prototype: A New Approach to Protein Structure Mining.” *Nucleic Acids Research* 42 (Database issue): D310–14. doi:10.1093/nar/gkt1242.

Argos, P. (1987). Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. Strategies for protein folding and a guide for site-directed mutagenesis. *Journal of Molecular Biology*, *197*(2), 331–48. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3681998

Armen, Roger, Darwin O V Alonso, and Valerie Daggett. 2003. “The Role of Alpha-, 3(10)-, and Pi-Helix in Helix-->coil Transitions.” *Protein Science : A Publication of the Protein Society* 12 (6): 1145–57. doi:10.1110/ps.0240103.

Balaji, S, S Sujatha, S S Kumar, and N Srinivasan. 2001. “PALI-a Database of Phylogeny and ALIgnment of Homologous Protein Structures.” *Nucleic Acids Research* 29 (1): 61–65.

Bansal, M, S Kumar, and R Velavan. 2000. “HELANAL: A Program to Characterize Helix Geometry in Proteins.” *Journal of Biomolecular Structure & Dynamics* 17 (5): 811–19. doi:10.1080/07391102.2000.10506570.

Bashford, D, C Chothia, and A M Lesk. 1987. “Determinants of a Protein Fold. Unique Features of the Globin Amino Acid Sequences.” *Journal of Molecular Biology* 196 (1): 199–216.

Bikadi, Z., Demko, L., & Hazai, E. (2007). Functional and structural characterization of a protein based on analysis of its hydrogen bonding network by hydrogen bonding plot. *Archives of Biochemistry and Biophysics*, *461*(2), 225–234. doi:10.1016/j.abb.2007.02.020

Bowie, J., J. Reidhaar-Olson, W. Lim, and R. Sauer. 1990. “Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions.” *Science* 247 (4948): 1306–10.

de Brevern, A G, C Etchebest, and S Hazout. 2000. “Bayesian Probabilistic Approach for Predicting Backbone.” *Proteins* 287 (January): 271–87.

Bystroff, Christopher, and David Baker. 1998. “Prediction of Local Structure in Proteins Using a Library of Sequence-Structure motifs11Edited by J. Thornton.” *Journal of Molecular Biology* 281: 565–77. doi:10.1006/jmbi.1998.1943.

C. Lacey, S. Cole. 1993. “Merger Rates in Hierarchical Models of Galaxy Formation.” *Mon. Not. Roy. Astron. Soc.* 262.

Camproux, A C, P Tuffery, J P Chevrolat, J F Boisvieux, and S Hazout. 1999. “Hidden Markov Model Approach for Identifying the Modular Framework of the Protein Backbone.” *Protein Engineering* 12 (12): 1063–73. doi:10.1093/PROTEIN/12.12.1063.

Camproux, A. C., P. Tuffery, L. Buffat, C. André, S. Hazout, and J. F. Boisvieux. 1999. “Analyzing Patterns between Regular Secondary Structures Using Short Structural Building Blocks Defined by a Hidden Markov Model.” *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* 101: 33–40.

Camproux, A.C, R Gautier, and P Tufféry. 2004. “A Hidden Markov Model Derived Structural Alphabet for Proteins.” *Journal of Molecular Biology* 339 (3): 591–605. doi:10.1016/j.jmb.2004.04.005.

Carpentier, Mathilde, Sophie Brouillet, and Joël Pothier. 2005. “YAKUSA: A Fast Structural Database Scanning Method.” *Proteins: Structure, Function, and Bioinformatics* 61 (1): 137–51. doi:10.1002/prot.20517.

Chandonia, John-Marc, Naomi K. Fox, and Steven E. Brenner. 2017. *SCOPe: Manual Curation and Artifact Removal in the Structural Classification of Proteins – Extended Database*. *Journal of Molecular Biology*. Vol. 429. doi:10.1016/j.jmb.2016.11.023.

Cheng, Hua, R. Dustin Schaeffer, Yuxing Liao, Lisa N. Kinch, Jimin Pei, Shuoyong Shi, Bong-Hyun Kim, and Nick V. Grishin. 2014. “ECOD: An Evolutionary Classification of

Protein Domains." Edited by Arne Elofsson. *PLoS Computational Biology* 10 (12). Public Library of Science: e1003926. doi:10.1371/journal.pcbi.1003926.

Cheng, Hua, R. Dustin Schaeffer, Yuxing Liao, Lisa N. Kinch, Jimin Pei, Shuoyong Shi, Bong-Hyun Kim, and Nick V. Grishin. 2014. "ECOD: An Evolutionary Classification of Protein Domains." Edited by Arne Elofsson. *PLoS Computational Biology* 10 (12). Public Library of Science: e1003926. doi:10.1371/journal.pcbi.1003926.

Chothia, C, and A M Lesk. 1986. "The Relation between the Divergence of Sequence and Structure in Proteins." *The EMBO Journal* 5 (4). European Molecular Biology Organization: 823–26.

Chou, P Y, and G D Fasman. 1974. "Prediction of Protein Conformation." *Biochemistry* 13 (2): 222–45. http://www.ncbi.nlm.nih.gov/pubmed/4358940.

Cohen, B. I., Presnell, S. R., & Cohen, F. E. (1993). Origins of structural diversity within sequentially identical hexapeptides. *Protein Science : A Publication of the Protein Society*, *2*, 2134–2145. doi:10.1002/pro.5560021213

Colloc'h, Nathalie, Catherine Etchebest, Etienne Thoreau, Bernard Henrissat, and Jean-Paul Mornon. 1993. "Comparison of Three Algorithms for the Assignment of Secondary Structure in Proteins: The Advantages of a Consensus Assignment." *"Protein Engineering, Design and Selection"* 6 (4). Oxford University Press: 377–82. doi:10.1093/protein/6.4.377.

Cowan-Jacob, Sandra W, Gabriele Fendrich, Paul W Manley, Wolfgang Jahnke, Doriano Fabbro, Janis Liebetanz, and Thomas Meyer. 2005. "The Crystal Structure of a c-Src Complex in an Active Conformation Suggests Possible Steps in c-Src Activation." *Structure (London, England : 1993)* 13 (6): 861–71.

Craveur, Pierrick, Agnel P. Joseph, Jeremy Esque, Tarun J. Narwani, Floriane Noël, Nicolas Shinada, Matthieu Goguet, et al. 2015. "Protein Flexibility in the Light of Structural Alphabets." *Frontiers in Molecular Biosciences* 2. doi:10.3389/fmolb.2015.00020.

Crippen, Gordon M. 1978. "The Tree Structural Organization of Proteins." *Journal of Molecular Biology* 126 (3): 315–32. doi:10.1016/0022-2836(78)90043-8.

Crooks, Gavin E, Gary Hon, John-Marc Chandonia, and Steven E Brenner. 2004. "WebLogo: A Sequence Logo Generator." *Genome Research* 14 (6). Cold Spring Harbor Laboratory Press: 1188–90.

Dayhoff, Margaret O., Schwartz R.M., Orcutt B.C. 1978. "Atlas of Protein Sequence and Structure." National Biomedical Research Foundation.

de Brevern, A G, C Etchebest, and S Hazout. 2000. "Bayesian Probabilistic Approach for Predicting Backbone." *Proteins* 287 : 271–87.

de Brevern, Alexandre G, Aurélie Bornot, Pierrick Craveur, Catherine Etchebest, and Jean-Christophe Gelly. 2012. “PredyFlexy: Flexibility and Local Structure Prediction from Sequence.” *Nucleic Acids Research* 40 (Web Server issue): W317–22. doi:10.1093/nar/gks482.

de Brevern, Alexandre G. 2005. “New Assessment of a Structural Alphabet.” *In Silico Biology* 5 (3): 283–89.

Dessailly, Benoît H, Marc F Lensink, and Shoshana J Wodak. 2007. “Relating Destabilizing Regions to Known Functional Sites in Proteins.” *BMC Bioinformatics* 8 (1): 141.

Dokholyan, Nikolay V, Boris Shakhnovich, and Eugene I Shakhnovich. 2002. “Expanding Protein Universe and Its Origin from the Biological Big Bang.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (22): 14132–36. doi:10.1073/pnas.202497999.

Donohue, J. 1953. “Hydrogen Bonded Helical Configurations of the Polypeptide Chain.” *Proceedings of the National Academy of Sciences of the United States of America* 39 (6): 470–78. http://www.ncbi.nlm.nih.gov/pubmed/16589292.

Dror, O., H. Benyamini, R. Nussinov, and H. Wolfson. 2003. “MASS: Multiple Structural Alignment by Secondary Structures.” In *Bioinformatics*. Vol. 19. doi:10.1093/bioinformatics/btg1012.

Dudev, Minko, and Carmay Lim. 2007. “Discovering Structural Motifs Using a Structural Alphabet: Application to Magnesium-Binding Sites.” *BMC Bioinformatics* 8: 106. doi:10.1186/1471-2105-8-106.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. doi:10.1093/nar/gkh340.

Edwards, Hannah, and Charlotte M. Deane. 2015. “Structural Bridges through Fold Space.” *PLoS Computational Biology* 11 (9). doi:10.1371/journal.pcbi.1004466.

Etchebest, C., C. Benros, A. Bornot, A.-C. Camproux, and A. G. de Brevern. 2007. “A Reduced Amino Acid Alphabet for Understanding and Designing Protein Adaptation to Mutation.” *European Biophysics Journal* 36 (8). Springer-Verlag: 1059–69. doi:10.1007/s00249-007-0188-5.

Etchebest, Catherine, Cristina Benros, Serge Hazout, and Alexandre G. De Brevern. 2005. “A Structural Alphabet for Local Protein Structures: Improved Prediction Methods.” *Proteins: Structure, Function and Genetics* 59: 810–27. doi:10.1002/prot.20458.

Etchebest, Catherine, Cristina Benros, Serge Hazout, and Alexandre G. De Brevern. 2005. “A Structural Alphabet for Local Protein Structures: Improved Prediction Methods.” *Proteins: Structure, Function and Genetics* 59 (December 2004): 810–27. doi:10.1002/prot.20458.

Everitt, Brian S., Sabine Landau, Morven Leese, and Daniel Stahl. 2011. *Cluster Analysis*. Wiley Series in Probability and Statistics. Chichester, UK: John Wiley & Sons, Ltd. doi:10.1002/9780470977811.

Fetrow, Jacquelyn S., Michael J. Palumbo, and George Berg. 1997. "Patterns, Structures, and Amino Acid Frequencies in Structural Building Blocks, a Protein Secondary Structure Classification Scheme." *Proteins: Structure, Function and Genetics* 27 (2): 249–71. doi:10.1002/(SICI)1097-0134(199702)27:2<249::AID-PROT11>3.0.CO;2-M.

Finn, Robert D., Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, et al. 2014. "Pfam: The Protein Families Database." *Nucleic Acids Research* 42 (D1): 222–30. doi:10.1093/nar/gkt1223.

Fox, Naomi K., Steven E. Brenner, and John-Marc Chandonia. 2014. "SCOPe: Structural Classification of Proteins—extended, Integrating SCOP and ASTRAL Data and Classification of New Structures." *Nucleic Acids Research* 42 (D1). Oxford University Press: D304–9. doi:10.1093/nar/gkt1240.

Frishman, Dmitrij, and Patrick Argos. 1995. "Knowledge-Based Protein Secondary Structure Assignment." *Proteins: Structure, Function, and Genetics* 23 (4). Wiley Subscription Services, Inc., A Wiley Company: 566–79. doi:10.1002/prot.340230412.

Furnham, N., Holliday, G. L., De Beer, T. A. P., Jacobsen, J. O. B., Pearson, W. R., & Thornton, J. M. (2014). The Catalytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Research*, *42*(D1). doi:10.1093/nar/gkt1243

Garnier, J, J F Gibrat, and B Robson. 1996. "GOR Secondary Structure Prediction Method Version IV." *Methods Enzymol* 266: 540–53. doi:10.1002/146518.

Gelly, Jean-Christophe, Agnel Praveen Joseph, Narayanaswamy Srinivasan, and Alexandre G de Brevern. 2011. "iPBA: A Tool for Protein Structure Comparison Using Sequence Alignment Strategies." *Nucleic Acids Research* 39 (Web Server issue): W18–23. doi:10.1093/nar/gkr333.

Gerstein, M, A M Lesk, and C Chothia. 1994. "Structural Mechanisms for Domain Movements in Proteins." *Biochemistry* 33 (22): 6739–49.

Gerstein, M, and W Krebs. 1998. "A Database of Macromolecular Motions." *Nucleic Acids Research* 26 (18). Oxford University Press: 4280–90.

Ghouzam, Yassine, Guillaume Postic, Alexandre G. de Brevern, and Jean-Christophe Gelly. 2015. "Improving Protein Fold Recognition with Hybrid Profiles Combining Sequence and Structure Evolution." *Bioinformatics*, no. August: btv462. doi:10.1093/bioinformatics/btv462.

Gibrat, Jean Francois, Thomas Madej, and Stephen H. Bryant. 1996. “Surprising Similarities in Structure Comparison.” *Current Opinion in Structural Biology*. doi:10.1016/S0959-440X(96)80058-3.

Gowri, V S, Shashi B Pandit, P S Karthik, N Srinivasan, and S Balaji. 2003. “Integration of Related Sequences with Protein Three-Dimensional Structural Families in an Updated Version of PALI Database.” *Nucleic Acids Research* 31 (1): 486–88.

Griep, S., and U. Hobohm. 2010. “PDBselect 1992-2009 and PDBfilter-Select.” *Nucleic Acids Research* 38 (Database): D318–19. doi:10.1093/nar/gkp786.

Hansen, Jeffrey C, Xu Lu, Eric D Ross, and Robert W Woody. 2006. “Intrinsic Protein Disorder, Amino Acid Composition, and Histone Terminal Domains.” *The Journal of Biological Chemistry* 281 (4). American Society for Biochemistry and Molecular Biology: 1853–56.

Hark Gan, Hin, Rebecca A. Perlow, Sharmili Roy, Joy Ko, Min Wu, Jing Huang, Shixiang Yan, et al. 2002. “Analysis of Protein Sequence/Structure Similarity Relationships.” *Biophysical Journal* 83 (5): 2781–91.

Heinig, M., and D. Frishman. 2004. “STRIDE: A Web Server for Secondary Structure Assignment from Known Atomic Coordinates of Proteins.” *Nucleic Acids Research* 32 (Web Server). Oxford University Press: W500–502. doi:10.1093/nar/gkh429.

Higgins, D G, and P M Sharp. 1988. “CLUSTAL: A Package for Performing Multiple Sequence Alignment on a Microcomputer.” *Gene* 73 (1): 237–44. http://www.ncbi.nlm.nih.gov/pubmed/3243435.

Holland, Timothy A., Stella Veretnik, Ilya N. Shindyalov, and Philip E. Bourne. 2006. “Partitioning Protein Structures into Domains: Why Is It so Difficult?” *Journal of Molecular Biology* 361 (3): 562–90. doi:10.1016/j.jmb.2006.05.060.

Holm, L, and C Sander. 1993. “Protein Structure Comparison by Alignment of Distance Matrices.” *Journal of Molecular Biology* 233 (1): 123–38. doi:10.1006/jmbi.1993.1489.

Holm, L, and C Sander. 1994. “Parser for Protein Folding Units.” *Proteins* 19 (3): 256–68. doi:10.1002/prot.340190309.

Holm, Liisa, and Chris Sander. 1998. “Touring Protein Fold Space with Dali/FSSP.” *Nucleic Acids Research* 26 (1): 316–19. doi:10.1093/nar/26.1.316.

Hou, Jingtong, Se-Ran Jun, Chao Zhang, and Sung-Hou Kim. 2005. “Global Mapping of the Protein Structure Space and Application in Structure-Based Inference of Protein Function.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (10): 3651–56. doi:10.1073/pnas.0409772102.

Islam, S A, J Luo, and M J Sternberg. 1995. “Identification and Analysis of Domains in Proteins.” *Protein Engineering* 8 (6): 513–25.

Hartigan J. 1975. *Clustering Algorithms*. New York: John Wiley and Sons. http://www.bibsonomy.org/bibtex/1fb9537d9624e0e10d39d92208e77de8c.

Johnson, M S, and J P Overington. 1993. “A Structural Basis for Sequence Comparisons. An Evaluation of Scoring Methodologies.” *Journal of Molecular Biology* 233 (4): 716–38.

Jones, D. T. 1999. “Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices.” *J. Mol. Biol.* 292: 195–202. doi:10.1006/jmbi.1999.3091.

Jones, D. T., Buchan, D. W. A., Cozzetto, D., & Pontil, M. 2012. PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, *28*(2), 184–190. doi:10.1093/bioinformatics/btr638

Jones, S, M Stewart, A Michie, M B Swindells, C Orengo, and J M Thornton. 1998. “Domain Assignment for Protein Structures Using a Consensus Approach: Characterization and Analysis.” *Protein Science : A Publication of the Protein Society* 7 (2): 233–42. doi:10.1002/pro.5560070202.

Joseph, Agnel Praveen, Garima Agarwal, Swapnil Mahajan, Jean-Christophe Gelly, Lakshmipuram S Swapna, Bernard Offmann, Frédéric Cadet, et al. 2010. “A Short Survey on Protein Blocks.” *Biophysical Reviews* 2 (3): 137–47. doi:10.1007/s12551-010-0036-1.

Joseph, Agnel Praveen, N Srinivasan, and Alexandre G de Brevern. 2011. “Improvement of Protein Structure Comparison Using a Structural Alphabet.” *Biochimie* 93 (9). Elsevier Masson SAS: 1434–45. doi:10.1016/j.biochi.2011.04.010.

Joseph, Agnel Praveen, Narayanaswamy Srinivasan, and Alexandre G de Brevern. 2012. “Progressive Structure-Based Alignment of Homologous Proteins: Adopting Sequence Comparison Strategies.” *Biochimie* 94 (9). Elsevier Masson SAS: 2025–34. doi:10.1016/j.biochi.2012.05.028.

Kabsch, Wolfgang, and Sander, Christian. 1983. “Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.” *Biopolymers* 22 (12): 2577–2637. doi:10.1002/bip.360221211.

Kabsch, W., & Sander, C. (1984). On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(4), 1075–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6422466

Karchin, Rachel, Melissa Cline, Yael Mandel-Gutfreund, and Kevin Karplus. 2003. “Hidden Markov Models That Use Predicted Local Structure for Fold Recognition: Alphabets of

Backbone Geometry." *Proteins: Structure, Function, and Bioinformatics* 51 (4). Wiley Subscription Services, Inc., A Wiley Company: 504–14. doi:10.1002/prot.10369.

Karchin, Rachel, Melissa Cline, and Kevin Karplus. 2004. "Evaluation of Local Structure Alphabets Based on Residue Burial." *Proteins: Structure, Function, and Bioinformatics* 55 (3). Wiley Subscription Services, Inc., A Wiley Company: 508–18. doi:10.1002/prot.20008.

Karypis, George. 2006. "YASSPP: Better Kernels and Coding Schemes Lead to Improvements in Protein Secondary Structure Prediction." *Proteins: Structure, Function and Genetics* 64 (3): 575–86. doi:10.1002/prot.21036.

Katoh, Kazutaka, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. 2005. "MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment." *Nucleic Acids Research* 33 (2). Oxford University Press: 511–18. doi:10.1093/nar/gki198.

Kendrew, J C, G Bodo, H M Dintzis, R G Parrish, H Wyckoff, and D C Phillips. 1958. "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis." *Nature* 181 (4610): 662–66. doi:10.1038/181662a0.

Kendrew, J C, R E Dickerson, B E Strandberg, R G Hard, D R Davies, D C Phillips, and V C Shore. 1960. "Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 A. Resolution." *Nature* 185 (4711): 422–27.

Kim, Paul M, Xin Duan, Alex S Huang, Cindy Y Liu, Guo-li Ming, Hongjun Song, and Solomon H Snyder. 2010. "Aspartate Racemase, Generating Neuronal D-Aspartate, Regulates Adult Neurogenesis." *Proceedings of the National Academy of Sciences of the United States of America* 107 (7): 3175–79. doi:10.1073/pnas.0914706107.

Kneissl, Benny, Sabine C Mueller, Christofer S Tautermann, and Andreas Hildebrandt. 2011. "String Kernels and High-Quality Data Set for Improved Prediction of Kinked Helices in α-Helical Membrane Proteins." *Journal of Chemical Information and Modeling* 51 (11): 3017–25. doi:10.1021/ci200278w.

Kohonen, Teuvo. 1982. "Self-Organized Formation of Topologically Correct Feature Maps." *Biological Cybernetics* 43 (1). Springer-Verlag: 59–69. doi:10.1007/BF00337288.

Kolbeck, Bjoern, Patrick May, Tobias Schmidt-Goenner, Thomas Steinke, and Ernst-Walter Knapp. 2006. "Connectivity Independent Protein-Structure Alignment: A Hierarchical Approach." *BMC Bioinformatics* 7: 510. doi:10.1186/1471-2105-7-510.

Kolodny, Rachel, Patrice Koehl, Leonidas Guibas, and Michael Levitt. 2002. "Small Libraries of Protein Fragments Model Native Protein Structures Accurately." *Journal of Molecular Biology* 323 (02): 297–307. doi:10.1016/S0022-2836(02)00942-7.

Konagurthu, A S, J C Whisstock, P J Stuckey, and A M Lesk. 2006. "MUSTANG: A Multiple Structural Alignment Algorithm." *Proteins* 64 (3): 559–74. doi:10.1002/prot.20921.

Konagurthu, Arun S, Arthur M Lesk, and Lloyd Allison. 2012. “Minimum Message Length Inference of Secondary Structure from Protein Coordinate Data.” *Bioinformatics (Oxford, England)* 28 (12): i97–105. doi:10.1093/bioinformatics/bts223.

Koolman, Jan, and Klaus-Heinrich Roehm. 2005. *Coloured Atlas of Biochemistry*. 2nd Editio. Thieme.

Kosloff, Mickey, and Rachel Kolodny. 2008. “Sequence-Similar, Structure-Dissimilar Protein Pairs in the PDB.” *Proteins: Structure, Function and Genetics* 71: 891–902. doi:10.1002/prot.21770.

Krissinel, E., and K. Henrick. 2004. “Secondary-Structure Matching (SSM), a New Tool for Fast Protein Structure Alignment in Three Dimensions.” *Acta Crystallographica Section D: Biological Crystallography* 60. International Union of Crystallography: 2256–68. doi:10.1107/S0907444904026460.

Ku, Shih-Yen, and Yuh-Jyh Hu. 2008. “Protein Structure Search and Local Structure Characterization.” *BMC Bioinformatics* 9 (1): 349. doi:10.1186/1471-2105-9-349.

Lam, Hubert, Dong-Chan Oh, Felipe Cava, Constantin N Takacs, Jon Clardy, Miguel A de Pedro, and Matthew K Waldor. 2009. “D-Amino Acids Govern Stationary Phase Cell Wall Remodeling in Bacteria.” *Science (New York, N.Y.)* 325 (5947). American Association for the Advancement of Science: 1552–55. doi:10.1126/science.1178123.

Lander, E S, a Heaford, a Sheridan, L M Linton, B Birren, a Subramanian, a Coulson, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921. doi:10.1038/35057062.

Law, Eleanor C., Henry R. Wilman, Sebastian Kelm, Jiye Shi, Charlotte M. Deane, HR Wilman, J Shi, et al. 2016. “Examining the Conservation of Kinks in Alpha Helices.” Edited by Hendrik W. van Veen. *PLOS ONE* 11 (6). Public Library of Science: e0157553. doi:10.1371/journal.pone.0157553.

Lecomte, Juliette TJ, David A Vuletich, and Arthur M Lesk. 2005. “Structural Divergence and Distant Relationships in Proteins: Evolution of the Globins.” *Current Opinion in Structural Biology* 15 (3): 290–301.

Lesk, A M, and G D Rose. 1981. “Folding Units in Globular Proteins.” *Proceedings of the National Academy of Sciences of the United States of America* 78 (7): 4304–8.

Levitt, Michael, and Jonathan Greer. 1977. “Automatic Identification of Secondary Structure in Globular Proteins.” *Journal of Molecular Biology* 114 (2). Academic Press: 181–239. doi:10.1016/0022-2836(77)90207-8.

Low, B. W., and R. B Baybutt. 1952. “The P-Helix -A Hydrogen Bonded Configuration of the Polypeptide Chain.” *J Am Chem Soc* 74 (5806).

Lu, Shin Yee, and King Sun Fu. 1978. “A Sentence-to-Sentence Clustering Procedure for Pattern Analysis.” *IEEE Transactions on Systems, Man and Cybernetics* 8 (5): 381–89. doi:10.1109/TSMC.1978.4309979.

Magnan, Christophe N., and Pierre Baldi. 2014. “SSpro/ACCpro 5: Almost Perfect Prediction of Protein Secondary Structure and Relative Solvent Accessibility Using Profiles, Machine Learning and Structural Similarity.” *Bioinformatics* 30 (18): 2592–97. doi:10.1093/bioinformatics/btu352.

Mahajan, S. 2013. Applications of a structural alphabet for protein structure analysis, prediction and fold recognition. PhD. thesis. http://www.theses.fr/2013LARE0032

Mahajan, Swapnil, Alexandre De Brevern, Bernard Offmann, Swapnil Mahajan, Alexandre De Brevern, Bernard Offmann, and Narayanaswamy Srinivasan. 2014. “Correlation between Local Structural Dynamics of Proteins Inferred from NMR Ensembles and Evolutionary Dynamics of Homologues of Known Structure.” *Journal of Biomolecular Structure and Dynamics* 32 (5): 751–58.

Mahajan, Swapnil, Alexandre G. de Brevern, Yves-Henri Sanejouand, Narayanaswamy Srinivasan, and Bernard Offmann. 2015. “Use of a Structural Alphabet to Find Compatible Folds for Amino Acid Sequences.” *Protein Science* 24: 145–53. doi:10.1002/pro.2581.

Margraf, Thomas, Gundolf Schenk, and Andrew E. Torda. 2009. “The SALAMI Protein Structure Search Server.” *Nucleic Acids Research* 37 (SUPPL. 2). doi:10.1093/nar/gkp431.

Marti-Renom, M A, M S Madhusudhan, and A Sali. 2004. “Alignment of Protein Sequences by Their Profiles.” *Protein Sci* 13 (4): 1071–87. doi:10.1110/ps.03379804.

Martin, A.C.R. and Porter C.T. “ProFit.” http://www.bioinf.org.uk/software/profit/.

Martin, Juliette, Guillaume Letellier, Antoine Marin, Jean-François Taly, Alexandre G de Brevern, Jean-François Gibrat, L Pauling, et al. 2005. “Protein Secondary Structure Assignment Revisited: A Detailed Analysis of Different Assignment Methods.” *BMC Structural Biology* 5 (1). BioMed Central: 17. doi:10.1186/1472-6807-5-17.

McDonald, I. K., & Thornton, J. M. (1994). Satisfying Hydrogen Bonding Potential in Proteins. *Journal of Molecular Biology*, *238*(5), 777–793. doi:10.1006/jmbi.1994.1334

McLachlan, A.D. 1982. “Rapid Comparison of Protein Structures.” *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*.

Micheletti, Cristian, Flavio Seno, and Amos Maritan. 2000. “Recurrent Oligomers in Proteins: An Optimal Scheme Reconciling Accurate and Concise Backbone Representations in

Automated Folding and Design Studies." *Proteins: Structure, Function and Genetics* 40 (4): 662–74. doi:10.1002/1097-0134(20000901)40:4<662::AID-PROT90>3.0.CO;2-F.

Monzon, Alexander Miguel, Ezequiel Juritz, María Silvina Fornasari, and Gustavo Parisi. 2013. "CoDNaS: A Database of Conformational Diversity in the Native State of Proteins." *Bioinformatics (Oxford, England)* 29 (19). Oxford University Press: 2512–14.

Murray, Robert, Darryl Granner, Peter Mayes, and Victor Rodwell. 2003. *Harper's Illustrated Biochemistry*. 26th ed.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, *247*, 536–540. doi:10.1016/S0022-2836(05)80134-2

Murzin, Alexey G. 1994. "New Protein Folds." *Current Opinion in Structural Biology* 4 (3): 441–49. doi:10.1016/S0959-440X(94)90114-7.

Nagl, S B, J Freeman, and T F Smith. 1999. "Evolutionary Constraint Networks in Ligand-Binding Domains: An Information-Theoretic Approach." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 90–101.

Najmanovich, Rafael, Natalja Kurbatova, and Janet Thornton. 2008. "Detection of 3D Atomic Similarities and Their Use in the Discrimination of Small Molecule Protein-Binding Sites." *Bioinformatics (Oxford, England)* 24 (16): i105–11.

Nelson, David, and Michael Cox. 2004. *Lehninger's Principles of Biochemistry*. 4th Editio. W H Freeman.

Nepomnyachiy, Sergey, Nir Ben-Tal, and Rachel Kolodny. 2014. "Global View of the Protein Universe." *Proceedings of the National Academy of Sciences*, 201403395. doi:10.1073/pnas.1403395111.

Nguyen, Minh N., and M. S. Madhusudhan. 2011. "Biological Insights from Topology Independent Comparison of Protein 3D Structures." *Nucleic Acids Research* 39 (14). doi:10.1093/nar/gkr348.

Nick Pace, C., and J. Martin Scholtz. 1998. "A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins." *Biophysical Journal* 75 (1): 422–27. doi:10.1016/S0006-3495(98)77529-0.

Notredame, Cédric, Desmond G Higgins, and Jaap Heringa. 2000. "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment." *Journal of Molecular Biology* 302 (1): 205–17. doi:10.1006/jmbi.2000.4042.

Orengo, Ca, Ad Michie, S Jones, Dt Jones, Mb Swindells, and Jm Thornton. 1997. "CATH - a Hierarchic Classification of Protein Domain Structures." *Structure* 17 (March): 1093–1109. doi:10.1016/S0969-2126(97)00260-8.

Ortiz, Angel R., Charlie E.M. Strauss, and Osvaldo Olmea. 2009. “MAMMOTH (Matching Molecular Models Obtained from Theory): An Automated Method for Model Comparison.” *Protein Science* 11 (11): 2606–21. doi:10.1110/ps.0215902.

Pascual-García, Alberto, David Abia, Ángel R. Ortiz, and Ugo Bastolla. 2009. “Cross-over between Discrete and Continuous Protein Structure Space: Insights into Automatic Classification and Networks of Protein Structures.” *PLoS Computational Biology* 5 (3). doi:10.1371/journal.pcbi.1000331.

Pauling, L, R B Corey, and H R Branson. 1951. “The Structure of Proteins; Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain.” *Proceedings of the National Academy of Sciences of the United States of America* 37 (4): 205–11.

Postic, Guillaume, Yassine Ghouzam, Romain Chebrek, and Jean-Christophe Gelly. 2017. “An Ambiguity Principle for Assigning Protein Structural Domains.”

Prestrelski, S J, D M Byler, and M N Liebman. 1992. “Generation of a Substructure Library for the Description and Classification of Protein Secondary Structure. II. Application to Spectra-Structure Correlations in Fourier Transform Infrared Spectroscopy.” *Proteins* 14 (4): 440–50. doi:10.1002/prot.340140405.

Pundir, Sangya, Michele Magrane, Maria J. Martin, Claire O’Donovan, and UniProt Consortium. 2015. “Searching and Navigating UniProt Databases.” In *Current Protocols in Bioinformatics*, 50:1.27.1–1.27.10. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471250953.bi0127s50.

Qi, Guoying, Richard Lee, and Steven Hayward. 2005. “A Comprehensive and Non-Redundant Database of Protein Domain Movements.” *Bioinformatics (Oxford, England)* 21 (12): 2832–38.

Ramachandran, G N, and V Sasisekharan. 1968. “Conformation of Polypeptides and Proteins.” *Advances in Protein Chemistry* 23: 283–438.

Reddy, Chilamakuri CS, Khader Shameer, Bernard O Offmann, and Ramanathan Sowdhamini. 2008. “PURE: A Webserver for the Prediction of Domains in Unassigned Regions in Proteins.” *BMC Bioinformatics* 9 (1): 281. doi:10.1186/1471-2105-9-281.

Redfern, Oliver C., Andrew Harrison, Tim Dallman, Frances M. G. Pearl, and Christine A. Orengo. 2007. “CATHEDRAL: A Fast and Effective Algorithm to Predict Folds and Domain Boundaries from Multidomain Protein Structures.” *PLoS Computational Biology* 3 (11): e232. doi:10.1371/journal.pcbi.0030232.

Richards, F M, and C E Kundrot. 1988. “Identification of Structural Motifs from Protein Coordinate Data: Secondary Structure and First-Level Supersecondary Structure.” *Proteins* 3 (2): 71–84. doi:10.1002/prot.340030202.

Rooman, M J, J Rodriguez, and S J Wodak. 1990. “Automatic Definition of Recurrent Local Structure Motifs in Proteins.” *Journal of Molecular Biology* 213 (2): 327–36. doi:10.1016/S0022-2836(05)80194-9.

Rosano, G. L., & Ceccarelli, E. A. (2014). Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in Microbiology*, *5*, 172. doi:10.3389/fmicb.2014.00172

Rose, Peter W, Andreas Prlić, Chunxiao Bi, Wolfgang F Bluhm, Cole H Christie, Shuchismita Dutta, Rachel Kramer Green, et al. 2015. “The RCSB Protein Data Bank: Views of Structural Biology for Basic and Applied Research and Education.” *Nucleic Acids Research* 43 (Database issue): D345–56. doi:10.1093/nar/gku1214.

Rost, B, and C Sander. 1993. “Improved Prediction of Protein Secondary Structure by Use of Sequence Profiles and Neural Networks.” *Proc Natl Acad Sci U S A* 90 (16): 7558–62.

Rost, Burkhard, and Chris Sander. 1994. “Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure.” *Proteins: Structure, Function, and Genetics* 19 (1). Wiley Subscription Services, Inc., A Wiley Company: 55–72. doi:10.1002/prot.340190108.

Russell, Robert B., and Geoffrey J. Barton. 1992. “Multiple Protein Sequence Alignment from Tertiary Structure Comparison: Assignment of Global and Residue Confidence Levels.” *Proteins: Structure, Function, and Bioinformatics* 14 (2): 309–23. doi:10.1002/prot.340140216.

S. Hubbard and J. Thornton. (n.d.). Naccess V2.1.1 - Atomic Solvent Accessible Area Calculations.

Saldaño, Tadeo E., Alexander M. Monzon, Gustavo Parisi, and Sebastian Fernandez-Alberti. 2016. “Evolutionary Conserved Positions Define Protein Conformational Diversity.” *PLOS Computational Biology* 12 (3): e1004775.

Šali, Andrej, and Tom L. Blundell. 1990. “Definition of General Topological Equivalence in Protein Structures. A Procedure Involving Comparison of Properties and Relationships through Simulated Annealing and Dynamic Programming.” *Journal of Molecular Biology* 212 (2): 403–28. doi:10.1016/0022-2836(90)90134-8.

Sander, Oliver, Ingolf Sommer, and Thomas Lengauer. 2006. “Local Protein Structure Prediction Using Discriminative Models.” *BMC Bioinformatics* 7 (1): 14. doi:10.1186/1471-2105-7-14.

Schneider, R, A de Daruvar, and C Sander. 1997. “The HSSP Database of Protein Structure-Sequence Alignments.” *Nucleic Acids Research* 25 (1). Oxford University Press: 226–30. doi:10.1093/NAR/25.1.226.

Schneider, Thomas D., and R.Michael Stephens. 1990. “Sequence Logos: A New Way to Display Consensus Sequences.” *Nucleic Acids Research* 18 (20): 6097–6100.

Schuchhardt, J, G Schneider, J Reichelt, D Schomburg, and P Wrede. 1996. “Local Structural Motifs of Protein Backbones Are Classified by Self-Organizing Neural Networks.” *Protein Engineering* 9 (10): 833–42. doi:10.1093/PROTEIN/9.10.833.

Seemayer, S., Gruber, M., & Soding, J. (2014). CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, *30*(21), 3128–3130. doi:10.1093/bioinformatics/btu500

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11). Cold Spring Harbor Laboratory Press: 2498–2504. doi:10.1101/gr.1239303.

Shatsky, Maxim, Ruth Nussinov, and Haim J. Wolfson. 2002. “Flexible Protein Alignment and Hinge Detection.” *Proteins: Structure, Function and Genetics* 48 (2): 242–56. doi:10.1002/prot.10100.

Shatsky, M, R Nussinov, and H J Wolfson. 2004. “A Method for Simultaneous Alignment of Multiple Protein Structures.” *Proteins* 56 (1097-0134 (Electronic)): 143–56. doi:10.1002/prot.10628.

Shindyalov, Ilya N, and Philip E Bourne. 1998. “Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path.” *Protein Engineering* 11 (9): 739–47.

Siddiqui, Asim S., and Geoffrey J. Barton. 2008. “Continuous and Discontinuous Domains: An Algorithm for the Automatic Generation of Reliable Protein Domain Definitions.” *Protein Science* 4 (5): 872–84. doi:10.1002/pro.5560040507.

Sievers, Fabian, and Desmond G. Higgins. 2014. “Clustal Omega.” In *Current Protocols in Bioinformatics*, 3.13.1–3.13.16. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471250953.bi0313s48.

Sowdhamini, R, and T L Blundell. 1995. “An Automatic Method Involving Cluster Analysis of Secondary Structures for the Identification of Domains in Proteins.” *Protein Science : A Publication of the Protein Society* 4 (3): 506–20. doi:10.1002/pro.5560040317.

Suresh, V, K Ganesan, and S Parthasarathy. 2012. “A Protein Block Based Fold Recognition Method for the Annotation of Twilight Zone Sequences.” *Protein and Peptide Letters* 655: 249–54.

Suresh, V, Liang Liu, Donald Adjeroh, and Xiaobo Zhou. 2015. “RPI-Pred: Predicting ncRNA-Protein Interaction Using Sequence and Structural Information.” *Nucleic Acids Research* 43 (3): 1370–79. doi:10.1093/nar/gkv020.

Taylor, William R., and Christine A. Orengo. 1989. “Protein Structure Alignment.” *Journal of Molecular Biology* 208 (1): 1–22. doi:10.1016/0022-2836(89)90084-3.

Tung, Chi-Hua, Jhang-Wei Huang, and Jinn-Moon Yang. 2007. “Kappa-Alpha Plot Derived Structural Alphabet and BLOSUM-like Substitution Matrix for Rapid Search of Protein Structure Database.” *Genome Biology* 8 (3). BioMed Central: R31. doi:10.1186/gb-2007-8-3-r31.

Tyagi, M, A G de Brevern, N Srinivasan, and B Offmann. 2008. “Protein Structure Mining Using a Structural Alphabet.” *Proteins* 71 (2): 920–37. doi:10.1002/prot.21776.

Tyagi, M. 2006. *New perspectives for protein structure analysis and mining using sequences of a structural alphabet*. PhD. thesis. http://www.theses.fr/2006LARE0007

Tyagi, M., P. Sharma, C. S. Swamy, F. Cadet, N. Srinivasan, a. G. de Brevern, and Bernard Offmann. 2006. “Protein Block Expert (PBE): A Web-Based Protein Structure Analysis Server Using a Structural Alphabet.” *Nucleic Acids Research* 34: 119–23. doi:10.1093/nar/gkl199.

Tyagi, Manoj, Aurélie Bornot, Bernard Offmann, and Alexandre G. de Brevern. 2009. “Analysis of Loop Boundaries Using Different Local Structure Assignment Methods.” *Protein Science* 18 (9): 1869–81. doi:10.1002/pro.198.

Manoj Tyagi, Venkataraman S. Gowri, Narayanaswamy Srinivasan, Alexandre G. de Brevern, & Bernard Offmann. 2006. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins : Struct. Func. Bioinf. 65(1) :32-39. doi:10.1002/prot.

Unger, R, D Harel, S Wherland, and J L Sussman. 1989. “A 3D Building Blocks Approach to Analyzing and Predicting Structure of Proteins.” *Proteins-Structure Function And Bioinformatics* 5 (4): 355–73. doi:10.1002/prot.340050410.

Varadi, Mihaly, and Peter Tompa. 2015. “The Protein Ensemble Database.” *Advances in Experimental Medicine and Biology* 870: 335–49.

Vehlow, C., Stehr, H., Winkelmann, M., Duarte, J. M., Petzold, L., Dinse, J., & Lappe, M. (2011). CMView: Interactive contact map visualization and analysis. *Bioinformatics*, *27*(11), 1573–1574. doi:10.1093/bioinformatics/btr163

Venter, J C, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, et al. 2001. “The Sequence of the Human Genome.” *Science* 291 (5507): 1304–51. doi:10.1126/science.1058040.

Vinayagam, A, J Shi, G Pugalenthi, B Meenakshi, T L Blundell, and R Sowdhamini. 2003. "DDBASE2.0: Updated Domain Database with Improved Identification of Structural Domains." *Bioinformatics (Oxford, England)* 19 (14): 1760–64.

Voet, Donald, and Judith Voet. 2011. *Biochemistry*. 4th Editio. Wiley.

Wagner, Ingrid, and Hans Musso. 1983. "New Naturally Occurring Amino Acids." *Angewandte Chemie International Edition in English* 22 (11). Hüthig & Wepf Verlag: 816–28. doi:10.1002/anie.198308161.

Wang, G., and R. L. Dunbrack. 2003. "PISCES: A Protein Sequence Culling Server." *Bioinformatics* 19 (12). Oxford University Press: 1589–91. doi:10.1093/bioinformatics/btg224.

Weaver, T M. 2000. "The Pi-Helix Translates Structure into Function." *Protein Science : A Publication of the Protein Society* 9 (1): 201–6. doi:10.1110/ps.9.1.201.

Wernisch, L, M Hunting, and S J Wodak. 1999. "Identification of Structural Domains in Proteins by a Graph Heuristic." *Proteins* 35 (3): 338–52.

Wetlaufer, D B. 1973. "Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins." *Proceedings of the National Academy of Sciences of the United States of America* 70 (3). National Academy of Sciences: 697–701.

Williamson, Michael P., Timothy F. Havel, and Kurt Wüthrich. 1985. "Solution Conformation of Proteinase Inhibitor IIA from Bull Seminal Plasma by 1H Nuclear Magnetic Resonance and Distance Geometry." *Journal of Molecular Biology* 182 (2). Academic Press: 295–315. doi:10.1016/0022-2836(85)90347-X.

wwPDB. (2008). Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description. Version 3.30.

Xu, W, A Doshi, M Lei, M J Eck, and S C Harrison. 1999. "Crystal Structures of c-Src Reveal Features of Its Autoinhibitory Mechanism." *Molecular Cell* 3 (5): 629–38.

Xu, Ying, Dong Xu, and Harold N Gabow. 2000. "Protein Domain Decomposition Using a Graph-Theoretic Approach." *BIOINFORMATICS* 16 (12): 1091–1104. http://.

Yang, Jiaan. 2008. "Comprehensive Description of Protein Structures Using Protein Folding Shape Code." *Proteins: Structure, Function, and Bioinformatics* 71 (3): 1497–1518. doi:10.1002/prot.21932.

Ye, Y., and A. Godzik. 2003. "Flexible Structure Alignment by Chaining Aligned Fragment Pairs Allowing Twists." *Bioinformatics* 19 (Suppl 2). Oxford University Press: ii246–55. doi:10.1093/bioinformatics/btg1086.

Ye, Y., and A. Godzik. 2004. "FATCAT: A Web Server for Flexible Structure Comparison and Structure Similarity Searching." *Nucleic Acids Research* 32 (Web Server): W582–85.

Young, V R. 1994. “Adult Amino Acid Requirements: The Case for a Major Revision in Current Recommendations.” *The Journal of Nutrition* 124 (8 Suppl): 1517S – 1523S. http://www.ncbi.nlm.nih.gov/pubmed/8064412.

Zhang, Yang, and Jeffrey Skolnick. 2005. “TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score.” *Nucleic Acids Research* 33 (7): 2302–9. doi:10.1093/nar/gki524.

Zhou, Hongyi, Bin Xue, and Yaoqi Zhou. 2007. “DDOMAIN: Dividing Structures into Domains Using a Normalized Domain-Domain Interaction Profile.” *Protein Science* 16 (5): 947–55. doi:10.1110/ps.062597307.

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Iyanar VETRIVEL

Exploring new applications of a structural alphabet in protein structure analysis

Nouvelles applications autour de l'usage d'un alphabet structural pour l'analyse des structures des protéines

Abstract

Structural alphabets are a collection of repetitive structural motifs found in protein structures. They help in approximating protein structure as a 1D string with minimal loss of structural information. Here I have employed a widely used structural alphabet called Protein Blocks (PB) for various applications like predicting, comparing and analyzing protein structures.

PBs were used to study the structural variations in proteins with identical primary structure and also during the course of molecular dynamics (MD) simulations. The results from these analyses were summarized in the form of substitution matrices and showed striking similarities to previously established matrices for homologous proteins and NMR ensembles.

I improved kPRED, a knowledge-based prediction of protein backbone in terms of PBs, by taking into consideration the neighboring local structures. The new version of the algorithm also privileges structural information from homologous proteins when available reaching an average accuracy of 66.3% on a benchmark dataset.

The scope of the PENTAdb database has been expanded to cover the entire protein structure space in an automated manner. I show that the effect of this 950% increase in the contents of PENTAdb improves our understanding of the sequence-structure relationship at the pentapeptide level.

I also used PB-ALIGN, a fast and efficient protein structure comparison tool, to compare all protein structures in PDB in an all-vs-all manner and to investigate PB-based structural similarities. This generated a huge collection of alignment data and I discuss its use for functional annotation and identification of possible evolutionary relationships.

Key words: Structural alphabet, Protein Blocks, Protein structure comparison, Protein structure prediction, Pentapeptide, Protein sequence-structure relationship.

Résumé

Les alphabets structuraux sont des collections de motifs structuraux répétés dans les structures de protéines. Durant ma thèse, j'ai utilisé l'un d'eux, les blocs protéiques (BPs), pour diverses applications comme la prédiction, la comparaison et l'analyse de structures de protéines.

J'ai d'abord étudié les variations structurales observées au sein de structures ayant des séquences protéiques identiques et au cours de simulations de dynamique moléculaire. Les résultats de ces analyses ont été synthétisés sous la forme de matrices de substitution et ont montré des similitudes surprenantes avec les précédentes matrices établies pour les protéines homologues ou les ensembles RMN.

La méthode kPRED qui permet de prédire le squelette des protéines sous forme de BPs a été amélioré en prenant en compte les structures locales voisines. La nouvelle version privilégie aussi les informations provenant des protéines homologues lorsqu'elles sont disponibles, atteignant une précision moyenne de 66,3% sur un jeu de données test.

J'ai aussi étendu la portée de la base de données PENTAdb pour couvrir l'ensemble des structures de la PDB. Je montre que l'effet de cette augmentation de 950% dans le contenu de PENTAdb améliore notre compréhension de la relation séquence-structure au niveau d'un pentapeptide.

J'ai également utilisé PB-ALIGN, un outil de comparaison rapide et efficace de structures des protéines, pour comparer toutes les structures protéiques dans la PDB entre elles et pour étudier les similitudes structurales. J'ai généré une grande collection de données d'alignement et j'explore son usage pour l'annotation fonctionnelle et l'identification de possibles relations évolutives.

Mots clés: Alphabet structural, Blocs protéiques, Comparaison des structures, Prédiction de structures protéiques, Pentapeptides, Relation séquence-structure