DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE NANTES

THESE DE DOCTORAT DE

COMUE UNIVERSITE BRETAGNE LOIRE

L'UNIVERSITE DE NANTES

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : Bioinformatique structurale

Par
Surbhi DHINGRA

New Method for Generating Protein Atomic Models using Structural Alphabets in the Absence of Homologs of Known Structures

Thèse présentée et soutenue à Bangalore, le 12 Juin 2020
Unité de recherche : UFIP, UMR CNRS 6286

Rapporteurs avant soutenance :

Anne-Claude CAMPROUX	Professeure, Université de Paris
Juliette MARTIN	Chargée de Recherche HDR CNRS, Université de Lyon

Composition du Jury :

Examinateurs : Srinivasan NARAYANASWAMY	Professeur, Indian Institute of Science (IISc)
Dir. de thèse : Bernard OFFMANN	Professeur, Université de Nantes
Co-dir. de thèse : Frédéric CADET	Professeur, Université de La Réunion
Co-encad de thèse : Sowdhamini RAMANATHAN	Professeure, National Center for Biological Sciences (NCBS)

Acknowledgements

Compiling of a PhD thesis is never an outcome of an individualistic perspective but a culmination of constant effort put together by multiple individuals. I would like to focus on these people who have helped me bring this project to a conclusion.

Let's begin with the person behind inculcating an idea and turning it to a physical project. My sincere gratitude towards Prof. Bernard Offmann for designing such an intriguing research work and being a constant source of guidance until the end of this work. A heartfelt appreciation of Prof. Ramanthan Sowdhamini, who along with being a constant inspiration throughout the run of the thesis, in the past few months have become a undeterred source of strength in the collective difficult times. Thanks a lot for taking extra steps to provide me with a comfortable environment to complete my thesis writing. I would also like to show my gratitude towards Prof. Frédéric Cadet for constantly reading the drafts and suggesting ways for its improvements.

Throughout the run of this project, I have been lucky enough to get a substantial amount of external guidance. Specifically coming from Prof. Alexandre G. de Brevern, Dr. Stéphane Telechea and Dr. Yves-Henri Sanejouand. My sincere acknowledgment to them for being a source of counsel and support.

A heartfelt gratitude towards the examiners, Prof. Anne-Claude Camproux and Dr. Juliette Martin for agreeing to be a part of the committee and taking time out from their routine to indulge in the work discussed below at length.

Lastly, I would like to show my gratitude towards the entire UFIP group for welcoming me and making me a part of the laboratory and constantly supporting me through the thick and thins. Especially, to my colleagues including Alexandre Demeyer, Ennys Gheyouche, Johann Hendrickx, Laura Marin, Thomas Chabot, Typhaine Violo, Udita Chandola, Xue Zhao who have been wonderful colleagues to work and share this experience with while gaining generous amount of laughs on numerous occasions.

A special thanks to Mohak Sharda, Saurabh Gupta and Shilpa Nair for being a light in dimmest of nights.

In the end, 2020 being a challenging year for all, I hope that humanity will soon recover from its plight and in not so far ahead future regain normalcy.

The following institutions are thanked for their generous support.









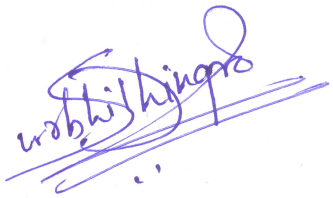


Declaration

I hereby certify that the work documented in this thesis is carried out by me over the duration of this project. This project has been done under the supervision of Prof. Bernard Offmann at the laboratory UFIP located inside Université de Nantes, Nantes, France; and has been co-supervised by Prof. Ramanathan Sowdhamini, NCBS, Bangalore, India and Prof. Frédéric Cadet, Université de La Réunion and Université de Paris, INSERM, Paris, France.

I also conform that this work has not been submitted for any other degree or qualification except the one specified here.

Date : 04-May-2020



Surbhi Dhingra

In dedication to the place and people around whom I have been able to complete this work, ***UFIP****.*

Life is not a kiwi or a kaki. But one day it shall be...:)

- Cafeteria, UFIP

Synopsis (version française)

Les protéines sont à la base de l'exécution de diverses fonctions essentielles à la « vie moléculaire » au sein d'une unité cellulaire. Les briques élémentaires dont la combinaison conduit à la formation de ces biomolécules sont les acides aminés. L'assemblage de ces briques selon un ordre dicté par l'information génétique conduit à son tour à la formation de la structure primaire de la protéine, sa séquence. Une quantité massive de données de séquences s'est accumulée ces dernières décennies avec le séquençage des génomes et leur déchiffrement. Des progrès similaires ont été observés s'agissant de l'élucidation de la structure secondaire des protéines soit par des approches expérimentales ou prédictives.

Le troisième niveau de représentation de la structure des protéines est appelée sa structure tertiaire. Ici, la chaîne protéique se replie en une conformation stable qui, en combinaison avec d'autres chaînes, peut former une structure quaternaire. Ce repliement est généralement associée à sa fonction. Si nos connaissances des mécanismes de formation de la structure tertiaire des protéines ont énormément progressé, la résolution du « problème du repliement des protéines » reste un défi majeur en biologie. De nombreuses recherches ont été menées dans cette direction avec un énorme succès avec la conception d'algorithmes rapides et efficaces pouvant modéliser des protéines à partir de structures connues (modélisation par homologie ou comparative).

Un autre volet traite de la prédiction de structures de protéines lorsqu'aucun homologue connu n'est documenté dans la base de données des structures protéiques. Ce domaine de recherche a connu un essor considérable au cours des dernières décennies avec le développement de plusieurs approches prometteuses. Dans une certaine mesure, ils ont également aidé à cartographier les courbes d'énergie associées au repliement protéiques. Ce domaine de la prédiction de structures des protéines est généralement connu sous le nom de stratégies de modélisation sans matrice (MSM) ou en anglais *Template Free Modelling* (TFM) ou encore modélisation *ab initio*.

Les premiers efforts dans ce domaine se sont concentrés sur le développement d'algorithmes reposant sur des principes de physique. Mais avec le temps, le champ s'est élargi avec un afflux continu d'idées visant à élaborer de nouvelles stratégies. Au cours des deux dernières décennies, des évolutions constantes ont été observées dans ce domaine mais la longueur des chaînes pouvant être prédites reste un enjeu. **Dans la première partie du chapître 1 de cette thèse**, nous rendons ainsi compte de l'évolution du domaine de la prédiction *ab initio* des structures des protéines et des stratégies développées dans ce domaine au cours de ces dernières années.

Le projet de thèse actuel vise à explorer une nouvelle stratégie permettant de contourner certains des inconvénients rencontrés lors de la conception des protocoles de prédiction des structures des protéines. **L'idée principale derrière ce projet est centrée sur l'exploration de l'utilisation des blocs protéiques (BP), un**

type d'alphabet structural (AS), pour construire une nouvelle approche de prédiction de la structure des protéines en l'absence d'homologues structuraux connus. Les BPs sont des outils majeurs qui ont été largement étudiés et utilisés pour développer des applications autour de l'analyse des structures des protéines, notamment au sein de notre laboratoire. **Nous avons ainsi fourni dans une deuxième partie du chapitre 1**, leurs principales caractéristiques, leur utilisation et un aperçu général de la stratégie qui a été conçue et suivie pour leur utilisation dans l'exécution du projet de thèse en cours.

En bref, les BPs sont un ensemble de 16 prototypes structuraux qui ont été standardisés pour approximer chaque partie du squelette local de la structure protéique via des descripteurs d'angles dièdres (ϕ et ψ) avec une déviation moyenne de 0,42 Å. Chaque bloc est long de 5 résidus d'acides aminés et est capable de définir une partie locale de la conformation, y compris les régions de transition entre les principaux types de structure secondaire. Plusieurs études ont également indiqué le potentiel des AS dans l'élaboration de nouveaux protocoles de prédiction de structure *ab initio*.

Avoir les BPs comme descripteurs structuraux pouvant sous-diviser une structure protéique en plus petits segments qui se chevauchent offre un moyen de réduire la recherche au sein de l'espace conformationnel de solutions structurales en nombre gérable. En d'autres termes, au lieu de travailler à la minimisation globale de l'énergie d'une chaîne protéique de pleine longueur, avec les BP, il est possible de se concentrer sur la minimisation de petits segments locaux de séquence au sein de la protéine réduisant ainsi considérablement la dimension de l'espace de recherche ainsi que le coût en terme de calculs à effectuer.

Comme détaillé à la fin du chapitre 1, l'hypothèse principale derrière ce projet consiste à supposer que les BPs, en tant que type de descripteur de structure locale, peuvent se révéler être un outil efficace pour l'échantillonnage de l'espace conformationnel et l'identification de courts prototypes structuraux pertinents pour le développement d'un nouveau protocole de modélisation en l'absence d'homologie de séquence avec des structures connues. À cet égard, **l'objectif global de la thèse a consisté à établir un workflow pour la prédiction** *ab initio* **des structures protéiques en s'appuyant sur la recherche de ces fragments structuraux et leur assemblage**.

Le chapitre 2 de cette thèse est dédié à la première brique du workflow susmentionné. Il rend compte de la façon dont les BPs ont été utilisés pour créer des bibliothèques pertinentes de fragments structuraux nécessaires à la réalisation de la modélisation moléculaire dans une étape ultérieure des travaux. Les détails du développement d'une procédure pour générer des bibliothèques de fragments pour une séquence de protéine cible sont fournis dans ce chapitre. En bref, il repose sur le fait que les BPs sont capables de décrire les caractéristiques de la structure locale d'une structure protéique sous forme d'une séquence de BPs en 1D facilement compréhensible (procédure d'assignation des blocs protéiques). De même, grâce à des approches basées sur les connaissances, une méthodologie connue sous le nom de PB-kPRED a été normalisée dans un travail antérieur effectué dans l'équipe pour prédire les caractéristiques structurales et locales d'une protéine directement à partir de sa séquence primaire et l'afficher sous forme d'une séquence de BPs. Avec ces deux applications en main, la construction de la bibliothèque de fragments a été réalisée en utilisant un outil d'alignement local par paire de séquences de BPs

appelé PB-ALIGN. Une base de données représentative de structures protéiques de la PDB a été conçue selon les exigences de ce type de travaux notamment par la réduction maximale de la redondance de l'information. Une longueur minimale de 7 BPs consécutifs (c'est-à-dire 11 résidus d'acides aminés) a été choisie comme seuil de sortie pour l'outil d'alignement local de BPs, cette longueur étant suffisante pour décrire complètement tout type de structure secondaire. La procédure élaborée dans ce cadre a été exécutée sur un ensemble de données de validation comprenant 43 protéines monomériques de structures connues d'une longueur comprise entre 70 à 500 résidus et englobant les 4 principales classes structurales de SCOP. La procédure a été exécutée deux fois, une fois avec les séquence de BPs attribuées (assignation des blocs protéiques) comme contrôles positifs et la seconde fois avec les séquences BPs prédites produites à partir du serveur PB-kPRED.

Les fragments identifiés par comparaison des séquences de BPs avec la base de données de structures ont été traités comme des « hits » potentiels. Pour chaque séquence de protéines de requête, un très grand nombre de fragments a été à chaque fois obtenu. Pour réduire leur redondance et rassembler les fragments les plus représentatifs, seules les séquences de BPs identiques entre la protéine requête et les fragments identifiés ont été retenues pour former une bibliothèque de fragments.

La qualité de la bibliothèque de fragments a été évaluée sur la base des deux critères suivants : la précision globale et la couverture. Au total, une précision moyenne meilleure que 2 Å de *rmsd* a été obtenue pour 66,70 % des fragments pour BPs prédits et ces fragments couvraient en moyenne 93,9 % des positions de la requête.

Une recherche de critères objectifs de qualité pour détecter les meilleurs fragments, notamment les plus proches de ceux de la protéine requête, a été effectuée. Cela comprenait le calcul de la similarité des séquences d'acides aminés et de l'identité de la structure secondaire partagée entre la requête et le fragment identifié et ce pour chaque position. Nos résultats ont montré que les meilleurs fragments ont tendance à partager un degré plus élevé de similitude structurale au niveau secondaire avec la protéine cible que la partie de la séquence d'acides aminés de celle-ci. Cela peut être expliqué par la nature des BPs qui sont la représentation des repliements locaux des protéines montrant ainsi une corrélation plus élevée avec l'identité de la structure secondaire. A travers la procédure mise en place, nous avons ainsi démontré notre capacité à produire une bibliothèque de fragments structuraux pertinents à partir de protéines non apparentées.

L'étape que nous avons ensuite abordée a consisté à travailler à l'assemblage de fragments dans un protocole de modélisation moléculaire pour la prédiction de la structure des protéines *ab initio*. **Ceci a été largement discuté dans les chapitres 3 et 4** de cette thèse.

Comme indiqué dans le chapitre 3, plusieurs expériences ont été conçues et testées à l'aide de l'outil de modélisation comparative *Modeller*. Les expériences ont été conçues pour forcer les limites fondamentales de *Modeller* en ce qui concerne la nécessaire similitude de la séquence d'acides aminés partagée entre la protéine cible et les matrices structurales ainsi que la nécessité d'avoir des matrices de pleine longueur pour générer des modèles de bonne qualité. Sachant que le protocole de génération de fragments développé dans la première partie de la thèse (chapitre 2) produisait des fragments d'une longueur d'environ 7 BPs (soit 11 résidus d'acides aminés), il était presque impossible de fournir à *Modeller* de longues matrices qui

englobent la toute la longueur de la requête ou une proportion élevée de celle-ci. Au lieu de cela, de courts fragments ont été obtenus et utilisés. Ils ont été manuellement disposés dans le fichier d'alignement de *Modeller* de telle sorte que les descriptions de coordonnées disponibles couvraient finalement l'intégralité de longueur de la protéine requête. Ceci a été réalisé en traitant chaque fragment comme une pseudo-matrice et a été fourni à *Modeller* sous la forme d'un fichier d'alignement de matrices multiples.

Pour mieux comprendre l'impact de l'alignement des fragments sur la qualité de la modélisation, deux expériences ont été exécutées. La première consistait à fournir à *Modeller* des alignements de fragments qui partageaient des régions chevauchantes entre deux fragments adjacents. La seconde consistait en des alignements de fragments adjacents qui n'avaient aucun chevauchement entre eux. Conformément à nos attentes, les modèles construits à l'aide de fragments se chevauchant étaient plus proches de la structure native de la protéine que ceux produits à l'aide de fragments sans chevauchement.

Une dernière expérience a été menée en utilisant des alignements de fragments chevauchants mais où les fragments sont issus du pipeline de génération de fragments avec comme point de départ, une séquence de BPs prédite par PB-kPRED. Un modèle déviant de 3,25 Å a été généré. Cela nous a donc permis d'établir un protocole standardisé pour assembler des fragments au sein de *Modeller* en vue du développement d'un workflow complet pour la modélisation *ab initio* des structures protéiques.

Fort de ces expériences préliminaires d'assemblage de fragments (chapitre 3), une évaluation rigoureuse de l'ensemble du protocole standardisé de modélisation *ab initio* a été réalisée. **Le chapitre 4 de la thèse** rapporte cette évaluation. Il plonge également au coeur des différentes parties du protocole utilisé et dans l'interprétation des résultats obtenus. Le protocole a été vérifié sur jeu de 10 petites protéines sélectionnées dans la liste des cibles de la récente compétition CASP13. Une séquence de BPs a été affectée à chacune d'elle (ou prédite pour) et a été utilisée pour générer une bibliothèque de fragments.

De chacune des bibliothèques ainsi générés, des ensembles de fragments ont été sélectionnés de telle sorte que chacun des fragments adjacents partageait entre eux autant que possible un chevauchement. Une attention particulière a été portée afin que soit utilisé autant que possible plusieurs fragments par position dans le fichier d'alignement afin de consolider la possibilité d'échantilloner des conformations alternatives. De plus, deux contraintes externes ont également été ajoutées aux scripts *Modeller* pour guider la modélisation du repliement protéique. D'une part, des prédictions de structures secondaires sous la forme d'hélices α et de brins β ont été transmises comme contraintes principales pour guider le repliement au niveau local. D'autre part, des prédictions de contacts inter-résidus longues distances ont été transférées sous forme de contraintes de distance aux scripts de modélisation pour guider le repliement au niveau global.

Les modèles générés via ce protocole ont été analysés pour leur similitude structurale avec la structure d'origine en utilisant le score global de test de distance (GD_TS) abondamment utilisés ainsi que le *rmsd*. Les meilleurs modèles déterminés à partir de ces deux critères ont à leur tour été affinés à l'aide du serveur web GalaxyRefine2. Cela a fait passer le score GDT_TS moyen des meilleurs modèles de 52,36 à 56,36.

Le dernier chapitre de cette thèse (Chapitre 5) récapitule le travail effec-

tué et fournit une conclusion générale. Il discute des jalons atteints et spécule sur les perspectives ouvertes par cette thèse. En bref, l'objectif de la thèse a été de définir un nouveau processus de prédiction de la structure des protéines *ab initio* en utilisant des blocs protéiques, un type d'alphabet structural. Nous avons pu formuler un protocole préliminaire pour la prédiction de structure et l'avons porté à un niveau de précision moyenne à élevée pour de petites protéines. À l'avenir, une fonction de score pourrait être conçue pour regrouper les fragments afin d'améliorer la sélection des meilleurs représentants. Le protocole de modélisation pourrait également inclure des routines d'optimisation plus sophistiquées notamment celles natives de *Modeller* pour augmenter la précision des modèles.

Mots clés: Protéines, Prédiction des structures des protéines, Modélisation sans matrice (MSM), modélisation *ab initio*, alphabets structuraux, Blocs protéiques.

Synopsis (English version)

Protein form the basis of carrying out various essential functions to thrive the molecular life within a single celled unit. The individual units whose combination leads to the formation of these biomolecules are called amino acids. A combination of these micro-units in turn leads to the formation of the primary structure of the protein which is its sequence. A massive amount of research has been successfully implemented into deciphering proteins in this state. Similar gains in research has been observed in the secondary structure determination of the proteins. This combines description of secondary structural units, depiction of secondary structures in 1D sequences as well as their predictions possible through its primary sequence.

The third stage of protein description is called its tertiary structure. Here, the protein chain folds into stable conformation which in combination with other chains can form a quaternary structure. This last phase of protein state is usually associated to its function. A gap exists in our knowledge of protein sequence formation and its mechanism of folding. This had been identified as the Protein folding problem. A massive amount of research has been driven towards this direction with enormous success gained in designing fast and efficient algorithms that can model protein with known structures (Homology/Comparative Modelling).

The other side of this coin deals with the prediction of new protein models or models which have no known homologues documented in the structural database. This field of research has seen a rise in the past few decades with several promising approaches put forward to efficiently determine protein structures of novel proteins. To some extend they have also helped in mapping out the energy curves associated with protein folding. This field of structure prediction is usually characterized under the name of Template Free Modelling (TFM) strategies.

Initial efforts in this field were focused on development of algorithms that relied on physics based principles of protein folding. But with time the field broadened by continuous influx of ideas revolving around newer strategies to determine the protein folding paradigm. In the past two decades a constant evolution has been witnessed in this field, although it is still lagging behind majorly by the length up to which a high resolution structure can be predicted. A detailed discussion on the evolution of the field of protein structure prediction and the strategies developed under it throughout the years **is documented in the first part of Chapter 1 of this thesis**.

The current thesis project was proposed with regard to the potential of a novel strategy that could ultimately work around some of the drawbacks encountered during the design of protein structure prediction protocols. **The main idea behind this project was centered around the exploration of the usage of Protein Blocks (PBs), a type of Structural Alphabet (SA), to build a novel approach of protein structure prediction in the context of absence**

of any known structural homologues. PBs are a major resource that has been extensively studied, optimized and used for developing protein structure related applications in our laboratory. We have detailed their characteristics, how they were designed and we provided a general overview of the strategy that was designed and followed for their use in the execution of the current thesis project.

In short, PBs are a set of 16 structural prototypes which have been standardized to approximate every part of the local backbone of the protein structure via dihedral angles (ϕ and ψ) descriptors up to an average deviation of 0.42 Å. Each block is 5 amino acid residues in length and is capable of defining a part of local conformation including the transition regions between the three major secondary structure types. Several studies have also indicated towards the potential of SAs in designing conformationally accessible *ab initio* structure prediction protocols, which is the major limitation of this field.

Having structural descriptors like PBs which can distribute (sub-divide) the definition of a protein structure into smaller stretches of overlapping segments offers a well-grounded way to reduce the huge dimension of the search of the conformational space into decipherable sizes. In other words, instead of working on achieving the overall energy minimization of a full length protein sequence, focusing on minimizing small local stretches of sequence within the protein significantly reduces the search space dimension along with the computational cost towards this objective.

As detailed in the end of chapter 1, the main hypothesis behind this project consists in assuming that PBs, as a type of local structure descriptor, can prove to be an efficient tool for developing template free modelling protocols, by attempting to resolve the efficient sampling of conformational space for short structural templates in absence of known sequence homology, a major set back experienced by the algorithms designed in this field. In that respect, the overall objective of the thesis consisted in establishing a workflow for the *ab initio* prediction of protein structures relying on the search for these structural fragments and their assembly.

Chapter 2 of this thesis is dedicated to the first brick of the above-mentioned workflow. It reports how PBs were utilized for building pertinent libraries of structural fragments needed to carry out molecular modelling in a subsequent stage of the work. The details of the development of a procedure to generate fragment libraries for a given target protein sequence is provided. In short, it relies on the fact that it has already been established that PBs are able to describe the local structure features of a protein structure in a readily comprehensible 1D PB sequence format (Protein Block Assignment). Similarly, through knowledge-based approaches, a methodology known as PB-kPRED was standardized in an earlier work done in the team to predict the local structural features of a protein directly from its primary sequence and display it in for of a PB sequence. Having these two applications at hand, the fragment library construction was achieved using a PB-based local pairwise sequence alignment tool called PB-ALIGN. A structural database was setup by clustering the sequence data available from the PDB database. This helped in establishing a representative database of protein structures designed around the requirements of the work undertaken in this thesis while reducing the redundancy of information at an initial stage. A minimum length of 7 consecutive PBs (i.e 11 amino acid residues) was settled upon as an output produced by the local PB alignment tool as it was an appropriate length to completely describe any type of secondary

structure. The pipeline was ran on a validation dataset comprising of 43 monomeric query proteins ranging in length from about 70 to 500 residues and encompassing the 4 main SCOP structural classes. The same dataset of proteins was ran twice, once with PB sequence assigned to the test proteins which acted as a positive control in the standardization process and the second was with the predicted PB sequences produced from PB-kPRED web server.

Local hits produced by the pairwise comparison of the test PB sequences against the representative database were treated as potential fragment hits. A large pool of hits were generated for each query protein sequence and to reduce the redundancy of data produced as well as to gather the best possible representative fragments per query protein, the alignment data comprising of PBs hits identical to the PB sequence of the query were further considered. They formed the whole set of fragment library generated for each target protein sequence.

The quality of fragment library was confirmed on the basis of the following two criteria: the overall precision and the coverage which highlighted the overall efficiency of the pipeline. Altogether, an average precision better than 2Å *rmsd* was achieved for 66.70% of the fragments for the predicted PB runs and these fragments covered on average 93.9% of the query positions.

An objective scheme was tested to detect the fragments closest to the query protein. It involved calculation of amino acid sequence similarity and secondary structure identity shared between the query and the template hit for any position. Our results showed that the fragment hits tend to share higher degree of secondary structural similarity with the target protein than the amino acid sequence part of it. This stood true due to the nature of PBs which are the structural depiction of a protein and being optimized for the dihedral definitions of the same, thereby showing higher correlation with secondary structure identity. With these, we demonstrated our ability to produce a library of meaningful structural fragments from unrelated proteins.

The next evident step that we addressed was to work towards assembling fragments within a molecular modelling protocol for *ab initio* protein structure prediction. **This has been extensively discussed in Chapters 3 and 4** of this thesis.

As reported in Chapter 3, several experiments were designed and tested on the prominently known comparative modelling tool called Modeller. The experiments were designed challenging the basic limitations of Modeller with respect to the shared amino acid sequence similarity between the target protein and the structural templates as well as the necessity to have full length templates for generating good quality models. Since the fragment generation protocol developed in the earlier part of the work (Chapter 2) produced fragments which were majorly around 7 PBs in length (i.e 11 amino acid residues), it was close to impossible to provide very long templates to Modeller that encompass the full length of the query or a high proportion of it. Instead, short fragments were used. They were arranged in the alignment file such that the available coordinate descriptions eventually covered the maximum length of the query protein. This was achieved by treating each fragment as a pseudo-template which was delivered to Modeller in the form of multiple template alignment file. On the top of it, to figure out the impact of gaps while delivering fragments in this format, two subsets of the same experiments were run. One composed of fragment hits that shared a flanking region (sequence overlap)

between two adjacent fragments and the second one where the fragments had no sequence overlaps between them. As per our notion, models built using overlapping fragments were closer to native structure of the protein than the ones produced using non-overlapping fragments. A final experiment was conducted using fragments generated from the PB based pipeline employed for predicting the protein structure using an overlapping fragment protocol. A model deviating by 3.25Å was generated by following the protocol with overlapping fragments. This let us to establish a standardized protocol for assembling fragments within Modeller towards our objective to develop a workflow for *ab initio* modelling of protein structures.

Taking heed from these preliminary fragment assembly experiments (Chapter 3), a rigorous evaluation of the entire standardized *ab initio* modelling protocol was performed. **Chapter 4 of the thesis** reports this evaluation. It also dives deep into the different parts of the protocol used and into the interpretation of the results obtained through the runs. Confining its applicability on a set of small proteins, the protocol was verified on 10 small proteins that were selected from the target list of the recently conducted CASP13 biannual competition. A PB sequence were assigned to each of the query in the test dataset which were in turn used to generate fragment library for each protein.

For each target protein, sets of fragments were selected from the fragment library such that each of the adjacent fragments shared as much as possible a sequence overlap. It was also made a point to use multiple fragments per position within the alignment file to consolidate the possibility of encountering alternate conformations. In addition to these measures, two external restraints were also added to the Modeller scripts to guide the modelling of the protein fold. Secondary structure predictions in the form of α-helices and β-strands were passed as the primary restraints to enforce the secondary structure type at the predicted positions. Similarly, to guide the protein folding pattern within the models, predicted inter-residue contacts were transferred in the form of distance restraints to the modelling scripts.

The models generated via the above protocol were analysed for their structural similarity to the original structure utilising the abundantly used global distance test score (GD_TS) criteria along with *rmsd* calculations. The best models determined through these two criteria were in turn refined using the GalaxyRefine2 web server. This increased the average GDT_TS score of the best models from 52.36 to 56.36.

The last chapter of this thesis (Chapter 5) recapitulates the work done and provides a general conclusion. It discusses on the milestone reached and speculates the further perspectives opened by this thesis. In short, the aim of the thesis was to define a new process for *ab initio* protein structure prediction by making use of protein blocks, a type of structural alphabet. We were able to formulate a preliminary protocol for structure prediction and took it up to a level of medium to high prediction accuracy of small proteins. In the future, a scoring function could be designed to cluster fragments in order to get the best representatives at each position. The modelling protocol could also be further optimized to increase the model prediction accuracy of Modeller.

Keywords: Protein, Protein Structure Prediction, Template Free Modelling (TFM), *ab initio* modelling, Structural Alphabets, Protein Blocks.

List of Publications

- A glance into the evolution of template-free protein structure prediction methodologies. Accepted in *Biochimie*.
  Surbhi Dhingra, Ramanathan Sowdhamini, Frédéric Cadet, Bernard Offmann
  Preprint available at `https://arxiv.org/abs/2002.06616`

- Customised fragment libraries for *ab initio* protein structure prediction using a structural alphabet. Submitted to *Bioinformatics* (Oxford University Press).
  Surbhi Dhingra, Ramanathan Sowdhamini, Yves-Henri Sanejouand, Frédéric Cadet, Bernard Offmann
  Preprint available at `https://arxiv.org/abs/2005.01696`

- Fragment based *de novo* modelling of small proteins using external restraints in Modeller. *In preparation.*
  Surbhi Dhingra, Ramanathan Sowdhamini, Frédéric Cadet, Bernard Offmann

Contents

1 Introduction **1**
- 1.1 Preamble . . . 1
- 1.2 Why predict protein structure? . . . 2
  - 1.2.1 Sequence-structure-function paradigm . . . 2
  - 1.2.2 Evolution of Template Free Modelling protocols . . . 3
- 1.3 Structural Alphabets . . . 21
- 1.4 Protein Blocks . . . 22
  - 1.4.1 Introduction . . . 22
  - 1.4.2 Applications . . . 23
- 1.5 Hypothesis . . . 26

2 Fragment generation protocol **28**
- 2.1 Why fragment generation? . . . 28
- 2.2 Our strategy . . . 30
  - 2.2.1 Protein Blocks based fragment generation pipeline . . . 31
- 2.3 Other results and in-depth discussion . . . 40
  - 2.3.1 Fragment quality assessment . . . 40
  - 2.3.2 Sequence similarity and Secondary structure identity distributions . . . 41
  - 2.3.3 ROC Curve Analysis . . . 47
- 2.4 Uniqueness of the pipeline . . . 47

3 Modeller for assembling fragments: preliminary experiments **50**
- 3.1 Fragment-based modelling . . . 50
- 3.2 Modeller for protein structure prediction . . . 51
- 3.3 Experiments performed using Modeller . . . 52
  - 3.3.1 Fragments from the same protein . . . 53
  - 3.3.2 Fragments from close sequence homologs . . . 54
  - 3.3.3 Fragments from multiple homologous proteins . . . 56
- 3.4 Fragments from non-homologous proteins . . . 57
- 3.5 Inference . . . 59
  - 3.5.1 Strengths and weaknesses . . . 60

4 Fragment based prediction of protein structure **62**
- 4.1 External restraints . . . 62
  - 4.1.1 Secondary structure prediction . . . 63
  - 4.1.2 Contact predictions . . . 64
- 4.2 Model Refinement . . . 65
- 4.3 Modelling Protocol . . . 67

4.4 Concluding Remarks 81

5 Conclusions and Perspectives **83**
5.1 Conclusions 83
5.2 Perspective 88

A Annexures **90**
A.1 Chapter 2 - Supplementary files 90
A.2 Chapter 4 - Supplementary files 109

List of Figures

1.1 The sequence-structure-function relationship. 2
1.2 Principle of precise encoding of a protein 3D structure information into a simplified 1D representation using a structural alphabet. 22
1.3 Representation of Protein Blocks. 23
1.4 Dihedral angles distribution of each protein block (a to p). 24
1.5 Principle of structure comparison by PB sequence alignment. 25
1.6 General overview of the template-free modelling protocol developed in this thesis 26

2.1 Histograms for SCOP Class all α 42
2.2 Histograms - SCOP class - all-β 43
2.3 Histograms - SCOP class $\alpha + \beta$ 45
2.4 Histograms - SCOP class - a/b 46
2.5 ROC curve 48

3.1 Query as template fragment based modelling 54
3.2 Single template fragment based modelling 55
3.3 Multiple template fragment based models. 57
3.4 Cartoon representation of fragments superimposed over original structure 58
3.5 Modeller alignment format 59
3.6 1CRN - Structure Prediction 60

List of Tables

1.1 List of few examples of different types of structural alphabets (SAs) documented in the literature. 21

2.1 Fragment generation protocols. 29
2.2 SCOP Class - RMSD distribution 40
2.3 SCOP Class - all α 41
2.4 SCOP Class - all β. 43
2.5 SCOP class $\alpha + \beta$ 45
2.6 SCOP class α / β 46

3.1 Fragmented arrangement of the single template protein 55
3.2 Description of templates chosen for modelling the protein 56

4.1 Refinement Protocols 66

List of Abbreviations

1D	One dimensional
3D	Three dimensional
AUC	Area under the curve
CASP	Critical Assessment of protein Structure Prediction
CATH	Class Architecture Toplogy/fold Homologous superfamily
CM	Comparative modelling
Cryo-EM	Cryogenic electron microscopy
DMP	DeepMetaPSICOV
FM	Free modelling
GDT_TS	Global distance test total score
HM	Homology modelling
HMM	Hidden markov model
MD	Molecular dynamics
MELD	Modelling Employing Limited Data
MS	Mass spectroscopy
NMR	Nuclear Magnetic Resonance
NOE	Nuclear overhauser effect
PB	Protein Blocks
PBA	Protein Block assignment
PBP	Protein Block prediction
PDB	Protein data bank
PSIPRED	PSI-blast based secondary structure PREDiction
PSP	Protein structure prediction
RCSB	Research Collaboratory for Structural Bioinformatics
REMD	Replica exchange molecular dynamics
RMSD	Root mean square deviation
RMSDA	Root mean square deviation on angle
ROC curve	Receiver operating characteristic curve

SA	Structural alphabets
SCOP	Structural Classification of Proteins
SI	Sequence identity
SOM	Self Organising Maps
SS	Sequence similarity
SSE	Secondary structure elements
SSI	Secondary structure identity
SW	Structural words
TBM	Template based modelling
TFM	Template free modelling
TrEMBL	Translation of EMBL nucleotide sequence database
TM-score	Template modeling score
UniProtKB	Universal Protein Resource Knowledgebase
X-Ray	X-radiation

To avoid intellectual sterility, it is wise to periodically reexamine our most deeply held presuppositions.

- Alice S. Rossi

1 Introduction

1.1 Preamble

Proteins, as large complex biomolecules, form one of the critical entities needed for the functioning of most life forms. They contribute to the major part of the work carried out in the single cellular unit of life including construction and maintenance of its structure, generating immunity response, carrier of information in the form of messenger proteins, transportation of small molecules etc. Each protein is formed by a combination of 20 recurrent standard amino acids connected via peptide bonds to form long chains which together have functional implications. The genetically encoded combination of these 20 amino acids to form a protein sequence is responsible for its distinctive three dimensional structure in a given environment and henceforth its function. This depicts the well known sequence-structure-function triangular relationship. Keeping this in mind, a huge amount of scientific labour has been put into determining its properties.

Great successes have been achieved in deconstructing the transcriptional, translational and post-translational pathways at the basis of the conception of proteins. With the emergence of genome sequencing techniques, astronomical data have been gathered about the diversity of protein sequences in terms of their amino acid content, length range and numbers.

Research further down the same path deals with the structural analysis of proteins to answer the basic questions like how proteins fold into a specific structure, the effects of mutations on the folding of the structure, number of recurrent folds existing in the nature etc. Knowledge of these aspects of proteins help in understanding how they function as macromolecules and makes it possible to use this information for fundamental understanding of biological diversity and processes and for various applications like drug discovery, protein engineering etc.

Initially a major part of structural data at atomic level was generated by using the existing experimental techniques like X-ray crystallography and nuclear magnetic resonance (NMR). Eventually, the increase in the structural data available for different kinds of proteins and the evolution of sequencing platforms sparked

initiatives for the development of computational based protein structure determination. The general aim of these techniques was to produce 3D models for proteins sharing enough sequence similarities. This gave rise to computational approaches for protein structure prediction encompassing homology modelling and comparative modelling approaches resulting in a marked availability of structural data. Despite the progress in this field, a major part of documented sequence data has no known correspondence to known structural information. To map these territories, a new field of computational structure prediction appeared with the aim to determine protein structure by solely relying on its sequence.

1.2 Why predict protein structure?

1.2.1 Sequence-structure-function paradigm

The studies revolving around proteins have mainly been focused on three aspects: (i) determination of its sequence by the combination of various amino acid, (ii) ability to determine a stable fold from its sequence and (iii) deciphering the underlying functioning associated with the macromolecule. This is termed as the *sequence-structure-function* paradigm (Whisstock et al. 2003; Feig 2017) and is illustrated in Figure 1.1.

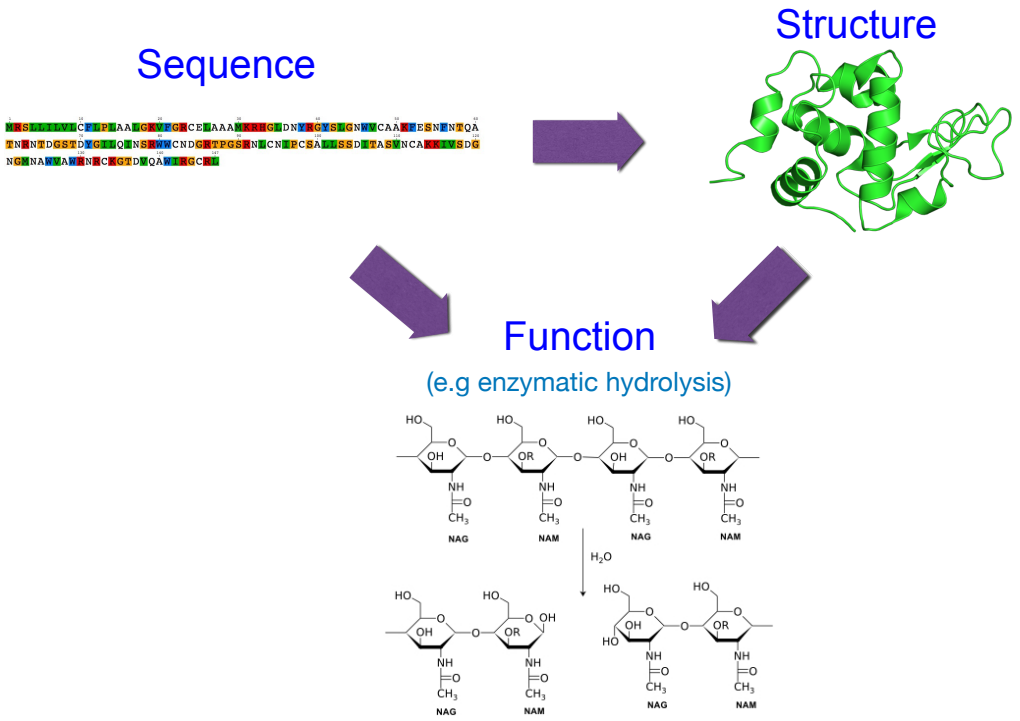


Figure 1.1: The sequence-structure-function relationship.

Shown is the example of lysozyme, a hydrolytic enzyme that degrades polysaccharides.

Since Anfinsen's seminal work, it has already been established that a protein

sequence contains all the needed information of how it folds into its native three-dimensional (3D) structure in a given environment (Anfinsen 1972, 1973). It has also been confirmed that proteins fold into limited number of structural motifs conferring them specific functionalities (Govindarajan et al. 1996, 1999).

The sequence part of proteins has been much worked on and explained in much details. Now, with the help of sequencing techniques, the whole transcriptome and proteome of an organism can be gathered with minimum of sequencing efforts. Similar achievements have been made in determining protein structures based on the advancements made in homology modelling protocols (Webb et al. 2016; Waterhouse et al. 2018).

Despite all the progress made in the above described fields, the answer to the basis behind the second genetic code (Kolata 1986), i.e the folding of proteins, is far from being completely solved.

Eventually it also let to the establishment of another branch of protein structure prediction which focused on building models for proteins with no known structural homologues available. A deeper dive into this domain has been taken in the next section in the form of a review paper.

1.2.2 Evolution of Template Free Modelling protocols

The evolution of Template Free Modelling strategies in the scientific literature has been documented in the attached review paper that was deposited in the ArXiv preprint server, peer-reviewed and **accepted in the Biochimie (Elsevier) international journal**.

A glance into the evolution of template-free protein structure prediction methodologies

Surbhi Dhingra[1], Ramanathan Sowdhamini[2], Frédéric Cadet[3,4,5], and Bernard Offmann*[1]

[1]Université de Nantes, CNRS, UFIP, UMR6286, F-44000 Nantes, France
[2]Computational Approaches to Protein Science (CAPS), National Centre for Biological Sciences (NCBS), Tata Institute for Fundamental Research (TIFR), Bangalore 560-065, India
[3]University of Paris, BIGR—Biologie Intégrée du Globule Rouge, Inserm, UMR_S1134, Paris F-75015, France
[4]Laboratory of Excellence GR-Ex, Boulevard du Montparnasse, Paris F-75015, France
[5]DSIMB, UMR_S1134, BIGR, Inserm, Faculty of Sciences and Technology, University of La Reunion, Saint-Denis F-97715, France

Abstract

Prediction of protein structures using computational approaches has been explored for over two decades, paving a way for more focused research and development of algorithms in comparative modelling, *ab intio* modelling and structure refinement protocols. A tremendous success has been witnessed in template-based modelling protocols, whereas strategies that involve template-free modelling still lag behind, specifically for larger proteins ($>$ 150 a.a.). Various improvements have been observed in *ab initio* protein structure prediction methodologies overtime, with recent ones attributed to the usage of deep learning approaches to construct protein backbone structure from its amino acid sequence. This review highlights the major strategies undertaken for template-free modelling of protein structures while discussing few tools developed under each strategy. It will also briefly comment on the progress observed in the field of *ab initio* modelling of proteins over the course of time as seen through the evolution of CASP platform.

This paper is dedicated to the memory of Anna Tramontano (1957-2017) who was an Italian computational biologist and chair professor of biochemistry at the Sapienza University of Rome.

Declarations of interest: none
Keywords: protein structure prediction; ab-initio; template-free modelling; artificial intelligence

*corresponding author : bernard.offmann@univ-nantes.fr

Abbreviations: Critical Assessment of protein Structure Prediction (CASP), Template-based modelling (TBM), Template-free modelling (TBM), Fragment-based approaches (FBA), Artificical Intelligence (AI).

1 Introduction

Proteins are complex biomolecules that play a crucial role in building, strengthening, maintaining, protecting and repairing a living entity. Each protein folds into a specific three-dimensional structure owing to its amino acid composition. This in turn corresponds to a specific function, collectively termed as sequence-structure-function paradigm [1]. The relationship between protein sequence and its corresponding secondary and tertiary structure is termed as second genetic code [2]. A major gap exists in our knowledge of the science behind protein folding based on its sequence. Research focused in deciphering the second genetic code has been budding for past few decades by means of various schemes.

Advent of genomics has led to the availability of large deposit of sequence data online. This helps in easy classification of proteins and in approximating their functional annotation. A considerable amount of this classification is based on shared sequence similarity (and conserved domain search) between two or more sequences. Currently, UniProtKB/TrEMBL database is enriched with around 170 million sequence data [3]. Yet protein functionality remains unclear primarily due to the lack of structural description at the atomic levels. The equivalent structural database, RCSB [4] (https://www.rcsb.org) documents around 160,000 structures to date belonging to well defined protein families. There is also an ever increasing gap between protein sequence and structure data availability due to considerable growth observed in sequencing techniques.

Scientific community has always relied on experimental approaches to deliver high resolution protein structures. Structural data deposited in data banks are only accountable when verified through experiments like X-Ray [5], NMR [6] etc. Time and again these techniques have been proven to be most efficient in getting relevant spatial characterisation of a protein. On the other hand, they also have remained stagnant in terms of improvements due to being heavily restricted by time and manpower requirements [7]. A recent introduction of Cryo-EM has fostered an acceleration of protein structure determination process [8]. The core of this technique lies in photographing frozen molecules to determine their structure. Nonetheless, the approach is relatively new and usually generates lower resolution structures than those benchmarked by other experimental techniques.

Twentieth century has witnessed a blooming era for scientific community indulging in computational approaches for approximation of protein structures. Anfinsen in 1972 laid the foundation for protein structure prediction by correctly refolding ribonuclease molecule from its sequence [9]. As stated in the paper, "the native conformation is determined by the totality of inter-atomic interactions and hence by the amino acid sequence in a given environment" [10]. In other words, a protein attains its conformational nativity when its environment is at its lowest Gibbs free energy levels. Another statement put forward in their work was that a protein structure is only stable and functional in the environment it was chosen during natural selection. Despite knowing the physical environment requirement for folding a protein sequence, it remains a challenge to fold them into their functional form. Therefore, limiting the understanding of the sequence-structure-function paradigm [11, 12].

Computational approaches for protein structure prediction can broadly be categorized into two groups: Template-Based Modelling (TBM) [13, 14] and Template Free Modelling/Free-Modelling (FM) [15]. A representative flowchart of the categorization is illustrated in Figure 1. This classification has been adopted by well-known biennial competition of protein structure prediction, Critical Assessment of protein Structure Prediction (CASP) [16] [17–20]. Results

from this competition benchmark the improvement in the field of computational protein structure prediction [19, 21]. Majority of progress witnessed in this field is in construction of protein models using templates sharing high sequence similarities with unknown protein. The basis behind the approach is that similar sequence tend to fold in a similar manner. This tendency of proteins to envelop into similar folds reduces with shared sequence similarity, though there exist cases of proteins having same folds even when their shared sequence similarity is low.

TBM, as the name suggests, makes use of template to predict 3D models. Single or multiple homologous protein sharing high sequence similarity are aligned to the unknown protein sequence predicting likely models [13]. Structures predicted through TBM usually have a good resolution and might fall into same functional classes. But there is little progress made when it comes to predicting new protein folds or structures. TBM is an effective approach as long as the query shares at-least 30% sequence identity with the template [22]. On the basis of shared sequence identity, it can be classified into Homology Modelling (HM) [23–25], Comparative Modelling (CM) [26, 27] and Threading approaches (fold-recognition) [28–31]. Each sub-class follows similar methodology into prediction of protein three-dimensional organisation from its primary one-dimensional sequence. One might argue that HM and CM are two terms for one and the same approach. It is true to a great extent except that homology modelling is defined when template shares an ancestry with the query being modelled whereas in case of CM, the query sequence has no identified evolutionary relationship with the template but only shared sequence similarity. So far, comparative modelling has been the most successful computational protein structure prediction approach available [22]. The third category of TBM is fold-recognition/threading which follows the idea of picking template structures based on their fit with the protein sequence in question. It is basically a comparison of 1D protein sequence to 3D template structure.

2 *Ab initio* Protein Structure Prediction

A significant amount of sequence data does not share homology with well-studied protein families. This called for development of approaches which could help predicting protein structures with minimal or no known information. Such approaches fall into the second major class of computational protein structure prediction called "Template-Free modelling/Free-modelling" (TFM/FM). The word "free" used in the name indicates the initial take on such algorithms to rely on physical laws to determine protein structures. Though, most of the algorithms developed around it are guided by structural information. In this review we will touch into the evolution of Free-modelling and the approaches that have been used to predict 3D models. Throughout this review Free-modelling, *ab initio* modelling and *de novo* modelling will be used interchangeably to discuss template-free modelling approaches.

Template-free modelling comprises of algorithms/pipeline/methods for generating protein models with no known structural homologs available. Mainly these approaches focused on using physics based principles and energy terms to model proteins. The nomenclature remains debatable as in several cases, information from known structures is used in one way or another. This review is considering the following definition as best suited to describe our understanding of TFM: "*Ab initio* protein structure prediction or Free modelling (FM) can most appropriately be defined as an effort to construct 3D structure without using homologous proteins as template" [14, 22, 32–34]. FM approaches majorly depend on designing algorithms with ability to rapidly locate global energy minimum and a scoring function capable of selecting best available conformation from the several generated models [35–37].

The aim of free-modelling protocols is to predict the most stable protein spatial arrangement with lowest free energy. The major challenge faced while developing *ab initio* approaches is searching conformational space which is usually huge considering the dynamic nature of

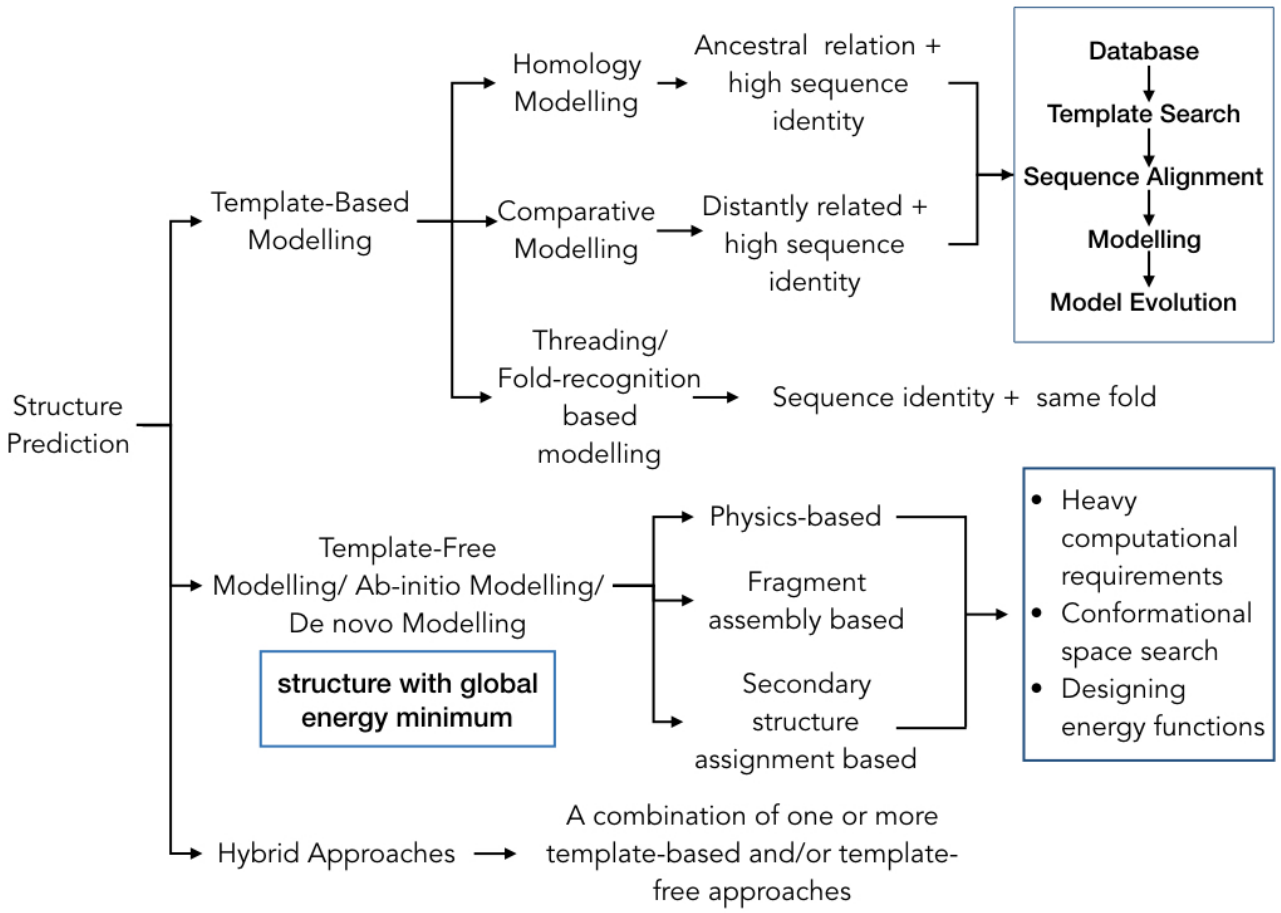


Figure 1: Computation Protein Structure Prediction Approaches.

This figure provides a broad classification of few computational protein structure prediction strategies developed and used to determine protein structure. The two major classifications of these methodologies lie under the domain of Template-Based Modelling and Template-free modelling. Each of these categories can further be split into a set of strategies based on the basic principle followed by the parent approaches for structure prediction.

proteins. Since, these approaches involve building the protein structure from scratch, focus is laid on building effective energy functions to minimise conformational search space and facilitate accurate folding [22, 34]. *Ab initio* algorithms can also be influenced from experimental data available in the form of abstract NMR restraints, predicted residue-residue contact maps, Cryo-EM density maps etc. [38–40].

2.1 Strategies for *Ab initio* Prediction of Protein Structure

Free modelling has witnessed a major bloom in the past era owing to several strategies developed for structure prediction, few of them have been stated in the Table 1. Initially the scientific community resorted to use pure physics based laws, MD simulations etc. to explore the atomic dynamics of protein molecules. The prediction horizon expanded with time into utilizing restraints like Cα-Cα distance, dihedral angles, solvent interactions, side-chain atoms, contact map information and more from available structures. The newer fundamentals involved building saturated library of structural information in the form of small fragments, secondary structural elements, motifs, foldons etc. Below we have broadly classified the *ab initio* protein

Table 1: Strategies available for protein structure prediction. First two columns provide the name and a brief description of few algorithms developed overtime for *ab inito* protein structure prediction. Third column states the length which for most tools is indicative state-of-the-art results observed throughout the field,i.e, able to achieve a GDT_TS score of 30% and above. For most part high accuracy models generation is still limited to small protein (< 150 residue long).

Algorithms/ Servers	**Strategy/Approach**	**Maximum predicted length**	**Reference**
Rosetta	Fragment-assembly using MC simulations, all-atom energy function to determine structure and clustering of models	~ 250	[41–48]
Quark	Fragment-based assembly using REMC simulation guided by knowledge based potentials	upto 150	[7, 49]
EdaRose	Utilising EdaFold algorithm for fragment based assembly with cluster based and energy based variations	upto 150	[20]
Chunk-Taser	Hybrid approach using restraints derived from super secondary structure chunks as well as by threading the templates	< 200	[50]
UNRES	Physics-based conformational space search using UNRES energy function	< 80	[51]
BCL::FOLD	Assembling secondary structural elements using MC sampling and knowledge-based energy functions	~ 200	[15, 40, 52]
SS-Thread	Prediction of contacting pairs of α-helix and β-strand	< 200	[53]
UniCon3D	Using foldons and probabilistic models to capture local backbone structural preference and side chain conformation search space	~ 200	[34]
Bhageerath-H	Hybrid approach involving combination of several tools developed in the lab with the goal to reduce conformational search space	~ 350	[37]
SmotifTF	Fragment-Based approach developed using saturated library of super secondary structure fragments	~ 150	[54]
Touchstone	Secondary and tertiary restraints prediction through threading-based approaches	~ 150	[55, 56]
PconsFold	Evolutionary based structure prediction pipeline using PconsC contact predictions on Rosetta folding protocols	~ 250	[57]
EdaFold	Fragment-based all atom energy function to produce atomic models	upto 150	[58, 59]
Astro-Fold	folding amino acid sequences using first-principle based approaches	~ 150	[60, 61]
DMFold	Deep learning based protein modelling by incorporating predicted structural constraints like inter-atomic distance bonds, hydrogen bonds and torsion angles	upto 200	[62]

structure prediction approaches based on the core methodologies used to develop them.

2.2 Physics Based Approaches

These formed the basis of initial algorithms built under the emerging field. The main idea behind developing these physics-based approaches is to rely on MD simulations to trace the folding path of the proteins. The philosophy backing their design is to obtain lowest energy conformation model by folding the protein sequence using quantum mechanism and coulomb potential [63,64]. But due to high computational requirements, the field majorly relies on inter atomic interactions and force fields to solve the protein folding problem.

Free energy calculations have been explored from the very beginning of computational protein structure prediction evolution. It is believed that these approaches can go beyond documented structures and capture novel folds and patterns by exploring the inherent dynamic motion of proteins [65,66]. Despite the availability of better computing, physics based approach continues to lag behind due to the amount of time required to reach the native state along-with the meddling of erroneous force-field that restrict the model to attain it [12, 67–69].

MELD (Modelling Employing Limited Data) [65] is a recently developed physics-based protein structure prediction approach which uses Bayesian law to tap into atomic molecular dynamics of proteins for structural modelling. It has proven to be effective in determining high resolution structures of proteins up to 260 residues long [65]. Similar effort was made by David Shaw's group where they utilised different sets of restraints to reduce the MD simulation runs and prevent the model from getting trapped in non-native energy state [70]. H. Nguyen et al demonstrated that the combination of an implicit solvent and a force field can result in near-native models in-case of small proteins (less than 100 amino acids) [71]. Another group showed that simulation time can be reduced and energy landscapes can be managed using residue-specific force field (RSFF1) in explicit solvent and Replica exchange molecular dynamics (REMD) [72].

2.3 Fragment Based Approaches (FBA)

It is by far the most successful strategy used for template-free prediction of protein structure. This approach revolves around the construction of fragment libraries of varied lengths, where each fragment represents a pseudo-structure. The idea is to map information from protein fragments instead of using entire templates for constructing protein model. Segments of query sequences are replaced by the fragment coordinates recorded in the fragment library or by its predicted fold. Since, it is computationally exhaustive to go through all possible protein fold conformations for a structure built from scratch, fragmenting the sequence limits the number of folding patterns thus reducing the computational expense. Bowie and Eisenberg introduced Fragment-Based assembly approach to predict protein structures [41]. They used fragments of length 9 to 25 from a database of known proteins and an energy function (composed of 6 terms) that can guide building of energetically stable models [41]. This attempt set path for the evolution of computational 3D-modelling of protein structures using fragments.

Through the years several fragment-based approaches have been developed; few of which have done exceedingly well and remain the best options for *ab initio* protein structure prediction to date. The basic idea behind these algorithms remains the same and typically varies with fragment type, length and scoring functions used to generate energetically minimised stable structure. Rosetta [43, 44], one of the most renowned fragment based approach, uses fragment libraries of length 3 and 9. It follows a Monte Carlo simulation based strategy to predict globally minimised protein models. The scoring function used in Rosetta is based on Bayesian separation of total energy into individual components. Its highest achievement has been noted during

CASP11, where it correctly predicted the structure of a 256 amino acid long sequence [73].

SmotifsTF [54] produces library of supersecondary structure fragments known as Smotifs to built probable models. The fragment library construction and utilisation is based on fragment assembly protocols. The fragment collection is governed by weak sequence similarities generating fragments on average of 25 residues in length. QUARK [49] has more dynamic fragment length range of up to 20 residues which are assembled using replica-exchange Monte Carlo simulations guided by knowledge-based force-field. It has also been ranked as the best predictor in FM category for both CASP9 [18] and CASP10 [19] competition and was among the two dominant tools in CASP11 [73].

The energy functions or scoring functions used in FBA are directed by micro-state interactions existing within known protein structures. These energy terms or functions are also termed as "Knowledge Based Potentials" [74]. FBA algorithms sought out to optimize these energy functions. Though and on one hand, FBA based algorithms have witnessed the most success in the biannual CASP competition by designing algorithms around the principle that certain local structures are favoured by local amino acid sequence. On the other hand, this limits their ability to search for alternate conformation of the proteins within a single run which reduces their probability of discovering a novel fold.

2.4 Secondary Structural Elements Based Approaches

Algorithms employing the use of SSEs for building protein models usually focus on assembling the core backbone of the protein with an exception of loop regions leading to model refinement protocols. BCL::FOLD [15] is one of such algorithms built with the objective to overcome the size and complexity limits faced by most approaches. In the later edition, restraints recovered from sparse NMR data were also incorporated in the pipeline aiding in rapid identification of protein topology [40]. This was benchmarked on protein data set upto a length of 565 containing both soluble and membrane proteins. The algorithm was tested on 20 CASP11 targets, out of which it was able to produce a GDT_TS score of 30% on average for twelve [52]. The average GDT_TS score was accounted for 36%. The study was conducted by using targets belonging to different categories offered by CASP, for example T0 (regular targets), TP (predicted residue-residue contacts) and TS (NMR-NOE restraints) etc. This study also pointed out that better structures were predicted for proteins dominated by α-helix than β-strands. The prediction accuracy also decreased with the size of the protein.

Another algorithm based on the similar principle is SSThread [53]. It predicts contacting pairs of α-helices and β-strands from experimental structures, secondary structure prediction and contact map predictions. The overlapping pairs are then assembled into a core structure leading to the prediction of loop regions. The contact pairing strategy employed by SSThread has been shown to be better in predicting β-strand pairs then all α pairs.

2.5 Deep Learning Based Approaches

Quite recently neural network based deep learning approaches have seen a boom in the field of protein structure prediction.

So far, deep learning approaches for PSP have vastly been used as one of the component in the entire pipeline rather than implicitly being implemented as the driving force. Majority of its use revolve around prediction of residue-residue contacts, which are primarily derived through co-evolutionary approaches and/or by building sequence alignment profiles [75–77].

Recent work done by Al Quraishi [78] focused on building a pure deep learning based prediction approach. It was designed as a one step algorithm for prediction of protein structure relying on end-to-end deferential deep learning strategy. The emphasis was laid on not using

any co-evolutionary data or information from existing templates for protein model construction. Instead, the algorithm relied on data derived solely from protein sequence in question and evolutionary profile of individual residues within the sequence. This method achieved state-of-the-art results as observed in case of *ab initio modelling* protocols.

Another tool that has shown prominence in CASP13 is DeepMind's AlphaFold [79]. It uses a two step process for protein structure determination, which also involves the use of co-evolutionary profiles to guide model building. Through this methodology high-accuracy models were constructed for 24 out of 43 test proteins achieving a TM-score of 0.7 and above in the template-free modelling domain.

The community is still beginning to explore the benefits of deep learning approaches into PSP. The major step back that can be encountered by these techniques would be related to lack of availability of structural data. Since, these approaches are based on training the algorithm based on the certain patterns followed by available data. Structure prediction field has always been slower than the sequence equivalent which translates into lower availability of data that can be used to train the algorithm. Hence, though deep learning based approaches can be better implemented in sequence domain of protein biology, it will take other advances in structural biology to push forward deep learning based approaches. Other problem that such approaches can be prone to is over training of the algorithms.

3 Hybrid Approaches

With the advancements made in computational approaches to protein structure prediction, the line between individual methodology is diminishing. Now the structure prediction community is moving forward towards the use of "Hybrid Approaches", which do not strictly rely on pure template based or template-free prediction criteria but on the amalgamation of both. Bhageerath [80] is one such homology/*ab initio* hybrid protocol. It is available in the form of a web-server called Bhageerath-H [37]. The main focus of the pipeline is on reduction of conformational search space. Out of thousands of predicted models, top 5 are selected based on physio-chemical metric (pcSM) scoring function (specific to this algorithm). Efficiency of this software was put to test by using CASP10 targets with promising prediction results. After the assessment of its shortcomings, an updated version was presented in CASP12 meeting as BhageerathH+ [81].

In another study, Quark [49] and fragment-guided molecular dynamic (FG-MD) were added to I-Tasser pipeline [11, 82] to improve on the existing protocol [33, 83]. The basic idea was to introduce *ab initio* generated structures from QUARK into LOMETS [84] to find any hit with existing homologous template with a good TM-score. Top hits are then passed into I-Tasser pipeline for atomic refinement to obtain a structure with low rmsd. This combination produced better results for FM targets in CASP10 and CASP11 experiments than QUARK alone [33,85]. MULTICOM_NOVEL approach is one more example of a hybrid algorithm which was constructed by combining various complementary structure prediction pipelines including MULTICOM server, I-Tasser, RaptorX [14], Rosetta etc.

Chunk-Tasser can also be put into this category as it utilizes both chunks of folded secondary structural fragments along with fold-recognition to assemble protein structures [50].

On similar grounds, an initiative was undertaken in 2014 to combine methods of the best known protein structure prediction techniques and to come up with a pipeline which could generate better structures. This initiative came to be known as WeFold, where 13 labs collaborated to merge their algorithms forming 5 major branches [86]. The outcome was promising and the authors of this study discussed on further improvements to be made in prediction protocols as a result of this 'coopetition' [86].

4 Evolution of CASP and its contribution

CASP has been a contributing factor for the work done in the field of computational protein structure prediction. It is a biennial competition being conducted for around two decades serving as a platform to judge the accuracy of prediction pipelines. It has grown overtime into a protein structure prediction platform to qualify prediction strategies coming under domains like template-based modelling, template-free modelling, refinement protocols, contact prediction etc. [12, 17, 87, 88].

To keep a track of advancement in PSP techniques, CASP prepares a list of unpredicted protein sequences in each category every two years. This provides an uniformity in assessing the advancement perceived in each area of structure prediction. The protein sequence list provided for blind testing of *ab initio* modelling approaches often constitutes of protein sequences with "soon to be released" structures. Best models are determined on the basis of several criteria, one of them being a local-global alignment score called GDT_TS score (Global Distance Test) [89]. It calculates the Cα distance between residues from model and template protein at defined rmsd cut-off values. Henceforth determining both local and global similarities between two protein molecules.

The initial achievement in protein tertiary structure prediction was observed in CASP4, but mainly for small proteins ($\leq$ 120 residues). In later years, the *ab initio* prediction field remained stagnant for about a decade until the introduction of better contact prediction approaches in CASP11 competing pipelines with promising improvements in prediction accuracy [90]. Similar trend was observed in CASP12 with the inclusion of alignment-based contact prediction methods [91].

Recently conducted CASP13 demonstrated further improvement on average GDT_TS score due to the employment of deep learning approaches in structure prediction [78]. This served as an encouragement to dig further into deep learning based approaches to solve the protein folding problem.

5 Conclusion

Template-based prediction in general are quicker than experimental methods, at least in providing initial spatial arrangement of the protein. One of the major drawback of these approaches is the redundancy of information, i.e., no new fold or family can be discovered as it relies on building models from existing structures. In addition, these methods fail to establish the structural integrity of a protein sequence with decreasing sequence or structure identity.

This review peeks into few methods and possibilities of free-modelling techniques developed and available for the prediction of protein structure. *Ab initio* protein structure prediction still bare influence from PDB structures for optimizing the parameters of protein folding. This information helps them reduce the conformational space sampling requirements by maximizing the efficiency of energy functions. Most of the algorithms are still directed by a combination of knowledge-based potentials and physics-based approaches [92].

To date free-modelling has been been well adapted for protein sequences length of 150 residues or below [16] [92, 93]. Few instances have seen algorithms overdoing themselves and going beyond the length restrictions to predict structure for longer proteins. CASP11 witnessed major success in *ab initio* protein structure prediction for a structure of length 256 a.a. [90].

The major challenges faced by this field starts with finding an efficient way to explore the conformational search space as a protein sequence can fold into indefinite forms. Thereby, reducing the plausible folding possibilities to the best probable fold is a hard task to achieve.

The length limitation could indicate towards the design strategy of the algorithms, many of which rely on defining domain boundaries prior to structure prediction. But, this would not

be a very strong case of argument as most reliable predictions still lie under the length of 150 residue, though single domain boundary could expand upto a length 200 to 250 residues. The only exception to this case seen over the past years of CASP competition where a protein of length 256 a.a. was accurately predicted [90].

One could also argue that as a community, we are still at the domain-level of structure prediction given the length limitations. Algorithms have been built that specifically target solving the structure of single-domain proteins [57, 94] or choices of restraints that are limited to single-domain proteins only [56]. Many of the designed algorithms tend to be validated on a data set of single-domain proteins as well [75].

Even so, in the case of small protein structure predictions, not all of them deliver model close to its nativity. It is just that chances of having a good structure is more if the length of the protein is around 150 or less.

There is also the case of stagnancy in the field for the majority of years and been pointed out a lot. One of the example is the discussion provided by the end of CASP10 [95]. The study pointed out that the results from CASP10 lay closer to what were observed during CASP5. It also reflected on the fact that it might be due to the gradual increase in the complexity of CASP targets along with the inclusion of multi-domain targets provided for modelling.

Another drawback faced by few of these algorithms can be the run time which can vary a lot depending on the size of the protein and the internal functioning of the algorithm. With the inclusion of artificial intelligence, the time scale for modelling has been reduced to milliseconds but generally, a single prediction can take from somewhere from few minutes to hours to days for an algorithm to complete.

Most of the algorithms still rely on manual intervention to complete the runs and so the human error should also be considered.

The point of PSP is not just high accuracy structure determination but also to ascertain the basis behind this biological process (protein folding). Thereby, finally answering questions like "why good protein become faulty and cause disease".

De novo protein structure prediction still requires a lot of improvement, but at the same time it promises a better prospect of structure prediction in future. It brings with it a hope of predicting newer folds at a faster pace when compared to experimental approaches which can remain stuck for years altogether due to numerous reasons. In general computational structure prediction techniques though have a room for improvement are still quick when compared to traditional approaches [16]. If considering Template-Based modelling approaches, few limitations still persist whereas *ab initio* approaches can move a step ahead and might help understanding the basic principles of protein folding [93].

Acknowledgements

The authors are most thankful to Yves-Henri Sanejouand for critical reading of the manuscript. SD is thankful to Conseil Régional de La Réunion and Fonds Social Européen for providing a PhD scholarship under tier 234275, convention DIRED/20161451. BO is thankful to Conseil Régional Pays de la Loire for support in the framework of GRIOTE grant.

Conflict of interest

Authors declare no competing interests.

Authors contributions

Conception of the work was done by SD, RS, FC and BO. SD, RS and BO performed literature data collection and analysis. All four authors SD, RS, FC and BO wrote the manuscript. All authors approved the final version of the article.

References

[1] James C Whisstock and Arthur M Lesk. Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics*, 36(3):307–340, 2003.

[2] Gina Kolata. Trying to crack the second half of the genetic code. *Science*, 233:1037–1040, 1986.

[3] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, and Amos Bairoch. UniProtKB/Swiss-Prot. *Methods in molecular biology (Clifton, N.J.)*, 406:89–112, 2007.

[4] Peter W. Rose, Andreas Prlić, Ali Altunkaya, Chunxiao Bi, Anthony R. Bradley, Cole H. Christie, Luigi Di Costanzo, Jose M. Duarte, Shuchismita Dutta, Zukang Feng, Rachel Kramer Green, David S. Goodsell, Brian Hudson, Tara Kalro, Robert Lowe, Ezra Peisach, Christopher Randle, Alexander S. Rose, Chenghua Shao, Yi Ping Tao, Yana Valasatava, Maria Voigt, John D. Westbrook, Jesse Woo, Huangwang Yang, Jasmine Y. Young, Christine Zardecki, Helen M. Berman, and Stephen K. Burley. The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*, 45(D1):D271–D281, 2017.

[5] Max F. Perutz, Michael G. Rossmann, Ann F. Cullis, Hilary Muirhead, Georg Will, and A. C. T. North. Structure of Haemoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature*, 185(4711):416–422, 1960.

[6] Xavier Morelli, Alain Dolla, Myrjam Czjzek, P. Nuno Palma, Francis Blasco, Ludwig Krippahl, Jose J.G. Moura, and Françoise Guerlesquin. Heteronuclear NMR and soft docking: An experimental approach for a structural model of the cytochrome c553-ferredoxin complex. *Biochemistry*, 39(10):2530–2537, 2000.

[7] Dong Xu and Yang Zhang. Ab Initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment. *Scientific reports*, 3:1895, 2013.

[8] Ewen Callaway. The Revolution Will Not Be Crystallized. *Nature*, 525:172–174, 2015.

[9] Christian B. Anfinsen. The Formation and Stabilization of Protein Structure. *Biochemical Journal*, 128(4):737–749, 1972.

[10] Christian B Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, 1973.

[11] Ambrish Roy, Alper Kucukural, and Yang Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738, 2010.

[12] Michael Feig. Computational protein structure refinement: almost there, yet still so far to go. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(3), 2017.

[13] Andras Fiser. Template-based protein structure modeling. *Methods in molecular biology (Clifton, N.J.)*, 673:73–94, 2010.

[14] Morten Källberg, Haipeng Wang, Sheng Wang, Jian Peng, Zhiyong Wang, Hui Lu, and Jinbo Xu. Template-based protein structure modeling using the RaptorX web server. *Nature Protocols*, 7(8):1511–1522, 2012.

[15] Mert Karaka, Nils Woetzel, Rene Staritzbichler, Nathan Alexander, Brian E. Weiner, and Jens Meiler. BCL::Fold - De Novo Prediction of Complex and Large Protein Topologies by Assembly of Secondary Structure Elements. *PLoS ONE*, 7(11), 2012.

[16] Ling-Hong Hung, Shing-Chung Ngan, and Ram Samudrala. *De Novo Protein Structure Prediction*, pages 43–63. Springer New York, New York, NY, 2007.

[17] Moshe Ben-David, Orly Noivirt-Brik, Aviv Paz, Jaime Prilusky, Joel Sussman, and Yaakov Levy. Assessment of casp8 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics*, 77(Suppl. 9):50–65, 2009.

[18] Lisa Kinch, Shuo Yong Shi, Qian Cong, Hua Cheng, Yuxing Liao, and Nick V. Grishin. CASP9 assessment of free modeling target predictions. *Proteins: Structure, Function, and Bioinformatics*, 79(Suppl. 10):59–73, 2011.

[19] Chin Hsien Tai, Hongjun Bai, Todd J. Taylor, and Byungkook Lee. Assessment of template-free modeling in CASP10 and ROLL. *Proteins: Structure, Function, and Bioinformatics*, 82(Suppl. 2):57–83, 2014.

[20] David Simoncini, Thomas Schiex, and Kam Y.J. Zhang. Balancing exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 85(5):852–858, 2017.

[21] Lisa Kinch, Wenlin Li, Richard Schaeffer, Roland Dunbrack, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Nick Grishin. Casp 11 target classification. *Proteins: Structure, Function and Bioinformatics*, 84:20–33, 2016.

[22] Bee Yin Khor, Gee Jun Tye, Theam Soon Lim, and Yee Siew Choong. General overview on structure prediction of twilight-zone proteins. *Theoretical Biology and Medical Modelling*, 12(1):15, 2015.

[23] Konstantin Arnold, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede. The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, 2006.

[24] Lorenza Bordoli, Florian Kiefer, Konstantin Arnold, Pascal Benkert, James Battey, and Torsten Schwede. Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols*, 4(1):1–13, 2008.

[25] Alexander Miguel Monzon, Diego Javier Zea, Cristina Marino-Buslje, and Gustavo Parisi. Homology modeling in a dynamical world. *Protein Science*, 26:2195–2206, 2017.

[26] Andras Fiser, Roberto Sánchez, Francisco Melo, and Andrej Fiser. Comparative Protein Structure Modeling. *Computational Biochemistry and Biophysics*, 2001.

[27] Benjamin Webb and Andrej Sali. Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 54(1):5.6.1–5.6.37, 2016.

[28] Burkhard Rost, Reinhard Schneider, and Chris Sander. Protein fold recognition by prediction-based threading. *Journal of Molecular Biology*, 270(3):471–480, 1997.

[29] Jeffrey Skolnick and Daisuke Kihara. Defrosting the frozen approximation: PROSPECTOR–a new approach to threading. *Proteins: Structure, Function, and Bioinformatics*, 42(3):319–331, 2001.

[30] William R Taylor and Inge Jonassen. A structural pattern-based method for protein fold recognition. *Proteins: Structure Function, and Bioinformatics*, 56(2):222–34, 2004.

[31] Jinbo Xu, Feng Jiao, and Libo Yu. Protein structure prediction using threading. *Methods in molecular biology (Clifton, N.J.)*, 413:91–121, 2008.

[32] David E. Kim, Ben Blum, Philip Bradley, and David Baker. Sampling Bottlenecks in De novo Protein Structure Prediction. *Journal of Molecular Biology*, 393(1):249–260, 2009.

[33] Wenxuan Zhang, Jianyi Yang, Baoji He, Sara Walker, Hongjiu Zhang, Brandon Govindarajoo, Jouko Virtanen, Zhidong Xue, Hong-Bin Shen, and Yang Zhang. Integration of quark and i-tasser for ab initio protein structure prediction in casp11. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl. 1):76–86, 2015.

[34] Debswapna Bhattacharya, Renzhi Cao, and Jianlin Cheng. UniCon3D: De novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics*, 32(18):2791–2799, 2016.

[35] Paul IW de Bakker, Nicholas Furnham, Tom L. Blundell, and Mark A. DePristo. Conformer generation under restraints. *Current Opinion in Structural Biology*, 16(2):160–165, 2006.

[36] Adam Liwo, Cezary Czaplewski, Stanisław Ołdziej and Harold A Scheraga. Computational techniques for efficient conformational sampling of proteins Adam. *Structure*, 18(2):134–139, 2008.

[37] B Jayaram, Priyanka Dhingra, Avinash Mishra, Rahul Kaushik, Goutam Mukherjee, Ankita Singh, and Shashank Shekhar. Bhageerath-H: A homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. *BMC Bioinformatics*, 15(Suppl. 16):S7, 2014.

[38] Peter M Bowers, Charlie E Strauss, and David Baker. De novo protein structure determination using sparse NMR data. *Journal of Biomolecular NMR*, 18(4):311–318, 2000.

[39] Maya Topf, Matthew L. Baker, Marc A. Marti-Renom, Wah Chiu, and Andrej Sali. Refinement of protein structures by iterative comparative modeling and cryoEM density fitting. *Journal of Molecular Biology*, 357(5):1655–1668, 2006.

[40] Brian Weiner, Nathan Alexander, Louesa Akin, Nils Woetzel, Mert Karakas, and Jens Meiler. Bcl::fold-protein topology determination from limited nmr restraints. *Proteins: Structure Function, and Bioinformatics*, 82:587–595, 2014.

[41] James. U. Bowie and David Eisenberg. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proceedings of the National Academy of Sciences*, 91(10):4436–4440, 1994.

[42] Kim T. Simons, Charles Kooperberg, Enoch Huang, and David Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, 1997.

[43] Kim T. Simons, Rich Bonneau, Ingo Ruczinski, and David Baker. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Structure, Function, and Bioinformatics*, 37(Suppl. 3):171–176, 1999.

[44] Richard Bonneau, Jerry Tsai, Ingo Ruczinski, Dylan Chivian, Carol Rohl, Charlie E M Strauss, and David Baker. Rosetta in CASP4: Progress in ab initio protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 45(Suppl. 5):119–126, 2001.

[45] Srivatsan Raman, Robert Vernon, James Thompson, Michael Tyka, Ruslan Sadreyev, Jimin Pei, David Kim, Elizabeth Kellogg, Frank Dimaio, Oliver Lange, Lisa Kinch, Will Sheffler, Bong Hyun Kim, Rhiju Das, Nick V. Grishin, and David Baker. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Structure, Function, and Bioinformatics*, 77(Suppl. 9):89–99, 2009.

[46] Hahnbeom Park, Frank DiMaio, and David Baker. Casp11 refinement experiments with rosetta. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl. 1):314–322, 2015.

[47] Sergey Ovchinnikov, David Kim, Ray Wang, Yuan Liu, Frank DiMaio, and David Baker. Improved de novo structure prediction in casp11 by incorporating co-evolution information into rosetta. *Proteins: Structure, Function, and Bioinformatics*, 84:67–75, 2015.

[48] Rebecca Faye Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V Shapovalov, P. Douglas Renfrew, Vikram Khipple Mulligan, Kalli Kappel, Jason W. Labonte, Michael Steven Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, 2017.

[49] Dong Xu and Yang Zhang. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, 80(7):1715–1735, 2012.

[50] Hongyi Zhou and Jeffrey Skolnick. Ab initio protein structure prediction using chunk-TASSER. *Biophysical Journal*, 93(5):1510–1518, 2007.

[51] S Ołdziej, C Czaplewski, A Liwo, M Chinchio, M Nanias, J A Vila, M Khalili, Y A Arnautova, A Jagielska, M Makowski, H D Schafroth, R Kaźmierkiewicz, D R Ripoll, J Pillardy, J A Saunders, Y K Kang, K D Gibson, and H A Scheraga. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7547–7552, 2005.

[52] Axel W. Fischer, Sten Heinze, Daniel K. Putnam, Bian Li, James C. Pino, Yan Xia, Carlos F. Lopez, and Jens Meiler. CASP11 - An evaluation of a modular BCL: Fold-based protein structure prediction pipeline. *PLoS ONE*, 11(4), 2016.

[53] Kevin J. Maurice. SSThread: Template-free protein structure prediction by threading pairs of contacting secondary structures followed by assembly of overlapping pairs. *Journal of Computational Chemistry*, 35(8):644–656, 2014.

[54] Brinda Vallat, Carlos Madrid-Aliste, and Andras Fiser. Modularity of Protein Folds as a Tool for Template-Free Modeling of Structures. *PLoS Computational Biology*, 11(8):1–16, 2015.

[55] Daisuke Kihara, Hui Lu, Andrzej Kolinski, and Jeffrey Skolnick. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proceedings of the National Academy of Sciences of the United States of America*, 98(18):10125–10130, 2001.

[56] Yang Zhang, Andrzej Kolinski, and Jeffrey Skolnick. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophysical Journal*, 85(2):1145–64, 2003.

[57] Mirco Michel, Sikander Hayat, Marcin J. Skwark, Chris Sander, Debora S. Marks, and Arne Elofsson. PconsFold: Improved contact predictions improve protein models. *Bioinformatics*, 30(17):482–488, 2014.

[58] David Simoncini, Francois Berenger, Rojan Shrestha, and Kam Y J Zhang. A probabilistic fragment-based protein structure prediction algorithm. *PLoS ONE*, 7(7):1–11, 2012.

[59] David Simoncini and Kam Y J Zhang. Efficient Sampling in Fragment-Based Protein Structure Prediction Using an Estimation of Distribution Algorithm. *PLoS ONE*, 8(7):1–10, 2013.

[60] John L. Klepeis and Christodoulos A. Floudas. ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical journal*, 85(4):2119–2146, 2003.

[61] Ashwin. Subramani, Yinan Wei, and Christodoulos A. Floudas. ASTRO-FOLD 2.0: An enhanced framework for protein structure prediction. *AIChE Journal*, 58(5):1619–1637, 2012.

[62] Joe G. Greener, Shaun M. Kandathil, and David T. Jones. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nature Communications*, 10(1):1–13, 2019.

[63] Adam Liwo, Jooyoung Lee, Daniel R Ripoll, Jaroslaw Pillardy, and Harold A Scheraga. Protein structure prediction by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences of the United States of America*, 96(10):5482–5485, 1999.

[64] Corey Hardin, Taras V. Pogorelov, and Zaida Luthey-Schulten. Ab initio protein structure prediction. *Current Opinion in Structural Biology*, 12(2):176–181, 2002.

[65] Alberto Perez, Joseph A Morrone, Emiliano Brini, Justin L Maccallum, and Ken A Dill. Blind protein structure prediction using accelerated free-energy simulations. *Science Advances*, 2:e1601274, 2016.

[66] Alpan Raval, Stefano Piana, Michael P. Eastwood, and David E. Shaw. Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations. *Protein Science*, 25(1):19–29, 2016.

[67] Richard Bonneau and David Baker. Ab initio protein structure prediction: progress and prospects. *Annual Review of Biophysics and Biomolecular Structure*, 30(1):173–189, 2001.

[68] Carlos Simmerling, Bentley Strockbine, and Adrian E. Roitberg. All-atom structure prediction and folding simulations of a stable protein. *Journal of the American Chemical Society*, 124(38):11258–11259, 2002.

[69] Alpan Raval, Stefano Piana, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, 80(8):2071–2079, 2012.

[70] Alpan Raval, Stefano Piana, Michael P. Eastwood, and David E. Shaw. Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations. *Protein Science*, 25(1):19–29, 2016.

[71] Hai Nguyen, James Maier, He Huang, Victoria Perrone, and Carlos Simmerling. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *Journal of the American Chemical Society*, 136(40):13959–13962, 2014.

[72] Fan Jiang and Yun Dong Wu. Folding of fourteen small proteins with a residue-specific force field and replica-exchange molecular dynamics. *Journal of the American Chemical Society*, 136(27):9536–9539, 2014.

[73] Lisa Kinch, Wenlin Li, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Nick Grishin. Evaluation of free modeling targets in casp11 and roll. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl. 1):51–66, 2015.

[74] Evandro Ferrada and Francisco Melo. Effective knowledge-based potentials. *Protein Science*, 18(7):1469–1485, 2009.

[75] Jörg Schaarschmidt, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Alexandre Bonvin. Assessment of contact predictions in casp12: Co-evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*, 86(Suppl. 1):51–66, 2017.

[76] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Computational Biology*, 13(1):1–34, 2017.

[77] Shaun M. Kandathil, Joe G. Greener, and David T. Jones. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1092–1099, 2019.

[78] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell Systems*, 8(4):292 – 301.e3, 2019.

[79] Andrew Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577:1–5, 2020.

[80] B. Jayaram, Priyanka Dhingra, Bharat Lakhani, and Shashank Shekhar. Bhageerath - Targeting the near impossible: Pushing the frontiers of atomic models for protein tertiary structure prediction. *Journal of Chemical Sciences*, 124(1):83–91, 2012.

[81] Rahul Kaushik, Ankita Singh, Debarati Dasgupta, Amita Pathak, Shashank Shekhar, and B. Jayaram. BhageerathH+: A hybrid methodology based software suite for protein tertiary structure prediction. In *CASP12 Proceedings*, pages 25–26, 2016.

[82] Jianyi Yang and Yang Zhang. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Research*, 43(W1):W174–W181, 2015.

[83] Dong Xu, Jian Zhang, Ambrish Roy, and Yang Zhang. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins: Structure, Function, and Bioinformatics*, 79(Suppl. 10):147–160, 2011.

[84] Sitao Wu and Yang Zhang. LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, 35(10):3375–3382, 2007.

[85] Yang Zhang. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins: Structure, Function, and Bioinformatics*, 82(Suppl. 2):175–187, 2014.

[86] George A. Khoury, Adam Liwo, Firas Khatib, Hongyi Zhou, Gaurav Chopra, Jaume Bacardit, Leandro O. Bortot, Rodrigo A. Faccioli, Xin Deng, Yi He, Pawel Krupa, Jilong Li, Magdalena A. Mozolewska, Adam K. Sieradzan, James Smadbeck, Tomasz Wirecki, Seth Cooper, Jeff Flatten, Kefan Xu, David Baker, Jianlin Cheng, Alexandre C.B. Delbem, Christodoulos A. Floudas, Chen Keasar, Michael Levitt, Zoran Popović, Harold A. Scheraga, Jeffrey Skolnick, and Silvia N. Crivelli. WeFold: A coopetition for protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 82(9):1850–1868, 2014.

[87] Ralf Jauch, Hock Chuan Yeo, Prasanna R Kolatkar, and Neil D Clarke. Assessment of CASP7 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):57–67, 2007.

[88] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function, and Bioinformatics*, 86:7–15, 2018.

[89] Adam Zemla. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research*, 31(13):3370–3374, 2003.

[90] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP) - progress and new directions in Round XI. *Proteins: Structure, Function, and Bioinformatics*, 84:4–14, 2016.

[91] Luciano A. Abriata, Giorgio E. Tamò, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Matteo Dal Peraro. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 86:97–112, 2018.

[92] Yang Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, 2008.

[93] Jooyoung Lee, Peter L Freddolino, and Yang Zhang. Ab initio protein structure prediction. In *From protein structure to function with bioinformatics*, pages 3–35. Springer, 2017.

[94] David E. Kim, Dylan Chivian, and David Baker. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, 32(WEB SERVER ISS.):526–531, 2004.

[95] Andriy Kryshtafovych, Krzysztof Fidelis, and John Moult. CASP10 results compared to those of previous CASP experiments. *Proteins: Structure, Function, and Bioinformatics*, 82(Suppl. 2):164–174, 2014.

1.3 Structural Alphabets

The most challenging part of protein structure description are the loop regions which are much more structurally diverse than just "coil" or the regular secondary structures. The correct classification of these irregular regions and there participation in protein fold has been a major focus in past studies (Fetrow 1995; Efimov 1996; Oliva et al. 1997) along with the sub-description of the regular secondary structures. Nevertheless, most often, the general classification of protein secondary structure is boiled down to three states, α-helix and β-strand and coils. However, this kind of description provides a restrictive overview of the local structure and 3D dynamics of a protein.

To compensate for the lack of protein structure description in terms of secondary structure, local protein structure libraries have been explored to a great extend. These libraries consist of representative structural prototypes that are capable of approximating the local structure of every part of a protein backbone. Such structural libraries are collectively termed as "Structural Alphabets" (SAs). Basically, SAs are collections or libraries of small structural prototypes that can approximate every part of a protein backbone and represent it in the form of a 1D sequence (Offmann et al. 2008). Several such local libraries have been determined based on different criteria settled upon by the different teams. Few examples of available SAs are shown in Table 1.1.

Table 1.1: List of few examples of different types of structural alphabets (SAs) documented in the literature.

SA name	**Learning Method**	**Distance Used**	**Prototype number**	**Prototype length(s)**	**Reference**
Building Blocks	K-means	rmsd on C-α	103	6	(Unger et al. 1989)
Structural building blocks	HMM	C-α distance	27	4	(Camproux et al. 1999, 2004)
unassigned	Simulated annealing k-means	cRMS deviation	25 - 300	4 - 7	(Kolodny et al. 2002)
Protein Blocks	Unsupervised Classification (SOM)	Dihedral angles	16	5	(de Brevern et al. 2000)
Oligon	Clustering algorithm	cRMS deviation	28, 202, 932 & 2561	4, 5, 6 & 7	(Micheletti et al. 2000)
Centroids	Hypercosine clustering	Hypercosine C-α	28 to 16336	7	(Hunter et al. 2003)
Kappa-alpha	Near Neighbour Clustering	Kappa and alpha angles	23	5	(Tung et al. 2007)
Structural representatives	Leader algorithm and k-means clustering	Cα-Cα distance	27	7	(Sander et al. 2006)

One of the initial type of these structural libraries constructed was Structural Building Blocks (SBBs) (Camproux et al. 1999). These SA type was optimized into 12 representative SAs each 4 residues long using HMM profiling.

By definition the design of SAs enabled descriptive study of protein architecture which could range from a simplistic description of local protein backbone to a more challenging structure prediction protocols.

SAs have been extensively used in loop mining (Regad et al. 2010) and loop modelling protocols (Camproux et al. 2001), analysis of structural variability (Mahajan et al. 2014; Regad et al. 2017), protein flexibility (Craveur et al. 2015; Dong et al. 2016) and even detection of outer membrane proteins (Martin et al. 2008) to cite a few. A recent study have also exploited SAs in determining the structural asymmetry that can exist between the monomeric units of the same protein. The study was performed on a homodimer protein, HIV-2 protease (PR2) and pointed out that the asymmetry in its monomers could exsit due to its flexibility allowing its open and closed form as well for probable binding to different ligands (Triki et al. 2018).

The SAs have been designed in the aim of representing as precisely as possible the local three dimensional conformations of a protein in a simplistic and readable manner (Figure 1.2). It also promotes structure analysis and comparison on the basis of this simplified 1D representation. The 1D depiction of these protein carries more sophisticated definition of its tertiary structure and due to the ability of each SA type defines structure features at local level (single fragment), continuous work on SAs can eventually help in better understanding of protein architecture.

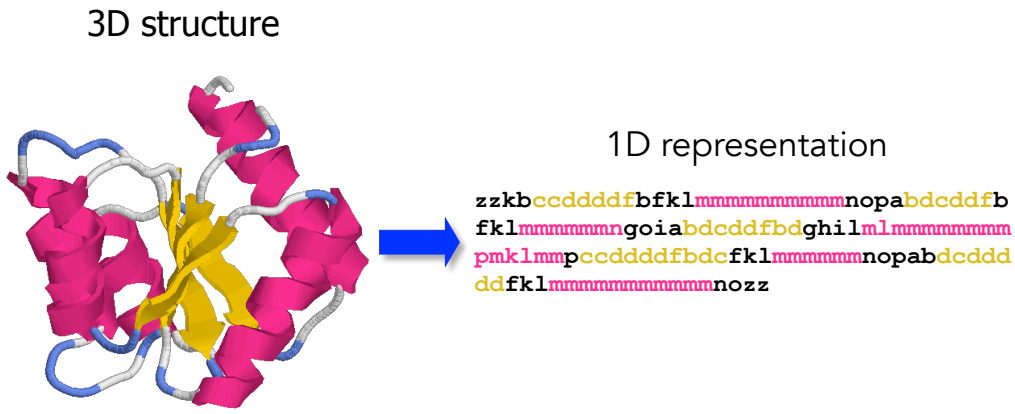


Figure 1.2: Principle of precise encoding of a protein 3D structure information into a simplified 1D representation using a structural alphabet.

In general, SAs hold a huge potential in the field of structural biology and can be readily used for distant homology search as well as in modelling *ab inito* proteins using its ability to define the conformational search space.

1.4 Protein Blocks

1.4.1 Introduction

Protein Blocks (PBs) are a type of structural alphabet that were designed to attain good local structure approximation of protein backbone as well as to predict the local structure of protein from its sequence (de Brevern et al. 2000). These were determined by building a library of fragments of 5 residues long from a database of 342 representative protein structures. The dihedrals (ϕ and ψ) corresponding

to each isolated pentamer were used to quantify the structural differences among themselves by calculations of root mean square deviation on angle (*rmsda*). These structural differences were clustered together by using Kohonen algorithm, which is an unsupervised clustering algorithm known as Self Organising Maps (SOM) (Kohonen 1982). By the end of the learning and clustering process, 16 representative structural prototypes were procured (see Figure 1.3).

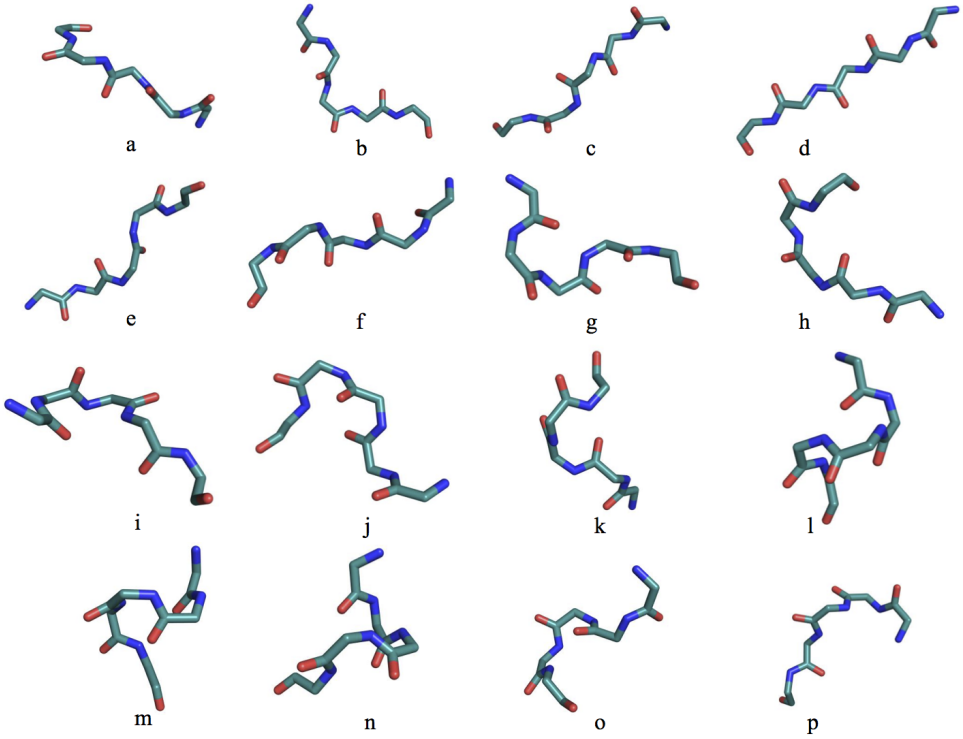


Figure 1.3: Representation of Protein Blocks.

Shown are the 16 pentapeptide units that represent each one PB. They have been annotated alphabetically by the letters a to p (de Brevern et al. 2000).

These 16 structural prototypes have been annotated from a to p in alphabetical order. Each of this designation is specific to a type of local fold. They are described by 8 standard consecutive dihedral angles (Figure 1.4). For instance, PB a to c primarily represent approximation of β-strand N-caps region; PB d approximates to central β-strand; PB e to f represent β-strand C-cap region; PB g through j are for coils; PB k to l are specific to N-cap α-helix; PB m is primarily for central α-helix and lastly PB n to p constitute C-caps region of α-helix.

It was shown that PBs are able to approximate on average to a precision of 0.42 Å of the local parts of a protein 3D structure (de Brevern 2005).

1.4.2 Applications

Each of the 16 PB represent the backbone orientation of a pentapeptide defined by 8 dihedral angles. The approximation of protein 3D structure into 1D PB sequence is done by the comparison of these 8 consecutive dihedral angles with the standard torsion angles to find a match with least Root Mean Square Deviation on Angular values (RMSDA), thereby, assigning a PB. Same protocol is followed for a sliding

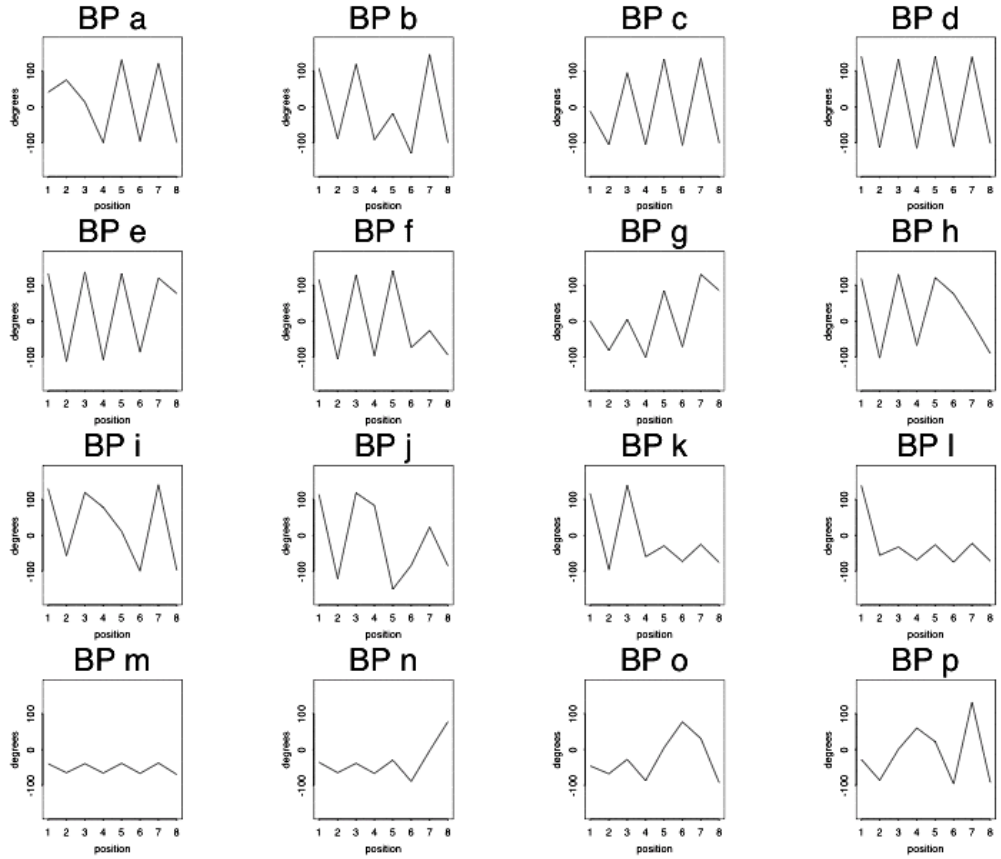


Figure 1.4: Dihedral angles distribution of each protein block (a to p).

Each of the PBs is defined by eight dihedral angles (ψ and ϕ) which are represented in this figure. Each of these set of dihedrals are able to describe a part of secondary structure.

window of 5 residues to generate a PB string description of a protein structure. Over the years, several applications have been developed that revolve around the use of PBs. Its initial application involved PB-assignment (PBA), wherein a protein 3D structure can be represented in a readable sequence space using an in-house software or by the use of PBE web server (Tyagi et al. 2006a). Via PBA, an approximated representation of protein 3D structure could be achieved in the form of a 1D PB sequence of length (n-4), where 'n' is the length of the protein. This loss of length is compensated by the addition of 'ZZ' at the beginning and the end of the PB sequence (see Figure 1.2 above).

They have also been used for protein local structure prediction from its sequence by using overlapping Structural Words (SWs) (de Brevern et al. 2007). One of the work also confirmed its efficacy in determining the local structure of long fragments (Benros et al. 2006, 2009). Another known application of PBs is focused on structure alignment using PB sequences (Tyagi et al. 2006b). Here, the structure comparison is reduced to a sequence alignment problem, though the sequence compared are PB sequences (Figure 1.5).

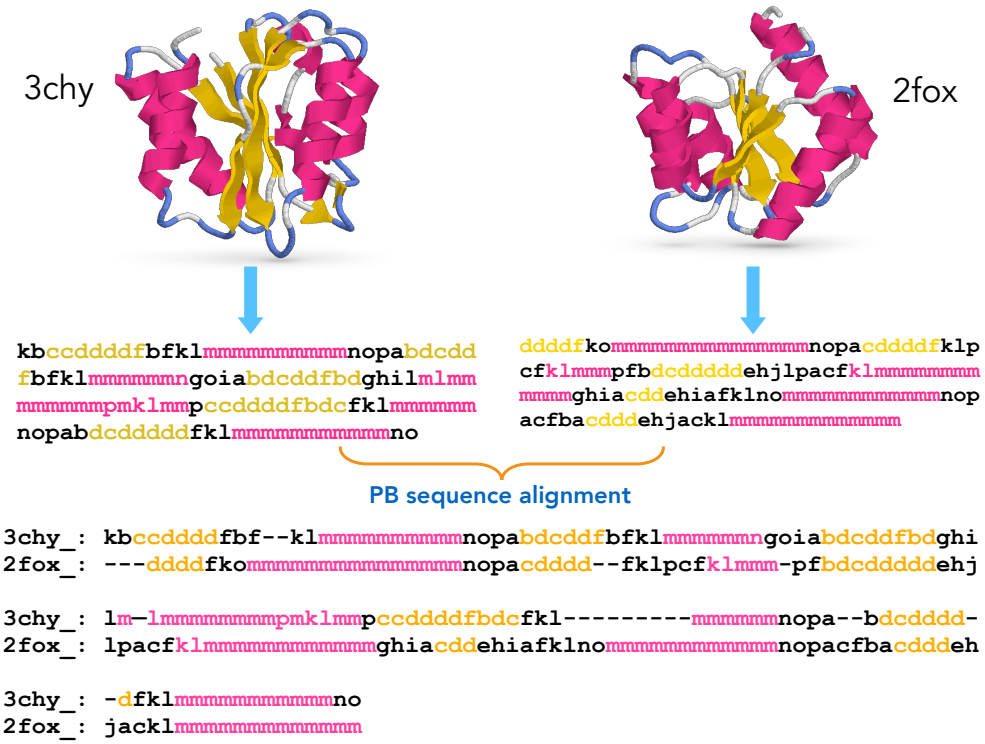


Figure 1.5: Principle of structure comparison by PB sequence alignment.

PB-kPRED

Recently, a knowledge-based local structure predicting approach has been developed to predict most probable backbone conformation of protein local structure from its sequence (Vetrivel et al. 2017). This prediction of PB sequence from protein primary sequence can be termed as Protein Blocks Prediction (PBP).

This algorithms approximates the information present in the structural database in the form of all possible PB pentamer fragments (PENTAdb) and uses this data to apply knowledge based prediction of protein backbone in the form of predicted PB sequences. This tool on average achieved an accuracy between 40.8% to 66.3% on a data set of 15,544 protein queries. Lower prediction accuracy was calculated for cases with no structural homologue available in the PENTAdb database. The accuracy of prediction is calculated on the basis of a scoring function optimised for this algorithm. Usually, an accuracy score lying above 1 is considered to be a reliable prediction.

PB-ALIGN

To facilitate easy protein structure comparison, the methodology of PB-ALIGN was developed, available on web server called Protein Block Expert (PBE) (Tyagi et al. 2006a). It uses approach akin to sequence alignment to compare protein structures using a string of PB sequences. This program has been defined such that it can perform both local and global level of structure alignment. It is built on a PB substitution matrix and a dynamic programming algorithm to align PB sequences.

Since, PBs are capable of representing a large part of protein local structure, their alignment offers a quick way to detect structural similarities between proteins.

One of the benefit of using PB-ALIGN tool is that it can be run recursively for local alignment runs such that multiple local regions can be extracted for a given PB sequence. This aids in isolating multiple domains, discontinuous domains or discontinuous domains etc. constituting a single protein structure.

1.5 Hypothesis

To boost the research in protein structure prediction field, a whole new area of research has emerged over the past few decades which focuses on development of computational protein structure prediction (PSP) algorithms. Under this field, a branch focuses on building algorithms specifically for prediction of protein structures with limited to no information known. This field is primarily termed as *Template-Free Modelling* (TFM). Slowly but steadily, success have been achieved by implementation of new ideas into solving unknown structures. Every two years, the advancements on this field are brought together and judged on a public forum known as the Critical Assessment of Structure Prediction (CASP) (Moult et al. 2003, 2016, 2018).

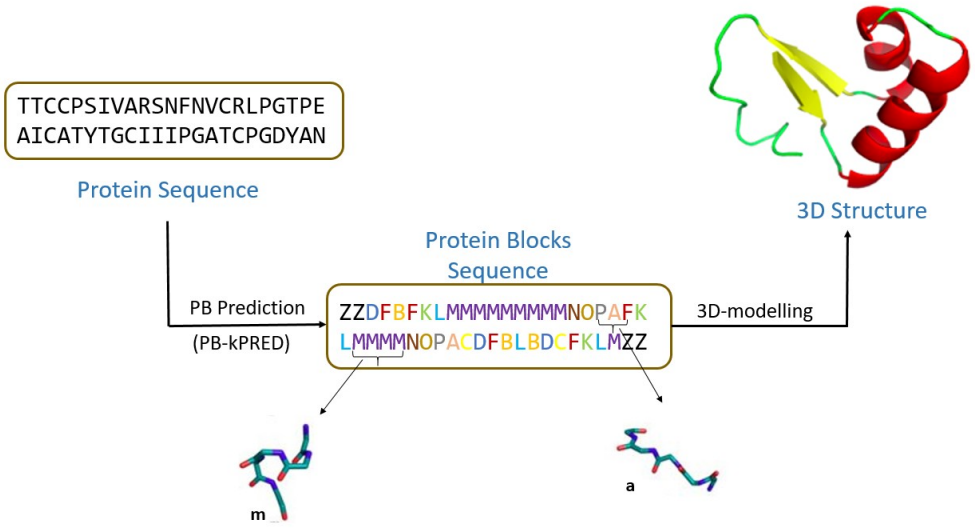


Figure 1.6: General overview of the template-free modelling protocol developed in this thesis

In the this thesis, we have tried to build one such TFM approach using PBs as a core element. In the first part of our contribution, we built upon the work done by Vetrival et. al. ((Vetrivel et al. 2017)) who developed a knowledge-based algorithm to predict probable PBs sequence from a stretch of protein sequence, a process known as PB prediction (PBP). These PB sequences more or less precisely predict the local conformation of a query protein. The second part of the project focused on building a complete workflow that could take information from the predicted PB sequence to a predicted protein structure at atomic level. Figure 1.6 provides a general overview of the basic idea behind this thesis project. This is further detailed below.

The main idea behind this work was laid on the premises that PBs are 1D representation of a 3D structure of protein. We hypothesized that the depiction of

a protein primary sequence in the form of a predicted PB sequence should provide enough information to guide the prediction of its structure at atomic level. In order to reach this objective, we built a workflow comprised of two complementary pipelines: one is a fragment generation pipeline (described in Chapter 2) and second being a pipeline for assembling fragments to model of protein structures (discussed in Chapters 3 and 4).

The PBs fragment generation pipeline aided in pulling long structurally similar stretches of proteins from a defined structural database. The information was retrieved in the form of identical PBs stretches shared between the predicted PB sequence for the query protein and assigned PB sequences in the representative structural database.

These identical PB stretches were subsequently utilised in a fragment assembly protocol towards the 3D modelling of proteins. Such an approach falls withing the category of *fragment based prediction* (Simons et al. 1997; Bonneau et al. 2001; Dong Xu et al. 2012; Alford et al. 2017) or *fragment assembly*, which has been demonstrated over the years to be the most successful and reliable procedure into *ab initio* protein structure prediction. We inclined towards forcing the homology modelling tool Modeller (Šali et al. 1993) for carrying out our fragment assembly predictions. This was decided based on earlier research done on Modeller as well as sets of experiments designed and performed to gauge the extend to which Modeller can fulfil Free modelling requirements. Preliminary experiments towards this have been documented in Chapter 3 of this thesis.

We further demonstrated to which extent we were able to optimise the current pipeline by running it on a dataset of 10 small proteins. While doing so, the modelling protocol was also tweaked by addition of two external restraints to the modelling script. All of these results have been discussed in Chapter 4 of the thesis.

Finally, Chapter 5 of this thesis dives into the implication of this research in the field of computational protein structure prediction and the possibilities of its improvement in the coming future.

We might sell ourselves short by engaging in ideological obedience.

- Anonymous

2

Fragment generation protocol

The project began with the solidification of a fragment generation pipeline defined around the usage of Protein Blocks. Since, fragment based modelling methodologies have by far shown the best outcomes when dealing with TFM protocols. Construction of a new methodology of fragment mining to further increase the efficiency of these protocols was in favour with the long term goal of the current project.

Given the nature of PBs, building a pipeline to extract local structural features in the form of representative fragment libraries was feasible given the already standardized PB-align and PB sequence prediction (PB-kPRED) algorithms. A fragment mining approach was built focusing on the usage of the previously stated algorithms into construction viable fragment libraries. This pipeline was validated on a dataset of 43 monomeric proteins and its efficiency was compared to already available fragment generation tools.

We have summarised the constant trends observed throughout these strategies as well as the strengths and limitations of the current pipeline in an article submitted to Bioinformatics journal. All these aspects of the work have been discussed at large in the sections below.

2.1 Why fragment generation?

The field of in-silico protein structure prediction via FM protocols is highly limited by the amount of conformational space that can be explored in-order to achieve an energetically minimised structure. The amount of computational power required to find the native fold of the protein is also massive and computationally expensive. To reduce the sea of probable protein conformations, Bowie and Eisenberg in 1994 published a study that focused on using multiple fragments to construct protein models (Bowie et al. 1994). This study became a hallmark of the various methodologies developed under fragment assembly approaches for building protein structures. The experiment was designed such that fragments of length 9-25 were procured from a database of known protein structures each of which shared a level of correspondence with the segment of protein sequence that was to be modelled.

By reducing the search space to local conformations, these fragments served as a way to narrow down the conformational search space while simultaneously guiding the protein towards its most probable fold. This was followed by a 'guided energy function' to bring the model closer to its native structure. Using these two criteria, they were able to model small proteins with a deviation of 2.5 - 4 Å from its native structure.

Table 2.1: Fragment generation protocols.

Strategies	**Fragment length**	**Methodology**	**Precision**	**Coverage**	**Reference**
LRFragLib	7-10	Logistic regression based fragment library	79.96% (1.5 Å)	>90%	(T. Wang et al. 2016)
SA-Frag	6-27	Structural alphabet profile-profile comparison	86.7% (2 Å)	86%	(Shen et al. 2013)
NNMAKE	9	Fragment scoring function to pick up fragments from a database; picking and collecting; final selection	38.17%	∼92%	(Gront et al. 2011)
FRAGSION	any length	Database free method for generating fragment libraries using IOHMM	24% (1-4 Å)	64%	(Bhattacharya et al. 2016a)
HHFrag	6-21	HMM based algorithm built on the tool HHPred	62 ± 16	76% (1.4 Å)	(Kalev et al. 2011)
Flib	6-20	Uses a combination of exhaustive and random libraries	∼35% (1.0 Å)		(Oliveira et al. 2015)

The table documents few of the fragment generation protocols that have been developed over the past years. It also iterates on general methodology used for developing the pipelines as well as the length of the fragments standardized in each of the algorithms.

This is when fragment assembly based prediction of protein structure became popular and to date have remained a state-of-the-art in the field of free modelling. This approach has been laid down on the basic principle that the protein sequences exhibit certain structural features locally (Han et al. 1996). Thereby, following this sequence-structure co-relation, small closely (sequence) related fragments of proteins representing a certain local structure feature can be combined together to build a protein structure. Thus, reducing the problem to the level of fragments which in turn translate to smaller conformations to search for and thereby reducing the folding problem into manageable chunks. One could also say that this method explores the sequence-structure relationships (Han et al. 1996) at a local level, thereby simplifying the folding problem.

As the research in this field gained motion, several studies were conducted to determine the optimal size of the fragments and their number required per position to achieve native-like conformation of the target protein sequence. Studies have indicated that the fragment of length of 10 residues long is of optimal size and that a minimum of 100 fragments should be explored per position to construct a model close to its nativity. (Bystroff et al. 1996; D. Xu et al. 2013).

Most of the strategies developed under this category rely on sequence comparison for determining viable fragments and exploration of secondary structure and torsion angle information to promote fragment selection for assembly (Gront et al. 2011). Studies have also been developed that have already departed from this concept and

have derived fragment libraries based on shared structural features (Shen et al. 2013). Table 2.1 describes few of the fragment generation pipelines available for *ab initio* modelling of proteins.

High precision and coverage have unanimously been used to determine the efficiency of these pipelines in generating close-to-native fragments. Precision in this context is defined as the number of fragments lying under certain rmsd cut-off value divided by the total number of fragments generated for that query sequence. Similarly, coverage is defined by having at least one fragment hit per position for a given query sequence.

The principle followed by these approaches lowers the conformational search space bar by having limited number of fragments per position to construct *ab initio* protein models. On the hind-side, this also collectively serves as a limitation for the protocols, as these algorithms are inherently limited by their ability to search large and complex conformational spaces. This also restricts them from identifying alternative conformations for a protein sequence. Furthermore in some cases the quality of fragments is affected by sequence database used to generate the libraries, specifically when it comes to inclusion of limited loop enrichment data.

2.2 Our strategy

A deeper discussion on the design, functioning and validation of the pipeline has been done in the next section. But, to give a gist of the pipeline, it depends on the local pairwise comparison of PB sequences to construct fragment libraries. The pipeline differs from the commonly developed fragment generation algorithms in a sense that it does not rely on primary sequence to derive structurally similar hits. Instead it uses PBs as a way to explore the conformational search space. Hits resembling the PB sequence (assigned or predicted) of the test protein are considered as potential fragments. Technically, it is not following the same principle as laid down by Bowie and Eisenberg, as they heavily relied on sequence similarity to select protein fragments. In this study, we have instead used one dimensional representation of a protein structure (PB sequence) to explore the conformational space boundaries. Multiple hits can be retrieved from this pipeline from which only a percentile of hits (identical hits) are considered in the final library of fragment data. This helps in reducing the space dimensions as well as help to recover the best structural representatives per position for the query sequence. The chosen minimum fragment length (7 PBs depicting 11 residues) corresponds to what has been confirmed to be the most appropriate for fragment based assembly of proteins (Section 2.1).

The quality of fragments generated is judged on the basis of precision and coverage achieved by the current pipeline. These two criteria have been readily employed by successive fragment generation pipelines to access the efficacy of the algorithms. Precision calculation in this case helps to quantify the amount of data (positive cases) that the algorithm can achieve at any given time. In our case precision is calculated for fragment hits lying below 2Å rmsd from the native structure. Thereby, quantifying the amount of fragment data generated being of good quality. Secondly, we have used coverage as a quality marker as it is able to provide an estimation of the amount of sequence space that can be enveloped by the fragment data. In other words, provide an estimation of the percentage of information that can be retrieved

for a primary sequence.

2.2.1 Protein Blocks based fragment generation pipeline

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original paper

OXFORD

Structural bioinformatics

Customised fragment libraries for *ab initio* protein structure prediction using a structural alphabet

Surbhi Dhingra [1,2], Ramanathan Sowdhamini [2], Yves-Henri Sanejouand [1], Frédéric Cadet [3,4,5,6], Bernard Offmann [1,*]

[1]Université de Nantes, CNRS, UFIP, UMR6286, F-44000 Nantes, France
[2]Computational Approaches to Protein Science (CAPS), National Centre for Biological Sciences (NCBS), Tata Institute for Fundamental Research (TIFR), Bangalore 560-065, India
[3]University of Paris, BIGR—Biologie Intégrée du Globule Rouge, Inserm, UMR_S1134, Paris F-75015, France
[4]Laboratory of Excellence GR-Ex, Boulevard du Montparnasse, Paris F-75015, France and
[5]DSIMB, UMR_S1134, BIGR, Inserm, Faculty of Sciences and Technology, University of La Reunion, Saint-Denis F-97715, France
[6]PEACCEL, Protein Engineering ACCELerator, 6 Square Albin Cachot, box 42, 75013 Paris, France

*To whom correspondence should be addressed : bernard.offmann@univ-nantes.fr

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Computational protein structure prediction has taken over the structural community in past few decades, mostly focusing on the development of Template-Free modelling (TFM) or *ab initio* modelling protocols. Fragment-based assembly (FBA), falls under this category and is by far the most popular approach to solve the spatial arrangements of proteins. FBA approaches usually rely on sequence based profile comparison to generate fragments from a representative structural database. Here we report the use of Protein Blocks (PBs), a structural alphabet (SA) to perform such sequence comparison and to build customised fragment libraries for TFM.
Results: We demonstrate that predicted PB sequences for a query protein can be used to search for high quality fragments that overall cover above 90% of the query. The fragments generated are of minimum length of 11 residues, and fragments that cover more than 30% of the query length were often obtained. Our work shows that PBs can serve as a good way to extract structurally similar fragments from a database of representatives of non-homologous structures and of the proteins that contain less ordered regions.
Availability: Data and scripts are available for download at https://frama.link/z2wh7rKC
Contact: bernard.offmann@univ-nantes.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A switch has been observed in the field of protein structure prediction (PSP) in the past few decades wherein research is being encouraged in building computational approaches, mostly focusing on Free modelling (FM) or Template-free Modelling (TFM) protocols (Maurice, 2014; Khor *et al.*, 2015). These approaches are validated through the biannual competition of CASP (Moult *et al.*, 1995) intending into resolving sequence-structure-function paradigm. The core methodologies around which the free modelling protocols have been developed are fragment-based assembly methods (Rohl *et al.*, 2004; Zhang *et al.*, 2005; Xu and Zhang, 2012), threading approaches (Roy *et al.*, 2010), physics-based methods (Klepeis and Floudas, 2003; Ołdziej *et al.*, 2005) and quite recently machine learning approaches (AlQuraishi, 2019). We have recently reviewed the progress made in the field lately (Dhingra *et al.*, 2020).

From these, fragment-based approaches (FBA) have been explored the most for the construction of *ab initio* protein models. Such methodologies rely on assembling structural fragments covering short or long stretches of amino acid sequences from a representative structural database sharing sequence similarity with a query protein. The fundamental behind FBA is that local protein sequence patterns follow a general trend of structural features (Simons *et al.*, 1997). This lead to the stipulation that the

© The Author 2020. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

local conformations for a given protein sequence can be recovered from fragments sharing local sequence similarity with local regions in existing protein structures (Kandathil *et al.*, 2018). FBA proceeds by collecting such local conformations in a fragment library and assembling them to construct potential structural models, mostly by using knowledge-based scoring functions. In short, a number of fragments are generated for each position within the target protein sequence which are then reduced to the best representative, based on different scoring criteria. The fragment lengths vary with the algorithm in question, usually lying within the range of 20 residues (de Oliveira *et al.*, 2015). Nonetheless, accurate models have been constructed using fragments as short as 3 residues long (Bonneau *et al.*, 2001; Gront *et al.*, 2011).

In general, fragment based approaches are beneficial in restricting the dimension of conformational search space by limiting the number of fragments used per position. This also serves as a major check point for such algorithms, as they inherently fall behind in exploring alternative conformations for the same sequence (Kandathil *et al.*, 2016). Recently, efforts have been made in overcoming this drawback by redesigning fragment search heuristics (Kandathil *et al.*, 2018).

Algorithms have also been developed for fragment mining using non-traditional methods. One such approach is SA-Frag (Shen *et al.*, 2013), which uses a type of structural alphabet to construct fragment libraries. It builts local profile comparison between target and template structures based on predicted SA sequences. Though, this study placed SA in the conversation of PSP, it is still not on par with the available sequence counterparts (Abbass and Nebel, 2015). Yet, this algorithm has left an open space to dig more into the ways different SAs can be exploited for structure prediction.

In the current work, we have evaluated the potential of Protein Blocks (PBs) (de Brevern *et al.*, 2000), a structural alphabet, in constructing effective fragment libraries for *ab initio* protein structure modelling. There are several types of structural alphabets (SAs) (Unger *et al.*, 1989; Camproux *et al.*, 2004; Li *et al.*, 2008) available that focus on clustering protein backbone conformations into a limited set of representative local stretches. Protein Blocks (PBs) emerged with a goal to obtain a good local structure approximation of protein 3D structures when converted into 1D PB sequences and good prediction of local structures directly from amino acid sequence (de Brevern *et al.*, 2000; Joseph *et al.*, 2010). This structural alphabet was obtained after an unsupervised clustering using a Self Organizing Map (SOM) (Kohonen, 1982; Kohonen *et al.*, 2001). Protein blocks constitute a set of 16 structural prototypes lettered *a* to *p*. Each of the PB is 5 residues in length depicting 2(*M*-1) dihedral angles, where *M* (here *M*=5) is the number of residues constituting the prototype. Using PBs, any protein structure of *n* residues long can be converted into a PB sequence of length *n*-4. PBs have been shown to approximate on average at 0.42Å every local conformation of protein structures (de Brevern, 2005).

The present work makes use of the available applications of PBs. One being translation of 3D protein structure coordinates into a string of readable 1D PB sequence by a procedure commonly named as PB assignment (PBA). Another being the prediction of the probable backbone conformation of a protein sequence in absence of secondary structure and sequence alignment profiles using a knowledge-based scoring function. This algorithm is available in the form of web-based tool called PB-kPRED (Vetrivel *et al.*, 2017) for the methodology loosely termed as PB prediction (PBP).

We have utilised this ability of PBs to approximate protein backbone into mining fragments for *ab initio* modelling of protein structures. The fragments are retained in the form of PB sequence hits from a non-redundant, non-homologous protein database. The quality of fragments are assessed for PB sequences coming from both PB assignment (PBA) and PB prediction (PBP).

The significance of this work has been centered around the idea of extensive conformational space search using a form of structural prototypes as an initial step. PBs are shown to be beneficial in such cases by primarily excavating the template database for local backbone conformations relying primarily on its local conformation.

2 Methods

2.1 Curated Template Database

A database of non-redundant protein structures was set up. Protein chain entries were downloaded from RCSB database (www.rcsb.org) (Berman *et al.*, 2000) based on the following criteria: (a) Experimental type: X-ray crystallography, (b) Resolution: $\leq$ 3Å, (c) R-factor: $\leq$ 0.2 and (d) length of the protein: $\geq$ 40 residues. These accounted for a total of 23,989 unique protein chains. The hits were further clustered at 30% sequence identity using the KClust algorithm (Hauser *et al.*, 2013), resulting in a total of 7632 protein chains. Additionally, any protein with chain breaks were removed from the clustered sequence set, finally resulting in 5391 unique chains.

PB sequences were assigned for each of these 5391 templates in the final database using an in-house script which alongside generates equivalent DSSP output files and torsion angle files. Secondary structure assignments for each protein chain were congregated using Pdb-tools (Rodrigues *et al.*, 2018). All these files along with protein sequence corresponding to each template constituted our curated database.

2.2 Query Dataset

The query dataset used in this work is derived from a previous publication that focused on building fragment libraries (de Oliveira *et al.*, 2015). This dataset is comprised of 43 query protein structures ranging from 59 to 508 residues in length and is provided in Table 1. Each of them is a monomer and has been further categorized into four main SCOP family classes, specifically, all alpha, all beta, alpha plus beta (α+β) and alpha and beta (α/β).

Each query structure was further transformed into a PB sequence using two concurrent procedures:

- PB-Assignment (PBA) procedure: in this case the query PB sequence is retrieved directly from the PDB coordinate files using an in-house PB assignment script, following the basic principle of PB generation, i.e conversion of 3D information into 1D sequence as described before (de Brevern *et al.*, 2000). In short, the dihedral angles of all constitutive overlapping pentapeptides from a query are used to classify each of them in one of the 16 PB classes *a* to *p*. Secondary structure representations for the target protein were obtained using DSSP assignment protocol (Kabsch and Sander, 1983).
- PB-Prediction (PBP) procedure: here the potential query PB sequences were manually predicted using the PB-kPRED web tool (Vetrivel *et al.*, 2017). This step was done to account for the loss of information endured due to probabilistic estimation of PBs per position for a given sequence. PB-kPRED features a prediction quality criterion in the form of a prediction score and a standardized accuracy percentile. Generally, a score of < 1 concludes unreliable prediction. The prediction accuracies and kPRED scores for the query dataset are shown in Table 1. Secondary structure representations for the target protein were estimated by the secondary structure prediction tool Psipred (McGuffin *et al.*, 2000).

Table 1. **The query dataset and its characteristics.** The Table summarises the lengths in terms of number of residues and the estimated accuracies of PB predictions as observed for each protein from the query dataset. A PB-kPRED prediction score (last column) of 1 or above has been determined as an indicator of reliable PB sequence prediction (Vetrivel et al., 2017). All the proteins in the dataset, except for one, laid above this score cut-off value.

PDB id	SCOP Class	Length (AA)	Length (PDB)	Accuracy (%)	kPred score
1AIL	all-α	73	70	63.6	1.5
1RRO	all-α	108	108	73.1	1.99
1U61	all-α	138	127	65.2	1.43
1SL8	all-α	191	181	77.5	2.23
1QUU	all-α	250	248	63.8	1.51
1T5J	all-α	313	301	65.8	1.61
1PO5	all-α	476	465	Too low	Too low
1MHN	all-β	59	59	69.2	1.79
1TEN	all-β	90	90	72.1	1.94
2G1L	all-β	104	103	70.5	1.86
1IFR	all-β	121	113	67.1	1.68
1BFG	all-β	146	126	83.3	2.53
2FR2	all-β	172	161	70.4	1.85
1EE6	all-β	197	197	79.5	2.33
1UAI	all-β	224	223	71.1	1.9
2C9A	all-β	259	259	68.9	1.77
1O4Y	all-β	288	270	66.1	1.63
1HG8	all-β	349	349	81.9	2.46
1NKG	all-β	508	508	82.7	2.5
1VJW	α+β	60	59	59.4	1.28
1MWP	α+β	96	96	89.7	2.86
1GNU	α+β	117	117	80.6	2.39
1R9H	α+β	135	118	85.8	2.66
206L	α+β	164	162	81.9	2.45
2FS3	α+β	282	280	82.1	2.47
1DZF	α+β	215	211	85	2.62
1DXJ	α+β	242	242	74.2	2.1
1MAT	α+β	264	263	83.9	2.56
1JKS	α+β	294	280	80.8	2.36
1MC4	α+β	370	369	77.3	2.22
2FKF	α+β	462	455	80.3	2.37
1H75	α/β	81	76	74.3	2.06
1IU9	α/β	111	111	68.8	1.77
1E6K	α/β	130	130	77.7	2.24
1P90	α/β	145	123	64.8	1.56
1FTG	α/β	168	168	79.8	2.34
1QCY	α/β	193	193	75.9	2.14
2A14	α/β	263	257	63.4	1.48
1IZZ	α/β	283	276	77.6	2.23
1QUE	α/β	303	303	82.8	2.51
1KRM	α/β	356	349	81.2	2.42
3BSG	α/β	414	404	77.2	2.21
1PGN	α/β	482	473	72.7	1.98

2.3 Fragment Mining

A procedure was setup to mine non-homologous fragments from the curated template database. To avoid any bias in the benchmarking of our method, the primary step of the procedure involved looking for sequence homologs in the template database. Any template protein sharing more than 30% sequence identity with any the query protein sequence was considered for removal: pairwise sequence identity was calculated for each query against the entire curated template database using Needleman and Wunch global alignment algorithm (Needleman and Wunsch, 1970). Any hit sharing more than the defined sequence identity cut-off with the target sequence was removed from the template database for that run. A local PB alignment tool, PB-Align (Tyagi *et al.*, 2006) was used to generate alignment hits for minimum length of 7 PBs corresponding to 11 amino acid long sequence stretches. PB alignment hits whereby the query PB sequence was identical to the template fragment were collected as viable fragments, since it should correspond to identical backbone conformations. Further fragment quality assessment was performed on all identical hits using the criteria discussed below. The overall process of fragment generation is summarised in Figure 1.

2.4 Fragment Quality Assessment

Two sets of experiments were conducted to assess the quality of fragments generated for each protein of the query dataset: one corresponds to real life scenario and considered PB sequences obtained from the PBP procedure, and the other corresponds to the positive control which considered PB sequences obtained with the PBA procedure.

For every position, fragment quality was assessed by fitting fragment hits over query coordinates for that position using the McLachlan algorithm for rapid comparison of protein structures (McLachlan, 1982) implemented in the ProFit program (www.bioinf.org.uk/software/profit/). This criterion gauged the "precision" of the procedure in generating viable fragments. Here, precision is defined as the number of fragments lying under a *rmsd* of 2Å divided by the total number of fragments generated for the query protein. The second main criterion that was examined was coverage which was defined as the number of positions in the query sequence for which at least one fragment was obtained by the fragment mining procedure. Other assessment criteria that were tested include calculation of amino acid sequence identity and similarity and secondary structure identity between each fragment hit against the target sequence.

3 Results

3.1 Template database

The final template database consisted of 5391 protein chains clustered at < 30% sequence identity and having a resolution of < 3Å. The examination of the secondary structural elements distribution in the curated database showed that it is populated with 44.76% helices, 27.82% of strands and 27.42% coils and loops. In terms of PBs, the database corresponded to 29.9% PB *m* the central part of alpha-helices, 19.13% PB *d* the central part of a strand and 51% of other PBs (or coils), which approximated to the general protein secondary structure distribution (regular and irregular) (Joseph *et al.*, 2010).

Our database contained templates from 10 out of 12 SCOP representative classes, with the majority being associated to the 4 main SCOP classes. Out of 5391 proteins, 1962 found no corresponding SCOP hit. This might be due to delay in synchronisation of structural annotation data across platforms.

3.2 Fragment mining and generation

The fragment generation procedure is illustrated in Figure 1. Any protein from our curated template database that could have been a probable sequence homologue with a protein from our que query dataset was excluded on the fly from the template database. In this scenario, any template sharing >30% identity with the query sequence was considered as a *homologue*. This ensures the premises of Free Modelling and the legitimacy of our pipeline in picking up good fragments in absence of sequence homologues.

An average of ∼60k hits were generated by local PB-align for each query sequence with least being 34,655 for 1AIL and maximum being 86,866 for 1HG8. The fragments were further filtered on the basis of PB sequence identity. All the PB hits that were 100% identical to the query

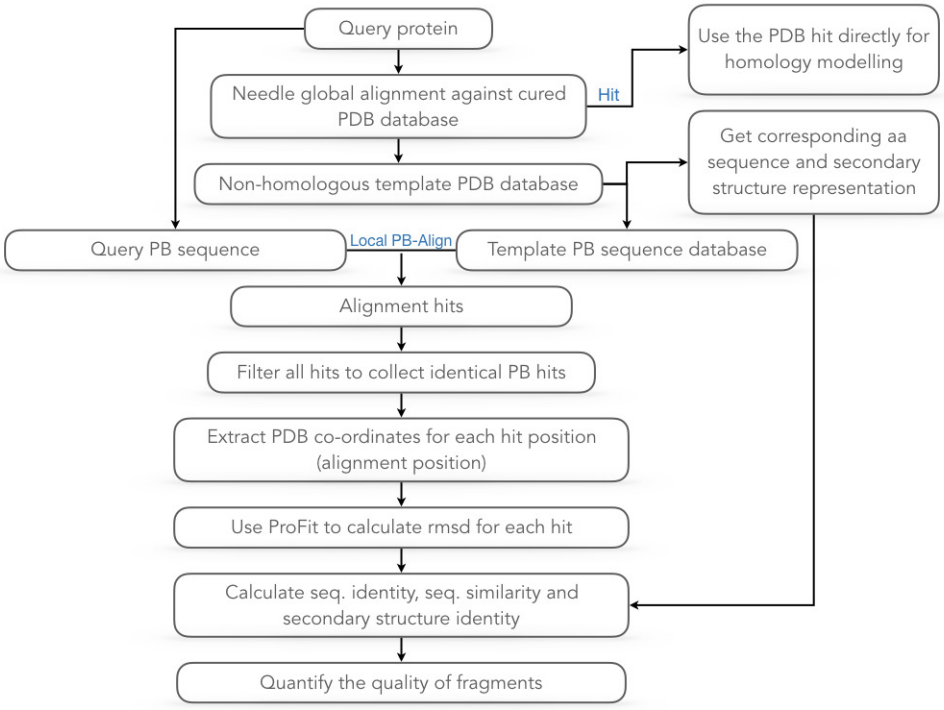


Fig. 1. Fragment Generation Procedure. Shown is a schematic representation of the general flow of fragment generation carried out using protein blocks as primary source. Local PB alignment between query and template PB sequence effectuates fragments of minimum length of 11 amino acid residues. The alignment hits are then analysed based on several criteria to procure the quality of the generated fragments.

PB sequence, whether this was assigned (PBA) or predicted (PBP), were chosen as best representative fragments. This functioned as a preliminary filtering step by limiting the fragment search area to the best PB fit. In doing so, the average number of fragments reduced to ∼13k hits per query without significantly affecting the overall coverage. The lowest number of identical hits was documented for query protein 1EE6 with 3761 hits and maximum for 1MWP with 21732 hits. Detailed counts of the total hits before and after filtering are provided in Supplementary Tables 1 and 2 for PBA and PBP respectively.

As seen in Figure 2, the maximum number of fragments laid below 15 residues in length. This remained the case for all the query proteins, whatever their total length. The longest fragments obtained were 65 and 61 residues long for PBA and PBP schemes respectively.

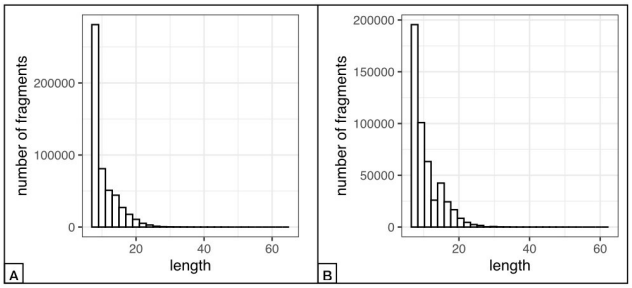


Fig. 2. Fragment length distribution : Protein Block Assignment (left) and Protein Block Prediction (right). In both the cases maximum number of fragments are lying at the assigned minimum fragment length of 7 residues. The histogram is plotted for fragment length against number of fragments generated. There is an exponential drop in number of fragments generated with the increase in fragment length.

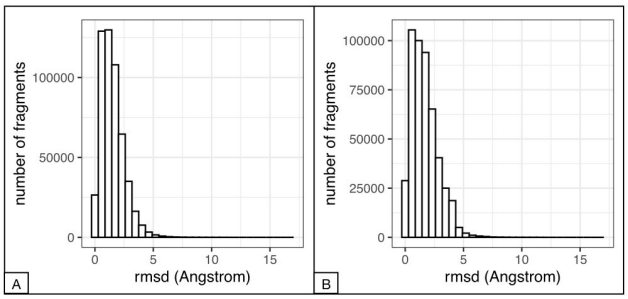


Fig. 3. Rmsd distribution : Protein Block Assignment (left) and Protein Block Prediction (right)). The graph is plotted for total number of fragments generated from all the queries in test dataset against the rmsd observed when these fragments are fitted over original query structure. For both the cases of PB production, maximum number of fragments lie below the rmsd cut-off value of 2.5 Å.

The overall effectiveness of the fragment generating pipeline was judged on the basis of *rmsd* calculations and coverage attained by the fragments for all the queries. Figure 3 shows an overview of rmsd distribution as observed in all the cases for both PBA and PBP schemes. It is clear from the graph that most of the fragment hits lie below the *rmsd* cut-off of 2.5Å which approximates to 85% of the generated fragments.

A barplot depicting the overall percent of sequence space covered (coverage) for each query protein is shown in Figure 4. Higher coverage was observed in case of PBA with an average of 96.5% in comparison to PBP with an average of 93.9%. For 6 cases, PBP gave better percentage coverage than PBA cases. Protein queries belonging to all-β SCOP class showed lowest percentage coverage when compared to queries from the other three main SCOP classes.

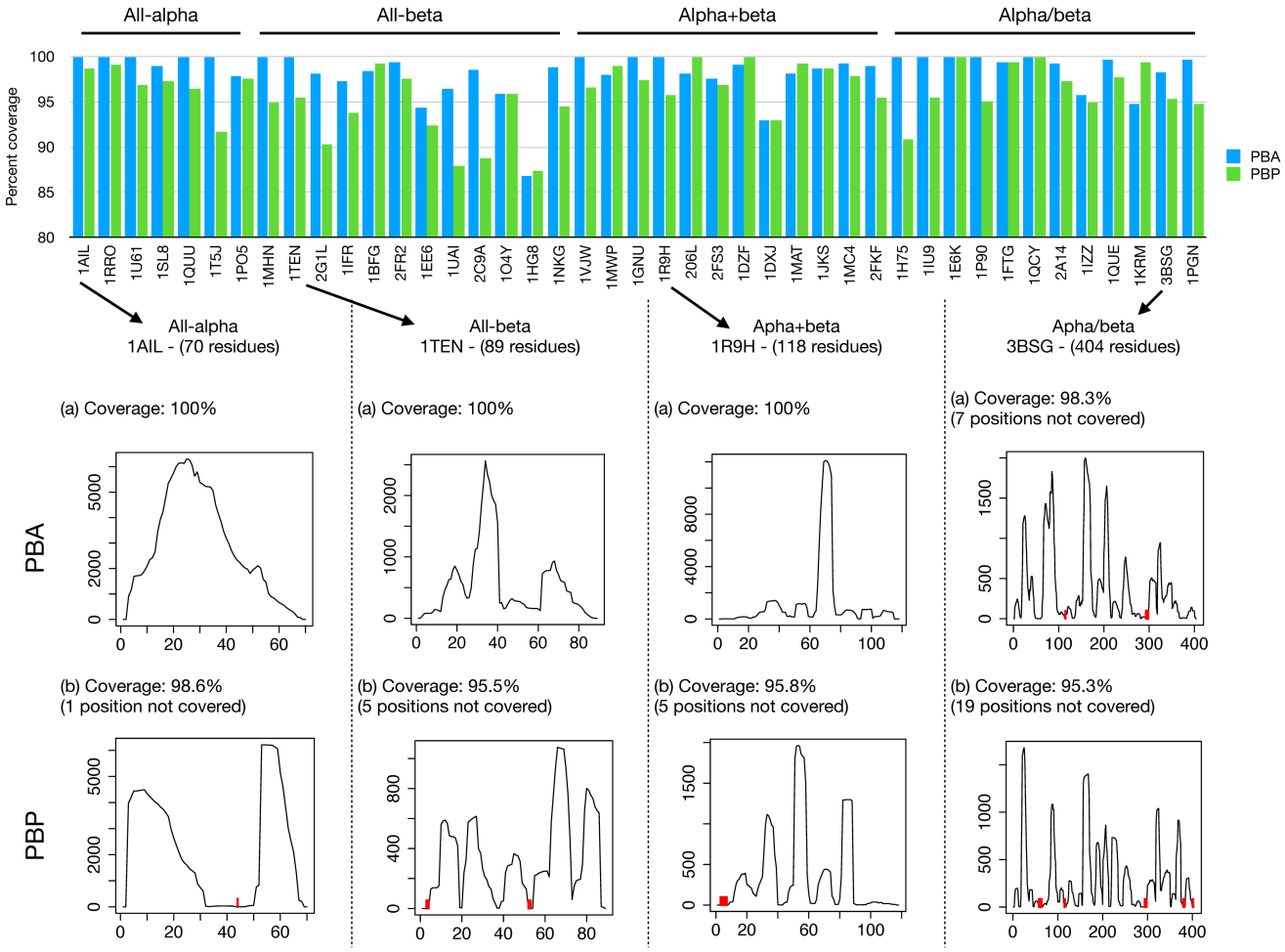


Fig. 4. Coverage analysis of fragments generated. *Top*: barplot showing, for each query protein from our query dataset, the percentage of positions covered by a fragment (y axis is from 80-100% and x-axis are the labels for the PDB codes of the queries). Results obtained with PBA and PBP procedures are shown in blue and green respectively. Query proteins are grouped according to their SCOP classes. The queries with lowest coverage belongs to SCOP β class. *Bottom*: detailed coverage results illustrated for 4 query examples for both PBA and PBP schemes. The x-axis represents the residue positions and the y-axis is the raw count of the number of fragments that covered each of the positions. Positions not covered by any fragment are shown in red on the plots.

The number of hits observed per position also varied between the two cases for each query. This is illustrated in Figure 4 for four examples drawn one from each SCOP class and in Supplementary Figures 1 and 2 for all the queries. Both coverages for PBA and PBP schemes are shown. In many queries, the coverage was heterogenous with some positions that were covered by a low number of fragments and some positions with a high number of hits. This is well illustrated by the spikes in the number of hits along the query protein length. These highly covered stretches mostly correspond to regions with canonical secondary structures stretches. Noteworthily, on average, lower number of hits were observed per position in case of PBP without affecting the overall coverage of the query sequence. This can be explained by the loss of accuracy procured in case of PB predictions (Figure 4 and Supplementary Figures 1 and 2).

Table 2. Precision of the pipeline in generating fragments at three *rmsd* cut-off values.

SCOP Class	PBA scheme (*rmsd* Å)			PBP scheme (*rmsd* Å)		
	$\leq$ **1.5**	$\leq$ **2.0**	$\leq$ **2.5**	$\leq$ **1.5**	$\leq$ **2.0**	$\leq$ **2.5**
all α	66.2	80.8	89.9	51.7	67.5	80.0
all β	46.1	71.8	84.5	28.9	47.7	64.4
α + β	55.1	73.6	85.4	51.4	68.7	80.3
α / β	61.3	75.4	85.6	58.1	73.4	83.9
Average	57.2	75.4	86.4	47.5	64.3	77.2

The table describes the precision gained by the fragment generation pipeline as seen at three different *rmsd* cutoff values of 1.5Å, 2Å and 2.5Å. It also compares the precision that could be achieved at the best case scenario of exact PB sequence (PBA) when compared to predicted PB sequences (PBP). At the *rmsd* cut-off of 2Å, the pipeline is able to retain an average precision of 64%. Lowest precision has been noted for hits belonging to SCOP class all-β proteins.

3.3 Assessment of fragment quality

The quality assessment was primarily done by fitting fragment hits onto the original structures using a protein structure least square fitting program. It was measured in terms of precision, which is defined as the percentage of number of hits lying under a given *rmsd* cut-off. Table 2 summaries the precision percentage values obtained for each SCOP class analysed in the current work at three *rmsd* cut-off values. An overall precision of 75.4% and 64.3% was calculated for PBA and PBP respectively at the

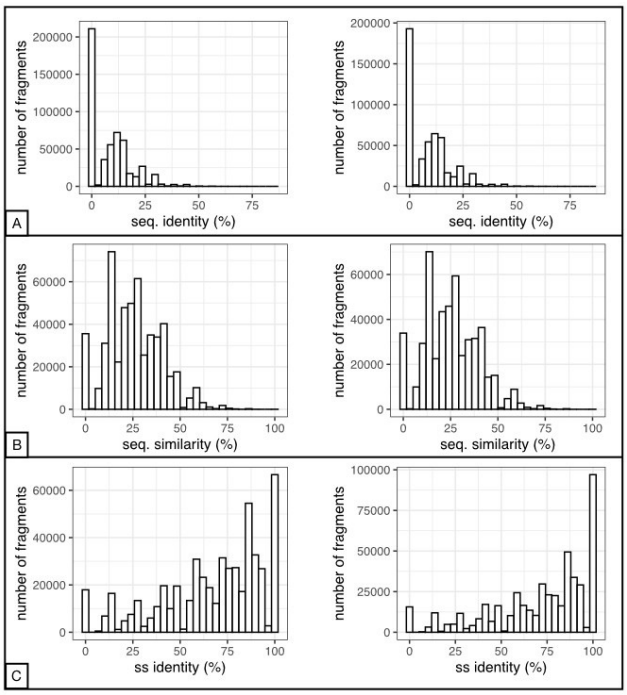


Fig. 5. Qualitative analysis of fragment hits.. The graphs above exhibit the distribution of sequence identity, sequence similarity and secondary structure identity as observed in case of fragments generated for two working cases of PB generation, PBA (left) and PBP (right). [A] The graphs shows the distribution of amino acid sequence identity. [B] Amino acid sequence similarity distribution as observed between query sequences from test dataset and the fragments generated by the pipeline. [C] Secondary structure identity shared between query sequences and fragments generated by the pipeline. Both the cases of PB generation, i.e PBA and PBP, follow a similar trend of identity and similarity distribution.)

rmsd cut-off of 2Å. Least precision was quantified for SCOP class all-beta reaching up to 71.8% and 47.7% for PBA and PBP schemes respectively. This could be due to the reason that by nature β-strands are amenable to distortions such as twists which are harder to model while also showing a broader range of the ϕ,ψ distribution for their conformations. Rest of the classes had higher precision levels lying above 65% collectively. With more stringent *rmsd* cut-off values, the precision dropped.

Amino acid sequence identity and similarity and secondary structure identity were calculated for all fragments against corresponding position in the query sequence. Figure 5 depicts their overall distributions. Note that amino acid sequence identity distribution is dominated by fragments sharing no identity (0%) with the query sequence. On the other hand, an increase in sequence similarity is observed in the plots, but the graphs are still skewed towards low values, indicating low similarity correspondence between query and template amino acid sequences. This was not the case for secondary structure identity: the distribution shifted towards right with maximum hits reaching up to 100% identity. This attributes to the nature of PB sequences which depict 1D sequences associated to the local backbone conformation, thus indicating towards it being an objective criteria for assessing and qualifying fragments.

Interestingly, 40% of the fragments lying under *rmsd* of 2Å didn't share any sequence identity with the query sequence and 90% shared less than 37.50% sequence identity. This result could be appreciated as it aligns with the specifics of template database being non-homologous, thereby lowering the chances of finding the exact same amino acid sequence hit altogether for the variable lengths of fragments extracted. The aggregation was consistent in both PBA and PBP schemes: 50% of the data shared <25% sequence similarity with query sequence with 75% of the data reaching ∼36% sequence similarities in case of PBP. Noteworthily, an average secondary structure identity of 85% was noted for fragments lying below the 2Å mark with most being 100% identical for both PBA and PBP.

A ROC curve analysis (Figure 6) performed for the aforementioned elements against *rmsd* (2Å) confirmed that secondary structure identity criteria was indeed the best amongst the three tested criteria for prioritising fragments. The sensitivity and specificity curves for individual SCOP classes are further detailed in Supplementary Figures 2.

A visual inspection of the fragments in Pymol (DeLano *et al.*, 2002) shows that the global fold of the target structures can be retained by the fragments. Sample of this visualization is illustrated in Figure 7.

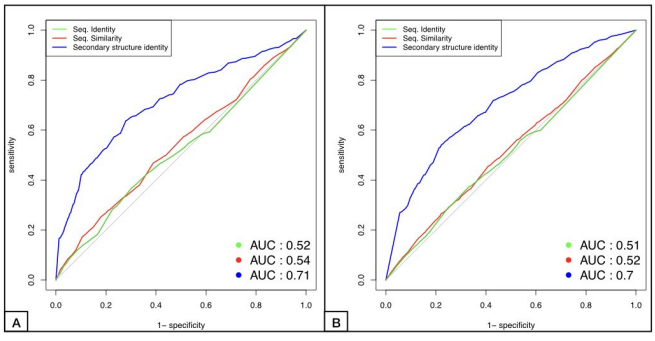


Fig. 6. ROC analysis using three objective criteria. A) and B) represent results for PBA and PBP schemes respectively. For all the fragments lying below the rmsd cut-off of 2Å, ROC curve was plotted to visualize effective criteria for choosing viable fragments. The three criteria are: (i) amino acid sequence identity (green), (ii) amino acid sequence similarity (red) and (iii) secondary structure identity (blue). Sequence identity and similarity lie on the mark of 50% depicting no direct correlation with the quality of the fragments generated. SS identity, on the other hand, has an AUC of approximately 70%, ensuring its role in recuperating good fragments.

4 Discussion

In the current paper, protein blocks have shown to hold a potential for effective fragment mining towards the goal of *ab initio* protein structure prediction. The work concludes that a huge amount of good quality fragments can be extracted from structural database represented in the form of PBs in absence of sequence homology. It is a very simple pipeline that mimics amino acid sequence pairwise alignment algorithms to detect, thanks to PB alignment, local structural hits within the working database of structural templates. This shifts the fragment-based approaches from relying on amino acid sequence to directly accessing structural patterns from a curated database. Huge pool of structural stretches of varying lengths were recovered for each query in the dataset equating to congregated fragments data. This importantly included loop regions that connect regular secondary structures.

The primary requirement of the study was to effectively depict amino acid sequence into reliable PB sequences while retaining the probable local conformation. It was attained by using a previous work done in our group ((Vetrivel *et al.*, 2017)) for optimizing possible PB per position for each overlapping pentapeptide within a protein sequence (PB-kPRED webserver). The calculated PB prediction scores for all the target proteins from the dataset were above the standardized threshold, with an exception of one (1PO5). The average accuracy of PB-prediction was noted to be 74.98 % for 42 out of 43 targets in the dataset, which was ample to guide valid fragment picking through the pipeline.

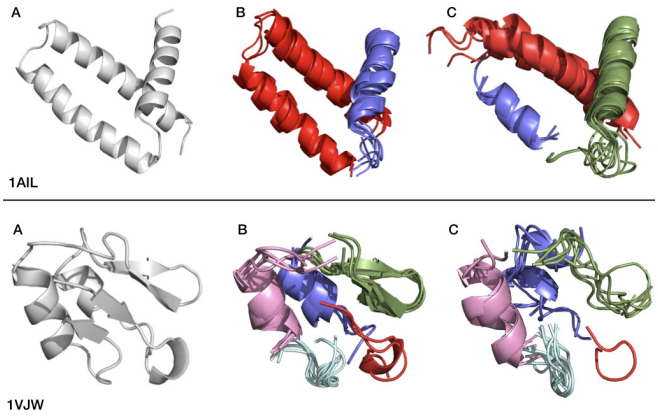


Fig. 7. Superimposition of fragments generated onto the original query structures. The figure shows the original structure of two queries from the test dataset, namely, 1AIL and 1VJW, along with the clusters of best fragments retained for each. Here, (A) is original structure, (B) is the fragments clusters obtained for PBA scheme and (C) is the fragments clusters obtained for PBP scheme .

Prior to running fragment extraction protocol, it was insisted upon abiding by the constraints of free modelling approaches to avoid biasing the aftermath of fragment quality assessment. In line with it, any potential homologous template sequence was removed from the curated database as a preliminary step. This consequently produced personalised database for each query sequence. This step also ensured objectivity of the pipeline when inspecting the quality of fragments generated through PB-predicted sequences.

From our analysis, it was noted that fragments generated via this pipeline show no correspondence between amino acid sequence identity or similarity and *rmsd*. On the other hand, secondary structure juxtaposition between the template fragments with the predicted secondary structure for the query showed excellent relevance. The data unanimously favoured secondary structure identity comparison, with a large proportion of fragments that hit 100% secondary structure identity. A similar trend was observed when examining the correlation between *rmsd* and secondary structure identity. Higher secondary structure identity corresponded to lower *rmsd* between template fragments and query (data not shown). This is not surprising since PB itself is a representation of the backbone.

The quality of fragments was measured by fitting fragment coordinates onto original structure and was termed here as "precision". This was evaluated for the 4 main SCOP classes, all agreeing with the cumulative results of the pipeline. Significant discrepancies were noted for class all-beta. Here the accuracy of fragment prediction did not reach the mark of other classes. In general, it has been noticed that alpha folds are easier to map back to than beta strands. This was even seen in the current study where the 'all-beta' queries showed overall higher *rmsd* and more scattered secondary structure identity distribution.

The pipeline was tested against the best possible case of PB representation, i.e PB assignment (PBA) and was compared to real life scenario of approximating PB sequence for the given query through PB prediction (PBP). This accounted for the effects of missed data (information) on the process of fragment picking. Our results shown that the overall quality of fragments retained in case of PBP was not compromised.

The quality of fragments was judged on the basis of two common criteria used in earlier work done on fragment library generation pipelines, i.e precision and coverage. Though, the overall efficiency of our pipeline is on a par with other algorithms available like HHFrag and NNMake that report precision of 62.16% and 38.17% respectively (de Oliveira *et al.*, 2015). Higher precision of HHFRAG can be attributed to the presence of sequence homologues in the working database (de Oliveira *et al.*, 2015) which is not the case here. SAFrag, another SA based fragment mining pipeline, has shown a precision of 86.7% (Shen *et al.*, 2013). Their SAFrag strategy was built on the grounds similar to HHFrag and uses HMM-based profile-profile comparison to generate fragment hits. It makes use of structural alphabets that describe 27 states. The pipeline carried out profile comparison based on the pre-segregated or partitioned query sequence segments into sub-sets of varying lengths of 6 to 27 residues. These sub-sets were then used to search for similar profiles within the established structural data banks. It has to be noted that SAFrag used two types of structural data banks: PDB25 and PDB50 for fragment generation. The higher precision and coverage of this algorithm can be attributed to the grounds that during the fragment search they accept the inclusion of the target structure as well as data coming from structural homologs as well in their template database.

The coverage attained in our work lies on the higher range when compared to these other algorithms (HHFrag - 71%±13, NNMake - ∼92%) (de Oliveira *et al.*, 2015): lowest coverage was observed in case of all-beta class, while the rest showed a coverage of above 90%. It was also observed that fragments predominantly belonging to SS class of alpha-helix were predicted more accurately than other secondary structure classes.

The primary results produced by our procedure are promising and can be worked on towards its further refinement. Improvements could be achieved by both repopulating the core of our curated structural database to increase the variability in fragment selection and the PB-PENTAPEPT database underlying PB-kPRED. The latter is expected to increase the accuracy of PB prediction which in turn will influence the precision and coverage. Another way of improving the precision could be decreasing the minimum length of the fragment, which was not done initially as the coverage achieved was still good.

The procedure that was developed in this study drifts from conventional fragment mining approaches by using PB sequences instead of amino acid sequences for producing fragment libraries. We have also challenged the idea of relying on sequence similarity for identifying reasonable fragments. It further discriminates from SA-Frag for it focused on building pairwise PB sequence alignment profiles between the target protein and templates from a database without any imposed length constraints and without taking advantage of sequence homologues or distantly related proteins.

5 Conclusion

Our work shows that structural alphabets are useful to find fragments for prediction of protein structures in template-free modelling approaches. SAs are indeed more liberating when compared to amino acids sequence based fragment libraries due to their potential in accessing a more expanded conformational space. This aids in broadening the search space without losing the potential to capture the native fold of the target protein. They can successfully be implemented in foraging potential structural homologues for proteins sharing otherwise low or no aminoc acid sequence similarity. We show here that fragments covering the entire length of small proteins can be readily generated easing the task for structure prediction protocols.

Since, PBs are local conformations represented in the form of 1D sequence, the tenacity of pulling fragments sharing similar fold is higher when compared to the possibility of that happening with amino acid sequence comparison. Longer fragments representing entire domains can be extracted using the protocol. In general, PBs hold a promising step towards protein structure prediction protocols using the local conformations as a starting point.

Acknowledgements

The authors thank Dr Alexandre G. de Brevern and Prof Narayanaswamy Srinivasan for fruitful discussions on this work.

Funding

This work has been supported by the Conseil Régional de La Réunion and Fonds Social Européenin the form of a PhD scholarship to SD under tier number 234275, convention number DIRED/20161451. BO is thankful to Conseil Régional Pays de la Loire for support in the framework of GRIOTE grant.

Competing interests

F.C. is linked to Peaccel. SD, RS, YHS and BO declare no competing interests.

References

Abbass, J. and Nebel, J.-C. (2015). Customised fragments libraries for protein structure prediction based on structural class annotations. *BMC bioinformatics*, **16**(1), 136.

AlQuraishi, M. (2019). AlphaFold at CASP13. *Bioinformatics*, **35**(22), 4862–4865.

Berman, H. M. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Research*, **28**(1), 235–242.

Bonneau, R. *et al.* (2001). Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, **45**(S5), 119–126.

Camproux, A.-C. *et al.* (2004). A hidden Markov model derived structural alphabet for proteins. *Journal of Molecular Biology*, **339**(3), 591–605.

de Brevern, A. G. (2005). New assessment of a structural alphabet. *In Silico Biology*, **5**(3), 283–289.

de Brevern, A. G. *et al.* (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins: Structure, Function, and Bioinformatics*, **41**(3), 271–287.

de Oliveira, S. H. *et al.* (2015). Building a better fragment library for *de novo* protein structure prediction. *PloS One*, **10**(4).

DeLano, W. L. *et al.* (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography*, **40**(1), 82–92.

Dhingra, S. *et al.* (2020). A glance into the evolution of template-free protein structure prediction methodologies. *arXiv preprint arXiv:2002.06616*.

Gront, D. *et al.* (2011). Generalized fragment picking in Rosetta: design, protocols and applications. *PloS One*, **6**(8).

Hauser, M. *et al.* (2013). kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinformatics*, **14**(1), 248.

Joseph, A. P. *et al.* (2010). A short survey on protein blocks. *Biophysical Reviews*, **2**(3), 137–145.

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

Kandathil, S. M. *et al.* (2016). Toward a detailed understanding of search trajectories in fragment assembly approaches to protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, **84**(4), 411–426.

Kandathil, S. M. *et al.* (2018). Improved fragment-based protein structure prediction by redesign of search heuristics. *Scientific Reports*, **8**(1), 1–14.

Khor, B. Y. *et al.* (2015). General overview on structure prediction of twilight-zone proteins. *Theoretical Biology and Medical Modelling*, **12**(1), 15.

Klepeis, J. and Floudas, C. (2003). ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical Journal*, **85**(4), 2119–2146.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, **43**(1), 59–69.

Kohonen, T. *et al.*, editors (2001). *Self-Organizing Maps*. Springer-Verlag, Berlin, Heidelberg, 3rd edition.

Li, S. C. *et al.* (2008). Designing succinct structural alphabets. *Bioinformatics*, **24**(13), i182–i189.

Maurice, K. J. (2014). SSThread: Template-free protein structure prediction by threading pairs of contacting secondary structures followed by assembly of overlapping pairs. *Journal of computational chemistry*, **35**(8), 644–656.

McGuffin, L. J. *et al.* (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, **16**(4), 404–405.

McLachlan, A. D. (1982). Rapid comparison of protein structures. *Acta Crystallographica Section A*, **38**(6), 871–873.

Moult, J. *et al.* (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, **23**(3), ii–iv.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443–453.

Ołdziej, S. *et al.* (2005). Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proceedings of the National Academy of Sciences*, **102**(21), 7547–7552.

Rodrigues, J. P. *et al.* (2018). Pdb-tools: a swiss army knife for molecular structures. *F1000Research*, **7**.

Rohl, C. A. *et al.* (2004). Protein structure prediction using Rosetta. In *Methods in enzymology*, volume 383, pages 66–93. Elsevier.

Roy, A. *et al.* (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, **5**(4), 725.

Shen, Y. *et al.* (2013). Detecting protein candidate fragments using a structural alphabet profile comparison approach. *PloS one*, **8**(11).

Simons, K. T. *et al.* (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, **268**(1), 209–225.

Tyagi, M. *et al.* (2006). A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins: Structure, Function, and Bioinformatics*, **65**(1), 32–39.

Unger, R. *et al.* (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins: Structure, Function, and Bioinformatics*, **5**(4), 355–373.

Vetrivel, I. *et al.* (2017). Knowledge-based prediction of protein backbone conformation using a structural alphabet. *PloS One*, **12**(11).

Xu, D. and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, **80**(7), 1715–1735.

Zhang, Y. *et al.* (2005). TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins: Structure, Function, and Bioinformatics*, **61**(S7), 91–98.

2.3 Other results and in-depth discussion

2.3.1 Fragment quality assessment

The number of fragment hits congregated after running the PB fragment generation pipeline was huge despite reducing the redundancy of hits by considering just the identical PB hits for the final analysis. The quality assessment of fragments was done to demonstrate the efficiency of the pipeline in capturing local structural features close to that of the target protein sequence. This qualification criteria also influenced the choice of fragment per position that can be used for modelling the protein.

The initial qualification measure was a simple calculation of *rmsd* as observed between the query protein and the template hit for that position. This was done for each pair of identical PB alignment hit tabulated for the query and template PB sequences. The template co-ordinate files equivalent to the sequence hit for each position were built from the representative PDB database standardized for the current work and saved separately. These partial coordinate definitions were generated by using a PDB manipulation tool called PDB_Tool (v 4.80) (S. Wang et al. 2008, 2013). For the sake of continuation, we will call these truncated template coordinate files as 'pseudo-templates'. Once created, these pseudo-template files were used along with the original query structure to calculate the rmsd per position for the query sequence. The calculations were carried out by using a protein structure comparison algorithm called ProFit ((http://www.bioinf.org.uk/software/profit/).

The analysis of the rmsd distribution (Table 2.2) revealed that for the three out of four SCOP classes, up to 80% of the fragment data generated lay under 2.5 Å deviation from the original query structure. Thereby, confirming that most of the fragment data retrieved by the pipeline are of good quality. But, this distribution dropped significantly (64.4% at 2.5 Å) in the case of hits belonging to SCOP class - all beta. These results are in line with the results observed in the sections discussed below where the fragments hits belonging to this SCOP class did not follow the general trend of the analysis.

Table 2.2: SCOP Class - RMSD distribution

SCOP Class	**$\leq$ 1.5 Å**	**$\leq$ 2.0 Å**	**$\leq$ 2.5 Å**
all alpha	51.7	67.5	80
all beta	28.9	47.7	64.4
a/b	58.1	73.4	83.9
a+b	51.4	68.7	80.3

This table shows the percentage of fragment data that lies below three rmsd cut-off values of 1.5, 2 and 2.5 Å respectively for each SCOP class within the validation dataset used in the study. Three out of four SCOP classes show similar progression in the range of fragment hits covered under each rmsd cut-off. All beta SCOP class showed the least percentage of fragment data collection at the highest rmsd cut-off of 2.5 Å chosen for the current work.

2.3.2 Sequence similarity and Secondary structure identity distributions

The other measures considered to prioritize fragment selection for modelling purposes was principally based on shared sequence similarity and secondary structure identity between the query and the fragment hits. These calculations helped to visualize the measure that could be used in the modelling protocol as a basis of picking up reliable fragments for modelling purposes.

Histograms to visualize the data distribution of four variables, including *rmsd*, sequence identity (SI), sequence similarity (SS) and secondary structure identity (SSI) were plotted for this purpose. Along with the overall qualitative analysis conducted on the validation dataset, analysis corresponding to each SCOP class within the dataset was also carried out to map the discrepancies that exist in the data distribution. The cumulative results of the data distribution based on the four criteria were in agreement with what was observed in the qualitative results of validation dataset (See paper in section above). An exception was seen in SCOP class all-beta, though the overall quality of fragment data was maintained. In this following sections we will shed some light upon the results belonging to each SCOP class and the exceptions observed in the quality trend of the fragment generated for each of these cases.

SCOP class all α

All the proteins from the validation dataset belonging to this SCOP class have been tabulated in Table 2.3. This table also illustrates the accuracy and score of PB predicted sequences along with the coverage attained for each target protein. Six out of seven cases showed acceptable prediction score of >1, except for one protein in the list, i.e., 1PO5. This is also the only protein in the entire validation dataset that showed no approximation of the prediction score or accuracy from the PB prediction tool. It has to be noted that despite the lack of prediction accuracy and score, the predicted PB sequence was able to recover fragments covering 97.6 % of the protein sequence.

Table 2.3: SCOP Class - all α

PDBId	Length	Accuracy	Score	Coverage
1AIL	170	63.6	1.5	98.6
1RRO	108	73.1	1.99	99.1
1U61	127	65.2	1.43	96.9
1SL8	181	77.5	2.23	97.2
1QUU	248	63.8	1.51	96.4
1T5J	301	65.8	1.61	91.7
1PO5	465	Too low	Too low	97.6

The table demonstrates the PB prediction accuracy and score calculated for each protein belonging to all-alpha SCOP class. It also jots the percentage coverage attained by the fragment generation pipeline for each target query.

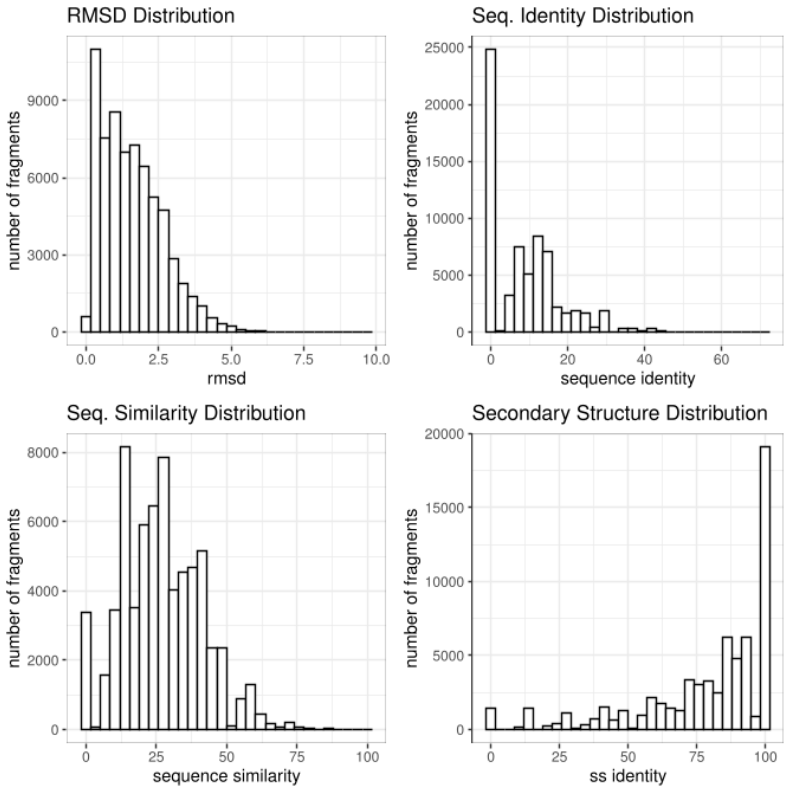


Figure 2.1: Histograms for SCOP Class all α

The following figure visualises the fragment data distribution in terms of rmsd (top left), sequence identity (top right), sequence similarity (bottom left) and secondary structure identity (bottom right) histograms for the structural class containing all-α proteins.

The results obtained for the qualitative analysis of fragments generated by the pipeline were most consistent for this SCOP class. As seen in the Figure 2.1, most of the fragment data retrieved for this class had low *rmsd*. When considering the SI, the distribution showed that most of the fragment hits did not share any sequence identity with the target protein sequence. SS results were also in continuation with the cumulative results, in which the data was distributed within the similarity range of 0 to 50 percent. Though the peaks were observed at $<$30% for the fragment data.

SSI distribution histogram was quite straightforward in depicting that most fragment hits had identical shared secondary structure distribution to the query sequences. In general the fragments retrieved for this structural class were of good quality and agreed well with the hypothesis that secondary structure comparison would give a better idea in to the relationship between the fragments and query protein.

SCOP class all β

This SCOP class was represented by 12 test proteins with the length distribution lying in the same range as the rest of the targets. As seen in Table 2.4, the accuracy and prediction scores were reliable across the target protein list with all proteins attaining high query coverage. Regardless, this SCOP class showed the most variation in the quality analysis conducted based on the four criteria.

Altogether, higher *rmsd* was recorded for the fragments extracted for proteins

Table 2.4: SCOP Class - all β.

PDBId	Length	Accuracy	Score	Coverage
1MHN	59	69.2	1.79	94.9
1TEN	90	72.1	1.94	95.5
2G1L	103	70.5	1.86	90.3
1IFR	113	67.1	1.68	93.8
1BFG	126	83.3	2.53	99.2
2FR2	161	70.4	1.85	97.5
1EE6	197	79.5	2.33	92.4
1UAI	223	71.1	1.9	87.9
2C9A	259	68.9	1.77	88.8
1O4Y	270	66.1	1.63	95.9
1HG8	197	81.9	2.46	88.4
1NKG	508	82.7	2.5	94.5

The table shows the length, PB prediction accuracy and score as well as the coverage attained for each protein in this subset of the validation dataset.

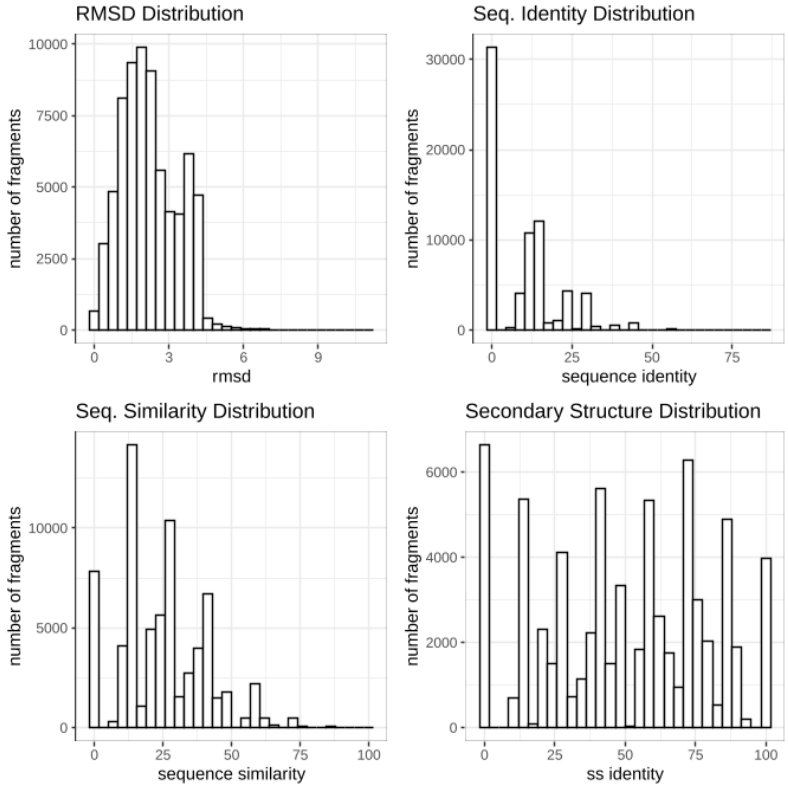


Figure 2.2: Histograms - SCOP class - all-β

The following figure visualises the fragment data distribution in terms of rmsd (top left), sequence identity (top right), sequence similarity (bottom left) and secondary structure identity (bottom right) histograms for the structural class containing all-β proteins.

belonging to this group, indicating that the fragments recuperated for this SCOP

class are further apart from that of the original structure. Though, significant amount of the data (64.43%) still reached a precision of 2.5 Å and could be used for modelling of proteins (documented in Table 2.2). In the case of all-beta (Table 2.2 and Figure 2.2), the *rmsd* values maybe higher despite high PB prediction accuracy - since the (phi,psi) range for beta-conformation is quite broad when seen on the map. Besides, beta-strands acquire structural variations like twists which could add to the discrepancies in its detection by the pipeline.

Despite the statistics seen above, sequence identity distribution was still on point by the data peaking at 0% identity. Although, sequence similarity distribution was a bit more erratic then what was observed in other cases. A major drift in observation were noted for SSI distributions. As seen in the "Secondary Structure Distribution" histogram (Figure 2.2 (bottom left), there is no consensus in the identity distribution, with peaks at multiple points. The highest peak observed in this graph accounts to 0% SSI, stipulating that a percentage of fragments did not share any backbone similarity with the query sequence. These results also contradict with the inherent nature of PBs, which are the sequence representation of any protein structure. These variations could be mapped back to few conditions of data generation. First being, the inefficiency of the PB prediction pipeline to correctly ascertain beta topology from the protein sequence. But, the results from Table 2.4 contradict with this assumption, as all the test protein sequences have optimal accuracy and score for PBP. Another reason could be the difficulty of SS prediction strategy to assign SSEs to the protein sequence. Though, this hypothesis can also be rejected since, PB assignment results also showed similar distribution of SSIs. In this case the DSSP assigned secondary structure distributions of query sequence were used to carry out the SSI calculations.

SCOP class $\alpha + \beta$

As shown in Table 2.5, all the test proteins belonging to this SCOP class showed high PB prediction accuracy and score except for one case (1VJW). This was lowest accuracy percentage recorded for the entire validation dataset of 43 proteins, though the score is still above the defined cut-off value. The overall query coverage attained by the identical fragment hits was also high. Two cases of this class even achieved a 100% percentage coverage across the length of the query sequence.

The other quality criteria tested for this class also aligned well with the overall quality assessment done for the fragment generation pipeline. The RMSD distribution and SSI distribution in this class is in line with that observed for all alpha structural class. Thereby, the proteins in this subset were in agreement with the general qualitative analysis trend established by the research work.

SCOP class α / β

Similar to SCOP class all-alpha and a+b, this class also aligned well with the consensus results for the fragment generation pipeline. All the predicted sequence showed reliable accuracy and scores. The fragment generation pipeline was able to achieve 100% coverage for two proteins (1E6K and 1QCY) from the 12 representing this SCOP class. The results are tabulated in Table 2.6.

Similarly, the histograms constructed for each of the four qualitative measures also aligned well with the overall conclusion the fragment generation pipeline (Figure

Table 2.5: SCOP class $\alpha + \beta$

PDBId	Length	Accuracy	Score	Coverage
1VJW	59	59.4	1.28	96.6
1MWP	96	89.7	2.86	99
1GNU	117	80.6	2.39	97.4
1R9H	118	85.8	2.66	95.8
206L	162	81.9	2.45	100
2FS3	280	82.1	2.47	96.8
1DZF	211	85	2.62	100
1DXJ	242	74.2	2.1	93
1MAT	263	83.9	2.56	99.2
1JKS	280	80.8	2.36	98.6
1MC4	369	77.3	2.22	97.8
2FKF	455	80.3	2.37	95.4

The table shows the length, PB prediction accuracy and score as well as the coverage attained for each protein in this subset of the validation dataset.

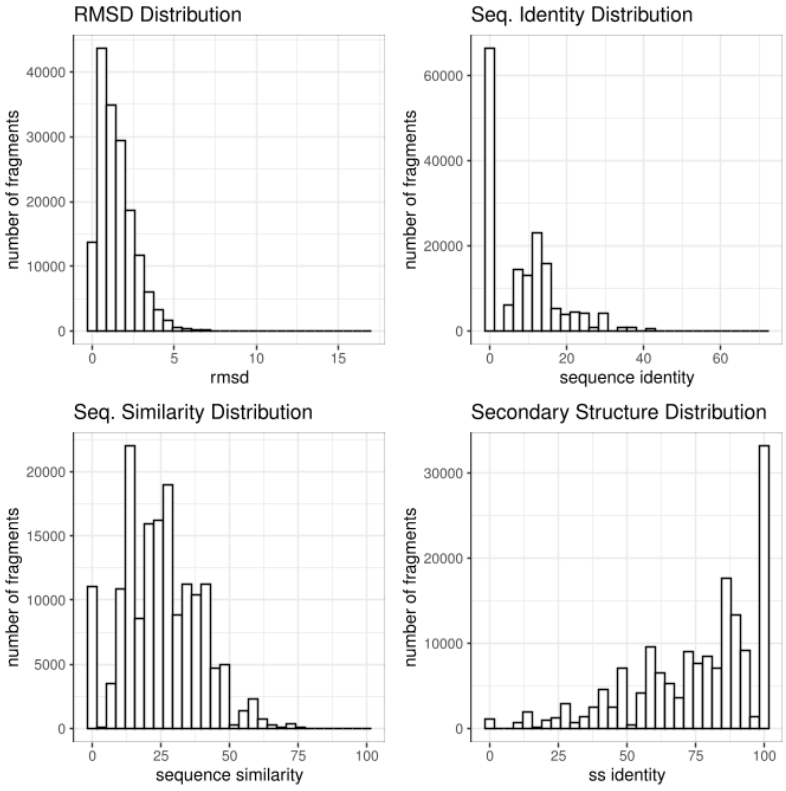


Figure 2.3: Histograms - SCOP class $\alpha + \beta$

The following figure visualises the fragment data distribution in terms of rmsd (top left), sequence identity (top right), sequence similarity (bottom left) and secondary structure identity (bottom right) histograms for the structural class containing alpha and beta proteins.

Table 2.6: SCOP class α / β

PDBId	Length	Accuracy	Score	Coverage
1H75	76	74.3	2.06	90.8
1IU9	111	68.8	1.77	95.5
1E6K	130	77.7	2.24	100
1P90	123	64.8	1.56	95.1
1FTG	168	79.8	2.34	99.4
1QCY	193	75.9	2.14	100
2A14	257	63.4	1.48	97.3
1IZZ	276	77.6	2.23	94.9
1QUE	303	82.8	2.51	97.7
1KRM	349	81.2	2.42	99.4
3BSG	404	77.2	2.21	95.3
1PGN	473	72.7	1.98	94.7

The table shows the length, PB prediction accuracy and score as well as the coverage attained for each protein in this subset of the validation dataset.

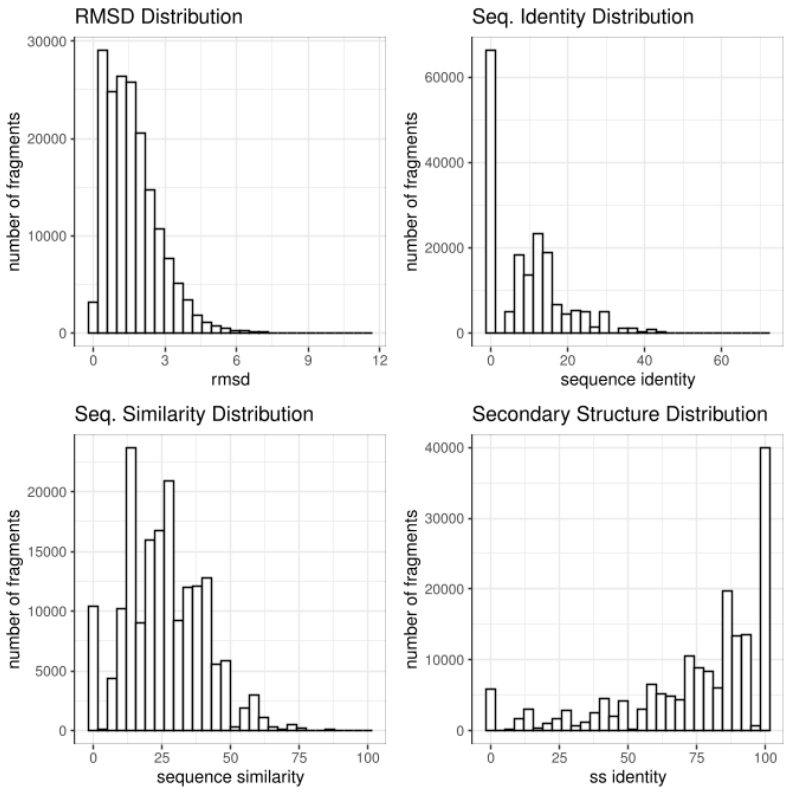


Figure 2.4: Histograms - SCOP class - a/b

The following figure visualises the fragment data distribution in terms of rmsd (top left), sequence identity (top right), sequence similarity (bottom left) and secondary structure identity (bottom right) histograms for the structural class containing alpha or beta proteins.

2.4). Further approving on the hypothesis that the fragments generated through the current PB based pipeline will have higher structural correspondence with the query sequence then the sequence itself. The histogram show that the maximum number of fragment hits share identical secondary structure with the query sequence in question, despite losing accuracy due to secondary structure prediction done on the query sequence.

2.3.3 ROC Curve Analysis

A sensitivity-specificity analysis was also performed in this study to determine if there is any co-relation between the *rmsd* and any of the three remaining factors (SI, SS, SSI) studied during this work.

The plots constructed for all the fragment hits lying under *rmsd* of 2.0 Å for the entire validation dataset have been shown in the paper above. Along with this, individual ROC plots belonging to each SCOP class were also analysed. All of these plots can been seen in Figure 2.5, in the order from top right to bottom left indicating ROC plots for SCOP class all-α, all-β, a+b and a/b respectively.

As seen in the Figure 2.5, sequence identity (green line in the plot) and sequence similarity (red line in the plot) lie at $sim0.50$ mark of the AUC. Thereby, ascertaining that these two factors are independent of the *rmsd* distribution calculated for the fragment data. In other words, these variables have no co-relation with the quality of fragment data produced by the pipeline and the closeness of the fragment hit to query protein.

As per our presumptions, SSI calculations do seem to be higher for the fragments lying under the defined rmsd cut-off. This was a continued trend even true for the SCOP class - all β, which otherwise did not show consistent results in the previous analysis. Instead, the drift from the norm was observed in SCOP class, a+b, where the co-relation dropped to around 0.62 between the calculated SSIs and the rmsd of the fragments. However, it is higher in comparison to the protein primary sequence counterparts.

2.4 Uniqueness of the pipeline

The PB based fragment generation pipeline built during this work has been standardized to generate fragments for *ab initio* protein structure prediction protocols. The idea being that predicted PB sequence carry enough structural information to derive similar structural chunks from a structure representing database.

The pipeline follows a well-known sequence alignment format as a way to extract fragment hits. The only difference being that the sequence alignment is carried out between PB sequence representations for the query protein and the assigned PB sequences for the structural database. The highlight of the pipeline can also be reflected by simplicity of its fragment generation protocol which makes use of local pairwise alignment to gather fragment hits.

Traditionally, all the fragment generation pipeline use sequence comparison as a way to determine viable fragments. Instead in our pipeline, we have tried to follow the fragment based ideology but the generated fragments do not rely on the sequence relationships as much as they do on the locally shared structural features which could be extracted by using PB sequences.

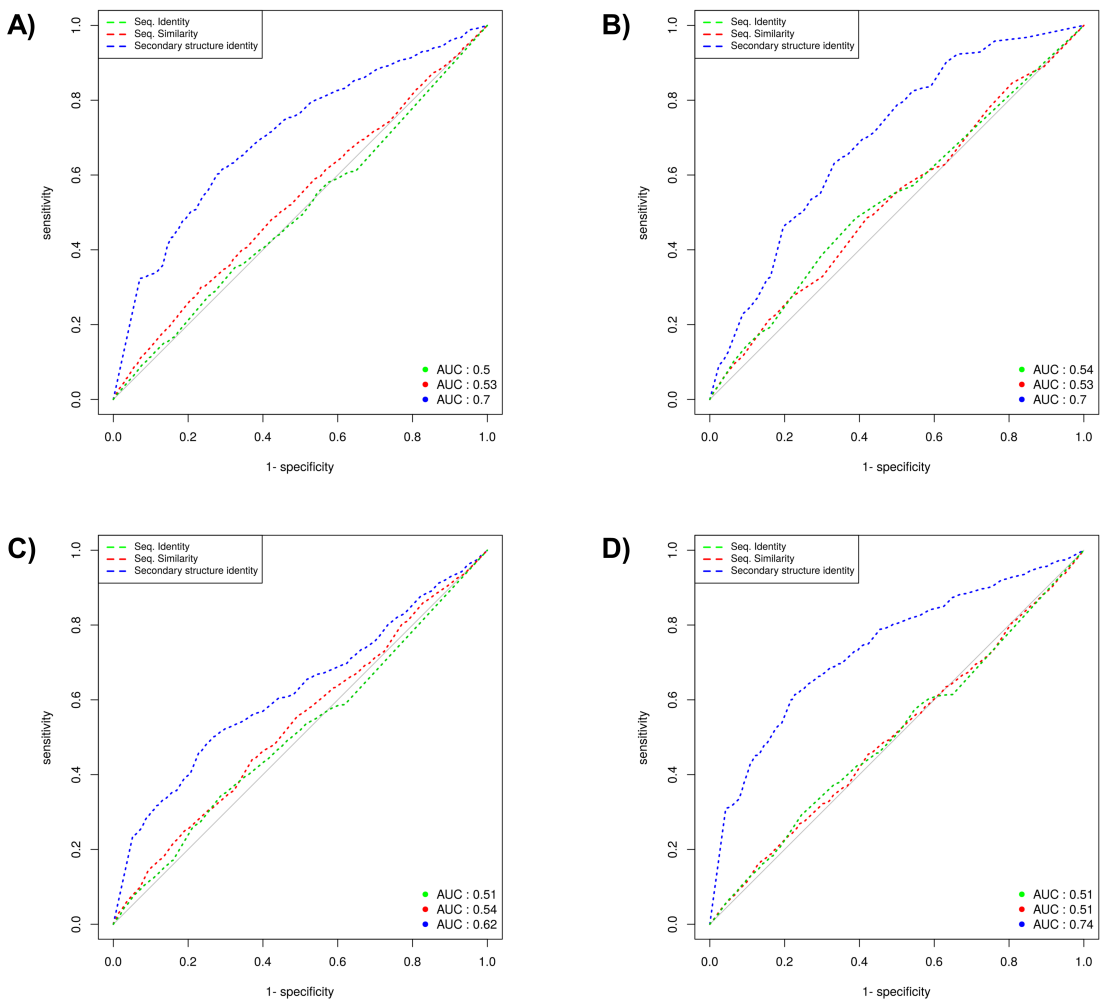


Figure 2.5: ROC curve

This figure depicts the ROC curve plotted for the four SCOP class (all-α, all-β, α+β and α/β from top left to bottom right respectively) against three criteria of sequence identity, sequence similarity and secondary structure identity shared between the fragments lying under the *rmsd* cut-off of 2 Å. Sequence identity curve is marked by colour green, similarly, sequence similarity by colour red and secondary structure identity by colour blue. The calculated area under the curve (AUC) have been marked in the far right corner of the individual graphs.

The pipeline was validated on a dataset of 43 proteins. Since, this dataset is taken from one of the fragment generation study, it helped in easy comparison of the efficiency of the pipeline between the two. We showed that on an average our pipeline achieved a precision of 64% at an rmsd cut-off of 2 Å, which is on the higher side when compared to the other algorithms available for the same purposes. Similarly, an average coverage of 93.9% was attained by the pipeline, reassuring that structural features can be extracted for almost the entire length of the protein sequence.

We also wanted to test our hypothesis that the fragment data generated by the current pipeline should be closely reflective of the structural features of the protein sequence. This was reinstated by SSIs calculations and ROC plots constructions for

fragments lying under 2 Å *rmsd* value. A point to note in this protocol, in continuation with what is noticed in other fragment generation protocols as well, is the limitation of reaching high precision for proteins with higher β-strand representations. Few possibilities owing to this situation have been discussed in the previous sections, though a closer re-evaluation of the outlook needs to be considered to ascertain the reason for this trend.

One more benefit that can be pointed out in regards to the pipeline is the loop region enrichment. Most of the template-free protocols revolve around construction of the regular secondary structures (α-helix and β-strand) regions before optimising the irregular seindary structures (loops) of the models. But, in our case enough fragments are extracted to correctly define the loop regions in the models along with the basic secondary structure constructions. This is possible due to the general definition of PBS, where 4 out of 16 blocks are used to describe the loops and coils regions of the proteins, thereby expanding the search dimension in this usually restricted area.

In general, PBs have been effectively used in optimizing the fragment generation pipeline that could either be used in gathering distant structural homologues or in construction of fragment library for a given protein target.

The name that can be named is
not the eternal name.

- Lao Tzu

3

Modeller for assembling fragments: preliminary experiments

Once the fragment generation pipeline was developed and settled, we needed to find a way to use these fragments into modelling of protein structures. Modeller, a well-known comparative modelling tool was considered for our modelling purposes. A set of experiments were designed inline with the premises of fragment-based modelling protocols and used into the protocols of Modeller to built native like models.

These experiments were tested and analysed for the three cases of template derivation to aid in standardization of the protocol, starting with usage of target protein sequence as the fragmented template. The second case had the same experimental design with the exception of fragments coming from a close sequence homologue and lastly, the third set of experiments were performed using fragments coming from multiple homologous templates. The experimental design and the results from each of the cases have been discussed in detail in the sections below.

3.1 Fragment-based modelling

Since the introduction of this concept by Bowie and Eisenberg in 1994, fragment-based modelling has been a staple in the field of *ab initio* protein structure prediction. Exploiting sequence-structure relationship locally, helps in reducing the conformational search space barriers which are normally huge in cases of PSP using entire sequence stretch and more often than not are the limiting factor of these algorithms.

A successful implementation of fragment assembly into template-free modelling has led to a rise in the precise structure determination of small proteins. The standard (and ideal) example to describe the effectiveness of fragment-based approaches is the algorithm Rosetta which fragments of length 3 and 9 as the local descriptors of the target protein sequence. In the entirety of CASP competition, Rosetta has had several shining moments. These include approximation of a 70 residue long protein structure to 1.6 Å during CASP6 competition. Similarly, Rosetta was able to built a model for 112 residue long protein at rmsd of 1.8 Å in CASP7 (Zhang

2008). Quite recently in CASP11, it correctly predicted a model for 256 residue long protein. The success emulated by this approach further pushed the development of approaches following a similar design.

Another successful approach developed under the umbrella of fragment-assembly is TASSER (Zhang et al. 2004). It is based on collection and assembly of fragments of various length using threading based approaches. Several versions of this approach have been developed in years and implemented into structure prediction through fragment assembly.

The main advantage behind using fragment-based approach to predict protein structures is the substantial reduction in the available conformational search space. The use of fragments akin to specific position in the query protein having strong local structural feature limits the conformational variation widely. In cases where the local features do not match up, fragments can easily be switched by more favourable ones. Thereby, vastly improving the structure prediction accuracy via this scheme.

3.2 Modeller for protein structure prediction

Modeller (Šali et al. 1993) is a well established tool in the field of homology and/or comparative modelling of protein structures. It is based on modelling the three-dimensional structure of a protein sequence by satisfaction of its spatial restraints.

In principle the user have to provide a sequence alignment file containing the query and the template protein sequence. Usually, the template chosen for modelling is a sequence homologue of the target protein, in which case the procedure will come under homology modelling. Although, modelling can also be performed using distant templates sharing high sequence similarity with the target protein. In such cases the protocol can be termed as comparative modelling.

Once aligned, Modeller uses a simple script to generate model(s) based on the user requirements. The model is built based on the extraction and usage (superimposition) of spatial restraints from the template structure on to the spatial arrangement of the sequence to be modelled. The output model(s) is the result of the satisfaction of these spatial restraints. This procedure of extraction and implementation of the restraints from a known structure is similar to what is used in structure determination done using NMR-derived restraints. The ease in exploitation of these restraints comes from the assumption that the aligned target and template sequence will share similar topology. The restraints can be passed in the form of angles, distances, dihedral angles, pairs of dihedral angles, non-bonded atom-atom contacts etc. Furthermore, the final model(s) are internally optimized in the Cartesian space using conjugate gradient and MD with simulated annealing methods.

In the current study we are exploiting this inherent property of Modeller to exploit spatial restraints into producing *ab initio* protein models. Though, it might not be the first time that Modeller has been employed for similar purposes. A homology/*ab initio* hybrid approach is already available to the best of our knowledge that makes use of Modeller to model its fragments. The algorithm named as Bhageerath-H (Jayaram et al. 2014) has been built on the usage of knowledge derived from sequence and structures along with a separate *ab initio* folding program, Bhageerath (Jayaram et al. 2012) to built a hybrid approach to PSP. The pipeline starts with the secondary structure prediction of the target protein sequence, followed by the search of its sequence homologues accompanied with the searches for distant structural ho-

mologues of the target sequence. Mostly, the homology based modelling approaches donot include these distantly related structural homologues during normal sequence based database search. Bhageerth-H takes care of this loophole and improves the rate of structure prediction by addition of such structural homologues if available. At this point, the use of Modeller comes in, where fragments of various lengths are modelled using the template-target alignments acquired through the above followed protocols. Any missing link in between the modelled fragments are filled using the Bhageerath *ab initio* protocol (specifically at loop regions).

Another work that influenced our use of Modeller as the main modelling tool in the current pipeline has been implemented in the form of a web server called ORION (Ghouzam et al. 2016). Similar to Bhageerath-H, its algorithmic design is based on better description of the local protein structure by inclusion of distantly related structural homologues using the descriptors called Protein Blocks. As discussed in Chapter 1 of this thesis, PBs are the one dimensional sequence description of a tertiary protein structure. Therefore, by using a pairwise comparison of PB sequences, distant structurally similar templates can be easily identified. ORION uses a combination of sequence/structure profiles to narrow down viable templates from a structural databank to locate remote homologues and use them for modelling structures.

Our approach in the current study lies somewhere in between both the above stated approaches. Instead of looking for full length distant homologues for the target protein, it uses PBs as a way to generate structurally similar fragments that could be directly used in the modelling protocol. Similarly, instead of modelling fragments individually likewise in Bhageerath-H, the idea behind our PSP pipeline is to use the fragments generated by our PB-based fragment generation pipeline into building full length multiple template alignments for the target protein sequence and solely relying on Modeller to built close enough models for the target protein sequence.

3.3 Experiments performed using Modeller

The experiments performed using Modeller followed a fragment-based protein structure assembly protocol. Three level of experiments were designed to check the potential of this comparative modelling tool into TFM field.

It followed a trend of increasing the number of templates used to derive information for modelling of test protein. The introductory model building experiment used fragments derived from the query sequence itself to reconstruct the protein structure. The second set of experiments focused on using fragments derived from a close sequence homologue. In line with it while increasing the difficulty, the third level of experiments involved use of fragments coming from multiple sequentially homologous templates to construct the protein models.

Each of this level of experiment was further divided into sub categories, where first part of experiments were performed using fragments that did not have any sequence overlap among themselves and were categorised under "non-overlapping fragment" category. On the other hand, the second set of experiments for each level were performed using the same templates but fragmented such that the adjacent fragments shared a region of sequence overlap amongst themselves. These sets of experiments were categorized under "overlapping fragments" category. The reason

behind using the fragments in these two ways was to confirm the effects of the continuity or discontinuity in transfer of information that could affect the overall quality of models generated by Modeller.

In the sections below each of these experiments have been explained at vast using Phospholipase A2 (3ELO) as the test protein. It is a monomeric protein that is 133 residues long and comes into all-alpha SCOP class and in mainly alpha class under CATH annotation. As the classification suggests, the structure is dominated by α-helical secondary structure with a minor region folding into β-sheets. The structure of this protein has been crystallized at 1.6 Å. This was a randomly chosen test subject for the experiments and served as an acceptable representative of a protein with complicated enough fold with a good depiction of secondary structure elements including loops.

All of the experiments stated below followed a general flow in terms of use of Modeller.

- The fragments were created manually for each template protein and their representative co-ordinate files were built using the tool ProFit.
- To make the pseudo-template files readable by Modeller, it was necessary to add 'TER' at the end of each fragmented PDB file. This addition made the co-ordinate files appear to be complete.
- The alignment file used by'Modeller to access the template information were also edited manually to position the fragments and gaps as per the experimental design.
- Basic Modeller multiple template modelling script was used to generated 5 models for each test case.

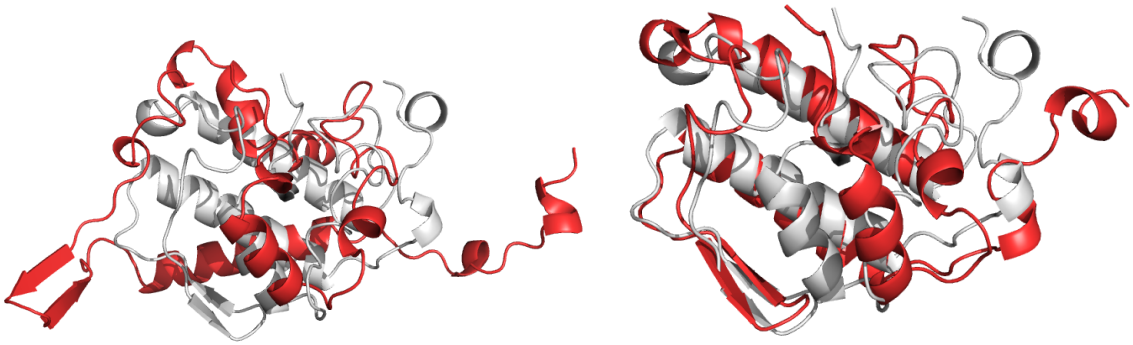
All of these experiments have been extensively discussed in the coming subsections.

3.3.1 Fragments from the same protein

To see the effectiveness of Modeller in building purely *ab initio* models, a number of experiments were performed to ascertain the strengths and limitations of the tool. The first set of experiments were designed to see if Modeller is able to reconstruct the model given the target query sequence itself is fragmented and provided in the form of fragments in the alignment file.

The primary experiment was designed such that the efficiency of Modeller can be documented in case where the query and template sequence are completely identical to each other. The only glitch or change in the protocol was that instead of providing the template sequence in the form of one single sequence alignment, it was delivered in the form of fragments, i.e., the template was shredded such that it acted like multiple fragments instead.

The sub-experiments under this case involved building two alignment files, one including fragments that shared no sequence overlap between two adjacent aligned sequences. The second sub-experiment had fragments with shared sequence space in the alignment file. In this case each fragment shared a sequence overlap of 7



(a) Non-overlapping fragments based model (b) Overlapping fragments based model

Figure 3.1: The figure shows the models constructed based on the two study cases of fragment utilization tested in this study. Here, both the models are predicted using the query sequence itself as template provided in the form of overlapping and non-overlapping chunks. The original structure and the model is shown in colour grey and red respectively.

residues. This experiment was designed to test the effects of the lack of flow of information in the alignment file on the modelling ability (and quality) of Modeller.

The result for this experiment is shown in Figure 3.1. Part (a) of this figure shows the best model (9.4 Å) built by Modeller using non-overlapping fragments. It can be seen that despite fragments being completely identical to query sequence, it was challenging to obtain structurally similar model using this approach. It could reflect at the nature of Modeller which is based on Cartesian system, therefore, the reassignment of the geometry even at this basic level was not manageable by tool.

For the second case, same scenario was tested but with fragments sharing some degree of sequence overlap between two adjacent fragments. In this case on generating 5 models, the best model had an rmsd of 5.142 (Figure 3.1 b). Although, the best model did not satisfy the quality score, the superimposition itself shows that the model has folded close to the geometry of the original structure.

3.3.2 Fragments from close sequence homologs

In the second set of experiments, the same ideology was followed but instead of constructing fragments from the same sequence, a close sequence homologue was used for model reconstruction.

The sequence homologue chosen in this context was 1HN4, which shared 81% sequence identity with the target protein while having a 96% of query coverage. The template protein sequence was divided into a total of 7 fragments for both the cases of non-overlapping and overlapping fragment used. The gaps that were observed according to the full length template alignment of the query sequence to the template were kept as it is in the experiments performed after fragmenting the template. Therefore, the first template fragment (1hn4_1) had 2 gaps added to it agreement with the full-length template alignment. All of these details have been documented in Table 3.1.

Similar modelling script was followed as used in the above experiments while replacing the template fragments (1hn4_1 to 1hn4_7) provided to Modeller for protein modelling. From the 5 models generated for non-overlapping fragment category, the best model deviated by 6.3 Å from the original structure of the test protein (3elo).

Table 3.1: Fragmented arrangement of the single template protein

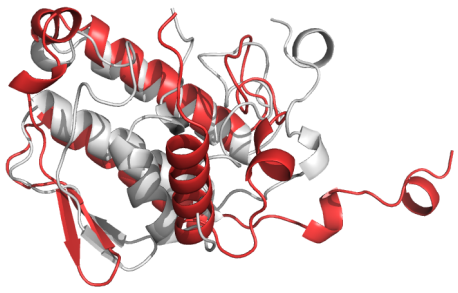
Single Template	Seq. Identity	Query coverage	Overlapping fragments		Non-overlapping fragments
			Fragment length	length of overlap	Fragment length
1hn4_1	81%	96%	22+2 gaps	7	17+2 gaps
1hn4_2	81%	96%	21	7	19
1hn4_3	81%	96%	21	7	19
1hn4_4	81%	96%	21	7	19
1hn4_5	81%	96%	21	7	19
1hn4_6	81%	96%	21	7	19
1hn4_7	81%	96%	21	7	19

The table describes the number of fragments into which the template sequence has been split along with the length of the fragment used in each study case (overlapping and non-overlapping fragments).

Despite the huge deviation, it can be noted in the Figure 3.2 a, that the overall geometry of the secondary structures have been well conserved in the model although the fold is distorted. This can again affirm the notion that Modeller is unable to guide the modelling in case the line of information travel is discontinuous, i.e., the fragmented templates do not share an overlap.



(a) Model built using non-overlapping fragments



(b) Model built using overlapping fragments

Figure 3.2: The following figure shows the models built for the two study cases (overlapping and non-overlapping fragments) in this study using fragments derived from a single homologous query protein sequence. The original structure of the test protein is shown in colour grey and the predicted model is shown in the colour red.

For the second experiment that was conducted under the same conditions, the non-overlapping templates were replaced with the ones sharing an overlap. Each adjacent pair of fragmented template shared a consensus 7 residues overlap among themselves. The number of fragments used were again 7 with the length of each fragments coming down to 21 residues, except for the first fragment that was 22 residues long. The best model constructed for this experiment had a recorded rmsd of 5.7 Å. This was an improvement over the previous model constructed under the same experiment but not essentially by a large scale. Though by looking at the model (Figure 3.2 b), it can be noted that the geometry of the fold is much more conserved then in the case of non-overlapping fragments.

Through these experiments, a clear improvement in the quality of the models can be seen when the same fragmented template sequence is used with and without residue overlap.

3.3.3 Fragments from multiple homologous proteins

This experiment was the closest to the actual utility of Modeller that was being standardized. In this case instead of relying on one template sequence, chunks of several templates were used to built model resembling the query protein structure.

Several close sequence homologues of the test protein were picked up from blast results and used for conducting the experiments. The PDB id of the templates along with the shared sequence identity have been documented in Table 3.2. Each chosen template covered a minimum of 90% of the query sequence space considering that the sequence identity varied between 77% to 81% for each template.

Table 3.2: Description of templates chosen for modelling the protein

Multiple Template			Overlapping fragments		Non-overlapping fragments
	Seq. Identity	Query coverage	Fragment length	length of overlap	Fragment length
1HN4	81%	96%	23+2 gaps	7	17+2 gaps
1FX9	81%	92%	25	7	19
2PHI	81%	92%	25	7	19
2BP2	77%	95%	25	7	19
1BVM	79%	90%	25	7	19
4BP2	77%	94%	25	7	19
1Y6O	80%	93%	25	7	19

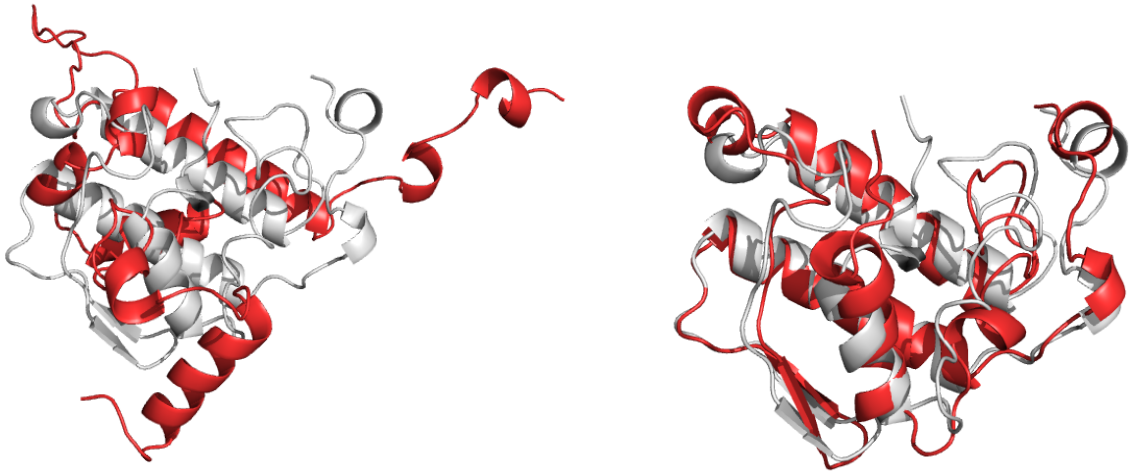
This table describes the PDB id, sequence identity and query coverage of the proteins that share sequence homology with the query protein (3ELO) used in this study. It also describes the length of the fragments constructed for carrying out modelling for the two test cases undertaken during this work.

Similar to the previous experiments, chunks of the template sequence were used to built the alignment file for Modeller and two sets of experiments were performed. The best models determined for each of the experiment had a deviation of 17.1 Å and 2.4 Å respectively for non-overlapping and overlapping template fragments. By far this was the largest gap in the deviation of models from the native structure that was confirmed between the two cases of protein structure modelling.

This suggested towards the possibility of decrease in the quality of models generated by Modeller with the increase in the sequence diversity of the templates used with no source for the continuation of transfer of information as seen in case of non-overlapping fragments. Our experimental designs further helped us to narrow down at least one of the essential drawback that we would encounter for cases where our PB generation pipeline is not able to produce high coverage fragment libraries. This lack of information will directly affect the quality of the models produced by the pipeline.

Figure 3.3 a, shows that the overall fold of the model constructed for this case was bad and hinted towards the increased distortion in model building that Modeller could experience with the increasing size of the protein. As the bigger proteins have more complex fold and that the fragments lacking overlap at any region will affect the globular fold of the protein.

On the other hand, the best model (rmsd 2.4 Å) constructed for the case where the information flow was maintained, i.e., using the overlapping fragment, the overall global fold of the protein was retained along with its secondary structure distributions. This was keeping in mind that no extra restraints or information was provided to Modeller to aid in model building.



(a) Model built from non-overlapping fragments

(b) Model built from overlapping fragments

Figure 3.3: This figure shows the best models built from templates coming from several homologous protein sequences for the two cases in study. The original structure of the protein (3elo) is represented by colour grey and the model by colour red. Same templates were used in the same order to construct these models with the only difference being in the alignment file generation in which one had template regions sharing no sequence overlap and in the second case the adjacent templates had a sequence overlap of 7 residues each.

3.4 Fragments from non-homologous proteins

The certainty attained from the previous experiment regarding the quality of the models that could be generated while having no influence from restraints and just relying on overlapping fragmented templates prompted us to use this strategy to try and built an *ab initio* based model to verify the route we were on.

Crambin (1CRN) was chosen as a model protein to conduct this experiment. It is a small protein of length 46 aa with the structure crystallised at a resolution of 1.5 Å. It belongs to the SCOP class of small proteins. The reason behind choosing this protein to validate the experiments conducted so far is that it is a small protein with distinctive secondary structures, therefore being an appropriate model to begin with. Plus, given its small size and less complicated fold, it would be easy to model the structure correctly.

To follow the premises of template-free modelling, a PB based fragment library was built using the PB sequence assigned to its structure. A description on this fragment generation pipeline has already been iterated in Chapter 2 of this thesis. A total of 50,553 fragments were generated for this query PB sequence with the identical hits boiling down to 15,452 fragments.

From this set of hits, 5 fragments were randomly chosen as pseudo-templates, while following the premises of the fragments sharing an overlap between two adjacent templates. The PB sequence description of the query protein and that of the fragments can be seen in Figure 3.4. It can be confirmed from this figure that the query and chosen templates had identical PB sequence for the position they were secluded for.

PDB co-ordinate files corresponding to the position of fragment hits chosen for

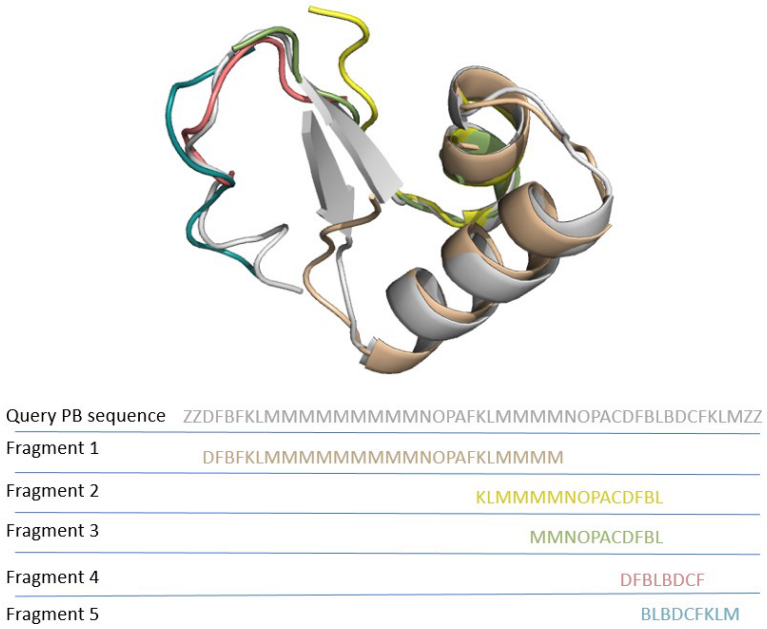


Figure 3.4: Cartoon representation of fragments superimposed over original structure

The following figure shows the fragments chosen to model the protein Crambin superimposed on its known structure. It also depicts the alignment of the PB sequences both for the query protein and the template in the 1D readable format.

this experiment were constructed using PDB_tools. These pseudo PDB files were then used to frame Modeller alignment file according to its format. The structure of the alignment file was built using the overlapping fragments in accordance with the requirements of Modeller to produce an *ab initio* based structure is shown in Figure 3.5. This alignment file specifies the PDB id and chain of the template used along with the starting position of the template coordinate and its length.

Each of these fragments were passed as individual templates in the modelling script to construct 100 models of the protein Crambin (given its sequence). The quality of models built by Modeller were judged on the basis of global distance score (GDT_TS), TM-score and *rmsd* of the models built for this protein.

GDT score is the "Global Distance Test" score calculated for a pair of proteins sharing high sequence coherence with different tertiary structures. It is most commonly used to compare predicted protein models to the experimentally determined protein structures. The scoring function was first established by Adam Zemla in 2003 (Zemla 2003). It is calculated in the form of a distance cut-off of the amino acid positions in comparison to the experimental structures. In CASP the average distance cut-off are at 1, 2, 4 and 8Å respectively.

TM-score or Template modelling is score is designed to to measure similarities between two proteins with different tertiary structures. It indicates a difference between two protein structures on a scale of [0,1], where 1 is the perfect match. A TM-score of >0.5 postulates to the structures having similar folds and this correlation quickly increases with the score. On the other hand, a TM-score of <0.2 indicates the unrelatedness of the protein structures being compared.

```
>P1;3WHR_A_fit
structureX:3WHR_A_fit:  48 :A:+28  :A:::-1.00:-1.00
MVATAHPLASEIGADVLKKGGNAIDAAV-----------------*

>P1;5WHX_B_fit
structureX:5WHX_B_fit:  43 :B:+15  :B:::-1.00:-1.00
---------------------YSELCLQMGIHFFRD---------*

>P1;4GBJ_A_fit
structureX:4GBJ_A_fit:  41 :A:+11  :A:::-1.00:-1.00
-------------------------LTKLGATVVEN---------*

>P1;4H2B_A_fit
structureX:4H2B_A_fit: 252 :A:+8   :A:::-1.00:-1.00
--------------------------------NPPSKEVP-----*

>P1;4RJX_A_fit
structureX:4RJX_A_fit: 129 :A:+11  :A:::-1.00:-1.00
----------------------------------TAPERNAAYEL*

>P1;1crn
sequence:1crn:0    : :0    : :::-1.00:-1.00
TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN*
```

Figure 3.5: Modeller alignment format

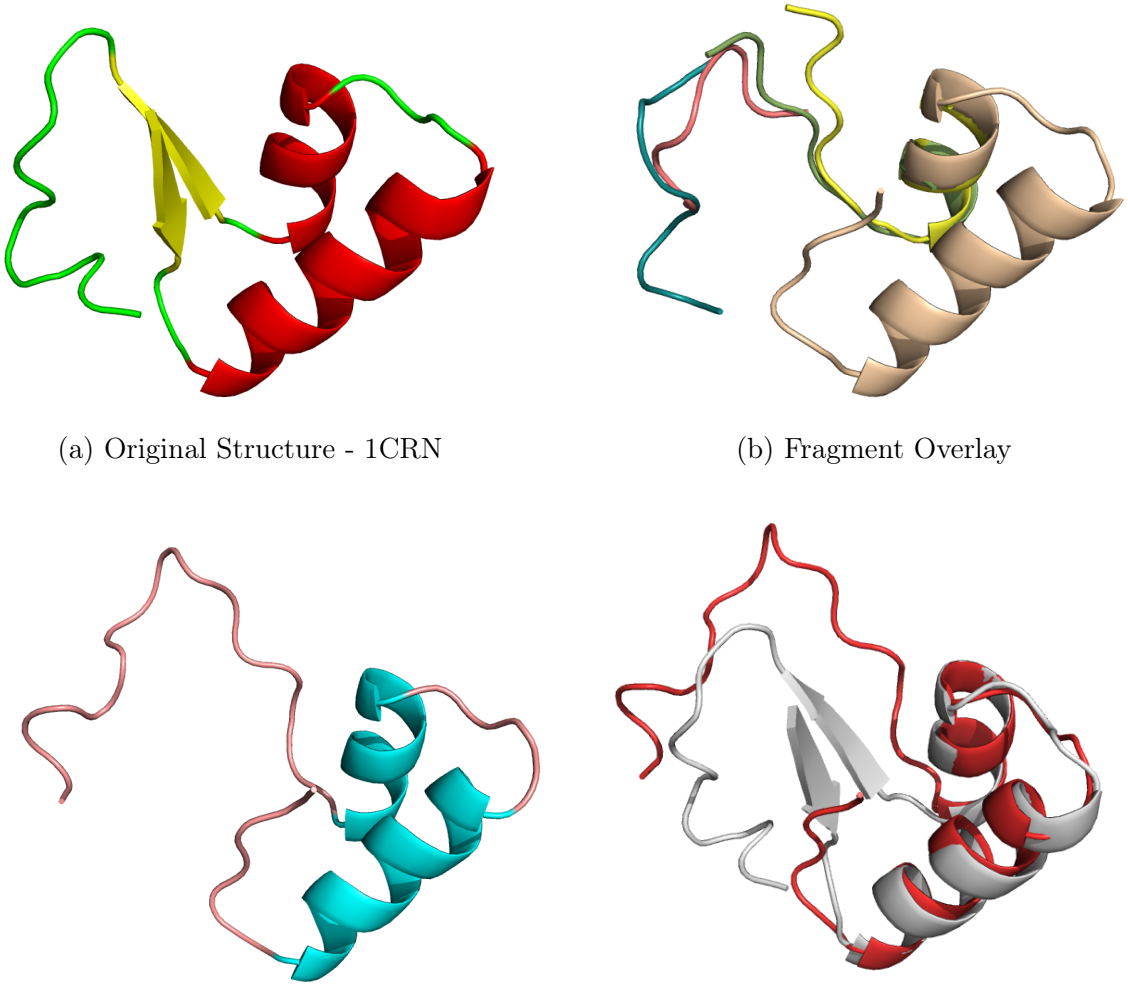
This figure is the depiction of the alignment file built that is mimicking Modeller multi-template alignment file design. Each of the chosen fragment is arranged according to the position of the hit determined by the fragment generation pipeline. The alignment file also contains the details of the starting position and the length of the fragment used along with the PDB id of the protein it is extracted from.

The best model had a GDT_TS score of 0.76, along with the TM-score of 0.59 and it deviated by 3.3 Å from its native structure. Figure 3.6 c and d show the modelled protein and a superimposition of the best model over the original structure. These results are promising in affirming that the quality of models generated by the pipeline being standardized here could be good while depicting the potential of the current work in progress.

3.5 Inference

Modeller, despite being developed for comparative modelling purposes can be used for *ab initio* protein structure prediction under certain conditions that rely on cheap imitation of its multiple template based protein structure prediction protocol. The fragments generated by the pipeline developed and discussed in Chapter 2, were used as pseudo-templates to guide the folding of the unknown protein structure. It was also confirmed that the quality of model generated is affected by the way the fragments are passed into the modelling script. It was observed through a series of experiments that the cases where the fragments had a flow of information amongst themselves by having a sequence overlap to some degree achieved better modelling results than the cases where fragments had no overlapping regions.

Through the last experiment performed using fragments extracted from a non-homologous structural database, it was confirmed that Modeller is able to construct models even if the proteins share low sequence identity with the query sequence. This could be due to the structural similarity shared between the test protein sequence



(a) Original Structure - 1CRN

(b) Fragment Overlay

(c) 1CRN Model

(d) Original structure vs Model

Figure 3.6: The figure represents the progression of model building using a PB based fragment generation pipeline into Modeller multiple-template modelling protocol. Figure (a) shows the original structure of protein which is being modelled (1CRN), (b) Is the overlay of fragments chosen for constructing the structure, (c) Is the model generated after running Modeller and (d) shows the superimposition of the best model (red) generated onto the original structure (grey) of Crambin.

and the fragment library generated for it. It was also identified that a model can be constructed with as low as one fragment depiction per position. Though in the later experiments (Chapter 4) the number of fragments used per position have been increased to consider the possibility of capturing alternative conformations.

3.5.1 Strengths and weaknesses

Through the sets of experiments performed above few definitive points can be put forward regarding the trends that were seen throughout this chapter. We can infer that Modeller as a protein structure prediction tool is very easy to understand and manipulate as long as the rules of its modelling protocols are followed. Given its ease of comprehension, the tool in itself is easy to use with minimum efforts needed to produce the files required for structure prediction. Despite its claim of loss of

modelling accuracy with increase in gaps, pseudo-templates (fragments) can readily be used for generation of protein models (at least for small proteins). In our case, the overlap of sequence shared between two adjacent fragments serve a purpose of long-range contact to aid in global folding of the protein. To further improve upon the quality of models, external restraints can be added to constrict the globular fold of the protein. Furthermore, its internal optimisation packages can also be used to get an energetically minimized structure.

On the contrary, while using Modeller for template-free modelling, it is almost compulsory to use fragments which have shared overlaps to generate decent quality models. In the positions where the fragments were not overlapped, the model showed inability of Modeller to successfully determine the correct fold. Given this situation, there is no assurance of generating fragment libraries that always have fragments hits which might share a sequence overlap. This claim can be confirmed by the cases seen in validation dataset used in Chapter 2 for fragment quality analysis. Despite giving high coverage the percentage of queries which reached 100% coverage were 4 out of 43 test protein sequence, which in this case translates into the introduction of gaps in the alignment file.

The prediction of protein structures through template-free protocols using Modeller can be limited by the size of the protein. As bigger proteins tend to have more complicated folds and would require a larger number of fragments to built a model. So far there is no defined limitation of number of templates that can be used to built a protein model. Inherently, the sequence similarity shared between the query and the template protein also affects the quality of models produced by the tool. The effects of the loss of sequence similarity between the two are not yet evident on the current pipeline and requires more intensive research in this area.

This is not a homecoming, since this home has never before existed. It is the creation of home.

- Alice Miller,
The Drama of the Gifted Child

4

Fragment based prediction of protein structure

Having established and analysed results from the last two Chapters of the thesis (Chapter 2 and 3), this chapter serves as a basis of bringing these two protocols together to work towards the final goal of the project that was to introduce a new methodology of *de novo* protein structure prediction that uses of PBs to do so.

In regards to this, the current chapter will have an in-depth discussion on the other elements that were added to the prediction pipeline as well as provide a refined picture of the pipeline standardized so far. This includes a detailed description of the two external restraints added to the modelling protocol that were carefully chosen to help promote proper global folding of the protein structure and the way these restraints were added to the modelling scripts.

To increase the quality of the predicted models, a refinement protocol was also added as the last step to the pipeline. The changes observed in the post refined models in comparison to the native structure of the protein as well as the pre-refinement model were also documented and will be discussed in the coming sections.

4.1 External restraints

The emergence of knowledge-based protein structure prediction strategies into Free modelling sphere also bought along with it the concept of utilising structural restraints into constraining the spatial dimensions of the proteins. The major lagging for the TFM approaches have been attributed to the lack of effectively guided dive into space search, along with the ways to look for alternative conformations and designing of efficient energy functions.

It has been demonstrated that use of external restraints do help in limiting the dimensionality of the search space (Kolinski et al. 1998; Maurice 2014; Santos et al. 2017). Many algorithms have been developed that make use of secondary structural restraints, solvent accessibility and distance restraints etc. for doing the same.

It has also been documented that the use of sparse distance restraints can be

effective in guiding the protein fold (Aszódi et al. 1995; Kolinski et al. 1998; Bowers et al. 2001). These restraints are usually derived from the early stages of Nuclear Magnetic Resonance (NMR) experiments and are termed as Nuclear Overhauser Effect (NOE) restraints. Other restraints readily used in PSP are distance restraints applicable in the form of contact predictions or the once derived from mass spectroscopy (MS) cross-linking (XL) data (Fischer et al. 2016).

From the beginning of CASP biennial competition the importance of co-evolutionary based inter-residue restraints have also been emphasised on. Focused research has been conducted to improve upon the accuracy of residue-residue contact predictions. After analysing the possibilities of external restraints that could be used in the modelling protocol, we settled on exploitation of two vastly employed restraints in this field, i.e, predicted secondary structural restraints and contact predictions. Each of these restraints served as a form of intra-residue and inter-residue distance constraints respectively. All these aspects are further detailed in the following subsections and in the accompanying paper.

4.1.1 Secondary structure prediction

Hydrogen bond formation in the neighbouring residues of a protein (mainly between the amino and carboxyl groups) results in the formation of some repetitive patterns in the local backbone of the protein, termed as secondary structures (SS). These basically act like small range intra-chain contacts that form the basis of protein folding and their formation is essential to define local structure of the protein backbone.

These backbone definitions have readily been employed to solve the tertiary structure of protein via *de novo* protein structure predictions as well. In most of the TFM algorithms predicted secondary structures are used as initial restraints (Simons et al. 1997, 1999; Kim et al. 2004; Rohl et al. 2004; Koga et al. 2012) . Studies have also implicated that the pre-restriction of secondary structure type does effect the overall fold of the predicted tertiary structure (Simons et al. 1997, 1999).

Algorithms have been designed that focus on usage of secondary structure descriptions to reduce the conformational search space (Kihara et al. 2001). Similarly, BCL::FOLD, an *ab initio* protein structure prediction algorithm was developed under the same idea of reduction of sampling space by usage of predicted secondary structure assembly done by Monte Carlo algorithm (Karaka et al. 2012). On the same ground, Chunk-Tasser, a fragment based approach for PSP has individual chunks (fragments) defined by three successive secondary structures along with their adjoining loops (Zhou et al. 2007).

With plenty of research indicating towards the nature and implications of the secondary structure elements (SSEs) used as restraints in structure building algorithms, it was an easy choice to make. In our pipeline, we used the PSIPRED (David T Jones 1999; D. W. A. Buchan et al. 2019) web server to recover predicted SSs. The predicted intra-residue contacts were passed as position specific restraints into the modelling script to favour a SSE type for that position in the query. In the fragment building pipeline (Chapter 2) these predictions were also used to facilitate fragment selection by further narrowing down the fragment traffic based on its shared secondary structure identity.

4.1.2 Contact predictions

A pair of spatially close residues in a protein structure are know as "contacts" or "residue-residue contacts" (Adhikari et al. 2016). They are capable of providing a simplified representation of a protein structure by capturing its fold. A usual defined cut-off between the residue pair is within 8Å shared mostly between carbon-alpha or carbon-beta atoms of the amino acids. Along with this, a sequential separation is also defined within the protein sequence to ensure that the spatially close residues do not lie in the sequence proximity of each other. Studies have indicated that use of Cβ atom distance is more effective in reconstructing protein structures than cα atoms (Duarte et al. 2010).

The potential of evolutionary data to recover residue-residue contacts was recognized and appreciated from the very beginning of CASP competition (Goebel et al. 1994). It also defined a contact as a pair of residues whose Cβ atoms are at less than 8Å distance apart and are at least 5 residues away in the sequence. It was hypothesized that sufficient quality and quantity of such predictions could directly be used to infer the tertiary fold of a protein (Izarzugaza et al. 2007). Emphasis was laid on the benefits of long range contacts in *ab initio* protein folding as they aided in constraining the global fold of the protein. CASP7 assessment posed a question on the amount of improvement needed in contact prediction algorithms for it to be useful in *de novo* modelling of proteins (Izarzugaza et al. 2007). Eventually, it was realised that the constant failing of contact prediction algorithms was the methodological flaw (Burger et al. 2010), though until CASP11 no major improvement was documented in this field. Analysis carried out on prediction protocol that participated in CASP10 showed that the prediction accuracy was similar to what were observed in CASP9 (Monastyrskyy et al. 2014). Most encouraging result was observed in CASP11, with an accurate model generation of 256 residue long protein (Monastyrskyy et al. 2016). This was attributed to progress made in prediction of residue-residue contacts, specifically focusing on the accuracy of long-range contacts predictions. This marked a point of success in the run time of CASP that was beyond what was previously noticed, both in terms of protein size and prediction accuracy.

CASP12 assessment dwelled upon the progressive use of alignment-based contact predictions methods to model hard targets (Schaarschmidt et al. 2017). They also pointed out the use of these algorithm into solving structures of complexes and to look for alternate conformation which has already been pointed out as a major drawback of fragment-based modelling protocols. CASP12 pondered upon the possibility of contact predictions becoming main stream in free modelling protein structure prediction strategies.

As per the discussion above we can infer that for the longest of time, the importance of residue-residue contacts have been emphasised on in the field of PSP. It has also been quoted that just the information of correctly predicted contacts should be enough to guide proper folding of proteins. Keeping this in mind, our second choice of restraints for the study in hand was residue-residue contacts data.

DeepMetaPSICOV (DMP)

In the current study we have made use of a deep learning based approach to residue-residue contact prediction called DeepMetaPSICOV (DMP) (Kandathil et al. 2019),

designed based on fully convolutional residual network. It is a combination of the two previously developed methods from the same group, namely MetaPSICOV (David T. Jones et al. 2015; D. W. Buchan et al. 2018) and DeepCov (David T. Jones et al. 2018). It has internal features similar to that used in MetaPSICOV including sequence profile, predicted secondary structure, solvent accessibility etc. The predicted scores for each residue pair is based on that used in DeepCOV, where an average of predicted values for the residue pairs (i,j) and (j,i) is considered. The final prediction score is an average of 5 independently run DMP models.

The contact prediction techniques used in this tool are similar to what are used in image analysis. These include loop sampling, feature interpolation and flipped input feature tensors and contact maps. Appropriate sequence alignment profile generation has been ensured by using two different strategies encompassing HHblits and Psi-BLAST search on Swiss-Prot database.

The algorithm was trained on a dataset of 6729 proteins. It was tested on dataset of proteins coming from both CASP12 and CAP13 FM targets list along with membrane proteins and achieved a mean precision of 66.18% with this reaching up to 90% for 16 FM domain. Given the accuracy of prediction of residue-residue contacts of this algorithm we settled on utilising this tool to generate contact prediction files for our modelling test dataset.

4.2 Model Refinement

Model refinement is usually the last step in most of the computational protein prediction protocols with the primary two steps being, prediction of protein structure through TBM or FM protocols and second being assessment of the quality of models generated. In theory, the refinement step helps in bringing the predicted models closer to its native structure.

CASP8 introduced refinement category in the competition as a separate domain to solely focus on developing and testing strategies aiding in further improving upon the models built by prediction algorithms. This ensured a more dedicated research into pushing computational pipeline to take the modelled protein closer to its nativity. In line with Anfinsen's hypothesis (Anfinsen 1972, 1973), the native state of the protein is to be found at its lowest free energy. Henceforth, the refinement protocols are designed such that they focus on getting the models to its lowest free energy levels.

Only the starting model for a given protein is provided to the refinement community by different teams and the aim is to produce an improved model by the end of the refinement run. Several algorithms have been developed under this CASP category. This thesis is documenting a few in Table 4.1.

The recent bloom in the use of deep learning approaches in solving protein structure has also been implemented in developing structure refinement protocols. One of such recently developed algorithm is GalaxyRefine2 (Lee et al. 2019), which is an upgrade of GalaxyRefine web server(Heo et al. 2013).

The algorithm is improved upon such that it iteratively accumulates the conformational changes by movements done both locally and globally for any given model. Thereby, promoting larger changes to the protein structure, accounting up to two fold times to its predecessor. GalaxyRefine2 has been successfully tested on CASP12 (Hovan et al. 2018) and CASP13 refinement targets (Lee et al. 2019).

Table 4.1: Refinement Protocols

Refinement protocol	**Approach**	**Reference**
3Drefine	It is a two-step refinement process which is built on optimization hydrogen bonding network and energy minimization at atomic level	(Bhattacharya et al. 2013, 2016b)
ModRefiner	It is a two step refinement protocol based on all-atom level energy minimization	(Dong Xu et al. 2011)
GalaxyRefine	The protocol performs side-chain repacking followed by MD simulation based structure relaxation	(Heo et al. 2013)
PREFMD	A molecular dynamics based refinement program	(Heo et al. 2018)
PRINCETON_TIGRESS	Implicit solvent based MD refinement protocol	(Khoury et al. 2014)
KoBaMIN	It is based on minimization of knowledge-based potentials derived from solved PDB structures	(Rodrigues et al. 2012)

This web server takes input in the form of a monomeric PDB co-ordinate file and is currently limited to the refinement of protein upto 300 residue in length. It generates 10 refines models for each input PDB file in a run time varying from 6-10 hours. It was observed that currently our modelling pipeline is able to generate reasonable models up to the protein length of 100 amino acids; use of this refinement protocol fit well with our requirements to further improve upon the quality of the predicted models.

4.3 Modelling Protocol

The main idea behind the set up of this section is to use the pipeline as a whole for structure prediction on a set of ten small proteins chosen from the CASP13 target list. The publication goes into details of how the elements tested through out the thesis starting from the inculcation of an idea have been brought together to examine the validity of the work established so far. It also aims to make the structure prediction community aware of the possible strategy that can be used generate at least state-of-the-art prediction results at this established stage of the pipeline with further room for improvement.

The modelling protocol has been detailed in the form of a publication that focuses on two key elements of the pipeline which are PB based fragments and use of Modeller as fragment assembly tool. First step of the modelling protocol is fragment library generation for the query protein using an assigned PB sequence. This is followed by fragment selection based on secondary structure identity and coverage criteria leading to fragment assembly done in Modeller. Along with the selected fragments the modelling script is enriched with external restraints provided in the form of predicted secondary structure definitions and residue-residue contact predictions derived from PSIPRED and DMP web servers respectively. Both of the above stated restraints serve as an intra and inter residue contact essential for facilitating the global fold of the protein.

The modelling script is run to produce 10k initial models which lie in a spectrum of folding patterns discerned by Modeller. Best model from the bunch is chosen based on the GDT and TM-scores calculated for each model against the native structure of the protein. Additionally, the best model chosen based on the scores is further refined using the GalaxyRefine2 tool. This tool generates 10 refined models which are reevaluated for their nativity based on GDT and TM-score calculations. The final model presented is considered as the best optimized model for the query protein sequence in question. A detailed description of the experiments carried out on the modelling protocol is stated in the publication below.

ORIGINAL ARTICLE

Journal Section

Fragment-based *ab initio* modelling of small protein using external restraints in Modeller

Surbhi Dhingra[1]* | Ramanathan Sowdhamini[2] | Frederic Cadet[3,4,5] | Bernard Offmann[1]*

[1]Université de Nantes, CNRS, UFIP, UMR6286, F-44000 Nantes, France

[2]Computational Approaches to Protein Science (CAPS), National Centre for Biological Sciences (NCBS), Tata Institute for Fundamental Research (TIFR), Bangalore 560-065, India

[3]University of Paris, BIGR—Biologie Intégrée du Globule Rouge, Inserm, UMR_S1134, Paris F-75015, France

[4]Laboratory of Excellence GR-Ex, Boulevard du Montparnasse, Paris F-75015, France

[5]DSIMB, UMR_S1134, BIGR, Inserm, Faculty of Sciences and Technology, University of La Reunion, Saint-Denis F-97715, France

Correspondence
Prof G. Bernard Offmann PhD, Université de Nantes, UFIP UMR CNRS 6286, Nantes, 44322, France
Email: bernard.offmann@univ-nantes.fr

Funding information
Conseil Régional de La Réunion, Fonds Social Européen, Tier number 234275, convention number DIRED/20161451; Conseil Régional des Pays de la Loire, GRIOTE project

Several methodologies have been developed in the last few decades towards the modelling of proteins with no known structures. The progress made by the computational structure prediction field is commendable with constant efforts being put in to it to further increase the efficiency of these predictive algorithms. Towards this end, we report our results on a new fragment mining and assembly methodology developed in this work which is applicable to *ab initio* modelling of small proteins.

Its novelty relies on the fact that, in a primary step that makes use of Protein Blocks (PBs), a structural alphabet, for mining structural databases and building meaningful libraries of structural fragments. In a second step, these fragments are assembled to generate atomic models through the exploitation of the multiple template modelling protocol in-built in the well-known Modeller software. In addition to PB-based fragments, additional restraints in the form of predicted intra and inter residue contacts from PSIPRED and DeepMetaPSICOV respectively were also provided to Modeller to further guide the protein folding. The current pipeline has been tested on a dataset of 10 small proteins taken from the recent CASP13 target list.

Through this work, we were able to model small pro-

* Equally contributing authors.

teins reaching an average GDT score of 56.36. This study also served as an opportunity to evaluate Modeller in the field of fragment-based *ab initio* prediction under the guidelines of template-free modelling protocols.

KEYWORDS
Protein Blocks, Modeller, fragment-based *ab initio* modelling, protein structure prediction

1 | INTRODUCTION

Modelling of proteins that have negligible to no resemblance to already documented structures are put in the box of template-free modelling (TFM), *ab initio* modelling or free modelling (FM) protocols [1, 2]. To date this remains the most challenging problem in the field of structure prediction. Over the years, instances of progress have been marked by the introduction of new parameters into these approaches as monitored through Critical Assessment of protein Structure Prediction (CASP) experiments. Until CASP10 most of the good quality predictions were limited to small proteins, lying under 120 residue in length. A major breakthrough was seen in CASP11, where a good quality model prediction was possible due to the introduction of improved contact prediction strategies employed in the modelling protocols [3, 4]. The results of CASP13 were marked by the utilization of deep learning artificial intelligence based methods to predict protein structures, marking another leap in the protein structure prediction field [5].

Fragment-based approaches (FBA) have been very popular among *ab initio* methodologies for modelling protein structures. Most of them rely on assembling fragments covering short or long stretches of amino acid sequences from a representative structural database sharing sequence similarity with a query protein [6, 7, 8]. However, methods can be built that excavate extensive amount structural fragment data by not relying solely on protein sequences. This will help circumvent the limitation of sequence similarity to extract knowledgeable data from structural databases.

Recently, a fragment generation method was developed by our team towards this end [9]. It makes use of a Structural Alphabet (SA) [10, 11] type called Protein Blocks (PBs) [12] anticipating its use in a standardized TFM protocol as discussed below. From the analysis, it was confirmed that a massive amount of fragment data that share close structural resemblance to sequence counterpart can be extracted from a representative structural database. The quality of fragments generated using PB-frag approach was shown to be reliable with maximum fragments lying under *rmsd* of 2.5Å. The pipeline has been standardized according to the guidelines of FM protocol, where it pulls fragment hits of a minimum length of 7 PB residues from a representative structural database. The study also concluded that high secondary structure similarity between the query protein sequence and the fragment hits is a confirmation of the fragments sharing high structural resemblance with the query sequence at that position, thereby, aiding in determining potential good quality fragments.

In theory, PBs are 1D sequence representatives of the 3D protein backbone. Owing to this information, we hypothesized that the fragments generated using PBs should also follow the same trend, thereby, providing varied lengths of folds potentially resembling that of the protein sequence in question. These fragments can then be stitched together to construct a protein model.

In the current work, we have tried to gauge the potential of these PBs based fragments into free modelling of proteins by forcing a template-based modelling tool called Modeller [13]. At the same time we also wanted to verify the limits to which Modeller an be propelled into TFM by creating an environment that mimicks its multiple-template

modelling protocol. To this end models were generated by delivering PB derived fragments in a single alignment file to Modeller. The test were conducted on a small set of proteins with known structures. The model quality was assessed on the scoring criteria usually employed for ranking template-free modelling protocols. The best model generated for each target protein was further refined using a structure refinement protocol which employs an iterative optimization approach to manipulate protein structure.

2 | MATERIAL AND METHODS

The pipeline put together in this study constructs protein models by stitching together PB based fragments using multiple-template modelling protocol of Modeller (v9.19). The idea was to replace the full-length templates used in the basic (standardized) homology modelling protocol with fragments available for each position in the query sequence. The overall flow of the pipeline is shown in the Figure 1.

2.1 | Fragment Selection

Ten small proteins with known structures were chosen from the recent CASP13 target list. These were used as a working dataset to build the current pipeline (see Table 1.)

PB sequence were assigned to each query protein in the working dataset using an in-house scripts. These PB sequences were then used to generate fragment libraries by following the protocol described in our earlier study [9]. From the pool of fragment hits generated for each query PB sequence, a minimum of one fragment was selected per position until the maximal length of test protein was covered. The choice of fragments was based on shared secondary structure identity between the test protein and the fragment hits as well as the highest attainable coverage for the query sequence. These criteria were based on the earlier study performed for qualifying the PB based fragments. The secondary structure predictions for the query sequence were derived from PSIPRED v4.0 web server [14] while those for the fragments were the ones already assigned to the structures in the curated template database. The selected fragments were passed into Modeller as viable 'pseudo-templates' to predict protein structures. In this study, we have used the term 'pseudo-templates' to describe the truncated pdb coordinate files that correspond to the fragment hits and used in place of full length template files in Modeller.

2.2 | Assembling Restraints

The primary restraint the was provided to Modeller were the fragment hits derived from the PBs based fragment generation pipeline. As much as possible, overlapping fragments were chosen for modelling. This was done to perpetuate a contact throughout the fragments used, simultaneously ensuring introduction of none to minimum number of gaps between two adjacent fragments. In this case, fragments sharing at least a single residue sequence superimposition with each other were considered to have an overlap (bridge/connection). For the cases where the fragments did not share any sequence overlap were deemed as gaps in the sequence alignment.

Two external restraints were chosen to assist protein folding, namely, predicted secondary structure restraints and contact predictions. The secondary structure predictions for the target protein sequence were derived from PSIPRED v4.0 online web server [14]. These distributions were documented and passed into the Modeller script to enforce the secondary structure type in the model. The percentage distribution data is available in Supplementary Table 2. Contact predictions were derived from DeepMetaPSICOV v1.0 web server [15]. The procured contacts were

further filtered to recover all contacts lying above the prediction accuracy of 75%. The algorithm of DeepMetaPSICOV predicts medium to long-range contacts for a given protein sequence. In our case, we deemed any contact lying at a residue-residue distance of 10 residues or above as long-range contacts and the rest were considered as medium-range contacts. Henceforth, the contacts could be enumerated into medium-range and long-range contacts predicted for each query from the working dataset at a given accuracy.

2.3 | Modelling Protocol

The sequence alignment files were constructed based on the multiple template modelling protocol of Modeller. The coordinate files corresponding to individual fragments were extracted from full length pdb coordinate files using PDB_tool (v4.80) [16, 17] and were saved as 'pseudo-template' coordinate files. These co-ordinate files were then used to construct sequence alignments, which were manually edited for proper placement of fragments and gaps in accordance to hit position and the length of the query protein sequence. Sample alignment files can be found in Supplementary Figure 1 and 2.

The subsequent step was the construction of the modelling script with chosen fragments passed as pseudo-templates and inclusion of the two external restraints. Secondary structure restraints were passed as α-helix or β-strand based on the predictions made for each position within the target protein sequence. Few example of the command passed are shown below.

```
rsr.add(secondary_structure.alpha(self.residue_range('7:','10:')))
rsr.add(secondary_structure.strand(self.residue_range('12:','13:')))
rsr.add(secondary_structure.alpha(self.residue_range('14:','16:')))
rsr.add(secondary_structure.strand(self.residue_range('24:','29:')))
```

Similarly, the contact predictions were delivered as distance restraints in the modelling script, lying at a mean distance of 8Å between Cβ-Cβ except for Glycine, which was calculated for Cα atom in agreement with the distance standardised between two residues as suggested in DeepMetaPSICOV. Although, the distance restraint were granted some flexibility by introducing a std dev of 3Å for each restraint passed. The example of the command passed is listed below:

```
rsr.add(forms.gaussian(group=physical.xy_distance, feature=features.distance(at['CB:2'],
at['CB:6']), mean=8.0, stdev=3.0))
rsr.add(forms.gaussian(group=physical.xy_distance, feature=features.distance(at['CB:6'],
at['CA:10']),mean=8.0, stdev=3.0))
```

A total of 10k models were generated for each run (each query) followed by quality assessment performed on the basis of GDT_TS score, rmsd and TM-score calculated using the tool, TM-score [18]. Best model was determined on the basis of highest GDT_TS score calculated for all models against the original structure of test protein. Z-score was then calculated for the original structure as well as the best model (initial model) using the protein structure analysis tool, ProSA [19, 20].

2.4 | Refinement of Models

The web server, GalaxyRefine2 was further used to refine the best predicted model for each target protein from the working dataset [21, 22]. This tool was used as it has been demonstrated to be an effective refinement tool in recent CASP13 analysis [23]. The output of the web server is in the form of a table listing Clash scores, Molprobity score, Ramachandran Clashes etc. in comparison to the initial model submitted as well as the downloadable coordinate files of the 10 refined models. All the refined models were re-assessed using the same scoring criteria as above (GDT_TS score, TM-score and rmsd). This accounted for any change in the quality of refined models in comparison to the best model generated by Modeller.

3 | RESULTS

Fragment libraries were generated for each target protein sequence producing on average ~60,000 hits across the working dataset. From this, the number of identical PB hits averaged to ~14,000 fragments (results have been tabulated in Supplementary Table 1). The total number of fragments selected for final modelling varied with the total length and structural composition of query proteins. The fragment length ranged between 7-31 PB residues which corresponds to 11 to 35 residues throughout the chosen fragment list.

All the fragments chosen for modelling of the target protein sequence shared an overlap between two adjacent fragments except for one target protein which is 6d7y. The lack of sequence overlap can be translated into the introduction of gap in the sequence alignment file. This information was taken into account as per Modeller documentation which states that the quality of model generated decreases with the increase in the number of gaps. While taking care of the minimal introduction of gaps; the usage of overlaps also served as a bridge to maintain the contact between all the fragments, thereby, guiding the protein fold (orientation).

Two external restraints were chosen to assist protein folding, namely, secondary structural restraints and predicted contacts respectively. When compared to the secondary structure distribution observed in the original structures from the working dataset, it was noted that the predictions were more or less close to that of original structure with the exception of few cases showing bigger variation. In this study, we focused on getting larger number of contacts despite losing accuracy with a lower confidence level in order to increase our chances of retaining long-range contacts. Alongside, contact predictions have also been shown to be valuable for constricting the conformational space. The total number of contacts recovered for each test case above the confidence level of 75% along with the number of potential long-range contacts have been documented in the Table 2.

3.1 | Model Building

Table 2 summarises all the parameters that were considered while building the models from PB generated fragments. We were able to attain an average coverage of 98% for all the test proteins in the working dataset. The number of contacts predicted above the defined cut-off level varied despite the length of the protein sequences. Alignment files were constructed according to the syntax followed by Modeller multiple-template modelling protocol. Here, full length templates were substituted by fragments chosen for each position within the query sequence. The final modelling script was enriched with the two external restraints and PB based fragments to generate 10k models.

The GDT_TS scores calculated for the initial models ranged from as low as 35.1 to the maximum of 88.6 respectively (see Table 3). An overall increase in GDT_TS score was observed for all the refinement cases except for two (6d7y and 6cp9). In these two cases, target protein, 6d7y did not show any difference in the global distance score cal-

culation and the second one, which is 6cp9, suffered from a reduction of the calculated score. The rmsd calculations between the original structure and the inital models ranged from 1.40 to 16.3, with 60% of test cases lying under rmsd of 10Å. A slight improvement in rmsd was observed post refinement, with models lying between 0.9 to 14Å. The two smallest protein models were the one with the best of GDT_TS score, TM-score and rmsd (see Table 3). This could be attributed to the shorter length and in general less complicated structure of the protein.

Z-scores were calculated individually for the original structure, its initial model and the refined model on the ProSA web server. The pre and post refined models for all the proteins from the working dataset were in the allowed region of the graph as per the regions documented for known structures. The descriptive graphs for all the cases can be seen in the Supplementary Figure 3. Figure 2 shows the superimposition of the best model generated by Modeller along with the best refined model on to the known structure of the protein. It is clear from the figure that the best model regeneration was in the case of targets, 6fxa (Figure 2.a) and 5w9f (Figure 2.b) respectively.

4 | DISCUSSION

This pipeline set up during this work tested the effectiveness of a new PBs based fragment generation pipeline into *ab initio* protein structure prediction done using a homology modelling tool. It also explored the degree to which Modeller can be pushed into free modelling of small proteins.

We were able to generate models with GDT_TS score as close to 100 (describing the best fit on the original structure) as well as models going as low as 34.65 in GDT_TS score calculations. From the 10 protein models constructed to built this pipeline, 60% of our refined models achieved an acceptable GDT_TS score of >50, with the rest not lying far off from the accepted threshold. As a starting point, these results seem to be promising and can be improved upon by modifications done at the modelling level.

The usefulness of long-range inter-residue contacts in PSP have long been noticed and discussed in CASP biennial competition [24, 25, 26]. Recent improvement in prediction strategies have been attributed to improved contact predictions as documented in CASP11 results [27]. We used one of the deep learning approaches that has been vastly discussed in recent CASP13 for the prediction of contacts. The choice of providing contact prediction as restraints was to promote protein folding by restricting the conformational freedom of modelled structures.

We played against the sequence similarity requirements of Modeller. Since, the templates were provided in the form of fragments derived from PBs, they offered higher structural identity in comparison to its sequence counterpart. We were able to demonstrate that models can be build using basic comparative modelling protocol while abiding to its technical rules. Through our pipeline, we were also able to construct average to good quality models for small proteins listed in our working dataset.

The modelling process gets tougher with increase in the size of the protein sequence. To this end, there is a possibility to construct better models by providing additional external restraints to Modeller. The probability of getting the better fold is also limited by the number of gaps introduced in the alignment file. Wherein, the probability of encountering gaps increases with the length of the protein sequences. It was also noticed that the over representation of one kind of secondary structure type did end up affecting the morphology of the predicted models (see figure 2.c). We would also like to point out that the size of the proteins modelled in the current work usually come under the average domain lengths known for proteins. It might be the case that so far the pipeline is able to model at domain level only.

Finally, we tested to what amount a structure can be mapped back to, given a PB sequence derived from its original structure. This was demonstrated in terms of calculated rmsd and GDT scores. From the Table 3 it can be

seen that for 2 cases we reached a GDT_TS score of 93.57 and 89.63 respectively, which correspond to the predicted structures being very close to the native state. All the fragments generated came from a representative database which did not share any sequence homology with the query sequence. The free modelling approach was tested in a comparative modelling environment where each fragment was treated as a complete template.

5 | CONCLUSION

Ab initio protein structure prediction is an ever growing field with new strategies being developed everyday aiming to solve the sequence-structure-function paradigm. Amidst this era of immense leap in PSP, we tried to step back and excavate the potential of a comparative modelling tool into the field of unknown structure determination. We tried to maximize the raw material available to us into solving the structure prediction dogma. We questioned the effectiveness of a standardised comparative modelling tool when provided with half of its resources (resources with false pretence) and forced it into *ab initio* modelling domain. Overall, the initial results were encouraging and further research can be done to increase the efficiency of the pipeline, specifically focusing on doing so for larger proteins.

6 | EXTRA INFORMATION

Supporting information will be included with the published article. For submission any supporting information should be supplied as separate files but referred to in the text. Appendices will be published after the references. For submission they should be supplied as separate files but referred to in the text.

acknowledgements

conflict of interest

references

[1] Khor BY, Tye GJ, Lim TS, Choong YS. General overview on structure prediction of twilight-zone proteins. Theoretical Biology and Medical Modelling 2015;12(1):15. http://www.tbiomed.com/content/12/1/15.

[2] Feig M. Computational protein structure refinement: almost there, yet still so far to go. Wiley Interdisciplinary Reviews: Computational Molecular Science 2017;7(3).

[3] Kinch L, Li W, Monastyrskyy B, Kryshtafovych A, Grishin N. Evaluation of free modeling targets in CASP11 and ROLL. Proteins 2015 12;84 Suppl 1.

[4] Abriata LA, Tamò GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. Proteins: Structure, Function and Bioinformatics 2018 mar;86:97–112.

[5] AlQuraishi M. End-to-end differentiable learning of protein structure. Cell systems 2019;8(4):292–301.

[6] 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. Journal of Molecular Biology 1997;268(1):209–225.

[7] Kandathil SM, Handl J, Lovell SC. Toward a detailed understanding of search trajectories in fragment assembly approaches to protein structure prediction. Proteins: Structure, Function, and Bioinformatics 2016;84(4):411–426.

[8] Kandathil SM, Garza-Fabre M, Handl J, Lovell SC. Improved fragment-based protein structure prediction by redesign of search heuristics. Scientific Reports 2018;8(1):1–14.

[9] Dhingra S, Sowdhamini R, Sanejouand YH, Cadet F, Offmann B. Customised fragment libraries for textitab initio protein structure prediction using a structural alphabet. arXiv preprint —— 2020;.

[10] Unger R, Harel D, Wherland S, Sussman3 JL. A 3D Building Blocks Approach to Analyzing and Predicting Structure of Proteins 1989;5:355–373.

[11] Camproux AC. A Hidden Markov Model Derived Structural Alphabet for Proteins. Journal of Molecular Biology 2004 06;339:591–605.

[12] de Brevern A, Etchebest C, Hazout S. Bayesian Probabilistic Approach for Predicting Backbone Structures in Terms of Protein Blocks. Proteins 2000 12;41:271–87.

[13] Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology 1993;234(3):779–815.

[14] Buchan DWA, Jones DT. The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic acids research 2019 jul;47(W1):W402–W407.

[15] Kandathil SM, Greener JG, Jones DT. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins: Structure, Function and Bioinformatics 2019 dec;87(12):1092–1099.

[16] Wang S, Zheng WM. CLePAPS: Fast Pair Alignment of Protein Structures based on Conformational Letters. Journal of bioinformatics and computational biology 2008 05;6:347–66.

[17] Wang S, Ma J, Peng J, Xu J. Protein structure alignment beyond spatial proximity. Scientific reports 2013 03;3:1448.

[18] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function and Genetics 2004 dec;57(4):702–710.

[19] Sippl MJ. Recognition of Errors in Three-Dimensional Structures of Proteins 1993;17:355–362.

[20] Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 2007 Jul;35(Web Server issue):W407–410.

[21] Lee GR, Heo L, Seok C. Simultaneous refinement of inaccurate local regions and overall structure in the CASP12 protein model refinement experiment. Proteins: Structure, Function and Bioinformatics 2018 mar;86:168–176.

[22] Lee GR, Won J, Heo L, Seok C. GalaxyRefine2: simultaneous refinement of inaccurate local regions and overall protein structure. Nucleic acids research 2019 jul;47(W1):W451–W455.

[23] Read RJ, Sammito MD, Kryshtafovych A, Croll TI. Evaluation of model refinement in CASP13. Proteins: Structure, Function and Bioinformatics 2019 dec;87(12):1249–1262.

[24] Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins 1999;Suppl 3:149–170.

[25] Lesk AM, Conte LL, Hubbard TJP. Assessment of novel fold targets in CASP4: Predictions of three-dimensional structures, secondary structures, and interresidue contacts. Proteins: Structure, Function and Genetics 2001;45(SUPPL. 5):98–118.

[26] Izarzugaza JM, Gra?a O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7. Proteins 2007;69 Suppl 8:152–158.

[27] Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. New encouraging developments in contact prediction: Assessment of the CASP11 results. Proteins 2016 09;84 Suppl 1:131–144.

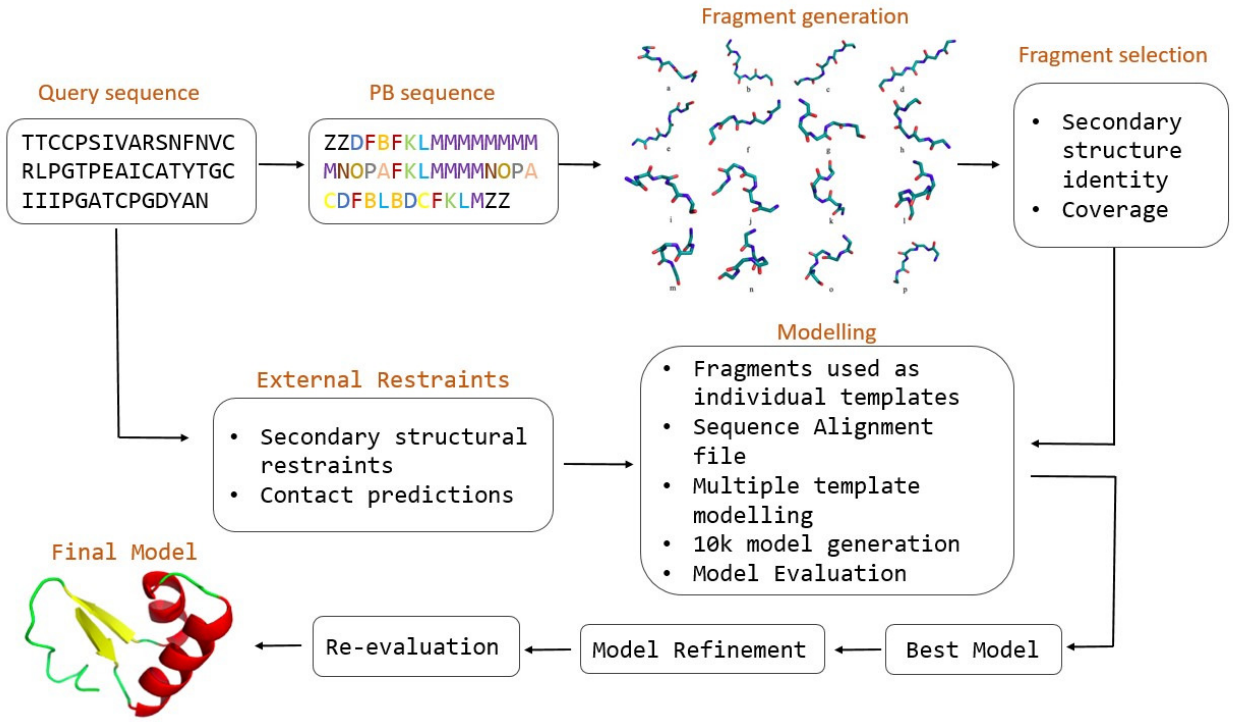


FIGURE 1 This figure is the visual representation of the steps followed into modelling proteins using fragments derived from PBs based pipeline. The initial step involves PB sequence generation followed by their use in the modeling pipeline. Protein structure construction is carried out in Modeller using its multiple-template modelling protocol. External restraints are added to the modelling script in the form of secondary structure restraints and distance restraints. Best model is chosen based on GDT_TS scores calculated for each model against original structure. The refinement of the best models is carried out to get 10 refined models which are re-evaluated using same scoring criteria.

TABLE 1 This table lists all the protein chosen for building a working dataset of small proteins derived from the latest CASP13 target list. It also documents the length of the proteins along with the experiment type used to infer each of the structure.

PDB Id	Chain	Protein Biology	Experiment Type	Length
6fxa	A	TP901-1 CI repressor	X-Ray Diffraction	35
5w9f	A	Denovo designed peptides	Solution NMR	41
6gnx	B	Telomere repeats-binding bouquet formation protein 2	X-Ray Diffraction	52
6f45	B	bacteriophage S16 long tail fiber distal subunit	X-Ray Diffraction	72
6qek	A	Magnetosome protein	X-Ray Diffraction	79
6msp	A	Denovo designed protein	Solution NMR	80
6qfj	A	Magnetosome membrane protein MamB	X-Ray Diffraction	81
6btc	A	LP1413 - SCCmec type IV-encoded DNA binding protein	X-Ray Diffraction	96
6d7y	A	Hemagglutinin	X-Ray Diffraction	96
6cp9	B	Contact-dependent growth inhibition toxin	X-Ray Diffraction	116

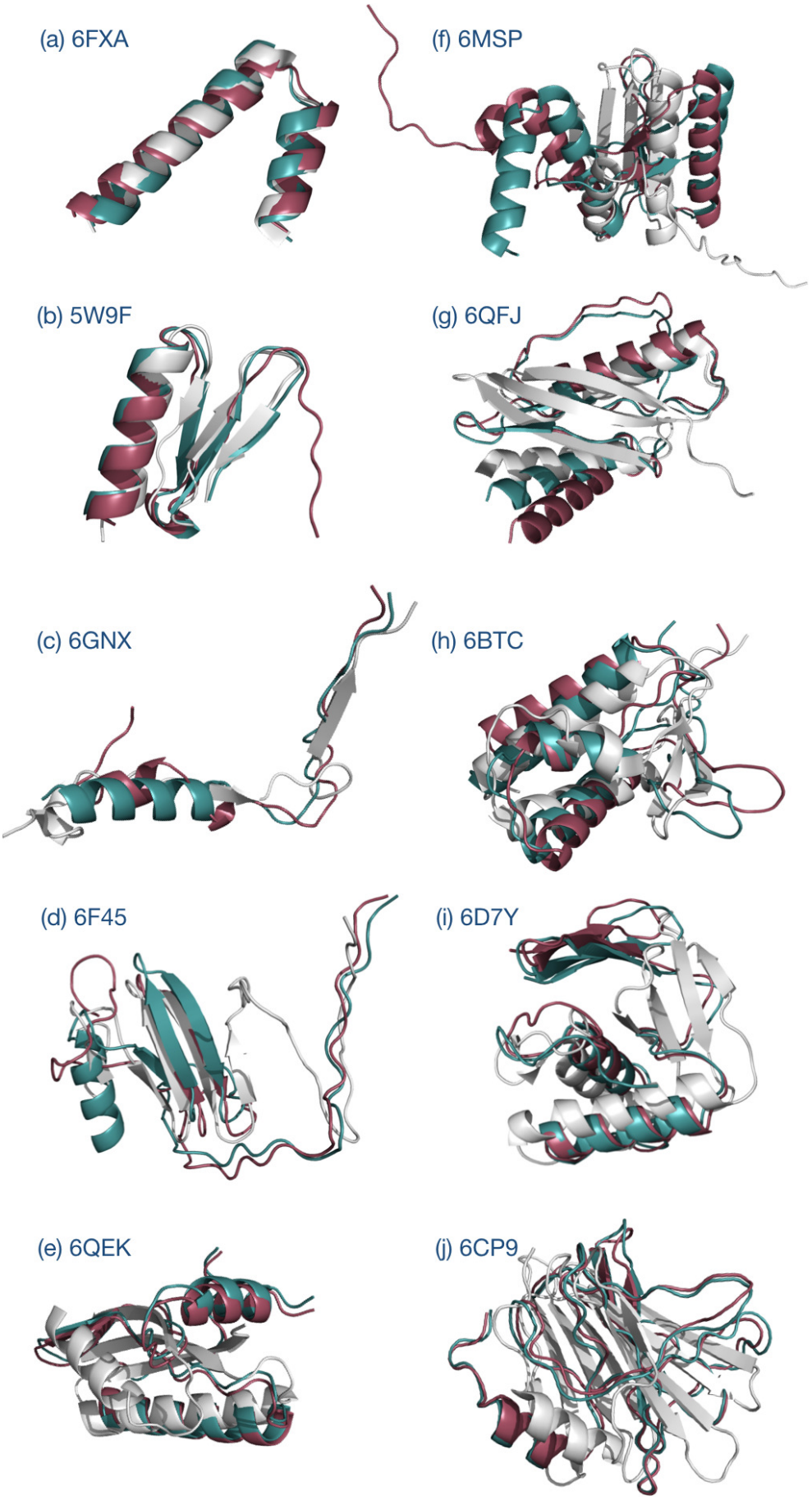


FIGURE 2 shows superimposition of initial model (pink) and refined model (green) on to the original protein structure (white) of the protein for all the targets in the working dataset. The structures have been arranged with the increasing length of the protein (a to j).

TABLE 2 Summaries all the parameters passed into Modeller for generating models. These include the number of fragments used for modelling, number of internal gaps observed in the alignment file, percentage coverage attained from the fragments used for each target protein, total number of predicted contacts lying above the 75% confidence level cut-off and the number of probable long-range contacts from the previously filtered list of contacts for each target protein within the working dataset.

PDB Id	Number of Fragments	Number of Gaps	Coverage	Total number of contacts	Long range contacts
6fxa	4	0	100	73	0
5w9f	10	0	100	43	9
6gnx	13	0	100	27	6
6f45	31	0	94.4	28	8
6qek	20	0	97.44	333	94
6msp	28	0	98.99	89	10
6qfj	21	0	97.62	325	69
6btc	25	0	97.62	184	11
6d7y	26	2	96.63	90	19
6cp9	32	0	97.37	231	64

TABLE 3 This table summaries the protein structure prediction scores as observed for the cases of pre and post refinement models. The scores include GDT_TS score and TM-score as well as rmsd distribution of each model. The last column summaries Z-score distribution as observed for the original structure, initial model and refined models.

PDB Id	RMSD		GDT_TS		TM-Score		Z-score		
	Initial	Refined	Initial	Refined	Initial	Refined	Original	Initial	Refined
6fxa	1.40	0.99	88.57	93.57	0.61	0.77	-1.56	-1.72	-2.12
5w9f	3.6	1.2	74.39	89.63	0.57	0.79	-7.17	-3.44	-6.7
6gnx	7.7	5.8	46.77	51.61	0.27	0.28	0.38	-1.16	-0.23
6f45	11.14	10.4	42.36	44.1	0.39	0.39	-2.07	-2.77	-2.93
6qek	9.7	9.72	52.56	54.17	0.49	0.52	-5.27	-2.47	-2.92
6msp	16.27	13.93	44.59	45.1	0.45	0.45	-4.42	-2.69	-4.3
6qfj	8.9	7.70	49.7	58.93	0.48	0.58	-5.18	-1.01	-3.17
6btc	5.35	4.61	52.08	54.17	0.44	0.49	-4.52	-2.85	-4.62
6d7y	11.5	11.15	37.64	37.64	0.37	0.35	-2.38	-3.29	-3.66
6cp9	13.4	13.51	35.09	34.65	0.39	0.37	-4.66	-0.37	-2.53

4.4 Concluding Remarks

This chapter compiled the two protocols established and tested in the former two chapters (2 and 3) of the thesis to perform experiments on a stitched together pipeline arranged toward achieving the goal of the work in hand. Towards this goal, we discussed in the beginning of the chapter the importance of the two external restraints that have been chosen to guide protein folding during the assimilation of the protocol. We have also shone some light towards the refinement step added at the end of the modelling protocol and its importance in guiding the modelled proteins further towards nativity.

In the later part of the chapter the methodology and results obtained by the study have been documented in the form of a paper discussing the outcome of the experiment performed on 10 small proteins. We have documented the number of fragments that were generated for these test proteins using our PB based fragment generation pipeline. Thereby, confirming that a good amount of fragment data can be produced via the pipeline established in Chapter 2 of the thesis and that the representative database built is enriched enough to deliver structural information even for the recently added structures in the PDB database.

Likewise, to promote protein folding, restraints were added to the modelling pipeline in the form of distance restraints and rigid secondary structure definitions. A combination of these restraints along with the fragment data followed by the refinement step generated high to medium resolution structures for the test proteins. We were able to benchmark an acceptable GDT and TM score for these test cases, with the average being 56.36 and 0.5 respectively. Both these scoring function lie at the range which is generally considered as a point where modelled protein and the native structure have similar folds. The average documented GDT_TS score estimated for CASP13 results was 65.7 which is not far off from the average prediction scores we have been able to reach on the validation dataset (56.36) used in this study.

We were also able to demonstrate through a simple Z-score calculations that the best models predicted through our pipeline were lying in the known territory of the structural features captured so far. Therefore, confirming that the folded proteins were all in the allowed regions of the protein packing. At this stage we would like to emphasize that this *ab initio* modelling protocol is still on its way to maturity and hence can be worked upon more to increase the average GDT_TS score calculations. It would also be beneficial to add a scoring criteria at the end of the preliminary modelling output (10k models) to pick up the top 5 or 10 models predicted when the pipeline is applied to blind testing of the PSP.

The results demonstrated during this experiment are also representative of the best case scenario where no loss of prediction accuracy is procured due to the initial prediction of PB sequence from its primary sequence. But, documenting these variation will also give a clearer picture of how well the pipeline will perform during blind testing. Despite the scope of improvements needed, the results obtained by the initial runs are encouraging to keep working towards the development on the modelling front of the project and participate with this pipeline in the next CASP competition in 2022. It will give enough time for the protocol to mature with all its strengths and weaknesses being well worked on. At this point, it is safe to say that under the influence of PBs derived information delivered in the form of fragments,

Modeller was able to handle protein folding at an acceptable level and further exploration of its internal design would help to better implement the fragment based protocol on this comparative modelling tool.

As time goes on, you will understand. What lasts, lasts; What doesn't, doesn't. Time solves most things. And what time can't solve, you have to solve yourself

- Haruki Murakami, *Dance Dance Dance*

5 Conclusions and Perspectives

This chapter deals with the concluding remarks on the work established so far in this project as well as discusses on the perspective of the possible modifications or improvements that can be implemented to the current protocol to further enhance its structure building properties.

While doing so this chapter will also summaries the parts described in each chapter of this thesis in terms of implementation of PBs directly or the knowledge derived through the exploitation of these structural alphabets into constructing a modelling pipeline for *ab initio* protein structure prediction.

We will shed some light on the benevolence of this research in the structure prediction field. As well as the use of Modeller in *de novo* PSP field. To the best of our knowledge, sole dependency on Modeller for prediction of structures for *ab initio* protocols have not been done so far.

5.1 Conclusions

Development in the field of computational protein structure prediction was an evident outcome after achieving a tremendous success in making its complementary part (genome sequencing) readily accessible.

Though generally this factor is iterated over as one of the primary reason for the development in computational structure prediction field, the need to decipher protein structure is emphasised specifically to understand the mechanism behind proteins folding and their function. In another words, to decrypt the sequence-structure-function paradigm. So much that it has become central to the advances in molecular biology in regards to drug discovery, mutation studies and protein engineering etc.

The go-to experimental techniques for structure generation provide us with high resolution 3D models of proteins that can directly be used to answer major molecular queries. Although, the time consumed in generating each protein structure varies a lot and hence has always been a major drawback of this field. Despite that the quality of structure data generated through these techniques surpass the ones

detected by computational approaches.

On one hand, the success achieved in the development of efficient sequencing techniques made evident the capacity of computational biology in accelerating the biological research by many folds. On the other hand, proteins being dynamic macromolecules and their structures having variable degree of flexibility poses a far more difficult challenge than its sequence counterpart. Development of prediction algorithms that correctly identify protein folds is far from being over. Nevertheless, several computational techniques have been standardized that can predict protein structures having known structural homologues available in structural database, consequently filling the sequence-structure gap to a great degree.

Invariably, these techniques also reduce the time factor that is associated as a limiting component of experimental protein structure determination. Owing to the fact that structure determined through homology/comparative modelling techniques are influenced by the known structural databank, the protein atomic data accumulated through these approaches more often than not advocate structural redundancy.

When this redundancy is set aside, the altogether variability in structure information documented is still very low. To increase the variability quotient of the structural data documented, past few eras have seen an increasing amount of efforts put in the development of computational protein structure prediction field called Template-free modelling/*ab initio* modelling which focuses on predicting structure of proteins with limited to no structural data available.

Though, the efficiency of these techniques are limited by several factors, specifically when dealing with large proteins or proteins with complex folds. Leading to an allover reduction of the credibility of the field in providing high resolution structures which can then be used to solve other biological problems. A continuous advancement in this field helps to gain newer insights into the protein folding mechanism and the possibility to generate new folds. This in general is limited due to the redundancy of structural data available in PDB database, making the allover information less intuitive in solving the second genetic code.

Globally, scientific communities have been working on designing algorithms using various approaches such that this gap between the sequence and structure is reduced. Thus, making computational structure prediction field a main stream in producing good quality structure which can be used in other research. This is also driven by the fact that these approaches are much more efficient in terms of time consumption and man power requirements. Hence emphasising on the need to put continuous efforts to further develop similar protocols using newer methodological approaches. Regardless, of the improvements seen over the years in this field, it is still far from reaching its full potential, specifically if focused on template-free modelling protocols. Keeping this in mind, the current project on protein structure prediction was designed to introduce a new aspect into this field via testing the scope (potential) of a Structural Alphabet type called Protein Blocks in determining the correct protein folds.

A major portion of the work established around SAs deals with the classification and characterization of proteins (Le et al. 2009; Dong et al. 2016), defining the protein backbone locally (Vetrivel et al. 2017), protein fold recognition and distant similarity search for comparative modelling of proteins (Ghouzam et al. 2016), protein structure alignments (Gelly et al. 2011; Joseph et al. 2011), loop modelling (Camproux et al. 2001) etc.

Not a lot of research has been done in determining the capability of structural alphabet to tap into the *ab initio* structure prediction field. Though research focused around the ability of SAs to determine structural variability or the ability to register alternative conformation of the proteins (Mahajan et al. 2014; Regad et al. 2017) as well as its ability to predict protein flexibility (Dong et al. 2016) have been extensively conducted . Thereby making it a beneficial tool in the field of *de novo* modelling as these algorithms are usually unable to determine the alternate conformations of the target protein.

Along with this, a fragment generation pipeline developed in 2013 made use of SAs to generate fragment coordinate files superimposition according to the primary sequence in question (Shen et al. 2013). It produced high precision, accuracy and coverage of structural fragment predicted at varying length of 6 to 27 residues. Another research which used fragment based assembly to do homology modelling of proteins was based on the use of Proteins Blocks for doing the same (Brevern et al. 2003).

Having the basis established for the dimensionality of SAs over the past several years, the premises of the current project were laid down. This project proposed to discover the scope of SAs in modelling medium to low resolution structures of proteins with no known structural homology. The sequence-structure relationship has long been explored to fold proteins in the same order but the research focused into utilising distant structural ques to predict *ab initio* models is still in the path to progress.

We had a tool that has been disposable to serve our needs in the form of Protein Blocks which are a type of SA. These have been standardized and used for structural manipulation for past several years and here the idea was to explore the dimensions in which PBs can be used objectively for developing TFM protocol. To this end, the project commenced with the utilization of Protein Blocks to set up two independent pipelines. One dealt with the construction of a PB based fragment generation pipeline and second was optimising the deliverance of fragments into a modelling tool to predict optimal structures.

The initial question posed was the way PBs could be utilised in carrying forward the structure prediction pipeline. Given the success achieved by fragment based protein structure prediction protocols and garnering the influence from the study published in 2013 on building high coverage fragment libraries using SAs type (Shen et al. 2013), we decided to go down a similar path.

To do this a PB based fragment generation pipeline was first optimised which was in turn used for testing fragment based PSP protocol standardized on Modeller. The pipeline relied on pulling similar structural signatures from a non-homologous database using a pairwise comparison between the target protein and the established structural database. In the duration of this study we have shown that PBs can be a good detector of distant structural features already available in the structural pool while functioning as a helpful indicator of local structural similarity between the target sequence and structural templates through an accessible pairwise comparison of PB sequences. This is in line with the generally admitted idea of "structural convergence", i.e despite the large diversity in protein sequences, the way they fold locally is quite limited (Tomii et al. 2012). Our aim was to capture this sequence-structure relationship without relying on sequence similarity.

In that respect, this pipeline employs local pairwise alignment of PB sequences to

look for identical structural motifs in a representative structural database populated with unrelated proteins. Each of these structural motif is considered as a fragment hit. The quality of the fragments were tested at different *rmsd* cut-offs to confirm the precision of the pipeline. We were also able to confirm that high coverage fragment libraries can be fabricated through the protocol discussed in Chapter 2. This was possible through the ability of PBs to provide a more descriptive representation of a protein structure in the form of 16 local conformations in contrast to secondary structural elements which are typically limited to 3 structural states (α-helices, β-strands and coils). Subsequently providing a more precise description of protein local backbone and the ability to use these descriptions for the construction of protein models.

The inherent nature of PBs was also significant in the construction of the pipeline by demonstrating its ability to retain similar folds from unrelated protein structures which can be combined to generate new models. Considering the aim of the current work was to develop a new method for *de novo* PSP, the fragment generation pipeline was standardized under the approved rules set around such protocols.

The pipeline showed the lowest precision values for regions comprised of β-strands. This could be because of the reason that in general β-strand geometry is closer to loops and therefore more difficult to recognize. All in all, our pipeline stood on par with the other well documented fragment generation strategies available alongside showing the highest coverage generally retrieved by such protocols.

After having standardized a fragment library generation protocol, the obvious next step was to work on the fragment assembly pipeline for model building. In accordance with our strategy to design a fragment based protein structure prediction approach, the strengths and limitations of Modeller were assessed in view of its use as an *ab initio* structure prediction tool. Therefore, several experiments were designed and tested on Modeller as discussed in Chapter 3 of the thesis to maintain a continuity in the development of the pipeline from Chapter 2 and 4.

The experiments were conducted to explore the ways fragments can be delivered to Modeller inorder to construct models closest to its nativity. Through these experiments, it was recognized that fragments linked through sequence overlap tend to generate lower resolution models in comparison to the ones derived from fragments sharing no overlaps. This also indicated towards the reliance of Modeller to have a continuity in the flow of information in order to construct better models, which in this case was achieved by sequence overlaps shared between two adjacent fragments.

In the later stages of analysis, the results from an experiment performed on a small protein following fragment-based protein structure prediction protocol was discussed in the chapter. This served as an establishing ground to carry forward the analysis on Modeller using a bigger dataset with additional restraints. In the experiment carried out on the test protein Crambin (1CRN), fragment library was generated using the PB fragment generation protocol and 5 random fragments were picked up for model generation.

Fragments and target protein alignment files were built and passed into the Modeller script to generate 100 models. The best model constructed by following this procedure had a global distance score of 76 out of 100 with a recorded *rmsd* of 3.25 Å. As an initial trial this served as a promising result given in general analysis, GDT score of 50 or above usually indicates the atomic coordinates of the models lying close to that observed in its native structure. This was also confirmed by the

visualization of model superimposition onto its original structure where the model retained the global fold of the protein.

In Chapter 4, we discussed how additional restraints can be used in the modelling protocol to facilitate accurate protein folding. This was demonstrated using two types of external restraints added to the modelling scripts and the results were tested on a dataset of 10 small proteins taken from CASP13 target list. The restraints considered in this study have time and again been proven to be effective in guiding protein fold. These include predicted secondary structure descriptors of the target protein sequence and the deep learning based contact predictions generated by recently standardized method called DeepMetaPSICov. The contact predictions were passed in the form of distance restraints in the modelling script whereas the secondary structural restraints were used to promote a secondary structure type in the defined region.

A more dedicated research on protein structure refinement has been effectuated in the past few years leading to the development of specific pipelines for post template-free modelling purposes. This also encouraged segregation of the structure determination problem into sections to promote a committed development in each sector. In line with this, an additional refinement step was also added in the end of the pipeline as a measure to bring the best model closer its nativity.

The pre-refinement modelling runs produced best models lying at the *rmsd* range of 1.4 to 17 Å with 40% of the proteins in the test dataset having GDT_TS score of 50 or above. This range shifted to a minimum of 1 Å to the maximum of 13.9 Å in the refined models and 60% of the test data lay above the GDT_TS score of 50. To summarise, chapter 4 of this thesis dealt with implementation of a template-free modelling pipeline and its application onto a dataset of small proteins. We evaluated the quality of protein models generated when using fragments derived from a PB based pipeline into a comparative modelling tool, followed by the refinement process.

Though, we would like to point out that these are still preliminary results with the validation tests done on a dataset of small proteins. But, on the hind side, this is the area of *ab initio* PSP that our results can easily be correlated to, as most of the prediction approaches are efficient in modelling high accuracy structures upto a certain length (<150 residues). So, to have achieved an overall decent GDT_TS score range on a relatively non-standardized pipeline is a good starting point.

The pipeline is still at a nascent stage of development and therefore has a lot of scope for further manipulations or additions to increase the quality of output generated presently. This work can also be considered as the first step towards *de novo* modelling, of protein structures using a SA type. The results obtained, though fascinating, are limited by the extend to which Modeller has been explored in the given time frame of the project and can be exploited more in the future. One must also keep in mind that Modeller as an approach has specifically been designed to do comparative modelling of proteins, was still able to achieve an acceptable result at modelling the entire length of protein sequence using the information coming from a set of overlapping fragments.

In short, throughout this thesis work, we were able to come up with a PB-based fragment generation pipeline which will soon be made available to users in the form of a webserver (PB-Frag) to generate fragment libraries based on the query sequence provided. We were also able to analyse preliminary results on the modelling pipeline set up so far, relying on fragment assembly using a comparative modelling tool called

Modeller.

5.2 Perspective

The work done during the time frame of this thesis provides a glimpse into the fundamental properties of PBs and how they can be implemented into building *ab initio* protein structure prediction protocols. Nevertheless, the work is still at its early phase of development and can be improved upon by introduction of additional steps in order to increase the efficiency of the pipeline established so far.

One of the aspect that can be worked on is the PB prediction accuracy for protein sequences with no known homologues. Since, PB prediction accuracy will have a direct effect on the quality of fragment library generated by the PB-based fragment generation pipeline. Naturally, the protein targets provided (chosen) for modelling in TFM protocols are the ones which have no identified homologous structure available in the PDB database. This essentially falls in the B8 category of PB-kPRED algorithm, where PDB dataset shares $<30\%$ sequence identity with the test protein sequences resulting in an average PB prediction accuracy of $\sim 39\%$. In the current scenario, this will lead to generation of fragment data further from the actual structural description of the protein sequence. Further optimization of the fragment generation pipeline is required in terms of development of a scoring function to determine the best representative fragment hits per position.

Similar sets of experiments as conducted in Chapter 4 of the thesis have to be conducted on PB sequence derived from PB prediction protocol which was not feasible at the moment due to the low accuracy PB sequence predictions accumulated for the modelling dataset used in this chapter. On the same grounds, one drawback that can be faced by the use of Modeller for *de novo* prediction is its inability to model all types of secondary structural sub-classes thereby restricting the possibility to model all possible search spaces of the protein structure.

Although the thesis work has already included distance based contacts to promote protein folding, yet a more intensive search into employing long-range contacts for three dimensional protein structure arrangements have to be investigated. Since, it has already been shown to be a driving factor in protein folding. For example, development of a protocol to determine PB based long range contacts might add a new dimension to this protein folding protocols.

Additional restraints can be included in the modelling script to determine their contribution in improving the protein fold. These can include distance based restraints employed in the form of di-sulphide bridges and lys-lys bonds derived from mass-spec data etc. Inspiration can also be drawn from the CASP target delivery format, where more focused pipelines can be set up which rely on using restraints derived from the preliminary results of several experiments like NMR based, mass spectrometry based and cryo EM based restraints etc.

Another territory that can be explored is the development of protocol solely based on predicted local backbone information produced by using the knowledge-based prediction of PB sequence from its primary sequence. In other words, folding protein sequence based on the dihedral definitions of each PB character.

An in-depth research can be done on detecting the range of alternative conformations covered by the fragment generation pipeline within hits generated for each

position for a query protein sequence. This will help us estimate the amount of conformational search space covered by the fragment generation pipeline.

One type of proteins that were not modelled (tested) during the compilation of this work belong to the class of intrinsically disordered proteins. It will be interesting to know if the set up of the current pipeline is able to correctly model such proteins as they have remained a bottleneck of modelling field in general.

Together, we form a necessary paradox, not a senseless contradiction.

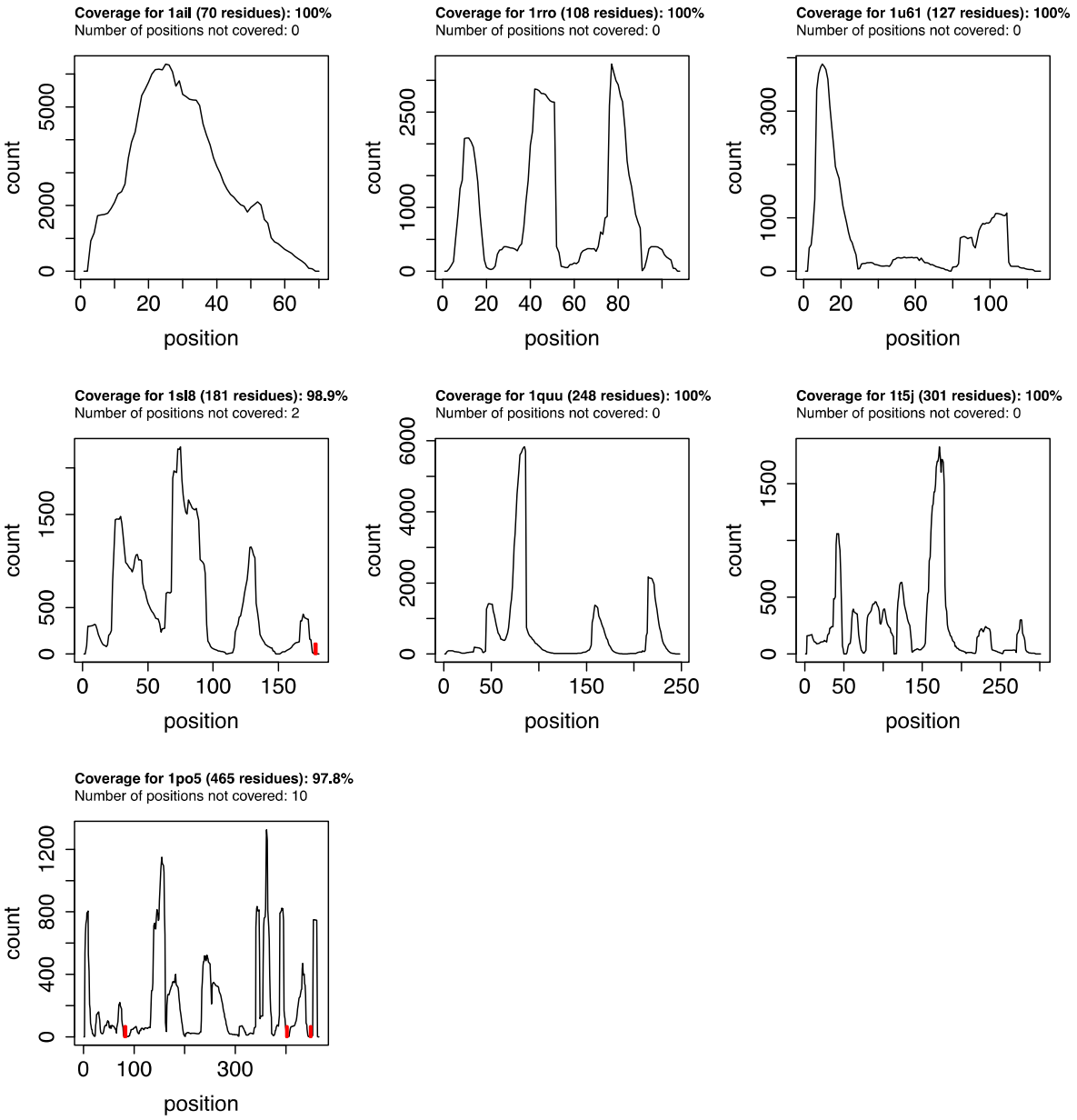
- Criss Jami, *Healology*

A Annexures

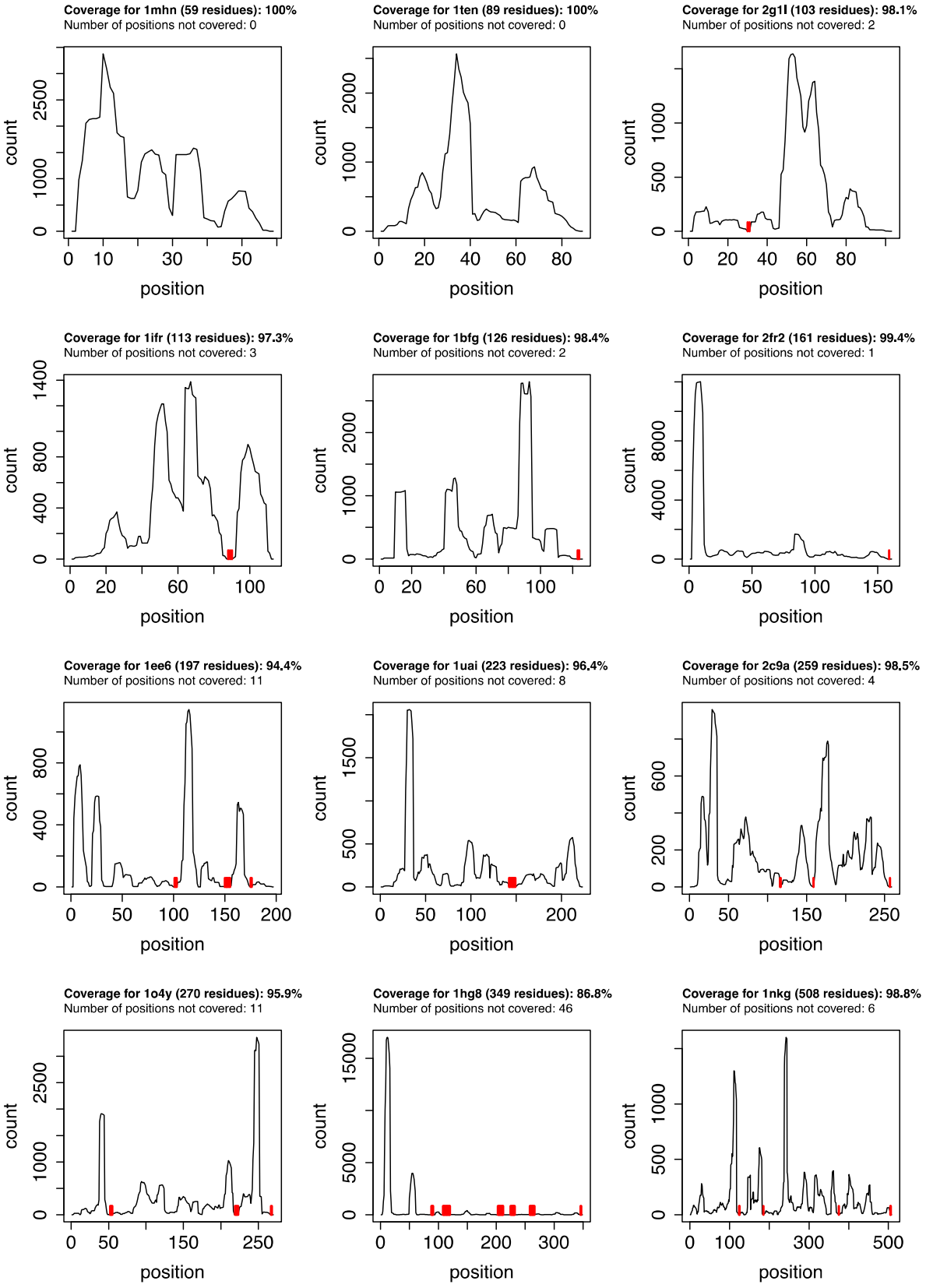
A.1 Chapter 2 - Supplementary files

Supplementary Figures 1. Coverage density plots.
The plots under this heading show the distribution of the number of fragments per position. The analysis has been performed for both treatment of the query proteins: (1) Protein Block Assignment (PBA) and (2) Protein Block Prediction (PBP). The graphs are further classified into section : (a) SCOP Class – all α, (b) SCOP Class – all β, (c) SCOP Class - α+ β and (d) SCOP Class - α/β . The percentage of coverage for each test protein is marked above the plot. Along with it the positions for each protein with no fragment hits are marked in red and the number of residues with no hits is also shown above each graph.

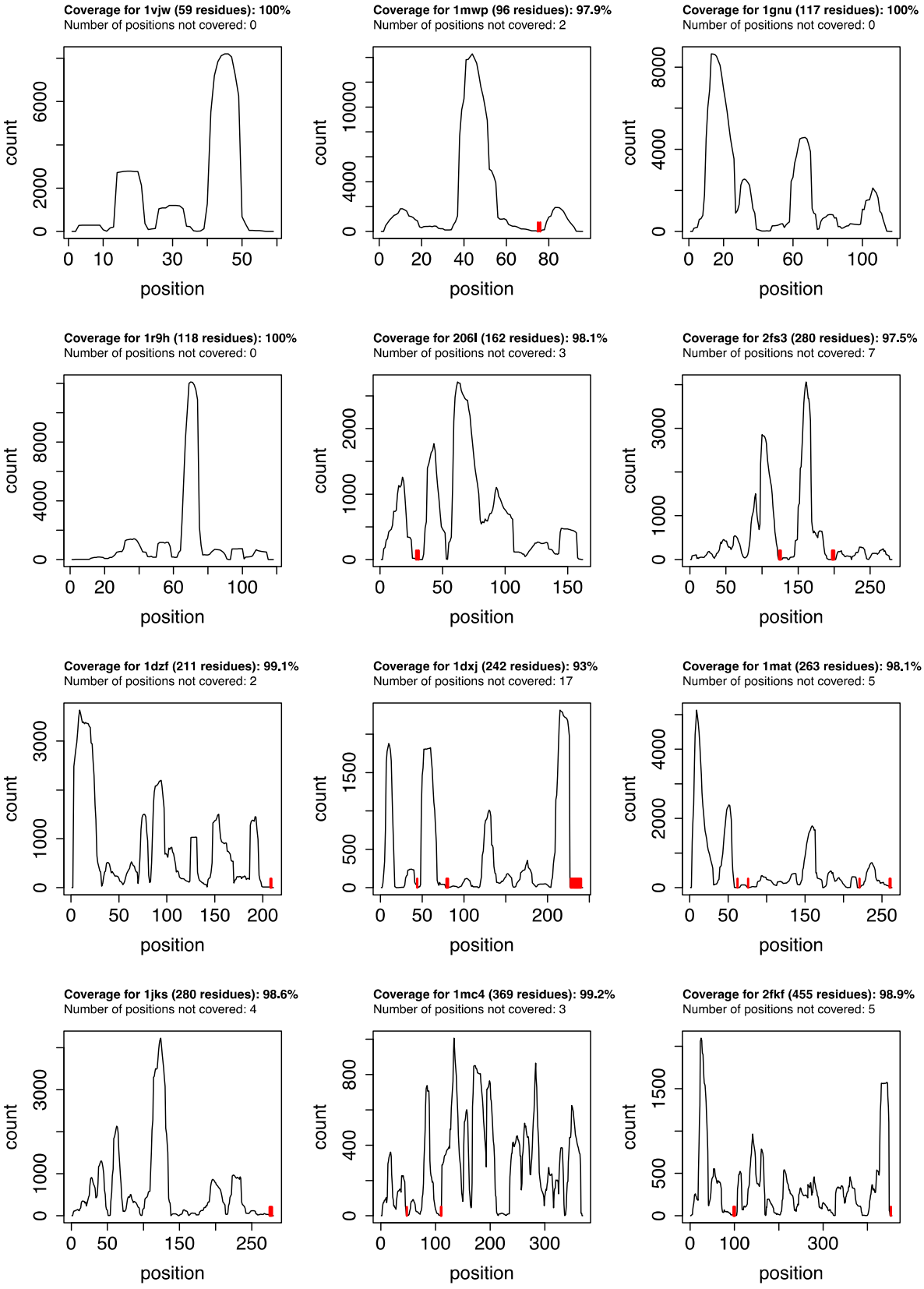
Supplementary Figure 1.1a
Protein Block Assignment (PBA) - SCOP Class – all α



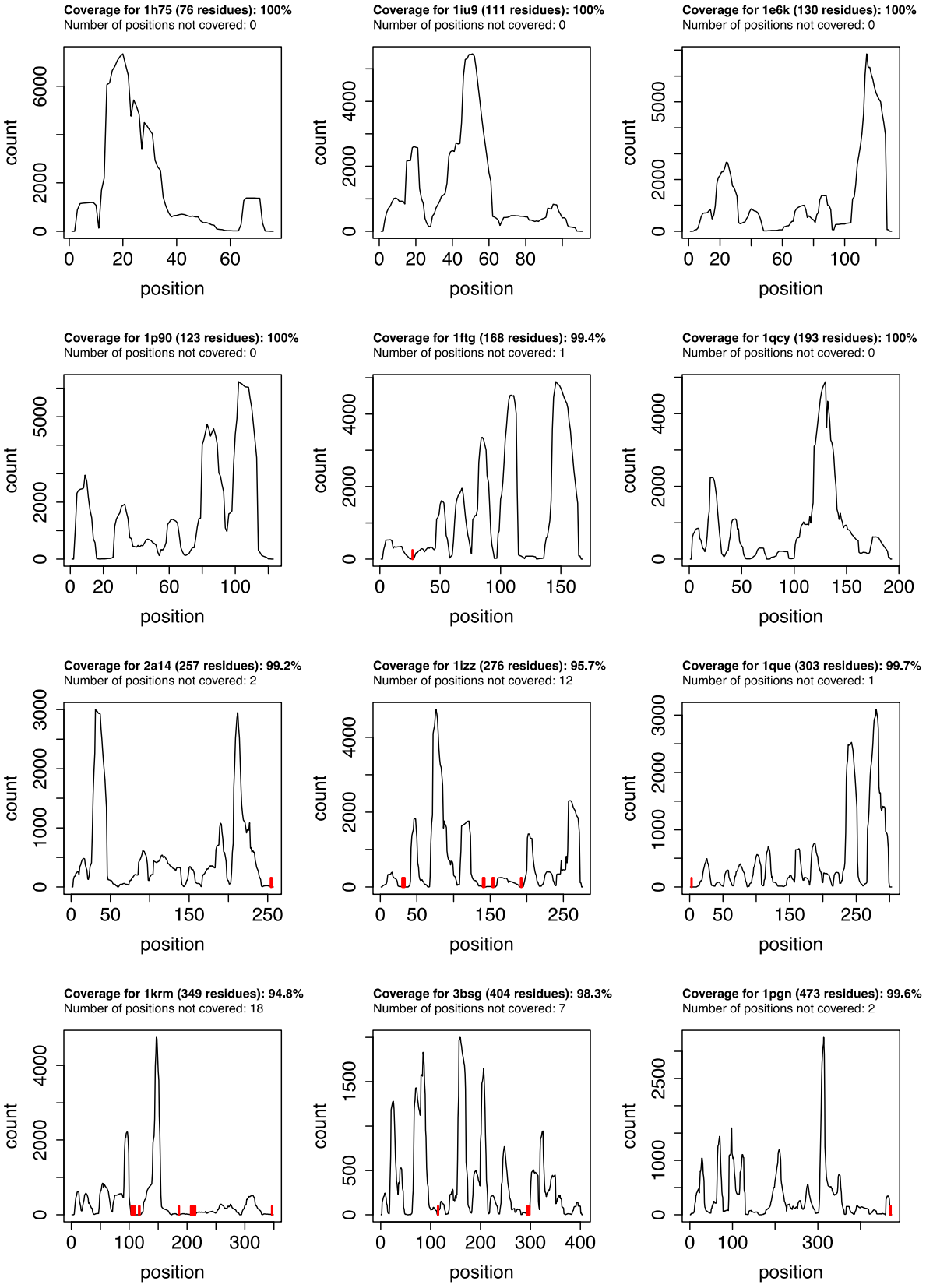
Supplementary Figure 1.1b
Protein Block Assignment (PBA) – SCOP Class – all β



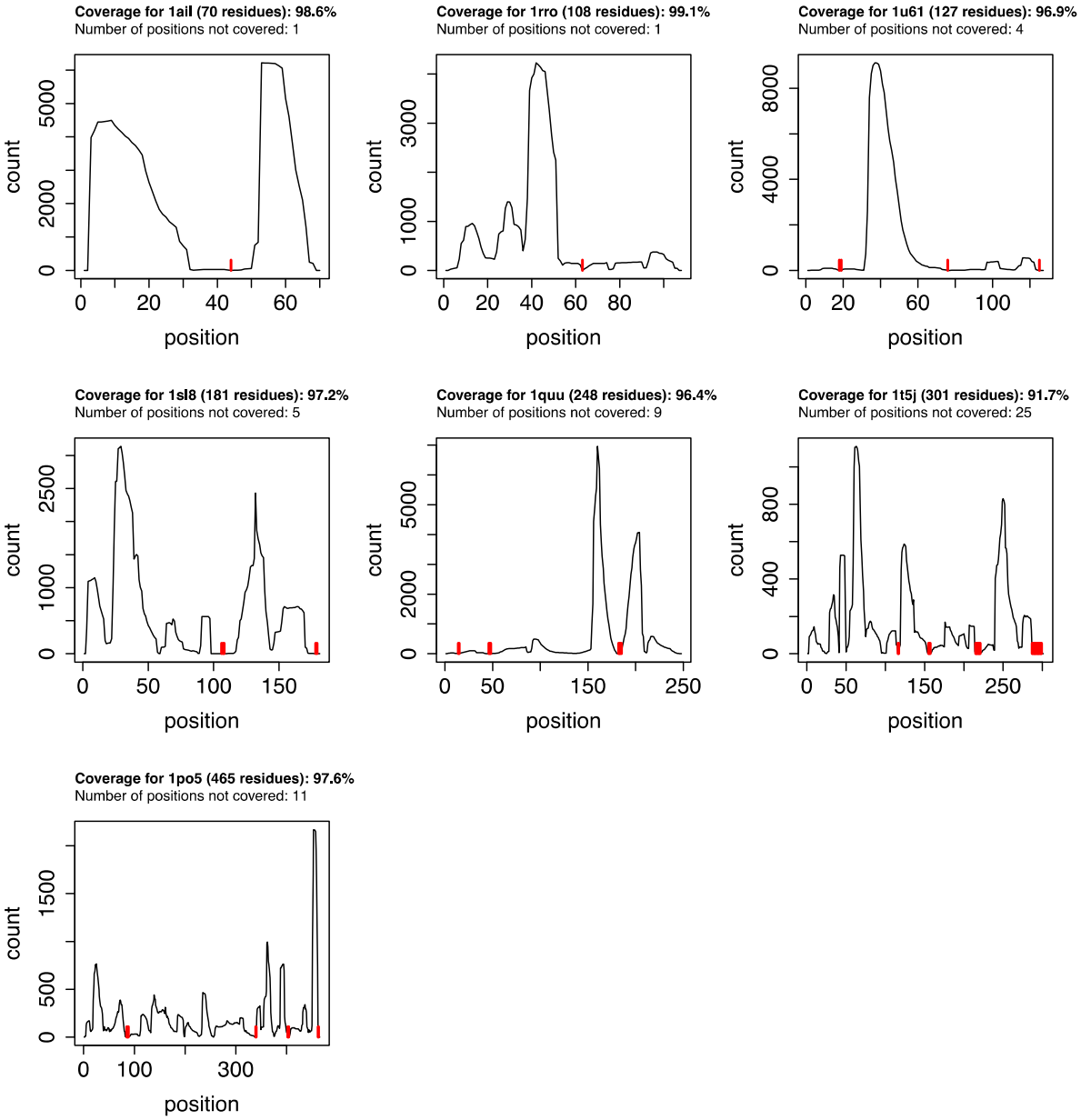
Supplementary Figure 1.1c
Protein Block Assignment (PBA) – SCOP Class – α+β



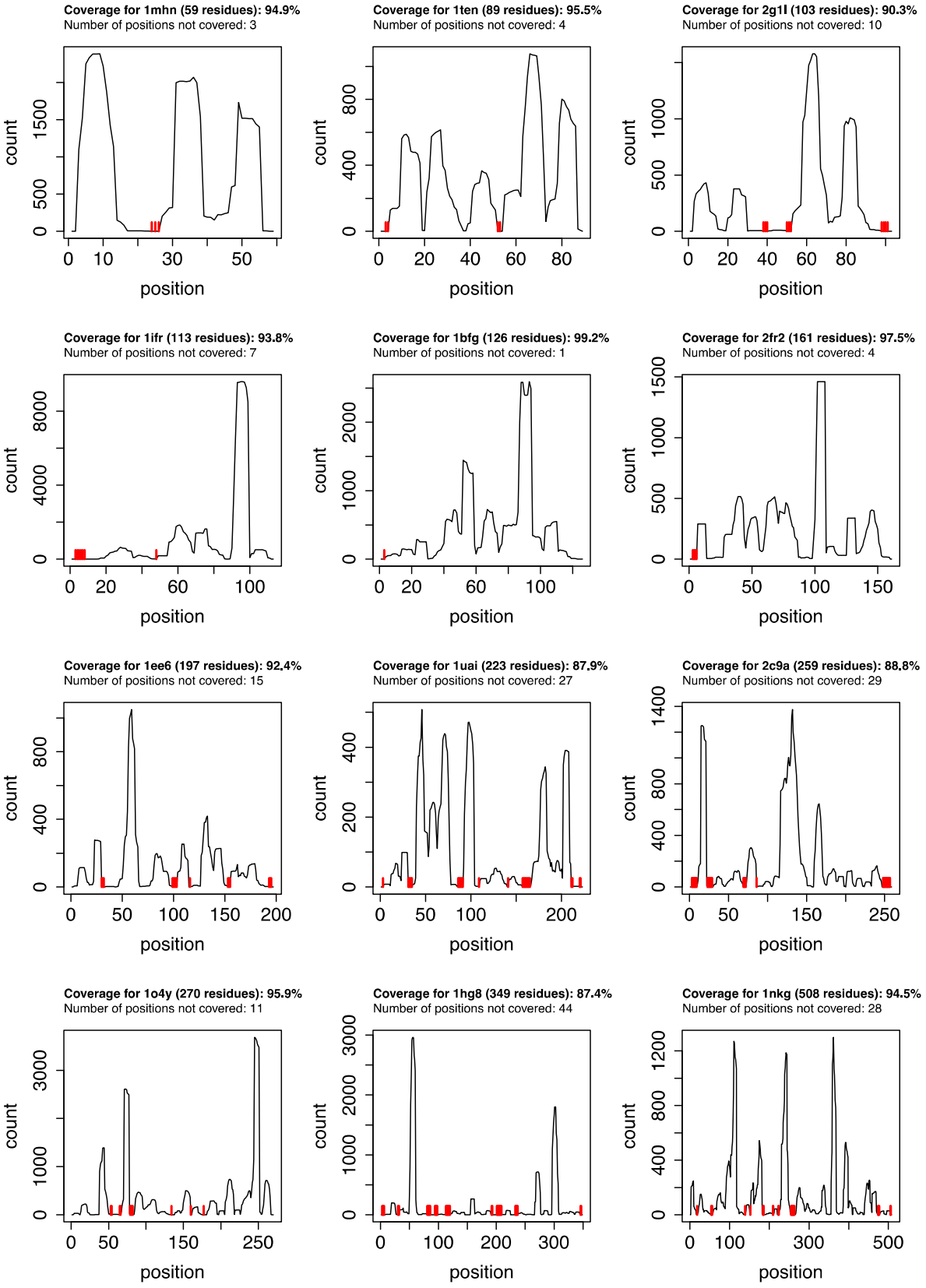
Supplementary Figure 1.1d
Protein Block Assignment (PBA) – SCOP Class – α/β



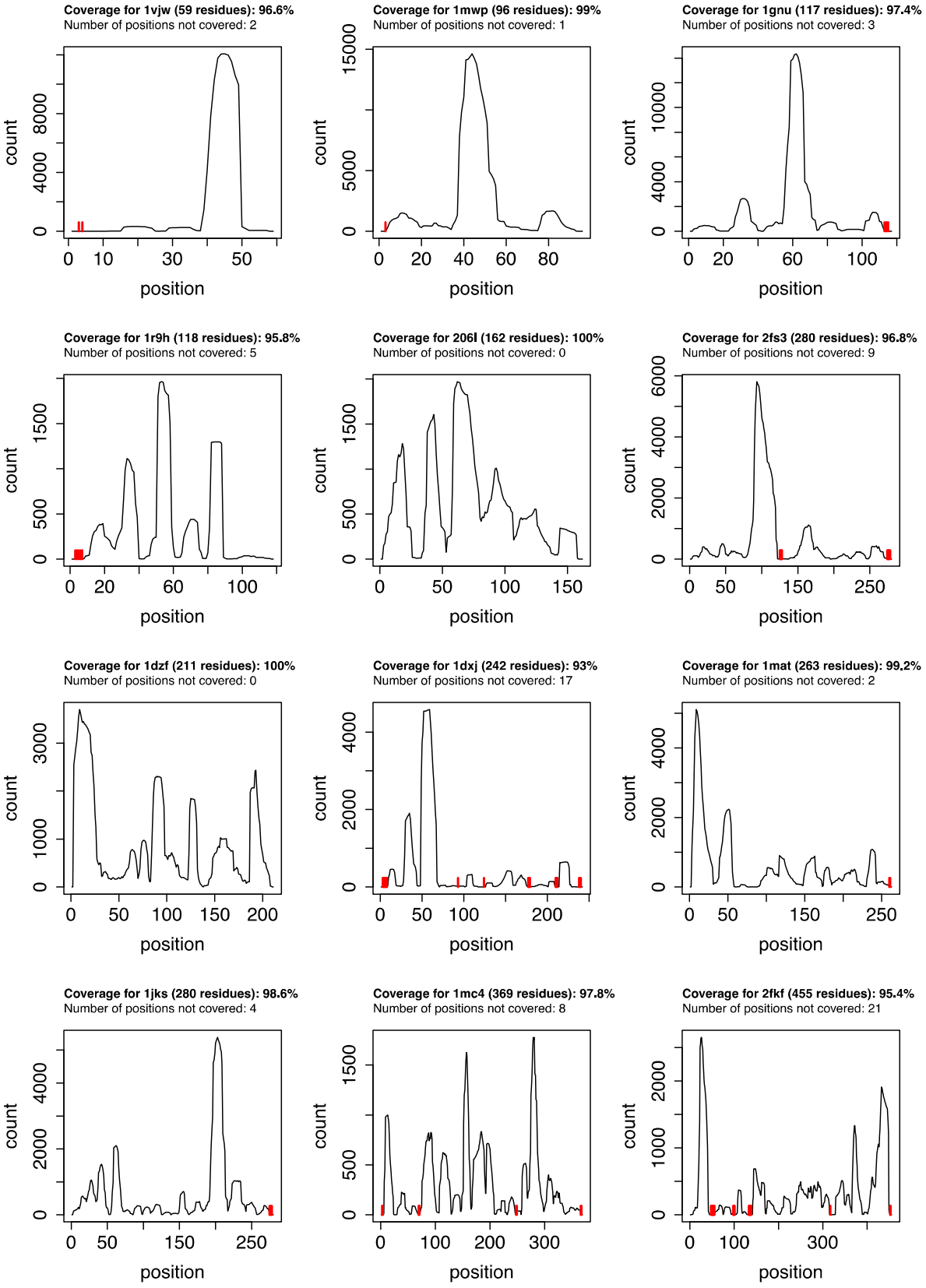
Supplementary Figure 1.2a
Protein Block Prediction (PBP) - SCOP Class – all α



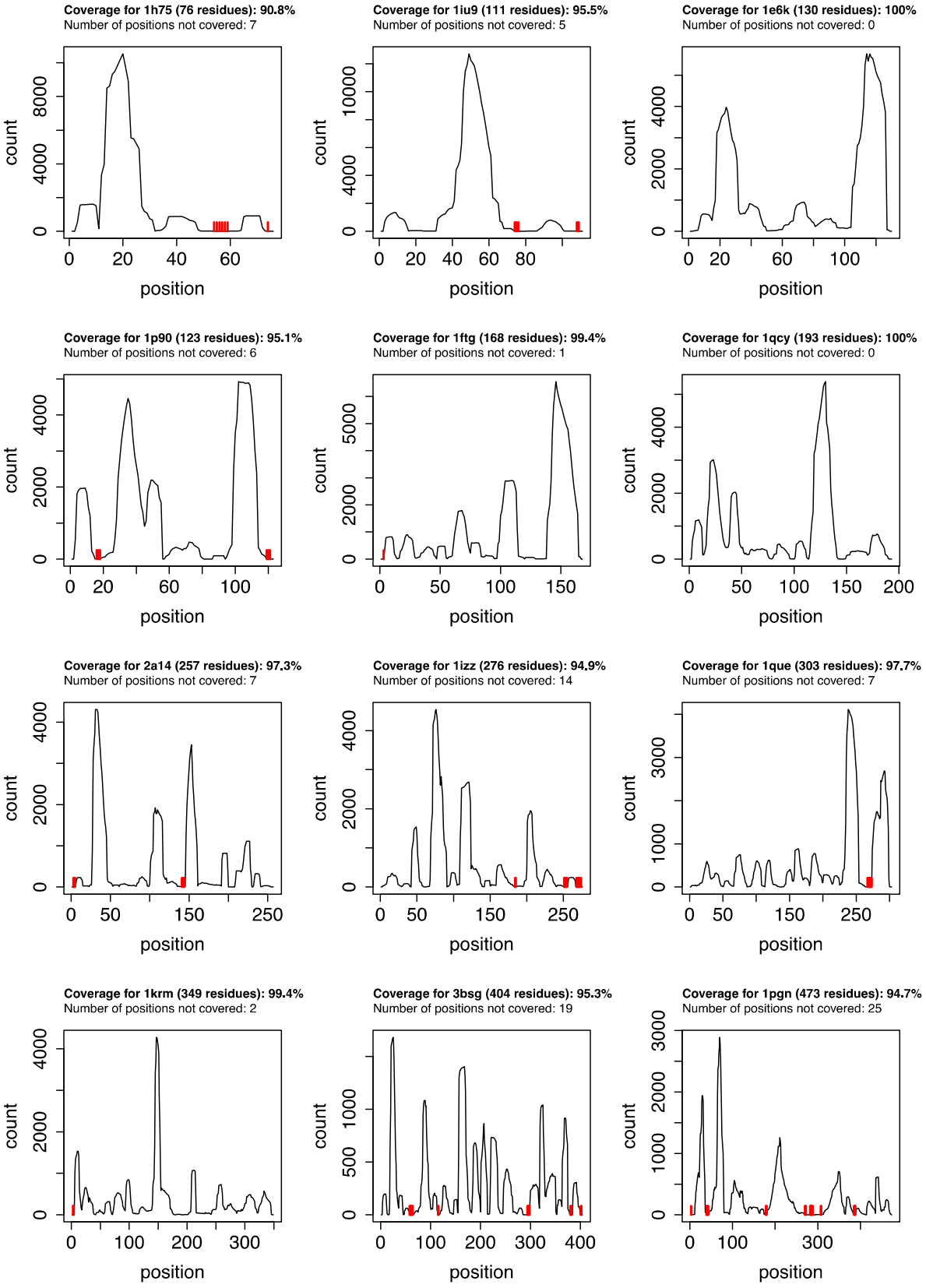
Supplementary Figure 1.2b
Protein Block Prediction (PBP) - SCOP Class – all β



Supplementary Figure 1.2c
Protein Block Prediction (PBP) - SCOP Class – α+β



Supplementary Figure 1.2d
Protein Block Prediction (PBP) - SCOP Class – α/β

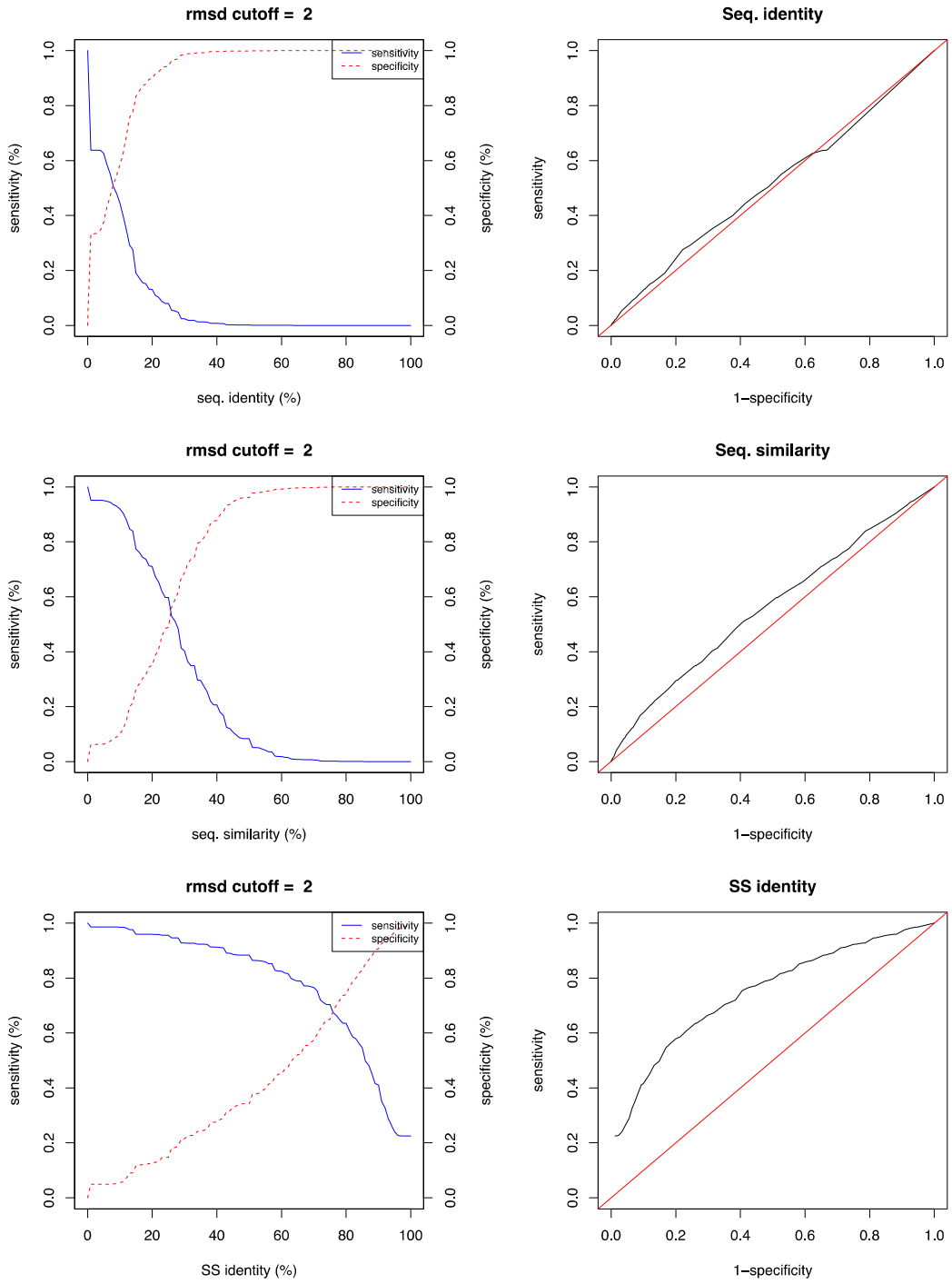


Supplementary Figures 2. Sensitivity and specificity Plots.

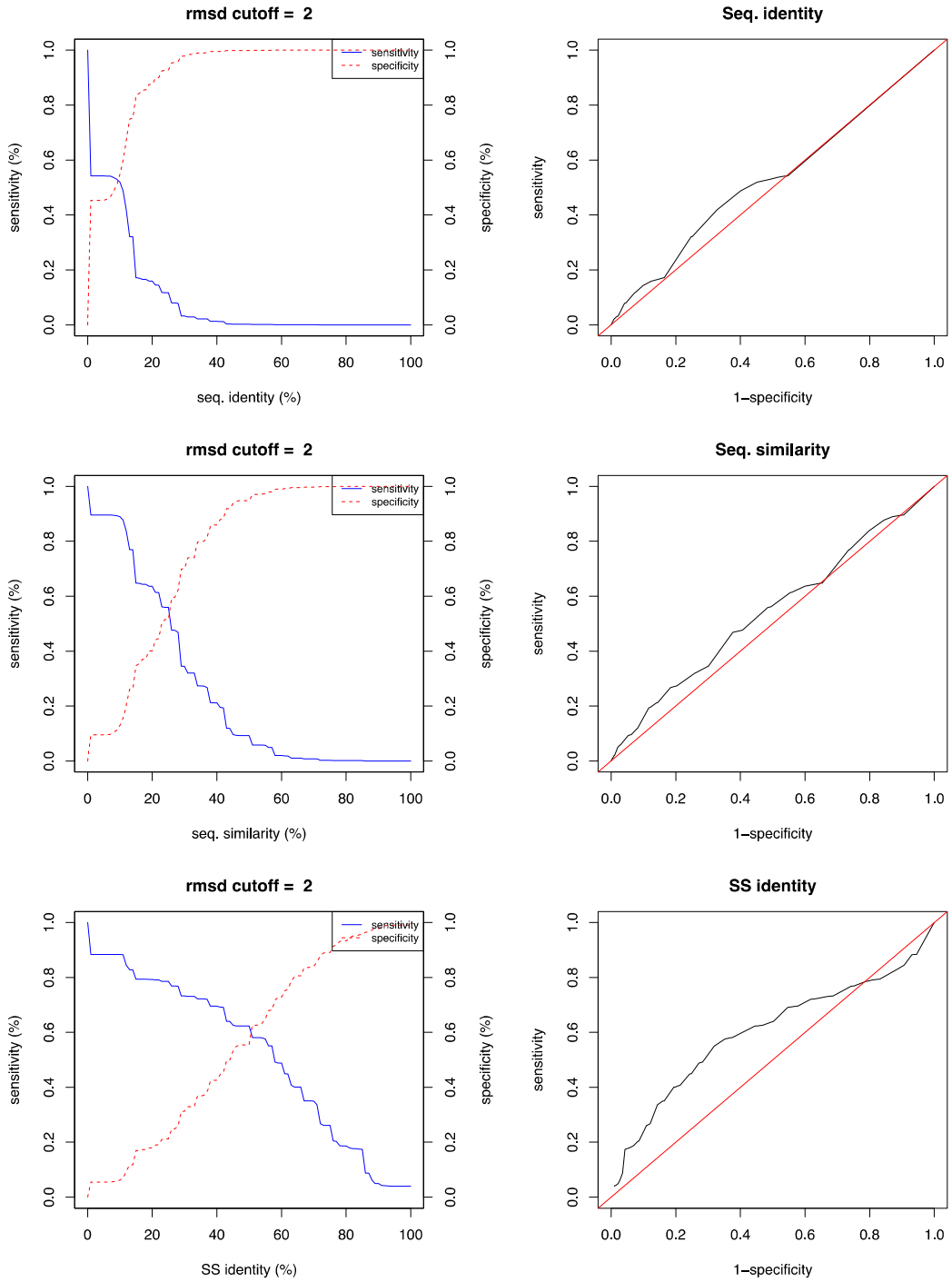
The plots under this heading show the co-relation between *rmsd* (2 Å) and three chosen criteria for prioritizing fragment selection, i.e., protein sequence identity, protein sequence similarity and secondary structure identity. The analysis has been performed for both treatment of the query proteins: (1) Protein Block Assignment (PBA) and (2) Protein Block Prediction (PBP). The graphs are further classified into section : (a) SCOP Class –all α, (b) SCOP Class – all β, (c) SCOP Class - α+β and (d) SCOP Class - α/β.

Supplementary Figure 2.1a

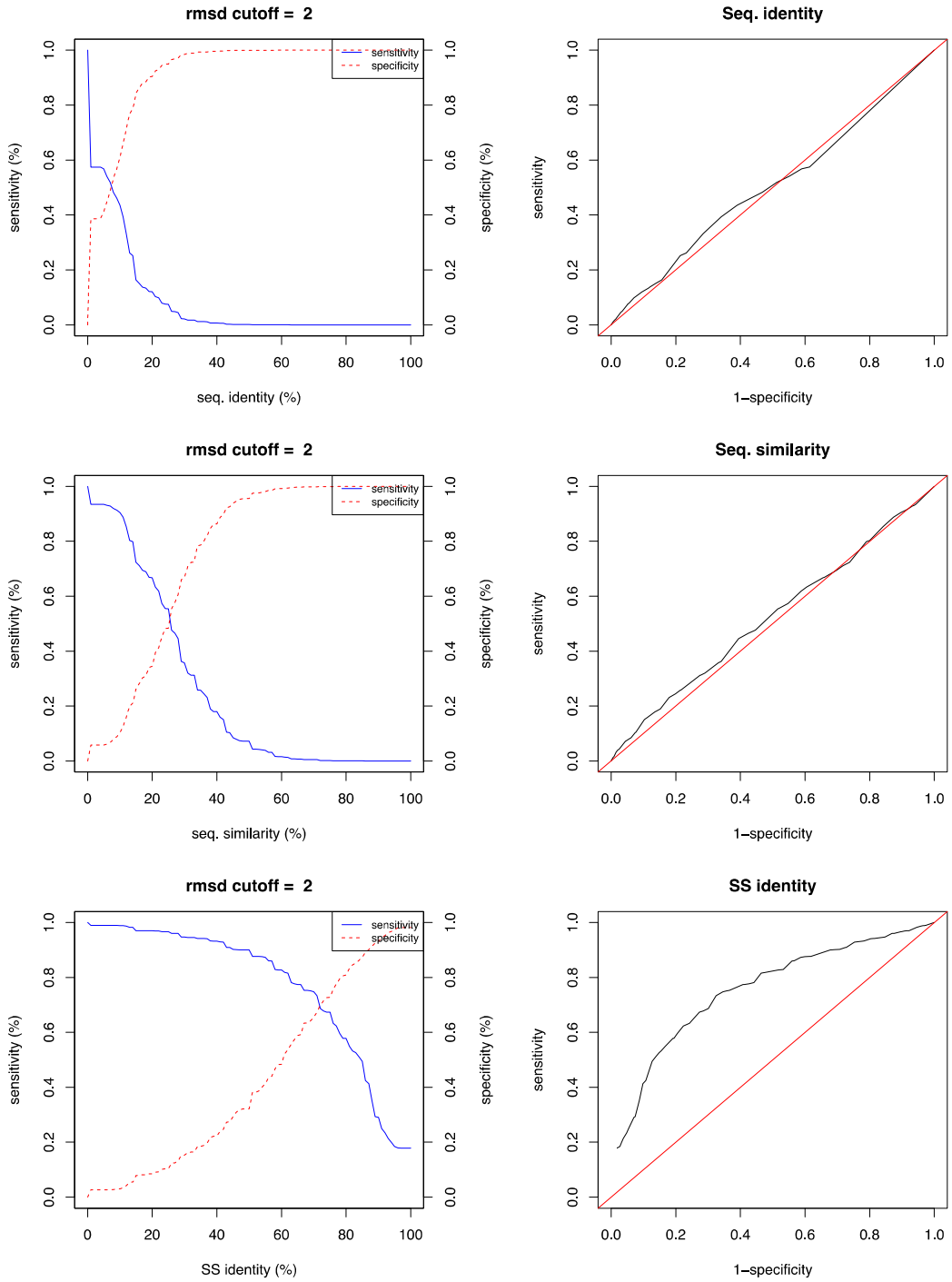
Protein Block Assignment (PBA) - SCOP Class – all α



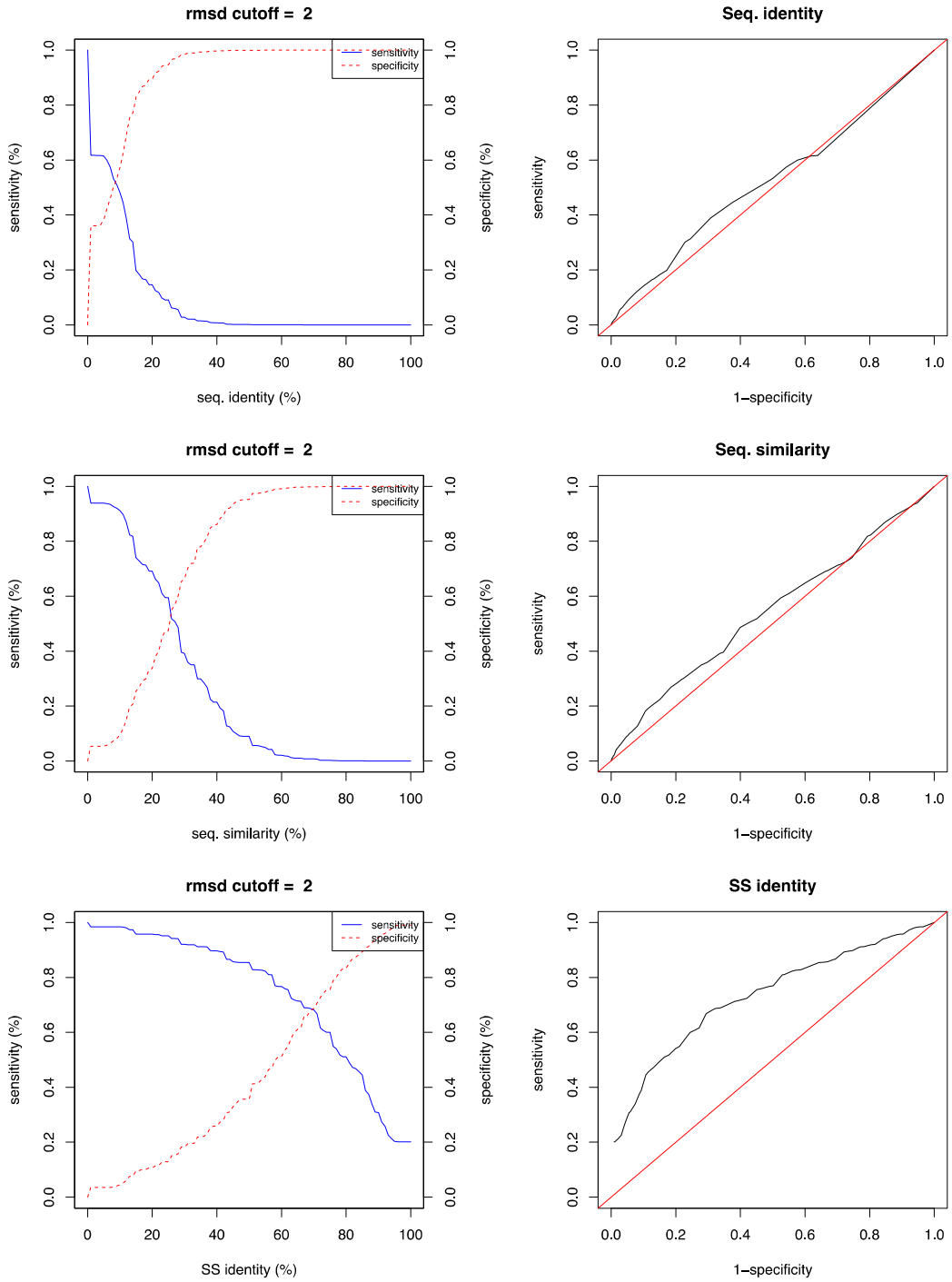
Supplementary Figure 2.1b
Protein Block Assignment (PBA) – SCOP Class – all β



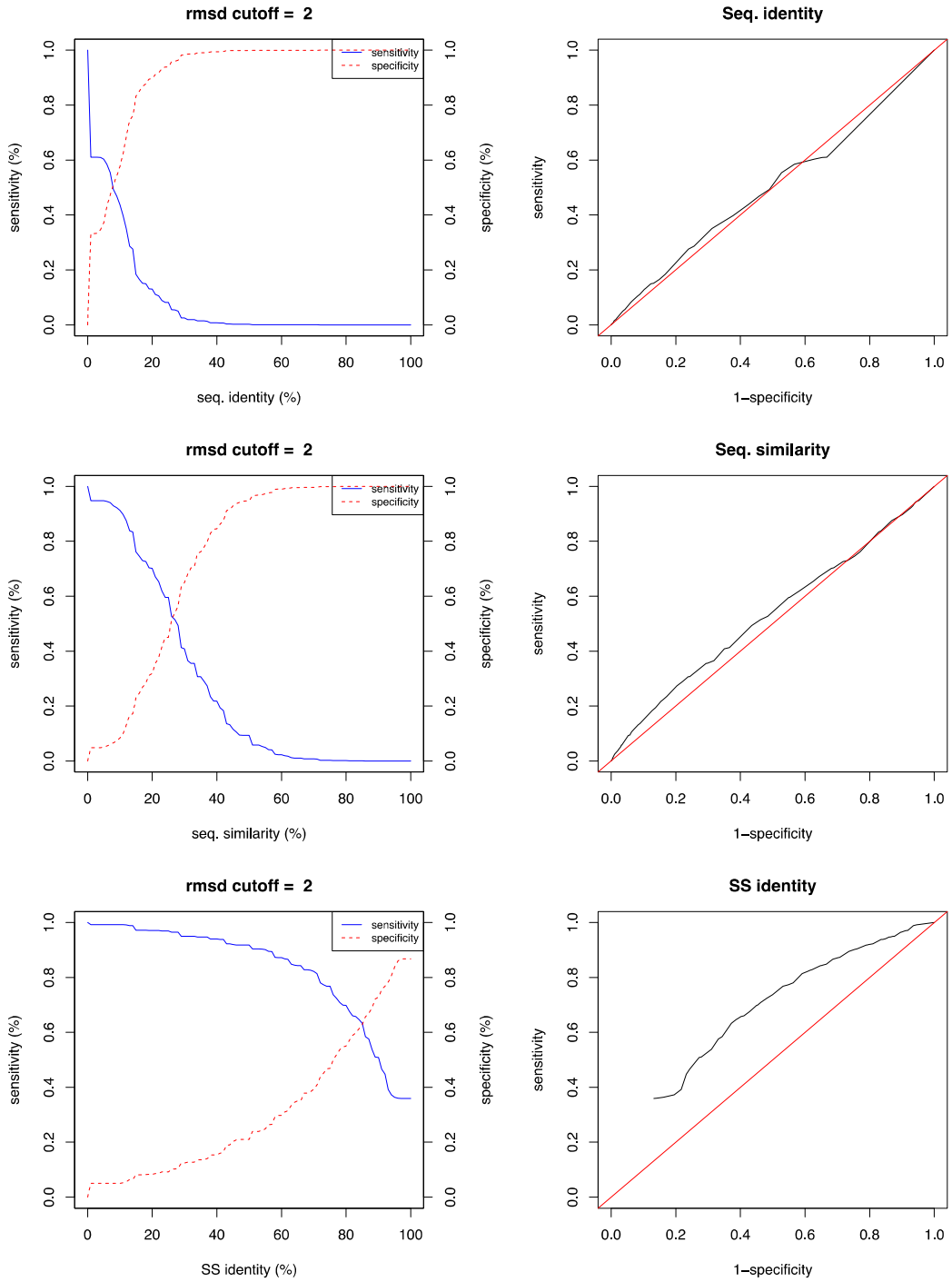
Supplementary Figure 2.1c
Protein Block Assignment (PBA) – SCOP Class – α+β



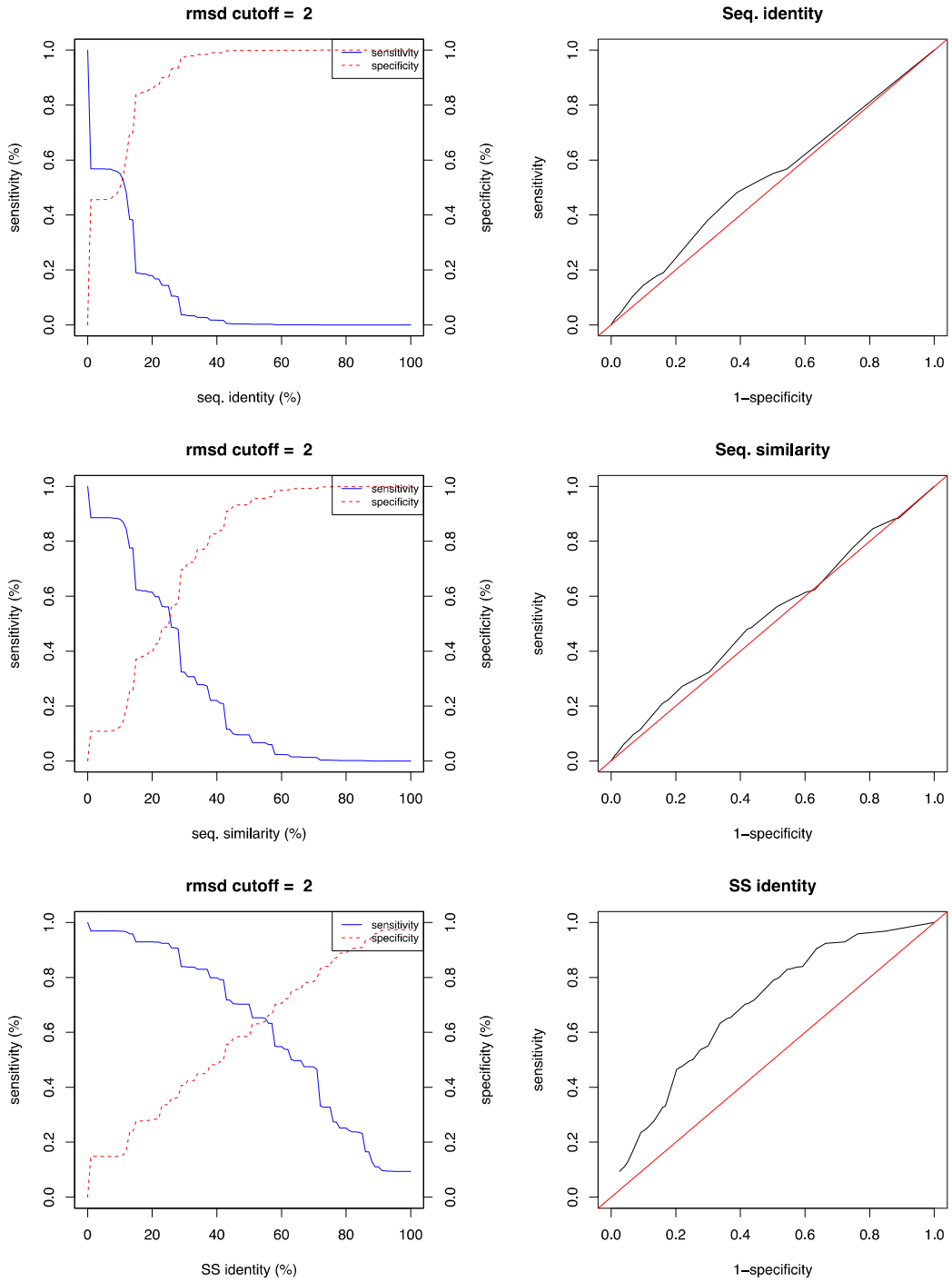
Supplementary Figure 2.1d
Protein Block Assignment (PBA) – SCOP Class – α/β



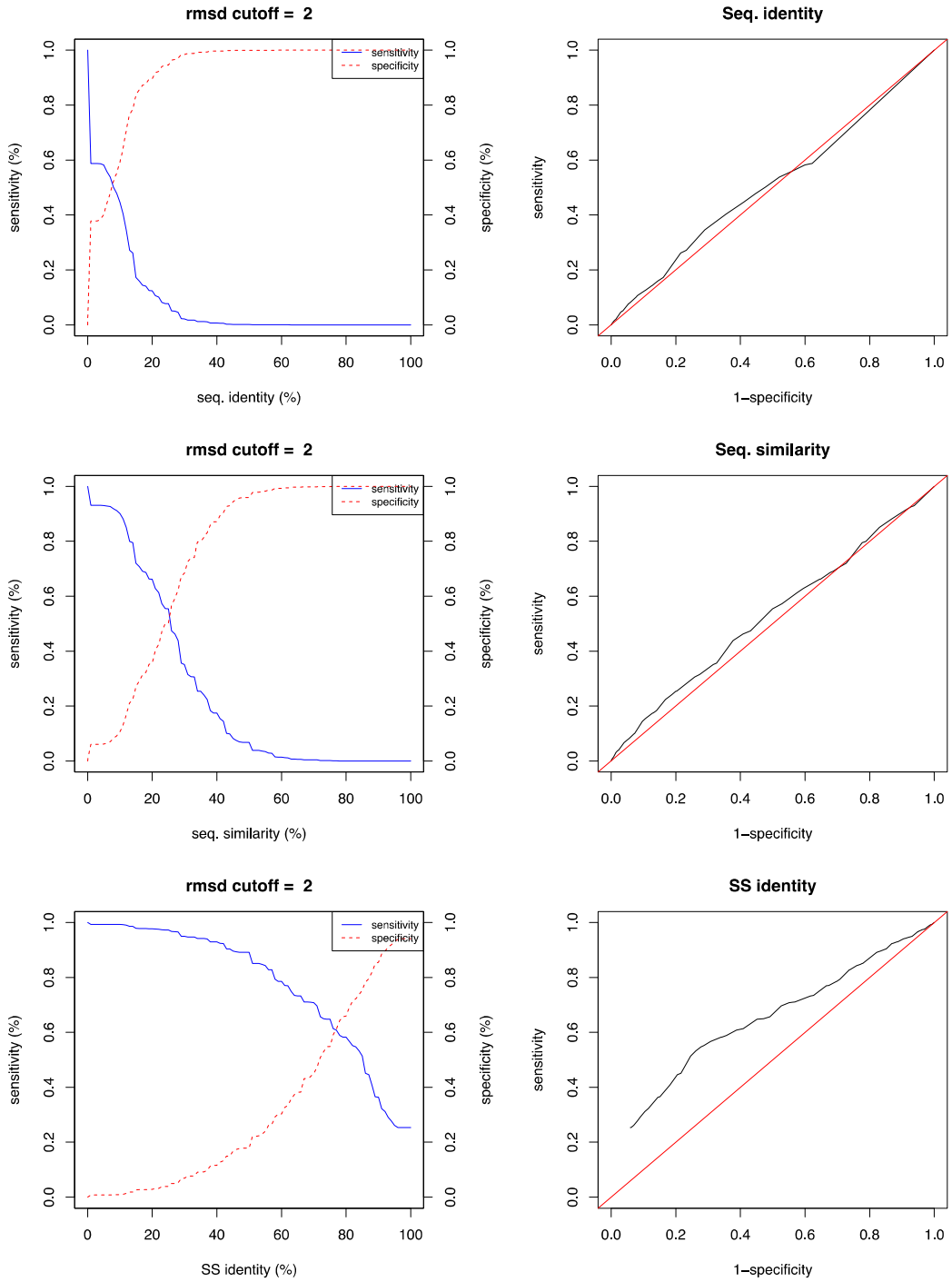
Supplementary Figure 2.2a
Protein Block Prediction (PBP) - SCOP Class – all α



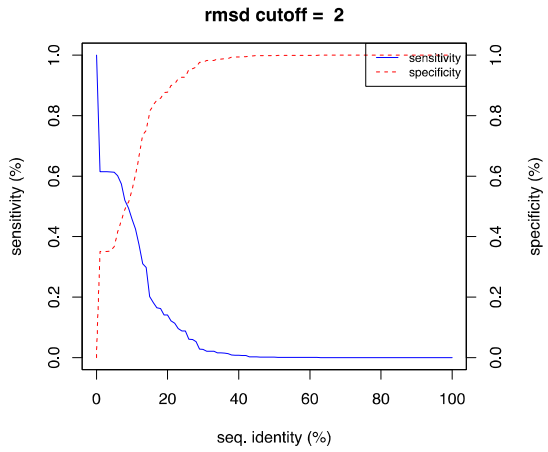
Supplementary Figure 2.2b
Protein Block Prediction (PBP) - SCOP Class – all β

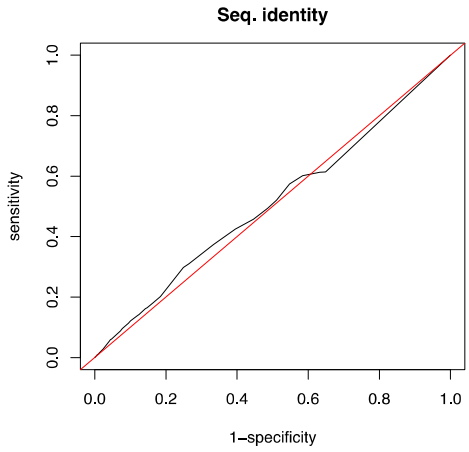


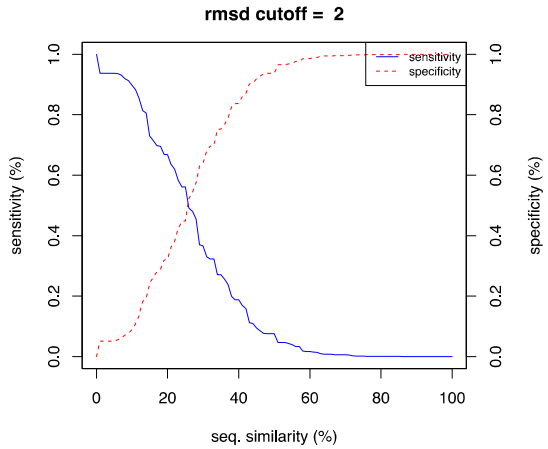
Supplementary Figure 2.2c
Protein Block Prediction (PBP) - SCOP Class – α+β

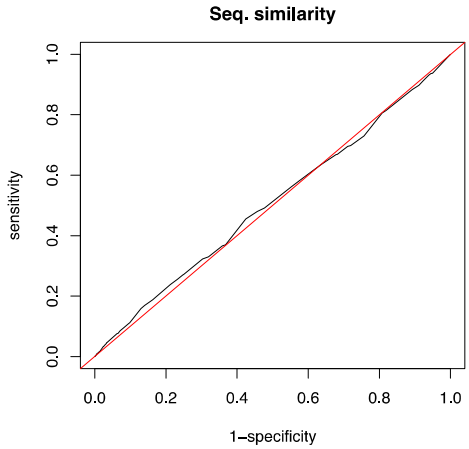


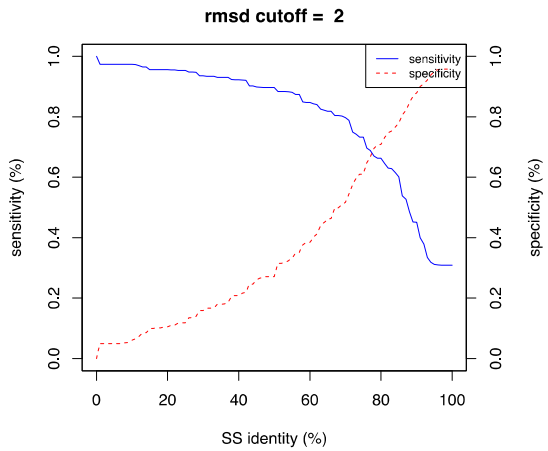
Supplementary Figure 2.2d
Protein Block Prediction (PBP) - SCOP Class – α/β

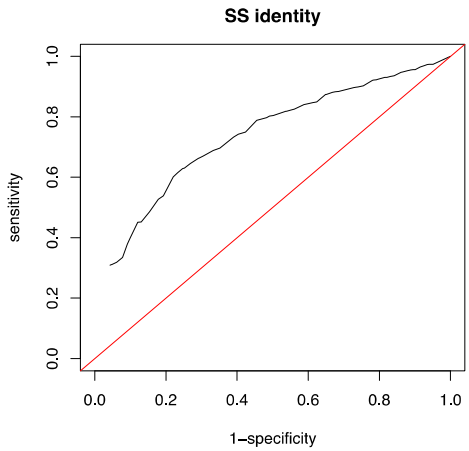












Supplementary Table 1. This table provides the fragment hit counts obtained after Protein Block Assignment (PBA) for each protein from the query dataset.

PB Assignment

	PDB ID	Length (AA)	Length (PDB)	All Hits	Identical PB Hits	Longest Hits	Coverage (%)
All Alpha	1AIL	73	70	34655	14402	150	100
	1RRO	108	108	49734	9728	415	100
	1U61	138	127	38640	6490	358	100
	1SL8	191	181	44102	8195	402	98.9
	1QUU	250	248	44824	11293	225	100
	1T5J	313	301	51463	6480	658	100
	1PO5	476	465	63837	9992	1569	97.8
All Beta	1MHN	59	59	45693	7763	172	100
	1TEN	90	89	41975	5675	232	100
	2G1L	104	103	44320	4135	343	98.1
	1IFR	121	113	43452	5637	270	97.3
	1BFG	146	126	57287	7341	343	98.4
	2FR2	172	161	78928	18226	392	99.4
	1EE6	197	197	42734	3761	471	94.4
	1UAI	224	223	45133	5718	437	96.4
	2C9A	259	259	44350	5378	717	98.5
	1O4Y	288	270	63734	10210	791	95.9
	1HG8	349	349	86866	22834	505	86.8
	1NKG	508	508	55584	8251	1841	98.8
Alpha and Beta	1VJW	60	59	72137	12525	235	100
	1MWP	96	96	73965	21732	461	97.9
	1GNU	117	117	73953	19932	583	100
	1R9H	135	118	76827	17704	558	100
	206L	164	162	55886	9290	861	98.1
	2FS3	282	280	64493	13195	674	97.5
	1DZF	215	211	70425	15982	850	99.1
	1DXJ	242	242	59053	8455	791	93
	1MAT	264	263	62926	13142	1024	98.1
	1JKS	294	280	66566	13929	885	98.6
	1MC4	370	369	61250	9888	739	99.2
	2FKF	462	455	61147	13037	1300	98.9
Alpha/Beta	1H75	81	76	69484	13346	299	100
	1IU9	111	111	66589	13346	342	100
	1E6K	130	130	71746	16321	621	100
	1P90	145	123	73913	20568	363	100
	1FTG	168	168	72925	18987	711	99.4
	1QCY	193	193	69619	15430	532	100
	2A14	263	257	64677	12697	835	99.2
	1IZZ	283	276	71107	16494	770	95.7
	1QUE	303	303	62765	12752	742	99.7
	1KRM	356	349	67668	12602	779	94.8
	3BSG	414	404	67000	14272	832	98.3
	1PGN	482	473	66032	14479	1019	99.6

Supplementary Table 2. This table provides the fragment hit counts obtained after Protein Block Prediction (PBP) for each protein from the query dataset.

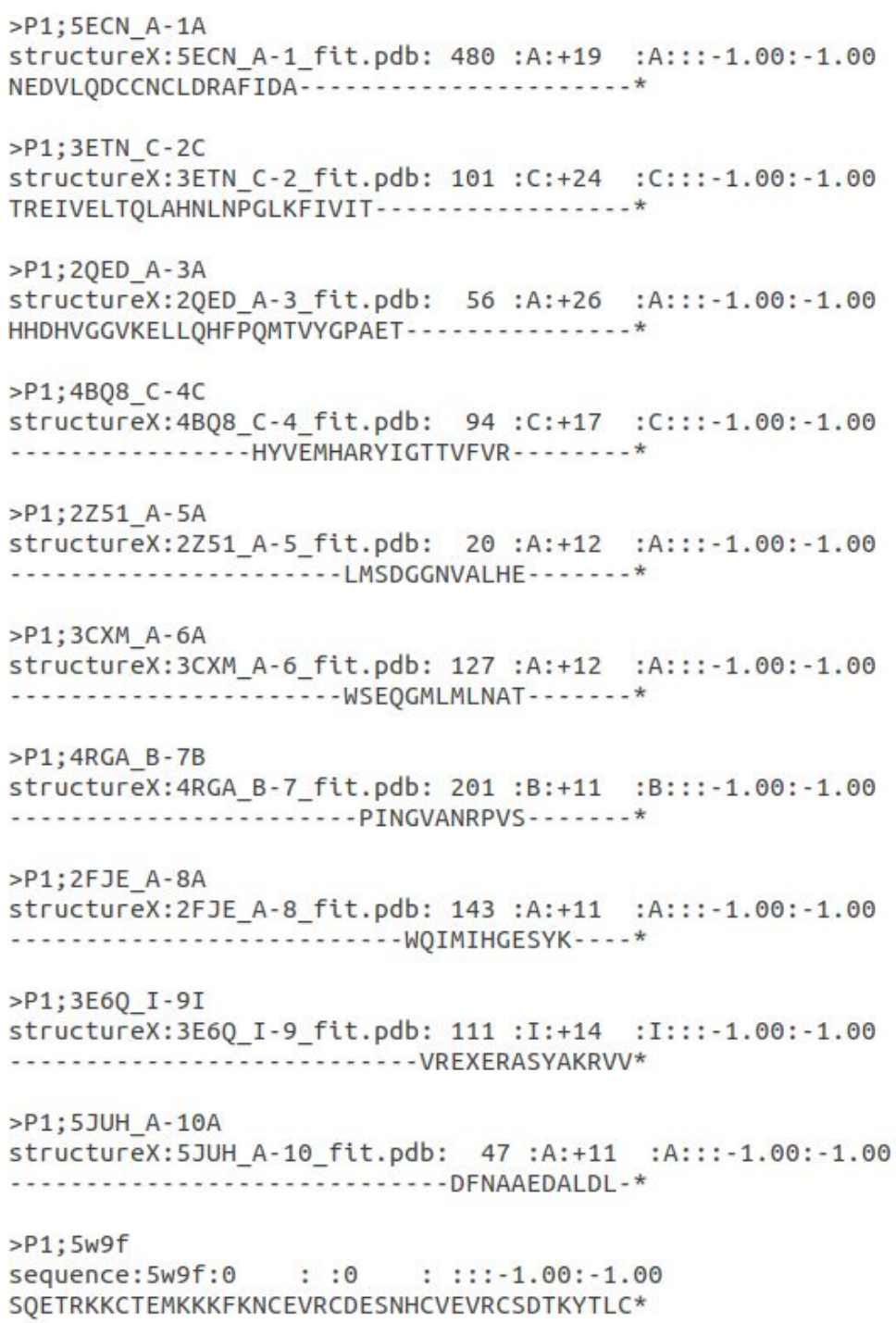
PB Prediction

	PDB ID	Length (AA)	Length (PDB)	All Hits	Identical PB Hits	Longest Hits	Coverage (%)
All Alpha	1AIL	73	70	38447	11501	414	98.6
	1RRO	108	108	40842	7661	451	99.1
	1U61	138	127	56844	10419	248	96.9
	1SL8	191	181	50381	10150	984	97.2
	1QUU	250	248	53552	12696	441	96.4
	1T5J	313	301	54456	4973	668	91.7
	1PO5	476	465	62554	9430	1591	97.6
All Beta	1MHN	59	59	70551	6379	380	94.9
	1TEN	90	90	69989	3775	859	95.5
	2G1L	104	103	45732	3721	317	90.3
	1IFR	121	113	66507	15170	946	93.8
	1BFG	146	126	53359	7121	482	99.2
	2FR2	172	161	60031	4934	2542	97.5
	1EE6	197	197	49351	3073	792	92.4
	1UAI	224	223	47480	2926	740	87.9
	2C9A	259	259	59190	5551	705	88.8
	1O4Y	288	270	67912	12023	1869	95.9
	1HG8	349	349	50990	6720	659	87.4
	1NKG	508	508	56424	8486	1530	94.5
Alpha+Beta	1VJW	60	59	61326	12716	157	96.6
	1MWP	96	96	73836	21621	308	99
	1GNU	117	117	82325	21173	780	97.4
	1R9H	135	118	51549	5452	636	95.8
	206L	164	162	53082	8458	736	100
	2FS3	282	280	64098	12754	1236	96.8
	1DZF	215	211	70028	17204	824	100
	1DXJ	242	242	63698	9378	500	93
	1MAT	264	263	63583	13768	909	99.2
	1JKS	294	280	69244	16160	1134	98.6
	1MC4	370	369	66492	11564	1209	97.8
	2FKF	462	455	64106	14040	1236	95.4
Alpha/Beta	1H75	81	76	72386	14483	729	90.8
	1IU9	111	111	66329	18860	1258	95.5
	1E6K	130	130	72078	15837	1205	100
	1P90	145	123	74454	15671	2383	95.1
	1FTG	168	168	70585	14923	587	99.4
	1QCY	193	193	69674	15537	687	100
	2A14	263	257	66681	14408	2781	97.3
	1IZZ	283	276	71995	14779	1563	94.9
	1QUE	303	303	64904	14981	889	97.7
	1KRM	356	349	67838	13370	605	99.4
	3BSG	414	404	67614	11980	1006	95.3
	1PGN	482	473	65827	11766	1414	94.7

A.2 Chapter 4 - Supplementary files

Supplementary Figure 1

Image of the alignment file generated for the test protein 5w9f.



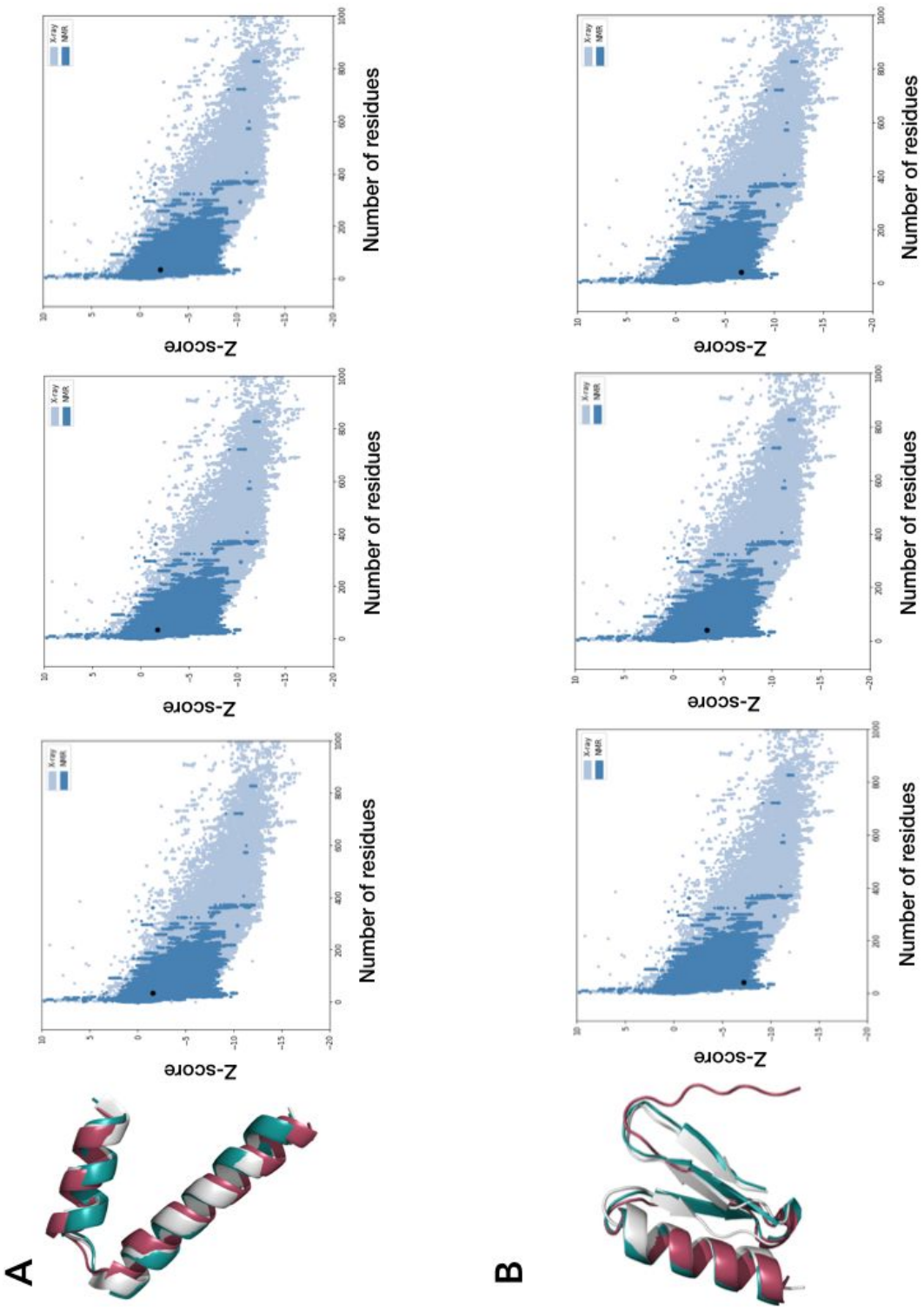
```
>P1;5ECN_A-1A
structureX:5ECN_A-1_fit.pdb: 480 :A:+19  :A:::-1.00:-1.00
NEDVLQDCCNCLDRAFIDA---------------------*

>P1;3ETN_C-2C
structureX:3ETN_C-2_fit.pdb: 101 :C:+24  :C:::-1.00:-1.00
TREIVELTQLAHNLNPGLKFIVIT-----------------*

>P1;2QED_A-3A
structureX:2QED_A-3_fit.pdb:  56 :A:+26  :A:::-1.00:-1.00
HHDHVGGVKELLQHFPQMTVYGPAET---------------*

>P1;4BQ8_C-4C
structureX:4BQ8_C-4_fit.pdb:  94 :C:+17  :C:::-1.00:-1.00
----------------HYVEMHARYIGTTVFVR--------*

>P1;2Z51_A-5A
structureX:2Z51_A-5_fit.pdb:  20 :A:+12  :A:::-1.00:-1.00
----------------------LMSDGGNVALHE-------*

>P1;3CXM_A-6A
structureX:3CXM_A-6_fit.pdb: 127 :A:+12  :A:::-1.00:-1.00
----------------------WSEQGMLMLNAT-------*

>P1;4RGA_B-7B
structureX:4RGA_B-7_fit.pdb: 201 :B:+11  :B:::-1.00:-1.00
-----------------------PINGVANRPVS-------*

>P1;2FJE_A-8A
structureX:2FJE_A-8_fit.pdb: 143 :A:+11  :A:::-1.00:-1.00
-------------------------WQIMIHGESYK----*

>P1;3E6Q_I-9I
structureX:3E6Q_I-9_fit.pdb: 111 :I:+14  :I:::-1.00:-1.00
--------------------------VREXERASYAKRVV*

>P1;5JUH_A-10A
structureX:5JUH_A-10_fit.pdb:  47 :A:+11  :A:::-1.00:-1.00
-----------------------------DFNAAEDALDL-*

>P1;5w9f
sequence:5w9f:0    : :0    : :::-1.00:-1.00
SQETRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLC*
```

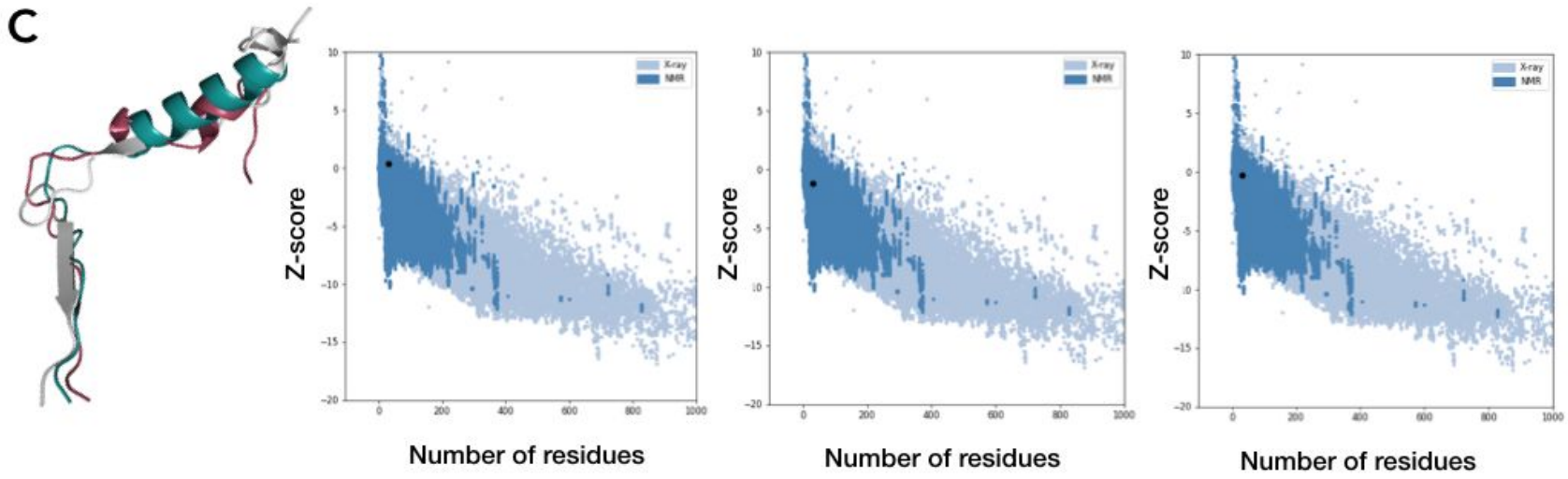
Supplementary Figure 2

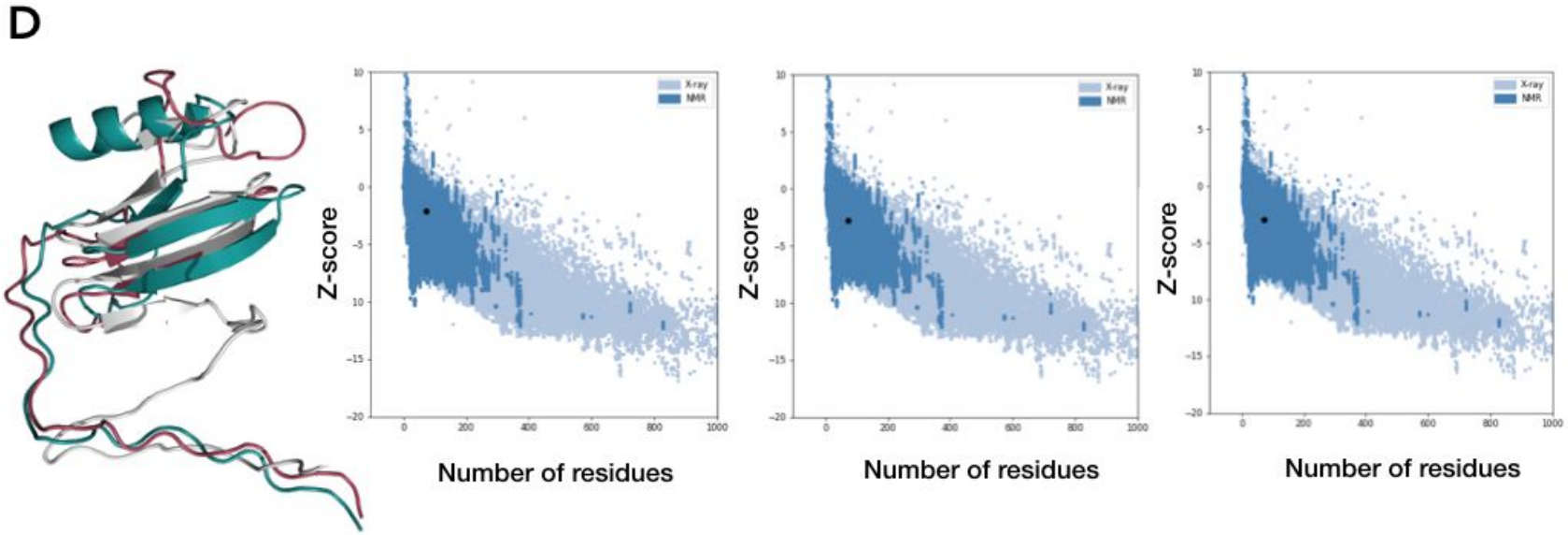
Image of the alignment file generated for the test protein 6gnx.

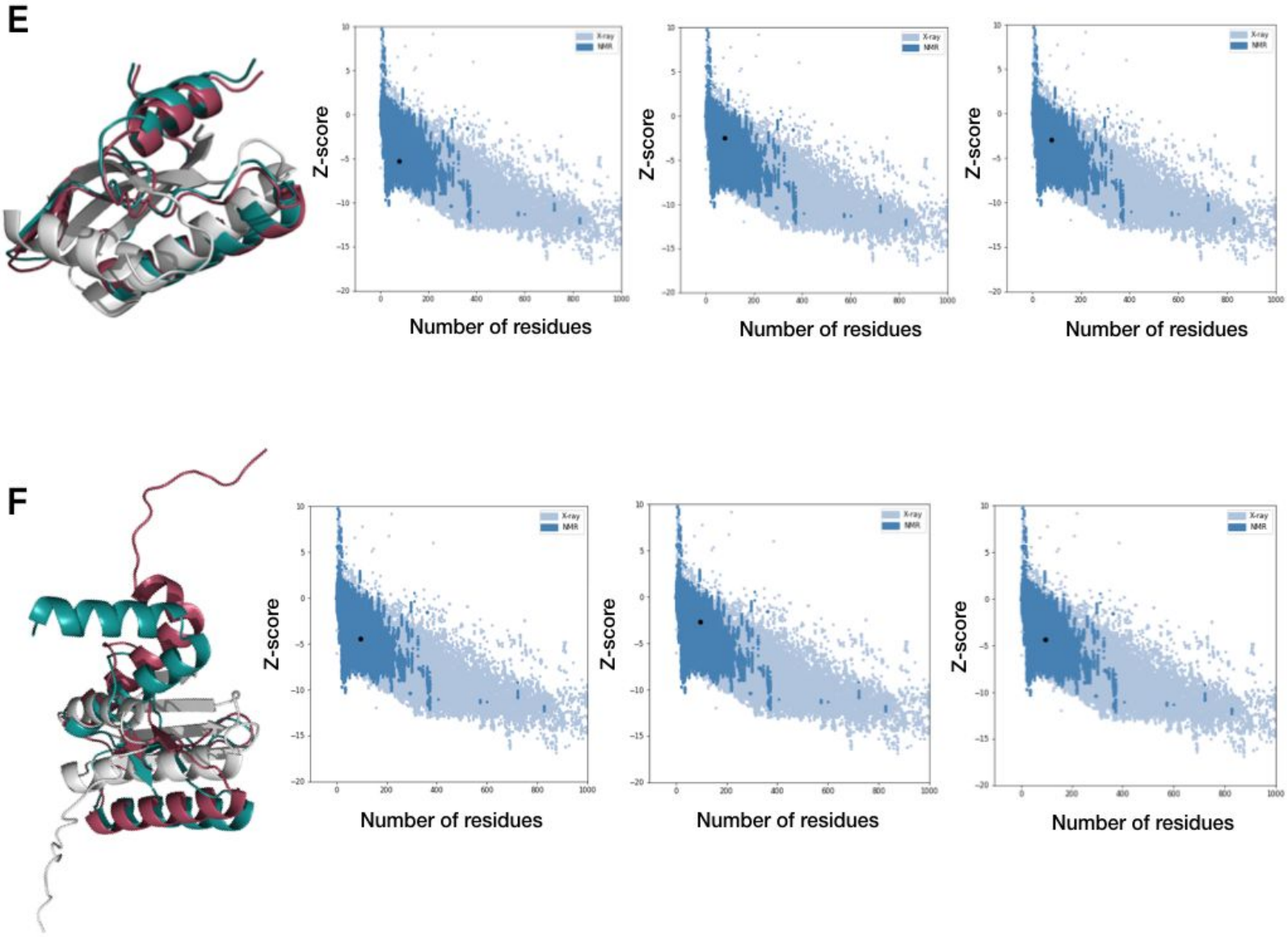
```
>P1;6DFP_A-1A
structureX:6DFP_A-1_fit.pdb: 228 :A:+12  :A:::-1.00:-1.00
AVGAASIGVTWG-------------------*

>P1;1ION_A-2A
structureX:1ION_A-2_fit.pdb: 191 :A:+11  :A:::-1.00:-1.00
-------VPLLAVIPEDP-------------*

>P1;3C7G_A-3A
structureX:3C7G_A-3_fit.pdb: 362 :A:+13  :A:::-1.00:-1.00
-------ETFAWNGRILTEK-----------*

>P1;6GNU_A-4A
structureX:6GNU_A-4_fit.pdb: 383 :A:+15  :A:::-1.00:-1.00
-------GAGDGHLRYYFYNWA---------*

>P1;4HXF_B-5B
structureX:4HXF_B-5_fit.pdb:  51 :B:+15  :B:::-1.00:-1.00
-------------SRKFIEDASMPRISP---*

>P1;5Z5I_A-6A
structureX:5Z5I_A-6_fit.pdb: 172 :A:+13  :A:::-1.00:-1.00
----------------GGSYPEAPHLYKV--*

>P1;5ZQJ_A-7A
structureX:5ZQJ_A-7_fit.pdb: 182 :A:+12  :A:::-1.00:-1.00
-----------------IKLTEAPHLYYI--*

>P1;1D4T_A-8A
structureX:1D4T_A-8_fit.pdb:  45 :A:+11  :A:::-1.00:-1.00
------------------VLYHGYIYTYR--*

>P1;1V5I_B-9B
structureX:1V5I_B-9_fit.pdb:  36 :B:+13  :B:::-1.00:-1.00
------------------YNMPGMKGFAGEL*

>P1;1M1F_B-10B
structureX:1M1F_B-10_fit.pdb:   1 :B:+11  :B:::-1.00:-1.00
-------------------MERGEIWLVSL-*

>P1;4KMR_A-11A
structureX:4KMR_A-11_fit.pdb:   1 :A:+11  :A:::-1.00:-1.00
--------------------GRSGVIGLAVP-*

>P1;4BLW_B-12B
structureX:4BLW_B-12_fit.pdb: 215 :B:+12  :B:::-1.00:-1.00
-------------------GIRKILQIELAV*

>P1;5LIR_A-13A
structureX:5LIR_A-13_fit.pdb:   1 :A:+11  :A:::-1.00:-1.00
--------------------FQGKKIGVVLS*

>P1;6gnx
sequence:6gnx:0    : :0    : :::-1.00:-1.00
GYISIDAKKFLGELHDFIPGTSGYLAYHVQN*
```

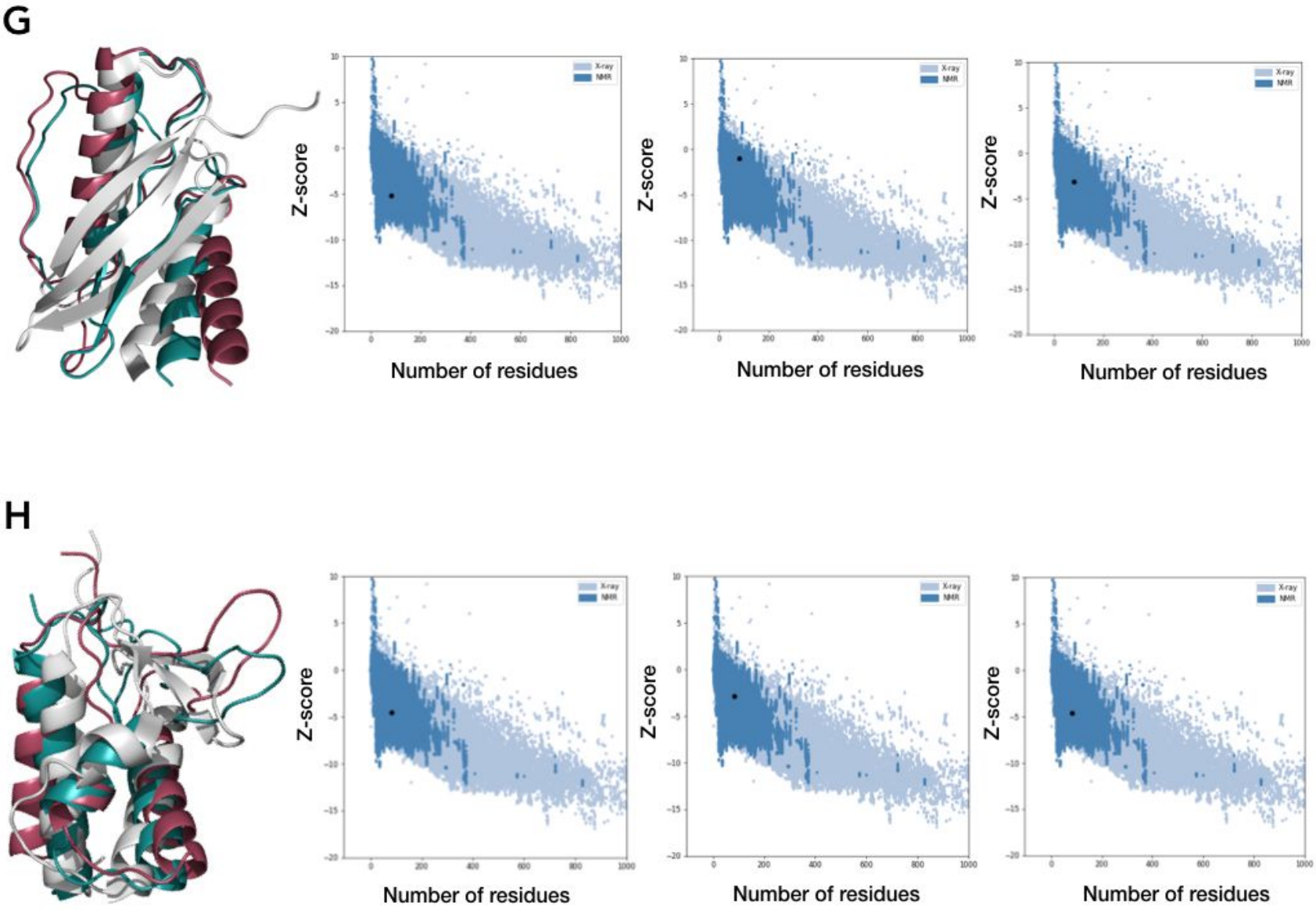
Supplementary Figure 3. Z-score calculations for the models designed for each protein in the query dataset.



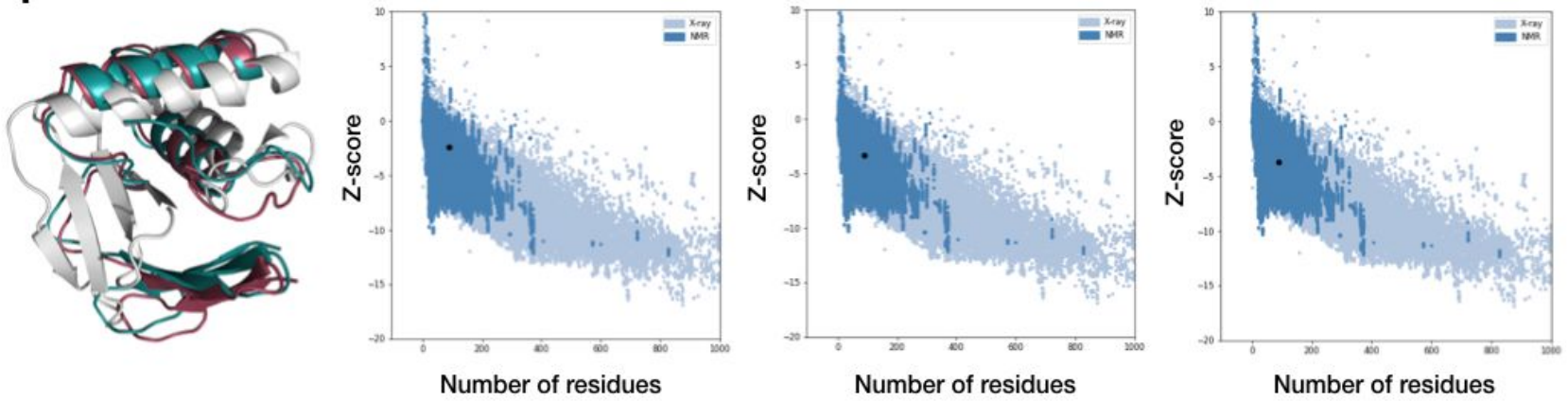

C
Z-score
Number of residues
X-ray
NMR


D
Z-score
Number of residues
X-ray
NMR

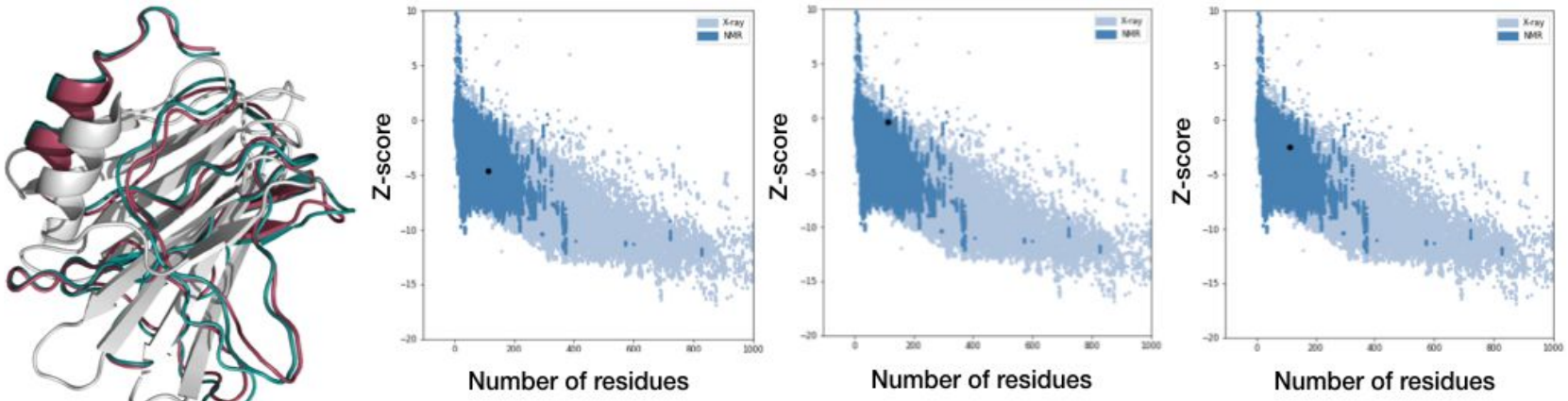

E
Z-score
Number of residues
X-ray
NMR
F
Z-score
Number of residues
X-ray
NMR


G
Z-score
Number of residues
X-ray
NMR
H
Z-score
Number of residues
X-ray
NMR

I


X-ray
NMR
Z-score
Number of residues
X-ray
NMR
Z-score
Number of residues
X-ray
NMR
Z-score
Number of residues

J


X-ray
NMR
Z-score
Number of residues
X-ray
NMR
Z-score
Number of residues
X-ray
NMR
Z-score
Number of residues

Supplementary Table 1. This table provides the fragment hit count obtained after the Protein Block Assignment (PBA) for each protein from the test dataset

PDB-ID	All hits	Identical PB hits
6fxa	35918	18719
5w9f	64375	18129
6gnx	39865	5001
6f45	38463	1787
6qek	65166	16205
6msp	68421	16602
6qfj	64983	16585
6btc	68094	12356
6d7y	67023	12493
6cp9	71111	16657

Supplementary Table 2. Secondary strutcure distribution as observed in the original structures along with the predicted secondary structure distribution as calculated by PSIPRED.

PDB-ID	Assigned Secondary Structures			Predicted Secondary Structures		
	% alpha	% beta	% coil	% alpha	% beta	% coil
6fxa	80	0	20	85.71	0	14.29
5w9f	29	26	45	34.15	31.71	34.14
6gnx	5	18	77	13.46	15.38	71.16
6f45	0	17	83	0	45.83	54.17
6qek	35	37	28	34.18	31.65	34.17
6msp	29	29	42	48.75	32.5	18.75
6qfj	38	35	27	37.04	34.57	28.39
6btc	39	15	49	40.6	14.58	44.82
6d7y	34	25	41	25	20.83	54.17
6cp9	6	56	38	6	49.12	44.88

Nothing in Biology Makes Sense Except in the Light of Evolution.

- Theodosius Dobzhansky, *in American Biology Teacher*

Bibliography

Adhikari, Badri and Jianlin Cheng (Aug. 2016). "Protein Residue Contacts and Prediction Methods". In: *Methods in molecular biology (Clifton, N.J.)* 1415, pp. 463–476. DOI: 10.1007/978-1-4939-3572-7_24.

Alford, Rebecca Faye, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V Shapovalov, P. Douglas Renfrew, Vikram Khipple Mulligan, Kalli Kappel, Jason W. Labonte, Michael Steven Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray (2017). "The Rosetta all-atom energy function for macromolecular modeling and design". In: *Journal of Chemical Theory and Computation*, acs.jctc.7b00125. DOI: 10.1021/acs.jctc.7b00125. arXiv: 106054. URL: http://pubs.acs.org/doi/abs/10.1021/acs.jctc.7b00125.

Anfinsen, Christian B (July 1972). "The formation and stabilization of protein structure". In: *Biochem. J.* 128.4, pp. 737–749.

— (1973). "Principles that Govern the Folding of Protein Chains". In: *Science* 181.4096, pp. 223–230. DOI: 10.1126/science.181.4096.223.

Aszódi, András, M.J. Gradwell, and W.R. Taylor (Sept. 1995). "Global Fold Determination from a Small Number of Distance Restraints". In: *Journal of molecular biology* 251, pp. 308–26. DOI: 10.1006/jmbi.1995.0436.

Benros, Cristina, Alexandre G de Brevern, Catherine Etchebest, and Serge Hazout (Mar. 2006). "Assessing a novel approach for predicting local 3D protein structures from sequence". In: *Proteins: Structure, Function, and Bioinformatics* 62.4, pp. 865–880.

Benros, Cristina, Alexandre G de Brevern, and Serge Hazout (2009). "Analyzing the sequence-structure relationship of a library of local structural prototypes". In: *Journal of Theoretical Biology* 256.2, pp. 215–226. DOI: 10.1016/j.jtbi.2008.08.032.

Bhattacharya, Debswapna, Badri Adhikari, Jilong Li, and Jianlin Cheng (Feb. 2016a). "FRAGSION: ultra-fast protein fragment library generation by IOHMM sampling". In: *Bioinformatics* 32.13, pp. 2059–2061. DOI: 10.1093/bioinformatics/btw067.

Bhattacharya, Debswapna and Jianlin Cheng (Jan. 2013). "3Drefine: Consistent Protein Structure Refinement by Optimizing Hydrogen-Bonding Network and Atomic-Level Energy Minimization." In: *Proteins: Structure, Function, and Bioinformatics* 81.1, pp. 119–131. DOI: 10.1002/prot.24167.

Bhattacharya, Debswapna, Jackson Nowotny, Renzhi Cao, and Jianlin Cheng (Apr. 2016b). "3Drefine: an interactive web server for efficient protein structure refinement". In: *Nucleic Acids Research* 44.W1, W406–W409. DOI: 10.1093/nar/gkw336.

Bonneau, R., J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. Strauss, and D. Baker (2001). "Rosetta in CASP4: progress in ab initio protein structure prediction". In: *Proteins: Structure, Function, and Bioinformatics* Suppl 5, pp. 119–126.

Bowers, Peter, Charlie Strauss, and David Baker (Jan. 2001). "De novo protein structure determination using sparse NMR data". In: *Journal of Biomolecular NMR* 18, pp. 311–318. DOI: 10.1023/A:1026744431105.

Bowie, J. U. and D. Eisenberg (1994). "An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function." In: *Proceedings of the National Academy of Sciences* 91.10, pp. 4436–4440. DOI: 10.1073/pnas.91.10.4436. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.91.10.4436.

Brevern, A. G. de and S. Hazout (2003). "'Hybrid protein model' for optimally defining 3D protein structure fragments". In: *Bioinformatics* 19.3, pp. 345–353. DOI: 10.1093/bioinformatics/btf859.

Buchan, D. W. A. and D. T. Jones (July 2019). "The PSIPRED Protein Analysis Workbench: 20 years on". In: *Nucleic Acids Research* 47.W1, W402–W407.

Buchan, Daniel W.A. and David T. Jones (2018). "Improved protein contact predictions with the MetaPSICOV2 server in CASP12". In: *Proteins: Structure, Function and Bioinformatics* 86.July 2017, pp. 78–83. DOI: 10.1002/prot.25379.

Burger, Lukas and Erik Van Nimwegen (2010). "Disentangling direct from indirect co-evolution of residues in protein alignments". In: *PLoS computational biology* 6.1. DOI: 10.1371/journal.pcbi.1000633.

Bystroff, Christopher, Kim Simons, Karen Han, and David Baker (Sept. 1996). "Local sequence-structure correlation in proteins". In: *Current Opinion in Biotechnology* 7, pp. 417–21. DOI: 10.1016/S0958-1669(96)80117-0.

Camproux, Anne-Claude, AG Brevern, Serge Hazout, and Pierre Tufféry (2001). "Exploring the use of a structural alphabet for structural prediction of protein loops". In: *Theoretical Chemistry Accounts* 106, pp. 28–35.

Camproux, Anne-Claude, Romain Gautier, and Pierre Tuffery (July 2004). "A Hidden Markov M Derived Structural Alphabet for Proteins". In: *Journal of Molecular Biology* 339, pp. 591–605. DOI: 10.1016/j.jmb.2004.04.005.

Camproux, Anne-Claude, 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". In: *Protein Engineering* 12.12, pp. 1063–1073. DOI: 10.1093/protein/12.12.1063.

Craveur, Pierrick, Agnel Praveen Joseph, Jeremy Esque, Tarun JaiRaj Narwani, Floriane Noel, Nicolas Shinada, Matthieu Goguet, Léonard Sylvain, Pierre Poulain, Olivier Bertrand, Guilhem Faure, Joseph Rebehmed, Amine Ghozlane, Lakshmipuram Swapna, Ramachandra Bhaskara, Jonathan Barnoud, Stéphane Téletchéa, Vincent Jallu, Jiri Cerny, Bohdan Schneider, Catherine Etchebest, Narayanaswamy Srinivasan, Jean-Christophe Gelly, and Alexandre de Brevern (2015). "Protein flexibility in the light of structural alphabets". In: *Frontiers in Molecular Biosciences* 2, p. 20. ISSN: 2296-889X. DOI: 10.3389/fmolb.2015.00020. URL: https://www.frontiersin.org/article/10.3389/fmolb.2015.00020.

de Brevern, Alexandre G (2005). "New assessment of a structural alphabet". In: *In Silico Biol. (Gedrukt)* 5.3, pp. 283–289.

de Brevern, Alexandre G, Catherine Etchebest, Christina Benros, and Serge Hazout (2007). ""Pinning strategy": A novel approach for predicting the backbone

structure in terms of protein blocks from sequence". In: *Journal of Biosciences* 32.1, pp. 51–70. DOI: 10.1007/s12038-007-0006-3.

de Brevern, Alexandre G, Catherine Etchebest, and Serge Hazout (2000). "Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks". In: *Proteins: Structure, Function, and Bioinformatics* 41.3, pp. 271–287.

Dong, Qiwen, Kai Wang, Bin Liu, and Xuan Liu (Aug. 2016). "Characterization and Prediction of Protein Flexibility Based on Structural Alphabets". In: *BioMed Research International* 2016, pp. 1–7. DOI: 10.1155/2016/4628025.

Duarte, Jose, Sathyapriya Rajagopal, Henning Stehr, Ioannis Filippis, and Michael Lappe (May 2010). "Optimal contact definition for reconstruction of Contact Maps". In: *BMC bioinformatics* 11, p. 283. DOI: 10.1186/1471-2105-11-283.

Efimov, A. V. (1996). "A structural tree for α-helical proteins containing α-α-corners and its application to protein classification". In: *FEBS Letters* 391.1-2, pp. 167–170. DOI: 10.1016/0014-5793(96)00720-X.

Feig, Michael (2017). "Computational protein structure refinement: almost there, yet still so far to go". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 7.3, e1307. DOI: 10.1002/wcms.1307.

Fetrow, Jacquelyn S (1995). "Omega loops: nonregular secondary structures significant in protein function and stability." In: *The FASEB Journal* 9.9, pp. 708–717. URL: https://doi.org/10.1096/fasebj.9.9.7601335.

Fischer, Axel W., Sten Heinze, Daniel K. Putnam, Bian Li, James C. Pino, Yan Xia, Carlos F. Lopez, and Jens Meiler (2016). "CASP11 - An evaluation of a modular BCL: Fold-based protein structure prediction pipeline". In: *PLoS ONE* 11.4, pp. 1–19. DOI: 10.1371/journal.pone.0152517.

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". In: *Nucleic Acids Research* 39.SUPPL. 2, pp. 18–23. DOI: 10.1093/nar/gkr333.

Ghouzam, Yassine, Guillaume Postic, Pierre Edouard Guerin, Alexandre G. De Brevern, and Jean Christophe Gelly (2016). "ORION: A web server for protein fold recognition and structure prediction using evolutionary hybrid profiles". In: *Scientific Reports* 6.June, pp. 1–10. DOI: 10.1038/srep28268. URL: http://dx.doi.org/10.1038/srep28268.

Goebel, U., Constantin Sander, Reinhard Schneider, and Aurelio Valencia (1994). "Correlated mutations and residue contacts in proteins." In: *Proteins: Structure, Function, and Bioinformatics* 18.4, pp. 309–17.

Govindarajan, Sridhar and Richard A Goldstein (1996). "Why are some proteins structures so common?" In: *Proceedings of the National Academy of Sciences* 93.8, pp. 3341–3345.

Govindarajan, Sridhar, Ruben Recabarren, and Richard A. Goldstein (1999). "Estimating the total number of protein folds". In: *Proteins: Structure, Function, and Bioinformatics* 35.4, pp. 408–414. DOI: 10.1002/(sici)1097-0134(19990601)35:4<408::aid-prot4>3.3.co;2-1.

Gront, Dominik, Daniel W. Kulp, Robert M. Vernon, Charlie E.M. Strauss, and David Baker (Aug. 2011). "Generalized fragment picking in rosetta: Design, protocols and applications". In: *PLoS ONE* 6.8. DOI: 10.1371/journal.pone.0023294.

Han, Karen F and David Baker (1996). "Global properties of the mapping between local amino acid sequence and local structure in proteins". In: *Proceedings of the National Academy of Sciences* 93.12, pp. 5814–5818.

Heo, Lim and Michael Feig (2018). "PREFMD: a web server for protein structure refinement via molecular dynamics simulations". In: *Bioinformatics* 34.6, pp. 1063–1065.

Heo, Lim, Hahnbeom Park, and Chaok Seok (June 2013). "GalaxyRefine: Protein structure refinement driven by side-chain repacking". In: *Nucleic Acids Research* 41. DOI: 10.1093/nar/gkt458.

Hovan, Ladislav, Vladimiras Oleinikovas, Havva Yalinca, Andriy Kryshtafovych, Giorgio Saladino, and Francesco Luigi Gervasio (2018). "Assessment of the model refinement category in CASP12". In: *Proteins: Structure, Function, and Bioinformatics* 86.S1, pp. 152–167. DOI: 10.1002/prot.25409.

Hunter, Cornelius and Shankar Subramaniam (Mar. 2003). "Protein local structure prediction from sequence". In: *Proteins: Structure, Function, and Bioinformatics* 50, pp. 572–9. DOI: 10.1002/prot.10310.

Izarzugaza, José M.G., Osvaldo Graña, Michael L. Tress, Alfonso Valencia, and Neil D. Clarke (2007). "Assessment of intramolecular contact predictions for CASP7". In: *Proteins: Structure, Function, and Bioinformatics* 69.SUPPL. 8, pp. 152–158. DOI: 10.1002/prot.21637.

Jayaram, B., Priyanka Dhingra, Bharat Lakhani, and Shashank Shekhar (2012). "Bhageerath - Targeting the near impossible: Pushing the frontiers of atomic models for protein tertiary structure prediction". In: *Journal of Chemical Sciences* 124.1, pp. 83–91. DOI: 10.1007/s12039-011-0189-x.

Jayaram, B., Priyanka Dhingra, Avinash Mishra, Rahul Kaushik, Goutam Mukherjee, Ankita Singh, and Shashank Shekhar (2014). "Bhageerath-H: A homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins". In: *BMC Bioinformatics* 15.16, pp. 1–12. DOI: 10.1186/1471-2105-15-S16-S7.

Jones, David T (1999). "Protein secondary structure prediction based on position-specific scoring matrices". In: *Journal of Molecular Biology* 292.2, pp. 195–202. DOI: 10.1006/jmbi.1999.3091.

Jones, David T. and Shaun M. Kandathil (2018). "High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features". In: *Bioinformatics* 34.19, pp. 3308–3315. DOI: 10.1093/bioinformatics/bty341.

Jones, David T., Tanya Singh, Tomasz Kosciolek, and Stuart Tetchner (2015). "MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins". In: *Bioinformatics* 31.7, pp. 999–1006. DOI: 10.1093/bioinformatics/btu791.

Joseph, Agnel Praveen, N. Srinivasan, and Alexandre G. de Brevern (2011). "Improvement of protein structure comparison using a structural alphabet". In: *Biochimie* 93.9, pp. 1434–1445. DOI: 10.1016/j.biochi.2011.04.010. URL: http://dx.doi.org/10.1016/j.biochi.2011.04.010.

Kalev, Ivan and Michael Habeck (Sept. 2011). "HHfrag: HMM-based fragment detection using HHpred". In: *Bioinformatics* 27.22, pp. 3110–3116. DOI: 10.1093/bioinformatics/btr541.

Kandathil, Shaun M., Joe G. Greener, and David T. Jones (Dec. 2019). "Prediction of interresidue contacts with DeepMetaPSICOV in CASP13". In: *Proteins: Structure, Function and Bioinformatics* 87.12, pp. 1092–1099. DOI: 10.1002/prot.25779.

Karaka, Mert, Nils Woetzel, Rene Staritzbichler, Nathan Alexander, Brian E. Weiner, and Jens Meiler (2012). "BCL::Fold - De Novo Prediction of Complex and Large Protein Topologies by Assembly of Secondary Structure Elements". In: *PLoS ONE* 7.11. DOI: 10.1371/journal.pone.0049240.

Khoury, George A, Phanourios Tamamis, Neesha Pinnaduwage, James Smadbeck, Chris A Kieslich, and Christodoulos A Floudas (May 2014). "Princeton_TIGRESS: protein geometry refinement using simulations and support vector machines". In: *Proteins: Structure, Function, and Bioinformatics* 82.5, pp. 794–814. ISSN: 0887-3585. DOI: 10.1002/prot.24459. URL: https://doi.org/10.1002/prot.24459.

Kihara, Daisuke, Hui Lu, Andrzej Kolinski, and Jeffrey Skolnick (2001). "TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints". In: *Proceedings of the National Academy of Sciences* 98.18, pp. 10125–10130. URL: www.pnas.orgcgidoi10.1073pnas.181328398.

Kim, David E., Dylan Chivian, and David Baker (2004). "Protein structure prediction and analysis using the Robetta server". In: *Nucleic Acids Research* 32.WEB SERVER ISS. Pp. 526–531. DOI: 10.1093/nar/gkh468.

Koga, Nobuyasu, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione, and David Baker (2012). "Principles for designing ideal protein structures". In: *Nature* 491.7423, pp. 222–227. DOI: 10.1038/nature11600.

Kohonen, Teuvo (Jan. 1982). "Self-organized formation of topologically correct feature maps". In: *Biological Cybernetics* 43.1, pp. 59–69. DOI: 10.1007/BF00337288. URL: https://doi.org/10.1007/BF00337288.

Kolata, Gina (1986). "Trying to crack the second half of the genetic code". In: *Science* 233, pp. 1037–1040.

Kolinski, Andrzej and Jeffrey Skolnick (1998). "Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model." In: *Proteins: Structure, Function, and Bioinformatics* 32.4, pp. 475–494. DOI: 10.1002/(SICI)1097-0134(19980901)32:4<475::AID-PROT6>3.0.CO;2-F. URL: http://www.ncbi.nlm.nih.gov/pubmed/9726417.

Kolodny, Rachel, Patrice Koehl, Leonidas Guibas, and Michael Levitt (2002). "Small libraries of protein fragments model native protein structures accurately". In: *Journal of Molecular Biology* 323.2, pp. 297–307. DOI: 10.1016/S0022-2836(02)00942-7.

Le, Quan, Gianluca Pollastri, and Patrice Koehl (Mar. 2009). "Structural Alphabets for Protein Structure Classification: A Comparison Study". In: *Journal of Molecular Biology* 387, pp. 431–50. DOI: 10.1016/j.jmb.2008.12.044.

Lee, Gyu Rie, Jonghun Won, Lim Heo, and Chaok Seok (July 2019). "GalaxyRefine2: simultaneous refinement of inaccurate local regions and overall protein structure". In: *Nucleic Acids Research* 47.W1, W451–W455. DOI: 10.1093/nar/gkz288.

Mahajan, Swapnil, Alexandre G. 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". In: *Journal of Biomolecular Structure and Dynamics* 32.5, pp. 751–758. DOI: 10.1080/07391102.2013.789989.

Martin, Juliette, Alexandre G de Brevern, and Anne-Claude Camproux (2008). "In silico local structure approach: a case study on outer membrane proteins". In: *Proteins: Structure, Function, and Bioinformatics* 71.1, pp. 92–109.

Maurice, Kevin J. (Mar. 2014). "SSThread: Template-free protein structure prediction by threading pairs of contacting secondary structures followed by assembly of overlapping pairs". In: *Journal of Computational Chemistry* 35.8, pp. 644–656. DOI: 10.1002/jcc.23543.

Micheletti, Cristian, Flavio Seno, and Amos Maritan (Sept. 2000). "Recurrent oligomers in proteins: An optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies." In: *Proteins: Structure, Function, and Bioinformatics* 40, pp. 662–74. DOI: 10.1002/1097-0134(20000901)40:4<662::AID-PROT90>3.0.CO;2-F.

Monastyrskyy, Bohdan, Daniel D'Andrea, Krzysztof Fidelis, Anna Tramontano, and Andriy Kryshtafovych (Feb. 2014). "Evaluation of residue-residue contact prediction in CASP10". In: *Proteins: Structure, Function, and Bioinformatics* 82. DOI: 10.1002/prot.24340.

— (2016). "New encouraging developments in contact prediction: Assessment of the CASP 11 results". In: *Proteins: Structure, Function, and Bioinformatics* 84.Suppl 1, pp. 131–144.

Moult, John, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano (May 2016). "Critical assessment of methods of protein structure prediction (CASP) - progress and new directions in Round XI". In: *Proteins: Structure, Function, and Bioinformatics* 84, pp. 4–14. DOI: 10.1002/prot.25064.

— (2018). "Critical assessment of methods of protein structure prediction (CASP)—Round XII". In: *Proteins: Structure, Function, and Bioinformatics* 86, pp. 7–15.

Moult, John, Krzysztof Fidelis, Adam Zemla, and Tim Hubbard (2003). "Critical assessment of methods of protein structure prediction (CASP)-round V". In: *Proteins: Structure, Function, and Bioinformatics* 53.S6, pp. 334–339.

Offmann, Bernard, Manoj Tyagi, and Akexandre de Brevern (2008). "Local Protein Structures". In: *Current Bioinformatics* 2.3, pp. 165–202. DOI: 10.2174/157489307781662105.

Oliva, Baldomero, Paul Bates, Enrique Querol, Francesc Avilés, and Michael Sternberg (Apr. 1997). "An automated classification of the structure of protein loops". In: *Journal of Molecular Biology* 266, pp. 814–30. DOI: 10.1006/jmbi.1996.0819.

Oliveira, Saulo de, Jiye Shi, and Charlotte Deane (Apr. 2015). "Building a Better Fragment Library for De Novo Protein Structure Prediction". In: *PloS one* 10, e0123998. DOI: 10.1371/journal.pone.0123998.

Regad, Leslie, Jean-Baptiste Chéron, Dhoha Triki, Caroline Senac, Delphine Flatters, and Anne-Claude Camproux (2017). "Exploring the potential of a structural alphabet-based tool for mining multiple target conformations and target flexibility insight". In: *PloS one* 12.8.

Regad, Leslie, Juliette Martin, Gregory Nuel, and Anne-Claude Camproux (Feb. 2010). "Mining protein loops using a structural alphabet and statistical exceptionality". In: *BMC Bioinformatics* 11, p. 75. DOI: 10.1186/1471-2105-11-75.

Rodrigues, João, Michael Levitt, and Gaurav Chopra (May 2012). "KoBaMIN: A knowledge-based minimization web server for protein structure refinement". In: *Nucleic Acids Research* 40, W323–8. DOI: 10.1093/nar/gks376.

Rohl, Carol A., Charlie E.M. Strauss, Kira M.S. Misura, and David Baker (2004). "Protein Structure Prediction Using Rosetta". In: *Methods in Enzymology* 383.February 2004, pp. 66–93. DOI: 10.1016/S0076-6879(04)83004-0.

Šali, Andrej and Tom L Blundell (1993). "Comparative protein modelling by satisfaction of spatial restraints". In: *Journal of Molecular Biology* 234.3, pp. 779–815.

Sander, Oliver, Ingolf Sommer, and Thomas Lengauer (2006). "Local protein structure prediction using discriminative models". In: *BMC Bioinformatics* 7, pp. 1–13. DOI: 10.1186/1471-2105-7-14.

Santos, Karina B, Gregório K Rocha, Fábio L Custódio, Helio JC Barbosa, and Laurent E Dardenne (2017). "Improving de novo protein structure prediction using contact maps information". In: *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, pp. 1–6.

Schaarschmidt, Joerg, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Alexandre M.J.J. Bonvin (2017). "Assessment of contact predictions in CASP12: coevolution and deep learning coming of age". In: *Proteins: Structure, Function, and Bioinformatics* October 2017, pp. 51–66. DOI: 10.1002/prot.25407. URL: http://doi.wiley.com/10.1002/prot.25407.

Shen, Yimin, Géraldine Picord, Frédéric Guyon, and Pierre Tuffery (2013). "Detecting protein candidate fragments using a structural alphabet profile comparison approach". In: *PloS One* 8.11.

Simons, Kim T., Rich Bonneau, Ingo Ruczinski, and David Baker (1999). "Ab initio protein structure prediction of CASP III targets using ROSETTA". In: *Proteins: Structure, Function, and Bioinformatics* 37.SUPPL. 3, pp. 171–176. DOI: 10.1002/(SICI)1097-0134(1999)37:3+<171::AID-PROT21>3.0.CO;2-Z.

Simons, Kim T., Charles Kooperberg, Enoch Huang, and David Baker (1997). "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions". In: *Journal of Molecular Biology* 268.1, pp. 209–225. DOI: 10.1006/jmbi.1997.0959. URL: http://linkinghub.elsevier.com/retrieve/pii/S0022283697909591.

Tomii, Kentaro, Yoshito Sawada, and Shinya Honda (2012). "Convergent evolution in structural elements of proteins investigated using cross profile analysis". In: *BMC bioinformatics* 13.1, p. 11.

Triki, Dhoha, Mario Enrique Cano Contreras, Delphine Flatters, Benoit Visseaux, Diane Descamps, Anne-Claude Camproux, and Leslie Regad (2018). "Analysis of the HIV-2 protease's adaptation to various ligands: characterization of backbone asymmetry using a structural alphabet". In: *Scientific Reports* 8.1, pp. 1–13.

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". In: *Genome Biology* 8.3. DOI: 10.1186/gb-2007-8-3-r31.

Tyagi, Manoj, Priyanka Sharma, CS Swamy, Frédéric Cadet, Narayanaswamy Srinivasan, Alexandre G de Brevern, and Bernard Offmann (2006a). "Protein Block

Expert (PBE): a web-based protein structure analysis server using a structural alphabet". In: *Nucleic Acids Research* 34.suppl_2, W119–W123.

Tyagi, Manoj, Gowri Venkatraman, Narayanaswamy Srinivasan, Alexandre de Brevern, and Bernard Offmann (Oct. 2006b). "A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications". In: *Proteins: Structure, Function, and Bioinformatics* 65, pp. 32–9. DOI: 10.1002/prot.21087.

Unger, Ron, Dov Harel, Scot Wherland, and Joel L. Sussman (1989). "A 3D building blocks approach to analyzing and predicting structure of proteins." In: *Proteins: Structure, Function, and Bioinformatics* 5.4, pp. 355–73.

Vetrivel, Iyanar, Swapnil Mahajan, Manoj Tyagi, Lionel Hoffmann, Yves Henri Sanejouand, Narayanaswamy Srinivasan, Alexandre G. De Brevern, Frédéric Cadet, and Bernard Offmann (2017). "Knowledge-based prediction of protein backbone conformation using a structural alphabet". In: *PLoS ONE* 12.11, pp. 1–20. DOI: 10.1371/journal.pone.0186215.

Wang, Sheng, Jianzhu Ma, Jian Peng, and Jinbo Xu (2013). "Protein structure alignment beyond spatial proximity". In: *Scientific reports* 3, p. 1448. DOI: 10.1038/srep01448.

Wang, Sheng and Wei-Mou Zheng (May 2008). "CLePAPS: Fast Pair Alignment of Protein Structures based on Conformational Letters". In: *Journal of Bioinformatics and Computational Biology* 6, pp. 347–66. DOI: 10.1142/S0219720008003461.

Wang, Tong, Yuedong Yang, Yaoqi Zhou, and Haipeng Gong (Dec. 2016). "LR-FragLib: an effective algorithm to identify fragments for de novo protein structure prediction". In: *Bioinformatics* 33.5, pp. 677–684. DOI: 10.1093/bioinformatics/btw668.

Waterhouse, Andrew, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian Heer, Tjaart Beer, Christine Rempfer, Lorenza Bordoli, Alba Lepore, and Torsten Schwede (May 2018). "SWISS-MODEL: homology modelling of protein structures and complexes". In: *Nucleic Acids Research* 46, W296–W303. DOI: 10.1093/nar/gky427.

Webb, Benjamin and Andrej Sali (2016). "Comparative protein structure modeling using MODELLER". In: *Current protocols in bioinformatics* 54.1, pp. 5–6.

Whisstock, James C. and Arthur M. Lesk (Aug. 2003). "Prediction of protein function from protein sequence and structure". In: *Quarterly Reviews of Biophysics* 36.3, pp. 307–340. DOI: 10.1017/S0033583503003901.

Xu, D. and Y. Zhang (Feb. 2013). "Toward optimal fragment generations for ab initio protein structure assembly". In: *Proteins: Structure, Function, and Bioinformatics* 81.2, pp. 229–239.

Xu, Dong and Yang Zhang (Nov. 2011). "Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization". In: *Biophysical Journal* 101, pp. 2525–34. DOI: 10.1016/j.bpj.2011.10.024.

— (2012). "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field". In: *Proteins: Structure, Function and Bioinformatics* 80.7, pp. 1715–1735. DOI: 10.1002/prot.24065.

Zemla, Adam (July 2003). "LGA: a method for finding 3D similarities in protein structures". In: *Nucleic Acids Research* 31.13, pp. 3370–3374. ISSN: 0305-1048. DOI: 10.1093/nar/gkg571. eprint: https://academic.oup.com/nar/article-

pdf/31/13/3370/9487387/gkg571.pdf. URL: https://doi.org/10.1093/nar/gkg571.

Zhang, Yang (2008). "Progress and challenges in protein structure prediction". In: *Current Opinion in Structural Biology* 18.3, pp. 342–348. DOI: 10.1016/j.sbi.2008.02.004.

Zhang, Yang and Jeffrey Skolnick (June 2004). "Automated Structure Prediction of Weakly Homologous Proteins on a Genomic Scale". In: *Proceedings of the National Academy of Sciences* 101, pp. 7594–9. DOI: 10.1073/pnas.0305695101.

Zhou, Hongyi and Jeffrey Skolnick (Sept. 2007). "Ab initio protein structure prediction using Chunk-TASSER". In: *Biophysical Journal* 93, pp. 1510–1518. DOI: 10.1529/biophysj.107.109959.

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE NANTES

Titre : Nouvelle approche pour la modélisation à l'échelle atomique de protéines à l'aide d'alphabets structuraux en l'absence d'homologues de structures connues.

Mots clés : Modélisation ab initio, Alphabet Structuraux, Blocs Protéiques et Modeller

Résumé : La modélisation des structures protéiques en l'absence d'homologie est appelée modélisation sans matrice, modélisation libre ou modélisation *ab initio*. À ce jour, ce type de modélisation reste le plus grand défi du domaine.
Durant cette thèse, nous avons apporté une contribution méthodologique pour relever ce défi en explorant l'utilisation d'un alphabet structurel connu sous le nom de blocs protéiques (BP). Il est maintenant possible de prédire le squelette carbonné d'une protéine sous forme de séquence de BPs de n'importe quelle protéine en utilisant des outils comme PB-kPRED. Notre travail a consisté à trouver une méthode pour passer de la séquence de BPs prédite à la structure 3D.
Nous avons donc développé une stratégie qui part d'une requête sous forme de séquences de BP prédites et recherche dans une base de données de structures non homologues des fragments de BPs similaires.
Notre stratégie tente ensuite d'assembler des fragments sélectionnés pour générer des modèles 3D. Pour cela, nous avons tenté d'exploiter les propriétés inhérentes de Modeller en la contraignant à utiliser des fragments de BPs et à utiliser des prédictions de structures secondaires et de contact comme contraintes externes supplémentaires.
Nous avons validé les résultats de cette stratégie sur un sous-ensemble de petites protéines cibles issues de CASP13 sur la base des scores couramment utilisés (GDT_TS et TM-score) pour les protocoles de modélisation gratuits. Nous avons pu montrer que des modèles de bonne qualité peuvent être générés pour de petites protéines en utilisant une combinaison de fragments à base de PB et de Modeller. Ce travail ouvre de nouvelles perspectives d'application des blocs protéiques pour solutionner l'épineux problème du repliement des protéines.

Title : New Method for Generating Protein Atomic Models using Structural Alphabets in the Absence of Homologs of Known Structures

Keywords : Template-free modelling, Structural Alphabets, Protein Blocks and Modeller

Abstract : Modelling of protein structures in the absence of known homology is termed as template-free modelling (TFM), free-modelling (FM) or *ab initio* modelling. To date this remains the biggest challenge in the field.
In the current work, we have made a methodological contribution to address this challenge by exploring the use of a structural alphabet known as protein blocks (PBs). With the help of PBs it is now possible to predict the PB sequence of any protein sequence using tools like PB-kPRED. Our work was motivated by the need to find a method to move forward from predicted PB sequence to 3D structure.
We hence developed a strategy that starts from a predicted query PB sequences and searches into a non-homologous structural database for similar PB stretches or fragments.
Our strategy then tries to assemble selected fragments to generate meaningful 3D models. For that, we attempted to exploit the inherent property of Modeller by constraining it to use PB fragments along with secondary structures and contact predictions as additional restraints to predict protein models.
We validated the results of our strategy on a subset of small proteins from CASP13 target list on the basis of the commonly used scores (GDT_TS and TM-score) for free modelling protocols. We were able to show that good quality models can be generated for small proteins using a combination of PB based fragments and Modeller. This work opened a new explorative path for protein blocks down the lane of solving the ever so perplexing protein folding dogma.