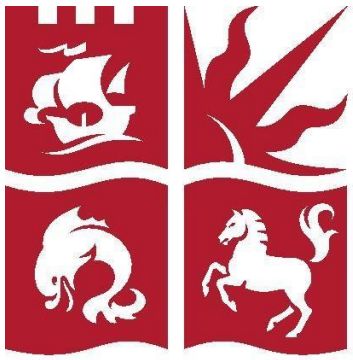
Author:
Neary, Tim E

Title:
Conformational Control of Modular Proteins



University of BRISTOL

Conformational Control of Modular Proteins

Timon Neary

September 2019

Dissertation submitted to the University of Bristol in accordance with requirements for the award of the degree of Masters by Research in Biochemistry in the Faculty of Life Sciences

Abstract

Repeat proteins represent an important class of proteins which, in nature, have been used for a variety of functions. Many repeat proteins in nature are used as dynamic protein scaffolds. Repeat proteins display unique folding properties whereby, each repeat domain folds largely independently of the overall protein, only interacting with its nearest neighbours. Accordingly, each repeat domain adopts different conformations depending on the context it is found in. Recently improvements in methods targeted at rapidly designing de novo repeat proteins have been developed, however, incorporating the dynamic aspects of these proteins into the design has largely been ignored. Here we simulated and analysed a series of repeat protein domains to assess their dynamics and classify the resulting ensemble of structures into a small representative set of conformations. To accomplish this, we utilised metrics such as RMSD and developed Dynamatch, a python tool which maps structural perturbations to rigid body transforms, to extract dynamic information from each domain. Analysing structural conformations from each domain we demonstrate that they were most commonly found in conformations deviating little from the reference structure. Comparison of the behaviours of modules in different contexts was shown to have an impact on the conformational set they were able to sample. We aim to use information gained on the dynamic properties of these repeat domains to better guide the design of repeat proteins with an emphasis on enabling design towards more functional proteins.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

“This project is my own work except where indicated. All text, figures, tables, data or results, which are not my own work, are indicated and the sources acknowledged. In addition, I confirm that the hardcopy and the e-submission are identical.”

Signed:_____ Dated:_____

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

I would like to thank my supervisor, Dr. Fabio Parmeggiani for his support and guidance throughout the project, without which this project would not have come to be.

I would also like to thank Dr Richard Sessions and the members of his lab for providing me with immeasurable support they provided me both inside and out of the lab.

Special thanks to the Advanced Computing Research Centre (ACRC) and BrisSynBio at the University of Bristol for allowing me access to their computer clusters to enable my research.

Introduction

Repeat proteins, RPs, represent a large proportion of the total protein space, comprising a hugely diverse range of functions, particularly in eukaryotes. Tandem RPs are formed from several repeat domains distributed adjacent to each other within a protein sequence [1]. By making a small number of interactions per repeat domain, RPs can interact with large macromolecular structures through the additive interactions of each of its repeats. This enables RPs to become particularly useful in protein-protein interactions, DNA-protein interactions or simply as protein scaffolds [2,3]. Indeed, many protein design attempts have utilised this cooperative binding capacity to design RP to bind extended recognition motifs [4,5]. Tandem RPs are thought to have arisen from duplication of repeat domain sequences within a gene, though the specific mechanism has yet to be deciphered. However, sequence conservation between duplications is often low, partially due to the low number of conserved residues required to ensure a correct fold [6]. Which has enabled particular repeat domains to be modified to enable a modular system whereby individual repeat domains may have designated roles. In nature this may enable more RPs to adapt to different stimuli [6], however, this modularity has also been exploited in protein design methods to enable the design of novel protein architectures [7].

Protein folding in globular proteins, however, is more complex whereby each residue contributes in a non-obvious manner to the overall protein. Interactions may be long or short range and can contribute in energetically favourable or repulsive interactions. Overall, interactions between residues tend to be energetically favoured compared to random interactions which creates an ensemble of similar structures at the lowest energy states [8]. Combinations of small-scale thermal motions, such as rotamer sampling, and larger scale domain movements such as loop movements can be used to describe the pathways to move between these structures, giving a protein its folding landscape [9]. Due to the minimal differences in energy between these states, it is necessary to consider folded proteins as an ensemble of these like structures which are able to interconvert. Unlike in globular proteins in which a combination of long- and short-range interactions between each residue coalesce to determine the fold and overall stability of the whole protein, each domain repeat in an RP folds individually then acts on the adjacent repeats [3,10]. This model for describing the folding of RPs is known as the Ising model. Originally the Ising equation was used to describe the spin states in ferromagnetic metals, whereby the spin states of dipoles is influenced by the spin present in neighbouring atoms. Later, this dependence on neighbours was shown to be necessary to explain the folding of consensus designed tandem repeat proteins [11,12]. The Ising model describes the folding energy of consensus tandem repeat proteins as an additive score of each repeat domain, which in turn is the summation of the intrinsic energy of the individual repeats and the interaction energy of its interfaces [3]. The α-solenoid class of RPs, whereby each repeat domain is comprised entirely of α-helices, demonstrate an extension to the model. These RPs can be very flexible which enables then to act as dynamic protein scaffolds, able to flex about or in response to substrates [13]. This makes them of interest in the design of flexible nanoscale scaffolds, particularly if this flexibility can be modulated through the use of external stimuli.

In nature, the modularity of RPs has been exploited to enable them to act as allosteric modulators for multi-subunit protein complexes. Utilising small molecule or protein binding RPs are able to alter their flexibility either wholly or partially across the structure. A study of the Rap proteins demonstrated that binding of their respective small molecule ligand, PhrC, was able to reduce the conformational space that the C-terminal domain was able to sample,

locking it into a closed conformation. Despite minimal changes to the internal structure of each of its repeat domains, while in the closed conformation, Rap displayed a very large global conformational change which enables it to activate downstream effector proteins [3]. Hence, alterations made in part of the protein were able to be propagated across the repeat domain array. Using the Rosetta software suite [14], a number of approaches have been developed to enable the rational design of targeted ligand binding. Metalloproteins have an diverse range of functions and as such common targets for ligands are metals, either as single metal ions or larger cofactors. The design of these binding sites tends to be fairly simple whereby the orientation and distance of coordinating ligands can be sufficient to enable metal binding [15]. Additionally, metals have been shown to be able to mediate other interactions, such as oligomerisation [16], expanding their potential use in controlling the construction of nanoscale structures. The design of more complex ligand binding sites has also been achieved, Lansu et al. have developed a method to locate high affinity binding targets even against proteins with no known structure. Using the structures of homologous proteins they were able to approximate the structure of the atypical opioid receptor MRGPRX2 and then screen this structure against a ligand library to assess its binding [17]. They were able to discover a specific high affinity ligand using this approach. However, the design of small molecule ligand binding sites often requires deep hydrophobic cavities in the protein, making this approach difficult for RPs whose individual repeat domains may only be a few hundred amino acids. Recently, Maguire et al. were able to develop a new protocol to rapidly search across a protein to form self-contained hydrogen bonding networks [18]. The introduction of this model could enable the design of specific pH sensitive networks. In addition, it better enables the design of extended hydrogen bond networks to act as secondary coordination shells. These networks have been seen as a critical to achieve high affinity binding, indeed large structural changes in binding sites have been shown to better enable the protein to achieve these extended networks [19].

More recently, methods to enable the rapid design of RPs without the need for extensive computational simulations have developed. These methods have exploited the modular nature of RPs to better enable rapid *de novo* designed for RPs with specific geometries. Designed helical repeats (DHRs), comprised entirely α-helix turn α-helix turn motifs, were designed using Rosetta which sampled a range of helix and loop lengths [7]. These DHRs were then expanded on to create modular units (modules) [20]. Junctional regions, also α-helix turn α-helix turn motifs, were introduced between pairs of DHR helices to allow for different RP geometries. Modules lacking a joining region were termed base modules while those containing a joining region were termed junction modules. The terminology of the modules was altered such that DHR## now corresponds to D##, where ## indicates the number of the DHR, and junctional modules were in the form D##_j#_D##, where each D## can correspond to any DHR and j# corresponds to some junction. Later, Elfin was developed, by Joy Yeh, to utilise a genetic algorithm to place each module, in sequence, in 3D space next to each other and allowed for the rapid design of RPs [21,22]. These designed RPs were able to adopt novel geometries and can be built for specific 3D structures. Elfin is available on GitHub at https://github.com/joy13975/elfin.

Here we describe an approach to broaden the dynamical understanding of these designed modules. We aim to use this understanding to extrapolate the global dynamics of RPs to enable the design of more functional RPs. As part of this, we created an RP library and performed simulations to elucidate the ensemble of conformational states available to each module.

Methods

Repeat protein database construction

To analyse the dynamics of all modules in every environment, a library of RPs was designed to include all possible environments for all modules. The construction of the RP database utilised 34 modules from the overall repeat module database, described in [7,20].

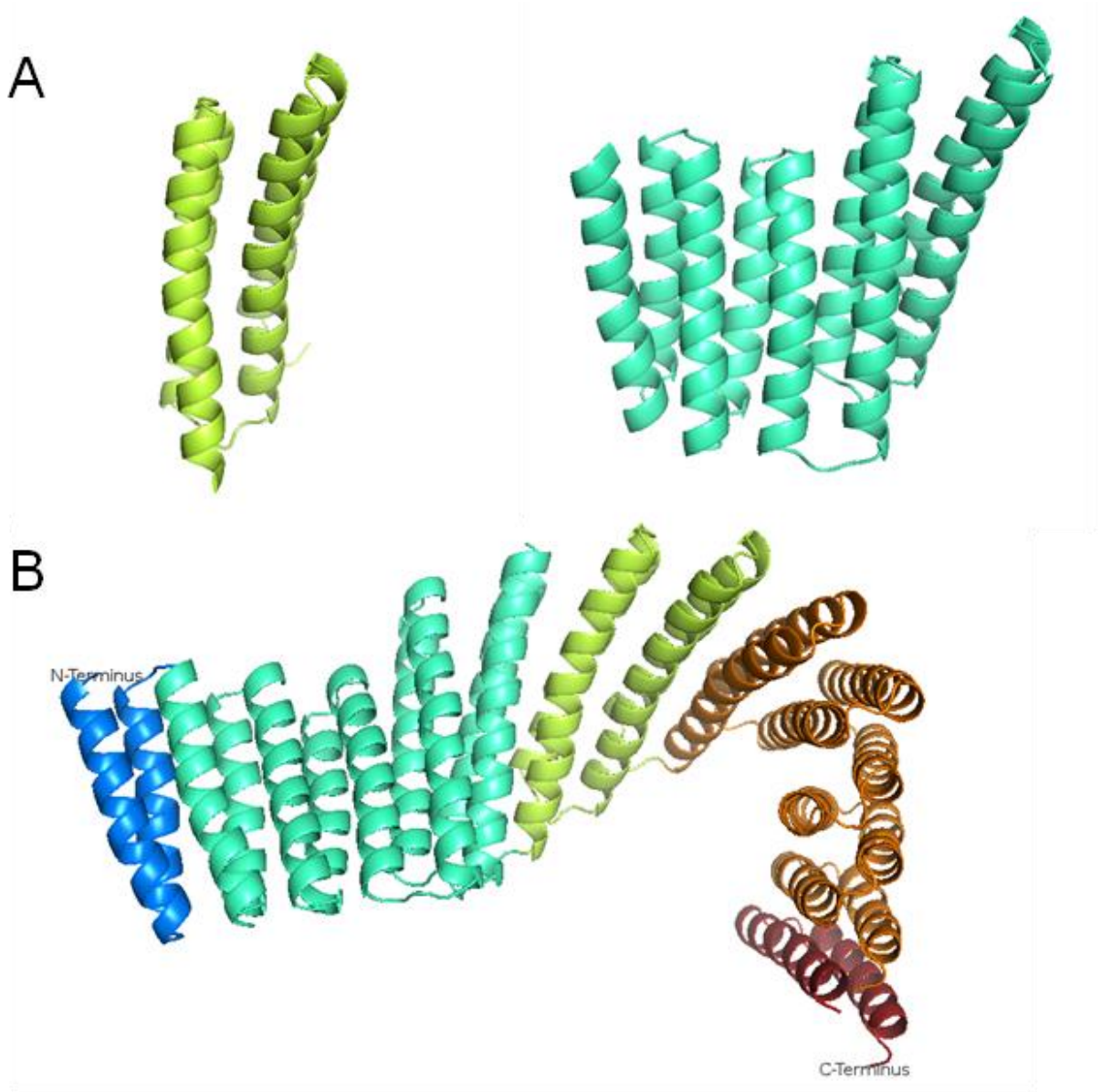


Figure 1 – Representative structures of individual modules and full RPs from the RP library. A, left) Structure of the simple module, D79, taken from rp78. A, right) Structure of the complex module D54_j1_D79, taken from rp78. B) Full structure of rp78 (NcapD14, D14_j2_D79, D79_j2_D14, D14_j1_D14, CcapD14). Its N and C terminus have been labelled.

Using an exhaustive combinatorial approach, 644 RPs were designed to accommodate all possible valid combinations of each set of 3 modules. A full table of each RP and its constituent modules can be found in figure S1. When creating the combinations of modules, their order significant, hence, 3 modules in the sequence [A, B, C] would be considered different to the form [A, C, B]. Module combinations were considered invalid if adjacent modules had incompatible interfaces. This created all valid contexts for each module, where a module's context is defined as the identity of its adjacent modules and their order. The standard nomenclature for a module context is as follows: ["N-terminal module", "Central module of interest", "C-terminal module"]. For example, a context for the module D79 is ["D79", "D79", "D79_j2_D14"]. In this case D79 is found towards the N-terminus of the central D79 module and D79_j2_D14 is found towards the C-terminus. To allow for the designed proteins to be soluble in water, capping module were added to the N and C termini of the protein. Therefore, all RPs in this report contain 3 modules and 2 capping repeats. In cases where the central

module in a context is located adjacent to capping repeats, the capping module may be named or simply labelled “CAP”. Capping domains were not considered for individual analysis for their dynamic properties. RPs were expressed as a single polypeptide chain where the first amino acid of the next module directly follows the last amino acid of the previous. Henceforth, any specifically referenced RPs will be followed by their respective modules, from N- to C-terminus in parentheses. E.g. rp78 (NcapD14, D14_j2_D79, D79_j2_D14, D14_j1_D14, CcapD14).

To generate the protein structure model for each of the RPs the Elfin program was used. As Elfin uses a low-resolution model to place modules in 3D space, it can create gaps between interfaces in adjacent modules. To combat this, each structure was relaxed by Rosetta, using the Rosetta relax application [23–25], to optimise packing of side chains between modules and generate a low energy reference structure from which all further analysis of that RP could be based from. For each RP an ensemble of low energy decoy structures was created using a further 100 rounds of the Rosetta relax application. Rosetta version 3.9 and the ref2015 score function were used to relax the RPs at all stages. No criteria were used to reject decoys once they were generated. Therefore, for each RP, a single reference structure and 100 decoy structures were generated, these were then used to create the RP database.

ppmp - determining conformations of modules in context

A common method for detecting perturbations in proteins structure is to use the RMSD as calculated against a reference structure. RMSD has proven to be useful metric in this sense as it quantifies structural changes across whole protein structures using a single value. As such it was used to classify the ensemble of structure present in the RP library. For every decoy, the RMSD of each module compared to a reference structure, was calculated using Rosetta. The RMSD is calculated as the square root of the sum of the square of the deviation of each C_α atom for each residue in a module divided by the total number of residues, see equation 1.

$$RMSD(v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||v_i - w_i||^2}$$

Where n is the number of residues and w and v correspond to the reference and decoy set of (x, y, z) coordinates for each atom (in sequence).

Ppmp (Predicting Perturbations of Modular Proteins), a python based analysis tool developed by M., Ramcharan et al., as an undergraduate engineering project at the University of Bristol, was used to quantify the changes in structure between the reference and the decoys with RMSD [26]. The original source code can be downloaded at https://github.com/vlamacko/ppmp. Ppmp generates histograms for each distribution which are then visualised as kernel density approximations. Kernel density approximations convert individual points to small gaussian distribution to estimate a probability density. This attempts to reduce the overall noise seen in the data. The bandwidth for each kernel, critical to creating good estimates, is determined automatically at run time to best fit the data [27,28].

The Rosetta RMSD values for each RP’s decoys were converted to a series of RMSD distributions for each context and a series of general statistics were calculated from these distributions. I modified the code, improving its error detection, introducing more general

statistics, and altering the distribution graph output. Specifically, ppmp now outputs the maximum peak RMSD, standard deviation, and the total number of gaussian peaks which comprise each individual distribution. To determine the number of peaks present in each distribution scikit-learn's GaussianMixture class was used. Given a user defined number of gaussian peaks, it attempts to best fit those peaks to a given distribution using a maximum likelihood algorithm. In ppmp, a series of GaussianMixtures is fitted to each distribution and where the best performing number of gaussians is assumed to be the number of peaks present in the distribution. The full dataset of each distribution can be found in the supplementary materials.

Dynamatch - describing protein motions as rigid body transformations

As RMSD reduces the perturbations in protein structure to a single value, information is lost. This loss of information can make it difficult to accurate differentiate between structures as similar RMSD values may correspond to like structures. To combat this a more information rich classification method was developed.

Dynamatch is a python tool which aims to characterise conformational changes observed in modular proteins as a series of rigid body transforms applied to the secondary structure elements of the module. Figure 2 demonstrates the overall methodology for the Dynamatch program. Protein structures are reduced to a set of 3D datapoints corresponding to the C_α coordinates of each residue in the structure. Each decoy is split into its constituent modules and then each module is compared against a common reference structure. To ensure that translations between rigid bodies were consistent, the decoy structure was first superimposed against the reference structure. In each case the x-ray crystallography structure generated in [7] was used as the reference. From these crystallography structures the amino acids which comprise the secondary structure elements (SSEs) of the module are determined using the mkdssp program [29]. The secondary structure definition generated for the reference was used for all other decoy structures. Loops were not considered as SSEs during comparison or analysis.

To generate rigid body transforms of each SSE, a plane and centre of mass was defined. The centre of mass was calculated as the mean of the x, y, and z coordinates respectively for all C_α atoms comprising the SSE. To define a plane for each element, three axes were calculated. The first axis was calculated as the first principle component of the 3D coordinate set for the SSE. To ensure a valid comparison can be made between each plane, N- to C-terminal directionality is ensured for this axis, i.e. the axis points towards the C-terminal residue in the SSE. A temporary second axis is calculated as the vector between the centre of mass and the middle residue. The middle residue of an SSE is defined as the residue which falls halfway, rounded down, along the primary sequence of the SSE. For example, for a helix formed from 21 amino acids, the 10^{th} residue would be the middle residue. The third axis is calculated as the normal of the first two axes. The second axis is then recalculated as the negative of the normal of the first and third axes. Therefore, the rigid body transform for a single element is calculated as the translation from the reference centre of mass to the decoy centre of mass and the rotation which maps the reference plane to the decoy's.

Protein structures were clustered using the rigid body transforms. For each module a 6n x m matrix, M, is constructed which contains the x, y, and z translation and X, Y, Z rotation for each centre of mass/plane for each SSE for a module. Where: n = the number of SSEs, and m = the total number of decoys for a module across the RP library. To reduce the dimensionality

of the data, allowing for efficient clustering, principle component analysis, PCA, is performed on M, to reduce the data to 3-dimensional space. Clustering is then performed on the PCA transformed M. The clustering algorithm used for each module is determined at run time. Three different algorithms (KMeans, Agglomerative, and OPTICS as implemented by scikit-learn [30,31]) are assessed, and the gap statistic is used as a measure of fitness to determine which clustering algorithm to use [32]. The user guide for all clustering methods implemented by scikit-learn can be found at https://scikit-learn.org/stable/modules/clustering.html#clustering. Specifically, the gap statistic is calculated for a range of parameters for each clustering algorithm and then compared. The best performing parameters for an algorithm are selected and compared against the best parameters for the other algorithms. The method with the largest gap statistic (fittest) is selected with its optimised parameters.

The clusters then represent specific conformational states for a given module. Each cluster/conformation is described by a range of metrics, including: mean average structure, calculated as the mean x, y, and z coordinates for each C_α atom; distributions for the rigid body transforms for each SSE; and average deviation/fluctuation of each C_α atom with respect to the reference structure.

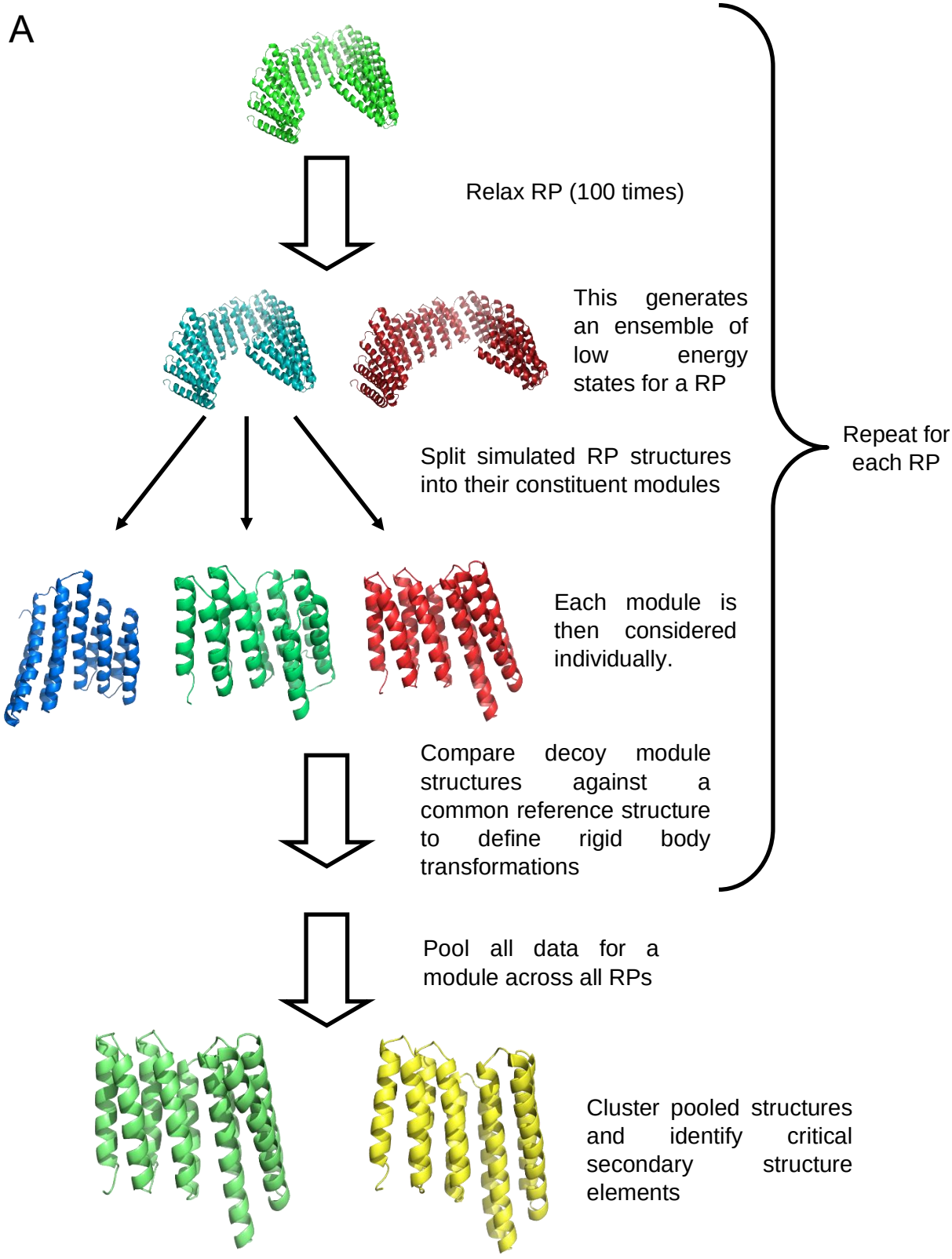


Figure 2 - Schematic overview of the methodology of the Dynamatch program. When splitting RPs into its modules, information is stored regarding the context of each module. This information is used during the pooling and analysis process. For each module, all contexts in which it can be found across all RPs are pooled and analysed.

Validation of computational methods

To assess the validity of the models generated by Rosetta relax a small number of RPs were selected and simulated using the orthogonal computational method of Molecular Dynamics, MD. In total 6 RPs were selected for additional simulations, rp2, rp110, rp209, rp281, rp282, and rp628. The structures used for the simulations were those generated after the initial relaxation step of Elfin generated RPs. Each protein was simulated using Gromacs 5.1 in the Amber99sb-ildn forcefield modelled with explicit waters (tip3p) [33–35]. Gromacs is available from http://www.gromacs.org/. The simulation was conducted at 1 atm pressure and 300 K. Each RP was simulated for a total of 100 ns over 5 different trajectories. A series of RMSD calculations were calculated for each time point to determine whether the protein was able to achieve an equilibrated state. Structures found at equilibrium were then compared against structures generated from Rosetta relax.

Controlling Conformations using selective conformational stabilisation

Rosetta Match [36] was used to design a Zinc binding site that could be hosted by one of the RPs in a subset of the RP library [match paper], using the procedure described in the documentations. RPs were chosen to be a part of the subset if they demonstrated significant deviations in structure, specifically, if they demonstrated a maximum RMSD of greater than 5. The match algorithm was applied to each conformational set for the RPs. Hits which were found in only one conformation for a given RP were then selected for further scrutiny. 113 different hits were then relaxed, using Rosetta, 10 times. A series of metrics, including RMSD, and Rosetta score were then used to determine hits most likely to bind Zinc and demonstrate conformational shifts when binding Zinc. The best 15 hits were then manually inspected to determine the best hit. Criteria for selection included exposure of the binding site to solvent, location of the binding site in the RP and potential entropic caused by the arrangement of the side chains in the binding site.

Experimental methods, protein expression and purification

Additionally, rp628 was selected to be expressed, purified and characterised. A 6His tagged rp628 was over-expressed in BL21 E. coli cells. Purification was achieved in Nickel affinity purification columns using 20mM TRIS, 500mM NaCl, 200mM Imidazole as the elution buffer. A dialysis step was performed to remove the imidazole. A high salt concentration was found to be necessary to prevent aggregation, lower than 500mM NaCl was found to reversibly trigger rp628 to precipitate. A final size exclusion chromatography step, using the Superdex 200, was used to remove the remaining impurities.

Results and Discussion

To efficiently understand the dynamic properties of each module with respect to the context it was necessary to extract a small subset of structures from the overall ensemble which

represented the range of movement exhibited by the module. In addition, it was necessary to discern how the set of conformations a module sampled in a context related to the overall set. Critically, due to the large number of protein structures, 644,000, it was necessary to reduce the number of structures which needed to be manually inspected and provide information regarding global trends.

To assess the dynamic properties of each module in all possible environments, the RP database was constructed, as described in Methods, such that all valid contexts for each module were represented. By assessing the local energy minima of each module in each context we hypothesised that it should be possible to model the dynamics of the overall RPs by applying the specific dynamics of each module in the context(s) in which they are found in the protein. Which would enable high throughout design of larger protein structures with detailed understanding of the kinetics of the system without the need to undergo expensive *in silico* simulations.

A coarse grain approach to conformational classification

To locate the energy minima for each of the modules, in context, all 644 designed modular proteins were relaxed by Rosetta's relax application. Despite Rosetta's high performance, due to the length of the proteins, $\bar{x}$ = 647 amino acids, the average time to relax a single protein on 1 core was approximately 30 minutes. Therefore, relaxing each protein once would require 322 CPU hours. While significantly faster than a MD simulation, it still requires a large amount of resources to process. As such, to compromise on computational expense whilst also attempting to sample all local minima, each RP was relaxed 100 times, producing 100 decoy structures. Comparison of these decoy structures to the reference structure was used as the basis from which the dynamics of each module was extracted.

ppmp, developed by M., Ramcharan, et al., was used to quantify the changes in structure between the reference and the decoys [26]. Specifically, it was used to determine the distribution of conformational states sampled by modules in each context and the extent by which they differed. ppmp uses a coarse grain metric, RMSD, between module structures to define changes in structure. Global RMSD distributions were created by amalgamating all data across all contexts into a single distribution, shown in figure 3. The global distributions demonstrate the differences in flexibility between modules, whereby some modules can sample conformational states significantly more different than others.

For all modules, there is a clear preference for a single conformational state similar to the reference as seen by the approximate right-skewed normal distributions they exhibit. This would indicate that, irrespective to context, modules have a preference to a small subset of conformations which deviate only slightly to the original structure. In particular, except for D81, all base module distributions display smaller deviations in structure than junction modules. Thus, suggesting that they are more rigid. The abnormal distribution of D81 is likely due to the initial placement of the modules by Elfin. As stated before, Elfin uses a coarse grain model of each module to orientate and place them next to each other in 3D space. After conversion to full atom structures, some atoms overlap each other. Hence during relaxations with Rosetta large changes in structure occur to relieve these clashes. Overlaying all context distributions onto a single graph can help to identify potential reasons for this observation.

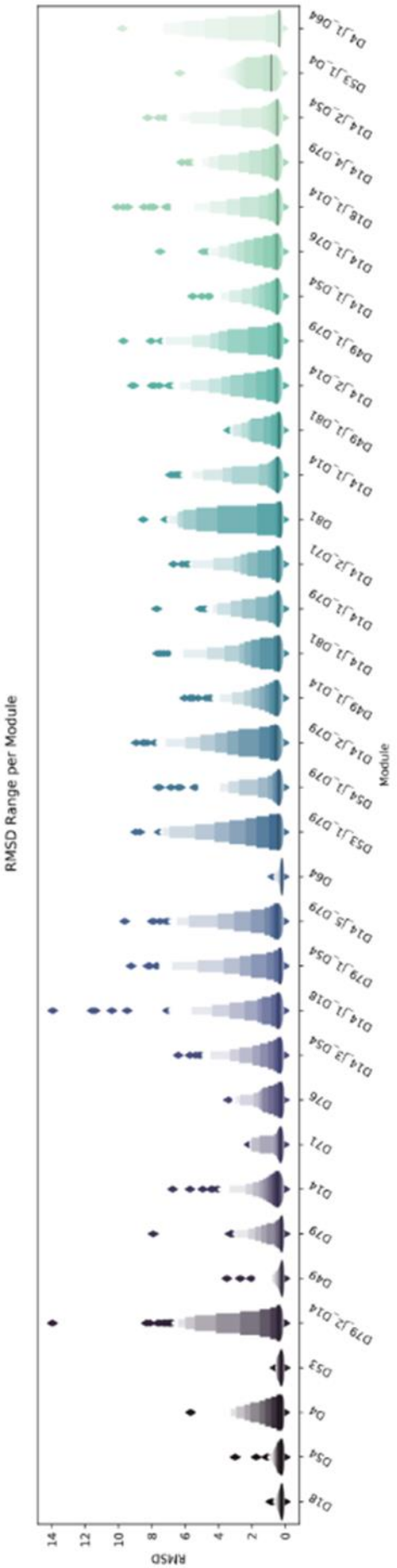


Figure 3 – Level plot of the overall distributions of each module. Bar width indicates a higher frequency density at the given RMSD value. Diamonds located at higher RMSD values indicate outliers for a given distribution.

The global distribution of each module is comprised of the distributions of the different module contexts. By inspecting the individual context distributions it is possible to identify additional trends which are not immediately visible when observing the global distributions. In general, most contexts deviate only slightly from the reference structure (as seen in the four examples in figure 4) and often display similar behaviours as denoted by the similarly shaped distributions. In contrast a smaller number of distributions display significantly different behaviours, these changes can be seen as much larger shifts in the peak values, much broader distributions, or distinct multimodal properties. Indeed, all four examples in figure 4 are consistent with this observation.

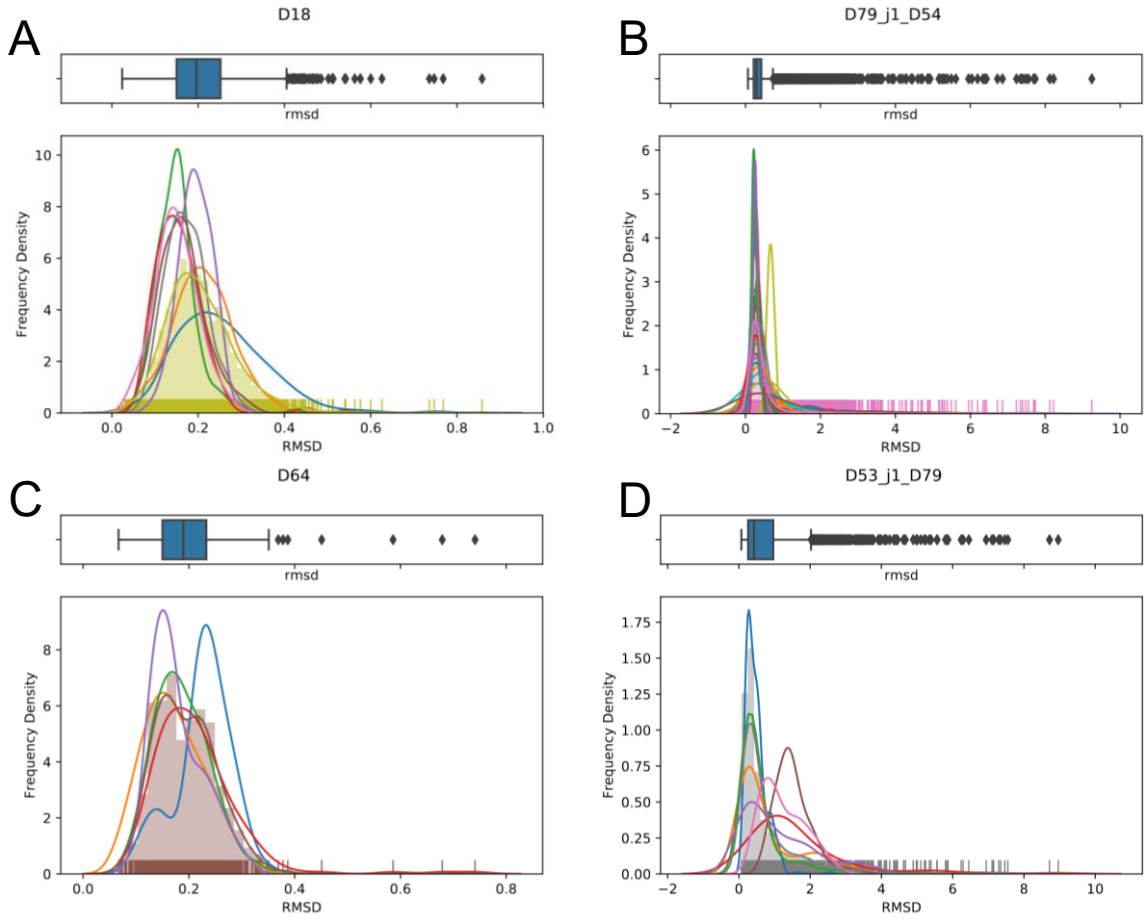


Figure 4 – Graphs representing the RMSD distributions for a single module. The box plot above the Kernel density estimates/histogram are calculated for the overall distribution only. The diamonds represent outliers in the overall distribution. For kernel density approximation/histogram graph each coloured line represents a single context distribution. The pips at the bottom are individual data points and the bars represent the histogram for the overall distribution. The line in the same colour as the histogram bars represents the kernel density approximation for the global distribution. A) RMSD distributions for the D18 module, the global distribution is coloured yellow in this instance. B) RMSD distributions for the D79_j1_D54 module, the global distribution is coloured pink in this instance. C) RMSD distributions for the D54 module, the global distribution is coloured maroon in this instance. D) RMSD distributions for the D53_j1_D79 module, the global distribution is coloured grey in this instance.

Analysis of table 1 which displays the series of average statistics for each module provides quantitative data supporting our conclusions. The average maximum peak RMSD for each module indicates that most structures deviate little from their starting point with an average RMSD across all modules of 0.41 Å. Indeed, only 5 modules display an average maximum peak above 0.5, D14_j1_D54, D14_j1_D79, D53_j1_D4, D53_j1_D79, D81. Surprisingly,

while most of these modules contain D53 interfaces D53 itself is very rigid with a maximum peak of 0.21 and standard deviation of 0.20.

It is likely that the small changes in RMSD observed in the majority of contexts are caused by minor rearrangements made to the structure to accommodate alternative packing triggered during Rosetta relax. Analysis of the average standard deviation, which in almost all cases is observed to be greater than the average maximum peak, may suggest that the data are very disperse perhaps and each module is able to adopt a wide range of continuous structures. However, when the average number of peaks is considered (>1 in all modules) the unusually high standard deviation may instead be due to the spread of data across multiple gaussian peak distributions, which would instead suggest that modules adopt a small number of discrete conformations rather than a continuous set. Manual inspection of each distribution also suggests that this is the most likely scenario. That said, in some contexts where the distribution is observed to be unimodal with a broad spread a more continuous set of conformations is more likely. In the three cases where the average standard deviation is lower, the modules in question have a very small range of RMSD values across all its contexts.

Module	Average Max peak RMSD (Å)	Average Standard Deviation (Å)	Average Number of Peaks	Total number of clusters
D14	0.46	0.62	1.55	1
D14_j1_D14	0.44	0.81	1.73	2
D14_j1_D18	0.35	1.11	1.75	5
D14_j1_D54	0.54	0.84	1.85	2
D14_j1_D76	0.45	1.11	2.00	2
D14_j1_D79	0.44	1.05	1.89	5
D14_j1_D81	0.42	1.24	1.62	3
D14_j2_D14	0.48	1.21	1.85	1
D14_j2_D54	0.48	0.83	1.55	3
D14_j2_D71	0.40	1.20	1.92	1
D14_j2_D79	0.68	2.02	1.93	2
D14_j3_D54	0.34	0.87	2.00	3
D14_j4_D79	0.49	1.10	1.93	5
D14_j5_D79	0.49	1.50	2.07	4
D18	0.17	0.15	3.75	3
D18_j1_D14	0.44	1.24	1.70	5
D4	0.30	0.59	2.13	3
D49	0.17	0.25	2.78	2
D49_j1_D14	0.48	1.01	1.83	3
D49_j1_D79	0.39	1.98	1.86	4
D49_j1_D81	0.35	0.88	2.00	3
D4_j1_D64	0.38	1.65	2.60	3
D53	0.21	0.20	2.14	3
D53_j1_D4	0.91	1.10	1.60	4
D53_j1_D79	0.63	1.90	2.00	1
D54	0.18	0.28	2.71	4
D54_j1_D79	0.35	1.01	2.00	3
D64	0.18	0.14	3.00	3
D71	0.29	0.42	2.40	2
D76	0.21	0.35	2.00	2
D79	0.23	0.51	2.62	1
D79_j1_D54	0.32	1.30	2.23	2
D79_j2_D14	0.46	1.29	1.96	2
D81	0.71	1.81	1.86	2

Table 1 - Depicts the general statistics for each module. For each statistic the value was calculated for each context, then contexts with the same central module were pooled and the mean was

calculated across all contexts. The max peak was calculated as the RMSD for the largest frequency density. The standard deviation was calculated across all data points in the distribution. The number of peaks was calculated as the number of gaussian peaks comprising the distribution, see methods for more details. For each module, its contexts were clustered using the scikit-learn KMeans clustering algorithm. Each context was considered as a 2D point using its maximum peak and standard deviation. The total number of clusters corresponds to the optimal number of clusters found for that module when the gap statistic was used to measure fitness.

As has been eluded to previously, inspecting individual distributions lead to the classification of the behaviours exhibited by each module on a contextual level. Individual context distributions can be broadly split into three separate categories based on their overall shape. Contexts were classified by calculating the standard deviation and number of gaussian peaks which comprise the context distribution. The types of distributions are detailed below using a representative context for each category.

Single sharp peak

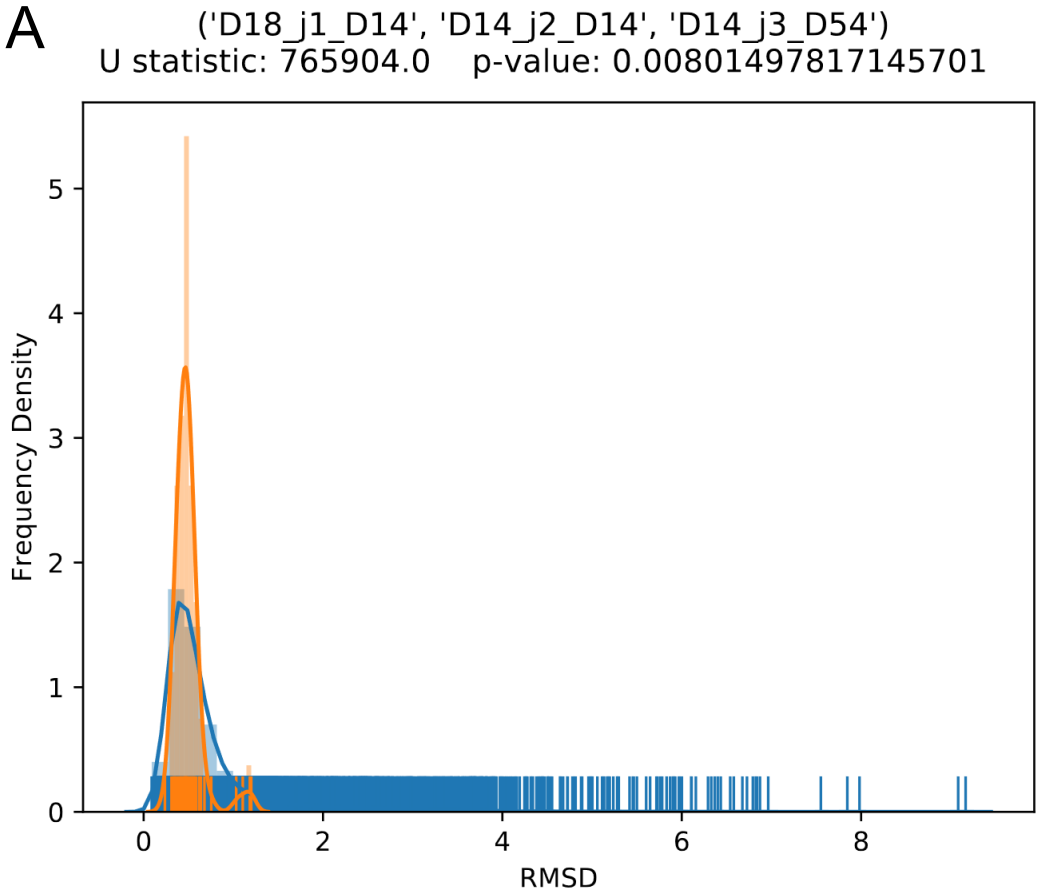


Figure 5 – Graph representing the histogram data for the ["D18_j1_D14", "D14_j2_D14", "D14_j3_D53"] context and the global D14_j2_D14 RMSD distributions. In this graph the colour orange represents the context distribution while blue represents the overall distribution. The marks located at the bottom of the graph represent individual data points. The bars represent the frequency density bars as calculated for the corresponding histogram. The lines represent the kernel density approximation of the histograms. The U statistic and p-value are, respectively, the Mann-Whitney U test statistic and probability as calculated for the two distributions.

Distributions were classified as single sharp peaks (SSP) if they had only a single gaussian peak and an overall standard deviation of less than 0.5 (see figure 5). This distribution type was found in a total of 18.4 % all contexts. Distributions of this type represent a single

conformational state which often only deviate slightly from the original structure, RMSD $\bar{x}$ = 0.37. Despite the rigidity of simple modules as suggested by the overall distributions, see figure 3, only 29.2 % of distributions of this type are made up of simple distributions. This may suggest that despite limited range of movement displayed by simple modules a series of different conformational shifts may be possible. Additionally, no single module is comprised entirely of SSP distributions or even those of equivalent RMSD peaks.

Single broad peak

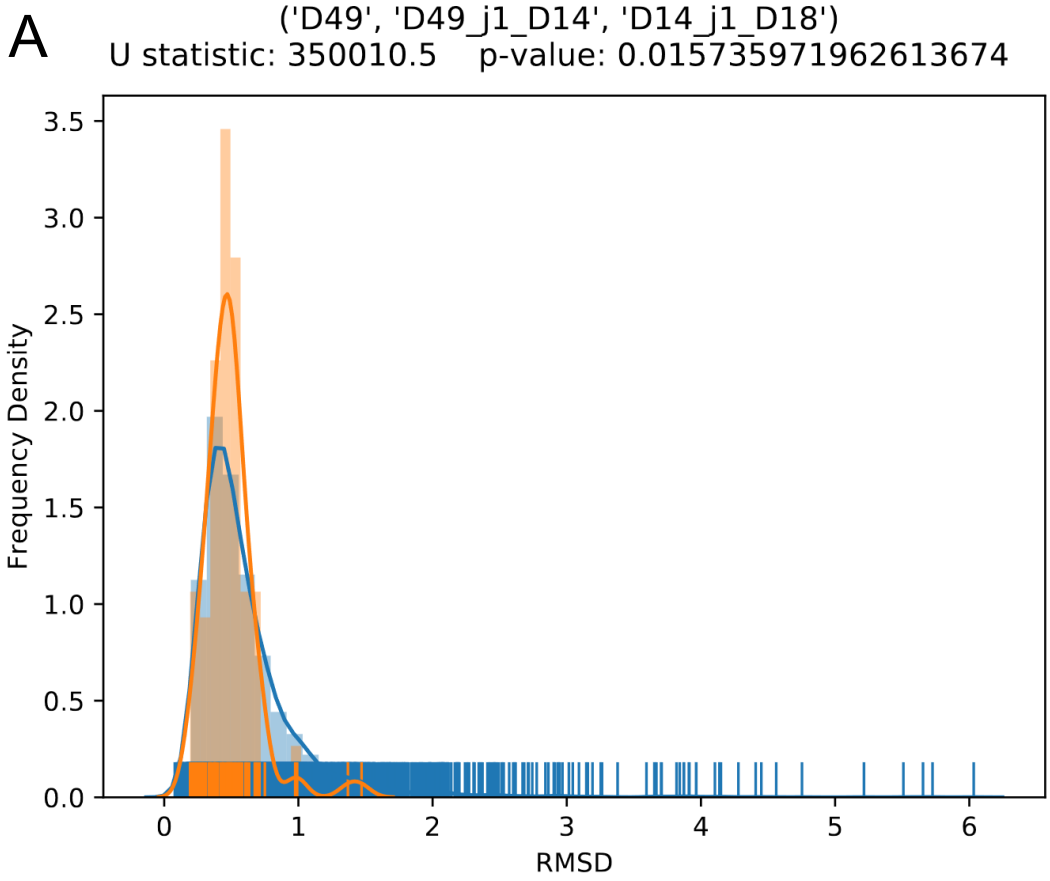


Figure 6 – Graph representing the histogram data for the ["D49", "D49_j1_D14", "D14_j1_D18"] context and the global D49_j1_D14 RMSD distributions. In this graph the colour orange represents the context distribution while blue represents the overall distribution. The marks located at the bottom of the graph represent individual data points. The bars represent the frequency density bars as calculated for the corresponding histogram. The lines represent the kernel density approximation of the histograms. The U statistic and p-value are, respectively, the Mann-Whitney U test statistic and probability as calculated for the two distributions.

Single broad peak (SBP) distributions (see figure 6) have a single gaussian peak and an overall standard deviation of greater or equal to 0.5. Of all three distributions this classification is the least common, found in only approximately 14.6 % of all contexts. In these cases, a continuous set of conformations can be observed in the central module. Rather, more accurately, it is difficult to precisely define a series of discreet conformations by comparing structures using only the RMSD data alone. Surprisingly, 87.4 % of all contexts under this classification contain D14 interface helices. Exactly what may cause this is unknown.

Multimodal peak

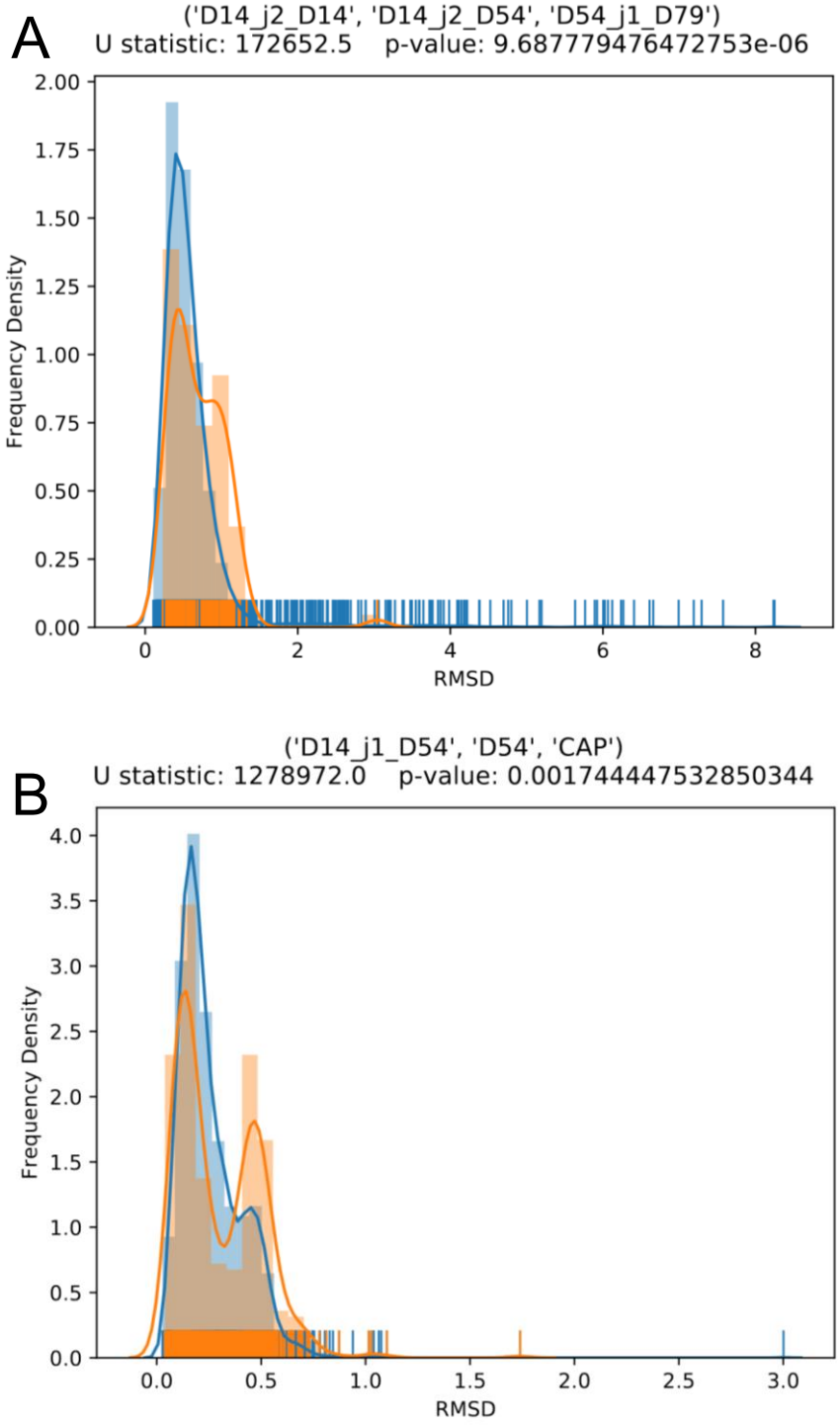


Figure 7 – A) Graph representing the histogram data for the ["D14_j2_D14", "D14_j2_D54", "D54_j1_D79"] context and the global D14_j2_D54 RMSD distributions. B) Graph representing the histogram data for the ["D14_j1_D54", "D54", "CAP"] context and the global D54 RMSD distributions. In both cases, the colour orange represents the context distribution while blue represents the overall distribution. The marks located at the bottom of the graph represent individual data points. The bars represent the frequency density bars as calculated for the corresponding histogram. The lines represent the kernel density approximation of the histograms. The U statistic and p-value are, respectively, the Mann-Whitney U test statistic and probability as calculated for the two distributions.

Multimodal peak (MMP) distributions (see figure 7) were found to contain more than one gaussian peak. This classification also contains distributions where multiple peaks partially overlap creating a single distorted peak (see figure 7A). Potentially due to the broader description used for this distribution type but, approximately 67.0 % all contexts fall under this category. In both cases, the different peaks represent significantly different conformational states.

Owing to the combinatorial construction of the RP library, all possible module contexts are represented, however, contexts describing the middle module of an RP are only found once in the entire library, for example the context [“D14”, “D14”, “D14”] is only found once in rp211. Whereas contexts containing capping domains can be found in multiple proteins, for example [“D14_j1_D54”, “D54”, “Cap”] is found in rp489, rp247, rp60, and rp163. As such a single context distribution (containing a capping domain) may be comprised of multiple RP decoy data. Indeed, 77.8 % of all contexts containing capping domains are MMPs. Given that each RMSD calculation used the input structure as a reference when different RPs are overlaid it may result in different peak RMSDs, this effect may also contribute to some broad peaks observed in SBP as well. To evaluate the effect this may be having on the dataset overall it would be necessary to recalculate all the RMSD values again using a single reference structure for each module.

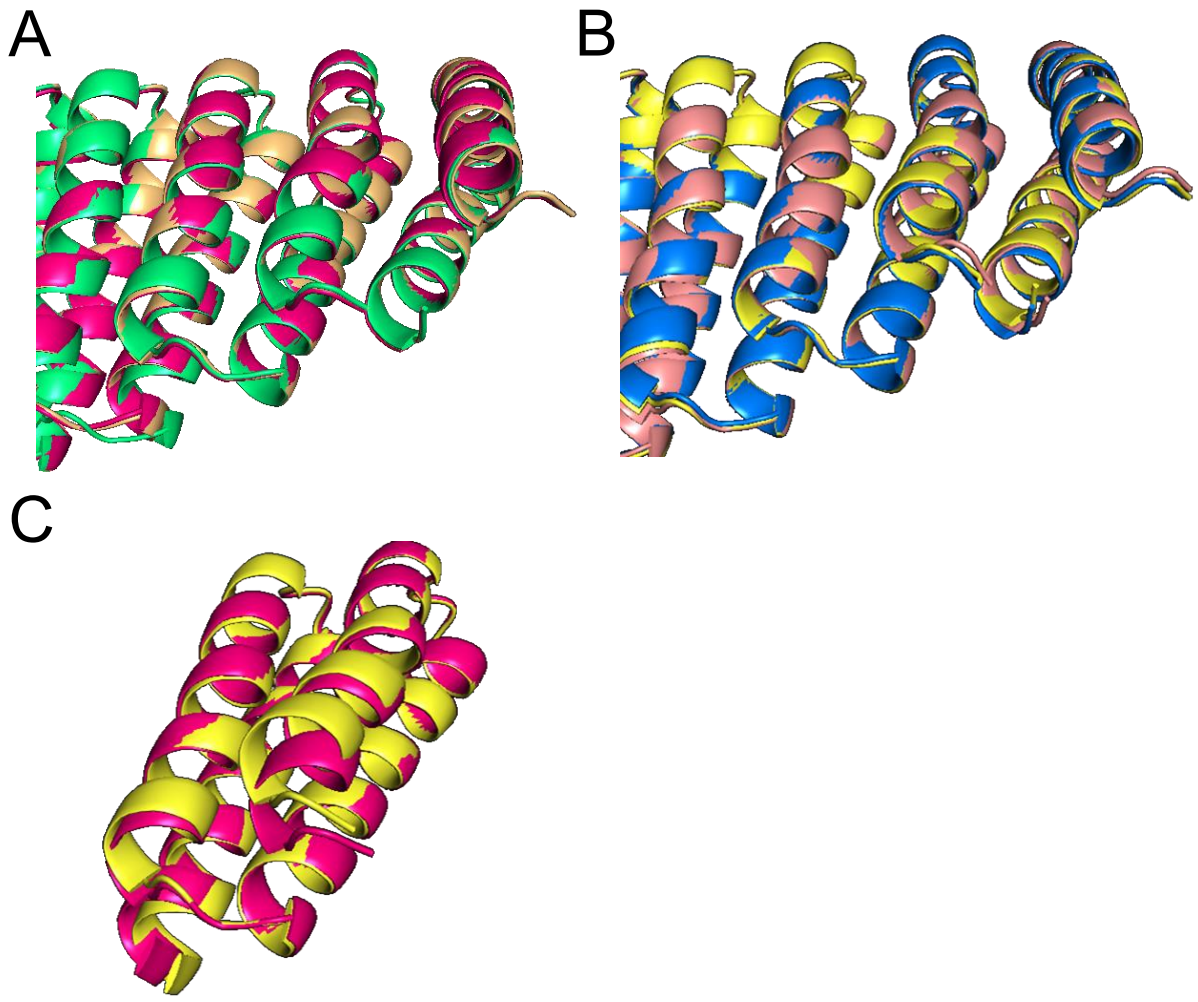


Figure 8 – Representative structures taken from different RPs containing [“D14_j1_D54”, “D54”, “Cap”]. Structures were superimposed against each other using only D54 in each RP. Each colour corresponds to a different RP. A) Structures taken from decoys where the D54 RMSD was

approximately 0.2. Here green = rp489, pink = rp247, and brown = rp60. B) Structures taken from decoys where the D54 RMSD was approximately 0.5. Here blue = rp247, yellow = rp489, pink = rp163. C) A representative structure from both RMSD peaks was selected. Only the structure for D54 from each RP is visible. Here pink corresponds to a decoy from rp489 where RMSD = 0.2 and yellow corresponds to a decoy from rp247 where RMSD = 0.5.

Despite the potential issues resulting from using different references for different distributions some observations can be made. Figure 8 depicts structures pertaining to the peaks displayed by the [“D14_j1_D54”, “D54”, “Cap”] context (see figure 7B) from different proteins. A trend develops whereby each protein containing this context displays preference for one or either RMSD peak. Critically, when structures are taken from each protein and the D54 modules are overlaid it is possible to see that the conformational states indicated by the two peaks are identical. Analysing the RMSD distributions for each protein containing this context separately reveals that some RPs only sample one conformational state. For example, rp60 (NcapD18, D18_j1_D14, D14_j1_D54, D54, CcapD54) is found entirely the 0.5 Å state while rp163 (NcapD14, D14_j2_D14, D14_j1_D54, D54, CcapD54) is only found at 0.2 Å. Other RPs appear to have no preference, being found at either, such as rp247 (NcapD14, D14, D14_j1_D54, D54, CcapD54) and rp489 (NcapD49, D49_j1_D14, D14_j1_D54, D54, CcapD54).

The preference observed by certain modules may corroborate findings found in natural repeat proteins, such that the folds and dynamics of individual repeats are determined by their neighbours and intrinsic folding enthalpies [3]. It suggests the modular domains described in this report may also follow this principle and critically that using only the primary sequence of modules may give an accurate description of the dynamics of the system. However, as some modules demonstrate an inability to adopt one of these conformations it also suggests that an additional factor is determining which conformations D54 is able to adopt. It is possible there is some cascade effect occurring based on the composition of the module to the N-terminal of D14_j1_D54. This N-terminal module may restrict the conformations that can be sampled by D14_j1_D54 and in the process also restrict D54 from adopting all its conformations. This may suggest to accurately determine the overall dynamics of a new RP the overall module sequence must be considered to account for any cascade effects.

Of note, on figure 7B, context [“D14_j1_D54”, “D54”, “CAP”], was classified as an SBP rather than an MMP despite the obvious two gaussian peaks which can be observed. This highlights a potential flaw in the method used to determine the number of peaks for each context. As such, it is necessary to manually inspect each curve to assess whether the distribution type assigned is valid. For future work, one potential improvement to this method, rather than to use an external library which uses a maximal likelihood algorithm to fit a number of gaussian peaks simultaneously, it may be more accurate to recursively fit a gaussian peak to the maximal point of the dataset, then subtract this peak from the dataset. This process would then be repeated until the maximal peak was below a threshold and the number of iterations would indicate the number of gaussian peaks.

Despite the apparent efficiency in using RMSD to differentiate between conformations, extracting an individual representative structure from these distributions has proven to be difficult in some situations. One example can be found when examining a context distribution for D53_j1_D4, [“D53”, “D53_j1_D4”, “D4_j1_D64”]. In this case the RMSD distribution is comprised entirely of decoys from rp282 and displays an SBP shape. Despite this, when 3 decoys with an RMSD of approximately equal RMSD are compared at least three distinct conformations can be observed (see figure 9). Thus, highlighting the issues with using a 1D metric to attempt to finely define 3D structures. This is likely to be a particular problem when RMSD values are larger and when modules are longer, as is the case in complex modules,

as all atoms in the module contribute to the metric. The issues this situation present are further exacerbated when the modular nature of the RPs are considered. Any number of modules may be placed in sequence to generate a protein, as such, in longer proteins, small changes in structure can be propagated resulting in much larger changes across the entire structure. In particular this problem highlights that obtaining a small set of conformations from a large ensemble using only RMSD is difficult as such a more fine tuned method was developed to better differentiate between structures.

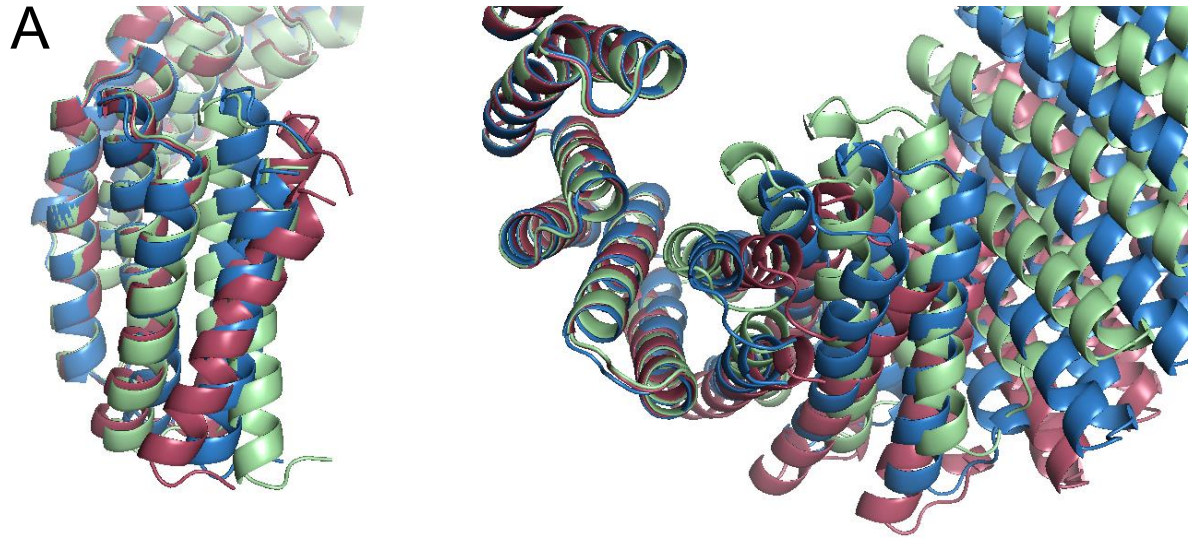


Figure 9 - A) Overlaid structures of 3 different decoys of rp282. All 3 decoys were overlaid over each other using the D53_j1_D4 module as reference. The RMSD for the decoys are as follows; blue = 2.725, green = 2.865, red = 2.634. Left) The model depicts only residues consisting of D53_j1_D4 and its N-terminal modules. The nearest helices correspond to the D4 C-terminal interface helices of D53_j1_D4. Right) A partial image of the total protein structure. The structure has been rotated to give an overhead view of the structure compared with left.

Dynamatch – defining conformations using rigid body transforms

Due to ppmp's coarse grain representation of conformational changes in protein structure and the necessity to be able to derive a conformational state directly form the large RP library, a more accurate description of the differences between conformational states was required. Hence, Dynamatch was developed which could both discern subtle differences in protein structure and highlight the most significant changes in structure. Dynamatch describes perturbations in decoy structures as a series of rigid body transformations to better enable structure clustering whilst also reducing the complexity of the original datasets to maintain a useful speed and efficiency. By converting the change in structure between the reference and decoy structure to a 6N dimensional point, where N indicates the number of secondary structure elements in the module, it is possible to accurately cluster protein structures to generate a small set of conformations. See Methods for further information. To improve the performance of the clustering algorithms each 6N dimensional point is reduced to a 3-dimensional point using principle component analysis. This dimensionality reduction reduces the number of calculations needed by each clustering algorithm dramatically reducing their computational expense.

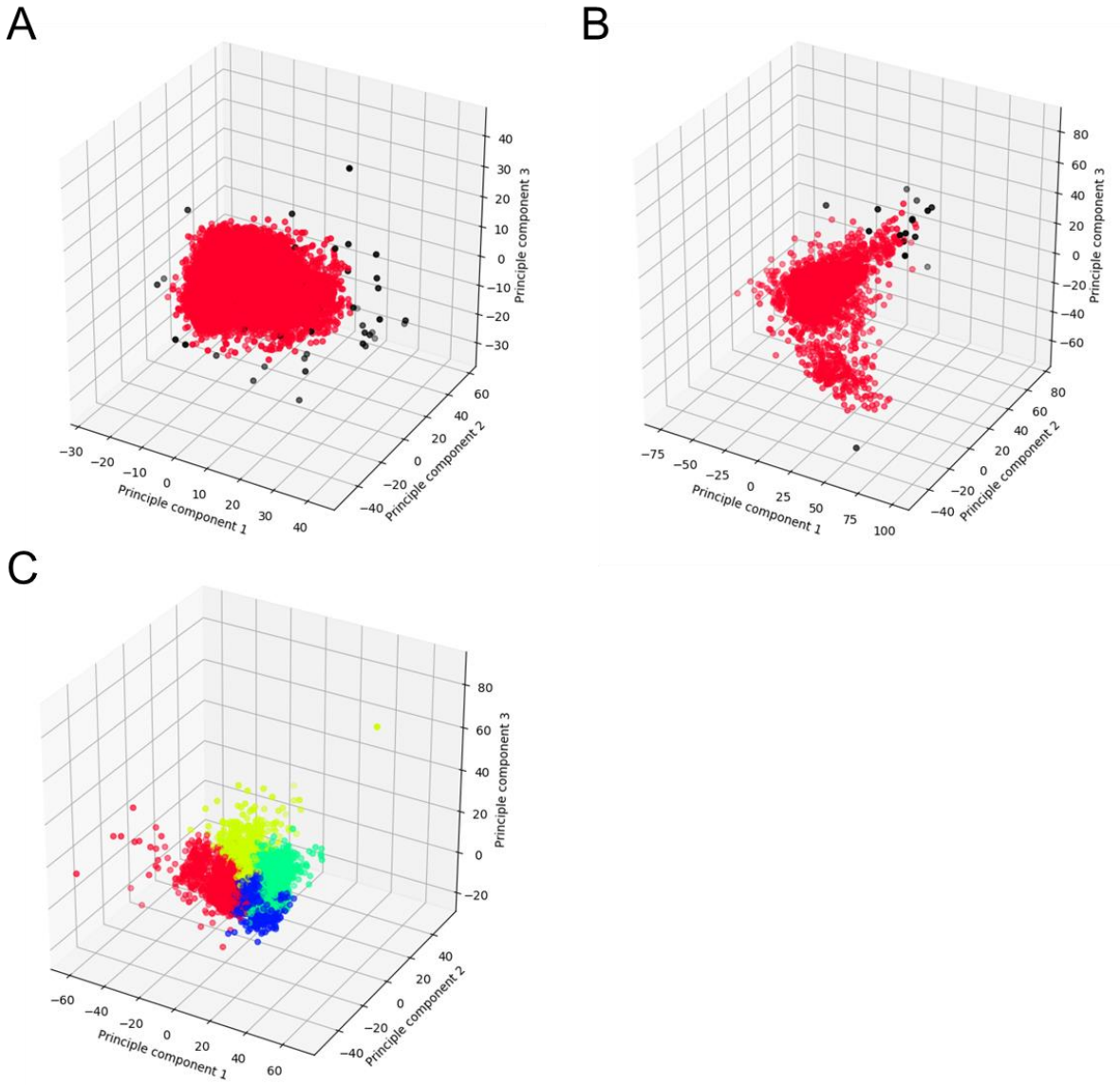


Figure 10 – PCA transformed datasets points were plotted against the first 3 principle axis of each dataset. The colour of each point indicates the cluster to which the point was assigned to after clustering. Points in black correspond to outliers, as determined during the clustering algorithm. A) represents the PCA transformed dataset for the module D14, which has a disperse cloud-like shape. B) Represents the PCA transformed dataset for module D79_j2_D14, which has a incorrectly clustered dataset. C) Represents the PCA transformed dataset for the D49_j1_D14 module, which has a correctly clustered dataset.

To accurately determine a small subset of conformations which are representative of the whole set the optimal clustering algorithm must be chosen. Unfortunately, the most appropriate clustering algorithm is largely determined by the shape of the dataset and the shape of the distributions for each module was found to be varied. Therefore, no single clustering algorithm would be appropriate for use on all datasets. Figure 10 demonstrates how different distributions are possible across the modules. These distributions, which when clustered, can be defined as Disperse Cloud-Like (DCL), Incorrectly Clustered (IC), and Efficiently Clustered (EC). Of particular interest are modules such as D14 which demonstrate DCL distributions. These distributions may be indicative of potential issues with underlying dataset or the methods used to reduction methods used on them. Additionally, a continuous dataset further

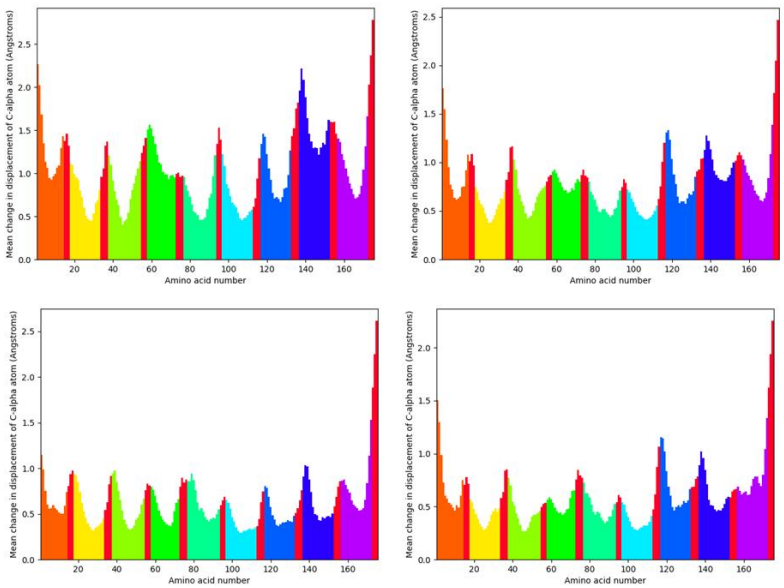
complicates clustering as it reduces the signal to noise ratio making it very difficult to select clusters without arbitrarily dividing data.

One explanation for these continuous distributions may also be a result of the PCA dimensionality reduction. PCA uses a covariance matrix to generate a series of linear principle axis (created from eigenvectors) which minimise the unexplained variance of each point. As a result, PCA has been shown to be a poor algorithm for dimensionality reduction for certain shaped high dimensional distributions. One potential method to alleviate this issue is to use kernel PCA, KPCA, instead. KPCA uses kernel methods to first map the dataset to an equal or higher dimensional feature space before extracting the new feature space's principle components [37]. By first mapping the data to feature space first it can properly separate data that would not otherwise be linearly separable.

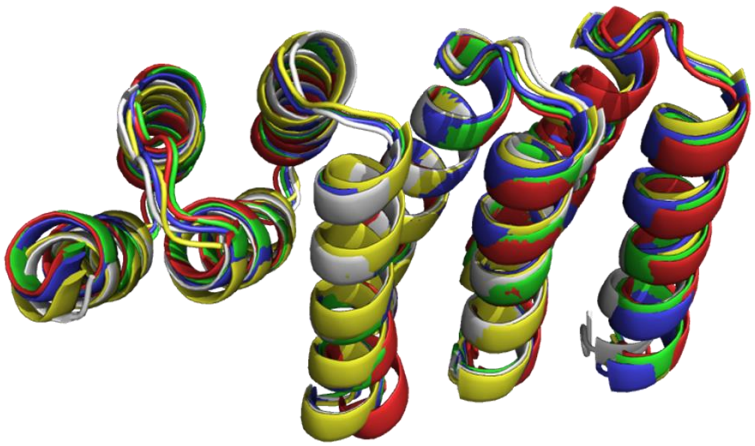
Another issue potentially giving rise to DCL distributions is the definition used for SSEs. Currently, each helix is defined as a single point with a plane, which reduces a large proportion of the data stored in the helix. For example, this definition would struggle to efficiently differentiate between a kink in a helix and a rotation about the centre of mass. Therefore, each point would map to a similar point in space, making separation of these points difficult without another means to differentiate between them. To better capture these differences helices could be defined as a polynomial line function which could then be compared instead. Unfortunately, due to the complexity of the data assessing the effectiveness of the strategies listed above would be difficult. However, future work to improve the clustering setup could implement some of these methods. A final consideration would be to use supervised machine learning instead of clustering, however the effectiveness of this is questionable due to the relatively small sample sizes present (compared to the sample sizes required for accurate machine learning methods).

54.5 % of modules were found to have 1 cluster. But, based on ppmp data this would be unusual as every module was shown to have a small number of distributions which altered significantly from the majority, indicating modules are able to exhibit at least a small subset of conformational states. Manually inspecting these distributions reveals that they demonstrate distributions similar to D14 or D79_j2_D14, figures 10A and 10B. These modules represent either DCL or IC distributions. Unlike DCL (figure 10A), IC (figure 10B) distributions highlight potential problems with the selection of the best clustering parameters. IC are most prevalent when sets of dense clusters are found in close proximity, as can be seen in figure 10B. Analysing the gap statistic across all clustering algorithms shows that in each clustering algorithm 1 cluster is favoured. Additionally, there is a minimal difference between the best and second-best parameter set. In general cluster analysis, determining the fitness for a cluster size of 1 is impossible as fitness metrics will often use inter-cluster variance to calculate fitness. The gap statistic, however, uses the dispersion between all points in a cluster allowing it to calculate a fitness for a cluster size of 1. For this reason, the gap statistic was originally chosen as a fitness metric. However, given the poor clustering of IC distribution it may demonstrate how using the gap statistic alone may not be appropriate and special consideration must be given to parameters leading to a cluster size of 1. One potential solution is to remove such parameters or to reduce the gap statistic in cases where 1 cluster is found. Overall, this issue can be amended through a series/combination of methods to mitigate the impact of single clusters.

A



B



C

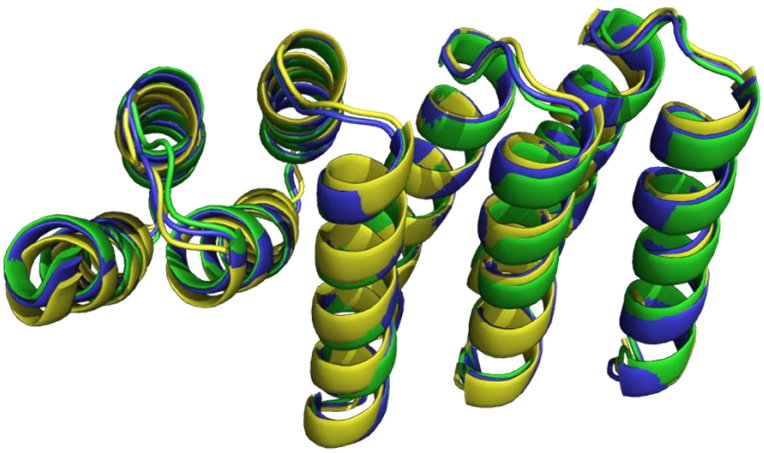


Figure 11 – A) Mean deviation of each amino acid in the D49_j1_D14 module, each graph corresponds to the mean distribution for each of the four clusters. In order from left to right, top to bottom, the graphs correspond to the mean per amino acid distributions for c0, c1, c2, and c3. The

different colours correspond to different helices where red corresponds to amino acids found in loop regions. B) Overlaid structures for the mean averaged structures for D49_j1_D14 generated for each of the 4 clusters. Structures were superimposed on each other using the D49_j1_D14 crystal structure. The colours correspond to the following colours: c0 = red, c1 = green, c2 = yellow, c3 = blue, D49_j1_D14 crystal structure = white. C) Overlaid structures for the mean averaged structures from c1, c2, and c3. Structures were overlaid using c3 as a reference. The colours correspond to the following colours: c1 = green, c2 = yellow, c3 = blue.

Despite the issues presented by clustering, analysing EC distributions give new insights into the behaviours exhibited by the modules. We analysed the mean per amino acid deviation across entire modules. In each conformation different amino acids deviate more than the average across clusters. Specifically, the amino acids are centred about sets of helices indicating an overall helix motions. Extracting structures which deviate the least from these average structures from each cluster can help to identify reasons for these differences. Calculating the RMSD of each cluster, c0 = 1.18, c1 = 0.74, c2 = 0.61, c3 = 0.59, finds that they roughly lie in accordance with the general ppmp peak distributions for the D49_j1_D14 module. Critically, however, despite the RMSD of c1, c2, c3 being similar when being compared against the D49_j1_D14 crystal structure, when they are compared against each other the RMSD is significantly different. When the structure for cluster 3 is used as a reference, the RMSD for c1 = 0.44, c2 = 0.80. Manual inspection of each of the structures gives a structural overview of why this is occurring, see figure 11. For cluster 0 the larger changes in structure are a result of a kinking of helix 5 which is then propagated along the module, manifesting as shifts in the terminal 3 helices. Cluster 2 also exhibits a much larger deviation again making it easy to determine the differences being used for classification. In this case a large concerted shift in the first 3 helices of the structure can be seen. These helices have been rotated against the helices 4 and 5 in a concerted manner. For clusters 1 and 3 it is difficult to extract exactly what the critical structural differences are. In this instance minor changes to the curvature of the helices, particularly at the termini, may be the cause with more notable differences being seen in helices 7 and 8. In this instance it is critical to note that Dynamatch clusters structures based on the overall deviation in structure rather than a single motion. So, it is possible that the classification was made due to a pattern of specific perturbations which are only observed in one cluster and other deviations are less important.

Dynamatch can efficiently capture subtle alterations in structure but also much larger motions in specific helices, as seen in figure 12 for module D53_j1_D4. For D53_j1_D4 the RMSD for each cluster was found to be very similar, c0 = 0.48, c1 = 0.96, c2 = 0.87, c3 = 0.80. Excluding c0, all clusters have an RMSD within 0.2 Å of each other and as can be seen from figure 12 both c0 and c3 display significant shifts in a small subset of helices. Both clusters demonstrate large shifts in helix 9 and in c3 helix 8 has also been significantly shifted. However, it was unable to extract all the different structures seen present in rp282, which was discussed previously. In the case of rp282 the perturbations of D53_j1_D4 all occurred in a similar direction, i.e. the rotation of helices largely occurring in the same directed manner but what separated structures from each other was the magnitude of this movement. This may be a failing in PCA clustering as each structure is reduced in a linear fashion.

Despite the minor differences observed the small deviations in structure are significant. While perturbations in a single module may prove to be minor in the singular context of the module, when these perturbations are accumulated across the length of a whole RP it can lead to very large changes to the overall geometry of the protein (see figure 9). However, determining the exact boundary between conformational states is difficult particularly as proteins are unlikely often act in discrete steps. One method may use a metric such as RMSD between the conformations gathered and discard states which are too similar, but it may discard structures where only a small number of helices shift. As such, it stands that the conformations which

are extracted by Dynamatch are acceptable even if some would be considered the same as the reduction in the total ensemble is still extremely high with an average of < 1 % of structures being extracted for each module. Therefore, Dynamatch is still able to dramatically reduce the volume of structures which need to be processed.

The ability for Dynamatch to discern between these minor changes is also promising for future functionalisation studies as very minor alterations in structure can be critical in facilitating a functional protein. Based on this evidence it suggests that Dynamatch can efficiently classify different types of motions even between very similar structures, such as rotations in different directions (with respect to the reference) or significant kinking of helices.

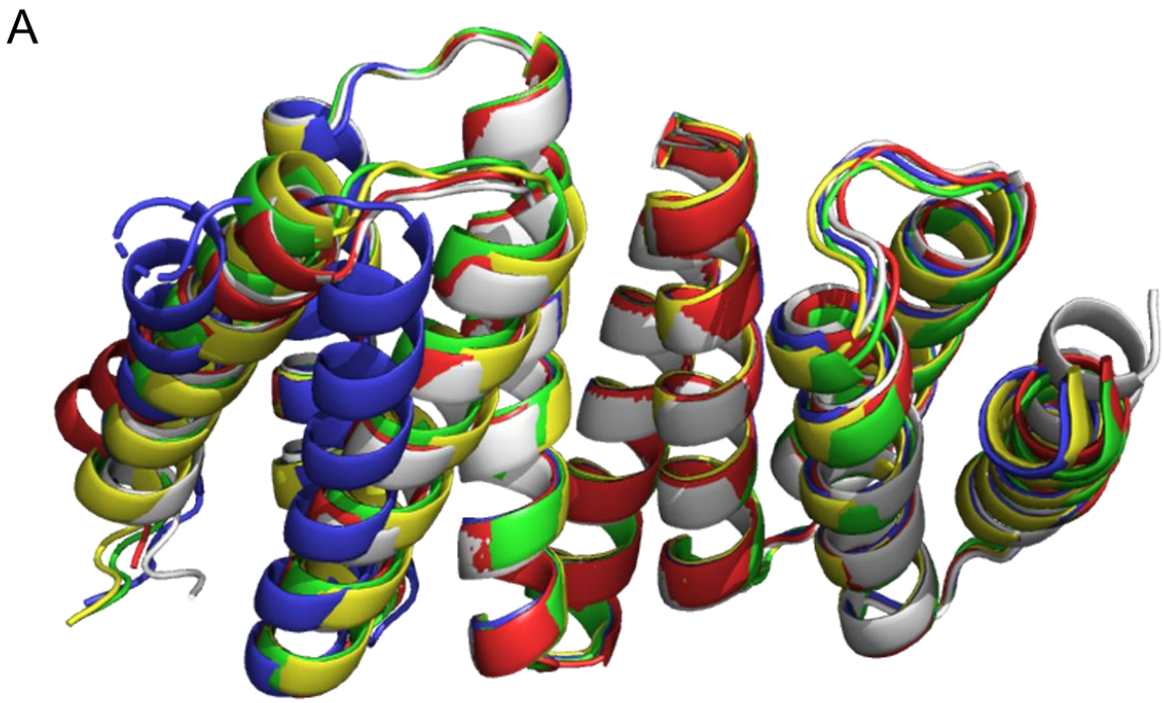


Figure 12 – A) overlaid structures of the mean averaged structures for D53_j1_D4 generated for each of its 4 clusters The structures were superimposed against the D53_j1_D4 crystal structure. The colours correspond to the following colours: c0 = red, c1 = yellow, c2 = green, c3 = blue, D53_j1_D4 crystal structure = white.

As eluded to previously, Dynamatch can provide information regarding the range of conformations each module can adopt, in contexts. This is especially useful when considering design aspects with larger structures as it would then be possible to provide whole protein dynamics by the additive effects of each module. Surprisingly nearly every RP containing D49_j1_D14 was found in nearly every cluster described above and only one RP was found in only one cluster, rp470 (NcapD49, D49_j1_D14, D14_j1_D14, D14, CcapD14). Equally structures from every context are present in at least two of the different clusters. Similarly with D53_j1_D4, most proteins are found in nearly every cluster as are contexts. The 3 exceptions to this are rp282 (NcapD53, D53, D53_j1_D4, D4_j1_D64, CcapD64) and ['D53', 'D53_j1_D4', 'CcapD4'] which are found exclusively in c0 and rp281 (NcapD53, D53, D53_j1_D4, D4, CcapD4) found only in c1. This promiscuity is corroborating the overrepresentation of MMP distributions found during ppmp analysis. However, it also highlights the issues with using RMSD as the sole determinant for differentiating structures. As despite having similar RMSD values the conformations can fall into different clusters. Indeed, the trend is observed in all modules where more than 1 cluster was calculated.

Looking more generally at the structural changes observed in each of the clusters it appears that larger RMSD changes arise from concerted shifts in a subset of helices. These shifts may be propagated along the module as was the case in c0 of D49_j1_D14 or in other cases may be large movements of the terminal helices of the module, as is the case in c0 and c3 of D53_j1_D4. Where the deviation is cause either by a rotation in helix 9 or a cooperative rotation of helices 8 and 9 in c0 and c1, respectively. In accordance with this, in nearly all modules the largest average helix motions are seen in the terminal helices, see table 2. Additionally, increases in the centre of mass deviation (see table 2) in these terminal helices tend to coincide with increases in the neighbouring helices perhaps suggesting a cooperative motion of several helices.

	Average Helix CoM deviation (Angstroms)									
Module	helix 1	helix 2	helix 3	helix 4	helix 5	helix 6	helix 7	helix 8	helix 9	helix 10
D14	0.884	0.442	0.634	0.424						
D14_j1_D14	0.705	0.523	0.341	0.277	0.226	0.276	0.282	0.252	0.299	0.569
D14_j1_D18	0.913	0.645	0.558	0.407	0.407	0.493	0.503	0.335	0.562	
D14_j1_D54	0.953	0.359	0.334	0.452	0.339	0.287	0.367	0.445	0.352	
D14_j1_D76	1.049	0.687	0.588	0.438	0.522	0.481	0.428	0.469	0.675	
D14_j1_D79	0.764	0.702	0.660	0.420	0.434	0.289	0.476	0.441	0.643	
D14_j1_D81	0.896	0.725	0.668	0.373	0.306	0.407	0.274	0.347	0.483	
D14_j2_D14	0.999	1.019	0.500	0.539	0.432	0.565	0.598	0.552	0.535	0.910
D14_j2_D54	1.515	0.812	0.692	0.446	0.660	0.530	0.630	0.578	0.688	
D14_j2_D71	0.676	0.502	0.406	0.327	0.331	0.348	0.213	0.363	0.452	
D14_j2_D79	1.102	1.277	0.998	0.927	0.565	0.556	0.691	0.557	0.795	0.979
D14_j3_D54	1.437	0.517	0.615	0.545	0.704	0.487	0.380	0.398	0.599	
D14_j4_D79	1.722	1.002	0.788	0.906	1.005	0.542	0.756	0.911	0.549	0.803
D14_j5_D79	1.077	0.731	0.733	0.716	0.447	0.578	0.436	0.604	0.865	
D18	0.294	0.309	0.205	0.280						
D18_j1_D14	0.861	0.263	0.535	0.422	0.452	0.672	0.545	1.197	1.650	
D4	0.703	0.517	0.573	0.833						
D49	0.204	0.107	0.180	0.282						
D49_j1_D14	0.694	0.287	0.331	0.655	0.483	0.272	0.428	0.624	0.546	
D49_j1_D79	0.851	0.588	0.612	0.456	0.566	0.649	0.544	0.716	0.643	
D49_j1_D81	0.454	0.243	0.390	0.285	0.466	0.425	0.478	0.401	0.357	
D4_j1_D64	0.681	0.902	0.450	0.436	0.490	0.398	0.324	0.400	0.406	
D53	0.124	0.188	0.283	0.204						
D53_j1_D4	0.913	0.671	0.450	0.587	0.544	0.639	0.828	0.968	0.869	
D53_j1_D79	0.895	0.496	0.550	0.514	0.589	0.617	0.623	0.739	0.680	0.924
D54	0.304	0.465	0.686	0.224						
D54_j1_D79	0.647	0.434	0.394	0.338	0.416	0.440	0.555	0.770	0.547	
D64	0.172	0.143	0.197	0.160						
D71	0.232	0.089	0.226	0.245						
D76	0.398	0.455	0.381	0.542						
D79	0.513	0.282	0.406	0.298						
D79_j1_D54	0.775	0.488	0.366	0.350	0.329	0.514	0.277	0.423	0.661	
D79_j2_D14	0.505	0.656	0.389	0.465	0.459	0.306	0.340	0.322	0.763	

Table 2 – Table containing the average deviation in the centre of mass of each helix with respect to the reference. The average was calculated as the mean across all decoys of the module.

Due to the varied nature of the behaviours exhibited by each module assessments of the efficiency of the clustering as well as the precise differences between each cluster need to made before overarching statements can be accurately made across the entire set of modules. However, the use of Dynamatch can dramatically reduce the workload required to analyse large datasets of differing structures. In particular it can highlight key helices which may become the subject of future design work or help eliminate modules which too much motility or instability.

Validation of Rosetta computational models

Validation of Rosetta models with Molecular Dynamics

A total of 6 RPs were selected from the RP library to undergo MD simulations. Proteins were manually selected based on the modules they contained. The RPs chosen can be seen in table 3.

RP	Module composition of RP
rp2	NcapD64, D64, D64, D64, CcapD64
rp110	NcapD14, D14_j2_D14, D14_j2_D14, D14_j1_D81, CcapD81
rp209	NcapD14, D14, D14, D14_j2_D79, CcapD79
rp281	NcapD53, D53, D53_j1_D4, D4, CcapD4
rp282	NcapD53, D53, D53_j1_D4, D4_j1_D64, CcapD64
rp628	NcapD79, D79_j2_D14, D14_j1_D14, D14_j1_D18, CcapD18

Table 3 – Table containing the names and module composition of the RPs selected for MD simulations.

RPs were selected to ensure a variety of contexts, and distribution types were represented. In particular, rp2 was selected as it contained only simple modules and both rp281 and rp282 were selected as they both contained D53_j1_D4 but it displayed significantly different behaviours between the two. Therefore, validation of these proteins should help to determine the level of accuracy with which Rosetta is able to reproduce the dynamics of systems with only slight variations. Calculating the RMSD of the protein against the first frame of the simulation. By analysing the RMSD time course it is possible to determine at what point the simulation reaches equilibrium. Potentially due to the accuracy of Rosetta modelling, but most RPs can be seen to be in equilibrium throughout the simulation, as determined by the general fluctuations present in each time course. Of note, rp2 and rp209 display equilibration steps which can be seen by the steady increase in RMSD over the initial time steps. Therefore, only the last 80 ns and 60 ns should be used for equilibrium analysis of rp2 and rp209 respectively.

A

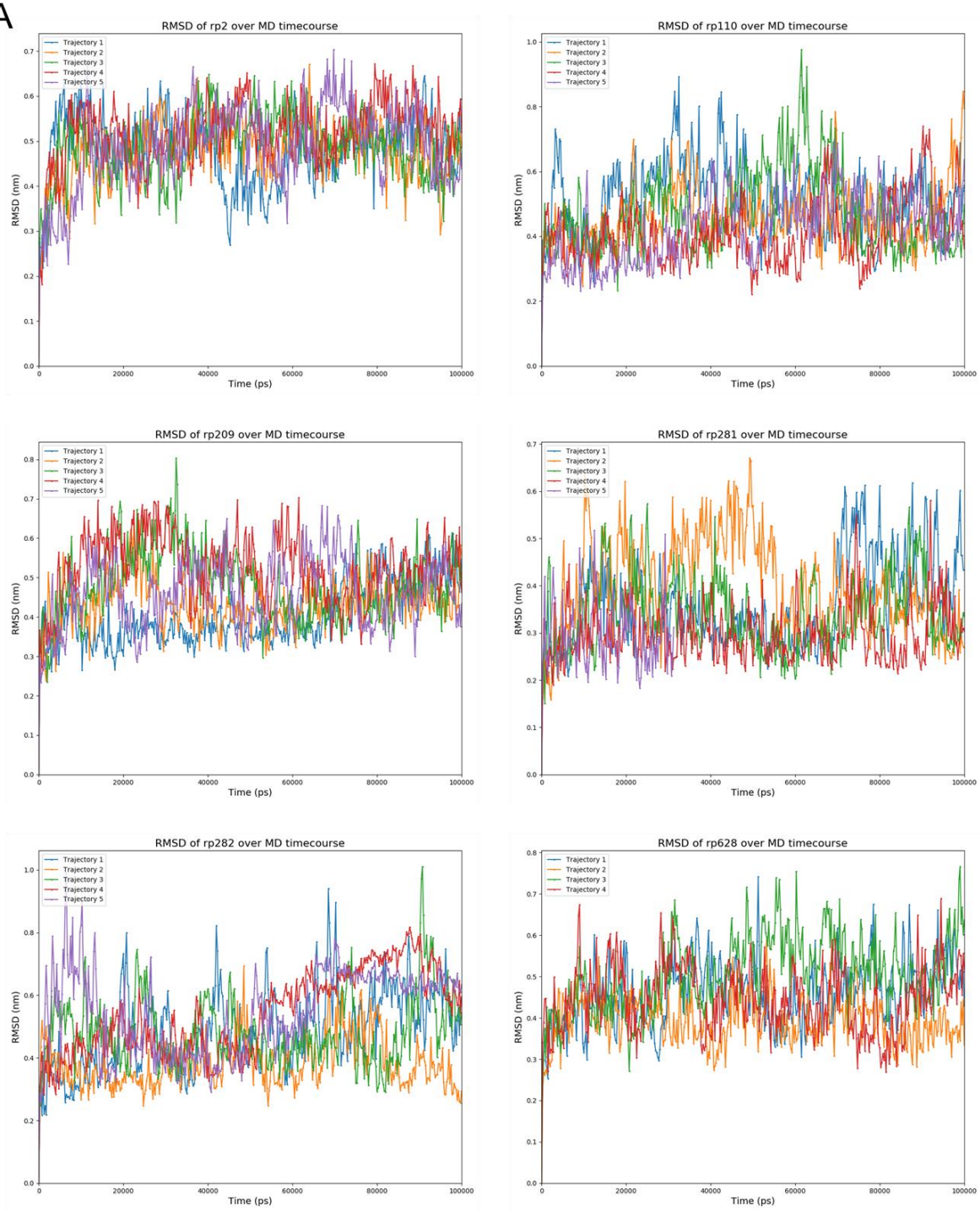


Figure 13 – A) RMSD of each RP across the time course of the MD simulations. RMSDs were calculated using the first frame of the simulation as a reference. From left to right, top to bottom the graphs represent the RMSD calculated for rp2, rp110, rp209, rp281, rp282, rp628.

Due to limited time, only minimal examination of each time course was achieved, with an emphasis placed on identifying unusual patterns or behaviours. As such, manual inspection of the RMSD plots (see figure 13) were used to attempt to identify abnormalities within the simulations. In particular sustained changes in RMSD may indicate a more permanent conformational change in the protein. From inspection of the RMSD across the time course of each MD simulation an interesting observation can be seen for rp282, whereby there is a sustained increase in RMSD towards the end of the time course in trajectories 4 and 5.

Inspection of the simulation leads to the conclusion that a pair of salt bridges is able to form which locks the conformation of rp282 into a closed toroid-like shape. Despite the nearly identical sequence homology with rp281, this same effect is not replicated in simulations of this RP. It seems likely that the initial salt bridge formed between Glu35 and Arg624 (figure 14A) in the second and final helices of rp282 may act to hold together the terminal ends of the RP.

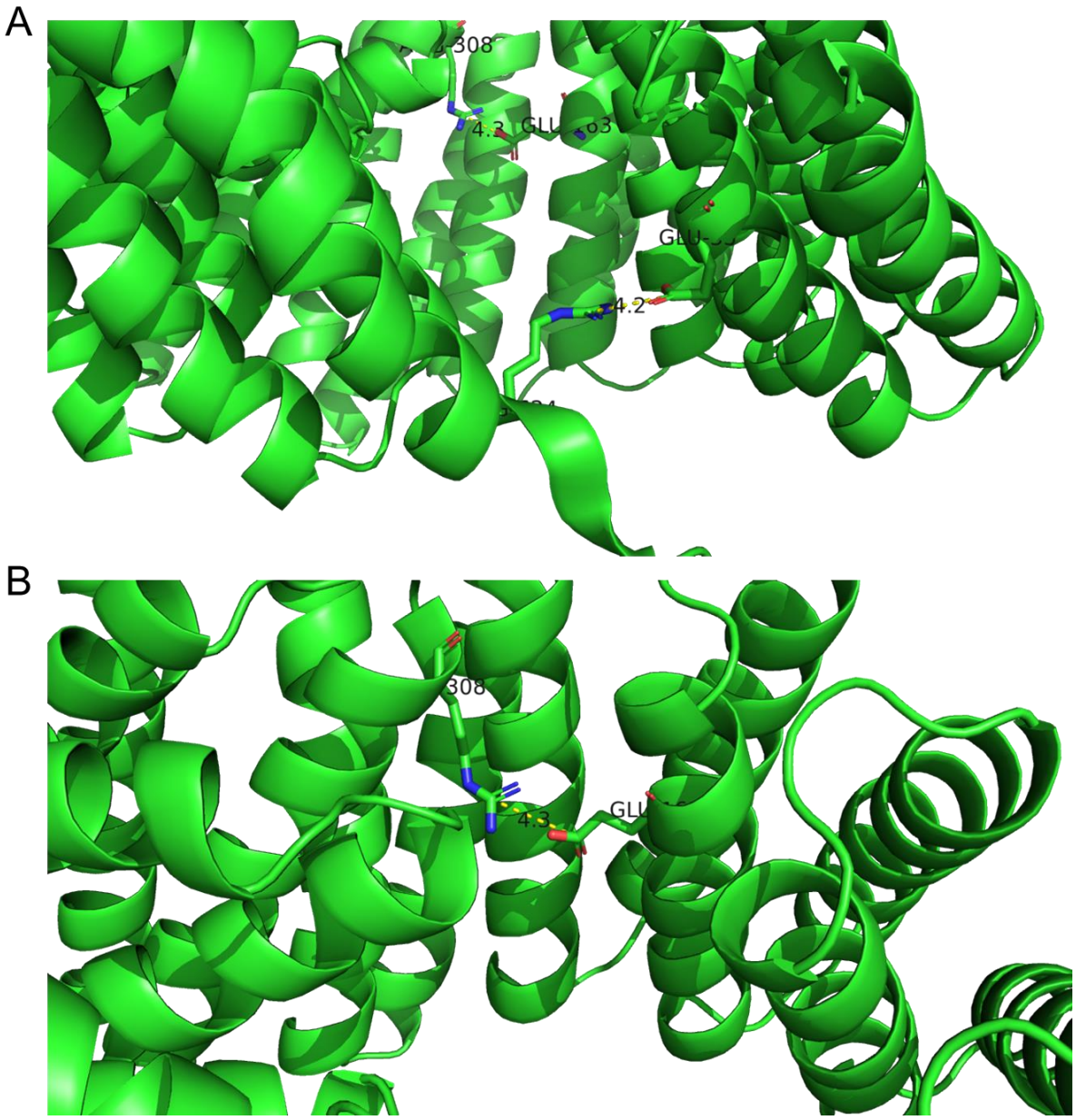


Figure 14 – Structure of rp282 from the final frame of trajectory 4. A) Depicts a partial structure of rp282, residues Glu35, Arg624, Glu163, and Arg308 are visible and the Euclidian distance between the two pairs has been labelled. B) Depicts a partial structure of rp282, residues Glu163 and Arg308 are visible and the Euclidean distance between the two residues has been labelled.

The formation of this salt bridge seems to be mediated largely by an additional salt bridge formed between Glu163 and Arg308 (see figure 14B), which are both part of the module D53_j1_D4. This salt bridge can also be seen in c0 of D53_j1_D4 (see figure 12). Importantly, Dynamatch was able to demonstrate that rp281 was unable to adopt this conformation. Indeed, ability of Rosetta and Dynamatch to be able to identify this critical structure provides

evidence for the usefulness of this approach. Critically, however, the salt bridge which fully stabilises the toroid conformation (between Glu35 and Arg624) is formed between the NCapD53 and CCapD64and highlights the importance in considering the whole protein in addition to each module. This will be especially important in much larger proteins were modules which may not be able to be placed adjacent to each other in primary sequence will be brought in close proximity spatially allowing them to interact in ways which are not considered currently using Dynamatch.

Future analysis utilising comparisons, using Dynamatch and ppmp, of rp628 simulations in MD and Rosetta will lead to more definitive explanations for whether the protein ensemble being sampled by Rosetta can be reasonably expected to be present *in vitro*.

Experimental Validation and Controllable Dynamics – Expression, Purification and Characterisation

In an attempt to control the dynamics of the modular protein system a molecular switch was designed using the each of the modules. Using Rosetta Match a Zinc binding site was designed against all RPs which exhibited a large enough dynamic range. Hits were the filtered such that only RPs where Zinc binding occurred in only one conformation remained. After a final manual inspection of the best 15 hits, rp628 (NcapD79, D79_j2_D14, D14_j1_D14, D14_j1_D18, CcapD18) was found to be the most promising target. The three mutations were introduced to enable Zinc binding. The “wild type”, WT, rp628 and triple mutant, 3M, rp628 were then chosen for characterisation to assess Zinc binding and its ability to modulate the predicted dynamics of rp628.

At the time of writing, all RPs which have been successfully expressed have also been purified, with initial data suggesting the RPs are fully folded and stable. Consistent with previous data, a 6His tagged WT and 3M rp628 were able to be expressed and purified. Similar to the behaviour displayed by the modules when they were purified in [7] rp628 of both forms were found to run at a lower weight band in SDS-PAGE gels, see figure 15. The initial Ni chromatography purification was found to be insufficient to remove all impurities, see figure 15A. Normally an ion exchange chromatography would be used to further purify the protein, however, rp628 was found to reversibly precipitate in salt concentrations lower than 500 mM NaCl. As such only SEC was used to further purify the protein. Due to the high ratio of 260/280 nm absorption of found at peaks 1 and 3-5 during the size exclusion, these peaks were determined to be largely nucleic acid contaminants. The was then corroborated by the SDS-PAGE gel whereby no proteinaceous bands were found, see figure 15B. Peak 2 was found to be rp628. While peak 2 appears to be unimodal indicating rp628 was stable and folded slight bands can be seen in figure 15B. These bands likely resulted from partial digestion of rp628 by residual protease contaminants found in the Ni chromatography column. This partial digestion is may have occurred in minor structural elements, such as loop regions, as the protein is able to maintain it’s overall fold as indicated by the single peak during SEC. Additionally the supernatant band in figure 15B has a noticeably weak intensity. This is likely caused by the low overall concentration of protein as the solid fraction was diluted into 1 mL of buffer.

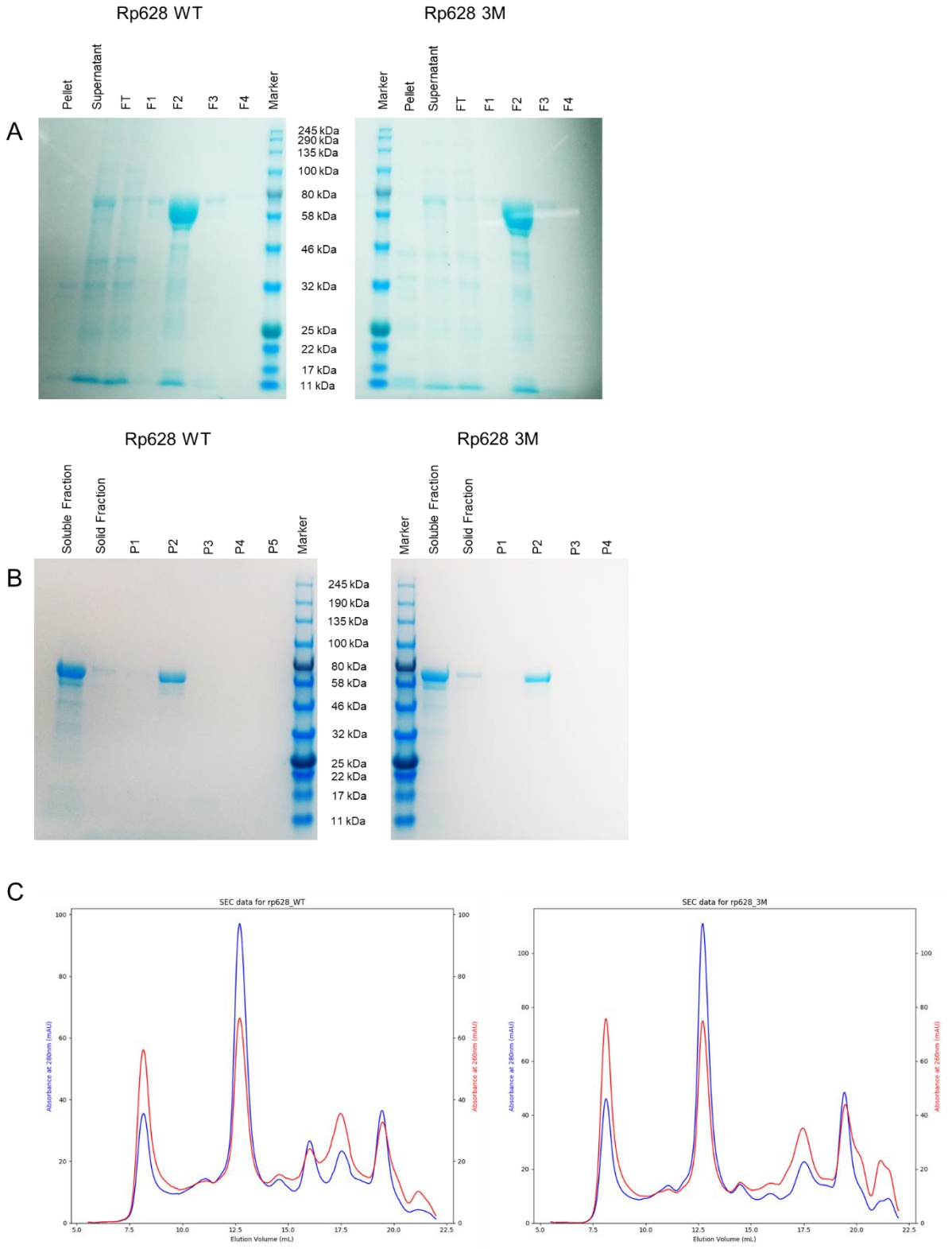


Figure 15 - rp628 is approximately 80kDa. Due to rp628's long linear shape it was found to run below its expected weight on the SDS-PAGE gel. A) SDS-PAGE gel for the Ni chromatography purification of WT and 3M rp628. The lanes, in order, are as follows: solid fraction of E. coli pellet; soluble fraction of E. coli pellet; flow through from Ni column purification; the next 4 columns are the 1.5 mL fractions eluted from the Ni column. B) SDS-PAGE gel for size exclusion chromatography fraction for the purification of WT and 3M rp628. The lanes, in order, are as

follows: soluble fraction following dialysis; solid fraction following dialysis; the next 5 columns correspond to the 5 peaks described in figure 15C. C) 260/280 nm absorption of size exclusion elution. The line in blue corresponds to the 280 nm absorption while the red line corresponds to 260 nm absorption. The graph on the left is rp628 WT and on the right is rp628 3M.

The purified protein will be sent to the Diamond Light Source, UK, where the protein can then be characterised. Currently no crystal structures have been obtained for the expressed proteins.

Conclusions

Through analysis of each module we have found that each module demonstrates a general pattern of behaviours, in which modules are able to adopt a varied range of conformations. The range of conformations which each module can adopt appear to be influenced by the neighbouring modules and in accordance with this the effect cascading effect of modules in longer RPs deserves investigation.

Using rigid body transforms to map structural transformations of modules presents a new tool to enable the rapid characterisation of modular protein dynamics and marks a significant increase in precision over using coarse grain metrics such as RMSD. The ability of Dynamatch to select representative conformational states from an ensemble has proven to be effective in both the reduction of structures which need to be examined and the range of structures it is able to differentiate between. It has demonstrated considerable practical improvements over ppmp in efficiency in which structures can be elucidated. The success found using a combination of Rosetta modelling and ensemble clustering using structural perturbations underscores its ability to efficiently reduce workloads in larger pipeline projects. In particular, it serves as a useful step in bringing about a more comprehensive understanding of the dynamics of new *de novo* RPs at a conceptual level.

Further work in aiding the ability of edge cases in clustering and definitions of SSEs should be considered to improve the performance of the approach outlined and enable better differentiation between similar structures. Comparison of select RPs will be able to provide more definitive evidence for whether Rosetta based structures can be reasonably expected to be sampled *in vitro*.

Bibliography

1. Mary Rajathei, D. & Selvaraj, S. Analysis of sequence repeats of proteins in the PDB. *Comput. Biol. Chem.* **47**, 156–166 (2013).

2. Backe, P. H. *et al.* A new family of proteins related to the HEAT-like repeat DNA glycosylases with affinity for branched DNA structures. *J. Struct. Biol.* **183**, 66–75 (2013).

3. Albert, P.-R., Marie, S. & S., I. L. Folding cooperativity and allosteric function in the tandem-repeat protein class. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170188 (2018).

4. Reichen, C. *et al.* Computationally Designed Armadillo Repeat Proteins for Modular Peptide Recognition. *J. Mol. Biol.* **428**, 4467–4489 (2016).

5. Parmeggiani, F. *et al.* A General Computational Approach for Repeat Protein Design. *J. Mol. Biol.* **427**, 563–575 (2015).

6. Björklund, Å. K., Ekman, D. & Elofsson, A. Expansion of Protein Domain Repeats. *PLOS Comput. Biol.* **2**, e114 (2006).

7. Brunette, T. *et al.* Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).

8. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).

9. Bahar, I., Cheng, M. H., Lee, J. Y., Kaya, C. & Zhang, S. Structure-Encoded Global Motions and Their Role in Mediating Protein-Substrate Interactions. *Biophys. J.* **109**, 1101–1109 (2015).

10. Aksel, T. & Barrick, D. Chapter 4 Analysis of Repeat-Protein Folding Using Nearest-Neighbor Statistical Mechanical Models. *Methods Enzymol.* **455**, 95–125 (2009).

11. Kajander, T., Cortajarena, A. L., Main, E. R. G., Mochrie, S. G. J. & Regan, L. A New Folding Paradigm for Repeat Proteins. *J. Am. Chem. Soc.* **127**, 10188–10190 (2005).

12. Millership, C., Phillips, J. J. & Main, E. R. G. Ising Model Reprogramming of a Repeat Protein's Equilibrium Unfolding Pathway. *J. Mol. Biol.* **428**, 1804–1817 (2016).

13. Kappel, C., Zachariae, U., Dölker, N. & Grubmüller, H. An unusual hydrophobic core confers extreme flexibility to HEAT repeat proteins. *Biophys. J.* **99**, 1596–1603 (2010).

14. Leaver-Fay, A. *et al.* Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).

15. Guffy, S. L., Der, B. S. & Kuhlman, B. Probing the minimal determinants of zinc binding with computational protein design. *Protein Eng. Des. Sel.* **29**, 327–338 (2016).

16. Salgado, E. N., Lewis, R. A., Faraone-Mennella, J. & Tezcan, F. A. Metal-Mediated Self-Assembly of Protein Superstructures: Influence of Secondary Interactions on Protein Oligomerization and Aggregation. *J. Am. Chem. Soc.* **130**, 6082–6084 (2008).

17. Lansu, K. *et al.* In silico design of novel probes for the atypical opioid receptor MRGPRX2. *Nat. Chem. Biol.* **13**, 529 (2017).

18. Maguire, J. B., Boyken, S. E., Baker, D. & Kuhlman, B. Rapid Sampling of Hydrogen Bond Networks for Computational Protein Design. *J. Chem. Theory Comput.* **14**, 2751–2760 (2018).

19. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).

20. Parmeggiani, F. & Huang, P.-S. Designing repeat proteins: a modular approach to protein design. *Curr. Opin. Struct. Biol.* **45**, 116–123 (2017).

21. Yeh, C.-T., Brunette, T., Baker, D., McIntosh-Smith, S. & Parmeggiani, F. Elfin: An algorithm for the computational design of custom three-dimensional structures from modular repeat protein building blocks. *J. Struct. Biol.* **201**, 100–107 (2018).

22. Yeh, J. Elfin: An Algorithm for Computational Protein Design using a Protein LEGO Strategy. (University of Bristol, 2017).

23. Nivón, L. G., Moretti, R. & Baker, D. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS One* **8**, e59004 (2013).

24. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55 (2014).

25. Tyka, M. D. *et al.* Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405**, 607–618 (2011).

26. Gus Breese, Ben Davies, Mark James, Vladimír Macko, M. R. Internal Dynamics of Modular Protein Structures. (University of Bristol, 2017).

27. Sheather, S. J. Density Estimation. *Stat. Sci.* **19**, 588–597 (2004).

28. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **33**, 1065–1076 (1962).

29. Touw, W. G. *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **43**, D364–D368 (2014).

30. Arthur, D. & Vassilvitskii, S. K-means++: The Advantages of Careful Seeding. in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035 (Society for Industrial and Applied Mathematics, 2007).

31. Ankerst, M., Breunig, M. M., Kriegel, H.-P. & Sander, J. OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Rec.* **28**, 49–60 (1999).

32. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. in (2000). doi:https://doi.org/10.1111/1467-9868.00293

33. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**–**2**, 19–25 (2015).

34. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).

35. Mark, P. & Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **105**, 9954–9960 (2001).

36. Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S. & Baker, D. De Novo Enzyme Design Using Rosetta3. *PLoS One* **6**, e19230 (2011).

37. Lee, J.-M., Yoo, C., Choi, S. W., Vanrolleghem, P. A. & Lee, I.-B. Nonlinear process monitoring using kernel principal component analysis. *Chem. Eng. Sci.* **59**, 223–234 (2004).