Title page

The full title of the manuscript:

Coil conversion to β-strand induced by dimerisation

The short title of the manuscript:

Coil conversion to β-strand by dimerisation

Key words:

Protein-Protein Interaction, Protein Conformational Change, Protein Dimerisation, Intermolecular β-strand Interfaces, Dimorphic Sequences, Protein Structure

The names and affiliations of all authors:

Johanna Laibe, Faculty of Science, Engineering & Computing, Kingston University, London

Aaron Caffrey, Faculty of Science, Engineering & Computing, Kingston University, London

Melanie Broutin, Nice Sophia Antipolis University Engineering School

Solene Guiglion, Nice Sophia Antipolis University Engineering School

Barbara Pierscionek, School of Science & Technology, Nottingham Trent University

Jean-Christophe Nebel, Faculty of Science, Engineering & Computing, Kingston University, London

The institution at which the work was performed:

Kingston University, London

Contact information for the author responsible for correspondence:

Jean-Christophe Nebel

Faculty of Science, Engineering and Computing, Kingston University, London

Kingston-upon-Thames, Surrey, KT1 2EE, UK

J.Nebel@kingston.ac.uk

Coil conversion to β-strand induced by dimerisation

Johanna Laibe[1], Aaron Caffrey[1], Melanie Broutin[2], Solene Guiglion[2], Barbara Pierscionek[3] and Jean-Christophe Nebel[1]

[1]Faculty of Science, Engineering and Computing, Kingston University, London, Kingston-upon-Thames, Surrey, KT1 2EE, UK
[2]Department of Bioengineering, Nice Sophia Antipolis University Engineering School, Templiers Campus, 06410 Biot, France
[3]School of Science & Technology, Nottingham Trent University, 50 Shakespeare Street, Nottingham, NG1 4FQ, UK

Abstract. Most molecular processes in living organisms rely on protein–protein interactions, many of which are mediated by β-sheet interfaces; this study investigates the formation of β-sheet interfaces through the conversion of coils into β-strands. Following an exhaustive search in the Protein Data Bank, the corresponding structural dimorphic fragments were extracted, characterised and analysed. Their short strand lengths and specific amino acid profiles indicate that dimorphic β-strand interfaces are likely to be less stable than standard ones and could even convert to coil interfaces if their environment changes. Moreover, the construction of a simple classifier able to discriminate between the sequences of dimorphic and standard β-strand interfaces suggests that the nature of those dimorphic sequences could be predicted, providing a novel means of identifying proteins capable of forming dimers.

1 Introduction

Since most molecular processes rely on protein–protein interactions, knowledge of those interactions is extremely valuable for biomedical research and drug design. Despite the availability of high-throughput proteomics approaches[1], protein interactomes are still largely incomplete. Consequently, the development of bioinformatics methods allowing the prediction of such interactions is a very active field of investigation as reported in a recent

review[2]. Protein–protein interactions through β-sheet interfaces have been of particular interest[3,4], predominantly resulting from their potential to cause aggregation[5]. In addition, a variety of features have been identified that discriminate between "central strands, bordered on both sides by other β-strands, and edge strands, bordered on only one side by another β-strand"[5–7]. It has also been reported that dimerisation of members of the met-repressor-like family - which share a similar ribbon-helix-helix structure - can bind DNA following their homo-dimerisation process where the coil becomes a β-strand to form a β-sheet interface[8] (Figure 1). Although such types of conformational changes have often been associated with pathologies[9], this case suggests they may also lead to formation of functional dimers. Since no study has thus far performed a systematic analysis of those strands that exist only as result of dimerisation (i.e. β-sheet interfaces formed of only two intermolecular-β-strands), this work investigates secondary structure alteration resulting from dimerisation. More specifically, this work focuses on coil sequences forming intermolecular β-strand interfaces. Following an exhaustive search in the Protein Data Bank[10], the structures and corresponding sequences of dimorphic fragments were extracted, characterised and investigated.

2 Materials and methods

2.1 Dataset generation

Since very few proteins displaying such a 'dimorphic' property have been reported in the literature, with the notable exception of the met-repressor-like family, an exhaustive search was conducted for all dimer structures available in the Protein Data Bank (PDB)[10]. Representatives at 30% sequence identity of all dimers annotated as biological assemblies were retrieved from the PDB (as of January 13, 2016). This identity constraint was selected to prevent biased results due to the presence of homologous proteins. A filter was then designed to identify those dimers that interact through a β-strand-β-strand interface. In line

with the definition used at the CAPRI community-wide experiment (Critical Assessment of Prediction of Interactions)[11], interface residues were defined as amino acids containing heavy atoms located within 5Å from their counterparts on residues in another chain. Through analysis of the 'SHEET' records located within PDB structure files, the β-sheets containing those interface residues were identified and are referred to as β-sheet interfaces. Further analysis of the 'SHEET' records also allowed establishment of the number of strands per chain comprising the β-sheet interfaces and their general topologies. Four types of β-sheet interfaces were defined: a) dimorphic: those composed of only two β-strands – one from each chain, b) standard: those composed of two existing β-sheets - at least two strands from each chain, c) hybrid: those composed of one β-strand from one chain and one β-sheet from the other, and d) ambiguous: those not fitting into categories a, b, or c. Manual curation was then applied to identify and amend any interface β-strands whose automated classification was either incorrect or ambiguous. Eventually, four non-homologous sets of β-strand-β-strand interfaces were generated: a) 146 dimorphic, b) 205 standard, c) 218 hybrid, and d) 286 ambiguous (Figure 2). Interestingly, none of the dimorphic interfaces belong to a membrane protein. This dataset is available in 'Supplementary Material'.

2.2 Interface properties

A range of properties was extracted from the β-strand interfaces. After categorising them according to the nature of the dimer (i.e. homodimer or heterodimer), and their configuration (i.e. parallel or antiparallel), homodimer interfaces were computationally classified according to the sequences of their interacting β-strands to detect the presence of symmetry. First, interfaces composed of two identical sequences were annotated as presenting a one-site symmetry. Second, dimers displaying two separate β-strand interfaces, where each interface utilises the same pair of distinct sequences, were annotated as depicting a two-site symmetry.

Third, the remaining interfaces were annotated as asymmetrical. Figure 3 illustrates the types of interface symmetries that were encountered in the selected homodimer sets. Finally, after collecting those qualitative properties, the amino acid compositions and the β-strand lengths was calculated for each dimer in the datasets.

2.3 Dimorphic β-strand interface stability

To further study the stability of dimorphic β-strand interfaces, the PDB was queried to extract homologous dimers (i.e. sequence identity ≥ 30% and E-value < 1.10^{-6}) which do not display the dimorphic β-strand interfaces (i.e. where the interface sequences remain in their original coil conformations). To highlight structural differences between the interfaces that are present in both coil and β-strand structural conformations, the homologous PDB structures were aligned using Pymol[12]. In addition, Pymol has also been used to produce the figures displaying protein structures within this article.

Among those homologous dimer structures where the interface was found present in both coil and β-strand conformations, structures determined using NMR spectroscopy were of particular interest, since their multiple model records allows for quantitative analysis to be performed on the mobility of their residues. In this study, the analysis of residue mobility was performed utilising the MOBI webserver, which identifies “regions with different conformations among all the models in a NMR solved PDB structure ensemble” and calculates the average RMSD for each residue within that PDB file[13].

2.4 Property discriminative power

To evaluate quantitatively whether the observed property differences between dimorphic and standard β-strand interfaces allows discrimination between those two types of interfaces, a

supervised machine learning model was employed, i.e. Support Vector Machine (SVM), to create a binary classifier.

2.4.1 Features and encoding

Each dimorphic and standard β-strand interface sequence in the dataset was encoded as a feature vector of 20 numeric values which are associated to the presence of each amino acid type in the interface sequence. This 20 feature set is known as the OAAC (overall amino acid composition) which represents the occurrence frequency of each amino acid type within a sequence divided by the sequence length. Additionally, the OAAC values were square rooted since it has been shown to improve predictive performance[14]. Although additional sequence features, such as physicochemical properties[15,16] and evolutionary information[17,18], could have be exploited this was out of the scope of this particular investigation which was only aimed at demonstrating that dimorphic and standard β-strand interfaces fit within two distinct classes.

2.4.2 Nested cross-validation

The popular open source SVM library, LIBSVM (version 3.22) was selected to build the binary classifier[19]. In particular, it supports the Radial Basis Function (RBF), its default kernel, which is known to perform well on a variety of classification tasks[19]. Nested cross-validation was used prevent overfitting, class bias, or performance bias; *n*-fold inner-cross-validation was used for model selection, while *k*-fold outer-cross-validation was used to estimate the generalised classifier performance.

2.4.2.1 Inner-cross-validation

The RBF kernel relies on only two training parameters: c (cost), which sets a compromise between misclassification and model simplicity (i.e. its generalisation capability), and γ

(gamma), which limits the influence of each individual training sample. The optimal values for these kernel parameters are initially unknown until they are found through utilisation of a cross-validation model selection function, whereby multiple potential classification models are prospectively created with the same training data, each however with differing parameter values for c and γ. Thereafter, the top performing model in terms of n-fold cross-validation accuracy is selected as the best model, which is then used to predict the unseen independent data in an outer-cross-validation procedure.

In the case of this classifier, the n-fold inner-cross-validation model selection on the training dataset was performed utilising LIBSVM's implementation of the efficient and effective grid search function[20]. However, since in the default implementation of LIBSVM, cross-validation performance is measured using the accuracy metric, which for imbalanced datasets can be misleading as it is not a class specific measure, the training dataset classes were balanced through down-sampling the majority class to avoid any potential class bias.

2.4.2.2 Outer-cross-validation

For the k-fold outer-cross-validation procedure, the inner-cross-validation model selection procedure is repeated k times. For each k, the whole dataset is shuffled randomly before partitioning, using the same partitioning percentage ratios each time, into the training and testing datasets. This procedure results in k training datasets, k independent testing datasets, k selected models, and lastly, k sets of independent testing performance results, which are then used to estimate generalised performance.

2.4.3 Evaluation of performance

From the nested cross-validation procedure described above, generalised performance is estimated according to the arithmetic mean and standard deviation for each performance

metric over the *k* independent testing dataset prediction results. The classifier's performance is evaluated, first, by providing the confusion matrix (i.e. the table visualising the total number of instances that have had their class correctly and incorrectly predicted), accuracy, the harmonic mean average of precision and recall (F1 score), and the Matthews correlation coefficient (MCC) which offers a balanced measure of the quality of binary classification even if the classes are of different sizes[21]. Note that with MCC the coefficient varies between -1 and +1, where -1, 0, and +1, indicate, respectively, total disagreement with observation, random, and perfect predictions. Additionally specified is the non-interpolated Average Precision, calculated using all thresholds, synonymous with Area Under the Precision-Recall Curve (AUC PR), which is particularly relevant when dealing with imbalanced data[22].

3 Results and discussion

3.1 Experimental results

Classification of the dimer interfaces according to their dimeric nature reveals that, across the PDB, β-strand-β-strand interfaces generally tend to be formed as part of homodimers (88%), as seen in Table 1. Moreover, in line with previous work[4], data shows that those interfaces are more likely to adopt an antiparallel configuration (80%). Indeed, Watkins and Arora suggested that, in protein complex interactions where binding energy is critical, such orientation is favoured since it offers better hydrogen bonding geometry and improved energetics[4]. Since a meaningful analysis can only be produced if the data is relatively homogenous, ambiguous interfaces are not considered any further in this study.

Among the homodimers, over 97% of the β-strand interfaces display some symmetry, see Table 2. On one hand, while antiparallel dimorphic and standard interfaces tend to exhibit one-site symmetry (over 73%), parallel interfaces show two-site symmetry (over 80%).

Conversely, hybrid interfaces, which cannot form interfaces with a one-site symmetry, display two-site symmetry in around 96% of cases regardless of their parallelism.

As Table 3 reveals, there are important dissimilarities in terms of amino acid composition (Figure 4) and β-strand length (Figure 5) between dimorphic and standard β-strand interfaces. First, aromatic amino acids show a clear preference for a type of interface; the basic histidine favours standard β-strand interfaces, whereas tryptophan, tyrosine and, even, phenylalanine, prefer dimorphic interfaces. Furthermore, among other charged amino acids, glutamic acid is more common in standard β-strand interfaces, whilst arginine and lysine are more often found in dimorphic interfaces. In addition to showing differently charged amino acid profiles, dimorphic interfaces display a much more important imbalance towards positively charged residues. Regarding small amino acids, the three smallest, glycine, alanine and serine, are more frequent in standard interfaces, while proline is overrepresented in dimorphic interfaces. Finally, dimorphic interfaces are much shorter than standard interfaces, containing on average, three fewer amino acids.

To further investigate differences between dimorphic and standard β-strand interfaces, CATH[34] annotations (where available) were associated with each interface, allowing for the creation of topological profiles for each of these classes. As Figure 6 shows, not only do dimorphic and standard β-strand interface profiles differ significantly from the profile of all CATH domains, but they also display quite different class preferences. While standard β-strand interfaces are essentially associated to alpha-beta domains (78%), with a sizeable group of mainly beta domains (21%), no single CATH class hosts the majority of dimorphic β-strand interfaces (i.e. 45% associated to mainly alpha domains, 20% to mainly beta domains and 34% to alpha-beta domains). Interestingly, around 25% of both dimorphic and standard β-strand interfaces display a 3-layer (aba) sandwich architecture.

3.2 Qualitative analysis

A study of salt bridge compositions within β-sheets revealed that arginine and histidine have much higher propensities than lysine[23]. More specifically, glutamic acid-histidine interactions have the highest propensity followed by the glutamic acid-arginine interactions. Since standard β-strand interfaces are particularly rich in histidine and glutamic acid, in comparison with dimorphic interfaces, whilst displaying a better charge balance and a standard arginine frequency, those interfaces offer an environment that is particularly favourable for the formation of salt bridges, which can further stabilise their β-strand interactions (Figure 7A). Conversely, due to the charge imbalance, the dimorphic interface leaves many charges available. Such a consequence is consistent with observations by Richardson and Richardson[24] who studied mechanisms used to protect edge β-strands from further β-sheet interactions that might lead to aggregation. Note that dimorphic interfaces, unlike standard β-strand interfaces, are formed only from edge β-strands. Among the most common strategies, they reported the presence on edge β-strands of not only inward-pointing charged residues (Figure 7B.a), but also of proline residues (Figure 7B.b), which are also overrepresented on dimorphic interfaces (Figure 4).

Although the contribution of aromatic-aromatic interactions to the formation of secondary structures has been a topic of investigation for many years[25], it is difficult to explain the higher frequency of tryptophan, tyrosine and phenylalanine in dimorphic interfaces. Since it was proposed that such interactions could play a role of β-sheet stabilisation in absence of inter-strand hydrogen bonding[26], relationship between β-strand length and presence of aromatic residues was investigated. It was observed that strands with the highest frequency of aromatic residues tend to be short, which is consistent with the idea that the presence of the aromatic residues provides short β-sheets additional stability by creating aromatic-aromatic

interactions (Figure 8). Although that high frequency appeared to be a feature of dimorphic interfaces, as seen in Table 3, this may be a consequence of the fact that dimorphic interfaces are, on average, much shorter than standard β-strand interfaces. Indeed, many short standard β-strand interfaces also display a high frequency of tryptophan, tyrosine, and phenylalanine.

An experiment conducted on peptides composed of antiparallel β-sheets shows that strands of length 7 are more stable than those of length 5 or 9[27]. Moreover, it has been reported that a strand length of 7 allows a peptide to display optimal antimicrobial activity[28]. Computer simulation analysing interactions of over 50,000 β-strand interfaces concur with these findings[4]: the average strand length of strong interfaces is 5.9, whereas it is 4.4 for weak ones. In this study, whilst standard β-strands have an average strand length of 7.1, dimorphic strands are found to be much shorter, with an average length of only 4.1. This suggests a lower stability of dimorphic antiparallel β-sheets. This hypothesis is further supported by experiments comparing the stability of β-sheets when increasing the number of β-strands from 2 to 3 which demonstrated that a higher number of strands lead to higher stability[29]. As a consequence, dimorphic interfaces are expected to be less stable than standard interfaces.

3.3 Dimorphic β-strand interface stability

Analysis of stability of dimorphic β-strand interfaces was conducted by comparing the dimorphic interfaces of three distinct proteins (histone, transcription factor and Tip-alpha) with the corresponding interfaces of their homologous dimers that do not display those β-strands.

The PDB contains 3 homologous dimeric entries (with sequence similarity over 85%) of histones from the archaea species Methanothermus fervidus. Although all three share identical interaction sequences, the histone HMfA, 1B67, and the recently published structure

of histone-based chromatin[30], 5T5K, display two β-strand dimorphic interfaces exhibiting a two-site symmetry, whereas the histone B, 1BFM, shows one dimorphic β-strand and one coil-based interfaces (Figure 9). Analysis on the hydrogen bonding of the amino acids involved in those interfaces showed that interaction with a third molecule creates an additional 'indirect' hydrogen bond via that molecule between the two monomers supporting the formation of the β-sheet: while one of 1B67's dimorphic interfaces interacts with SO_4^{2-} and both of 5T5K's dimorphic interfaces interact with DNA, 1BFM's do not bind to any ligand.

Since 1BFM was resolved using NMR (33 individual models are available), the mobility of its residues was measured using the MOBI web server[13]. Figure 10 reveals that, in addition to the terminal regions that, as expected, are highly mobile, the two regions comprising the residues involved in the dimorphic β-strand interface also display high mobility.

Structure of the N-terminal domain of AbrB-like Transcription Factors is known in its unbound form, 1YSF, and bound to DNA, 2K1N. While 1YSF displays a β-strand dimorphic interface exhibiting a one-site symmetry, 2K1N has a coiled based interface instead (Figure 11).

Taking advantage that both 1YSF and 2K1N were produced by NMR (with 22 and 10 models respectively), mobility was investigated using MOBI[13]. As Figure 12 shows, the chains within 2K1N are much more mobile than those of 1YSF. Previous study of the mobility of AbrB identified the need of conformational change and concerted motions to enable its interaction with a DNA target[31]. As a consequence, this suggests that, although the dimorphic β-strand interface stabilises the protein structure, the interface is able to adopt a coil conformation and gain in flexibility when AbrB is involved in the DNA binding process. The importance of that dimorphic segment, Arg-Val, was further highlighted by mutagenic

analysis of AbrB that identified the arginine which is present in that interface as critical to DNA binding[32].

Structures of the dimeric fragments (34-192) of the tumor necrosis factor alpha inducing protein Tip-alpha (2WCQ & 2WCR) were resolved in very different environments of pH values 4 and 8.5, respectively. As Figure 13 shows, they display distinctive configurations where 2WCR relies on the presence of a dimorphic interface.

3.4 Property discriminative power

Since comparison of the amino acid compositions between dimorphic and standard β-strand interfaces revealed different amino acid profiles, a binary SVM classifier was built based on the overall amino acid composition (OAAC) of their interface sequences.

Inner-cross-validation was performed utilising the LIBSVM grid search function with its default parameters, which performs model selection on training data using *5*-fold inner-cross-validation. This cross-validation was chosen as it segments the training set into 4/5 parts (i.e. 80%) for training and 1/5 parts (i.e. 20%) for a validation testing, which is repeated 5 times to cover the whole training dataset, which ensures that each instance (i.e. interface sequence) is predicted for validation once only.

Outer-cross-validation was performed on all selected models to prevent any bias that could have been caused by specific partitioning of the dataset into training and testing. *1000*-fold outer-cross-validation was selected to measure the generalised performance metrics. This showed that, for all calculated performance metrics, average performance was stable within 2% when $k \geq 36$ folds, within 1% when $k \geq 134$ folds, and in within 0.1% (in 92% of cases) when $k \geq 750$ folds, compared with the final average performance at $k=1000$.

The 146 dimorphic and 205 standard β-strand interfaces are composed, respectively, of 165 and 224 unique interface amino acid sequences. Due to the relatively small size of the β-strand interface datasets, a large proportion needed to be allocated to the training set. For each outer-cross-validation iteration, both classes were randomly shuffled, and then, as commonly chosen, partitioned into a training dataset made of 80% of the sequences, and a test dataset with the remaining 20%. In addition, since it has been shown that class size imbalance in the training dataset can negatively affect a classifier's performance[33], class imbalance was addressed by down-sampling the majority class to the size of the minority class[33]: the standard β-strand sequences class was down-sampled to match the size of the dimorphic β-strand sequences class. Moreover, LIBSVM's grid search function is more reliable with a balanced training set due to using the accuracy metric for model selection (i.e. correct prediction of the majority class alone could result in misleadingly high accuracy). The training dataset is therefore comprised of 132 dimorphic β-strand sequences and 132 standard β-strand sequences. Meanwhile, the class ratios within independent testing sets are not altered, thus keeping their original imbalance of 33 dimorphic and 45 standard β-strand sequences, which should represent what is seen in nature, or at least in the PDB depiction of it.

The confusion matrix shown in Table 4 provides the average performance of the classifier: out of the 33 dimorphic sequences, 23.39 are classified correctly, whereas out of the 45 standard β-strand sequences correct classification occurs for 32.46 sequences. As Table S1 reports, the SVM classifier shows good accuracy of $\overline{0.72}$ ($\sigma = 0.05$), and F1 score of $\overline{0.68}$ ($\sigma = 0.05$) for the dimorphic β-strand class, and $\overline{0.74}$ ($\sigma = 0.05$) for the standard β-strand class. In addition, the associated MCC (Matthews correlation coefficient) is $\overline{0.43}$ ($\sigma = 0.10$), which is usually interpreted as indicating the classifier is a moderate to strong predictor. Finally, Average Precision is $\overline{0.77}$ ($\sigma = 0.05$) for the dimorphic β-strand class and $\overline{0.82}$ ($\sigma =$

0.05) for the standard β-strand class. Consequently, this quantitative analysis has confirmed that dimorphic and standard β-strand interfaces display distinct amino acid profiles. Note that, for both dimorphic and standard β-strand interfaces, the type of CATH class to which an interface sequence is associated does not appear to affect the quality of its classification (study not shown).

4 Conclusion

This investigation has revealed that many dimers rely on dimorphic β-strand interfaces, listing a non-redundant set of 146 examples. Whereas their nature and parallelism do not differ from standard β-strand interfaces', their average β-strand length and their amino acid profile are quite distinct. Not only does analysis of those features indicate that dimorphic β-strand interfaces are likely to be less stable than standard β-strand interfaces, but they also tend to take advantage of strategies preventing further β-sheet interactions that would increase interface stability. The study of these interfaces that are found in both dimorphic and coil forms shows that the presence of a binding molecule, or a change of environment pH, can affect the structural conformation of such interfaces. Whereas a dimorphic β-strand interface adds some stability to the structure of a protein, its intrinsic flexibility allows it to return to the coil configuration required for that protein to perform other aspects of its function.

The construction of a classifier based only on amino acid profiles has shown that sequences involved in either dimorphic or standard β-strand interfaces are sufficiently different to allow for some automatic discrimination between them using machine learning methods. This suggests that, with the usage of additional features, including structural ones, a robust classifier could be designed to predict whether a monomer has the potential to form a dimer through the conversion of one of its coils into a β-strand. Such a tool would provide a high-

throughput means to enrich knowledge of protein interactomes and would also support the analysis of individual proteins within a given environment.

5 Supplementary Material

The supplement contains the following information:

1. List of all β-strand interfaces in the dataset (as shown in Table 1 and Table 2).

2. Table S1 which presents additional performance results from the SVM classifier experiment.

References

1. Braun P, Gingras AC. History of protein-protein interactions: From egg-white to complex networks. Proteomics. 2012;12(10):1478–1498. doi:10.1002/pmic.201100563

2. Esmaielbeiki R, Krawczyk K, Knapp B, Nebel JC, Deane CM. Progress and challenges in predicting protein interfaces. Briefings in Bioinformatics. 2016;17(1):117–131. doi:10.1093/bib/bbv027

3. Feverati G, Achoch M, Vuillon L, Lesieur C. Intermolecular β-strand networks avoid hub residues and favor low interconnectedness: A potential protection mechanism against chain dissociation upon mutation. PLoS ONE. 2014;9(4):1–16. doi:10.1371/journal.pone.0094745

4. Watkins AM, Arora PS. Anatomy of β-strands at protein-protein interfaces. ACS Chemical Biology. 2014;9(8):1747–1754. doi:10.1021/cb500241y

5. Westerlund I, von Heijne G, Emanuelsson O. LumenP-A neural network predictor for protein localization in the thylakoid lumen. Protein Science. 2009;12(10):2360–2366. doi:10.1110/ps.0306003

6. Minor DL, Kim PS. Context is a major determinant of beta-sheet propensity. Nature. 1994;371(6494):264–267. doi:10.1038/371264a0

7. Parisien M, Major F. A new catalog of protein ??-sheets. Proteins: Structure, Function and Genetics. 2005;61(3):545–558. doi:10.1002/prot.20677

8. Golovanov AP, Barillà D, Golovanova M, Hayes F, Lian LY. ParG, a protein required for active partition of bacterial plasmids, has a dimeric ribbon-helix-helix structure. Molecular Microbiology. 2003;50(4):1141–1153. doi:10.1046/j.1365-2958.2003.03750.x

9. Dobson CM. The structural basis of protein folding and its links with human disease. Philosophical Transactions of the Royal Society B: Biological Sciences. 2001;356(1406):133–145. doi:10.1098/rstb.2000.0758

10. Berman HM. The Protein Data Bank. Nucleic Acids Research. 2000;28(1):235–242. doi:10.1093/nar/28.1.235

11. Janin J, Wodak S. The Third CAPRI Assessment Meeting Toronto, Canada, April 20-21, 2007. Structure. 2007;15(7):755–759. doi:10.1016/j.str.2007.06.007

12. The PyMOL Molecular Graphics System, Version 1.6, Schrödinger, LLC. 2015. The {PyMOL} Molecular Graphics System, Version~1.6. 2015.

13. Martin AJM, Walsh I, Tosatto SCE. MOBI: A web server to define and visualize structural mobility in NMR protein ensembles. Bioinformatics. 2010;26(22):2916–2917. doi:10.1093/bioinformatics/btq537

14. Feng ZP, Zhang CT. Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. International Journal of Biological Macromolecules. 2001;28(3):255–261. doi:10.1016/S0141-8130(01)00121-0

15. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: Amino acid index database, progress report 2008. Nucleic Acids Research. 2008;36(SUPPL. 1):D202–D205. doi:10.1093/nar/gkm998

16. Du P, Gu S, Jiao Y. PseAAC-General: Fast building various modes of general form of chou's pseudo-amino acid composition for large-scale protein datasets. International Journal of Molecular Sciences. 2014;15(3):3495–3506. doi:10.3390/ijms15033495

17. Ramsay L, Macaulay M, Degli Ivanissevich S, MacLean K, Cardle L, Fuller J, Edwards KJ, Tuvesson S, Morgante M, Massari A, et al. A simple sequence repeat-based linkage map of Barley. Genetics. 2000;156(4):1997–2005. doi:10.1093/nar/25.17.3389

18. Xu R, Zhou J, Wang H, He Y, Wang X, Liu B. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. BMC Systems Biology. 2015;9(1):S10–S10. doi:10.1186/1752-0509-9-S1-S10

19. Chang C-C, Lin C-J. Libsvm. ACM Transactions on Intelligent Systems and Technology. 2011;2(3):1–27. doi:10.1145/1961189.1961199

20. Silva MFM, Leijoto LF, Nobre CN. Algorithms Analysis in Adjusting the SVM Parameters: An Approach in the Prediction of Protein Function. Applied Artificial Intelligence. 2017;31(4):316–331. doi:10.1080/08839514.2017.1317207

21. Matthews BW. Comparison of Predicted and Observed Secondary Structure of T4 Phase Lysozyme. Biochim. Biophys. Acta. 1975;405(2):442–451. doi:https://doi.org/10.1016/0005-2795(75)90109-9

22. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432

23. Donald JE, Kulp DW, DeGrado WF. Salt bridges: Geometrically specific, designable interactions. Proteins: Structure, Function and Bioinformatics. 2011;79(3):898–915. doi:10.1002/prot.22927

24. Richardson JS, Richardson DC. Natural -sheet proteins use negative design to avoid edge-to-edge aggregation. Proceedings of the National Academy of Sciences.

2002;99(5):2754–2759. doi:10.1073/pnas.052706099

25. Thomas A, Meurisse R, Charloteaux B, Brasseur R. Aromatic side-chain interactions in proteins. I. Main structural features. Proteins: Structure, Function and Genetics. 2002;48(4):628–634. doi:10.1002/prot.10190

26. Budyak IL, Zhuravleva A, Gierasch LM. The role of aromatic-aromatic interactions in strand-strand stabilization of β-sheets. Journal of Molecular Biology. 2013;425(18):3522–3535. doi:10.1016/j.jmb.2013.06.030

27. Stanger HE, Syud FA, Espinosa JF, Giriat I, Muir T, Gellman SH. Length-dependent stability and strand length limits in antiparallel beta -sheet secondary structure. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(21):12015–20. doi:10.1073/pnas.211536998

28. Dong N, Ma Q, Shan A, Lv Y, Hu W, Gu Y, Li Y. Strand length-dependent antimicrobial activity and membrane-active mechanism of arginine- and valine-rich β-hairpin-like antimicrobial peptides. Antimicrobial Agents and Chemotherapy. 2012;56(6):2994–3003. doi:10.1128/AAC.06327-11

29. Kung VM, Cornilescu G, Gellman SH. Impact of Strand Number on Parallel β-Sheet Stability. Angewandte Chemie - International Edition. 2015;54(48):14336–14339. doi:10.1002/anie.201506448

30. Mattiroli F, Bhattacharyya S, Dyer PN, White AE, Sandman K, Burkhart BW, Byrne KR, Lee T, Ahn NG, Santangelo TJ, et al. Structure of histone-based chromatin in Archaea. Science. 2017;357(6351):609–612. doi:10.1126/science.aaj1849

31. Sullivan DM, Bobay BG, Kojetin DJ, Thompson RJ, Rance M, Strauch MA, Cavanagh

J. Insights into the Nature of DNA Binding of AbrB-like Transcription Factors. Structure. 2008;16(11):1702–1713. doi:10.1016/j.str.2008.08.014

32. Vaughn JL, Feher V, Naylor S, Strauch MA, Cavanagh J. Novel DNA binding domain and genetic regulation model of Bacillus subtilis transition state regulator AbrB. Nature Structural Biology. 2000;7(12):1139–1146. doi:10.1038/81999

33. Wei Q, Dunbrack RL. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. PLoS ONE. 2013;8(7):e67863. doi:10.1371/journal.pone.0067863

34. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2017. doi: 10.1093/nar/gkw1098

Figure legends

Figure 1: Representative of the met-repressor-like family (1MNT). The dimer is formed by joining a ribbon-helix-helix pattern (RHH) from each chain. In the process, RHH converts into a β-strand-helix-helix pattern.

Figure 2: Examples of each β-strand interface type: a) dimorphic 2BA3), b) standard (2P24), c) hybrid (3GLA) and d1) & d2) ambiguous (2GE7 & 2O38). Note: PDB chains A and B are coloured in green and blue respectively.

Figure 3: Symmetries encountered in the dimorphic, standard and hybrid homodimer β-strand interfaces. Note: Blue and yellow arrows denote interface strands with distinct sequences.

Figure 4: Percentage variations of amino acid frequency between dimorphic and standard β-strand interfaces. Positive and negative values show preference for, respectively, standard and dimorphic β-strand interfaces.

Figure 5: Distribution of β-strand lengths forming dimorphic and standard β-strand interfaces.

Figure 6: Comparison of topological profiles of dimorphic β-strand interfaces, standard β-strand interfaces and domains in CATH (based on PDB release: July 01, 2017). In addition to the percentage of sequences belonging to the four CATH classes, the five main CATH architectures (identified by their CATH codes) are highlighted.

Figure 7: A. Standard β-strand interface including a salt bridge between E226-A and H232-B (4YWJ) **B.** Dimorphic interfaces displaying common strategies to prevent further β-sheet interactions: presence of a) inward-pointing charged (K35, R36 & D42 in 1P94) or b) proline residues (P34 in 1HUL).

Figure 8: Stabilisation of a short (5 amino acid long) dimorphic interface by aromatic-aromatic interactions (W228, Y229, W230 & F231 on both chains of 1H8G).

Figure 9: Structural alignment of 1BFM (chain A in green and chain B in blue) and 1B67 (in red) highlighting 1BFM's loss of a dimorphic interface and the interaction of 1B67's dimorphic interface with SO_4^{2-} creating an 'indirect' hydrogen bond (yellow lines).

Figure 10: Residue mobility in terms of average RMSD of 1BFM model. Solid and dashed rectangles highlight residues involved in, respectively, the dimorphic β-strand interfaces and the surrounding coils.

Figure 11: Structural alignment of 2K1N (chain A in green and chain B in blue) and 1YSF (in red) highlighting 2K1N's loss of a dimorphic interface.

Figure 12: Residue mobility in terms of average RMSD of 1YSF and 2KIN models. Solid and dashed rectangles highlight residues involved in, respectively, the dimorphic β-strand interface (Arg-Val) and the surrounding coil. Note that 2KIN's sequence is used as reference for residue numbering.

Figure 13: Structural alignment of 2WCQ (chain A in green and chain B in blue) and 1WCR (chain A in dark red and chain B in red) highlighting dimers' different configurations and the absence of a dimorphic interface in 2WCQ.

Tables

Table 1: β-strand interface classification according to their nature, i.e. homo- or hetero-, and parallelism.

Dimer nature & parallelism		**Dimorphic**	**Standard**	**Hybrid**	**Ambiguous**	**All**
Homodimers	Antiparallel	111	153	123	229	**616**
	Parallel	19	35	73	13	**140**
	All	**130**	**188**	**196**	**242**	**756**
Heterodimers	Antiparallel	11	8	18	34	**71**
	Parallel	5	9	4	10	**28**
	All	**16**	**17**	**22**	**44**	**99**
Total	Antiparallel	122	161	141	263	**687**
	Parallel	24	44	77	23	**168**
	All	**146**	**205**	**218**	**286**	**855**

Table 2: Symmetry in homodimer β-strand interfaces.

	Dimorphic			**Standard**			**Hybrid**		
Symmetry type	**One-site**	**Two-site**	**None**	**One-site**	**Two-site**	**None**	**One-site**	**Two-site**	**None**
Antiparallel	82	26	3	121	32	0	0	116	7
Parallel	3	16	0	5	28	2	0	73	0
All	**85**	**42**	**3**	**126**	**60**	**2**	**0**	**189**	**7**

Table 3: Strand length, charged residue frequency, and residues showing large frequency differences between dimorphic and standard β-strand interfaces.

	Dimorphic interfaces	Standard β-strand interfaces
Average strand length (standard deviation)	4.1 (2.0)	7.1 (2.9)
Percentage of charged amino acids:		
Positively charged, including histidine	Arg: 8.1, Lys: 6.2, His: 0.9	Arg: 5.1, Lys: 4.8, His: 2.6
Negatively charged	Glu: 3.1, Asp: 2.5	Glu: 5.2, Asp: 2.1
Amino acids displaying at least a 25% increase of frequency between dimorphic and standard β-strand interfaces	Pro, Trp, Arg, Lys, Tyr	His, Ala, Glu, Gly

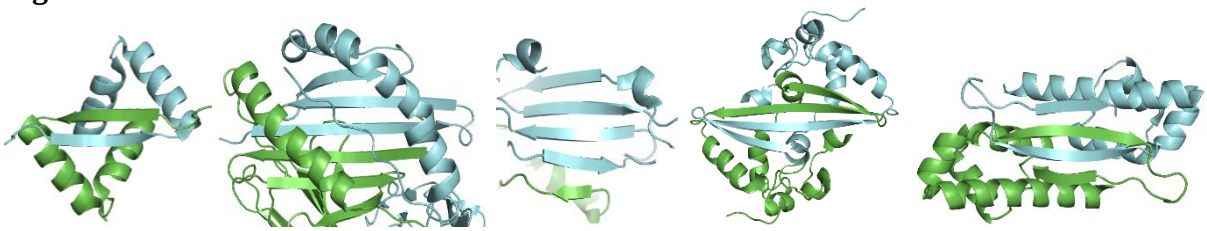
Table 4: Confusion matrix reporting the classifier's performance in discriminating between dimorphic and standard β-strand interface sequences, with dimorphic β-strand interfaces specified as the positive class.

		Predicted class	
		Dimorphic	**Standard**
Actual class	**Dimorphic**	TP: $\overline{23.39}$ ($\sigma = 2.59$)	FN: $\overline{9.61}$ ($\sigma = 2.59$)
	Standard	FP: $\overline{12.54}$ ($\sigma = 3.09$)	TN: $\overline{32.46}$ ($\sigma = 3.09$)

Figures

Figure 1:

Figure 2:



a) b) c) d1) d2)

Figure 3:

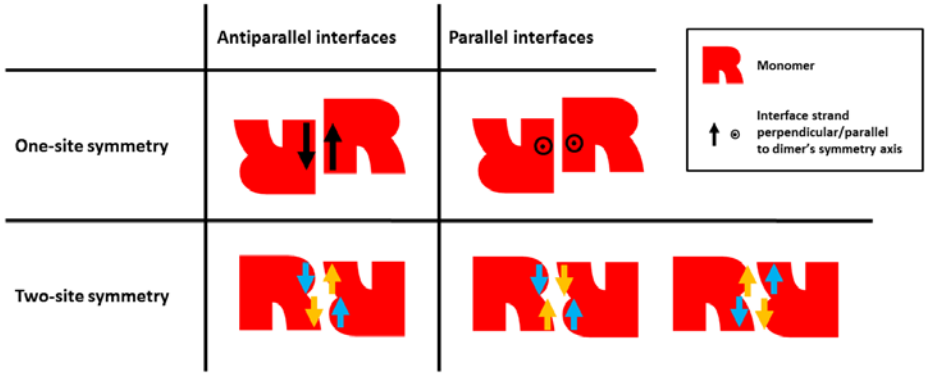


Figure 4:

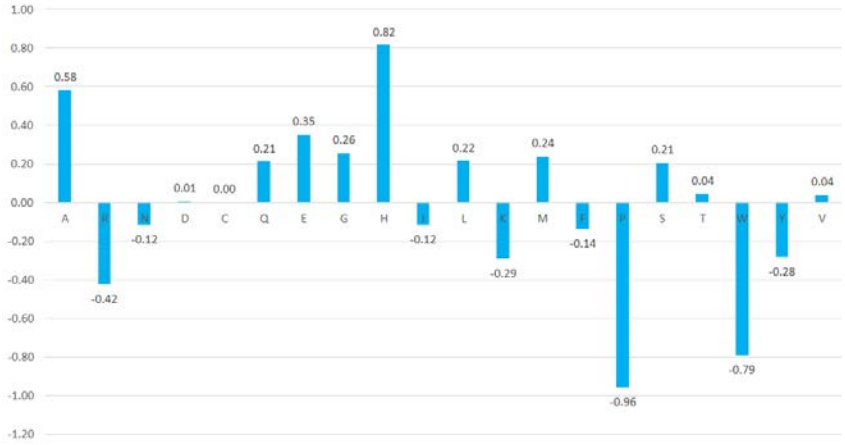


Figure 5:

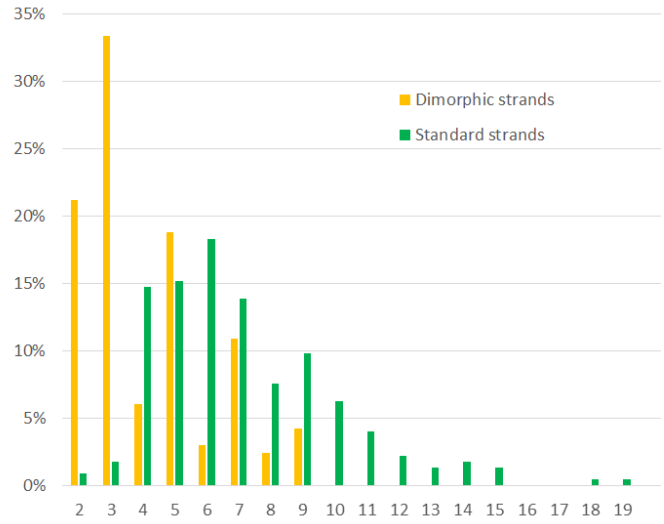


Figure 6:

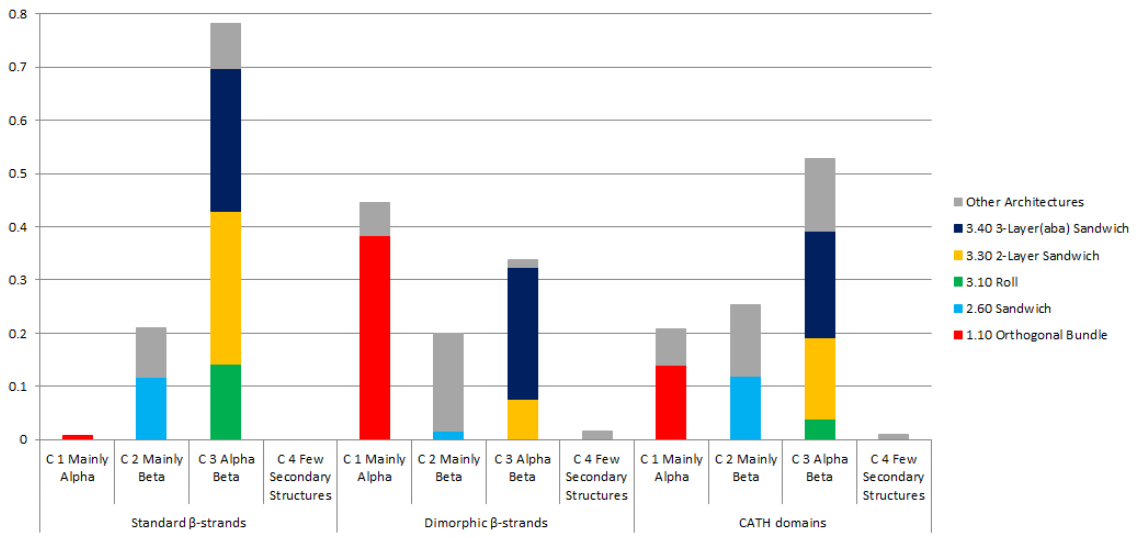
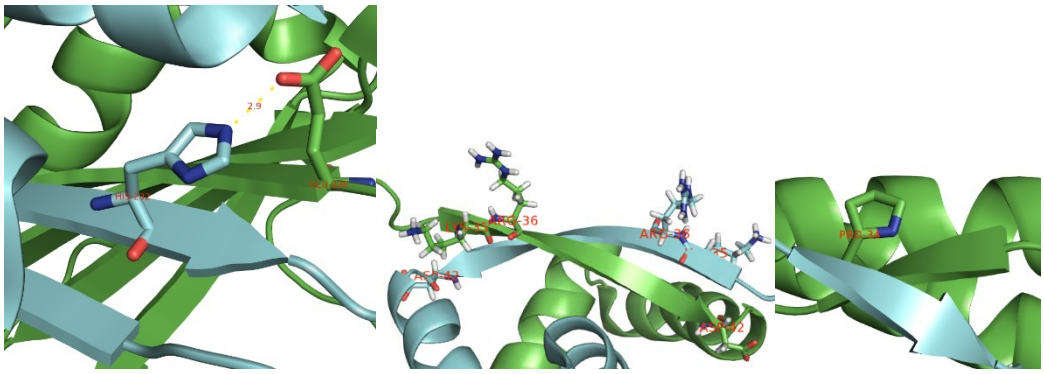


Figure 7:



A) B.a) B.b)

Figure 8:

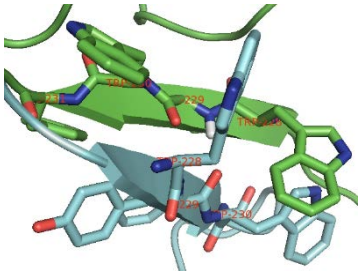


Figure 9:

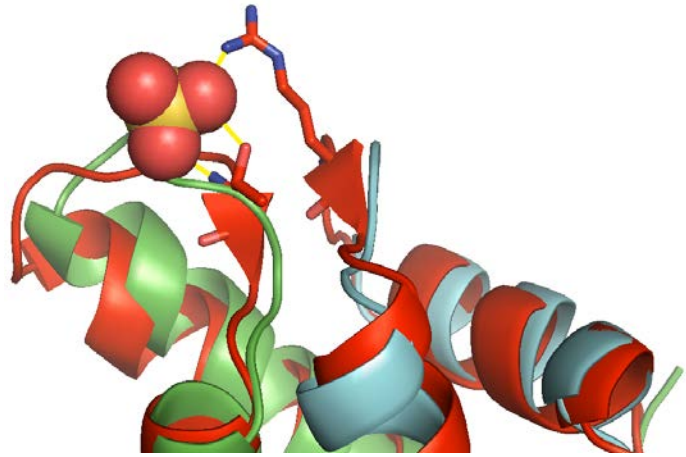


Figure 10:

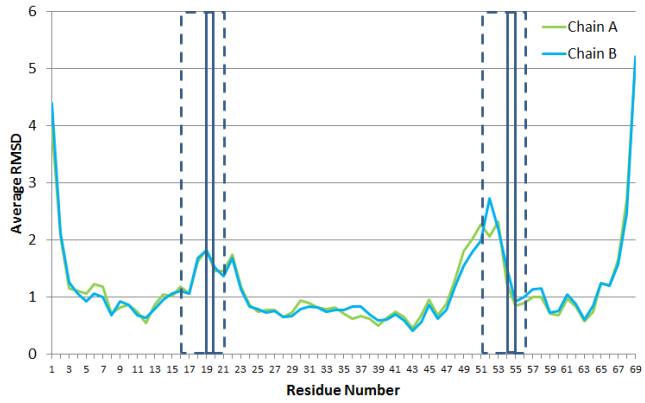


Figure 11:

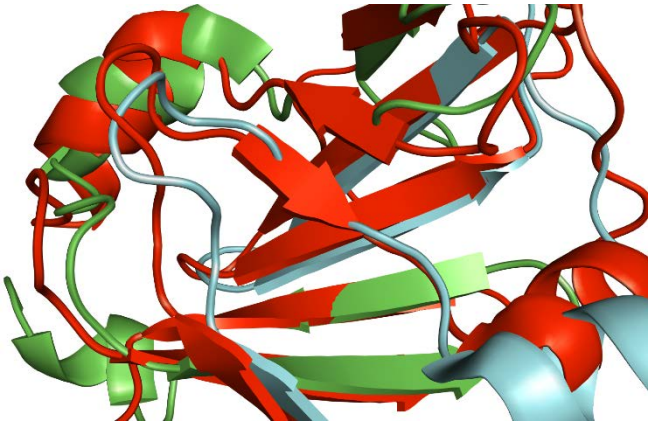


Figure 12:

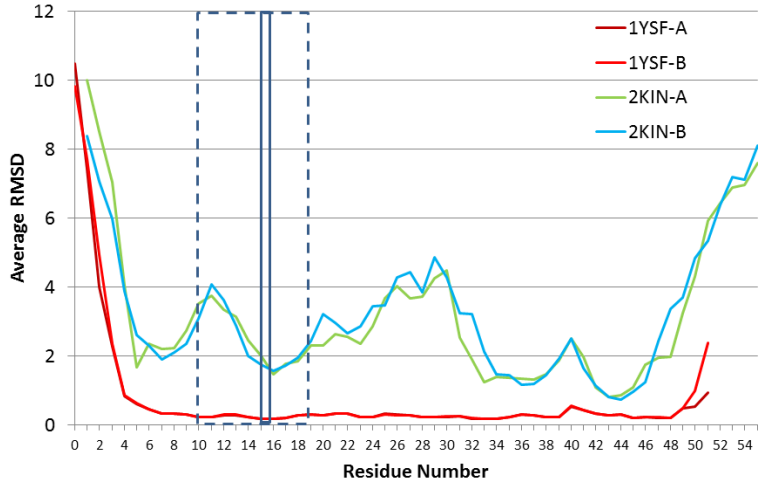


Figure 13:

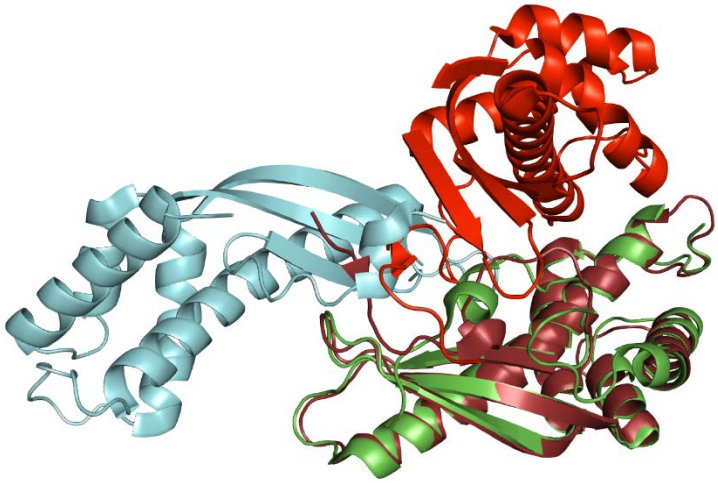


Table S1: Classifier generalised performance estimate metrics showing the arithmetic mean and standard deviation from the outer-cross-validation procedure. Note: (a) ACC, MCC, and AUC-ROC are classifier performance measures. (b) TPR, TNR, PPV, F1, and AUC-PR are class performance measures.

Metric	**Formula**	**Positive class: Dimorphic β-strands**	**Negative class: Standard β-strands**
Recall (TPR)	$\frac{TP}{P}$	$\overline{0.71}$ ($\sigma = 0.08$)	$\overline{0.72}$ ($\sigma = 0.07$)
Specificity (TNR)	$\frac{TN}{N}$	$\overline{0.72}$ ($\sigma = 0.07$)	$\overline{0.71}$ ($\sigma = 0.08$)
Precision (PPV)	$\frac{TP}{TP + FP}$	$\overline{0.65}$ ($\sigma = 0.06$)	$\overline{0.77}$ ($\sigma = 0.05$)
F1	$2 \times \frac{Precision \times Recall}{Precision + Recall}$	$\overline{0.68}$ ($\sigma = 0.05$)	$\overline{0.74}$ ($\sigma = 0.05$)
AUC PR (AP)	$\sum_{k=1}^{n} p(k)\Delta r(k)$	$\overline{0.77}$ ($\sigma = 0.05$)	$\overline{0.82}$ ($\sigma = 0.05$)
Accuracy (ACC)	$\frac{TP + TN}{P + N}$	$\overline{0.72}$ ($\sigma = 0.05$)	
MCC	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$	$\overline{0.43}$ ($\sigma = 0.10$)	
AUC ROC	$\frac{1}{P \times N} \sum_{i=0}^{P} \sum_{k=i+1}^{N} N(k) > N(i)$	$\overline{0.79}$ ($\sigma = 0.05$)	