Data Mining the Protein Data Bank to Identify and Characterise Chameleon Coil Sequences that Form Symmetric Homodimer β-sheet Interfaces

Johanna Laibe[1], Melanie Broutin[2], Aaron Caffrey[1], Barbara Pierscionek[1] 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
k1552417@kingston.ac.uk

Abstract. A protein's environment may affect its secondary structure. In this study, the focus is on homodimers with symmetric β-sheet interfaces resulting from the conversion of coil sequences into β-strands. All homodimers in the Protein Data Bank relying on those chameleon sequences have been identified. Initial analysis based on sequential and structural features has revealed that many of those dimers display specific properties which could contribute to their detection. Such result is important since it could provide some insight on dimerisation and possibly aggregation mechanisms.

Keywords: Proteins, Homodimerisation, Intermolecular β-strand Interfaces, Chameleon Sequences, Aggregation

1. Introduction

A protein consists of a chain of amino acids which generally folds spontaneously into a unique three-dimensional conformation corresponding to its global energy minimum [1]. Failure of adopting that structure may lead to loss of function and even harmful effects [6]. As winners of the Paracelsus challenge [18] have shown, a limited number of mutations can dramatically change a protein conformation: a protein which adopts a four helix conformation was designed while retaining 50% identity of a predominantly β-sheet protein [5]. Similarly, it was demonstrated that mutation of a single amino acid could be sufficient to convert a β-strand into an α-helix [23]. In addition to mutations, a protein's environment may also affect its secondary structure. For example, it has been shown that the prion protein, PrP^{C}, changes its conformation and forms aggregates when interacting with one of its isoforms PrP^{SC} [17]. Those β-sheet aggregates are called amyloid fibrils [4] and have been linked to several human diseases including Alzheimer's, Parkinson's and Creutzfeldt–Jakob's [7].

This study investigates secondary structure alteration resulting from homodimerisation. More specifically, it focuses on coil sequences forming symmetric intermolec-

ular β-strand interfaces. Following exhaustive search in the Protein Data Bank [3], properties of those 'chameleon' fragments were analysed. This led to the identification of specific features which should contribute to their detection and provide some insight on dimerisation and possibly aggregation mechanisms.

2. Methodology

Since very few proteins displaying that 'chameleon' property have been reported in the literature, with the notable exception of the Met-repressor like family, where all members share a similar ribbon-helix-helix structure that forms a homodimer interface by conversion of their ribbon into a β-strand [9], see figure 1, an exhaustive search was conducting using the Protein Data Bank [3]. This was performed according to the following process.

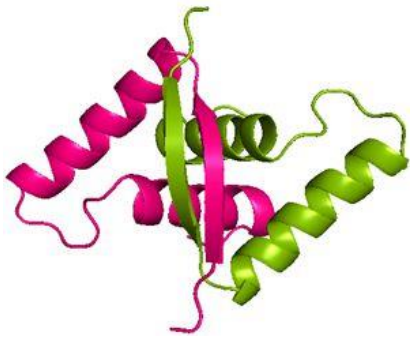


Fig. 1: Met-repressor like family interface (PDB 2P24): this symmetric interface is formed by the interaction of a ribbon-helix-helix pattern (RHH) from each chain. In the process, RHH converts to the β-strand-helix-helix pattern.

Firstly, the whole PDB was filtered to remove entries that don't contain two identical protein chains. Models with sequences with more than 30% identity were also discarded so that the set did not contain homologous proteins.

Secondly, homodimers interacting through at least an interface composed of a β-sheet were identified. This was performed by detecting the presence of amino acids belonging to β-strands from different chains whose C-alphas are within 5Å from each other, i.e. the interaction distance used by the CAPRI community-wide experiment (Critical Assessment of Prediction of Interactions) [10] which corresponds to the distance between two carbons alpha in a hydrogen bond.

Thirdly, for each remaining homodimer interacting through a β-sheet, information available in the 'SHEET' field of the PDB file was extracted to collect the interacting β-strand sequences, their nature, i.e. parallel or anti-parallel, and the number of strands forming the sheet involved in the interface. All anti-parallel interfaces of homodimers were then classified into two categories: the 'chameleon' interfaces, which are formed of exactly two β-strands each of them belonging to a different chain, i.e.

the corresponding fragments would have a coil structure in the monomer form, and the 'standard' interfaces, which are formed of a β-sheet composed of at least four β-strands where each chain provides at least two β-strands, i.e. the corresponding fragments would already belong to a β-sheet in the monomer form. Although the existence of 'hybrid' interfaces, i.e. formed by one strand from one chain and two or more strands from the other chain, was also detected, they were not considered further in this study since their mixed environment would not be useful in identifying discriminative properties of chameleon fragments.

Finally, since analysis of the nature of the remaining interaction strands revealed that 90% of 'chameleon' interfaces are anti-parallel, and, among them, 70% are symmetric, it was decided to focus this study on those interfaces. In this work, a β-sheet interface was classified as symmetric, if both strands have the same amino acid sequence. Eventually, this process produced a dataset of 249 anti-parallel symmetric homodimer interfaces from non-homologous proteins: it comprises 80 'chameleon' and 169 'standard' interfaces.

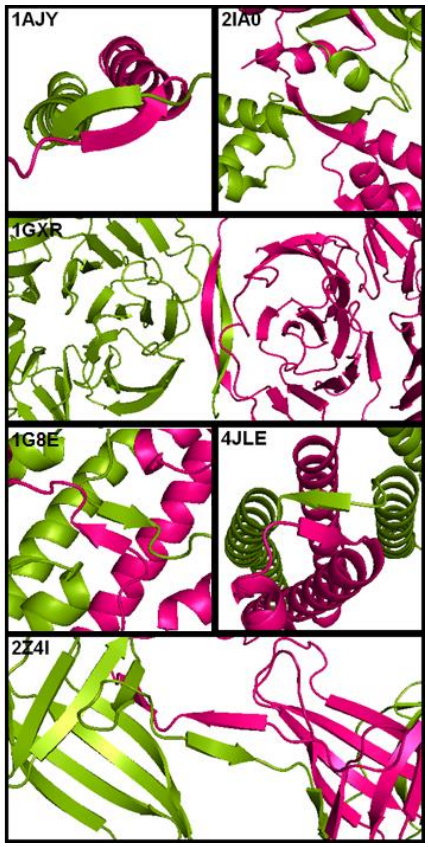


Fig. 2: Example of homodimers displaying symmetric anti-parallel 'chameleon' interfaces

To analyse differences between chameleon and standard fragments, a set of properties was calculated for the two classes of interfaces under consideration. Firstly, since many protein interfaces (~1/3) display a recognizable hydrophobic core [13], hydrophobicity of those protein interfaces was estimated. This was performed by calculating the grand average of hydropathy (GRAVY) value [11].

For each strand $\boldsymbol{S_i}$ of length $\boldsymbol{n_i}$, its GRAVY values, $\boldsymbol{G_i}$, is defined as:

$$\boldsymbol{G_i} = (\Sigma_j \boldsymbol{H}_{ij}) / \boldsymbol{n}_i \quad (1)$$

where $\boldsymbol{H}_{ij}$ is the hydropathy value of amino acid $\boldsymbol{j}$ in the strand $\boldsymbol{S}_i$.

Secondly, given that β-sheets are created by interaction of β-strands through backbone hydrogen bonds, interface hydrogen bond propensity may be informative about interface type. Using the structural information associated to each homodimer in its PDB file, all hydrogen bonds were retrieved from each β-sheet interface using the RING software with a 3.5Å threshold and the 'Closest' and 'Multiple' parameters, so that all atoms and multiple interactions are considered per residue pair, respectively [15].

Since a backbone residue can form up to 2 hydrogen bonds with an adjacent strand, for each strand $\boldsymbol{S_i}$ of length $\boldsymbol{n_i}$, its hydrogen bond propensity, $\boldsymbol{HB_i}$, is defined as:

$$\boldsymbol{HB_i} = (\Sigma_j \boldsymbol{B}_{ij}) / 2\boldsymbol{n}_i \quad (2)$$

where $\boldsymbol{B}_{ij}$ is the number of backbone hydrogen bonds formed by amino acid $\boldsymbol{j}$ in the strand $\boldsymbol{S}_i$.

Thirdly, as experiments have shown that stability of antiparallel β-sheets is affected by their length [19], average strand length was calculated for each set. Finally, propensities of all amino acids were calculated.

	Chameleon interfaces	Standard interfaces
Average hydrophobicity	0.59	0.52
Average hydrogen bond propensity	0.43	0.42
Average strand length	5.0	7.9

Table 1: Average hydrophobicity, hydrogen bond propensity and strand length of chameleon and standard interfaces

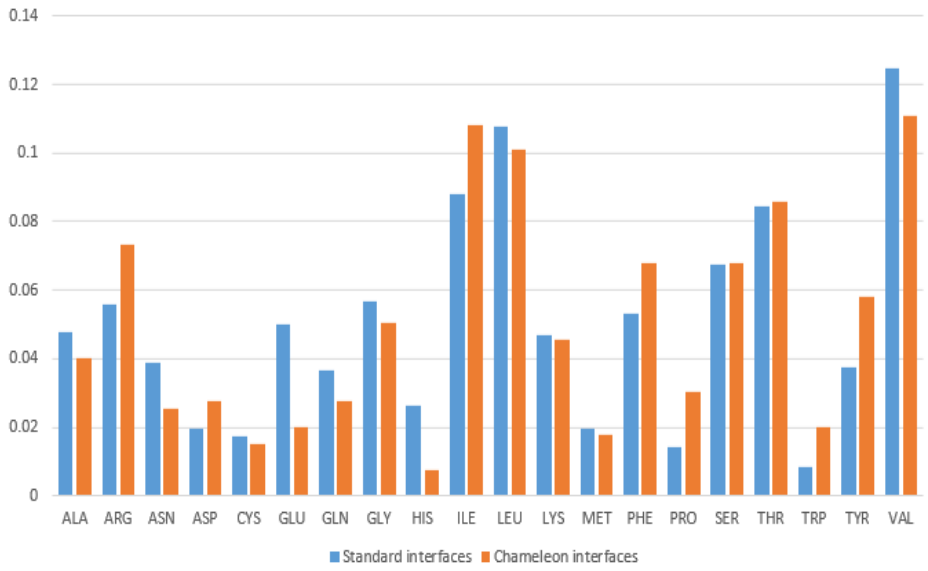


Fig. 3: Amino acid propensities of chameleon and standard interfaces

While Table 1 presents average hydrophobicity, hydrogen bond propensity and strand length of chameleon and standard anti-parallel homodimer β-sheet interfaces, Figure 3 show their amino acid propensities. One observes that neither average hydrophobicity nor average hydrogen bond propensity is affected by the interface type. On the other hand, chameleon interfaces are much shorter than standard interfaces which are three residues longer in average. Moreover, there are significant differences in their amino acid propensity profiles in particular for aromatic and charged amino acids.

To explore combinations of features whish may allow discriminating chameleon fragments, unsupervised clustering was performed using different sets of features. More specifically, data were processed using a general purpose clustering tool, CLUTO [16], which has been used in a variety of bioinformatics applications [8], [2], [14], [12]. In order to give each feature equal weight, a normalization process is applied. For each feature $\boldsymbol{F}$, its values, $\boldsymbol{F}_i$, are normalised between 1 and -1 [20] as:

$$\boldsymbol{F}_{i\_normalised} = 2\,(\boldsymbol{F}_i - \boldsymbol{F}_{min})\,/\,(\boldsymbol{F}_{max} - \boldsymbol{F}_{min}) - 1 \qquad (3)$$

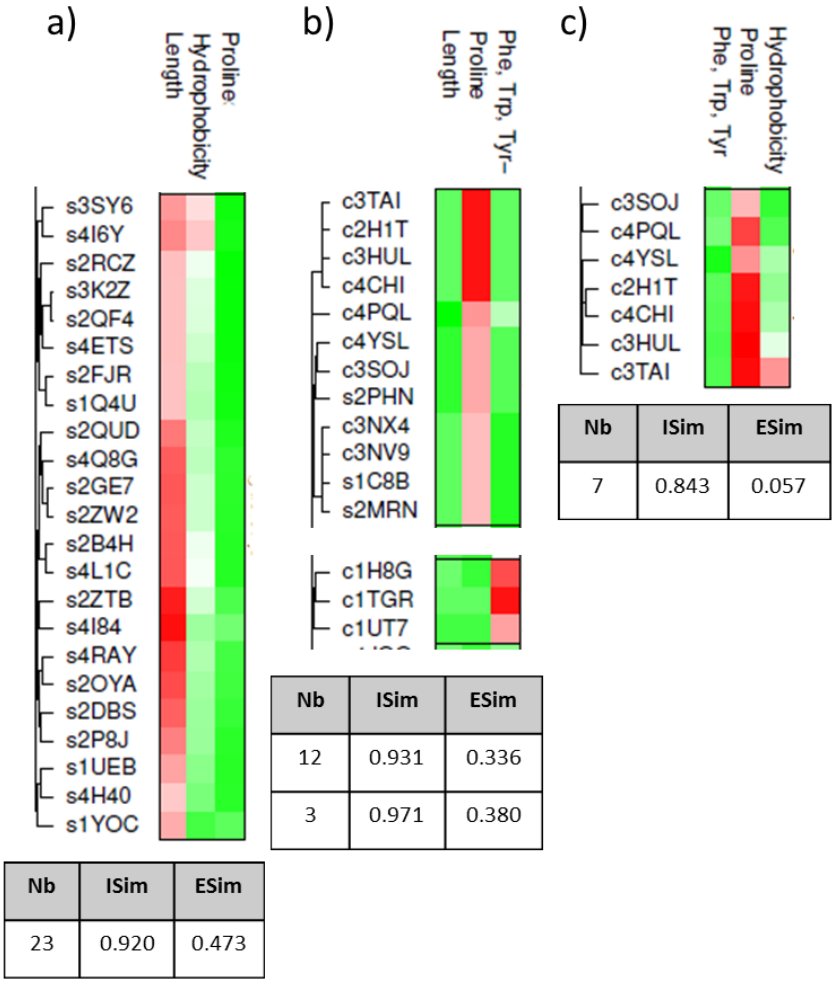
where $\boldsymbol{F}_{max}$ and $\boldsymbol{F}_{min}$ represent the maximum and minimum values of the feature $\boldsymbol{F}$.

Using hierarchical partitional clustering, CLUTO produces a binary tree representing similarities between interface profiles and identifies specific clusters within the tree. Note that the quality of each cluster is estimated by its internal similarity (ISim), i.e. the average similarity between the interfaces of the cluster, and its external similarity (ESim), i.e. the average similarity between the interfaces of the cluster and all the other interfaces. The "ideal" cluster would have: ISim=1.0 and ESim=0.0.

In addition, CLUTO displays feature values for each interface using a colour palette: shades of green and red indicate feature values between -1 and 1 respectively.

3. Results

Informed by results presented in Table 1 and Figure 3, all interfaces of interest were clustered using CLUTO and a combination of features including length, hydrophobicity, proline, aromatic (without histidine) and charged amino acid propensities. Figure 4 shows the most disciminative clusters produced using subsets of those properties.



Nb	ISim	ESim
23	0.920	0.473

Nb	ISim	ESim
12	0.931	0.336
3	0.971	0.380

Nb	ISim	ESim
7	0.843	0.057

Fig. 4: Good quality interface clusters created by CLUTO according to different sets of properties. The prefix added to PDB ids specifies if an interface is chameleon, “c”, or, standard, “s”.

Based on length, hydrophobicity and proline propensity, an homogeneous cluster of relatively good quality allows to discriminate 23 “standard” interfaces, see Figure 4.a. All those interfaces display a high length, low proline propensity and relatively average hydrophobicity. Usage of length, proline and non charged aromatic amino acids (Phe, Trp and Tyr) reveals two good quality clusters, see Figure 4.b, populated mainly of “chameleon” interfaces – 12 “chameleon” and only 3 “standard”: both are composed of short interfaces, but one has a high proline propensity, see Figure 5, while the other one has a high aromatic propensity, see Figure 6.

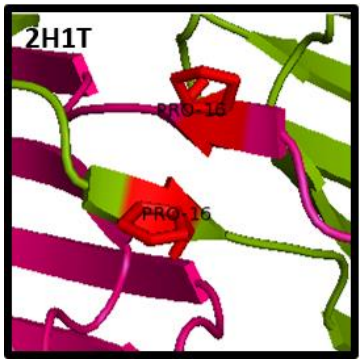


Fig. 5: Example of chameleon interface involving a proline.

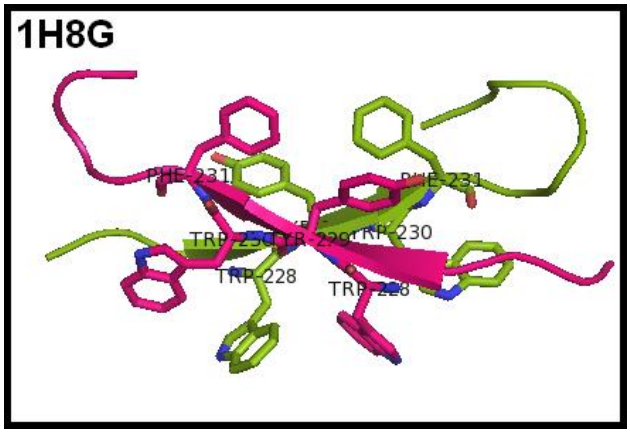


Fig. 6: Example of chameleon interface suppoted by pi-pi interactions between aromatic amino acids.

Interstingly, if length is substituted by hydrophobicity, the high proline propensity group is reduced from 12 to 7 members, but is only composed of “chameleon” proteins, see Figure 4.c. Note that among the only 3 “standard” interfaces with high proline content classifed in a largely chameleon cluster, one of them, PDB id 1C8B, displays a sheet structure which is “almost” chamelon, since the non interface strands are much shorter than the interface ones, See Figure 7.

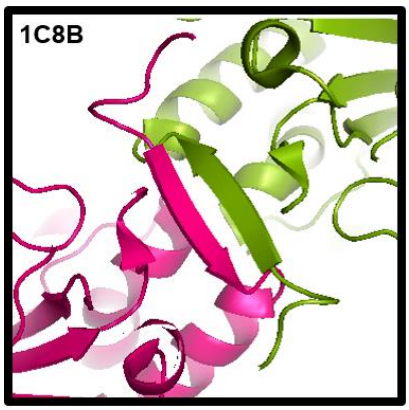


Fig. 7: "Standard" interfaces where non interface strands are much shorter than the interface ones.

Since usage of strand length proved useful to produce the clusters shown on Figure 4.a and 4.b, it was also used as sole feature to discriminate between chameleon and standard interfaces: whereas among interfaces based on strands of length 3 amino acids, 91% of them, i.e. 30, are chameleon, all strands of size 10 or more form standard interfaces, i.e. 48.

This initial analysis of chameleon interfaces has revealed that a many of those chameleon dimers (45%) display properties, i.e. short length, high aromatic or proline propensity, allowing to discriminate them from standard ones. Moreover, this study suggests that there are unlikely to form long β-strands since none of them was composed of 10 or more residues. There is no doubt that more advanced machine learning approaches, such as support vector machines, neuron networks and decision trees [21, 22], would allow to combine the identified features and others to further characterise chameleon interfaces. Since, many chameleon fragments have been associated to human diseases through aggregation [4,7,17], the ability to detect a specific class of chameleon fragments, i.e. those able to form symmetric homodimer β-sheet interfaces, should contribute, not only, to a better insight about homodimerisation, but also in aggregation mechanisms.

4. Conclusion

This study has identified in the Protein Data Bank all symmetric homodimers relying on β-sheet interfaces involving the conversion of coil sequences into β-strands. Initial comparison with standard intermolecular β-strand interfaces has revealed that many of those chameleon dimers display specific properties which should contribute to their detection. When possible, this could provide some insight on homodimerisation and possibly aggregation mechanisms.

5. References

[1] Anfinsen, C.B., et al.: The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. USA. **47**, 1309-1314 (1961).

[2] Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., Kämper, J.: Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics. **21(**7), 1069-1077 (2005)

[3] Berman, H.M., et al.: The Protein Data Bank. Nucleic Acids Res. **28**, 235-242 (2000).

[4] Chiti, F., Dobson, C.M.: Protein misfolding, functional amyloid, and human disease. Annu. Rev. Bioch. **75,** 333–366 (2006).

[5] Dalal, S., Balasubramanian, S., Regan, L.: Protein alchemy: changing β-sheet into α-helix. Nature structural biology. **4.7**, 548-552 (1997).

[6] Dobson, C.M.: The structural basis of protein folding and its links with human disease. Phil Trans R Soc Lond B. **356**, 133-145 (2001).

[7] Eisenberg, D., Jucker, M.: The amyloid state of proteins in human diseases. Cell. **148**, 1188-1203 (2012).

[8] Glazko, G.V., Mushegian, A.R.: Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. Genome Biol. **5**, R32 (2004).

[9] Golovanov, A.P., Barilla, D., Golovanova, M., Hayes, F., Lian, L.-Y.: Parg, A protein required for active partition of bacterial plasmids, has a dimeric Ribbon-Helix-Helix structure. Molecular Microbiology. **50**, 1141-1153 (2003).

[10] Jauch, R., Yeo, H.C., Kolatkar, P.R., Clarke, N.D.: Assessment of CASP7 structure predictions for template free targets. Proteins Struct. Funct. Bioinforma. **69**, 57–67 (2007).

[11] Kyte, Jack, Russell, F.: Doolittle. A Simple Method For Displaying The Hydropathic Character Of A Protein. Journal of Molecular Biology **157.1**, 105-132 (1982).

[12] Lanara, Z., Giannopoulou, E., Fullen, M., Kostantinopoulos, E., Nebel, J.-C., Kalofonos, H.P., Patinos, G.P., Pavlidis, C.: Comparative study and meta-analysis of meta-analysis studies for the correlation of genomic markers with early cancer detection. Human Genomics. 7:14, (2013)

[13] Larsen, T.A., Olson, A.J., Goodsell, D.S.: Morphology of protein–protein interfaces. Structure. **6**, 421-427 (1998).

[14] Nebel, J.-C., Herzyk, P., Gilbert, D.R.: Automatic generation of 3D motifs for classification of protein binding sites. BMC Bioinformatics. 8:32 (2007).

[15] Piovesan, D., Minervini, G., Tosatto, S.C.E.: The RING 2.0 web server for high quality residue interaction networks. Nucleic Acids Research. (2016).

[16] Rasmussen, M.D., Deshpande, M.S., Karypis, G., Johnson, J., Crow, J.A., Retzel, E.F.: wCLUTO: a Web-enabled clustering toolkit. Plant Physiol. **133**(2), 510–516 (2003).

[17] Roostaee, A., Cote, S., Roucou, X.: Aggregation and amyloid fibril formation induced by chemical dimerization of recombinant prion protein in physiological-like conditions. Journal of Biological Chemistry. **284(**45), 30907-30916 (2009).

[18] Rose, G., Creamer, T.: Protein folding: predicting predicting. Proteins: Structure, function, and genetics. **19(**1), 1-3 (1994).

[19] Stanger,,H.E., Syud, F.A., Espinosa, J.F., Giriat, I., Muir, T., Gellman, S.H.: Length-dependent stability and strand length limits in antiparallel β-sheet secondary structure. Proc. Natl Acad. Sci. USA. **98**, 12015–12020 (2001).

[20] Su, C.-H., Pal, N.R., Lin, K.-L., Chung, I.-F.: Identification of Amino Acid Propensities That Are Strong Determinants of Linear B-cell Epitope Using Neural Networks. PLoS ONE **7**(2), e30617 (2012). doi:10.1371/journal.pone.0030617

[21] Vapnik, V.N., Kotz, S.: Estimation of Dependences. Based on Empirical Data, Springer, ISBN 0-387-30865-2 (2006).
[22] Winston, P.H.: Artificial Intelligence. 3rd ed. Addison-Wesley (1992).
[23] Yang, W.Z., et al.: Conversion of a B-strand to an α-Helix induced by a single-site mutation observed in the crystal structure of Fis mutant Pro26ala. Protein Sci. **7**(9), 1875-1883 (1998).