Analysis of protein chameleon sequence characteristics.

Amine Ghozlane, Agnel Joseph, Aurélie Bornot, Alexandre de Brevern

Analysis of protein chameleon sequence characteristics

Amine Ghozlane[1], Agnel Praveen Joseph[1,2], Aurélie Bornot[1] & Alexandre G. de Brevern[1*]

[1] Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM UMR-S 726, DSIMB, Université Paris Diderot- Paris 7, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France.
[2] Molecular Biophysics Unit, Indian Institute of Science Bangalore 560 012, India.

E-mails :amine.ghozlane@edu.univ-paris-diderot.fr
agnel@mbu.iisc.ernet.in
aurelie.bornot@univ-paris-diderot.fr
alexandre.debrevern@univ-paris-diderot.fr

* Corresponding author:
mailing address: Dr. de Brevern A.G., Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM UMR-S 726, DSIMB, Université Paris Diderot- Paris 7, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France.
E-mail : alexandre.debrevern@univ-paris-diderot.fr
Tel: (33) 1 44 49 30 38
Fax: (33) 1 47 34 74 31

Abstract

Conversion of local structural state of a protein from an α-helix to a β-strand is usually associated with a major change in the tertiary structure. Similar changes were observed during the self assembly of amyloidogenic proteins to form fibrils, which are implicated in severe diseases conditions, *e.g.*, Alzheimer disease. Studies have emphasized that certain protein sequence fragments known as chameleon sequences do not have a strong preference for either helical or the extended conformations. Surprisingly, the information on the local sequence neighborhood can be used to predict their secondary at a high accuracy level. Here we report a large scale-analysis of chameleon sequences to estimate their propensities to be associated with different local structural states such as α-helices, β-strands and coils. With the help of the propensity information derived from the amino acid composition, we underline their complexity, as more than one quarter of them prefer coil state over to the regular secondary structures. About half of them show preference for both α-helix and β-sheet conformations and either of these two states is favored by the rest.

Background.

Repetitive secondary structures like α-helices and β-strands have been viewed as key building blocks of proteins. These local protein structures are stabilized mainly by hydrogen bonds within the protein backbone. In 1984, Kabsch and Sander identified identical fragment sequences of limited length found in both α-helices and β-strands, namely chameleon sequences [1]. This suggests that only local sequence composition and the order of amino acids are not sufficient to predict the secondary structure accurately [2]. The number of examples supporting the above speculation has strikingly increased in the recent past [3]. Elegant experimental studies have shown the importance of non-local interactions to guide the formation of the α-helix or β-strand, e.g. the IgG-binding domain of protein G (GB1) [4]. Chameleon sequences have also been designed, *e.g.* MATa2 and MCM1 DNA complexes [5]. Studies have emphasized that these

chameleon sequences, have no strong preference for either α-helical or β-strand conformations [6]. Nonetheless, the information on the local sequence neighborhood can be used to predict their secondary at a high accuracy level [3, 7]. Here, we have analyzed chameleon sequences to estimate their propensities to form not only the regular secondary structures like α-helix or β-strand , but also coil [8].

Description.

Unlike the previous studies that focused only on limited parts of the Protein DataBank [9], all the protein structures available in 2007 (~40.000 protein structures) have been used. Secondary structures have been assigned for these proteins using the DSSP algorithm [10]. Only those proteins with complete side-chain co-ordinates and without multiple breaks in the chain were considered, leading to a final number of 14,692,070 amino acid residues associated to a given secondary structure. The 8 secondary structural assignments made by DSSP were reduced to the 3 classical states: helix includes α, 3_{10} and π-helices, strand has only the β-strand assignments, and coil covering the rest of the assignments (β-bridges, turns, bends, and coil). Default parameters of the program have been used.

In the second step, we searched for chameleon sequences of length *L*, *L* ranging from 4 to 8 amino acids. A fragment is considered as a chameleon sequence if all the residues in this fragment are associated at least once to the helical conformation and also, at least once to the β-strand. Thus, numerous chameleon sequences have been located: 63,228 (for *L*=4 residues), 34,408 (for *L*=5), 2,423 (for *L*=6), 179 (for *L*=7) and 64 (for *L*=8). As the dataset is large and complete when compared to the ones used in previous studies, more examples were found, especially for the longer fragments [3].

Our main goal is to check whether the chameleon sequences don't have any strong preference for either helical or strand conformations [6], and also to extend the questioning to the preference of chameleon sequences for the coil state, a question not directly tackled in the previous works. For

this purpose, we have used a simple methodology. We have used a non-redundant databank containing proteins with not more than 20% pairwise sequence identity. The selected chains have X-ray crystallographic resolutions less than 1.6 Å, with a R-factor less than 0.25 (details can be found in [11]). Using this non-redundant databank, the propensity of an amino acid *k* to be associated to a given secondary structure state *i*, namely p_i^k, has been computed. *i* corresponds to α-helix, β-strand or the coil state, while *k* corresponds to one of the 20 amino acids:

$$p_i^k = \frac{f_i^k}{f^k}$$

with f_i^k the frequency of amino acid *k* to occur in the secondary structure state *i*, and f^k the frequency of occurrence of amino acid *k* in the databank. Then for each chameleon sequence X^S, an adequacy score S_i was computed as:

$$S_i\ (X^S) = \prod_{k=aa^S}^{k=aa^{S+L}} (p_i^k)$$

Hence, each chameleon sequence X^S is associated to a score S_α, S_β and S_{coil} As these scores are propensity products, a score S_i of 1.0 corresponds to the random value. If S_i is higher than one, this chameleon sequence is found preferentially associated with the secondary state *i* and vice versa. This measure is crude but gives some basic insights into the behaviors of chameleon sequence.

Figure 1a shows a plot of S_α versus S_β for the 63,228 chameleon sequences (for *L*=4 residues). Adequacy scores greater than 4.0 were set to a maximum value of 4.0. The figure shows that 53.7% and 47.3% of the chameleon sequences have S_β and Sα scores greater than 1.0 respectively. Thus, each square delineated by the red lines are quite equivalent. S_β scores go far beyond S_α scores, as 16% of the S_β scores are greater than 2.0, 5.3% than 3.0 and 2.7% than 4.0, while only 5.1% of the S_α scores are greater than 2.0 and 0.2% than 3.0.

21.6% of the chameleon sequences have S_β and S_α scores greater than one, with an average S_{coil} of 0.42 (*i.e.* less than two times the random value). For 25.7% of these fragments, α-helix is

statistically preferred over β-strand, with an average S_{coil} of 0.68, while for 24.7%, only β-strand is preferred (average S_{coil}. of 0.65). Interestingly, 27.9% of the chameleon sequences have S_α and S_β less than 1.0, *i.e.*, the coil state is favored.

Figure 1b shows the chameleon sequence fragment MLIL that have S_α and S_β scores greater than 2.0 (shown as the blue dot in Figure 1a). In type-1 beta-hydroxysteroid 2 dehydrogenase, this chameleon sequence forms the central β-strand of a β-sheet composed of 5 β-strands (Figure 1b left), while in hyperthermophilic tungstoperin enzyme 2 aldehyde ferredoxin oxidoreductase, this sequence is in the middle of a long α-helix (Figure 1b right).

With this simple approach, we have underlined that chameleon sequences have no strong preference for either α- or β-conformation. We have also found that very different chameleon sequences exist, some showing a higher preference for either helical or strand conformations, some showing preference for both, while some sequences favor the coil state over the regular secondary structures. These observations again support the idea that non-local factors [2, 3] have a major influence over the secondary structure that an amino acid sequence adopts. Supplementary information can be found on our website: http://www.dsimb.inserm.fr/~joseph/chameleon/

Acknowledgements

This work was supported by grants from the Ministère de la Recherche, Université Paris Diderot – Paris 7, Université de Saint-Denis de la Réunion and the French Institute for Health and Medical Care (INSERM). APJ has a grant from CEFIPRA number 3903-E and AB has a grant from Ministère de la Recherche.

References.

[1] W. Kabsch & C. Sander, *Proc Natl Acad Sci U S A*, (1984) **81**: 1075 [PMID: 6422466].
[2] B.I. Cohen *et al*., *Protein Sci*. (1993) **2**: 2134 [PMID : 8298461].
[3] J.T. Guo *et al*., *Proteins*, (2007) **67**: 548 [PMID: 17299764].
[4] D. L. Minor & P.S. Kim, *Nature*, (1996) **380**: 730 [PMID: 8614471].
[5] K. Takano *et al*., *Proteins*, (2007) **68**: 617 [PMID: 17510955].
[6] M. Mezei, *Protein Eng*. (1998) **11**: 411 [PMID: 9725618].
[7] I. Jacoboni *et al*., *Proteins*, (2000) **41**: 535 [PMID: 11056040].
[8] B. Offmann *et al*., *Current Bioinformatics*, (2007) **3**: 165 [http://www.bentham.org/cbio/contabs/cbio2-3.htm#2].
[9] H.M. Berman *et al*., *Nucleic Acid Res*., (2000) **28:** 235 [PMID: 10592235].
[10] W. Kabsch & C. Sander, *Biopolymers*, (1983)**22:** 2577 [PMID: 6667333].
[11] G. Faure *et al*., *Biochimie*, (2008) **90:** 626 [PMID : 18086572].

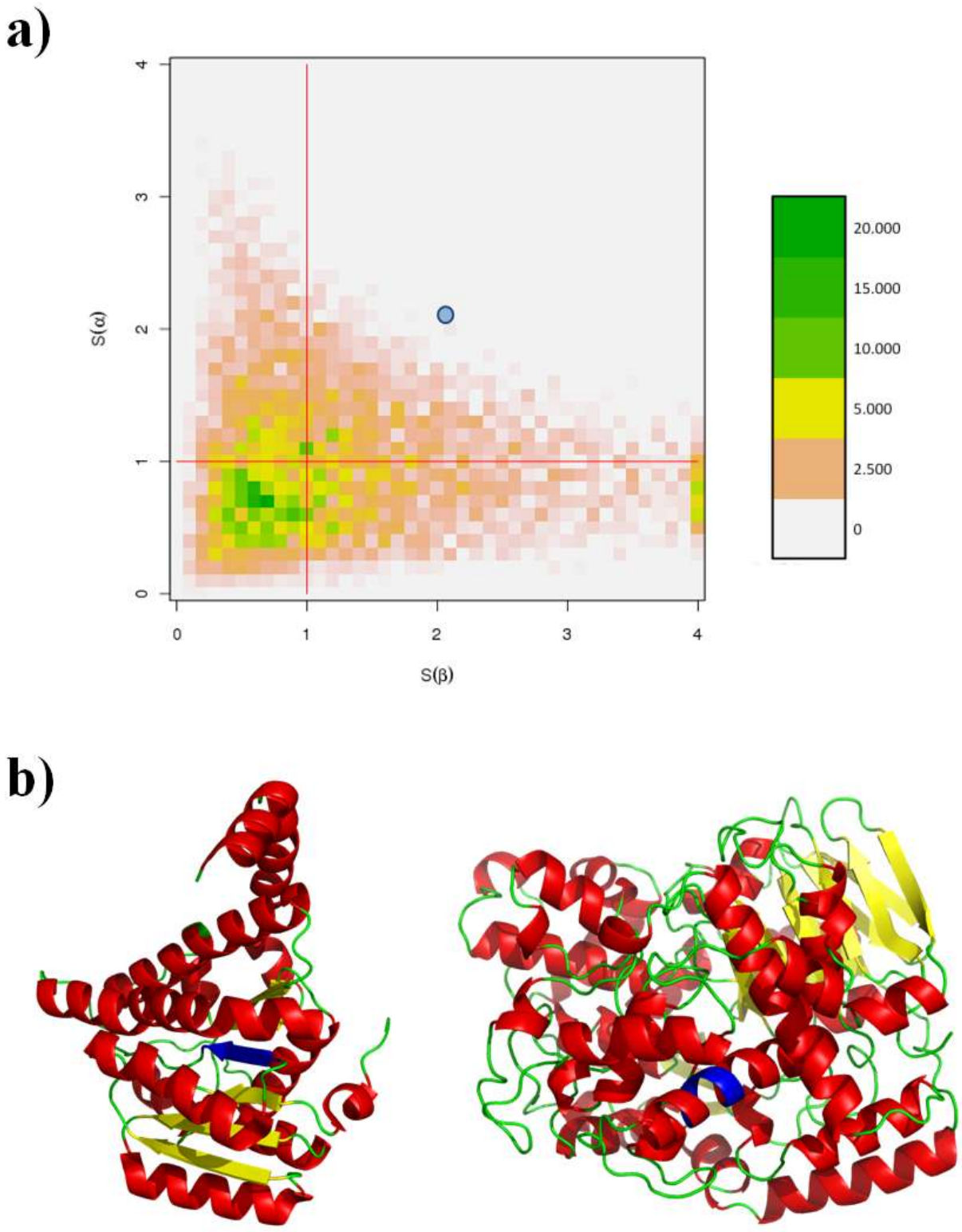

Figure 1. (a) Distribution of adequacy scores S(α) and S(β) of chameleon sequence fragment of length 4. The legend gives the occurrence number of observed fragments. (b) example of the chameleon sequence fragments MLIL found (left) in a β-strand of Guinea pig 11 beta-hydroxysteroid 2 dehydrogenase type 1 (PDB code 1xse) and in an α-helix of a hyperthermophilic tungstoperin enzyme 2 aldehyde ferredoxin oxidoreductase (PDB code 1aor). The blue point in (a) represents the scores of example (b).