Foldability and chameleon propensity of fold-switching protein sequences

Mihaly Mezei

Department of Pharmacological Sciences,
Icahn School of Medicine at Mount Sinai,
New York, NY 10029, USA.

E-Mail: Mihaly.Mezei@mssm.edu
Tel.: +1-212-659-5475

Keywords: Protein fold switching, foldability, chameleon propensity

Short title: On fold-switching sequences

.

Abstract

It has been shown recently by Porter & Looger that a significant number of proteins exist that can form more than one stable fold. This note examines the sequences of these fold-switching proteins by (a) calculating their foldability scores recently introduced by the present author and (b) comparing the propensity of chameleon sequences in fold switchers and in non-fold switchers.

It has been found that the average foldability score of the fold switchers indicates weaker foldability. As for the propensity of chameleon sequences of length 5-7 it was found, somewhat surprisingly, that there is only a very small difference between the fold switchers and the non-fold switchers. Furthermore, when comparing the amino acid propensities in chameleon sequences and in fold switchers, for several amino acids there was even an opposing trend in the deviation of their propensities from the overall amino acid propensities.

1. **Introduction and background.**

While most proteins are known to fold into a unique structure, Porter & Looger[1] have shown that fold switching is not an extreme rarity – they found 95 protein pairs that contain the same sequence but stabilize in different folds. Their work showed that one diagnostic of fold switching is the poor performance of secondary structure prediction algorithms. The present work aimed to find other, purely sequence based, diagnostics of fold switching. First, a recently developed foldability score[2] was calculated for the fold switching sequences. Next, the propensity of chameleon sequences, generated recently[3] from the sequences in the Protein Data Bank (PDB)[4], were obtained; both in sequences that are fold switching and in sequences that are not.

2. **Methods.**

The PDB ids of the fold-switching proteins were obtained from the work by Porter & Looger[1]. The sequences corresponding to the PDB ids were downloaded from the PDB. When the two sequences of a fold-switching pair were of different length, the shorter one was used.

The chameleon sequence set was generated by the program Cham[3] using the same sequence set that was used in Ref[3], referred to as the PDB set. The number of chameleons of length 4, 5, 6 and 7 in that set are 118981, 187803, 36669, and 1822, resp., representing 74.4%, 5.9%, 0.057% and 0.00014%, resp., of their respective n-tuplet space. Since about three quarter of all possible 4-residue sequences are chameleons and there are hardly any chameleons over length 7, only length 5, 6, and 7 were used.

The foldability scores SC_3 and SC_4 of a sequence were obtained from the ratio of propensities of triplets and quadruplets in a filtered version (making sure that no pairs in the filtered set are more than 50% similar) of the PDB set to the propensities of the same triplets and quadruplets expected from the overall amino acid propensities. The scores SC_3 and SC_4 of the fold-switching sequences were calculated with the program Fold[2]. The amino acid

propensities in the chameleon sequences were calculated with the program Cham and in the fold-switching sequences with the program Fold.

The propensities of the chameleon sequences in a set of protein sequences were obtained by the following algorithm:

1. For each protein sequence of length L_s and each chameleon of length L_{ch}, combine the list of chameleons of length L_{ch} with all L-L_{ch} subsequences of length L_{ch} of the protein considered.
2. Sort the combined list
3. Scan the sorted list. For any stretch of n_{mat} identical items check if one of them is from the chameleon list.
4. If a chameleon was among the members of the stretch, n_{mat}-1 chameleons were part of the protein sequence.

This calculation has been added as a new option to the program Fold, that can be downloaded from the URL http://inka.mssm.edu/~mezei/fold.

The comparison of the propensities of chameleons in sequences that are fold switching and those that are not used the 4735 sequences that served for the test of the folding prediction[2]. These 4735 sequences, referred to as the PDB test set, were not part of the PDB set thus they were not used for the generation of the statistics on which the scores SC_3 and SC_4 are based, nor for the chameleon set generation.

The overall amino acid propensities were obtained as the average of the propensities in various organism classes[5].

3. **Results and Discussion.**

The triplet and quadruplet average folding scores and standard deviations of the fold-switching proteins are shown in Table I along with the corresponding results for the (filtered) PDB set, a set of Intrinsically Disordered Proteins (IDP)[6], as well as for two randomly generated sets, using either the uniform distribution or the amino-acid

propensity-weighted distribution. For a graphical representation of the foldability scores Fig. 1 presents a ‘foldability meter’ showing where the various sequence sets fall on the foldability scale.

There is a clear progression in the foldability scores from the PDB test set to the uniform random set, with the fold-switching set in the middle, albeit well within the foldability range. This finding of weakened foldability score can be interpreted in the same vein as the observation of Porter & Looger: the capability of a sequence of folding into two different stable conformations tends to ‘confuse’ the otherwise reasonably reliable prediction algorithms.

The propensities of finding chameleon sequences in the fold-switching set and in a non-folding set are compared in Table II. To get a sense of the precision of the results, the standard deviation of 100-sequence averages from the 4735 sequences were also calculated. While there is a slight increase of the frequency of chameleon sequences in the fold-switching set when compared with the frequencies of the large set from the PDB, the results clearly show that, contrary to what one would expect, the short chameleon sequences are not a significant factor in the evolution of fold switching proteins.

As a final test, the propensities of each amino acid to be in a chameleon sequence were compared to their propensities to be in a fold-switching sequence. The results of the comparison are given in Table III. Even worse than for the case of the chameleon sequence propensities, there is no correlation between the residue propensities of chameleon and fold-switching sequences. In fact, for several amino acids their propensity varies from the overall value in the opposite direction in a significant extent.

In conclusion, it is found that (a) the fold switching proteins have, on average, lower foldability scores than all the proteins in the PDB but not low enough to predict them to be not folding and (b) the short chameleon sequences that are quite abundant in the proteins in the PDB appear not to have a significant role in fold switching.

Acknowledgments

Prof. George Rose is thanked for helpful suggestions as well as for critical reading of an early version of the manuscript. This work was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Conflict of Interest

The author has no conflict of interest.

Table I: Foldability scores of the different sequence sets

	$\langle SC_3 \rangle$	S.D.	$\langle SC_4 \rangle$	S.D.
Fold-switching set	0.028	0.060	0.061	0.093
PDB set	0.047	0.062	0.094	0.097
IDP set	0.044	0.067	0.070	0.100
Propensity-weighted set	-0.043	0.032	-0.082	0.048
Uniform random set	-0.051	0.034	-0.109	0.052

Table II: Chameleon sequence propensities

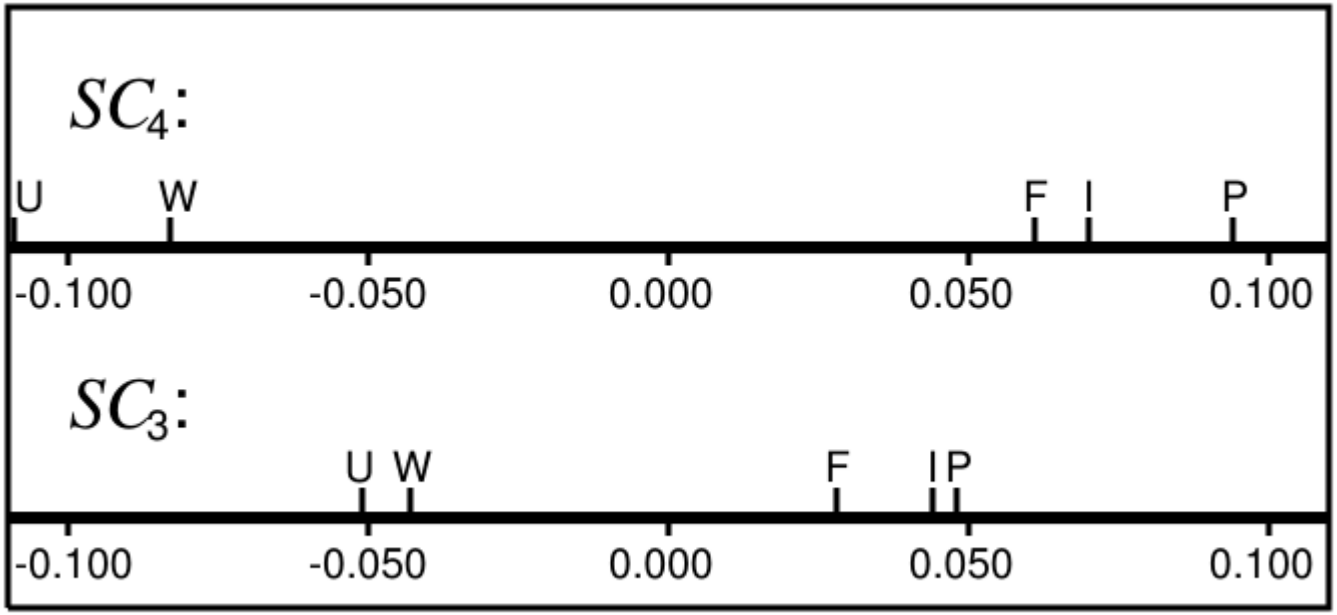
L_{ch}	N_{ch}	Fold switching %chameleons	Non fold switching %chameleons	S.D. of %chameleons of 100 sequence averages
5	187803	15.6	13.7	0.61
6	36669	1.2	0.8	0.41
7	1822	0.1	0.0	0.02

Table III. Amino-acid excess propensities

	%fold switching/%AA	%chameleon/%AA	%PDB/%AA	%AA
GLY	1.07	0.76	1.13	6.40
ALA	1.94	1.28	1.14	7.05
VAL	1.06	1.80	1.06	6.45
LEU	0.87	1.27	0.93	9.97
ILE	0.91	1.51	0.89	6.32
SER	0.81	0.70	0.80	8.06
THR	1.09	1.13	0.98	5.48
ASP	1.21	0.71	1.22	4.72
GLU	1.24	1.08	1.21	5.62
ASN	0.95	0.54	0.83	5.16
GLN	1.20	0.99	1.06	3.56
LYS	1.18	0.90	0.99	5.90
HIS	1.05	0.84	1.32	2.20
ARG	0.96	1.15	1.09	4.71
PHE	0.84	1.17	0.83	4.73
TYR	1.00	0.33	0.97	3.44
TRP	1.00	0.86	0.97	1.33
CYS	0.78	0.58	0.69	1.88
MET	0.82	0.98	0.99	2.36
PRO	0.93	0.22	0.99	4.67

Figure caption

Foldability mater showing the place of the average foldability scores SC_3 and SC_4 on the foldability scale.
U:randomly generated with uniform distribution; W:randomly generated with amino-acid propensity distribution;
F: fold-switching set; I: Intrinsically Disordered (IDP) set; P: PDB test set.



References

1. Porter LL, Loogera LL. Extant fold-switching proteins are widespread. Proceedings of the National Academy of Sciences of the United States of America 2018;115:5968–5973.
2. M.Mezei. On predicting foldability of a protein from its sequence Proteins 2020; 88:355-356.
3. Mezei M. Revisiting chameleon sequences in the Protein Data Bank. Algorithms 2018;11.
4. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, Bourne PE. The Protein Data Bank. Nucleic Acids Research 2000;28:235-242.
5. Gaur RK. Amino acid frequency distribution among eukaryotic proteins. The IIOAB Journal 2014;5:6-11.
6. Damiano Piovesan, Francesco Tabaro, Ivan Mičetić Marco, Necci Federica Quaglia, Christopher J. Oldfield, Maria Cristina Aspromonte, Norman E. Davey, Radoslav Davidović, Zsuzsanna Dosztányi, Arne Elofsson, Alessandra Gasparini, András Hatos, Andrey V. Kajava, Lajos Kalmar, Emanuela Leonardi, Tamas Lazar, Sandra Macedo-Ribeiro, Mauricio Macossay-Castillo, Attila Meszaros, Giovanni Minervini, Nikoletta Murvai, Jordi Pujols, Daniel B. Roche, Edoardo Salladini, Eva Schad, Antoine Schramm, Beata Szabo, Agnes Tantos, Fiorella Tonello, Konstantinos D. Tsirigos, Nevena Veljković, Salvador Ventura, Wim Vranken, Per Warholm, Vladimir N. Uversky, A. Keith Dunker, Sonia Longhi, Peter Silvio, Tosatto CE. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Research 2016;45:D219–D227.