Protein Engineering vol.11 no.6 pp.411–414, 1998

SHORT COMMUNICATION

Chameleon sequences in the PDB

https://en.wikipedia.org/wiki/K-mer

Mihaly Mezei

Department of Physiology and Biophysics, Mount Sinai School of Medicine, CUNY, New York, NY 10029, USA

The Brookhaven Protein Data Bank has been searched for sequences that can be found both in helix and sheet conformation. The longest such sequences consist of seven residues.

Introduction

There has been recent interest in exploring the possibilities of a given amino acid sequence assuming different secondary structures in different contexts. An 11-residue segment of a synthetic protein was successfully designed that formed an α helix in one context and a β sheet in another, prompting the authors to dub it a 'chameleon' sequence (Minor and Kim, 1996). On a larger scale, a 56-residue protein domain was successfully converted to a different fold by changing no more than half of the residues (Dalal *et al*., 1997). The fact that this achievement was prompted by a challenge supplemented by a monetary prize by some of the leaders in the protein folding field (Rose and Creamer, 1994) indicates the strength of the belief that the secondary structure of a protein is essentially determined locally by the primary structure.

Theory and computations

Given the level of challenge involved in the studies discussed above, it is natural to turn to the Brookhaven Protein Data Bank (PDB; Bernstein *et al*., 1997) to search for naturally occurring chameleon sequences. In fact, this has already been done earlier. The first searches (Kabsch and Sander, 1984; Argos, 1987) found five-residue chameleons in the PDB and six-residue chameleons elsewhere, while a subsequent work (Cohen *et al*., 1993) reported six-residue chameleons in the PDB. The rapid growth in the number of entries in the PDB suggests, however, that revisiting such a search might be a fruitful endeavor.

The present search included all files included in the distribution of April, 1997. It relied on the PDB annotation of helices and sheets. Also, unlike Cohen *et al*. (1993), a sequence was required to have a uniform SHEET conformation in one structure and a uniform HELIX in the other in order to qualify. This explains why none of the six-residue chameleons found by Cohen *et al*. (1993) made the present list.

The use of the PDB annotation necessitated a screening process since several instances were found where a segment labeled HELIX overlapped or even was completely included in a segment labeled SHEET (or vice versa). The screening yielded two lists labeled HELIX and SHEET, containing sequences of varying length in helix or sheet conformations, respectively. The two lists were checked against each other for the occurrence of chameleons of various length. The longest chameleon sequences were also checked visually by molecular graphics, allowing the screening out of any that was mistakenly labeled as HELIX in the PDB. Since all entries were involved in the search, a large number of duplicate chameleons were obtained, either coming from repeated subunits or from structures that are minor variants of each other. Therefore, all duplicates were reduced to a single representative.

Results and discussion

The longest chameleon sequences found are seven residues long. There are three of them. The sequences are as follows: LSLAVAG corresponding to residues 455–461 and 67–73 in the structures of yeast topoisomerase II (1bgw) and methylamine dehydrogenase (1mda), respectively; LITTAHA corresponding to residues 121–127 and 835–841 in the structures of a cyclodextrin glycosyltransferase (1cgu) and beta-galactosidase (1bgl), respectively; and KGLEWVS corresponding to residues 192–198 and 43–50 in the structures of triacylglycerol hydrolase (1thg) and an immunoglobulin fragment (1igm), respectively. Also, there are 38 additional 6-residue long chameleon sequences and 940 five-residue chameleon sequences. Table I lists the six and seven-residue long chameleon sequences and their locations and Figures 1–3 display the protein backbones containing the seven-residue chameleons, highlighted with stick representation.

The hydrogen-bonded and hydrophobic partners of the seven-residue chameleon sequences were examined by the programs Hbplus (McDonald and Thornton, 1994) and Ligplot (Wallace *et al*., 1995), hoping for the emergence of some further insight. However, no clear pattern emerged: they were equally likely to form hydrogen bonds from the backbone and from the side chains in either conformation and no regularity in the partners was observed. The position within the protein was also varied: some of the seven-residue chameleon sequences (given in Table I) are buried, some are on the surface of the protein and there is no correlation between the secondary structure and the position. The distribution of the amino acids in the chameleon sequences was also examined and presented in Table II, showing a prevalence of alanines, leucines and valines.

The seven-residue chameleons represent two full turns of helix and over three repeating units of sheets. This indicates that given the right surroundings, these chameleons can repeat themselves to greater length. This suggests that there should be no intrinsic limitation against chameleons of even greater length. However, the fact that no chameleon longer than seven residues can be found may be considered to contradict this conclusion. To study this issue further, an estimate was obtained for the probability of no two *k*-mer pairs being identical, if one is chosen from the HELIX list and the other from the SHEET list. Assuming that the HELIX and SHEET lists contains n_i^H and n_i^S segments of length *i*, respectively, there are N_k^H and N_k^S possible *k*-mers, where

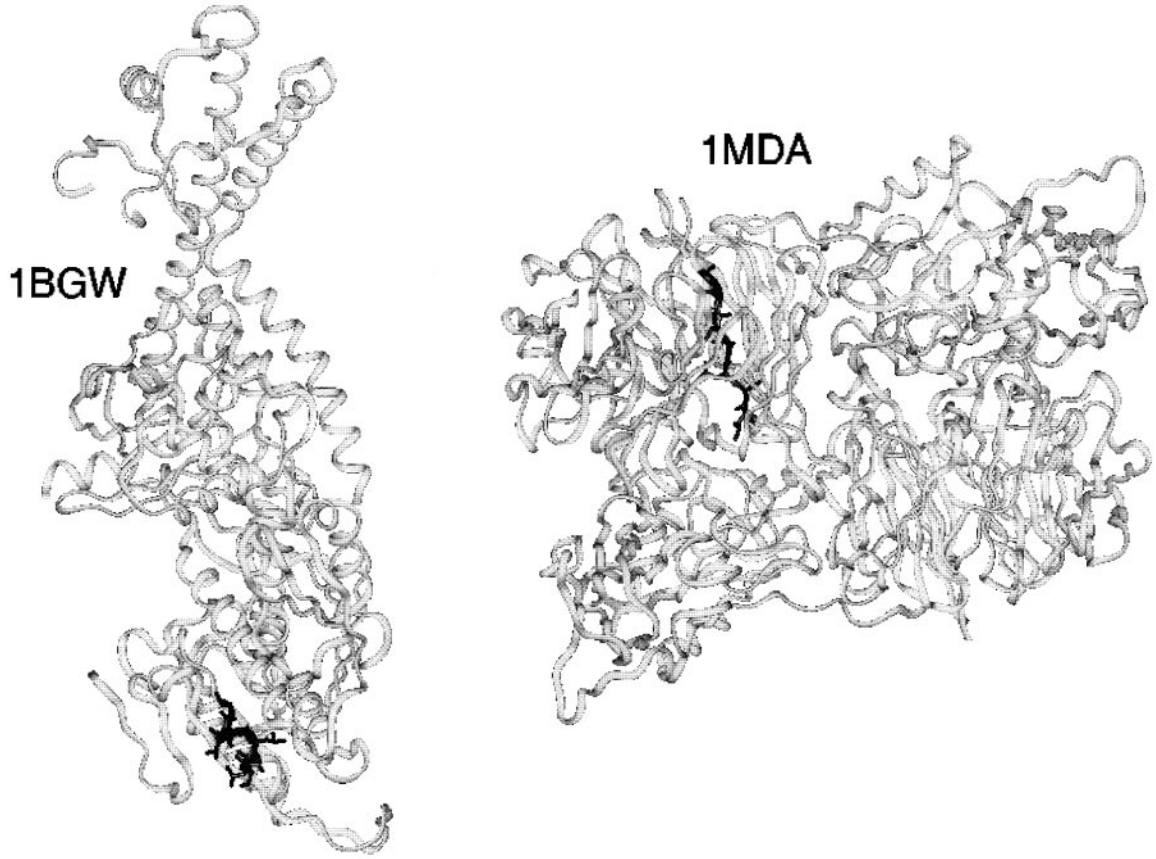

Fig. 1. Rendering of the protein backbones 1mda (**Right**) and 1bgw (**Left**) containing the sequence LSLAVAG (rendered black) in sheet and helix conformations, respectively. The chameleon residues are also shown in stick representation.

$$N_k^H = \sum_{i=k}^{\infty} n_i^H\,(i-k+1) \quad \text{and} \quad N_k^H = \sum_{i=k}^{\infty} n_i^H\,(i-k+1)$$

Assuming further that the various amino acids occur with equal probability in either list, it can be shown that to a good approximation the probability of not finding two identical k-mers at all in the two lists, P_{norep} is

$$P_{norep} = \exp(-N_k^H\, N_k^S / 20^k)$$

Similar expressions were also derived earlier (Wilson *et al.*, 1985). Table III gives the values of P_{norep}, N_k^H and N_k^S showing that the probability of repeats rapidly decreases as k is increased. Replacing the crude assumption of uniform occurrence of amino acids by a more realistic one would only shift the threshold value of k where P_{norep} becomes very close to one. Thus it can be concluded that not finding longer than seven-residue chameleons is not an indicator of an intrinsic limitation against longer ones.

The existence of these chameleon sequences of nontrivial length is also in accord with the recent demonstration of the statistical significance of multi-body terms in determining protein structure (Munson and Singh, 1997). Furthermore, the prevalence of leucines, valines and alanines found among these chameleons is nicely in accord with the fact that the residues in the most attractive corner of their visualized four-body potential are valine, isoleucine, leucine and alanine (indicating that any pair of these residues has the most chance to produce favorable interactions under a variety of circumstances). The non-uniform distribution of chameleon residues also appears to confirm the earlier suggestion (Kabsch and Sander, 1984) that the structural adaptability should vary from sequence to sequence.

References

Argos,P. (1987) *J. Mol. Biol.*, **197**, 331–348.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F. Jr., Brice,M.D., Rogers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M., (1977) *J. Mol. Biol.*, **112**, 535–542.
Cohen,B.I., Presnell, S.R. and Cohen,F.E. (1993) *Protein Sci.*, **2**, 2134–2145.
Dalal,S., Balasubramanian,S. and Regan,L. (1997) *Nature Struct. Biol.*, **4**, 548–552.
Kabsch,W. and Sander,C., (1984) *Proc. Natl Acad. Sci. USA*, **81**, 1075–1078.
McDonald,I.K. and Thornton,J.M. (1994) *J. Mol. Biol.*, **238**, 777–793.
Minor,D.L.,Jr. and Kim,P.S. (1996) *Nature*, **380**, 730–734.
Munson,P.J. and Singh,R.K. (1997) *Protein Sci.*, **6**, 1467–1481.
Rose,G.D. and Creamer,T.P. (1994) *Proteins Struct. Funct. Genet.*, **19**, 1–3.
Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1995) *Protein Engng.*, **8**, 127–1340.
Wilson,I.A., Haft,H.H., Getzoff,E.D., Tainer,J.A., Lerner, R.A. and Brenner,S. (1985) *Proc. Natl Acad. Sci. USA*, **82**, 5255–5259.

Received August 11, 1997; revised January 13, 1998; accepted January 23, 1998

Note added in proof

A recent study, using a general dissimilarity criterion, found 8-residue long dissimilar sequences [Sudarsanum, S. (1998) *Proteins Struct. Funct. Genet.*, **30**, 228–231].

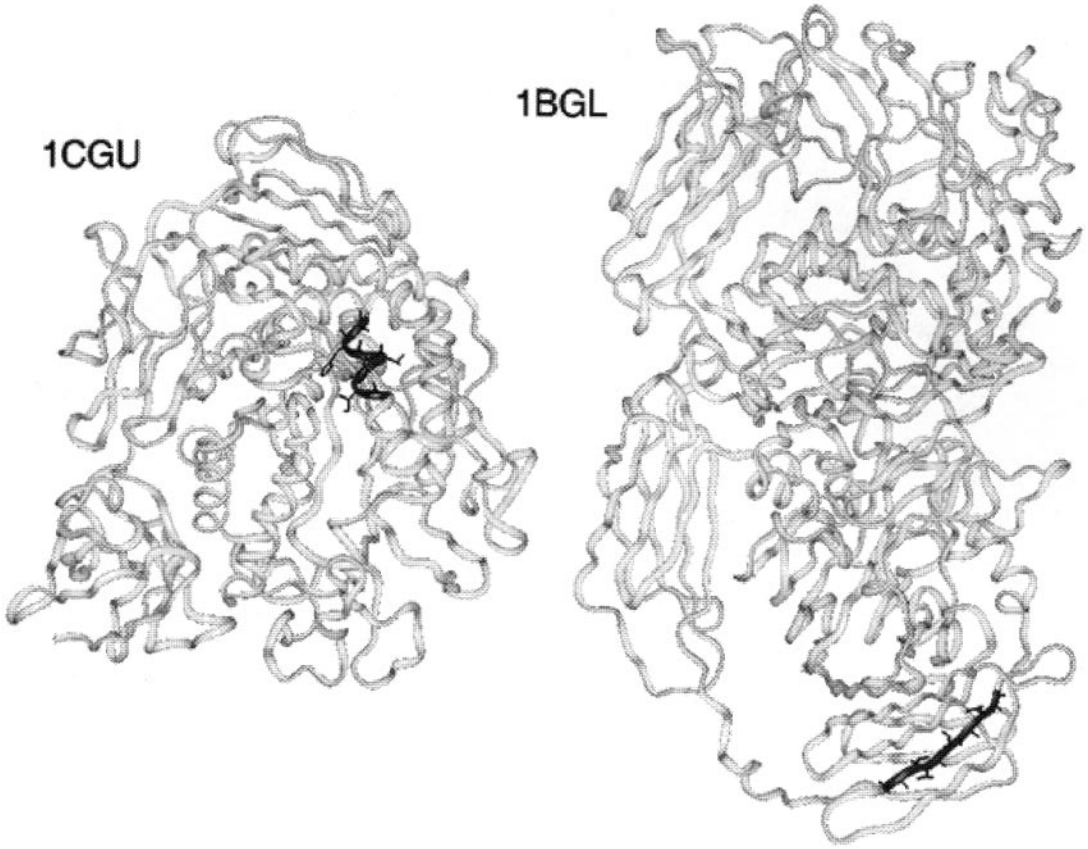

Fig. 2. Rendering of the protein backbones 1bgl (**Right**) and 1cgu (**Left**) containing the sequence LITTAHA (rendered black) in sheet and helix conformations, respectively. The chameleon residues are also shown in stick representation.

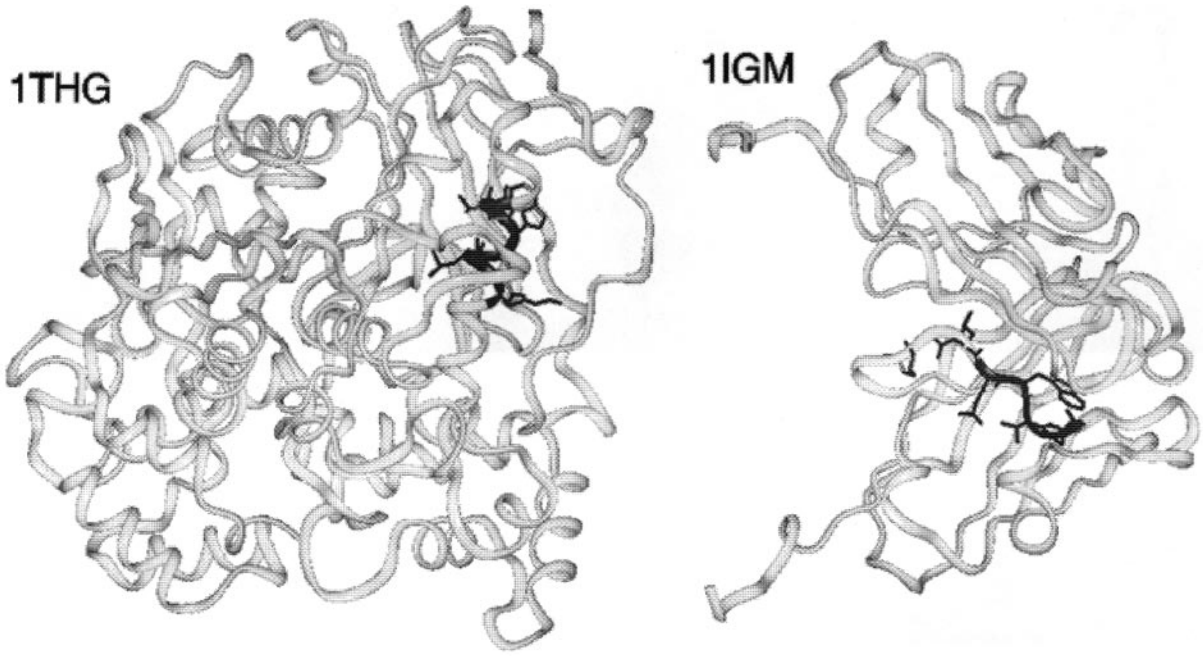

Fig. 3. Rendering of the protein backbones 1igm (**Right**) and 1thg (**Left**) containing the sequence KGLEWVS (rendered black) in sheet and helix conformations, respectively. The chameleon residues are also shown in stick representation.

Table I. List of chameleon sequences of length six and seven

Helix PDBid	Start residue	Sheet PDBid	Start residue	Length	Helix type	Sequence
1aam	287	1mda	245 H	6	α	MKAAID
1aat	331	1gff	236 1	6	α	LLMRSE
1ala	252	1cgp	36 A	6	α	AETLYY
1ann	150	1ace	142	6	α	VLVSLS
1atf	31	1hle	347 A	6	α	AAAATA
1bac	91	1ids	155 A	6	α	PLLLLD
1bgc	150	1dst	43	6	α	GGVLVA
1dgd	314	1rhd	27	6	α	GLRVLD
1dnp	17 A	1ktq	322	6	α	LALAAA
1ebg	109 A	1eaa	578	6	α	AILGVS
1fct	8	1gba	86 A	6	α	TFAARV
1gdd	122	1rds	84	6	α	ELAGVI
1gnc	51	1mda	59 H	6	α	LGHSLG
1han	16	1bms	79 A	6	α	VAAWRS
1hbi	136 A	1eip	103 A	6	α	AKLVAV
1hlb	36	1bll	198 E	6	α	TDVFIR
1hrs	98	1doi	72	6	α	AAIVLE
2hwd	68 1	2rhn	68 1	6	α	ESFLGR
1hyt	6	1lna	6 E	6	α	TVGVGR
1ign	547 A	1cov	129 3	6	α	KFLLAY
1kau	6 C	1qbe	86 A	6	α	RQAYAD
1kif	330 A	1cyn	137 A	6	3_{10}	FGKVLE
1kny	160 B	1mrj	178	6	3_{10}	TFLPSL
1lea	14	1pbn	131	6	α	LIRDHI
1mhc	64 A	1wit	74	6	α	KLKVKN
1ola	375 A	1thm	178	6	α	AIAVAS
1out	59 B	1amy	389	6	α	KVAAHG
1out	113 B	1ebd	375 A	6	3_{10}	VIAAKF
1ouu	68 C	1gfm	15	6	α	GKAVGL
1ron	26	1rne	180	6	α	HYINLI
1rpa	228	2kai	43 A	6	α	GGVLVN
1spb	79 S	1sbn	88 E	6	α	ASLYAV
1tml	203	1fug	90 A	6	α	AVLSAI
2tmv	130 P	1ing	424 A	6	α	VELIRG
1vsg	70 A	1atp	67 E	6	α	NHYAMK
1ygp	586 A	1bfg	72	6	α	RYLAMK
1bgw	455	1mda	67 H	7	α	LSLAVAG
1cgu	121	1bgl	835 A	7	α	LITTAHA
1thg	192	1igm	43 H	7	α	KGLEWVS

Table II. Frequency of occurrence of amino acids in chameleon sequences

Residue	N_5	P_5^{excess}	N_6	P_6^{excess}	N_7	P_7^{excess}
ALA	686	2.92	46	3.74	4	2.85
CYS	27	0.11	0	0.00	0	0.00
ASP	123	0.52	6	0.49	0	0.00
GLU	272	1.16	9	0.73	1	0.71
PHE	161	0.69	7	0.57	0	0.00
GLY	223	0.95	21	1.71	4	2.86
HIS	53	0.23	6	0.49	1	0.71
ILE	373	1.59	14	1.14	1	0.71
LYS	263	1.12	13	1.06	1	0.71
LEU	714	3.03	39	3.17	4	2.86
MET	65	0.28	4	0.33	0	0.00
ASN	96	0.41	4	0.33	0	0.00
PRO	31	0.13	2	0.16	0	0.00
GLN	118	0.50	1	0.08	0	0.00
ARG	185	0.79	11	0.89	1	0.71
SER	252	1.07	13	1.06	3	2.14
THR	249	1.06	8	0.65	3	2.14
VAL	650	2.77	31	2.52	4	2.86
TRP	28	0.12	3	0.24	1	0.71
TYR	131	0.56	8	0.65	0	0.00

N_k: number of times the residue occurred in a *k*-mer chameleon; P_k^{excess}: $N_k \times 20/(k \times n_k)$ where n_k is the number of *k*-mers.

Table III. Estimate of the probability of no chameleons

k	N_k^H	N_k^H	P_{norep}
5	294585	117643	0.31
6	258009	81664	0.979
7	224142	54026	0.99988

k, length of sequence; N_k^H, number of helix *k*-mers; N_k^S, number of sheet *k*-mers; P_{norep}, probability of no *k*-mers in the helix list is identical with a *k*-mer in the sheet list, assuming uniform residue distributions.