International Journal of Biological Macromolecules 49 (2011) 218–222

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac

Certain heptapeptide and large sequences representing an entire helix, strand or coil conformation in proteins are associated as chameleon sequences

Neethu Krishna, Kunchur Guruprasad*

Bioinformatics, Centre for Cellular and Molecular Biology (CCMB), Uppal Road, Hyderabad 500 007, India

ARTICLE INFO

Article history:
Received 15 March 2011
Accepted 25 April 2011
Available online 5 May 2011

Keywords:
Chameleon sequences
Heptapeptides
Protein Data Bank
Protein sequence-structure analysis

ABSTRACT

Helices, strands and coils in proteins of known three-dimensional structure, corresponding to heptapeptide and large sequences ('probe' peptides), were scanned against peptide sequences of variable length, comprising seven or more residues that correspond to a different conformation ('target' peptides) in protein crystal structures available from the Protein Data Bank (PDB). Where the 'probe' and 'target' peptide sequences exactly match, they correspond to 'chameleon' sequences in protein structures. We observed ~548 heptapeptide and large chameleon sequences that included peptides in the coil conformation from 53,794 PDB files that were analyzed. However, after excluding several chameleon peptides based on the quality of protein structure data, redundancy and peptides associated with cloning artifacts, such as, histidine-tags, we observed only ten chameleon peptides in structurally different proteins and the maximum length comprised seven amino acid residues. Our analysis suggests that the quality of protein structure data is important for identifying possibly, the 'true chameleons' in PDB. Majority of the chameleon sequences correspond to an entire strand in one protein that is observed as part of helix sequence in another protein. The heptapeptide chameleons are characterized with a high propensity of alanine, leucine and valine amino acid residues. The total hydropathy values range between −11.2 and 22.9, the difference in solvent accessibility between 2.0 Å^2 and 373 Å^2 units and the difference in total number of residue neighbor contacts between 0 and 7 residues. Our work identifies for the first time heptapeptide and large sequences that correspond to a single complete helix, strand or coil, which adopt entirely different secondary structures in another protein.

1. Introduction

The analysis of features relating amino acid sequences to structures continues to be a problem of interest to the structural biologists and bioinformaticians. In this context, we were interested in analyzing heptapeptide and large sequences characterized by identical amino acid sequence but with a different conformation in proteins of known three-dimensional structure. Identical pentapeptide sequences [1] and hexapeptide sequences [2,3] are known to adopt different conformations in protein tertiary structures. A synthetic protein comprising 11-residue segment that formed alpha-helix in one context and a beta-sheet in another stabilized by non-local interactions was successfully designed and dubbed as 'chameleon' sequence [4]. Later, a challenge by leaders in protein folding [5] prompted [6] to convert a much larger, i.e., 56-residue protein domain to a different fold by changing no more than half the number of residues. Subsequently, three instances of identical heptapeptide sequences (then known to be the longest) in naturally occurring proteins were observed in both helix and sheet conformation along with thirty-six newly identified hexapeptide sequences [7]. In later studies [8], four identical octapeptide sequences and eight new heptapeptide chameleon sequences were identified. In a recent study comprising 6962 proteins, 2 octapeptides and 52 heptapeptides of helix in strand (HS) conformation and 7 heptapeptides and 37 hexapeptides of helix in sheet (HE) conformation were reported [9]. Another study on 'redundant' sets of experimental models of protein structures, showed the presence of "dual-personality" (DP) fragments that differentiate between regular fold and intrinsically disordered fragments [10].

Different properties have been linked to chameleon sequences. For instance, chameleon sequences have been suggested to play a role in the structural fold conservation and functional diversity of alternative splicing protein isoforms [9]. Chameleon sequences have been implicated in the context of theories on immune recognition [11], or in the induction of amyloid-related fatal diseases [12]. Chameleon sequences are known to inherently possess the property of "conformational contagion", i.e., to take on alpha-helix or beta-sheet conformation depending on the

* Corresponding author. Tel.: +91 40 2719 2779; fax: +91 40 2716 0591x0311.
E-mail addresses: kunchur.guruprasad@gmail.com, guru@ccmb.res.in (K. Guruprasad).

0141-8130/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijbiomac.2011.04.017

sequentially neighboring secondary structure if little other non-local interaction occurs [13]. In the case of 'redundant' proteins, the "dual-personality" fragments are often targets of regulation [10].

In the present study, we were interested in identifying whether there were heptapeptide and large sequences in the PDB that 'entirely' correspond to a helix, strand or coil in one protein that adopt a different secondary structure in another protein. This would represent a different type of chameleon peptide and contrast with, previous studies, which have reported heptapeptide and large chameleon sequences, where even the 'probe' sequence corresponded to subsequence of larger helix sequence [9]. The present analysis included the search for chameleon sequences that represent the coil conformation as 'probe'. Further, we have used high resolution, well refined protein crystal structure data, including the region corresponding to the chameleon peptide, thus ensuring the quality of data used for identifying these specific types of chameleon peptides from the PDB.

2. Materials and methods

The list of PDB [14] codes selected for our analyses were obtained from ftp://ftp.wwpdb.org/pub/pdb/derived_data/index/entries.idx. Only high resolution crystal structures, i.e., refined at ≤2.5 Å resolution was used for the analyses. The amino acid sequences corresponding to helices and strands in these proteins were defined according to the PROMOTIF program [15] and were readily available from PDBsum [16] http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/. These were retrieved using a Python script. The PROMOTIF program identifies helices and strands using a slightly modified method described in the DSSP program [17], in order to include one extra residue at the end of each strand and helix where possible (according to the IUPAC convention rule 6.3 that is most commonly used among crystallographers). The heptapeptide and large sequences corresponding to the coils in protein structure were later extracted. Three files were prepared corresponding to heptapeptide and large sequences; one for helices, the second for strands and the third for coils. From these three files, another equivalent set of files were prepared that corresponded only to the representative (or non-redundant) helix, strand and coil sequences. These were selected based on identical peptide sequence length and amino acid sequence. These representative sequences were used for identifying chameleon sequences in the PDB.

A PERL program was developed, – in order to search the above helix sequences within the strand and coil sequences, by sliding the helix (or 'probe') sequence along the strand (or 'target') sequence one residue at a time. Likewise, sequences that correspond to entire strands were searched against the helix and coil sequences and the sequences corresponding to entire coils were searched against the helix and strand sequences. Thereby, in each case the 'probe' peptide was of a fixed sequence length but the 'target' peptides in the dataset analyzed were of variable sequence length. Where the 'probe' sequence exactly matched the 'target' sequence, it represented 'chameleon' sequences in the corresponding proteins identified by their PDB codes. Redundant chameleon peptides were excluded by examining the peptide sequence length, amino acid sequence, secondary structure conformation and protein superfamily according to the Structural Classification of Proteins (SCOP) [18].

The final list of heptapeptide and large 'chameleons' in the PDB were selected by examining the atomic temperature factor, or *B*-factor for every atom in the chameleon peptide. The *B*-factor is a measure of the dynamic disorder caused by the temperature-dependent vibration of the atom, as well as the static disorder resulting from subtle structural differences in different unit cells throughout the crystal; a *B*-factor of less than 30 Å^2 for a particular atom usually indicates confidence in its atomic position, but a *B*-factor of higher than 60 Å^2 likely indicates that the atom is disordered. We selected only chameleon peptides where ~99% of the chameleon peptide atoms had *B*-values <40 in both the proteins representing the chameleon peptide. Further, we excluded chameleon peptide sequences associated as histidine-tags and synthetic constructs. Finally, for each chameleon peptide, we noted the peptide sequence length, peptide sequence, PDB code/chain, SCOP name (or the protein name) and location of the chameleon sequence along the protein chain. The schematic representations of the proteins comprising the chameleon peptides; 'probe' peptide (green) and 'target' peptide (red) were drawn using PyMol [19].

Table 1
Dataset.

Item	Total
PDB files analyzed	53,794
Helices	600,421
Strands	619,173
Coils	1,256,996
Non-redundant helices	181,975
Non-redundant strands	137,953
Non-redundant coils	77,520
Heptapeptide and large helices	330,115
Heptapeptide and large strands	167,534
Heptapeptide and large coils	40,497
Non-redundant heptapeptide and large helices	132,663
Non-redundant heptapeptide and large strands	56,713
Non-redundant heptapeptide and large coils	12,916
Heptapeptide and large chameleon sequences (in PDB crystal structures with resolution ≤2.5 Å)	80

The total hydropathy scores corresponding to the chameleon peptide sequences were evaluated using the values described in [20], and the solvent accessibilities and residue neighborhood contacts were determined using the programs AREAIMOL and NCONT, respectively, available in the CCP4 suite of programs [21]. We compared these values with corresponding values computed for heptapeptide and large chameleon sequences reported in [9], that were selected based on the crystal structure resolution and *B*-factors as used in the present work. However, for calculating the amino acid propensity values, we combined the heptapeptide chameleon sequences identified in the present work along with those selected from chameleon peptides previously reported [9]. The propensity values were calculated by taking the ratio of the frequency distribution of amino acid residues in chameleon sequences versus frequency distribution of amino acid residues in chameleon peptide containing proteins.

3. Results and discussion

The dataset used for analyzing heptapeptide and large chameleon sequences in the PDB (as described in Section 2) is shown in Table 1. The heptapeptide and large chameleon sequences reported in the present work, represents a new type of chameleon peptide that has not been hitherto reported in the literature. We observed ~548 heptapeptide and large chameleon sequences from our analysis of the type where an entire helix, strand or coil conformation, in one protein is observed with a different conformation in another protein. However, by selecting non-redundant chameleon peptides from protein crystal structure data with resolution ≤2.5 Å and *B*-factor for chameleon peptide atoms ≤40.0, we observed only ten chameleon peptides in different proteins as shown in Fig. 1(A) (supplementary data). Our analysis suggests that the largest chameleon peptide observed in the PDB in this manner is restricted to a maximum of seven amino acid residues. The chameleon peptide type, PDB code/chain, location in the protein chain, chameleon peptide sequence, peptide length, protein superfamily (or protein name) and the crystal structure resolution are indicated. These proteins cannot be structurally superimposed and often belong to different species. Majority of these chameleon peptides represent an entire strand in one protein that corresponds to subsequence of a larger helix sequence in another protein. Two examples of heptapeptide chameleons associated with the coil conformation were also observed. For instance, the heptapeptide chameleon sequence; TPIVTLY in coil conformation in the lyase protein (PDB_code: 2ZSJ/A) corresponds to a strand in the hydrolase protein (PDB_code: 2E9L/A).

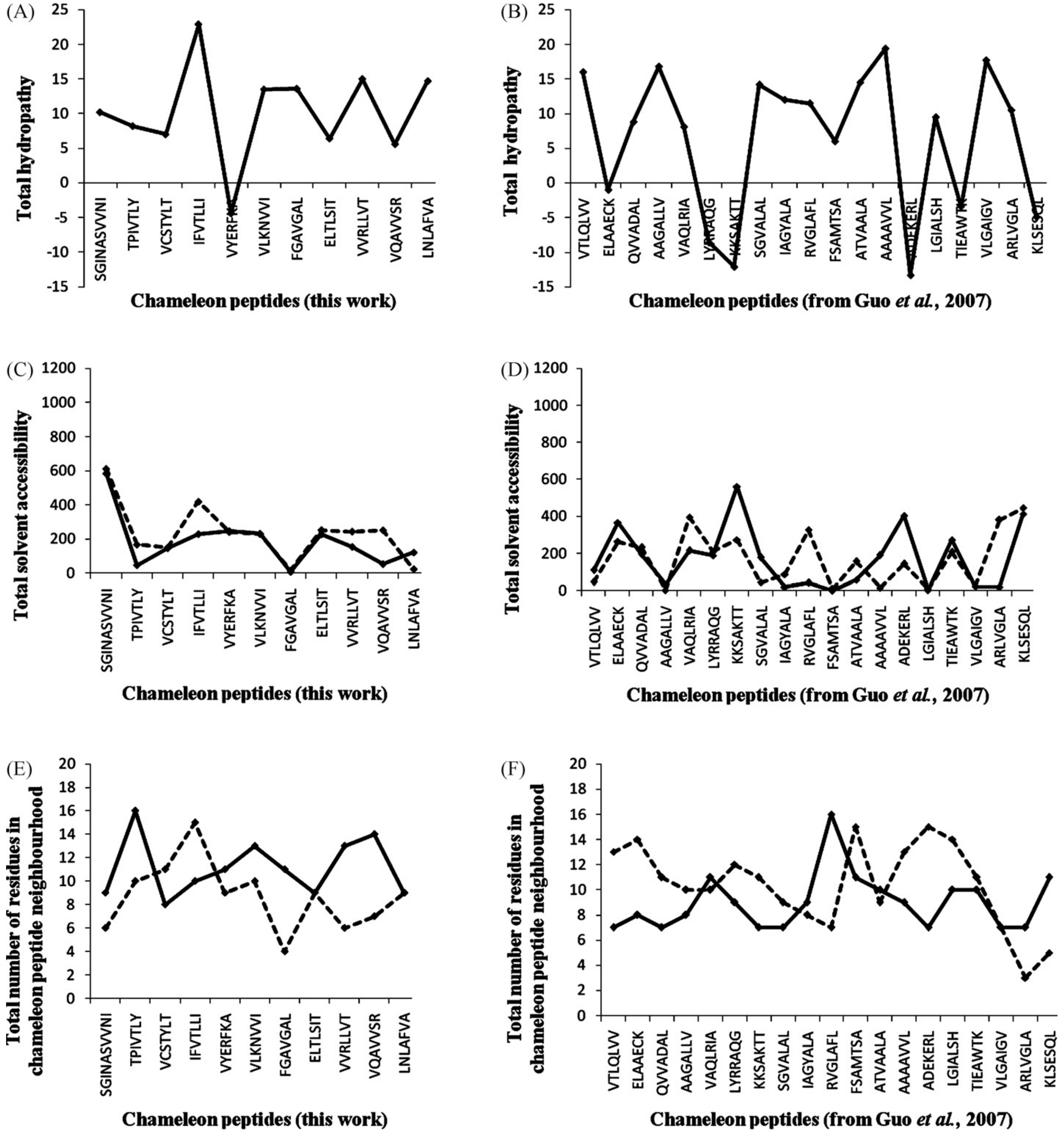

Fig. 1. Total (A) hydropathy, (C) solvent accessibility, (E) residue neighbor contacts for heptapeptide and large chameleon sequences identified (in this work) and equivalent values shown in figures (B), (D), (F), respectively, evaluated for heptapeptide and large chameleon sequences selected from high quality protein crystal structures selected from [9]. The values shown in (C), (D), (E) and (F) along the *Y*-axis are for the chameleon peptide sequence in the corresponding protein pairs (represented by dash and continuous line). (G) Amino acid propensity values for combined heptapeptide and large chameleon sequences (this work) and chameleon-HS and chameleon-HE sequences selected from [9] as shown along the *X*-axis in (B).

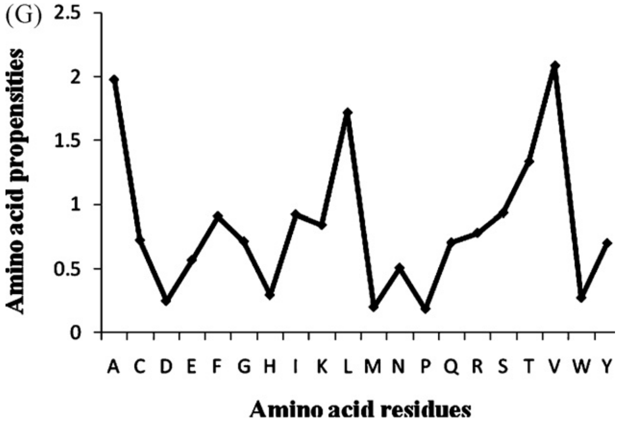

Fig. 1. (Continued)

In another earlier study [10], a systematic analysis on 'redundant' proteins identified "dual-personality" fragments in proteins that are implicated in the regulation of protein function. However, in the present study, the heptapeptide chameleons identified were associated with a 'non-redundant' dataset of proteins in the PDB. While it is known that the conformation of chameleon peptides in structurally different proteins depends upon their context in protein three-dimensional structure, the heptapeptide chameleon sequences that have been identified, possibly reflect an intrinsic feature of such adaptability. The size of such examples is likely to grow as more structures become available in the PDB.

We observed one chameleon peptide comprising ten amino acid residues that is associated with structurally similar proteins as shown in Fig. 1(B) (supplementary data). The chameleon decapeptide; SGINASVVNI, in the L-chain (PDB code: 1ZV8) is in coil conformation and in the A-chain (PDB code: 1ZV7), the peptide represents subsequence of helix sequence. Entry of the SARS coronavirus into its target cell requires large-scale structural transitions in the viral spike (S) glycoprotein in order to induce fusion of the virus and cell membranes [22]. This suggests that chameleon peptides and their associated conformations observed in structurally similar proteins may be important for the corresponding protein function. The different protein environments, such as interaction with solvent or ligand or both contribute to identical peptides with different conformations in structurally similar proteins.

The values determined for total hydropathy for chameleon peptide sequences (Fig. 1A), the difference in solvent accessibility in the corresponding protein pairs (Fig. 1C), and the difference in total number of residue neighbor contacts (Fig. 1E) are consistent with equivalent values evaluated for the heptapeptide and large chameleon sequences selected from high quality protein crystal structures reported previously [9] (Fig. 1B, D and F, respectively). The values for chameleon peptides identified in the present work are in the range: −11.2 to 22.9 for total hydropathy, 2.0 to 373.8 for difference in solvent accessibility and 0 to 7 for the difference in total number of residue neighbor contacts. The equivalent values evaluated for chameleon heptapeptides selected from those reported in [9] are in the range: −13.4 to 19.4 for total hydropathy, 2.8 to 362 for difference in solvent accessibility values and 0 to 9 for difference in total number of residue neighbor contacts. The aliphatic amino acid residues; alanine, leucine and valine are associated with a high propensity, whereas, aspartic acid, histidine, methionine, asparagine, proline and tryptophan are associated with low propensity values as shown in Fig. 1G.

A wild chameleon sequence fused to the C-terminal alpha-helix or beta-sheet in foreign stable proteins from hyperthermophiles has been shown to form the same conformation as the host secondary structure. This "conformation contagion" property of chameleon sequence has been proposed as a new nonlocal determinant factor in protein structure and misfolding related to protein conformational diseases [13]. From Fig. 1(A), we see that certain sequences in the helix conformation ('VCSTYLT' in PDB code: 1B2R/A) constitute the latter half of the strand conformation as in (PDB code: 2HJR/A-chain). Likewise, a strand sequence ('VQAVVSR' in PDB code: 2O2G/A) is observed as latter half of helix conformations as in (PDB code: 1QNO/A). These may be due to a chameleon sequence taking on a satellite state through contagion by the power of a secondary structure as proposed by the "conformational contagion" hypothesis [13]. In the same context, what would happen to a chameleon sequence that adopts a coil conformation in protein structure when fused at the end of a helix or beta-strand that has little interaction with rest of the protein remains to be understood.

In summary, proteins do contain heptapeptide sequences that are 'entirely' in helix, strand or coil conformation that are observed as subsequence in a different conformation in another protein. These types of chameleon peptides are different from those hitherto reported in the literature. We did not, however, identify situations where a whole strand sequence was observed as a whole helix sequence, or a whole helix sequence as whole coil, and so on, corresponding to the heptapeptide and large sequences in protein structures. Most importantly, our analysis suggests that there are far fewer heptapeptide and large chameleon sequences in the PDB, when strict quality criterion are applied for the selection of crystal structures that include the *B*-factor corresponding to the chameleon peptide atoms; these criterion may be important for possibly recognizing the "true chameleons" in the PDB. Such large heptapeptide sequences that occur as chameleons in known protein three-dimensional structures possibly reflect their inherent characteristics to such adaptability.

4. Conclusions

The Protein Data Bank contains certain heptapeptide and large sequences representing an entire helix, or strand or coil conformation as chameleon sequences. Although, ∼548 instances were observed in the PDB, by applying a strict criterion for selecting chameleon peptides from protein crystal structures defined at ≤2.5 Å resolution, *B*-factor ≤40.0 and by excluding redundant examples, ten chameleon peptides were observed from 53,794 PDB files analyzed. A majority of these chameleon sequences represent an entire strand in one protein observed as part of helix in another protein and the largest chameleon peptide comprised seven amino acid residues in structurally different proteins. The choice of good quality protein datasets in the PDB may be important for recognizing "true chameleons" in proteins, particularly, in light of their functions implicated in the literature. These peptides possibly reflect the inherent features of their sequences to structural adaptability in protein structures.

Acknowledgements

The corresponding author acknowledges the Council of Scientific and Industrial Research and the Department of Biotechnology, Government of India for support. We thank Prof. Sir Tom Blundell, University of Cambridge, U.K., for useful suggestions. We are grateful to the authors who contributed to protein three-dimensional structures deposited in the Protein Data Bank referred in this work.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ijbiomac.2011.04.017.

References

[1] W. Kabsch, C. Sander, Proc. Natl. Acad. Sci. U.S.A. 81 (1984) 1075–1078.
[2] P. Argos, J. Mol. Biol. 197 (1987) 331–348.
[3] B.I. Cohen, S.R. Presnell, F.E. Cohen, Protein Sci. 2 (1993) 2134–2145.
[4] D.L. Minor Jr., P.S. Kim, Nature 380 (1996) 730–734.
[5] G.D. Rose, T.P. Creamer, Proteins 19 (1994) 1–3.
[6] S. Dalal, S. Balasubramanian, L. Regan, Nat. Struct. Biol. 4 (1997) 548–552.
[7] M. Mezei, Protein Eng. 11 (1998) 411–414.
[8] S. Sudarsanam, Proteins 30 (1998) 228–231.
[9] J.T. Guo, J.W. Jaromczyk, Y. Xu, Proteins 67 (2007) 548–558.
[10] Y. Zhang, B. Stec, A. Godzik, Proteins 15 (2007) 1141–1147.
[11] I.A. Wilson, D.H. Haft, E.D. Getzoff, J.A. Tainer, R.A. Lerner, S. Brenner, Proc. Natl. Acad. Sci. U.S.A. 82 (1985) 5255–5259.
[12] H. Tidow, T. Lauber, K. Vitzithum, C.P. Sommerhoff, P. Rösch, U.C. Marx, Biochemistry 43 (2004) 11238–11247.
[13] K. Takano, Y. Katagiri, A. Mukaiyama, H. Chon, H. Matsumura, Y. Koga, S. Kanaya, Proteins 68 (2007) 617–625.
[14] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Res. 28 (2000) 235–242.
[15] E.G. Hutchinson, J.M. Thornton, Protein Sci. 5 (1996) 212–220.
[16] R.A. Laskowski, Nucleic Acids Res. 29 (2001) 221–222.
[17] W. Kabsch, C. Sander, Biopolymers 22 (1983) 2577–2637.
[18] A. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, J. Mol. Biol. 247 (1995) 536–540.
[19] W.L. DeLano, DeLano Scientific, San Carlos, CA, USA, 2002. http://www.pymol.org.
[20] J Kyte, R. Doolittle, J. Mol. Biol. 157 (1982) 105–132.
[21] Collaborative Computational Project, Number 4, Acta Crystallogr. D 50 (1994) 760–763.
[22] Y. Deng, J. Liu, Q. Zheng, W. Yong, M. Lu, Structure 14 (2006) 889–899.