Available online at www.sciencedirect.com

ScienceDirect

Journal of Molecular Graphics and Modelling 26 (2007) 652–655

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

# Conservation of closed loops

Boon K. Yew [a], Sree V. Chintapalli [a], Graham G.C. Upton [b], Christopher A. Reynolds [a,*]

[a] Department of Biological Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom

[b] Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom

Accepted 20 March 2007

Available online 24 March 2007

**Abstract**

The closed loop hypothesis of Berezovsky and Trifonov implicates the closure of loops of length 25–35 through hydrophobic interactions at the 'locks' as a key event in protein folding. The hypothesis is supported by published analyses of nine major superfolds. Here, we have generated multiple sequence alignments for the nine superfolds with PDB codes lthb, 1ilb, 256b, 2rhe, 1aps, 2stv, 4fxn (2fox), lubq and 7tim and have analysed the degree of conservation at the loop ends. Seventy percent of these loop ends are found to be well conserved and the peak in the distribution of distances between these well conserved regions lies at around 25 residues; both observations are consistent with the Berezovsky and Trifonov's hypothesis.

*Keywords:* Conservation; Entropy; Hydrophobicity; Closed loops; Protein folding; Leventhal paradox

## 1. Introduction

Berezovsky and Trifonov have presented a theory on protein folding, that if true, elegantly avoids a problem with the Levinthal paradox [1] since the basic folding unit is proposed to be a closed loop of about 25–30 amino acids, which is sufficiently small to fold within a reasonably short time [2]. The structure of these closed loops has been determined for proteins in nine major superfolds [3] and here we use a variety of measures for determining the degree of conservation at the important regions believed to be involved in loop closure.

Selected evidence that proteins consist of a sequential set of non-overlapping closed loops of 25–35 residues is briefly summarised below. Firstly, the autocorrelation function of hydrophobic residues shows a peak at 25–30 residues [4]. Moreover, hydrophobicity plots show a maximum at the closed loop ends and so the hydrophobic residues at the end of the closed loops are referred to as locks [4]. Arguments from polymer science (experimental and theoretical) imply that for mixed unstructured polypeptide chains, where the persistence length, *a*, is around 4–5 monomer units, the closed loop length will be around 3.5*a*, which is 10–25 amino acids or 20–50 after correction for the presence of rigid alpha helices [5,6]. The distribution of the length of loops in general, defined as having a $C_\alpha$–$C_\alpha$ distance of less than 10 Å, shows a maximum at 25–35 residues. An approximation to this distribution is shown in Fig. 1, which was generated from the data in reference [5]. This shows a definite effect that appears to arise from a mixture, with the principal distribution being geometric-like and the possibly more interesting but less frequent loops having lengths coming from a Gaussian-like distribution centered near 25. We propose that the geometric-like distribution may arise simply because proteins inevitably contain loops; we propose that the more interesting Gaussian-like distribution may not have a trivial origin and may therefore be functionally important. The Gaussian-like distribution accounts for about 5% of all loops and about 20% of all loops in the interesting range of 25–35 residues. (Thus, examples of loops contrary to the Berezovsky and Trifonov hypothesis do not necessarily disprove the hypothesis as they may belong to the ~80% in the geometric-like distribution.) The unit of 25–35 residues is supported by studies of the distribution of insertions and deletions and of recombinatorial swapping of protein sequence segments and of ancestral exons [6]. Studies on proteins show that the number of neighbours as a function of sequence distance reaches a maximum at about 27 amino acids [6]. Independent segments within a protein can be characterised by substantially higher internal versus external van der Waals interaction energies. A

[*] Corresponding author.
*E-mail address:* reync@essex.ac.uk (C.A. Reynolds).

1093-3263/$ – see front matter © 2007 Published by Elsevier Inc.
doi:10.1016/j.jmgm.2007.03.011

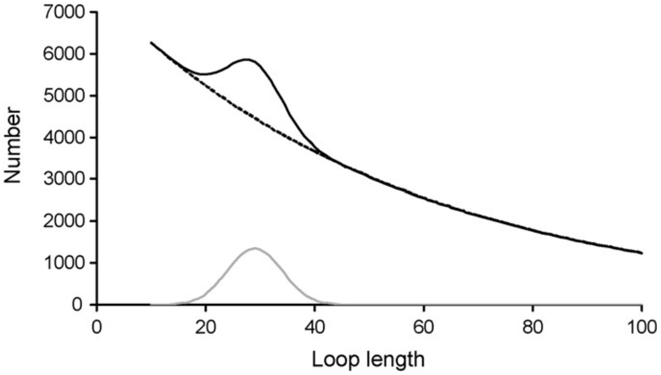

Fig. 1. The distribution of loop lengths, taken from reference [5]. The distribution (solid black line) has been approximated by combining an exponential decay (black dotted line) with a Gaussian distribution (grey solid line).

scheme was proposed that can distinguish a hierarchy of domains dependant upon the energy cut off: at the lowest level the 25–30 residue segments were identified, at the higher level traditional protein domain boundaries were identified. Berezovsky and Trifonov report that there is no relation between secondary structure and the loop ends reported, but rather that 25% of closed loop end residues reside in the middle of α-helix or β-sheet sections and so their loops do not correspond to regions of protein structure that have traditionally been assigned as coil (or loop). Further evidence for the closed loops comes from the autocorrelation of specific hydrophobic tripeptides (e.g. SGG, SAL) within 113 prokaryote genomes (327,129 protein sequences). A statistically significant number of peptide signatures (14,638) of length about 25 amino acids were identified, thus denoting many more potential closed loop peptides [7,8].

Here, we take a slightly different approach to conservation. We have taken multiple sequence alignments of the nine key superfolds and have analysed the locks using entropy and several related methods. The results are generally consistent with Berezovsky and Trifonov's hypothesis.

## 2. Methodology

The nine superfolds are represented by PDB codes 1thb, 1ilb, 256b, 2rhe, 1aps, 2stv, 4fxn (2fox), 1ubq and 7tim. The closed loops for these proteins were identified essentially as the set of non-overlapping loops of around ~25 amino acids that form the strongest interactions, as judged by the number and closeness of the interactions near the loop ends; these are published elsewhere [5]. Homologous sequences were identified using BLAST with default parameters [9] and a multiple sequence alignment was generated using Clustal-X [10], again using default paramters. The identity between the sequences varied between 30 and 99%; the multiple sequence alignments are given as supporting information. Conservation was determined from this multiple sequence alignment using entropy, variability and maximum proportion. Entropy was defined using $S_j = \Sigma N_i \log N_i$ where $S_j$ is the entropy at position $j$ and $N_i$ is the fraction of amino acid $i$ at position $j$ [11]. We define the quantities variability ($V_j$) and maximum proportion ($M_j$) as, respectively, the number of different amino acids and the proportion for the most abundant amino acid at position $j$. The hydrophobicity scale chosen was the amino acid octanol water partition coefficient data of White and Wimley [12]. Conservation and hydrophobicity data were normalised such that 0.0 was 100% conserved and hydrophobic while 1.0 was not conserved and hydrophilic. A window of five residues was used to smooth the hydrophobicity and conservation data; the use of such windows is common in hydropathy analysis [13].

## 3. Results and discussion

The plots of conservation against residue number for 1aps, 256b, 7tim, 2rhe and 1ubq are shown in Fig. 2; the plots of entropy against residue number for the remaining four proteins are shown in Fig. 3. The closed loops are shown as black horizontal bars.

Predominantly, the minima (i.e. the most conserved regions) lie adjacent to the ends of the closed loops. This is particularly apparent for the start of the first loop in 1aps and for most of the loops in 7tim. However, there are some minima that do not coincide with closed loop ends (e.g. loop 1 of 256b and loop 6 of 7tim) and there are some that correspond to a junction between two loops (e.g. for loops 2 and 3 of 256b)—this region not only forms locks for the two loops but also contacts the heme.

The three alternative measures of conservation generally give similar results. Since entropy takes into account both maximum proportion and variability, it usually gives intermediate values. The agreement between the three measures is particularly clear for 7tim (except in the case of the eighth loop).

Over the nine proteins, taking 0.4 as a reasonable, albeit arbitrary, cut off, 70% of the loop ends are found to be conserved. This percentage is similar to the 74% of closed loop ends found to be hydrophobic [14]. Generally, as shown in Fig. 3 for 2fox, 1thb, 1ilb and 2stv, the plots of hydrophobicity tend to parallel those of conservation (entropy). A notable exception occurs around residues 92–96 of 1thb where a high degree of conservation does not correspond with a high degree of hydrophobicity. The product of hydrophobicity and entropy is similarly low in the region of the loop ends. Considering the minima in the graphs for conservation, hydrophobicity and conservation × hydrophobicity, it is possible to determine the distribution of distances between conserved and hydrophobic regions (from the set of distances along the sequence between any two minima). The resultant plot for entropy in Fig. 4 (grey) shows that the majority of the conserved residues are separated by ~35 residues (i.e. they lie in the bin for 31–40 residues), which clearly contains the standard closed loop size (25–35 residues). Similar conclusions can be drawn from the hydrophobicity and conservation × hydrophobicity plots. The difference in peak position between the entropy curve and those for hydrophobicity and entropy × hydrophobicity arises primarily from the limitations in the size of the data set (nine proteins), from the use of a bin of size 10 and also from the

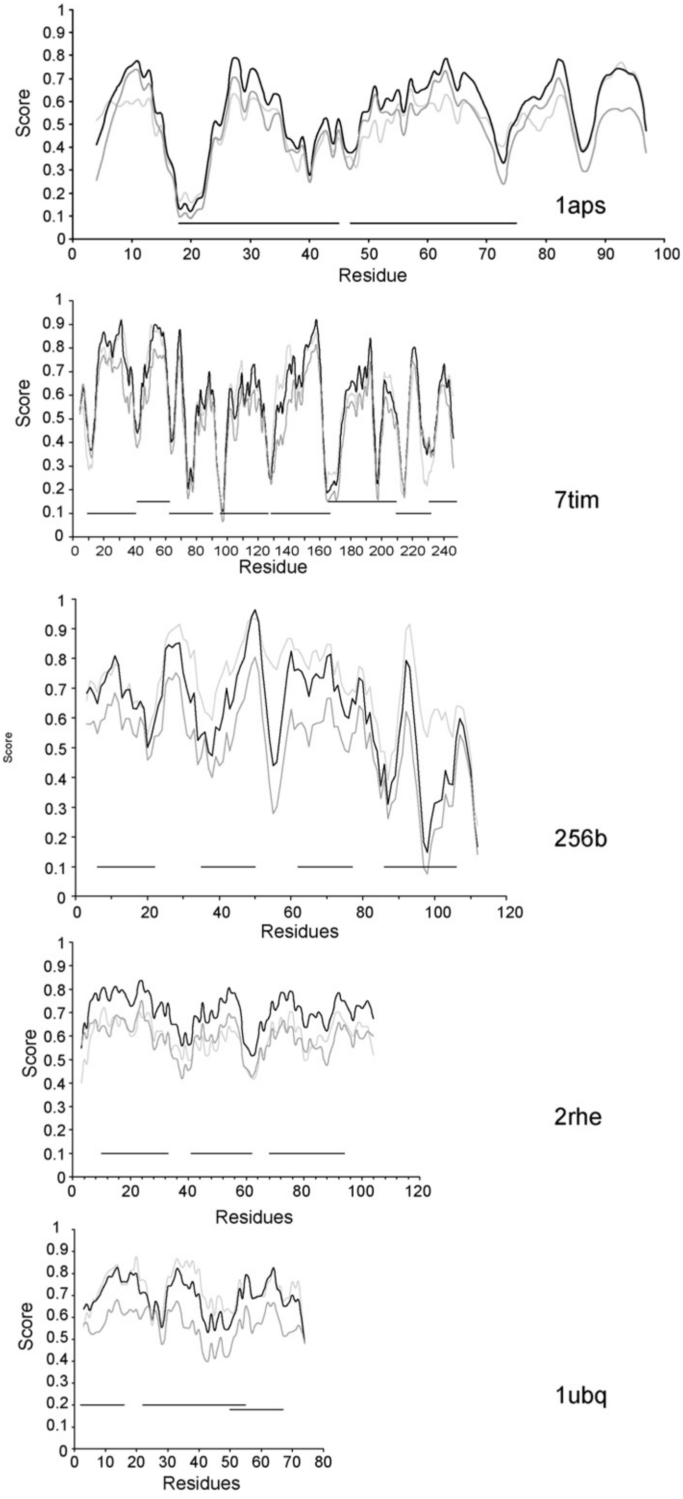

Fig. 2. Plots of conservation, determined using entropy (solid black line), variability (light grey line) and maximum proportion (dark grey line) against residue number for 1aps, 7tim, 256b, 2rhe and 1ubq. The closed loops are shown as black horizontal bars.

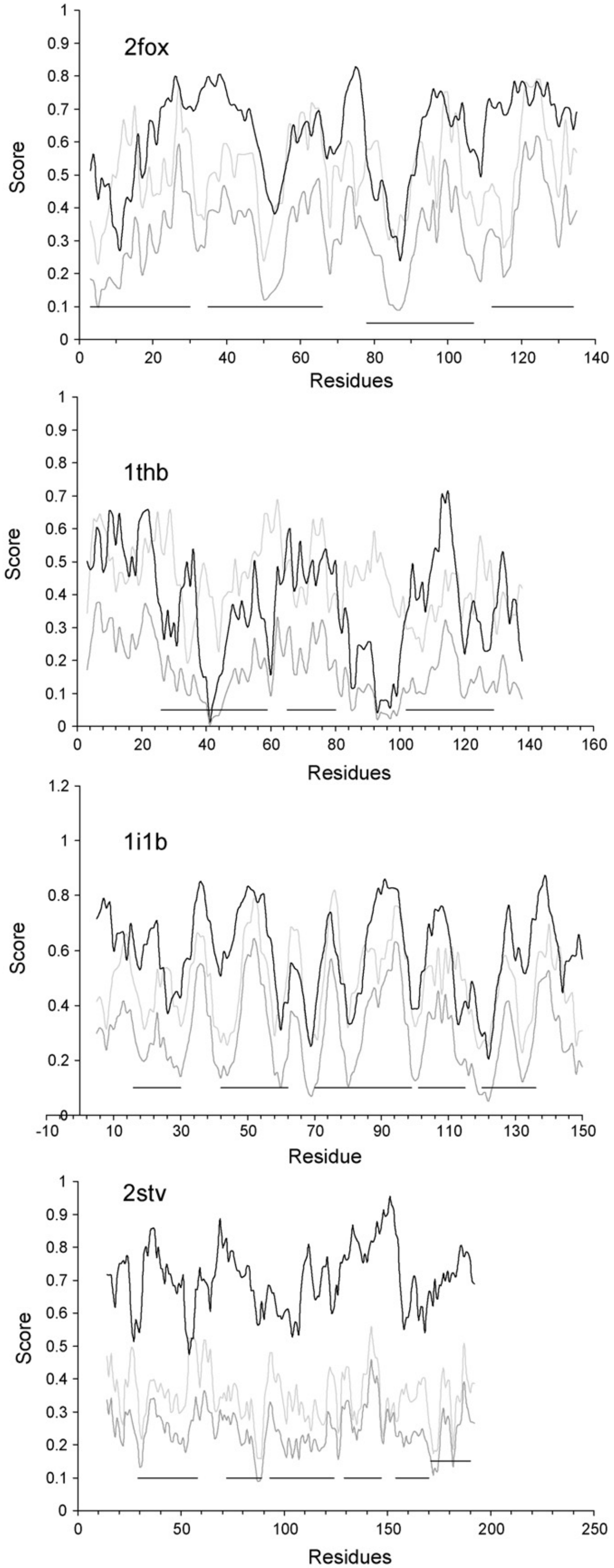

Fig. 3. Plots of conservation (black), hydrophobicity (light grey) and conservation × hydrophobicity (dark grey) against residue number for 2fox, 1thb, 1ilb and 2stv. The closed loops are shown as black horizontal bars.

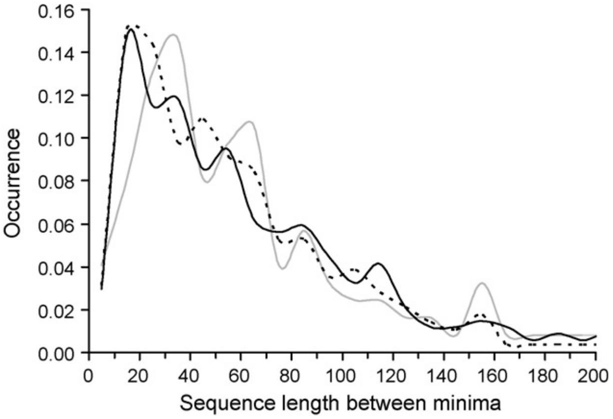

Fig. 4. Distribution of length (measured in residues) between hydrophobic and conserved residues for the nine major superfold proteins. Conserved residue to residues length distribution is given in grey, hydrophobicity in black and the product of the two (conservation × hydrophobicity) in black (dotted).

presence of hydrophilic and non-conserved residues within the window of five largely hydrophobic residues.

## 4. Conclusions

The observation of conserved residues at the loop ends is consistent with the Berezovsky and Trifonov's hypothesis of structurally important hydrophobic residues at the ends of closed loops. In addition, the distribution of sequence lengths between conserved and between hydrophobic regions is consistent with closure of loops of length ∼25–35. Both of these observations are consistent with the hypothesis that loop closure through hydrophobic interactions may play an important role in protein folding.

## References

[1] L. Stryer, J.L. Tymoczko, J.M. Berg, Biochemistry, W.H. Freenab, New York, 2002.

[2] I.N. Berezovsky, E.N. Trifonov, Loop fold structure of proteins: resolution of Levinthas paradox, J. Biomol. Struct. Dyn. 20 (2002) 5–6.

[3] I.N. Berezovsky, Discrete structure of van der Waals domains in globular proteins, Protein Eng. 16 (2003) 161–167.

[4] I.N. Berezovsky, V.M. Kirzhner, A. Kirzhner, E.N. Trifonov, Protein folding: looping from hydrophobic nuclei, Proteins Struct. Funct. Genet. 45 (2001) 346–350.

[5] I.N. Berezovsky, A.Y. Grosberg, E.N. Trifonov, Closed loops of nearly standard size: common basic element of protein structure, FEBS Lett. 466 (2000) 283–286.

[6] E.N. Trifonov, I.N. Berezovsky, Evolutionary aspects of protein structure and folding, Curr. Opin. Struct. Biol. 13 (2003) 110–114.

[7] E. Aharonovsky, E.N. Trifonov, Protein sequence modules, J. Biomol. Struct. Dyn. 23 (2005) 237–242.

[8] E. Aharonovsky, E.N. Trifonov, Sequence structure of van der Waals locks in proteins, J. Biomol. Struct. Dyn. 22 (2005) 545–553.

[9] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[10] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 25 (1997) 4876–4882.

[11] M.K. Dean, C. Higgs, R.E. Smith, P.D. Scott, R.P. Bywater, T.J. Howe, C.A. Reynolds, Entropy in the alignment and dimerization of class C G-protein coupled receptors, in: R.H. Templer, R. Leatherbarrow (Eds.), Biophysical Chemistry: Membranes and Proteins, Royal Society of Chemistry, 2002, pp. 85–93.

[12] S.H. White, W.C. Wimley, Hydrophobic interactions of peptides with membrane interfaces, Biochim. Biophys. Acta 1376 (1998) 339–352.

[13] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.

[14] B.K. Yew, G.J.G. Upton, C.A. Reynolds, A new strategy for the determination of closed loops, submitted for publication.