Introduction

A protein chain trajectory makes many returns to itself, thus forming closed loops (Berezovsky et al., 2000; Berezovsky and Trifonov, 2001a,b), as the analysis of protein crystal structures shows. The loops of nearly standard size (25–30 residues) dominate and the proteins actually consist of assemblies of these standard elements. Their size is confirmed by an independent study on the protein sequences of 23 complete bacterial proteomes, one 32-residue prototype was studied here in detail. Numerous sequence segments related to the prototype are identified in the crystal structures of proteins of a PDB_SELECT database. Analysis of the respective chain trajectories for the cases with different degrees of sequence conservation confirms that the majority of the segments correspond to the closed loops. In the evolutionary diversification of the prototypes the secondary structure yields first, while the sequence is still moderately conserved. The last feature to go is the chain return property. Apparently, the opening of the loops would severely destabilize the protein fold, which explains their conservation.

Keywords: closed loops/protein chain return persistence/protein evolution

Materials and methods

The protein sequences of the following complete prokaryotic genomes were used for the calculations: Archaea, A.pernix, A.fulgidas, M.thermoaerophilicum and Pabysisi; and Eubacteria, A.aeolicus, B.burgdorferi, C.jejuni, C.pneumoniae, C.trachomatis, D.radiodurans, E.coli, H.influenzae, H.pylori, M.tuberculosis, M.pneumoniae, N.meningitidis, P.rubrazeki, S.synechocystis, T.maritima, T.pallidum, U.urealyticum, V.cholerae and X.fastidiosa. The sequences were provided by the National Center for Biotechnology Information, via Entrez Browser. The sequences were used without any filtering. The search for the most frequent motifs of the size 30 residues consisted of the following steps. (i) Every 30 amino acid long segment from an exhaustive collection of about one million different segments taken from the proteome of E.coli is matched to all protein sequences of 23 complete bacterial proteomes. The number of matching fragments is counted. In this first step the threshold is taken as equal 11 matching residues, to ensure a sufficiently large number of well matching (>37% match) fragments. (ii) The collected matching fragments (of the order of several hundred for a successful sequence motif) are used for derivation of an initial matrix of distribution of 20 amino acid residues in 30 positions. (iii) The initial matrix is used for the next round of collecting the matching fragments. In this case the comparison is made between the matrix and the tested sequence, rather than sequence-to-sequence as in the initial stage. The similarity of each 30-residue long tested sequence from the bacterial proteomes is calculated as an average of matching normalized matrix elements. This average should exceed a certain minimal value: the threshold of similarity in which case the sequence fragment is included in the family for calculation of the next round matrix. Inclusion of sequence fragments of lower similarity would cause instability of the matrix, that is, its divergence in the iteration process. The range of the threshold values...
corresponding to stable solutions may vary (between 0.4 and 0.8) depending on the amino acid compositions of the initial 30-residue sequence and of the whole sequence ensemble. It is noteworthy that the convergence of the iterated matrices to a stable solution is the sole criterion to consider the respective pattern as being present in the sequences, irrespective of the numbers of the matching segments in natural and random (shuffled) sequences. This also means that the choice of the initial match and the subsequent matrix thresholds is empirical.

(iv) From a total of about one million different tested segments are chosen those which show the highest final scores after several rounds (typically, 5–10 rounds) of convergent iteration of the matrices. (v) The sequence dimension of the resulting matrix is adjusted by inspecting frequencies of the amino acid residues below and beyond the initial 30-residue range. For example, the matrix for the sequence prototype analyzed in this work has the sequence dimension of 32 residues. In the histogram shown in Figure 1 it corresponds to the dark gray area with high frequency values at the edges. If the dimension is taken beyond the 32 residues, the respective frequency values will drop to the background level. On the other hand, the choice of a shorter range will result in high values beyond it, indicating that the choice is not optimal. The size of the family of the ‘descendants’ of this prototype is 978 fragments. The similarity threshold is taken as equal to 0.4.

**Results and discussion**

Table I presents a converged matrix (available in electronic form upon request) with the ‘consensus sequence’ LSGGQRQRVAILARALEPKLLLLLEPTSA. Columns of the matrix correspond to the amino acids, the rows to their positions in the sequence and the highest frequency elements (bold face) to the consensus sequence. The matrix and the sequence provide a generalized description of sequences of the 23 bacterial proteomes. The actual number of related sequences is much higher, since a high discriminating threshold of the 23 bacterial proteomes.

Figure 1 displays frequencies of the ‘consensus’ residues beyond the 32 amino acid prototype sequence. The typical match is 2–5 residues. For respective random sequences the expected match is 30/20 = 1.5 residues; the higher match of 2–5 residues is due to non-uniformity of the composition.) Figure 2 includes four cases with the highest matches observed [from 11 to 26 residues, respectively; Figure 2(A)–(D)], and also as a representative set of lower sequence match structures. Among the segments with a match of nine residues [Figure 2(H)–(L)] and data not shown), 12 display a closed loop structure and 10 segments have a non-loop appearance [as in Figure 2(K) and (L)]. The four higher sequence match structures (11–26 residues) are all of the type α-turn–β. With the match 9–10 residues (28 cases in total), this structural motif still dominates, appearing nine times. In other cases these are either loops of different structures [e.g. Figure 2(F), (H) and (I)] or non-loop sections (seven and 12 times, respectively). Thus, of 32 structures, the highest sequence match cases (four) correspond to standard α–turn–β elements; nine more cases have lower sequence match, but still retain the α–turn–β structure; seven structures of lower sequence match have lost the α–turn–β motif, while retaining their loop property; finally, 12 elements of lower sequence match have lost both the typical structural motif and the chain return property.

**Fig. 1.** Histogram of the most frequent (‘consensus’) elements of the prototype matrix.

**Fig. 2.** Chain trajectories of the matching segments extracted from the PDB_SELECT database (Hobohm and Sander, 1994). A, protein 1b0u, chain A (match 26); B, 1fl2, chain B (match 15); C, 1cs1, chain A (match 12); D, 1lfh, chain A (match 11); E, 1qap, chain A (match 10); F, 1di1, chain A (match 10); G, 1ds1, chain A (match 10); H, 1nf1, chain A (match 9); I, 8ohm, chain A (match 9); J, 1d4c, chain C (match 9); K, 8gsv, chain A (match 9); L, 1guq, chain A (match 9).
As the data above demonstrate, the prototype α–turn–β structure survives even after 72% of the presumed original prototype sequence has been lost. However, the chain return property is still conserved in the low-match structures, where the prototype secondary structure is lost. Apparently, the return property is still maintained owing to either marginal sequence conservation or supporting influence of the remaining parts of the fold or both. Unfolding of the closed loop would cause severe changes in the path of the protein chain, while sequence variations and changes in the details of secondary structure would only be of local influence and, in general, of less importance for the overall protein fold.

This work introduces an important dimension in protein evolutionary studies. Highly diverged protein segments with barely recognizable sequence similarity and with no structural resemblance may still be related as soon as segments appear as closed loops in the folds. For possible generalization of the above observations, other sequence prototypes have to be analyzed in a similar way (work in progress). Many of the closed loops observed in the crystallized proteins resemble the structural types described in the early study by Levitt and Chothia (Levitt and Chothia, 1976). It remains to be seen what would be a complete spectrum of the closed loop structures.

Acknowledgements

We are grateful to Mrs. A. Weinberg for editing of the text. I.N.B. is a Post-Doctoral Fellow of the Feinberg Graduate School, Weizmann Institute of Science. V.M.K. is supported by the Ministry of Absorption.

References


Received May 28, 2002; revised September 6, 2002; accepted October 1, 2002