Protein Science (1997), 6:1587–1590. Cambridge University Press. Printed in the USA.

FOR THE RECORD

Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns

KAREN F. HAN,[1,3] CHRISTOPHER BYSTROFF,[2] AND DAVID BAKER[2]

[1]Graduate Group in Biophysics, University of California San Francisco School of Medicine, San Francisco, California 94143-0448
[2]Department of Biochemistry, University of Washington, Seattle, Washington 98195

(RECEIVED March 18, 1997; ACCEPTED April 18, 1997)

Abstract: We have used cluster analysis to identify recurring sequence patterns that transcend protein family boundaries. A subset of these patterns occur predominantly in a single type of local structure in proteins. Here we characterize the three-dimensional structures and contexts in which these sequence patterns occur, with particular attention to the interactions responsible for their structural selectivity.

Keywords: cluster analyses; protein folding; sequence motifs; structure production

The traditional approach to characterizing the mapping between amino acid sequence and local structural properties is to decide first on the important structural properties and then investigate their associated amino acid probability distributions. The prediction of protein secondary structure (Chou & Fasman, 1978; Presnell et al., 1992; Rost & Sander, 1993) and residue environments (Bowie et al., 1991) are examples of this approach: relationships between sequence and the predefined structural properties are identified using the database of sequences whose structures are known, and then used to predict the structural characteristics of new sequences. This is supervised learning where correlations between two variables are sought from a large set of examples. An alternative approach is unsupervised learning, where patterns are sought in a data set without reference to correlations with other variables. Such an approach is less useful for prediction because groupings are not chosen to optimize the prediction of the second variable from the first. However, unsupervised learning has the advantage that the important properties need not be specified in advance and thus new patterns and groupings can be identified more readily.

We have used unsupervised learning methods to identify recurring amino acid sequence patterns (Bystroff et al., 1996). Sequence segments ranging from 3 to 15 residues in length from a nonredundant subset (Hobohm & Sander, 1994) of the HSSP database of multiple sequence alignments (Sander & Schneider, 1991) were partitioned into groups of related sequences (Han & Baker, 1995). In a subsequent study (Han & Baker, 1996), the structural correlates of the sequence patterns were investigated. Many of the sequence patterns were found to occur primarily in one or two types of secondary structural elements in proteins, whereas virtually no such patterns were found in a control data set in which the sequence structure relationships of the segments were randomized. However, this connection of sequence patterns with secondary structural elements did not fully capitalize on the power of unsupervised learning approaches noted in the previous paragraph: the potential to identify new structural properties and groupings. Toward this end, in this paper, we investigate the three-dimensional structures adopted by a particularly interesting subset of the sequence motifs. We find that many of the motifs not only occur in well-defined three-dimensional structures, but also in well-defined protein contexts.

Results: The six sequence–structure motifs selected for detailed characterization span a wide range of local structures (Fig. 1). In principle, a set of closely related short sequences might consistently adopt the same structure in proteins for a variety of reasons, including specific conserved side-chain–side-chain or side-chain–main-chain interactions or favorable side-chain–solvent interactions. These factors are considered for each of the six motifs in turn.

To illustrate the approach, we begin with motif I, the well-studied N-terminal helix capping box (Harper & Rose, 1993). The sequence pattern shown in Figure 1 is a concise representation of the mean frequency of occurrence of the amino acids at each position in the motif (Fig. 1 legend). The red triangles depict the positions of protein α-carbons surrounding the superimposed segments. Specific interactions account for the pronounced structural preference of this motif. Three interactions break the N-terminal propagation of the α-helix: the side chain at position 7, usually Glu or Gln, makes a hydrogen bond with the backbone nitrogen at position 4; the side chain at position 4, usually Ser, Thr, Asp, or Asn, makes a hydrogen bond with the backbone nitrogen at position 7; and a hydrophobic contact occurs between the side chains at positions 3 and 8. There is also an aspect of "negative design";

Reprint requests to: David Baker, Department of Biochemistry, University of Washington, Seattle, Washington; e-mail: dabaker@u.washington.edu.
[3]Present address: Northwestern University Medical School, Box 182, Chicago, Illinois 60611.

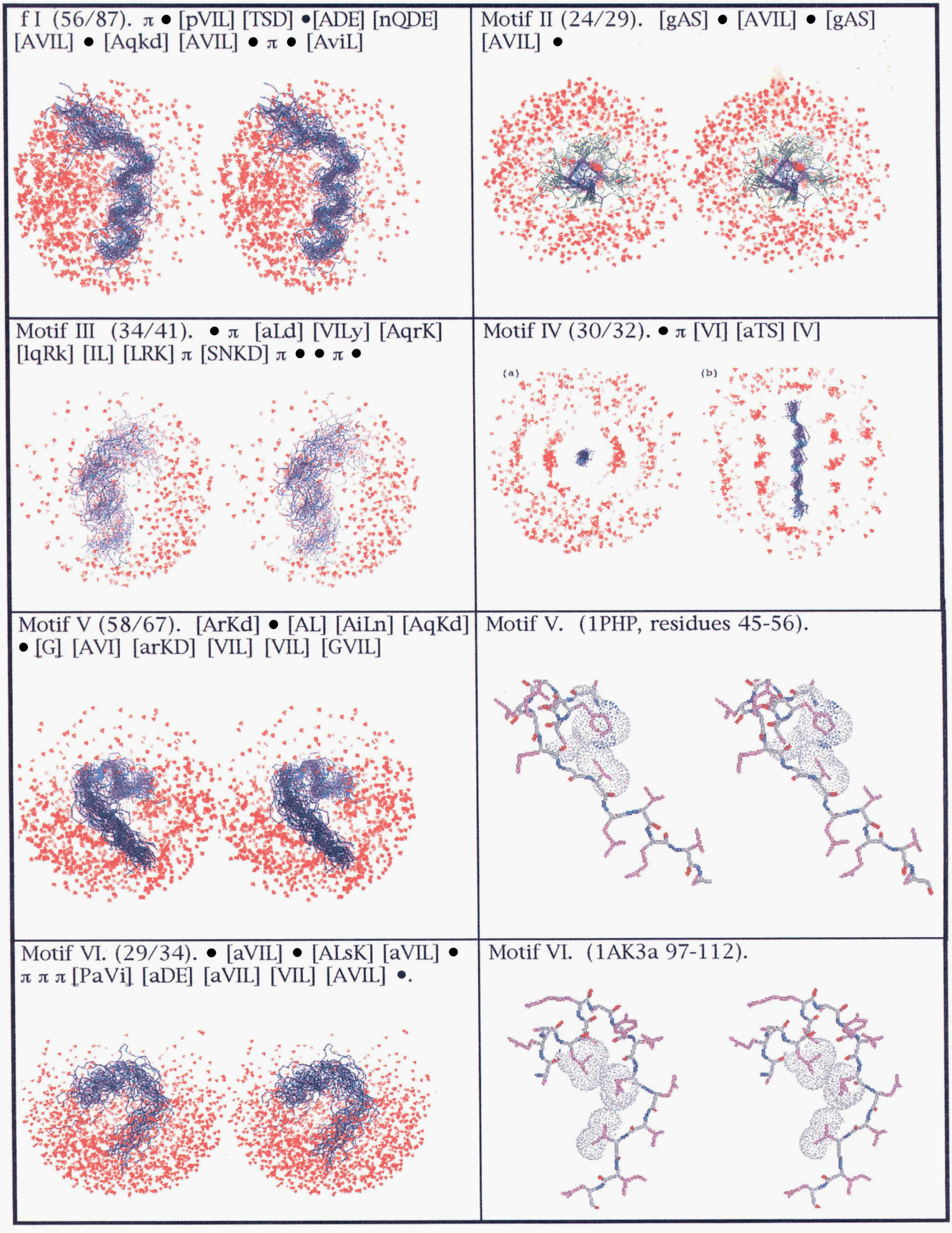

Fig. 1. *See caption on facing page.*

the five-residue separation between the hydrophobic residues 3 and 8 breaks the periodicity of an amphipathic helix.

Motif II occurs predominantly in buried helices. There is an intriguing pattern of small residues on one side of the helix and large residues on the other; the small residues may be constrained by packing interactions. Also, because small residues are less hydrophobic than large nonpolar residues, the sequence pattern has the amphipathic periodicity consistent with helix formation and stabilization, but can still be buried. Such a sequence might fold to a helix originally due to the amphipathicity, and only later be buried. The identification of this motif is interesting in light of the results of Benner et al. (1994), who showed that buried helices are particularly difficult to distinguish from surface ones.

Motif III is an amphipathic helix terminated by a strongly polar segment. The motif usually occurs on the surface of the protein (Fig. 1). Helix termination here appears to be brought about by negative design. Positions 4 and 7 are consistently nonpolar, but a conserved nonpolar residue at position 10 or 11 to continue the amphipathic α-helix pattern is conspicuously absent. Instead, positions 9–11 are either strongly polar, contain nonpolar side chains in positions that are out-of-register with the preceding turn of the helix, or contain a proline. All disfavor the formation of an additional turn of helix by positions 9–12.

Motif IV is a short segment of buried amphipathic strand. The β-branched nonpolar side chains at positions 3 and 5 restrict the backbone dihedral angles and provide one conserved nonpolar contact. The striking superposition of the C alpha atoms surrounding the segments (Fig. 1) suggests that the motif frequently occurs in an interior strand of a five- to seven-stranded beta sheet.

Motif V is an extension of the Schellman α-helix C-terminal capping motif (Schellman, 1980; Aurora et al., 1994), which continues from the helix C cap into a buried beta strand (Fig. 1). There are at least five specific conserved contacts surrounding the conserved glycine at position 7. Backbone nitrogens at positions 7 and 8 make hydrogen bonds with the backbone oxygens of positions 4 and 3, respectively. The nonpolar side chain at position 8 interacts with the nonpolar side chain at position 3, and sometimes with the side chain at position 6. The nonpolar side chain at position 10 interacts with the nonpolar side chains of positions 3 and 4, creating a small hydrophobic cluster around position 3. Glycine at position 7 allows the backbone to adopt a left-handed turn at the end of the helix ($\phi > 0$) (84%). This occurs frequently even when position 7 is not a glycine (38%). The turn is perhaps the tightest, which places a nonpolar side chain of a β-strand (position 10) between two nonpolar side chains of a preceding α-helix (positions 3 and 4).

Motif VI is a surface α-helix (positions 1–6) which turns into a buried β-strand (positions 11–15) that usually folds back on the helix (65%). Positions 2, 5, 12, and 14 prefer nonpolar side chains, whereas positions in the turn region (7–9) prefer polar side chains. The turn is longer than that of motif V, with about six residues between the last nonpolar side chain of the α-helix (position 5) and the first nonpolar side chain of the β-strand (position 12). The motif is stabilized by a pair of sequential hydrophobic interactions involving residue 10 (Fig. 1). This occasionally results in a small cavity that is not filled by other residues in the protein; when the cavity is filled, the pairing strand often contributes its nonpolar residue to the hydrophobic cluster. Most members in this cluster that do not have the helix-turn-strand structure maintain the conservative sequential hydrophobic interaction.

Motifs V and VI illustrate quite different mechanisms for generating helix to strand transitions. In motif V, a glycine is required to make the turn sufficiently tight to bring the three nonpolar residues into contact and reinforce the helix stop signal. In contrast, in motif VI, the only required feature in the turn is the conserved hydrophobic residue (or proline) at position 10; the two nonpolar contacts involving this residue appear to drive the change in chain direction.

Discussion: Our previous studies led to the identification of sequence patterns that occur primarily in a single type of local structure in proteins. The above analysis shows that, in most cases, the strong structural preferences of the sequence patterns can be accounted for by conserved side-chain contacts, matching of sequence amphipathicity to secondary structure periodicity, and residue conformational preferences. Negative design also appears to play an important role, particularly in terminating secondary structural elements.

Because of the important role "folding initiation sites" play in some models of protein folding (Avbelj & Moult, 1995), it is tempting to speculate that the sequence patterns might adopt the corresponding local structure independent of the rest of the chain. The interest in folding initiation sites stems from their possible relevance to the resolution of the Levinthal paradox: the greater the degree to which local interactions constrain the conformation of the chain, the smaller the conformational space that must be searched during folding. Such restrictions could also play an important role in computer search strategies; for example, in a recent protein structure prediction algorithm, conformations highly populated when only local interactions were considered were kept "frozen" when longer-range interactions were introduced subsequently (Srinivasan & Rose, 1995).

Certainly, a property that might be expected of the sequence of a folding initiation site is that it occurs consistently in the same structure in proteins, and this is the case for the sequence patterns described here (the patterns may be viewed as being "frozen" in known protein structures). A possible caveat is that the sequence

Fig. 1 (*facing page*). Three-dimensional structures and contexts associated with sequence patterns. The ratio for each motif is the total number of segments in the cluster (denominator) and the number of segments superimposed in the figure (numerator); the remaining segments have quite different local structures. To depict structural context, the positions of alpha-carbon atoms that surround up to 30 randomly selected instances of each of the motifs are shown (red triangles). The sequence patterns are described using a simple shorthand. Letters within brackets indicate the prominent amino acids at a single position: upper-case, frequencies greater than 0.1; lower-case, frequencies between 0.07 and 0.1. Positions at which more than seven different amino acids occurred with frequencies greater than 0.05 are represented by π (polar), ϕ (hydrophobic), and · for average hydrophobicities (sum of the frequencies of A, V, I, L, M, P, F, and W) of less than 0.35, greater than 0.65, and between 0.35 and 0.65, respectively. For example, the first residue in motif I is usually polar, and the fourth residue is frequently T, S, or D. All images except for motif IV are stereo pairs.

patterns may reflect optimization of the stability of a local structural element not in isolation, but in frequently occurring protein contexts, and sequences that consistently adopt a particular conformation (extended beta strand, for example) in protein structures may not preferentially adopt that conformation in isolated peptides.

The motifs described in this paper were selected based on the correlation between sequence patterns and secondary structure. No structural information was used in the clustering process that generated the clusters, and the strong associations between sequence and structure are not artifacts of the selection process: no such associations were observed in randomized data sets subjected to the same clustering procedure (Han & Baker, 1996). However, as expected for an unsupervised learning procedure, the structures for a number of the motifs (particularly those involving beta strands) are relatively poorly defined. Current efforts are directed at refining the patterns using structural information to increase both their structural consistency and their predictive power.

Methods: Profile segments from the HSSP database of multiple sequence alignments for proteins of known three-dimensional structure were clustered using the K-means algorithm and a sequence-based distance measure as described (Han & Baker, 1996). Clusters in which a large fraction of the segments had similar secondary structures were chosen for further characterization (Table 3 in Han & Baker, 1996).

Because of the importance of nonlocal interactions, not all segments in the same cluster have similar structures. For each motif, the RMS deviation (RMSD) in backbone atom positions between all pairs of segments was calculated. A cutoff RMSD was selected such that at least 80% of the segments in the group were within the cutoff distance from the central segment, the segment with the lowest sum of RMSDs to all of the other segments. Segments within the cutoff distance are shown superimposed on the central segment in Figure 1; the remaining segments would be false positives for a classifier based on the motifs. The average RMSD from the central segment for the segments within the cutoff was 1.8, 0.6, 2.9, 0.6, 1.6, and 4.5 Å, for motifs I–VI, respectively. The number of segments within molecular graphics inspection was determined using Insight II (Biosym technologies) or MidasPlus (UCSF). The nomenclature for beta turns is that of Thornton and coworkers (Hutchinson & Thornton, 1996).

Acknowledgments: We thank R. Schneider and C. Sander for the HSSP database and D. Teller, K. Zhang, R. Klevit, and members of the Baker laboratory for critical reading of the manuscript. K.F.H. was supported by a Howard Hughes Medical Institute Predoctoral Fellowship. This work was partially supported by the National Science Foundation, Science and Technology Center Cooperative Agreement BIR-9214821, and young investigator awards to D.B. from the NSF and the Packard Foundation.

References

Aurora R, Srinivasan R, Rose GD. 1994. Rules for alpha-helix termination by glycine. *Science 264*:1126–1130.

Avbelj F, Moult J. 1995. Determination of the conformation of folding initiation sites in proteins by computer simulation. *Proteins Struct Funct Genet 23*:129–141.

Benner S, Badcoe I, Cohen M, Gerloff D. 1994. Bona fide prediction of aspects of protein conformation. Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. *J Mol Biol 235*:926–958.

Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Bystroff C, Simons KT, Han KF, Baker D. 1996. Local sequence-structure correlations in proteins. *Curr Opin Biotechnol 7*:417–421.

Chou P, Fasman G. 1978. Empirical predictions of protein conformation. *Annu Rev Biochem 47*:251–276.

Han KF, Baker D. 1995. Recurring local sequence motifs in proteins. *J Mol Biol 251*:176–187.

Han KF, Baker D. 1996. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci USA 93*:5814–5818.

Harper ET, Rose GD. 1993. Helix stop signals in proteins and peptides: the capping box. *Biochemistry 32*:7605–7609.

Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci 3*:522–524.

Hutchinson E, Thornton J. 1996. PROMOTIF—A program to identify and analyze structural motifs in proteins. *Protein Sci 5*:212–220.

Presnell S, Cohen B, Cohen F. 1992. A segment-based approach to protein secondary structure prediction. *Biochemistry 31*:983–993.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Schellman C. 1980. Protein folding. *Protein folding: Proceedings of the 28th Conference of the German Biochemical Society, held at the University of Regensburg, Regensburg, West Germany, September 10–12, 1979*. New York: Elsevier/North-Holland.

Srinivasan R, Rose G. 1995. LINUS: A hierarchic procedure to predict the fold of a protein. *Proteins Struct Funct Genet 22*:81–99.