Predicting Coiled Coils from Protein Sequences

ANDREI LUPAS, MARC VAN DYKE, JEFF STOCK*

The probability that a residue in a protein is part of a coiled-coil structure was assessed by comparison of its flanking sequences with sequences of known coiled-coil proteins. This method was used to delineate coiled-coil domains in otherwise globular proteins, such as the leucine zipper domains in transcriptional regulators, and to predict regions of discontinuity within coiled-coil structures, such as the hinge region in myosin. More than 200 proteins that probably have coiled-coil domains were identified in GenBank, including α- and β-tubulins, flagellins, G protein β subunits, some bacterial transfer RNA synthetases, and members of the heat shock protein (Hsp70) family.

COILED COILS ARE FORMED BY TWO or three α helices in parallel and in register that cross at an angle of approximately 20°, are strongly amphipathic, and display a pattern of hydrophilic and hydrophobic residues that is repeated every seven residues (*1*). The seven positions of the heptad repeat are designated a through g, a and d being generally hydrophobic. The sequences of coiled-coil proteins exhibit common patterns of amino acid distribution that appear to be distinct from those of other proteins (*1, 2*). Here we show that, by comparing the sequence of a given protein with that of known coiled-coil proteins, it is possible to identify regions of coiled-coil structure.

The sequences of the coiled-coil domains from tropomyosins, myosins, and keratins deposited in GenBank (*3*) provided a coiled-coil database. The frequency of occurrence of each amino acid at each position of the coiled-coil heptad repeat was tabulated separately for the three groups of proteins. These values were averaged and divided by the overall frequency of occurrence of each amino acid in GenBank to establish relative frequencies in coiled coils (Table 1). The relative frequencies were used for evaluation of sequences by means of a gliding window of 28 residues. This length was chosen because the shortest peptides still exhibiting a stable coiled-coil structure in solution are between four and five heptads long (*4*). A preliminary score for a residue within the window was computed by (i) assigning the sequence in the window a heptad repeat frame, (ii) assigning each residue in the window the relative frequency obtained from Table 1, and (iii) taking the geometric average over the frequencies in the window (*5*). Because a window can be assigned seven different heptad repeat frames and a residue can occupy 28 positions in the gliding window, there are 196 preliminary scores for each residue; of these, the highest one was taken as the score for that residue (*6*). If a residue was closer than 28 residues from an end, then fewer preliminary scores were computed because the window was not allowed to glide past the terminus.

We used this algorithm to analyze databases containing (i) sequences of known globular proteins, (ii) randomly generated sequences, and (iii) all the sequences in GenBank (*3, 7*). The score distributions obtained from these three databases were similar to one another but different from the distribution obtained from the coiled-coil database (Fig. 1). In the region of overlap between scores obtained for globular and coiled-coil sequences (Fig. 1A), all the globular sequences with scores above 1.1 corresponded to long amphipathic α helices, whereas most of the coiled-coil sequences with scores below 1.3 were either from helices 1A and 2A in keratins, which are short and probably stabilized by the vicinity of the long helices 1B and 2B, or from regions that are likely to represent discontinuities in coiled-coil structure (Fig. 2).

The score distributions in Fig. 1 allowed an estimate of the probability that a residue with a given score would be in a coiled coil. To accomplish this, we approximated the globular and coiled-coil score distributions by Gaussian curves, G_g and G_{cc}, respectively (*8*), to yield finite distribution values for scores lying outside those obtained with our limited databases. The ratio of coiled coil to globular residues in GenBank was then estimated by approximation of the GenBank score distribution with scaled globular and coiled-coil Gaussians. The best fit indicated a ratio of 1:30. The probability P of forming a coiled coil of a score S is then, as shown in Fig. 1D,

$$P(S) = G_{cc}(S)/[30G_g(S) + G_{cc}(S)]$$

We first applied this coiled-coil prediction algorithm to the analysis of proteins with established coiled-coil structures. It has been a matter of debate whether the myosin coiled coil contains a region of low structural stability that can lead to the formation of a hinge (*9*). Fluorescence depolarization and viscoelastic studies have indicated that the rod is rigid, and fast Fourier transform

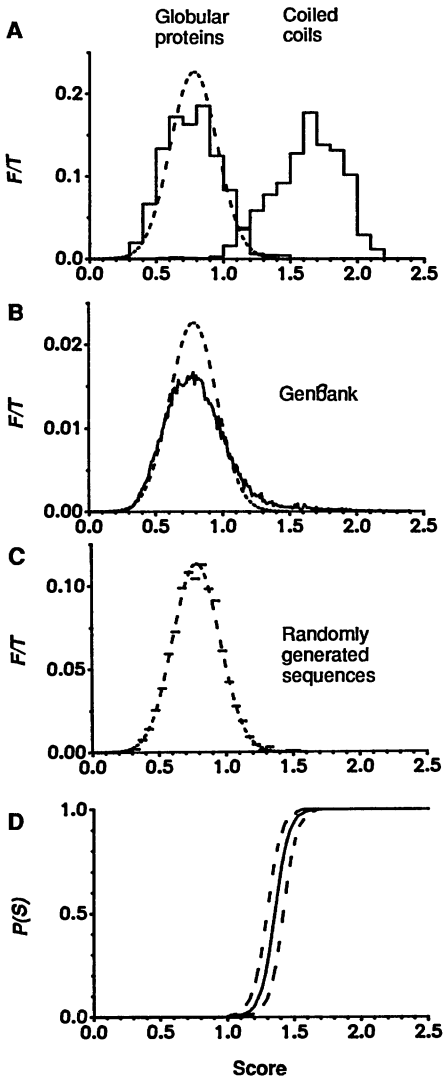

Fig. 1. (**A**) Distribution of scores in globular proteins (29,661 scores; mean, 0.77; σ, 0.20) and coiled-coil sequences (16,968 scores; mean, 1.63; σ, 0.24). In panels A, B, and C, the fraction of scores in the histogram interval (0.1 for A, 0.01 for B, 0.05 for C) is shown on the y-axis, and the score distribution of random sequences is shown as a Gaussian curve (dotted line). To evaluate the coiled-coil sequences we did not use the combined residue frequencies obtained from myosins, tropomyosins and keratins but rather omitted the frequencies obtained from the group of proteins under study. Obtained scores were normalized and then averaged to yield the histogram shown. The distribution of scores for the three groups of proteins were: myosins (8444 scores; mean, 1.70; σ, 0.19); tropomyosins (3293 scores; mean, 1.75; σ, 0.21); keratins (5762 scores; mean, 1.44; σ, 0.23). (**B**) Distribution of scores in GenBank (2,002,907 scores; mean, 0.80; σ, 0.19). (**C**) Distribution of scores in randomly generated sequences (49,169 scores; mean, 0.78; σ, 0.18); the vertical lines of the histogram are omitted to show the close fit between the histogram and its Gaussian curve. For all sets of data, scores of 0 were eliminated before statistical analysis. The fraction of scores of 0 was 0.090 in globular proteins, 0.059 in random sequences, 0.117 in GenBank, 0 in coiled coils. (**D**) Plot of the probability of forming coiled coils, $P(S)$, versus score. The dotted lines show plots of $P(S)$ versus score for ratios of coiled coil to globular residues of 1:10 and 1:100.

Department of Molecular Biology, Princeton University, Princeton, NJ 08544.

*To whom correspondence should be addressed.

analysis revealed an essentially unbroken continuity of heptad repeats over the entire length of the rod. Nevertheless, several electron microscopy studies have shown the existence of a bend, and the region containing this bend has been found to be highly susceptible to proteolysis. Thermal denaturation and scanning calorimetry have also indicated the existence of regions of low stability toward the center of the rod. The hinge region was clearly delineated when myosins were analyzed with the coiled-coil prediction algorithm (Fig. 2). Thus, this algorithm appears to provide a method for locating regions of low stability in extended coiled-coil structures.

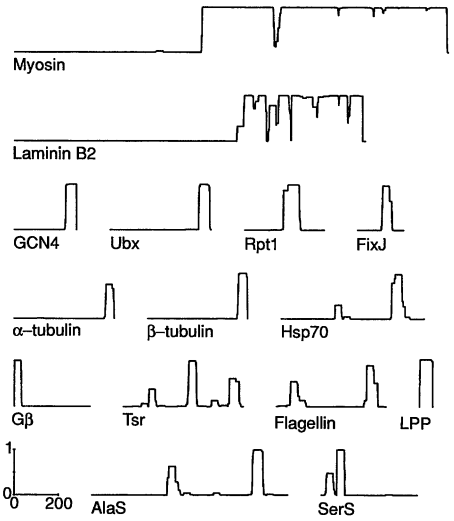

Fig. 2. Probabilities of forming coiled coils for various proteins. The scale at the lower left shows a probability of 1 on the *y*-axis and a segment of 200 residues on the *x*-axis. The GenBank accession number or source for the sequence, the residue numbers for regions with *P(S)* >0.5, and the highest score in these regions are given in parentheses for the following proteins. Rat myosin (X04267 X05004; 840–1933, 2.16, with a discontinuity between 1155 and 1193); myosin was scored with the use of residue frequencies obtained from tropomyosins and keratins only. Mouse laminin B2 (J02930; 1027–1559, 1.89, with discontinuities around 1134 and 1239); this protein forms a three-stranded coiled coil with laminins A and B1. Yeast GCN4 (K02205; 233–(380, 1.91); the leucines of the leucine zipper are at positions 253, 260, 267, and 274. *Drosophila melanogaster* Ubx Ib [(*23*); 343–386, 1.71]; the homeobox domain extends from 295 to 354. Mouse Rptl (J03776; 172–243, 2.20); the zinc finger domain extends from the NH_3-terminus to approximately residue 140. *Rhizobium meliloti* FixJ (J03174; 111–152, 1.56). Human α-tubulin (K00558; 414–447, 1.41). Human β-tubulin (X02344; 403–443, 1.61). Human Hsp70 (M1177 M15432; 496–542, 1.54, with a minor peak at 242–269, 1.31); similar score profiles are obtained for many other members of the Hsp70 family, but in some proteins, such as rat BiP, the relative height of the two peaks is inverted. Human G-protein β-subunit (M16538; 1–31, 1.70). *Escherichia coli* Tsr (J01718; 292–327, 1.38; 476–518, 1.61); the methylated residues are 297, 304, 311, 493, and 503. *Salmonella typhimurium* phase I flagellin (M11332; 63–94, 1.37; 404–451, 1.47). *Escherichia coli* outer membrane lipoprotein (J01645; 1–58, 1.70). *Escherchia coli* alanyl tRNA synthetase (J01581; 346–374, 1.39; 720–766, 1.73); the region identified as essential for oligomerization lies between residues 699 and 808. *Escherichia coli* seryl tRNA synthetase (X05017; 26–53, 1.35; 71–106, 1.90); the antiparallel coiled coil is formed by residues 27–64 and 69–100 (*16*).

We also used the algorithm to examine coiled-coil regions in otherwise globular proteins. Results for the leucine zipper protein GCN4 show a distinct region of high scores that coincides with the leucine zipper domain (Fig. 2). A number of other DNA binding proteins that have been proposed to dimerize by means of a leucine zipper motif (*10*) were also analyzed. Of these, proteins with a basic DNA binding region adjacent to the leucine repeat obtained high scores in the region of the repeat. Other DNA binding proteins containing a leucine heptad repeat, such as the retinoblastoma protein RB (*11*), the homeobox-containing protein Oct2 (*12*), the lupus KU antigen proteins (*13*), and the bacterial transcriptional activator MetR (*14*), had no residues with scores greater than 1.1, corresponding to a probability of forming coiled coils of less than 1%. The occurrence of a leucine heptad repeat is not by itself a reliable indicator of coiled-coil structure (*15*).

A search of the GenBank database for proteins predicted to contain coiled-coil segments yielded more than 200 proteins that contain residues with *P(S)* values higher than 0.99. These included all proteins known to contain coiled coils and no proteins, such as globins or immunoglobulins, that are known not to contain coiled coils. Besides proteins that form double- or triple-stranded parallel coiled coils, the algorithm identified proteins that contain other types of α-helical bundles, however. These include seryl tRNA synthetase (Fig. 2), which forms a double-stranded antiparallel coiled coil (*16*); spectrin, which forms triple-stranded antiparallel bundles (*17*); and the variable surface glycoprotein of trypanosomes, which forms an elongated four-helix bundle

Table 1. Frequency of residue in GenBank and relative occurrence of residues at the seven positions of the coiled-coil heptad repeat.

Residue	Frequency in GenBank (%)	Relative occurrence at position						
		a	b	c	d	e	f	g
Leu	9.33	3.167	0.297	0.398	3.902	0.585	0.501	0.483
Ile	5.35	2.597	0.098	0.345	0.894	0.514	0.471	0.431
Val	6.42	1.665	0.403	0.386	0.949	0.211	0.342	0.360
Met	2.34	2.240	0.370	0.480	1.409	0.541	0.772	0.663
Phe	3.88	0.531	0.076	0.403	0.662	0.189	0.106	0.013
Tyr	3.16	1.417	0.090	0.122	1.659	0.190	0.130	0.155
Gly	7.10	0.045	0.275	0.578	0.216	0.211	0.426	0.156
Ala	7.59	1.297	1.551	1.084	2.612	0.377	1.248	0.877
Lys	5.72	1.375	2.639	1.763	0.191	1.815	1.961	2.795
Arg	5.39	0.659	1.163	1.210	0.031	1.358	1.937	1.798
His	2.25	0.347	0.275	0.679	0.395	0.294	0.579	0.213
Glu	6.10	0.262	3.496	3.108	0.998	5.685	2.494	3.048
Asp	5.03	0.030	2.352	2.268	0.237	0.663	1.620	1.448
Gln	4.27	0.179	2.114	1.778	0.631	2.550	1.578	2.526
Asn	4.25	0.835	1.475	1.534	0.039	1.722	2.456	2.280
Ser	7.28	0.382	0.583	1.052	0.419	0.525	0.916	0.628
Thr	5.97	0.169	0.702	0.955	0.654	0.791	0.843	0.647
Cys	1.86	0.824	0.022	0.308	0.152	0.180	0.156	0.044
Trp	1.41	0.240	0	0	0.456	0.019	0	0
Pro	5.28	0	0.008	0	0.013	0	0	0

(*18*). All these structures consist of long, amphipathic α helices that pack at a 20° angle, thus sharing many structural requirements with coiled-coil helices.

The search revealed several transcriptional activators that do not contain a basic DNA binding region or a leucine zipper but appear to have a zipperlike coiled-coil region. These include the homeobox-containing protein Ubx, the zinc finger protein Rpt1, and the prokaryotic transcriptional activator FixJ (Fig. 2). Ubx and Rpt1 have a topology similar to leucine zipper proteins, with the DNA binding domain followed at its COOH-terminus by the predicted coiled coil, whereas in FixJ the coiled coil precedes the DNA binding COOH-terminal domain (*19*).

In several proteins (Fig. 2), predicted coiled-coil segments lie in areas that are thought to play a functionally important role. For instance, in *Escherichia coli* alanyl tRNA synthetase, a predicted coiled coil lies in a segment that has been identified by deletion analysis as being responsible for oligomerization (*20*); in bacterial chemoreceptors, two predicted coiled-coil regions coincide with the methylated domains that have been implicated in sensory adaptation (*21*); and, in bacterial flagellins, the predicted coiled-coil domains are at the NH_2- and COOH-termini of the proteins in regions that are thought to mediate the polymerization of flagellin into the flagellar filament (*22*). The score profiles presented in Fig. 2 indicate that zipperlike coiled coils occur in proteins that mediate many different types of biological processes.

REFERENCES AND NOTES

1. C. Cohen and D. A. D. Parry, *Trends Biochem. Sci.* **11**, 245 (1986); *Proteins Struct. Funct. Genet.* **7**, 1, (1990).
2. D. A. D. Parry, *Biosci. Rep.* **2**, 1017 (1982).
3. The database was constructed from a translated version of GenBank (release 57, September to November 1988, 2,268,298 residues). Redundant entries were eliminated. Myosin sequences were aligned with nematode myosin [A. D. McLachlan and J. Karn, *J. Mol. Bio.* **164**, 605 (1983)], and the rod sequences were extracted (8444 residues). Keratin sequences were aligned with the 59-kD epidermal keratin from mouse [P. M. Steinert, R. H. Rice, D. R. Roop, B. L. Truss, A. C. Steven, *Nature* **302**, 794 (1983); P. M. Steinert and D. A. D. Parry, *Annu. Rev. Cell Biol.* **1**, 41 (1985)], and the four coiled-coil domains were extracted (5762 residues). All tropomyosin sequences (3293 residues) were extracted. These three groups of proteins were selected because their coiled-coil structure is well established and their sequences from several different organisms are known.
4. E. K. O'Shea, R. Rutkowski, P. S. Kim, *Science* **243**, 538 (1989); T. G. Oas, L. P. McIntosh, E. K. O'Shea, F. W. Dahlquist, P. S. Kim, *Biochemistry* **29**, 2891 (1990).
5. This procedure is an extension of a method proposed by Parry (*2*). We obtained the geometric average by multiplying the frequencies for the 28 residues in the window and taking the 28th root. This average was chosen because it gives high and low frequencies equal weight. In an arithmetic average, the score is dominated by high frequencies.
6. Each score is derived from a 28-residue sequence with a specified heptad repeat frame. For sequences with residues having low scores, corresponding frames change frequently. Within regions with high scores, a continuous frame is generally maintained. No frame changes were observed in tropomyosin, three changes were observed in the myosin rod, and one change was observed in keratin helix 2B. These changes corresponded to the three skip residues predicted in myosin and to the stutter predicted in keratin helix 2B (*3*). Although frame changes are generally accompanied by significant changes in score, this often does not happen when the frame continues unbroken after a skip residue or "stutter." In those cases, changes in frame are the only indicators of local discontinuities.
7. Using the coordinates obtained from the Protein Data Bank (Brookhaven National Laboratories, January 1989) we constructed the database of globular proteins. We eliminated all multiple entries or mutant forms of the same protein and the proteins tropomyosin and influenza hemagglutinin, which contain coiled coils. Our final database contained 150 proteins and 32,588 residues. We used the random number generator of a Phoenix computer to construct the database of random sequences. This database has the same overall amino acid composition as GenBank (Table 1) and contains 52,224 residues.
8. $G(x) = (\sigma \cdot 2\pi)^{-1} e^{-0.5((x-m)/\sigma)2}$ where x = score, m = mean, σ = standard deviation. Justification for approximating score distributions with Gaussian curves is taken from the close fit of the score distribution for random sequences to its Gaussian curve (Fig. 1C).
9. A. McLachlan and J. Karn, *Nature* **299**, 226 (1982); W. F. Harrington and M. E. Rodgers, *Annu. Rev. Biochem.* **53**, 35 (1984).
10. We analyzed the leucine zipper proteins c-Myc, Jun, Fos, and C/EBP [W. H. Landschulz, P. F. Johnson, S. L. McKnight, *Science* **240**, 1759 (1988)], as well as CREB [J. P. Hoeffler, T. E. Meyer, Y. Yun, J. L. Jameson, J. F. Habene, *ibid.* **242**, 1430 (1988)], Cys3 [J. H. Fu, J. V. Paietta, D. G. Mannix, G. A. Marzluf, *Mol. Cell. Biol.* **9**, 1120 (1989)], Opaque2 [H. Hartings *et al.*, *EMBO J.* **8**, 2795 (1989)], Bsg25D (P. D. Boyer, P. A. Mahoney, J. A. Lengyel, *Nucleic Acids Res.* **15**, 2309 (1987)], v-Maf [M. Nishizawa, K. Kataoka, N. Goto, K. T. Fujiwara, S. Kawai, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 7711 (1989)], and HBP1 [T. Tabata *et al.*, *Science* **245**, 965 (1989)]; their high scores ranged between 1.72 (v-Maf) and 1.88 (Opaque2) (in all cases, $P(S) > 0.99$). The region of high scores often included sequences outside the leucine repeat and did not always overlap the entire repeat.
11. F. D. Hong *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 5502 (1989).
12. R. G. Clerc, L. M. Corcoran, J. H. LeBowitz, D. Baltimore, P. A. Sharp, *Genes Dev.* **2**, 1570 (1988).
13. M. Yaneva, J. Wen, A. Ayala, R. Cook, *J. Biol. Chem.* **264**, 13407 (1989).
14. M. E. Maxon, J. Wigboldus, N. Brot, H. Weissbach, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 7076 (1990).
15. To evaluate the reliability of coiled-coil prediction from a fourfold leucine heptad repeat, we extracted all the proteins from GenBank that contained this motif. After eliminating redundant entries we obtained 194 proteins, only 70 of which had $P(S)$ values larger than 0.5.
16. S. Cusack, C. Berthet-Colominas, M. Haertlein, N. Nassar, R. Leberman, *Nature* **347**, 249 (1990).
17. D. W. Speicher and V. T. Marchesi, *ibid.* **311**, 177 (1984).
18. D. Freymann *et al.*, *FASEB J.* **4**, A2166 (1990); D. Freymann, P. Metclaf, M. Turner, D. C. Wiley, *Nature* **311**, 167 (1984).
19. M. David *et al.*, *Cell* **54**, 671 (1988).
20. M. Jasin, L. Regan, P. Schimmel, *ibid.* **36**, 1089 (1984).
21. J. Stock, G. Lukat, A. Stock, *Annu. Rev. Biophys. Chem.* **20**, 109 (1991).
22. O. V. Fedorov, A. S. Kostyukova, M. G. Pyatibratov, *FEBS Lett.* **241**, 145 (1988); G. Kuwajima, *J. Bacteriol.* **170**, 3305 (1988).
23. K. Kornfeld *et al.*, *Genes Dev.* **3**, 243 (1989).
24. A VAX Pascal program implementing the described algorithm is available from the authors upon request. We thank J. M. Lupas for help with the manuscript and D. Welsh for computer assistance. Supported by NIH grant AI20980.

8 August 1990; accepted 13 February 1991

Self-Assembled Organic Monolayers: Model Systems for Studying Adsorption of Proteins at Surfaces

KEVIN L. PRIME AND GEORGE M. WHITESIDES*

Self-assembled monolayers (SAMs) of ω-functionalized long-chain alkanethiolates on gold films are excellent model systems with which to study the interactions of proteins with organic surfaces. Monolayers containing mixtures of hydrophobic (methyl-terminated) and hydrophilic [hydroxyl-, maltose-, and hexa(ethylene glycol)–terminated] alkanethiols can be tailored to select specific degrees of adsorption: the amount of protein adsorbed varies monotonically with the composition of the monolayer. The hexa(ethylene glycol)–terminated SAMs are the most effective in resisting protein adsorption. The ability to create interfaces with similar structures and well-defined compositions should make it possible to test hypotheses concerning protein adsorption.

UNDERSTANDING THE MECHANISM of protein adsorption at surfaces (*1, 2*) is an important element of research in protein chromatography (*3*), clinical diagnostics (*4*), biomedical materials (*5*), and cellular adhesion (*6*). No system is available that permits the structure and properties of the interface to be controlled in detail sufficient for the investigation of hypotheses concerning protein adsorption at the molecular level. We report a study of protein adsorption at interfaces between SAMs and aqueous buffer solutions. The results indicate that the organic interfaces prepared by the self-assembly of long-chain

Department of Chemistry, Harvard University, Cambridge, MA 02138.

*To whom correspondence should be addressed.