Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 8259–8263, August 1995
Biochemistry

Predicting coiled coils by use of pairwise residue correlations

BONNIE BERGER*, DAVID B. WILSON*, ETHAN WOLF*, THEODORE TONCHEV*, MARIA MILLA†, AND PETER S. KIM†

*Mathematics Department and Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139; and †Howard Hughes Medical Institute, Whitehead Institute, Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142

Communicated by Gian-Carlo Rota, Massachusetts Institute of Technology, Cambridge, MA, February 24, 1995

ABSTRACT **A method is presented that predicts coiled-coil domains in protein sequences by using pairwise residue correlations obtained from a (two-stranded) coiled-coil database of 58,217 amino acid residues. A program called PAIRCOIL implements this method and is significantly better than existing methods at distinguishing coiled coils from α-helices that are not coiled coils. The database of pairwise residue correlations suggests structural features that stabilize or destabilize coiled coils.**

The two-stranded, parallel coiled-coil motif consists of two right-handed α-helices wrapped around each other with a slight left-handed superhelical twist. Coiled coils have traditionally been associated with fibrous proteins such as keratin, myosin, and tropomyosin (reviewed in ref. 1). They attracted particular attention when the "leucine zipper" motif (2), found in several DNA-binding proteins, was shown to correspond to a short coiled coil (3, 4).

The simple, repeating units of structure in coiled coils makes them particularly amenable to computer-based recognition methods. Traditionally, coiled coils have been identified by the occurrence of hydrophobic residues spaced every four and then three residues apart. This pattern defines a heptad repeat, $(\mathbf{abcdefg})_n$, in which generally hydrophobic residues occur at positions **a** and **d**. The interaction between two α-helices in a coiled coil involves these hydrophobic residues, as well as the predominantly charged residues at the **e** and **g** positions (4).

Parry (5) proposed a method for using heptad-repeat positional information to identify coiled coils from protein sequences. The frequency of each of the 20 amino acid residue types in each of the seven heptad-repeat positions was compiled in a 20 × 7 table from a database of known coiled coils. These table entries were incorporated into residue scores to predict coiled-coil domains. A residue score was computed as follows. Each window of 28 residues was given a score in which each residue's probability in the window was multiplied (5, 6), and the 28th root was taken (6). A residue score was taken to be the maximum window score over all windows containing the residue and over all heptad-repeat registers. A residue's score was related to the likelihood of it being in a coiled coil. Lupas *et al.* (6) normalized the residue probabilities by the probability of the corresponding residue occurring in Genpept (the protein sequence database translated from GenBank) and wrote a computer program to identify coiled coils, called NEWCOILS, based on Parry's algorithm. This program has been useful in identifying coiled-coil candidates, including a region in the influenza virus hemagglutinin that is thought to act in a "spring-loaded" manner when the virus infects cells (7, 41).

Although Parry's algorithm and the NEWCOILS program are quite successful, this approach leads to a significant number of "false positives." For example, because lysine residues are frequently found at all positions in coiled coils, the polylysine sequence $(\text{Lys-Lys-Lys})_n$ scores highly even though it is not a coiled coil. Moreover, when tested with the Brookhaven x-ray crystal structure database, a significant fraction of the sequences predicted to form coiled coils with this approach are known to fold into structures that are not coiled coils.

Here we propose an extension of Parry's algorithm that takes advantage of pairwise residue correlations in known coiled coils. Pairwise correlation analysis has traditionally been useful for breaking codes (8, 9). Pairwise correlations have been identified for amino acids in proteins that are physically close (10, 11) and, more specifically, have been identified in zinc fingers (12), α-helices (13), β-sheets (14, 15), and α-helix capping interactions (16). Recently, pairwise mutation correlations have been used to infer spatial proximity between residues (17) and to analyze divergent evolution of protein sequences (18).

The PAIRCOIL program,‡ which implements the pairwise correlation method presented here, does not produce any obvious false positives or false negatives when tested with the Brookhaven database. The program is particularly useful for eliminating false positives (most commonly, amphipathic α-helices that are not coiled coils) and in turn clarifying many true positives. Examples of proteins with sequences that score well with NEWCOILS but that would appear to be false positive predictions, based on a corresponding low score with the PAIRCOIL program, include the α subunit of a heterotrimeric guanine nucleotide-binding protein (G protein), α- and β-tubulin, α-farnesyltransferase, and transcription factor NF-κB.

METHODS

The Algorithm. A database of known coiled-coil sequences from myosins, tropomyosins, and intermediate filament (IF) proteins was generated (see *The Databases*). Each sequence in the database was used to tabulate the frequency of occurrence of each pair of amino acids at each pair of positions in the heptad repeat.

The pairwise frequency values in the coiled-coil database are used to estimate the probability that a given residue pair exists in a given pair of heptad-repeat positions in a coiled coil (see *Estimating Probabilities*). To obtain normalized pair probabilities, the coiled-coil probabilities for each residue pair in each pair of heptad-repeat positions distance i apart are divided by the corresponding distance-i probabilities for sequences in Genpept. Normalized single probabilities are computed in a similar manner.

These normalized probabilities are used to compute a score S_k for the kth residue, which corresponds to the likelihood that this residue is in a coiled coil. To compute a residue score S_k, the maximum window score over all 30-residue windows containing the kth residue is taken. A window score is the maximum over the seven possible heptad-repeat positions of the sum of the residue propensities in the window. A residue propensity for a given heptad-repeat position incorporates the correlations between that residue and the residues that follow

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: PDB, Brookhaven Protein Data Bank; PIR, Protein Identification Resource. IF, intermediate filament.

‡The PAIRCOIL program is available upon request: e-mail at paircoil@theory.lcs.mit.edu.

at structurally relevant distances $i = 1$, $i = 2$, and $i = 4$. For normalized probabilities P, the propensity of the kth residue is

$$\frac{1}{3} \ln \frac{P(k, k+1)P(k, k+2)P(k, k+4)}{P(k+1)P(k+2)P(k+4)}.$$

In other words, the product of the normalized pair probabilities of residue pairs $(k, k+1)$, $(k, k+2)$, and $(k, k+4)$ is divided by the product of the normalized single probabilities of residues $k+1$, $k+2$, and $k+4$; the residue's propensity is a third of the logarithm of this quantity, which is equivalent to taking the geometric mean of the three quantities $\ln \frac{P(k, k+i)}{P(k+i)}$. If residue $k+i$ in residue pair $(k, k+i)$ does not exist, then rather than multiplying by the normalized pair probability of $(k, k+i)$ and dividing by the normalized single probability of residue $k+i$, the geometric mean of the other n existing pairs is taken (so the fraction out front is $1/n$). If no pairs exist, then the propensity is the logarithm of the normalized single probability of residue k. PAIRCOIL scores windows of length at least 28 when a 30-position window containing the given residue does not exist. Windows containing a proline or an unknown or unusual residue are not scored. [Prolines can occur within coiled coils but are nearly always in the first turn of the α-helix (19).]

PAIRCOIL uses the dynamic programming algorithm in ref. 20 to quickly produce the same result as the algorithm described above: a run through the entire Genpept sequence database takes approximately 15 min on a Sun SPARC 10 computer. The proposed method used in PAIRCOIL reduces to the previous method used in NEWCOILS when there are no pairwise dependencies between the amino acids in the window. To convert a NEWCOILS score to a corresponding PAIRCOIL score, take the logarithm and multiply by 28. Mathematical motivation for the proposed method is given in ref. 20. The distances $i = 1$, $i = 2$, and $i = 4$ were chosen empirically. It is not surprising that correlations between residues distances 1 and 4 apart were found to be useful, given that coiled coils are amphipathic helices with a repeating structure of 3.5 residues per turn. The algorithm used in PAIRCOIL to compute residue propensities uses the geometric mean over these distances on the basis that it is unlikely that a non-coiled-coil sequence would score well using all three distances. A window length of 30 residues was also chosen empirically: this window length is consistent with the finding that short, stable coiled coils are approximately four heptads long (3, 21, 22).

The Databases. The coiled-coil database was constructed from Genpept [a translated version of GenBank (release 73, September 1992)] and the Protein Identification Resource (PIR, release 34, September 1992). Redundant and unverified entries were eliminated. Alignments were done with an algorithm similar to that described previously (23), and the coiled-coil domains were extracted. Sequences that appeared to include errors or to align poorly were excluded from the database.

Myosin sequences were aligned with nematode myosin (MWKW) (24–27). The first and last 14 residues of the rod region were not included in the database because of uncertainties as to the coiled-coil boundaries. The myosin hinge region (residues 1161–1177) was excluded. The 10 residues before and 11 residues after each skip, insertion, or deletion were not included because of uncertainties as to where the skips were. Paramyosin has an extra skip, so the corresponding region in the myosins was not included (28, 29). The regions included in the database align with nematode residues 864–1160, commencing in **d**; 1212–1386, **a**; 1409–1583, **a**; 1606–1808, **a**; and 1831–1929, **a**.

Tropomyosin sequences were aligned with the rabbit skeletal muscle tropomyosin (TMRBA) (30); the horse platelet tropomyosin and similar sequences were aligned with the rabbit sequence (31). The first and last 14 residues were not included because they are not helical; residues 183–196 were excluded because this region has an excess of negative charges which could cause electrostatic repulsion (32, 33, 40). The regions included in the database align with rabbit residues 15–182, **a**; and 196–269, **a**.

IF proteins [i.e., keratin, vimentin, desmin, glial fibrillary acidic protein, neurofilaments, and lamins (34)] were aligned as in ref. 19. The first and last 14 residues of coiled-coil regions were not included because of uncertainties as to the coiled-coil boundaries. The 2A segment was not included because it was too short. The 1A segment and the last four heptads of the 2B segment of all IF proteins were not included because they are conserved segments involved in higher-order formation (35). Residues 54–67 of the 2B segment were not included because this region contained a stutter (36). The 2B segment of the lamins was also excluded. For both the tropomyosins and IF proteins, the 7 residues before and after each insertion and deletion were not included. Regions that were less than 28 residues long were also excluded.

The PDB-minus database was constructed from the Brookhaven Protein Data Bank (PDB, February 1994), with known coiled coils excluded and entries with high sequence similarity removed as follows. The PILEUP (37) multiple-alignment program was run on the PDB and a relational tree was generated. The number of protein "classes" in the PDB was reduced to 286, and one sequence was chosen from each class for PDB-minus.

The PIR-minus database was constructed from the PIR (release 38.09, 1994), with the myosins, tropomyosins, and IF proteins removed.

Estimating Probabilities. Probabilities for residue pair A and B in positions k and $k+i$ in the database are computed as follows. Let $f(A, B)$ be the number of times A and B occur in positions k and $k+i$ in the set of T such positions in the database. The probability of residue A in heptad-repeat position k is initially set to $P(A) = f_k(A)/T_k$, where $f_k(A)$ is the number of times A occurs in position k in the T_k such positions in the coiled-coil database. If $f_k(A) = 0$, then to compensate for the limited size of the positive database, the zero probability is adjusted to $P(A) = \min\{1/(5T_k), 1/20\}$. Thus, for the zero-frequency residues, $P(A)$ can be no greater than 1/5 of the probability of any nonzero frequency residue. The value 1/5 was chosen empirically. The 1/20 upper bound arises because 1/20 would be the probability if each residue were equally likely at a given position. $P(A)$ values are then normalized by the total probability mass to obtain a probability function. The pairwise probability of A and B in positions k and $k+i$ is computed similarly, except that, prior to normalization, the zero probabilities are updated to be $P(A, B) = \min\{1/(5T), 1/400, P(A) \cdot P(B)\}$, where T is the number of position pairs $(k, k+i)$ in the database.

RESULTS

There are striking pairwise residue correlations in the coiled-coil database (Fig. 1). For example, a negative correlation is evident for L_aL_d (Leu at position **a** followed by Leu at **d**), a positive correlation for I_aL_d, and a less strong positive correlation for V_aL_d. Indeed, it was found experimentally for the GCN4 leucine zipper (38) that when there is a leucine residue in the **d** register, I_a favors two-stranded coiled coils, L_a favors three-stranded coiled coils, and V_a is associated with both.

The improvement of the proposed method over the previous method is evident in Fig. 2, which shows histograms of residue scores as computed by NEWCOILS and PAIRCOIL. For PAIRCOIL, a particular coiled-coil test protein was excluded from the

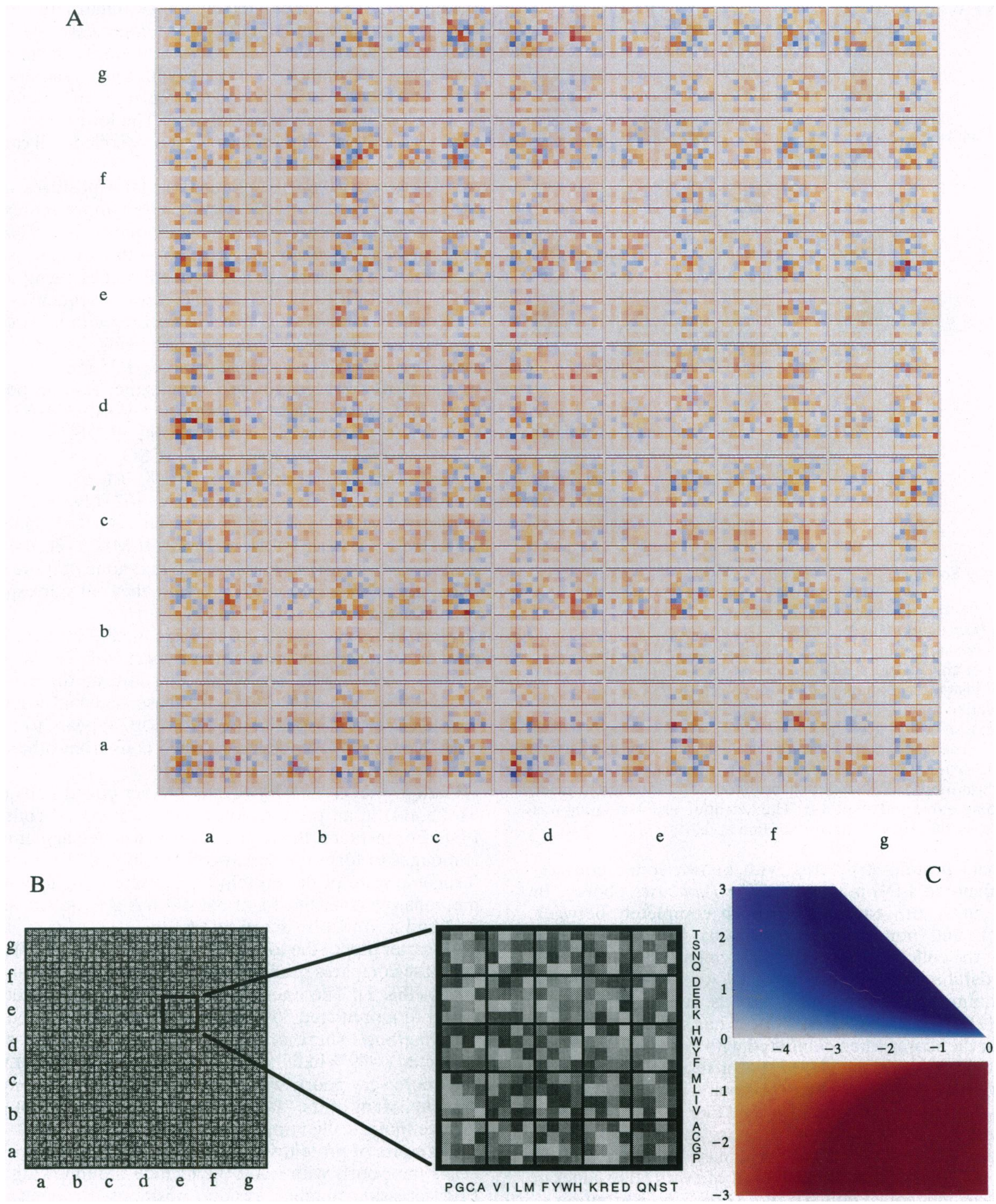

FIG. 1. Pairwise correlations of amino acid residues in the two-stranded coiled-coil database. (*A*) Data are organized into a 7 × 7 array of heptad-repeat correlation boxes. Each correlation box is a 20 × 20 array of squares, where each square represents a particular residue pair (see *B*). The color of the square indicates the correlation: negative correlations are shown in red, positive correlations in blue, and zero correlation in white. Less significant correlations, where the product of the single residue frequencies is small, are shown in paler colors (see *C*). For example, there is a positive correlation when Ile is at the **a** position followed by Leu at **d** (I_aL_d). Interesting pairwise correlations stand out as intense colorful regions. Examples include L_aL_d, K_bK_f, and E_gE_e. (*B*) The boxes in *A* are arranged so that the register of the first residue (closest to the amino terminus) determines the *x* coordinate, and the register of the second residue determines the *y* coordinate. (If the two registers are the same, then the residues are distance 7 apart.) Four horizontal and vertical reference lines are drawn on the box to make it easier to identify the amino acid pair of each square. (*C*) How to interpret the colors in *A*. The *x* axis is the logarithm of the product of the single residue frequencies, and the *y* axis is the correlation (i.e., the logarithm of the corresponding pair frequency divided by this product). The color of a square is determined by these two values as in this picture. Thus, a particular pair correlation is judged to be significant (i.e., dark) only if there are sufficient data to yield a high product of the single frequencies in that pair. The upper right region of the picture is blank because this corresponds to an impossibly high pair frequency.

coiled-coil database at the time it was scored. For NEWCOILS, this was not possible, since scoring was done with the table given in ref. 6.

There is no overlap in the scores computed by PAIRCOIL when the histogram scores for the coiled-coil database are plotted against those for the PDB database of three-

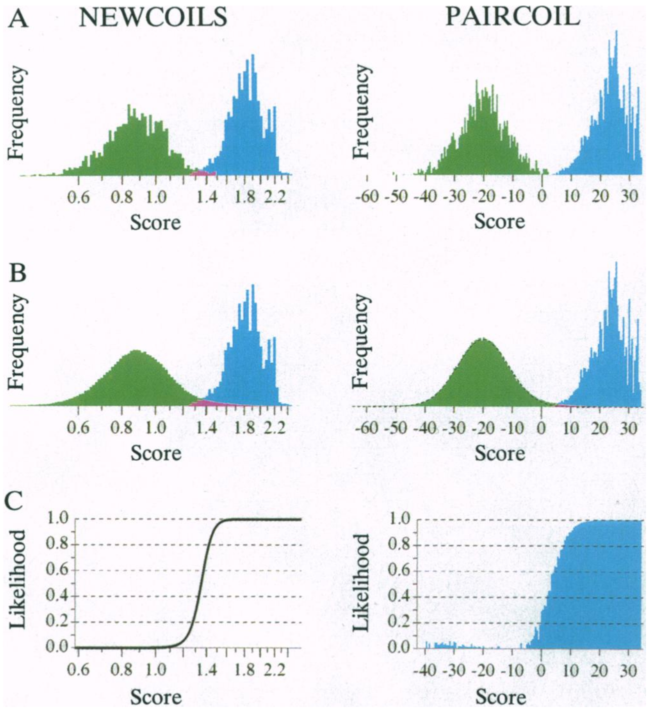

FIG. 2. (*A* and *B*) Histograms of residue scores as computed by NEWCOILS (*Left*) and PAIRCOIL (*Right*). The coiled-coil scores (58,217 residues) were superimposed on the scores of the PDB-minus (63,116 residues) (*A*) and PIR-minus (7,322,501 residues) (*B*) databases. The coiled-coil histogram is in blue, PDB-minus and PIR-minus in green, and the overlap in pink. (*C*) Estimates of the likelihood that a residue with a given NEWCOILS score (6) or PAIRCOIL score (see text) is in a coiled coil. The histograms of scores in *Left* were calculated on a logarithmic scale to be consistent with the histograms in *Right*. NEWCOILS was modified to score only proline-free windows to be more directly comparable with PAIRCOIL. The height of each histogram was normalized so that they all have the same area.

dimensional protein structures, with known coiled coils removed, denoted PDB-minus (see *The Databases* above). In contrast, NEWCOILS cannot distinguish completely between coiled-coil and non-coiled-coil domains (Fig. 2*A*). When scores for the coiled-coil database are compared with those for the PIR database of protein sequences, with entries removed for the common two-stranded coiled coils, denoted PIR-minus (see *The Databases* above), NEWCOILS has significantly more overlap in the scores when compared with PAIRCOIL (Fig. 2*B*).

An estimate of the likelihood that a residue with a given PAIRCOIL score is in a coiled coil was made by noting that the PIR and PIR-minus histograms looked like a Gaussian distribution with some extra probability mass added on the right-hand tail (Fig. 2*B Right*). This extra mass was attributed to coiled coils. It was estimated that 1 out of every 50 residues in the PIR was in a coiled coil. To fit a Gaussian distribution to the histogram data, the mean was calculated so that the extra probability mass to the right of the mean would correspond to 1/50 of the total mass of the PIR. The standard deviation was then computed by using only scores below that mean, where a Gaussian distribution better fits the histogram data. The histograms for the PIR-minus and PIR correspond well at values below this mean. The likelihood $l(x)$ that a residue with a given PAIRCOIL score x is in a coiled coil was estimated as the ratio of the extra histogram mass above the Gaussian at that score to the total histogram mass at that score (Fig. 2*C Right*). A least-squares fit line was then used to approximate the likelihood data in the linear region from 10% to 90%. The percent likelihood for a given PAIRCOIL score x can be calculated with the equation $l(x) = 7.45x + 25.84$, where $x \in [-2, 9]$. A PAIRCOIL score of 3.24 corresponds to 50% likelihood.

Lupas *et al.* (6) computed likelihood estimates for NEWCOILS (Fig. 2*C Left*) by approximating the residue score distributions of coiled coils and non-coiled coils with Gaussian curves. The probability of forming a coiled coil with a given score was then calculated by using an assumed 1:30 ratio of coiled-coil to non-coiled-coil residues in Genpept. The lowest score associated with a region identified to be a new coiled-coil candidate in ref. 6 was 1.35 (50% likelihood).

PAIRCOIL is useful for eliminating false positives in NEWCOILS. The $(\text{Lys-Lys-Lys})_n$ example cited above scores highly with NEWCOILS (score = 1.43, corresponding to >82% likelihood of being a coiled coil) but low with PAIRCOIL (score = −5.47, corresponding to <10% likelihood of being a coiled coil). There are ≥14 out of 286 distinct sequences in the PDB-minus database that have high scores with NEWCOILS but low scores with PAIRCOIL. The following lists all protein sequences with a NEWCOILS score over 1.35 (50% likelihood) in the PDB-minus database (PDB name, residue positions, NEWCOILS score, PAIRCOIL score): 1ADA, 304–331, 1.35, −11.32; 2TS1, 290–317, 1.40, −5.74; 9LDB, 40–67, 1.45, −3.18; 1YPI, 137–164, 1.47, −4.29; 1CSG, 14–41, 1.36, −4.35; 1EMD, 191–218, 1.39, −3.46; 1FLX, 41–68, 1.37, −5.53; 1APK, 53–80, 1.47, −4.52; 1GPA, 102–129, 1.42, −1.41; 3BLM, 99–126, 1.37, 0.73; 1LE2, 33–60, 1.49, 0.81; 256B, 1–28, 1.41, 0.43; 2HPD, 197–224, 1.36, 1.0; 1LMB, 9–36, 1.47, 1.26. The x-ray crystal structures indicate that none of these regions correspond to coiled coils although they all correspond to α-helical regions.

Thus, two-thirds of the structures in the PDB that NEWCOILS predicts to form coiled coils do not form coiled coils (14 false positives and 6 true positives). In contrast, there were no sequences in the PDB-minus database that had a PAIRCOIL score ≥3.24 (50% likelihood). PAIRCOIL appears to be especially useful for distinguishing coiled coils from other α-helices.

Both PAIRCOIL and NEWCOILS predict coiled coils (≥50% likelihood) in all proteins known to form coiled coils in the PDB. For instance, the region of influenza hemagglutinin that is thought to form the coiled-coil "spring" that results in the fusogenic state of the protein (7, 41) scores highly with both methods (residues 54–81 of 3HMG has a PAIRCOIL score of 9.43 and a NEWCOILS score of 1.62).

A scatter plot of the top-scoring sequences in the PIR-minus database compares predictions made with PAIRCOIL and NEWCOILS (Fig. 3). The lines in Fig. 3 indicate scores that correspond to a predicted 50% likelihood of being a coiled coil in each method. There are many sequences that are strongly predicted (>99% likelihood) to be coiled coils with NEWCOILS that score very poorly with PAIRCOIL, whereas the converse is not true. Many of the "false positive" predictions of NEWCOILS involve biologically important proteins.

Examples of protein sequences that score highly with NEWCOILS but poorly with PAIRCOIL include the following (name, PIR accession number, residue positions, NEWCOILS score, PAIRCOIL score): human α-tubulin, A23035, 414–441, 1.41, −4.79; human β-tubulin, A26561, 407–434, 1.62, −0.54; human G-protein α subunit $G_s\alpha$, RGHUA2, 10–37, 1.44, −6.41; *Bacillus subtilis* threonyl-tRNA synthetase, YSBST1, 227–254, 1.47, −5.02; herpes simplex virus UL14 protein, WMBE21, 97–124, 1.62, −4.98; *Escherichia coli* SelB, EFECSB, 116–143, 1.60, −7.67; bovine ephemeral fever virus glycoprotein G, VGVNBE, 463–490, 1.55, −3.78; bovine α-farnesyltransferase, A41013, 284–311, 1.52, −2.49; *B. subtilis* FliM, B39136, 7–34, 1.56, −5.53; rat nucleolin, JH0148, 240–267, 1.54, −7.44; human α-prothymosin, TNHUA, 40–67, 1.52, −4.52; chicken acetylcholine receptor α_2, ACCH2N, 385–412, 1.49, −11.56; human transcription factor TFIID, A34830, 52–79, 1.53, −11.72; human NF-κB, A41645, 167–194, 1.43, −6.87; human papillomavirus E2 protein, W2WLE, 1–28, 1.40, −9.95. It

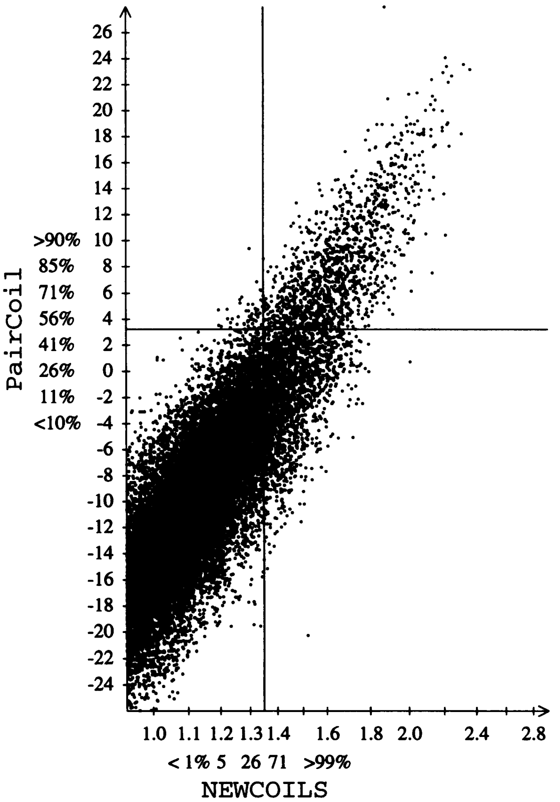

FIG. 3. Scatter plot of protein scores (the maximum score in a particular protein) for each sequence in the PIR-minus database, when scored by NEWCOILS (*x* axis) and PAIRCOIL (*y* axis). All points above and to the right of the axes are shown. Both axes are labeled with a score and its corresponding likelihood. A horizontal line is drawn at a PAIRCOIL score of 3.24 (50% likelihood), which is greater than the score of any non-coiled coil in the PDB. A vertical line is drawn at a NEWCOILS score of 1.35 (50% likelihood). Many sequences that score highly for NEWCOILS are quite low-scoring for PAIRCOIL, but the converse is not true.

seems likely that most of these sequences do not form coiled coils although they may form α-helices.

DISCUSSION

Our results indicate that there are pairwise correlations in coiled coils and that these correlations can help distinguish two-stranded coiled-coil from non-coiled-coil domains. It is possible that there are sequence features of the two-stranded coiled coils in our database that are not general features of coiled coils. The relative success, however, of coiled-coil prediction methods that utilize databases based on these known coiled-coil sequences (5, 6) suggests that inherent biases are not great.

Indeed, it seems likely that data of the type in Fig. 1 will provide new insights into the types of interactions that stabilize or destabilize coiled coils in general. For example, there are some interesting "asymmetries" in the database for two-stranded coiled coils (Fig. 1): L_aL_d is negatively correlated whereas L_dL_a is not correlated, and L_aE_d is positively correlated whereas E_dL_a is negatively correlated. These observations do not reflect true asymmetries, since position **a** followed by **d** is a (k, $k + 3$) spacing of residues whereas position **d** followed by **a** is a (k, $k + 4$) spacing.

Some of the apparent asymmetries in Fig. 1 can be rationalized. Examples of these include correlations between charged residues. Whereas E_bK_f and E_bR_f are positively correlated, K_fE_b and R_fE_b are negatively correlated or not correlated, respectively. These observations are consistent with previous observations (13) that oppositely charged residues in α-helices of proteins are frequently found with a (k, $k + 4$) spacing but not a (k, $k + 3$) spacing.

The methods described here can be applied readily to three-stranded coiled coils with a new database. It also seems likely that these pairwise-correlation methods can enhance other structure prediction methods, such as profile methods that classify amino acid residues by hydrophobicity, size, and solvent accessibility (see e.g., ref. 39).

We thank Russell Schwartz, David Krikorian, and Doug Muir for computational assistance; Dan Kleitman, Mona Singh, Chave Carr, and Tom Alber for numerous discussions; and A. Lupas and J. Stock for sharing their original database. D.B.W. and E.W. are Department of Defense graduate fellows. This research was supported by the National Institutes of Health (GM44162) and the Massachusetts Institute of Technology–State Street Bank Science Partnership Fund.

1. Cohen, C. & Parry, D. A. D. (1990) *Proteins Struct. Funct. Genet.* **7,** 1–15.
2. Landschulz, W. H., Johnson, P. F. & McKnight, S. L. (1989) *Science* **243,** 1681–1688.
3. O'Shea, E. K., Rutkowski, R. & Kim, P. S. (1989) *Science* **243,** 538–542.
4. O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. (1991) *Science* **254,** 539–544.
5. Parry, D. A. D. (1982) *Biosci. Rep.* **2,** 1017–1024.
6. Lupas, A., van Dyke, M. & Stock, J. (1991) *Science* **252,** 1162–1164.
7. Carr, C. M. & Kim, P. S. (1993) *Cell* **73,** 823–832.
8. Shannon, C. E. (1949) *Bell Syst. Tech. J.* **28,** 656–715.
9. Andreassen, K. (1988) *Computer Cryptology: Beyond Decoder Rings* (Prentice-Hall, Englewood Cliffs, NJ).
10. Sippl, M. J. (1990) *J. Mol. Biol.* **213,** 859–883.
11. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Nature (London)* **358,** 86–89.
12. Desjarlais, J. R. & Berg, J. M. (1992) *Proteins Struct. Funct. Genet.* **12,** 101–104.
13. Maxfield, F. R. & Scheraga, H. A. (1975) *Macromolecules* **8,** 491–493.
14. von Heijne, G. & Blomberg, C. (1978) *Biopolymers* **17,** 2033–2037.
15. Lifson, S. & Sander, C. (1980) *J. Mol. Biol.* **139,** 627–639.
16. Harper, E. T. & Rose, G. D. (1993) *Biochemistry* **32,** 7605–7609.
17. Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994) *Proteins Struct. Funct. Genet.* **18,** 309–317.
18. Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1994) *Biochem. Biophys. Res. Commun.* **199,** 489–496.
19. Conway, J. F. & Parry, D. A. D. (1988) *Int. J. Biol. Macromol.* **10,** 79–98.
20. Berger, B. (1995) *J. Computat. Biol.* **2,** 125–138.
21. Lumb, K. J., Carr, C. M. & Kim, P. S. (1994) *Biochemistry* **33,** 7361–7367.
22. Lau, S. Y. M., Taneja, A. K. & Hodges, R. S. (1984) *J. Biol. Chem.* **259,** 13253–13261.
23. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
24. McLachlan, A. D. & Karn, J. (1982) *Nature (London)* **299,** 226–231.
25. McLachlan, A. D. & Karn, J. (1983) *J. Mol. Biol.* **164,** 605–626.
26. Lu, R. C. & Wong, A. (1985) *J. Biol. Chem.* **260,** 3456–3461.
27. Dibb, N. J., Maruyama, I. N., Krause, M. & Karn, J. (1989) *J. Mol. Biol.* **205,** 603–613.
28. Kagawa, H., Gengyo, K., McLachlan, A. D., Brenner, S. & Karn, J. (1989) *J. Mol. Biol.* **207,** 311–333.
29. Laclette, J. P., Landa, A., Arcos, L., Willms, K., Davis, A. E. & Shoemaker, C. B. (1991) *Mol. Biochem. Parasitol.* **44,** 287–296.
30. Stone, D. & Smillie, L. B. (1978) *J. Biol. Chem.* **253,** 1137–1148.
31. Lau, S. Y. M., Sanders, C. & Smillie, L. B. (1985) *J. Biol. Chem.* **260,** 7257–7263.
32. Talbot, J. A. & Hodges, R. S. (1982) *Biochemistry* **15,** 225–230.
33. Zot, A. S. & Potter, J. D. (1987) *Annu. Rev. Biophys. Chem.* **16,** 535–559.
34. Steinert, P. M. & Roop, D. R. (1988) *Annu. Rev. Biochem.* **57,** 593–625.
35. Steinert, P. M., Marekov, L. N., Fraser, R. D. B. & Parry, D. A. D. (1993) *J. Mol. Biol.* **230,** 436–452.
36. Dowling, L. M., Crewther, W. G. & Parry, D. (1986) *J. Biochem.* **236,** 705–712.
37. GGC (1994) *Program Manual for the Wisconsin Package* (Univ. of Wisconsin, Madison, WI).
38. Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. (1993) *Science* **262,** 1401–1407.
39. Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991) *Science* **253,** 164–170.
40. O'Shea, E. K., Rutkowski, R. & Kim, P. S. (1992) *Cell* **68,** 699–708.
41. Bullough, P. A., Hughson, F. M., Skehel, J. J. & Wiley, D. C. (1994) *Nature (London)* **371,** 37–43.