Protein Science (1997), 6:1179–1189. Cambridge University Press. Printed in the USA.

MultiCoil: A program for predicting two- and three-stranded coiled coils

ETHAN WOLF,[1,2] PETER S. KIM,[1] AND BONNIE BERGER[2]

[1]Howard Hughes Medical Institute, Whitehead Institute, MIT, 9 Cambridge Center, Cambridge, Massachusetts 02142
[2]Mathematics Department and Laboratory for Computer Science, MIT, Cambridge, Massachusetts 02139

(RECEIVED December 26, 1996; ACCEPTED March 5, 1997)

Abstract

A new multidimensional scoring approach for identifying and distinguishing trimeric and dimeric coiled coils is implemented in the MultiCoil program. The program extends the two-stranded coiled-coil prediction program PairCoil to the identification of three-stranded coiled coils. The computations are based upon data gathered from a three-stranded coiled-coil database comprising 6,319 amino acid residues, as well as from the previously constructed two-stranded coiled-coil database. In addition to identifying coiled coils not predicted by the two-stranded database programs, MultiCoil accurately classifies the oligomerization states of known dimeric and trimeric coiled coils. Analysis of the MultiCoil scores provides insight into structural features of coiled coils, and yields estimates that 0.9% of all protein residues form three-stranded coiled coils and that 1.5% form two-stranded coiled coils. The MultiCoil program is available at http://theory.lcs.mit.edu/multicoil.

Keywords: coiled coil; protein folding; statistical prediction

The coiled-coil motif is composed of right-handed α-helices wrapped around each other with a slight left-handed superhelical twist. Homo-oligomers form from monomers of the same α-helical sequence, whereas hetero-oligomers form from distinct α-helical monomers. Two-stranded, three-stranded, and four-stranded versions of coiled coils are all possible (Cohen & Parry, 1990; Harbury et al., 1993). Two-stranded coiled coils have drawn particular interest because of the "leucine-zipper" motif (Landshulz et al., 1989), found in several DNA-binding proteins. Three-stranded coiled coils have been identified in influenza hemagglutinin and the envelope proteins of Moloney murine leukemia virus and HIV-1, and are thought to play a role in membrane fusion (Carr & Kim, 1993; Bullough et al., 1994; Fass et al., 1996; Chan et al. 1997).

Coiled-coil motifs are particularly amenable to computer-based prediction schemes because of the characteristic repeating pattern of hydrophobic residues spaced every four and then three residues apart. This pattern forms a heptad repeat $(\mathbf{abcdefg})_n$ of amino acids in which positions **a** and **d** tend to be hydrophobic and positions **e** and **g** are predominantly charged residues. The interactions between amino acids at these heptad positions are essential to the formation of the coiled-coil structure. The specificity of these interactions and the amount of packing space required for the residues at each position vary depending on the oligomerization state of the coiled coil (Harbury et al., 1993, 1994, 1995).

Reprint requests to: Peter S. Kim, Howard Hughes Medical Institute, Whitehead Institute, MIT, 9 Cambridge Center, Cambridge, Massachusetts 02142; e-mail: doka@wi.mit.edu.

Because of the regular structure of coiled coils, statistics-based prediction programs have been successful in identifying coiled coils (Lupas et al., 1991; Berger et al., 1995), taking advantage of trends of amino acids that occur preferentially in particular heptad-repeat positions in a database of known coiled coils. The NEWCOILS and COILS programs of Lupas et al. predict regions likely to form coiled coils, based upon a scheme suggested by Parry (1982). This scheme calculates the frequency of each amino acid in each of the seven heptad-repeat positions from the database. The "single probabilities" derived from these frequencies provide the basis for the scores computed by the program. The NEWCOILS program has been quite successful, but has been shown to produce "false positives" by classifying noncoiled-coil α-helical regions incorrectly to be coiled coils (Berger et al., 1995). The program PairCoil of Berger et al. successfully predicts coiled coils, while significantly reducing the number of false positives by using a scoring method based upon "pairwise probabilities." Pairwise probabilities are computed from the frequencies of each pair of amino acids in each pair of coiled-coil heptad-repeat positions in the database. The method takes advantage of correlations and anticorrelations between pairs of residues in coiled coils.

These prediction programs have typically been based on databases consisting of two-stranded coiled coils. In this paper, we describe the construction of a three-stranded coiled-coil database. The MultiCoil program introduced here uses this three-stranded database and a previously constructed two-stranded database (Berger et al., 1995) to extend the PairCoil program for identifying and distinguishing between dimeric and trimeric coiled coils. The Multi-

Coil program simultaneously computes PairCoil scores based on each database to obtain a multidimensional score vector. The strength of a residue's dimeric score relative to its trimeric score in this multidimensional scoring space determines its classification as a two-stranded, three-stranded, or noncoiled coil. The MultiCoil program then converts the multidimensional score into a mathematically justified estimate of the probability that the residue is in a trimeric, dimeric, or noncoiled coil. Because the multidimensional score is used simultaneously for both the prediction of coiled coils and the classification of their oligomerization state, MultiCoil implements a single unified scoring algorithm.

Another scheme for distinguishing dimeric coiled coils from trimeric coiled coils has been proposed by Woolfson and Alber (1995). In their approach, a coiled-coil predictor is first used to locate potential coiled-coil regions under loose criteria dominated by hydrophobic considerations, and these regions are then classified as trimeric or dimeric. Woolfson and Alber show that all sequences in a selected dimeric data set score above zero, whereas all sequences in a selected trimeric data set score below zero. This compares with the results presented here for the MultiCoil program, which correctly classifies the oligomerization states of the test coiled-coil data sets. However, unlike MultiCoil, the program of Woolfson and Alber does not give confidence probabilities to the classifications, and coiled-coil predictors based upon hydrophobic content without considering pairwise interactions between heptad-repeat positions are prone to false positive predictions (Berger et al., 1995).

In this paper, the performance of the MultiCoil program is measured, both in terms of its ability to identify and classify sequences of known two- and three-stranded coiled coils, as well as its performance on databases of sequences with unknown structures. The program correctly classifies the oligomerization state of all sequences from the known databases, without returning any false positives. As in PairCoil (Berger, 1995; Berger et al., 1995), the use of pairwise residue correlations for the MultiCoil computations provides significantly better performance than was achieved by using single-frequency methods with MultiCoil. In addition to classifying individual sequences, MultiCoil scores give a measure of which pairwise residue interaction distances are most influential in differentiating among dimers, trimers, and noncoiled coils. These results support packing models previously proposed for residue-specific interactions within coiled coils that affect the oligomerization state (Harbury et al., 1993, 1994, 1995). The distribution of the MultiCoil scores is consistent with the hypothesis that trimeric coiled coils allow for more freedom in packing than dimeric coiled coils (Woolfson & Alber, 1995). A maximum likelihood approach is also used to estimate the fraction of all residues in proteins that fold as trimeric coiled coils and the fraction that fold as dimeric coiled coils.

Results

Scoring dimensions

The results reported in this paper were obtained using MultiCoil scores based upon pairwise interactions for residues distances 3, 4, and 5 apart with the trimeric table, and distances 2, 3, and 4 apart with the dimeric table. These distances were chosen as the scoring dimensions that best balanced the performance of the program for both accurately locating coiled coils and for distinguishing dimeric and trimeric coiled coils (see Table 1). Fitting Gaussians to the score distributions on each of the known data sets, distances 3 and 4 had the smallest overlap between both coiled-coil databases and the noncoiled-coil data. For distinguishing dimers from trimers, distance 5 had a small overlap when using the dimeric table, and distance 2 was best for the trimeric table. These optimal distances may have structural roots: for example, distances 3 and 4 include the **a** to **d** and **d** to **a** interactions essential for coiled-coil formation, whereas distances 5 and 2 may be useful for distinguishing trimers from dimers based upon **g** to **e** and **e** to **g** dependencies.

Table 1. *Three most relevant scoring dimensions for distinguishing pairs of data classes using the MultiCoil program*[a]

Sequence sets to distinguish		
Two-stranded coiled coils and PDB-minus Sequences	Three-stranded coiled coils and PDB-minus Sequences	Three-stranded coiled coils and Two-stranded coiled coils
Scoring dimension with the smallest overlap between Gaussians fit to sequence set scores		
Dimeric distances 4, 3, and 1	Trimeric distances 4, 3, and 2	Dimeric distances 3, 5, and 4, and Trimeric distances 2, 7, and 6

[a]A scoring dimension is determined by the table (dimeric or trimeric) and the scoring distance (1–7) used. The entries were used to determine three distances for each table, which were used to compute the results reported here. For each pair of data sets in the first row, the second row lists the scoring dimensions that best separated the data sets. Relevant distances had the smallest pairwise overlap of the Gaussians fit to the scores from that scoring dimension for the database sequences. Because it is to be expected that the dimeric scoring dimensions are most relevant to the two-stranded database and that the trimeric scoring dimensions are most relevant to the three-stranded database, the scoring dimensions were considered accordingly. The distances are listed in increasing order of overlap between the two Gaussians.

Score distribution of dimers, trimers, and noncoiled-coil sequences

Using the OWL database (release December 17, 1995) (Akrigg et al., 1988; Bleasby & Wootton, 1990) to approximate the space of all proteins, a maximum log likelihood analysis of the distribution of MultiCoil scores on the known dimeric, trimeric, and noncoiled-coil databases resulted in an estimate that 1.5% of all residues occur in dimeric coiled coils and 0.9% occur in trimeric coiled coils. Running on the Protein Identification Resource (PIR, release 38.09, 1994) gave comparable estimates of 2.25% dimeric and 0.8% trimeric. These estimates agree well with the estimate given in Berger et al. (1995) that 1 in every 50 residues is in a coiled coil, and the estimate of Lupas et al. (1991) that 1 in every 30 residues is in a coiled coil. A three-dimensional surface representing the magnitude of the log likelihood versus the values for the estimated fraction of residues occurring in dimeric and trimeric coils (P_{dim} and P_{trim}, respectively) is shown from various perspectives in Figure 1. The rate at which the plots fall away from the maximum value as the dimeric and trimeric probabilities are varied gives an indication of the certainty of the estimated probabilities. There are caveats to the estimates obtained from the MultiCoil scores. First, the databases (OWL and PIR), which were scored as approximations of the set of all proteins, are by no means complete

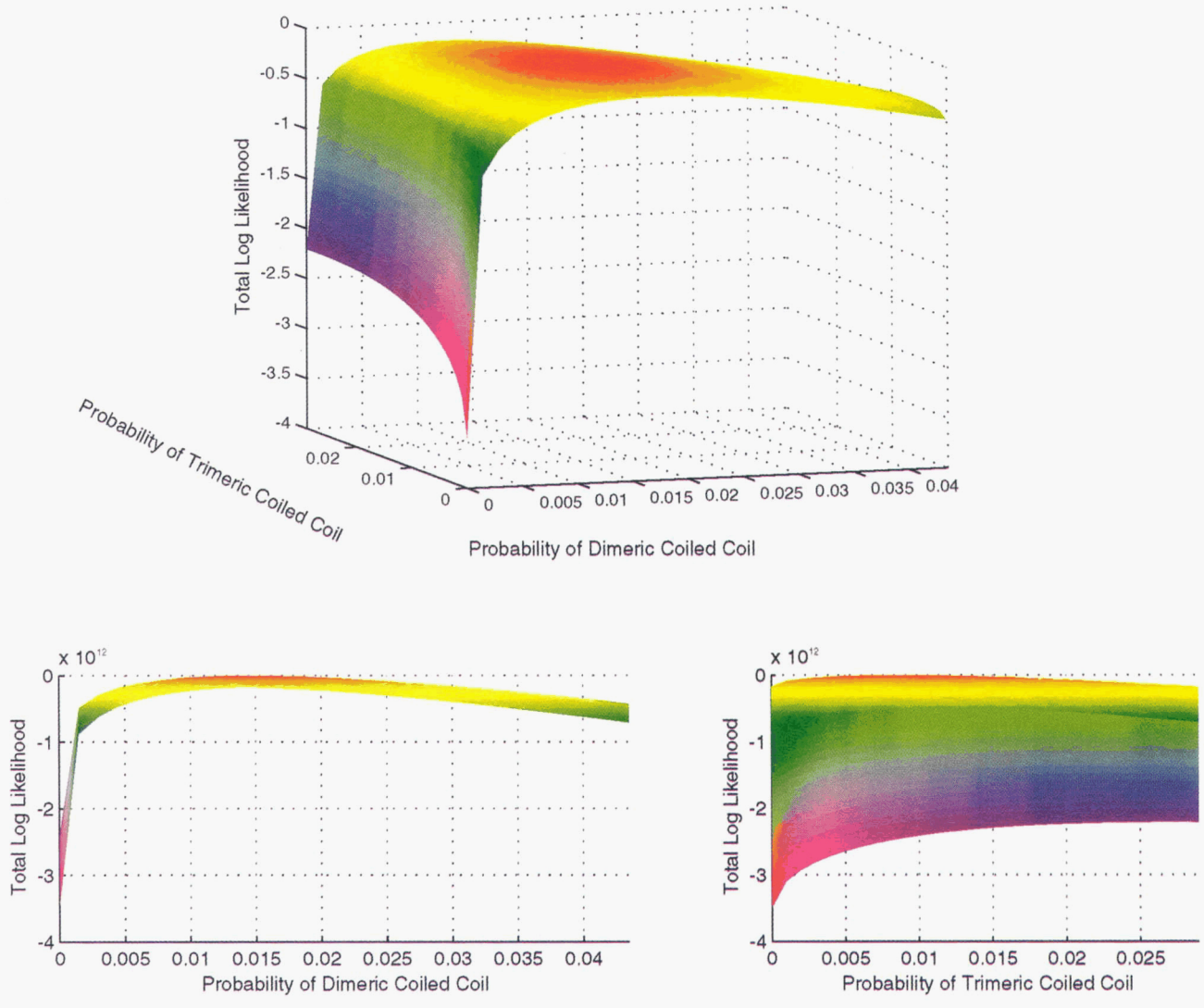

Fig. 1. Plots of the log likelihood surface from various perspectives, as the probability of being in a dimeric (P_{dim}) and a trimeric (P_{trim}) coiled coil vary. The rate at which the plots fall away from the maximum with changes in the probabilities indicates the confidence in the result of the maximum log likelihood analysis. The color is proportional to the surface height (red is the highest point), and represents the total log likelihood (minus the maximum value) for values of P_{dim} and P_{trim}. The upper panel shows the entire surface, and the lower panels show the variation of the log likelihood with the two probability parameters. The total log likelihood is maximized at $P_{dim} = 0.015$ and $P_{trim} = 0.009$.

and may be biased. Second, the databases used to represent dimers and trimers are also incomplete. Additional data (especially in the case of trimers) could affect the results easily. Nevertheless, given the limited data available currently, it appears that approximately 3% of protein residues will be in coiled coils.

Separation of data sets and classification: Advantages of multidimensional scores

The simplistic method of independently running the PairCoil program using frequencies gathered from the dimeric and trimeric databases was found to be insufficient to distinguish between trimeric and dimeric coiled coils. The overlap of the scores when running on the three known databases is shown in Figure 2 (top). This overlap decreases when the scores are viewed as a multidimensional vector. Figure 2 (bottom) plots the distribution of scores in a two-dimensional space of dimeric PairCoil scores versus trimeric PairCoil scores, due to the difficulty of visualizing the actual 14-dimensional space used by MultiCoil. Each sequence from a coiled-coil data set was removed from the probability table for that data set before scoring in order to decrease bias. A cursory examination reveals that the scores for each data set fall into distinct clusters that have a Gaussian-like density.

Classification of sequences in the known databases

The MultiCoil program predicts potential coiled coils with a probability divided into a dimeric and a trimeric portion. The sum of these two probabilities is the total probability for forming a coiled coil. Taking the ratio of the trimeric (or dimeric) probability to the total coiled-coil probability gives the probability that the predicted coiled coil is trimeric (or dimeric). Values greater than 0.5 for this

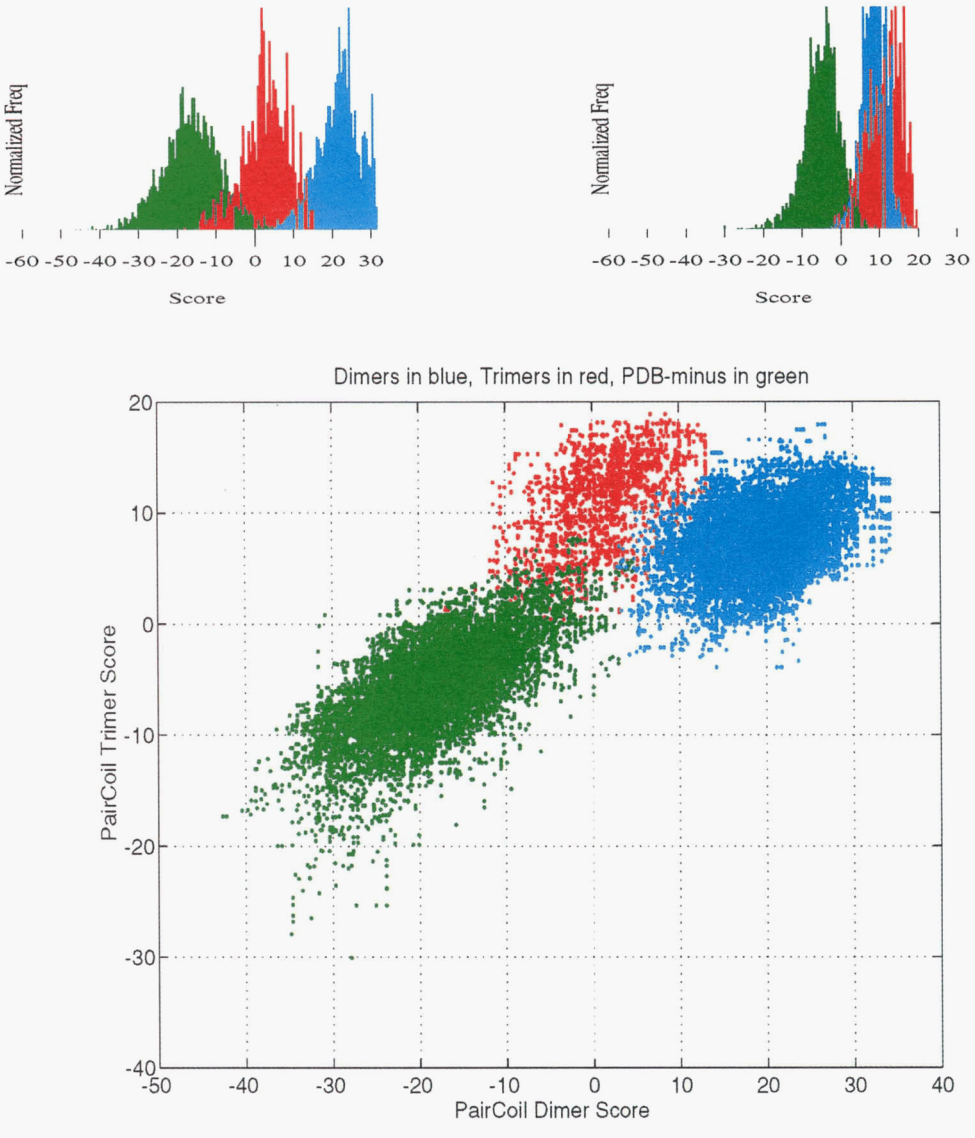

Fig. 2. Distribution of residue scores for the data sets of two-stranded coiled coils (58,191 residues) [blue], three-stranded coiled coils (6,319 residues) [red], and PDB-minus (34,940 residues) [green]. Top left: Histogram representation of the distribution of the residue scores from the PairCoil program when run with probabilities estimated from the two-stranded database. Top right: Distribution when run with probabilities estimated from the three-stranded database. Bottom: Improved separation with multidimensional scores. The *x*-dimension score was computed by PairCoil with probabilities from the two-stranded database, and the *y*-dimension score was computed using the three-stranded probability table. PairCoil scores based on the two-stranded database were computed using distances 3, 4, and 5. Scores with probabilities estimated from the three-stranded database were computed using distances 2, 3, and 4. Distances were combined using the method of Berger et al. (1995). Heights of histograms were normalized.

trimeric oligomerization ratio indicate a predicted trimer, whereas values less than 0.5 indicate a dimer. Care should be taken not to confuse the trimeric (or dimeric) oligomerization ratio with the trimeric (or dimeric) probability. The oligomerization ratio is a useful device for classifying coiled coils as dimeric or trimeric, whereas the probability measures the strength of identification for the coiled coil.

The MultiCoil program correctly classified all sequences in the coiled-coil databases as dimeric or trimeric. The *x* coordinate of Figure 3 shows the distribution of the trimeric oligomerization ratios for the sequences in the two coiled-coil databases. The *y* coordinate plots the total coiled-coil probability for each sequence, representing the confidence with which each sequence is located as a coiled coil. Each sequence was removed from its probability table before scoring. The worst (highest) trimeric oligomerization ratio by a dimer was F1;A27040: chicken neurofilament triplet M protein, which scored 58% dimeric probability and 19% trimeric probability (25% trimeric oligomerization ratio). The worst (lowest) oligomerization ratio by a trimer was L25541: human S laminin, which scored 30% dimeric probability and 38% trimeric probability (56% trimeric oligomerization ratio). In addition, the probabilities for the sequences in the database of noncoiled-coils (PDB-minus) are also plotted. The highest total coiled-coil probability from the PDB-minus was 33% (17% dimeric probability, 16% trimeric probability, residues 31–61) by 1LE2: lipoprotein. The structure of lipoprotein has been shown to include a four-α-

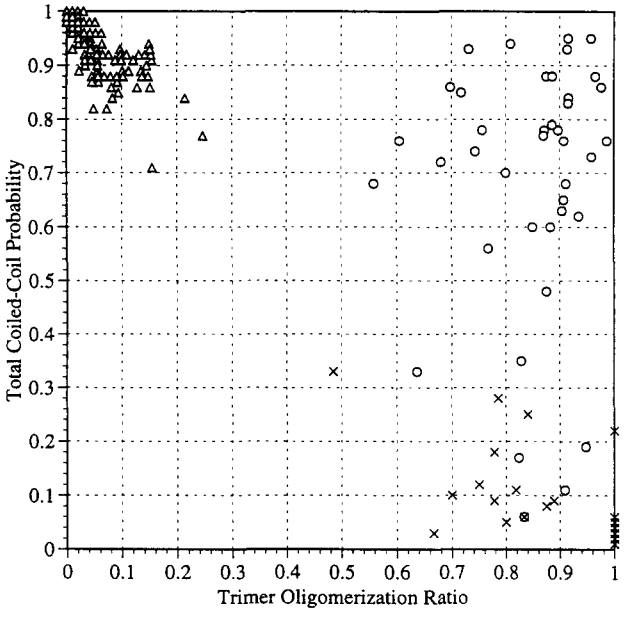

Fig. 3. Separation between sequences in the two-stranded database (△) and sequences in the three-stranded database (○) given by the MultiCoil trimeric oligomerization ratio. The *y*-axis represents the predicted coiled-coil probability for each sequence. Larger *y* values indicate greater confidence that the sequence contains a coiled coil. The *x*-axis represents the trimeric oligomerization ratio for each sequence. Values greater than 0.5 represent trimeric predictions by the program. The further the value from 0.5, the greater the confidence in the prediction. Sequences from the noncoiled-coil data set PDB-minus (×) are also plotted.

helical bundle, which bears structural similarities to coiled coils, but is not a coiled coil (Wilson et al., 1991). The other scores above 20% from the PDB-minus were: (1DPI: DNA polymerase I [Klenow fragment], 6% dimeric probability, 22% trimeric probability, residues 556–584); (1NGI: hydrolase, 4% dimeric probability, 21% trimeric probability, residues 248–276); (1CSG: cytokine, 0% dimeric probability, 22% trimeric probability, residues 14–41). The solved structures of these regions all include α-helices. All dimers scored above the total probability predicted for lipoprotein. From the trimeric database, all but the lamprey gamma fibrinogen strand, the heat shock factor proteins, and chicken fibrinogen beta strand scored above lipoprotein. The lamprey fibrinogen strands align only moderately well with the other fibrinogens in the database, which may account for the moderate scores of lamprey fibrinogen. It has been hypothesized that a number of heat shock factor proteins can form both homo-trimeric and hetero-dimeric coiled coils (Rabindran et al., 1993). The scores for these low-scoring trimers were: *Kluyveromyces lactis* heat shock factor: 19% (1% dimeric probability, 18% trimeric probability); *Saccharomyces cerevisiae* heat shock factor: 17% (3% dimeric probability, 14% trimeric probability); lamprey fibrinogen gamma chain: 11% (1% dimeric probability, 10% trimeric probability); and chicken fibrinogen beta chain: 6% (1% dimeric probability, 5% trimeric probability). All other sequences scored above 50% total coiled-coil probability, except for: *Xenopus laevis* gamma fibrinogen: 48% (6% dimeric probability, 42% trimeric probability); lamprey fibrinogen alpha2 chain: 35% (6% dimeric probability, 29% trimeric probability); lamprey fibrinogen beta chain: 33% (12% dimeric probability, 21% trimeric probability).

Thus, using a cutoff of 50% for the total coiled-coil probability, there are no false positive sequences and no false negatives in the dimeric data set, and only four trimeric sequences that score below the highest scoring negative (33%). All of the coiled coils in the database have a nonzero predicted chance of forming a coiled coil (with the lowest probability at 6%). For the 253 sequences in the PDB-minus database, 201 scored 0% coiled-coil probability, with 236 scoring below 6%, and all but four scoring below 20%. It is expected that the performance will improve even more as other three-stranded coiled-coils are discovered and the three-stranded database is increased in size.

MultiCoil on "unknown" test data

The MultiCoil program was run on the envelope, spike, and glycoproteins obtained from the PIR and GenPept [a translated version of GenBank (release 73, September 1992)], as well as on a set of dimeric coiled coils with the leucine-zipper motif (Hurst, 1994). A bound of 0.5 on the maximum residue coiled-coil probability over the sequence was used as a cutoff for finding coiled coils. Figure 4 shows the distribution of the sequence probabilities using the 0.5 cutoff (note that a sequence probability can be below 0.5 even though the maximum scoring residue is above 0.5; see methods). Using the 0.5 cutoff, 103 of 1,013 sequences in the envelope protein database were found as coiled coils (83 of which were classified as trimeric) and 41 of 53 sequences were found from the leucine zipper database (39 of which were classified as dimeric). Using a cutoff of 0.1, 191 of the envelope proteins and 50 of the leucine zippers were found.

It has been hypothesized that many of the envelope proteins are trimeric coiled coils (Blacklow et al., 1995; Lu et al., 1995; Fass et al., 1996; Chan et al., 1997). The distribution of MultiCoil scores supports this hypothesis. In general, the leucine zipper se-

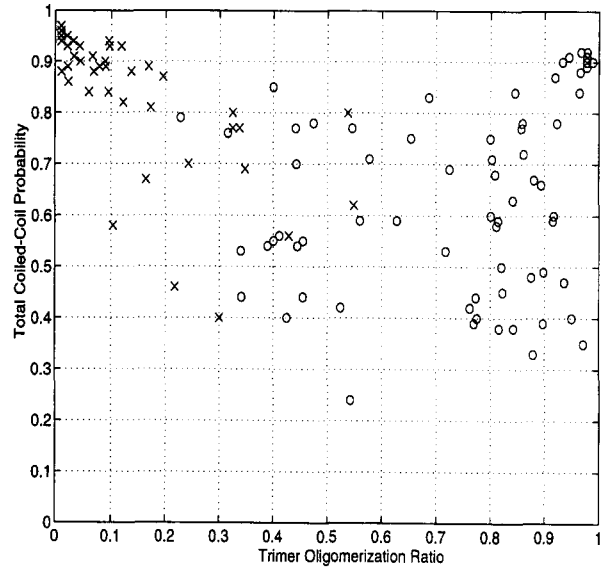

Fig. 4. Distribution of the predicted sequence probabilities and oligomerization ratios for MultiCoil run on a database of sequences containing the leucine-zipper motif (×) and for the envelope spike proteins and glycoproteins (○), which are hypothesized to include a number of trimeric coiled coils. The *y*-axis represents the predicted coiled-coil probability for each sequence. The *x*-axis represents the trimeric probability ratio for each sequence.

quences tend to score as dimeric. It is of interest that, as the total coiled-coil probability for the sequence increases, the clustering of the envelope proteins tends more toward the trimeric region, whereas the leucine zipper data set tends to cluster more tightly in the dimeric region of the space.

Performance on other types of coiled-coils

Despite the fact that the MultiCoil program was designed for the location of parallel dimeric and trimeric coiled coils, it can also provide insight into other structures. The five-stranded coiled coil in COMP (accession L32317) (Malashkevich et al., 1996) was predicted at 54% (14% dimeric, 40% trimeric) from residues 33–68. In contrast, the dimeric-based programs PairCoil and COILS (using windows of size 28 and the MTK matrix) scored the region under 5% (the MTIDK matrix of COILS scored it at 91%). However, the antiparallel tetramer ROP (accession P03051) (Munson et al., 1994) was not located as a coiled coil by MultiCoil or the other prediction programs. For the antiparallel trimer spectrin (accession X14519) (Pascual et al., 1996), all programs found coiled coils, however, none of the programs predicted correctly all three helices observed in the structure. Using a 0.5 bound, the MultiCoil score for the entire sequence was 66% (43% trimer, 23% dimer), with a maximum residue score of 88% (71% trimer, 17% dimer). For the antiparallel dimer seryl-tRNA synthetase (accession X91257) (Funjinaga et al., 1993; Oakley & Kim, 1997), the MultiCoil program only located a coiled coil with probability 3%, whereas PairCoil scored 38% and COILS scored 84%. However, when the MultiCoil dimeric scoring distances were changed to match the PairCoil scoring distances (1, 2, and 4) and only the best trimeric distance 4 for locating coiled coils was used, MultiCoil located the coiled coil with score 35% (15% dimeric, 20% trimeric). The fact that other distances give higher scores is to be expected, because the MultiCoil scoring dimensions were optimized for locating and distinguishing dimeric and trimeric parallel, not antiparallel coiled coils.

Predictions for protein sequences derived from the GCN4 leucine zipper

Figure 5 gives an example of the residue probabilities across the yeast GCN4 sequence as displayed by the MultiCoil program. GCN4 forms a two-stranded coiled coil in its final 30 residues (O'Shea et al., 1989, 1991), and these residues are classified as a two-stranded coiled coil by MultiCoil with a score of 80% (82% dimeric oligomerization ratio). A number of mutations in the residues in the **a** and **d** positions of the GCN4 leucine zipper have been shown to change the preferred oligomerization of the coiled coil (Harbury et al., 1993). Table 2 lists several of these mutations, the actual oligomerization state of the sequence, and the coiled-coil probability predicted by the MultiCoil program, along with the predicted trimeric oligomerization ratio. Although none of the GCN4 mutations scores exceedingly strongly in any oligomerization state, using 50% as a cutoff, they are all classified correctly, excluding the tetrameric coiled coil. The **a** → V, **d** → L mutation, which forms both trimers and dimers, has a trimeric oligomerization ratio of 31%.

Table 2. *MultiCoil predictions for various mutations of GCN4:leucine zipper versus the actual oligomerization states*[a]

Substitutions in the GCN4 leucine zipper:	Oligomerization	Total MultiCoil probability	Trimeric oligomerization ratio
Unmodified	Dimeric	0.80	18%
a → I, **d** → L	Dimeric	0.88	23%
a → I, **d** → I	Trimeric	0.27	81%
a → L, **d** → I	Tetrameric	0.54	76%
a → V, **d** → L	Forms trimers and dimers	0.81	31%
a → L, **d** → V	Trimeric	0.76	70%

[a]The first column gives the residues that were placed in registers **a** and **d** in the leucine-zipper region of the mutated proteins. The second column gives the actual oligomerization state of the protein. The third column gives the sum of the MultiCoil dimeric and trimeric predicted probabilities, representing the confidence with which the sequence is predicted to form a coiled coil. The final column of trimeric oligomerization ratios represents how confidently the program predicts the sequence to form a trimer, given that it forms a coiled coil. Values near 0% are dimeric predictions, and values near 100% are trimeric predictions.

Discussion

Comparison with other methods

Because the MultiCoil scorer uses the PairCoil scorer as a subprocess, many of the regions found by PairCoil (with the dimeric database) are also found by MultiCoil. In fact, many of the probabilities for regions predicted by PairCoil showed striking agreement with the total residue probabilities predicted by MultiCoil, despite the fact that the two programs compute probabilities from scores in different ways. MultiCoil generalizes the PairCoil program to locate more trimeric coiled coils and to give an idea of each coiled coil's bias for a particular oligomerization state.

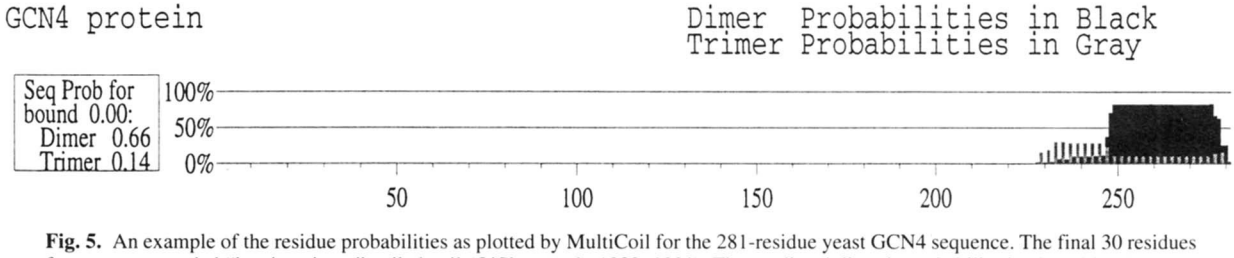

Fig. 5. An example of the residue probabilities as plotted by MultiCoil for the 281-residue yeast GCN4 sequence. The final 30 residues form a two-stranded "leucine zipper" coiled coil (O'Shea et al., 1989, 1991). The predicted dimeric probability is plotted in black and the trimeric probability is plotted in stripes (that appear gray in the program). MultiCoil probability predictions for classifying the entire sequence are also shown at the left of the figure, using a bound of 0 in the computations.

Differences between the data sets

The distribution of scores in Figure 2 indicates that the separation between the PDB-minus and the dimers is better defined than the separation between the scores for the PDB-minus and the trimers. This observation supports the hypothesis that trimers have fewer constraints on them than dimers because they have more packing freedom than dimers (Woolfson & Alber, 1995). An additional factor that contributes to this overlap is the small number of sequences that have been shown to contain three-stranded coiled coils, resulting in the small size of the three-stranded database. Hence, the trimers are more difficult to distinguish from noncoiled coils than dimers.

In addition, the PairCoil scores using the trimeric database for both coiled-coil data sets shown in Figure 2 (top right) range over roughly the same values. The similar range of scores indicates that many of the properties needed to form dimers are captured by the statistical distribution of the three-stranded coiled-coil data set. However, the reverse is not true. The PairCoil scores for trimers using the dimeric database overlap significantly with the noncoiled-coil scores and are noticeably lower than the scores for the two-stranded coiled coils. Thus, the dimeric table fails to pick out some of the statistical features that are important for distinguishing three-stranded coiled coils from noncoiled-coil sequences. This is consistent with the fact that a number of coiled coils in the three-stranded database are not identified or are not identified in the correct register by the PairCoil program.

Improving performance

The best way to improve the performance of the MultiCoil program does not involve changing the program, but instead, expanding the trimeric data set. The three-stranded coiled-coil database consists of only 6,319 residues, a small fraction of the size of the 58,191-residue two-stranded database. The current lack of a large database of trimeric coiled coils limits the accuracy of the statistical information that can be extracted from the database. As more trimeric coiled-coil structures are discovered, their addition to the database should improve the separation between the PDB-minus and the trimeric data set. The learning algorithm of Berger and Singh may provide another method to increase the size and diversity of the databases (Berger & Singh, 1997).

Materials and methods

Databases and probability estimates

A database of noncoiled-coil sequences was constructed from sequences with solved crystal structures in the Brookhaven Protein Data Bank (PDB, February 1994), and is represented by PDB-minus (Berger et al., 1995). Two databases of dimeric and trimeric coiled coils were also constructed. The construction of the two-stranded coiled-coil database is discussed in Berger et al. The three-stranded coiled-coil database was constructed from 42 sequences. The proteins in the three-stranded database have been characterized in the following references: laminins (families A, B1, and B2 chains) (Pikkarainen et al., 1987, 1988; Sasaki & Yamada, 1987; Sasaki et al., 1987, 1988; Montell & Goodman, 1988, 1989; Hunter et al., 1989, 1992; Engel, 1992; Kallunki et al., 1992; Nomizu et al., 1992; Beck et al., 1993; Porter et al., 1993; Gerecke et al., 1994; Vuolteenaho et al., 1994; Wewer et al., 1994), fibrinogens (families α, β, and γ chains) (Chung et al., 1981, 1983a, 1983b; Crabtree & Kant, 1982; Rixon et al., 1983; Crabtree et al., 1985; Strong et al., 1985; Bohonus et al., 1986; Koyama et al., 1987; Wang et al., 1989; Pastori et al., 1990; Weissbach & Grieninger, 1990; Pan & Doolittle, 1992), *K. lactis* heat shock factor (Peteranderl & Nelson, 1992), *S. cerevisiae* heat shock transcription factor (Sorger & Nelson, 1989), influenza virus hemagglutinin (Bullough et al., 1994), Moloney murine leukemia virus (Fass et al., 1996), fibritin encoded by wac gene for bacteriophage T4 and K3 (Efimov et al., 1994), gp17 leg protein in bacteriophages T7 and T3, and macrophage scavenger receptor protein (Conway & Parry, 1991).

Additional coiled-coil regions for fibrinogen and laminin sequences from species not discussed in the papers cited above were identified by using sequence alignments obtained from the program PILEUP (Genetics Computer Group, 1994). Fibrinogen strands were aligned with the corresponding human sequences characterized by Conway and Parry (1991). The laminin A chains were aligned with Conway and Parry's trimeric coiled-coil predictions for mouse A chain. Strands from the B1 and B2 laminin families were aligned with the corresponding human sequences.

The guidelines for determining which regions of the sequences to include were as follows. The seven residues before and the seven residues after skips, deletions, and register shifts in alignments were cut because of uncertainties as to the exact location of such structural changes. The seven residues before and the seven residues after each proline were cut from any coiled-coil regions because prolines are helix breakers (Conway & Parry, 1988). Only coiled coils of at least 28 residues were included in the frequency counts, consistent with the finding that short, stable coiled coils are approximately four heptads long (Lau et al., 1984; O'Shea et al., 1989; Lumb et al., 1994).

The frequencies of amino acids in heptad-repeat positions in the databases were used to estimate the corresponding single and pair probabilities for use in the MultiCoil algorithm, as discussed in Berger et al. (1995). Based on empirical measures of performance, the frequencies used to estimate probabilities were adjusted so that zero frequency events were given frequency 1/3 when estimating the probabilities. This method is discussed in Berger et al., where a value of 1/5 was used instead of 1/3. Single probabilities for residues *B*, *Z*, and *X* were calculated as follows. The probability of *B* was computed by averaging the probabilities of the asparagine and aspartic acid residues. The probability of *Z* was obtained by averaging the probabilities of glutamine and glutamic acid. An unknown residue, *X*, was given probability 1/20 in all databases, because there are 20 amino acids that could occur in the unknown position. Pairwise probabilities for these residues were computed analogously.

The algorithm

The MultiCoil program can be divided into a scoring phase and a conversion phase, in which the scores obtained are converted into probabilities that the scores identify a dimeric, trimeric, or noncoiled-coil structure. MultiCoil uses the PairCoil scorer implemented in Berger et al. (1995) as a subprocess to compute scores for each residue in a sequence. Recall that the PairCoil program predicts a region to be a coiled coil based on window scores, where a window consists of w contiguous residues in the sequence with an assigned heptad-repeat position. The window score is computed from the pairwise probabilities estimated from the databases. A score for each residue in a particular heptad-repeat position is obtained by

taking the maximum score over the windows containing that residue in that register. The PairCoil score for the residue is the maximum of the seven possible heptad-repeat position scores. The results presented here are based on scoring windows of size 28 (four heptads). Scoring windows of size 21 and 14 were found to predict less accurately. The pairwise interaction distances used to compute the scores are variable parameters to PairCoil, ranging between one and seven. Distance d corresponds to computations based on pairwise probabilities for residues lying distance d apart in the sequence, as well as on the single probabilities. The multidimensional MultiCoil scorer performs a separate run of PairCoil for each of the seven possible pairwise interaction distances. The result is a vector of seven different predictive scores based on different models of interactions between the residues in the coiled coil. Two *types* of PairCoil scores are computed for each of these interaction distances. Dimeric PairCoil scores are computed from the residue probabilities estimated from the dimeric database, and trimeric PairCoil scores are computed from the residue probabilities estimated from the trimeric database. Thus, MultiCoil computes a 14-dimensional score vector composed of seven dimeric scores and seven trimeric scores for each residue.

Converting scores to probabilities

To convert a set of scores into probabilities, the scores for each class (dimers, trimers, and noncoiled coils) were assumed to be Gaussian distributed in each of the 14 dimensions. The expected range of scores for the data classes is therefore characterized by 14 means and by the 14 × 14 covariance matrix of a multivariate Gaussian distribution. For each class, let μ_{class} denote this vector of the means along each of the 14 scoring dimensions. Similarly, let Σ_{class} be the covariance matrix for the class. These Gaussian parameters were determined experimentally by running the MultiCoil scoring program on the databases of known sequences for each of the two-stranded, three-stranded, and PDB-minus databases discussed in the first section of Materials and methods.

The positive databases of dimers and trimers were handled specially when computing the Gaussian parameters in order to reduce biased scores, because the program uses probabilities estimated from these databases to score the sequences. Each residue in a sequence from a positive database was scored in two different ways, and these scores were averaged in order to obtain a final score. The sequence was first scored using a probability table created by removing that single sequence from the database. The sequence was then scored again using a probability table in which the entire family of sequences with which the test sequence was aligned in creating the database was also removed. This method was used in order to maintain the characteristics of the residue probabilities computed from the database, while making the test sequence score in a fashion similar to a sequence "unknown" to the database. For the dimers, the families consisted of: four families of keratins; intermediate filaments comprising vimentin, desmin, and glial fibrillary acidic protein; one family for each of neurofilament H, L, and M; lamins; myosins; tropomyosins; and paramyosins (Berger et al., 1995). For the trimers, the families were: laminin A chains (including alpha chains and merosin); laminin B1 chains (including beta and S laminins); laminin B2 chains (including gamma); fibrinogen alpha chains; fibrinogen beta chains; fibrinogen gamma chains; and each of the other individual sequences described previously in the the first section of Materials and methods.

Having fixed the Gaussians for each of the three data sets, an arbitrary score vector can be classified into one of these three classes of dimeric, trimeric, and noncoiled-coil structures. When run on an arbitrary sequence, the MultiCoil program computes a 14-tuple of score values x for each position in the sequence. By standard statistical analysis (James, 1985), the probability that score x belongs to each class is computed as follows. Let $v_i(x)$ denote the value at the score vector x of the multivariate Gaussian determined for class i (where the classes range over $i = dim$ for dimers, $i = trim$ for trimers, and $i = non$ for noncoiled coils). This value is given by the matrix computation

$$v_i(x) = \frac{1}{(2\pi)^{n/2}\,|\Sigma_i|^{1/2}}\, e^{-0.5[x-\mu_i]^T \Sigma_i^{-1}[x-\mu_i]},$$

where $n = 14$ is the number of score dimensions, $|\Sigma_i|$ is the determinant of Σ_i, and y^T is the transpose of vector y.

The probability of being in class i is the fraction of the total Gaussian value from all three classes that is contributed by that particular class, where the Gaussian for each class is weighted by an initial probability of being in that class. For these initial probabilities, let P_i denote the probability that a residue chosen at random from any protein sequence lies within class i. Then, the total Gaussian weight for a score x is

$$\text{Total-gauss}(x) = P_{dim} * v_{dim}(x) + P_{trim} * v_{trim}(x) + P_{non} * v_{non}(x).$$

This gives

$$Pr[x \text{ belongs to class } i] = \frac{P_i * v_i(x)}{\text{total-gauss}(x)}.$$

Therefore, to convert scores to probabilities, estimates for the initial probabilities that a random protein residue is in a dimeric or trimeric coiled coil are needed (i.e., estimates for P_{dim} and P_{trim}). Note that, for simplicity, it is assumed that each residue can be classified as dimeric, trimeric, or noncoiled coil. This means that the fraction of noncoiled-coil residues is $P_{non} = 1 - P_{dim} - P_{trim}$. Making this assumption means that other higher-order oligomerization states of coiled coils will be classified into one of these three classes.

Estimating the fraction of residues that are dimeric and trimeric

The values of P_{dim}, P_{trim}, and P_{non}, along with the three Gaussians, define a distribution on the space of score vectors. The Gaussians have already been fixed by fitting to the MultiCoil scores for the known databases. Varying the initial probabilities P_{dim}, P_{trim}, and P_{non} changes the distribution these Gaussians define for all residue scores. A second distribution for the score vectors is defined by running the MultiCoil scorer on all residues in all protein sequences. The two distributions should be fairly similar when P_{dim} and P_{trim} are fixed at their actual values. To this effect, the two unknown parameters P_{dim} and P_{trim} were estimated using a maximum log likelihood analysis (Hartigan, 1975). The MultiCoil scoring program was run on the OWL database (release December 17, 1996) (Akrigg et al., 1988; Bleasby & Wootton, 1990) of proteins in order to approximate the distribution of score vectors on all proteins. Letting P_{dim} and P_{trim} vary, the two distributions agree best when the total log likelihood is maximized, where the total log likelihood is given by

$$\sum \log(P_{dim} * v_{dim}(x) + P_{trim} * v_{trim}(x) + P_{non} * v_{non}(x))$$

and the sum is over all score vectors x for residues in OWL.

Coiled-coil and sequence probabilities

The methods presented in the previous sections compute probability predictions for each residue in the sequence. In general, residue probabilities can vary significantly over the length of a predicted coiled coil. It is desirable to combine the information into one score for the whole coiled coil or sequence. Coiled-coil probabilities and sequence probabilities are defined here in order to automate the process of classifying sequences as dimeric, trimeric, or noncoiled coil. The first step in this process involves defining when the residue probabilities indicate that the residue is likely to be in a coiled coil. To this effect, a bound between 0 and 1 is set so that residues with total coiled-coil probability above the bound are marked as coiled coil (where the residue's total coiled-coil probability is the sum of its dimeric and trimeric probabilities).

The coiled-coil scores are computed as follows. All residues with a predicted coiled-coil total probability above the bound are marked. Contiguous regions of marked coiled-coil residues are grouped as one coiled coil. Weighted averages for the dimeric and trimeric residue probabilities across the whole coiled coil are then computed to give the entire coiled coil a single dimeric and a single trimeric probability. The weight for a residue is taken to be the residue's total coiled-coil probability. The weighted average causes the strongest predicted coiled-coil residues to have the most influence on the predicted oligomerization state for the entire coiled coil. Formally, the dimeric probability for the entire coiled coil is computed from the MultiCoil probabilities as

$$\frac{\sum_{\text{res in coil}} Pr[\text{res is coiled-coil}] * Pr[\text{res is dimeric}]}{\sum_{\text{res in coil}} Pr[\text{res is coiled-coil}]},$$

and the trimeric probability is

$$\frac{\sum_{\text{res in coil}} Pr[\text{res is coiled-coil}] * Pr[\text{res is trimeric}]}{\sum_{\text{res in coil}} Pr[\text{res is coiled-coil}]}.$$

Classification probabilities for the entire sequence are computed similarly. That is, the weighted average of all sequence residues above the bound is taken. Taking the ratio of the trimeric (or dimeric) probability to the total coiled-coil probability gives the probability that the sequence is trimeric (or dimeric), assuming that the residues that score above the bound are, in fact, part of a coiled coil. This ratio represents a probability for differentiating trimeric coiled coils from dimeric coiled coils, and is called the trimeric (or dimeric) oligomerization ratio.

Using fewer dimensions for improved performance

By scoring along a subset of the 14 dimensions with Gaussian submatrices along those dimensions, the MultiCoil score can be tailored more specifically to the problem considered here, distinguishing dimeric from trimeric coiled coils and separating both of these classes of sequences from the noncoiled-coil sequences. Reducing the number of dimensions scored also increases the speed of the program and decreases the effects of over-fitting to spurious statistical data. Each scoring dimension defines a one-dimensional Gaussian distribution of scores for the three structural classes. In a sense, the dimensions that have the least area of overlap between their Gaussians are the best dimensions for classifying sequences correctly. The Gaussian overlaps were examined pairwise in order to determine which dimensions were best for distinguishing each pair of structural classes (Table 1).

Acknowledgments

Thanks to Marcos Kiwi, Dan Kleitman, Mona Singh, and David Wilson for helpful discussions, and to Ravi Sundaram for occasional programming help. E.W. received support from a Department of Defense graduate fellowship. This research was supported by the National Institute of Health (GM44162 to P.S.K.), an NSF Career Award (to B.B.), and the Massachusetts Institute of Technology–State Street Bank Science Partnership Fund.

References

Akrigg D, Bleasby A, Dix N, Findlay J, North A, Parry-Smith D, Wootton J, Blundell T, Gardner S, Hayes F, Islam S, Sternberg M, Thornton J, Tickle I, Murray-Rust P. 1988. A protein sequence/structure database. *Nature 335*:745–746.

Beck K, Dixon T, Engel J, Parry D. 1993. Ionic interactions in the coiled-coil domain of laminin determine the specificity of chain assembly. *J Mol Biol 231*:311–323.

Berger B. 1995. Algorithms for protein structural motif recognition. *J Comput Biol 2*:125–138.

Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS. 1995. Predicting coiled coils by use of pairwise residue correlations. *Proc Natl Acad Sci USA 92*:8259–8263.

Berger B, Singh M. 1997. An iterative method for improved protein structural motif recognition. *Proc First Annual Int Conf on Computational Molecular Biology (RECOMB)*. ACM Press. pp 37–46.

Blacklow S, Lu M, Kim PS. 1995. A trimeric subdomain of the simian immunodeficiency virus envelope glycoprotein. *Biochemistry 34*:14955.

Bleasby AJ, Wootton J. 1990. Construction of validated, non-redundant composite protein sequence databases. *Protein Eng 3*:153–159.

Bohonus V, Doolittle R, Pontes M, Strong D. 1986. Complementary DNA sequence of lamprey fibrinogen beta chain. *Biochemistry 25*:6512–6516.

Brown JH, Cohen C, Parry DAD. 1996. Heptad breaks in α-helical coiled coils: Stutters and stammers. *Proteins Struct Funct Genet 26*:134–145.

Bullough PA, Hughson FM, Skehel JJ, Wiley DC. 1994. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature 371*:37–43.

Carr CM, Kim PS. 1993. A spring-loaded mechanism for the conformational change of influenza hemagglutinin. *Cell 73*:823–832.

Chan DC, Fass D, Berger JM, Kim PS. 1997. Core structure of gp41 from the HIV envelope glycoprotein. *Cell 89*:263–267.

Chung DW, Chan WY, Davie E. 1983a. Characterization of complementary deoxyribonucleic acid coding for the gamma chain of human fibrinogen. *Biochemistry 22*:3250–3256.

Chung DW, Que B, Rixon MW, Mace M Jr, Davie E. 1983b. Characterization of complementary deoxyribonucleic acid and genomic deoxyribonucleic acid for the beta chain of human fibrinogen. *Biochemistry 22*:3244–3250.

Chung DW, Rixon M, MacGillivvray R, Daie E. 1981. Characterization of a cDNA clone coding for the beta chain of bovine fibrinogen. *Proc Natl Acad Sci USA 78*:1466–1470.

Cohen C, Parry DAD. 1990. α-Helical coiled coils and bundles: How to design an α-helical protein. *Proteins Struct Funct Genet 7*:1–15.

Conway JF, Parry DAD. 1988. Intermediate filament structure: 3. Analysis of sequence homologies. *Int J Biol Macromol 10*:79–98.

Conway JF, Parry DA. 1991. Three-stranded α-fibrous proteins: The heptad repeat and its implications for structure. *Int J Biol Macromol 13*:14–16.

Crabtree GR, Comeau C, Fowlkes D, Fornace AJ Jr, Malley J, Kant J. 1985. Evolution and structure of the fibrinogen genes. *J Mol Biol 185*:1–19.

Crabtree GR, Kant JA. 1982. Organization of the rat gamma-fibrinogen gene. *Cell 31*:159–166.

Efimov VP, Nepluev IV, Sobolev BN, Zurabishvil TG, Schulthess T, Lustig A, Engel J, Haener M, Aebi U, Venyaminov SY, Potekhin SA, Mesyanzhinov VV. 1994. Fibritin encoded by bacteriophage t4 gene wac has a parallel triple-stranded α-helical coiled coil structure. *J Mol Biol 242*:470–486.

Engel J. 1992. Laminins and other strange proteins. *Biochemistry 31*:10645–10651.

Fass D, Harrison SC, Kim PS. 1996. Retrovirus envelope domain at 1.7 Å resolution. *Nature Struct Biol 3*:465–469.

Funjinaga M, Berthet-Colominas C, Yaremchuk AD, Tukalo MA, Cusack S. 1993. Refined crystal structure of the seryl-trna synthetase from *Thermus thermophilus* at 2.5 Å. *J Mol Biol 204*:222–233.

Genetics Computer Group. 1994. *PILEUP: Program Manual for the Wisconsin Package, August 1994*. Genetics Computer Group, 575 Science Drive, Madison, Wisconsin 53711, USA.

Gerecke DR, Wagman DW, Champliaud M, Burgeson R. 1994. The complete primary structure for a novel laminin chain, the laminin b1k chain. *J Biol Chem 269*:11073–11080.

Harbury PB, Kim PS, Alber T. 1994. Crystal structure of an isoleucine-zipper trimer. *Nature 371*:80–83.

Harbury PB, Tidor B, Kim PS. 1995. Repacking protein cores with backbone freedom: Structure prediction for coiled coils. *Biochemistry 92*:8408–8412.

Harbury PB, Zhang T, Kim PS, Alber T. 1993. A switch between two-, three-, and four-stranded coiled coils in gcn4 leucine zipper mutants. *Science 262*:1401–1406.

Hartigan J. 1975. *Clustering algorithms*. New York: John Wiley and Sons. p 116.

Hunter DD, Shah V, Merlie J, Sanes J. 1989. A laminin-like adhesive protein concentrated in the synaptic cleft of the neuromuscular junction. *Nature 338*:229–234.

Hunter I, Schulthess T, Engel J. 1992. Laminin chain assembly by triple and double stranded coiled-coil structures. *J Biol Chem 267*:6006–6011.

Hurst HC. 1994. bzip proteins. *Protein Profile 1*:123–168.

James M. 1985. *Classification algorithms*. New York: John Wiley and Sons. p 20.

Kallunki P, Sainio K, Eddy R, Byers M, Kallunki T, Sariola H, Beck K, Hirvonen H, Shows T, Tryggvason K. 1992. A truncated laminin chain homologous to the b2 chain. *J Cell Biol 119*:679–693.

Koyama T, Hall L, Haser W, Tonegawa S, Saito H. 1987. Structure of a cytotoxic t-lymphocyte-specific gene shows a strong homology to fibrinogen beta and gamma chains. *Proc Natl Acad Sci USA 85*:1609–1613.

Landshulz W, Johnson P, McKnight S. 1989. Leucine repeats and an adjacent DNA binding domain mediate the formation of functional cfos-cjun heterodimers. *Science 243*:1681–1688.

Lau SYM, Taneja AK, Hodges RS. 1984. Synthesis of a model protein of defined secondary and quaternary structure. Effect of chain length on the stabilization and formation of two-stranded alpha-helical coiled-coils. *J Biol Chem 259*:13253–13261.

Lu M, Blacklow S, Kim PS. 1995. A trimeric structural domain of the HIV-1 transmembrane glycoprotein. *Nature Struct Biol 2*:1075–1082.

Lumb KJ, Carr CM, Kim PS. 1994. Subdomain folding of the coiled coil leucine zipper from the bzip transcriptional activator gcn4. *Biochemistry 33*:7361–7367.

Lupas A, van Dyke M, Stock J. 1991. Predicting coiled coils from protein sequences. *Science 252*:1162–1164.

Malashkevich VN, Kammerer RA, Efimov VP, Schulthess T, Engel J. 1996. The crystal structure of a five-stranded coiled coil in COMP: A prototype ion channel? *Science 274*:761–764.

Montell DJ, Goodman CS. 1988. *Drosophila* substrate adhesion molecule: Sequence of laminin b1 chain. *Cell 53*:463–473.

Montell DJ, Goodman CS. 1989. *Drosophila* laminin: Sequence of b2 subunit. *J Cell Biol 109*:2441–2453.

Munson M, O'Brien R, Sturtevant JM, Regan L. 1994. Redesigning the hydrophobic core of a four-helix-bundle protein. *Protein Sci 3*:2015–2022.

Nomizu M, Utani A, Shiraishi N, Yamada Y, Roller P. 1992. Synthesis and conformation of the trimeric coiled-coil segment of laminin. *Int J Peptide Protein Res 40*:72–79.

Oakley MG, Kim PS. 1997. Protein dissection of the antiparallel coiled coil from *E. coli* seryl tRNA synthetase. *Biochemistry*. Forthcoming.

O'Shea EK, Klemm JD, Kim PS, Alber T. 1991. X-ray structure of the gcn4 leucine zipper, a two-stranded, parallel coiled coil. *Science 254*:539–544.

O'Shea EK, Rutkowski R, Kim PS. 1989. Evidence that the leucine zipper is a coiled coil. *Science 243*:538–542.

Pan Y, Doolittle RF. 1992. cDNA sequence for a second fibrinogen alpha chain in lamprey. *Proc Natl Acad Sci USA 89*:2066–2070.

Parry D. 1982. Coiled-coils in alpha-helix-containing proteins: Analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. *Biosci Rep (England) 2*:1017–1024.

Pascual J, Pfuhl M, Rivas G, Pastore A, Saraste M. 1996. The spectrin repeat folds into a three-helix bundle in solution. *FEBS Lett 383*:201–207.

Pastori RL, Moskaitis J, Smith LH Jr, Shoenberg D. 1990. Estrogen regulation of *Xenopus laevis* gamma-fibrinogen gene expression. *Biochemistry 29*:2599–2605.

Peteranderl R, Nelson HCM. 1992. Trimerization of the heat shock transcription factor by a triple-stranded alpha-helical coiled-coil. *Biochemistry 31*:12272–12276.

Pikkarainen T, Eddy R, Fukushima Y, Byers M, Shows T, Pihlajaniemi T, Saraste M, Tryggvason K. 1987. Human laminin b1 chain. *J Biol Chem 262*:10454–10462.

Pikkarainen T, Kallunki T, Tryggvason K. 1988. Human laminin b2 chain. *J Biol Chem 263*:6751–6758.

Porter BE, Justice M, Copeland N, Jenkins N, Hunter D, Merlie J, Sanes J. 1993. S-laminin: Mapping to mouse chromosome 9 and expression in the linked mutants tippy and ducky. *Genomics 16*:278–281.

Rabindran SK, Haroun RI, Clos J, Wisniewski J, Wu C. 1993. Regulation of heat shock factor trimer formation: Role of a conserved leucine zipper. *Science 259*:230–234.

Rixon MW, Chan WY, Davie E, Chung DW. 1983. Characterization of complementary deoxyribonucleic acid coding for the alpha chain of human fibrinogen. *Biochemistry 22*:3237–3244.

Sasaki M, Kato S, Kohno K, Martin G, Yamada Y. 1987. Sequence of the cDNA encoding the laminin b1 chain reveals a multidomain protein containing cysteine-rich repeats. *Proc Natl Acad Sci USA 84*:935–939.

Sasaki M, Kleinman HK, Huber H, Deutzmann R, Yamada Y. 1988. Laminin, a multidomain protein. *J Biol Chem 263*:16536–16544.

Sasaki M, Yamada Y. 1987. The laminin b2 chain has a multidomain structure homologous to the b1 chain. *J Biol Chem 262*:17111–17117.

Sorger PK, Nelson HCM. 1989. Trimerization of a yeast transcriptional activator via a coiled coil motif. *Cell 59*:807–813.

Strong D, Moore M, Cottrell A, Bohonus V, Pontes M, Evans B, Riley M, Doolittle R. 1985. Lamprey fibrinogen gamma chain: Cloning, cDNA sequencing, and general characterization. *Biochemistry 24*:92–101.

Vuolteenaho R, Nissinen M, Sainio K, Byers M, Eddy R, Hirvonen H, Shows T, Sariola H, Engvall E, Tryggvason K. 1994. Human laminin m chain (merosin). *J Cell Biol 124*:381–394.

Wang Y, Patterson J, Gray J, Yu C, Cottrell B, Shimizu A, Graham D, Riley M, Doolittle R. 1989. Complete sequence of the lamprey fibrinogen alpha chain. *Biochemistry 28*:9801–9806.

Weissbach L, Grieninger G. 1990. Bipartite mrna for chicken alpha-fibrinogen potentially encodes an amino acid sequence homologous to beta- and gama-fibrinogens. *Proc Natl Acad Sci USA 87*:5192–5202.

Wewer UM, Gerecke D, Durkin M, Kurtz K, Mattei M, Champliaud M, Burgeson R, Albrectsen R. 1994. Human beta2 chain of laminin (formerly s chain). *Genomics 24*:243–252.

Wilson C, Wardell MR, Weisgraber KH, Mahley RW, Agard DA. 1991. Three-dimensional structure of the ldl receptor-binding domain of human apolipoprotein e. *Science 252*:1817–1822.

Woolfson DN, Alber T. 1995. Predicting oligomerization states of coiled coils. *Protein Sci 4*:1596–1607.

Appendix: Optimizing the performance of MultiCoil

For the results reported here, the scoring dimensions used were distances 3, 4, and 5 with the dimeric database, and distances 2, 3, and 4 with the trimeric database. These distances were chosen to locate and distinguish two- and three-stranded coiled coils simultaneously. Other sets of distances may be more suited to other applications. For example, if a region is known to be a coiled coil and MultiCoil is used only as a dimeric/trimeric distinguisher, distance 4 may not be the most relevant trimeric distance (see Table 1). The MultiCoil program allows for the scoring dimensions to be modified at the user's discretion. Changing the scoring dimensions changes which of the characteristics of the structures in the known databases are emphasized for classification.

Additional improvement in performance can be obtained by tailoring the parameters P_{dim} and P_{trim} to the sequences being tested. These parameters represent the a priori probabilities that a random residue in a sequence is in a dimeric or trimeric coiled coil. They were optimized for the purpose of identifying coiled-coil sequences from a database of random sequences, without returning noncoiled-coil sequences. However, if the program is being run on a sequence that is known to contain a coiled coil, more accurate probability predictions may be obtainable by raising P_{dim} and P_{trim} so as to incorporate this initial knowledge. For a sequence known to contain a coiled coil, eliminating false positives is less important than classifying the type of coiled coil, and the parameters may be adjusted accordingly.

Finally, the bound used to determine which residues are classified as coiled coils when computing the sequence scores and the coiled-coil scores may be adjusted. For scanning through a large database, the bound may be set relatively high in order to return only the most probable coiled coils. When running MultiCoil on a set of sequences thought to contain coiled coils, the bound may be set lower in order to classify even sequences with a weak propensity for coiled-coil formation.

Interpreting scores

The sequence and coiled-coil scores are the most straightforward scores to interpret. If the total coiled-coil probability is high (e.g., above 75%), then there is high confidence that the region forms some sort of a coiled coil. If most of that probability is contributed by either the predicted dimeric or the predicted trimeric probability, then the region is predicted to form that oligomerization state. In sequence regions where there is a significant amount of probability contributed by both oligomerization states, there are several interpretations. The region could very well be dimeric or trimeric (with scores falling in the area of overlap between the dimeric and trimeric data sets in Fig. 2). Scoring in this overlap area could also indicate that the bias toward one particular oligomerization state is not overwhelming, as with the GCN4 mutation, which forms both dimers and trimers (Table 2).

Interpreting very weakly predicted coiled-coil regions is another difficult problem. As can be seen from Figures 2 and 3, the scores for known trimers and known noncoiled coils overlap. Hence, a coiled-coil probability under 50% does not mean that the region is definitely not in a coiled coil. It merely means that there is a greater chance for that region to be in a noncoiled-coil structure than in a coiled-coil structure. The program user must interpret the results on a case by case basis. Weakly predicted regions may actually form coiled coils, or may just possess some properties of coiled coils, but not fit the exact structure of a coiled coil. For example, ENV_SIVM1, Simian immunodeficiency virus (MM142-83 isolate), has a predicted coiled-coil region at residues 636–679 of the envelope polyprotein with 58% total coiled-coil probability (44% trimeric, 14% dimeric). ENV_HV1H2, Human immunodeficiency virus type 1 (HXB2 isolate), scores 10% total coiled-coil probability (9% trimeric, 1% dimeric) at residues 539–573 of the envelope polyprotein, and scores 15% total coiled-coil probability (15% trimeric, 0% dimeric) at residues 628–671. The core of HIV gp41 forms a bundle of six alpha-helices (Blacklow et al., 1995; Lu et al., 1995; Chan et al., 1997). Lipoprotein scores 33% total coiled-coil probability and forms a four α-helical bundle (Wilson et al., 1991).

It is especially important to consider the strength with which the coiled-coil region is located when determining the confidence to place in the oligomerization state predictions. It was found in trial runs that the accuracy of the oligomerization ratio predictions decreased as the total predicted coiled-coil probability decreased. Because the low scores for the trimeric database tend to overlap with the high scores for the PDB-minus, many of the weakly predicted coiled coils tended to score as trimeric.

Interpreting residue scores

The MultiCoil residue scores (Fig. 5) indicate the regions of a coiled coil that contribute most to the predicted oligomerization state. In the case of sequences that are not classified strongly in one state by the coiled-coil scorer, the distribution of residue scores might provide additional information to help predict the structure of the region.

An additional point of interest in drawing conclusions from the program predictions occurs when examining the registers predicted by MultiCoil. Multiple heptad-repeat register can have a high score. The scores output by MultiCoil come from the maximum scoring register along each dimension, but the probabilities predicted for the other registers can be obtained as well. It has been suggested that register shifts in coiled coils occur over several positions in the sequence, where these residues simultaneously have characteristics of two heptad-repeat positions (Brown et al., 1996). Multiple high-scoring registers in the MultiCoil predictions may suggest the presence of a register shift. Additionally, in some cases, one predicted register may give the coiled coil an apparent dimeric structure, whereas another register may predict a trimeric coiled coil. Knowledge of these alternative predictions may be useful in understanding the structure of the protein.