Available online at www.sciencedirect.com
ScienceDirect
Journal of Structural Biology 155 (2006) 140–145
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

Comparative analysis of coiled-coil prediction methods

Markus Gruber, Johannes Söding, Andrei N. Lupas *

Max-Planck Institute for Developmental Biology, Spemannstr.35, 72076 Tübingen, Germany

Received 19 December 2005; accepted 22 March 2006
Available online 31 March 2006

Abstract

In this study we compare commonly used coiled-coil prediction methods against a database derived from proteins of known structure. We find that the two older programs COILS and PairCoil/MultiCoil are significantly outperformed by two recent developments: Marcoil, a program built on hidden Markov models, and PCOILS, a new COILS version that uses profiles as inputs; and to a lesser extent by a PairCoil update, PairCoil2. Overall Marcoil provides a slightly better performance over the reference database than PCOILS and is considerably faster, but it is sensitive to highly charged false positives, whereas the weighting option of PCOILS allows the identification of such sequences.

Keywords: Coiled-coil prediction; COILS; Marcoil; MultiCoil; PairCoil2; Sequence analysis

1. Introduction

Coiled-coils are built by two or more α-helices that wind around each other to form a supercoil. Their structure is better understood than that of any other fold and, uniquely among proteins, their coordinates can be calculated from parametric equations (Crick, 1953). In essence they are built of sequence elements of three and four residues whose hydrophobicity pattern and residue composition is fundamentally compatible with the structure of amphipathic α-helices (Hicks et al., 1997; Gruber and Lupas, 2003). By far the most common arrangement of elements in coiled coils is the regular alternation of threes and fours, referred to as the heptad repeat. In this repeat the individual positions are labelled *a–g*, with positions *a* and *d* being generally hydrophobic (for a recent review of coiled-coil structure see Lupas and Gruber, 2005).

The prediction of coiled-coils from protein sequences was pioneered by Parry (1982), who showed that each heptad position has a characteristic residue distribution and proposed to score the coiled-coil-forming propensity of a sequence by its match to a position-specific scoring matrix derived from these distributions. To our knowledge this represents the first application of sequence profiles for structure prediction.

The first widely-used prediction program, COILS (Lupas et al., 1991; Lupas, 1996), extended this approach by substituting residue preferences for frequencies, introducing a scanning window, and scaling scores against reference databases to obtain probabilities.

Three subsequent programs, PairCoil (Berger et al., 1995), MultiCoil (Wolf et al., 1997), and LearnCoil (Berger and Singh, 1997) were built on similar concepts, but increased the amount of informational input by incorporating pairwise residue correlations into the scoring matrix (PairCoil has recently been updated to PairCoil2 (McDonnell et al., 2006) trained with a larger set of coiled-coil data.). MultiCoil goes beyond generic coiled-coil prediction by additionally differentiating between two- and three-stranded coiled-coils. LearnCoil can be trained iteratively on a specific set of target proteins and is available for histidine kinases (Singh et al., 1998) and viral membrane fusion proteins (Singh et al., 1999).

A further increase in informational input was achieved by Marcoil (Delorenzi and Speed, 2002), which calculates

* Corresponding author. Fax: +49 7071 601 349.
E-mail address: Andrei.Lupas@tuebingen.mpg.de (A.N. Lupas).

1047-8477/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2006.03.009

posterior probabilities from a hidden Markov model and is currently the only windowless method, removing a limitation of previous prediction programs.

We have recently also increased the informational input into COILS by developing a version, PCOILS (Gruber et al., 2005), which substitutes profile-profile comparisons for the sequence-profile comparison step in COILS (http://protevo.eb.tuebingen.mpg.de/coils). The profiles are either derived from a multiple sequence alignment provided as input or are generated by an automated procedure based on PSI-BLAST (Altschul et al., 1997).

In this study, we have undertaken a comparative analysis of these prediction programs employing a strict benchmark set of coiled coil superfamilies that were unknown at the time when the individual programs were trained (except for PairCoil2, which has just been released). By taking advantage of the large number of coiled-coil structures solved to atomic resolution in recent years, we developed a benchmark database of over 16,000 residues. This database includes a large number of comparatively short coiled-coils, thus representing a more realistic benchmark than the databases of long coiled-coils such as tropomyosin, myosin and intermediate filaments, which have hitherto been used, for instance in the most recent comparison provided by McDonnell et al. (2006). An essential difference of this database over previously used testing databases is its more objective nature. By using the intersection between the SCOP coiled-coil class and the output of the SOCKET program, we have eliminated arbitrary decisions from the database generation process. Our results show that COILS and MultiCoil provide similar performance, but that MultiCoil is too restrictive in assigning probabilities. Both programs are substantially outperformed by Marcoil and PCOILS, with Marcoil providing the best performance overall. The recent PairCoil update, PairCoil2, is less performant than either Marcoil or PCOILS, despite including most proteins of the test set in its training database.

2. Materials and methods

2.1. Benchmark database

A challenge in constructing a benchmark database is that many sequences of known coiled-coil structures are already included in the matrices used by the methods, and therefore it would not be surprising if those structures could easily be detected. This is why we only considered coiled-coil superfamilies whose member structures are recent enough that they could not be included in the training sets of any program, with the exception of PairCoil (this program was not available at the time of submission of this article and was included in revision). We thus obtained coiled-coil superfamilies from the latest SCOP database (version 1.69) that were not yet included in the SCOP 1.55 release (i.e. after July 2001), 19 in total (Murzin et al., 1995, http://scop.mrc-lmb.cam.ac.uk/scop). Myosins and SNAREs were excluded, because at least one program already used them for training. For each superfamily, a representative structure was taken and analyzed with SOCKET (Walshaw and Woolfson, 2001, http://www.lifesci.sussex.ac.uk/research/woolfson/html/coiledcoils/socket). Only sections that were clearly identified as being engaged in knobs-into-holes packing were considered (superfamilies h.4.10, h.4.11, h.4.14, h.4.16, and h.6.1 were thus excluded, since SOCKET could not find this packing in any member structure). For the database of non-coiled coils, one tenth of the families present in the SCOP alpha and beta protein class (a/b) were chosen at random (superfamilies c.37.1, c.49.2, c.67.1, and c.93.1 were discarded, as they contain coiled-coil domains). To enrich these databases with reliable homologs, searches of the nonredundant protein sequence database were made with PSI-BLAST (two iterations, *E*-value cutoff of 10^{-4}). Prior to the searches the database was reduced to a set with a maximum of 90% sequence identity and low-complexity regions were masked out with Pfilt (Jones and Swindells, 2002). The identified sequences were aligned and again filtered to sequences that had a minimum coverage of 20% and a minimum sequence identity of 40% to the query sequence by hhfilter (Soding, 2005). Finally, an input profile from this alignment was compiled with hhmake for use in the profile-version of COILS. The final database contained 16,449 residues in the positive set and 1,287,148 residues in the negative set.

2.2. Analyzing the performance of COILS and PCOILS

Three parameters can be set by the user in COILS: the window size (14, 21, or 28), the profile matrix (MTK or MTIDK), and a weighting option for core residues (on or off). The significance of the window size is well-understood: generally a longer window provides higher statistical significance, but as soon as the window size exceeds the size of the coiled-coil region, extraneous residues are included and distort the results. Since our benchmark set contains many coiled coils that are shorter than 28 residues and a window of 14 residues would not properly reflect the performance of COILS, we used a window of 21 residues throughout.

The two matrices in COILS were compiled over a decade ago on a very limited dataset of 'certain' coiled coils (myosins, tropomyosins, and keratins (intermediate filaments type I and II) in the case of MTK and myosins, paramyosins, tropomyosins, intermediate filaments type I–V, desmosomal proteins and kinesins in the case of MTIDK). To study whether matrices compiled from a larger and more diverse dataset would improve the performance of COILS, we generated a matrix derived from coiled coils of known structure (the PDB matrix) and a converged matrix built by iterative searches over the nonredundant protein sequence database (the iterated matrix). For the PDB matrix we used proteins that we had identified as coiled coils by visual inspection and SOCKET analysis in preparation for a recent review article (Lupas and Gruber, 2005) (note that we did not use this set to build the reference database of coiled coils for this study since it is biased by our view of

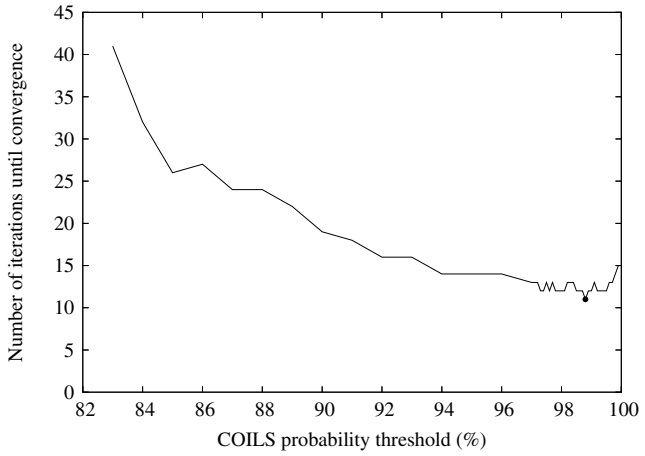

Fig. 1. Convergence of COILS matrix in an iterative process to find an ideal matrix.

the structure database; rather, we used the intersection of SCOP and SOCKET as a 'safer' and more impartial set.). For the iterated matrix, we scanned the nonredundant sequence database with COILS, using the PDB matrix and retained all residues above a given probability threshold (for this purpose the sequence database was filtered to 70% sequence identity with both regions of low complexity and regions with extremely biased amino acid compositions masked out by Pfilt (Jones and Swindells, 2002)). From the retained residues we calculated a new matrix and repeated this procedure until the matrix converged (convergence was defined as a root mean square deviation of less than 0.05% relative to the preceding matrix). We found that the procedure converged fastest at a probability cutoff of 98.8% (Fig. 1). The converged matrix at this cutoff was the one we used in this study.

The third parameter that can be set in COILS is the weighting of the core positions *a* and *d* relative to the other positions (2.5:1 versus 1:1). In the unweighted mode, the two core positions are outweighed by the solvent-exposed coat positions, introducing a bias into the method towards hydrophilic, charge-rich sequences. Occasionally this leads to high predicted coiled-coil probabilities in the obvious absence of heptad periodicity and coiled-coil-forming potential. COILS therefore allows the user to assign the same weight to the two hydrophobic positions *a* and *d* as to the five hydrophilic positions *b*, *c*, *e*, *f*, and *g*. We analyzed the performance of both COILS and PCOILS in the weighted and unweighted mode.

2.3. Comparison of the different tools

In this study we compared COILS, MultiCoil, Marcoil, PairCoil2, and PCOILS by analyzing coverage versus reliability. Coverage was defined as $true\ positives/(true\ positives + false\ negatives)$ and shows the proportion of coiled-coil residues that were correctly predicted. Reliability was defined as $true\ positives/(true\ positives + c \times false\ positives)$ with *c* being a correction factor that imposes the assumption that in nature non-coiled-coil residues occur 20 times more often than coiled-coil residues.

Moreover, the quality of the coiled-coil probabilities reported by the programs were analyzed (Fig. 3), with the exception of Paircoil2 which does not provide probabilities. For the benchmark set, the density of the reported probabilities was analyzed separately for positives and negatives in discrete 10% steps. The probability $P_{observed}$ was then recalculated as $density_{positives}/(density_{positives} + 20 \times density_{negatives})$, with 20 imitating the frequency of non-coiled-coil residues as compared to coiled-coil residues.

3. Results and discussion

3.1. Analyzing the performance of COILS and PCOILS

3.1.1. The ideal matrix

Fig. 2(a) illustrates the effect of the choice of matrix on the performance of COILS and PCOILS. In COILS, the iterated matrix, which is the most recent one and also includes the most sequence data, performs best and the MTK matrix, which is the oldest one and derived from the smallest dataset performs worst. However, the difference is not as large as one might have expected, suggesting that much of the positional information inherent in coiled-coil sequences can be gleaned from a dataset as small as a few thousand residues. In PCOILS, the difference in performance between matrices seems largely erased, with the two new matrices barely having an edge at high reliabilities. We propose that this is due to the additional information added by the profiles, which effectively reduces the idiosyncrasies of the older matrices resulting from their limited training set.

3.1.2. Weighting

When originally introduced into COILS, weighting appeared to slightly decrease the performance of the program (Lupas, 1996). We do not confirm this finding with the structure-based database used here. If anything the weighted mode very slightly outperforms the unweighted one for all matrices tested. This is not the case for PCOILS, where weighting clearly lowers performance for the PDB and iterated matrices at high reliabilities. The reason for this seems to be that some coiled-coil proteins such as Colicin E3 or Arfaptin have unusual core residues, for which the scores in the matrices differ, especially in the PDB matrix with its few data. This effect is multiplied if the core residues are weighted. Since the justification for introducing the weighting option was the detection of highly charged false positives (see Section 2), but such segments are not represented in our database, we show the case of the KEKE motif in PA28 (PDB code: 1AVO) as an example of weighted versus unweighted prediction (Fig. 4). PCOILS provides a better detection of the KEKE motif as a false positive and in both methods the iterated matrix performs substantially worse than the other three matrices. Given the observations described here, we have set weighting and the

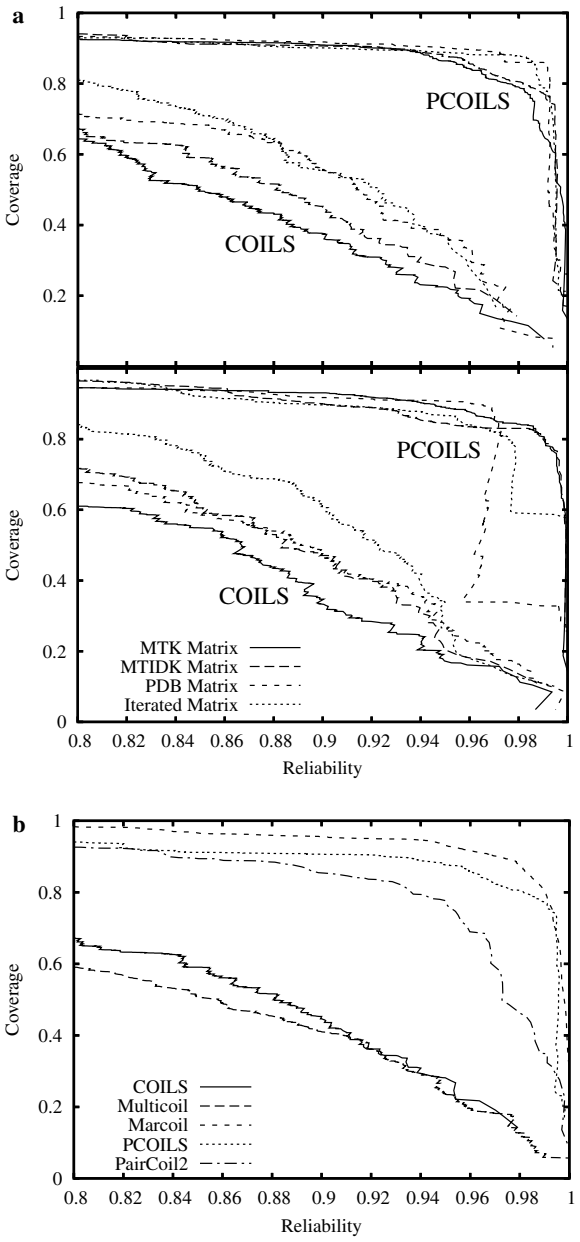

Fig. 2. Comparison of CC prediction tools for reliabilites of at least 80% on a set of known CC structures. (a) Unweighted (upper panel) and weighted (lower panel) performance of COILS and PCOILS. (b) General benchmark of all programs (COILS and PCOILS unweighted with the MTIDK matrix). Note that PairCoil2 included sequences from the test set in its training data.

MTIDK matrix as the default options in our web server (http://protevo.eb.tuebingen.mpg.de/coils).

3.1.3. COILS versus PCOILS

Regardless of the matrix used, PCOILS performs significantly better than COILS. At a reliability of 90% it assures a coverage of 90–93%, whereas COILS ranges between 35% and 61% depending on the parameters used. As noted above, PCOILS also becomes largely independent of the scoring matrix. However, this increase in performance is obtained at the expense of speed; whereas COILS is a very fast program, PCOILS is limited by the speed of PSI-BLAST.

3.2. Comparison of the different tools

COILS and MultiCoil show comparable performances in coverage versus reliability (Fig. 2b), with COILS maintaining a slight edge at reliabilities under 90%. From an informational standpoint, this is somewhat surprising: MultiCoil uses the same principles as COILS, but employs a much more elaborate matrix containing pairwise residue correlations and more recent data. This matrix can however only provide an improved performance to MultiCoil if enough sequences are available to fill its many more fields with statistically significant data. The improved performance of PairCoil2 suggests that this is now the case. COILS and MultiCoil do differ in one point, however, namely the way in which they assign probabilities (Fig. 3). While in COILS the reported probabilities approximately correspond to the effective probabilities, MultiCoil is too restrictive. At a reported probability of 20%, for example, the actual probability is already at 56%. This means that MultiCoil was tuned to provide 'safe' positive predictions and thus systematically underpredicts coiled-coil segments. On the positive side MultiCoil has few problems with highly charged false positives. With the recent release of PairCoil2, the PairCoil/MultiCoil developers have abandoned the use of probabilities and only report scores. With this decision, it has become impossible to compare their results with that of other prediction programs and is therefore missing from Figs. 3 and 4. It will remain to be seen whether this prediction format is accepted by the users.

COILS, MultiCoil, and to a lesser extent PairCoil2 are outperformed by Marcoil and PCOILS, with Marcoil providing a slightly higher performance than PCOILS over practically the entire range of reliabilities. We interpret this as being due to the fact that Marcoil operates without a scanning window, thus being able to calculate posterior probabilities from its hidden Markov model. This increase in flexibility allows it to customize its prediction to the size of the coiled coil under study. In contrast, PCOILS uses a

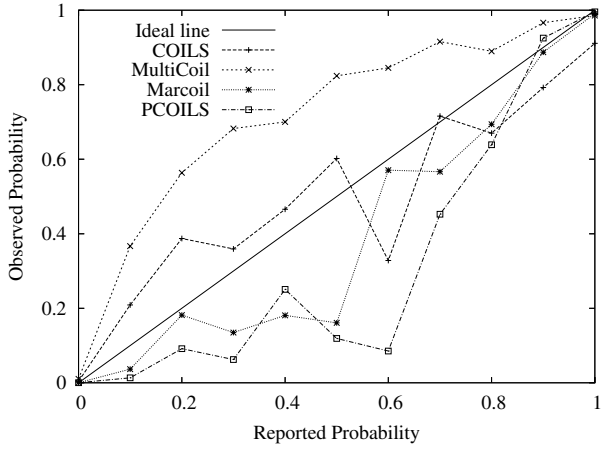

Fig. 3. Information quality of the given probabilities. Probabilities provided by MultiCoil are generally too low, while probabilities from both Marcoil and PCOILS are too high.

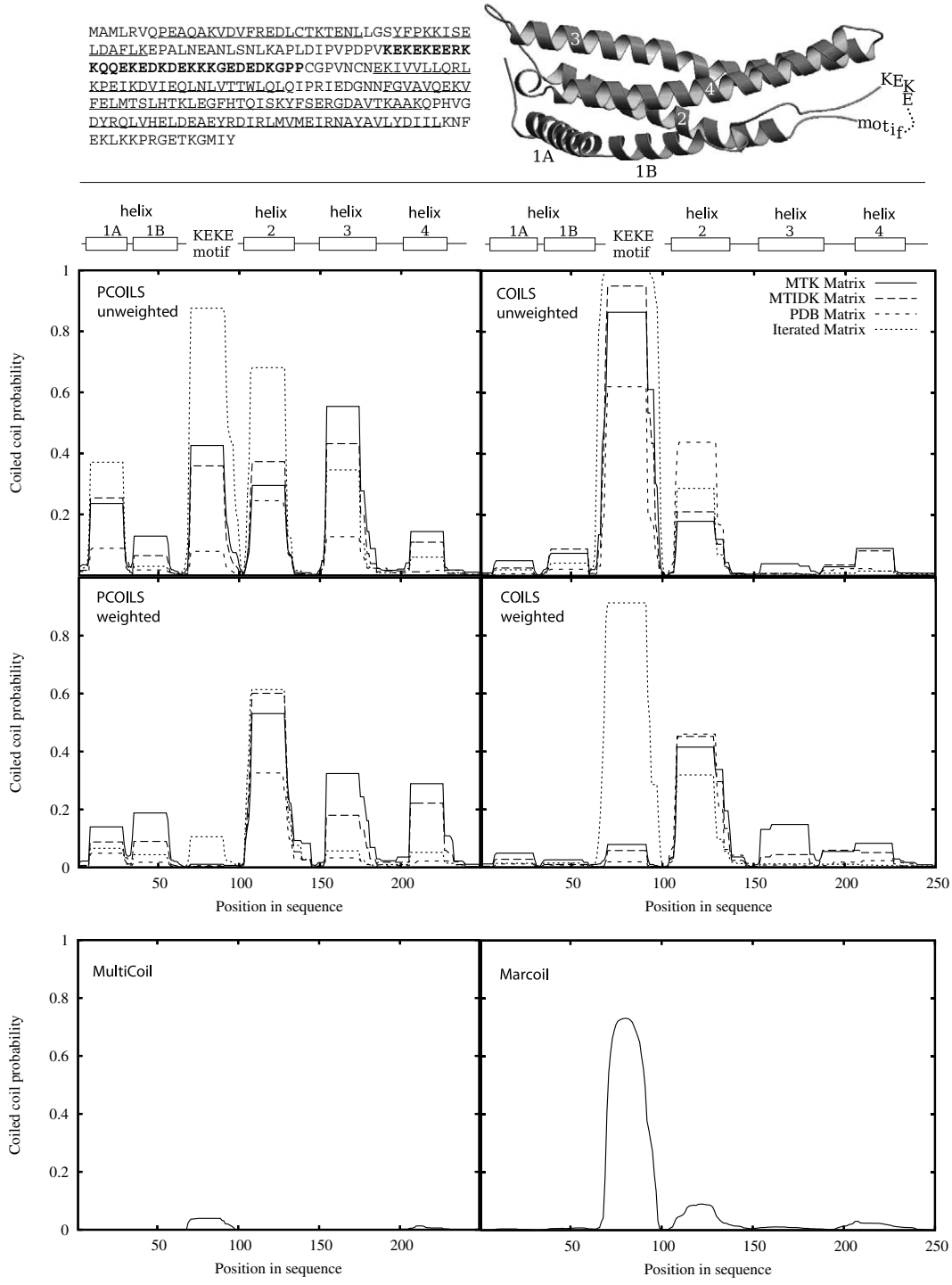

Fig. 4. Behavior of the prediction programs in case of highly charged sequence fragments as in residues 70–100 in the α-subunit of PA28 (gi accession number of the sequence: 186513, helices are underlined, the charged fragment is in bold letters, PDB code of the corresponding structure: 1AVO).

fixed window size and only operates optimally when this is the same as the size of the coiled coil. Both Marcoil and PCOILS assign probabilities too loosely; on the positive side this means that they provide 'safe' negative predictions.

3.3. Implications for users and outlook

Marcoil and PCOILS are prediction tools of a new generation, whose underlying methods significantly outperform those of older developments. Even with an updated training database, the PairCoil algorithm does not reach the performance of these two programs. Although Marcoil and PCOILS are approximately comparable, Marcoil has the substantial advantage that it is much faster, since it does not rely on time-consuming PSI-BLAST runs. In both programs, the assignment of probabilities is overly optimistic, but once this fact is known to the user, this can be taken into account. A disadvantage of Marcoil lies in its suscepti-

bility to highly charged false positives (Fig. 4), which can be dealt with effectively with the weighting option in PCOILS.

As the two best-performing programs obtain their advantage from entirely different approaches, Marcoil from calculating posterior probabilities and thereby making the sliding window obsolete and PCOILS from exploiting additional biological information, it would be desirable to combine these two traits for the next-generation coiled-coil prediction tools. Furthermore, the problem of reliably predicting coiled coils that deviate globally from a heptad periodicity remains to be addressed. In addition, although tools with a probability assignment that is too optimistic (PCOILS) or too pessimistic (MultiCoil) can help to exclude false negatives or false positives, respectively, it will be necessary in the future to provide some solid statistics, which draw a more accurate picture.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Berger, B., Singh, M., 1997. An iterative method for improved protein structural motif recognition. J. Comput. Biol. 4, 261–273.

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

Crick, F.H.C., 1953. The Fourier transform of a coiled-coil. Acta Crystallogr. 6, 685–689.

Delorenzi, M., Speed, T., 2002. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18, 617–625.

Gruber, M., Lupas, A.N., 2003. Historical review: another 50th anniversary—new periodicities in coiled coils. Trends Biochem. Sci. 28, 679–685.

Gruber, M., Soding, J., Lupas, A.N., 2005. REPPER-repeats and their periodicities in fibrous proteins. Nucleic Acids Res. 33, 239–243.

Hicks, M.R., Holberton, D.V., Kowalczyk, C., Woolfson, D.N., 1997. Coiled-coil assembly by peptides with non-heptad sequence motifs. Fold Des. 2, 149–158.

Jones, D.T., Swindells, M.B., 2002. Getting the most from PSI-BLAST. Trends Biochem. Sci. 27, 161–164.

Lupas, A., 1996. Prediction and analysis of coiled-coil structures. Methods Enzymol. 266, 513–525.

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

Lupas, A.N., Gruber, M., 2005. The structure of alpha-helical coiled coils. Adv. Protein Chem. 70, 37–78.

McDonnell, A.V., Jiang, T., Keating, A.E., Berger, B., 2006. Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics 22, 356–358.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.

Parry, D.A., 1982. Coiled-coils in α-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. 2, 1017–1024.

Singh, M., Berger, B., Kim, P.S., 1999. LearnCoil-VMF: computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins. J. Mol. Biol. 290, 1031–1041.

Singh, M., Berger, B., Kim, P.S., Berger, J.M., Cochran, A.G., 1998. Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. Proc. Natl. Acad. Sci. USA 95, 2738–2743.

Soding, J., 2005. Protein homology detection by HMM–HMM comparison. Bioinformatics 21, 951–960.

Walshaw, J., Woolfson, D.N., 2001. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. J. Mol. Biol. 307, 1427–1450.

Wolf, E., Kim, P.S., Berger, B., 1997. MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci. 6, 1179–1189.