FEBS Letters 584 (2010) 1623–1627

journal homepage: www.FEBSLetters.org

Folded-unfolded cross-predictions and protein evolution: The case study of coiled-coils

Balázs Szappanos [a,b,1], Dániel Süveges [b], László Nyitray [b], András Perczel [a,c], Zoltán Gáspári [a,*]

[a] Eötvös Loránd University, Institute of Chemistry, Structural Chemistry and Biology Laboratory, Pázmány Péter s. 1/A, H1117 Budapest, Hungary
[b] Eötvös Loránd University, Department of Biochemistry, Pázmány Péter s. 1/C, H1117 Budapest, Hungary
[c] ELTE-MHAS Protein Modeling Research Group, Pázmány Péter s. 1/A, H1117 Budapest, Hungary

ARTICLE INFO

Article history:
Received 20 January 2010
Revised 13 March 2010
Accepted 15 March 2010
Available online 19 March 2010

Edited by Robert B. Russell

Keywords:
Coiled-coil
Intrinsically disordered protein
Structure prediction
Protein evolution

ABSTRACT

Here we report a thorough analysis of cross-predictions between coiled-coil and disordered protein segments using various prediction algorithms for both sequence classes. Coiled-coils are often predicted to be unstructured, consistent with their obligate multimeric nature, whereas reverse cross-predictions are rare due to the regularity of coiled-coil sequences. We propose the simultaneous use of the programs COILS and IUPRED to achieve acceptable prediction accuracy and minimize the extent of cross-predictions. The relevance of observed cross-predictions might be that disordered sequences can adopt coiled-coil conformation relatively easily during protein evolution.

1. Introduction

Coiled-coil structures, formed by α-helices wrapped around each other, have underlying sequences with characteristic regularity in the form of typical seven-residue (heptad) repeats. In the heptad pattern, *abcdefg*, the *a* and *d* positions are usually occupied by hydrophobic residues, while residues at *e* and *g* positions are often polar/charged (Fig. 1) [1,2]. Although simple and elegant in design, coiled-coils are highly diverse in stability and structural specificity [3,4]. Disordered sequences (Intrinsically Disordered Proteins, IDPs) do not adopt a well-defined three-dimensional structure in their functional monomeric state (Fig. S1), although might get folded upon interaction with binding partners [5]. Both types of segments account for a considerable fraction of prokaryotic and even more of eukaryotic proteomes [6,7]. These estimates are based on various predictions that use primary amino acid sequences as inputs. Coiled-coils can be regarded as a division of IDPs as they are disordered in their monomeric state and the stability of the superhelices formed varies over a wide range [3,8]. Proper identification of coiled-coil regions is of high importance both for structural and functional annotations: coiled-coil motifs are specifically associated with a number of cellular processes ranging from organization of the cytoskeleton and membrane fusion to transcriptional regulation [4,9]. In spite of this, to our knowledge, no systematic attempt has been made to assess the rate and significance of cross-predictions between coiled-coil and IDP segments. Many sequences characterized as Charged Single α-Helices (CSAHs), sequences adopting stable helical conformation in water, are predicted to be both IDPs and coiled-coils by various programs [10]. This raises the question whether cross-prediction occurs regularly between coiled-coils and disordered sequences. In this paper we describe a systematic evaluation of different recognition algorithms and discuss the relevance of cross-predictions in the light of structural transitions during protein evolution.

2. Methods

Detailed description of the methods and databases used can be found in the Supplementary data. We have tested six coiled-coil and seven IDP predictor programs: AMPHISEARCH [11], COILS [12], MARCOIL [13], MULTICOIL [14], PAIRCOIL2 [15] and PCOILS [16] as well as DISEMBL (hot loops, missing coordinates) [17], FOLDINDEX [18], GLOBPLOT2 [19], IUPRED [20], RONN [21], VSL2B [22]. An in-house coiled-coil database (ccDB) was built based on the PDB Select archive (2007 October release 25% filtered list, [23] and the SOCKET program [24], similarly to the recently published CC + database [25]. For disordered sequences, the DisProt database (version 4.5, [26]) was used. For comprehensive analysis, we used Swiss-Prot

* Corresponding author. Fax: +36 1 3722620.
E-mail address: szpari@chem.elte.hu (Z. Gáspári).
[1] Present address: Evolutionary Systems Biology Group, Institute of Biochemistry, Biological Research Center, Temesvári krt. 62, H6701 Szeged, Hungary.

0014-5793/$36.00
doi:10.1016/j.febslet.2010.03.026

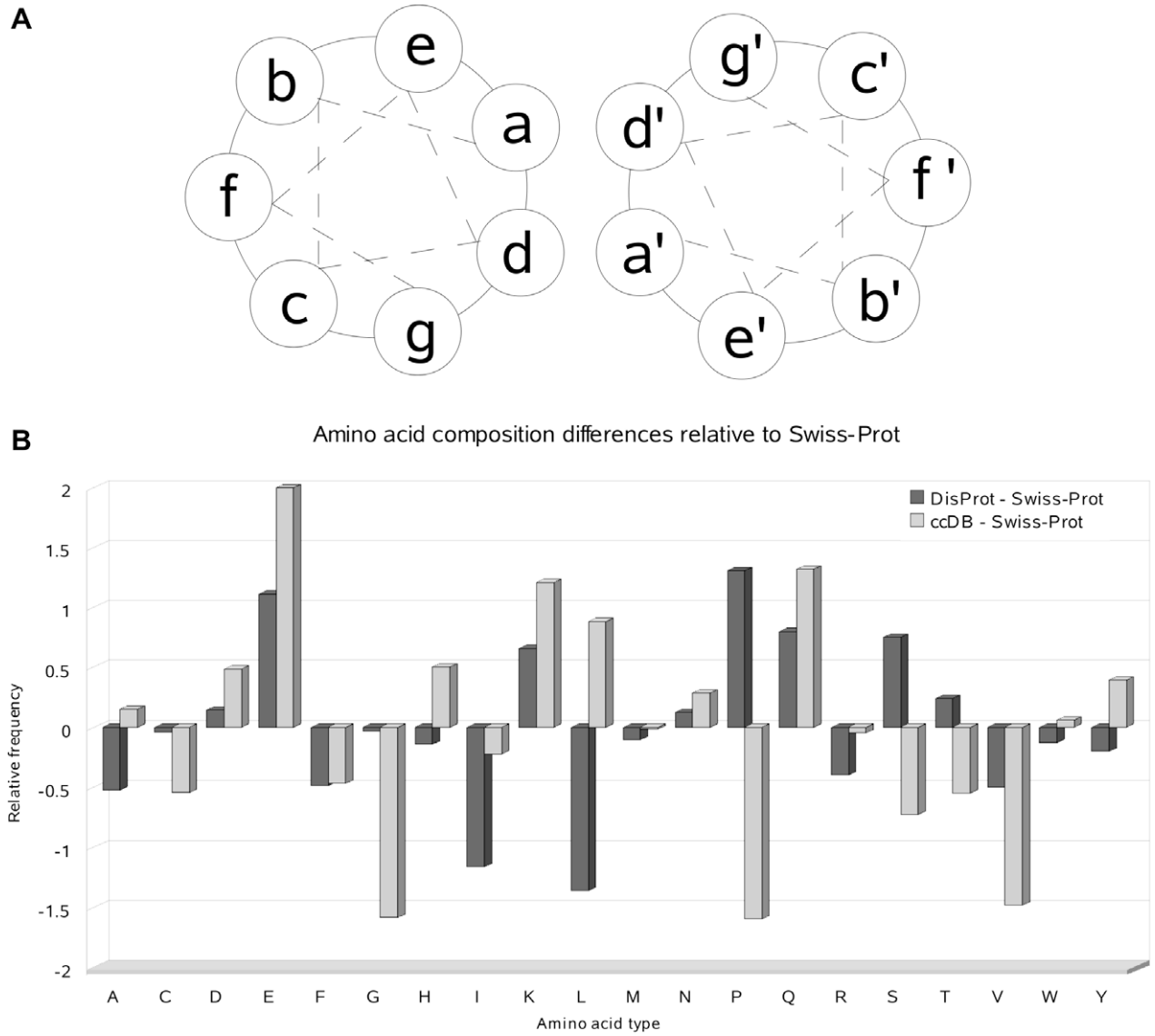

Fig. 1. (A) Helical wheel diagram of a two-stranded parallel coiled-coil with the heptad positions denoted *a–f* and *a′–f′* in the two strands, respectively. (B) Differences of amino acid frequencies in coiled-coil and disordered sequences relative to Swiss–Prot based on the databases used in the study.

(version 55) filtered with the program CD-HIT [27] to exclude sequences with over 90% similarity to each other. To assess prediction performance, we used the sensitivity (TP/(TP + FN)) and specificity (TN/(FP + TN)) measures, where TP stands for true positives, TN for true negatives, FP for false positives and FN for false negatives. We have used identical criteria for the analysis on the ccDB and DisProt databases for both types of programs, i.e. when running coiled-coil predictions on DisProt, disordered residues predicted to be in coiled-coils were treated as TPs (i.e. we tested the performance of coiled-coil predictor programs as IDP predictor algorithms in this case). Each program was also evaluated against randomized predictions (see Supplementary data). All programs were tested on both the ccDB and DisProt databases, as well as on Swiss-Prot. To assess cross-predictions, we used the Segment Overlap measure (SOV, [28]), evaluating the overlaps of predicted segments both with respect to the output of each coiled-coil and also to that of each IDP predictor program.

3. Results

Most programs investigated are based on more specific features than amino acid composition, in particular, coiled-coil predictor algorithms rely on heptad repeats as a signature of such sequences. Nevertheless, a birds-eye picture focusing only on the amino acid compositions shows characteristic differences between coiled-coils, IDPs and CSAHs (Figs. 1 and S1), as well as between these and the average proteins in the Swiss-Prot (version 55) and PDB SELECT (2007 October release) databases. To obtain a reliable picture of the predictive power of the seven IDP and six coiled-coil predictor algorithms tested, we performed benchmark tests on the DisProt database [26] and a coiled-coil sequence set termed ccDB (see Section 2).

The performance of coiled-coil predictors is much more balanced than that of disordered predictors. Generally, coiled-coil prediction programs have relatively low sensitivity (i.e. they recognize less than half of the coiled-coil residues in our ccDB database) and high specificity (they do not mispredict residues in other structures as coiled-coils). In contrast, IDP predictions vary much more in their performance on the DisProt database, with generally higher sensitivity and lower specificity values than coiled-coil predictions on the ccDB dataset (Fig. 2, Tables S1–S3). Cross-predictions were evaluated by multiple methods, the simplest of which is swapping the data sets and the programs (i.e. a true positive (TP) hit is counted when a disorder predictor program finds a coiled-coil res-

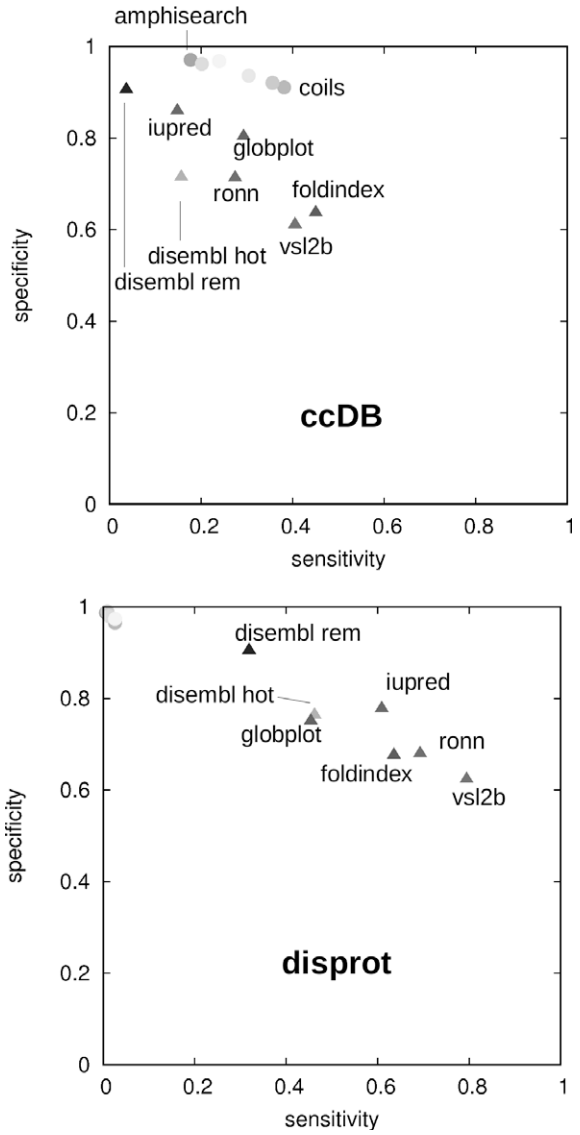

Fig. 2. Performance of the predictors used in this study. Top: sensitivity–specificity plot of all programs evaluated as coiled-coil prediction algorithms on the ccDB database. Bottom: sensitivity–specificity plot of all programs evaluated as IDP prediction algorithms on the DisProt database. Data points corresponding to coiled-coil prediction programs are shown as circles whereas those of IDP prediction algorithms as triangles. To ensure readability, not all points are labeled. 'Disembl hot' and 'disembl rem' stand for DisEmbl algorithms based on 'hot loops' and 'missing coordinates', respectively. Higher specificity and sensitivity values indicate more accurate prediction.

idue in the ccDB database and when a coiled-coil prediction algorithm detects a disordered one in the DisProt dataset). Disorder prediction algorithms are found to be comparable in sensitivity but not in specificity to coiled-coil predictors on the ccDB database (Fig. 2). In contrast, virtually no residues are recognized as coiled-coils by their respective predictor programs in the DisProt database.

Our general tests on a homology-filtered version of the SWISS-PROT database also show that cross-prediction of coiled-coils as unstructured regions occur in a remarkable number of cases, whereas the reverse is much more rare (Fig. 3). In particular, regions predicted to be in coiled-coils by PAIRCOIL2 and PCOILS often overlap with IDP regions recognized by FOLDINDEX, RONN and VSL2B both on the ccDB and Swiss-Prot databases. On the other hand, GLOBPLOT and IUPRED predicted segments that have the least overlap with predicted coiled-coil regions.

4. Discussion

Our benchmarks reveal that, in general, coiled-coils are predicted to be disordered sequences at a comparable rate as IDP segments in DisProt while only few disordered segments are predicted to form coiled-coils (Fig. 2 and Table S1). This latter observation is in line with cross-predictions reported [29] using the programs PONDR VL-XT [30] and COILS. While our results are not unexpected given the higher regularity and thus better predictability of coiled-coils, they nevertheless emphasize the need for caution in interpreting coiled-coil and IDP predictions on whole proteomes.

With respect to cross-predictions, it is worthwhile to investigate which predictor pairs yield the best results when applied together. To this end, both the performance of the programs in the benchmarks and the cross-predictions produced should be evaluated. We propose the use of the COILS-IUPRED predictor pairs based on the relatively high sensitivity of Coils (over 0.37; the specificity of all coiled-coil predictor programs are high, Fig. 2) and robust performance of IUPRED in the benchmark tests (the most specific among methods with sensitivity over 0.5). Furthermore, these two programs yield relatively few cross-predictions (Fig. 3). This recommendation is also supported by the performance of these programs relative to randomized predictions (Tables S2 and S3).

The Janus-faced prediction of certain protein segments both as coiled-coil and disordered motifs could be, at least in some cases, valid with important structural and functional consequences [31]. Short coiled-coils could be unstable, monomeric and hence disordered; however, binding of a homodimeric protein nearby could promote dimerization and switch the structure of the disordered segment to a coiled-coil dimer without direct interaction. The LC8 dynein light chain (DYNLL) has recently been suggested to function in such a chaperon-like manner inducing coiled-coil dimerization [32–34]. Partner-binding induced folding is a key process both in coiled-coil formation and in the function of molecular recognition elements (MoREs) in IDPs. Our analysis involving the ANCHOR server for the identification of protein binding sites in disordered regions [35] revealed that coiled-coil prediction programs practically do not recognize such regions (Table S4), thus, the majority of the observed cross-predictions is unlikely to correspond to MoREs.

Disordered proteins are thought to be able to evolve through tandem repeat expansion [36], and expanded repeats can then be disrupted by mutations, causing divergence even in closely related genomes [37]. Moreover, IDP sequences tolerate a relatively high number of amino acid substitutions [38], providing rapid evolvability [39]. In line with this, cross-predictions observed in this study might indicate that disordered segments can relatively easily be converted to coiled-coils by amino acid substitutions, "switching" them to a different structural class. The most remarkable difference between the two segment types in terms of amino acid composition is in the abundance of leucine and proline residues (Fig. 1). Interestingly, these two residues are coded by neighbouring boxes in the standard genetic code (the CUN and CCN boxes, respectively), which raises the possibility that the two segment types could be interconverted by suitably positioned point mutations (transitions). Naturally, amino acid abundance alone does account neither for homomeric amino acid stretches nor heptad repeats characteristic for IDPs and coiled-coils, respectively. We note that the amino acid distribution in segments predicted to be both coiled-coils and IDPs in Swiss-Prot represent, on average, a transition between those characteristic of coiled-coils and CSAHs (Fig. S1), although the five longest such segments could not be identified as CSAHs (Table S5). Moreover, CSAHs might be considered specific "monomeric coiled-coils", and indeed they are often

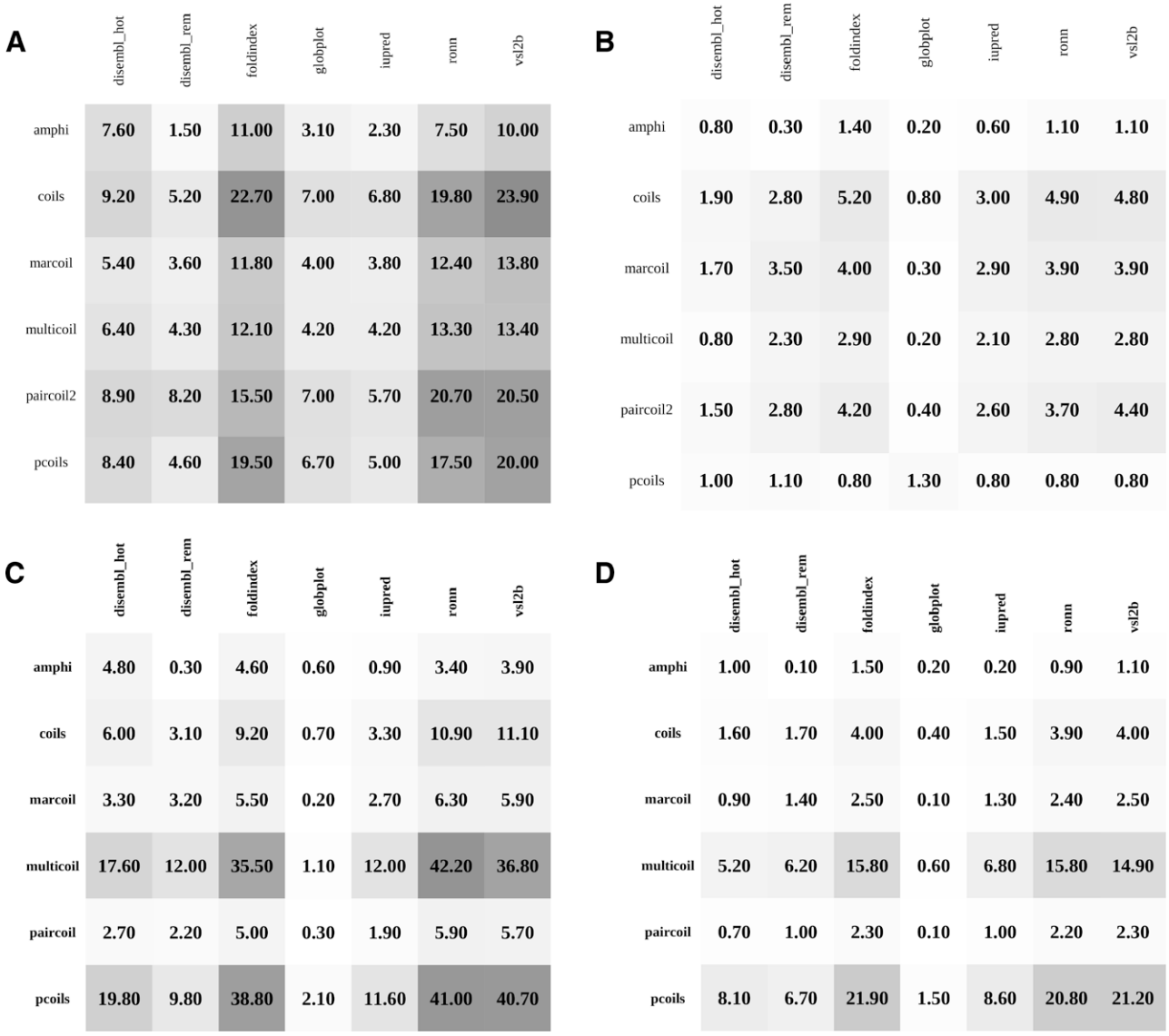

Fig. 3. Segment overlaps of cross-predictions. (A) Overlaps with predicted coiled-coil sequences on the ccDB database, (B) overlaps with predicted IDP sequences on the Disprot database, (C) overlaps with predicted coiled-coil sequences on the Swiss-Prot database, (D) overlaps with predicted IDP sequences on the Swiss-Prot database. 'Disembl_hot' and 'disembl_rem' stand for DisEmbl algorithms based on 'hot loops' and 'missing coordinates', respectively, whereas 'amphi' denotes AmphiSearch. Weighted SOV(x) values are shown (see Supplementary data). Lower values and lighter box colors correspond to lower degree of cross-predictions.

mispredicted as coiled-coil segments by coiled-coil predictors. Moreover, they were found to replace coiled-coil segments in several homologous proteins [10], suggesting interconversion during protein evolution.

We conclude that there is no simple scenario for the interpretation of all cross-predictions. They might be mispredictions, indicative of CSAHs or otherwise functionally relevant. Such segments provide an example of the twilight zone between protein order and disorder [40] and some of them might even represent an evolutionary transition between different protein structural states [41].

Acknowledgements

This work was supported by the Hungarian Scientific Research Fund (OTKA F68079, K72973, K61784 and NI68466) and ICGEB (CRP/HUN09-03). Z.G. also acknowledges a János Bolyai Research Fellowship.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2010.03.026.

References

[1] Lupas, A.N. and Gruber, M. (2005) The structure of alpha-helical coiled coils. Adv. Protein Chem. 70, 37–78.
[2] Woolfson, D.N. (2005) The design of coiled-coil structures and assemblies. Adv. Protein Chem. 70, 79–112.
[3] Grigoryan, G. and Keating, A.E. (2008) Structural specificity in coiled-coil interactions. Curr. Opin. Struct. Biol. 18, 477–483.
[4] Rose, A. and Meier, I. (2004) Scaffolds, levers, rods and springs: diverse cellular functions of long coiled-coil proteins. Cell Mol. Life Sci. 61, 1996–2009.
[5] Tompa, P. (2002) Intrinsically unstructured proteins. Trends Biochem. Sci. 27, 527–533.
[6] Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C. and Brown, C.J. (2000) Intrinsic protein disorder in complete genomes. Genome Inform. Ser. Workshop Genome Inform. 11, 161–171.
[7] Liu, J. and Rost, B. (2001) Comparing function and structure between entire proteomes. Protein Sci. 10, 1970–1979.

[8] Li, Y., Brown, J.H., Reshetnikova, L., Blazsek, A., Farkas, L., Nyitray, L. and Cohen, C. (2003) Visualization of an unstable coiled coil from the scallop myosin rod. Nature 424, 341–345.
[9] Burkhard, P., Stetefeld, J. and Strelkov, S.V. (2001) Coiled coils: a highly versatile protein folding motif. Trends Cell Biol. 11, 82–88.
[10] Suveges, D., Gaspari, Z., Toth, G. and Nyitray, L. (2009) Charged single alpha-helix: a versatile protein structural motif. Proteins 74, 905–916.
[11] Morii, H., Takenawa, T., Arisaka, F. and Shimizu, T. (1997) Identification of kinesin neck region as a stable alpha-helical coiled coil and its thermodynamic characterization. Biochemistry 36, 1933–1942.
[12] Lupas, A., Van Dyke, M. and Stock, J. (1991) Predicting coiled coils from protein sequences. Science 252, 1162–1164.
[13] Delorenzi, M. and Speed, T. (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18, 617–625.
[14] Wolf, E., Kim, P.S. and Berger, B. (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci. 6, 1179–1189.
[15] McDonnell, A.V., Jiang, T., Keating, A.E. and Berger, B. (2006) Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics 22, 356–358.
[16] Gruber, M., Soding, J. and Lupas, A.N. (2005) REPPER-repeats and their periodicities in fibrous proteins. Nucleic Acids Res. 33, W239–W243.
[17] Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. (2003) Protein disorder prediction: implications for structural proteomics. Structure 11, 1453–1459.
[18] Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I. and Sussman, J.L. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21, 3435–3438.
[19] Linding, R., Russell, R.B., Neduva, V. and Gibson, T.J. (2003) GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res. 31, 3701–3708.
[20] Dosztanyi, Z., Csizmok, V., Tompa, P. and Simon, I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol. 347, 827–839.
[21] Yang, Z.R., Thomson, R., McNeil, P. and Esnouf, R.M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21, 3369–3376.
[22] Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P. and Dunker, A.K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 61 (Suppl. 7), 176–182.
[23] Hobohm, U. and Sander, C. (1994) Enlarged representative set of protein structures. Protein Sci. 3, 522–524.
[24] Walshaw, J. and Woolfson, D.N. (2001) Socket: a program for identifying and analysing coiled-coil motifs within protein structures. J. Mol. Biol. 307, 1427–1450.
[25] Testa, O.D., Moutevelis, E. and Woolfson, D.N. (2009) CC+: a relational database of coiled-coil structures. Nucleic Acids Res. 37, D315–D322.
[26] Sickmeier, M. et al. (2007) DisProt: the database of disordered proteins. Nucleic Acids Res. 35, D786–D793.
[27] Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.
[28] Rost, B., Sander, C. and Schneider, R. (1994) Redefining the goals of protein secondary structure prediction. J. Mol. Biol. 235, 13–26.
[29] Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z. and Dunker, A.K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol. 323, 573–584.
[30] Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J. and Dunker, A.K. (2001) Sequence complexity of disordered protein. Proteins 42, 38–48.
[31] Gazi, A.D., Bastaki, M., Charova, S.N., Gkougkoulia, E.A., Kapellios, E.A., Panopoulos, N.J. and Kokkinidis, M. (2008) Evidence for a coiled-coil interaction mode of disordered proteins from bacterial type III secretion systems. J. Biol. Chem. 283, 34062–34068.
[32] Barbar, E. (2008) Dynein light chain LC8 is a dimerization hub essential in diverse protein networks. Biochemistry 47, 503–508.
[33] Hodi, Z., Nemeth, A.L., Radnai, L., Hetenyi, C., Schlett, K., Bodor, A., Perczel, A. and Nyitray, L. (2006) Alternatively spliced exon B of myosin Va is essential for binding the tail-associated light chain shared by dynein. Biochemistry 45, 12582–12595.
[34] Hodi, Z. et al. (2007) The LC8 family of dynein light chains: multifunctional chaperon-like proteins. FEBS J. 274, 106.
[35] Meszaros, B., Simon, I. and Dosztanyi, Z. (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput. Biol. 5, e1000376.
[36] Tompa, P. (2003) Intrinsically unstructured proteins evolve by repeat expansion. Bioessays 25, 847–855.
[37] Gaspari, Z., Ortutay, C. and Toth, G. (2007) Divergent microsatellite evolution in the human and chimpanzee lineages. FEBS Lett. 581, 2523–2526.
[38] Tóth-Petróczy, Á., Mészáros, B., Simon, I., Dunker, A.K., Uversky, V.N. and Fuxreiter, M. (2008) Assessing conservation of disordered regions in proteins. Open Proteom. J. 1, 46–53.
[39] Pal, C., Papp, B. and Lercher, M.J. (2006) An integrated view of protein evolution. Nat. Rev. Genet. 7, 337–348.
[40] Szilagyi, A., Gyorffy, D. and Zavodszky, P. (2008) The twilight zone between protein order and disorder. Biophys. J. 95, 1612–1626.
[41] Tokuriki, N. and Tawfik, D.S. (2009) Protein dynamism and evolvability. Science 324, 203–207.