Protein secondary structure: category assignment and predictability

Claus A. Andersen[a], Henrik Bohr[b], Søren Brunak[a,]*

[a] *Center for Biological Sequence Analysis, BioCentrum, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark*
[b] *Department of Physics, Building 307, Technical University of Denmark, DK-2800 Lyngby, Denmark*

Received 9 July 2001; revised 4 September 2001; accepted 5 September 2001

First published online 27 September 2001

Edited by Amy M. McGough

Abstract In the last decade, the prediction of protein secondary structure has been optimized using essentially one and the same assignment scheme known as DSSP. We present here a different scheme, which is more predictable. This scheme predicts directly the hydrogen bonds, which stabilize the secondary structures. Single sequence prediction of the new three category assignment gives an overall prediction improvement of 3.1% and 5.1% compared to the DSSP assignment and schemes where the helix category consists of α-helix and 3_{10}-helix, respectively. These results were achieved using a standard feed-forward neural network with one hidden layer on a data set identical to the one used in earlier work. © 2001 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Sequence–structure relationship; Hydrogen bond prediction

1. Introduction

Up until now most protein secondary prediction setups have used the assignment scheme developed in the seminal work of Kabsch and Sander [1] known as DSSP (Dictionary of Secondary Structure of Proteins). The DSSP program converts atomic coordinates, as determined by X-ray crystallography or nuclear magnetic resonance, into backbone hydrogen bonds between donors and acceptors, and transforms subsequently repetitive bonding patterns into categories of helices, sheets and turns.

Pauling et al. predicted in 1951 the idealized protein secondary structures: α-helices [2] and β-sheets [3], based on the conformations of intra-backbone hydrogen bonds. In natural proteins, around 68% of the protein hydrogen bonds are intra-backbone hydrogen bonds [4], the rest being side-chain–backbone and inter-side-chain bonds, which do not form a basis apt for secondary structure definitions. The energy stored in a good hydrogen bond is approximately 2 kcal/mol, exceeding by a factor of 10 the van der Waals energy [5], which means that the energies stored in hydrogen bond interactions are dominant at the secondary structure level, except for S–S bridges.

The conventional strategies towards prediction are not designed to exploit directly the systematics in the recognition between hydrogen donors and acceptors, but are designed to approximate the mapping between sequence space and a set of *visually* appealing structural categories. Each structural category of protein secondary structure is assigned *consecutively* to the linear sequence. This means that amino acid residues are assigned to a particular category even if they are not involved in maintaining the structure by backbone hydrogen bonds. In order to maximize the predictability of the assignment scheme it is essential not to contaminate a category with examples not sharing the prevailing and common characteristics at the sequence level of a given structural class, which we will show is the case for the visually appealing structural categories assigned by DSSP.

2. Materials and methods

2.1. Neural network algorithm and performance evaluation

A standard three layer feed-forward neural network has been employed for structure prediction [6,7] ($N_{\text{input}} = 13 \times 21$; $N_{\text{hidden}} = 20$; $N_{\text{output}} = 3$) to verify the importance of predicting the hydrogen bonds. The amino acids were sparsely encoded and represented by binary strings of 21 bits, i.e. 20 bits for the amino acid type and 1 bit to specify N- and C-terminal ends. An input window of 13 amino acids was found to give optimal performance with the hydrogen bond prediction scheme, which is identical to the optimal window size when predicting the DSSP assignments [8]. Seven-fold cross-validation has been performed in the prediction performance measurements. During training all categories were presented equally often, termed balanced training.

2.2. Sequence data

As usual in comparisons of prediction schemes a set of proteins with sequence similarity below 25% was constructed from the entries in PDB. This set is identical to the classical one used by other groups [8,9], having 23 345 amino acids in 126 proteins or protein chains. It contains 30 anomalies such as chain breaks, GLX amino acids, etc.

2.3. Information content

To investigate characteristics at the sequence level we used a first order correlation analysis tool as quantified by the Kullback–Leibler information on a selected set of aligned sequence windows. The Kullback–Leibler information is defined as

$$I(i) = -\sum_k p_k^i \log_2 \left(\frac{p_k^i}{p_k^{'}} \right)$$

where k runs over the amino acid alphabet, p_k^i is the propensity for amino acid k at position i in a multiple alignment of sequence segments, and $p_k^{'}$ is the background propensities for the data set. This signal to noise ratio analysis can be visualized using sequence logos of the amino acids [10]. The logos are shown with a window of size 13, which is the window size used in the optimal neural networks. The height at each sequence position is proportional to the information content at that position.

*Corresponding author. Fax: (45)-45931585.
E-mail address: brunak@cbs.dtu.dk (S. Brunak).

Abbreviations: PDB, Protein Data Bank of 3D coordinates; DSSP, Dictionary of Secondary Structure of Proteins

PII: S0014-5793(01)02910-6

2.4. Correlation coefficient

As a consequence of the imbalance between the categories of secondary structure, mere percentages of correctly predicted window configurations can be bad indicators of predictive performance. An alternative measure which takes the relation between correctly predicted positives (TP) and negatives (TN) as well as false positives (FP) and negatives (FN) into account is Matthews correlation coefficient [11],

$$C_{\mathrm{cat}} = \frac{(\mathrm{TP}\cdot\mathrm{TN})-(\mathrm{FP}\cdot\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (1)$$

where the positive category is either α-helix, β-sheet or coil.

2.5. Hydrogen bond

The hydrogen bond definition used is the Coulomb energy (see Fig. 1) implemented in DSSP:

$$E = f\delta^{+}\delta^{-}\left(\frac{1}{r_{\mathrm{NO}}}+\frac{1}{r_{\mathrm{HC'}}}-\frac{1}{r_{\mathrm{HO}}}-\frac{1}{r_{\mathrm{NC'}}}\right) \quad (2)$$

where $f = 332$ Å kcal e^{-2} mol^{-1} is the dimensional factor and $\delta^{+}\delta^{-}$ are the polar charges given in elementary charges e. A cut-off level has been set for the weakest acceptable H bond at $E < -0.5$ kcal/mol in DSSP.

3. Results

The aim was to make a new 'clean' hydrogen bond following scheme, and compare it to the conventional three category DSSP scheme using exactly the same cross-validation approach and data as previous studies [8].

3.1. Assigning the structural elements

In order to apply Paulings idealized definitions on structural data further specifications are needed. The backbone hydrogen bonds sustaining an α-helix connect amino acid i with the amino acid four positions down (acceptor) and/or up (donor) in the protein sequence. This gives us two hydrogen bonding categories for α-helices, i.e. acceptors and donors. By inspecting a logo of the $i{-}4 \leftarrow i$ category (see Fig. 2a) it is directly seen that the information peaks at the donor position, making us choose to predict this category. In the $i{-}4 \leftarrow i$ category (II4), it is not clear, a priori, what the requirement should be. We have tested three definitions having just one, two and three consecutive $i{-}4 \leftarrow i$ hydrogen bonds as a minimum requirement to end an α-helix, by making a neural network prediction of each of them. The neural network with the best performing definition had two consecutive $i{-}4 \leftarrow i$ hydrogen bonds in its α-helix definition which is in agreement with Kabsch and Sander's definition in DSSP.

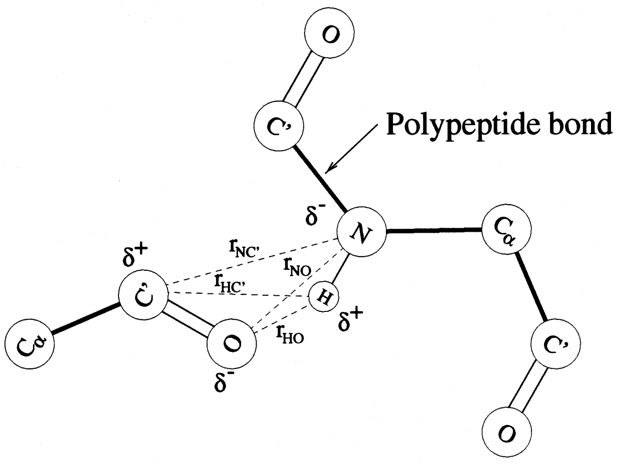


Fig. 1. The distances used to calculate the Coulomb H bond.

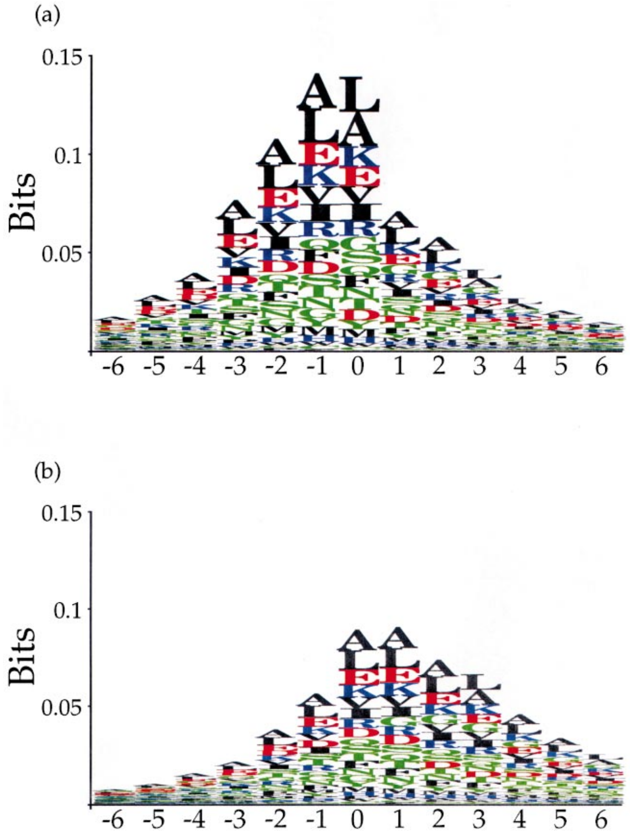
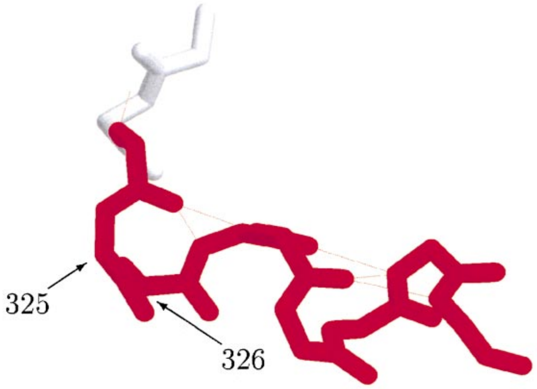
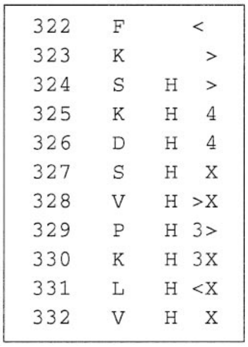


Fig. 2. Kullback–Leibler information content in 13 amino acid windows. a: Amino acids having $i{-}4 \leftarrow i$ hydrogen bonds forming α-helices at position 0. b: DSSP 'H' assigned amino acids at position 0. The signal in panel a is stronger than that in panel b, and the top is centered over the assigned amino acid, showing that the first order correlations are higher and that the most correlated amino acid with regard to the assignment is the one assigned. The logos are basically in accordance with other statistical analyses [21], e.g. showing the well known fact that alanine and leucine occur at a markedly higher rate in α-helices.

We expected the single $i{-}4 \leftarrow i$ hydrogen bonds to contain a weaker sequence-to-structure signal, because they can be obtained by a simple turn of the backbone, which is not related to a helix (e.g. hairpin turns or some β-turns). Initiating the α-helix with two instead of three hydrogen bonds is harder to explain. The α-helix length distribution is bell shaped except for length 4 helices, which are observed excessively often [12,13]. These short helices contain only two successive $i{-}4 \leftarrow i$ hydrogen bonds, but are probably a mixture of normal helices (with a standard sequence-to-structure signal) and random double turns. The optimality of the two bond definition indicates that the cost would be too high if these short helices were categorized as coil.

The II4 category is not a sub-category of the DSSP 'H', since in the C-terminal the II4 category ends one residue after the 'H' assignment and only hydrogen bonding amino acids are included in the II4 category. Every amino acid in an ideal (fully hydrogen bonded) α-helix longer than five residues has $i \rightarrow i{+}4$ and/or $i{-}4 \leftarrow i$ hydrogen bonds, but not all α-helices are ideal. This means that disruptions in the hydrogen bond-





```
322   F      <
323   K        >
324   S    H   >
325   K    H   4
326   D    H   4
327   S    H   X
328   V    H  >X
329   P    H  3>
330   K    H  3X
331   L    H  <X
332   V    H   X
```

Fig. 3. This example shows two amino acids (#325 and #326), which have been assigned ‘H’ by DSSP, but without hydrogen bonds in the helix. On the left is a view of the backbone structure and hydrogen bonds, and on the right a cut-out of the corresponding DSSP assignments ($>$ indicates a $i \rightarrow i+(3 \text{ or } 4)$ hydrogen bond, $<$ indicates an $i-(3 \text{ or } 4) \leftarrow i$ hydrogen bond and X indicates both). The protein shown has the PDB name 8adh.

ing pattern inside a helix are ignored for the ‘H’ assignment in DSSP (see Fig. 3). It is therefore clear that the II4 assignment can be converted into the DSSP ‘H’ assignment, but *not* back again.

Corresponding to the DSSP ‘G’ (3_{10}) and ‘I’ (π) helices, other categories, II3 and II5, containing residues having $i \rightarrow i+3(i+5)$ hydrogen bonds may be defined. In earlier work ‘G’ and ‘H’ were most often merged into one category all together, which has contaminated the common characteristics seen in α-helices. They are in fact distinct types: the 3_{10}-helices are short having an average length of approximately 3.5 amino acids (using the DSSP assignment scheme) and constitute approximately 12% of the helix class; the α-helices have an average length of approximately 11 amino acids and constitute almost all the remaining 88%. A comparison of the logos for 3_{10}- (Fig. 4) and α-helices (Fig. 2b) shows that they are distinct, which explains why the signal in a mixed helix category is much harder to predict than that of α-helices. This has also been shown previously [14].

Predicting the actual hydrogen bonds instead of a visually appealing category can also be done for β-sheets. We have tried two options: one, where amino acids are required to have two hydrogen bonds in the sheet structure (E-hard) and a second one also including amino acids having one bond only in the sheet (E-med). The two distinct types of β-sheets, parallel and anti-parallel, have different bonding patterns, but the above stated definitions can be applied to both. Disruptions in β-sheets (β-bulges) are of some concern as they occur for approximately 4% of the amino acids assigned ‘E’ by DSSP and on average in approximately 15% of the DSSP assigned sheets. The latter amino acids will not be assigned as sheet in the hydrogen bonding scheme applied here, since they do not have hydrogen bonds in the β-sheet conformation. However, this assignment method is distinct from and not convertible into the one used in DSSP.

The neural network is trained to predict the hydrogen bonding scheme described above, but before the prediction performance is calculated we performed simple post-processing, for reasons of comparison. For α-helix prediction, single $i-4 \leftarrow i$ predictions were removed and the rest converted into the conventional ‘H’ assignment scheme. For β-sheet prediction gaps between two assignments were filled out (i.e. E-E → EEE). A similar post-processing was performed on predictions of the DSSP assignments, i.e. lone assignments removed and holes patched.

3.2. Statistical comparison

To see how often non-hydrogen bonding amino acids are assigned to a structure in proteins, we made some statistics on the data set. The conventional ‘H’ assignment of α-helices by

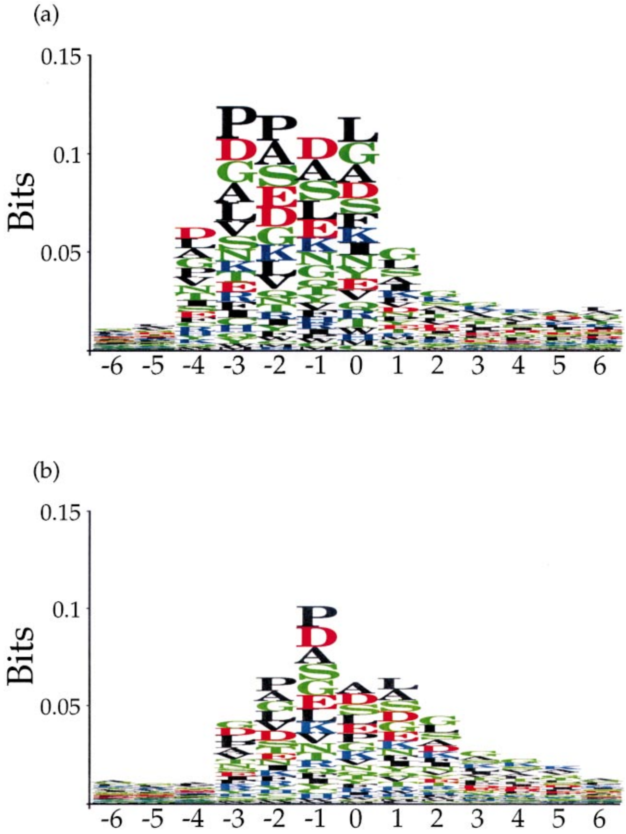


Fig. 4. The Kullback–Leibler information logo of (a) $i-3 \leftarrow i$ hydrogen bonds forming 3_{10}-helices and (b) DSSP ‘G’ 3_{10}-helices is shown here. The amino acid distribution is seen to be dominated by proline (P) in the beginning of the helix (position −2, −1) and aspartic acid (D) closer towards the helix center, which is quite different from the amino acid distributions in α-helices (see Fig. 2).

Table 1
Prediction results

	Q_3^{tot}	C_α	Q_α^{p}	Q_α^{o}	C_β	Q_β^{p}	Q_β^{o}	C_{coil}	$Q_{\text{coil}}^{\text{p}}$	$Q_{\text{coil}}^{\text{o}}$
II4-E-hard	64.9	0.52	62	71	0.36	42	55	0.42	78	65
II4-E-med	66.5	0.51	67	61	0.26	52	24	0.39	68	84
II4-E-B	64.1	0.52	62	72	0.37	53	50	0.42	72	67
H-E-B	63.4	0.45	58	58	0.34	44	44	0.40	77	76
R&S	60.6	0.43	62	58	0.36	45	57	–	–	–
R&S$_{\text{pro}}$	72.5	0.64	78	71	0.53	65	62	–	–	–
F&A$_{\text{pro}}$	74.8	0.61	–	–	0.45	–	–	0.44	–	–
Jones$_{\text{Xpro}}$	76.2	0.70	–	–	0.62	–	–	0.56	–	–

Q_3^{tot} is the correctly predicted assignments relative to the total number of assignments. Q_α^{o} is the percentage of residues correctly predicted to be α-helical relative to those observed to be α-helical. Q_α^{p} is the percentage of residues correctly predicted to be α-helical relative to those predicted to be α-helical. $C_{\alpha,\beta,\text{coil}}$ are the correlation coefficients. By optimizing the neural network with regard to the Q_3^{tot} (II4-E-med) instead of the correlation coefficient sum (II4-E-hard) it should be noticed that the helix and sheet predictions deteriorate considerably. R&S refers to the equivalent non-profile based results obtained by Rost and Sander [15], R&S$_{\text{pro}}$ to the corresponding results with profiles [8], and F&A to Frishman and Argos [16] having the highest percentage performance, but at the cost of lower correlation coefficients. Note that Frishman and Argos exclude four amino acid long α-helices as well as 3_{10}-helices from the helix category.

DSSP contains approximately 20% residues without hydrogen bonds from residues $i \rightarrow i+4$ and/or $i-4 \leftarrow i$. For the β-sheet category, residues without one or two hydrogen bonds in the sheet amount to approximately 33% of the DSSP 'E' category due to the composition of single strands and the occurrence of β-bulges.

The randomness in the conventional 'H' helix logos assigned by the DSSP program is much higher than the corresponding randomness in logos representing the context of residues with $i-4 \leftarrow i$ hydrogen bonds in the same helices (see Fig. 2). In the figure the information is seen to peak at the amino acids having $i-4 \leftarrow i$ hydrogen bonds.

3.3. Prediction performance

Table 1 shows that the best hydrogen bonding scheme measured by the correlation coefficient (II4-E-hard) is superior to the standard DSSP assignment (H-E-B) by 0.11 for the correlation coefficient sum and by 1.5% in Q_3^{tot} scores. As the standard DSSP assignment (H-E-B) we used the DSSP 'H' assignment for the helix category, even though others have mixed the DSSP α-helix 'H' and 3_{10}-helix 'G' [9,15]. The 'H' assignment is used to enable a fair comparison between the standard assignment scheme and the one presented here, because the inclusion of 3_{10}-helices significantly deteriorates the prediction of the helix category. The standard β-sheet assignment includes both DSSP 'E' and 'B-' (but not 'B-B') in the sheet category, which was used earlier [9,15].

We have used the hydrogen bond assignment in three variants: one where both the α-helix and β-sheet category are based on the hydrogen bonds (II4-E-hard, II4-E-med) and another where only the α-helix category is based on the hydrogen bond scheme (II4-E-B). Our performances should be compared with similar results, which do not make use of profile alignment techniques and jury networks, such as R&S [15]. The latter does nevertheless not give a completely fair comparison, since the helix category comprises both 'G' and 'H' helices. A fair comparison can be made between our results and H-E-B. The II4-E-B is seen to raise the Q by 11% compared to II4-E-hard, but gets a lower Q_{tot} due to the increase in wrong coil predictions ($Q_{\text{coil}}^{\text{p}}$), since the coil category is more than twice as large as the β-sheet category.

The prediction performance of Rost and Sander's PHD [8] secondary structure is also shown in Table 1, with and without the use of sequence profiles (R&S$_{\text{pro}}$, R&S). For comparison we also list the results reported by Frishman and Argos [16] (F&A$_{\text{pro}}$) and the performance obtained by Jones' PSIpred method [17] (Jones$_{\text{Xpro}}$ with extended profiles taken from EVA [18]), which is similar to other recent results [19,20]. The exploitation of evolutionary information in the form of sequence profiles has raised the prediction performance by 5–10%.

4. Discussion and conclusion

By following the hydrogen bond scheme originally assigned by Pauling we have included structurally important information, which was otherwise ignored. The new assignment scheme's primary success is the α-helix $i-4 \leftarrow i$ hydrogen bond assignment, which does not occlude the missing hydrogen bonds in helices.

There are a number of ways of boosting the percentage scores of the prediction of secondary structures such as employing jury networks, structure-to-structure nets and using profiles from alignment techniques. We have chosen to focus on a secondary structure definition which assigns the helix and sheet structures in a different way, which is impossible to derive from the original DSSP output categories. This means that a neural network given the DSSP assignment cannot distinguish which amino acids have hydrogen bonds and which have not. This should imply that our augmented performance will propagate through the different optimization techniques used and thus raise the final prediction results in a full scale implementation.

References

[1] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[2] Pauling, L., Corey, R.B. and Branson, H.R. (1951) Proc. Natl. Acad. Sci. USA 37, 205–211.
[3] Pauling, L. and Corey, R.B. (1951) Proc. Natl. Acad. Sci. USA 37, 729–740.
[4] Bordo, D. and Argos, P. (1994) J. Mol. Biol. 243, 504–519.
[5] Creighton, T.E. (1993) Proteins: Structures and Molecular Properties. W.H. Freeman, New York.
[6] Qian, N. and Sejnowski, T.J. (1988) J. Mol. Biol. 202, 865–884.
[7] Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Nørskov, L., Olsen, O.H. and Petersen, S.B. (1988) FEBS Lett. 241, 223–228.
[8] Rost, B. and Sander, C. (1994) Proteins 19, 55–72.
[9] Riis, S.K. and Krogh, A. (1996) J. Comp. Biol. 3, 163–183.
[10] Schneider, T.D. and Stephens, R.M. (1990) Nucleic Acids Res. 18, 6097–6100.
[11] Matthews, B.W. (1975) Biochim. Biophys. Acta 405, 442–451.

[12] Andersen, C.A. (1998) Master's Thesis, Technical University of Denmark, Lyngby.
[13] Colloc'h, N., Etchebest, C., Thoreau, E., Henrissant, B. and Mornon, J.P. (1993) Protein Eng. 6, 377–382.
[14] Frishman, D. and Argos, P. (1996) Protein Eng. 9, 2.
[15] Rost, B. and Sander, C. (1993) J. Mol. Biol. 232, 584–599.
[16] Frishman, D. and Argos, P. (1997) Proteins 27, 329–335.
[17] Jones, D.T. (1999) J. Mol. Biol. 292, 195–202.
[18] Rost, B. (2001) J. Struct. Biol. 134, 204–218.
[19] Baldi, P., Brunak, S., Frasconi, P., Pollastri, G. and Soda, G. (1999) Bioinformatics 15, 937–946.
[20] Petersen, T.N., Lundgaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G.P. and Lund, O. (2000) Proteins 41, 17–20.
[21] Padmanabhan, S., Marqusee, S., Ridgeway, T., Laue, T.M. and Baldwin, R.L. (1990) Nature 344, 268–270.