Article No. mb981792 *J. Mol. Biol.* (1998) **279**, 665–684

JMB

Role of Main-chain Electrostatics, Hydrophobic Effect and Side-chain Conformational Entropy in Determining the Secondary Structure of Proteins

F. Avbelj* and L. Fele

National Institute of Chemistry
Hajdrihova 19, Ljubljana
SI 1115, Slovenia

**Corresponding author*

The physiochemical bases of amino acid preferences for α-helical, β-strand, and other main-chain conformational states in proteins is controversial. Hydrophobic effect, side-chain conformational entropy, steric factors, and main-chain electrostatic interactions have all been advanced as the dominant physical factors which determine these preferences. Many attempts to resolve the controversy have focused on small model systems. The disadvantage of such systems is that the amino acids in small molecules are largerly exposed to the solvent. In proteins, however, the amino acids are in contact with the solvent to a different degree, causing a large variability of strengths of all interactions. The estimates of mean strengths of interactions in the actual protein environment are therefore essential to resolve the controversy. In this work the experimental protein structures are used to estimate the mean strengths of various interactions in proteins. The free energy contributions of the interactions are implemented into the Lifson-Roig theory to calculate the helix and strand free energy profiles. From the profiles the secondary structures of proteins and peptides are predicted using simple rules. The role of hydrophobic effect, side-chain conformational entropy, and main-chain electrostatic interactions in determining the secondary structure of proteins is assessed from the abilities of different models, describing stability of secondary structures, to correctly predict α-helices, β-strands and coil in 130 proteins. The three-state accuracy of the model, which contains only the free energy terms due to the main-chain electrostatics with 40 coefficients, is 68.7%. This accuracy is approaching to the accuracy of currently the best secondary structure prediction algorithm based on neural networks (72%); however, many thousands of parameters have to be optimized during the training of the neural networks to reach this level of accuracy. The correlation coefficient between the calculated and the experimental helix contents of 37 alanine based peptides is 0.91. If the hydrophobic and the side-chain conformational entropy terms are included into the helix-coil transition parameters, the accuracy of the algorithm does not improve significantly. However, if the main-chain electrostatic interactions are excluded from the helix-coil and strand-coil transition parameters, the accuracy of the algorithm reaches only 59.5%. These results support the dominant role of the short-range main-chain electrostatics in determining the secondary structure of proteins and peptides. The role of the hydrophobic effect and the side-chain conformational entropy is small.

Keywords: electrostatic screening; helix-coil transition theory; protein folding; Lifson-Roig theory; prediction of secondary structure

Introduction

Amino acids exhibit distinct preferences for α-helical, β-strand, β-sheet, and other main-chain conformational states in proteins. Understanding of physical background for these preferences is crucial in solving the protein folding problem. Although a considerable effort has been directed to

0022–2836/98/230665–20 $25.00/0

elucidate the nature of forces that dominate the preferences of amino acids, the issue remains controversial.

Hydrophobic interactions (Blaber *et al*., 1993, 1994), side-chain conformational entropy (Creamer & Rose, 1992; Creamer & Rose, 1994), steric effects (Yun & Hermans, 1991; Hermans *et al*., 1992; Bai *et al*., 1993; Bai & Endglander, 1994), and main-chain electrostatics (Avbelj & Moult, 1995b), have all been suggested as the dominant physical factors, which determine the preferences of amino acids for particular main-chain conformational states.

Hydrophobicity has been proposed to determine primarily the preferences of residues for the α-helix, because it has been demonstrated that the side-chain hydrophobic surface area buried along the α-helix correlates with the relative free energies of unfolding of the phage T4 lysozyme mutants (Blaber *et al*., 1993, 1994).

The side-chains of large β branched residues in the α-helix are more constrained compared to the side-chains of small amino acid like Ala (Piela *et al*., 1987; McGregor *et al*., 1987; Padmanabhan *et al*., 1990; Padmanabhan & Baldwin, 1991; Yun & Hermans, 1991; Hermans *et al*., 1992). It has been suggested (Creamer & Rose, 1992; Creamer & Rose, 1994) that the associated side-chain conformational entropy cost may be primarily responsible for the differences between propensities of amino acids for the α-helix. Creamer & Rose (1992) and Creamer & Rose (1994) used the Monte Carlo simulations of small model systems to calculate the side-chain conformational entropy in α-helical and unfolded states. They have shown that the differences in side-chain conformational entropies between these two states ($T\Delta A_{\alpha}$) correlate with the experimental α-helix forming tendencies for eight non-polar amino acids. In contrast, Blaber *et al*. (1994) found that for all 20 amino acids the correlation between $T\Delta A_{\alpha}$, obtained from the rotamer distributions in the experimental protein structures, and the α-helix forming tendencies is much weaker with the average coefficient of only 0.39.

The steric strain between the side-chain of residues and the bulky α-helix backbone may give an unfavorable energy contribution (Yun & Hermans, 1991; Hermans *et al*., 1992). However, there are no data that would demonstrate a clear role of the strain for the α-helix preferences. Bai *et al*. found the correlation between the hydrogen exchange rates of model peptides and the β-sheet propensities of 13 amino acids (Bai *et al*., 1993; Bai & Englander, 1994). They propose that the side-chains can modulate the strength of the main-chain hydrogen bonds by the side chain steric blocking effect. According to their hypothesis, the β-sheet preferences depend on the side-chain steric clash, which interferes with peptide to solvent hydrogen bonding and thus increases the stability of intramolecular hydrogen bonding.

The main-chain electrostatics has been shown to quantitatively explain the preferences not only for α-helical, but also for β-sheet, β-strand, and other main-chain conformational states of 20 amino acids (Avbelj & Moult, 1995b). It has been demonstrated that the stability of a conformational state is primarily determined by the balance of strengths between the local and the short-range non-local main-chain electrostatic interactions. The strengths of these interactions depend on the amino acid side-chains involved, because they are screened to a different degree by the solvent and polar protein groups. The electrostatic screening model was developed, which enables the calculation of the free energy contributions of the local and the non-local main-chain electrostatic interactions in proteins by scaling the point charge electrostatic interaction energies with the residue-dependent mean field screening coefficients. The screening coefficients, two parameters per each amino acid type (a total of 40 parameters) are derived by fitting the potentials of mean force calculated from a large set of high resolution experimental protein structures.

According to the electrostatic screening model, the free energy contributions of the main-chain hydrogen bonding considerably stabilize the folded state of a protein (Avbelj & Moult, 1995b). This view is in striking contradiction with the widely accepted hydrophobic collapse model of protein folding (Kauzmann, 1959; Dill *et al*., 1995), according which the hydrophobic interactions represent the main driving force of protein folding and the free energy contribution of the hydrogen bonding is considered to be small (Klotz & Franzen, 1962; Yang & Honig, 1995; Sippl, 1996). The role of hydrogen bonding is only to provide the specificity in folding, because there is a considerable free energy cost for burying unsatisfied hydrogen binding groups.

However, the experimental data have shown that the contribution of the main-chain hydrogen bonding to the stability of folded proteins is large (Myers & Pace, 1996). Scholtz *et al*. (1991) measured ΔH of coil to α-helix transition of a 50-residue peptide. The enthalpy change was estimated as −1.3 kcal/mol per residue. This contribution was primarily assigned to the main-chain–main-chain interactions. Murphy & Gill (1990, 1991) have determined the energetics of peptide and other groups of atoms from the solubilities of small model cyclic peptides. The free energy contribution of the peptide groups to protein stability was found to be comparable to the contribution of non-polar atoms. Habermann & Murphy (1996) have demonstrated that the contribution of main-chain hydrogen bonding to protein stability is larger than the side-chain hydrogen bonding. Studies of some model compounds also indicate that the enthalpy change of hydrogen bonding in protein folding is large and negative (Shellman, 1955). These data agree with the hypothesis that was introduced many years ago by Pauling *et al*., according to which the electrostatic interactions considerably stabilize the α-helices and β-sheets (Pauling *et al*., 1951; Pauling & Corey, 1951). Brandt & Flory (1965a,b) have shown that

the local main-chain electrostatics is a crucial factor in determining the end to end distances in peptides.

If the main-chain electrostatic interactions determine the preferences of amino acids, then we would expect that these preferences depend on the nature of the solvent. Indeed, it has been demonstrated that the low dielectric mediums like: trifluoroethanol, methanol, sodium dodecyl sulfate, and membranes, have a strong influence on the amino acids preferences (Tanford *et al.*, 1960; Nelson & Kallenbach, 1986; Buck *et al.*, 1993; Jasanoff & Fersht, 1994; Waterhous & Johnson, 1994; Blanco *et al.*, 1994; Schonbrunner *et al.*, 1996; Luo & Baldwin, 1998). For example, trifluoroethanol promotes α-helix formation and stabilizes β-sheet structures (Blanco *et al.*, 1994; Schonbrunner *et al.*, 1996). It has been suggested that trifluoroethanol increases the strength of intramolecular hydrogen bonds (Nelson & Kallenbach, 1986; Schonbrunner *et al.*, 1996). Luo & Baldwin (1998) measured the thermal helix-coil transitions for alanine based peptides in different concentrations of trifluoroethanol. Their results confirmed that the hydrogen bond strength increases with trifluoroethanol molarity in the same manner as the mean α-helix propensity. Other explanations for the solvent effects have also been proposed (Thomas & Dill, 1993; Jasanoff & Fersht, 1994; Bodkin & Goodfellow, 1995); however, the side-chain conformational entropy or the steric effects have not been involved in these mechanisms. Trifluoroethanol decreases the strength of hydrophobic interactions (Thomas & Dill, 1993; Schonbrunner *et al.*, 1996) causing partial denaturation of proteins, therefore, the hydrophobic effect is probably not responsible for the α-helix promoting character of this solvent.

The preferences of amino acids form a basis for various secondary structure prediction algorithms (Chou & Fasman, 1974b; Garnier *et al.*, 1978, 1996; Gibrat *et al.*, 1987; Holley & Karplus, 1989; Kneller *et al.*, 1990; Rost & Sander, 1993, 1994). Most secondary structure prediction methods are based on probabilistic approaches rather tan on an explicit physiochemical model. Currently the most successful secondary structure prediction methods with ≈72% accuracy are neural networks in which evolutionary information in the form of multiple sequence alignments has been included (Rost & Sander, 1993, 1994). Unfortunately, physical reasons for the high accuracy achieved by these methods are hidden in the complexity of neural network algorithms.

Statistical mechanical helix-coil transition theories (Zimm & Bragg, 1959; Lifson & Roig, 1961; Qian & Schellman, 1992) have been used for studying the helix content of polypeptides (Finkelstein & Ptitsyn, 1976; Ptitsyn & Finkelstein, 1983; Wojcik *et al.*, 1990; Padmanabhan *et al.*, 1990; Lyu *et al.*, 1990; Chakrabartty *et al.*, 1991, 1994; Kemp *et al.*, 1991; Scholtz *et al.*, 1991; Finkelstein *et al.*, 1991; Qian & Schellman, 1992; Doig *et al.*, 1994; Munoz & Serrano, 1994, 1997; Stapley *et al.*, 1995; Rohl *et al.*, 1996) and much less frequently for the prediction of α-helices in proteins (Lewis *et al.*, 1970; Finkelstein & Ptitsyn, 1976; Ptitsyn & Finkelstein, 1983; Qian, 1996). According to the helix-coil transition theories, the nucleation of an α-helix is unfavorable, because it requires the spatial fixing of the ϕ and ψ of the three consecutive residues to the α conformation before the first hydrogen bond is formed. Propagation of the α-helix, which represents the addition of one residue to the already existing α-helix, is favorable, because of the large stabilizing contribution from hydrogen bonding and the spatial confinement of one residue only. Nucleation and propagation parameters, σ and s, respectively, have been obtained for all residue types from the helix contents of mutated peptides. The transition parameters derived by the host-guest studies and the alanine based peptides disagree in size and range order (Vila *et al.*, 1992; Padmanabhan *et al.*, 1994). The helix-coil transition theory has been used for the prediction of α-helices in proteins. In some studies the protein α-helix probabilities were found to be very small (<6%) (Lewis *et al.*, 1970; Qian, 1996). Much larger α-helix probabilities were obtained in the model which includes various short-range and long-range interactions of side-chains (Finkelstein & Ptitsyn, 1976; Ptitsyn & Finkelstein, 1983). The long-range interactions have been approximated with the average hydrophobic template (Ptitsyn & Finkelstein, 1983). This model has been used to predict α-helices and β-strands in proteins.

In this work, the role of hydrophobic effect, side-chain conformational entropy, and main-chain electrostatic interactions in determining the secondary structure of proteins is examined by a new experiment in which the accuracies of secondary structure prediction algorithms based on the different models for the stability of secondary structures are compared. The mean free energy contributions of these interactions are incorporated into the Lifson-Roig transition theory (Lifson & Roig, 1961; Qian & Schellman, 1992; Doig *et al.*, 1994) to evaluate the helix free energy profiles of proteins. Further, we develop a strand-coil transition theory utilizing the mathematics of the Lifson-Roig algorithm to obtain the strand free energy profiles of proteins. The mean strengths of hydrophobic effect, conformational entropy, and main-chain electrostatics in the actual protein environment are derived from a set of 328 high-resolution (resolution <2.0 Å and R factor <20%) X-ray protein structures from the Protein Data Bank (Bernstein *et al.*, 1977). The helix and strand free energy profiles are averaged over a large number of homologous sequences for each protein. From these profiles, the secondary structures are predicted using simple rules. The statistical mechanical method is used to predict the secondary structure of 130 proteins and 37 peptides. There are no homologous pairs of sequences between the set of 130 proteins used in the predictions and the set of 328 proteins used to obtain the

mean strengths of the most important interactions. The influence of all possible combinations of the free energy terms on the accuracy of secondary structure prediction algorithm are examined. The results obtained by the three representative models describing the stability of secondary structures are shown. In the first model (model I) the helix and strand free energy profiles are calculated using the electrostatic screening model only. In the second model (model II) the combination of those free energy terms which generate the most accurate prediction of secondary structure is presented. The free energy terms due to the hydrophobic effect and the side-chain conformational entropy are included in the electrostatic screening model (model I) to describe the helix-coil transition. In the third model (model III) solely the free energy terms due to the hydrophobic effect and the conformational entropy are utilized.

Results and Discussion

Mean strengths of interactions in actual protein environment

The controversial role of hydrophobic effect, side-chain conformational entropy, steric factors, and main-chain electrostatic interactions in determining the secondary structures in proteins have been studied primarily using the small model systems (Creamer & Rose, 1992, Creamer & Rose, 1994; Herman *et al*., 1992; Yun *et al*., 1991; Vila *et al*., 1992; Yang & Honig, 1995; Wang & Purisima, 1996). The disadvantage of small model systems is that the amino acids in such systems are largely exposed to the solvent. In proteins, however, the amino acids are in a very complex environment. Some amino acids are completely exposed to the solvent, as in the denatured state, some are completely buried in the protein core, but most of the amino acids are exposed to the solvent to a different degree (Avbelj, 1992). The strengths of hydrophobic effect, side-chain conformational entropies, main-chain electrostatics, and other interactions depend strongly on amount of the solvent in close proximity (Warshel & Russell, 1984; Doig & Sternberg, 1995; Chothia, 1974), therefore it is very difficult to elucidate the role of these interactions in determining the secondary structures in proteins using such small systems. The estimates of the mean strengths of interactions in the actual protein environment are therefore essential to resolve the controversy. Here the free energy contributions of the hydrophobic effect, side-chain conformational entropies, and main-chain electrostatic interactions in the actual protein environment are estimated (Avbelj, 1992) from a set of 328 high-resolution X-ray protein structures (see Methods).

Models for stability of secondary structures

The mean free energy contributions of the most important interactions in the actual protein environment, derived from the experimental protein structures (see above and Methods), are used in the Lifson-Roig helix-coil transition theory (Lifson & Roig, 1961; Qian & Schellman, 1992; Doig *et al*., 1994) to describe the stability of α-helices in proteins (Lewis *et al*., 1970; Qian, 1996). Further, we develop a strand-coil transition theory utilizing the mathematics of the Lifson-Roig algorithm to describe the stability of β-strands in proteins. The β-strand formation is treated using the Lifson-Roig theory because of the similar interdependence of residue conformations, i.e. the general cooperativity, as it occurs in the helix-coil transition (see section Strand-coil transition below).

The helix-coil transition theory considers the equilibrium between two conformational states in proteins: α-helix and coil. In this work, however, the equilibrium between three conformational states of amino acids in proteins is studied. The three states are: α-helix, β-strand and coil. We define α-helical and β-strand states in a protein as those states which are assigned as α-helices and β-strands, respectively, by the modified Kabsch & Sander (1983) DSSP algorithm (see Methods). The state of the remainder of amino acids in a protein is defined as coil. The coil state in the helix-coil transition is taken to be identical to the coil state in the strand-coil transition. The advantage of using the coil state of folded proteins as the standard state is that an amino acid in such coil state is in the restricted state (Avbelj, 1992). In contrast, the properties of the coil state of denatured proteins are still far from being understood. The definition of the conformational states in this work is therefore different from the definition of states usually use in studying the helix-coil transition of peptides. As a consequence the parameters v and w used in this study are not directly comparable to the parameters used by other algorithms.

We tested all possible models for the stability of α-helices and β-strands. Various combinations of the free energy terms due to the main-chain electrostatics, the side-chain and main-chain conformational entropy, and the hydrophobic interactions are incorporated in the helix-coil and the strand-coil transition parameters to evaluate the helix and strand free energy profiles (equation (20)). The results obtained by the following three representative models for the stability of secondary structures are presented in this work.

Model I, electrostatic screening model. The short-range main-chain electrostatic interactions (Avbelj & Moult, 1995b) are used in this model.

Model II. This model represents the combination of those free energy terms which generate the most accurate prediction of secondary structure. The free energy terms due to the hydrophobic effect and the side-chain conformational entropy are added into the electrostatic screening model (model I) to describe the helix-coil transition. If the conformational entropy and the hydrophobic terms are included into the strand-coil transition, the accu-

racy of the prediction algorithm is considerably reduced.

Model III. This model contains solely the free energy terms due to the hydrophobic effect and the conformational entropy.

Helix-coil transition

The helix free energy profiles are calculated by the Lifson-Roig theory using equation (20). The helix-coil transition parameter v_i in equations (16), (17) and (19) represents the equilibrium constant for formation of an α conformation in a coil, where at least one flanking residue of the amino acid i is in the coil conformation (Lifson & Roig, 1961; Qian & Schellman, 1992). The formation of an α conformation in a coil is unfavorable because of the main-chain conformational entropy cost of confinement an amino acid in the α conformation and because of the repulsive local main-chain electrostatic interactions (Qian & Schellman, 1992). The helix-coil transition parameter w_i in equations (16), (17) and (19) is the equilibrium constant for formation of an α conformation in a coil, where both flanking residues of the amino acid i are also in the α conformation (Lifson & Roig, 1961; Qian & Schellman, 1992). The parameter w_i depends on the short-range main-chain electrostatic interactions (hydrogen bonds), the main-chain and side-chain conformational entropy cost of confinement an amino acid i in the α conformation, and the hydrophobic interactions of side-chain atoms in the α-helix (Qian & Schellman, 1992). The three representative models for the stability of secondary structures (models I, II, and III; see above) contain the most interesting combinations of the free energy contributions due to these interactions.

Model I: electrostatic screening model

This model contains the short-range non-local and the local main-chain electrostatic interactions only (Avbelj & Moult, 1995b). The main-chain conformational entropies are included implicitly in the screening coefficients (Avbelj & Fele, 1998).

From the definition of the transition parameters v_i and w_i the following equations are obtained:

$$-kT \ln v_i = \gamma^i_{\text{local}}(\bar{E}^h_{\text{local}} - \bar{E}^c_{\text{local}}) \quad (1)$$

$$-kT \ln w_i = \frac{1}{2}(\gamma^{i-2}_{\text{nonlocal}} + \gamma^{i+2}_{\text{nonlocal}})\bar{E}_{\text{hb}} + \gamma^i_{\text{local}}(\bar{E}^h_{\text{local}} - \bar{E}^c_{\text{local}}) \quad (2)$$

where γ^i_{local} represents the local screening coefficient of an amino acid i (Table 1). The $\gamma^{i-2}_{\text{nonlocal}}$ and $\gamma^{i+2}_{\text{nonlocal}}$ are the non-local screening coefficients for residues i−2 and i+2, respectively (Table 1). The $\bar{E}^h_{\text{local}}$ and $\bar{E}_{\text{hb}}$ are the average local and nonlocal electrostatic energies in the α-helix. $\bar{E}^c_{\text{local}}$ is the average local electrostatic energy of an amino acid in the coil state.

Table 1. The electrostatic screening coefficients $\gamma^r_{\text{nonlocal}}$ and γ^r_{local}

Residue	$\gamma^r_{\text{nonlocal}}$	γ^r_{local}	*r.s.d.* (kcal/mol)
Gly	−0.103	0.251	0.20
Ala	0.285	0.169	0.09
Val	0.325	0.500	0.10
Ile	0.404	0.491	0.10
Leu	0.367	0.394	0.10
Phe	0.331	0.429	0.09
Pro	−0.325	–	0.24
Met	0.358	0.347	0.12
Trp	0.232	0.382	0.11
Cys	0.183	0.321	0.01
Ser	0.078	0.279	0.15
Thr	0.126	0.354	0.12
Asn	0.116	0.230	0.19
Gln	0.257	0.290	0.11
Tyr	0.241	0.402	0.09
His	0.176	0.264	0.10
Asp	0.102	0.237	0.11
Glu	0.345	0.339	0.10
Lys	0.244	0.283	0.10
Arg	0.275	0.330	0.09

$\gamma^r_{\text{nonlocal}}$ and γ^r_{local} mainly represent the attenuation of the electrostatic energies E_{nonlocal} and E_{local}, respectively, due to the screening by water and protein dipoles. The residual standard deviations of the fit (*r.s.d.*) are also shown.

The term $\gamma^i_{\text{local}}(\bar{E}^h_{\text{local}} - \bar{E}^c_{\text{local}})$ in equations (1) and (2) represents the free energy cost of an amino acid being in α conformation due to the local main-chain electrostatic interactions. This term is positive because the peptide dipoles flanking a residue in the α conformation are parallel. The term $\frac{1}{2}(\gamma^{i-2}_{\text{nonlocal}} + \gamma^{i+2}_{\text{nonlocal}})\bar{E}_{\text{hb}}$ in equation (2) represents the free energy contribution of the short-range non-local main-chain electrostatic interactions to the stability of an amino acid. This term is negative and is mainly due to the main-chain hydrogen bonding which depends on the screening coefficients of two residues $i-2$ and $i+2$. Note that the parameter w_i depends on the nature of amino acids i, $i-2$, and $i+2$, although it is assigned to the residue i. The Lifson-Roig transition parameters w_i are therefore sequence dependent.

The α-helices in globular proteins must contain amino acids with small γ^r_{local} (polar and charged residues) to reach the high energy α conformation and a few residues with large $\gamma^r_{\text{nonlocal}}$ (non-polar residues) at positions $i-2$ and $i+2$ to stabilize the α-helical turns with hydrogen bonds. This arrangement of polar and non-polar residues in α-helices resembles a pattern of non-polar residues called a ''helical wheel'', utilized in many helix prediction algorithms (Shulz & Schirmer, 1979; Lim, 1974).

Model II

The parameters w_i in this model contain the terms due to the side-chain conformational entropy and the hydrophobic effect in addition to the free energy terms used in the electrostatic screening model (model I). The parameters v_i and w_i are

defined by the following equations:

$$-kT \ln v_i = \gamma^i_{\text{local}}(\bar{E}^h_{\text{local}} - \bar{E}^c_{\text{local}}) \quad (3)$$

$$-kT \ln w_i = \frac{1}{2}(\gamma^{i-2}_{\text{nonlocal}} + \gamma^{i+2}_{\text{nonlocal}})\bar{E}_{\text{hb}} + \gamma^i_{\text{local}}(\bar{E}^h_{\text{local}} - \bar{E}^c_{\text{local}}) - K\bar{A}^h_{i-2,i+2} - K\bar{A}^h_{i-1,i+2} + T\Delta S^i_\alpha \quad (4)$$

The terms $-K\overline{A}^h_{i-2,i+2}$ and $-K\overline{A}^h_{i-1,i+2}$ represent the free energy contribution of the hydrophobic interactions between side-chains at spacings of $i-2$, $i+2$ and $i-1$, $i+2$, respectively (the spacings of i, $i+4$ and i, $i+3$ in the Table 2). The term $T\Delta S^i_\alpha$ represents the free energy difference due to the change in the side-chain conformational entropy between the α-helical and the unfolded states for amino acid i. $\overline{A}^h_{i-2,i+2}$, $\overline{A}^h_{i-1,i+2}$, and $T\Delta S^i_\alpha$ are calculated from the set of 328 experimental protein structures (see Methods and Tables 2 and 3). T is temperature, K is constant in equation (13).

Model III

This model contains only the terms due to the main-chain and side-chain conformational entropies and the terms due to the hydrophobic interactions. The free energy terms due to the main-chain electrostatic interactions are excluded. The parameters v_i and w_i are defined by the following equations:

$$-kT \ln v_i = T\Delta S^i_{\alpha-\text{main}} \quad (5)$$

$$-kT \ln w_i = -K\bar{A}^h_{i-2,i+2} - K\bar{A}^h_{i-1,i+2} + T\Delta S^i_\alpha + T\Delta S^i_{\alpha-\text{main}} \quad (6)$$

The terms $T\Delta S^i_{\alpha-\text{main}}$ and $T\Delta S^i_\alpha$ are the free energy differences due to the change in the main-chain and side-chain conformational entropies, respectively, between the α-helical and the unfolded states for amino acid i (Table 3). The terms $-K\overline{A}^h_{i-2,i+2}$ and $-K\overline{A}^h_{i-1,i+2}$ are the free energy contributions due to the hydrophobic interactions between side-chains at spacings of i, $i+4$ and i, $i+3$, respectively. $\overline{A}^h_{i-2,i+2}$ and $\overline{A}^h_{i-1,i+2}$, $T\Delta S^i_{\alpha-\text{main}}$ and $T\Delta S^i_\alpha$ are calculated from the set of 328 experimental protein structures (see Methods and Tables 2 and 3). T is temperature. K is constant in equation (13).

Strand-coil transition

It is uncommon to consider the stability of β-strands separately from the stability of β-sheets; however, the electrostatic screening model provides a simple physical background for the development of the strand-coil transition theory. The physical reason for the stability of β-strands originates from the antiparallel alignment of the main-chain dipole moments.

According to the electrostatic screening model, the energetics of residues in the β conformation

Table 2. The average pairwise hydrophobic accessible surface contact areas ($\overline{A}_{ij}$) (Å^2) between side-chains in α-helices and β-strands calculated from the set to 328 X-ray protein structures

i	Ala	Val	$i+4$ Ile	or Leu	$i+3$ Phe	or Met	$i+2$ Trp	Cys	Tyr
Ala	1.8	9.0	20.2	12.4	14.3	13.3	5.3	6.9	8.1
	7.4	7.6	10.4	9.4	7.6	12.0	10.2	7.8	8.5
	0.3	9.8	12.1	9.5	15.6	12.7	4.1	10.7	17.5
Val	11.6	31.5	36.8	37.5	28.2	33.7	28.7	27.1	22.3
	8.8	10.1	18.9	16.9	11.2	20.8	15.0	10.4	11.4
	8.3	25.5	27.0	24.8	40.9	35.8	26.3	20.0	23.1
Ile	13.7	32.9	42.6	39.8	33.1	38.1	34.7	26.2	38.1
	10.8	12.0	24.3	20.7	18.3	20.1	20.6	12.2	15.3
	11.7	26.4	27.4	24.7	34.2	34.6	45.4	14.8	19.7
Leu	7.5	24.2	38.7	39.5	32.9	46.6	16.4	25.3	24.1
	17.0	23.1	34.4	32.8	28.6	30.8	28.2	20.1	29.4
	7.6	27.7	33.9	31.7	35.6	28.4	37.0	9.9	21.3
Phe	18.8	42.8	53.6	52.7	46.0	60.5	39.2	53.5	52.1
	11.7	17.4	31.5	29.7	35.2	31.8	49.1	17.9	23.5
	15.2	25.5	34.0	25.7	41.2	32.3	39.8	11.9	33.0
Met	4.4	26.1	30.7	33.0	43.4	39.3	1.6	13.2	21.0
	12.3	23.8	30.2	32.4	33.3	42.5	43.4	13.7	30.5
	10.4	28.3	37.7	39.2	38.3	44.9	16.5	8.0	24.8
Trp	7.5	43.3	52.2	45.5	50.1	65.3	82.8	–	33.6
	21.1	47.5	50.2	50.2	18.2	33.1	28.9	–	22.1
	6.3	24.9	19.4	39.6	22.8	41.0	48.2	24.5	14.4
Cys	4.3	18.9	29.5	30.7	24.3	28.8	41.7	8.1	5.5
	8.4	11.8	13.3	16.6	13.4	11.1	29.7	3.5	14.2
	4.3	22.3	20.6	18.3	24.4	14.2	13.3	20.1	20.4
Tyr	12.9	38.3	41.2	39.0	38.2	49.8	36.8	57.7	15.6
	10.9	17.5	22.3	22.7	28.8	29.4	40.8	15.3	20.4
	7.0	20.0	23.5	15.3	26.2	7.8	29.7	4.8	28.7

In each cell, the top value is the surface contact area at spacings i, $i+4$ in α-helices ($\overline{A}^h_{i,i+4}$), the medium line is the area at spacings i, $i+3$ in α-helices ($\overline{A}^h_{i,i+3}$), and the bottom line is the area at spacings i, $i+2$ in β-strands ($\overline{A}^s_{i,i+2}$).

Table 3. The conformational entropies for amino acids multiplied with temperature (kcal/mol)

Residue	TS_{α}	TS_{coil}	$T\Delta S_{\alpha}$	$TS_{coil}^{Pickett}$	$T\Delta S_{\alpha-main}$	$T\Delta S_{\beta-main}$
Gly	0.000	0.000	0.00	0.00	0.74	0.59
Ala	0.000	0.000	0.00	0.00	0.54	0.43
Val	−0.172	−0.541	0.37	−0.51	0.46	0.43
Ile	−0.481	−0.926	0.45	−0.89	0.45	0.41
Leu	−0.696	−0.763	0.07	−0.78	0.53	0.49
Phe	−0.409	−0.544	0.13	−0.58	0.56	0.51
Pro	0.000	0.000	0.00	0.00	0.41	0.28
Met	−1.452	−1.540	0.09	−1.61	0.57	0.46
Trp	−0.633	−0.909	0.28	−0.97	0.52	0.43
Cys	−0.535	−0.572	0.04	−0.55	0.54	0.53
Ser	−1.686	−1.695	0.01	−1.71	0.54	0.39
Thr	−1.363	−1.618	0.25	−1.63	0.47	0.35
Asn	−1.436	−1.708	0.27	−1.57	0.67	0.39
Gln	−1.929	−2.107	0.18	−2.11	0.59	0.55
Tyr	−0.858	−1.019	0.16	−0.98	0.55	0.49
His	−0.794	−0.895	0.10	−0.96	0.62	0.57
Asp	−0.959	−1.318	0.36	−1.25	0.56	0.31
Glu	−1.547	−1.763	0.22	−1.81	0.55	0.45
Lys	−1.849	−1.973	0.12	−1.94	0.60	0.51
Arg	−1.991	−2.120	0.13	−2.03	0.58	0.48

The temperature T is 300 K. TS_{α} and TS_{coil} represent the side-chain conformational entropies of amino acids in the α-helical and the coil states, respectively. These entropies are calculated using equation (14) from the rotamer distributions in the set of 328 experimental protein structures. $T\Delta S_{\alpha}$ is the difference between TS_{α} and TS_{coil}. $TS_{coil}^{Pickett}$ is the side-chain conformational entropy of the coil amino acids calculated from the experimental protein structures by Picket & Sternberg (1993). $T\Delta S_{\alpha-main}$ is the free energy difference due to the main-chain conformational entropy between the α-helical and the coil states. $T\Delta S_{\beta-main}$ is the corresponding difference between the β-strand and the coil states.

depends on the conformation of the first neighbor residues. The reason for this is that $\overline{E}^{i}_{local}$ of a residue i does not depend only on the conformation of residue i, but also on the conformations of both flanking residues $i-1$ and $i+1$ (triplets). The average local main-chain electrostatic energy decreases gradually from −1.7 kcal/mol for an isolated amino acid with β conformation in coil, −2.3 kcal/mol for an amino acid at ends of an uninterrupted sequence of amino acids in β conformation, to −3.0 kcal/mol for an amino acid in the interior of β-strand. These average electrostatic energies are calculated from 328 experimental protein structures. Thus, both flanking residues must be in the β conformation in order to reach the minimum value of $\overline{E}^{i}_{local}$ of an amino acid. More neighboring residues having large γ^{r}_{local} (for example Val) would therefore reinforce each other in the ability to form β-strands.

The energetics of such systems, whose free energy of a residue depends on the conformational states of both neighbor residues (triplets), is cooperative (Poland & Scheraga, 1967). The β-strand formation must therefore be treated by taking into account the interdependence of residue conformations, i.e. the general cooperativity. Lifson-Roig theory is suitable for treating energetics of such triplet states (Lifson & Roig, 1961; Poland & Scheraga, 1967; Qian & Schellman, 1992). The advantage of using the Lifson-Roig theory, comparing to other transition theories, is that it is independent on the model for the stability of secondary structures (Qian & Schellman, 1992). The basic equations for the partition function of the strand-coil transition are the same as the equations for the helix-coil transition. Although the helix-coil and strand-coil transitions can be treated by the same mathematical formalism, there is a large difference between these two transitions. Forming on isolated β conformation in a coil is favorable, but the formation of one isolated α conformation is a coil is unfavorable. Consequently, there is no energetically unfavorable nucleation for the β-strand formation.

We define β conformation and β-strand state for the strand-coil transition analogous to the definitions of α conformation and α-helical state in the helix-coil transition theory (see Methods). The residue is classified as being in β conformation, if the torsion angles ϕ and ψ are in the β (extended) region of the Ramachandran plot. Note that the β conformation does not mean a residue is in the β-strand state. We define α-helical and β-strand states in a protein as those states which are assigned as α-helices and β-strands, respectively, by the modified Kabsch & Sander (1983) DSSP algorithm (see Methods). The state of the remainder of amino acids in a protein is defined as coil.

Analogous to v and w in the helix-coil transition, new statistical weights o and t, respectively, are introduced. The strand-coil transition parameter o_i is defined as the equilibrium constant for formation of a β conformation in a coil, where at least one flanking residue of the amino acid i is in the coil conformation. The parameter o_i depends on the favorable local main-chain electrostatic interactions and the main-chain conformational entropy cost of

confinement of an amino acid in the β conformation. The parameter t_i is the equilibrium constant for formation of a β conformation in a coil, where both flanking residues of the amino acid i are also in β conformation. The parameter t_i involves the favorable local main-chain electrostatic interactions, the favorable hydrophobic side-chain–side-chain interactions in the β-strands, and the main-chain and side-chain conformational entropy cost of confinement of an amino acid in the β conformation. The three representative models (models I, II, III: see above) contain various free energy terms describing the stability of β-strands.

The partition function for a hetero polymer (Z), the probability to find the ith amino acid in a chain of n residues in a β-strand ($p(i)_{\mathrm{str}}$), the probability that the ith amino acid in a chain of n residues has a weighting of o ($p(i)_{\mathrm{stnuc}}$), and the strand free energy profiles (G^i_{str}) can be defined by equations (16), (17), (19), and (20), substituting the equilibrium constants w and v with t and o, respectively, and replacing subscripts hb and nuc with str and stnuc, respectively.

Model I: electrostatic screening model

An important consequence of the electrostatic screening model is a considerable stability of an amino acid with the large $\gamma^r_{\mathrm{local}}$ (β branched residues) in the β-strand state, even if it is not a part of a β-sheet. The reason is in the large stabilizing free energy contribution of the local electrostatic interactions arising from the antiparallel alignment of the CO and NH dipole moments in β-strands which are protected from screening by bulky side-chain. Strands in β-sheets are further stabilized by the large contribution of the long-range non-local electrostatic interaction, i.e. hydrogen bonding which are ignored in the present treatment of the strand-coil transition.

Using equation (15) the following relations for o and t are obtained:

$$-kT\ln o_i = \gamma^i_{\mathrm{local}}(\bar{E}^{ee}_{\mathrm{local}} - \bar{E}^{c}_{\mathrm{local}}) \tag{7}$$

$$-kT\ln t_i = \gamma^i_{\mathrm{local}}(\bar{E}^{ei}_{\mathrm{local}} - \bar{E}^{c}_{\mathrm{local}}) \tag{8}$$

where $\gamma^i_{\mathrm{local}}$ is the screening coefficient of residue i for local electrostatic interactions (Table 1). The $\overline{E}^{ee}_{\mathrm{local}}$ and $\overline{E}^{ei}_{\mathrm{local}}$ are the average local electrostatic energies of an amino acid at both ends and in the interior of the β-strand, respectively (see Methods). The $\overline{E}^{c}_{\mathrm{local}}$ is the average local electrostatic energy in the coil state (see Methods). The terms $\gamma^i_{\mathrm{local}}(\overline{E}^{ee}_{\mathrm{local}} - \overline{E}^{c}_{\mathrm{local}})$ and $\gamma^i_{\mathrm{local}}(\overline{E}^{ei}_{\mathrm{local}} - \overline{E}^{c}_{\mathrm{local}})$ are the favorable free energy contributions due to the local main-chain electrostatic interactions.

Model II

The strand-coil transition parameters o_i and t_i for this model are equal to those in the model I.

$$-kT\ln o_i = \gamma^i_{\mathrm{local}}(\bar{E}^{ee}_{\mathrm{local}} - \bar{E}^{c}_{\mathrm{local}}) \tag{9}$$

$$-kT\ln t_i = \gamma^i_{\mathrm{local}}(\bar{E}^{ei}_{\mathrm{local}} - \bar{E}^{c}_{\mathrm{local}}) \tag{10}$$

The side-chain conformational entropy and the hydrophobic terms are not included in this model (see below).

Model III

This model contains the free energy contributions due to the main-chain conformational entropy and the hydrophobic interactions only. Although the main-chain electrostatic interactions are excluded from this model, the formation of the β-strands is cooperative and can be treated by the Lifson-Roig theory. The reason is in the hydrophobic interactions between amino acids at spacings of $i-1$, $i+1$, which is causing the dependence of the free energy of a residue on the conformational states of both neighbor residues (see above).

The parameters o_i and t_i are defined by the following equations:

$$-kT\ln o_i = T\Delta S^i_{\beta-\mathrm{main}} \tag{11}$$

$$-kT\ln t_i = -K\bar{A}^s_{i-1,i+1} + T\Delta S^i_{\beta-\mathrm{main}} \tag{12}$$

The term $T\Delta S^i_{\beta-\mathrm{main}}$ represents the free energy difference due to the change in the main-chain conformational entropy between β-strand and the unfolded state for amino acid i (Table 3). The term $K\bar{A}^s_{i-1,i+1}$ is the free energy contribution due to the hydrophobic interactions between side-chains at spacings of $i-1$, $i+1$ in β-strands (the spacings of i, $i+2$ in the Table 2).

Helix and strand free energy profiles

The helix free energy profile represents free energy difference between α-helical and coil states of amino acids as a function of sequence. Analogously, the strand free energy profile represents difference between free energies of amino acids in β-strand and coiled states as a function of sequence.

The helix and strand free energy profiles are calculated for a set of 130 protein chains (equation (20)). We have chosen the same set of proteins as used earlier by Rost & Sander (1993). Figures 1 to 4 show the helix and strand free energy profiles calculated by using model I for 4 representative experimental proteins structures. These free energy profiles are obtained as averages from the large number of homologous proteins (see Methods). Calmodulin (3cln, Figure 1) and apo-plastocyanin (2pcy, Figure 2) are constituted entirely from α-helices and β-strands, respectively. Staphylococcal nuclease (2sns, Figure 3) contains both α-helices and β-strands. Agglutinin (9wga, Figure 4) is without secondary structure (see Methods). The assign-

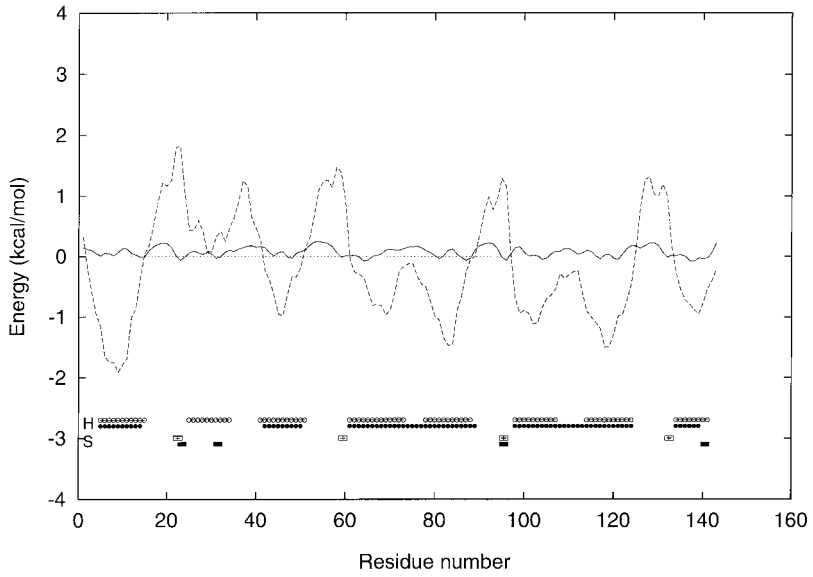

Figure 1. The helix and strand free energy profiles of calmodulin (3cln). The helix and strand free energy profiles are calculated using model I and averaged over homologous sequences. The strand free energy profile is plotted with the continuous line. The helix free energy profile is plotted with the broken line. The assignment of the α-helix and β-strand amino acids calculated by the modified Kabsch & Sander (1983) DSSP algorithm is marked by the open circles and open squares, respectively. The predicted assignment of the α-helix and β-strand amino acids is labeled by the filled circles and filled squares, respectively.

ment of the α-helical, β-strand, and coil states by the modified Kabsch & Sander (1983) DSSP algorithm is also shown.

Figures 1 to 4 show a strong correlation between the negative peaks of helix and strand free energy profiles with the occurrences of α-helices and β-strands in proteins, respectively. The strand free energy profiles vary considerably less than the helix free energy profiles. The reason for much larger oscillations of the helix free energy profiles around zero comparing to the strand free energy profiles is in the absence of energetically unfavorable nucleation in the β-strand formation. These correlations are utilized by our secondary structure prediction method.

Role of interactions in secondary structure of proteins

The role of various interactions in determining the secondary structure of proteins is assessed from the accuracies of various stability models (models I, II, III) to predict α-helices, β-strands and coil in 130 proteins are predicted using simple rules (see Methods). The lowest free energy rule is applied for predicting a conformational state of amino acids from the free energy profiles. The cut-

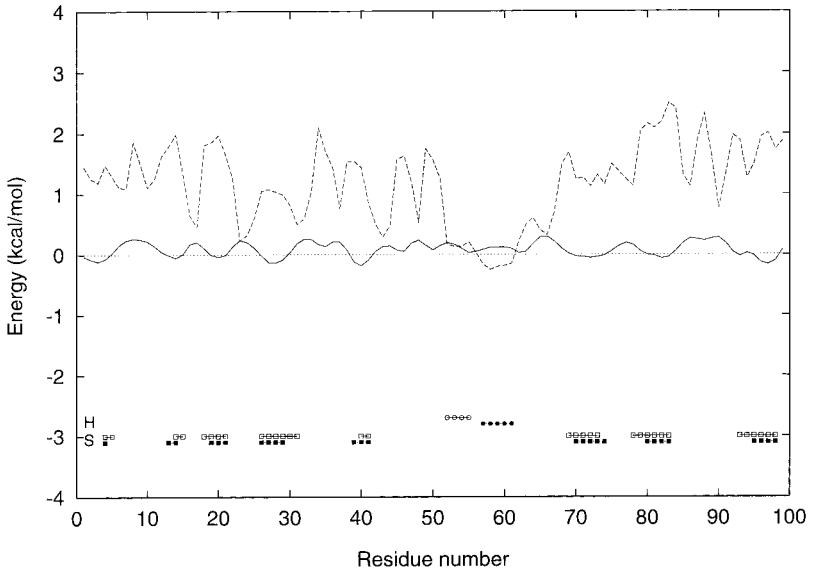

Figure 2. The helix and strand free energy profiles of apo-plastocyanin (2pcy). See legend to Figure 1.

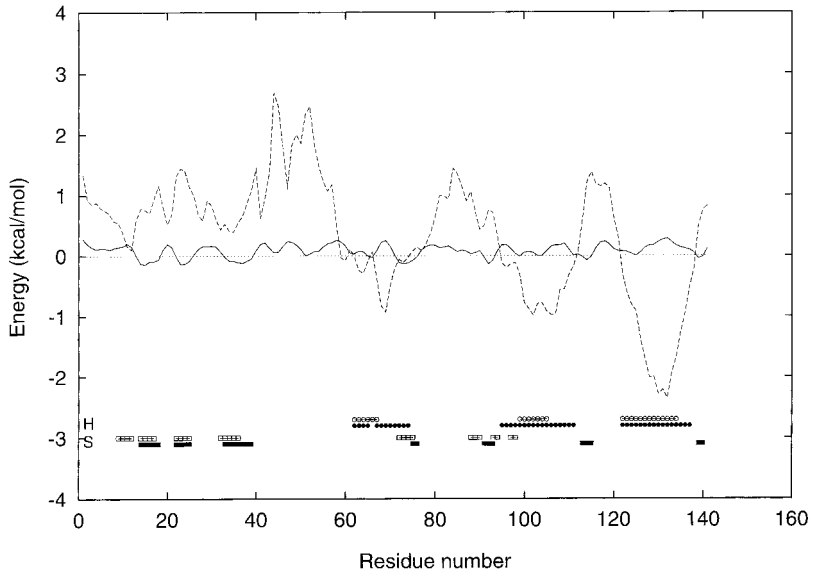

Figure 3. The helix and strand free energy profiles of staphylococcal nuclease (2sns). See legend to Figure 1.

off free energies are defined as sequence-independent thresholds for the prediction. For example, if the helix free energy of a residue is smaller than the α-helix cutoff and the strand free energy is larger than the β-strand cutoff, a residue is predicted to be in the α-helical state (see Methods for details and the additional requirement). The α-helix and β-strand cutoffs are fixed to zero. The results of the prediction are shown in Table 4 and in Figure 1 to 4. See also the reference Rost & Sander (1993) for the definition of the accuracy measures.

Using the main-chain electrostatic terms in the model I, the three state accuracy (Q_{total}) for predicting α-helices, β-strands and coil is 68.7% for 130 proteins, which is similar to the accuracy of neural network algorithms (Rost & Sander, 1993, 1994). Helical residues are predicted with the accuracies of $Q_\alpha = 56.8\%$ and $Q_\alpha^{\text{pred}} = 70.3\%$. The α-helices are predicted better than β-strands. For most methods the Q_β value is below 45%, with this prediction method; however, the accuracies of Q_β and Q_β^{pred} are 52.8% and 55.3%, respectively. This occurs despite the fact that the free energy contributions of the long-range hydrogen bonding, usually present in β-sheets, are not included in the model.

If the hydrophobic effect and the side-chain conformational entropy are included in the parameters for the helix-coil transition (model II) the Q_{total} value improves by only a small amount (from 68.67% to 68.73%). The role of the hydrophobic effect and the side-chain conformational entropy in determining the secondary structure in proteins is therefore smaller than the role of the main-chain electrostatics. These two types of interactions are important in stabilizing some α-helices (Avbelj & Fele, 1998). We used Pickett & Sternberg (1993) definition of rotamer classes. If a slightly different

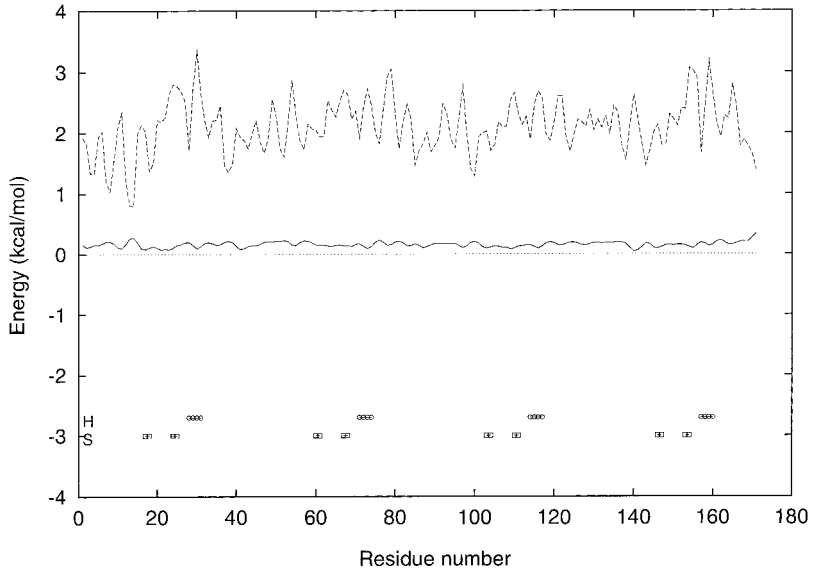

Figure 4. The helix and strand free energy profiles of agglutinin (9wga$_A$). See legend to Figure 1.

Table 4. Prediction accuracies

Type of model	Q_{total} %	Q_α %	Q_α^{pred} %	C_α	Q_β %	Q_β^{pred} %	C_β	Q_{coil} %	Q_{coil}^{pred} %	C_{coil}
Model I	68.67	56.8	70.3	0.57	52.8	55.3	0.51	83.4	69.1	0.43
Model II	68.73	60.2	66.8	0.57	49.7	56.5	0.50	84.0	69.7	0.44
Model III	59.45	43.0	52.5	0.41	42.7	47.0	0.41	83.0	63.7	0.31
Model Ia	63.13	55.5	55.4	0.47	50.7	45.7	0.43	75.6	68.4	0.36
Model Ib	67.38	61.1	62.0	0.47	52.8	50.3	0.46	74.5	70.6	0.38
Single GOR IV_{dc}	58.87	42.6	44.0	0.38	42.2	43.0	0.38	81.9	62.0	0.26
Single GOR IV_{ns}	65.67	55.7	57.3	0.49	50.9	51.4	0.49	81.8	68.7	0.40
GOR IV_a	63.10	62.1	55.9	0.52	60.3	39.7	0.45	68.3	72.1	0.37
GOR IV	68.16	67.1	67.0	0.61	61.6	44.9	0.49	74.0	73.8	0.42
GOR IV^{ori}	67.25	67.9	67.6	0.61	69.2	41.7	0.51	69.4	76.9	0.44

Q_{total} is the overall three state accuracy. Q_α, Q_α^{pred}, Q_β, Q_β^{pred}, and Q_{coil}, and Q_{coil}^{pred} are conditional probabilities of correct prediction. C_α, C_β, and C_{coil} are the Matthews correlation coefficients. See Rost & Sander, (1993) for the definition of these accuracy measures. Model Ia represents the predictions based on model I where the homologous sequences are excluded. Model Ib represents the predictions based on model I where the homologous sequences are excluded and the α-helix cutoff is adjusted to fit the predicted with the experimental helix contents. Single GOR IV_{dc}, Single GOR IV_{ns}, GOR IV_a and GOR IV represent the predictions using information calculated from the set of 328 high-resolution protein X-ray structure (see Methods). Single GOR IV_{dc} represents the GOR predictions based on the single residue information in which the decision constants are optimized (Garnier *et al.*, 1978). Single GOR IV_{ns} represents the GOR predictions based on the single residue information in which the decision and run constants are optimized (Garnier *et al.*, 1978). GOR IV_a represents the prediction with GOR IV algorithm using complete parameter set (Garnier *et al.*, 1996) in which the homologous sequences are excluded. GOR IV represents the prediction with GOR IV algorithm using complete parameter set (Garnier *et al.*, 1996). GOR IV^{ori} represents the prediction with GOR IV using complete original parameter set based on 267 protein structures (Garnier *et al.*, 1996).

definition of rotamer classes (Creamer & Rose, 1992) is used, the accuracy of the prediction algorithm does not change significantly. However, the accuracy of prediction algorithm is considerably reduced, if the side-chain conformational entropy and the hydrophobic terms are included into the strand-coil transition (model II: from 68.73% to below 50%).

The accuracy of the algorithm in which the main-chain electrostatic terms are excluded from the stability model (model III) reaches only 59.5%. The hydrophobic and the conformational entropy effects are used in this model. To achieve this level of accuracy, the constant K (in equations (1) to (12)) has to be increased from the value of 0.015 kcal/(mol·Å^2) used in model II to quite unrealistic seven times larger value of 0.105 kcal/(mol·Å^2). These results strongly support the dominant role of the short-range main-chain electrostatics in determining the secondary structure of proteins and peptides and suggest that the model for describing the stability of the secondary structures with solely the hydrophobic and the conformational entropy terms (model III) is not plausible.

Significance of new secondary structure prediction algorithm

The prediction method based on the electrostatic screening model of amino acid preferences is assessed by comparing the accuracies of the new methods with the accuracies of the GOR IV algorithm (Garnier *et al.*, 1978, 1996; Gibrat *et al.*, 1987). The predictions with the GOR IV algorithm are performed using identical data sets and assignments of the secondary structures as used by the newly developed method. The information parameters are calculated from the set of 328 high-resolution X-ray protein structures (see Methods). The homologous sequences are used. Table 4 shows the results of secondary structure predictions obtained by using different levels of the GOR IV algorithm. The accuracy of the GOR IV method, with information from $i-8$ to $i+8$ amino acids included (Table 4, GOR IV), is 68.2%. The accuracy of the GOR IV method in which the information parameters are calculated from the smaller original set of 267 protein structures (Table 4, GOR IV; Garnier *et al.*, 1996) is 67.3%. To reach this level of accuracy 17320 information parameters have to be obtained from the experimental protein structures. In contrast, the electrostatic screening model (model I) is able to achieve the same level of accuracy (Table 4, 68.7%) with only 40 parameters.

One may argue that the algorithm for predicting the secondary structures with model I does not differ significantly for the probabilistic methods like the Chou & Fasman (1974a,b) algorithm or the single residue GOR methods (Garnier *et al.*, 1978, 1996; Gibrat *et al.*, 1987), because the screening coefficients are derived from the experimental proteins structures. The screening coefficients ($\gamma^r_{nonlocal}$ and γ^r_{local}) in the electrostatic screening model (model I) would therefore correspond to the α-helical (P_α) and β-sheet (P_β) conformational parameters used in the Chou-Fasman algorithm or the single residue information parameters used in the GOR methods. The accuracy of the single residue GOR IV algorithm (Garnier *et al.*, 1978) with the optimized decision constants (Table 4, Single GOR IV_{dc}) is 58.9%. It is well known that averaging the single residue information over n_s neighbor amino acids can considerably improve the predictions (Garnier *et al.*, 1978). The n_s are run constants and measure the cooperativity between residues. If the decision and the run constants (Garnier *et al.*, 1978) are optimized in the single residue GOR IV algor-

ithm (Table 4, Single GOR IV$_{ns}$, the accuracy improves to 65.7%, which is well below the 68.7% achieved by the electrostatic screening model (model I). Note that the number of empirical parameters in the electrostatic screening is smaller (40 parameters; two per residue type) than in the single residue GOR method (60 parameters; three per residue type). This result shows that the reason for the accuracy achieved by the new method is not a consequence of the averaging of single residue information over a neighbor amino acids with the Lifson-Roig algorithm (Garner *et al.*, 1978).

The accuracy of the new secondary structure prediction algorithm is approaching to the accuracy of currently the best secondary structure prediction algorithm based on neural networks (≈72%; Rost & Sander, 1993, 1994). The accuracies of these two methods are not directly comparable because of the different contents of secondary structures due to the modified assignments of secondary structures (see Methods). There are important differences between the neural network algorithms and the new statistical mechanical method, which have to be pointed out. First, the main disadvantage of the neural network algorithms is that they do not provide any physical insight into forces which determine the protein secondary structure. On the other hand, our method based on the simple physical model for amino acid preferences and statistical mechanics allows us to identify the forces that are responsible for the stability of α-helices and β-strands in a particular protein. Second, only 40 screening coefficients are needed to reach the accuracy of 68.7% with the electrostatic screening model (model I). In contrast, many thousands of parameters (5000 to 15,000) have to be optimized during the training of the neural networks to obtain this level of accuracy (Rost & Sander, 1993, 1994). Third, in our method the range of interactions of an amino acid i is limited to short-range interactions with amino acids at spacings between $i-4$ and $i+4$, while in the neural network algorithms the windows usually contain 13 to 17 amino acids (Rost & Sander, 1993, 1994). Fourth, the new method based on the Lifson-Roig theory is very fast (it takes less than a second of computer time per protein) and the simple code can be easily implemented into any computer program for predicting the three-dimensional structure of proteins.

It is interesting to examine the secondary structure prediction ability of an early version of the electrostatic screening model, which was used to predict the three-dimensional structure of small peptides (Avbelj & Moult, 1995a). The strength of hydrogen bonds, which are the non-local main-chain electrostatic interactions, was assumed to be equal for all residue pairs with the screening coefficients γ_{nonlocal} of 0.38. The ability of the old free energy function to correctly predict the secondary structure in proteins and peptides is rather limited. The three-state accuracy Q_{total} is ≈51%. The Matthews (1975) correlation coefficient for predicting α-helices is 0.22, which is much smaller than the correlation coefficient of 0.57 obtained with the new free energy function (Table 4).

The influence of averaging the free energy profiles over homologous sequences on the accuracy of prediction algorithm is also examined. If the free energy profiles are not averaged over homologous sequences, the accuracy Q_{total} for 130 protein chains decreases from 68.7% and 63.1% (see Table 4, model Ia). Similar change in the accuracy has been obtained also by using the neural network (Rost & Sander, 1993) and the GOR IV algorithms (see Table 4, GOR IV$_a$). In order to investigate the reason for this change in the accuracy, the α-helix cutoff is iteratively adjusted to fit the predicted helix content to the experimental one. The Q_{total} value of such algorithm increases from 63.1 to 67.4 % (see Table 4, model Ib), which indicates that the change in the accuracy using the homologous sequences is predominantly due to the better predictions of α-helices. One possible reason may lay in solvent effects. It has been shown that the solvent has a considerable influence on the propensities of amino acids for different conformational states (Tanford *et al.*, 1960; Nelson & Kallenbach, 1986; Buck *et al.*, 1993; Jasanoff & Fersht, 1994; Waterhous & Johnson, 1994; Blanco *et al.*, 1994; Schonbrunner *et al.*, 1996; Luo & Baldwin, 1998). The reason is in different screening abilities of different solvents. The solvent compositions used in preparations of proteins varies considerably from protein to protein, therefore the zero cutoff approach used for the majority of predictions described in this work (except model Ia in Table 4) may not be entirely appropriate for the prediction of secondary structure of all proteins.

The present treatment of the helix-coil transition is valid only for α_R helices with the hydrogen bonds between amino acids at spacings i and $i+4$. The new secondary structure prediction algorithm can be further improved to predict the 3_{10} and π-helices, as well as for the predictions of other features in the proteins structures, like β-turns, etc.

Helix contents of alanine based peptides

Alanine-based peptides exhibit partial helix formation in water (Padmanabhan *et al.*, 1990). Chakrabartty *et al.* (1994) determined the helix contents of peptides homologous to the sequence K(AAAAK)$_3$. The helix contents were measured by UV circular dichroism spectroscopy. The experimental data have shown that the helix contents of peptides strongly depend on the type of blocking groups at both ends of a sequence (Chakrabartty *et al.*, 1994). The electrostatic screening parameters of blocking groups cannot be determined from the potentials of mean force based on the experimental proteins structures, therefore the new prediction method is applied to only those peptides with identical blocking groups. From the set of 58 peptides (Chakrabartty *et al.*, 1994) 37 peptides were

Table 5. Sequences and helix contents of peptides

No.	Sequence	f_{helix}^{exp}	f_{helix}^{calc}	f_{helix}^{Agadir}	f_{helix}^{SCINT}
1	Ac-YGKAAAAKAAAAKAAAAK-CONH2	0.737	0.552	0.666	0.703
2	Ac-YGGKAAAAKAAAAKAAAAK-CONH2	0.677	0.511	0.613	0.605
3	Ac-YGGGKAAAAKAAAAKAAAAK-CONH2	0.616	0.477	0.555	0.613
4	Ac-YGGKAAAAKAAAAKAAGAK-CONH2	0.465	0.412	0.407	0.484
5	Ac-YGGKAAGAKAAAAKAAAAK-CONH2	0.373	0.379	0.346	0.303
6	Ac-YGGKAAAAKAAGAKAAAAK-CONH2	0.256	0.298	0.280	0.218
7	Ac-YGGKAAAAKALAAKAAAAK-CONH2	0.546	0.469	0.571	0.586
8	Ac-TGGKAAAAKAMAAKAAAAK-CONH2	0.507	0.485	0.540	0.552
9	Ac-TGGKAAAAKAQAAKAAAAK-CONH2	0.507	0.459	0.513	0.548
10	Ac-YGGKAAAAKAIAAKAAAAK-CONH2	0.460	0.441	0.470	0.503
11	Ac-YGGKAAAAKACAAKAAAAK-CONH2	0.452	0.409	0.384	0.447
12	Ac-YGGKAAAAKASAAKAAAAK-CONH2	0.426	0.373	0.435	0.479
13	AcYGGKAAAAKAFAAKAAAAK-CONH2	0.426	0.435	0.451	0.421
14	Ac-YGGKAAAAKANAAKAAAAK-CONH2	0.397	0.412	0.402	0.426
15	Ac-YGGKAAAAKATAAKAAAAK-CONH2	0.295	0.364	0.350	0.357
16	Ac-YGGKAAAAKAPAAKAAAAK-CONH2	0.000	0.237	0.056	0.103
17	Ac-YGGKALAAKALAAKALAAK-CONH2	0.492	0.369	0.442	0.424
18	Ac-YGGKAMAAKAMAAKAMAAK-CONH2	0.364	0.411	0.374	0.307
19	Ac-YGGKAQAAKAQAAKAQAAK-CONH2	0.312	0.371	0.316	0.284
20	Ac-YGGKASAAKASAAKASAAK-CONH2	0.223	0.221	0.187	0.170
21	Ac-YGGKAIAAKAIAAKAIAAK-CONH2	0.216	0.302	0.199	0.193
22	Ac-YGGKANAAKANAAKANAAK-CONH2	0.163	0.294	0.149	0.124
23	Ac-YGGKAVAAKAVAAKAVAAK-CONH2	0.093	0.222	0.136	0.077
24	Ac-YGGKATAAKATAAKATAAK-CONH2	0.079	0.193	0.129	0.055
25	Ac-KALAAKALAAKAAAAKGGY-CONH2	0.487	0.428	0.460	0.532
26	Ac-KAMAAKAMAAKAAAAKGGY-CONH2	0.403	0.454	0.436	0.453
27	Ac-KAQAAKAQAAKAAAAKGGY-CONH2	0.395	0.421	0.430	0.435
28	Ac-KAAAAKAIAAKAAAAKGGY-CONH2	0.484	0.441	0.470	0.525
29	Ac-KAAAAKASAAKAAAAKGGY-CONH2	0.448	0.374	0.437	0.499
30	Ac-KAAAAKAVAAKAAAAKGGY-CONH2	0.433	0.397	0.433	0.431
31	Ac-KAAAAKANAAKAAAAKGGY-CONH2	0.428	0.412	0.405	0.446
32	Ac-KAAAAKATAAKAAAAKGGY-CONH2	0.402	0.365	0.355	0.377
33	Ac-YCGQAAAAQAAAAQAAAAQ-CONH2	0.518	0.519	0.436	0.523
34	Ac-YGGQAAAAQAQAAQAAAAQ-CONH2	0.378	0.469	0.322	0.370
35	Ac-YGGQAQAAQAQAAQAQAAQ-CONH2	0.266	0.380	0.172	0.145
36	Ac-KAAAAKAWAAKAAAAK-CONH2	0.504	0.480	0.549	0.569
37	Ac-KAAAAKAYAAKAAAAK-CONH2	0.507	0.475	0.531	0.643

The experimental helix contents f_{helix}^{exp} were measured by Chakrabartty *et al.* (1994). The helix contents f_{helix}^{calc} are calculated using equation (18). The helix contents f_{helix}^{Agadir} and f_{helix}^{SCINT} are calculated using Agadir (Munoz & Serrano, 1994, 1997) and SCINT (Stapley *et al.*, 1995; Rohl *et al.*, 1996) algorithms, respectively. The correlation coefficients between f_{helix}^{calc}, f_{helix}^{Agadir} and f_{helix}^{SCINT} with f_{helix}^{exp} are 0.91, 0.96 and 0.95, respectively.

selected containing Ac and $CONH_2$ as blocking groups. Table 5 shows the sequences, the experimental helix contents (f_{helix}^{exp}) and the calculated helix contents using three algorithms (f_{helix}^{calc}, f_{helix}^{Agadir}, and f_{helix}^{SCINT}). The f_{helix}^{calc} represents the helix contents calculated by the electrostatic screening model (model I) using equation (18). The f_{helix}^{Agadir} and the f_{helix}^{SCINT} are helix contents calculated by the Agadir (Munoz & Serrano, 1994; Munoz & Serrano, 1997) and the SCINT (Stapley *et al.*, 1995; Rohl *et al.*, 1996) algorithms, respectively. The Agadir and SCINT algorithms have been developed for the prediction of the helix contents in peptides using the helix-coil transition theory.

The correlation coefficient between (f_{helix}^{exp}) is 0.91 (Figure 5). This relatively high correlation coefficient indicates that the electrostatic interaction have a dominant role in determining the helix contents in these peptides. The correlation coefficients between f_{helix}^{Agadir} and f_{helix}^{SCINT} with f_{helix}^{exp} are 0.96 and 0.95, respectively. The f_{helix}^{Agadir} and the f_{helix}^{SCINT} correlate with the experimental data better than the f_{helix}^{calc}. The absolute values of f_{helix}^{Agadir} and f_{helix}^{SCINT} are also more accurate than the absolute values of f_{helix}^{calc}. However, there is an important difference between the new prediction method and the Agadir, SCINT, and ALB (Finkelstein & Ptitsyn, 1976; Ptitsyn & Finkelstein, 1983; Finkelstein *et al.*, 1991) algorithms. A number of parameters representing the side-chain interactions are included in those algorithms. The prediction method based on the electrostatic screening model (model 1) contains only 40 parameters representing the electrostatic interactions between the main-chain polar atoms.

The helix-coil transition parameters w_i and v_i, calculated by the electrostatic screening model (model I), are different from those obtained by host-guest and peptide studies (Wojcik *et al.*, 1990; Padmanabhan *et al.*, 1990; Lyu *et al.*, 1990; Chakrabartty *et al.*, 1991, 1994; Kemp *et al.*, 1991; Doig *et al.*, 1994). The reason lays in different definition of conformational states (see above). Probably the most important deficiency of the peptide helix-coil transitions studies is the assumption that w_i parameters depend only on the nature of one amino acid, i. In contrast, the w_i parameters calcu-

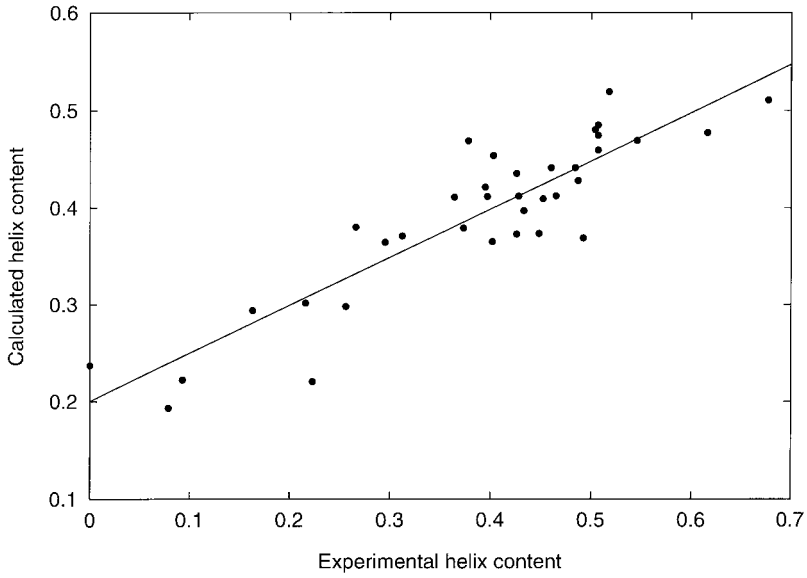

Figure 5. Correlation between experimental and calculated helix content. The experimental helix contents were measured by Chakrabartty *et al*., 1994). The fitted line represented with the equation $f_{\text{helix}}^{\text{calc}} = 0.496 f_{\text{helix}}^{\text{exp}} + 0.200$. The correlation coefficient is 0.91.

lated by the electrostatic screening model are sequence-dependent. They depend on the sequence of amino acids in a particular protein or peptide, because the strength of main-chain hydrogen bond depends on the nature of both amino acids involved (see above).

Implications for protein folding

In the first approximation, the screening coefficients are considered to be independent of the residue burial, however, there are indications that such dependence exists. The screening coefficients seem to be smaller for exposed residues than for buried amino acids. The dependence of the screening coefficients on the residue burial would explain the two state behavior of proteins in unfolding experiments. Kiefhaber *et al*. (Kiefhaber & Baldwin, 1995; Kiefhaber *et al*., 1995) demonstrated that in the process of unfolding or ribonuclease A, the entire main-chain hydrogen bond network disintegrates in a single rate-limiting step, resulting in the overall unfolding of protein. We suggest that the reason for this behavior might be in the weakening of the local and non-local main-chain electrostatic interactions due to the solvent screening. Breaking some of the hydrogen bonds would allow water to enter into the protein interior, which would destabilize other main-chain electrostatic interactions in a cooperative manner.

Conclusions

We show that the short-range main-chain electrostatic interactions (between amino acids at spacings i, $i+4$) are crucial in determining the secondary structure of proteins and peptides. The role of the hydrophobic effect and the side-chain conformational entropy is small. The three-state accuracy of the secondary structure prediction method, based on the electrostatic screening model and the Lifson-Roig theory (model I), is 68.7%. The short-range interactions used in the electrostatic screening model (model I) account for almost the entire accuracy achieved by currently the best secondary structure prediction algorithms (72%). The disadvantage of the neural network algorithms is that they do not provide any physical insight into forces which determine the protein secondary structure. In contrast the physical background of the new method is simple and well defined. Only two residue type dependent screening coefficients, a total of 40 parameters, are needed for relatively accurate predictions. Many thousands of parameters have to be optimized during the training of the neural networks to reach this level of accuracy. The accuracy of predicting strands by our method is better than in many other algorithms, although the long-range hydrogen bonds, usually present in β-sheets, are ignored. This result supports the hypothesis that β-strands are stabile structural elements even if they are not part of a β-sheet, because of the favorable local main-chain electrostatic interactions. The electrostatic screening model is also able to predict the helix contents in small peptides. The correlation coefficient between the calculated and the experimental helix contents for 37 alanine-based peptides is 0.91. The new secondary structure prediction method is very fast (it takes less than a second of computer time per protein) and the code can be easily implemented into any computer program for predicting the three-dimensional structure of proteins.

Methods

Secondary structure assignment

The modified Kabsch & Sander (1983) DSSP assignment of the secondary structures in proteins is used. The secondary structure of a residue is classified into eight

types. These classes are grouped into three larger types: α-helix (*H*), β-strand (*E*), and coil (*T*, *S*, *G*, *I*, *B*, and the rest of amino acids). The *G* and *I* classes, which represent amino acids in 3_{10} and π-helices, are considered to be coil, because the number of amino acids with this secondary structure in the database is too small for the reliable calculation of the mean pairwise hydrophobic accessible surface contacts areas and the conformational entropies.

Mean pairwise hydrophobic accessible surface contacts areas between side-chains in helices and strands

The mean pairwise hydrophobic accessible surface contact areas between side-chains ($\overline{A}_{ij}$) are used to estimate the average free energy contribution of the hydrophobic effect in the stabilization of secondary structures in proteins. The pairwise hydrophobic accessible surface contact area (A_{ij}) between side-chains *i* and *j* is defined as the surface area loss due to the contact between the hydrophobic atoms of two amino acids *i* and *j*. It is a sum of two terms: $A_i + A_j$. A_i is obtained as a difference between the non-polar accessible surface area of residue *i* in the presence of first two neighboring residues on each side, reduced for the accessible surface area of this residue in the presence of residues *j*. The contacts with other amino acids in a protein are ignored. The hydrophobic atoms are defined as the side-chain carbon and sulphur atoms of nine amino acids: Ala, Val, Ile, Leu, Phe, Met, Cys, Trp, and Tyr. The hydrophobic surface areas are calculated using Lee & Richards (1971) algorithm with Chothia (1976) radii.

The mean pairwise hydrophobic surface contact areas between side-chains in α-helices and β-strands are calculated as averages from the set of 328 high-resolution (resolution <2.0 Å and *R* factor <20%) X-ray structures of proteins from the Protein Data Bank with total of 95,413 amino acids. The modified Kabsch & Sander (1983) DSSP assignment of the secondary structures in proteins is used. The proteins can be accessed from the Protein Data Bank under the following codes: 1aaj, 1aal, 1aap, 1aba, 1abk, 1acb, 1acf, 1ads, 1afg, 1ahc, 1ake, 1alk, 1amp, 1ank, 1aoz, 1apm, 1arb, 1arp, 1ast, 1asz, 1bam, 1bbh, 1bit, 1bmd, 1bns, 1brn, 1brs, 1btl, 1byb, 1caa, 1cbs, 1ccr, 1cdc, 1cdg, 1cel, 1cew, 1cge, 1cgo, 1cgt, 1chm, 1chn, 1cho, 1cka, 1clc, 1cmb, 1cot, 1cpc, 1 cpm, 1cpn, 1crl, 1csh, 1csn, 1ctf, 1cth, 1cus, 1daa, 1ddt, 1dfn, 1drf, 1dsb, 1dts, 1ebh, 1edt, 1emy, 1enx, 1erl, 1esl, 1ezm, 1fas, 1fba, 1fdd, 1fdn, 1fgv, 1fia, 1flp, 1flr, 1flv, 1fna, 1frd, 1frp, 1frr, 1fus, 1fut, 1fxd, 1gbs, 1gcs, 1gia, 1gky, 1glq, 1glt, 1gma, 1gof, 1gox, 1gpb, 1gpr, 1hag, 1hbg, 1hfc, 1hil, 1hle, 1hml, 1hmt, 1hne, 1hpg, 1hpi, 1hpm, 1hsl, 1htr, 1huw, 1hyl, 1hyp, 1iag, 1icm, 1ida, 1ids, ligd, 1ilk, 1isa, 1isu, 1knb, 1knt, 1lcf, 1jct, 1lec, 1lga, 1lib, 1lki, 1lld, 1lmb, 1lst, 1lte, 1lts, 1lzl, 1mba, 1mdc, 1mfa, 1mfe, 1mjc, 1 mol, 1mpp, 1mrj, 1nar, 1nba, 1nci, 1nco, 1ndc, 1nfp, 1nhk, 1noa, 1npk, 1nsc, 1ntn, 1ofv, 1olb, 1onc, 1opa, 1opg, 1ova, 1oyb, 1pbp, 1pda, 1pga, 1pgb, 1pgs, 1pgx, 1php, 1pii, 1pk4, 1pmy, 1pne, 1poa, 1poc, 1poh, 1ppa, 1ppb, 1ppf, 1ppo, 1prn, 1pso, 1ptf, 1ptq, 1ptx, 1rcf, 1rds, 1rec, 1ris, 1rnh, 1rop, 1rro, 1rsy, 1rtp, 1sac, 1sar, 1sat, 1sbp, 1scs, 1sct, 1sem, 1sgt, 1sha, 1shb, 1shf, 1shg, 1slt, 1smr, 1sri, 1st3, 1tad, 1tag, 1tca, 1ten, 1thg, 1thm, 1thv, 1thw, 1tib, 1tml, 1ton, 1top, 1tph, 1trb, 1trk, 1tsp, 1ukz, 1wfa, 1wht, 1wtl, 1xib, 1xnb, 1xso, 1xya, 1xyn, 1yma, 2act, 2apr, 2ayh, 2bbk, 2cba, 2cdv, 2cga, 2ci2, 2cmd, 2cmm, 2cpl, 2cst, 2cut, 2cy3, 2cyr, 2dnj, 2dri, 2ebn, 2end, 2fb4, 2fbj, 2fcr, 2fgf, 2fx2, 2gct, 2gst, 2had, 2hbe, 2hpd, 2hpe, 2hpr, 2hts, 2imm, 2imn, 2kau, 2lig, 2mcg, 2mcm, 2mlt, 2mnr, 2msb, 2mye, 2nac, 2nad, 2pgd, 2pia, 2pkc, 2plt, 2por, 2psg, 2rhe, 2rn2, 2scp, 2sga, 2sil, 2spc, 2tgi, 2tir, 2trx, 2zta, 351c, 3app, 3bcl, 3c2c, 3chy, 3cox, 3dni, 3est, 3hhb, 3mcg, 3mds, 3pga, 3pte, 3rp2, 3rub,3tgl, 4blm, 4enl, 4fgf, 4pep, 4pti, 4q21, 5cha, 5can, 5p21, 5pal, 5rub, 6rlx, 6rxn, 7aat, 7fab, 7pcy, 8dfr, 8fab, 8pti.

The largest mean surface contact areas are between pairs of side-chains at spacings of *i*, $i+4$ ($\overline{A}^h_{i,i+4}$) and *i*, $i+3$ ($\overline{A}^h_{i,i+3}$) in α-helices and between pairs of side-chains at spacings of and *i*, $i+2$ ($\overline{A}^s_{i,i+2}$) in β-strands (Table 2). Similar results have been obtained from the Monte Carlo simulations of small model systems (Creamer & Rose, 1995). The pairwise contact areas of an amino acid in α-helices are not symmetric, because the C_β atoms point towards the N terminus of the α-helix (Creamer & Rose, 1995). Creamer & Rose (1995) have also shown that interactions in triples may exceed the sum of pairwise interactions in α-helices. These effects are ignored in the present treatment of the helix-coil transition.

The free energy contribution due to the hydrophobic interactions between side-chains *i* and *j* ($\Delta\overline{G}^{\text{hydro}}_{ij}$) in α-helices or β-strands is proportional to the areas $\overline{A}_{ij}$ (Chothia, 1974):

$$\Delta G^{\text{hydro}}_{ij} = -K\bar{A}_{ij} \tag{13}$$

The proportionality constant *K* was estimated to be between 0.014 and 0.024 kcal/(mol·Å^2) (Chothia, 1974; Eisenberg & McLachlan, 1986; Murphy & Gill, 1990, 1991; Eriksson *et al.*, 1992). The constant *K* is calculated from the potentials of mean force base on the set of 328 experimental protein structures and is found to be 0.015 kcal/(mol·Å^2) (Avbelj, 1992; Avbelj & Fele, 1998).

Conformational Entropy

The differences in conformational entropies between folded and denatured states are derived from the probability distributions of rotamers in the experimental protein structures (McGregor *et al.*, 1987; Pickett & Sternberg, 1993; Blaber *et al.*, 1994; Doig & Sternberg, 1995) using equation (McQuarrie, 1976):

$$S = -R\sum_i p_i \ln p_i \tag{14}$$

where p_i represents the probability of an amino acid being in rotamer class *i* and *R* is the gas constant. The conformational entropies of amino acids in the α-helical and β-strand states are calculated from the rotamer distributions of amino acids in the α-helices and β-strands in proteins, respectively. The amino acids within four residues of the N or C terminus of the helix and within one residue of both ends of the strands are considered as coil. The conformational entropies of amino acids in the unfolded state are calculated from the rotamer distribution in those states which are not considered to be part of the α-helices or β-strands. The conformations of these states are assumed to represent the conformations in the unfolded state (Pickett & Sternberg, 1993; Doig & Sternberg, 1995). This definition of the coil state differs from the definition used by Pickett & Sternberg (1993) in which the coil state includes the amino acids in β-strands.

The secondary structures in proteins are assigned by the modified Kabsch & Sander (1983) DSSP algorithm. The side-chain as well as main-chain conformational entropies of amino acids are calculated. The Pickett &

Sternberg (1993) definition of rotamer classes is used. The main-chain conformational entropies are calculated using four rotamers for torsion angles ϕ and ψ in the Ramachandran plot: α_R $(-180 < \phi < 0;\ -90 < \psi < +90)$, β $(-180 < \phi < 0;\ -90 > \psi > +90)$, α_L $(-0 < \phi < 180;\ -90 < \psi < +90)$, and β' $(-0 < \phi < 180;\ -90 > \psi > +90)$. The effect of residue burial on the rotamer distribution is ignored. The temperature T is 300 K. The results are shown in Table 3.

The difference in side-chain conformational entropies between α-helical and denatured states ($T\Delta S_x$) calculated in this study (Table 3) moderately correlates with the results obtained by the Monte Carlo simulations (Creamer & Rose, 1992). The correlation coefficient for eight non-polar amino acids is 0.69. The side-chain conformational entropies calculated here correlate with the results obtained by other authors. For example, the correlation coefficient between the side-chain conformational entropies of coil state (TS_{coil}) and those by Pickett & Sternberg (1993) ($TS_{coil}^{Pickett}$) is 0.997 (see Table 3).

Electrostatic screening model

The electrostatic screening model is explained in detail elsewhere (Avbelj & Moult, 1995b). Here, we describe only the main points which follow from the model. The residue-dependent strengths of the main-chain electrostatic interactions have been shown to correlate with the preferences of 20 amino acids, not only for the α-helical but also for the β-strand and other main-chain conformational states in the experimental protein structures (Avbelj & Moult, 1995b). In contrast, the hydrophobic effect and the conformational entropies correlate with the experimental data only for the α-helix forming tendencies. In the electrostatic screening model of amino acid preferences, the stability of a main-chain conformational state of an amino acid in a proteins depends primarily on the strengths of local and short-range non-local main-chain electrostatic interactions. The strength of local and non-local electrostatic interactions is related to the electrostatic screening with solvent and protein groups. The local main-chain electrostatic interactions are primarily due to the interaction of the main-chain CO and NH groups within an amino acid. The non-local main-chain electrostatic interactions are predominantly due to the main-chain hydrogen bonding. Note the differences between local–non-local and short-range–long-range interactions. Short-range interactions are interactions between amino acids less than four residues apart in the sequence. Long-range interactions are interactions between amino acids distant in the sequence.

The relative free energy G of a residue i as a function of the main-chain conformation is proportional to the point charge main-chain electrostatic interaction energies, with coefficients $\gamma^r_{nonlocal}$ and γ^r_{local} dependent on amino acid type r (Avbelj & Moult, 1995b):

$$G^i = \gamma^r_{nonlocal} E^i_{nonlocal} + \gamma^r_{local} E^i_{local} + C \qquad (15)$$

E^i_{local} and $E^i_{nonlocal}$ denote the local and non-local main-chain electrostatic energies of a residue i. The coefficients $\gamma^r_{nonlocal}$ and γ^r_{local} represent the attenuation of the electrostatic energies $E_{nonlocal}$ and E_{local}, respectively, due to the electrostatic screening. C is an undefined constant.

According to the electrostatic screening model (Avbelj & Moult, 1995b), the strengths of the main-chain electrostatic interactions depend on the screening by solvent and other protein dipoles in the vicinity. In the first approximation the screening coefficients are assumed to depend only on the types of amino acids involved. However, there are indications that the screening coefficients may depend also on the residue burial for at least some residues. The data base of protein structures is currently too small to clarify this point.

The screening coefficients are determined from the potentials of mean force as described elsewhere (Avbelj & Moult, 1995b; Avbelj & Fele, 1998). E_{local} and $E_{nonlocal}$ are calculated using Coulomb's law with a dielectric constant of 1.0. Point atomic charges for the main-chain atoms N, H_N, C, and O are −0.28, +28, +0.38, and −0.38 electrons, respectively (Avbelj & Moult, 1995b). Interactions between atoms within the NH and CO dipoles are ignored. Interactions between dipoles are included in the electrostatic energy, if the distance between the N or C atoms is smaller than 6.5 Å. The potentials of mean force are calculated from the set of 328 experimental protein structures dividing the E_{local} range between −4.4 kcal/mol and 3.0 kcal/mol into 31 equal bins and counting the population of each residue type in these bins. The base lines are excluded from the fitting procedure. The screening coefficients $\gamma^r_{nonlocal}$ and γ^r_{local} and the residual standard deviations of the fit are shown in Table 1.

Helix-coil transition

The Lifson-Roig theory (Lifson & Roig, 1961; Qian & Schellman, 1992) has been used in studying the helix-coil transition of polypeptides. In this theory each residue in a polypeptide sequence is considered to exist in one of two states: α conformation or coil conformation. A residue is classified as being in α conformation, if the torsion angles ϕ and ψ are in the α_R ($\phi \approx -60^\circ$, $\psi \approx -40^\circ$) region of the Ramachandran plot. The remainder of the conformational space is considered as coil. Note that the α conformation does not mean a residue is in the α-helical state. In the α-helical state the CO of residue $i-2$ is hydrogen-bonded to the NH of residue $i+2$. The torsion angles of three consecutive residues $i-1$, and $i+1$ must therefore be confined to the α conformation to be able to form a hydrogen bond. This is the origin of the cooperative nature of α-helix formation (Lifson & Roig, 1961; Qian & Schellman, 1992).

The equilibrium between three conformational states of amino acids in proteins (α-helix, β-strand and coil) is studied here, therefore we modified the definition of conformational states. The α-helical and β-strand states in a protein are those states which are assigned as α-helices and β-strands, respectively, by the modified Kabsch & Sander (183) DSSP algorithm (see above). The state of the remainder of amino acids in a protein is defined as coil.

The Lifson-Roig theory gives each residue in a sequence a statistical weight, depending on its own state and the states of the two residues to either side (triplets). These weights are 1, v, and w. The statistical weight of coil has been arbitrary set to 1. The helix-coil transition parameter v represents the equilibrium constant for formation of an α conformation in a coil. The parameter v is smaller than one, because fixing a residue in the α conformation is unfavourable. The parameter w is the equilibrium constant for formation of an α conformation and a hydrogen bond in a coil.

The partition function for heteropolymer is a sum of all possible products, obtained by the matrix method (Lifson & Roig, 1961; Qian & Schellman, 1992; Doig *et al.*, 1994):

$$Z = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \prod_{j=1}^{n} \begin{vmatrix} w_j & v_j & 0 \\ 0 & 0 & 1 \\ v_j & v_j & 1 \end{vmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \tag{16}$$

The probability $p(i)_{\mathrm{hb}}$ that the ith amino acid in a chain of n residues is hydrogen-bonded is given by:

$$p(1)_{\mathrm{hb}} = \frac{\delta \ln Z}{\delta \ln w_i} = \frac{1}{Z} \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \prod_{j=1}^{j=i-1} \begin{vmatrix} w_j & v_j & 0 \\ 0 & 0 & 1 \\ v_j & v_j & 1 \end{vmatrix} \begin{vmatrix} w_i & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{vmatrix} \prod_{j=i+1}^{j=n} \begin{vmatrix} w_j & v_j & 0 \\ 0 & 0 & 1 \\ v_j & v_j & 1 \end{vmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \tag{17}$$

The helix content $f_{\mathrm{helix}}^{\mathrm{calc}}$ is defined as:

$$f_{\mathrm{helix}}^{\mathrm{calc}} = \frac{\delta \ln Z}{\delta \ln w} = \sum_{i}^{n} p(i)_{\mathrm{hb}}/n \tag{18}$$

the probability $p(i)_{\mathrm{nuc}}$ that the ith amino acid in a chain of n residues has a weighting of v is given by:

$$p(i)_{\mathrm{nuc}} = \frac{\delta \ln Z}{\delta \ln v_i} = \frac{1}{Z} \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \prod_{j=1}^{j=i-1} \begin{vmatrix} w_j & v_j & 0 \\ 0 & 0 & 1 \\ v_j & v_j & 1 \end{vmatrix} \begin{vmatrix} 0 & v_i & 0 \\ 0 & 0 & 0 \\ v_i & v_i & 0 \end{vmatrix} \prod_{j=i+1}^{j=n} \begin{vmatrix} w_j & v_j & 0 \\ 0 & 0 & 1 \\ v_j & v_j & 1 \end{vmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \tag{19}$$

We define the helix free energy profile from the ratio between probability of hydrogen-bonded residues $p(i)_{\mathrm{hb}}$ and probability of coil residues $p(i)_{\mathrm{coil}}$ at the ith position in the sequence:

$$G_{\mathrm{hb}}^{i} = -kT \ln \frac{p(i)_{\mathrm{hb}}}{p(i)_{\mathrm{coil}}} \tag{20}$$

where $p(i)_{\mathrm{coil}}$ is:

$$p(i)_{\mathrm{coil}} = 1 - p(i)_{\mathrm{hb}} - p(i)_{\mathrm{nuc}} \tag{21}$$

Representative set of proteins used in predictions

We have chosen the set of 130 proteins used by Rost & Sander (1993) for testing the prediction algorithms. These are no homologous pairs of sequences between this set of proteins and the set which is used to obtain the mean strengths of the most important interactions. The protein chains can be accessed from the Protein Data Bank under the following codes: 1acx, 1azu, 1bbp_A, 1bds, 1bmv_1, 1bmv_2, 1cbh, 1cc5, 1cdt_A, 1crn, 1xse_I, 1eca, 1etu, 1fc2_C, 1fsl_H, 1fdx, 1fkf, 1fnd, 1fxi_A, 1gdl_O, 1gpl_A, 1hip, 1il8_A, 1l58, 1lmb_3, 1mcp_L, , 1mrt, 1ovo_A, 1paz, 1ppt, 1prc_C, 1prc_H, 1prc_L, 1prc_M, 1pyp, 1r09_2, 1rbp, 1rhd, 1s01, 1sh1, 1tgs_I, 1tnf_A, 1ubq, 1wsy_A, 1wsy_B, 256b_A, 2aat, 2ak3_A, 2alp, 2cab, 2ccy_A, 2cyp, 2fxb, 2gbp, 2gcr, 2gls_A, 2gn5, 2hmz_A, 2i1b, 2lbh, 2lrn_A, 2ltn_B, 2mev_4, 2mhu, 2or1_L, 2pab_A, 2pcy, 2phh, 2rsp_A, 2sns, 2sod_B, 2stv, 2tgp_I, 2tmv_P, 2tsc_A, 2utg_A, 2wrp_R, 3ait, 3b5c, 3blm, 3cd4, 3cla, 3cln, 3ebx, 3gap_A, 3hmg_A, 3hmg_B, 3icb, 3pgm, 3rnt, 3sdh_A, 3tim_A, 4bp2, 4cms, 4cpa_I, 4cpv, 4fxn, 4gr1, 4pfk, 4rhv_1, 4rvh_3, 4rhv_4, 4rxn, 4sgb_I, 4ts1_A, 4xia_A, 5cyt_R, 5er2_E, 5hvp_A, 5ldh, 5lyz, 6can, 6cpa, 6cpp, 6cts, 6dfr, 6hir, 6tmn_E, 7cat_A, 7icd, 7rsa, 8abp, 8adh, 9api_A, 9api_A, 9api_B, 9ins_B, 9pap, 9wga_A. The average contents of the α-helical and β-strand states in this set of proteins with 24,436 residues are 28.5% and 20.4%, respectively.

Calculating helix and strand free energy profiles

The average values of electrostatic energies $\overline{E}_{\mathrm{local}}^{h}$, $\overline{E}_{\mathrm{local}}^{ee}$, and $\overline{E}_{\mathrm{local}}^{ei}$, used in equations (1) to (10) (see Results and Discussion), are calculated from 328 experimental structures and the values obtained are: 1.475 kcal/mol, −2.335 kcal/mol, and −2.984 kcal/mol, respectively. The average local electrostatic energy of the coil state ($\overline{E}_{\mathrm{local}}^{c}$) of −2.030 kcal/mol is used in the calculations of both helix and strand free energy profiles. The constant K (in equation (13)) is calculated from the potentials of mean force based on the set of 328 experimental protein structures and is found to be 0.015 kcal/(mol·Å^2) (Avbelj, 1992; Avbelj & Fele, 1998). In order to get the correct balance between α-helices and β-strands, some parameters have to be optimized depending on the model. In model I the helix and strand free energy profiles are calculated using the electrostatic screening model. The values of $\overline{E}_{\mathrm{hb}}$ are optimized to −5.395 kcal/mol and −3.650 kcal/mol for the predictions of protein and peptide secondary structures, respectively. The smaller value of $\overline{E}_{\mathrm{hb}}$ for peptides can be rationalized if we take into account that the amino acids are more exposed to the solvent in peptides than they are in proteins. Decreasing the $\overline{E}_{\mathrm{hb}}$ value for peptides is analogous to linearly decreasing the screening coefficients of nonlocal interactions. (See also sections: Electrostatic screening model and Implications for protein folding.) In model II the free energy terms due to the hydrophobic effect and the side-chain conformational entropy are added in the electrostatic screening model (model I) to describe the helix-coil transition. The value of $\overline{E}_{\mathrm{hb}}$ is optimized to −5.555 kcal/mol. In the model III solely the hydrophobic and entropy effects are utilized. The value of constant K (in equation (13)) is optimized to 0.105 kcal/(mol·Å^2).

The free energy profiles of homologous sequences of a protein chain are taken from the HSSP database of the homology-derived protein structures (Sander & Schneider, 1991). A total of 4549 homologous sequences of 130 protein chains are used. The insertions and deletions in the aligned homologous sequences are ignored.

Secondary structure prediction algorithm

The conformational state of an amino acid is predicted from the helix and strand free energy profiles and the α-helix and β-strand cutoffs. The cutoffs are used as sequence-independent thresholds for the prediction. The following rules are applied. (a) An amino acid is in the coil state, if the helix and strand free energies are both larger than the corresponding cutoffs. (b) An amino acid is in the α-helical state, if the helix free energy is smaller than the α-helix cutoff and strand free energy larger than

the β-strand cutoff. (c) An amino acid is in the β-strand state, if the strand free energy is smaller than the β-strand cutoff and helix free energy larger than the α-helix cutoff. (d) If the helix and strand energies of a residue are both smaller than the corresponding cutoffs, then the areas of the free energy peaks below the cutoff lines are compared. If the area of the negative peak of helix energy profile is larger than the area of the negative peak of strand free energy profile, all residues within this peak are predicted to be α-helical, and *vice versa*. Note the difference between this algorithm and the ''winner-takes-all'' procedure usually used in the prediction algorithms (Rost & Sander, 1993), in which residues are treated individually.

The requirement that the predicted α-helices and β-strands must contain at least three and two consecutive amino acids, respectively, has been applied.

Acknowledgments

We are grateful to D. Hadži, J. Moult, D. Kocjan, and R. Jerala for reading the manuscript and helpful suggestions. We thank T. M. Klinger for the computer program of the Lifson-Roig algorithm. This work was supported by the Ministry of Science and Technology of Slovenia and the PECO grant by the Commission of the European Communities.

References

Avbelj, F. (1992). Use of potential of mean force to analyse free energy contributions in protein folding. *Biochemistry*, **31**, 6290–6297.

Avbelj, F. & Moult, J. (1995a). The conformation of folding initiation sites in proteins determined by computer simulation. *Proteins: Struct. Funct. Genet.* **23**, 129–141.

Avbelj, F. & Moult, J. (1995b). Role of electrostatic screening in determining protein main-chain conformational preferences. *Biochemistry*, **34**, 755–764.

Avbelj, F. & Fele, L. (1998). Prediction of the three-dimensional structure of proteins using the electrostatic screening model and hierarchic condensation. *Proteins: Struc. Funct. Genet.* **31**, 74–96.

Bai, Y. & Englander, S. W. (1994). Hydrogen bond strength and β-sheet propensities: the role of a side chain blocking effect. *Proteins: Struct. Funct. Genet.* **18**, 262–266.

Bai, Y., Milne, J. S., Mayne, L. & Englander, S. W. (1993). Primary structure effects on peptide group hydrogen exchange. *Proteins: Struct. Funct. Genet.* **17**, 75–86.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Jr, E, F. N., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tusami, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Blaber, M., Zhang, X. & Matthews, B. W. (1993). Structural basis of amino acid α-helix propensity. *Science*, **260**, 1637–1640.

Blaber, M., Zhang, X., Lindstrom, J. D., Pepiot, S. D., Basse, W. A. & Matthews, B. W. (1994). Determination of α-helix propensity within the context of a folded protein. *J. Mol. Biol.* **235**, 600–624.

Blanco, F. J., Jimenez, M. A., Pineda, A., Rico, M., Santoro, J. & Nieto, J. L. (1994). NMR solution structure of the isolated *n*-terminal fragment of protein – gbl. *Biochemistry*, **33**, 6004–6014.

Bodkin, M. J. & Goodfellow, J. M. (1995). Hydrophobic solvation in aqueous trifluoroethanol solutions. *Biopolymers*, **39**, 43–50.

Brant, D. A. & Flory, P. J. (1965a). The configuration of random polypeptide chains. ii. Theory. *J. Am. Chem. Soc.* **87**, 2791–2800.

Brant, D. A. & Flory, P. J. (1965b). The role of dipole interactions in determining polypeptide conformation. *J. Am. Chem. Soc.* **87**, 663–664.

Buck, M., Radford, S. E. & Dodson, C. M. (1993). A partially folded state of hen egg while lysozyme in trifluoroethanol. *Biochemistry*, **32**, 668–678.

Chakrabartty, A., Shellman, J. A. & Baldwin, R. L. (1991). Large differences in the helix propensities of alanine and glycine. *Nature*, **351**, 586–588.

Chakrabartty, A., Kortemme, T. & Baldwin, R. L. (1994). Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Proteins Sci.* **3**, 843–852.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature*, **248**, 338–339.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1–14.

Chou, P. Y. & Fasman, G. D. (1974a). Conformational parameters for amino acid in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**, 211–222.

Chou, P. Y. & Fasman, G. D. (1974b). Prediction of protein structure. *Biochemistry*, **13**, 222–235.

Creamer, T. C. & Rose, G. D. (1994). α-Helix-forming propensities in peptides and proteins. *Proteins: Struct. Funct. Genet.* **19**, 85–97.

Creamer, T. P. & Rose, G. D. (1992). Side-chain entropy opposes alpha-helix formation by rationalizes experimentally determined helix-forming propensities. *Proc. Natl Acad. Sci. USA*, **89**, 5937–5941.

Creamer, T. P. & Rose, G. D. (1995). Interactions between hydrophobic side chains within α-helices. *Protein Sci.* **4**, 1305–1314.

Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. & Chan, H. S. (1995). Principles of proteins folding – a perspective from simple exact models. *Protein Sci.* **4**, 561–602.

Doig, A. J. & Sternberg, M. J. E. (1995). Side-chain conformational entropy in protein folding. *Protein Sci.* **4**, 2247–2251.

Doig, A. J., Chakrabartty, A., Klinger, T. M. & Baldwin, R. L. (1994). Determination of free energies of n-capping inα-helices by modification of the lifson-roig helix-coil theory to include n- and c-capping. *Biochemistry*, **33**, 3396–3403.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature* **319**, 199–203.

Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M. & Matthews, E. P. B. B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**, 178–183.

Finkelstein, A. V. & Ptitsyn, O. B. (1976). A theory of protein molecule self-organization. *J. Mol. Biol.* **103**, 15–24.

Finkelstein, A. V., Badretdinov, A. Y. & Ptitsyn, O. B. (1991). Physical reasons for secondary structure

stability: α-helices in short peptides. *Proteins: Struct. Funct. Genet.* **10**, 287–299.
Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.
Garnier, J., Gibrat, J. F. & Robson, B. (1996). Gor method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266**, 540–553.
Gibrat, J. F., Garnier, J. & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* **198**, 425–443.
Habermann, S. M. & Murphy, K. P. (1996). Energetics of hydrogen bonding in proteins: a model compound study. *Protein Sci.* **5**, 1229–1239.
Hermans, J., Anderson, A. G. & Yun, R. H. (1992). Differential helix propensity of small apolar side chain studied by molecular dynamics simulation. *Biochemistry*, **31**, 5646–6553.
Holley, L. H. & Karplus, M. (1989). Proteins secondary structure prediction with a neural network. *Proc. Natl Acad. Sci. USA,* **86**, 152–156.
Jasanoff, A. & Fersht, A. R. (1994). Quantitative determination of helical propensities from trifluoroethanol titration curves. *Biochemistry,* **33**, 2129–2135.
Kabsch, W. & Sander, C. (1983). Dictionary of proteins structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577–2637.
Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Advan. Protein Chem.* **14**, 1–64.
Kemp, D. S., Boyd, J. G. & Muendel, C. C. (1991). The helical s constant for alanine in water derived from template-nucleated helices. *Nature,* **352**, 451–454.
Kiefhaber, T. & Baldwin, R. L. (1995). Kinetics of hydrogen bond breakages in the process of unfolding of ribonuclease a measured by pulsed hydrogen exchange. *Proc. Natl Acad. Sci. USA,* **92**, 2657–2661.
Kiefhaber, T., Labhardt, A. M. & Baldwin, R. L. (1995). Direct nmr evidence for an intermediate preceding the rate limiting step in the unfolding of ribonuclease A. *Nature,* **375**, 513–515.
Klotz, I. M. & Franzen, J. S. (1962). Hydrogen bonds between model peptide groups in solution. *J. Am. Chem. Soc.* **84**, 3461–3466.
Kneller, D. G., Cohen, F. E. & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **214**, 171–182.
Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
Lewis, P. N., Go, N., Go, M., Kotelchuck, D. & Scheraga, H. A. (1970). Helix probability profiles of denatured proteins and their correlation with native structure. *Proc. Natl Acad. Sci. USA,* **65**, 810–815.
Lifson, S. & Roig, A. (1961). On the theory of helix-coil transition in polypeptides. *J. Chem. Phys.* **34**, 1963–1974.
Lim, V. I. (1974). Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.* **88**, 873–894.
Luo, P. & Baldwin, R. L. (1998). Analysis of the stabilization of peptide helices by trifluoroethanol. *Protein Sci.* in the press.
Lyu, P. C., Liff, M. I., Marky, L. A. & Kallenbach, N. R. (1990). Side chain contribution to the stability of alpha-helical structure in proteins. *Science,* **250**, 669–673.
Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta,* **405**, 442–451.
McGregor, M. J., Islam, S. A. & Sternberg, M. J. E. (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol.* **198**, 295–310.
McQuarrie, D. A. (1976). In *Statistical Mechanics,* chapt. 13, Harper and Row, New York.
Munoz, V. & Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. *Nature Struct. Biol.* **1**, 399–409.
Munoz, V. & Serrano, L. (1997). Development of the multiple sequence approximation with the agadir model of α-helix formation. *Biopolymers,* **41**, 495–509.
Murphy, K. P. & Gill, S. J. (1990). Group additivity thermodynamics for dissolution of solid cyclic dipeptides into water. *Thermochim. Acta,* **172**, 11–20.
Murphy, K. P. & Gill, S. J. (1991). Solid model compounds and the thermodynamics of protein unfolding. *J. Mol. Biol.* **222**, 699–709.
Myers, J. K. & Pace, C. N. (1996). Hydrogen bonding stabilizes globular proteins. *Biophys. J.* **71**, 2033–2039.
Nelson, J. W. & Kallenbach, N. R. (1986). Stabilization of the ribonuclease s-peptide α-helix by trifluoroethanol. *Proteins: Struct. Funct. Genet.* **1**, 211–217.
Padmanabhan, S. & Baldwin, R. L. (1991). Straight-chain non-polar amino acids are good helix-formers in water. *J. Mol. Biol.* **219**, 135–137.
Padmanabhan, S., Marqusee, S., Ridgeway, T., Laue, T. M. & Baldwin, R. L. (1990). Relative helix-forming tendencies of non-polar amino acids. *Nature,* **344**, 268–270.
Padmanabhan, S., York, E. J., Gera, L., Stewart, J. M. & Baldwin, R. L. (1994). Helix-forming tendencies of amino acids in short (hydroxybutyl)-l-glutamine peptides: an evaluation of the contradictory results from host-guest studies and short alanine-based peptides. *Biochemistry,* **33**, 8604–8609.
Pauling, L. & Corey, R. B. (1951). *Proc. Natl Acad. Sci. USA,* **37**, 729–740.
Pauling, L., Corey, R. B. & Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of polypeptite chain. *Proc. Natl Acad. Sci. USA,* **37**, 205–211.
Pickett, S. D. & Sternberg, M. J. E. (1993). Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.* **231**, 825–839.
Piela, L., Nemethy, G. & Sheraga, H. A. (1987). Conformational constraints of amino acid side chain in alpha-helices. *Biopolymers,* **26**, 1273–1286.
Poland, D. & Scheraga, H. A. (1967). *Poly-Alpha-Amino Acids: Protein Models for Conformational Studies* (Fasman, G. D., ed.), pp. 391–497, Marcel Decker, New York.
Ptitsyn, O. B. & Finkelstein, A. V. (1983). Theory of protein secondary structure and algorithm of its prediction. *Biopolymers,* **22**, 15–25.
Qian, H. (1996). Prediction of α-helices in proteins based on thermodynamic parameters from solution chemistry. *J. Mol. Biol.* **256**, 663–666.
Qian, H. & Schellman, J. A. (1992). Helix-coil theories: a comparative study for finite length polypeptides. *J. Phys. Chem.* **96**, 3987–3995.

Rohl, C. A., Chakrabartty, A. & Baldwin, R. L. (1996). Helix propagation and n_{cap} propensities of the amino acids measured in alanine-based peptides in 40 volume procent trifluoroethanol. *Protein Sci.* **5**, 2623–2637.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure a better than 70 accuracy. *J. Mol. Biol.* **232**, 584–599.

Rost, B. & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Struct. Funct. Genet.* **19**, 55–72.

Sander, C. & Schneider, R. (1991). Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Scholtz, J. M., Marqusee, S., Baldwin, R. L., York, E. J., Stewart, J. M., Santoro, M. & Bolen, D. W. (1991). Calorimetric determination of the enthalpy change for the α-helix to coil transition of an alanine peptide in water. *Proc. Natl Acad. Sci. USA,* **88**, 2854–2858.

Schonbrunner, N., Wey, J., Engels, J., Georg, H. & Kiefhaber, T. (1996). Native like β-structure in a tfe induced partially folded state of the all-β-sheet protein tendamistat. *J. Mol. Biol.* **260**, 432–445.

Shellman, J. A. (1955). Thermodynamics f urea solutions and the heat of formation of the peptide hydrogen bonds. *C. V. Trav. Lab. Carlsberg Ser. Chim.* **29**, 223–229.

Shulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure, chapt. 8,* Springer Verlag Inc., New York.

Sippl., M. J. (1996). Helmholtz free energy of peptide hydrogen bonds in proteins. *J. Mol. Biol.* **260**, 644–648.

Stapley, B. J., Rohl, C. A. & Doig, A. J. (1995). Addition of side chain interaction in modified lifson-roig helix-coil theory: application to energetics of phenylalanine methionine interactions. *Proteins Sci.* **4**, 2383–2391.

Tanford, C., De, P. K. & Taggart, V. G. (1960). The role of the α-helix in the structure of proteins: optical rotary dispersion of β-lactoglobulin. *J. Am. Chem. Soc.* **82**, 6028–6034.

Thomas, P. D. & Dill, K. A. (1993). Local and nonlocal interactions in globular proteins and mechanism of alcohol denaturation. *Protein Sci.* **2**, 2050–2065.

Vila, J., Williams, R. L., Grant, J. A., Wojcik, J. & Scheraga, H. A. (1992). The intrinsic helix-forming tendency of l-alanine. *Proc. Natl Acad. Sci. USA,* **89**, 7821–7825.

Wang, J. & Purisima, E. O. (1996). Analysis of thermodynamic determinants in helix propensities of nonpolar amino acids through a novel free energy calculation. *J. Am. Chem. Soc.* **118**, 995–1001.

Warshel, A. & Russell, S. T. (1984). Calculation of electrostatic interactions in biological systems and in solutions. *Qart. Rev. Biophys.* **17**, 283–422.

Waterhous, D. V. & Johnson, W. C. (1994). Importance of environment in determining secondary structure in proteins. *Biochemistry,* **33**, 2121–2128.

Wojcik, J., Altmann, K.-H. & Scheraga, H. A. (1990). Helix-coil stability for the naturally occuring amino acids in water. XXIV. Half-cysteine parameters from random poly(hydroxybutylglutamine-co-*S*-methylthio-L-cysteine). *Biopolymers,* **30**, 121–134.

Yang, A. & Honig, B. (1995). Free energy determinants of secondary structure formation: I. α-Helices. *J. Mol. Biol.* **252**, 351–365.

Yun, R. H. & Hermans, J. (1991). Conformational equilibria of valine studied by dynamics simulation. *Proteins Eng.* **4**, 761–766.

Yun, R. H., Anderson, A. & Hermans, J. (1991). Proline in alpha-helix: stability and conformation studied by dynamics simulation. *Proteins: Struct. Funct. Genet.* **10**, 219–228.

Zimm, B. H. & Bragg, J. K. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* **31**, 526–535.

Edited by J. Thornton

(Received 29 September 1997; received in revised form 19 February 1998; accepted 12 March 1998)