PROTEINS: Structure, Function, and Genetics 31:74–96 (1998)

Prediction of the Three-Dimensional Structure of Proteins Using the Electrostatic Screening Model and Hierarchic Condensation

Franc Avbelj* and Ljudmila Fele
National Institute of Chemistry, Ljubljana, Slovenia

ABSTRACT **We describe a method for predicting the three-dimensional (3-D) structure of proteins from their sequence alone. The method is based on the electrostatic screening model for the stability of the protein main-chain conformation. The free energy of a protein as a function of its conformation is obtained from the potentials of mean force analysis of high-resolution x-ray protein structures. The free energy function is simple and contains only 44 fitted coefficients. The minimization of the free energy is performed by the torsion space Monte Carlo procedure using the concept of hierarchic condensation. The Monte Carlo minimization procedure is applied to predict the secondary, super-secondary, and native 3-D structures of 12 proteins with 28–110 amino acids. The 3-D structures of the majority of local secondary and super-secondary structures are predicted accurately. This result suggests that control in forming the native-like local structure is distributed along the entire protein sequence. The native 3-D structure is predicted correctly for 3 of 12 proteins composed mainly from the α-helices. The method fails to predict the native 3-D structure of proteins with a predominantly β secondary structure. We suggest that the hierarchic condensation is not an appropriate procedure for simulating the folding of proteins made up primarily from β-strands. The method has been proved accurate in predicting the local secondary and super-secondary structures in the blind ab initio 3-D prediction experiment. Proteins 31:74–96, 1998.** © 1998 Wiley-Liss, Inc.

Key words: Monte Carlo minimizations in torsion space; prediction of secondary structure; protein folding

INTRODUCTION

Prediction of the correct three-dimensional (3-D) structure of proteins, knowing only their sequence, is one of the most challenging problems in molecular biology. Many globular proteins can spontaneously refold in vitro after being completely unfolded; therefore the amino acid sequence of a protein contains all information required to generate its 3-D structure.[1] One-domain proteins fold into their stable structures in approximately 1 second. As shown by Levinthal,[2] there must be some kind of pathway by which the folding is directed from any accessible unfolded state to the native conformation, because only a small fraction of conformations available to the protein can be searched for randomly in such a short time. It has been shown recently that the native state is not necessarily the lowest free energy state of a protein,[1] but it is the lowest free energy state along the single or multiple pathways in the folding process.[3–5] For the ultimate goal, which is the prediction of the native 3-D structure of proteins from their amino acid sequence alone, the folding pathway has to be properly taken into account in the 3-D structure prediction algorithms.

Several mechanisms of the protein folding process with different pathways have been proposed.[6–8] In the framework model of the protein folding, the secondary structures are formed early in the folding process and are driven by the short-range interactions, which then coalesce to yield the tertiary structure of a native state.[6] In the diffusion-collision model, the microdomains, which are short enough to search for the most stable conformation,[7] diffuse together and collide with each other, thus forming the native conformation. In the hydrophobic collapse model, the long-range hydrophobic interactions cause the formation of the globular structure.[8] The secondary structures are formed as a consequence of the collapse.[9]

Many theoretical methods for the prediction of the 3-D structure of proteins have been developed.[10–23] Some of the methods work for the selected proteins; however, a reliable algorithm for predicting the native 3-D structure of proteins from their sequence alone has yet to be developed. Various approximations have been used to simplify the molecular system. The most common approximation is that an

Contract grant sponsor: Ministry of Science and Technology of Slovenia; Contract grant sponsor: Commission of the European Communities (PECO).
*Correspondence to: Franc Avbelj, National Institute of Chemistry, Hajdrihova 19, SI 1115 Ljubljana, Slovenia. E-mail: francl@kihp8.ki.si
Received 31 July 1997; Accepted 10 October 1997

amino acid is represented by one or more spheres.[10] The influence of solvent effects is included implicitly in the potentials of mean force obtained from the experimental structures of proteins.[24–26] The location of an amino acid has also been limited to points on various lattices.[9,14]

Although a considerable effort has been directed toward understanding the physical background of the folding process and the nature of pathways, the issues remain controversial. It is not clear which forces determine the secondary and the tertiary structures of proteins, nor even at what stage a particular type of interaction dominates the protein folding process.[27] The elucidation of the correct physical model for the energetics of secondary structures is a very demanding problem; therefore some authors begin their studies with predefined secondary structures.[17,18] The hydrophobic effect,[28,29] side-chain conformational entropy,[30,31] steric effects,[32–35] and main-chain electrostatics[36,37] have all been implicated as the physical factors that dominate the energetics of amino acids in the particular secondary structure.

The experimental data have shown that the secondary structures are formed early in the protein folding process.[38–41] It has been suggested that the secondary structures condense together in a hierarchic order. In this process, which is called the hierarchic condensation,[42] the neighbor secondary structures interact earlier than those parts of the structure distant in the sequence, creating the hierarchic organization of proteins.[42–44] Consequently, amino acids close in sequence are also close in space.[45]

Srinivasan and Rose[19] applied the hierarchic condensation to predict the native 3-D structure of proteins. The hierarchic condensation has been performed by gradually increasing the range of interacting residues during the simulation procedure. The range of interactions was allowed to grow from 6 residues at the onset to 48, in six amino acid increments. Parts of the protein close in the sequence interact in the folding process and form localized structures, which then associate with other neighbor structures to form larger compact structures. The stability of the secondary structures is determined by the side-chain conformational entropy, which is implicitly included in the algorithm.[30,31] Although a rather simple force field has been used, the secondary and super-secondary structures of five of seven globular proteins are predicted correctly, suggesting that the hierarchic condensation may indeed be one of the general mechanisms in protein folding.

Recently, a surprisingly simple, alternative view on the protein structure and the folding process has emerged from the electrostatic screening model of the amino acid preferences for the different main-chain conformations. According to this model, the short-range main-chain electrostatic interactions are crucial physical factors that determine the secondary structure in proteins.[36,37] The long-range electrostatic and hydrophobic interactions determine the super-secondary structure and the larger compact structures.[37] The electrostatic screening model of amino acid preferences has been tested on its ability to predict the secondary structure of proteins and peptides correctly.[37] The model was implemented in the Lifson-Roig theory to obtain the helix and strand free energy profiles. The α-helices and β-strands are predicted from the profiles using simple rules. The three-state accuracy (Q_{total}) of the method was found to be ≈70% for 130 proteins. If the hydrophobic effect and the side-chain conformational entropy terms are included in the model, the Q_{total} improves by only one point, to ≈71%. Similar accuracy has been achieved by the best current secondary structure prediction method based on the neural networks. The highest Q_{total} obtained by neural networks is ≈72%.[46,47] The disadvantage of the neural network algorithms is that they do not provide a physical insight into the forces that determine the protein secondary structure.

The simplified version of the electrostatic screening model has been tested on predicting the 3-D structure of small peptides with 11–14 amino acids, which were considered to be the nucleation sites in the protein folding process.[20] The torsion space Monte Carlo procedure was performed to minimize the free energy of the system. The root mean square (RMS) deviations of C_α atoms between the calculated and the experimental structures for several peptides were found to be as low as 1.0 Å. The simplified version of the electrostatic screening model has also been used to predict the 3-D structure of larger fragments by minimizing the free energy with the genetic algorithm.[21]

In this work, an improved version of the electrostatic screening model is applied to predict the 3-D structure of large peptides and proteins. One of the main goals is the correct prediction of the local 3-D structure of proteins, which includes the secondary and super-secondary structures. The hierarchic condensation is used in the torsion space Monte Carlo minimization procedure to find the local and the native 3-D structure of the proteins by minimizing their free energy. In the first phase of the minimization procedure only the short-range interactions are activated. Short-range interactions are interactions between amino acids less than four residues apart in the sequence. Most α-helices and β-strands are formed during the initial phase. The long-range interactions are gradually activated in later phases, which is causing the hierarchic condensation of α-helices and β-strands into super-secondary and the larger compact structures. Long-range interactions are interactions between the amino acids distant in the sequence. The Monte Carlo procedure is applied to predict the secondary, super-secondary,

TABLE I. Proteins Used in the Predictions*

Code	Name	PDB	No. res.	No. sim.	Range	Class
ENH[76]	Engrailed homeodomain	1enh	54	47	7–53	α
GTO[72]	rop cole1 repressor of primer	$1gto_A$	62	62	1–62	α
ICB[73]	Ca-binding protein	3icb	75	75	1–75	α
GFC[77]	C-terminal domain of grb2	1gfc	59	59	1–59	β
TEN[78]	Tenascin	1ten	90	85	807–891	β
HOE[79]	α-Amylase inhibitor	1hoe	74	67	5–71	β
PGA[80]	Protein G	1pga	56	56	1–56	α + β
UBQ[67]	Ubiquitin	1ubq	76	72	1–72	α + β
BRN[81]	Barnase	$1brn_L$	110	103	5–107	α + β
T30[51]	Domain 1 of protein G3	1fgp	66	66	1–66	α + β
T8[50]	Designed α-helical peptide	1coi	28	28	1–28	Peptide
PPT[70,71]	Avian pancreatic polypeptide	1ppt,1bba	36	36	1–36	Peptide

*The PDB entry codes[62] or the CASP2 target names[48] are shown. The disordered amino acids are not used in the simulations; therefore the number of amino acids in the simulations (No. sim.) is different from those in the PDB coordinate file (No. res). The actual range of amino acids used in the simulations (Range) and the SCOP protein class (Class) are also shown. The proteins are classified using the SCOP database.[52]

and native 3-D structures of 12 different proteins with 28–110 amino acids from their sequence alone. The 3-D structures of the majority of local secondary and super-secondary structures are predicted accurately.

The method has been proved accurate in predicting the local secondary and super-secondary structure in the blind ab initio 3-D prediction experiment (CASP2: Critical Assessment of Methods for Structure Prediction).[48] The experimental structures of proteins were unknown at the time the predictions were performed. Such blind predictions allow an objective assessment of the method and comparison with other contemporary algorithms. In the automatic assessment of the CASP2 results, the predictions were ranked according to the accuracy achieved. Two proteins in the selected set were the targets in the CASP2 experiment. The predictions of the local structure obtained by the new Monte Carlo method are classified on the top of the list in the automatic evaluation of 3-D predictions.[49]

MATERIALS AND METHODS

Criteria for the Selection of Proteins Used in Simulations

The method for predicting 3-D structures of proteins is tested on twelve proteins (Table I). The following criteria have been used in the selection:

1. The proteins contain less than 110 amino acids.
2. The folds of proteins are different.
3. The experimental structures for the proteins are of high quality. Proteins T8[50] and T30[51] were targets in the recent CASP2 experiment.[48] The experimental 3-D structures were unknown for these two proteins at the time when the predictions were performed.
4. The proteins are globular and have only one domain.
5. The proteins do not have *cis* peptide bonds.
6. The number of disulphide bridges is small (≤2).

In the selected set of proteins there are three proteins with the α structure, three proteins with the β structure, four proteins with the α + β structure, and two peptides. The folds of the selected proteins differ from each other. The proteins are classified by the SCOP database.[52] The disordered regions of the proteins are cut out to make the systems as small as possible. The starting and ending amino acids are replaced with Ace and N-M, respectively. The number of amino acids and the actual range of amino acids used in simulations are presented in Table I.

The Electrostatic Screening Model and Free Energy Profiles

The electrostatic screening model, which rationalizes the preferences of amino acids for the different main-chain conformations in terms of the main-chain electrostatics, is explained in detail elsewhere.[36,37] Here, we describe briefly the main points that follow from the model and are important for the understanding of the Monte Carlo minimization procedure used in this work.

In the electrostatic screening model of amino acid preferences, the stability of a main-chain conformational state of an amino acid in a protein depends primarily on the strengths of local and short-range nonlocal main-chain electrostatic interactions. The strength of local and nonlocal 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 nonlocal main-chain electrostatic interactions are due predominantly to the main-chain hydrogen bonding. Note the differences between local–nonlocal and short-range–long-range interactions (see above).

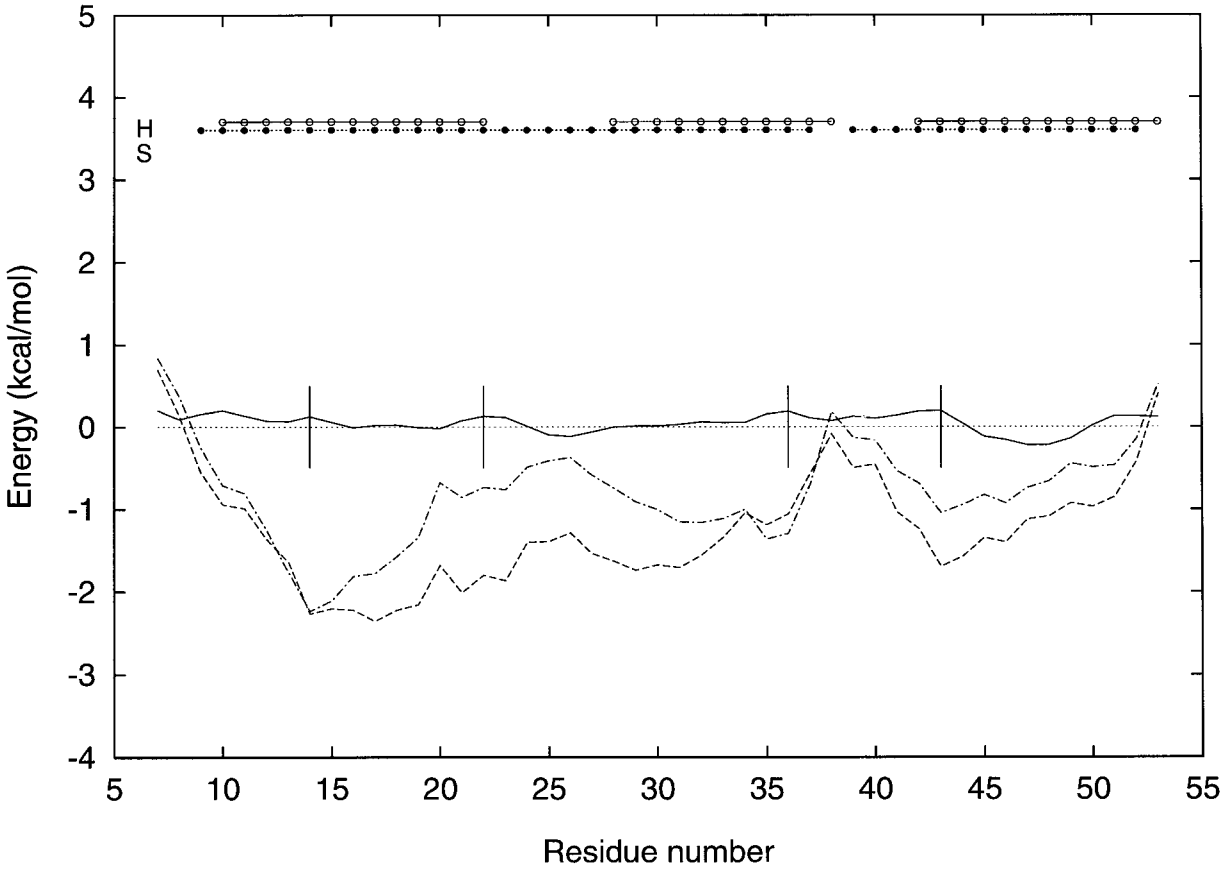

Fig. 1. The helix and strand free energy profiles of ENH. For Figures 1–12, the strand free energy profile is plotted with the solid line, and the helix free energy profiles obtained by Model I[37] and Model II[37] are plotted with the dashed and dash-dot-dash lines, respectively. Assignment of the α-helix and β-strand amino acids calculated by the modified Kabsch and Sander algorithm[65] 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 black circles and black squares, respectively. The vertical lines divide the sequence into folding units.

The helix and strand free energy profiles are the main-chain free energies of amino acids as a function of sequence.[37] They are calculated by incorporating the electrostatic screening model[36] in the Lifson-Roig transition theory.[53] These profiles can be used to predict α-helices, β-strands, and coil structures in proteins.[37] In addition, the helix and strand free energy profiles can also be used to predict turns. According to the electrostatic screening model, some regions of the proteins are more flexible than other regions, because both the free energy cost of escaping the β-conformation and the strength of short-range hydrogen bonds are small.[37] At these places, the chain is likely to fold back and form a turn or a loop. Turns are passive folding elements: the chain folds to obtain the compact structure.[54,55] The positions of turns can be predicted by finding the local maxima along the strand free energy profiles. These flexible regions divide the protein amino acid sequences into folding units. Formation of the folding units containing less than four amino acids is prevented by selecting only those maxima that are more than four amino acids apart in the sequence. If the local maxima in the strand free energy profile are too close to each other, the maximum with more positive strand free energy is chosen as the boundary between folding units. Figures 1–12 show the folding units for 12 proteins used in this work to perform the hierarchic condensation in the Monte Carlo procedure. The folding units often represent α-helices and β-strands. Two amino acids are excluded from each folding unit at both ends because it is quite common that three amino acids constitute a turn in proteins.

Form of the Free Energy Function

The free energy (ΔG) of a protein conformation relative to the standard restricted state[26] is a sum of the following contributions (Equation 1)[20]:

$$\Delta G = \Delta G_{\text{local}} + \Delta G_{\text{nonlocal}} + \Delta G_{\text{hydrophobic}} + \Delta G_{\text{desolvation}} + \Delta G_{\text{other}}. \quad (1)$$

ΔG_{local} and $\Delta G_{\text{nonlocal}}$ are contributions from local and the nonlocal main-chain electrostatic interactions, respectively. $\Delta G_{\text{hydrophobic}}$ and $\Delta G_{\text{desolvation}}$ are the contributions from the burial of the hydrophobic and hydrophilic accessible surface areas, respectively.

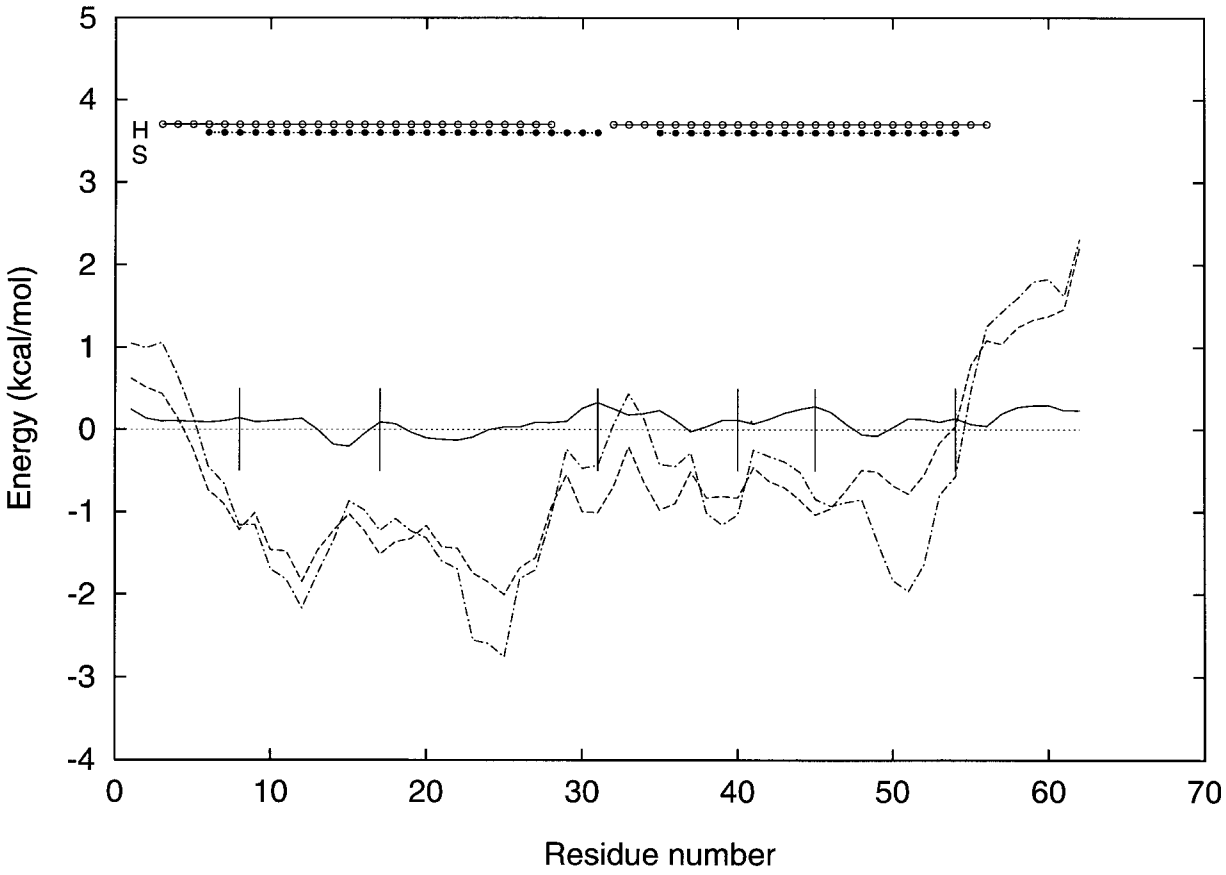

Fig. 2. The helix and strand free energy profiles of GTO.

The contributions of other interactions (ΔG_{other}), like the side-chain–side-chain and side-chain–main-chain electrostatic interactions, are ignored in the free energy function used in this work ($\Delta G_{\text{other}} = 0$).

The free energy contributions from local (ΔG_{local}) and nonlocal main-chain electrostatic interactions ($\Delta G_{\text{nonlocal}}$) are defined by Equations 2 and 3, respectively[36]

$$\Delta G_{\text{local}} = \sum_i \gamma^r_{\text{local}} E^i_{\text{local}} \tag{2}$$

$$\Delta G_{\text{nonlocal}} = \sum_i \gamma^r_{\text{nonlocal}} E^i_{\text{nonlocal}} \tag{3}$$

where the sums run over all amino acids i in the sequence. The E^i_{local} and E^i_{nonlocal} are the local and the nonlocal electrostatic energies, respectively. The coefficients $\gamma^r_{\text{nonlocal}}$ and γ^r_{local} are the screening coefficients dependent on amino acid type r of residue i. The coefficients $\gamma^r_{\text{nonlocal}}$ and γ^r_{local} represent the attenuation of the electrostatic energies E_{nonlocal} and E_{local}, respectively, due to the electrostatic screening (Table II). The screening coefficients are inversely related to the microscopic screening function described by Warshel and Russell.[56] The effective dielectric constant for the local and nonlocal electrostatic interactions is provided by the screening coefficients and depends on the amino acids involved. Electrostatic energies E_{local} and E_{nonlocal} are calculated using Coulomb's law with a dielectric constant of 1. Point atomic charges for the main-chain atoms N, H_N, C, and O are -0.28, $+0.28$, $+0.38$, and -0.38 electrons, respectively.[36] 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 hydrophobic contribution is defined as:

$$\Delta G_{\text{hydrophobic}} = \sum_i \sum_j \sigma_h \Delta A_{ij} \tag{4}$$

where σ_h is the free energy coefficient of burial of one Å^2 of the hydrophobic surface area (Table II).[57] ΔA_{ij} is the difference between the nonpolar accessible surface area of amino acid i in the presence of the first two neighboring residues on each side and the accessible surface area of this amino acid in the presence of all residues in the molecule. The sums run over all amino acids i in the sequence. The accessible surface areas are calculated using the Lee and Richards algorithm[58] with Chothia radii.[59] Hydrophobic atoms are defined as the side-chain carbon

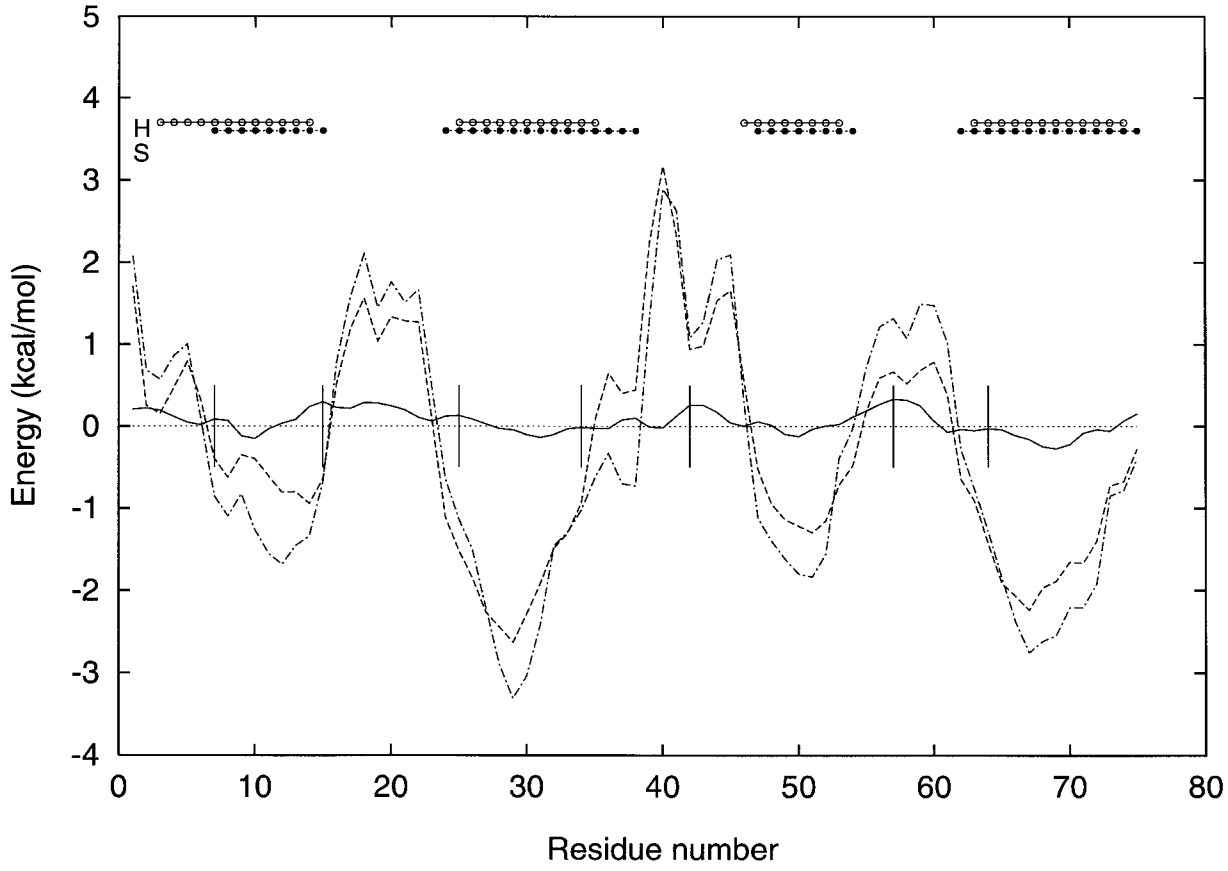

Fig. 3. The helix and strand free energy profiles of ICB.

and sulphur atoms of amino acids Val, Ile, Leu, Phe, Met, Cys, Trp, and Tyr.

The desolvation contribution of the polar atoms is defined as:

$$\Delta G_{\text{desolvation}} = \sum_i \sum_j \sigma_j \Delta A_{ij}^3 \qquad (5)$$

where σ_j is the free energy coefficient of the burial of the hydrophilic surface area of the group of atoms j (Table II). ΔA_{ij} is the difference between the accessible surface area of the group of polar atoms j in the random coil[20] and the accessible surface area of the group of atoms j in protein. The sums are over all amino acids i in the sequence.

Fitting the Potentials of Mean Force

The potential of mean force $[\Gamma(R)]$ is the free energy of a system as a function of the reaction coordinate R. It is related to the population of states along the reaction coordinate by the following equation[60,61]:

$$\Gamma(R) = -kT \ln [\, p(R)] + kT \ln [n(R)] + C. \qquad (6)$$

R represents a structural alternation in the system. $p(R)$ is the probability of finding the system in any of the states with a particular value of R, and $n(R)$ is the volume element or the normalization function. C is an undefined constant, T is temperature, and k is Boltzmann's constant. Potentials of mean force obtained from the experimental protein structures have been used to evaluate the free energy contributions of individual interactions in a protein environment.[20,26,36] It has been assumed that the observed distributions in the experimental protein structures are subject to the Boltzmann relationship between the population of the state of a system and the free energy of that state.

The electrostatic screening ($\gamma^r_{\text{nonlocal}}$, γ^r_{local}), hydrophobic (σ_h), and desolvation (σ_j) coefficients in the free energy function (see above) are calculated by fitting the potentials of mean force obtained from the experimental protein structures, as described previously.[20,26,36] The 40 screening coefficients $\gamma^r_{\text{nonlocal}}$ and γ^r_{local} are shown in Table II. The four hydrophobic and desolvation coefficients are shown in Table III. A total of 44 fitted coefficients is used in the free energy function (Equations 1–5).

The potentials of mean force depend on the normalization function used ($n(R)$; see Equation 6).[36] The normalization function is the probability distribution along the reaction coordinate. The probability distribution is calculated from the large number of

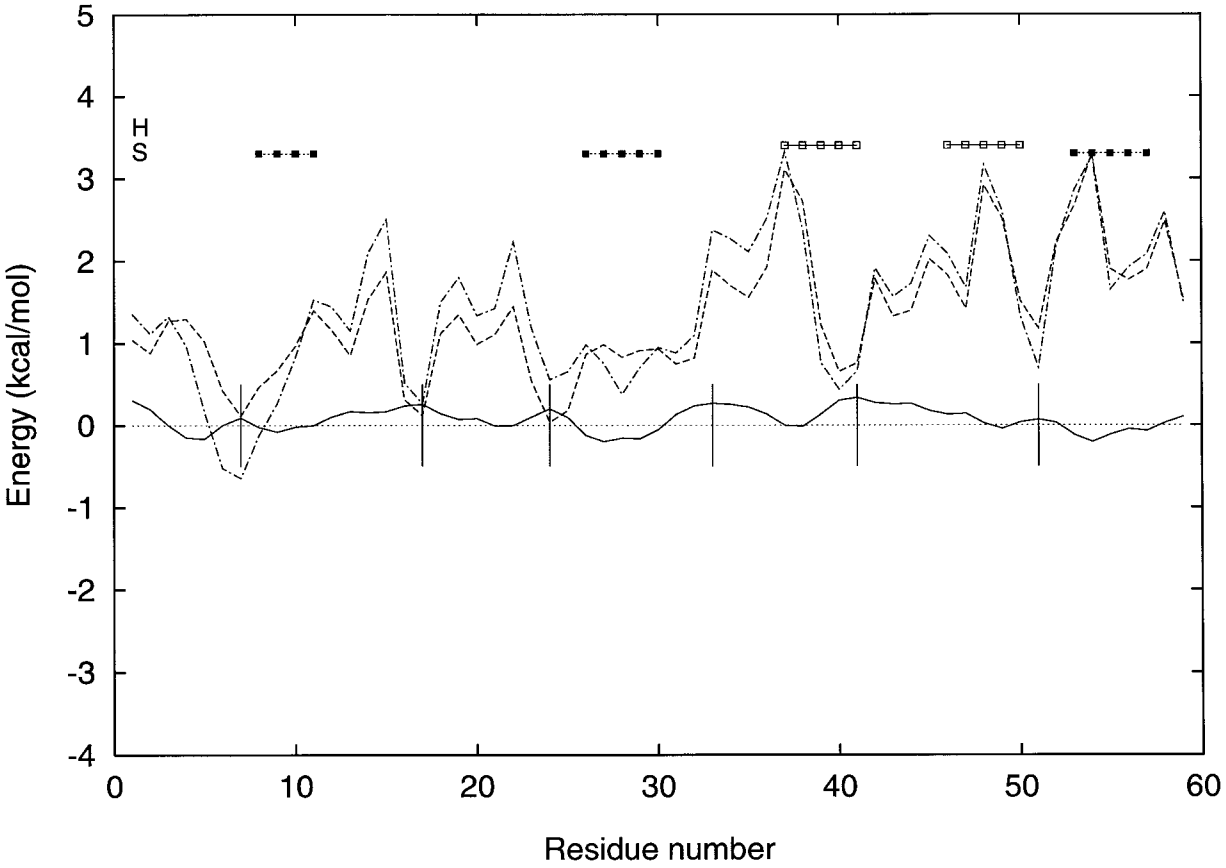

Fig. 4. The helix and strand free energy profiles of GFC.

compact random structures. These compact structures are generated in the procedure in which all interactions between atoms, with the exception of the steric overlap, are set to zero. As shown earlier,[36] the conformations in the random structures at both ends of the interval of the reaction coordinate E_{local} are less probable than those in the central region. These regions correspond to α-helices and β-strands. Such repeating structures are unlikely to occur in the compact random coil structures because of the unfavorable main-chain conformational entropy cost. If the normalization functions are excluded from the fitting procedure, the free energy contribution of the main-chain conformational entropy is implicitly included in the γ_{nonlocal}.

The potentials of mean force are obtained from the set of 443 high-resolution (resolution <2.0 Å and R factor <20%) x-ray crystal structures of proteins from the Protein Data Bank.[62] The proteins can be accessed from the Protein Data Bank under the following codes: 1aaj, 1aal, 1aap, 1aba, 1abk, 1acb, 1acf, 1aco, 1ads, 1afg, 1ahc, 1ake, 1alk, 1amp, 1ank, 1aoz, 1apm, 1arb, 1arp, 1ast, 1asz, 1bam, 1bbh, 1bbp, 1bgh, 1bit, 1bmd, 1bns, 1brs, 1btl, 1byb, 1caa, 1cbs, 1ccr, 1cdc, 1cdg, 1cdm, 1cel, 1cew, 1cge, 1cgo, 1cgt, 1chm, 1chn, 1cho, 1cka, 1clc, 1cmb, 1cnr, 1cot, 1cpc, 1cpm, 1cpn, 1crl, 1csh, 1csn, 1ctf, 1cth, 1cus, 1cyo, 1daa, 1ddt, 1dfn, 1drf, 1dsb, 1dts, 1ebh, 1edt, 1emy, 1enx, 1ept, 1erl, 1esl, 1ezm, 1fas, 1fba, 1fdd, 1fdn, 1fel, 1fgv, 1fia, 1fkb, 1fkd, 1flp, 1flr, 1flv, 1fna, 1fnc, 1frd, 1frp, 1frr, 1fus, 1fut, 1fxd, 1gbs, 1gca, 1gcs, 1gia, 1gky, 1glq, 1glt, 1gma, 1gof, 1gox, 1gp1, 1gpb, 1gpr, 1hag, 1hbg, 1hcb, 1hfc, 1hhl, 1hil, 1hle, 1hml, 1hmt, 1hne, 1hpg, 1hpi, 1hpm, 1hrc, 1hsl, 1htr, 1huw, 1hvc, 1hvi, 1hvr, 1hyl, 1hyp, 1ilb, 1iag, 1icm, 1ida, 1ids, 1igd, 1ilk, 1isa, 1isu, 1knb, 1knt, 1lcf, 1lcp, 1lct, 1lec, 1len, 1lga, 1lib, 1lki, 1lld, 1lmb, 1lmn, 1lra, 1lst, 1lte, 1lts, 1lz1, 1lz3, 1mba, 1mdc, 1mee, 1mfa, 1mfe, 1mjc, 1mol, 1mpp, 1mrj, 1nar, 1nba, 1nci, 1nco, 1ndc, 1nfp, 1nhk, 1noa, 1npc, 1npk, 1nsc, 1ntn, 1ofv, 1olb, 1onc, 1opa, 1opg, 1osa, 1ova, 1oyb, 1pal, 1paz, 1pbe, 1pbp, 1pca, 1pda, 1pga, 1pgb, 1pgs, 1pgx, 1php, 1pii, 1pk4, 1plc, 1pmy, 1pne, 1poa, 1poc, 1poh, 1ppa, 1ppb, 1ppe, 1ppf, 1ppn, 1ppo, 1prn, 1pso, 1ptf, 1ptq, 1ptx, 1pva, 1r69, 1ras, 1rcf, 1rcm, 1rdg, 1rds, 1rec, 1ris, 1rnh, 1rnv, 1rop, 1rro, 1rsy, 1rtp, 1s01, 1sac, 1sar, 1sat, 1sbp, 1scn, 1scs, 1sct, 1sel, 1sem, 1sgt, 1sha, 1shb, 1shf, 1shg, 1slt, 1smr, 1snc, 1spa, 1sri, 1st3, 1sxa, 1tad, 1tag, 1tca, 1tew, 1tgs, 1tgx, 1thg, 1thm, 1thv, 1thw, 1tib, 1tml, 1ton, 1top, 1tpf, 1tph, 1tpo, 1trb, 1trk, 1tsp, 1tta, 1tys, 1ukz, 1vfa, 1wfa, 1wgt, 1wht, 1wtl, 1xib, 1xnb, 1xso, 1xya, 1xyn, 1ycc, 1yma, 256b, 2act, 2ak3, 2alp, 2apr, 2ayh, 2aza, 2bbk, 2cba, 2ccy, 2cdv, 2cga, 2ci2, 2cmd,

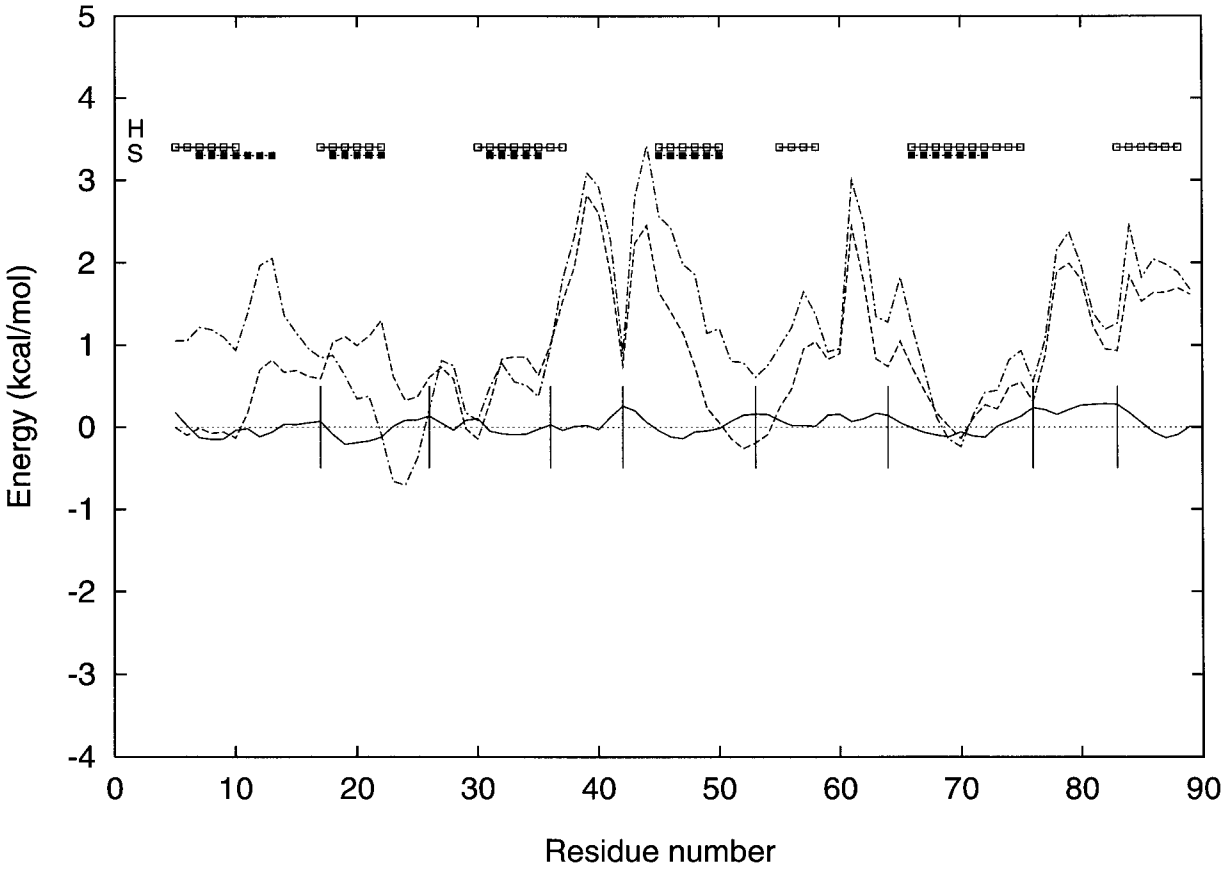

Fig. 5. The helix and strand free energy profiles of TEN.

2cmm, 2cpl, 2cpp, 2csc, 2cst, 2ctc, 2cut, 2cwg, 2cy3, 2cyp, 2cyr, 2dnj, 2dri, 2ebn, 2end, 2er7, 2fb4, 2fbj, 2fcr, 2fgf, 2fx2, 2gbp, 2gct, 2gst, 2had, 2hbe, 2hmq, 2hpd, 2hpe, 2hpr, 2hts, 2ihl, 2imm, 2imn, 2kau, 2lhb, 2lig, 2ltn, 2lyz, 2lzt, 2mcg, 2mcm, 2mlt, 2mnr, 2msb, 2mye, 2nac, 2nad, 2ohx, 2ovo, 2pgd, 2pia, 2pkc, 2plt, 2por, 2psg, 2ptc, 2rhe, 2rn2, 2rsp, 2scp, 2sec, 2sga, 2sic, 2sil, 2spc, 2st1, 2tgi, 2tir, 2trx, 2tsc, 2utg, 2wrp, 2zta, 351c, 3app, 3bcl, 3blm, 3c2c, 3chy, 3cla, 3cms, 3cox, 3dni, 3ebx, 3est, 3grs, 3hhb, 3il8, 3lzm, 3mcg, 3mds, 3ovo, 3pga, 3pte, 3ptn, 3rp2, 3rub, 3sdh, 3sgb, 3tgl, 4azu, 4blm, 4bp2, 4dfr, 4enl, 4fgf, 4fxn, 4ins, 4mt2, 4pep, 4pti, 4q21, 4sdh, 5cha, 5cna, 5cpv, 5cyt, 5p21, 5pal, 5rub, 5rxn, 5tim, 6cpa, 6rlx, 6rxn, 7aat, 7abp, 7fab, 7pcy, 7rsa, 7rxn, 8dfr, 8fab, 8pti, 8rsa, 8tln, 9ins, 9pap, 9rnt, 9wga.

Representation of the Protein Molecules

The protein molecule is represented by hard spheres with fixed bonded distances and angles. All heavy and polar hydrogen atoms are used explicitly. Solvent is implicitly included through the screening and desolvation coefficients (see the free energy function described above). Disulphide bonds and prosthetic groups are ignored.

The geometry of amino acids is generated using the distances and angles from the Discover residue library.[63] The distances and angles between bonded atoms are maintained constant throughout the simulation. Only the torsion angles are allowed to vary during the simulations. The torsion angles are selected from the library of torsion angles of 128,540 amino acids in 443 high-resolution protein structures, providing a biased sampling of likely angles. The ω peptide bond torsion angles are fixed at 180°.

The hard sphere repulsion is enforced by discarding geometries with steric clashes. The steric clash occurs if a pair of nonbond atoms is closer than 0.5 Å less than the sum of their van der Waals radii.[59] The clash distance for main-chain N and O atoms is 2.5 Å. Pairs of atoms related by a torsion angle (one to four pairs) are not checked for clashes because the torsion angles ϕ and ψ are always selected as pairs from the library of torsion angles, thus avoiding the forbidden zones in the ϕ-ψ plots.

The Simulation Procedure

The starting geometries of a protein for simulations are obtained by randomly selecting the torsion angles from the library of torsion angles. These different conformations are then relaxed by small variations of torsion angles to remove the steric overlaps between atoms. The simulation procedure is divided into several phases. The number of phases

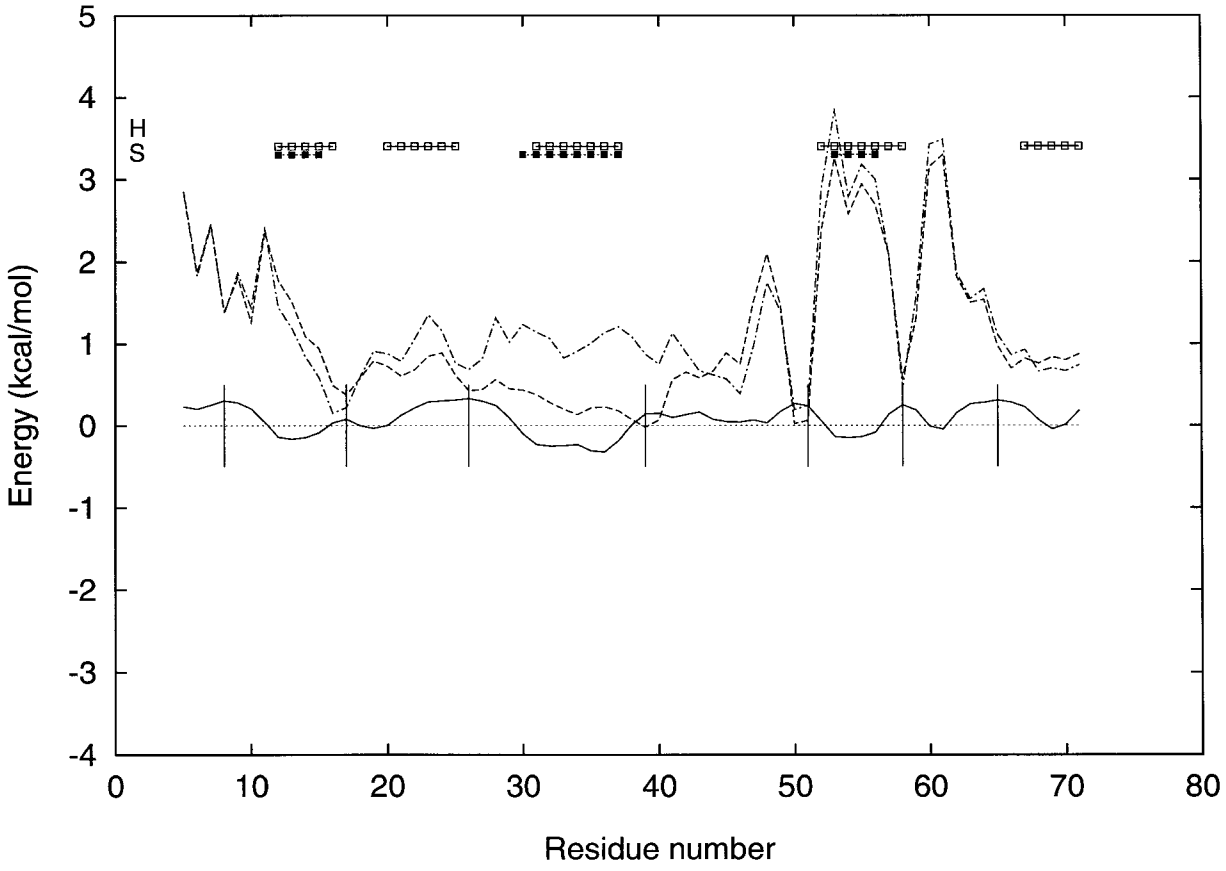

Fig. 6. The helix and strand free energy profiles of HOE.

depends on the number of folding units in the protein. To improve the sampling of the conformational space, a large number of independent Monte Carlo minimizations (between 200 and 400) with 2,000 steps is performed in each phase.

In the first phase of the minimization procedure, the free energies of residues are calculated from interactions of amino acids closer than four residues apart (i to $i + 4$). The 15 conformations with the lowest free energy obtained in this phase are used as the starting geometries for the second phase. In the second phase of the simulation procedure, the free energy of an amino acid in the folding unit i is calculated from the interactions with the amino acids in the folding units i, $i - 1$ and $i + 1$. In the next phases of minimization the procedure described above is repeated. The range of interacting residues is increased by one additional folding unit on both sides of an amino acid in each phase, until all interactions in the protein are included in the free energy function.

The Monte Carlo minimization of the free energy is performed by varying the torsion angles of the protein using various moves. To improve the convergence of the method the torsion angles of more amino acids variate in one step. The type of move and the amino acids involved are selected randomly from a set of moves available for the particular phase in simulation, and the resulting conformation is tested for the steric clashes. The conformations with the steric clashes are discarded.

The free energy of a conformation is calculated using Equations 1–5. If the free energy decreases, the new conformation is accepted. If the free energy increases, the Metropolis criterion[64] is used to decide whether to accept or reject the move. If 50 successive moves are rejected, the conformation is reset to that of the lowest free energy in the course of simulation. The temperature is 300 K. Overall 35% of the moves are accepted during the simulations. The searching for the clash-free conformations is the source of the largest computational cost.

Two minimization procedures are used in predicting the 3-D structure of the protein T30. The difference between these two procedures lies in the treatment of disulphide bridges. In the first procedure, which is also used for the other 11 proteins, the disulphide bridges are ignored. The distances between the sulphur atoms in the two pairs of disulphide bridges are minimized in the second procedure. The results calculated with the second procedure ($T30_{CASP2}$) were used in the CASP2 experiment.[48]

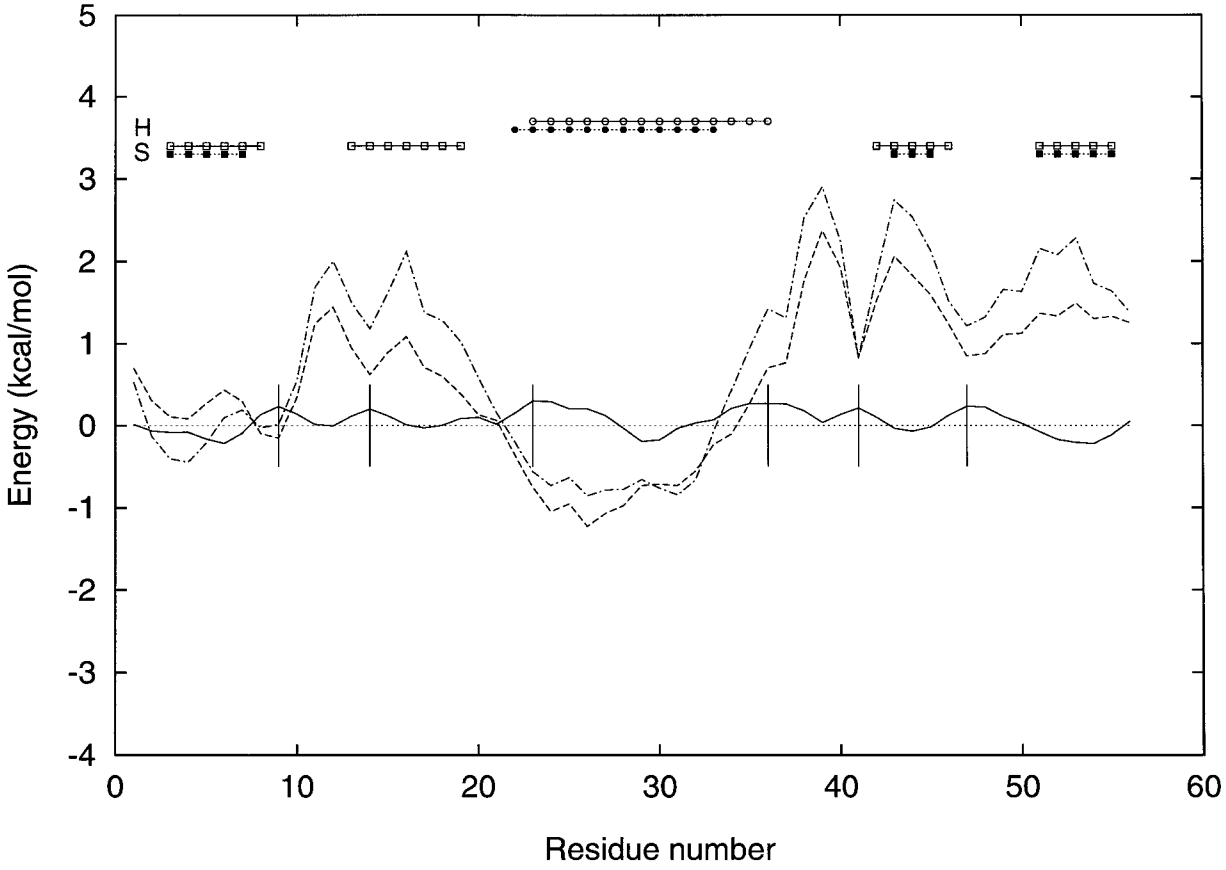

Fig. 7. The helix and strand free energy profiles of PGA.

Variations of Torsion Angles

Several torsion angles vary simultaneously in the majority of moves. The torsion angles are always selected from the library of torsion angles obtained from 443 high-resolution experimental protein structures. The features characteristic for proteins, like hairpin twist, can be formed with such moves. The moves are selected with different probabilities, which depend on the simulation phase.

The available moves in the first phase of generating the protein secondary structures are:

1. The formation of an α_R-helix. Three consecutive amino acids are selected randomly from the sequence. The torsion angles ϕ and ψ of these amino acids are selected randomly from the α_R region of the Ramachandran plot. These torsion angles are then fine tuned to obtain the lowest nonlocal electrostatic energy of the system.
2. The formation of the β-strand. The three consecutive amino acids are selected randomly from the sequence. The torsion angles ϕ and ψ of these amino acids are selected randomly from the β region of the Ramachandran plot. These torsion angles are then fine tuned to obtain the lowest local electrostatic energy of the system.
3. Variation of the side-chain conformation. One amino acid is randomly selected from the sequence. One of the randomly chosen side-chain torsion angle is then altered using the library of torsion angles.

For generating the super-secondary structures and compact proteins, the additional moves are used:

4. Changing the main-chain conformation. One amino acid is randomly selected from the sequence. The torsion angles ϕ and ψ of this amino acid are then changed using the library of torsion angles.
5. The formation of turn. The three consecutive amino acids are randomly selected from the sequence. The torsion angles ϕ and ψ of these amino acids are selected randomly from the library of torsion angles to obtain the strongest hydrogen bond in the turn.
6. The formation of hairpin. A group of consecutive amino acids is selected randomly from the sequence. First, the turn is formed in the middle of this group of amino acids as described above. Then the torsion angles of the two additional

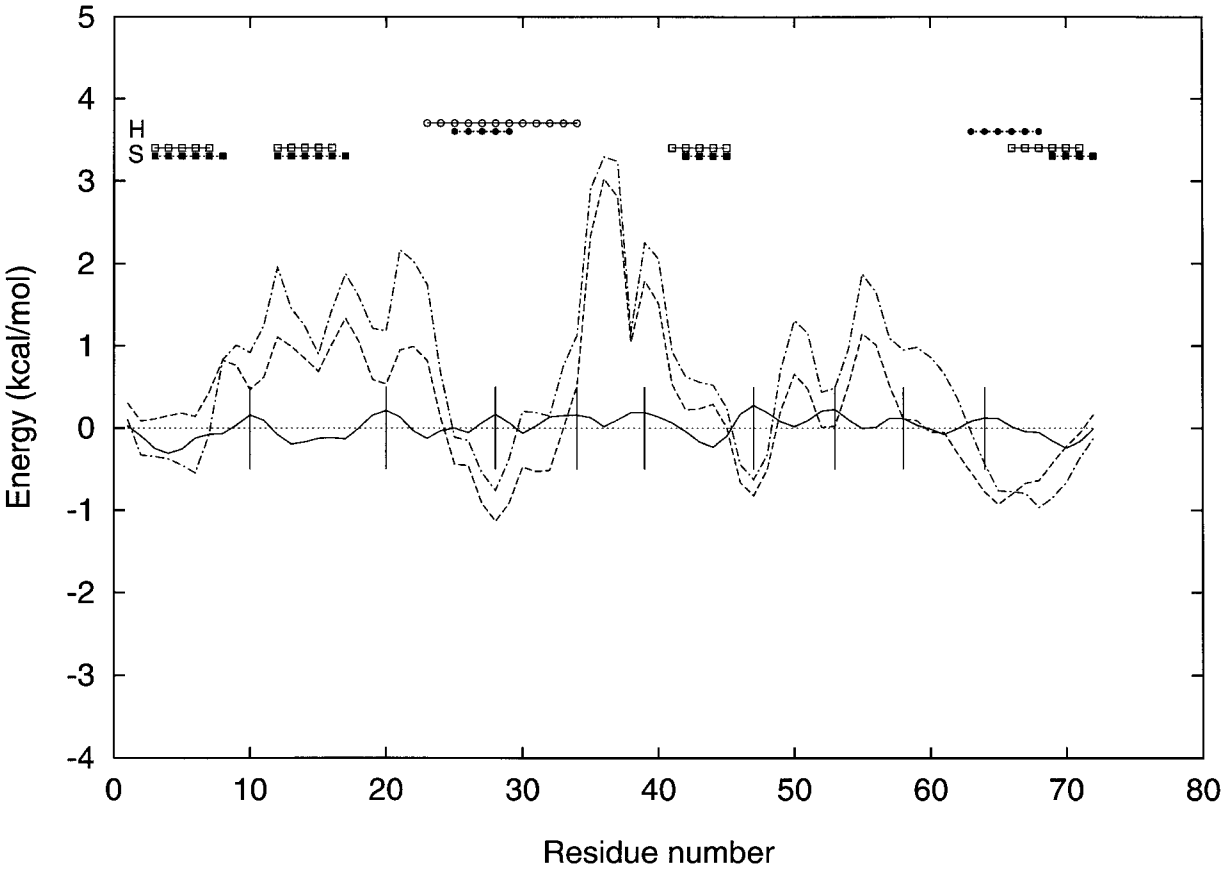

Fig. 8. The helix and strand free energy profiles of UBQ.

consecutive amino acids on both sides are selected from the β region of the Ramachandran plot. These torsion angles are then fine tuned to obtain the lowest main-chain electrostatic energy of the system.

Present Improvements of the Method

The method described in this work is an improved version of the method, which was used previously for the predictions of the 3-D structures of small peptides.[20] The new method permits 3-D predictions on larger peptides and proteins using the concept of hierarchic condensation. The strength of hydrogen bonds was assumed to be equal for all amino acid pairs with the screening coefficients $\gamma_{nonlocal}$ of 0.38. The screening coefficients for the main-chain hydrogen bonding interactions in this work are residue dependent, because it has been shown that the residue-dependent strengths of hydrogen bonds are the crucial physical factor for correctly predicting the secondary structures in proteins.[37] The present screening coefficients are calculated from a larger set of high-resolution protein structures. Only the side-chain carbon and sulphur of amino acids Val, Ile, Leu, Phe, Met, Cys, Trp, and Tyr are considered as the hydrophobic atoms. The calculation of the hydrophobic-accessible surface areas is improved considerably (see above). The number of coefficients used in the free energy function (Equations 1–5) is kept as small as possible. The main-chain–side-chain and side-chain–side-chain electrostatic interactions are ignored in this version of the free energy function. Only 44 fitted coefficients are used in the free energy function. To improve the convergence of simulations, a much larger number of independent Monte Carlo minimizations are performed, and more efficient variations of the torsion angles are introduced.

RESULTS AND DISCUSSION

Hierarchic Condensation and the Torsion Space Monte Carlo Minimizations

The new method for predicting the 3-D structure of proteins from their sequence only is tested on 12 proteins (Table I). The proteins are selected according to the requirements described in Materials and Methods. To find the lowest free energy conformation of a protein along the protein folding pathways (see Introduction), the free energy function (see Equations 1–5) is minimized with the Monte Carlo method using the concept of hierarchic condensation. The

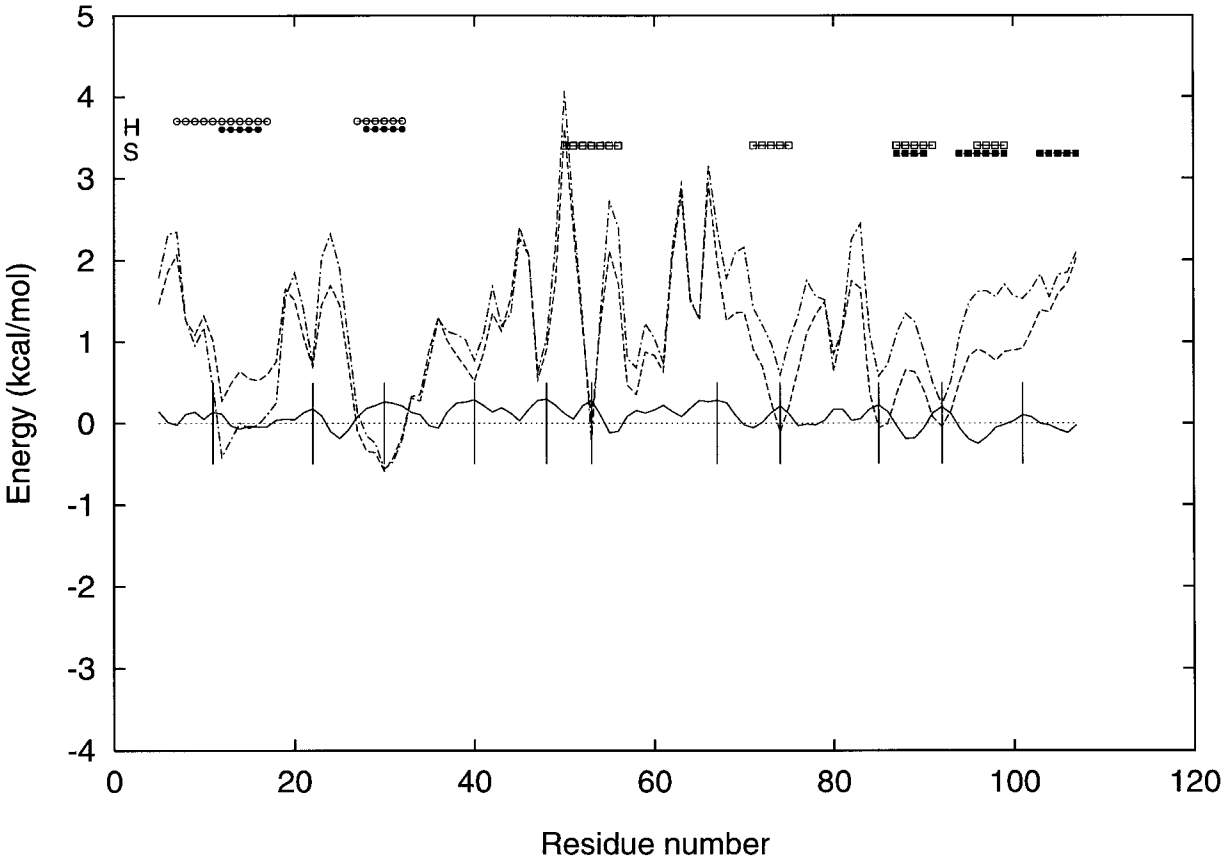

Fig. 9. The helix and strand free energy profiles of BRN.

free energy function is based on the electrostatic screening model of amino acid preferences for the different main-chain conformations.[36,37] In this model, the energetics of secondary structures depends primarily on the balance of strengths between the local and the short-range nonlocal main-chain electrostatic interactions.[36,37] The long-range interactions are important predominantly for the energetics of the super-secondary structures and larger compact structures.[37]

The hierarchic condensation is performed by increasing the range of interactions during the course of simulation. The range of interactions between amino acids is increased by the increments of several amino acids, which are grouped together as the folding units. A folding unit represents the stable and relatively rigid secondary structures separated from each other by the less stable and flexible regions in the sequence. The positions of the flexible regions in proteins are obtained by finding the local maxima along the strand free energy profiles (see Materials and Methods and Figs. 1–12).

The torsion space Monte Carlo minimization procedure starts from the random structures. In the first phase of the minimization, only short-range interactions between amino acids less than four residues apart are activated. As predicted by the electrostatic screening model, the majority of α-helices and β-strands are formed during this initial phase. In each next phase of the simulation, the range of interactions is increased by one additional folding unit on both sides of an amino acid, until all interactions in the protein are included in the free energy function. This procedure causes the hierarchic condensation of α-helices and β-strands into super-secondary structures and larger compact structures. The main-chain conformational state of each amino acid is free to change in any phase of the minimization procedure. Note the difference between the Monte Carlo minimization performed here and the ordinary Monte Carlo simulations, in which energy and not the free energy is used.

The lowest free energy conformation obtained in the last phase of the Monte Carlo minimization is then compared with the experimental structure. The accuracy of predicting secondary, super-secondary, and native 3-D structures is assessed from the RMS deviations between the predicted and experimental conformations.

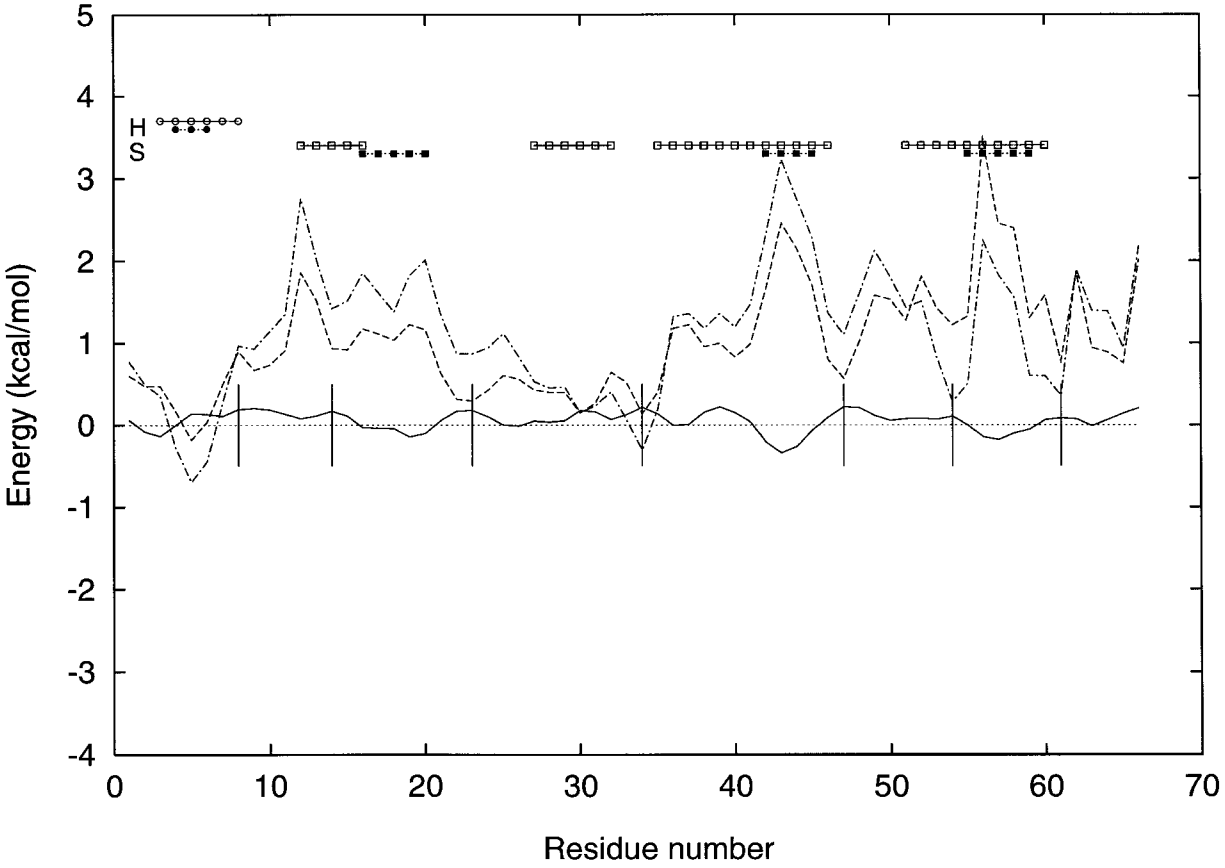

Fig. 10. The helix and strand free energy profiles of T30.

Prediction of Secondary Structures Using the Free Energy Profiles

The secondary structure prediction methods are based mainly on the probabilistic approaches. Recently, an explicit physico-chemical model for predicting the secondary structures has been introduced.[37] The electrostatic screening model of amino acid preferences was implemented in the Lifson-Roig theory to obtain the helix and strand free energy profiles. From these profiles α-helices and β-strands are predicted using simple rules.

Figures 1–12 present the helix and strand free energy profiles for 12 proteins. These figures show a strong correlation between the negative peaks of the helix and strand free energy profiles with the occurrence of α-helices and β-strands in the experimental protein structures, respectively.[37] The helix free energy profiles are calculated with two different models for the energetics of the α-helices (Figs. 1–12). In the first model (Model I[37]), only the short-range main-chain electrostatic interactions are included. The main-chain electrostatics, and the hydrophobic and conformational entropy effects are used in the second model (Model II[37]). Comparison of the helix free energy profiles obtained by these two models demonstrates that the hydrophobic and entropic effects have a small but not negligible influence on the stabilities of some α-helices.[37] The hydrophobic and entropic effects are particularly important for the stability of helix 1 in the native structure of barnase (BRN) (Fig. 9, Table IV).

The three-state accuracy of predicting α-helices, β-strands, and coil (Q_{total}) in 12 proteins from the free energy profiles is 72.6%. Model II is used to describe the energetics of α-helices in the calculations of these free energy profiles. Comparison between the experimental and the predicted positions of α-helices in 12 proteins shows a high degree of agreement (Table IV). Helical residues are predicted with the accuracies of $Q_\alpha = 66.0\%$ and $Q_\alpha^{\text{pred}} = 87.3\%$. The accuracies of predicting strands Q_β and Q_β^{pred} are 67.1% and 81.7%, respectively (see Rost and Sander[46] for the definition of these accuracy measures). The assignment of α-helices in the experimental protein structures is calculated by the algorithm of Kabsch and Sander.[65] The errors in the predictions are located mainly at both ends of the α-helices. The most plausible reason for the errors is the absence of the

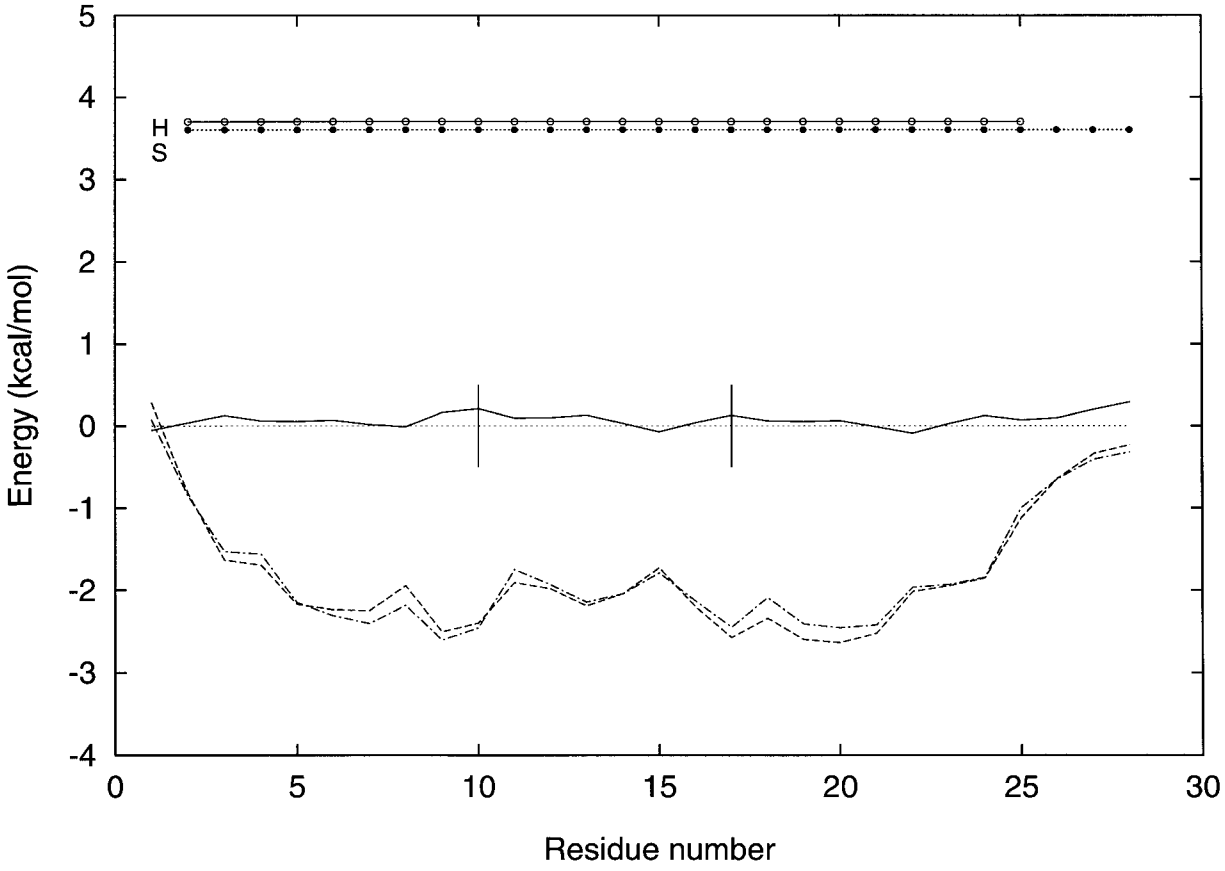

Fig. 11. The helix and strand free energy profiles of T8.

side-chain–main-chain electrostatic interactions from the free energy profiles. These interactions are important for the α-helix capping and ending effects.

Prediction of Secondary Structures Using the Monte Carlo Method

The 3-D structures of the majority of secondary structures are predicted correctly in all 12 proteins (Tables IV–VI). The average RMS deviation of C_α atoms between the experimental and the predicted secondary structures (folding units) is 1.28 Å (from Table V). The 3-D structures of α-helices are predicted more accurately than the 3-D structures of β-strands. The average RMS deviations in the running window with six amino acids (Table VI) are between 0.31 Å and 1.23 Å for the α proteins and between 1.73 Å and 1.98 Å for the β proteins. The RMS deviations depend strongly on the number of amino acids included in the comparison; therefore the method of running windows is used.[66] The RMS deviations between the predicted and the experimental structures are calculated for small fragments in this method, with fixed numbers of amino acids, which are defined by the window running along the protein sequence. The average number of amino acids in a folding unit (secondary structure) is 6.2; therefore the window with six amino acids is used for this comparison. Better accuracy in predicting α-helices in comparison with predicting β-strands is reasonable, because the region in the ϕ and ψ space of the β conformation is larger than the region of the α conformation.

Predictions of the α-helices with the Monte Carlo method are much more accurate than predictions based on the free energy profiles (see above). The accuracies of predicting α-helices by the Monte Carlo method for 12 proteins are Q_α = 85.9% and Q_α^{pred} = 90.9%. This improvement is primarily due to the long-range interactions included in the torsion space Monte Carlo minimization procedure. The long-range interactions are ignored in the free energy profiles as well as in the neural network algorithms. Table IV shows the locations of α-helices in the experimental and predicted conformations, obtained at the end of the last phase of the Monte Carlo minimizations. The errors in predicting α-helices by the Monte Carlo method are located mainly at both ends (see above). These errors are due to the absence

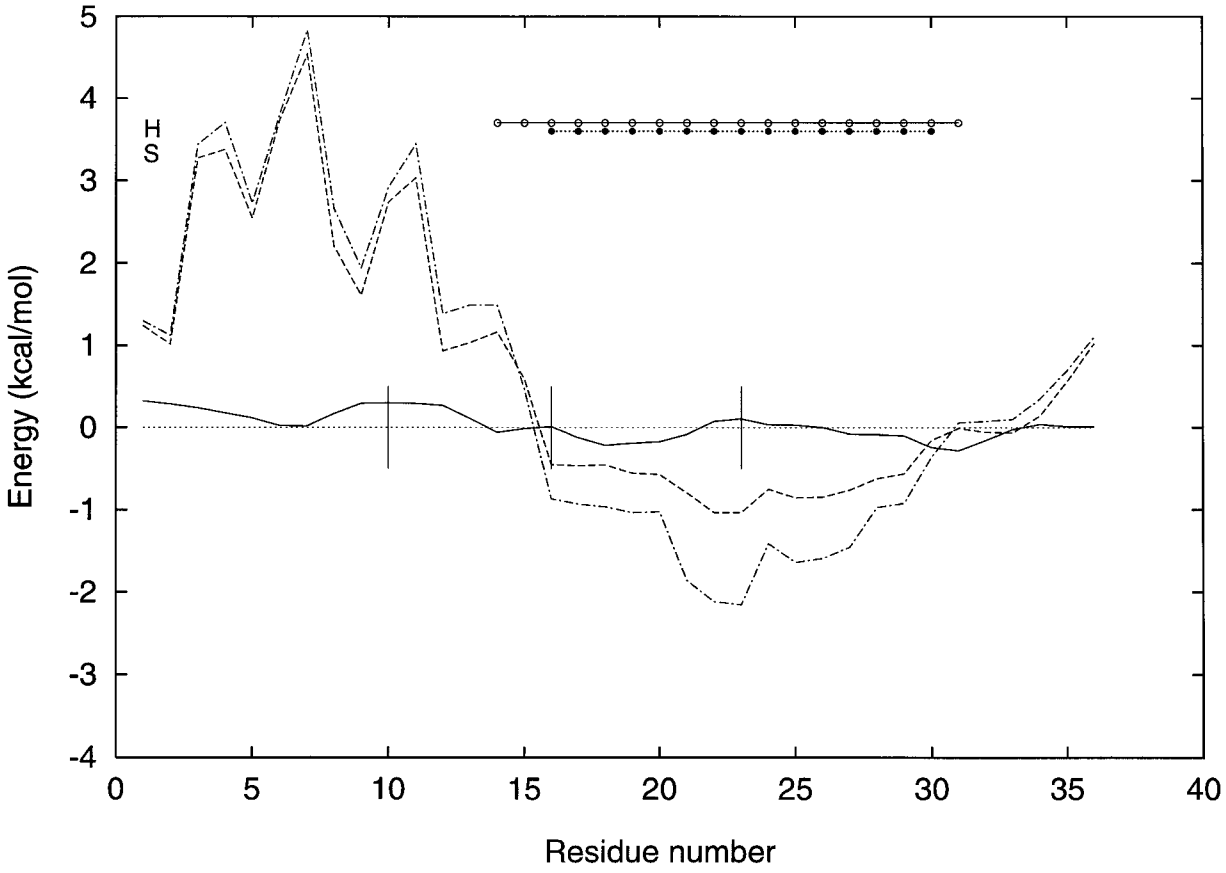

Fig. 12. The helix and strand free energy profiles of PPT.

of the side-chain–main-chain electrostatic interactions from the free energy function (see above and Equations 1–5). To keep the free energy function as simple as possible, only the contributions from main-chain–main-chain electrostatic interactions and from the desolvation of polar and nonpolar atoms are included in the free energy function. Addition of the side-chain–main-chain electrostatic interactions into the free energy function is expected to improve the results further.

The accuracy of a prediction method is realistically assessed by comparing the results of blind predictions obtained by the different algorithms as it been done in the CASP2 experiment. The average RMS deviation in the running window of six amino acids for the CASP2 target T8 is 0.31 Å. This result represents the best average accuracy in the window with six amino acids obtained in the CASP2 experiment (prediction T0008AB320).[49] The average RMS deviation of 2.27 Å in the running window of six amino acids for the CASP2 target $T30_{CASP2}$ is one of the best among the predictions for the proteins composed predominantly from the β-strands (prediction T0030AB794-10).[49,66] This deviation is larger

TABLE II. Screening Coefficients $\gamma^r_{nonlocal}$ and γ^r_{local}

Amino acid	$\gamma^r_{nonlocal}$	γ^r_{local}
Gly	−0.167	0.082
Ala	0.193	0.099
Val	0.427	0.455
Ile	0.419	0.435
Leu	0.368	0.334
Phe	0.337	0.384
Pro	−0.438	—
Met	0.393	0.319
Trp	0.236	0.309
Cys	0.209	0.228
Ser	0.040	0.140
Thr	0.210	0.299
Asn	0.192	0.160
Gln	0.294	0.226
Tyr	0.240	0.330
His	0.173	0.206
Asp	0.095	0.160
Glu	0.366	0.262
Lys	0.236	0.211
Arg	0.357	0.385

TABLE III. Hydrophobic and Desolvation Coefficients*

Atom type	σ
Hydrophobic carbon or sulphur	−0.015
Charged O of Asp and Glu	$0.51862 \cdot 10^{-5}$
Charged N of Lys	$0.42965 \cdot 10^{-4}$
Charged N of Arg	$0.15595 \cdot 10^{-5}$

*The units for σ_h and σ_j are kcal/(molÅ^2) and kcal/(molÅ^6), respectively.

TABLE IV. Comparison of the Positions of α-Helices in the Experimental and Predicted Structures Obtained by the Monte Carlo Minimization (MC)*

Protein	Helix No.	Experimental	Predicted with MC	Predicted with LR
ENH	1	10–22	10–37	9–37
ENH	2	28–38	—	—
ENH	3	42–53	43–51	39–52
GTO	1	2–28	3–28	6–31
GTO	2	32–56	32–54	35–54
ICB	1	3–14	6–15	7–15
ICB	2	25–35	22–35	24–38
ICB	3	46–53	46–56	47–54
ICB	4	63–74	63–73	62–75
PGA	1	23–36	20–36	22–33
UBQ	1	23–34	24–33	25–29
BRN	1	7–17	11–17	12–16
BRN	2	27–32	26–32	28–32
T30	1	3–8	Missing	4–6
T8	1	2–25	2–26	2–28
PPT	1	14–31	15–34	16–30

*The assignment of α-helices is calculated by the Kabsch and Sander algorithm.[65] The predictions of α-helices based on the negative helix free energy profiles calculated with the Model II[37] by the Lifson-Roig theory (LR) are also shown (see Figs. 1–12).

TABLE V. Average RMS Deviations of C_α Atoms (Å) for Secondary and Super-secondary Structures Between the Predicted and the Experimental Structures*

Protein	Second. (Å)	Super-sec. (Å)	Tot (Å)	No. second.	No. super-sec.
ENH	0.96	3.33	12.3	7.0	16.8
GTO	0.91	1.36	3.5	6.3	15.5
ICB	1.20	3.10	7.3	6.8	16.3
GFC	1.58	4.62	13.4	5.9	14.7
TEN	1.90	6.18	18.2	6.8	16.3
HOE	1.48	6.86	16.4	5.8	15.1
PGA	0.83	3.30	7.0	5.4	13.2
UBQ	0.68	2.93	14.6	4.5	11.3
BRN	1.85	4.28	19.7	5.8	14.8
$T30_{normal}$†	1.95	5.47	16.2	5.6	14.4
$T30_{CASP2}$‡	1.89	5.61	11.7	5.6	14.4
T8	0.57	0.97	1.7	7.3	16.0
PPT	1.46	3.10	4.2	6.8	14.3

*The RMS deviations between the predicted and the experimental structures of complete proteins (Tot) and the average number of amino acids in the secondary and super-secondary structures are also shown.
†$T30_{normal}$ corresponds to the predicted structure of T30 obtained by the minimization procedure identical to one used for other 11 proteins in which the disulphide bridges are ignored.
‡$T30_{CASP2}$ represents the predicted structure of T30 obtained by the procedure in which the distances between the sulphur atoms in the two pairs of disulphide bridges are minimized together with the free energy function.

than the RMS deviations of other proteins used in the present study. The reason lies in the absence of the N-terminal α-helix from the predicted conformation obtained by the Monte Carlo method. This α-helix, however, is correctly predicted by the free energy profiles (Table IV, Fig. 10). The most likely reason for this result is the constraints imposed by the disulphide bridges used in the prediction $T30_{CASP2}$ (see Materials and Methods).

Prediction of the Super-secondary Structure

The majority of the super-secondary structures are predicted accurately for all proteins (Fig. 13, Tables V, VI). The average RMS deviation of C_α atoms between the experimental and the predicted super-secondary structures is 3.8 Å (from Table V). The super-secondary structures of a protein are represented by two neighbor folding units. In each protein with n folding units there are $n - 1$ overlapping super-secondary structures.

The accuracies of predicting 3-D super-secondary structures in the α proteins are better than the accuracies of predicting super-secondary structures in the β proteins. The minimum RMS deviations for the fragments with 15 amino acids are between 0.38 Å and 1.24 Å for α proteins and between 2.01 Å and 2.12 Å for β proteins (Table VI). The average RMS deviations in the running window with 15 amino acids (Table VI) are between 0.60 Å and 3.14 Å for α proteins and between 6.78 Å and 8.16 Å for β proteins. The running window with 15 amino acids is used for these comparisons, because the average number of amino acids in the super-secondary structures is 14.9.

Figure 13 shows the x-ray and the predicted super-secondary structures of ubiquitin (UBQ).[67] The sequence of UBQ is divided into ten folding units (Fig. 8). The 3-D structures of the majority of the super-secondary structures are predicted accurately (see Fig. 13 and Tables IV–VI). A distinct similarity between the experimental and predicted structures exists even for the fragments with relatively large RMS deviations. For example, the large RMS deviation of 4.9 Å for the fragment with amino acids 1–18 is caused by forming the hairpin turn displaced by two residues from the position observed in the experimental structure. Figure 13A also shows that the hairpin twist of the fragment with amino acids

TABLE VI. Minimum and Average RMS Deviations of C_α Atoms (Å) i n Running Windows of 6, 15, 20, and 40 Amino Acids Between the Predicted and the Experimental Structures

	Minimum				Average			
Protein	6	15	20	40	6	15	20	40
ENH	0.11	0.37	2.97	9.36	0.93	3.14	4.84	10.95
GTO	0.06	0.38	0.62	2.55	0.61	1.14	1.49	2.78
ICB	0.15	1.24	2.70	3.83	1.23	3.11	3.91	5.63
GFC	0.93	2.01	4.28	8.38	1.83	4.94	6.78	12.13
TEN	0.35	2.08	5.02	8.21	1.73	5.72	7.57	12.48
HOE	0.37	2.12	4.31	9.18	1.98	6.15	8.16	12.23
PGA	0.15	0.42	1.91	5.07	1.47	3.38	4.05	5.65
UBQ	0.13	1.91	3.61	7.35	1.47	3.99	5.10	10.22
BRN	0.15	2.54	3.30	7.53	1.88	4.75	6.34	11.09
$T30_{normal}$*	0.32	2.25	3.73	9.51	2.21	5.76	7.39	10.75
$T30_{CASP2}$†	0.48	3.15	4.13	7.76	2.27	5.48	6.38	9.10
T8	0.10	0.38	0.69	—	0.31	0.60	0.88	—
PPT	0.06	0.34	0.52	—	1.13	2.29	2.86	—

*$T30_{normal}$ corresponds to the predicted structure of T30 obtained by the minimization procedure identical to one used for the other 11 proteins in which the disulphide bridges are ignored.
†$T30_{CASP2}$ represents the predicted structure of T30 obtained by the procedure in which the distances between the sulphur atoms in the two pairs of disulphide bridges are minimized together with the free energy function.

1–18 is well reproduced. To predict such a twist in hairpins, a large variability of torsion angles ϕ and ψ is necessary. This is achieved by using the library of torsion angles obtained from the experimental protein structures. Such similarities between the experimental and the predicted super-secondary structures exist in all other proteins studied in this work.

The accuracy of predicting the super-secondary structures with the Monte Carlo method can be objectively assessed by comparing the blind 3-D predictions of two proteins T8 and T30 with the predictions calculated by the other contemporary algorithms. The minimum RMS deviations of C_α atoms in the running windows of 15 and 20 amino acids for the CASP2 target T8 (T0008AB320) are 0.38 Å and 0.69 Å, respectively (Table VI). These accuracies are among the best obtained in the recent CASP2 experiment.[49,66] The 3-D predictions of the super-secondary structures of T30 have been classified at the top of the list in the automatic assessment of the CASP2 experiment (predictions T0030AB794-1 to T0030AB794-10).[49] The minimum RMS deviations of C_α atoms in the running window of 15 and 20 amino acids for $T30_{CASP2}$ are 3.15 Å and 4.13 Å, respectively (Table VI). These minimum RMS deviations are large compared with the minimum deviations of proteins with α-helices[66]; however, they are small compared with the minimum deviations of the proteins without α-helices. For the predicted structure obtained by the constrained disulphide bridges ($T30_{CASP2}$), the minimum RMS deviation of C_α atoms in the running window of 15 is larger in comparison with other predictions without such restrictions (Table VI).

The high accuracy of predictions of the local secondary and super-secondary structures achieved in this work (Fig. 13, Tables IV–VI) suggests that the nucleation does not occur at only a few regions in the protein but that the native structure originates from the entire protein sequence. These results support the hypothesis postulated by Lattman and Rose in which the stereochemical code for the folding process is distributed along the entire protein sequence and is not centralized to some discrete sites.[68,69] They have also suggested that the side-chain conformational entropy is the crucial physical factor for the local control of the protein folding process.[30,31,68,69] The results presented in this work, however, do not support this thesis. We show that the hydrophobic and predominantly the electrostatic interactions (Equations 1–5) represent the essential physical factors responsible for local control of the protein folding process.

Prediction of the Native 3-D Structure of Proteins

The native 3-D structures are predicted correctly for three proteins (T8, PPT, and GTO), composed predominantly of α-helices. The peptide T8, with 28 amino acids,[50] was a target in the ab initio prediction of the 3-D structure in the CASP2 experiment.[48] The helix free energy profile of T8 is more negative than the strand free energy profile along the entire sequence (Fig. 11). These profiles therefore indicate that this peptide is completely α-helical. The Monte Carlo minimizations confirm this result. The RMS deviation of C_α atoms between the experimental and the predicted structures is 1.7 Å (Table V, Fig. 14).

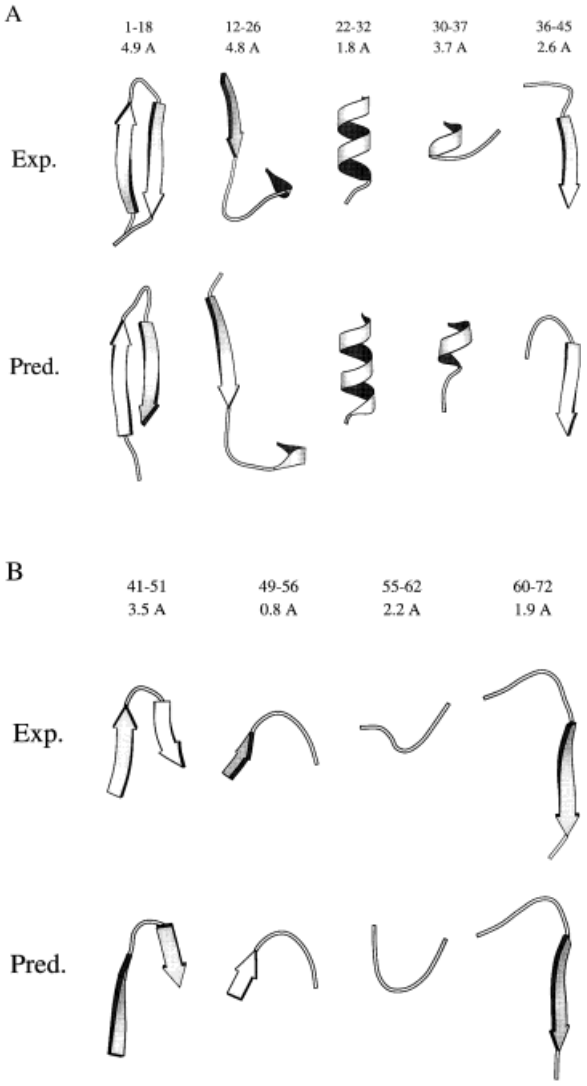

Fig. 13. **A** and **B.** Comparison between the experimental and the predicted super-secondary structures of UBQ. The nine super-secondary structures are obtained from pairs of ten neighbor folding units (Fig. 8). The average RMS deviation of C_α atoms between the experimental and the predicted super-secondary structures for UBQ is 2.93 Å (from Table V). The overlapping fragments contain between 8 and 18 amino acids. The 3-D structures of the majority of fragments are predicted accurately. This result suggests that the control in forming the native-like local structure is distributed along the entire protein sequence. The Molscript program was used to obtain this and the following figures.[82]

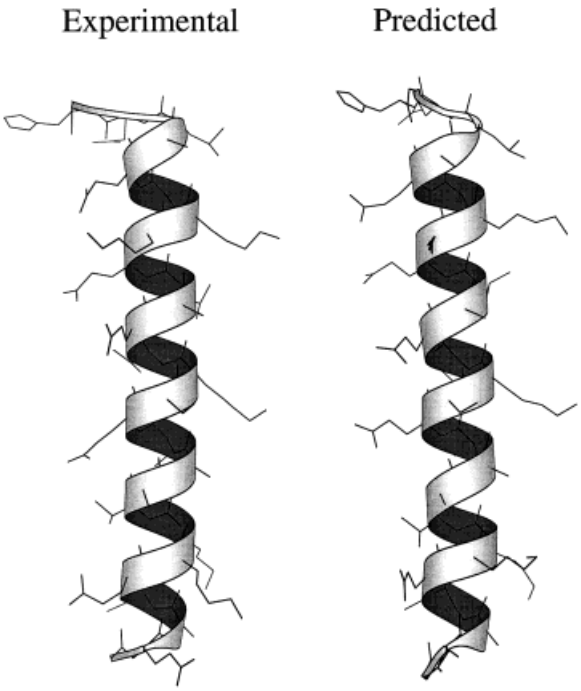

Fig. 14. Comparison between the experimental and the predicted structures of T8. The RMS deviation of C_α atoms between the experimental and the predicted structures is 1.7 Å (Table V). The largest deviations are at both ends of the α-helix. The minimum RMS deviations of predicting smaller fragments with 6, 15, and 20 amino acids can be as small as 0.10 Å, 0.38 Å, and 0.69 Å, respectively (see Table VI).

The experimental structures of the homologous peptides were known at the time of the CASP2 experiment; therefore, it cannot be considered as an entirely blind prediction. Nevertheless, the high accuracy of this prediction together with the correct free energy profiles is a clear indication that the new method has a strong prediction ability.

The experimental x-ray structure[70] and the nuclear magnetic resonance (NMR)[71] structure have been determined for the peptide PPT with 36 amino acids. The RMS deviation between the x-ray and NMR structures is 2.9 Å, which indicates that the 3-D structure of this peptide is flexible and depends on the environment. The free energy profiles of PPT (Fig. 12) predict the existence of an α-helix at the C-terminal end. The α-helix packed against the extended N-terminal end is correctly predicted by the Monte Carlo method (Fig. 15). The RMS deviations of C_α atoms between the predicted and the experimental x-ray and NMR structures are 4.2 Å and 3.8 Å, respectively.

Two α-helices are packed against each other in the experimental x-ray structure of GTO.[72] The protein GTO is a *rop* protein with a total of 62 amino acids. The free energy profiles predict the existence of two α-helices (Fig. 2). The packing of α-helices against each other is accurately predicted with the Monte Carlo minimization procedure (Fig. 16). The RMS deviation of C_α atoms between the experimental and the predicted structures is 3.5 Å.

The method fails to predict the native 3-D structures of other proteins (Table V). Although a certain similarity exists between predicted and experimental folds, the RMS deviations between these structures are large. For example, in the experimental structure of protein ICB with 85 amino acids the four α-helices are arranged in the bundle.[73] The free

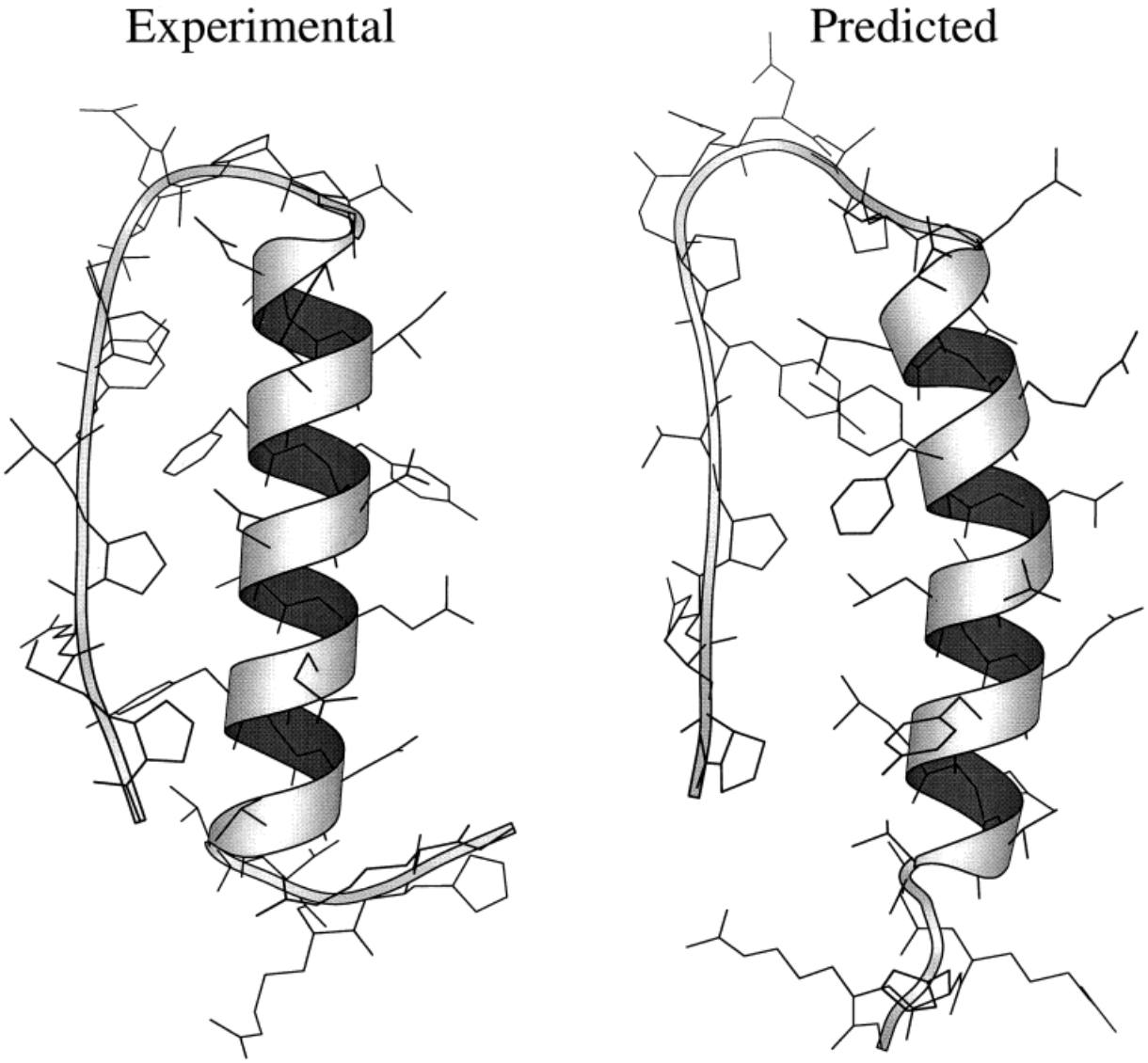

Fig. 15. Comparison between the experimental x-ray and the predicted structures of PPT. The x-ray and NMR structures are known for PPT. The RMS deviation of C_α atoms between the x-ray and the NMR structures is 2.9 Å. The RMS deviations between the predicted and the x-ray structures is 4.2 Å (shown in this figure). The RMS deviation between the predicted and NMR structures is 3.8 Å.

energy profiles indicate that this protein is composed of four α-helices (Fig. 3). The α-helix bundle in the experimental 3-D structure of ICB is correctly predicted by the Monte Carlo method; however, the RMS deviation of C_α atoms between the experimental and the predicted structures is 7.3 Å. Atoms in the predicted structure are not as tightly packed as atoms in the experimental structure (Fig. 17). The large RMS deviation between the experimental and predicted conformations is probably caused by the inaccuracies of predicting α-helices with the simplified free energy function used in this work (see above and Equations 1–5).

The β-strands are arranged in the β-barrel in the native structure of the protein T30.[51] This protein, with 66 amino acids, was the target for the ab initio 3-D prediction in the CASP2 experiment.[48] Two minimization procedures are used in predicting the 3-D structure of T30 (see Materials and Methods). The conformation obtained in the procedure, in which the distances between the sulphur atoms in the two pairs of disulphide bridges are minimized, ($T30_{CASP2}$) was used in the CASP2 experiment[48] (Tables V and VI). As a consequence of these constraints, the predicted structure of T30 is more compact than the predicted structures of other pro-

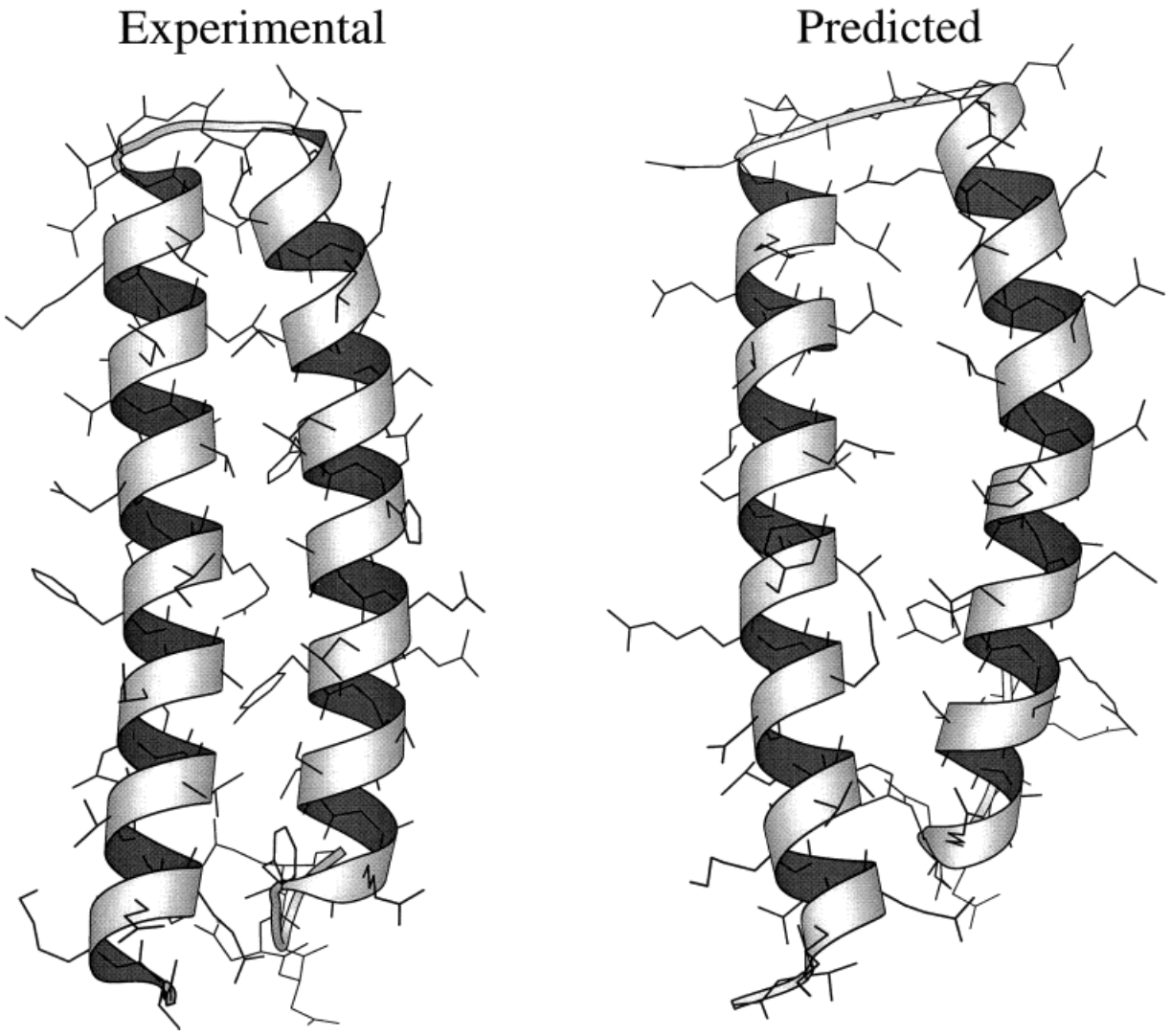

Fig. 16. Comparison between the experimental and the predicted structures of GTO. The RMS deviation of C_α atoms between the experimental and the predicted structures is 3.5 Å. Note the favorable interactions between the nonpolar amino acids in the experimental and the predicted structures.

teins used in this work. Although the fold of this protein was correctly predicted as the β-barrel, the RMS deviation between the experimental and the predicted structures is large (11.7 Å). The main reason lies in one of the hairpins, which was wrongly placed into the β-barrel. The predicted structure has a distinctive hydrophobic core and is similar to the structure of the SH3 domains.[48]

The method does not predict correctly the native 3-D structures for proteins composed predominantly from the β-strands. The packing of the atoms in the predicted structures is less dense than it is in the experimental structures. A plausible explanation for the deficiencies is the minimization procedure used in the Monte Carlo method. The minimization is performed by using the concept of hierarchic condensation. It has been suggested that during the folding process of β proteins having the β-barrel structure, the chain folds first at approximately the central position in the sequence, forming a long, two-stranded β-ribbon.[74] The process then continues by folding in half again, resulting in more long-range interactions between β-strands. This folding mechanism was successfully used for the prediction of β-strand connections and topologies in β proteins.[74,75] It is obvious that hierarchic condensation could not emulate such a process and generate the correct arrangement of β-strands in these proteins, because the long-range interactions between the N-terminal and C-terminal amino acids are already established in the first folding in half. We therefore suggest that hierarchic condensation is not the appropriate mech-

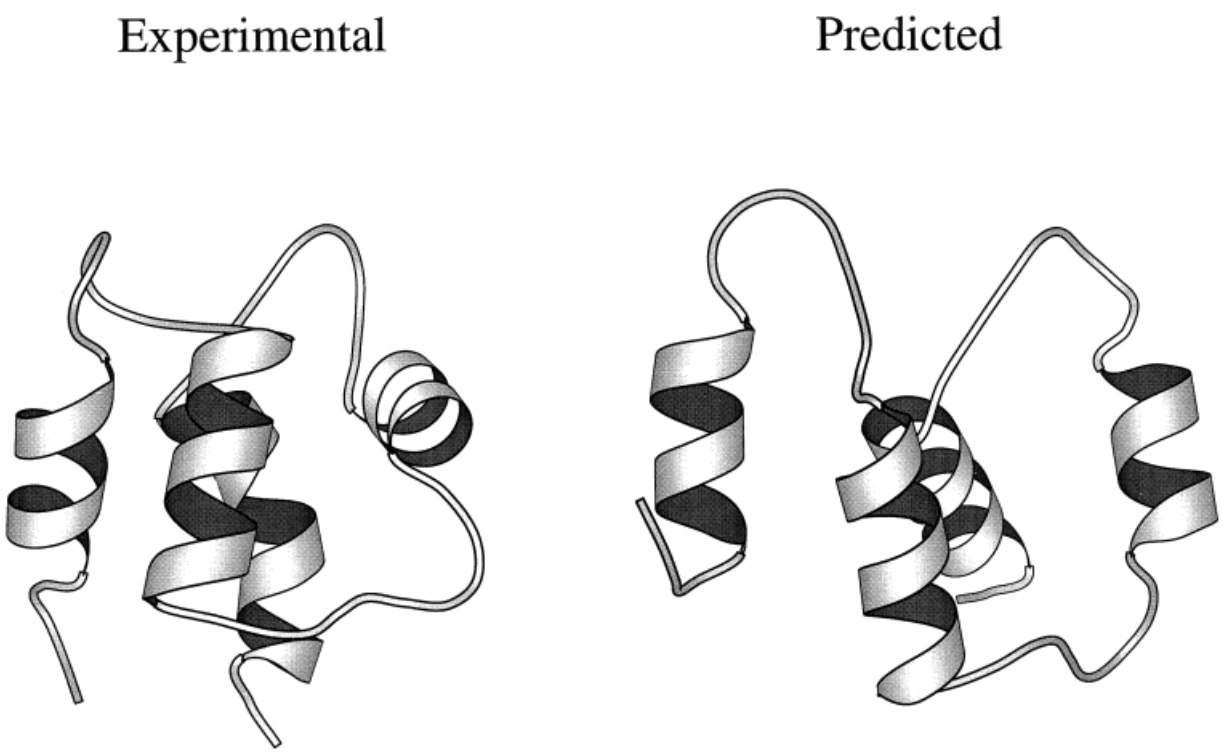

Fig. 17. Comparison between the experimental and the predicted structures of ICB. The α-helix bundle in the experimental 3-D structure of ICB is correctly predicted by the Monte Carlo method; however, the predicted structure is not as tightly packed as the experimental structure. The RMS deviation of C_α atoms between the experimental and the predicted structures is 7.3 Å.

anism for simulating the folding process of proteins composed predominantly from β-strands.

CONCLUSIONS

With the new prediction algorithm the 3-D structures of the majority of the local secondary and super-secondary structures are predicted correctly for 12 proteins. This result suggests that the native protein structure originates from the entire protein sequence. The hydrophobic and predominantly the electrostatic interactions (Equations 1–5) represent the essential physical factors responsible for local control of the protein folding process.

The native 3-D structure is predicted accurately for three proteins, composed predominantly from the α-helices. The hierarchic condensation may be a plausible model for simulating the folding mechanism for these proteins. The native 3-D structure of the other nine proteins is not predicted correctly, although a certain similarity between the predicted and the experimental folds does exist. Hierarchic condensation is not the appropriate mechanism for simulating the folding process of proteins made up predominantly from the β-strands.

The correct predictions of the local structure support the electrostatic screening model for amino acid preferences. The simple free energy function, with only 44 fitted coefficients, is adequate and essential for the performance of the Monte Carlo algorithm. The free energy function contains contributions from the local main-chain electrostatic interactions, the main-chain hydrogen bonding, and the desolvation of hydrophobic and polar atoms. The main deficiency of this free energy function is the absence of the side-chain–main-chain interactions, which are important for the predictions of both ends of helices in proteins.

The prediction ability of the new algorithm was tested in the blind ab initio predictions of 3-D structures in the recent CASP2 experiment. The predictions of the local 3-D structure of the two CASP2 targets T8 and T30 are quite good. They are classified at the top of the list in the automatic evaluation of 3-D predictions. The accuracies of the blind predictions are similar to those obtained for the proteins when the experimental structures are known in advance.

ACKNOWLEDGMENTS

We are grateful to D. Hadži and D. Kocjan for reading the manuscript and helpful suggestions.

REFERENCES

1. Anfinsen, C.A. Principles that govern the folding of protein chains. Science 181:223–230, 1973.
2. Levinthal, C. Are there pathways for protein folding? J. Chim. Phys. 65:44–45, 1968.

3. Harrison, S.C., Durbin, R. Is there a single pathway for the folding of a polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. 82:4028–4030, 1985.
4. Baker, D., Agard, D.A. Kinetics versus thermodynamics in protein folding. Biochemistry 33:7505–7509, 1994.
5. Baldwin, R.L. The nature of protein folding pathways: The classical versus new view. J. Biomol. NMR 5:103–109, 1995.
6. Kim, P.S., Baldwin, R.L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. Annu. Rev. Biochem. 51:459–489, 1982.
7. Karplus, M., Weaver, D.L. Protein folding dynamics. Nature 260:404–406, 1976.
8. Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Protein Chem. 14:1–64, 1959.
9. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., Chan, H.S. Principles of protein folding—a perspective from simple exact models. Protein Sci. 4:561–602, 1995.
10. Levitt, M., Warshel, A. Computer simulation of protein folding. Nature 253:694–698, 1975.
11. Hagler, A.T., Honig, B. On the formation of protein tertiary structure on a computer. Proc. Natl. Acad. Sci. U.S.A. 75:554–558, 1978.
12. Wilson, C., Doniach, S. A computer model to dynamically simulate protein folding. Proteins 6:193–209, 1989.
13. Honeycutt, J.D., Thirumalai, D. Metastability of the folded states of globular proteins. Proc. Natl. Acad. Sci. U.S.A. 87:3526–3529, 1990.
14. Skolnick, J., Kolinsky, A. Simulations of the folding of a globular protein. Science 250:1121–1125, 1990.
15. Crippen, G.M., Snow, M.E. A 1.8 angstrom resolution function for protein folding. Biopolymers 29:1479–1489, 1990.
16. Covell, D.G. Folding protein α-carbon chain into compact forms by Monte Carlo methods. Proteins 14:409–420, 1992.
17. Callaway, D.J.E. Solvent-induced organization: A physical model of folding myoglobin. Proteins 20:124–138, 1994.
18. Sun, S., Thomas, P.D., Dill, K.A. A simple protein folding algorithm using a binary code and secondary structure constraints. J. Mol. Biol. 235:600–624, 1994.
19. Srinivasan, R., Rose, G.D. Linus: A hierarchic procedure to predict the fold of a protein. Proteins 22:81–99, 1995.
20. Avbelj, F., Moult, J. The conformation of folding initiation sites in proteins determined by computer simulation. Proteins 23:129–141, 1995.
21. Pedersen, J.T., Moult, J. Ab initio structure prediction for small polypeptides and protein fragments using genetic algorithm. Proteins 23:454–460, 1995.
22. Boczko, E.M., Brooks, C.L. First-principles calculation of the folding free energy of a three helix bundle protein. Science 269:393–396, 1995.
23. Yue, K., Dill, K.A. Folding proteins with a simple energy function and extensive conformational searching. Protein Sci. 5:254–261, 1996.
24. Miyasawa, S., Jernigan, R.L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18:534–552, 1985.
25. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. J. Mol. Biol. 213:859–883, 1990.
26. Avbelj, F. Use of a potential of mean force to analyse free energy contributions in protein folding. Biochemistry 31: 6290–6297, 1992.
27. Dill, K.A. Dominant forces in protein folding. Biochemistry 29:7133–7155, 1990.
28. Blaber, M., Zhang, X., Matthews, B.W. Structural basis of amino acid α-helix propensity. Science 260:1637–1640, 1993.
29. Blaber, M., Zhang, X., Lindstrom, J.D., Pepiot, S.D., Basse, W.A., Mathews, B.W. Determination of α-helix propensity within the context of a folded protein. J. Mol. Biol. 235:600–624, 1994.
30. Creamer, T.P., Rose, G.D. Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. Proc. Natl. Acad. Sci. U.S.A. 89:5937–5941, 1992.
31. Creamer, T.C., Rose, G.D. α-helix-forming propensities in peptides and proteins. Proteins 19:85–97, 1994.
32. Hermans, J., Anderson, A.G., Yun, R.H. Differential helix propensity of small apolar side chain studied by molecular dynamics simulation. Biochemistry 31:5646–5653, 1992.
33. Yun, R.H., Hermans, J. Conformational equilibria of valine studied by dynamics simulation. Protein Eng. 4:761–766, 1991.
34. Bai, Y., Englander, S.W. Hydrogen bond strength and β-sheet propensities: The role of a side chain blocking effect. Proteins 18:262–266, 1994.
35. Bai, Y., Milne, J.S., Mayne, L., Englander, S.W. Primary structure effects on peptide group hydrogen exchange. Proteins 17:75–86, 1993.
36. Avbelj, F., Moult, J. Role of electrostatic screening in determining protein main chain conformational preferences. Biochemistry 34:755–764, 1995.
37. Avbelj, F., Fele, L. Role of main-chain electrostatics, hydrophobic effect, and side-chain conformational entropy in determining the secondary structure of proteins. submitted, 1997.
38. Roder, H., Elove, G.A., Englander, S.W. Structural characterization of folding intermediates in cytochrome *c* by hydrogen-exchange labelling and proton NMR. Nature 335:700–704, 1988.
39. Udgaonkar, J.B., Baldwin, R.L. NMR evidence for an early framework intermediate on the folding pathway of ribonuclease a. Nature 335:694–699, 1988.
40. Baum, J., Dobson, C.M., Evans, P.A., Hanley, C. Characterization of a partly folded protein by NMR methods: Studies on the molten globule state of guinea pig α-lactalbumin. Biochemistry 28:7–13, 1989.
41. Bycroft, M., Matouschek, A., Kellis, J.T., Serrano, L., Fersht, A.R. Detection and characterization of a folding intermediate in barnase by NMR. Nature 346:488–490, 1990.
42. Rose, G.D. Hierarchic organization of domains in globular proteins. J. Mol. Biol. 134:447–470, 1979.
43. Crippen, G.M. The three dimensional organization of proteins. J. Mol. Biol. 126:315–332, 1978.
44. Lesk, A.M., Rose, G.D. Folding units in globular proteins. Proc. Natl. Acad. Sci. U.S.A. 78:4304–4308, 1981.
45. Levitt, M., Chothia, T. Structural patterns in globular proteins. Nature 261:552–558, 1976.
46. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
47. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
48. CASP2. Critical assessment of methods for structure prediction. URL: http://PredictionCenter.llnl.gov/casp2, 1997.
49. CASP2. Critical assessment of methods for structure prediction: Ab-initio evaluation report. URL: http://PredictionCenter.llnl.gov/casp2/AbInitioEvaluation, 1997.
50. Ogihara, N.L., Weiss, M.S., Degrado, W.F., Eisenberg, D. The crystal structure of the designed trimeric coiled coil coil-v(a)l(d): Implications for engineering crystals and supramolecular assemblies. Protein Sci. 6:80–88, 1997.
51. Holliger, P., Riechmann, L. to be published, 1997.
52. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. Scop: A structural classification of protein database for the investigation of sequences and structures. J. Mol. Biol. 247:536–540, 1995.
53. Lifson, S., Roig, A. On the theory of helix-coil transition in polypeptides. J. Chem. Phys. 34:1963–1974, 1961.
54. Kabsch, W., Sander, C. Identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. U.S.A. 81:1075–1078, 1984.
55. Brunet, A.P., Huang, E.S., Huffine, M.E., Loeb, J.E., Weltman, R.J., Hecht, M.H. The role of turns in the structure of an alpha-helical protein. Science 250:1121–1125, 1990.
56. Warshel, A., Russell, S.T. Calculation of electrostatic inter-

actions in biological systems and in solutions. Q. Rev. Biophys. 17:283–422, 1984.
57. Chothia, C. Hydrophobic bonding and accessible surface area in proteins. Nature 248:338–339, 1974.
58. Lee, B., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
59. Chothia, C. The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 105:1–14, 1976.
60. McQuarrie, D.A. "Statistical Mechanics." New York: Harper and Row, 1976.
61. Beveridge, D.L., DiCapua, F.M. Free energy via molecular simulations: A primer. In: "Computer Simulations of Biomolecular Systems. Theoretical and Experimental Applications." van Gunsteren, W.F., Weiner, P.K. (eds.). Leiden: Escom, 1989:1–6.
62. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tusami, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
63. Dauber-Osguthorpe, P., Roberts, V.A., Osguthorpe, D.J., Wolff, D.J., Genest, M., Hagler, A.T. Structure and energetics of ligand binding to proteins. Proteins 4:31–47, 1988.
64. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, M.N., Teller, E. Equation of state calculation by fast computer machines. J. Chem. Phys. 21:1087–1092, 1952.
65. Kabsch, W., Sander, C. Dictionary of protein structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
66. Lesk, A.M. Casp2: Report on ab-initio predictions. Proteins Suppl. 1:151–166, 1997.
67. Vijay-Kumar, S., Bugg, C.E., Cook, W.J. Structure of ubiquitin refined at 1.8 angstroms resolution. J. Mol. Biol. 194:531–544, 1987.
68. Lattman, E.E., Rose, G.D. Protein folding—what is the question? Proc. Natl. Acad. Sci. U.S.A. 90:439–441, 1993.
69. Rose, G.D., Creamer, T.P. Protein folding: Predicting predicting. Proteins 19:1–3, 1994.
70. Blundell, T.L., Pitts, J.E., Tickle, I.J., Wood, S.P., Wu, C.-W. X-ray analysis (1.4-angstroms resolution) of avian pancreatic polypeptide, small globular protein hormone. Proc. Natl. Acad. Sci. U.S.A. 78:4175, 1981.
71. Li, X., Sutcliffe, M.J., Schwartz, T.W., Dobson, C.M. Sequence specific NMR assignments and solution structure of bovine pancreatic polypeptide. Biochemistry 31:1245–1253, 1992.
72. Predki, P.F., Agrawal, V., Brunger, A.T., Regan, L. Amino-acid substitutions in a surface turn modulate protein stability. Nat. Struct. Biol. 3:54–58, 1996.
73. Szebenyi, D.M.E., Moffat, K. The refined structure of vitamin D-dependent calcium-binding protein from bovine intestine. Molecular details, ion binding, and implications for the structure of other calcium-binding proteins. J. Biol. Chem. 261:8761–8777, 1986.
74. Ptitsyn, O.B., Finkelstein, A.V. In: "Protein Folding." Jaenicke R. (ed.). Amsterdam: Elsevier, 1980:101–115.
75. Richardson, J.S. Protein anatomy. Adv. Protein Chem. 34:168–339, 1981.
76. Clarke, N.D., Kissinger, C.R., Desjarlais, J., Gilliland, G.L., Pabo, C.O. Structural studies of the engrailed homeodomain. Protein Sci. 3:1779–1787, 1994.
77. Kohda, D., Terasawa, H., Ichikawa, S., Ogura, K., Hatanaka, H., Mandiyan, V., Ullrich, A., Schlessinger, J., Inagaki, F. Solution structure and ligand-binding site of the carboxy-terminal sh3 domain of grb2. Structure 2:1029–1040, 1994.
78. Leahy, D.J., Hendrickson, W.A., Aukhil, I., Erickson, H.P. Structure of a fibronectin type iii domain from tenascin phased by mad analysis of the selenomethionyl protein. Science 258:987–991, 1992.
79. Pflugrath, J.W., Wiegand, G., Huber, R., Vertesy, L. Crystal structure determination, refinement and the molecular model of the alpha-amylase inhibitor hoe-467a. J. Mol. Biol. 189:383–386, 1986.
80. Gallagher, T., Alexander, P., Bryan, P., Gilliland, G.L. Two crystal structures of the b1 immunoglobulin-binding domain of streptococcal protein g and comparison with NMR. Biochemistry 33:4721–4729, 1994.
81. Buckle, A.M., Fersht, A.R. Subsite binding in an rnase: Structure of a barnase-tetranucleotide complex at 1.76-a resolution. Biochemistry 33:1644–1653, 1994.
82. Kraulis, P.J. Molscript: A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.