PROTEINS: Structure, Function, and Genetics 23:566–579 (1995)

Knowledge-Based Protein Secondary Structure Assignment

Dmitrij Frishman and Patrick Argos
European Molecular Biology Laboratory, 69012 Heidelberg, Germany

ABSTRACT **We have developed an automatic algorithm STRIDE for protein secondary structure assignment from atomic coordinates based on the combined use of hydrogen bond energy and statistically derived backbone torsional angle information. Parameters of the pattern recognition procedure were optimized using designations provided by the crystallographers as a standard-of-truth. Comparison to the currently most widely used technique DSSP by Kabsch and Sander (Biopolymers 22:2577–2637, 1983) shows that STRIDE and DSSP assign secondary structural states in 58 and 31% of 226 protein chains in our data sample, respectively, in greater agreement with the specific residue-by-residue definitions provided by the discoverers of the structures while in 11% of the chains, the assignments are the same. STRIDE delineates every 11th helix and every 32nd strand more in accord with published assignments.** © 1995 Wiley-Liss, Inc.

Key words: protein structure analysis, hydrogen bond, torsional angle, α-helix, β-sheet

INTRODUCTION

Assignment of the secondary structural elements is an essential step in the characterization of three-dimensional protein structures and also serves as a departure point in many theoretical studies devoted to secondary structure prediction, modeling by homology, inverse protein folding, description of folding motifs, and the like (for a review, see ref. 1). Although intuitively the recognition of α-helices and β-sheets seems straightforward, an algorithmic solution is complicated by the fuzzy, often nonideal nature of these elements.

Several secondary structure assignment methods dependent on atomic resolution protein structures include detection of patterns in inter-C^{α} distances,[2] analysis of virtual bond angles and lengths between consecutive C^{α} atoms,[3] analysis of hydrogen bonding patterns,[4] comparison of interatomic distance matrices of structural fragments to idealized reference distance masks typical for a particular secondary structure type,[5] and quantification of the backbone curvature.[6] It is not surprising that techniques utilizing different approaches produce different assignments with disagreements up to 25%.[7] In fact, a detailed examination of 3 procedures by Colloc'h et al. showed complete agreement in only 64% of sequence sites in several proteins. This, however, does not automatically imply that all these methods deviate to the same extent from what one would call "intuitive reality." Colloc'h et al.[7] do not recommend any particular technique and suggest using a consensus assignment, but no evaluation is given.

Which method is the best? As noted by many authors,[4,5] there is no single and correct algorithm to assign secondary structural type and any method will be correct only within the framework of the definition upon which it relies. Nonetheless, different definitions aim at capturing the same reality, the typical appearance of secondary structural elements in hundreds of protein tertiary structures as reported in the Protein Data Bank[8] (PDB). This is reflected in the authors' assignments of helices, β-strands, and turns in the tertiary structures which they determined. In our opinion, these vast amounts of data provide the best and most complete standard-of-truth currently available. So, in lieu of asking which method is best, we think it appropriate to inquire: "Which criteria do cystallographers practically use for secondary structural assignment in newly determined protein structures and how can they be reproduced as best as possible in an automated algorithm?"

An extensive survey of papers devoted to protein three-dimensional structure determination reveals that crystallographers' assignments are based on consideration of hydrogen bonding using the definitions of Baker and Hubbard[9] (e.g., ref. 10), simplified distance criteria applied to donor and acceptor separation (e.g., refs. 11, 12), the more complex distance and geometric criteria by Presta and Rose[13] (e.g., ref. 14), hydrogen bonding patterns in combination with main-chain dihedral angles (e.g., refs. 15, 16), mainchain φ,ϕ angles only (e.g., ref. 17), the DSSP algorithm[4] with a stricter hydrogen bond definition (e.g., ref. 18), visual criteria (e.g., ref. 19), or a combination of several independent assignment

Received February 16, 1995; revision accepted July 13, 1995.
Address reprint requests to Dmitrij Frishman, European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, 69012 Heidelberg, Germany.

methods (e.g., ref. 20). In most cases crystallographers subject their assignments to careful visual inspection and subsequent modification if necessary. In spite of the considerable variety of approaches adopted, two main protein structural properties recur and play the most important role in structural element definition, namely, hydrogen bond patterns and backbone geometry generally expressed as mainchain dihedral angles.[21]

Analysis of the protein structure literature, both experimental and theoretical, shows that by far the most widely used automatic secondary structure assignment method is DSSP by Kabsch and Sander[4] which defines helices and sheets as repeating elementary hydrogen bonded patterns. In a large majority of cases, DSSP provides very good recognition of secondary structural elements and agrees well with intuitive visual criteria. Statistically, however, the agreement between the DSSP and crystallographers' assignments is between 70 and 100%, dependent on the structure quality and criteria used by the discoverers of the structure.[22] The purpose of this contribution is to create an automatic secondary structure assignment method which would reflect as well as possible known assignments contained in the current large collection of protein three-dimensional structures.[8]

METHODS

Outline of the Algorithm

In order to approximate as closely as possible the intuitive definition of α-helices and β-strands (as represented on the average by crystallographers' assignments), the weighted contribution of both the secondary structure forming hydrogen bonds and the backbone torsion angles must be considered. The quality of the elementary secondary structural units or patterns, four-residue turns for α-helices and bridges for β-sheets,[4] is expressed in terms of combined quantities which are a weighted product of the relevant hydrogen bond energies and statistically derived propensities of amino acid residues with given φ,ϕ values to occur in α-helices and β-sheets. Introduction of only one threshold for these quantities for each type of the hydrogen bonded pattern allows precise tuning of the recognition parameters since the patterns with corrupted torsional angles can still be accepted if they form strong hydrogen bonds and, vice versa, relatively weak hydrogen bonds can be compensated for by correct backbone geometry. Crystallographers' assignments as provided in hundreds of available coordinate sets are used systematically for tuning the thresholds in the recognition procedure. We refer to our technique as STRIDE for secondary STRuctural IDEntification.

Hydrogen Bond Energy

The hydrogen bond energy E_{hb} is calculated using the empirical energy function derived from the analysis of a large body of experimental data on hydrogen bond geometries in crystal structures of polypeptides, peptides, amino acids, and small organic compounds[23,24]:

$$E_{hb} = E_r \times E_t \times E_p$$

where E_r is the distance dependence of the hydrogen bond, and E_t and E_p describe its directional properties. The distance term is an 8–6 function:

$$E_r = \frac{C}{r^8} + \frac{D}{r^6}$$

where $C = -3E_m r_m^8$ kcal Å^8/mol, $D = -4E_m r_m^6$ kcal Å^8/mol, r is the distance between the donor and acceptor atoms participating in the hydrogen bond (see Fig. 1), and E_m and r_m are the optimal hydrogen bond energy and length, respectively. For mainchain–mainchain hydrogen bonds N—H···O, $E_m = -2.8$ kcal/mol and $r_m = 3.0$ Å.[23,24] The angular terms E_t and E_p have the following forms:

$$E_p = \cos^2 p$$

and

$$E_p = \begin{cases} (0.9 + 0.1 \sin 2t_i)\cos t_o, & 0 < t_i < 90° \\ K_1(K_2 - \cos^2 t_i)^3 \cos t_o, & 90° < t_i < 110° \\ 0, & t_i > 110° \end{cases}$$

where $K_1 = 0.9/\cos^6 110°$, $K_2 = \cos^2 110°$, and the angles t_i and t_o are respective angular deviations of the hydrogen atom from the bisector of the lone-pair orbital within the plane of the lone pair orbitals and from the plane of the lone pair orbitals (see Fig. 1).

For small separations between the interacting atoms, the distance potential E_r becomes repulsive and unfavorable energies result. This possibility exists for backbone N and O atoms due to errors in the X-ray or NMR determination of the protein structure. For the purposes of secondary structural assignment in this work, such distortions can usually be ignored unless the geometry of the hydrogen bond departs substantially from the norm in which case it can be accounted for by the angular dependence of the bond energy. Therefore, an additional energy functional constraint is included:

$$E_r = E_m \qquad \text{for } r < r_m$$

Torsional Angle Probabilities for α-Helices and β-Sheets

For each 20°-by-20° zone (i) on the Ramachandran map[25] of observed backbone dihedral angles in the many protein structures considered in this work, we calculated the probability P_i^α and P_i^β that the torsional angles for residues assigned in the α-helical or β-sheet state lie within the ith zone, i.e.,

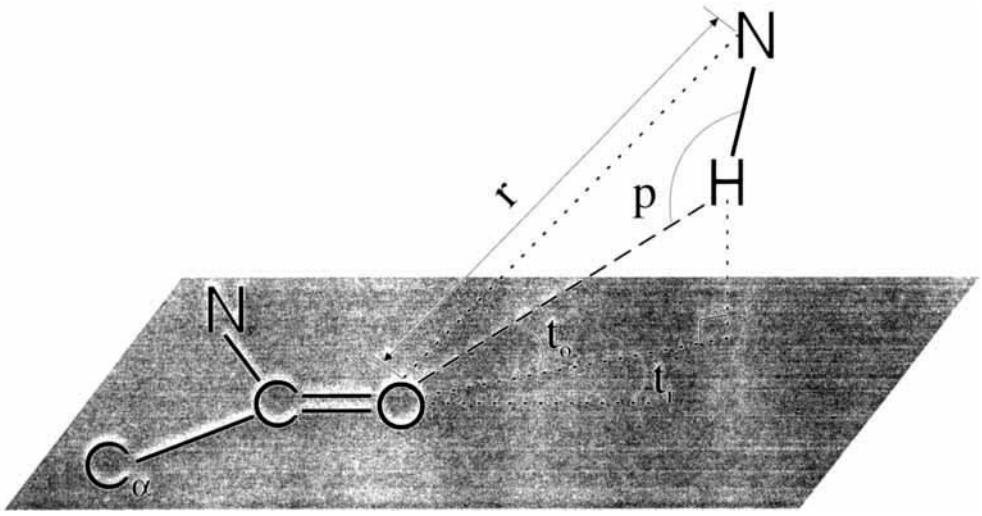


Fig. 1. An illustration of main-chain hydrogen bond geometry as adapted from Boobbyer et al.[23] The letter r refers to the donor–acceptor separation, the angle p indicates the departure of the hydrogen bond from linearity, t_i and t_o are deviations of the hydrogen atom from the bisector of the lone-pair orbitals within the plane of the lone pair orbitals and from the plane of the lone pair orbitals, respectively.

$$P_i^{\alpha} = \begin{cases} \dfrac{N_i^{\alpha}}{N_i^{\text{total}}} & \text{if } -180° < \varphi < 10° \text{ and } -120° < \phi < 45° \\ 0 & \text{otherwise} \end{cases}$$

and

$$P_i^{\beta} = \begin{cases} \dfrac{N_i^{\beta}}{N_i^{\text{total}}} & \text{if } -180° < \varphi < 0°, -180° < \phi < -120° \text{ or } 45° < \phi < 180° \\ 0 & \text{otherwise} \end{cases}$$

where N_i^{α} and N_i^{β} are the respective numbers of residues in the given φ,ϕ zone defined in the HELIX and SHEET PDB records as occurring in α-helix and β-sheet and N_i^{total} is the total number of residues with torsion angles falling within the zone (i) in our data sample. Note that P_i^{α} and P_i^{β} are set to zero outside of the generally accepted α-helical and β-sheet areas (e.g., refs. 26, 27).

The probability distributions are smoothed using digital binomial filtration.[28] The resulting plots of zonal P^{α} and P^{β} values versus the dihedral dihedral angular ranges are, respectively, shown in Figure 2a and b.

Recognition of α-Helices

We define a minimal α-helix which should include at least two consecutive hydrogen bonds between the residues k and $k+4$ (Fig. 3) such that

$$E_{\text{hb}}^{k,k+4}\left(1 + W_1^{\alpha} + W_2^{\alpha} \cdot \frac{P_k^{\alpha} + P_{k+4}^{\alpha}}{2}\right) < T_1^{\alpha}$$

If this condition is fulfilled for two consecutive hydrogen bonds between residue pairs $(k,k+4)$ and $(k+1,k+5)$, the central four residues $k+1,k+2$, $k+3$, and $k+4$ are assigned to the α-helical state, "H." The edge residues k and $k+5$ are included in this minimal helix if they satisfy the additional conditions that $P_k^{\alpha} < T_2^{\alpha}$ and $P_{k+5}^{\alpha} < T_3^{\alpha}$, respectively. In the above formulas, P_k^{α}, P_{k+1}^{α}, P_{k+2}^{α}, P_{k+3}^{α}, P_{k+4}^{α}, and P_{k+5}^{α} are respective torsional angle probabilities (vide supra) for the residues k, $k+1$, $k+2$, $k+3$, $k+4$, and $k+5$; and W_1^{α} and W_2^{α} and T_1^{α}, T_2^{α}, T_3^{α} are empirical weights and thresholds to be optimized.

Recognition of β-Sheets

A minimal β-sheet is defined by two consecutive hydrogen bonded β-bridges belonging to one of the possible types depicted in Figure 4. The quality of a β-bridge is determined by the strength of both its hydrogen bonds and by the average statistical propensity of its internal residues to be in β-strand conformations. Internal residues are either those that participate in two hydrogen bonds with both their main-chain carbonyl oxygen and peptide hydrogen or those flanked by two residues each participating in one hydrogen bond. The conformation of the latter is not taken into account since on the edges of β-strands abrupt changes of the backbone direction often occur. This leads to values for at least the φ angles on the N-terminal strand edge and for at least ϕ angles on the C-terminal strand edge that lie outside of the β-sheet zone on the Ramachandran map. Correspondingly, for a β-bridge to be recognized as such, the two hydrogen bonds involved must satisfy the following conditions (see Fig. 4):

$$\begin{cases} E_{\text{hb1}}(1 + W_1^{\beta} + W_2^{\beta} \cdot \text{CONF}_{\text{Antiparallel}}) < T_{\text{Antiparallel}}^{\beta} \\ E_{\text{hb2}}(1 + W_1^{\beta} + W_2^{\beta} \cdot \text{CONF}_{\text{Antiparallel}}) < T_{\text{Antiparallel}}^{\beta} \end{cases}$$

and, for parallel β-bridges,

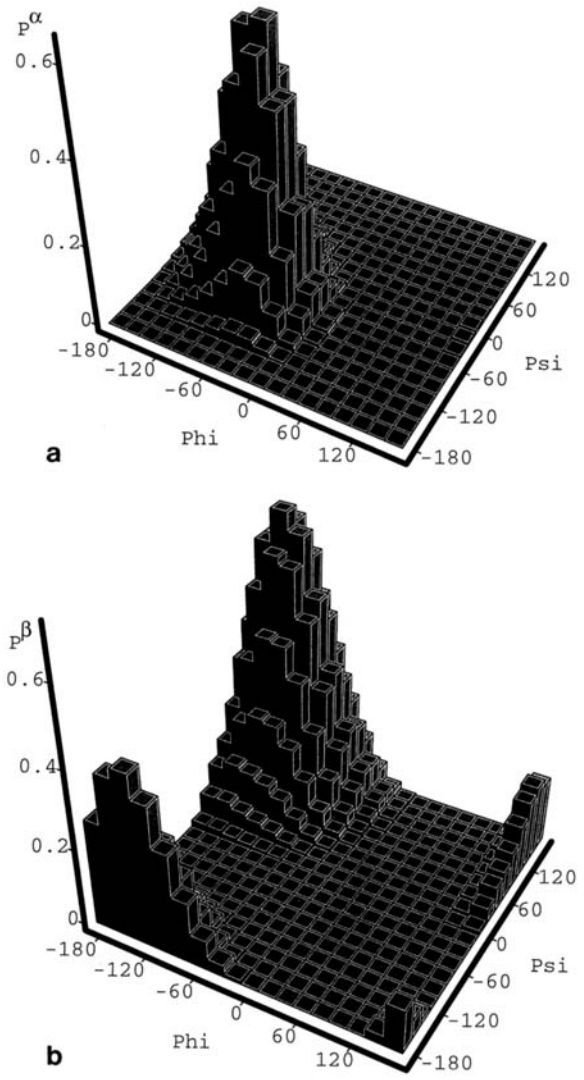


Fig. 2. Probabilities P^{α} (**a**) and P^{β} (**b**) for residues in α-helical and β-sheet secondary structural state, respectively, to have different torsional angles φ and ϕ. The histograms are given for 20°-by-20° zones.

$$\begin{cases} E_{hb1}(1 + W_1^{\beta} + W_2^{\beta} \cdot \mathrm{CONF}_{\mathrm{Parallel}}) < T^{\beta}_{\mathrm{Parallel}} \\ E_{hb2}(1 + W_1^{\beta} + W_2^{\beta} \cdot \mathrm{CONF}_{\mathrm{Parallel}}) < T^{\beta}_{\mathrm{Parallel}} \end{cases}$$

where E_{hb1} and E_{hb2} are energies of the first and second hydrogen bonds, respectively, and CONF = $\frac{(P^{\beta}_{Int1} + P^{\beta}_{Int2})}{2}$ if internal residues are present on both sides of the β-bridge (Fig. 4a,b,e) or CONF = $P_{Int}{}^{\beta}$ if only one residue is internal in a given β-bridge (Fig. 4c,d). $W_1{}^{\beta}$ and $W_2{}^{\beta}$ are empirical weights requiring optimization.

Adjacent bridges that fulfill the above criteria are merged into correspondingly antiparallel and parallel β-sheets with no more than four intervening residues between the bridges on one strand and no more than one residue on another stand. This latter definition for β-bulges is the same as that adopted by Kabsch and Sander[4] in DSSP. All residues within the merged adjacent bridges with possible bulges between them are assigned in an extended state, "E," with the exception of those bridges flanking the given β-sheet where only internal residues are assigned "E." In isolated β-bridges that have no suitable neighboring bridges for merging, internal residues are assigned the state "B." An exception are isolated bridges of the type III (Fig. 4c,d) where on one side there are two residues neither of which is internal. These two residues are assigned state "b." Isolated bridges involving such residues are rare.

Dataset

Representative sets of X-ray and NMR protein structures were gathered from a recent release of the PDB databank.[8] In correspondence with the goals of the present work, excluded were the protein chains that (1) list only C_{α} atoms, (2) contain no secondary structure assignment made by the authors, (3) contain obviously wrong secondary structure assignments (e.g., with long overlapping segments, unrealistically low or high secondary structure content, secondary structural element boundaries pointing to non-existing residues, etc.), (4) explicitly refer to existing automatic secondary structural assignment methods, most notably DSSP by Kabsch and Sander,[4] (5) are not yet published or in press, (6) have less than 70 residues, and (7) represent results of modeling studies.

From the remaining protein structures, three subsets were created: (1) X-ray structures at all resolutions as well as NMR structures (subset X + NMR), (2) X-ray structures with resolution better than 2.5 Å (subset X_HIGH), and (3) X-ray structures with resolution worse than 2.5 Å (subset X_LOW). Each of the three sets was made nonredundant using the program OBSTRUCT[29] such that no two chains in any set had sequence identity higher than 30%; the resulting nonredundant sets were referred to as X + NMR_30%, X_HIGH_30%, and X_LOW_30%. Finally, protein chains were excluded where, in the respective articles describing their structural determination, it is explicitly stated that DSSP[4] and, in one case, DEFINE_STRUCTURE[5] algorithms were used for secondary assignment. Thus, we made every possible effort to exclude from our dataset PDB entries with assignments of secondary structure made with existing automatic methods. Assignments made by eye or by manual application of certain consistent rules are also biased; however, such assignments are not inappropriate as standards of truth as long as the crystallographers subjected their classifications to careful visual inspection and manual modification if necessary.

The resulting dataset X + NMR_30% includes the

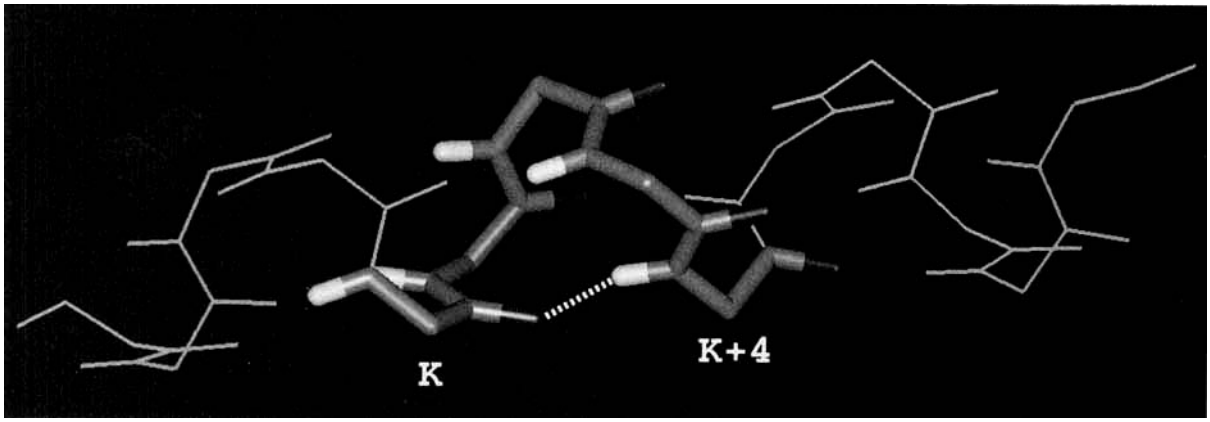


Fig. 3. Elementary α-helical pattern (shown in stick representation) including a hydrogen bond (dashed white line) between the hydrogen (white) associated with the peptide nitrogen of residue $K+4$ and the carbonyl oxygen (black) of residue K. The main chain is shown in a gray tone. In an ideal helix, the bond is continuously repeated between similarly separated residues.

following 226 protein chains:

1aak, 1ab2, 1abk, 1aca, 1ace, 1acp, 1acx, 1aep, 1agm, 1akeB, 1ald, 1alkB, 1apc, 1baa, 1bcx, 1bet, 1bllE, 1bmdA, 1bmv1, 1bw3, 1byc, 1cah, 1cc5, 1ccd, 1cdi, 1cdq, 1cgjE, 1cgt, 1chbH, 1cmbB, 1crl, 1csgB, 1ctm, 1cus, 1dlhA, 1draB, 1dsbB, 1eco, 1egl, 1ego, 1enk, 1fc1A, 1flv, 1fvcD, 1fxaA, 1gcg, 1gcs, 1gia, 1glaG, 1glv, 1gob, 1guhA, 1hbg, 1hbp, 1hcnA, 1hmf, 1hmy, 1hocA, 1hrhA, 1hsq, 1hstB, 1ikb, 1ipd, 1ithA, 1l97B, 1lab, 1lfb, 1lid, 1lldB, 1lmb4, 1lpe, 1mat, 1mbw, 1mdaH, 1mdaL, 1mec1, 1mec2, 1mpp, 1mrrA, 1ms2C, 1mup, 1nbvL, 1nnb, 1nnt, 1nrcB, 1nscA, 1pagB, 1pbxB, 1pda, 1pekE, 1pgd, 1pkp, 1pkt, 1pla, 1pmy, 1poeB, 1pou 1ppfE, 1prr, 1put, 1pyaC, 1pyp, 1rhd, 1ris, 1rmu2, 1rpa, 1serA, 1sgc, 1srdB, 1srnA, 1stb, 1sto, 1tfg, 1tlk, 1tlpE, 1tmuH, 1tnfC, 1tplB, 1troG, 1tssC, 1ttcA, 1ula, 1vsgB, 1vtmP, 1wsyA, 1wsyB, 1xllA, 1yat, 1ycc, 1zaaC, 2aaiA, 2achA, 2acu, 2azaA, 2bbvB, 2bopA, 2bpa2, 2bpp, 2btfA, 2btfP, 2ccyB, 2chsD, 2cna, 2cpl, 2dhc, 2dkb, 2dnjA, 2fgf, 2glsD, 2hmgB, 2hmgC, 2hmzB, 2hpdB, 2hsdB, 2hwd3, 2hwe1, 2ifb, 2int, 2lao, 2lbp, 2lh6, 2lhb, 2mcm, 2mhaB, 2mhbA, 2npx, 2pfkC, 2phh, 2phlA, 2pkaY, 2pna, 2por, 2prf, 2sas, 2scpB, 2sicI, 2snv, 2spcA, 2stv, 2taaA, 2trxB, 2tscB, 351c, 3aahA, 3bcl, 3c2c, 3ccp, 3chy, 3dfr, 3ecaC, 3fx2, 3gapB, 3hudB, 3hvtB, 3icb, 3ladA, 3mdeB, 3phv, 3rp2A, 3sdpA, 4ait, 4blmB, 4cla, 4cln, 4cpa, 4fisB, 4fxn, 4gpd2, 4lytB, 4mba, 4pad, 4rubB, 4rubV, 5cpy, 5enl, 5rubB, 5sicE, 6at1C, 6cts, 6q21D, 7timB, 8atcD, 8catB, 8rnt, 9aatA, 9abp,

where the first four symbols represent the structure identifier in the PDB database[8] and the last symbol is the protein chain code. Details of the selection process are available from the authors upon request.

Optimization of Recognition Parameters

To determine values for various weights and thresholds in pattern recognition, an exhaustive search was performed over all reasonable values and independently for α-helices and β-sheets. Those that give the best correspondence between our automatic assignment and designations by crystallographers were selected. As a measure of agreement, we used the percent of correctly assigned residues in two states over the entire dataset. Should several combinations of threshold give the same result, those that produce the best correlation coefficient Q_3[30] between our and crystallographers' assignments were adopted. The Q_3 correlation takes into account incorrect as well as correct assignments. The following optimal parametric values were established: $W_1^{\alpha} = W_2^{\alpha} = 1$, $T_1^{\alpha} = 230.0$, $T_3^{\alpha} = 0.06$, $W_1^{\beta} = W_2^{\beta} = 0.2$, $T_1^{\beta} = -240.0$, and $T_2^{\beta} = -310.0$.

3_{10} and π-Helices, Turns, and Solvent Accessibility

Ideally it would be useful to utilize the same rules, based on torsion angle preferences and hydrogen bond energy, for the assignment of other secondary structure elements, in particular 3_{10}-, π-, and left-handed α-helices. However, these structural types are relatively rare and the corresponding observed φ,ϕ statistics very sparse. Further, 3_{10}-helices are much more irregular than α-helices and their torsional angles are rather widely spread on the Ramachandran map.[31] Consequently, 3_{10}- and π-helices were delineated with the general rules of Kabsch and Sander,[4] but the definition for hydrogen bonds was that elaborated by Stickle et al.[46] For turn assignments, the nomenclature and definition proposed by Richardson[21] and extended by Wilmot and Thornton[32] was employed. Residue solvent ex-

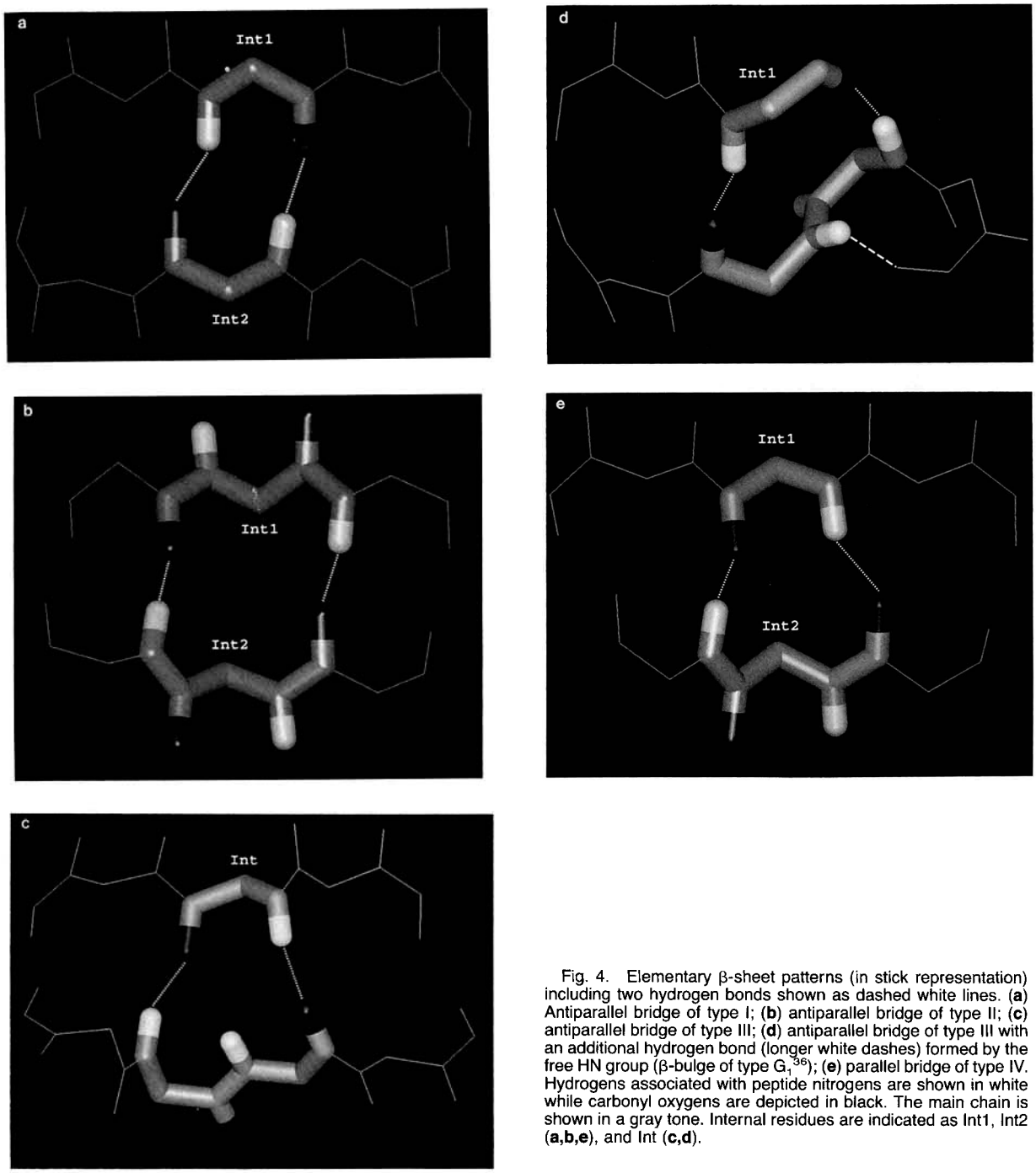


Fig. 4. Elementary β-sheet patterns (in stick representation) including two hydrogen bonds shown as dashed white lines. (**a**) Antiparallel bridge of type I; (**b**) antiparallel bridge of type II; (**c**) antiparallel bridge of type III; (**d**) antiparallel bridge of type III with an additional hydrogen bond (longer white dashes) formed by the free HN group (β-bulge of type G_1[36]); (**e**) parallel bridge of type IV. Hydrogens associated with peptide nitrogens are shown in white while carbonyl oxygens are depicted in black. The main chain is shown in a gray tone. Internal residues are indicated as Int1, Int2 (**a,b,e**), and Int (**c,d**).

posed area was calculated with the improved and fast technique developed by Eisenhaber and colleagues.[33,34]

RESULTS

The accuracy of the method STRIDE relative to the crystallographers' assignments and expressed as percent of correctly assigned residues in two states (α-helix or β-strand and coil) is 94.9% for helices and 92.6% for strands over all amino acids in the X+NMR_30% dataset. The correlation coefficient Q_3[30] which also accounts for over and under assignment gives, respectively, 88.3 and 79.8%.

Since the DSSP algorithm of Kabsch and Sander is undoubtedly the most widely used method for secondary structure assignment from atomic coordinates, we give a detailed account of the differences between our (STRIDE) and DSSP assignments with

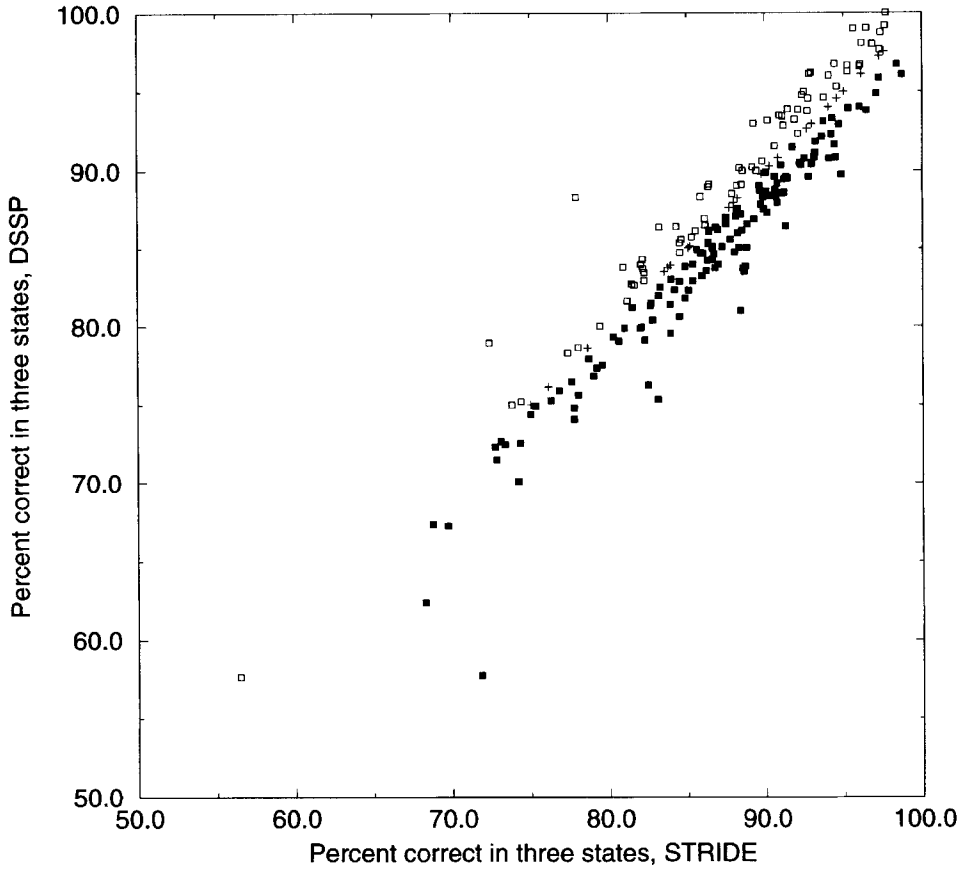


Fig. 5. Comparison between percentages of correctly assigned residues by our method STRIDE and by the DSSP procedure by Kabsch and Sander[4] with respect to the authors' assignments in three states (helix, extended, and coil). Filled and open squares denote, respectively, protein chains where STRIDE performs better and worse than DSSP relative to the designations of the crystallographers. Crosses denote the cases where STRIDE and DSSP yield the same assignments.

respect to those in PDB. As seen from Figure 5, assignments made by STRIDE are in general agreement with DSSP. Though the maximal difference in percent of correctly assigned residues in three states between STRIDE and DSSP does not exceed 14% for individual protein chains, STRIDE yields assignments closer to those given in PDB for nearly twice as many structures as DSSP. This is the case for 58% or 132 of the 226 chains in our data sample, while 11% or 24 were assigned the same by STRIDE and DSSP, leaving 31% or 70 chains where DSSP provided a better assignment. The significant differences between the two assignments become apparent if one excludes from consideration the majority of amino acid residue positions where STRIDE and DSSP agree (Table I). A total of 1223 residues are assigned by STRIDE differently from DSSP in the α-helical class: 716 of them better (true positives and negatives) and 507 worse (false positives and negatives). A true positive is constituted by a residue where STRIDE and the authors assign helix or strand while DSSP disagrees; a true negative is characterized by agreement between STRIDE and the authors in not making a helical or strand assignment whereas DSSP does. False positives and negatives are similarly defined except now assignments made by DSSP and the authors agree with STRIDE in disagreement. STRIDE outperforms DSSP for helical assignments at approximately every 6th (1223/209) residue where they disagree. For strands, STRIDE and DSSP give different assignments in 679 cases, and approximately every 7th residue is assigned by STRIDE closer to the PDB standard-of-truth than does DSSP. Out of 1308 α-helices assigned by crystallographers in the data sample, STRIDE assigns 432 better than DSSP and 301 worse than DSSP. For β-strands, the corresponding counts are 2102, 261, and 195. Thus, STRIDE assigns approximately every 11th helix and every 32nd strand more in register with the authors' assignments than DSSP.

For α-helices, this discrepancy becomes more pronounced if comparisons are performed separately for segments differing in STRIDE and DSSP assign-

TABLE 1. Comparison of STRIDE and DSSP Secondary Structure Assignments for Residue Positions Where the Assignments Disagree*

	True positives					True negatives					False positives					False negatives				
Description	Σ	+	−	~	Fig.	Σ	+	−	~	Fig.	Σ	+	−	~	Fig.	Σ	+	−	~	Fig.
Residues on helix edges	517	396	76	45	6a	35	22	2	11	6b	362	206	105	51	6d	45	21	14	10	6c
Internal helical residues	27	21	1	5	6e	2	1	1	0	—	2	2	0	0	—	8	2	2	4	—
Residues in whole helical segments	23	16	0	7	—	112	108	0	4	6f	4	4	0	0	—	86	46	25	15	6g
Total for helix residues	567					149					368					139				
Strand residues (without G_1 buldges)	282	273	4	5	6h	52	32	15	5	6i	231	208	12	11	6j	54	2	42	10	6k
Strand residues (only G_1 buldges)	55	55	0	0	—	0	0	0	0	—	5	5	0	0	—	0	0	0	0	—
Total for strand residues	337					52					236					54				

*The residues considered are contained in the dataset X + NMR_30% consisting of 226 protein chains (see Methods) for different categories of helical and strand residues. True positives, true negatives, false positives and false negatives are residue positions in which STRIDE, PDB and DSSP give assignments YYN, NNY, YNN and NYY, respectively, where Y denotes a residue assigned in an α-helical or extended state by the respective procedures and N denotes a residue not assigned in one of the two states. The table columns denoted as Σ, +, −, and ~ are respectively the total number of residue cases (Σ), the number of cases where on the basis of visual evaluation we agree with the STRIDE assignment (+), disagree with it in favour of DSSP (−), and cannot judge (~). The figure numbers illustrating appropriate examples for the several categories are given.

ments by less than four consecutive residues (see Table I) and for those with differences longer than 4 residues, the latter corresponding to missing or overpredicted individual helices. Four was chosen as a demarcation since it constitutes the length of a minimal helix (see Methods). In this latter case STRIDE assigns every 8th helix closer to that of the authors' than DSSP. For β-strands, consideration of missed elements is not possible since effectively their minimal length, in comparisons with DSSP, is 1 residue and not 2 as described in the Methods due to the necessity to account for individual β-bridges when comparing STRIDE and DSSP assignments. Very often, for example, when STRIDE finds two consecutive bridges, DSSP finds one. Consequently, for the sake of comparison, symbols "B" denoting individual β-bridges were considered "E" assignments (extended conformation) with the exception of isolated B's where the authors' assignment does not report β-strands.

Detailed comparison of the STRIDE and DSSP assignments for our data set is presented in Table I, including a visual evaluation of the assignment quality. The visual criteria for α-helical residues were similar to those used by Richardson and Richardson[35] who considered the extent to which the α-carbon in a given amino acid residue lies in the cylinder of the helix as well as the compact appearance of the helix. Spacially adjacent pairs of β-strands were required to be in good register and sufficiently parallel to each other. The following tendencies were noted:

- If we exclude residue positions where a visual judgment cannot be made regarding the performance of a given algorithm (columns denoted by ~ in Table I), the total numbers of residue positions where we favor assignments by STRIDE and DSSP are 845 and 226, respectively, for α-helices and for β-strands 575 and 73, respectively.
- At helix edges and in β-strands the differences between STRIDE and DSSP are typically true and false positives, i.e., cases where STRIDE assigns a residue to be in a secondary structural state whereas DSSP does not.
- At helix edges we agree visually with most of the true positives produced by STRIDE. For false positives, we favor in quite a few cases the DSSP assignment. Most of the latter participate in turns which are adjacent to helices and appear to constitute a separate structural entity.
- For missing helical segments with length four or more residues, most of the differences are true and false negatives and, especially for true negatives, we typically favor the STRIDE assignment. Thus, our algorithm is more conservative with respect to short and often irregular helical segments than DSSP.
- In contrast to DSSP, STRIDE assigns residues participating in bulges of type G_1[36] to the extended state (see Fig. 4d) which corresponds well to the authors' assignments and appears visually acceptable.

Many examples of differences between STRIDE and DSSP are presented in Figure 6 to allow the reader assessment of our judgments.

Recent representative studies show that there is a direct link between the structure resolution and its quality. In particular, the deviation of the backbone angles from their standard secondary structural values and distortions of the hydrogen bond geometry become more pronounced in badly resolved structures.[22,27,37] Furthermore, these features are not strongly restrained during structure refinement. It is not surprising, therefore, that our algorithm based on torsional angle and hydrogen bond statistics produces generally worse results on low resolution structures than on high resolution structures, i.e., weaker agreement with the PDB assignments.

We attempted to improve the assignment quality by incorporating in our technique dependence on the

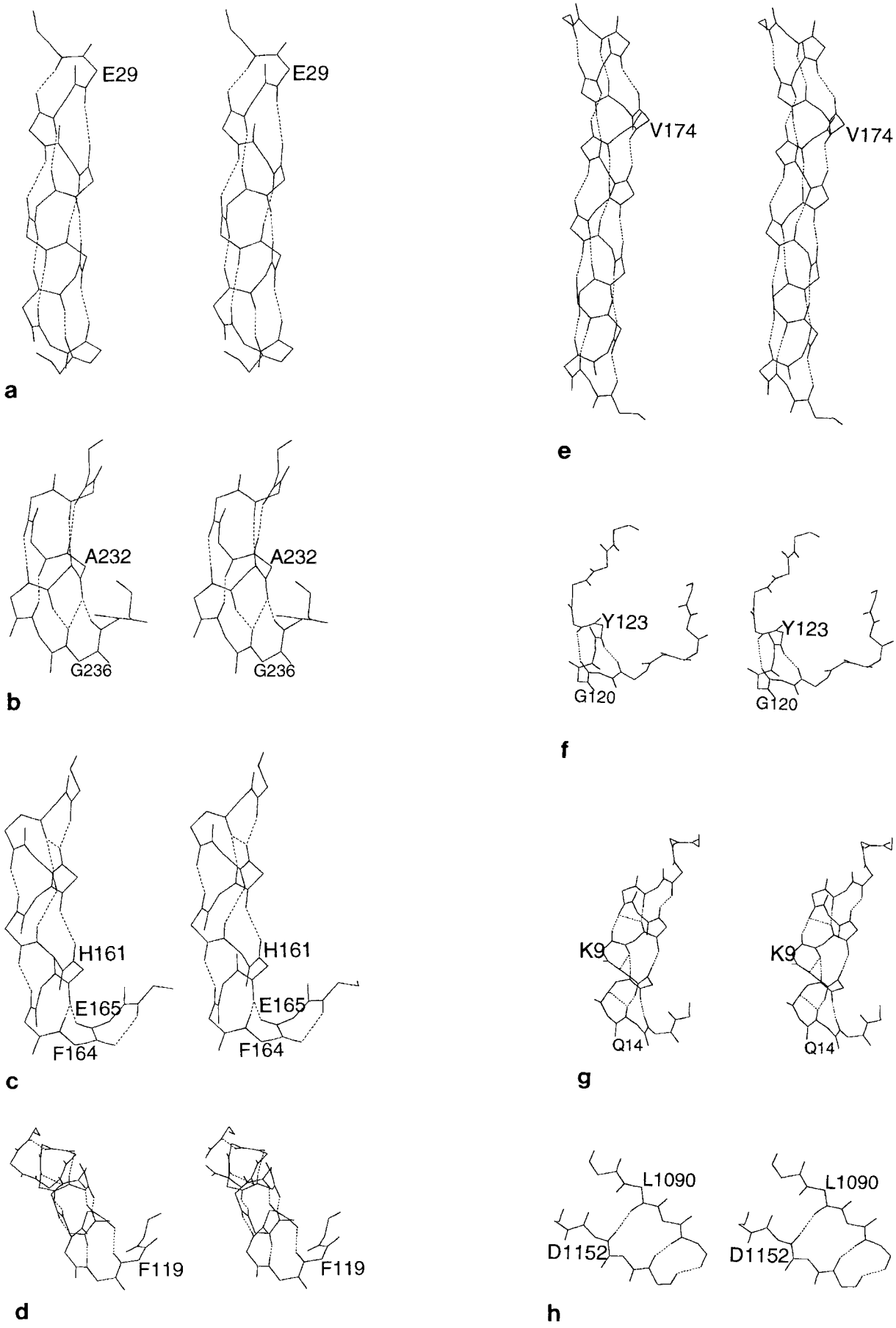
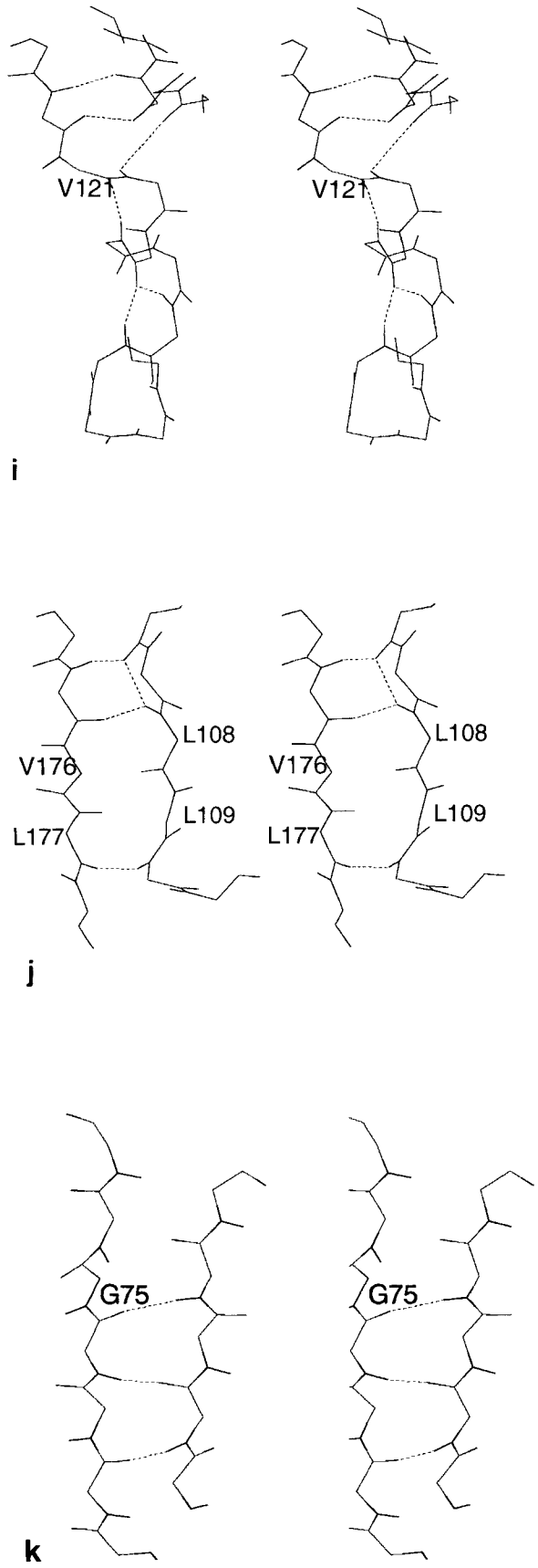

E29
E29
a
A232
A232
G236
G236
b
H161
E165
F164
H161
E165
F164
c
F119
F119
d
V174
V174
e
Y123
Y123
G120
G120
f
K9
K9
Q14
Q14
g
L1090
L1090
D1152
D1152
h

Fig. 6.



resolution. To this end, we derived optimal recognition thresholds separately for the datasets X_HIGH_30% and X_LOW_30% (see Methods). We then recalculated our assignment for the whole X+NMR_30% database such that for structures with resolution less or equal to 2.5 Å, greater than 2.5 Å, and for NMR structures the optimal thresholds derived from the datasets X_HIGH_30%, X_LOW_30%, and X+NMR_30%, respectively, were applied. Only very marginal gain in recogni-

Fig. 6. Examples of differences between STRIDE and DSSP secondary structural assignments. For each example the identification of the residue(s) involved are indicated in the respective captions within square brackets followed by STRIDE, PDB, and DSSP assignments (where "H" stands for α-helix, "E" for extended conformation, "T" for turn, "G" for 3_{10}-helix, and "C" for coil). In the figures, the residue types are indicated in single letter code followed by the PDB sequence position assignment. Our visual judgment is also indicated by +, −, or ~ where we favor STRIDE or DSSP assignments or cannot make a judgment, respectively. See Table I notes for the definition of true and false positives and negatives. For reference, hydrogen bonds are shown in broken lines as defined by DSSP. (**a**) True positive on a helix edge [Glu-29:HHT +]. Glu-29 of atypical homeodomain[53] (1LFB) is assigned as helix by STRIDE since it has acceptable torsional angles. (**b**) True negative on a helix edge [Gly-236:TTH +]. Gly-236 of alcohol dehydrogenase[54] (3HUD, chain B) forms a strong main-chain hydrogen bond with Ala-232 but lacks typical α-helical geometry. (**c**) False negative on a helix edge [Phe-164: CHH +]. Phe-164 of oxidoreductase[55] (4GPD, chain 2) has $\phi = 7°$, rather far from the standard values for α-helix but is included in the helix by DSSP (and by the crystallographers) since the next residue, Glu-165, forms an extremely strong hydrogen bond with residue His-161. STRIDE does not recognize this bond because the conformation of Glu-165 is considered unacceptable ($\varphi = 76°$). (**d**) False positive on a helix edge [Phe-119:HCC −]. Phe-119 of homotetrameric hemoglobin[56] (1TH, chain A) has backbone torsional angles on the very edge of allowed α-helical values ($\varphi° = 125$, $\phi = 30°$) but barely passes the test for T_3^{α} and is erroneously assigned to state "H." The number of such cases should decrease as more and more statistical data are incorporated into the recognition algorithm. (**e**) True positive in the middle of a long helix [Val-174:HHT +]. Val-174 of protein synthesis inhibitor[57] (1PAG, chain B) is part of an internal distortion. (**f**) True negatives in an entire helical segment [residues 120–123:CCH +] in cytochrome *F*[58] (1CTM). (**g**) False negative in entire helical segment [residues 9–14, CHH ~]. An example of a highly distorted helical segment in acyl carrier protein[59] (1ACP) partially missed by STRIDE but assigned as α-helix both by the crystallographers and by DSSP. The C-terminal region of the helix 4–14 is adjacent to the loop 16–36 which, according to Kim and Prestegard, is poorly defined. (**h**) True positive in a strand [Leu-1090: EEC +] of icosahedralvirus capsid protein[60] (1BMV, chain 1). The hydrogen bond between the nitrogen hydrogen of Leu-1090 and carboxyl oxygen of Asp-1152 is weak and not recognized by DSSP but is accepted by STRIDE since the backbone torsion angles of these residues fall into the β-strand region. (**i**) True negative in a strand [Val-121:CCE +]. Val-121 of methyltransferase[61] (1HMY) is assigned as coil by STRIDE since the elementary pattern of type II (hydrogen bonds 120–166 and 122–164) is rejected due to the distorted backbone geometry. (**j**) False positives in a strand [residues 108, 109, 176, and 177:ECC +]. In CD4 protein[62] (1CDI) two consecutive parallel patterns of type IV (residues 110, 109, 108, 177 and residues 108, 175, 176, 177) are accepted by STRIDE but rejected by DSSP and the crystallographers because of the bad quality of the hydrogen bond between the nitrogen hydrogen of Leu-177 and carboxyl oxygen of Leu-108. The general register of the corresponding β-sheet appears well preserved and the interacting strands are parallel to each other. (**k**) False negative in a strand [Gly-75:CEE −]. In the leucine binding protein[63] (2LBP) STRIDE fails to assign Gly-75 to the "E" state due to too strongly distorted geometry. Still the two strands are fairly parallel and therefore we favor the DSSP assignment.

tion was achieved (data not shown). This failure could be attributed to insufficient sample volume since the number of structures with resolution worse than 2.5 Å is rather limited. Also, it is not clear how exactly resolution-dependent stereochemistry translates into secondary structural features. It is interesting to note that we did not find any correspondence between the discrepancies in DSSP and STRIDE assignments and the quality of the structures. Although the quality of assignments made both by DSSP and STRIDE tends to decrease for poorly resolved structures, their relative performance was not affected.

DISCUSSION

The problem of defining the boundaries of secondary structure elements was characterized by Richardson and Richardson[36] as "trivial but difficult." While detection of the major part of α-helices and β-sheets is in fact a trivial task, the precise delineation of secondary structural edges and the correct handling of various experimental errors is challenging and difficult. Correspondingly, only a small fraction of residues in our data sample offers potential for improvement of the assignment quality relative to other methods. This, however, does not diminish the importance of the problem. For many practical purposes, such as development of secondary structure prediction methods or the engineering of protein structures, establishing the exact location of structural elements for training sets is essential.

As a standard-of-truth, we explicitly used the authors' assignments supplied in the PDB files. These assignments can be erroneous or incomplete; nevertheless, the overwhelming majority of the individual residues in the PDB database have been assigned to a secondary structural state on the basis of careful visual inspection and/or application of certain published and objective criteria. Important is that these assignments have been made by different scientists at different times and places and reflect statistically the consensus of hundreds of crystallographers regarding the form and shape of the main secondary structural elements. Many of the obviously erroneous assignments in PDB have been discarded automatically (see Methods); remaining mistakes should be independent from each other and will hopefully compensate each other in statistical tests as they act in opposite directions. In fact, crystallographic assignments were used for verification of automated algorithms (and thus, implicitly utilized as a standard-of-truth) by a number of authors in the past. Some of them give an extensive comparison of their assignments with the reported ones[2] while other authors evaluate the performance of their methods relative to researchers' assignments for just a few selected structures[4–6] or a small random selection from the protein structure databank.[3]

A major assumption of this work is that hydrogen bonding information is not itself sufficient to determine accurately the termini of helices and strands. Many authors have used for this purpose the backbone geometry. Thus, Richardson and Richardson[35] require that the first and last helix residue α-carbons lie within the cylinder defining the helix whereas Dasgupta and Bell[38] define N-cap and C-cap residues of a helix as those that do not possess torsional angles typical of α-helices. In the work of Presta and Rose[13] flanking helix residues are required to participate in $(i,i+4)$ hydrogen bonds and to have appropriate φ,ϕ values. Barlow and Thornton[31] also modify boundaries of DSSP α-helices if they have distorted geometry. Another known problem of the DSSP algorithm is that long helices with missing hydrogen bonds in the middle can be split into two separate helices in spite of the completely acceptable overall geometry (see Fig. 6e for an illustration).

Approaches to secondary structure delineation based on the combined use of hydrogen bonding and torsional angles, although not implemented as a consistent and generally applicable computer algorithm, have often appeared in a variety of studies (e.g., refs. 39, 40). Many simulation studies on helix formation constrain both hydrogen bonds and mainchain dihedral angles to achieve proper helix appearance (e.g., ref. 41). Colloc'h and Cohen[42] investigated the relative contribution of each of 7 different assignment methods to the accuracy of the consensus assignment of β-sheet regions (judged visually) and concluded that backbone torsion angles and hydrogen bonding, in this order, play the most significant roles in strand termination.

Using a product of weighted hydrogen bond energy and torsional terms is of course not the only possible formulation. An extra term added to the expression for hydrogen bond energy E_{hb}, which accounts for the compatibility of the residues participating in a given hydrogen bond with given secondary structural type, provides one main distinction between our method to define α-helices and β-sheets and the DSSP algorithm.[4] A second major difference regards the selection of secondary structural terminal residues through reliance on their torsional angles.

The functional form of the hydrogen bond energy E_{hb} adopted here[23,24] stresses the tendency of hydrogen bonds to be linear and planar and tolerates longer hydrogen bonds if they have otherwise good geometry.[4,9,43–46] The hydrogen bond energy function used for secondary structure definition by Kabsch and Sander[4] and based on electrostatic considerations is similar in spirit but less prohibitive, allowing in certain cases for unrealistic hydrogen bond geometries.

Although four-residue α-helices are in principle possible in our assignment when flanking residues do not satisfy torsional angle criteria, they actually

occur rarely. Many of the elements assigned by the Kabsch and Sander program as a four-residue helix are defined by our method as a turn or a short 3_{10}-helix on the basis of geometric considerations. This eliminates the known drawback of the DSSP algorithm which produces, relative to other assignment methods, a seemingly excessive number of short helices[7] that do not possess typical α-helical appearance. These helices often appear in peripheral loop regions and do not constitute the core secondary structures. Short helices have often been ignored in practical applications (e.g., ref. 47).

A characteristic feature of our algorithm involves different recognition thresholds for different types of secondary structure and different locations within them, including α-helices and their N- and C-terminal residues as well as antiparallel and parallel β-strands. This is in accordance with previous studies where, for example, researchers established different mean values of O...N distances for respective backbone–backbone hydrogen bonds in α-helices and β-sheets,[9] different occurrence statistics of individual amino acids at the ends of helices[35,48] and in parallel and anti-parallel β-strands,[49] and the increased stability of antiparallel over parallel sheets.[50]

It is noteworthy that the optimal values of weights $W_{1,2}^{\alpha}$ for α-helices are much higher than $W_{1,2}^{\beta}$ for β-sheets. This may indicate the β-strands in sheets are in general less sensitive to torsional angle spread than α-helices, in contrast to the conclusion of Colloc'h and Cohen.[42] Hydrogen bonds in β-sheets are known to be somewhat shorter than in α-helices[10] and therefore larger deviations of torsional angles have been tolerated in our definition.

Statistically, STRIDE tends to extend secondary structural elements rather than shrink them relative to the corresponding DSSP assignments (see Table I). This is in accord with the generally known property of DSSP to assign shorter segments than are apparent from visual analysis of the structure (e.g., ref. 51). In particular, α-helices are often extended by STRIDE at the expense of residues that in DSSP assignments appear as turns or 3_{10}-helical residues. Helix edges are often frayed and the flanking residues adopt hydrogen bonding configurations intermediate between α- and 3_{10}-helices.[52] Application of additional restrictions on backbone geometry helps to resolve this conflict in many cases in favor of the α-helical state.[16]

In Table I we have given statistics that show a visual preference for STRIDE secondary structure assignments over those of DSSP. Figure 6 illustrates many structural examples of our judgments. Nonetheless, the improvement of STRIDE over DSSP relative to PDB assignments has been objectively demonstrated, especially since STRIDE outperforms DSSP in nearly 70% of 226 protein folds tested here.

The intrinsic feature of our knowledge-based approach to secondary structural assignment is that further improvement of the recognition quality is possible (and envisaged). This can result from the availability of new protein structures and the consideration of more subtle properties related to secondary structure formation in proteins (such as individual residue preferences, side-chain–main-chain hydrogen bonding, etc.).

AVAILABILITY

The program STRIDE, compiled for most of the common computer platforms together with documentation and example files, is available by anonymous FTP from ftp.ebi.ac.uk (directories /pub/software/unix/stride, /pub/software/dos/stride, /pub/software/vms/stride, /pub/software/mac/stride). Data files with STRIDE secondary structure assignments for the current release of the PDB databank are in the directory /pub/databases/stride of the same site. Atomic coordinate sets can be submitted for secondary structure assignment either to WWW URL http://www.embl_heidelberg.de/stride/stride_info.html or through electronic mail to stride@embl_heidelberg.de. A mail message containing HELP in the first line will be answered with appropriate instructions.

ACKNOWLEDGMENTS

We thank R. Wade for help in implementing the hydrogen bond energy function; R. Abagyan, S. Hubbard and M. Totrov for useful advice; and G. Vogt for friendly assistance. Frank Milpetz implemented the STRIDE World Wide Web server and mail service. Figure 5 was prepared using the program XMGR by Paul Turner.

REFERENCES

1. Eisenhaber, F., Persson, B., Argos, P. Prediction of protein structure. Recognition of primary, secondary, and tertiary features from amino acid sequence. Crit. Rev. Biochem. Mol. Biol. 30:1–94, 1995.
2. Levitt M, Greer, J. Automatic identification of secondary structure in globular proteins. J. Mol. Biol. 114:181–239, 1977.
3. Ramakrishnan, C., Soman, K.V. Identification of secondary structures in globular proteins—a new algorithm. Int. J. Peptide Protein Res. 20:218–237, 1982.
4. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
5. Richards, F.M., Kundrot, C.E. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. Proteins 3:71–84, 1988.
6. Sklenar, H., Etchebest, C., Lavery, R. Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins 6:46–60, 1989.
7. Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., Mornon, J.-P. Comparison of three algorithms for the assignment of secondary structure in proteins: The advantages of a consensus assignment. Protein Eng. 6:377–382, 1993.
8. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based ar-

chival file for macromolecular structures. J. Mol. Biol. 112: 535–542, 1977.
9. Baker, E.N., Hubbard, R.E. Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol. 44:97–179, 1984.
10. Stehle, T., Ahmed, S.A., Claiborne, A., Schulz, G.E. Structure of NADH peroxidase from *Streptococcus faecalis* 10C1 refined at 2.16 Å resolution. J. Mol. Biol. 221:1325–1344, 1991.
11. Fan, Z.-c., Shan, L., Guddat, L.W., He, X.-min., Gray, W.R., Raison, R.L., Edmundson, A.B. Three-dimensional structure of an Fv from a human IgM immunoglobulin. J. Mol. Biol. 228:188–207, 1992.
12. Müller, C.W., Schulz, G.E. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap_5A refined at 1.9 Å resolution. J. Mol. Biol. 224: 159–177, 1992.
13. Presta, L.G., Rose, G.D. Helix signals in proteins. Science 240:1632–1641, 1988.
14. Eigenbrot, C., Randal, M., Presta, L., Carter, P., Kosiakoff, A.A. X-ray structures of the antigen-binding domains from three variants of humanized anti-$p185^{HER2}$ antibody 4D5 and comparison with molecular modeling. J. Mol. Biol. 229:969–995, 1993.
15. Benning, M.M., Wesenberg, G., Caffrey, M.S., Bartsch, R.G., Meyer, T.E., Cusanovich, M.A., Rayment, I., Holden, H.M. Molecular structure of cytochrome c2 isolated from *Rhodobacter capsulatis* determined at 2.5 Å resolution. J. Mol. Biol. 220:673–685, 1991.
16. McPhalen, C.A., Vincent, M.G., Jansonius, J.N. X-ray structure refinement and comparison of three forms of mitochondrial aspartate aminotransferase. J. Mol. Biol. 225: 495–517, 1992.
17. Bolognesi, M., Onesti, S., Gatti, G., Coda, A. *Aplysia limacina* myoglobin. Crystallographic analysis at 1.6 Å resolution. J. Mol. Biol. 205:529–544, 1989.
18. Newman, M., Watson, F., Roychowdhury, P., Jones, H., Badasso, M., Cleasby, A., Wood, S.P., Tickle, I.J., Blundell, T.L. X-ray analyses of aspartic proteinases. V. Structure and refinement at 2.0 Å resolution of the aspartic proteinase from *Mucor pusillus*. J. Mol. Biol. 230:260–283, 1993.
19. Öfner, C., Suck, D. Crystallographic refinement and structure of DNase I at 2 Å resolution. J. Mol. Biol. 192:605–632, 1986.
20. Weiss, M.S., Schultz, G.E. Structure of porin refined at 1.8 Å resolution. J. Mol. Biol. 227:493–509, 1992.
21. Richardson, J.S. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34:167–339, 1981.
22. Morris, A.L., MacArthur, M.W., Hutchinson, E.G., Thornton, J.M. Stereochemical quality of protein structure coordinates. Proteins 12:345–364, 1992.
23. Boobbyer, D.N.A., Goodford, P.J., McWhinnie, P.M., Wade, R. New hydrogen-bond potentials for use in determining energetically favorable binding sites in molecules of known structure. J. Med. Chem. 32:1083–1094, 1989.
24. Wade, R.C., Clark, K.J., Goodford, P.J. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. J. Med. Chem. 36:140–156, 1993.
25. Ramachandran, G.N., Sasisakharan, V.V. Conformation of polypeptides and proteins. Adv. Protein Chem. 23:283–855, 1968.
26. Gibrat, J.-F., Robson, B., Garnier, J. Influence of the local amino acid sequence upon the zones of the torsional angles φ and ϕ adopted by residues in proteins. Biochemistry 30: 1578–1586, 1991.
27. Laskowski, R., Moss, D.S., Thornton, J.M. Main-chain bond length and bond angles in protein structures. J. Mol. Biol. 231:1049–1067, 1993.
28. Jähne, B. "Digitale Bildverarbeitung." Springer-Verlag, 1989.
29. Heringa, J., Sommerfeldt, H., Higgins, D., Argos, P. OBSTRUCT: A program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. Comput. Appl. Biosci. 8:599–600, 1992.
30. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405:442–451, 1975.
31. Barlow, D.J., Thornton, J.M. Helix geometry in proteins. J. Mol. Biol. 201:601–619, 1988.
32. Wilmot, C.M., Thornton, J.M. β-Turns and their distortions: A proposed new nomenclature. Protein Eng. 3:479–493, 1990.
33. Eisenhaber, F., Argos, P. Improved strategy in analytic surface calculation for molecular systems: Handling of singularities and computational efficiency. J. Comput. Chem. 14:1272–1280, 1993.
34. Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., Scharf, M. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J. Comput. Chem. 16:273–284, 1995.
35. Richardson, J.S., Richardson, D.C. Amino acid preferences for specific locations at the ends of alpha helices. Science 240:1648–1652, 1988.
36. Richardson, J.S., Richardson, D.C. Principles and patterns of protein conformation. In: "Prediction of Protein Structure and the Principles of Protein Conformation." Fasman, G.D., ed. New York: Plenum Press, 1989:1–98.
37. McDonald, I.K., Thornton, J.M. Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. 238:777–793, 1994.
38. Dasgupta, S., Bell, J.A. Design of helix ends. In.t J. Peptide Protein Res. 41:499–511, 1993.
39. Edwards, M.S., Sternberg, M.J.E., Thornton, J.M. Structural and sequence patterns in the loops of βαβ units. Protein Eng. 1:173–181, 1987.
40. Harper, E.T., Rose, G.D. Helix stop signals in proteins and peptides: The capping box. Biochemistry 32:7605–7609, 1993.
41. Creamer, T.P., Rose, G.D. α-Helix-forming propensities in peptides and proteins. Proteins 19:85–97, 1994.
42. Colloc'h, N., Cohen, F.E. β-Breakers: An aperiodic secondary structure. J. Mol. Biol. 221:603–613, 1991.
43. Kroon, J., Kanters, J.A. Non-linearity of hydrogen bonds in molecular crystals. Nature (London) 248:667–669, 1974.
44. Pedersen, B. The geometry of hydrogen bonds from donor water molecules. Acta Cryst. B30:289–291, 1974.
45. Artymiuk, P.J., Blake, C.C.F. Refinement of human lysosyme at 1.5 Å resolution. Analysis of non-bonded and hydrogen-bond interactions. J. Mol. Biol. 152:737–762, 1981.
46. Stickle, D.F., Presta, L.G., Dill, K.A., Rose, G.D. Hydrogen bonding in globular proteins. J. Mol. Biol. 226:1143–1159, 1992.
47. Leszczynski, J.F., Rose, G.D. Loops in globular proteins: A novel category of secondary structure. Science 234:849–855, 1986.
48. Chou, P.Y., Fasman, G.D. Prediction of the secondary structure of proteins from their amino acid sequences. Adv. Enzym. 47:45–148, 1978.
49. Lifson, S., Sander, C. Antiparallel and parallel β-strands differ in amino acid residue preferences. Nature (London) 282:109–111, 1979.
50. Chou, K.-C., Pottle, M., Nemethy, G., Ueda, Y., Scheraga, H.A. Structure of β-sheets. Origin of the right-handed twist and of the increased stability of anti-parallel over parallel sheets. J. Mol. Biol. 162:89–112, 1982.
51. Schreuder, H.A., Prick, P.A.J., Wierenga, R.K., Vriend, G., Wilson, K.S., Hol, W.G.J., Drenth, J. Crystal structure of the p-hydroxybenzoate hydrolase-substrate complex refined at 1.9 Å resolution. J. Mol. Biol. 208:679–696, 1989.
52. Bally, R., Delettre, J. Structure and refinement of the oxidized P21 form of uteroglobin at 1.64 Å resolution. J. Mol. Biol. 206:153–170, 1989.
53. Ceska, T.A., Lamers, M., Monaci, P., Nicosia, A., Cortese, R., Suck, D. The X-ray structure of an atypical homeodomain present in the rat liver transcription factor LFB1/HNF1 and implications for DNA binding. EMBO. J. 12: 1805–1810, 1993.
54. Hurley, T.D., Bosron, W.F., Hamilton, J.A., Amzel, L.M. The structure of human β1β1 alcohol dehydrogenase: Catalytic effects of non-active-site substitutions. Proc. Natl. Acad. Sci. U.S.A. 88:8149–8153, 1991.
55. Murthy, M.R.N., Garavito, R.M., Johnson, J.E., Rossmann, M.G. Apo-D-glyceraldehyde-3-phosphate denydrogenase at 3.0 Å resolution. J. Mol. Biol. 138:859–872, 1980.
56. Kolatkar, P.R., Ernst, S.R., Hackert, M.L., Ogata, C.M., Hendrickson, W.A., Merritt, E.A., Phizackerley, R.P. Structure determination and refinement of homotet-

rameric hemoglobin from *Erechis caupo* at 2.5 Å resolution. Acta Cryst. 48B:191–199, 1992.
57. Monzingo, A.F., Collins, E.J., Ernst, S.R., Irvin, J.D., Robertus, J.D. The 2.5 Å structure of pokeweed antiviral protein. J. Mol. Biol. 223:705–715, 1993.
58. Martinez, S.E., Huang, D., Szczepaniak, A., Cramer, W.A., Smith, J.L. Crystal structure of chroloplast cytochrome f reveals a novel cytochrome fold and unexpected heme ligation. Structure 2:95–105, 1994.
59. Kim, Y., Prestegard, J.H. Refinement of the NMR structures for acyl carrier protein with scalar coupling data. Proteins 8:377–385, 1990.
60. Chen, Z., Stauffacher, C., Li, Y., Schmidt, T., Bomu, W., Kamer, G., Shanks, M., Lomonosoff, G., Johnson, J.E. Protein-RNA interactions in an icosahedral virus at 3.0 Å angstroms resolution. Science 245:154–159, 1989.
61. Cheng, X., Kumar, S., Posfai, J., Pflugrath, J.W., Roberts, R.J. Crystal structure of the HHAL DNA methyltransferase complexed with S-adenosyl-L-methionine. Cell 74: 299–307, 1993.
62. Ryu, S.-E., Truneh, A., Sweet, R.W., Hendrickson, W.A. Structures of an HIV and MHC binding fragment from human CD4 as refined in two crystal lattices. Structure 2:59–74, 1994.
63. Sack, J.S., Trakhanov, S.D., Tsigannik, I.H., Quiocho, F.A. Structure of the L-leucine binding protein refined at 2.4 Å resolution and comparison with the LEU/ILE/VAL-binding protein structure. J. Mol. Biol. 206:193–207, 1989.