Knowledge-Based Protein Secondary Structure Assignment

Dmitrij Frishman and Patrick Argos
European Molecular Biology Laboratory, 69062 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 Karplus and Shier (Biopolymers 22:2977–2937, 1983) shows that STRIDE and DSSP assign secondary structural states in 58 and 31% of 236 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 threedimensional 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-Cα distances, analysis of virtual bond angles and lengths between consecutive Cα atoms, analysis of hydrogen bonding patterns, comparison of interatomic distance matrices of structural fragments to idealized reference distance masks typical for a particular secondary structure type, and quantification of the backbone curvature. It is not surprising that techniques utilizing different approaches produce different assignments with disagreements up to 25%. In fact, a detailed examination of 3 procedures by Colleaux 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." Colleaux et al. 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, 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 (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 crystallographers practice 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 these-dimensional structure determination reveals that crystallographers' assignments are based on consideration of hydrogen bonding using the definitions of Baker and Hubbard (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 Przunte and Rose (e.g., ref. 14), hydrogen bonding patterns in combination with main-chain dihedral angles (e.g., refs. 15, 16), mainchain α,β angles only (e.g., ref. 17), the DSSP algorithm with a stricter hydrogen bond definition (e.g., ref. 18), visual criteria (e.g., ref. 19), or a combination of several independent assignment.

© 1995 WILEY-LISS, INC.
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.  

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 Karplus and Shindar  
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.  

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.

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, five-residue turns for α-helices and bridges for β-sheets,  
 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 ψ,φ values to occur in α-helices and β-sheets. Introduction of only one threshold for those 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 $E_{AB}$ 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.  

$$S_{AB} = E_{A} - E_{A} \times E_{B}$$

where $E_{A}$ is the distance dependence of the hydrogen bond, and $E_{A}$ and $E_{B}$ describe its directional properties. The distance term is an 8-6 function:

$$E_{D} = \left( \frac{C}{r^2} - \frac{D}{r^6} \right)$$

where $C = -35$ kcal/mol, $D = -46$, $r$ is the distance between the donor and acceptor atoms participating in the hydrogen bond (see Fig. 1), and $E_{A}$ and $E_{B}$ are the optimal hydrogen bond energy and length, respectively. For main-chain-mainchain hydrogen bonds $E_{A} = -10$ kcal/mol, $E_{B} = -1.8$ kcal/mol, and $r = 3.0$ Å.  

The angle terms $E_{A}$ and $E_{B}$ have the following forms:

$$E_{A} = -C \cos \theta$$

and

$$E_{B} = \left( 0.9 + 0.1 \sin 2\phi \right) \sin \theta$$

where $K_{1} = 0.8$ kcal/mol, $E_{D} = 10$ kcal/mol, and the angles $\theta$ and $\phi$ are respectively angular deviations of the hydrogen atom from the bisector of the bond pair orbital within the plane of the bond pair orbital and from the plane of the bond pair orbital (see Fig. 3). For small separations between the interacting atoms, the distance potential $E_{D}$ 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:

$$E_{A} = E_{B} \quad \text{for } \phi = \pi$$

Torsional Angle Probabilities for α-Helices and β-Sheets

For each 20°-by-20° zone (i) on the Ramachandran map of observed backbone dihedral angles in the many protein structures considered in this work, we calculated the probability $P_{i}$ and $P_{j}$ that the torsional angles for residues assigned in the α-helical or β-sheet state lie within the ith zone, i.e.,
where $N_r$ and $N_p$ are the respective numbers of residues in the given $\alpha$-zone defined in the HELIX and SHEET PDB records as occurring in $\alpha$-helix and $\beta$-sheet and $N_{total}$ is the total number of residues with torsion angles falling within the zone (1) in our data sample. Note that $F_r$ and $P_r$ are set to zero outside of the generally accepted $\alpha$-helical and $\beta$-sheet areas (e.g., refs. 25, 26).

The probability distributions are smoothed using digital binomial filtration. The resulting plots of mean $P_r$ and $P_p$ values versus the dihedral dihedral angular ranges, are respectively, shown in Figure 2a and b.

Recognition of $\alpha$-Helices

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

$$\beta_{k+4}^{h+4} < W_1 + W_2 + \frac{P_k + P_{k+4}}{2} < \frac{\beta_1}{2}$$

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

Recognition of $\beta$-Sheets

A minimal $\beta$-sheet is defined by two consecutive hydrogen bonded $\beta$-bridges belonging to one of the possible types depicted in Figure 4. The quality of a $\beta$-bridge is determined by the strength of both its hydrogen bonds and by the average statistical propensity of its internal residues to be in $\beta$-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 formation of the latter is not taken into account since on the edges of $\beta$-strands abrupt changes of the backbone direction often occur. This leads to values for at least the $\beta$ angles on the N-terminal strand edge and for at least 4 angles on the C-terminal strand edge that lie outside of the $\beta$-sheet zone on the Ramachandran map. Correspondingly, for a $\beta$-bridge to be recognized as such, the two hydrogen bonds involved must satisfy the following conditions (see Fig. 4):

$$\beta_{2}(1 + W_1 + W_2 + \frac{\beta_1^s}{2} + \frac{W_1^s}{2} + \frac{\beta_1^s}{2}) < \frac{W_1^s}{2} + \frac{\beta_1^s}{2}$$

and, for parallel $\beta$-bridges,
Fig. 2. Probabilities $P(a)$ and $P(b)$ for residues in a-$\alpha$-helix and $b$-sheet secondary structural state, respectively, to have different torsional angles $\phi$ and $\psi$. The $b$-sheet probabilities are given for 20-by-20 bins.

\[
\begin{align*}
\beta\text{sheet} & : \phi + \psi < 0 \\
\beta\text{hp} & : \phi + \psi < 0 \\
\beta\text{bridge} & : \phi + \psi < 0
\end{align*}
\]

If internal residues are present on both sides of the bridge (Fig. 4a,b), or $\text{CONF}_\text{bridge}$ is $P_{\text{bridge}}$, if only one residue is internal in a given bridge (Fig. 4c,d). $W_\text{p}$ and $W_\text{q}$ are empirical weights requiring optimization.

Adjacent bridges that fulfill the above criteria are merged into corresponding antiparallel and parallel $b$-sheets with no more than four intervening residues between the bridges on one strand and no more than one residue on another strand. This latter definition for $b$-bulges is the same as that adopted by Kaboch and Sandorfy in DSPP. All residues within the merged adjacent bridges with possible $b$-bulges between them are assigned in an extended state. $K$ with the exception of those bridges flanking the given $b$-sheet where only internal residues are assigned $E$. $K$ was the state in which all $b$-bridges that have suitable neighboring bridges for merging, internal res- idues are assigned the state $B$. An exception is isolated bridges of the state $F$ (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 database. In correspondence with the goals of the present work, included were the protein chains that (1) list only C, atoms, (2) contain no secondary structure assignment made by the au- thor, (3) contain obviously wrong secondary struc- ture assignments (e.g., with long overlapping seg- ments, 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 DSPP by Kaboch and Sandorfy, (5) are not yet published or in print, (6) have less than 70 residues, and (7) rep- resent results of modeling studies.

From the remaining protein structures, three sub- sets were created: (1) X-ray structures at all resolu- tions as well as NMR structures (subset X+NM), (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 non-redundant using the program OBSTRUCTW™ such that no two chains in any set had sequence identity higher than 30%; the resulting non-redundant sets were referred to as X+NM 30%, X_High 30%, and X_Low 30%. Finally, protein chains were excluded where, in the respective articles describing their structural deter- mination, it is explicitly stated that DSPP and, in one case, DEFINER STRUCTURE® algorithms were used for secondary assignment. Thus, we made ev- ery possible effort to exclude from our dataset PDB entries with assignments of secondary structure made with existing automatic methods. Assign- ments made by eye or by manual application of cer- tain 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+NM 30% includes the
following 226 protein chains:

\[ \text{K} \quad \text{K}+4 \quad \text{the cationic oxygen (black) of residue K. The main-chain is shown in a gray tone. In an ideal 3x, the bond is continuously repeated between similarly separated residues.} \]

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 \( \alpha \)-helices and \( \beta \)-sheets. That gives 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 \( Q^2 \) between our and crystallographers' assignments were adopted. The \( Q^2 \) correlation takes into account incorrect as well as correct assignments. The following optimal parametric values were established: \( W_{\alpha}^\alpha = W_{\beta}^\beta = 1, T_{\alpha}^\alpha = 250.0, T_{\beta}^\beta = 0.06, W_{\alpha}^\beta = W_{\beta}^\alpha = 0.5, T_{\alpha}^\beta = -240.0, \) and \( T_{\beta}^\alpha = -310.0. \)

\( \alpha \)-Helices, \( \beta \)-Turns, and Solvent Accessibility

Ideally it would be useful to use the same rules, based on torsion angle preferences and hydrogen bond energy, for the assignment of other secondary structure elements, in particular \( \beta \)-strands, \( \alpha \)-helices, and left-handed \( \alpha \)-helices. However, these structural types are relatively rare and the corresponding observed \( q,q \) statistics very sparse. Further, \( \beta \)-strands are much more irregular than \( \alpha \)-helices and their torsional angles are rather widely spread on the Ramachandran plot. Consequently, \( \beta \)-strands and \( \alpha \)-helices were delineated with the general rules of Kabsch and Sander, but the definition for hydrogen bonds was that elaborated by Bricogne et al. For turn assignments, the nomenclature and definition proposed by Richardson and extended by Wilmot and Thornton was employed. Residue solvent ex-
posed area was calculated with the improved and fast technique developed by Eisenhaber and colleagues.31,32

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-αNMR, 30% dataset. The correlation coefficient $R^2$ which also accounts for over and under assignment gives, respectively, 88.3 and 78.2%.

Since the DSSP algorithm of Kabach 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 Karplus and Shindyalov in respect to the authors' assignments in three states (kinked, coiled, and coil). Figure and open circles denote residues, respectively, where STRIDE performs better and worse than DSSP relative to the designations of the cryocryopographers. 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 112 of the 236 chains in our data sample, while 11% or 24 were assigned the same by STRIDE and DSSP, leaving 31% or 68 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 and residue positions where STRIDE and DSSP agree (Table 1). 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 62.8% (1223/199) residue where they disagree. For strands, STRIDE and DSSP give different assignments in 678 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, 201, and 139. 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

<table>
<thead>
<tr>
<th>Description</th>
<th>True positives</th>
<th>True negatives</th>
<th>False positives</th>
<th>False negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total for helix residues</td>
<td>567</td>
<td>149</td>
<td>368</td>
<td>139</td>
</tr>
<tr>
<td>Total for strand residues (without G, bulges)</td>
<td>1023</td>
<td>254</td>
<td>681</td>
<td>91</td>
</tr>
<tr>
<td>Total for strand residues (only G, bulges)</td>
<td>55</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*The residues considered are contained in the dataset XNMR359 consisting of 236 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, PDBe and DSSP give assignments TY, NY, YNN and NNY, respectively, where Y denotes a residue assigned in an α-helix 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 X, Y, Z, and – are respectively the total number of residue cases (X), the number of cases where on the basis of visual evaluation we agree with the STRIDE assignment (Y), disagree with it in favor of DSSP (Z), and cannot judge (--). The figure numbers illustrating approximate examples for the relevant 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-standards, consideration of missed elements is not possible since effectively their minimal length, in comparisons with DSSP, is 1 residue and not 3 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 β’s where the authors’ assignment does not report β-stands.

Data on the comparison of the STRIDE and DSSP as- signments 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 Rich- ardsen27 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 appear- ance of the helix. Spatially adjacent pairs of β-stands were required to be in good register and sufficiently parallel to each other. The following ten- dencies were noted:

- If we exclude residue positions where a visual judgment cannot be made regarding the perfect manance 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 236, respectively, for α-helices and for β-stands 575 and 73, respectively.
- At helix edges and in β-stands 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 consti- tute 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 nega- tives, we typically favor the STRIDE assignment. Thus, our algorithm is more conservative with re- spect to short and often irregular helical segments than DSSP.
- In contrast to DSSP, STRIDE assigns residues participating in bulges of type G,26 to the extended state (see Fig. 4d) which corresponds well to the authors’ assignments and appears visually accept- able.

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

Recent representational studies show that there is a direct link between the structure resolution and its quality. In particular, the deviation of the backbone angles from the standard secondary structural val- ues and distortions of the hydrogen bond geometry become more pronounced in badly resolved struc- tures.23,27,27 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 statis- tics produces generally worse results on low resolu- tion 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
resolution. To this end, we derived optimal recognition thresholds separately for the datasets X + NMR-30% and X + LOW-30% (see Methods). We then recalculated our assignment for the whole X + NMR-30% database such that for structure X with resolution less or equal to 2.5 Å, greater chains, 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-
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 stereochmistry 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 tend 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 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 obvious 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 while other authors evaluate the performance of their methods relative to researchers' assignments for just a few selected structures or a small random selection from the protein structure database.

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 required that the first and last helix residues α-carbons lie within the cylinder defining the helix whereas Daguupta and Bell 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, flanking helix residues are required to participate in i+1 (i-4) hydrogen bonds and to have appropriate φ, ψ values. Barlow and Thornton also modify boundaries of DSSP α-helices in order to capture 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). Collöbich and Cohen 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 B_H2O, 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. A second major difference regards the selection of secondary structural termini residues through reliance on their torsional angles.

The functional form of the hydrogen bond energy B_H2O adopted here for hydrogen bonds to be linear and planar and tolerate longer hydrogen bonds if they have otherwise good geometry. The hydrogen bond energy function used for secondary structure definition by Kabach and Sander 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 algo-

rithm which produces, relative to other assignment methods, a seemingly excessive number of short helixes that do not possess typical a-helical appear-

ance. These helixes often appear in peripheral loop regions and do not constitute the core secondary structure. Short helixes have often been ignored in protein structure determination.

A characteristic feature of our algorithm involves different recognition thresholds for different types of secondary structure and different locations within them, including a-helices and their N- and C-termini.

al residues as well as anti-parallel and parallel b-strands. This is in accordance with previous stud-

ies where, for example, researchers established differ-

ent mean values of O...N distances for respective b-sheet-b-helix hydrogen bonds in a-helices and b-sheets, different occurrence statistics of indi-

vidual amino acids at the ends of helices and in parallel and anti-parallel b-strands, and the in-

creased stability of antiparallel over parallel sheets.

It is noteworthy that the optimal values of weights W_i j for a-helices are much higher than W_i j for b-sheets. This may indicate that b-strands in sheets are in general less sensitive to torsional angle spread than a-helices, in contrast to the con-

clusion of Collett and Cohen. Hydrogen bonds in b-sheets are known to be somewhat shorter than in a-helices and therefore larger deviations of tor-

sional angles have been tolerated in our definition.

Statistically, STRIDE tends to extend secondary structural elements rather than shrink them rela-

tive to the corresponding DSSP assignments (see Ta-

ble 1). This is in accord with the generally known property of DSSP to assign short segments that are apparent from visual analysis of the structure (e.g., ref. 51). In particular, a-helices are often ex-

tended by STRIDE at the expense of residues that in DSSP assignments appear as turns or b-helical residues. Helix edges are often frayed and the flank-

ing residues adopt hydrogen bonding configurations intermediate between a and b-helices. Applica-

tion of additional restrictions on backbone geometry helps to resolve this conflict in many cases in favor of the a-helical assignment.

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. None-

thless, the improvement of STRIDE over DSSP rel-

ative to PDB assignments has been objectively dem-

onstrated, especially since STRIDE outperforms DSSP in nearly 70% of 226 protein folds tested here.

The intrinsic feature of our knowledge-based ap-

proach 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 con-

sideration of more subtle properties related to sec-

ondary structure formation in proteins (each as an in-


dividual residue preferences, side-chain–main-chain hydrogen bonding, etc.).

AVAILABILITY

The program STRIDE, compiled for most of the common computer platforms together with docu-

mentation and example files, is available by anony-

mous FTP from ftp.ebi.ac.uk/directories/pub/software/uniclstr/pdb/pdbview/pdb/soft-

ware/vms/stride, pub/software/mac/stride. Data files with STRIDE secondary structure assign-

ments 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 sec-

ondary structure assignment either to WWW URL http://

http://www.embl-heidelberg.de/stride/stride_info.html or through e-mail to stride@embl-

heidelberg.de. A mail message containing HELP in the first line will be answered with appropriate in-

structions.

ACKNOWLEDGMENTS

We thank R. Wade for help in implementing the b-hydrogen bond energy function; R. Abagyan, S. Hubbard and M. Totrov for useful advice; and G. Vogt for friendly assistance. Frank Milpetz imple-

mented the STRIDE World Wide Web server and mail service. Figure 5 was prepared using the pro-

gram XMGIR by Paul Turner.

REFERENCES


6. Sklar, H., Eichleib, C., Lavery, R. Describing protein structure: A general algorithm yielding complete helix-


7. Coll, C., Eichleib, C., leurs, K., Hentzen, B., Monot, J.-P. Comparison of these algorithms for the as-

signment of secondary structure in proteins: The advan-

ture of SADH polypeptide from Thermoactinomyces vulgaris D31 re-
10. Cane, E., et al. Structure of the complex be-
tween streptomyces frambuesiae maltoporin and the tri-
13. Busch, C., Sui, D. Crystallographic refinement and struc-
ture of DNA f at a 2.8 Å resolution. J. Mol. Biol. 192:625–
17. Jirasek, J., Wider, G., Peters, J. The structure of the complex between streptomyces frambuesiae maltoporin and the tri-
23. Jirasek, J., Wider, G., Peters, J. The structure of the complex between streptomyces frambuesiae maltoporin and the tri-
29. Jirasek, J., Wider, G., Peters, J. The structure of the complex between streptomyces frambuesiae maltoporin and the tri-
35. Jirasek, J., Wider, G., Peters, J. The structure of the complex between streptomyces frambuesiae maltoporin and the tri-
41. Jirasek, J., Wider, G., Peters, J. The structure of the complex between streptomyces frambuesiae maltoporin and the tri-


