Structure, Vol. 10, 175–184, February, 2002, ©2002 Elsevier Science Ltd. All rights reserved. PII S0969-2126(02)00700-1

Continuum Secondary Structure Captures Protein Flexibility

Claus A.F. Andersen,[1,2] Arthur G. Palmer,[1] Søren Brunak,[2] and Burkhard Rost[1,3,4]
[1] Department of Biochemistry and Molecular Biophysics
Columbia University
New York, New York 10032
[2] Center for Biological Sequence Analysis
BioCentrum
The Technical University of Denmark
DK-2800 Lyngby
Denmark
[3] Columbia University Center for Computational Biology and Bioinformatics (C2B2)
Russ Berrie Pavilion
New York, New York 10032

Summary

The DSSP program assigns protein secondary structure to one of eight states. This discrete assignment cannot describe the continuum of thermal fluctuations. Hence, a continuous assignment is proposed. Technically, the continuum results from averaging over ten discrete DSSP assignments with different hydrogen bond thresholds. The final continuous assignment for a single NMR model successfully reflected the structural variations observed between all NMR models in the ensemble. The structural variations between NMR models were verified to correlate with thermal motion; these variations were captured by the continuous assignments. Because the continuous assignment reproduces the structural variation between many NMR models from one single model, functionally important variation can be extracted from a single X-ray structure. Thus, continuous assignments of secondary structure may affect future protein structure analysis, comparison, and prediction.

Introduction

DSSP Assigns Secondary Structure through Hydrogen Bonds

Pauling and colleagues correctly predicted the idealized protein secondary structures of α helices [2], π helices [2], and β sheets [3] based on intrabackbone hydrogen bonds. Five decades later, we know that, on average, about half of the residues in proteins are located in helices and sheets [1]. Pauling and colleagues incorrectly predicted that 3_{10} helices would not occur in proteins, due to unfavorable bond angles; however, approximately 4% of all residues are observed in this conformation [4]. The DSSP (dictionary of secondary structure of proteins) program developed by Kabsch and Sander [5] assigns secondary structure as described by Pauling and colleagues for three helix types (3_{10}, G; α, H; π, I) and for two extended sheet types (antiparallel and parallel "E"). DSSP restricts helices to segments with at least two consecutive hydrogen bonds flanking the helix, and strands to segments with at least three hydrogen bonds within the same extended sheet. Shorter segments are categorized as turn "T" for helices and β bridge "B" for strands. The remaining two DSSP states are bends "S" and other (dubbed "L" in the following). DSSP allows for considerable deviations from the idealized hydrogen bond pattern in helices and strands; for example, four hydrogen bonds suffice to assign an eight residue α helix (">>44XX44<<" in DSSP) and β bulges of up to four residues are allowed within extended strands. These deviations are captured in DSSP output files, but not in the final discrete assignment of secondary structure states.

Discrete Secondary Structure Assignments Differ

DSSP is the most widely used assignment method. However, other methods have been used to assign secondary structure: based on C_α coordinates (DEFINE) [6], protein curvature (P curve) [7], phi/psi angles (Ramachandran) [8], expert knowledge (crystallographers' assignments in PDB), phi/psi angles and expert assignments (STRIDE) [9], and visual inspection of C_α traces [10]. Assignments from DSSP, DEFINE, and P curve agree for 63% of all residues [11]; DSSP and STRIDE agree for 96% [4]. All these methods use nonphysical thresholds in order to assign discrete secondary structure states.

Molecular Motion of Proteins in Solution Is Captured by NMR

Proteins do not have unique, rigid structures in solution. The degree of flexibility varies significantly between structural segments and at least some conformational fluctuations are essential for function. Recently, the correlation of local conformational variations with protein function has become an important part of experimental structural biology [12, 13]. In particular, NMR studies have emphasized the importance of structural changes over multiple length and time scales as observed, for instance, in calmodulin [14–16]. Protein structure determination by NMR spectroscopy finds many models, the ensemble, that are consistent with experimental constraints. The variations between these models result partially from experimental inconsistencies and incomplete data sets, but they are also believed to result partially from intrinsic fluctuations [17, 18]. NMR spin relaxation measurements are sensitive directly to conformational fluctuations [19]. In particular, the generalized order parameter S^2 describes the equilibrium distribution of bond vector orientations on pico- to nanosecond time scales. For example, $1\text{-}S^2$ is proportional to the variance of the

[4] Correspondence: rost@columbia.edu

Key words: protein secondary structure assignment; evaluation; protein motion; protein structure prediction; protein function; NMR spectroscopy; structure comparison

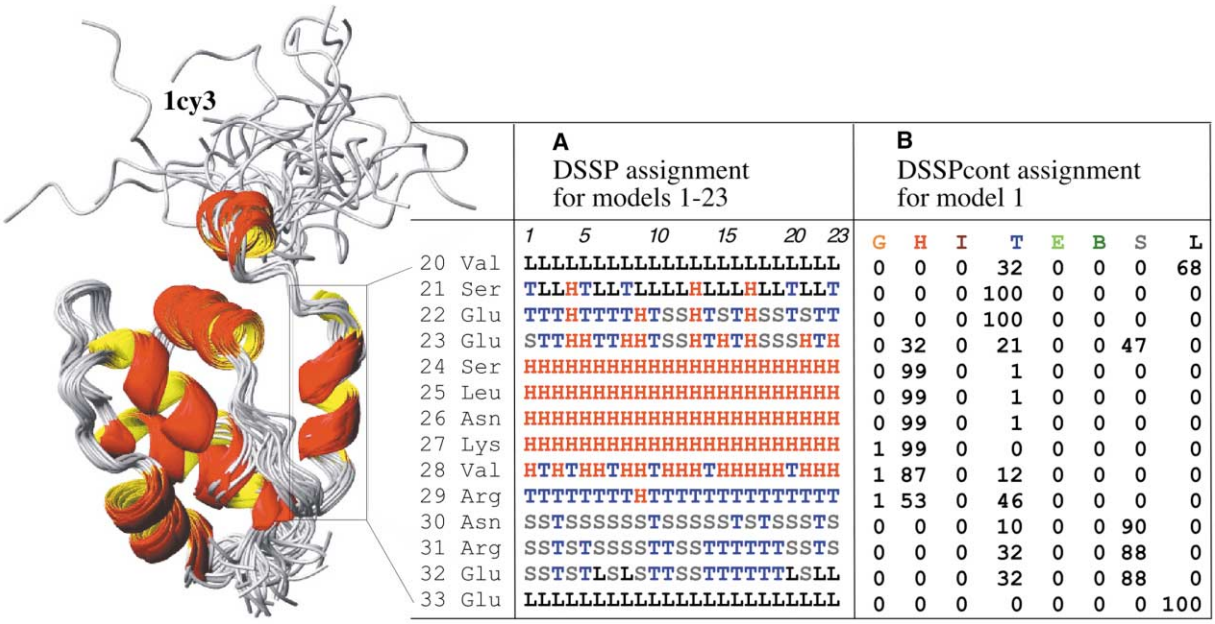

Figure 1. Default DSSP Assignment for 1c3y Fragment

The variations between the secondary structure assignments for different NMR models of the same protein illustrate the impact of fluctuations on structure and highlight the difficulty of predicting protein structure. 1cy3 structure from [59].

(A) The default DSSP [58] assignments for all 23 models of the THP12 carrier protein (PDB: 1c3y [59], residues). The structure models were calculated using $^{13}C/^{15}N$-labeled protein and 3D/4D NMR spectroscopy with 13 NOEs per residue.

(B) DSSPcont assignments for the first NMR model alone; the core of the helix (residues 24–28) are assigned as "H" by default DSSP although the entire α helix switched to a 3_{10} helix when applying a hydrogen bond threshold of -1 kcal/mol. A "fuzzy" helix capping, as seen here, is common and was observed for approximately one in four N caps and half the C caps in our data sets. Dissecting the continuous assignment shows that a 0.1 kcal/mol looser hydrogen bond threshold in the default DSSP would extend the helix by one residue (residue 29). If the default threshold instead had been tightened by 0.2 kcal/mol, the helix would lose one residue (residue 28).

angular distribution for small amplitude conformational fluctuations. Nonetheless, all currently successful secondary structure prediction methods implicitly assume the existence of one rigid protein structure. Typically, developers of structure prediction methods do not use ensembles of NMR structures at all, or use only one representative model.

Assignment Evaluation Based on Consistency

A fundamental question addressed here is how to evaluate and compare assignment schemes. A secondary structure assignment ought to neglect certain details of structures, while retaining others. We argue that a desirable feature of an assignment is consistency, that is, the difference between proteins with the same tertiary structure should be minimized. This means that a "good" secondary structure assignment scheme should minimize the influence of small structural variations due to noise in the experimental determination process and thermal fluctuations. We can therefore evaluate an assignment scheme by comparing assignments within structural families of proteins (different sequences, similar structures) or between different NMR models of a protein (same sequences, similar structures).

Continuum of Secondary Structure Assignment

We introduce a continuous assignment of secondary structure (DSSPcont). In our approach, we chose to rely on NMR models to develop DSSPcont; however, we also investigated structural homologs determined by X-ray crystallography. Both sequence variations and thermally induced conformational fluctuations can result in structural differences between structural homologs. These two effects are indistinguishable for structural homologs. The fuzzy helix capping depicted in Figure 1A illustrates the variability in secondary structure assigned by DSSP for different NMR models of the same protein. The second α helix has a well-defined core (residues 24–28), while the N and C caps of that helix are not well defined (fuzzy). Although strong capping signals have been reported for α helices [10, 20, 21], and β strands [4, 22], such caps are harder to predict than the core [23–26]. The fuzziness of the DSSP cap assignments (Figure 1A) indicates why caps are difficult to predict. Here, we show that the DSSPcont assignment successfully distinguishes between sharp and fuzzy caps. We found that secondary structure assignments varied less between different NMR models for the same protein than between X-ray structures of close homologs. The continuous assignment of secondary structure increases the assignment similarity in both cases. We also show that the variation between NMR models correlates with thermal motion, and that DSSPcont reproduces the variation observed between all models of a protein from the assignment based on a single model. Thus, DSSPcont captures information about thermal motion. The continuous assignment is publicly available (see Experimental Procedures).

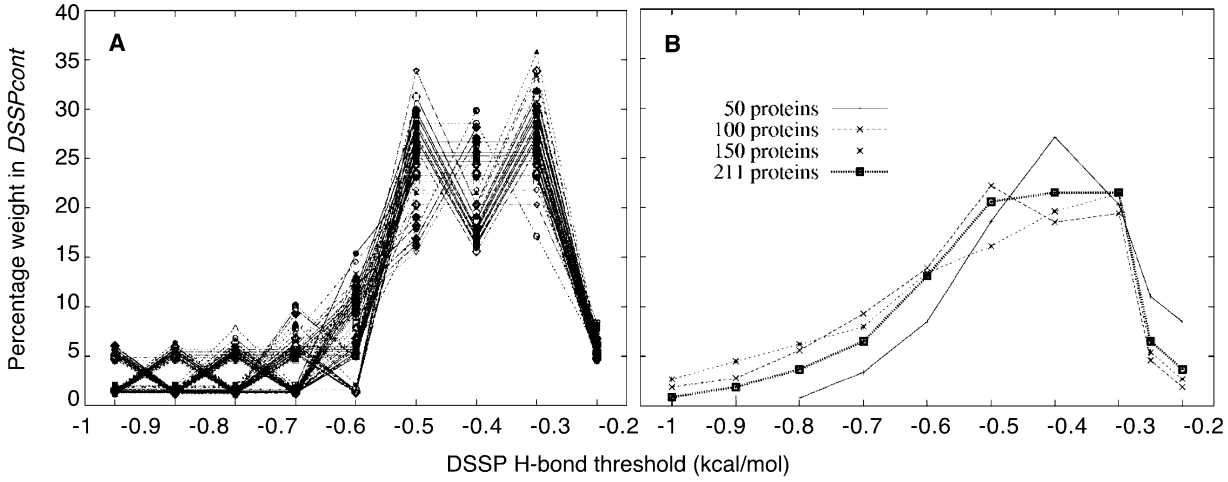

Figure 2. Optimal Weights

The hydrogen bond threshold of each DSSP version used to construct DSSPcont is plotted versus the weight given.

(A) The 100 best-scoring (lowest average difference) weighting schemes of an extensive grid search performed on ten proteins are shown. Their apparent similarity indicates a well-defined global optimum.

(B) We determined the optimal set of weights for DSSPcont using the entire set of 211 proteins through a stepwise gradient descent. Weights optimal for randomly chosen subsets of the 211 proteins were similar; that is, the optimal set of weights was robust.

Results

Continuous Assignment of Secondary Structure

Choosing Weights for the Hydrogen Bond Thresholds

We assigned a continuum of secondary structure by running DSSP with various hydrogen bond thresholds. We weighted the individual DSSP assignments by w^h for a given hydrogen bond threshold h. Thus, we calculated the DSSPcont values for the structural class c from the assigned state $s \in$ [G, H, I, T, E, B, S, L] and residue i:

$$\mathrm{DSSPcont}_{ic} = \sum_{h,s \in c} w^h \cdot \mathrm{DSSP}_i^h(s) \qquad (1)$$

where the discrete $\mathrm{DSSP}_i^h(s)$ assignment can be either 1 or 0, $\Sigma_h w^h \equiv 1$, and for example, three structural classes $c \in$ {GHI,EB,LST}. $\mathrm{DSSPcont}_{ic}$ describes the probability that a given residue i is in class c. To score a given weighting scheme, we used the different models reported in NMR structure ensembles and calculated the average difference between single model assignments and the mean assignment (Equation 2). The best weighting scheme consequently ensured that the assignment extracted as much information as possible from the single NMR model given.

Coarse-Grained Optimum Well Defined

The 100 best weighting schemes were all similar for helix {GHI}, strand {EB}, and other {LST} (Figure 2A). This similarity indicated that the weighting scheme had a well-defined stable global optimum. As expected, the most dominant weights were found close to the default DSSP hydrogen bond threshold of −0.5 kcal/mol. The weight for the −0.2 kcal/mol threshold was consistently low, while the adjacent threshold at −0.3 kcal/mol was consistently high (Figure 2A). This prompted us to insert another threshold at −0.25 kcal/mol.

Fine-Tuning the Weights

We found the optimal set of weights using the entire NMR data set containing 211 proteins by a stepwise gradient descent (Figure 2B). The final average difference over all states with respect to the mean assignment was 0.091. Hence, the DSSPcont assignment for a single model indeed reflected the structural variations between different NMR models of the same protein. Summing the weights w^h with hydrogen bonds ≤ -0.5 kcal/mol contributed 74% of the total weight. Thus, a helix or strand assigned by the DSSP default accounted for at least 74% of the probability in the DSSPcont assignment. 53% of the DSSPcont weight mass originated from hydrogen bond thresholds below −0.5 kcal/mol. Thus, helices or strands ignored by the default DSSP can maximally obtain 53% of the DSSPcont probability. The default DSSP hydrogen bond threshold (−0.5 kcal/mol) occurred near the center of the weighting scheme, with 53% probability weight for the weaker thresholds and 26% for the stronger thresholds. Thus, the conventional DSSP tends to under- rather than to overassign regular secondary structure.

DSSPcont Correlates with Variations between NMR Models

The continuous assignment for 1cy3 appeared to correlate well with variations between the NMR models (Figure 3B). Most strikingly, the transition from α helix to mixed α-helical/turn states observed for residues 23 and 28 in the NMR ensemble was captured by DSSPcont from one model alone (Figure 3B). To further define this correlation, we analyzed a complete database of homologous X-ray structures and NMR structural ensembles.

Properties of the Continuous DSSP Assignment

DSSP States Largely Maintained by DSSPcont

We found that all states were dominated by the original DSSP assignment when the DSSPcont assignments

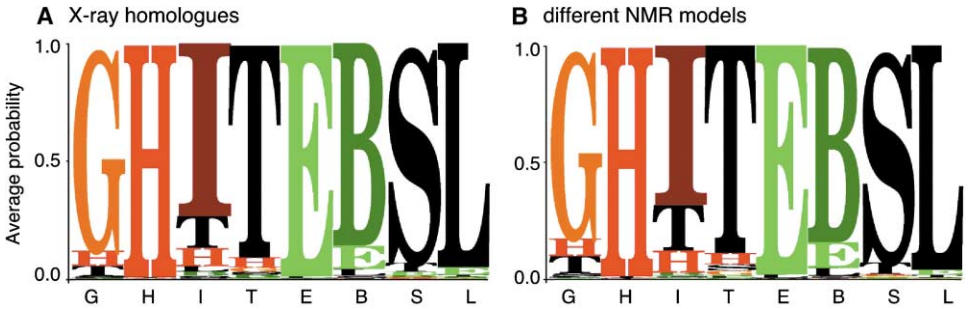

Figure 3. DSSPcont Assignments for DSSP States

For each of the eight default DSSP states (*x* axis), we compiled the average continuous assignment (letter height proportional to average) for both the X-ray (A) and the NMR (B) data sets. The DSSP states generally remained the same. Only the 3_{10} helix and β bridge states were markedly changed, but all changes occurred within their respective class (helix {GHI} and strand {EB}). X-ray and NMR data sets yielded similar results. Note that the low counts of the "I" state result in large variations. G, 3_{10} helix; H, α helix; I, π helix; T, turn; E, extended β strand; B, β bridge; S, bend; L, other/loop.

were mapped to the eight DSSP states (Figure 3). Thus, the exchanges between classes mutually cancelled (Table 1), leaving the average occurrence of each class practically unaltered. β bridges (B; Figure 3) were affected most: 10% of the probability mass was assigned as β strand. Consequently, the default DSSP β bridge often constitutes an ignored β strand that would have been assigned given a lower hydrogen bond threshold. Conversely, the B state received probability mass from the states L, S, and T, resulting in a 15% net increase of the overall probability mass for state B (Table 1).

Consistency Was Higher for DSSPcont than for Default DSSP

We compared the consistency of the default and continuous assignments by measuring the average difference (Adiff; Equation 2) and the assignment rmsd (Armsd; Equation 3). Throughout all states, DSSP appeared less consistent than DSSPcont under the Armsd score (Table 2). For the Adiff score, both assignments appeared similar due to small differences in the overall occurrences (Table 1). Large differences dominate the Armsd score (sum over squares), while many small differences dominate the Adiff score. Hence, the differences between two continuous assignments were common but small.

Flows between Classes Link Secondary Structure States

To compare two continuous assignments, that is, two probability vectors, we introduce the "flow" measure that describes the transformation of one DSSPcont vector into another (Figure 4). The average flow (Aflow matrix; Equation 6) links states that often are assigned differently for the same residue. The average probability flow between the eight states describes the web of links between the eight states (Figure 5). We observed important flows from the helix states (GHI) to turns (T), from turns to the bend state (S), and finally from bends to the loop/other state (L). Because the (GHITS) states all describe a spiral geometry of the backbone, a continuous transition appeared to exist from the helix conformation (GHI), through short helices with few hydrogen bonds (T), to spirals/bends without hydrogen bonds (S), and finally to nonregular conformations (L). This suggested a poor description of the energies involved, because we would expect a disfavored intermediate value [27]. Two effects may be involved: (1) the backbone-backbone hydrogen bonds do not include all energies involved and (2) the Coulomb hydrogen bond expression used in DSSP is too simple [4].

Constraints on Secondary Structure Prediction

About 2% of NMR Models Agree to Less than 80% in Q_3

Assume that all NMR models for one protein are, on average, equally accurate, and that we know only one model. How well can we then predict the secondary structure of all other models? The average Q_{tot} (Equation 5) prediction performance between NMR models using the default DSSP ranged from 93% for three classes (helix [GHI], strand [EB], and other [TSL]), to 89% for six classes (H, [GI], E, B, T, [SL]), to 85% for all eight DSSP states. How many of the inferred predictions were worse than that? One out of four NMR models achieved less than 80% prediction accuracy when comparing all eight DSSP states (Figure 6); one out of nine NMR models achieved 80% in six classes, and one out of 50 achieved less than 80% for three classes (Figure 6). We clearly do not expect to reach the accuracy of NMR experiments by methods predicting secondary structure from sequence.

Table 1. Percentages of Eight Secondary Structure States

	States							
Assignment method	G	H	I	T	E	B	S	L
Default DSSP	3.9	33.5	0.0	11.6	21.0	1.2	8.8	19.9
DSSPcont	3.8	34.0	0.0	11.4	21.6	1.4	8.5	19.3

The average propensities for the eight secondary structure states were compiled on 1534 nonhomologous X-ray and NMR protein chains (values given in percentages). The state propensities remained nearly unchanged between the default DSSP and DSSPcont assignments. The small differences observed show a small flow from the loop and tight helix states (GSL) to the more regular helix and strand structures (HEB).

Table 2. Consistency for Discrete and Continuous Assignments

Score	Set	Method	States							
			G	H	I	T	E	B	S	L
Adiff	X-ray	Default DSSP	2.5	5.7	0.0	4.9	3.5	0.8	4.5	7.4
		DSSPcont	2.5	5.7	0.0	4.8	3.7	0.8	4.3	7.3
	NMR	Default DSSP	1.4	2.7	0.0	5.7	2.0	0.6	6.3	5.8
		DSSPcont	1.3	2.6	0.0	5.7	2.1	0.7	6.2	5.9
Armsd	X-ray	Default DSSP	15.7	23.8	1.2	22.0	18.7	8.6	21.1	27.2
		DSSPcont	13.7	22.4	0.8	19.4	17.4	7.5	19.7	25.7
	NMR	Default DSSP	11.8	16.3	1.8	24.0	14.0	7.8	25.2	24.2
		DSSPcont	9.2	13.1	1.4	19.4	11.3	6.2	22.5	22.1

We compared the consistency of assignments by the average difference (Adiff; Equation 2) and the assignment rmsd (Armsd; Equation 3) to compare the consistency of default DSSP assignment with that of DSSPcont. We used two data sets: (1) X-ray homologs (according to FSSP [35], ZDali $\geq$ 10) and (2) different NMR models for the same protein. The regular secondary structure assignments were significantly more consistent between the NMR models of the same proteins than between the X-ray homologs. While Adiff penalizes many small differences, for Armsd, minor differences are less important (square). Thus, DSSPcont proved considerably more consistent, when giving less importance to minor differences (Armsd). All values are in percentages.

Current secondary structure prediction methods are gradually approaching a sustained mark around 80% [28, 29]. Furthermore, prediction accuracy has exceeded 80% for about 60% of all proteins, and reached 93% for about 4% of all proteins [29]. Hence, the best prediction methods appeared as accurate as the most extreme NMR model for most proteins, and reached the average of all NMR models for about 4% of all proteins.

Correlation Increases through DSSPcont

The accuracy measured by the Pearson correlation coefficient (Equation 4) showed the same trend as Q_{tot}: 2% of all NMR models fall below 0.65 for default DSSP (Figure 6D). This level is reached by today's best prediction methods [28, 29]. The disagreement decreased substantially when using DSSPcont; for example, reducing the percentage of models correlated <0.8 in three states from 11% for DSSP to 5% for DSSPcont (Figure 6D). We conclude that secondary structure prediction methods have reached a level of accuracy at which assignment inconsistencies have become important.

Protein Motion and Secondary Structure

Assignment Evaluation Minimizes Influence of Thermal Fluctuations

We have used the differences between good quality NMR models of the same protein as indicators of thermal fluctuations to evaluate secondary structure assignments. By comparing experimental backbone ^{15}N order parameter data relating the conformational fluctuations of proteins in solution to the C_α rmsd between NMR models, we were able to validate the stipulated correlation (Figure 7). As expected, not all of the variation between structural models reflected thermal fluctuations. For example, the high structural disorder in the C terminus of 1d5v was not reflected in the measured order parameters.

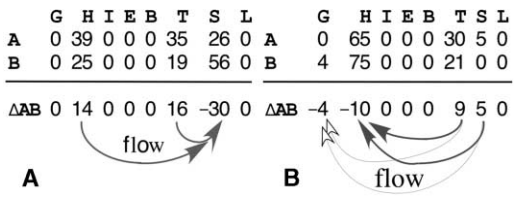

Figure 4. Flow between States

The flow measures the difference between two continuous assignments *A* and *B* by subtracting the vectors (ΔAB). The arrows describe the probability flows necessary to turn *A* into *B*, that is, the flows from positive values to negative values in ΔAB. These are treated as probability flows, as illustrated in (B), where the flow from state T (value 9) to state H (value −10) is relatively larger than that from state T to state G (value −4; see Equation 6). By analyzing the flow between continuous assignments, we can describe the overlap and/or exchange of assignment propensities for individual secondary structure states.

DSSPcont Reveals Protein Motion

The order parameter data enabled a comparison of the DSSPcont assignment with thermal fluctuations (Figure 8). By measuring the average propensity of regular structure (helix [GHI] and strand [EB]) for segment of three consecutive residues, DSSPcont indicated regions with medium to high degrees of motion given a single NMR model. Thus, DSSPcont can suggest regions of the polypeptide subject to conformational disorder from the coordinates of one NMR model or one X-ray structure alone.

Discussion and Conclusion

Continuous Assignment Captures Functional Variations

We have taken the diversity between different NMR models of the same protein at face value and shown how protein structure assignment can profit from this variety (Figure 1). The resulting DSSPcont assignment scheme extends Pauling's hydrogen bond energy-based definition of secondary structure by minimizing the influence of variations due to thermal fluctuations and noise (Table 2). In fact, the DSSPcont assignment correlates with intramolecular thermal fluctuations in solution (Figure 8).

Variation between NMR Models Correlates with Flexibility

We argue that "good" assignments differ only between regions in protein structures that are not conserved between close homologs or different NMR models and that distinguish between regions of thermal motion and less flexible regions. We show that these two objectives are closely related. The assignment consistency be-

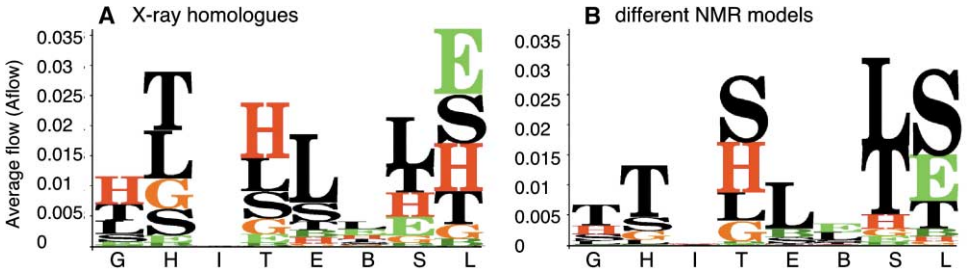

Figure 5. DSSPcont State Transitions

A residue assigned the state *c* by DSSP will typically have some probability for another state *s′* when using a continuous assignment. The other states *{s′}* are characterized by the average flow (Equation 6). For example, the most likely alternative assignment to α helix (H on *x* axis) is T for both X-ray homologs (A) and NMR models ([B]; letter height describes the average flow). This result was obtained because DSSP assigns T for single hydrogen-bonded helices. In particular, we noticed that the X-ray homologs were more inconsistent than the NMR models, as indicated not only by a larger average flow (*y* axis) for α helix (H) and β strand (E), but also because flows between helix and strand are observed only for the X-ray data set.

tween NMR models reflects in part the thermal fluctuations in proteins because the differences between NMR models correlate with independent NMR measurements of intramolecular flexibility (Figure 7). On this premise, we propose a novel assignment scheme (DSSPcont) that minimizes the influence of thermal motion and noise. This assignment had to be continuous because discrete assignments fail to capture thermal fluctuations [30, 31] inherent to protein structure and function. Overall, DSSPcont captures the structural variations between different NMR models of the same protein, as well as between close homologs, based on one NMR model or one X-ray structure (Figure 1; Table 2).

Variations between Homologs Tend to Maintain Assignment Classes

Overall, we observed a "continuous transition" from assignments describing various degrees of spiral backbone geometry (H→G→T→S) to nonregular conformations (Figure 5). FSSP (families of structurally similar proteins) finds homologs by focusing on the overall fold rather than on structural details. This could explain the surprisingly large flows (Figure 4) from helix to strand observed for homologs (Figures 5 and 6). In contrast, thermal fluctuations appeared to be the major contributor to DSSPcont assignments for different NMR models (Figure 8), suggesting that the experimental noise tends to cancel when averaging over many good NMR structures.

Continuous Assignment Affects Structure Analysis, Comparison, and Prediction

Continuous assignment of secondary structure is likely to improve methods that employ secondary structure assignments for structure comparison [32], threading [33], and prediction of conformational switches [34]. One major advantage of continuous assignment is that ragged helix caps and weak strand segments are less likely to be overlooked. Another advantage is that experimentally indiscernible differences are deemphasized. Secondary structure prediction methods are currently reaching within the realm of error inherent to the default DSSP assignment. This fact calls for a rethinking of the assignment scheme. Most importantly, thermal fluctuations are no longer ignored but have become an integral part of the DSSPcont assignments.

Biological Implications

The automatic assignment of protein secondary structure from three-dimensional coordinates of protein structures is an important and, in principle, a simple bioinformatics tool. Assignments are used to visualize structures, speed up expensive computational structural comparisons, and improve sequence searches. Hence, secondary structure assignments are important to assure the optimal yield of experimental structures and to cleverly select the targets for structural genomics. Although a conceptually simple task, the assignment of secondary structure is not always well defined. In fact, assignments vary between different NMR models of the same protein and between X-ray structures of homologs. Here we argue that such differences are not a problem of the assignment scheme, but rather that they carry important information if adequately processed. We show that the variations between different NMR models correlate with thermal disorder. Because the novel continuous assignment of secondary structure (DSSPcont) reproduces the observed variation between high-quality NMR models, it also correlates with mobility related to protein function. Thus, continuous secondary structure assignments can predict conformational variations from a single X-ray structure and thereby may assist predictions of functionally important residues. More generally, it may help to pave the way to automatically generate valid hypotheses from protein structures. Finally, the continuous assignment appears to describe ends of regular secondary structure segments (helices and strand) more accurately than discrete assignments. Often these caps carry important information about function and structure. Hence, the continuum may sharpen the tools that already profit from discrete assignments.

Experimental Procedures

The Continuous Assignment

The DSSPcont assignment was constructed by applying nine hydrogen bond thresholds from −0.2 kcal/mol in steps of 0.1 down to −1 kcal/mol (Equation 1). Testing three values per weight gives $3^9 = 19683$ test rounds and approximately 7 CPU days' testing for ten

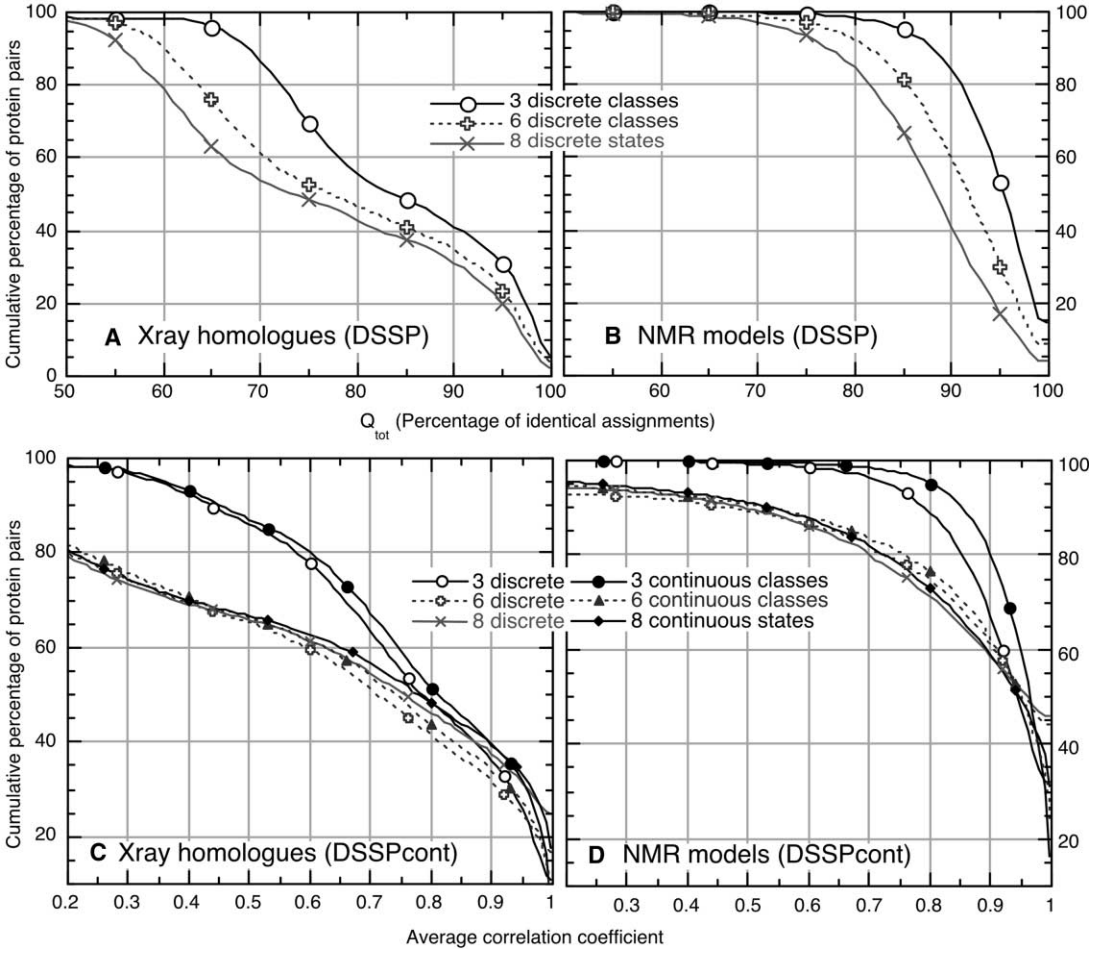

Figure 6. Agreement between NMR Models and Close Homologs

Using one NMR model to predict the default DSSP assignment of another highlights the problem of discrete assignments. The influence of small structural variations between NMR models from the same protein was substantial; only 67% of all models had more than 85% of the residues assigned to the same DSSP state ([B]; eight classes). When grouping the eight states into three classes, 95% of all models had more than 85% identical three-class assignments. Translated to the 20 NMR models usually deposited in PDB, this implies that one or two of them would have a three-class prediction below 85%.

(A and C) The agreement was considerably lower for X-ray structures of homologous proteins. This variation was most likely due to 3D misalignments and sequence-induced structural changes. These results suggest that the assignment problem will dominate more strongly as predictors increase in performance.

(D) The correlation was markedly higher for the continuous DSSP assignment: 11% of the protein pairs had an average correlation of 0.8 or worse using the DSSP assignment, while only 5% for DSSPcont (three classes).

proteins. We sampled the following nonnormalized values $w^h = 1$, 3, 5 $\forall$ h ≥ -0.5 kcal/mol and $w^h = 0.25$, 1, 2 $\forall$ h < -0.5 kcal/mol on ten NMR proteins with a total of 474 models (Figure 2A). To fine-tune the weighting scheme, we performed a simple gradient descent optimization for 50, 100, 150, and 211 proteins (Figure 2B).

Data Set Selection

Selecting Representative Protein Structures

We used representative protein structures according to FSSP (October 6, 2000) [35] in order to maximize the coverage of protein space. The FSSP data set contained 2361 structurally nonhomologous protein chains longer than 30 residues. All structures were downloaded from the Protein Data Bank [1]. We used three data sets to compare assignments: (1) a mixed data set (1534 representative high-resolution X-ray and NMR structures); (2) an X-ray data set with 145 representative high-resolution chains each having at least ten X-ray homologs (ZDali $\geq$ 10) yielding 3245 X-ray homologs; and (3) an NMR data set containing 211 chains from good NMR structures with at least ten models giving a total of 4639 NMR models. The NMR chains selected were either representative chains or substitutes for a representative chain (Z score $\geq$ 10). To ensure good quality X-ray structures, we discarded all structures with resolutions above 2.5 Å and chains having less than 70% of the residues in the most favored region of Ramachandran angles [36].

Missing Quality Criteria for NMR Structures

No well-established criteria exist for evaluating the quality of NMR structures, even with the experimental NMR data at hand. NMR quality assessment methods divide into two categories: self-consistency checks using coordinate data (PROCHECK-NMR [37], WHAT IF [38], and average pairwise rmsd between NMR models) and experimental validation using NMR data (crossvalidation of NMR data [39, 40], Monte Carlo noise simulation [41], NOE restraint violations [37], completeness of NOEs [42], and the number of NOEs per residue). Joint evaluation schemes have been presented [43, 44]. Many of the evaluation methods are available; however, the variety of data formats used to store NMR data are not easily interconvertible. This has left large scale quality assessment prohibitively time consuming and limited to experts [42].

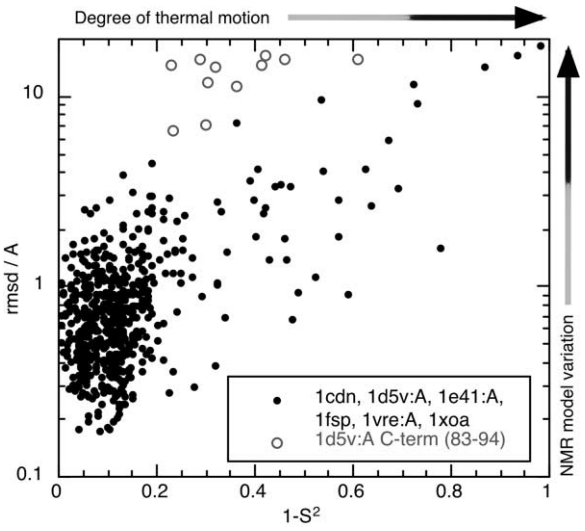

Figure 7. Thermal Fluctuations and Variations between Different NMR Models Correlate

The single-residue rmsd is plotted versus 1-S^2, in which S^2 is the generalized order parameter for backbone ^{15}N spins. This graph relates thermal fluctuations in solution to the differences between NMR models for the same protein. The rmsd difference between NMR models correlates to the conformational disorder measured by the order parameter (CC = 0.69). Note that the semiquantitative correlation between the variation among NMR models and intramolecular thermal motions of proteins in solution justifies our assignment evaluation scheme that measures the consistency between assignments for different NMR models. The dark circles are measurements for the following six proteins (given by their PDB identifiers): 1cdn [46], 1d5v [47], 1e41 [48], 1fsp [49], 1vre [50], and 1xoa [51]. Open circles highlight the notable exception to the observed correlation for the C terminus (residues 83–94) of 1d5v [47]. This region is completely disordered in the ensemble of structures; however, the ^{15}N heteronuclear NOEs are positive and the measured order parameters are high ($S^2 \in [0.4:0.8]$). Thus, the disorder in the ensemble is likely to be dominated by experimental limitations, rather than thermal motion. All data sets were kindly provided by the authors of the respective structures.

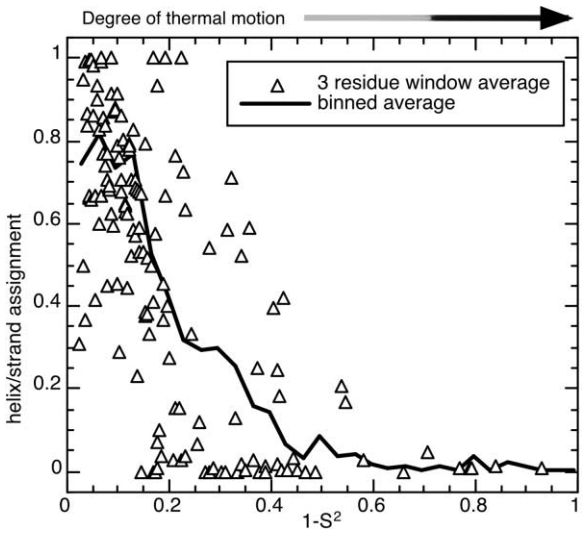

Figure 8. Protein Motion versus DSSPcont

Protein motion measured by the order parameter 1-S^2 is plotted against the DSSPcont assignment-grouping helices (GHI) and strands (EB). The points are averages over three consecutive residues and the line is a binned average of helix/strand assignments. Using one set of coordinates from an ensemble of NMR models, the continuous DSSP assignment reproduces the segments in a protein that experimentally had a high degree of motion due to thermal fluctuations in water. Note: data as in Figure 7.

Performing Large Scale Quality Evaluation of NMR Structures

To select good quality NMR structures, we constructed a ranking scheme based on the experimental methods described in the PDB header, the number of NOEs per residue (when available), the deposition date, and the percentage of residues in the most favored region of Ramachandran angles. The two latter measures have been shown to correlate to the completeness of NOEs [42] and more recently determined structures tend to use more powerful isotope-edited 3D and 4D spectroscopic methods. Our selection classified 39% of all NMR structures considered (≥10 models, ≥30 residues) to be of good quality. Good quality structures as defined by this protocol had an average of 79.3% residues with backbone dihedral angles in the most favored Ramachandran area and 17.3 NOEs per residue.

Order Parameter Data

Figure 7 displays results for seven NMR structures of good quality, for which we could obtain S^2 data: 1b2t[45], 1cdn [46], 1d5v [47], 1e41 [48], 1fsp [49], 1vre [50], and 1xoa [51] (variants of a single protein were not used to avoid overrepresenting a single protein fold). The rmsd was calculated between the heavy backbone atoms in all NMR model pairs, which were 3D aligned with respect to the helix/sheet core (HEcore) segments. The proteins were all selected to have physically reasonable order parameter data [52] satisfying $\langle S^2_{HEcor} \rangle < 0.95$ for an assumed N-H bond length of 1.02 Å. Data were obtained directly from the authors, the BioMagResBank [53], or the Indiana Dynamic database [54].

Programs Used

The DSSP version used was downloaded from CMBI, version as of April 2000 [55]. To calculate the rmsd between protein chains, we used the CE program [56]. Protein structures were visualized with MOLMOL [57].

Evaluation and Comparison

Adiff

Throughout all calculations we have ensured equal weighting of each protein (CC, Q_{tot}) or each residue (Adiff, Armsd, flow), irrespective of the number of NMR models or homologs found for a given structure. This is important because the number of pairs grows quadratically (*M(M-1)/2*) with the total number of models/homologs *M*. Some adaptations and extensions of the formulas given are therefore necessary. The average difference between two assignments *A*, *B* is defined as:

$$\mathrm{Adiff}_c = \frac{1}{N} \cdot \sum_i^N | A_{ic} - B_{ic} | \qquad (2)$$

where *N* is the number of residues, *i* counts over all the amino acids, and *c* is a given class. Adiff is also used as a single number to score an assignment scheme by comparing all NMR models *m* to the average assignment over all models (applying A_{ic} and $B_{ic} = \langle \mathrm{DSSPcont}_{ic} \rangle_m$) and by then summing over the classes.

Armsd

The root-mean-square difference between two assignments *A*, *B* is given by:

$$\mathrm{Armsd}_c = \sqrt{\frac{1}{N} \cdot \sum_i^N (A_{ic} - B_{ic})^2} \qquad (3)$$

CC

The Pearson correlation coefficient is defined as:

$$\mathrm{CC}_c = \frac{\sum_i (A_{ic} - \langle A_c \rangle) \cdot (B_{ic} - \langle B_c \rangle)}{\sqrt{\sum_i (A_{ic} - \langle A_c \rangle)^2 \cdot \sum_i (B_{ic} - \langle B_c \rangle)^2}} \qquad (4)$$

where *c* is the secondary structure class, *i* counts over all residues in a given protein, A_{ic} is the *predicted* value for residue *i* in class *c*

and B_{ic} is the respective assigned value, and $\langle A_c \rangle$, $\langle B_c \rangle$ denote the average values over all residues in all proteins. We defined the average in this particular way since some classes are missing in some proteins (e.g., in all α and all β proteins) yielding $CC = 0$ for these classes although the assignments *A* and *B* may be identical. The average correlation coefficient is then trivially: $\langle CC \rangle = 1/C\ \Sigma CC_c$, where *C* is the number of classes.

Q_{tot}

The total percentage of correct predictions Q_{tot} is:

$$Q_{tot} = \frac{1}{N} \sum_c^C TP_c \tag{5}$$

where *c* counts over all *C* classes, TP_c is the number of true positive predictions in class *c*, and *N* is the total number of residues.

Flow

The link between two continuous assignments *A* and *B* is measured by the flow $flow_{i-j}$ (Figure 4). It describes the probability flow from assignment *A* in state *i* to *B* in state *j*, when *A* and *B* disagree about the probability assigned for the two states:

$$flow_{i-j} = \sum_{\Delta AB_i > 0,\ \Delta AB_j < 0} \Delta AB_i \frac{-\Delta AB_j}{Tflow(AB)} \tag{6}$$

$$Tflow(AB) = \frac{1}{2} \sum_i |\Delta AB_i| \tag{7}$$

$$\Delta AB_i = A_i - B_i \tag{8}$$

where A_i is the probability of state *i* according to assignment *A*. Summing the matrix values in flow yields the total flow (Tflow*(AB)*) that reaches one for nonoverlapping assignments. The flow matrix describes the flows involved when turning the assignment found in *A* into the one found in *B*. Averaging the flow between all pairs *AB* for all residues, we finally get $Aflow_{i-j}$ (Figure 5).

Internet Resource

DSSPcont assignments are provided given a PDB identifier or a PDB file at http://cubic.bioc.columbia.edu/services/DSSPcont and at http://www.cbs.dtu.dk/services/DSSPcont. The program package can be downloaded through the same site.

Acknowledgments

This work was supported by a grant from The Technical University of Denmark (awarded to C.A.F.A.), the Danish National Research Foundation (awarded to S.B.), the National Science Foundation (MCB-9722392 awarded to A.G.P.), and the National Institutes of Health (P506M62413-01 and RO1-GM63029-01 awarded to B.R.). We thank Jinfeng Liu (Columbia University) for technical assistance and system maintenance. We thank the authors of the NMR structures 1b2t [45], 1cdn [46], 1d5v [47], 1e41 [48], 1fsp [49], 1vre [50], and 1xoa [51] for providing S^2 data. We are also grateful to Jürgen F. Doreleijers (BioMagResBank, University of Wisconsin) for comments regarding NMR quality assessment. Last but not least, we thank all those who deposit experimental data in public databases and those who maintain such databases.

Received: July 26, 2001
Revised: December 11, 2001
Accepted: December 13, 2001

References

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.
2. Pauling, L., Corey, R.B., and Branson, H.R. (1951). Two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. USA *37*, 205–211.
3. Pauling, L., and Corey, R.B. (1951). Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc. Natl. Acad. Sci. USA *37*, 729–740.
4. Andersen, C.A. (1998). Neural network assignment of protein secondary structure with increased predictability. Masters thesis, The Technical University of Denmark, Lyngby, Denmark.
5. Kabsch, W., and Sander, C. (1983). How good are predictions of protein secondary structure? FEBS Lett. *155*, 179–182.
6. Richards, F.M., and Kundrot, C.E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins *3*, 71–84.
7. Sklenar, H., Etchebest, C., and Lavery, R. (1989). Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins *6*, 46–60.
8. Ramachandran, G.N., and Sasisekharan, V. (1968). Conformation of polypeptides and proteins. Adv. Protein Chem. *23*, 284–438.
9. Frishman, D., and Argos, P. (1995). Knowledge-based protein secondary structure assignment. Proteins *23*, 566–579.
10. Richardson, J.S., and Richardson, D.C. (1988). Amino acid preference for specific locations at the end of α helices. Science *240*, 1648–1652.
11. Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., and Mornon, J.-P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng. *6*, 377–382.
12. Brunger, A.T., and Laue, E.D. (2000). New approaches to study macromolecular structure and function. Curr. Opin. Struct. Biol. *10*, 557.
13. van Heel, M. (1992). Unveiling ribosomal structures: the final phases. Curr. Opin. Struct. Biol. *10*, 259–264.
14. Lee, A.L., Kinnear, S.A., and Wand, A.J. (2000). Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. Nat. Struct. Biol. *7*, 72–77.
15. Barbato, G., Ikura, M., Kay, L.E., and Pastor, R.W. (2000). Backbone dynamics of calmodulin studied by nitrogen-15 relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. Biochemistry *31*, 5269–5278.
16. Evenas, J., Malmendal, A., and Akke, M. (2001). Dynamics of the transition between open and closed conformations in a calmodulin C-terminal domain mutant. Structure *9*, 185–195.
17. Bonvin, A.M.J.J., and Brunger, A.T. (1996). Do NOE distances contain enough information to assess the relative populations of multi-conformer structures? J. Biomol. NMR *7*, 72–76.
18. Chalaoux, F.R., O'Donoghue, S.I., and Nilges, M. (1999). Molecular dynamics and accuracy of NMR structures: effects of error bounds and data removal. Proteins *34*, 453–463.
19. Palmer, A.G. (2001). NMR probes of molecular dynamics: overview and comparison with other techniques. Annu. Rev. Biophys. Biomol. Struct. *30*, 129–155.
20. Harper, E.T., and Rose, G.D. (1993). Helix stop signals in proteins and peptides: the capping box. Biochemistry *32*, 7605–7609.
21. Aurora, R., and Rose, G.D. (1998). Helix capping. Protein Sci. *7*, 21–38.
22. Colloc'h, N., and Cohen, F.E. (1991). β-breakers: an aperiodic secondary structure. J. Mol. Biol. *221*, 603–613.
23. Brunak, S. (1991). Non-linearities in training sets identified by inspecting the order in which neural networks learn. In Neural Networks from Biology to High Energy Physics, O. Benhar, C. Bosio, P. Del Giudice, and E. Tabet, eds. (Elba, Italy: ETS Editrice Pisa), pp. 277–288.
24. Brunak, S., and Engelbrecht, J. (1996). Protein structure and the sequential structure of mRNA: α-helix and β-sheet signals at the nucleotide level. Proteins *25*, 237–252.
25. Riis, S.K., and Krogh, A. (1996). Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. J. Comput. Biol. *3*, 163–183.
26. Rost, B., and Sander, C. (1994). 1D secondary structure prediction through evolutionary profiles. In Protein Structure by Distance Analysis, H. Bohr and S. Brunak, eds. (IOS Press: Amsterdam, Oxford, Washington), pp. 257–276.
27. Sippl, M.J. (1996). Helmholtz free energy of peptide hydrogen bonds in proteins. J. Mol. Biol. *260*, 644–648.
28. Petersen, T.N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G.P., and Lund, O. (2000). Prediction of protein secondary structure at 80% accuracy. Proteins *41*, 17–20.

29. Rost, B. (2001). Protein secondary structure prediction continues to rise. J. Struct. Biol. *134*, 204–218.
30. Feher, V.A., and Cavanagh, J. (1999). Millisecond-timescale motions contribute to the function of the bacterial response regulator protein Spo0F. Nature *400*, 289–293.
31. Bax, A., and Tjandra, N. (1997). Are proteins even floppier than we thought? Nat. Struct. Biol. *4*, 254–256.
32. Przytycka, T., Aurora, R., and Rose, G.D. (1999). A protein taxonomy based on secondary structure. Nat. Struct. Biol. *6*, 672–682.
33. Rost, B. (1995). TOPITS: threading one-dimensional predictions into three-dimensional structures. In Third International Conference on Intelligent Systems for Molecular Biology, C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak, eds. (Menlo Park, CA: AAAI Press), pp. 314–321.
34. Young, M., Kirshenbaum, K., Dill, K.A., and Highsmith, S. (1999). Predicting conformational switches in proteins. Protein Sci. *8*, 1752–1764.
35. Holm, L., and Sander, C. (1998). Touring protein fold space with Dali/FSSP. Nucleic Acids Res. *26*, 318–321.
36. Morris, A.L., MacArthur, M.W., Hutchinson, E.G., and Thornton, J.M. (1992). Steriochemical quality of protein structure coordinates. Proteins *12*, 345–364.
37. Laskowski, R.A., Rullmann, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR *8*, 477–486.
38. Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. *8*, 52–56.
39. Brünger, A.T., Clore, M.G., Gronenborn, A.M., Saffrich, R., and Nilges, M. (1993). Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. Science *261*, 328–331.
40. Bonvin, A.M.J.J., and Brunger, A.T. (1995). Conformational variability of solution nuclear magnetic resonance structures. J. Mol. Biol. *250*, 80–93.
41. Shriver, J., and Edmondson, S. (1993). Defining the precision with which a protein structure is determined by NMR. Application to motolin. Biochemistry *32*, 1610–1617.
42. Doreleijers, J.F., Raves, M.L., Rullmann, J.A.C., and Kaptein, R. (1999). Completeness of NOEs in protein structures: a statistical analysis of NMR data. J. Biomol. NMR *14*, 123–132.
43. Doreleijers, J.F., Rullmann, J.A.C., and Kaptein, R. (1998). Quality assessment of NMR structures: a statistical survey. J. Mol. Biol. *281*, 149–164.
44. Doreleijers, J.F., Vriend, G., Raves, M.L., and Kaptein, R. (1999). Validation of nuclear magnetic resonance structures of proteins and nucleic acids: hydrogen geometry and nomenclature. Proteins *37*, 404–416.
45. Mizoue, L.S., Bazan, J.F., Johnson, E.C., and Handel, T.M. (1999). Solution structure and dynamics of the CX3C chemokine domain of fractalkine and its interaction with an N-terminal fragment of CX3CR1. Biochemistry *38*, 1402–1414.
46. Akke, M., Forsen, S., and Chazin, W.J. (1995). Solution structure of Cd^{2+}-calbinding D_{9k} reveals details of the stepwise structural changes along the Apo→$(Ca^{2+})^{II}{}_1$→$(Ca^{2+})^{II,I}{}_2$ binding pathway. J. Mol. Biol. *252*, 102–121.
47. van Dongen, M.J.P., Cederberg, A., Carlsson, P., Enerback, S., and Wikstrom, M. (2000). Solution structure and dynamics of the DNA-binding domain of the adipocyte-transcription factor FREAC-11. J. Mol. Biol. *296*, 351–359.
48. Berglund, H., Olerenshaw, D., Sankar, A., Federwisch, M., McDondald, H.Q., and Driscoll, P.C. (2000). The three-dimensional solution structure and dynamic properties of the human FADD death domain. J. Mol. Biol. *302*, 171–188.
49. Feher, V.A., Zapf, J.W., Hoch, J.A., Whiteley, J.M., McIntosh, L.P., Rance, M., Skelton, N.J., Dahlquist, F.W., and Cavanagh, J. (1997). High-resolution NMR structure and backbone dynamics of the *Bacillus subtilis* response regulator, Spo0F: implications for phosphorylation and molecular recognition. Biochemistry *36*, 10015–10025.
50. Volkman, B.F., Alam, S.L., Satterlee, J.D., and Markley, J.L. (1998). Solution structure and backbone dynamics of component IV *Glycera dibranchiata* monomeric hemoglobin-CO. Biochemistry *37*, 10906–10919.
51. Jeng, M.F., Campbell, A.P., Begley, T., Holmgren, A., Case, D.A., Wright, P.E., and Dyson, H.J. (1994). High-resolution solution structures of oxidized and reduced *Escherichia coli* thioredoxin. Structure *2*, 853–868.
52. Case, D.A. (1999). Calculations of NMR dipolar coupling strengths in model peptides. J. Biomol. NMR *15*, 95–102.
53. Goodman, J.L., Pagel, M.D., and Stone, M.J. (2000). Relationships between protein structure and dynamics from a database of NMR-derived backbone order parameters. J. Mol. Biol. *295*, 963–978.
54. Seavey, B.R., Farr, E.A., Westler, W.M., and Markley, J.L. (1991). A relational database for sequence-specific protein NMR data. J. Biomol. NMR *1*, 217–236.
55. Vriend, G., and Krieger, E. (2000). Centre for Molecular and Biomolecular Informatics CMBI version of DSSP, www.cmbi.kun.nl/gv/dssp/.
56. Shindyalov, I.N., and Bourne, P.E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. *11*, 739–747.
57. Koradi, R., Billeter, M., and Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. J. Mol. Graph. *14*, 51–55.
58. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers *22*, 2577–2637.
59. Rothemund, S., Liou, Y.C., Krause, E., and Sonnichsen, F.D. (1999). A new class of hexahelical insect proteins revealed as putative carriers of small hydrophobic ligands. Structure Fold Des. *7*, 1325–1332.