Dihedral-Angle Information Entropy as a Gauge of Secondary Structure Propensity

Shi Zhong, Jeremy M. Moix, Stephen Quirk, and Rigoberto Hernandez

Center for Computational and Molecular Science and Technology, School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332-0400

ABSTRACT Protein structural information can be uncovered using an information-theory-based entropy and auxiliary functions by taking advantage of high-quality correlation plots between the dihedral angles around a residue and those between sequential residues. A standard information entropy for a primary sequence has been defined using the values of the probabilities of the most likely dihedral angles along the sequence. The distribution of entropy differences relative to the standard for each protein in a reference set—a sublibrary of the Protein Data Bank at the 90% sequence redundancy level—appears to be nearly Gaussian. It gives rise to an auxiliary checking function whose value signals the extent to which the dihedral angle propensities differ from typical structures. Such deviations can arise either because of incorrect dihedral angle assignments or secondary structural propensities that are atypical of the structures in the reference set. This auxiliary checking function can be readily calculated at the public website, http://www.d2check.gatech.edu. Its utility is demonstrated here in an analysis displaying differences between experimentally and theoretically derived structures, and in the analysis of structures derived by homology modeling. A comparison of the new measure, D_2Check, to other checking functions based on backbone conformation—namely, PROCHECK and WHAT_CHECK—is also provided.

INTRODUCTION

The number of structures in the Protein Data Bank (PDB) (1) has increased dramatically during the past decade. More than 28,000 structures had been deposited as of October 2004 when the results were first collected for this study and the count stands at a little more than 37,000 as of July 2006. The accuracy of any new structure is of obvious importance because any error makes predictive methods more difficult to validate and creates problems for structural data-mining efforts (2–4). As the various computational methods mature, it becomes increasingly important to derive a varied set of scores or checking functions that assess and validate protein structures. Existing and new checking functions are also vital in the area of de novo structure prediction for validation. The Ramachandran plot (5) has provided a useful framework for discerning patterns in the dihedral angle correlations and has been successfully used as a guide during structure refinement. However, it is the work of Thornton and co-workers (6–10) that pioneered the field of structure validation (11–18) using scores based upon known statistical properties of the existing database. Although such checks are not foolproof as they rely on the working hypothesis that a new structure will interpolate within the known database, they at least raise the question of whether a new structure is atypical or is merely extrapolating outside of the database. In particular, Thornton and co-workers have proposed simple and effective ways to test the stereochemical quality of a proposed structure. Such approaches, based only on coordinates rather than on free energies or dynamical considerations, are easy to code and quick to process while still having significant merit, especially when used in conjunction with new measures.

In this study, an information-theory entropy is proposed based on the backbone dihedral angle distributions of the protein structure. It underlies an auxiliary robust checking function for evaluating the compatibility of a given protein structure with the experimentally derived structures in the PDB with respect to its dihedral angles. The 20 Ramachandran plots—i.e., ϕ_i-ψ_i distributions—for each of the naturally occurring amino acids are reconstructed using all of the nonredundant experimental protein structures available in the October 2004 PDB using a 90% sequence identity cutoff. In addition, the 400 ψ_i-ϕ_{i+1} distributions accounting for the statistics in the two dihedral angles between specified adjacent amino acids have also been constructed and are presented. The latter distributions have been seen to contain nontrivial structure and the present results—over the existing larger database—serve to validate prior conclusions (19–22). The information-theory entropy, S, is defined in terms of the probabilities (or likelihood) of particular pairs of dihedral angles along the protein given its primary structure. A standard entropy is defined using an ideal (but likely unattainable) structure in which every angle pair, ϕ_i-ψ_i and ψ_i-ϕ_{i+1}, takes on the value with maximum probability, where the index i labels a residue along a chain. The entropy difference, ΔS, is defined relative to the standard entropy of this structure, and has been calculated for all nonredundant protein structures in the PDB. A histogram of these entropy

Submitted May 16, 2006, and accepted for publication August 29, 2006.

Address reprint requests to Rigoberto Hernandez, Fax: 404-894-0594; E-mail: hernandez@chemistry.gatech.edu.

Stephen Quirk's present address is Kimberly-Clark Corporation, Atlanta, GA 30076-2199.

0006-3495/06/12/4014/10 $2.00

doi: 10.1529/biophysj.106.089243

differences leads to a nontrivial distribution. As a simple test of whether such a distribution is sensitive to differences between the theoretically and experimentally generated structures in the PDB, this distribution has been obtained for each cohort. The deviations in these distributions will be seen to emerge primarily from those theoretical structures that have been obtained using statistical information that ignores long-range correlation due to, for example, secondary structural elements.

Furthermore, the distribution in ΔS can be used to define auxiliary checking functions, herein called D_1 and D_2, which characterize the degree to which the dihedral angles of a given structure are compatible with the existing database (23). The ΔS distribution is peaked at a nonzero value because a typical structure contains a certain degree of correlation between distant residues due to secondary structural interactions. The use of the statistical distributions in the calculation of ΔS implies that this information is included in an averaged, or mean-field-like, sense. Thus D_2 can signal the existence of atypical structures whose unusual behavior is due to specific interactions between distant residues. Of course, deviations may also be due to incorrectly obtained structures, though such a determination is not available simply from the knowledge of D_2. It therefore complements the scores available in PROCHECK (6,7) and WHAT_CHECK (3) in that it includes the ψ_i-ϕ_{i+1} correlations, and it provides a simple check of the deviation from non-mean-field-like structure. Hence this measure can be used to guide modeling studies and to validate experimentally derived structures, while bolstering the tools that are available to guide the formation of de novo and engineered protein structures. In fact, D_2 provides an information-rich tool to guide experiments involving the replacement or redesign of large sections of protein structure (e.g., loop modeling). These new measures also complement the work of Shortle and coworkers (24–27), who focus on the propensities of a given residue's dihedral angles due to the nearby structure (through an energy-based scoring function) rather than on the mutual probability of given residue pairs. These subtle distinctions give rise to differences in the information that the respective checking functions or scores report. Thus the central result of this work is the construction of a new checking function D_2 that complements the existing checking functions by reporting on the extent to which the propensity of the dihedral angle deviations differ in a given protein from those of the reference database.

METHODS

ϕ_i-ψ_i and ψ_i-ϕ_{i+1} distributions

Dihedral angle analysis (28–31) of protein backbones is helpful in structure validation and modeling (6–11,13–18,32–40). Conventional Ramachandran plots $P_R(\phi_i, \psi_i)$ characterize the probability distribution for angles ϕ_i and ψ_i for each R of the 20 natural amino acids, where the two dihedral angles are defined by the backbone atom sequences, C(i – 1)-N(i)-CA(i)-C(i) and N(i)-CA(i)-C(i)-N(i + 1), respectively, as shown in Fig. 1. An extensive analysis of the Ramachandran plots using a fairly recent edition of the PDB has been reported by Hovmöller et al. (28).

Although useful, the information contained in a Ramachandran plot is not sufficient to construct a scoring function for high-accuracy protein structure validation. For example, flanking residues are known to affect the probability distribution in the dihedral angles of a given residue (24–27,41–47). As previously suggested, one defines the $P_{R_i,R_{i+1}}(\psi_i, \phi_{i+1})$ distributions—in which the angles are associated with the sequential residues—to complement the information in the Ramachandran plot (19–22). Since the ψ_i-ϕ_{i+1} plot accounts for the correlation between two adjacent residues, its use in structure assessment provides a nontrivial sequence-dependent measure of the likelihood that a given pair of residues will be connected by the specified dihedral angles. In principle, one could also account for the explicit correlations present between additional structural observables such as in the recent study by Esposito et al. (48) on the correlation between ψ and the angle ω describing the rotation of the peptide bond. However, only the correlation between ϕ and ψ around a residue and between bonded residues will be addressed, because, as shown below, this suffices to provide a different first-order estimate of protein structure than other scores presently available.

Data-mining the ψ_i-ϕ_{i+1} distributions

To obtain the 400 possible ψ_i-ϕ_{i+1} distributions labeled by each of the pairs of naturally occurring amino acids, a statistically representative sample of all possible proteins needs to be available. In this work (as with other similar studies), the sublibrary of deposited structures in the PDB are assumed to be representative of the protein space once it has been systematically pruned: DNA, RNA and complexes of proteins with DNA or RNA are removed. Model structures are discarded because of the unknown possibility that such theoretically derived structures may be of a different level of accuracy or representation. Additionally, structures with missing residues or containing unified atoms have been removed. (Although more aggressive pruning could have been done by discarding structures according to a more rigorous standard for its resolution, this was not done in this investigation.) After pruning the PDB subject to these criteria, the resulting library (called ''EXP'' throughout this work) includes a total of 24,444 experimentally derived structures.

The NR50, NR70, and NR90 sublibraries result from the intersection of the EXP library of October 2004 PDB structures with the nonredundant sequence databases posted in the PDB—as listed in the April 2005 update—at the 50%, 70%, and 90% sequence identity levels, respectively (49). The NR100 sublibrary is a subset of the EXP library in which a single arbitrarily-chosen structure is retained for each redundant sequence at 100% sequence identity. Note that, by definition, no two structures in a given database share a sequence identity greater than or equal to that of the database's defining percentage level. Hence, for example, the NR100 sublibrary will be smaller than the EXP library as the former includes only one structure for a given sequence. The subset, NR100T, of theoretically derived—that is, model—protein structures in the PDB at 100% sequence identity will also be investigated for confirmation of the relative level of

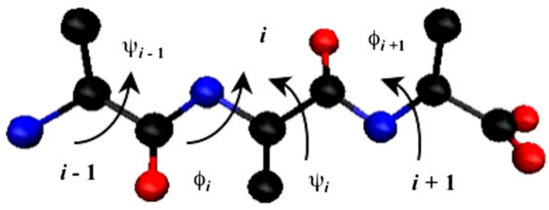

FIGURE 1 The backbone dihedral angles in a tripeptide ALA-ALA-ALA. (*Blue*, nitrogen; *black*, carbon; *red*, oxygen.)

information contained therein. The number of structures in each library is shown in Table 1.

All 400 ψ_i-ϕ_{i+1} and 20 Ramachandran plots have been generated for each of the five sublibraries, NR50, NR70, NR90, NR100, and EXP. Their construction is described explicitly in Supplement *A* in the Supplementary Material, and the results for the NR90 sublibrary are provided in Supplement *B* in the Supplementary Material. Typical one-dimensional distributions of the projections of the ϕ_i-ψ_i Ramachandran plots and the ψ_i-ϕ_{i+1} plots are displayed in Fig. 2 (for the procedure, see Supplement *A*, Supplementary Material). These results demonstrate the sequence dependence of the ψ_i-ϕ_{i+1} distribution, in accordance with the previous reports (19,21,50). Importantly, the dependence of ψ_i on the second residue and ϕ_{i+1} on the first residue obviously illustrates the impact of the distant residue identity on the absolute value of the maximum probability. The effects on glycine are particularly pronounced as the peak position of the distribution changes with the distant residue identity (Fig. 2 *b*). The torsion angles were extracted using a tool kit written in FORTRAN and verified within our group (S. Zhong and R. Hernandez, 2005. SiFiScore Toolkit, unpublished code). The 420 histogrammed distributions for NR90 have been saved into a single database which can, in turn, be used to calculate the dihedral-angle information entropy difference, ΔS, defined in Eq. 6 below.

The dihedral-angle information entropy

Given a proposed protein structure for a particular primary sequence, and the distribution functions described above, one can calculate an information-theory-based entropy for the angle pairs around and between the residues of the chain. In particular, for a given structure $\vec{q}$, the dihedral angle pairs across its n residues consist of the $(n-2)$ ϕ_i-ψ_i pairs and associated probabilities $P_{R_i}(\phi_i, \psi_i)$ at each site i for i ranging across 2 and $n-1$. Similarly, $\vec{q}$ gives rise to the $(n-1)$ ψ_i-ϕ_{i+1} pairs and associated probabilities $P_{R_i,R_{i+1}}(\psi_i, \phi_{i+1})$ between successive residues at i and $i+1$ for i ranging across 1 and $n-1$. For convenience, these two sets are interlaced into a single vector $\vec{\Upsilon}$ whose $2n-3$ entries are defined as

$$\Upsilon_{2i-1} \equiv (\psi_i, \phi_{i+1}) \quad \text{for} \quad 1 \leq i \leq n-1, \tag{1a}$$

$$\Upsilon_{2i} \equiv (\phi_{i+1}, \psi_{i+1}) \quad \text{for} \quad 1 \leq i \leq n-2. \tag{1b}$$

A Shannon entropy rooted in information theory (51) can now be rewritten as

$$S(\vec{q}) = -\sum_{k=1}^{2n-3} P_{\xi_k}(\Upsilon_k(\vec{q}))\ln P_{\xi_k}(\Upsilon_k(\vec{q})), \tag{2}$$

where the argument in $\vec{\Upsilon}$ specifies the angles according to the particular structure $\vec{q}$, and the residues are paired according to

$$\xi_{2i-1} \equiv (R_i, R_{i+1}) \quad \text{for} \quad 1 \leq i \leq n-1, \tag{3a}$$

$$\xi_{2i} \equiv R_{i+1} \quad \text{for} \quad 1 \leq i \leq n-2, \tag{3b}$$

corresponding to the structure of $\vec{\Upsilon}$. A standard information entropy for a given structure can be defined in terms of the most probable dihedral angles for a given primary sequence,

$$S^{\circ}(\vec{q}) = -\sum_{k=1}^{2n-3} \bar{P}_{\xi_k}\ln\bar{P}_{\xi_k}, \tag{4}$$

in which the maximal values are defined simply as

$$\bar{P}_{\xi_k}(\vec{q}) \equiv \max_{\Upsilon_k} P_{\xi_{k(\vec{q})}}(\Upsilon_k), \tag{5}$$

and depend on $\vec{q}$ only with respect to the specification of its primary sequence, $\vec{\xi}(\vec{q})$. The averaged entropy difference for a given structure relative to the standard can be written simply as

$$\Delta S(\vec{q}) = (S^{\circ}(\vec{q}) - S(\vec{q}))/(2n-3), \tag{6}$$

where $(2n-3)$ is the normalization factor.

Solis and Rackovsky (52,53) defined a similar information entropy to that of Eq. 2 for protein structure prediction. However, none of their measures emphasized the use of the ψ_i-ϕ_{i+1} distributions, and the possible correlation between neighboring amino acids that such distributions may display. Meanwhile, the GOR algorithm (54,55) uses the statistics of the multiple sequence alignment of segments of 17 or more residues in length to predict secondary structure assignments. The approach in this article is complementary to the GOR algorithm in that both recognize the need for studying multiple residue correlations: the latter emphasizes a larger segment while limiting the number of possibilities to the secondary structural motifs whereas the former—that is, the present approach—emphasizes segments limited to residue pairs while extending the accessible space to that of a discretization of the two-angle space with more than 5000 bins—that is, possible configurations.

Given the coordinates of a protein structure, the series of dihedral angles $\{\Upsilon_k\}$ can readily be computed. The probabilities entering in the sum of the structural entropy each depend on the relative probability that the measured dihedral angles are compatible with the corresponding residue(s) they connect. That is, the probabilities entering in Eq. 2 are $\{P_{R_i}(w_{k(i)}, v_{l(i)}), P_{R_i,R_{i+1}}(v_{l(i)}, w_{k(i+1)})\}$, where $\phi_i \in w_{k(i)}$ and $\psi_i \in v_{l(i)}$, given that $\{w_k\}$ and $\{v_l\}$ are the partitions in the angle space used to construct the histogrammed distributions. This procedure, while direct, discretizes the possible results. Smoother estimates of the dihedral-angle information entropy could be obtained using standard interpolating techniques. But this is not done here because the simpler discrete approach provides estimates of the structural entropy with sufficient accuracy to test the proposed checking functions.

A checking function for secondary structure propensity

Given the normalized probability distribution, $P(\Delta S)$, and a putative structure with well-defined dihedral angles, $\{(\phi_i, \psi_i), (\psi_i, \phi_{i+1})\}$, an integrated probability function for the entropy difference can be defined by merging the left and right cumulative distribution functions as

$$I(\vec{q}) = \begin{cases} \int_0^{\Delta S(\vec{q})} P(\Delta')d\Delta' & \text{if } \Delta S < \overline{\Delta S} \\ \int_{\Delta S(\vec{q})}^{\infty} P(\Delta')d\Delta' & \text{if } \Delta S \geq \overline{\Delta S} \end{cases}, \tag{7}$$

where $\overline{\Delta S}$ is the median value of ΔS. The integral I will, by definition, take the value of $\frac{1}{2}$ when evaluated at the median. The deviation relative to the median can thus be characterized by

$$D_1(\vec{q}) = \begin{cases} \ln(2I) & \text{if } \Delta S(\vec{q}) < \overline{\Delta S} \\ -\ln(2I) & \text{if } \Delta S(\vec{q}) \geq \overline{\Delta S} \end{cases}, \tag{8}$$

which takes the value of 0 for the median structure, and otherwise measures the distance away from the median structure in the distribution. When D_1 is negative (positive), it signals that the deviation is below (above) the median.

TABLE 1 The protein sublibraries in this work

Sublibrary	Structures	Peak/10^{-3}	Width/10^{-4}
EXP	24,444	4.17	0.96
NR100	11,157	4.24	0.92
NR90	2,762	4.37	0.98
NR70	2,176	4.74	1.03
NR50	1,768	4.85	1.04
NR100T	644	4.44	0.79

The name and number of proteins in the sublibraries used in this work are listed in the first and second columns, respectively. The peak and width of the $\Delta S^{(90)}$ distributions shown in panel *c* of Fig. 3 and evaluated using the dihedral angle distributions from NR90, are also listed.

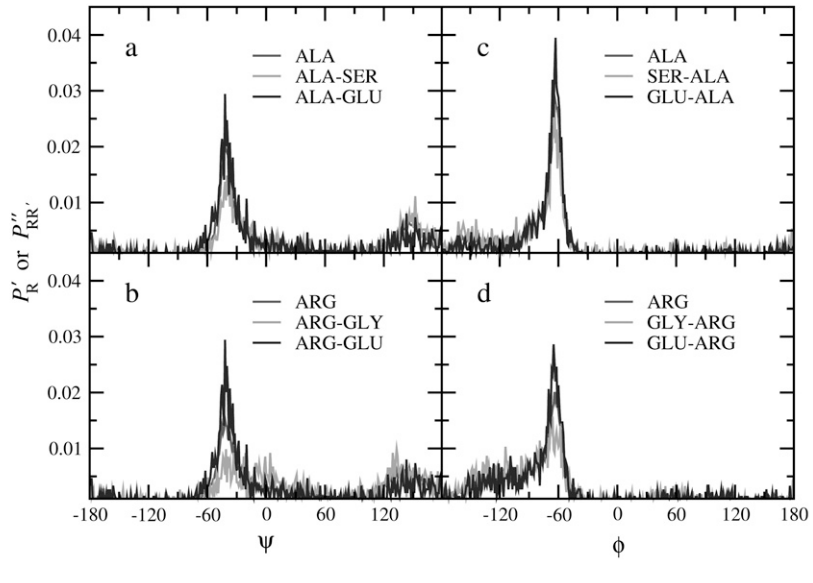

FIGURE 2 One-dimensional distributions of ψ and ϕ projected from ϕ_i-ψ_i and ψ_i-ϕ_{i+1} plots. (*a*) $P'_{\mathrm{ALA}}(\psi)$, $P''_{\mathrm{ALA,SER}}(\psi)$, $P''_{\mathrm{ALA,GLU}}(\psi)$; (*b*) $P'_{\mathrm{ARG}}(\psi)$, $P''_{\mathrm{ARG,GLY}}(\psi)$, $P''_{\mathrm{ARG,GLU}}(\psi)$; (*c*) $P'_{\mathrm{ALA}}(\phi)$, $P''_{\mathrm{SER,ALA}}(\phi)$, $P''_{\mathrm{GLU,ALA}}(\phi)$; and (*d*) and $P'_{\mathrm{ARG}}(\phi)$, $P''_{\mathrm{GLY,ARG}}(\phi)$, $P''_{\mathrm{GLU,ARG}}(\phi)$.

To make the D_1 checking function even more intuitive, a new checking function D_2 is defined to roughly describe the number of standard deviations away from the median structure through the expression

$$D_2(\vec{q}) = \begin{cases} \sqrt{2}\mathrm{erf}^{-1}(2I-1) & \text{if } \Delta S(\vec{q}) < \overline{\Delta S} \\ \sqrt{2}\mathrm{erf}^{-1}(1-2I) & \text{if } \Delta S(\vec{q}) \geq \overline{\Delta S} \end{cases}. \quad (9)$$

As described in Supplement *C* in the Supplementary Material, the D_2 checking function evaluated for a Gaussian distribution with zero mean and unit standard deviation is exactly equal to the number of standard deviations away from the median structure. Thus D_2 may be interpreted as a measure of the relative likelihood for ΔS in terms of deviations from the mean. It effectively uniformizes the distribution in the sense that it maps the original distribution precisely to the normal curve. In particular, values of $|D_2| > 3$ suggest that the specified structure in a group of structures whose cumulative likelihood, while possible, is <0.13%. To check the effectiveness of these new scores, D_1 and D_2 are calculated separately for the EXP and the NR100T libraries below.

RESULTS AND DISCUSSION

Dihedral angle distributions

The ψ_i-ϕ_i and ϕ_i-ψ_{i+1} dihedral angle distributions for all five libraries outlined in the section on Data-Mining the ψ_i-ϕ_{i+1} Distributions are presented and described in Supplements *A* and *B* in the Supplementary Material. In addition to their role in this work, they may be of use in homology-based methods for constructing proteins. For example, Srinivasan and coworkers (19–21) have used such distributions to predict backbone conformations of short polypeptides.

On the choice of the sequence database

To implement the checks presented in the section on A Checking Function for Secondary Structure Propensity, an underlying database must be selected. The EXP library would be a poor choice because it necessarily includes multiple copies of the same structure. Theoretically derived structures should also be ignored because they may differ from the experimental database. To choose which of the experimental subsets of the nonredundant sublibraries—NR50, NR70, NR90, or NR100—would be optimal, it is helpful to construct the corresponding dihedral-angle information entropy and their relative properties. In particular, the distributions of $\Delta S^{(X)}$—based on the NR*X* sublibrary—have been evaluated across all the structures in each of the five sublibraries: NR50, NR70, NR90, NR100, and EXP. The statistical error in $\Delta S^{(X)}$ decreases with increasing *X* because the size of the sublibrary increases with *X*. But at the same time, the bias due to redundancy is also increasing with *X*.

The distributions of $\Delta S^{(X)}$ are shown in Fig. 3. The EXP library and NR100 sublibrary contain several sets of structures with considerable sequence identity resulting in skewed distributions regardless of the choice of the checking function. As expected, the relatively small size of the sublibraries underlying the $\Delta S^{(50)}$ and $\Delta S^{(70)}$ measures leads to noisy distributions. Meanwhile, the distributions in $\Delta S^{(100)}$ appear to be broadened by the underlying redundancy in the NR100 sublibrary. The differences between the five sublibraries appear to be revealed—and perhaps converged—most sharply by panel *c*, which displays the distributions for $\Delta S^{(90)}$. One might be tempted to choose $\Delta S^{(70)}$ instead of $\Delta S^{(90)}$ because both scores reveal that the NR90 distribution is more like that of the redundant libraries. However, the better statistics of $\Delta S^{(90)}$ in light of the relatively small redundancy error, and the similarity in the peak positions between NR100 and NR90 as listed in Table 1, suggests that NR90 is an optimal choice. In light of this heuristic argument, NR90 is used in the remainder of this article as the underlying distribution in calculating ΔS and the associated checking functions; the superscript in $\Delta S^{(90)}$ is henceforth omitted.

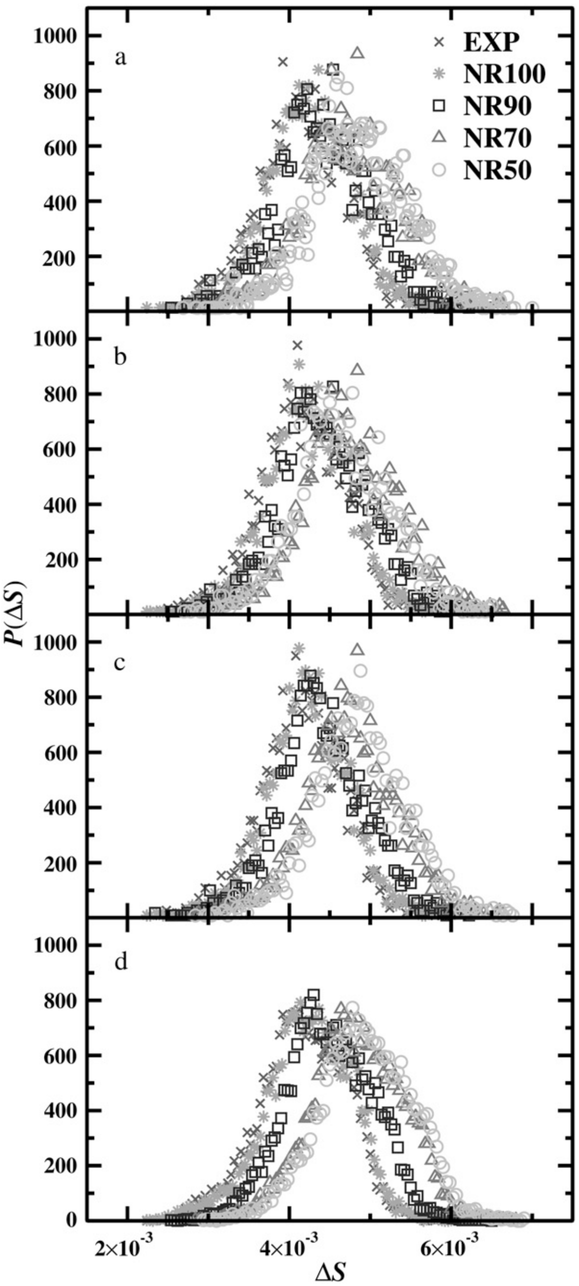

FIGURE 3 The distributions of $\Delta S^{(50)}$, $\Delta S^{(70)}$, $\Delta S^{(90)}$, and $\Delta S^{(100)}$ evaluated across several sublibraries are displayed in panels *a*, *b*, *c*, and *d*, respectively. In each panel the information entropy difference ΔS is evaluated across the NR50, NR70, NR90, and NR100 sublibraries, and the EXP library.

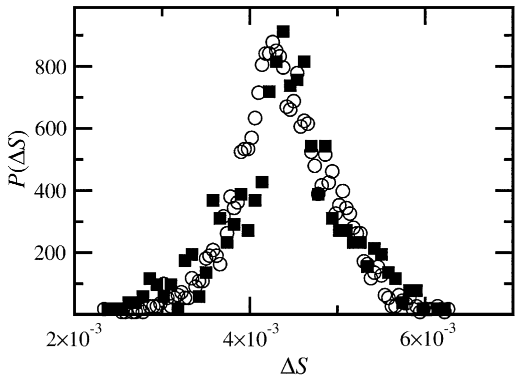

FIGURE 4 Distributions of ΔS evaluated across the 2762 experimental structures in the NR90 sublibrary (*circles*) and 644 theoretical structures in the NR100T sublibrary (*squares*). Note that, to make the results comparable, the distributions have been normalized by the bin size, i.e., 4×10^{-5} and 8×10^{-4}, respectively.

The distributions of ΔS for experimental and theoretical structures in NR90 and NR100T, respectively, are shown in Fig. 4. The mean value and standard deviation σ of ΔS of experimental structures are 4.38×10^{-3} and 5.74×10^{-4}, respectively, indicating that roughly 71% of the total structures have a ΔS between 3.81×10^{-3} and 4.95×10^{-3}, i.e., between $\langle \Delta S \rangle - \sigma$ and $\langle \Delta S \rangle + \sigma$. The mean value and standard deviation for the theoretical structures are 4.35×10^{-3} and 6.82×10^{-4}, respectively, and ~64% of the theoretical models have a ΔS within one standard deviation of the mean of the experimental models. The two distributions are surprisingly similar, particularly since the difference seen between the NR90 and NR100 distributions does not appear to persist for NR100T. The origin of this likely lies in the fact that the NR100T sublibrary does not have NR100's degree of sampling bias, because the latter contains many similar single-point mutants. However, on average, fewer theoretically determined structures are within a σ of the mean and this is a notable difference between the experimental and theoretical structures. This result is likely a consequence of the fact that many theoretical structures use rule sets for their construction which do not reflect the degree of correlation between distant residues present in nature. These observations indicate the insight that ΔS provides on the relative compatibility of a given structure with respect to the experimental NR90 sublibrary of the PDB.

One possible concern here is that the only standard for inclusion of a protein within any of these libraries with respect to the accuracy of the structure lies in the fact that the reported structure provides sufficient information to obtain all of its dihedral angles. One could use more rigorous criteria employing R-factors or other self-reported position error bars. Indeed several studies that have developed checking functions have used such rigorous criteria (3,6,7). However, we found that implementation of these criteria in constructing libraries nearly requires a file-by-file assessment since the requisite information is not coded uniformly through the PDB. Meanwhile our preliminary constructions of such libraries, while modifying the dihedral-angle distributions slightly, do not lead to appreciably distinct distributions in the information theory entropies or the various checking functions. Hence all the results reported here have been obtained using the simple rule for structure identification described above.

D_1 and D_2 checks

The distributions of D_1 calculated using Eq. 8 across the NR90 and NR100T sublibraries are shown in Fig. 5. The distributions are nearly Gaussian as suggested above. However, features seen above in Fig. 4 in assessing the relative compatibility between the NR90 and NR100T sublibraries are still visible in Fig. 5. The distributions in D_2 displayed in Fig. 6 retain these features as well, but the uniformizing procedure outlined in Supplement C (see Supplementary Material) now leads to a normal Gaussian distribution for the NR90 structures. Interestingly, the lack of correlation in some of the NR100T structures is exhibited by a shoulder on the left side of the NR100T distribution.

Although the definitions of D_1 and D_2 may appear cumbersome, their generalized forms are helpful so as to account for the fact that the probability distribution in ΔS is not symmetric. If it were symmetric, then the simpler arguments at the end of the previous section using a single characteristic σ would suffice. As remarked previously (and shown explicitly in Supplement C in the Supplementary Material), in the limit that the distribution in ΔS is Gaussian, the definition of D_2 reduces precisely to the number of standard deviations that a given structure differs from the median. In summary, Eqs. 8 and 9 define equivalent new checks, D_1 and D_2, for the compatibility of the dihedral angles of a given structure with the existing PDB set of nonredundant experimental structures, although D_2 is preferred because it takes on nontrivial values even for exponentially unlikely structures.

To illustrate the values of the D_1 and D_2 checks, it is helpful to examine a few representative structures arbitrarily chosen from the PDB. The HIV envelope glycoprotein (1g9nG) (56), the p53 DNA binding domain (1tupA) (57), and the G-protein α-1 chain (1gg2A) (58) are fairly common proteins whose structures have been resolved and deposited in the PDB. The D_1 values for these structures are -0.06, -0.25, and 0.23, respectively, which alone might not seem to provide a simple score of the structural quality. However, the D_2 values are -0.08, -0.32, and 0.33. These values are easily interpreted as they indicate that all three structures are within one standard deviation of the PDB database. That is, their dihedral angles with respect to correlation around a residue and between residues are typical of the structures in the NR90 sublibrary. But recall that their information entropy is consequently greater than their corresponding standard entropies. Thus, they evidently exhibit propensities for secondary structural interactions that are typical of the structures in the NR90 sublibrary.

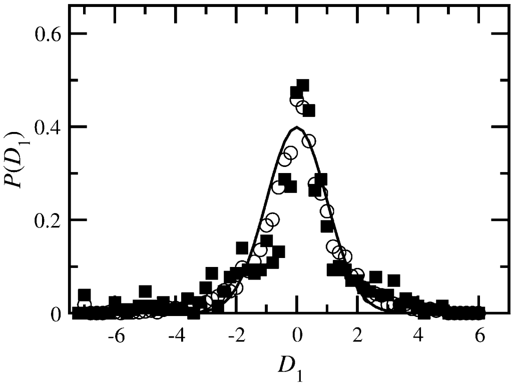

FIGURE 5 Distributions of D_1 evaluated across the NR90 (*circles*) and NR100T (*squares*) sublibraries. In all cases, D_1 is determined using Eq. 8 with $\overline{\Delta S}$ equal to the corresponding median value ($= 4.38 \times 10^{-3}$) of the experimental structures in the NR90 sublibrary. The solid line is a Gaussian distribution with zero mean value and unit standard deviation.

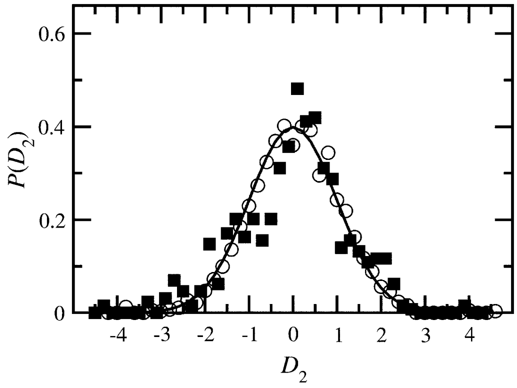

FIGURE 6 Distributions of D_2 evaluated across the NR90 (*circles*) and NR100T (*squares*) sublibraries. The median value of ΔS of experimental structures has been used. The solid line is the same Gaussian distribution as in Fig. 5.

Alternatively, the D_2 check can be used to identify protein structures whose angles are atypical with respect to the distribution of correlated angles in the PDB. Such atypical structures are not necessarily incorrect structures. Indeed, when D_2 is large and negative, the structures could be correct, but for whatever reason contain dihedral angles in the most probable positions independent of the sequence beyond their nearest neighbors. Alternatively when D_2 is large and positive, particularly strong correlations of distant residues may give rise to angles that adopt low probability configurations. Although correct structures exist that satisfy such limits, they are still atypical relative to the distribution because, as shown in Fig. 4, most of the experimental structures in the NR90 sublibrary have a structural entropy difference near the mean, $\overline{\Delta S}$. This raises the intriguing possibility that D_2 can be used to highlight atypical regions in proteins that are atypical due to some functional constraint. These regions could arise for reasons related to active site architectures or regions critical to forming protein-protein interactions. Hence the D_2 measure may serve a role in highlighting regions of interest when structures of unknown function or physiological role are solved as part of ongoing high throughput structural proteomics efforts. Long-range interactions through a protein structure are of course important to understanding catalysis, concerted movements, and even when seeking to understand the evolutionary history of proteins within a conserved family of proteins. Thus D_2 can highlight these potential regions within a structure too.

The role of D_2 in checking theoretical structures

All structures in the NR90 and NR100T sublibraries with a value of $|D_2| \geq 3$ are listed in Table 2. The number of such structures is 17 (0.6%) and 11 (1.7%) for the experimentally and theoretically derived structures, respectively. The structures in the larger EXP and model protein libraries have also been assessed according to the D_2 check. It was found that 264 (1.1%) and 66 (6.7%) structures are atypical out of the 24,444 experimental and 981 theoretical structures available, respectively. (All of the atypical structures and their D_1 and D_2 values are listed in Supplement *C* in the Supplementary Material.) The fact that in these sublibraries, the theoretical structures are much more likely to be atypical than the experimental structures, is a possible indicator that the former is somehow different from naturally occurring structures. More importantly, the primary difference manifests as a shoulder in the distributions in the negative D_2 region. This is the region that signals structures that are near to the structures with standard entropy. Thus the dihedral angles deviate little from the most likely angles, indicating that they have not been altered by secondary interactions. It should come as no surprise that some fraction of the theoretically derived structures contain dihedral angles that lack such information. However, the important result here is that D_2 is a reporter of such propensities.

TABLE 2 Atypical protein structures

Experimental			Theoretical		
PDB ID	D_1	D_2	PDB ID	D_1	D_2
1a2xB	7.0	3.9	1clgA	−7.9	−4.4
1a92A	−7.0	−3.9	1l1uT	7.0	3.9
1bb1B	7.9	4.4	1lh8A	−6.1	−3.1
1czqA	−6.7	−3.3	1llkA	−6.7	−3.3
1g6uA	−7.0	−3.9	1lp0A	−6.7	−3.3
1jekA	−7.0	−3.9	1m5gT	7.0	3.9
1jekB	7.0	3.9	1n1rA	−6.1	−3.1
1jrjA	−6.7	−3.3	1opvA	−6.1	−3.1
1l2pA	−7.0	−3.9	1sewA	−7.9	−4.4
1l2yA	−6.4	−3.1	1sr1	−6.7	−3.3
1motA	−6.7	−3.3	2clgA	−6.0	−3.0
1mz9A	−6.1	−3.1			
1n7sA	−7.0	−3.9			
1nyjA	−7.0	−3.3			
1pd7B	7.9	4.4			
1qr9A	−7.0	−3.9			
1sb0B	7.0	3.9			

The atypical structures—namely those structures whose $|D_2|$ value is $\geq$3.0—are listed according to their PDB ID, augmented by the chain ID. The corresponding D_1 and D_2 values are also provided. Note that there are 17 experimental (*left*) and 11 theoretical (*right*) structures. Perhaps noteworthy is the fact that the sequence identity between 1jekA and 1jekB is only 17.6%, and hence can lead to rather different values of these measures.

This can be further illustrated through a study of the D_2 check on a series of structures constructed by homology modeling. The success of the homology modeling package, MODELLER (37,59), has previously been shown using several template structures (TS) to construct model structures (MS) for the protein with PDB ID, 1fdx. This study has been reproduced here with the additional construction of the model structure based on the known 1fdx target structure as a template. The sequence identities (SI) and root mean-square deviations (RMSDs) are shown in Table 3, and illustrate the previously reported success. Namely, the greater the sequence identity between primary structures of the TS and target, the smaller the RMSD between the MS and the target. Although it should be noted that the RMSD is not zero even when the target structure is used as the TS. As also reported in Table 3, the D_2 check of the target structure (= −0.73) is far from zero, as is the value of this checking function for most of the TS values. However, the D_2 checks of all five predicted MS values are nearly zero, and all are evidently different from that of the corresponding TS and of the target structure. The D_2 check does not differentiate between these five MS values in terms of their relative fidelity to the target structure. Other scores or checking functions are needed for (and indeed some satisfy) this property. However, the consistently zero value in the D_2 checks of the MS values illustrates the fact that structures predicted by MODELLER, while often containing high fidelity to the target structure, leave out some property that would make them atypical of the PDB in the sense that is measured by the D_2 check. This property is the long-range correlation in the dihedral angles between non-neighbor residues. Although perhaps not surprising that MODELLER removes this propensity, it is nevertheless useful that D_2 check provides a quick verification of this removal and it evidently provides an independent check for what could be done to expand the functionality of programs such as MODELLER.

D_2 and other checking functions

A comparison between D_2 check, the torsion angle G-factor in PROCHECK (6,7), and the Ramachandran Z-score of WHAT_CHECK (3) has been made for several example

TABLE 3 Assessing templated model structures

Template	SI(%)	D_2(TS)	D_2(MS)	RMSD(Å)
1fdx	100.0	−0.73	−0.03	0.26
1fdn	66.7	−1.05	−0.08	0.69
5fd1	42.6	−0.48	−0.04	1.67
1fxd	35.2	−0.34	−0.04	5.27
2fxb	20.4	−0.12	−0.06	8.15

The values of D_2 for series of template structures (TS) and the corresponding model structures (MS) derived from them are shown for the 1fdx target. The sequence identity (SI), and the RMSD between the MS and target are also provided. Note that the use of the target as the TS results in a different MS than the target as indicated by a nonzero value in their RMSD.

structures in the PDB. The torsion angle G-factor is a log-odds score of the observed distributions of the ϕ_i-ψ_i combination. A low G-factor often indicates an unusual structure (6,7). The Ramachandran Z-score is the number of standard deviations that the score deviates from the expected value. It shows how ''normal'' the ϕ_i-ψ_i angles in a protein structure are. Z-scores above 4.0 and below -4.0 are very uncommon (3). The results are shown in Table 4 for six experimental structures—1tupA, 1g9nG, 1gg2A, 1stn, 1jekA, 1jekB, and 1n7sA—and three theoretical structures—1lluT, 1lp0, and 1lh8A—which have been chosen because they provide a range of D_2 values. Except for two structures, 1g9nG and 1gg2A, the Z-scores are compatible with the D_2 values in terms of the assessment that the structures are typical or not typical. However, most of the structures contain G-factors that are not compatible with their D_2 checks in terms of this assessment. (Note that, to run PROCHECK, a resolution for a structure must be specified. Although this is readily available for experimental structures, it is evidently not available for the theoretical structures. Nevertheless, the theoretical structures were run with varying resolutions—2.0, 2.5, and 3.0 Å—all resulting in the same values for the G-factors.) In summary, D_2 check differs from PROCHECK and WHAT_CHECK in their assessments of these protein structures, and evidently provides distinct information about the structures. In particular, as seen above, the use of the dihedral-angle correlation between neighboring residues in the D_2 checking function allows one to obtain a signal of the presence for propensities between residues beyond the nearest neighbor. It thereby complements the information from PROCHECK and WHAT_CHECK.

CONCLUSION

A dihedral-angle information entropy describing how a particular model protein is similar to naturally occurring proteins has been discussed in this work. Based on this entropy, new checking functions, D_1 and D_2, are proposed as a check of the likelihood of the compatibility of the dihedral angles of a given structure to the experimental structures in the PDB. The results for both experimentally and theoretically derived structures in the PDB indicate that this method is simple and effective.

TABLE 4 The D_2 check, G-factor, and Z-score values for 10 different protein structures available in the PDB

PDBID	D_2	G-factor	Z-score
1tupA	−0.32	−0.02	0.41
1g9nG	−0.08	−0.14	−4.47
1gg2A	0.33	0.07	−3.56
1stn	−0.8	0.06	−0.04
1jekA	−3.9	0.75	4.23
1jekB	3.9	0.67	5.41
1n7sA	−3.9	0.62	3.77
1lluT	3.9	0.05	−6.53
1lp0A	−3.3	0.12	3.36
1lh8A	−3.1	0.31	4.09

Generally speaking, the D_1 and D_2 checks signal the propensity for a protein to contain secondary structural interactions in comparison with the PDB. The overall structures found to be atypical by these checking functions may be classified as:

1. Weakly correlated (or mean-field-like) in the sense that residues beyond the nearest residue do not affect the dihedral angles; or
2. Strongly correlated in the sense that distant residues lead to large deviations in the dihedral angles away from the typical values; or
3. Incorrect in the sense that some of the angles may have been incorrectly assigned.

In particular, large negative values of D_2 check indicate structures that are perhaps too likely, while large positive values indicate structures that are perhaps too unlikely in comparison with the typical structures of the PDB database. The use of D_2 check at the residue level has been developed and will be discussed separately (S. Zhong, S. Quirk, and R. Hernandez, unpublished). D_2 check is complementary to existing scoring functions used in assessing structure predictions but provides a different form of stereochemical information. For example, it can be used in concert with other functions to identify important or unusual parts of a structure.

One criticism that could be levied against this work—and indeed against many bioinformatic tools based on a reference set—centers on the question of whether the chosen reference sublibrary of the PDB is representative of the protein universe. The recent work of Zhang et al. (60) suggests that the diversity of single-domain structures available in the PDB database is indeed representative of the protein universe. But there may be a danger that the distribution of such structures is skewed in some way. To reduce the presence of such biasing, the reference sublibrary selected in this work excluded structures that had >90% sequence redundancy. Meanwhile, the statistical information available from the current size of the database was sufficient only for bins with 5° windows. While both the coverage of the protein space and the accuracy of the distributions appear to be sufficient in the treatment performed here, one would expect that both would improve in the future as the PDB grows.

One additional result of this work is the confirmation that the ψ_i-ϕ_{i+1} plots contain correlation between dihedral angles of a given residue and the identity of the neighboring residue. This result validates previous observations (19–21,44,46, 47,50). It is seemingly in contradiction of the Flory isolated-pair hypothesis (61) in which it was assumed that the ϕ_i-ψ_i distribution of each residue in a protein backbone is independent of the neighbors' identities. However, the differences found here are sufficiently small that violations of the isolated-pair hypothesis are subtle. For this same reason, it is not surprising that Brooks and co-workers (62) found that the isolated-pair hypothesis holds very well upon averaging over the ensemble to obtain conformational entropies.

In summary, this work serves to increase the awareness of the effect of nearest-neighbor frequency on the pairwise dihedral distributions and introduces a useful series of checking functions that can be used to interpret both experimental and theoretical protein structures.

SUPPLEMENTARY MATERIAL

Supplements *A–C* can be found by visiting BJ Online at http://www.biophysj.org. Supplement *A* provides the detailed method and analysis of the construction of the distributions. Supplement *B* provides figures for all 420 distributions generated from the NR90 sublibrary. Supplement *C* provides an analysis of the uniformizing procedure discussed above, and a listing of the D_1 and D_2 scores of all atypical structures in the PDB.

We acknowledge the roles of Nga Anh Nguyen and Joanna Balsamo in the preliminary calculation of the dihedral angle libraries.

This work has been partially supported by National Science Foundation grants No. NSF 02-123320 and No. 04-43564. Additionally, R.H. is the Goizueta Foundation Junior Professor.

REFERENCES

1. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
2. Branden, C. I., and T. A. Jones. 1990. Between objectivity and subjectivity. *Nature.* 343:687–689.
3. Hooft, R. W. W., G. Vriend, C. Sander, and E. E. Abola. 1996. Errors in protein structures. *Nature.* 381:272.
4. Abola, E. E., A. Bairoch, W. C. Barker, S. Beck, D. A. Benson, H. Berman, G. Cameron, C. Cantor, S. Doubet, T. J. P. Hubbard, T. A. Jones, G. J. Kleywegt, A. S. Kolaskar, A. Van Kuik, A. M. Lesk, H. W. Mewes, D. Neuhaus, G. Pfeiffer, L. F. TenEyck, R. J. Simpson, G. Stoesser, J. L. Sussman, Y. Tateno, A. Tsugita, E. L. Ulrich, and J. F. G. Vliegenthart. 2000. Quality control in databanks for molecular biology. *Bioessays.* 22:1024–1034.
5. Ramakrishnan, C., and G. N. Ramachandran. 1965. Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophys. J.* 5:909–933.
6. Morris, A. L., M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton. 1992. Stereochemical quality of protein structure coordinates. *Proteins.* 12:345–364.
7. Laskowski, R. A., M. W. MacArthur, D. S. Moss, and J. M. Thornton. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 26:283–291.
8. MacArthur, M. W., and J. M. Thornton. 1993. Conformation analysis of protein structures derived from NMR data. *Proteins.* 17:232–251.
9. MacArthur, M. W., R. A. Laskowski, and J. M. Thornton. 1994. Knowledge-based validation of protein structure coordinates derived by x-ray crystallography and NMR spectroscopy. *Curr. Opin. Struct. Biol.* 4:731–737.
10. Laskowski, R. A., M. W. MacArthur, and J. M. Thornton. 1998. Validation of protein models derived from experiment. *Curr. Opin. Struct. Biol.* 8:631–639.
11. Brünger, A. T. 1992. Free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature.* 355:472–475.
12. Kleywegt, G. J., and T. A. Jones. 1996. Phi/psi-chology: Ramachandran revisited. *Structure.* 4:1395–1400.
13. Kleywegt, G. J. 1997. Validation of protein models from C^{α} coordinates alone. *J. Mol. Biol.* 273:371–376.
14. Kleywegt, G. J., and T. A. Jones. 1997. Model building and refinement practice. *Methods Enzymol.* 277:208–230.
15. Kleywegt, G. J. 2000. Validation of protein crystal structures. *Acta Crystallogr.* D56:249–265.
16. Hooft, R. W. W., C. Sander, and G. Vriend. 1997. Objectively judging the quality of a protein structure from a Ramachandran plot. *CABIOS.* 13:425–430.
17. Lovell, S. C., I. W. Davis, W. B. Arendall III, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. 2003. Structure validation by C_{α} geometry: ϕ, ψ, and C_{β} deviation. *Proteins.* 50:437–450.
18. Willard, L., A. Ranjan, H. Y. Zhang, H. Monzavi, R. F. Boyko, B. D. Sykes, and D. S. Wishart. 2003. VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.* 31:3316–3319.
19. Sudarsanam, S., R. F. DuBose, C. J. March, and S. Srinivasan. 1995. Modeling protein loops using a ϕ_{i+1},ψ_{i} dimer database. *Protein Sci.* 4:1412–1420.
20. Sudarsanam, S., and S. Srinivasan. 1995. Searching for protein loops in parallel. *CABIOS.* 11:591–593.
21. Sudarsanam, S., and S. Srinivasan. 1997. Sequence-dependent conformational sampling using a database of ϕ_{i+1} and ψ_{i} angles for predicting polypeptide backbone conformations. *Protein Eng.* 10:1155–1162.
22. Parker, J. M. R. 1999. The relationship between peptide plane rotation (PPR) and similar conformations. *J. Comput. Chem.* 20:947–955.
23. Ozer, G., J. Foley, S. Zhong, J. M. Moix, S. Quirk, and R. Hernandez. 2006. http://www.d2check.gatech.edu/.
24. Shortle, D. 2002. Composite of local structure propensities: evidence for local encoding of long-range structure. *Protein Sci.* 11:18–26.
25. Shortle, D. 2003. Propensities, probabilities, and the Boltzmann hypothesis. *Protein Sci.* 12:1298–1302.
26. Fang, Q. J., and D. Shortle. 2005. A consistent set of statistical potentials for quantifying local side-chain and backbone interactions. *Proteins.* 60:90–96.
27. Fang, Q. J., and D. Shortle. 2005. Enhanced sampling near the native conformation using statistical potentials for local side-chain and backbone interactions. *Proteins.* 60:97–102.
28. Hovmöller, S., T. Zhou, and T. Ohlson. 2002. Conformations of amino acids in proteins. *Acta Crystallogr.* D58:768–776.
29. Sheik, S. S., P. Ananthalakshmi, G. R. Bhargavi, and K. Sekar. 2003. CADB: conformation angles database of proteins. *Nucleic Acids Res.* 31:448–451.
30. Priestle, J. P. 2003. Improved dihedral-angle restraints for protein structure refinement. *J. Appl. Crystallogr.* 36:34–42.
31. Dayalan, S., S. Bevinakoppa, and H. Schroder. 2004. A dihedral angle database of short sub-sequences for protein structure prediction. The Second Asia-Pacific Bioinformatics Conference, Australian Computer Society, Inc., Sydney, NSW, Australia.
32. Vriend, G. 1990. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* 8:52–55.
33. Zheng, Q., R. Rosenfeld, C. DeLisi, and D. J. Kyle. 1994. Multiple copy sampling in protein loop modeling: computational efficiency and sensitivity to dihedral angle perturbations. *Protein Sci.* 3:493–506.
34. Mathiowetz, A. M., and W. M. Goddard III. 1995. Building proteins from C_{α} coordinates using the dihedral probability grid Monte Carlo method. *Protein Sci.* 4:1217–1232.
35. Cheng, B., A. Nayeem, and H. A. Scheraga. 1996. From secondary structure to three-dimensional structure: improved dihedral angle probability distribution function for use with energy searches for native structures of polypeptides and proteins. *J. Comput. Chem.* 17:1453–1480.

36. Fiser, A., R. K. Gian Do, and A. Šali. 2000. Modeling of loops in protein structures. *Protein Sci.* 9:1753–1773.

37. Marti-Renom, M. A., A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Šali. 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29:291–325.

38. Baker, D., and A. Šali. 2001. Protein structure prediction and structural genomics. *Science.* 294:93–96.

39. Fiser, A., M. Feig, C. L. Brooks III, and A. Šali. 2002. Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.* 35:413–421.

40. Jacobson, M. P., D. L. Pincus, C. S. Rapp, T. J. F. Day, B. Honig, D. E. Shaw, and R. A. Friesner. 2004. A hierarchical approach to all-atom protein loop prediction. *Proteins.* 55:351–367.

41. Wu, T. T., and E. A. Kabat. 1971. An attempt to locate the non-helical and permissively helical sequences of proteins: application to the variable regions of immunoglobulin light and heavy chains. *Proc. Natl. Acad. Sci. USA.* 68:1501–1506.

42. Kabat, E. A., and T. T. Wu. 1972. Construction of a three-dimensional model of the polypeptide backbone of the variable region of κ-immunoglobulin light chains. *Proc. Natl. Acad. Sci. USA.* 69: 960–964.

43. Wu, T. T., and E. A. Kabat. 1973. Attempt to evaluate influence of neighboring amino-acid $(n-1)$ and $(i+1)$ on backbone conformation of amino acid (n) in proteins' use in predicting three-dimensional structure of polypeptide backbone of other proteins. *J. Mol. Biol.* 75:13–31.

44. Pappu, R. V., R. Srinivasan, and G. D. Rose. 2000. The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl. Acad. Sci. USA.* 7:12565–12570.

45. Chakrabarti, P., and D. Pal. 2001. The interrelationships of side-chain and main-chain conformations in proteins. *Prog. Biophys. Mol. Biol.* 76:1–102.

46. Zaman, M. H., M. Y. Shen, R. S. Berry, K. F. Freed, and T. R. Sosnick. 2003. Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the Flory isolated-pair hypothesis for peptides. *J. Mol. Biol.* 331:693–711.

47. Betancourt, M. R., and J. Skolnick. 2004. Local propensities and statistical potentials of backbone dihedral angles in proteins. *J. Mol. Biol.* 342:635–649.

48. Esposito, L., A. De Simone, A. Zagari, and L. Vitagliano. 2005. Correlation between ω and ψ dihedral angles in protein structures. *J. Mol. Biol.* 347:483–487.

49. RCSB Protein Data Bank. 2006. http://www.rcsb.org/pdb/clusterStatistics.do.

50. DeWitte, R. S., and E. I. Shakhnovich. 1994. Pseudodihedrals: simplified protein backbone representation with knowledge-based energy. *Protein Sci.* 3:1570–1581.

51. Shannon, C. E. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27:379–423.

52. Solis, A. D., and S. Rackovsky. 2002. Optimally informative backbone structural propensities in proteins. *Proteins.* 48:463–486.

53. Solis, A. D., and S. Rackovsky. 2004. On the use of secondary structure in protein structure prediction: a bioinformatic analysis. *Polym.* 45:525–546.

54. Garnier, J., D. J. Osguthorpeb, and B. Robson. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120:97–120.

55. Kloczkowski, A., K. L. Ting, R. L. Jernigan, and J. Garnier. 2002. Combining the GOR algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins.* 49:154–166.

56. Kwong, P. D., R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson. 1998. Structure of HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature.* 393:648–659.

57. Cho, Y. J., S. Gorina, P. D. Jeffrey, and N. P. Pavletich. 1994. Crystal-structure of a p53 tumor suppressor DNA complex—understanding tumorigenic mutations. *Science.* 265:346–355.

58. Wall, M. A., D. E. Coleman, E. Lee, J. A. Iniguez-Lluhi, B. A. Posner, A. G. Gilman, and S. R. Sprang. 1995. The structure of the G-protein heterotrimer G(i-α-1)β(1)γ(2). *Cell.* 83:1047–1058.

59. Sali, A., and T. L. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234:779–815.

60. Zhang, Y., I. A. Hubner, A. K. Arakaki, E. Shakhnovich, and J. Skolnick. 2006. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. USA.* 103:2605–2610.

61. Flory, P. J. 1969. Statistical Mechanics of Chain Molecules. Wiley-Interscience, New York.

62. Ohkubo, Y. Z., and C. L. Brooks III. 2003. Exploring Flory's isolated-pair hypothesis: statistical mechanics of helix-coil transitions in polyalanine and the C-peptide from RNase A. *Proc. Natl. Acad. Sci. USA.* 100:13916–13921.