J|A|C|S
ARTICLES
Published on Web 05/21/2004

An Empirical Backbone−Backbone Hydrogen-Bonding Potential in Proteins and Its Applications to NMR Structure Refinement and Validation

Alexander Grishaev* and Ad Bax*

Contribution from the Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520

Received December 30, 2003; E-mail: grishaev@speck.niddk.nih.gov; bax@nih.gov

Abstract: A new multidimensional potential is described that encodes for the relative spatial arrangement of the peptidyl backbone units as observed within a large database of high-resolution X-ray structures. The detailed description afforded by such an analysis provides an opportunity to study the atomic details of hydrogen bonding in proteins. The specification of the corresponding potential of mean force (PMF) is based on a defined set of physical principles and optimized to yield the maximum advantage when applied to protein structure refinement. The observed intricate differences between hydrogen-bonding geometries within various patterns of secondary structure allow application of the PMF to both validation of protein structures and their refinement. A pronounced improvement of several aspects of structural quality is observed following the application of such a potential to a variety of NMR-derived models, including a noticeable decrease in backbone coordinate root-mean-square deviation relative to the X-ray structures and a considerable improvement in the Ramachandran map statistics.

Introduction

Determination of high-resolution protein structures by any experimental technique available today presents an example of an intrinsically ill-conditioned problem, as the number of the degrees of freedom necessary for defining such structures usually exceeds by far the number of experimentally attainable restraints. This situation, often encountered when interpreting experimental data in terms of an underlying model, is typically alleviated by regularizing the solution—that is, by creating an a priori bias toward models that are assumed to be "reasonable". For example, when interpreting X-ray data at lower than atomic resolution, regularization is performed by restraining interatomic bond lengths and angles to the values observed in small-molecule atomic resolution structures, where the data often overdetermine the solution. In the case of solution-state NMR of proteins, the problem is much more severe, owing to a smaller number of the experimental observables, as well as the local nature of the commonly used semiquantitative NOE restraints, and extensive regularization is therefore a prerequisite.

One approach to improve structural quality is to expand the number and nature of the observables restraining the structure. Residual dipolar couplings,[1] recorded in samples that are weakly aligned in the magnetic field, provide an example of these parameters; application of such restraints leads to a marked improvement of the overall structural quality.[2]

NMR structure refinement frequently also involves supplementing the usual semi-empirical force fields with terms derived from databases of high-resolution structures.[3] A major advantage of such a strategy,[4] first tested within the protein structure prediction field,[5] is the directness with which these pseudopotentials are extracted. In contrast, derivation of accurate energy terms based on first principles would require reproducing an extremely delicate balance between distinct physical forces, which often proves infeasible given the errors and approximations inherent in such calculations. An additional advantage of this so-called knowledge-based approach is a high "signal-to-noise" ratio attainable for these potentials, fueled by the fast growth of structural databases such as the RCSB Protein Data Bank.

Derivation of a potential of mean force starts from the observation of a correlation between certain internal variables within a set of high-quality structures. The degree with which a given structure follows such a correlation should increase with an increase of its overall quality. The correlation is then converted into a PMF via an inverted Boltzmann formula,

$$E(\Omega) \approx -kT \log[P(\Omega)]$$

(1) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9279−9283. King, H. C.; Wang, K. Y.; Goljer, I.; Bolton, P. H. *J. Magn. Reson., Ser. B* **1995**, *109*, 323−325. Tjandra, N.; Grzesiek, S.; Bax, A. *J. Am. Chem. Soc.* **1996**, *118*, 6264−6272. Bax, A.; Tjandra, N. *J. Biomol. NMR* **1997**, *10*, 289−29. Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111−1114.

(2) Tjandra, N.; Omichinski, J. G.; Gronenborn, A. M.; Clore, G. M.; Bax, A. *Nat. Struct. Biol.* **1997**, *4*, 732−738. Ottiger, M.; Tjandra, N.; Bax, A. *J. Am. Chem. Soc.* **1997**, *119*, 9825−9830. Bewley, C. A.; Gustafson, K. R.; Boyd, M. R.; Covell, D. G.; Bax, A.; Clore, G. M.; Gronenborn, A. M. *Nat. Struct. Biol.* **1998**, *5*, 571−578. Clore, G. M.; Starich, M. R.; Bewley, C. A.; Cai, M.; Kuszewski, J. *J. Am. Chem. Soc.* **1999**, *121*, 6513−6514.

(3) (a) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *Protein Sci.* **1996**, *5*, 1067−1080. (b) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Magn. Reson.* **1997**, *125*, 171−177. (c) Kuszewski, J.; Clore, G. M. *J. Magn. Reson.* **2000**, *146*, 249−254. (d) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 1, 2337−2338.

(4) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 248−256.

(5) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859−883.

10.1021/ja0319994 Not subject to U.S. Copyright. Publ. 2004 Am. Chem. Soc.

with the implicit hope that the application of the PMF would indeed result in an improvement of the structural accuracy. Here Ω denotes a set of generalized degrees of freedom, $P(\Omega)$ is the probability density function over Ω that describes the correlation, k is the Boltzmann constant, and T is usually assumed to be the ambient temperature.

Notwithstanding the obvious appeal of such potentials, their physical rigor is subject to considerable controversy. The applicability of Boltzmann statistics to structural databases has been claimed by considering the number of sequences that stabilize a fold of a given energy.[6] However, a strong theoretical argument against these constructs is rooted in the physical background of their derivation, which assumes that the probability density function $P(\Omega)$ is accumulated over either an ensemble or a trajectory for a single structure in question. In principle, this should lead to a separate PMF for every occurrence of the interaction within the macromolecule. In contrast, the sets of unrelated structures solved at a variety of experimental conditions, from which the PMFs are being derived, are clearly not in thermodynamic equilibrium with each other, and the extracted PMFs are not formulated to be site-specific.[7] Other complications are the entropy factor remaining in such free energy potentials and various biases inevitable in the databases.[8] These arguments indicate that extraction of successful potentials of mean force from a database of structures presents, in itself, a formidable problem.

This critique can be extended by considering that the PMF derived via Boltzmann inversion of the observed correlation function, such as those accumulated in the course of a simulation, will only reproduce the underlying potential when the system in question is a dilute medium where the average interparticle separation is much larger than the range of the interaction potentials. Clearly, dense biomolecular systems such as proteins are far from this scenario, bringing an additional degree of complication into the interpretation of the correlations they exhibit. Specifically, a large number of such correlations are expected to be indirect propagated consequences of other interactions. In that case, conversion of such correlations into the corresponding PMF may not necessarily lead to an improvement of the structural quality when the PMF is imposed.

The central question then becomes: how do we select correlations that can be expected to be the best candidates for conversion into a PMF? It seems reasonable to assume that such a correlation should be dominated by its internal degrees of freedom and largely decoupled from all external ones. For instance, extraction and application of the PMF describing covalent bond lengths and angles seem sensible in the absence of atomic-resolution data. In addition, the correlation is easier to investigate if it corresponds to a physical interaction of an established nature. Importantly, if the PMF is distance based, the interaction has to be short-range compared to the average separation between the partners. Third, the interaction has to contribute significantly to the proteins' stabilities and the details of their architectures. And finally, it has to be orthogonal to the terms encoded by the standard force fields. Most importantly, the ultimate indicator of the usefulness of any database-extracted PMF has to be the magnitude of structural accuracy improvement resulting from its application.

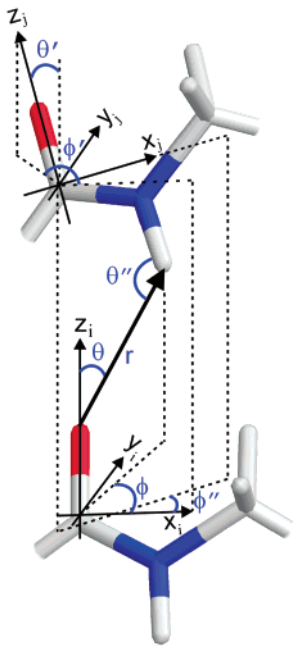

Figure 1. Variables that specity the relative arrangement of the atomic frames of the donor and acceptor peptidyl units. Oxygen atoms are shown in red, nitrogen atoms in blue, and hydrogen and carbon atoms in gray. Here, the z-axes coincide with the C−O vectors, and the x axes are within the C^{α}−C−O planes. Variables, r, θ, and ϕ, are the coordinates of the O···H^N vector in the spherical coordinate frame of the acceptor residue i; θ' is the angle between vectors z_i and z_j; ϕ' is the angle between the projection of the z_j vector onto the (x_i,y_i) plane and the x_i vector; and ϕ'' is the angle between the x_i vector and the projection of the x_j vector onto the (x_i,y_i) plane. Vertical dashed lines indicate projections onto the (x_i,y_i) plane, and skewed dashed lines show the direction of the projected vectors within the (x_i,y_i) plane.

An interaction that fits the above description is hydrogen bonding. Starting from the work of Pauling,[9] this phenomenon is considered to be one of the primary factors defining a protein's architecture. Correspondingly, there is a long history of implementation of potentials aimed at mimicking the hydrogen-bonding (H-bonding) interaction. This work has been carried out by both ab initio quantum mechanical calculations[10] and studies of the structural database statistics.[11]

A central problem in the investigation of the backbone−backbone H-bonding interaction is the specification of the relevant degrees of freedom out of the all possible internal variables describing the relative arrangement of two peptidyl units (Figure 1).

Neither the size of structural databases, nor the power of modern quantum chemistry methods are sufficient to create a detailed coverage of such multidimensional space. Consequently, all derivations of the H-bonding PMF (HB PMF) thus far have been based on the assumption of independence of most, or all, of the internal variables.

(6) Finkelstein, A. V.; Badretdinov, A. Y.; Gutin, A. M. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 142−150.
(7) Thomas, P. D.; Dill, K. A. *J. Mol. Biol.* **1996**, *257*, 457−469.
(8) Furuichi, E.; Koebl, P. *Proteins* **1998**, *31*, 139−149.
(9) Pauling, L.; Corey, R. B.; Branson, H. R. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 205−211. Pauling, L.; Corey, R. B. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 729−740.
(10) Steiner, T. *Angew. Chem., Int. Ed.* **2002**, *41*, 48−76.
(11) (a) Hooft, R.; Sander, C.; Vriend, G. *Proteins: Struct., Funct., Genet.* **1996**, *263*, 363−376. (b) McDonald, I. K.; Thornton, J. M. *J. Mol. Biol.* **1994**, *238*, 777−793. (c) Fabiola, F.; Bertram, R.; Korostelev, A.; Chapman, M. S. *Protein Sci.* **2002**, 1415−1423. (d) Kortemme, T.; Morozov, A. V.; Baker, D. *J. Mol. Biol.* **2003**, *326*, 1239−1259. (e) Lipsitz, R. S.; Sharma, Y.; Brooks, B. R.; Tjandra, N. *J. Am. Chem. Soc.* **2002**, *124*, 10621−10626.

An additional complication is that not all of these degrees of freedom are expected to be relevant to the H-bonding interactions; some of the apparent correlations may be indirect consequences of other phenomena characteristic of a dense protein environment. Finally, even if we were able to somehow select only those degrees of freedom that are responsible for the interaction, the probability density function (pdf) accumulated over the database would still contain the propagated correlation effects. For example, if we were to concentrate on the α-helical backbone geometry, the C═O···H^NN distance pdf would contain, along with the main $i/i + 4$ maximum at ∼2.0 Å, additional maxima at ∼2.6 Å ($i/i + 3$), ∼3.5 Å ($i/i + 2$), ∼4.5 Å ($i/i + 5$), and so on. Clearly, these secondary features should not be interpreted as characteristics of the H-bonding potential. In fact, one should not attempt to extract the energetics of such interactions directly from the database statistics, dominated in this case by the $i/i + 4$ interactions.

Faced with these problems, our goal is to establish an optimal projection of the full multidimensional data set onto a space of a lower dimensionality, while preserving the essential features of the H-bonding interactions. We can use several physical arguments to our advantage. First, specification of the relevant degrees of freedom has to capture the orbital overlap aspect of the underlying energetics, with the lone pair on the O atom interacting with the antibonding σ^* orbital at the H^N atom. Therefore, we will monitor the location of the donor H^N atom within the three-dimensional coordinate frame of the acceptor's C^α−(CO)−N group. Second, application of a proper, minimal H-bonding potential should maximize the quality of the resulting structures and reproduce the correlations of the remaining variables. After inspecting various variables we have concluded that, once the position of the H^N atom is fixed within the acceptor reference frame (i.e., for given values of r, θ, ϕ), the degrees of freedom likely to be relevant for the hydrogen bonding are those describing the orientation of the C−O or H^N−N vectors of the donor frame within the coordinate frame of the acceptor group. The observed nonrandom distributions of the ϕ'' angles are exemplified by the known right-handed twist of the β-strands. Such statistics are presumably mediated by the effects unrelated to the hydrogen bonding, i.e., the nonbonded interactions, already a part of the semiempirical force fields. These correlations appear to depend more on the type of the secondary structure than on the details of the particular H-bonding geometry (see Supporting Information). For this reason, the effect of ϕ'' was not considered when deriving the HB PMF. On the other hand, possible relevance of the θ' angle is underscored by a non-negligible effect of the dipolar interaction between the CO groups of the donor and acceptor peptidyl frames.[12] Third, a proper description of the θ'' angle, describing the linearity of the hydrogen bond at the H^N atom, is also considered to be important,[13] reflecting the collinearity of the antibonding σ^* orbital at the donor group with the N−H^N vector. Since our goal is the most compact description of the interaction, we will attempt to select a minimal subset of variables whose application maximizes the resulting structural accuracy and at the same time reproduces the observed correlations exhibited by the remaining degrees of freedom.

Results and Discussion

Database of Protein Structures. A protein structural database was constructed from the set of PDB entries solved by X-ray crystallography that conform to the following criteria: resolution better than 1.8 Å, R factor ≤ 0.25, free R factor ≤ 0.30, sequence length > 50 residues, and maximum pairwise primary sequence identity < 90%. The database was built by combining the set of 500 high-quality protein structures from Richardsons' lab[13] and the 90% homology PDB_SELECT list of January, 2003.[14] The entries were analyzed for continuity of the polypeptide chain; those with missing fragments of unknown length were removed. The final database comprises ∼1500 protein chains encompassing ∼350,000 amino acid residues. Several aspects of the database statistics are illustrated in Figure 2.

Most of the data results from structures of proteins of 100−400 residues solved at resolutions between 1.4 and 1.8 Å with the most typical R and free R factors being 0.18 and 0.23, respectively. A relatively large homology cutoff value, resulting in higher signal-to-noise, did not introduce any systematic bias in the derived data, as checked against those from a 25% sequence identity subset (data not shown). All hydrogen atoms were added by the REDUCE program[15] with the H^N atom in the standard geometry within the C−N−C^α plane.

Extraction of the HB PMF. After inspecting a large number of multidimensional distributions that correlate the relevant structural variables, we have concentrated on those that exhibit the simplest shapes. Such logic has led us to four potentials that were obtained by applying the inverse Boltzmann formula to the respective distribution functions accumulated over our database. The first potential, denoted by $E(r,\theta,\phi)$, describes a position of the amide donor hydrogen in the three-dimensional (3D) coordinate frame of the acceptor peptidyl unit. The bulk of the corresponding pdf was observed to occur at O···H^N distances below 2.3−2.4 Å, supporting the previously established[11] distance cutoff criteria. The other three PMF functions, denoted by $E(\theta'|\theta)$, $E(\phi'|\phi)$, and $E(\theta''|r)$ describe the strongest observed correlations of the angular variables θ', ϕ', and θ'' with the (r,θ,ϕ) degrees of freedom specified by the first function. Backbone−backbone CO/H^N hydrogen bonds were identified according to the following criteria: an O···H^N distance less than 3.0 Å, a CO/H^N angle larger than 110°, and the angle between the C−O vectors of the donor and acceptor frames larger than 110°. This definition is slightly more stringent than the commonly used requirement for the N−H^N···O angle to exceed 90°. However, there is >99% overlap between the two sets of hydrogen bonds (excluding bifurcation), selected by these criteria below the 2.3 Å distance cutoff. To ensure that the pdf is dominated by its own degrees of freedom, bifurcated hydrogen bonds with either partner showing possible interactions with other donors or acceptors, including side-chain atoms or crystallization water molecules, were not considered. Only those geometries were selected in which the donor N atom and all atoms defining the coordinate frame around the acceptor O atom

(12) Allen, F. H.; Baalham, C. A.; Lommerse, J. P. M.; Raithby, P. R. *Acta Crystallogr., Sect. B* **1998**, *54*, 320−329. Maccallum, P. H.; Poet, R.; Milner-White, E. J. *J. Mol. Biol.* **1995**, *248*, 374−384. Maccallum, P. H.; Poet, R.; Milner-White, E. J *J. Mol. Biol*. **1995,** *248*, 361−373.
(13) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 389−408.
(14) Hobohm, U.; Sander, C. *Protein Sci.* **1994**, *3*, 522.
(15) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. *J. Mol. Biol.* **1999** *285*, 1735−1747.

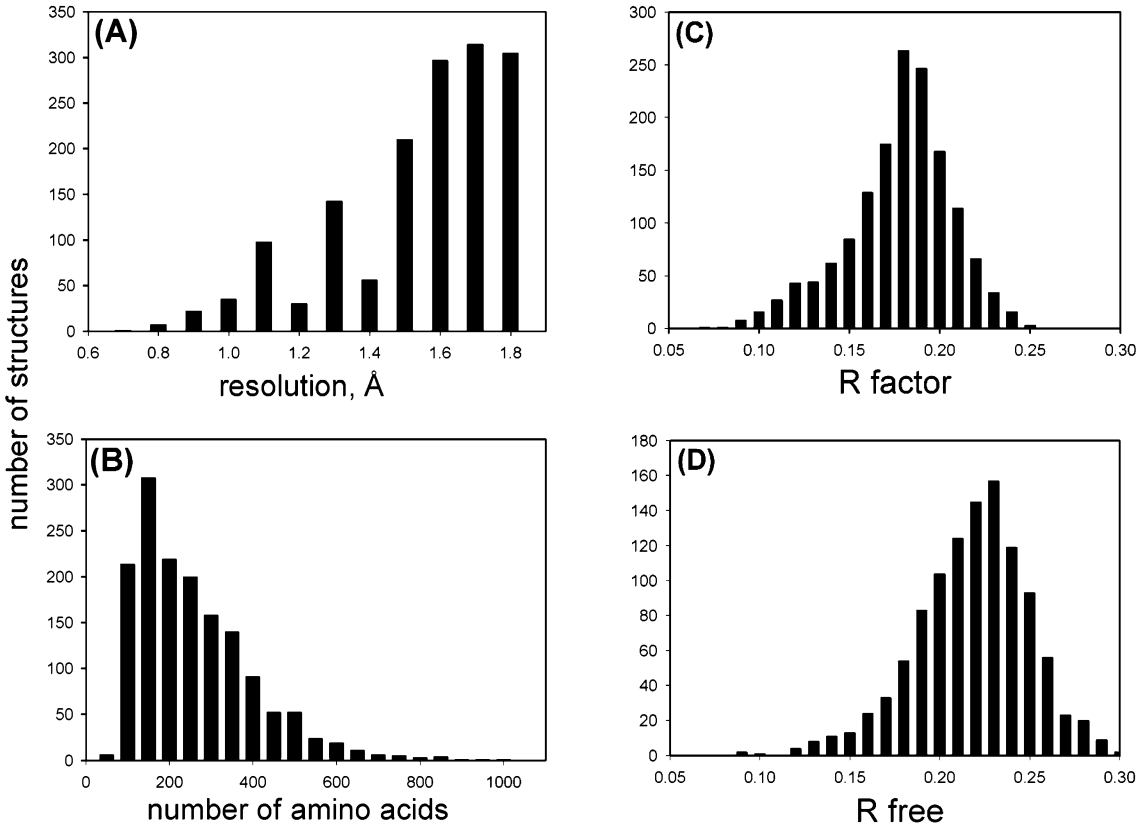

Figure 2. Histograms describing the statistics of the structural database. (A) crystallographic resolution; (B) protein primary sequence length; (C) working set *R* factor; (D) free *R* factor.

had *B*-factors less than 40, in line with previous recommendations for such work.[16]

Features of the $E(r,\theta,\phi)$ Function. H^N atomic positions were defined within a Cartesian coordinate frame of the $C^\alpha-C-O$ fragment with the *z*-axis formed by the acceptor's C−O vector and the *x*-axis by the component of the acceptor's $C^\alpha-C$ vector, orthogonal to the C−O (Figure 1). Such frame-dependent formulation of this multidimensional potential, made possible by the large size of our database, bears some resemblance to the previously reported database potential describing the base−base positional interactions observed in the nucleic acids.[17] One of the factors that led to the choice of the acceptor O atom as the origin of the coordinate system was the expected directionality of the hydrogen bond mediated by its lone pair orbitals. Another was direct observability of all three atoms used in the specification of such a system. The 3D distribution of the O$\cdot\cdot\cdot$H^N vector around the origin was accumulated on a cubic grid of 0.1 Å within a 3 Å box. Each data point was applied as a Gaussian mask. That is, the intensity contributed by each point to a given grid box was proportional to the value of the Gaussian function of the distance between the center of the box and the exact O$\cdots$H^N vector within the coordinate frame of the acceptor group. The width of the 3D Gaussian was set directly proportional to the crystallographic resolution of the structure in question: a 1 Å resolution corresponded to a 0.1 Å width. Initially, hydrogen bonds were classified according to the acceptor−donor (*i*/*j*) sequence separation: $j - i = 3$, $j - i = 4$, $|j - i| > 4$. These classes were further subdivided according to the H-bonding pattern or specific geometry. Examples of the relative H-bonding geometries of the several secondary structure classes are shown in Figure 3.

An additional lobe in the raw $i/i + 3$ distribution appearing as a secondary effect of the α-helical geometry was eliminated by the selection of the nonbifurcated geometries, as in these cases, $i/i + 4$ pattern would also be present. The resulting pure 3_{10} helical ($i/i + 3$) distribution exhibits two lobes of different intensities (∼90% right-handed), symmetrical with respect to the *zx* plane. To improve the definition of the less common left-handed geometry, the raw distribution was *y*-symmetrized by adding both (*x*,*y*,*z*) and (*x*,−*y*,*z*) points to the distribution whenever either one of those was observed within the database. The $i/i + 4$ distribution was divided into two classes: internal and N-terminal α-helix (if the $i + 1/j + 1$ hydrogen bond was also present), and C-terminal α-helix or isolated α-turn (all remaining cases). The combination of the two patterns within each such class was possible due to their apparent high geometric similarity. This classification scheme also agrees with the well-known capping patterns at the N- and C-termini of the α-helices.

Since most of the long-range hydrogen bonds occur in β-sheets, the $|j - i| > 4$ bonds were classified according to the sequence separation patterns characteristic of such structures. Those belonging to the parallel β-sheets were subdivided into internal (*i*/*j* flanked by both $j/i + 2$ and $j - 2/i$ hydrogen bonds) and edge (*i*/*j* flanked by either $j/i + 2$ or $j - 2/i$ only). The hydrogen bonds within the antiparallel β-sheets were separated

(16) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. *J. Mol. Biol.* **1999,** *285*, 1711−1733.
(17) Kuszewski, J.; Schwieters, C.; Clore, M. G. *J. Am. Chem. Soc.* **2001,** *123*, 3903−3918.

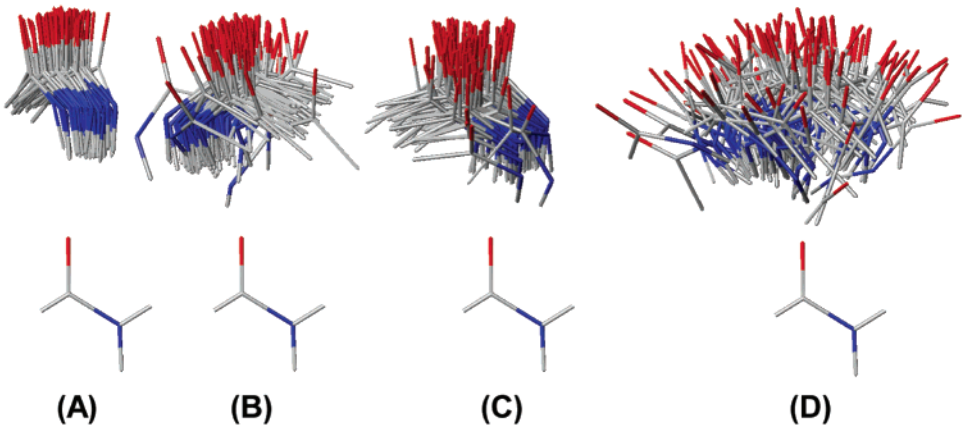

Figure 3. Relative geometries of 50 backbone−backbone hydrogen-bonded pairs, aligned at the acceptor frames, for the four most common classes of H-bonds in proteins: **(**A**)** central α-helix; **(**B**)** internal antiparallel β-sheet; **(**C**)** internal parallel β-sheet; and (D), long-range, isolated hydrogen bonds. Within the displayed $C^{\alpha}-(C{=}O)-(N-H)$ peptidyl units, O atoms are in red and N in blue. Pairs chosen correspond to the first 50 H-bond pairs of each type when searching through the ≤1.2-Å resolution database of Richardson et al.[13]

into internal or "short-cycle edge" (i/j flanked by j/i), and "long-cycle edge" (i/j flanked by $j-2,i+2$, but not by j/i). Again, all subdivisions were based on the apparent geometric similarities and differences between the distributions. The remaining long-range hydrogen bonds that exhibited none of the β-sheet flanking patterns described above were classified as "long-range, isolated". To increase the signal-to-noise ratio, some of the long-range distributions (Figure 4, panels C, D, E) were symmetrized with respect to the zx plane, as described above.

The resulting eight probability density functions were converted into the corresponding potentials via the inverse Boltzmann formula. An advantage of the PMF treatment is a possibility of establishing a potential's baseline over the areas of conformational space that show minimal variation of the pdf. These areas were defined when the $O\cdots H^N$ distance was between 3.0 and 3.2 Å. Setting the energy of the interaction potential at zero within such area makes the depths of the HB PMF equal to ∼−6 kT. Modest variation of the definition of the zero-energy region resulted in ∼0.5 kT change of the potential depths. Interestingly, assuming room temperature in the kT factor, these numbers appear quite similar to the literature estimates of the free energy of the hydrogen bond. To avoid problems with defining the zero energy on the basis of the poorly populated region in the H-bonding coordinate space, the minima of all potentials were assigned a value of −6 kT. All of the areas that exhibited energies above 0 kT were assigned an energy of zero. This appears reasonable considering that the smoothened distribution accumulated over our database is dominated by the Gaussian convolution for the energies that exceed this level. The statistics of the H-bonding distributions are listed in Table 1, and some of the slices through the corresponding PMF are shown in Figure 4.

Several interesting details emerge from these data. The optimum distances and θ angles corresponding to the potentials' minima, are similar for all classes except for 3_{10} helix, reflecting its restricted geometry. However, the 3D shapes of the distributions are quite different. The single-lobe $i/i+4$ distributions, heavily weighted by the α-helical geometry, are entirely right-handed. Antiparallel and parallel β-strands are characterized by the location of the H^N atom occurring on different sides of the zy plane, exhibiting very little overlap with each other. The maxima of these distributions are in the zx plane, consistent with the location of the carbonyl lone pair orbitals. The edge distributions of the β-strands (Figure 4, panels G and H) reveal additional lobes and appear related by a reflection with respect to the diagonal of the xy plane. They are the only distributions that exhibit a noticeable overlap between these two types of secondary structure. Three of the long-range distributions (Figure 4, panels C, D, and E) exhibit substantial symmetry with respect to the reflection in the zx plane which was exploited in the derivation of their PMF, as previously described. A particularly pronounced feature of all studied distributions is the almost complete absence of the H-bonding geometries in which the H^N atom is directly above the O atom.

There seems to be a general preference for long-range hydrogen bonds to be found on the acceptor's C^{α} side of the zy plane (with $|\phi| > 90°$). Interestingly, the antiparallel-to-parallel hydrogen-bond ratio in our database is numerically similar to the ratio of the long-range isolated hydrogen bonds in the "parallel" and "antiparallel" areas. Surprisingly, similar ratios characterize locations of several other donors with respect to the backbone carbonyl group. When translated into an energy difference between the "antiparallel" and "parallel" states, these numbers correspond to the inverse variance-weighted ∼0.38 kT (Table 2).

The difference in populations comes primarily from the smaller occupancy of the $\phi \approx 0°$ region compared to the $\phi \approx 180°$ region. A more pronounced preference for the antiparallel versus parallel β-sheet geometry, known from counting statistics within other structural databases,[18] is not incompatible with this number, being influenced by the cumulative effects of the formation of several consecutive hydrogen bonds as well as by a somewhat varying definition of what constitutes a β-sheet. As a side note, no such comparison from our data seems possible between the energies of an average "α-helical" and "β-sheet" hydrogen bond. The database counting statistics in these cases will be influenced by the entropy part of the free energy that would favor the formation of the α-helical structure as the one occurring between the partners separated by fewer residues.

Features of the $E(\theta'|\theta)$, $E(\phi'|\phi)$, and $E(\theta''|r)$ Functions. The functions describing the correlations between angular variables θ', ϕ', and θ'' with respect to the (r,θ,ϕ) variables were accumulated on rectangular grids with bin sizes of 3° in θ, θ', and θ'' dimensions, 5° in ϕ and ϕ' dimensions, and 0.05

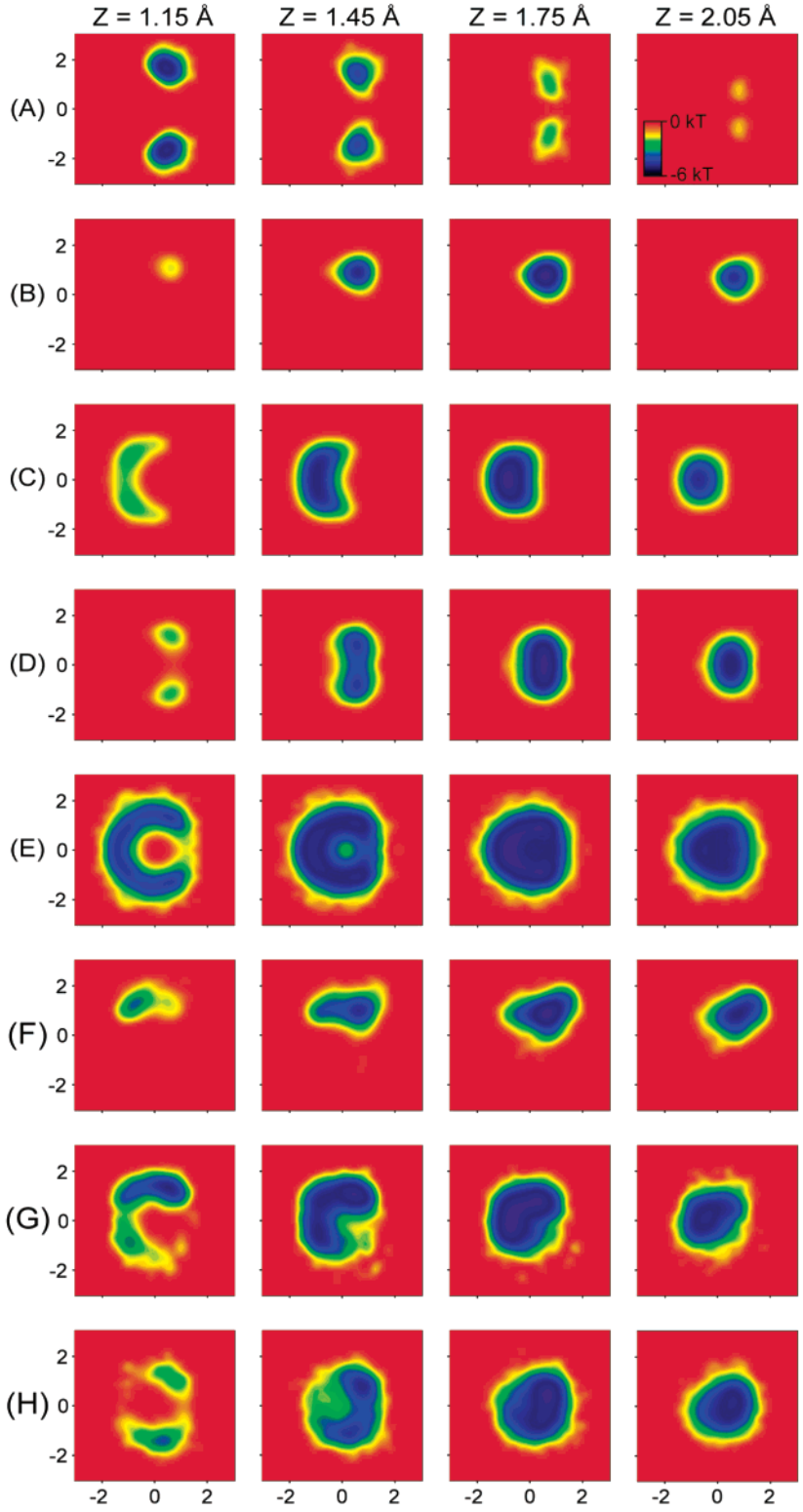

Figure 4. Slices through the three-dimensional PMF describing the location of the HN atom in reference to the coordinate frame of the CO donor group. The columns correspond to the *xy* slices at indicated distances above the O atom. The panels are: (A) 3_{10} helix; (B) central and N-terminal α-helix; (C) antiparallel *β*-sheet, central and "short-cycle" edge; (D) internal parallel *β*-sheet; (E) isolated, long-range hydrogen bond; (F) C-terminal α-helix and isolated α-turn; (G) antiparallel *β*-sheet, "long-cycle" edge; and (H), edge parallel *β*-sheet. The acceptor O atom is in the center of each square with the C−O vector pointing up. The grid markings are in Å.

Å in the *r* dimension. The $E(\theta''|r)$ functions were built separately for the $j - i = 3$, $j - i = 4$, and $|j - i| > 4$ sequence separation classes. The $E(\theta'|\theta)$ and $E(\phi'|\phi)$ functions were accumulated for each of the eight structural classes defined above for the $E(r,\theta,\phi)$ function. A reduction to only three classes of the $E(\theta''|r)$ function was possible due to a pronounced similarity between several of the eight classes. The accumulated $P(\theta',\theta)$ and $P(\theta'',r)$ functions were corrected for the volumes of the

Table 1. Statistics of the Backbone−Backbone Hydrogen-Bonding Interactions[a]

type of hydrogen bond	total number	r_{opt}, Å	θ_{opt},°	ϕ_{opt},°
3_{10} helix	2204	2.07	114	±79
α-helix, center and N-terminal	60836	1.98	152	54
α-helix, isolated turn and C-terminal	12863	2.05	149	53
antiparallel β-sheet, center and short cycle[b]	24671	1.95	154	180
antiparallel β-sheet, long cycle[c]	8934	1.94	163	165
parallel β-sheet, center	9138	1.95	164	0
parallel β-sheet, edge	9035	1.99	158	28
isolated long-range	11396	1.91	156	±169

[a] The geometric parameters in this table are those most likely to be observed within a particular class of our database, not the average values. [b] Short-cycle edge of antiparallel β-sheet: *i*/*j* H-bond flanked by *j*/*i*. [c] Long-cycle edge of antiparallel β-sheet: *i*/*j* flanked by $j-2$,$i+2$, but not by *j*/*i*.

conformational space inside each bin, dividing the raw distributions by the factors of $\sin(\theta')\cdot\sin(\theta)$ and $\sin(\theta'')\cdot r^2$, respectively. The resulting functions were then smoothened by the Gaussian convolutions of the widths equal to the bin sizes within the respective dimensions, and normalized with respect to the θ and r variables to produce the final $E(\theta'|\theta)$ and $E(\theta''|r)$. The $P(\phi'|\phi)$ function, for which no such corrections were necessary, was accumulated by the application of the Gaussian mask of 5° width, similar to the procedure used for the $E(r,\theta,\phi)$ distribution, followed by the normalization with respect to the ϕ variable. All resulting distribution functions were converted into their respective PMFs by the application of the inverse Boltzmann equation. The minima for all three types of potential for all respective classes was set to correspond to zero; the potentials were set to a constant value of 4 kT in regions that exhibited raw energies exceeding this value.

The final differences for any given class of angular potentials over the different types of H-bonding patterns are not as pronounced as those observed for the $E(r,\theta,\phi)$ function; however, the numerical differences are sufficient to warrant separation of these angular potentials. Figures 5 and 6 show examples of the three angular potentials. The features of the $E(\theta'|\theta)$ function, which are qualitatively the same for all our secondary structure classes, can be rationalized as arising from several effects. At $\theta \approx 145-180°$, the CO group linked to the donor is co-aligned with the acceptor's CO, consistent with both favorable dipolar interaction and the location of the antibonding orbital below the H-atom. At the values of θ below 145°, θ' becomes linearly correlated with θ, implying co-alignment of the O···H^N and H^N-N vectors. The correlation between the ϕ' and ϕ variables is always positive, again reflecting co-alignment of the O···H^N and H^N-N vectors.

Our formulation of the $E(\theta''|r)$ is similar to the H-bonding potential proposed by Lipsitz et al.[11e] However, our considerably larger database allows us to divide the distribution according to the donor−acceptor sequence separation, while using the PMF instead of a parametric fit.

Use of the HB PMF for Validation of Protein Structure. The usage of databases for validation is intrinsically more straightforward than for refinement. This is easily understood by considering our earlier discussion of the problems associated with conversion of the pdf into a properly formulated potential. In other words, deriving a number that describes the quality of a match between a structure and a given pdf is less challenging than establishing, on the fly, the direction of the force vector that would point from a less-than-ideal trial structure to the unknown correct geometry. On the other hand, if most of the structures already agree with the database pdf, the application of the corresponding PMF is unlikely to cause a noticeable structural improvement. Therefore, for evaluation purposes we require a set of structures that does not match the H-bonding pdf derived above.

Here, we consider the ensemble of recently solved NMR structures, which should reflect the average quality obtained in today's protein NMR structure determination. Specifically, we include all protein structures derived from solution-state NMR data that have been deposited into the RCSB Protein Data Bank between January and October 2003. We have excluded models containing non-natural amino acids or sugars, those having less than 20 residues, and those that were explicitly restrained by the data from previously solved X-ray structures, leaving us a total of 98 proteins. Only one model (the first one) was selected from each deposited bundle of structures.

The panels of Figure 7 show the distribution of the average PMF H-bonding energy per structure for two sets of models: the high-resolution X-ray database that the PMF was derived from and the set of NMR models described above. The statistics of these distributions are summarized in Table 3.

Several observations can be made from these results. The average HB PMF energy within the database (Figure 7) seems to behave like a self-averaging parameter for a given protein, with the rmsd of the distribution decreasing with the increase of the number of samples (sequence length). This allows us to establish small-protein (less than 250 residues) target values of -4.6 ± 0.3 kT and 0.51 ± 0.15 kT for the average $E(r,\theta,\phi)$ and $E(\theta''|r)$ energies per structure, by reference to the respective potentials' minima of −6.0 and 0.0 kT. In our view, these numbers should provide a faithful representation of the quality of the backbone packing. Another conclusion from studying this result is that the recently solved NMR structures differ significantly from high-resolution X-ray structures. In fact, ∼95% of the NMR structures in our sample have average H-bonding energies that are more than 2 standard deviations away from the values found in the high-resolution X-ray database, with the average NMR structure about 6 standard deviations separated from these database means. However, there clearly are also exceptions to this generalization. Notably, five out of the 98 considered structures have H-bonding parameters that resemble those of high-resolution X-ray models: PDB codes

Table 2. Positional Preference for Hydrogen Involved in a Long-Range ($|i-j| > 5$) Hydrogen Bond Relative to the H-Bond Accepting CO Group

donor	H^N	H_2O	Ser H^{γ}	Thr $H^{\gamma 1}$	Tyr H^{η}	Trp $H^{\epsilon 1}$	His $H^{\delta 1}$
% of H-bonds with $\|\phi\| > 90°$	61.0 ± 1.1	59.0 ± 0.5	63.6 ± 3.7	58.3 ± 3.9	65.0 ± 3.4	59.7 ± 5.1	48.5 ± 7.0
energy difference, kT	-0.45 ± 0.05	-0.36 ± 0.02	-0.56 ± 0.16	-0.34 ± 0.16	-0.62 ± 0.15	-0.39 ± 0.21	0.06 ± 0.28

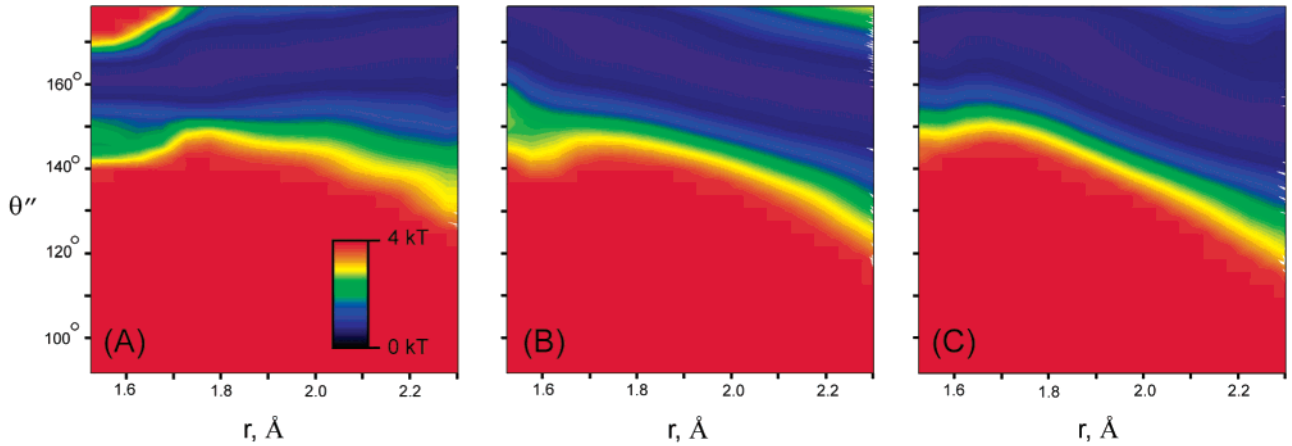

Figure 5. $E(\theta''|r)$ angular potential for the $j - i = 3$, $j - i = 4$, $|j - i| > 4$ classes of hydrogen bonds.

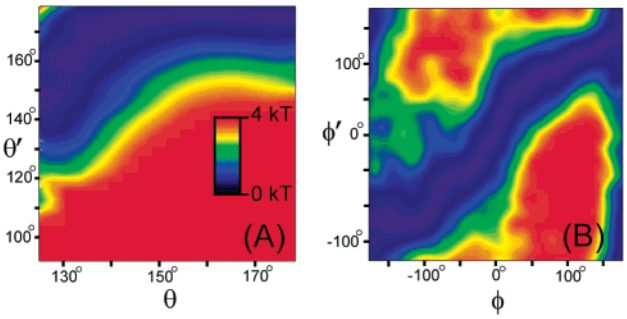

Figure 6. Potential energy surfaces for the "isolated, long-range" hydrogen bonds: (A) $E(\theta'|\theta)$; (B) $E(\phi'|\phi)$.

1NM4, 1NO8, 1OQP (chain A), 1ORL, and 1Q0W (chain A). It is the rest that are expected to benefit most from the refinement against the HB PMF potential. On the other hand, since none of the NMR structures deposited into the PDB thus far have been refined against such a potential, the average PMF energies may prove useful as independent measures of structural quality in addition to a variety of other indicators from popular validation packages such as PROCHECK[19] or WHATIF.[20] It will also be interesting to correlate "quality scores" afforded by our PMF functions with other measures of structural accuracy.

Optimization of the PMF in Structure Determination. By itself, demonstration of significant differences between the hydrogen-bond energies in structures solved by NMR and our X-ray reference data set does not guarantee that the application of a PMF during NMR refinement will actually improve the structural accuracy; thus far, the differences in such statistics simply suggest this to be a possibility. To test whether improved hydrogen-bond potentials resulting from application of our PMF force field could actually benefit structural accuracy, the HB PMF and their spatial derivatives were encoded in CNS[21] and XPLOR-NIH[22] packages. There are several problems that need to be resolved: selection of the relevant degrees of freedom, balancing the relative strengths of the terms within the resulting PMF, and balancing our PMF as a whole with respect to the rest of the semi-empirical X-PLOR/CNS force field. Our initial

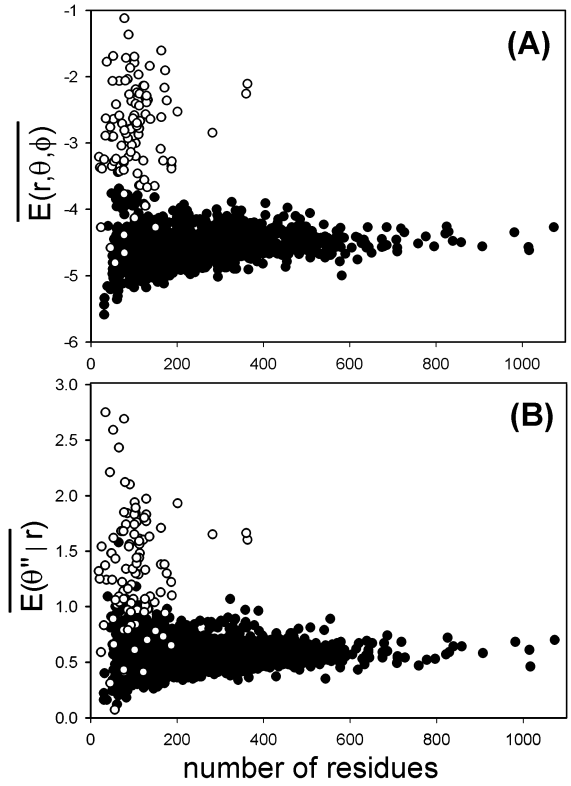

Figure 7. Average energies of H-bonding PMF per structure (units of kT). (A) $E(r,\theta,\phi)$; (B) $E(\theta''|r)$. Filled circles denote proteins in the high-resolution X-ray database. Open circles denote NMR structures solved in 2003, selected as discussed in the text.

Table 3. Statistics of the Backbone−Backbone Hydrogen-Bonding Interactions[a]

type of potential	$E(r,\theta,\phi)$	$E(\theta'\|\theta)$	$E(\phi'\|\phi)$	$E(\theta''\|r)$
X-ray database	-4.6 ± 0.2	0.66 ± 0.17	0.59 ± 0.14	0.53 ± 0.14
X-ray database, fewer than 250 residues	-4.6 ± 0.3	0.65 ± 0.19	0.59 ± 0.15	0.51 ± 0.15
NMR-2003 set	-2.8 ± 0.7	1.0 ± 0.3	1.2 ± 0.3	1.4 ± 0.5

[a] Average and standard deviations over each database of the average energy of a hydrogen bond per structure are reported in kT units.

(18) Ruczinski, I.; Kooperberg, G.; Bonneau, R.: Baker, D. *Proteins* **2002**, *48*, 85−97.
(19) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S., Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26*, 283.
(20) Vriend, G. *J. Mol. Graph.* **1990**, *8*, 52−56. Hooft, R. W. W.; Vriend, G.; Sander, S.; Abola, E. E. *Nature* **1996**, *381*, 272.
(21) Brunger, A. T.; Adams, P. D.; Clore, G. M.; Delano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J.-S.; Kuszewski, J.; Nilges, N.; Pannu, N. S.; Read, R.J.; Rice, L. M.; Simonson, T.; Warren, G. L. *Acta Crystallogr., Sect. D* **1998**, *54*, 905−921.
(22) Schwieters, C. D.; Kuszewski, J.; Tjandra, N.; Clore, G. M. *J. Magn. Res.* **2003**, *160*, 66−74.

tests were done on subsets of the experimental data (deposited NOE and dihedral angle restraints) for ubiquitin (PDB code 1D3Z). To bring the number of NOEs per residue to a value typical of an average NMR structure, we have randomly reduced the number of distance restraints to 20% of all deposited NOEs. An extensive fine-tuning of the potential was done by utilizing the NMR data sets of the B1 domain of protein G (GB1, PDB code 2GB1) and of Barstar (PDB code 1AB7). The same refinement protocol was performed for all test cases in this study: 20 structures were calculated resulting from 100 ps in vacuo Cartesian simulated annealing runs, with a linear temperature schedule from 1000 to 1 K and all atomic masses set to 25 amu. Nonbonded interactions were modeled by a simple repulsive-only term with all atomic van der Waals radii scaled down by a factor of 0.8. In addition to the HB PMF, the force field consisted of energy terms for bonds, angles, improper angles, nonbonded interactions, and experimental restraints—NOEs and dihedral angles. Soft square-well potentials were used for the distance restraints and quadratic, flat bottom potentials for the dihedral angle restraints. Initially, we were hoping to construct a single potential function that would describe every backbone−backbone hydrogen bond irrespective of its secondary structure pattern. However, during preliminary tests on the ubiquitin data we were unsuccessful in formulating such a pseudoenergy term that would yield a substantial improvement in structural quality. The origin of the problem was that the derived potential, similar to the "long-range" isolated one (Figure 4, panel E), was not sufficiently restrictive. For example, a substantial number of conformations with "antiparallel β" hydrogen bonds having $|\phi| < 90°$ or "parallel β" ones with $|\phi| > 90°$ were obtained, as opposed to the corresponding database distributions (panels C and D of Figure 4). This illustrates inherent difficulties in formulating a useful potential from a database, as a statistic sampled over the whole database might actually be a weighted average of several distinct classes. For this reason the potential was split according to the type of the secondary structure and the edge position inside each such pattern, which ultimately proved a viable solution. The recognition of H-bonding pairs and the type of hydrogen bonding (e.g., sheet, helix, etc.) is carried out in a fully automated manner, without user input. However, our software implementation also allows explicit definition of donor−acceptor pairs.

Separation of the H-bonding potential into distinct classes is entirely consistent with its statistical origin, reflecting the averaged effects of the additional degrees of freedom. This indicates that the features of our class-separated potentials should not be over-interpreted as to represent solely the effects of H-bonding. Even though the underlying H-bond potential is likely to be the same for all these classes, the apparently different local environments exert their influence on the variables that we monitor. In the ideal world, where both sampling and the empirical force field would truthfully reproduce all of the energetic aspects complementary to the hydrogen bonding, such separation would not have been necessary. However, given the approximate and simplistic nature of the force fields common in the structure calculation, the classification scheme provides a simple method to account simultaneously for both the H-bond potential and to partially overcome the deficiencies in the local force field.

At this stage, the main problem left involves balancing of the relative importance of each of the four possible terms: $E(r,\theta,\phi)$, $E(\theta'|\theta)$, $E(\phi'|\phi)$, and $E(\theta''|r)$. To establish such ordering, simulations on the ubiquitin data were performed with only one of the four PMF terms present. The accuracy of the resulting structures, measured as the backbone rmsd to the X-ray model, was compared to a reference calculation in which no H-bonding terms were present. This allowed ranking of the four terms in order of importance: $E(r,\theta,\phi) > E(\theta''|r) > E(\theta'|\theta) > E(\phi'|\phi)$. We then selected the $E(r,\theta,\phi)$ potential and ran simulations on the ubiquitin, GB1, and Barstar data, which, in addition to this term, included each of the remaining three, comparing it to a new reference calculation in which only $E(r,\theta,\phi)$ was active. A grid of potential strength values was sampled for each term to reproduce the average PMF energies compatible with our database. The results of such calculations confirm the previously established relative rankings already given above. At this stage, an increase of the structural accuracy was observed with the $E(r,\theta,\phi) + E(\theta''|r)$ potential with respect to the $E(r,\theta,\phi)$-only simulation; neither of the $E(\theta'|\theta)$ or $E(\phi'|\phi)$ terms produced such improvements when used in combination with the $E(r,\theta,\phi)$. Thus, the final potential is a function of four variables, arranged as $E(r,\theta,\theta'',\phi) = k_1E(r,\theta,\phi) + k_2E(\theta''|r)$. In a first round of applying this potential to the calculation of NMR structures from their original input restraints, the improvement in structural accuracy with respect to the X-ray models, as estimated from their backbone coordinate rms difference, was only modest, ranging from 0.07 Å (GB1) to 0.16 Å (ubiquitin).

The strength of the PMF with respect to the empirical force field was optimized to yield average energies of -4.6 ± 0.3 and 0.6 ± 0.2 kT for the $E(r,\theta,\phi)$ and $E(\theta''|r)$ functions within the NMR structures, respectively. Obtaining such energies required setting the relative $E(r,\theta,\phi)$:$E(\theta''|r)$ balance between 2:1 and 4:1, yielding the potential depths of 1.2−1.8 kcal/mol for $E(r,\theta,\phi)$ and 0.2−0.6 kcal/mol for $E(\theta''|r)$ with respect to their baselines. Combination of these values is not expected to match the literature estimates of the H-bond strength; rather, it is a consequence of balancing with the rest of the semiempirical force field, particularly affected by quality, nature, and the amount of the experimental restraints. Application of our potential within other molecular dynamics packages or with other types of the experimental data would require re-optimization of these force constants.

Although our analysis of the relevance of the $E(\theta'|\theta)$ potential, when applied in addition to the $E(r,\theta,\phi) + E(\theta''|r)$, did not result in a statistically meaningful improvement, it must be noted that the structural effects of such multidimensional potentials are rather subtle and somewhat variable, depending on the particular protein and the specifics of the experimental data set. We therefore cannot rule out that a further small improvement is attainable when including such terms combined with optimization versus a much larger data set.

Application of the PMF to Protein Structure Refinement. To further evaluate the effect of the PMF, we have applied it to the refinement of 10 protein structures previously solved by solution NMR, for which high-quality X-ray reference models were also available. The selected set of proteins ranges in size from 56 to 189 amino acids and represents a variety of

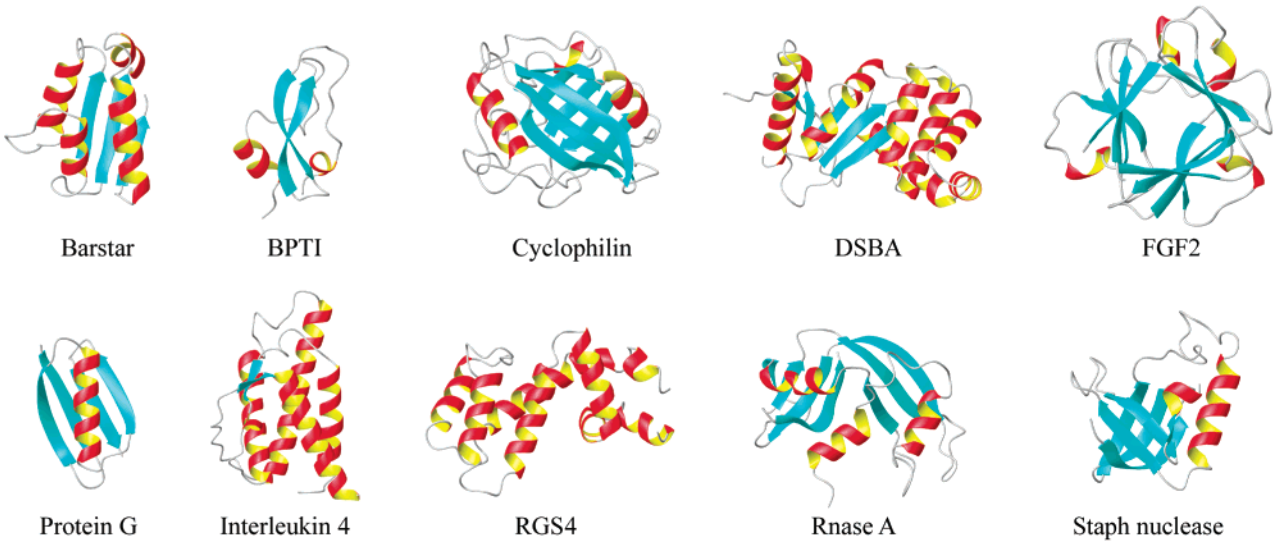

Figure 8. Structures of the 10 proteins used to test the application of the H-bonding PMF.

Table 4. Input Data Statistics for the Structure Refinement Test Cases

		X-ray[23]		NMR[24]	
protein	number of residues	PDB ID	resolution (Å)	PDB ID	NOEs/dihedrals per residue[c]
Barstar	89	1A19	2.76	1AB7	17.6/0.7
BPTI	58	5PTI	1.09	1PIT	11.1/2.0
DSBA	189	1FVK	1.70	1A24	9.5/1.2
GB1	56	1PGB	1.92	2GB1	16.5/1.7
FGF2	155	1BFG	1.60	1BLD	16.4/2.0
IL4	133	2INT	2.40	1BCN	6.9/1.3
cyclophilin	164	2CPL	1.63	1CLH	14.9/0.0
RGS4	129	1AGR	2.80	1EZY	15.1/0.0[a]
Rnase A	124	7RSA	1.26	2AAS	11.7/0.0[b]
SNase	103	2SNM	1.97	2SOB	8.0/0.5

[a] The deposited set of experimental restraints for RGS4 does not include 431 dihedral angles, 132 H^N-H^α *J*-couplings and 270 C^α and C^β shifts, listed as restraints in the original publication. [b] The deposited set of experimental restraints for Rnase A does not include 42 dihedral angles, listed as restraints in the original publication. [c] Except for a weaker NOE force constant (5 kcal/Å^2) and the absence of radius-of-gyration and Ramachandran terms, force constants are those listed in ref 3d.

architectures, from entirely α-helical to completely β-sheet (Figure 8). Statistics of the structures used in the test are shown in Table 4.

The amount and nature of experimental restraints used to generate these models is typical of the bulk of the NMR structures deposited in the PDB. The refinement protocol was the same as that applied for the optimization of the PMF definition.

A striking observation upon the initial application of the H-bonding potentials was that in many structures this led to a significant increase in persistent NOE restraint violations. Detection of such violations can be facilitated by using a very soft NOE potential with the force constant of ~5 kcal/Å^2, compared to the regular 20−50 kcal/Å^2. This increase in NOE violations was not unexpected since our PMF tends to move the models away from the original geometries that had been optimized for agreement with the NOE restraints. We also found that geometries associated with a large number of initial distance restraint violations generally exhibit worse-than-average PMF energies. The NOEs to be removed were taken from the output of the script that scanned the headers of the final coordinate files containing the summary of such violations. In no case were the distance or dihedral angle restraints files compared against those of the X-ray structure. Removal of the NOEs that were persistently violated led to a lowering of the H-bonding energies, accompanied by a decrease of the rmsd relative to that of the X-ray model, as well as by an increase of the number of backbone torsion angles within the "most favored" area of the Ramachandran plot, as defined by the PROCHECK package.[19] Therefore, we ran several cycles of structure refinement, removing all NOEs that were violated in more than 50% of the structures by more than 0.3 Å, until no changes in the restraints set could be made. The number of the removed NOEs ranged from 0 to ~100, depending on the test case, and the number of cycles ranged between 1 and 8. We also noticed that softening of the potential energy terms that enforce planarity of the peptide group led, on average, to a ~0.02 Å decrease of the rmsd to the X-ray structure and a ~3% increase of the number of residues within the most favored area of the Ramachandran map. Therefore, we have decreased the force constants enforcing such planarity from a standard 500 kcal/rad^2 to 25 kcal/rad^2, resulting in a ~1.7° standard deviation of the ω angle from the ideal *cis/trans* geometries. The improvement, however, was only seen when our HB PMF term was active, in line with results reported by Linge at al.[25] Further lowering of the planarity force constants consistently resulted in a decrease of both the Ramachandran map quality and the agreement with the X-ray structures. This effect of the ω dihedral angle description was not unexpected since the statistics leading to our PMF were accumulated on a set of X-ray structures that, on average, exhibited an approximately ~5° rmsd from $\omega = 180°$, much higher than the

(23) Ratnaparkhi, G. S.; Ramachandran, S.; Udgaonkar, J. B.; Varadarajan, R. *Biochemistry* **1998,** *37*, 6958−6966. Parkin, S.; Rupp, B.; Hope, H. *Acta Crystallogr., Sect. D* **1996**, *52*, 18. Ke, H.; Zydowsky, L. D.; Liu, J.; Walsh, C. T. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 9483. Martin, J. L.; Bardwell, J. C.; Kuriyan, J. *Nature* **1993**, *365*, 464. Ago, H.; Kitagawa, Y.; Fujishima, A.; Matsuura, Y.; Katsube, Y. *J. Biochem. (Tokyo)* **1991**, *110*, 360. Achari, A.; Hale, S. P.; Howard, A. J.; Clore, G. M.; Gronenborn, A. M.; Hardman, K. D.; Whitlow, M. *Biochemistry* **1992**, *31*, 10449. Walter, M. R.; Cook, W. J.; Zhao, B. G.; Cameron R., Jr.; Ealick, S. E.; Walter R. L., Jr.; Reichert, P.; Nagabhushan, T. L.; Trotta, P. P.; Bugg, C. E. *J. Biol. Chem.* **1992**, *267*, 20371. Tesmer, J. J.; Berman, D. M.; Gilman, A. G.; Sprang, S. R. *Cell* **1997**, *89*, 251. Svensson, L. A.; Sjolin, L.; Gilliland, G. L.; Finzel, B. C.; Wlodawer, A. *Proteins: Struct., Funct., Genet.* **1986**, 1, 370. Stites, W. E.; Gittis, A. G.; Lattman, E. E.; Shortle, D. *J. Mol. Biol.* **1991**, *221*, 7.

Table 5. Effect of the HB PMF on the Accuracy and Precision of NMR Structures

protein	NMR[23] (original deposition) rmsd to X-ray	NMR (no HB PMF) rmsd to X-ray	NMR (no HB PMF) rmsd to mean	NMR (with HB PMF) rmsd to X-ray	NMR (with HB PMF) rmsd to mean	residue range
Barstar	1.50 ± 0.11	1.50 ± 0.04	0.24	0.98 ± 0.05	0.28	1−89
BPTI	0.84 ± 0.06	0.89 ± 0.06	0.31	0.87 ± 0.06	0.27	2−56
cyclophilin	1.53 ± 0.09	1.22 ± 0.04	0.32	0.86 ± 0.04	0.30	5−9,15−35,41−56, 86−90,101−109, 124−136,158−164
DSBA	2.29 ± 0.14	2.23 ± 0.23	0.78	1.92 ± 0.19	0.71	6−187
FGF2	1.00 ± 0.05	1.04 ± 0.06	0.28	0.70 ± 0.03	0.24	29−152
GB1	1.16 ± 0.06	1.04 ± 0.08	0.27	0.65 ± 0.06	0.25	1−56
IL4	1.80 ± 0.12	1.74 ± 0.14	0.36	1.60 ± 0.11	0.34	7−39, 44−130
RGS4	1.95 ± 0.11	2.24 ± 0.22[a]	0.71	2.20 ± 0.13	0.51	6−132
Rnase A	1.09 ± 0.14	1.02 ± 0.09[b]	0.40	0.86 ± 0.05	0.31	5−123
SNase	2.64 ± 0.19	2.46 ± 0.20	1.19	1.67 ± 0.26	1.09	7−36, 54−96

[a] Calculations without restraints for 431 dihedral angles, 132 H^N-H^α *J*-couplings and 270 C^α and C^β shifts, used in the original publication. [b] Calculations without 42 dihedral angles, listed as restraints in the original publication.

ca. 0.2° characteristic of a typical NMR structure, calculated with 500 kcal/rad^2 planarity-enforcing force constants. These effects should be particularly noticeable when both the CO and the HN within the same peptidyl group are parts of a backbone−backbone H-bonding network, such as found in the central part of α-helices and internal β-strands.

Application of our HB PMF generally improves the structural accuracy as evaluated by the decrease in backbone rmsd with respect to the X-ray model (Table 5). In this table, the "no HB PMF" column corresponds to the results of a single cycle of the structure refinement without the HB potential. In all cases, the numbers we obtain are similar to the ones quoted in the original publications, even though the original refinements include software tools, force fields, and optimization protocols that are generally different from ours. The rmsd improvement with our HB PMF scheme ranges from 0.02 to 0.79 Å, with an average of 0.31 Å. Note that the bulk of this improvement comes from modifications of the experimental constraints, revealed in automated manner when using the HB PMF. This conclusion can be made either based on cases in which there were no violations of the experimental restraints to begin with, such as ubiquitin, and on comparison of the outcome of the last cycle of our refinement in the presence and absence of the HB PMF term. In these cases the rmsd improvement due to the PMF ranged between 0.02 and 0.16 Å. In our experience, the PMF has a larger impact on the accuracy when the structure in question has a substantial β-sheet content. A possible reason for this is a favorable effect of the larger primary sequence separation of the backbone fragments whose relative orientation is tightened by such restraints. For α-helical proteins, small improvements in local geometry may result, but the PMF has no effect on the relative packing of such helices, which generally dominates the rmsd to the X-ray structures. The effect of the PMF on the structural precision is small, with the backbone rmsd to the mean decreasing by only 0.07 Å on average.

Table 6. Effect of the HB PMF on Ramachandran Map Statistics

protein	X-ray	original NMR deposition	NMR refinement (no HB PMF)	NMR refinement (with HB PMF)
Barstar	78.8[a]	70.0	75.2[a]	86.7[a]
BPTI	91.3	80.9	71.0	82.6
DSBA	94.6	73.6	73.9	77.8
GB1	90.0	81.8	82.3	92.4
FGF2	93.5	69.3	67.7	78.0
IL4	91.1	73.9	77.7	80.4
cyclophilin	87.2	42.4	59.9	70.1
RGS4[b]	86.1	94.3	82.8	89.8
Rnase A[b]	90.4	85.0	64.3	75.8
SNase	85.6	49.1	47.9	71.5

[a] The reported entries are the average percentages in the most favored area as defined by PROCHECK. [b] See footnotes to Table 4.

A second important consequence of our PMF refinement is the improvement of the quality of the backbone Ramachandran map, as described by the percentages of residues inside the most favored, allowed, generously allowed, and disallowed areas. The effect of our potentials on these statistics is summarized in Table 6.

On average, application of the PMF leads to a 10% fraction increase of residues within the most favored area. We interpret this result as a consequence of propagated correlation effects that were discussed earlier. This outcome makes physical sense, given that a large percentage of residues within the secondary structure elements are affected by our potentials. Our results therefore strongly suggest that the position of a given amino acid residue within the Ramachandran (φ,ψ) plane is to a significant extent mediated by its H-bonding interactions.

We also note that even though our structures refined with the HB PMF exhibit the average values of the PMF per structure that are consistent with the X-ray database averages (Table 3), the distributions of PMF values for the individual hydrogen bonds are wider than those seen from the X-ray database (for the ∼12,000 hydrogen bonds from the sub-atomic resolution structures in our database, the averages and standard deviations are -4.6 ± 1.2 and 0.5 ± 0.8 kT, respectively). This indicates that the PMF exerts only a weak force during the NMR structure calculation and does not force a given hydrogen bond to adopt a near-ideal geometry if experimental restraints are not compat-

(24) Wong, K.-B.; Fersht, A. R.; Freund, S. M. V. *J. Mol. Biol.* **1997**, *268*, 494−511. Wagner, G.; Braun, W.; Havel, T. F.; Schaumann, T.; Go, N.; Wuthrich, K. *J. Mol. Biol.* **1982,** *155*, 347. Clubb, R. T.; Ferguson, S. B.; Walsh, C. T.; Wagner, G. *Biochemistry* **1994**, *33*, 2761. Schirra, H. J.; Renner, C.; Czisch, M.; Huber-Wunderlich, M.; Holak, T. A.; Glockshuber, R. *Biochemistry* **1998**, *37*, 6263−6276. Moy, F. J.; Seddon, A. P.; Bohlen, P.; Powers, R. *Biochemistry* **1996**, *35*, 13552−13561. Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. *Science* **1991**, *253*, 657. Powers, R.; Garrett, D. S.; March, C. J.; Frieden, E. A.; Gronenborn, A. M.; Clore, G. M. *Science* **1992**, *256*, 1673−1677. Moy, F. J.; Chanda, P. K.; Cockett, M. I.; Edris, W.; Jones, P. G.; Mason, K.; Semus, S.; Powers, R. *Biochemistry* **2000,** *39*, 7063−7073. Santoro, J.; Gozalez, C.; Bruix, M.; Neira, J. L.; Neito, J. L.; Herranz, J.; Rico, M. *J. Mol. Biol*. **1993**, *229*, 722−734. Alexandrescu, A. T.; Gittis, A. G.; Abeygunawardana, C.; Shortle, D. *J. Mol. Biol*. **1995**, *250*, 134−143.

(25) Linge, J. P.; Williams, M. A.; Spronk, C.; Bonvin, A.; Nilges, M. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 496−506.

ible with such a conformation. However, in the absence of sufficient experimental restraints, unusual hydrogen bond geometries will be disfavored by the PMF.

A possible critique of our method could be that the rmsd to the X-ray model and the Ramachandran statistics are nonoptimal parameters for evaluating improvement in structural accuracy. As described above, a considerable fraction of the improvement in these parameters obtained with our HB PMF-based refinement results from modification of the experimental restraints. To evaluate the effect of the PMF alone, we have applied it to a case in which no experimental restraints disagree with the imposed HB potential and where experimental dipolar couplings are available to evaluate accuracy. For test purposes, we use the most recent deposition of restraint data for the B1 domain of protein G, PDB code 3GB1.[26] The details of these tests are described in the Supporting Information.

Several conclusions can be drawn from these results. Inclusion of the dipolar couplings is clearly beneficial, both by conventional measures of structural accuracy—rmsd to the X-ray structure and Ramachandran statistics—as well as from the point of view of our PMF terms when using these for validation only (i.e. when the structures are calculated without the HB PMF). Supporting Information table 1 also shows that, in the absence of dipolar coupling data, the HB PMF improves all monitored aspects of structural quality: rmsd to 1PGB decreases by 0.16 Å, and dipolar coupling *Q*-factors improve by 0.135−0.057, with the biggest effect on the H^N-N couplings (which have the lowest experimental error). When the dipolar couplings are included in the structure calculation, the advantage of the HB PMF decreases, as witnessed by the rmsd to the X-ray decreasing only by 0.08 Å, and the absence of an improvement of the Ramachandran map statistics. Remarkably, the HB PMF does not have any adverse effect on how well the structure can fit to the dipolar couplings. Therefore, this test case confirms that the HB PMF has a positive effect on the NMR-derived measures of the structural accuracy. This improvement, however, diminishes when the amount of experimental data increases.

Concluding Remarks

In this study, we have formulated a multidimensional potential describing the features of backbone−backbone hydrogen bonding in protein structures; its proposed applications are structure refinement and validation. Our description is by no means complete; however, it may help to improve our understanding of this complex phenomenon. The obtained results seem to confirm a variety of the aspects of this interaction that were not entered explicitly into our formulation, while offering new insights on its action. For example, a pronounced dependence of the potentials on the ϕ dihedral angle is observed, as well as a significant energy difference between parallel and antiparallel arrangements of the peptidyl units.

In principle, application of the HB PMF should benefit any experimental NMR structure. On one hand, our multidimensional directional interaction potentials are considerably more restrictive than the common one-dimensional distance restraints. On the other, due to their spatial proximity component, they are complementary with respect to the orientation-dependent dipolar coupling or chemical shift anisotropy restraints recorded in weakly aligned media. In practice, the favorable effects of the HB PMF potential are most noticeable for NMR structures of intermediate quality, particularly those with a substantial β-sheet content.

In our evaluation of the effect of the HB PMF on previously deposited NMR structures, much of the improved agreement relative to structures solved independently by X-ray crystallography resulted from deletion of NMR input restraints that appeared incompatible with the HB PMF. In applications where the raw experimental data from which the restraints are extracted are still available, simple deletion of persistently violated restraints without inspection of the underlying data is of course unacceptable. However, in these cases, the HB PMF will prove useful in identifying restraints that are either too tight, or misassigned.

As demonstrated by our 3GB1 results, the application of the HB PMF is less beneficial for high-quality cases that are already very well defined by the available experimental information. It is also unlikely that it will be useful when serious mistakes in the protein fold or a significant number of resonance misassignments are present. As applied here, its intended use is simply as a refinement tool working in reference to an already reasonably well (and correctly)-defined structural model. Whether application of the PMF can be adapted to become beneficial at earlier stages of the NMR structure calculation process would require considerable further work and falls beyond the scope of the present study. Application of the HB PMF should be entirely compatible with the use of experimental NMR information on *J*-coupling interactions through the hydrogen bond.[27]

Another possible application of the HB PMF is validation of experimental structures by this new measure of quality. The HB PMF also is expected to become useful in improving all types of molecular models, with possible additional applications to low-resolution X-ray crystallography and ab initio or homology-based protein structure prediction.

Software Availability. The X-PLOR version of the software is a part of the Xplor-NIH package; the CNS version is available upon request from Alexander Grishaev.

Acknowledgment. We thank Drs. Marius Clore, Gerhard Hummer, and Attila Szabo for valuable discussions and comments.

Supporting Information Available: PMF setup parameters and the simulated annealing input file; validation of the effect of the HB PMF by dipolar couplings; cross sections through the raw data showing correlations between ϕ, ϕ', and ϕ''. This material is available free of charge via the Internet at http://pubs.acs.org.

JA0319994

(26) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 2337−2338.

(27) Dingley, A. J.; Grzesiek, S. *J. Am. Chem. Soc.* **1998**, *120*, 8293−8297. Cornilescu, G.; Ramirez, B. E.; Frank, M. K.; Clore, G. M.; Bax, A. **1999**, *121*, 6275−6279.