Improvements and Extensions in the Conformational Database Potential for the Refinement of NMR and X-ray Structures of Proteins and Nucleic Acids

JOHN KUSZEWSKI, ANGELA M. GRONENBORN,* AND G. MARIUS CLORE*

Laboratory of Chemical Physics, Building 5, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520

In a recent paper (*1*), we introduced the concept of a conformational database potential involving dihedral angle relationships as a method for improving the quality of structures generated from NMR and medium-resolution X-ray data. The aim of employing this potential in structure refinement by simulated annealing is to circumvent the poor discriminatory power of conventional nonbonded interaction terms (either attractive–repulsive or purely repulsive) between high and low probability local conformations. This is achieved by biasing the sampling during simulated annealing refinement to conformations that are energetically possible by limiting the choices of dihedral angles to those that are known to be physically realizable. In the previous work (*1*), the conformational database potentials were generated from two protein crystal structure databases: the PROCHECK database (*2*) and a backbone-dependent rotamer database (*3*) derived from 160 and 170 X-ray structures, respectively, refined at a resolution of 2.0 Å or better. In the present Communication, we extend and improve the protein conformational database potential by employing a new database of 70 diverse X-ray structures refined at 1.75 Å or better (*4*). Because of the higher quality of the structures present in this database, we are able to incorporate the dependence of the peptide covalent geometry on the ϕ and ψ backbone torsion angles, and to partition the ϕ/ψ, $\phi/\psi/\chi_1$, χ_1/χ_2, and $\chi_1/\chi_2/\chi_3$ distributions into different groups according to amino acid types. In addition, the same concept is used to generate conformational database potentials for nucleic acids.

Protein backbone and side-chain torsion angles, as well as backbone bond angles, were obtained from the database of 70 highly refined, high-resolution (<1.75 Å) protein crystal structures assembled by Karplus (*4*). Residues with temperature factors >25 Å^3 and those bordering a cis peptide bond were eliminated from the analysis. The resulting torsion angle correlations (see Table 1) were processed into potential of

mean force energy grids in the manner described previously (*1*). The potential of mean force is simply given by

$$E_{DB}(i) = -k_{DB}\log P_i, \qquad [1]$$

where P_i is the probability of occurence of a particular conformation, and k_{DB} a scale factor. Several features of the new conformational database potentials are noteworthy. In the original version (*1*), the ϕ/ψ distributions were partitioned into three groups: Gly, Pro, and all other residues. The higher resolution structure database, however, permits a further partitioning of the non-Gly and non-Pro residues into four groups: residues with a hydrogen bond donor or acceptor in the γ or δ position, residues preceeding a proline, β-branched residues, and the remainder. Subtle but consistent differences in the ϕ/ψ distributions are observed between these groups (see Figs. 1A and 1B). In the original application (*1*), the side-chain torsion angle correlations were limited to one-dimensional potential energy surfaces for the χ_1, χ_3, and χ_4 angles, two-dimensional potential energy surfaces of χ_1/χ_2, and three-dimensional potential energy surfaces of $\phi/\psi/\chi_1$ for each residue. In the current conformational database potentials, all one-dimensional potential energy surfaces have been eliminated with the exception of χ_4 for Lys and Arg, and have been replaced by two-dimensional χ_1/χ_2 and χ_2/χ_3 potential energy surfaces and three-dimensional $\phi/\psi/\chi_1$ (with distinct groupings; see Table 1) and $\chi_1/\chi_2/\chi_3$ potential energy surfaces. Examples of some two- and three-dimensional potential energy surfaces are shown in Figs. 1 and 2, respectively.

In the case of the Leu χ_1/χ_2 distributions, a number of potentially false rotamers may be present in the database owing to experimental ambiguity for the χ_1/χ_2 values of Leu residues in structures derived from medium resolution X-ray analyses. In particular, the well-staggered rotamer with χ_1/χ_2 values near $-60°/180°$ occupies approximately the same space as a poorly staggered rotamer with χ_1/χ_2

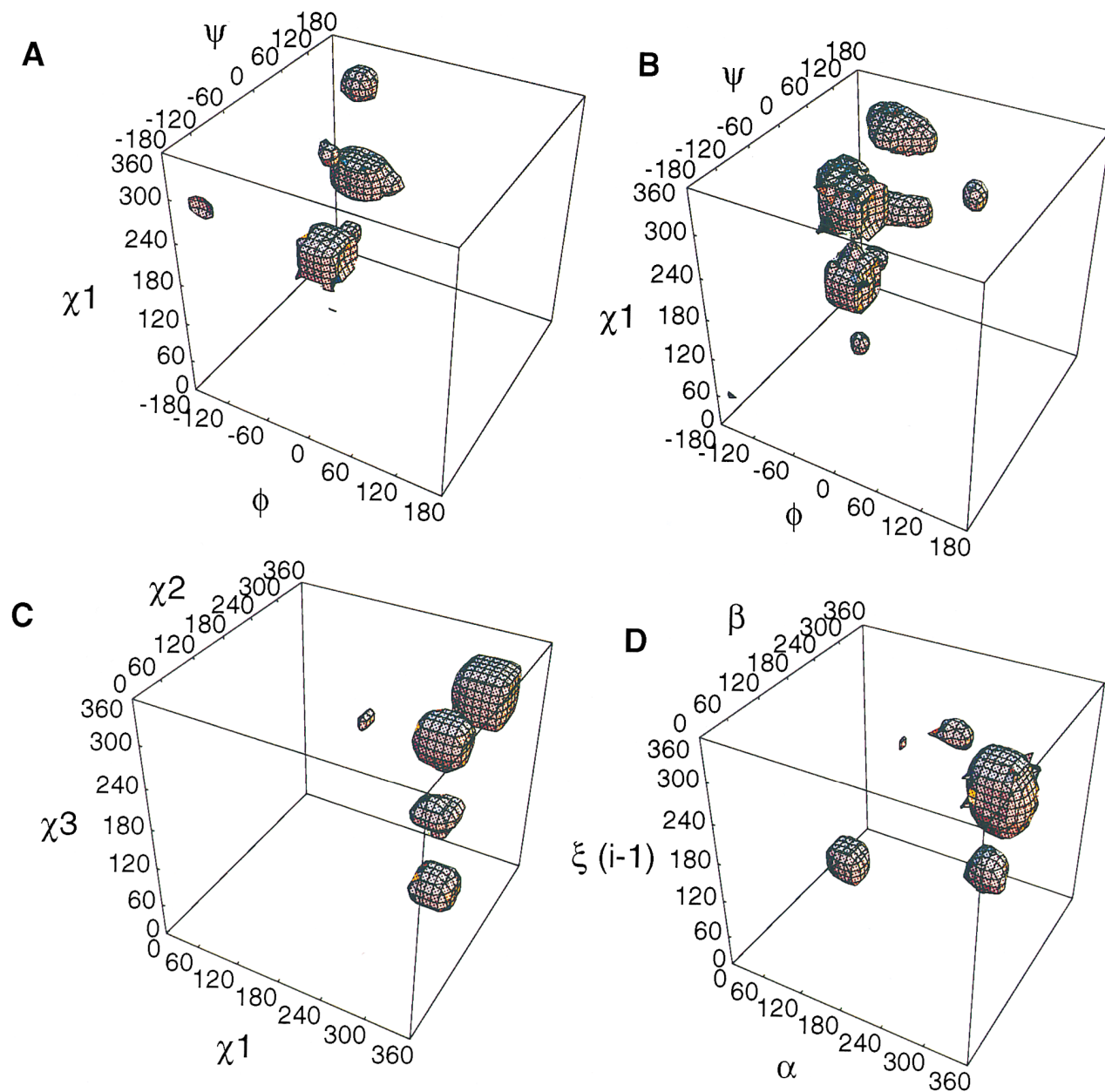* To whom correspondence should be addressed.
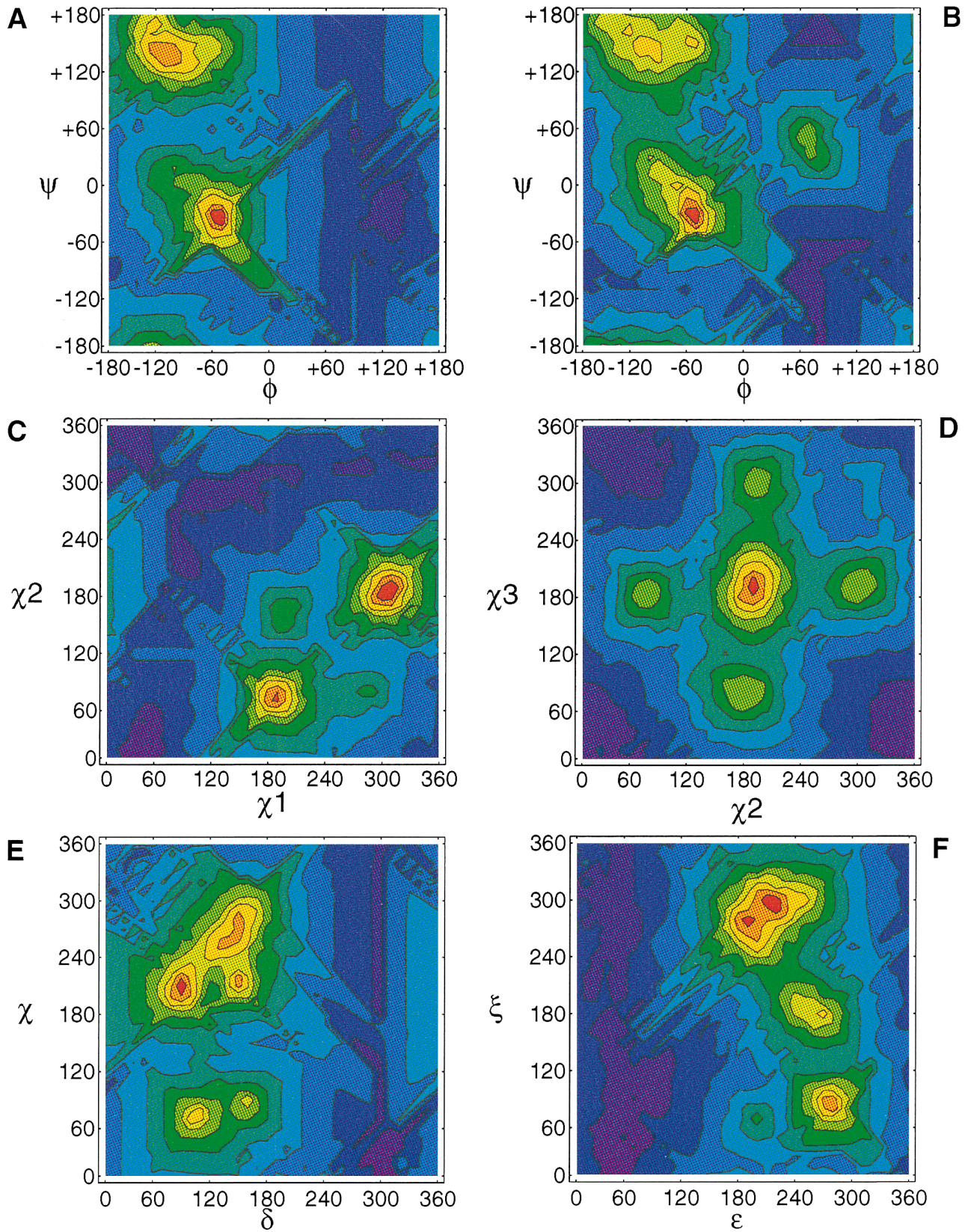
FIG. 2. Examples of three-dimensional potential energy surfaces. All surfaces are contoured at a level 40% higher than the minimum energy [i.e., at $0.4 (E_{max} - E_{min}) + E_{min}$]. (A) $\phi/\psi/\chi_1$ for Val; (B) $\phi/\psi/\chi_1$ for Arg, Lys, Met, Glu, and Gln; (C) $\chi_1/\chi_2/\chi_3$ for Met; and (D) $\alpha(i)/\beta(i)/\zeta(i-1)$ for nucleic acids.

values near $-90°/20°$, such that at a resolution worse than ~1.8 Å these two models are difficult to distinguish based on electron density alone (P. A. Karplus, personal communi-cation). Similarly, an alternate fit with χ_1/χ_2 values near $-150°/-140°$ exists for the well-staggered Leu rotamer with χ_1/χ_2 values near $180°/60°$. The most likely Leu side-chain

FIG. 1. Examples of two-dimensional potential energy surfaces. All surfaces are color-coded from red (minimum energy) to violet (maximum energy). (A) ϕ/ψ for Val and Ile; (B) ϕ/ψ for Ala, Arg, Cys (oxidized), Gln, Glu, Leu, Lys, Met, Phe, Trp, and Tyr; (C) χ_1/χ_2 for Leu; (D) χ_2/χ_3 for Lys; (E) δ/χ for nucleic acids; and (F) ϵ/ζ for nucleic acids.

TABLE 1
Summary of Database Potentials

A. Proteins

Torsion angles

One-dimensional

χ_4: Arg, Lys

Two-dimensional

ϕ/ψ: Gly, Pro, X-Pro, H-bonding,[a] Val/Ile, rest

χ_1/χ_2: Leu, Ile, Gln/Glu, Arg/Lys/Met, Asn, Asp, Cys(ox), His, Trp, Phe/Tyr

χ_2/χ_3: Met, Gln, Glu, Lys, Arg

Three-dimensional

$\phi/\psi/\chi_1$: Val, Ile, Phe/Tyr/Trp, Leu, X-Pro, Gln/Glu/Arg/Lys/Met, Cys(red)/His/Asp/Asn, Ser, Thr, Cys(ox), Pro

$\chi_1/\chi_2/\chi_3$: Gln, Glu, Arg, Lys, Met

$\phi_1/\psi_i/\phi_{i\pm1,2,3,4}$[b]

$\phi_i/\psi_i/\psi_{i\pm1,2,3,4}$[b]

Four-dimensional

$\phi_i/\psi_i/\phi_{i\pm1,2,3,4}/\psi_{i\pm1,2,3,4}$[b]

Backbone-dependent geometry[c]

τ_3, ϕ, ψ: Pro,X-Pro, rest

ω, ϕ, ψ: Gly, Pro, X-Pro, rest

ω_o, ϕ, ψ: Gly, Pro, X-Pro, rest

B. Nucleic acids

Torsion angles

Two-dimensional: α/β, α, γ, α/δ, α/ε, α/ζ, α/χ, β/γ, β/δ, β/ε, β/ζ, β/χ, γ/δ, γ/ε, γ/ζ, γ/χ, δ/ε, δ/ζ, δ/χ, ε/ζ, ε/χ, ζ/χ

Three-dimensional: $\alpha/\beta/\gamma$, $\beta/\gamma/\delta$, $\gamma/\delta/\varepsilon$, $\delta/\varepsilon/\chi$, δ/χ, $\gamma/\delta/\chi$, $\alpha/\beta/\zeta(i-1)$, $\alpha/\zeta(i-1)/\varepsilon(i-1)$

---

[a] These refer to residues with a hydrogen-bond donor or acceptor in the γ or δ position (Ser, reduced cysteine, Asp, Asn, Ser, and Thr).

[b] The scale factor used for the three-dimensional interresidue $\phi_i/\psi_i/\phi_{i\pm1,2,3,4}$, $\phi_i/\psi_i/\psi_{i\pm1,2,3,4}$ potentials and the four-dimensional interresidue $\phi_i/\psi_i/\phi_{i\pm1,2,3,4}/\psi_{i\pm1,2,3,4}$ potentials must be set to a value at least 10- and 6-fold lower, respectively, than that for the intraresidue potentials, since otherwise undesirable bias in the structures may be introduced. Typically, the final value of the scale factor for the intraresidue conformational database potentials is set to 1.0.

[c] The force constants for the backbone-dependent geometry potentials are treated separately from the scale factors for the conformational database potentials.

conformations are those with χ_1/χ_2 values corresponding to the well-staggered conformations, and it has been suggested that the two alternate-fit positions found in some X-ray structures are incorrect (P. A. Karplus, personal communication). Although these alternate-fit positions are rare in the high-resolution (1.75 Å or better) structure database, a small number, constituting about 3% of the total, are still present. We have therefore corrected our database potentials accordingly. In particular, the χ_1 and χ_2 angles for leucines with $240° \leq \chi_1 \leq 285°$ and $-10° \leq \chi_2 < 45°$ were changed by adding 40° to χ_1 and 160° to χ_2; the χ_1 and χ_2 angles for leucines with $190° < \chi_1 < 240°$ and $190° < \chi_2 < 240°$ were changed to subtracting 30° from χ_1 and 150° from χ_2 (P. A. Karplus, personal communication). After this correction, two minor clusters, in addition to the major well-staggered $t/g+$ and $g-/t$ conformations, are present. The two minor clusters are located in the t/t and $g-/g+$ χ_1/χ_2 regions and are centered around values of 180°/155° and −85°/+70°, respectively. These represent true alternate Leu sidechain conformations, and their small deviation from perfect staggering indicates some strain involved in adopting these conformations, consistent with their low frequency of occurrence (see Fig. 1C).

In addition to local conformation, it is also possible to include longer-order conformational information derived from the database by creating interresidue three-dimensional $\phi_i/\psi_i/\phi_{i\pm1,2,3,4}$ and $\phi_i/\psi_i/\psi_{i\pm1,2,3,4}$ and four-dimensional $\phi_i/\psi_i/\phi_{i\pm1,2,3,4}/\psi_{i\pm1,2,3,4}$ potential energy surfaces. If these terms are incorporated into the target function used for simulated annealing refinement, it is essential to ensure that the scale factor used for the three- and four-dimensional interresidue terms is at least 10- and 6-fold lower, respectively, than that used for the intraresidue conformational database potential terms, in order to prevent the introduction of undesirable bias in the resulting structures. Typically, for the current conformational database potential, the optimal scale factor for the intraresidue conformational database potential terms is 1.0.

Inspection of the database of high-resolution crystal structures also reveals a distinct dependence of the peptide covalent geometry on the backbone ϕ/ψ angles (4). In particular, the $\tau3$ (N–Cα–C) bond angle and the ω peptide bond torsion angle (Cα–N–C–Cα) vary by about ±5° from their standard values in a manner that is dependent on ϕ and ψ (4). These result in relatively large atomic RMS shifts which may impact the accuracy of NMR structures. Thus, for example, a $\tau3$ angle of 5° larger or smaller than the standard value of ~110° corresponds to a shift of 0.3 Å in the relative positions of the 1–3 related Cα atoms (4).

To incorporate the variations in backbone $\tau3$ and ω angles as a function of ϕ/ψ in simulated annealing refinement, we derived expectation values for the $\tau3$ and ω angles from the database as follows. For every $10° \times 10°$ ϕ/ψ bin, the number of examples present in the database is counted. If that number is greater than a predetermined cutoff (10 examples, in this work), the expectation values of $\tau3$ and ω at that bin are calculated as the mean of the $\tau3$ and ω angles of the examples seen in that bin. If the number of examples seen in the current bin is smaller than the cutoff, local average values of $\tau3$ and ω are calculated over that bin's neighbors, as described in (1). The expectation values were then incorporated into a new potential term E_{angledb} defined as

$$E_{\text{angledb}} = = k_{\tau 3} \sum \{\tau3 - \tau3_{\text{expectation}}(\phi, \psi)\}^2 + k_{\omega} \sum \{\omega - \omega_{\text{expectation}}(\phi, \psi)\}^2, \qquad [2]$$

TABLE 2
Effects of Conformational Database Refinement on the NMR Structures
of the Oligomerization Domain of p53 and Interleukin-4 (IL-4)[a]

	p53		IL-4	
	With database	No database	With database	No database
RMS deviations from experimental restraints[b]				
Interproton distances (Å)	0.027 ± 0.001	0.026 ± 0.002	0.030 ± 0.006	0.020 ± 0.005
Torsion angles (°)	0.037 ± 0.052	0.147 ± 0.056	0.196 ± 0.103	0.159 ± 0.088
$^3J_{HN\alpha}$ (Hz)	0.65 ± 0.04	0.36 ± 0.07		
$^{13}C\alpha$ (ppm)	0.81 ± 0.07	0.75 ± 0.09		
$^{13}C\beta$ (ppm)	0.88 ± 0.03	0.92 ± 0.04		
Deviations from idealized covalent geometry				
Bonds (Å)	0.004 ± 0.000	0.004 ± 0.000	0.004 ± 0.000	0.003 ± 0.000
Angles (°)	0.491 ± 0.021	0.520 ± 0.028	0.493 ± 0.016	0.287 ± 0.015
Improper torsions (°)[c]	0.615 ± 0.123	0.292 ± 0.068	0.466 ± 0.040	0.283 ± 0.040
Deviations from $\tau3$ and ω expectation values				
$\tau3$	0.578 ± 0.053	1.72 ± 0.09	0.61 ± 0.06	2.31 ± 0.08
ω	0.762 ± 0.071	1.26 ± 0.12	0.84 ± 0.06	1.47 ± 0.05
Structure quality				
% residues in most favorable region of Ramachandran plot[d]	99.4 ± 1.3	96.1 ± 1.9	89.0 ± 1.6	75.5 ± 3.9
Number of bad contacts/100 residues[d]	2.2 ± 1.4	4.04 ± 3.1	4.2 ± 1.8	11.4 ± 2.6
Packing score[e]	0.53 ± 0.07	0.17 ± 0.19	−0.36 ± 0.08	−0.65 ± 0.09
Torsion angle score[e]	0.76 ± 0.05	0.22 ± 0.23	0.25 ± 0.06	−0.41 ± 0.07
Backbone coordinate precision (Å)[f]	0.34 ± 0.08	0.32 ± 0.06	1.02 ± 0.10	1.13 ± 0.12
All atom coordinate precision (Å)[f]	0.71 ± 0.08	0.77 ± 0.09	1.65 ± 0.11	1.74 ± 0.12
Backbone RMS shifts (Å)[g]	0.21		0.39	
All atom RMS shifts (Å)[g]	0.32		0.55	
Backbone RMS versus X-ray (Å)[h]	0.53	0.60	1.88	1.90
ϕ deviation from X-ray (°)[i]	7.4	10.7	31.9	35.1
ψ deviation from X-ray (°)[i]	10.6	13.1	41.4	43.8

[a] Structures were calculated by simulated annealing (8) with the program XPLOR (9) modified to include coupling constant (10), secondary carbon shift (11), and conformational database (Ref. 1 and this work) refinement using the published experimental restraints for the oligomerization domain of p53 (5) and IL-4 (6). Each ensemble consisted of 30 simulated annealing structures. The interresidue conformational potentials (cf. Table 1) were not used in these calculations.

[b] The experimental restraints for the tetramerization domain of p53 (42 residues per subunit) comprise (5) 3752 interproton distance restraints, 284 torsion angle restraints, 144 $^3J_{HN\alpha}$ coupling constant restraints, and 292 secondary $^{13}C\alpha$ and 13Cβ chemical-shift restraints for the whole tetramer. The experimental restraints for IL-4 (6) (a monomer of 133 residues) comprise with 921 interproton distance restraints and 123 torsion angle restraints.

[c] The larger deviations in the improper torsion angles for the structures refined with the conformational database potential relative to those refined without it simply reflect the fact that the peptide bond torsion angle ω is no longer restrained to be absolutely planar but follows the expectation values of ω as a function of ϕ and ψ which vary over the range 175.6° to 185.9°.

[d] From the program PROCHECK (2).

[e] From the program WHATIF (12). The larger the value of the packing score, the better the packing. A packing score >-0.5 indicates a very good structure. A torsion angle score of less than −2 for any residue is poor.

[f] The coordinate precision is defined as the average RMS value between the individual simulated annealing structures and the mean coordinate positions.

[g] The atomic RMS shift is the RMS difference between the mean coordinates obtained with and without conformational database refinement.

[h] The values quoted are the RMS differences between the mean coordinates and the corresponding X-ray structures. The X-ray coordinates of the oligomerization domain of p53 and IL-4 are taken from (13) and (14), respectively.

[i] The values quoted are the average angular RMS differences between the 30 simulated annealing structures and the corresponding X-ray structures. In the case of IL-4, when residues where the precision in ϕ and ψ is $>20°$ are excluded, the RMS deviations in ϕ and ψ versus the X-ray structures are 13.1° and 13.0°, respectively, for the structures refined with the conformational database potential, and 15.8° and 18.2°, respectively, for the structures refined without the conformational database potential.
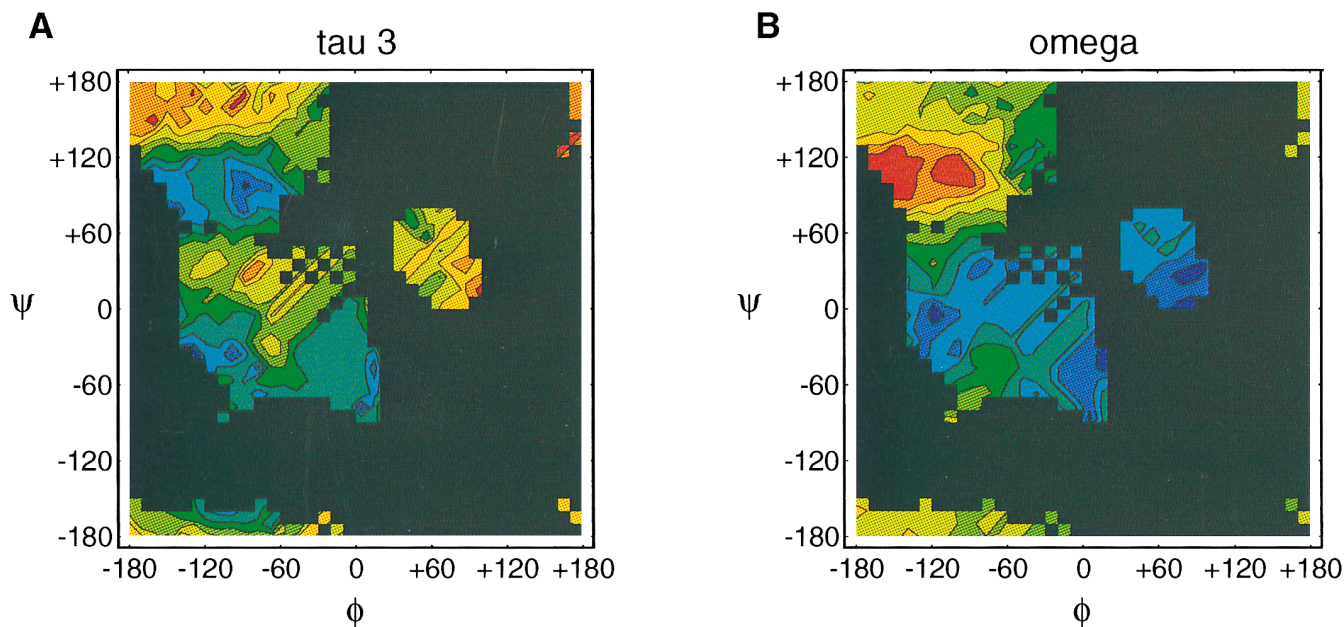
A   **B**

FIG. 3. Examples of two-dimensional expectation value surfaces for (A) ω (non-Pro, non-X-Pro, and non-Gly residues) and (B) $\tau 3$ (non-Pro and non X-Pro residues) as a function of ϕ and ψ. The surfaces are colored from red (minimum value of $\tau 3$ or ω) to violet (maximum value of $\tau 3$ or ω). For simplicity, only the populated regions of ϕ/ψ space are shown (i.e., regions with energies less than 8 kcal.mol^{-1} in the two-dimensional ϕ/ψ potential energy surfaces). (The unpopulated regions of ϕ/ψ space are colored in black.) The expected values of ω and $\tau 3$ range from 175.6° to 185.9° and from 105.4° to 116.6°, respectively.

where $\tau 3$ is the current value of the $\tau 3$ angle, ω the current value of the ω angle, $\tau 3_{\text{expectation}}(\phi, \psi)$ the expectation value of $\tau 3$ at the current values of ϕ and ψ, and $\omega_{\text{expectation}}(\phi, \psi)$ the expectation value of ω at the current values of ϕ and ψ, and $k_{\tau 3}$ and k_{ω} are force constants. Plots of two-dimensional expectation value surfaces for $\tau 3$ and ω as a function of ϕ and ψ are shown in Fig. 3. Since $\tau 3$ and ω should be dependent on ϕ and ψ but not the converse, the expression for the atomic forces F_{angledb} is reduced to

$$F_{\text{angledb}} = \partial \tau 3_{\text{expectation}}/\partial \tau 3 + \partial \omega_{\text{expectation}}/\partial \omega. \quad [3]$$

By setting the terms for $\partial \tau 3_{\text{expectation}}/\partial \phi$, $\partial \tau 3_{\text{expectation}}/\partial y$, $\partial \omega_{\text{expectation}}/\partial \phi$, and $\partial \omega_{\text{expectation}}/\partial \psi$ to zero in the expression for the partial derivative of Eq. (2), the expectation values of $\tau 3$ and ω may change instantly and arbitrarily as ϕ and ψ change. To circumvent instabilities, it is therefore essential to use an annealing protocol in which the force constants $k_{\tau 3}$ and k_{ω} are slowly increased from initially very small values (0.001 kcal $\cdot$ mol^{-1} $\cdot$ rad^{-2}) to their final values (1000 and 2000 kcal $\cdot$ mol^{-1} $\cdot$ rad^{-2}, respectively), while simultaneously reducing the force constants for the conventional harmonic potential for $\tau 3$ and ω from their initial large values (500 kcal $\cdot$ mol^{-1} $\cdot$ rad^{-2}) to their final small values (100 and 10 kcal $\cdot$ mol^{-1} $\cdot$ rad^{-2}, respectively).

We have tested the complete intraresidue conformational database potential using experimental NMR restraints from two systems: the refined high-resolution structure of the oligomerization domain of p53 (*5*) (a tetramer of 42 residues per subunit with 4472 experimental NMR restraints for the whole tetramer), and the initial low-resolution structure of interleukin-4 (*6*) (a monomer of 133 residues with 1044 experimental NMR restraints). In both cases, the use of the complete intraresidue conformational database potential does not affect the precision of the ensemble of simulated annealing structures or the agreement with the experimental NMR restraints which remain essentially unaltered, and results in small atomic RMS shifts which are within the errors of the coordinates. It does, however, result in a significant improvement in the quality of the Ramachandran plot, the nonbonded contacts, the internal packing, and the side-chain torsion angles, as well as in the agreement with the expectation values for $\tau 3$ and ω (Table 2).

Essentially, the same approach used to generate the protein conformational database potential can be applied to generate an analogous conformational database potential for nucleic acids. To this end, we created a database of 232 nucleic acid crystal structures refined at a resolution of 2 Å or better. The structures, which were derived from the NDB nucleic acids database (*7*), comprised the following: 40 A-DNA

structures, 63 B-DNA structures, 28 Z-DNA structures, 12 ''unusual'' DNA structures, 42 DNA–protein complexes, 14 DNA–groove binding drug complexes, 9 DNA–intercalating drug complexes, 9 DNA/RNA hybrids, 10 RNA structures, 2 ''unusual'' RNA structures, 1 RNA–protein complex, and 2 tRNA–protein complexes. Examples of two- and three-dimensional energy grids for the nucleic acids conformational database are shown in Figs. 1E and 1F and Fig. 2D, respectively. There is one significant difference between the protein and nucleic acids conformational database potentials that needs to be noted. In the case of the protein conformational database potential, the energy values for the various minima in the multidimensional potential energy surfaces provide a true reflection of the probability of occurence of particular conformations since protein structures in solution and the crystal state are essentially the same. In the case of nucleic acids, however, and in particular DNA, the frequency of occurence of different forms in the crystal state does not necessarily reflect their probability of occurence in solution. For example, in solution under physiological conditions, short DNA oligonucleotides are invariably B-form. In the crystal, however, A, B, or Z forms can occur depending on the crystallization conditions. As a result, the A and Z forms of DNA are overrepresented in the database, and the energy values for the different minima in the multidimensional potential energy surfaces comprising the nucleic acid conformational database potential do not necessarily reflect their probability of occurrence in solution. This does not, however, affect the positions of the various minima so that, as far as structure refinement is concerned, the nucleic acid conformational database potential still serves its primary function, namely biasing sampling to conformations that are physically realizable.

In conclusion, the protein and nucleic acids conformational database potentials permit one to take maximum advantage of the wealth of information present in high-resolution crystal structures of macromolecules. In this manner, experimentally determined NMR and medium-resolution X-ray structures can be improved by taking the accumulated knowledge on conformational preferences into account.

REFERENCES

1. J. Kuszewski, A. M. Gronenborn, and G. M. Clore, *Protein Sci.* **5,** 1067 (1996).

2. R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, *J. Appl. Crystallogr.* **26,** 283 (1993).

3. R. L. Dunbrack and M. Karplus, *J. Mol. Biol.* **230,** 543 (1993).

4. P. A. Karplus, *Protein Sci.* **5,** 1406 (1996).

5. G. M. Clore, J. Ernst, R. Clubb, J. G. Omichinski, W. M. Poindexter Kennedy, K. Sakaguchi, E. Appella, and A. M. Gronenborn, *Nature Struct. Biol.* **2,** 321 (1995).

6. R. Powers, D. S. Garrett, C. J. March, E. A. Frieden, A. M. Gronenborn, and G. M. Clore, *Science* **256,** 1673 (1992).

7. H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S. H. Hsieh, A. R. Srinivasan, and B. Schneider, *Biophys. J.* **63,** 751 (1992).

8. M. Nilges, G. M. Clore, and A. M. Gronenborn, *FEBS Lett.* **229,** 317 (1988).

9. A. T. Brünger, ''XPLOR Version 3.1: A System for X-ray Crystallography and NMR,'' Yale Press, New Haven, Connecticut.

10. D. S. Garrett, J. Kuszewski, T. J. Hancock, P. J. Lodi, G. W. Vuister, A. M. Gronenborn, and G. M. Clore, *J. Magn. Reson. B* **104,** 99 (1994).

11. J. Kuszewski, J. Qin, A. M. Gronenborn, and G. M. Clore, *J. Magn. Reson. B* **106,** 92 (1995).

12. G. Vriend and C. Sander, *J. Appl. Crystallogr.* **26,** 47 (1993).

13. P. D. Jeffrey, S. Gorina, and N. P. Pavletich, *Science* **267,** 1498 (1995).

14. A. Wlodawer, A. Pavlosky, and A. Gutschina, *FEBS Lett.* **309,** 59 (1992).