Determination of virtual-bond-angle potentials of mean force for coarse-grained simulations of protein structure and folding from *ab initio* energy surfaces of terminally-blocked glycine, alanine, and proline

2007 J. Phys.: Condens. Matter 19 285203

(http://iopscience.iop.org/0953-8984/19/28/285203)

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **19** (2007) 285203 (15pp) doi:10.1088/0953-8984/19/28/285203

Determination of virtual-bond-angle potentials of mean force for coarse-grained simulations of protein structure and folding from *ab initio* energy surfaces of terminally-blocked glycine, alanine, and proline

Urszula Kozłowska[1,2], **Adam Liwo**[1,2] **and Harold A Scheraga**[1]

[1] Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301, USA
[2] Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland

Received 20 September 2006, in final form 4 December 2006
Published 25 June 2007
Online at stacks.iop.org/JPhysCM/19/285203

Abstract

We determined the potentials of mean force corresponding to the bending of $C^\alpha \cdots C^\alpha \cdots C^\alpha$ virtual-bond angles (θ) for use in our united-residue (UNRES) force field. The potentials were determined by integrating the *ab initio* energy surfaces of terminally-blocked glycine, alanine, and proline calculated in our earlier work at the MP2/6-31G(d, p) level (Ołdziej *et al* 2003 *J. Phys. Chem.* A **107** 8035), where alanine represents all types of amino-acid residues except for glycine and proline. This resulted in 27 different free-energy surfaces. The potentials were found to depend both on θ and on the two virtual-bond dihedral $C^\alpha \cdots C^\alpha \cdots C^\alpha \cdots C^\alpha$ angles (γ_1 and γ_2) whose axes are the edges of θ, as well as on the types of all three consecutive amino-acid residues whose C^α atoms define the angle θ. The type of residue at the second and third position of a triad has a major influence on the potentials, while that in the first position is less important. Each surface was fitted well by a three-dimensional Fourier series in the trigonometric functions of multiplicities of $\theta/2$, γ_1, and γ_2; these analytical expressions can readily be implemented in the UNRES force field, thus replacing our earlier knowledge-based virtual-bond valence potentials.

M Supplementary data are available from stacks.iop.org/JPhysCM/19/285203

1. Introduction

Nowadays, *ab initio* all-atom simulations (i.e., those starting from completely unfolded structures) of protein structure and folding are too expensive, and can be applied successfully only to small proteins even when the solvent is treated implicitly [1–4]. Therefore, during the last decade, we have been developing a mesoscopic physics-based force field termed UNRES (for UNited RESidue) [5–12]. By contrast to most united-residue force fields, which are

largely knowledge-based potentials, UNRES was carefully derived, based on the physics of interactions, as a cluster–cumulant expansion [13] of the restricted free energy (RFE) function of a protein plus the surrounding solvent, in which the secondary degrees of freedom had been averaged out [7, 8, 10]. The force field is capable of *ab initio* prediction of the structures of proteins of different structural classes with good accuracy, as demonstrated in the CASP3–CASP6 blind-prediction experiments [14–16]; these results were obtained by predicting the native structure of a protein as the global minimum in the UNRES energy surface.

Recently [17–19], by developing the Langevin-dynamics formalism for UNRES, we extended its application to simulating protein-folding pathways. We found [19] that *ab initio* folding of real-size proteins can be simulated with this approach, and that UNRES provides a 4000-fold speed-up compared to all-atom simulations with explicit water and about a 200-fold speed-up compared to all-atom simulations with implicit water.

The present version of UNRES contains two types of local knowledge-based terms, both in functional form and parameterization, determined in our earlier work [6] from the statistics of the protein data bank (PDB) [20]; these are the virtual-bond-angle bending potentials and the potentials determining the energetics of side-chain rotamers. These terms do not determine the fold of the chain and, therefore, do not impair the overall characterization of UNRES as a physics-based force field. Nevertheless, these short-range terms determine the details of the geometry of the polypeptide chains particularly in the loop regions. It should be noted that the PDB statistics used to derive them is biased by long-range interactions which certainly impairs the accuracy of these local potentials and, consequently, the accuracy of the calculated structures. Moreover, we found that the functional forms that best fit the PDB statistics [6] may result in unstable forces in UNRES/MD simulations [17]. Therefore, in this work, we determined physics-based virtual-bond-angle bending potentials for UNRES by using our general approach [8] of factoring the RFE of the polypeptide chains into contributions coming from specific types of interactions together with the energy maps of terminally-blocked glycine, alanine, and proline calculated in our earlier work [9]. The new physics-based potentials will replace the statistical potentials determined in our earlier work [6] from PDB statistics.

2. Methods

2.1. The UNRES force field

In the UNRES model [5–12] a polypeptide chain is represented by a sequence of α-carbon (C^α) atoms linked by virtual bonds with attached united side chains (SC) and united peptide groups (p). Each united peptide group is located in the middle between two consecutive α-carbons. Only these united peptide groups and the united side chains serve as interaction sites, the α-carbons serving only to define the chain geometry, as shown in figure 1. The UNRES force field has been derived as an RFE function of an all-atom polypeptide chain plus the surrounding solvent, where the all-atom energy function is averaged over the degrees of freedom that are lost when passing from the all-atom to the simplified system (namely the degrees of freedom of the solvent, the dihedral angles χ for rotation about the bonds in the side chains, and the torsional angles λ for rotation (figure 2) of the peptide groups about the $C^\alpha \cdots C^\alpha$ virtual bonds) [7, 8]. The RFE is further decomposed into factors coming from interactions within and between a given number of united interaction sites [8]. Expansion of the factors into generalized Kubo cumulants [13] enabled us to derive approximate analytical expressions for the respective terms [7, 8], including the *multibody* or *correlation* terms, which are derived in other force fields from structural databases or on a heuristic basis [21]. The theoretical basis of the force field is described in detail in our earlier paper [8].

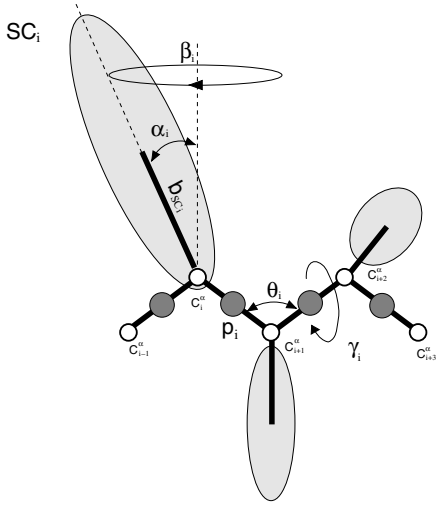

Figure 1. The UNRES model of the polypeptide chain. Dark circles represent united peptide groups (p), and open circles represent the C^α atoms, which serve as geometric points. Ellipsoids represent side chains, SCs, with their centres of mass at the b_{SC}. The p's are located half-way between two consecutive C^α atoms. The virtual-bond angles θ, the virtual-bond dihedral angles γ, and the angles α_{SC} and β_{SC} that define the location of a side chain with respect to the backbone are also indicated.

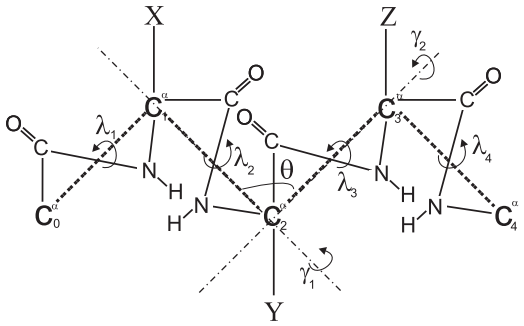

Figure 2. Illustration of a model system for the calculation of the potentials of mean force of the bending of the virtual-bond angles θ. The variables over which to integrate are the torsional angles of rotation of the peptide groups about the $C^\alpha \cdots C^\alpha$ virtual-bond axes: λ_1 (about $C_0^\alpha \cdots C_1^\alpha$), λ_2 (about $C_1^\alpha \cdots C_2^\alpha$), λ_3 (about $C_2^\alpha \cdots C_3^\alpha$), and λ_4 (about $C_3^\alpha \cdots C_4^\alpha$), while the virtual-bond angle θ and the virtual-bond dihedral angles γ_1 and γ_2 are the primary variables.

The energy of the virtual-bond chain is expressed by equation (1).

$$
\begin{aligned}
U = {} & w_{\mathrm{SC}} \sum_{i<j} U_{\mathrm{SC}_i \mathrm{SC}_j} + w_{\mathrm{SC}p} \sum_{i \neq j} U_{\mathrm{SC}_i p_j} + w_{pp}^{\mathrm{VDW}} \sum_{i<j-1} U_{p_i p_j}^{\mathrm{VDW}} + w_{pp}^{\mathrm{el}} \sum_{i<j-1} U_{p_i p_j}^{\mathrm{el}} \\
& + w_{\mathrm{tor}} \sum_i U_{\mathrm{tor}}(\gamma_i) + w_{\mathrm{tord}} \sum_i U_{\mathrm{tord}}(\gamma_i, \gamma_{i+1}) \\
& + w_b \sum_i U_b(\theta_i) + w_{\mathrm{rot}} \sum_i U_{\mathrm{rot}}(\alpha_{\mathrm{SC}_i}, \beta_{\mathrm{SC}_i}) + w_{\mathrm{corr}}^{(3)} U_{\mathrm{corr}}^{(3)} + w_{\mathrm{corr}}^{(4)} U_{\mathrm{corr}}^{(4)} \\
& + w_{\mathrm{corr}}^{(5)} U_{\mathrm{corr}}^{(5)} + w_{\mathrm{corr}}^{(6)} U_{\mathrm{corr}}^{(6)} \\
& + w_{\mathrm{turn}}^{(3)} U_{\mathrm{turn}}^{(3)} + w_{\mathrm{turn}}^{(4)} U_{\mathrm{turn}}^{(4)} + w_{\mathrm{turn}}^{(6)} U_{\mathrm{turn}}^{(6)} + w_{\mathrm{bond}} \sum_{i=1}^{n\mathrm{bond}} U_{\mathrm{bond}}(d_i). \qquad (1)
\end{aligned}
$$

Each term is multiplied by an appropriate weight, w_x. The term $U_{\mathrm{SC}_i SC_j}$ represents the mean free energy of the hydrophobic (hydrophilic) interactions between the side chains, which implicitly contains the contributions from the interactions of the side chain with the solvent. The term $U_{\mathrm{SC}_i p_j}$ denotes the excluded-volume potential of the side-chain–peptide-group interactions. The peptide-group interaction potential is split into two parts: the Lennard-Jones interaction energy between peptide-group centres ($U^{\mathrm{VDW}}_{p_i p_j}$) and the average electrostatic energy between peptide-group dipoles ($U^{\mathrm{el}}_{p_i p_j}$); the second of these terms accounts for the tendency to form backbone hydrogen bonds between peptide groups p_i and p_j. The terms U_{tor}, U_{tord}, U_b, and U_{rot} are the virtual-bond dihedral angle torsional terms, virtual-bond dihedral angle double-torsional terms, virtual-bond angle bending terms, and side-chain rotamer terms; these terms account for the local propensities of the polypeptide chain. The terms $U^{(m)}_{\mathrm{corr}}$ represent correlation or multibody contributions from the coupling between backbone–local and backbone–electrostatic interactions, and the terms $U^{(m)}_{\mathrm{turn}}$ are correlation contributions involving m consecutive peptide groups; they are, therefore, termed turn contributions. The multibody terms are indispensable for reproduction of regular α-helical and β-sheet structures [7, 8, 21]. The terms $U_{\mathrm{bond}}(d_i)$, where d_i is the length of the ith virtual bond and $n$bond is the number of virtual bonds, are simple harmonic potentials of virtual-bond distortions; they have been introduced recently [17] for molecular-dynamics implementation.

The internal parameters of $U^{\mathrm{VDW}}_{p_i p_j}$, $U^{\mathrm{el}}_{p_i p_j}$, U_{tor}, U_{tord}, $U^{(m)}_{\mathrm{corr}}$, and $U^{(m)}_{\mathrm{turn}}$ were recently derived by fitting the analytical expressions to the RFE surfaces of model systems computed at the MP2/6-31G** *ab initio* level [9, 10], while the parameters of $U_{\mathrm{SC}_i SC_j}$, $U_{\mathrm{SC}_i p_j}$, U_b, and U_{rot} were derived by fitting the calculated distribution functions to those determined from the PDB [6]; work is currently in progress to obtain these parameters from quantum mechanical *ab initio* calculations of the potentials of mean force of appropriate model systems. The ws are the weights of the energy terms, and they can be determined only by optimization of the potential-energy function, as described in our earlier work [11].

2.2. *Determination of physics-based U_b potentials*

As mentioned in sections 1 and 2.1, the current virtual-bond-angle bending potentials were derived in our earlier work [6] based on the distribution of virtual-bond angles determined from the PDB for all 20 types of amino-acid residues as functions of the angle θ and the neighbouring virtual-bond dihedral angles γ_1 and γ_2 (see figure 2 for definition). The dependence on γ_1 and γ_2 was introduced following an observation made earlier by Levitt [22] that there is a correlation between the angle θ and the neighbouring virtual-bond dihedral angles. The distributions were subsequently fitted with sums of Gaussian components by using the maximum-likelihood principle, and the potentials of mean force were calculated as the negatives of the logarithms of the distributions [6]. Such an approach is not free from a bias, because the statistics of the angles θ determined from the PDB are influenced by long-range interactions; the statistics were also insufficient to determine the potentials that would depend on the type of all three amino-acid residues that constitute a virtual-bond angle.

In this work, in order to determine physics-based U_b terms, we have implemented our earlier-developed formalism [8] in which the RFE of a polypeptide chain is factored into components, each of which corresponds only to interactions involved in a particular RFE term. The advantage of such an approach is twofold: first, there is no bias coming from other interactions which can appear in other approaches in which the complete energy of model systems is computed and, second, this approach enables us to use high-quality energy surfaces calculated at a quantum-mechanical level.

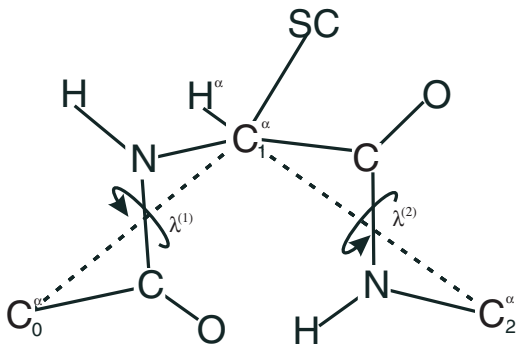

Figure 3. Definition of the dihedral angles $\lambda^{(1)}$ and $\lambda^{(2)}$ for rotation of the peptide groups about the C^α–C^α virtual bonds (dashed) of a peptide unit.

The model system used to calculate the potentials of mean force corresponding to the bending of virtual-bond angles θ is shown in figure 2; this system was used in our recent work [9] to determine the double-torsional potentials, U_{tord}. The variables to be averaged out are the angles $\lambda_1 - \lambda_4$ for rotation of the peptide groups about the $C_i^\alpha \cdots C_{i+1}^\alpha$ virtual-bond axes as well as other variables that do not belong to the UNRES degrees of freedom (i.e., the angles of rotation of the methyl groups, the distortions of the bond lengths and angles, and the out-of-plane distortions of the peptide groups). We will denote these other variables by $\mathbf{y}_1$ for peptide unit X, $\mathbf{y}_3$ for peptide unit Z, and $\mathbf{y}_2'$ for peptide unit Y; the 'prime' symbol indicates the fact that the space spanned by $\mathbf{y}_2'$ is orthogonal to the virtual-bond angle θ centred at peptide unit Y. Because U_b pertains to local-interaction terms, we compute the part of the free energy, $F_{XYZ}(\theta, \gamma_1, \gamma_2)$, of the system shown in figure 2 which arises from the local interactions of the peptide units X, Y, and Z, as given by equation (2).

$$\begin{aligned} F_{XYZ}(\theta, \gamma_1, \gamma_2) = -\beta^{-1} \ln\Big\{ & (2\pi)^{-4} (V_{\mathbf{y}_1} V_{\mathbf{y}_2'} V_{\mathbf{y}_3})^{-1} \\ & \times \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{\Omega_{\mathbf{y}_1}} \int_{\Omega_{\mathbf{y}_2'}} \int_{\Omega_{\mathbf{y}_3}} \exp\big\{-\beta\big[e_X(\lambda_1, \gamma_1 - \pi - \lambda_2, \mathbf{y}_1) \\ & + e_Y(\theta, \lambda_2, \gamma_2 - \pi - \lambda_3, \theta, \mathbf{y}_2') \\ & + e_Z(\lambda_3, \lambda_4, \mathbf{y}_3)\big]\big\} \, \mathrm{d}\lambda_1 \, \mathrm{d}\lambda_2 \, \mathrm{d}\lambda_3 \, \mathrm{d}\lambda_4 \, \mathrm{d}V_{\mathbf{y}_1} \, \mathrm{d}V_{\mathbf{y}_2'} \, \mathrm{d}V_{\mathbf{y}_3} \Big\} \end{aligned} \tag{2}$$

where $\beta = (RT)^{-1}$, T being the absolute temperature and R the universal gas constant, $\Omega_{\mathbf{y}_1}$, $\Omega_{\mathbf{y}_2'}$, and $\Omega_{\mathbf{y}_3}$ denote the regions of space corresponding to $\mathbf{y}_1$, $\mathbf{y}_2'$, and $\mathbf{y}_3$, whereas $\mathrm{d}V_{\mathbf{y}_1}$, $\mathrm{d}V_{\mathbf{y}_2'}$, and $\mathrm{d}V_{\mathbf{y}_3}$ denote the respective volume elements, e_X, e_Y, and e_Z denote the energy surfaces of peptide units X, Y, and Z (the symbols also indicate residue types). These energy surfaces are represented by the energy surfaces of terminally-blocked amino-acid residues, as in our earlier work [9]. The primary variables of these energy surfaces are the local angles for rotation about the $C^\alpha \cdots C^\alpha$ virtual-bond axes $\lambda^{(1)}$ and $\lambda^{(2)}$ first introduced by Nishikawa *et al* [23]; these angles are defined in figure 3. It should be noted that the angles $\lambda^{(1)}$ and $\lambda^{(2)}$ are defined only within a given peptide unit (X, Y, or Z; figure 2), as opposed to angles $\lambda_1 - \lambda_4$ of figure 2, which are defined for the entire model tripeptide. The relationship between the local angles $\lambda^{(1)}$ and $\lambda^{(2)}$ of residues X, Y, and Z and the angles $\lambda_1 - \lambda_4$ shown in figure 2 is given by equations (3)–(5) [23].

$$\lambda_X^{(1)} = \lambda_1 \quad \lambda_X^{(2)} = \gamma_1 - \pi - \lambda_2 \tag{3}$$

$$\lambda_Y^{(1)} = \lambda_2 \quad \lambda_Y^{(2)} = \gamma_2 - \pi - \lambda_3 \tag{4}$$

$$\lambda_Z^{(1)} = \lambda_3 \quad \lambda_Z^{(2)} = \lambda_4. \tag{5}$$

As in our earlier procedures for computations of the U_{tor} [8, 9], U_{tord} [9], and $U_{\text{corr}}^{(m)}$ [8, 10] potentials, we assume that the backbone and its closest surroundings (the C^β atoms for non-glycine and non-proline residues and the C^β, C^γ, and C^δ atoms for proline) contribute to U_b. Consequently, we distinguish three basic types of amino-acid residues: glycine, alanine, and proline, where alanine represents all non-glycine and non-proline residues. Hence, to calculate F of equation (2), we use the energy maps of N-acetyl-N'-methyl glycine, alanine, and proline, if a residue is not followed by proline and the energy maps of N-acetyl-N',N'-dimethyl glycine, alanine, and proline for those residues X and Y which are followed by a proline residue (to account for the replacement of a hydrogen atom with a carbon atom in a proline residue).

In our earlier work [9] we calculated non-adiabatic energy maps of terminally-blocked glycine, alanine, and proline using the *ab initio* quantum mechanics at the MP2/6-31G(d, p) level as functions of the $\lambda^{(1)}$ and $\lambda^{(2)}$ angles, i.e., each point of a map corresponds to an energy minimized with respect to all degrees of freedom except $\lambda^{(1)}$ and $\lambda^{(2)}$. To compute the integrals corresponding to the torsional and double-torsional potentials we assumed that the dominant contributions to the partition functions at $T = 298$ K come from the points of the maps. In other words, we neglected the integration over $\mathbf{y}$, and the integrals were reduced to summing over the non-adiabatic energy maps [9, 10]. This approach cannot be implemented here because we need to compute integrals corresponding to given values of the angle θ at the central residue Y (figure 2), which means that we cannot use the non-adiabatic energy map of Y in which, for each pair of $\lambda^{(1)}$ and $\lambda^{(2)}$ angles, θ takes a value of $\theta^*(\lambda^{(1)}, \lambda^{(2)})$ corresponding to an energy-minimized conformation (subject to given $\lambda^{(1)}$ and $\lambda^{(2)}$). On the other hand, numerical integration over $\mathbf{y}$ with energies computed in the *ab initio approach* would be prohibitively expensive. Therefore, we use the harmonic approximation of $e_Y(\lambda^{(1)}, \lambda^{(2)}, \theta, \mathbf{y}'_2)$ to compute the integral given by equation (2) for an arbitrary value of θ.

$$\begin{aligned} e_Y(\lambda^{(1)}, \lambda^{(2)}, \theta, \mathbf{y}') \approx{}& e_Y^*(\lambda^{(1)}, \lambda^{(2)}) + \tfrac{1}{2} H_{\theta\theta}^*(\lambda^{(1)}, \lambda^{(2)}) \Delta\theta^*(\lambda^{(1)}, \lambda^{(2)})^2 \\ &+ \mathbf{H}_{\theta\mathbf{y}'}^*(\lambda^{(1)}, \lambda^{(2)}) \Delta\mathbf{y}'^*(\lambda^{(1)}, \lambda^{(2)}) \Delta\theta^*(\lambda^{(1)}, \lambda^{(2)}) \\ &+ \tfrac{1}{2} \boldsymbol{\Delta}\mathbf{y}'^{*\mathrm{T}}(\lambda^{(1)}, \lambda^{(2)}) \mathbf{H}_{\mathbf{y}'\mathbf{y}'}^*(\lambda^{(1)}, \lambda^{(2)}) \boldsymbol{\Delta}\mathbf{y}'^* \end{aligned} \tag{6}$$

with

$$e_Y^*(\lambda^{(1)}, \lambda^{(2)}) = e_Y(\lambda^{(1)}, \lambda^{(2)}, \theta^*, \mathbf{y}'^*) \tag{7}$$

$$H_{\theta\theta}^*(\lambda^{(1)}, \lambda^{(2)}) = \frac{\partial^2 e_Y(\lambda^{(1)}, \lambda^{(2)}, \theta^*, \mathbf{y}'^*)}{\partial\theta^2} \tag{8}$$

$$H_{\theta y'_k}^*(\lambda^{(1)}, \lambda^{(2)}) = \frac{\partial^2 e_Y(\lambda^{(1)}, \lambda^{(2)}, \theta^*, \mathbf{y}'^*)}{\partial\theta \partial y'_k} \tag{9}$$

$$H_{y'_k y'_l}^*(\lambda^{(1)}, \lambda^{(2)}) = \frac{\partial^2 e_Y(\lambda^{(1)}, \lambda^{(2)}, \theta^*, \mathbf{y}'^*)}{\partial y'_k \partial y'_l} \tag{10}$$

$$\Delta\theta^*(\lambda^{(1)}, \lambda^{(2)}) = \theta - \theta^*(\lambda^{(1)}, \lambda^{(2)}) \tag{11}$$

$$\Delta\mathbf{y}'^* = \mathbf{y}' - \mathbf{y}'^*(\lambda^{(1)}, \lambda^{(2)}) \tag{12}$$

where the superscript 'T' denotes the transpose of a matrix or a vector, $\mathbf{H}$ denotes a Hessian matrix, and the asterisks indicate the values corresponding to the points on the non-adiabatic energy maps (i.e., $\mathbf{y}^*$ and θ^* denote these variables at a conditional minimum of the energy given the values of $\lambda^{(1)}$ and $\lambda^{(2)}$). For clarity we omitted the subscript 2 from $\mathbf{y}$ and the subscript Y from $\lambda^{(1)}$ and $\lambda^{(2)}$. The terms with the first derivatives of e_Y are not present in equation (6)

because $e_Y^*(\lambda^{(1)}, \lambda^{(2)})$ has been minimized with respect to all variables except for $\lambda^{(1)}$ and $\lambda^{(2)}$ which are held constant.

Use of a harmonic approximation in equation (6) is justified here because reaching a geometry corresponding to a given value of the virtual-bond angle θ from that which corresponds to θ^* (which, in turn, corresponds to the geometry optimized for given values of $\lambda^{(1)}$ and $\lambda^{(2)}$) requires mainly the distortions of the real valence angles at the C^α atoms. Because we are computing the statistical sums in equation (2) for $T = 298$ K, the thermally accessible states lie within $RT \approx 0.7$ kcal mol^{-1}. Also, the energy has a single minimum as a function of a given real valence angle. Consequently, assuming a typical value of the force constant of about 100 kcal mol^{-1} rad^{-2} in the expression for the bending energy of a real valence angle, we obtain about 7° distortion of a valence angle, which is acceptable (larger distortions for which anharmonic terms in the energy expansion are important give insignificant contributions to the statistical sums). It should also be noted that our use of the harmonic approximation is similar to computing the oscillation part of the partition function of rigid polyatomic molecules, which gives quite accurate values [24]. With $\lambda^{(1)}$ and $\lambda^{(2)}$ fixed, and neglecting the minor contributions from the rotation of the methyl groups, the systems considered can be treated as rigid molecules. It should be noted that we do not use the harmonic approximation to compute the statistical sum over the angles $\lambda^{(1)}$ and $\lambda^{(2)}$, but we evaluate these parts of the integrals in equation (2) numerically. For e_X and e_Z, we use the approximation of our earlier work [9, 10] (equations (13) and (14)).

$$e_X(\lambda^{(1)}, \lambda^{(2)}, \mathbf{y}_1) \approx e_X(\lambda^{(1)}, \lambda^{(2)}, \mathbf{y}_1^*) \tag{13}$$

$$e_X(\lambda^{(1)}, \lambda^{(2)}, \mathbf{y}_3) \approx e_X(\lambda^{(1)}, \lambda^{(2)}, \mathbf{y}_3^*). \tag{14}$$

After inserting equations (6), (13), and (14) into equation (2) and integrating over $\mathbf{y}_1$, $\mathbf{y}_2'$, and $\mathbf{y}_3$, we obtain equation (15).

$$\begin{aligned} F_{XYZ}(\theta, \gamma_1, \gamma_2) \approx -\beta^{-1} \ln \Bigg\{ & 2^{-4} \pi^{-(4+N/2)} V_{\mathbf{y}_2'}^{-1} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} (\det \mathbf{H}^*_{\mathbf{y}_2'^*\mathbf{y}_2'^*})^{-\frac{1}{2}} \\ & \times \exp\Big\{ -\beta \Big[e_X^* + e_Y^* + e_Z^* \\ & + \tfrac{1}{2} (\mathbf{H}_{\theta^*\theta^*} - \tfrac{1}{4} \mathbf{H}^*_{\theta^*\mathbf{y}_2'^*} \mathbf{H}^{*-1}_{\mathbf{y}_2'^*\mathbf{y}_2'^*} \mathbf{H}^*_{\theta^*\mathbf{y}_2'^*}{}^{\mathrm{T}}) \Delta\theta^{*2} \Big] \Big\} \, \mathrm{d}\lambda_1 \, \mathrm{d}\lambda_2 \, \mathrm{d}\lambda_3 \, \mathrm{d}\lambda_4 \Bigg\} \end{aligned} \tag{15}$$

where, for clarity, we omitted the variables $\lambda_1 - \lambda_4$ from e_X^*, e_Y^*, e_Z^*, $\mathbf{H}^*_{\mathbf{y}_2'^*\mathbf{y}_2'^*}$, $\mathbf{H}^*_{\theta^*\mathbf{y}_2'^*}$, and $\Delta\theta$.

As in our earlier work [9], we evaluate F_{XYZ} defined by equation (15) by numerical quadrature on a four-dimensional grid in $\lambda_1 - \lambda_4$. The bin lengths in $\lambda_1 - \lambda_4$ were 15°, consistent with the grid size of the calculated maps [9]. In this work we calculated the Hessian matrices of terminally-blocked amino-acid residues for the geometries corresponding to all points of the grid considered in our earlier work [9] at the quantum-mechanical *ab initio* RHF/6-31G(d, p) level, and transformed these matrices into internal coordinates. We used the GAMESS program [25] to carry out the quantum-mechanical calculations. Calculations at the MP2 level would be too expensive to carry out and given the fact that they have been carried out to estimate how the energy function behaves outside the non-adiabatic energy maps using the harmonic approximation (equation (6)) which neglects higher derivatives of the energy, and not to compare, for example, the theoretical and experimental IR frequencies, the RHF level is sufficient for the purpose of our work.

The U_b potentials can now be defined by equation (16) as the difference between the approximate $F_{XYZ}(\theta, \gamma_1, \gamma_2)$ and the quantity $\overline{F}_{XYZ}(\gamma_1, \gamma_2)$ defined by equation (17) calculated by integrating over all degrees of freedom of the tripeptide except the angles γ_1 and γ_2; $\overline{F}_{XYZ}(\gamma_1, \gamma_2)$ is a sum of torsional and double-torsional terms which have already been

introduced and parameterized using *ab initio* energy surfaces in our earlier work [9].

$$U_{b,XYZ}(\theta,\gamma_1,\gamma_2) = F_{XYZ}(\theta,\gamma_1,\gamma_2) - \bar{F}_{XYZ}(\gamma_1,\gamma_2) \tag{16}$$

$$\bar{F}_{XYZ}(\gamma_1,\gamma_2) = -\beta^{-1}\ln\Bigg\{(2\pi)^{-4}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\exp\{-\beta[e_X^* + e_Y^* + e_Z^*]\}\,\mathrm{d}\lambda_1\,\mathrm{d}\lambda_2\,\mathrm{d}\lambda_3\,\mathrm{d}\lambda_4\Bigg\}. \tag{17}$$

After the $U_{b,XYZ}(\theta,\gamma_1,\gamma_2)$ surfaces were calculated, we fitted each of them with a three-dimensional Fourier series in $\theta/2$, γ_1, and γ_2 defined by equation (18). The use of $\theta/2$ instead of θ as a basic variable is motivated by the fact that the squares of interatomic distances within a peptide unit (figure 3) are naturally expressed in terms of powers of $\cos(\theta/2)$ and $\sin(\theta/2)$. To keep the number of terms to an absolute minimum, we used only the terms with sines of the multiplicities of $\theta/2$, which resulted in as good fit as after including both sines and cosines of the multiplicities of $\theta/2$.

$$\begin{aligned}U_b(\theta,\gamma_1,\gamma_2) = {} & a_\circ + \sum_{l=1}^{10} a_l \sin(l\theta/2)\\ & + \sum_{l=1}^{4}\sum_{m=1}^{6}\sin(l\theta/2)[b_{lm}\cos(m\gamma_1) + c_{lm}\sin(m\gamma_1)\\ & + d_{lm}\cos(m\gamma_2) + e_{lm}\sin(m\gamma_2)]\\ & + \sum_{l=1}^{4}\sum_{m=2}^{4}\sum_{n=1}^{m-1}\sin(l\theta/2)\{f_{lmn}^{+}\cos[m\gamma_1 + (n-m)\gamma_2]\\ & + f_{lmn}^{-}\cos[m\gamma_1 - (n-m)\gamma_2]\\ & + g_{lmn}^{+}\sin[m\gamma_1 + (n-m)\gamma_2] + g_{lmn}^{-}\sin[m\gamma_1 - (n-m)\gamma_2]\}.\end{aligned} \tag{18}$$

We used linear least-squares fitting to determine the coefficients. Tables of the coefficients for all 27 model tripeptides are provided with the supplementary material (available at stacks.iop.org/JPhysCM/19/285203).

3. Results and discussion

Selected contour plots of $U_{b,XYZ}(\theta,\gamma_1,\gamma_2)$ for $\gamma_1 = 180^\circ$ (the left part of the model chain shown in figure 2 in extended conformation), $\gamma_1 = 60^\circ$ (right-handed α-helix) and $\gamma_1 = -60^\circ$ (left-handed α-helix) are shown in figure 4. Because alanine-type residues occur most frequently, we first show the effect of replacement of the X, Y, or Z residue in the AAA triad by glycine and proline; additionally, we also show plots for the APP, AGP, APG, and AGG triads. It can be seen from figure 4 that $U_{b,XYZ}(\theta,\gamma_1,\gamma_2)$ depends not only on the type of the central (Y) residue but also on those of the terminal (X and Z) residues. In the current UNRES force field, this dependence is neglected because U_b was parameterized [6] based on PDB statistics which were insufficient to determine the dependence of the potentials on the three different residue types.

For the AAA-type triad, which occurs most frequently in protein sequences, the $\theta - \gamma_2$ maps exhibit two low-energy regions, one centred about $\theta = 100^\circ$ and another one about $\theta = 140^\circ$. The second region does not appear for $\gamma_1 = -60^\circ$, while for $\gamma_1 = 60^\circ$ and $\gamma_1 = 180^\circ$ it appears for negative γ_2 values. This means that $\theta > 90^\circ$ occurs for the AAA triads when at least one neighbouring γ angle is extended (because $-90^\circ < \gamma_2 < 0^\circ$ correspond to left-handed helices which are energetically unfavourable). The first region (with θ around 100°)

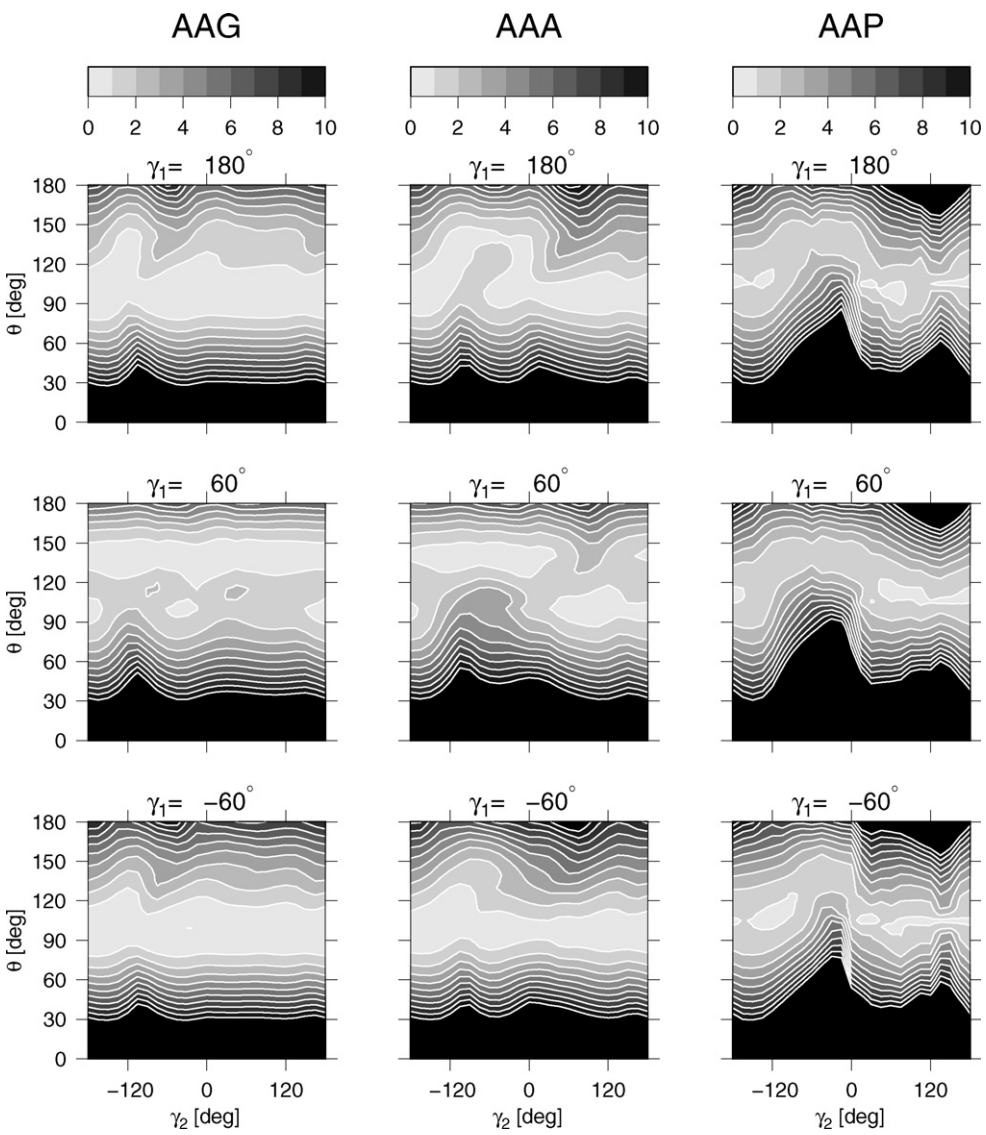

Figure 4. Selected contour plots of $U_{b,XYZ}(\theta, \gamma_1, \gamma_2)$ virtual-bond angle potentials determined in this work by using the *ab initio* energy surfaces of terminally-blocked glycine, alanine, and proline (equation (16)) in θ and γ_2 for $\gamma_1 = -60°$, 60°, and 180°. Residue types are shown above each panel.

corresponds to conformations with at least one γ angle corresponding to the folded α-helical region. This finding is consistent with the earlier observation made by Levitt [22] based on protein statistics.

We have expanded the comparison based on [22] by making scatter plots of the (γ_2, θ) pairs determined for the AAA-type triads from the database of proteins structures selected in our earlier work [6] to derive the statistical potentials (the statistics were insufficient to make a reasonable comparison for the other types of triads). We present three plots for γ_1 drawn for intervals centred at $-60°$, 60°, and 180°, respectively and width 60°. It can be seen that, in

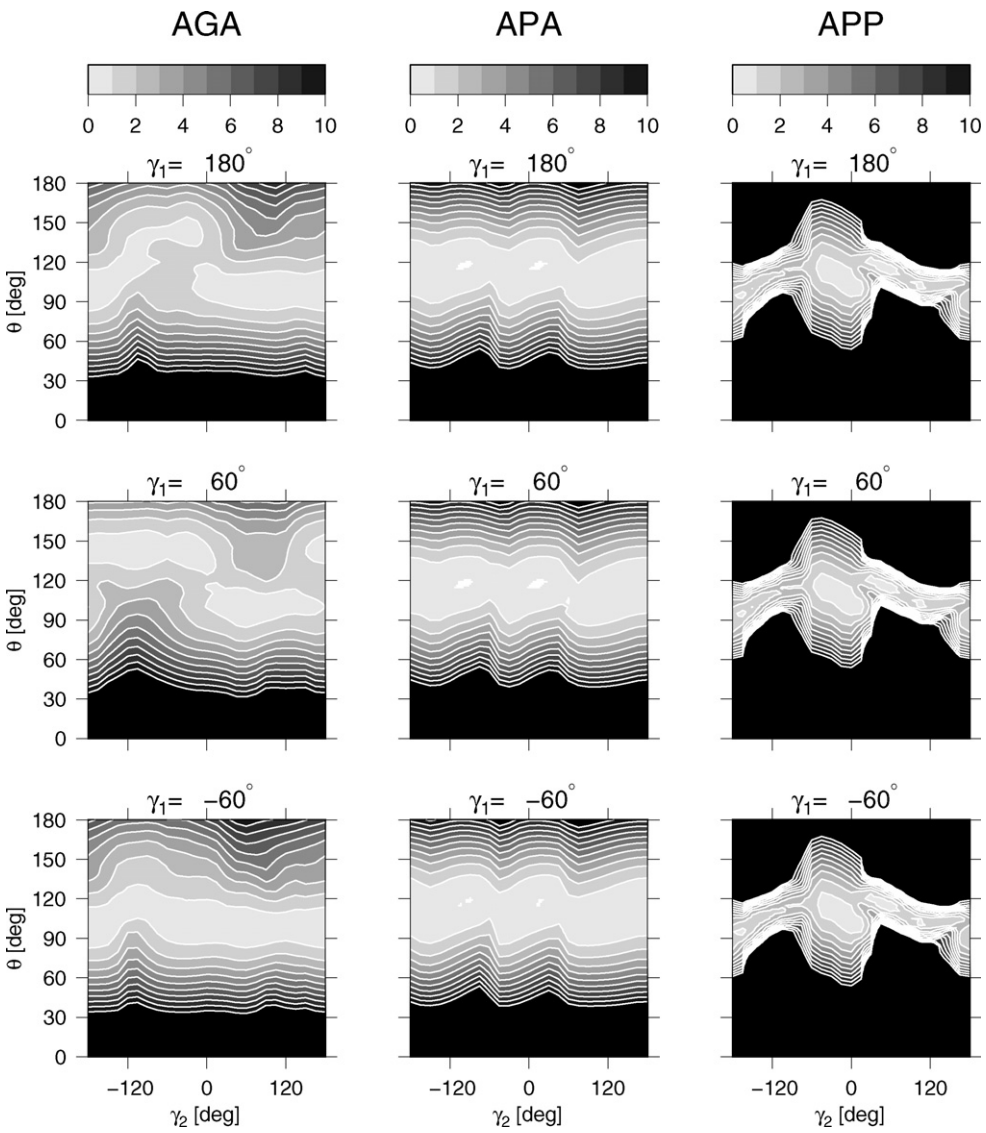

Figure 4. (Continued.)

figure 5, the density of points reflects the shapes of the sections (figure 4) of $U_b(\theta, \gamma_1, \gamma_2)$ of the AAA triad. For γ_2 centred at $-60°$ the points are scattered in γ_2 but concentrated about $\theta \approx 90°$ for positive values of γ_2 and around $\theta \approx 120°$ for negative values of γ_2, which follows the minimum valley of $U_b(\theta, \gamma_1, \gamma_2)$ for $\gamma_1 = -60°$. For $\gamma_1 = 60°$, there are three clusters of points reflecting the minima of the corresponding plot of $U_b(\theta, \gamma_1, \gamma_2)$ and the fact that the region of the minimum at positive γ_2 extends beyond $\gamma_2 = 180°$ (i.e., to the negative values of γ_2 in the plot of $U_b(\theta, \gamma_1, \gamma_2)$). The difference is, however, that the cluster of pairs with positive γ_2 and θ is much more populated than that of negative γ_2 and large θ values, which is caused by the predominance of α-helical structures in proteins. Finally, for $\gamma_1 = 180°$, the large cluster of points with θ spread from about 90° to about 150° and large and

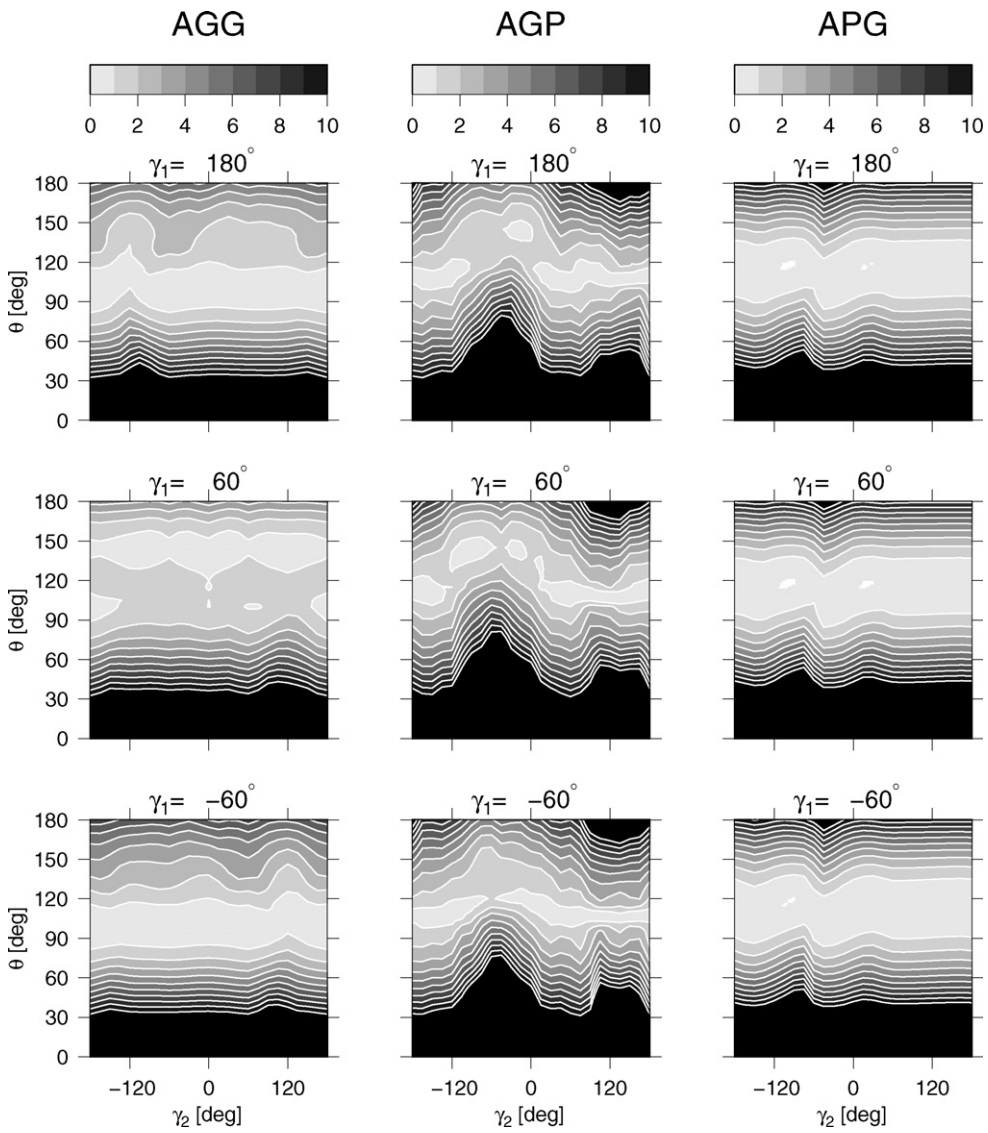

Figure 4. (Continued.)

negative γ_2 reflects the corresponding broad region of the minimum in the plot of $U_b(\theta, \gamma_1, \gamma_2)$. However, the region with large negative γ_2 is overrepresented in the scatter plot compared to that of the plot of $U_b(\theta, \gamma_1, \gamma_2)$ because residues with at least one extended γ angle occur most frequently in β-sheets in which both γ_1 and γ_2 are extended. In summary, the sections of the calculated $U_b(\theta, \gamma_1, \gamma_2)$ surface are similar to the data from PDB statistics; however, the latter (and also the statistical U_b potentials currently used in UNRES) are clearly influenced by long-range interactions which are present in regular α-helical and β-sheet structures. Consequently, replacing the current statistical U_b potentials with those derived in the current work will free the force field from the bias towards the local geometry of the regular structures present in the current potentials. This bias is probably one of the reasons that the current UNRES force field

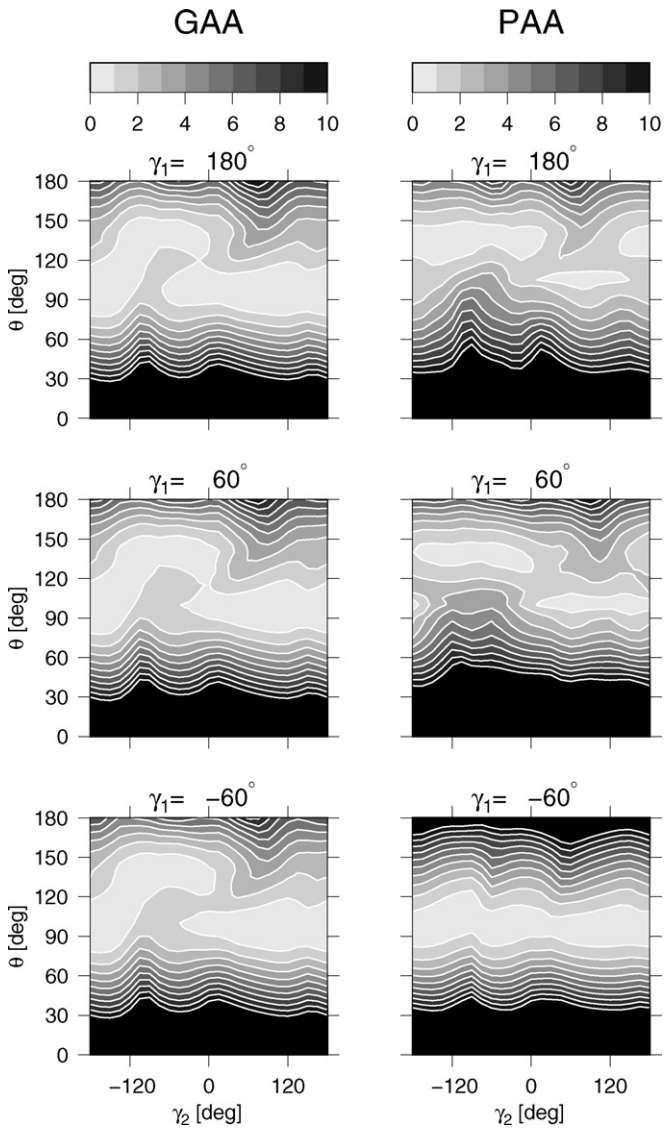

Figure 4. (Continued.)

is good at prediction of the structures of proteins containing predominantly regular elements but is less successful with proteins with large loops and segments with undefined secondary structure [12].

Replacing the central (Y) residue with glycine (the AGA triad) does not change the U_b surface qualitatively, which suggests that the local interactions within the third residue in the triad mainly influence the U_b surface. This conclusion is supported by the fact that the U_b surfaces for AAG, AGG, and also GGG (not shown) exhibit a common pattern different, however, from that of AAA, with less pronounced dependence on γ_2 and a large minimum region at θ around 100° for $\gamma_1 = -60°$ and $\gamma_1 = 180°$, and at θ around 140° for $\gamma_1 = 60°$. When the first residue is replaced with glycine, the U_b surface for $\gamma_1 = -60°$ becomes very

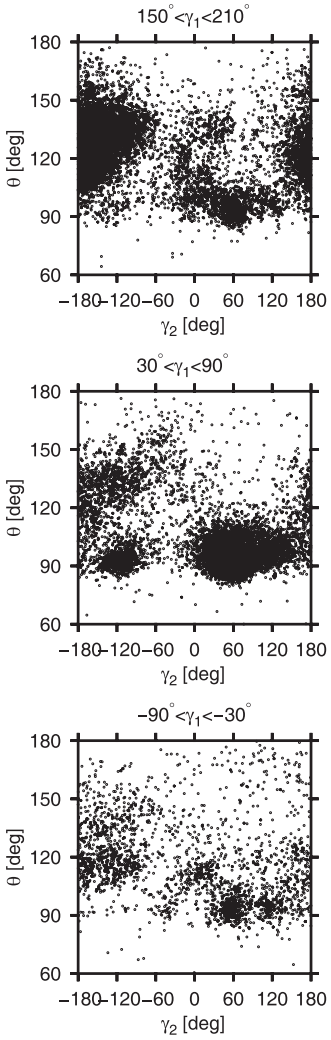

Figure 5. Scatter plots of γ_2, θ pairs of all non-proline triads using the protein database from [6] drawn for the γ_1 angles from 60°-wide intervals centred at values corresponding to those of figure 4.

similar to that for $\gamma_1 = 60°$ because of the fact that glycine is an achiral residue. However, a seemingly more drastic replacement of the alanine residue at the X position with a proline residue results in little change of the U_b surface with respect to that of the AAA triad.

Replacement of the alanine residue in the second (Y) or third (Z) position or both with proline results in major changes in the U_b surface. When proline appears in the second position, the U_b surface becomes virtually independent of γ_1 and little dependent on γ_2, and exhibits one minimum region around $\theta = 100°$; the surface observed for GPA (not shown) looks like that of APA. Some differences are observed for the APG (shown) and GPG (not shown) triads; however, the position of the minimum region and the weak dependence on the γ_1 and γ_2 angles remains the same. The surfaces obtained for proline in the second and the third positions are yet more different from all the others with two minima, one centred around $\theta = 90°$ and $\gamma_2 = 180°$

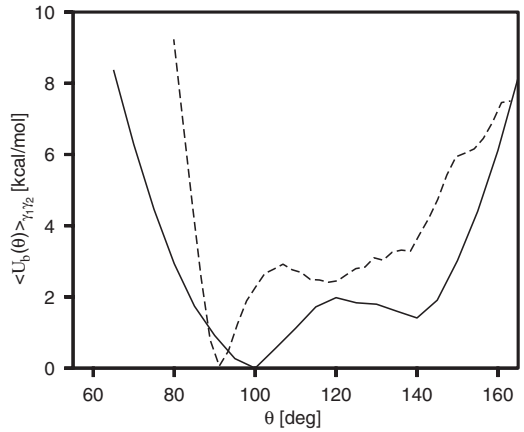

Figure 6. Potentials of mean force of virtual-bond-angle bending, Boltzmann-averaged over the virtual-bond dihedral angles γ_1 and γ_2 (solid line), compared with the statistical potential computed in our earlier work [6] from the PDB statistics, for Ala-type virtual-bond angles (dashed line).

and another one for $\theta = 100^\circ$ and $\gamma_2 = -30^\circ$. The surfaces are effectively independent of γ_1. The surfaces obtained for the PPP and GPP triads (not shown) are qualitatively the same as that shown for APP. When proline is present only in the third position, the U_b surface also exhibits a weaker dependence on γ_1 with a single curved minimum region with θ around 140° for $\gamma_2 \approx -60^\circ$ and θ around 100° for other values of γ_2. The great influence of proline in the second and third position on the U_b surface can easily be explained in terms of the severe restriction on the $\lambda^{(1)}$ angle of this amino-acid residue because of the presence of a pyrollidine ring. Consequently, the region of integration in equation (15) becomes severely restricted compared to those for triads in which alanine or glycine residues are present in the Y and Z positions.

In order to compare the potentials determined in this work with the statistical U_b potentials in the current UNRES [6], we computed the curve corresponding to U_b, Boltzmann-averaged over the γ_1 and γ_2 angles, $\langle U_b(\theta)\rangle_{\gamma_1,\gamma_2}$ (equation (19)), for the AAA triad.

$$\langle U_b(\theta)\rangle_{\gamma_1,\gamma_2} = -\beta^{-1} \ln \left\{ (2\pi)^{-2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \exp\left[-\beta U_b(\theta, \gamma_1, \gamma_2)\right] \mathrm{d}\gamma_1 \, \mathrm{d}\gamma_2 \right\}. \tag{19}$$

We plotted $\langle U_b(\theta)\rangle_{\gamma 1,\gamma_2}$ computed from equation (19) together with the corresponding statistical potential obtained from the data of [6] in figure 6. It can be seen from figure 6 that both $\langle U_b(\theta)\rangle_{\gamma_1,\gamma_2}$ and the statistical potential have two minima, one corresponding to the smaller and the other one to greater angles θ. As follows from the analysis presented earlier in this section and from the analysis of the statistical potential [6], the first minimum corresponds to folded and the second to the extended structures. It can be seen (figure 6) that the second minimum is not pronounced and the first is shifted to $\theta = 90^\circ$ and is narrower in the statistical potentials, which is caused by the fact that most of the data pertained to α-helical structures. The potentials determined in this work are not biased to any organized structure and are, therefore, expected to improve the performance of the UNRES force field. In particular, we expect that the new potentials will improve the force field for predicting the loop regions and parts with undefined secondary structure, where the bias of the statistical potentials is expected to have the greatest importance. The new U_b potentials will be incorporated into UNRES together with the new physics-based U_{rot} potentials which are now being determined in our laboratory from the energy surfaces of terminally-blocked amino-acid residues. After

this is accomplished, the energy-term weights in equation (1) will be redetermined by using our hierarchical-optimization method [11, 12], and the performance of the complete force field in the prediction of protein structures and folding pathways will be assessed on known proteins.

Acknowledgments

This work was supported by grants from the Polish Ministry of Science and Higher Education (3 T09A 134 27), the National Institute of Health (GM-14312) the National Science Foundation (MCB05-41633), and the NIH Fogarty International Center (TW7193). This research was conducted by using the resources of (a) our 800-processor Beowulf cluster at Baker Laboratory of Chemistry, Cornell University, (b) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, (c) the John von Neumann Institute for Computing at the Central Institute for Applied Mathematics, Forschungszentrum Jülich, Germany, (d) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk, (e) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and (f) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw.

References

[1] Vila J A, Ripoll D R and Scheraga H A 2003 *Proc. Natl Acad. Sci. USA* **100** 14812
[2] Jang S, Kim E, Shin S and Pak Y 2003 *J. Am. Chem. Soc.* **125** 14841
[3] Ripoll D R, Vila J A and Scheraga H A 2004 *J. Mol. Biol.* **339** 915
[4] Schug A and Wenzel W 2006 *Biophys. J.* **90** 4273
[5] Liwo A, Ołdziej S, Pincus M R, Wawak R J, Rackovsky S and Scheraga H A 1997 *J. Comput. Chem.* **18** 849
[6] Liwo A, Pincus M R, Wawak R J, Rackovsky S, Ołdziej S and Scheraga H A 1997 *J. Comput. Chem.* **18** 874
[7] Liwo A, Kaźmierkiewicz R, Czaplewski C, Groth M, Ołdziej S, Wawak R J, Rackovsky S, Pincus M R and Scheraga H A 1998 *J. Comput. Chem.* **19** 259
[8] Liwo A, Czaplewski C, Pillardy J and Scheraga H A 2001 *J. Chem. Phys.* **115** 2323
[9] Ołdziej S, Kozłowska U, Liwo A and Scheraga H A 2003 *J. Phys. Chem.* A **107** 8035
[10] Liwo A, Ołdziej S, Czaplewski C, Kozłowska U and Scheraga H A 2004 *J. Phys. Chem.* B **108** 9421
[11] Ołdziej S, Liwo A, Czaplewski C, Pillardy J and Scheraga H A 2004 *J. Phys. Chem.* B **108** 16934
[12] Ołdziej S, Łągiewka J, Liwo A, Czaplewski C, Chinchio M, Nanias M and Scheraga H A 2004 *J. Phys. Chem.* B **108** 16950
[13] Kubo R 1962 *J. Phys. Soc. Japan* **17** 1100
[14] Lee J, Liwo A, Ripoll D R, Pillardy J and Scheraga H A 1999 *Proteins: Struct. Funct. Genet.* **3** 204
[15] Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll D R, Kaźmierkiewicz R, Ołdziej S, Wedemeyer W J, Gibson K D, Arnautova Y A, Saunders J, Ye Y-J and Scheraga H A 2001 *Proc. Natl Acad. Sci. USA* **98** 2329
[16] Ołdziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila J A, Khalili M, Arnautova Y A, Jagielska A, Makowski M, Schafroth H D, Kaźmierkiewicz R, Ripoll D R, Pillardy J, Saunders J A, Kang Y K, Gibson K D and Scheraga H A 2005 *Proc. Natl Acad. Sci. USA* **102** 7547
[17] Khalili M, Liwo A, Rakowski F, Grochowski P and Scheraga H A 2005 *J. Phys. Chem.* B **109** 13785
[18] Khalili M, Liwo A, Jagielska A and Scheraga H A 2005 *J. Phys. Chem.* B **109** 13798
[19] Liwo A, Khalili M and Scheraga H A 2005 *Proc. Natl Acad. Sci. USA* **102** 2362
[20] Bernstein F C, Koetzle T F, Williams G J B, Meyer E F Jr, Brice M D, Rodgers J R, Kennard O, Shimanouchi T and Tasumi M 1977 *J. Mol. Biol.* **112** 535
[21] Kolinski A and Skolnick J 1992 *J. Chem. Phys.* **97** 9412
[22] Levitt M 1976 *J. Mol. Biol.* **104** 59
[23] Nishikawa K, Momany F A and Scheraga H A 1974 *Macromolecules* **7** 797
[24] McQuarrie D M 1976 *Statistical Mechanics* (New York: Harper Collins)
[25] Schmidt M W, Baldridge K K, Boatz J A, Elbert S T, Gordon M S, Jensen J H, Koseki S, Matsunaga N, Nguyen K A, Su S, Windus T L, Dupuis M and Montgomery J A Jr 1993 *J. Comput. Chem.* **14** 1347