Development of Novel Statistical Potentials Describing Cation−π Interactions in Proteins and Comparison with Semiempirical and Quantum Chemistry Approaches

Dimitri Gilis,*,† Christophe Biot,‡ Eric Buisine,§ Yves Dehouck,† and Marianne Rooman†

Unité de Bioinformatique Génomique et Structurale, Université Libre de Bruxelles, CP 165/61, 50 Avenue F Roosevelt, 1050 Bruxelles, Belgium, Laboratoire de Catalyse de Lille, UMR CNRS 8010, ENSCL, Bâtiment C7, Université des Sciences et Technologies de Lille, B.P. 90108, 59652 Villeneuve d'Ascq, France, and Laboratoire de Chimie Organique et Macromoléculaire, UMR CNRS 8009, Université des Sciences et Technologies de Lille, 59655 Villeneuve d'Ascq, France

Received September 12, 2005

Novel statistical potentials derived from known protein structures are presented. They are designed to describe cation−π and amino−π interactions between a positively charged amino acid or an amino acid carrying a partially charged amino group and an aromatic moiety. These potentials are based on the propensity of residue types to be separated by a certain spatial distance or to have a given relative orientation. Several such potentials, describing different kinds of correlations between residue types, distances, and orientations, are derived and combined in a way that maximizes their information content and minimizes their redundancy. To test the ability of these potentials to describe cation−π and amino−π systems, we compare their energies with those computed with the CHARMM molecular mechanics force field and with quantum chemistry calculations at the Hartree−Fock level (HF) and at the second order of the Møller−Plesset perturbation theory (MP2). The latter calculations are performed in the gas phase and in acetone, in order to mimic the average dielectric constant of protein environments. The energies computed with the best of our statistical potentials and with gas-phase HF or MP2 show correlation coefficients up to 0.96 when considering one side-chain degree of freedom in the statistical potentials and up to 0.94 when using a totally simplified model excluding all side-chain degrees of freedom. These potentials perform as well as, or better than, the CHARMM molecular mechanics force field that uses a much more detailed protein representation. The good performance of our cation−π statistical potentials suggests their utility in protein structure and stability prediction and in protein design.

INTRODUCTION

Noncovalent interactions play a fundamental role in the stabilization of the native structure of proteins and in molecular recognition. Many of these interactions involve aromatic amino acid side chains such as Phe, Tyr, or Trp engaged in $\pi-\pi$ stacking, amino−π, or cation−π interactions. The cation−π interaction is defined as the short-range interaction between a positively charged cation and π electrons of an aromatic group,[1] whereas the amino−π interaction corresponds to the interaction between an amino group carrying a partial positive charge and an aromatic group.[2,3] An interesting example of cation−π interactions can be found in the structure of the active conformations of Gi alpha (Protein Data Bank, PDB, code: 1GIA), where Lys 472 is at the heart of an aromatic cluster engaged in four cation−π interactions with two Trp's and two Tyr's.[4] Another typical example is the stair motif Gua$\therefore$Arg $\vee$ Gua ($\therefore$ denotes cation−π and $\vee$ denotes H-bond interactions) occurring in the DNA-binding domain of Tc3 transposase from *Caenorhabditis elegans* (PDB code: 1TC3) between Arg C236 and the two successive Gua's A7 and A8.[5]

Since cation−π and amino−π interactions are structurally important in proteins, their correct description in potentials used to evaluate the energy of an amino acid sequence adopting a given conformation is essential. The ability of molecular mechanics force fields to fulfill this task has already been investigated by comparing their performances to ab initio quantum chemistry calculations.[6−8] Although these force fields do not describe quadrupoles of molecular species, which are at the basis of the cation−π interaction, they reproduce with some success the energy of association between an aromatic ring and a positive charge. Database-derived potentials constitute another class of energy functions widely used in the field of protein structure prediction and design.[9−11] Compared to molecular mechanics potential energy functions, they present the advantage of being able to deal with a simplified representation of proteins and with implicit solvent effects. Misura et al.[12] recently proposed a database-derived potential depending on inter-residue angle and distance criteria between certain types of side chains. This potential describes hydrophobic packing, $\pi-\pi$ interactions, and cation−π interactions between Arg and the aromatic side chains merged in a single class. However, it considers neither Lys involving cation−π interactions nor

* Corresponding author phone: +32 2 650 36 15; fax: +32 2 650 35 75; e-mail: dgilis@ulb.ac.be.
† Université Libre de Bruxelles.
‡ Laboratoire de Catalyse de Lille, Université des Sciences et Technologies de Lille.
§ Laboratoire de Chimie Organique et Macromoléculaire, Université des Sciences et Technologies de Lille

10.1021/ci050395b CCC: $33.50
Published on Web 01/19/2006

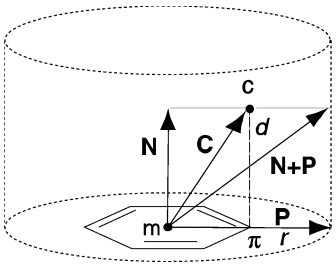

Figure 1. Geometric criteria defining cation−π interactions. **N** is the normal to the aromatic plane, m is the center of the ring, and c and π denote the (partial) positive charge and the atom of the aromatic cycle that is closest to c, respectively; they are separated by a distance *d*. *r* is twice the largest distance between the atoms of the ring and its center. **C** is the vector linking the center of the aromatic plane to c, and **P** is the vector of length *r* containing the center of the aromatic plane and π. We consider that there is a cation−π interaction between the ring and the charge if *d* is lower than or equal to 4.5 Å and if the angle between **N** and **C** is smaller than or equal to that between **N** and **N** + **P**.

amino−π interactions. Moreover, this potential requires the knowledge of the atomic details of the side chains.

Here, we develop database-derived residue pair potentials that depend on distances, angles, and amino acid types. Our approach allows the systematic description of the correlations between different sequence and structure descriptors and their combinations in a way that maximizes their information content and minimizes their redundancy.[13] Two versions of our potentials are computed. The first requires the knowledge of the atomic details of the side chains, or at least the positions of the geometric centers, to compute distances and angles. The other is developed at a residue level and is designed to be used in prediction algorithms. These potentials are derived for all residue types, but we focus here on cation−π and amino−π interactions. In particular, we compared the depth of the global minimum of the energy landscape computed with our potentials to the lowest energy obtained with quantum chemistry calculations for different aromatic side-chain groups interacting with side-chain moieties carrying a net or partial positive charge. We pursued the same analysis with a molecular mechanics force field. Our results are very encouraging, as some of our novel highly simplified database-derived potentials perform even better than the molecular force field.

METHODS

Data Set of Cation−π Partners. In the protein context, we focus on cation−π interactions between the aromatic amino acids Phe, Trp, and Tyr and the positively charged Arg and Lys or Asn and Gln that possess a partially charged (amino) group. Note that the interaction involving Asn or Gln and an aromatic side chain is often called an amino−π instead of a cation−π interaction, but for simplicity, we will call it a cation−π interaction as well.

We identified all the cation−π partners present in a data set of 141 well-resolved and refined protein chains sharing less than 20% sequence identity (see Wintjens et al.[14] for a list of these proteins), whose structures are taken from the Protein Data Bank.[15] Cation−π interactions were defined geometrically by a distance and an angle criterion (Figure 1).[8] The first criterion requires that the shortest distance between an aromatic ring atom and one of the atoms carrying the (partial) positive charge be lower than or equal to 4.5 Å. The angle criterion requires that the partial positive charge be above the plane corresponding to the aromatic ring, within a cylinder whose base includes the ring and has a radius equal to the ring diameter.

We found 315 pairs of residues in our data set whose geometry satisfies these criteria and were, thus, considered as forming cation−π interactions (Supporting Information, Table SI1). Among these, there are 8% Phe−Lys, 6% Phe−Asn, 7% Phe−Gln, 9% Phe−Arg, 7% Trp−Lys, 6% Trp−Asn, 4% Trp−Gln, 12% Trp−Arg, 14% Tyr−Lys, 9% Tyr−Asn, 5% Tyr−Gln, and 13% Tyr−Arg.

Quantum Chemistry Energy Calculations. The quantum chemistry energy calculations were performed with the Gaussian 03 suite of programs.[16] The cation−π systems were simplified for these calculations: Lys and Arg were represented by ammonium and guanidinium groups, respectively, and Gln and Asn were reduced to a formamide moiety. The aromatic amino acids Phe, Tyr, and Trp were represented as benzene, phenol, and indole rings, respectively. In a first step, all these chemical structures were optimized separately at the Hartree−Fock (HF) level using the 6-31G(d,p) basis set. These optimized structures of the cation−π partners were then superimposed onto the PDB structure using the U3BEST algorithm.[17] The structures so obtained were used for all quantum chemistry energy calculations.

In a second step, the energies of the 315 cation−π pairs were calculated at the second order of the Møller−Plesset perturbation theory (MP2).[18] We used a modified version of the standard 6-31G(d,p) basis set that has been shown to more accurately describe cation−π interaction energies.[19,20] In this basis set, denoted 6-31G(2d (0.8,0.2),p), the Gaussian α_d exponent of the d-polarization functions on the heavy atoms C, N, and O has the usual value of 0.8, with an additional α_d exponent equal to 0.2. The interaction energy of the cation−π complex A−B is defined as $\Delta E = E(\mathrm{A-B}) - E(\mathrm{A}) - E(\mathrm{B})$, where E(A−B) is the energy of the A−B complex and E(A) and E(B) are the energy of A and B taken separately. MP2 and HF energies were computed in the gas phase. The MP2 interaction energy, ΔE_{MP2}, can be decomposed into the HF interaction energy, ΔE_{HF}, and the electron correlation energy, ΔE_{corr}.

We also evaluated the interaction free energy of the cation−π systems in the presence of a solvent: $\Delta G_{\mathrm{MP2+solv}} = \Delta E_{\mathrm{MP2}} + \Delta\Delta G_{\mathrm{solv}}$, where the solvation free energy, $\Delta\Delta G_{\mathrm{solv}}$, is defined as $\Delta\Delta G_{\mathrm{solv}} = \Delta G_{\mathrm{solv}}(\mathrm{A-B}) - \Delta G_{\mathrm{solv}}(\mathrm{A}) - \Delta G_{\mathrm{solv}}(\mathrm{B})$. We calculated it by using the integral equation formalism (IEF) of the polarized continuum model (PCM) implemented in the Gaussian 03 program.[21] In IEF−PCM, the solvent is modeled as a polarizable continuum surrounding a cavity that contains the solute molecule. The PCM calculations have been performed at the HF/6-31G(2d (0.8,0.2),p) level. The combination of gas-phase energies estimated at the MP2 level and solvation contributions at the HF level is justified by the fact that HF and MP2 solvation contributions have been shown to be similar,[22] with the non-negligible advantage that HF is less computer-time-consuming. The $\Delta\Delta G_{\mathrm{solv}}$ energies were evaluated in acetone. We chose this solvent because its dielectric constant (ϵ = 20.7) corresponds to a protein environment intermediate between the core and the surface,[23] a region where most of the 315 cation−π partners of our data set are located. Indeed,

Table 1. Correlation between Quantum Chemistry Gas Phase Energy Calculations, a Molecular Mechanics Force Field, and Statistical Potentials[a]

CHARMM	HF	MP2
Part A		
complete side-chain model	0.85	0.89
simplified side-chain model	0.81	0.85
database-derived potentials	**HF**	**MP2**
Part B. $C^{\mu-\text{true}}$		
$\Delta W_{dsa1} + \Delta W_{ds2} + \Delta W_{as2}$	0.96	0.96
$\Delta W_{ds2} + \Delta W_{as2} + \Delta W_{da}$	0.94	0.96
$\Delta W_{dsa1} + \Delta W_{dsa2} + \Delta W_{ds2} + \Delta W_{as2} + \Delta W_{da}$	0.93	0.95
Part C. $C^{\mu-\text{average}}$		
$\Delta W_{dsa1} + \Delta W_{as2} + \Delta W_{da}$	0.94	0.87
$\Delta W_{dsa1} + \Delta W_{ds2} + \Delta W_{as2} + \Delta W_{da}$	0.91	0.86

[a] Correlation coefficients between the minimum energies found for each of the 12 cation$-\pi$ pairs, obtained with the quantum chemistry method at the HF or MP2 levels, and those obtained with the CHARMM force field (part A) or a database-derived potential (parts B and C). We evaluated the CHARMM interaction energy between the complete side chain and between the restricted set of atoms used to model the side chains in the quantum chemistry calculations. In the case of database-derived potentials, distances and angles are either computed between the true geometric center of the side chains, $C^{\mu-\text{true}}$ (part B), or between the average one, $C^{\mu-\text{average}}$ (part C). The p value is lower than or equal to 1% for all correlation coefficients.

the average solvent accessibility of their aromatic and (partially) charged residues is equal to 15% and 22%, respectively.

Molecular Mechanics Force Field Energy Calculations. We used the CHARMM27[24] molecular mechanics force field. In a first step, we extracted the coordinates of the cation$-\pi$ partners from the PDB. We built all the hydrogens and positioned the charges. Then, we carried out 10 steps of steepest-descent energy minimization. Performing a limited number of steps allowed relaxation of the conformation without breaking the cation$-\pi$ geometry. We checked the latter point by calculating the root-mean-square deviation (RMSD) between the initial and the relaxed conformations. The average RMSD on the 315 pairs is equal to 0.04 Å, with a standard deviation of 0.01 Å. We finally computed the in vacuo interaction energy between the aromatic and the (partially) charged side-chain groups, ΔE_{charmm}, disregarding interactions between backbone atoms. The interaction energy provided contains a van der Waals and an electrostatic term, the other terms being equal to zero. We also evaluated this interaction energy between the restricted set of atoms used to model the side chains in the quantum chemistry calculations, in view of considering the same structural models when computing the molecular mechanics and quantum chemistry energies. It is denoted as "simplified side-chain model" in Table 1.

In addition, we used different solvation models: the generalized Born model (GB),[25] the effective energy function (EEF1),[26] and a distance-dependent dielectric function (RDIE). GB is a continuum dielectric approximation that estimates the reaction field by a Coulomb potential. We used the five default parameters proposed for CHARMM27. EEF1 is an effective energy function accounting for the solvent exclusion volume of the atoms and using a distance-dependent dielectric constant to evaluate the charge−charge interactions in solution. RDIE is a more simple implicit solvation model in which the dielectric constant, ϵ, depends linearly on the distance, r:$\epsilon(r) = br$; b is a constant set to 1 (RDIE) or 4 (RDIE4).

Database-Derived Potentials. We developed and used several types of database-derived potentials that depend on distances or angles between residue pairs. These potentials were derived from a data set of 1279 proteins. To obtain this set, an initial set of 1522 high-resolution ($\leq$2 Å) X-ray structures of protein chains with less than 20% sequence identity was extracted in October 2003 from the website "Culling the PDB by Resolution and Sequence Identity" (http://dunbrack.fccc.edu/Guoli/pisces_download.php).[27] All structures containing more than 5% heteroatoms or non-natural residues were excluded, hence, reducing the set to 1403 protein chains. Finally, we excluded all proteins sharing more than 20% sequence identity with the 141 protein chains from which the data set of 315 cation$-\pi$ partners were extracted, leading to the final set of 1279 proteins. To ensure that this data set contains the proper (active) quaternary conformations of the selected proteins, the coordinates were taken from the "Protein Quaternary Structure" server (http://pqs.ebi.ac.uk).[28]

Only heavy backbone atoms N, C, O, and C^{α} were considered, and the side chains were represented by a centroid, C^{μ}. Two types of centroids were constructed. The first is the geometric center of the heavy side-chain atoms and is denoted $C^{\mu-\text{true}}$. Building $C^{\mu-\text{true}}$ requires knowledge of the atomic detail of the side chains. The second is the geometric average of all heavy side-chain atoms of a given amino acid type in a data set of known structures.[29] It is denoted $C^{\mu-\text{average}}$, and its positioning only requires knowledge of the coordinates of the backbone atoms and not the side-chain geometry. In the case of glycine, both $C^{\mu-\text{true}}$ and $C^{\mu-\text{average}}$ are identified with the C^{α} atom.

The derivation of potentials from a data set of known sequences and structures requires the subdivision of the sequences S into sequence elements s (e.g., single residues, residue pairs, and triplets) and the conformations C into structural states c (e.g., inter-residue distances and angles). For any given pair of s and c, at positions i and j along the sequence, the number of joint associations $n^{\text{obs}}(c_i,s_j)$ in the data set is computed and related to a difference in free energy, ΔW_{DB}, by using the Boltzmann law:[30−32]

$$\Delta W_{DB}(C,S) \cong \sum_{i,j} \Delta W_{DB}(c_i,s_j) \cong -kT \sum_{i,j} \ln \frac{n^{\text{obs}}(c_i,s_j)}{n^{\text{exp}}(c_i,s_j)} \qquad (1)$$

where k is the Boltzmann constant, T is room temperature, and $n^{\text{exp}}(c_i,s_j)$ is the number of associations between c_i and s_j that are expected if c and s are uncorrelated. This uncorrelated state defines a reference state, and ΔW_{DB} is the difference in free energy between the native and the reference states.

We focus here on directional pairwise interactions, which require the introduction of three descriptors specifying the sequence elements s and structural states c: the amino acid types of both residues, the spatial distance between them, and an angle describing their relative orientations. The 20 residue types were considered as sequence descriptors. The inter-residue distances were computed between the pseudo-atoms C^{μ} ($C^{\mu-\text{true}}$ or $C^{\mu-\text{average}}$). The distances between 3 and

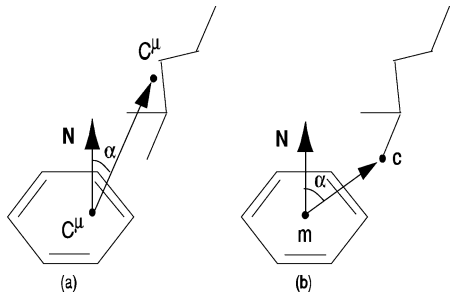

Figure 2. Geometric parameters used to represent the relative position of an aromatic side chain and a (partially) charged group. **N** is the normal to the aromatic plane, C^{μ} is the geometric center of the heavy side-chain atoms ($C^{\mu-\text{true}}$ or $C^{\mu-\text{average}}$). m is the center of the ring, and c corresponds to the atom closest to the plane that belongs to the group carrying the (partial) positive charge. α is the angle between (a) the normal to the plane, **N**, and the vector linking the C^{μ} of each residue and (b) **N** and the vector linking the center of the plane and c.

8 Å were grouped into 25 bins of 0.2 Å width, and two additional bins describe distances smaller than 3 Å and larger than 8 Å. Consecutive residues were not considered.

To specify the relative orientation between a side chain possessing a planar group and another side chain, we defined the angle α as the angle between the vector normal to the plane of the first side chain and the vector linking the C^{μ}'s of both residues (Figure 2). Note that, as the true and average C^{μ}'s occupy distinct positions, different distance and angle distributions, and therefore different energy landscapes, are associated with these two C^{μ} types. The atoms used to define the plane of each amino acid type are given in Table SI2 of the Supporting Information. The angles α and 180° − α were treated as equivalent, which reduces the possible α values to the range 0−90°. We divided these α values into bins of 10° width. An additional bin is used when the first side chain does not define a plane. This is the case for Ala, Cys, Gly, Lys, and Met. When both amino acids are planar, we considered two cases: the first side chain is the plane and the second is represented by the C^{μ}, and vice versa.

We derived six basic potentials, representing different correlations between the three descriptors presented above:

(a) A one-body distance potential describing the preference of a residue, s_i, to be separated by a distance d_{ij} from another residue whatever its type, x_j

$$\Delta W_{ds1}(s_i,d_{ij}) = -kT \ln\left[\frac{f(s_i,x_j,d_{ij})}{f(s_i,x_j)\, f(d_{ij})}\right] \qquad (2)$$

where f is the relative frequency, that is, the number of occurrences of an event divided by the total number of occurrences.

(b) A two-body distance potential describing the preference of a residue, s_i, to be separated by a distance d_{ij} from another residue, s_j, excluding the individual preferences of residues s_i and s_j described by ΔW_{ds1}

$$\Delta W_{ds2}(s_i,s_j,d_{ij}) = -kT \ln\left[\frac{f(s_i,s_j,d_{ij})\, f(s_i,x_j)\, f(x_i,s_j)\, f(d_{ij})}{f(s_i,x_j,d_{ij})\, f(x_i,s_j,d_{ij})\, f(s_i,s_j)}\right] \qquad (3)$$

(c) A one-body angular potential describing the propensity of an angle between a residue, s_i, and another residue whatever its type, x_j, to be equal to α_{ij}:

$$\Delta W_{as1}(s_i,\alpha_{ij}) = -kT \ln\left[\frac{f(s_i,x_j,\alpha_{ij})}{f(s_i,x_j)\, f(\alpha_{ij})}\right] \qquad (4)$$

(d) A two-body angular potential, based on the propensity of an angle between two residues s_i and s_j to be α_{ij}, excluding the individual preferences of residues s_i and s_j described by ΔW_{as1}

$$\Delta W_{as2}(s_i,s_j,\alpha_{ij}) = -kT \ln\left[\frac{f(s_i,s_j,\alpha_{ij})\, f(s_i,x_j)\, f(x_i,s_j)\, f(\alpha_{ij})}{f(s_i,x_j,\alpha_{ij})\, f(x_i,s_j,\alpha_{ij})\, f(s_i,s_j)}\right] \qquad (5)$$

(e) A sequence-independent potential:

$$\Delta W_{da}(d_{ij},\alpha_{ij}) = -kT \ln\left[\frac{f(d_{ij},\alpha_{ij})}{f(d_{ij})\, f(\alpha_{ij})}\right] \qquad (6)$$

(f) A one-body potential which characterizes the correlation between the amino acid type of one of the residues, independently of the type of the other residue, and the inter-residue distance and angle:

$$\Delta W_{das1}(s_i,d_{ij},\alpha_{ij}) = -kT \ln\left[\frac{f(s_i,x_j,d_{ij},\alpha_{ij})\, f(s_i,x_j)\, f(d_{ij})\, f(\alpha_{ij})}{f(s_i,x_j,\alpha_{ij})\, f(s_i,x_j,d_{ij})\, f(d_{ij},\alpha_{ij})}\right] \qquad (7)$$

(g) A two-body potential which characterizes the correlation between the type of amino acid of both residues and the inter-residue distance and angle:

$$\Delta W_{das2}(s_i,s_j,d_{ij},\alpha_{ij}) = -kT \ln\left[\frac{f(s_i,s_j,d_{ij},\alpha_{ij})\, f(d_{ij},\alpha_{ij})\, f(s_i,x_j,d_{ij})\, f(x_i,s_j,d_{ij})\, f(s_i,x_j,\alpha_{ij})\, f(x_i,s_j,\alpha_{ij})\, f(s_i,s_j)}{f(d_{ij})\, f(\alpha_{ij})\, f(s_i,x_j)\, f(x_i,s_j)\, f(s_i,x_j,d_{ij},\alpha_{ij})\, f(x_i,s_j,d_{ij},\alpha_{ij})\, f(s_i,s_j,d_{ij})\, f(s_i,s_j,\alpha_{ij})}\right] \qquad (8)$$

These elementary potentials represent distinct and nonredundant sequence−structure and structure−structure correlations, which can safely be summed to define new potentials, characterized by different reference states, in which the different correlations are added. For example, ΔW_{ds1} represents the preference of a given amino acid to be separated from any other amino acid by a certain distance, and ΔW_{ds2} the preference of two amino acid types to be separated by the same distance independently of the individual preferences of each amino acid. The sum of these two potentials yields the commonly used pairwise distance potential ΔW_{ds}:

$$\Delta W_{ds}(s_i,s_j,d_{ij}) = \Delta W_{ds1}(s_i,d_{ij}) + \Delta W_{ds1}(s_j,d_{ij}) + \Delta W_{ds2}(s_i,s_j,d_{ij}) = -kT \ln\left[\frac{f(s_i,s_j,d_{ij})}{f(s_i,s_j)\, f(d_{ij})}\right] \qquad (9)$$

More generally, the seven elementary potentials defined here describe each a single type of correlation. The procedure of defining such elementary potentials aims at disentangling the correlations between different sequence and structure descriptors and at combining them so as to maximize the information content and to minimize the redundancy of the

potentials. The generality of this procedure and its justification is described in Dehouck et al.[13]

Note finally that, when computing these potentials, a correction for sparse data is applied, which is a generalization of the correction proposed by Sippl:[32]

$$\frac{n^{\mathrm{obs}}(c,s)}{n^{\mathrm{exp}}(c,s)} \rightarrow \frac{\sigma}{\sigma + n^{\mathrm{obs}}(c,s)} + \frac{n^{\mathrm{obs}}(c,s)}{\sigma + n^{\mathrm{exp}}(c,s)} \qquad (10)$$

where σ is a parameter, taken to be equal to 20. This correction ensures that the potentials tend to be 0 when the number of observation in the data set is too small. It is applied to all the potentials defined by eqs 2−8. Note that eq 9 is not exact when this correction is used.

RESULTS AND DISCUSSION

We focused here on the development of database-derived potentials designed to suitably describe cation−π interactions and to deal with simplified protein representations. To test these potentials, we compared them to energies computed by quantum chemistry calculations and molecular mechanics force fields.

We considered 12 types of cation−π interactions, between the aromatic residues F, Y, and W and the residues R, K, Q, and N carrying a net or partially charged moiety, and used the three previously described approaches to sample the energy landscape. The most detailed method is provided by the quantum chemistry calculations where the energy of the system is estimated on the basis of the Schrödinger equation. Its drawback is that only small systems can be treated in a reasonable amount of computer time. In molecular mechanics force fields, macromolecules are described at the atomic level. Such force fields include several terms modeling the different types of interactions and contain parameters that have been fitted against experimental data or obtained by quantum chemistry calculations on small systems. Database-derived potentials correspond to the least explicit representation and can be easily adapted to simplified structural models. In this study, the proteins are represented by their heavy backbone atoms and a pseudoatom describing the side chain.

We assume that the quantum chemistry calculations (ΔE_{HF}, ΔE_{MP2}, and $\Delta G_{\mathrm{MP2+solv}}$) provide the most exact energies, especially ΔE_{MP2} in the gas phase and $\Delta G_{\mathrm{MP2+solv}}$ in the solvent. To compare these energies with those obtained using the molecular mechanics force field ($\Delta E_{\mathrm{charmm}}$) and the database-derived potentials (ΔW_{DB}), we computed the correlation coefficients, r, assuming a linear regression, between the energy values of the 12 cation−π pairs estimated by the different methods. The statistical significance of the correlation coefficient is provided by the p value, defined as the probability of obtaining the result or any more extreme result with noncorrelated data (null hypothesis). This value was computed with the R statistical package (http://www.r-project.org/).

Computing the complete energy landscape for the 12 cation−π pairs as a function of the distance and angle between the two partners is too computer-time-consuming for the quantum chemistry approach. Therefore, we used a data set of 315 cation−π systems extracted from 141 well-resolved and nonredundant protein structures (see Methods). They are characterized by different geometries and were taken as a representative sample of the conformational space of the 12 pairs. The minimal energy found for each pair was assumed to be close to the true global energy minimum of their energy landscape. We applied the same procedure to estimate the minimum of the energy landscape corresponding to the molecular mechanics force field included in the CHARMM27 package. In contrast, it is relatively straightforward and fast with database-derived potentials to construct the complete energy landscape for each cation−π pair by varying the inter-residue distance and angle; this is done by steps of 0.2 Å and 10°, respectively. In the case of statistical potentials, we were able to test the validity of sampling the energy landscape with the data set of 315 cation−π conformations. Indeed, we correlated the values of the true energy minima for the 12 pairs and those estimated from the 315 cation−π structures. We found correlation coefficients between 0.84 and 0.95 for the best performing potentials presented below. These high correlations support the validity of our assumption.

Note that, even if we disregard the limitations on computer time, comparing the complete energy landscapes computed by the three methods would not be relevant. Indeed, the description of the cation−π geometry at the atomic level is based on the distance between the aromatic plane and the atom carrying the positive charge and on the angle between the normal to the aromatic plane and the vector linking the center of the plane and the charge (Figure 2b). The representations at the residue level involve the distance between the C^{μ}'s and the angle between the normal to the plane and the vector linking both C^{μ} (Figure 2a). These two sets of descriptors cannot be compared: the correlation coefficients between the two distance descriptors and between the two angular descriptors are equal to 0.39 and 0.30, respectively. Therefore, the energy profiles obtained with the different approaches are not superimposable, and we limited ourselves to the comparison of the minimum energy values.

The correlation coefficients between the energy values obtained with the quantum chemistry approach, the molecular mechanics force field, and the statistical potentials are summarized in Table 1. The minimum energy values for the 12 cation−π pairs are given in Table 2, whereas the corresponding average energies and standard deviations are provided in Table SI3 of the Supporting Information. The geometries of the lowest energy conformers, computed with the quantum chemistry method and CHARMM, are supplied in Table SI4 of the Supporting Information. The correlation coefficient between the energies computed with the CHARMM force field and those obtained with quantum chemistry calculations is found to be between 0.81 and 0.89, the highest score corresponding to the energies calculated at the MP2 level. This result is in agreement with previous studies showing a good correspondence between ab initio and semiempirical calculations on cation−π systems.[6−8] The cation−π interaction energy calculated with CHARMM contains an electrostatic and a van der Waals contribution. The electrostatic component dominates, on average, for pairs including the charged amino acids Lys and Arg, whereas van der Waals forces are prevalent for pairs containing Asn or Gln. Previous ab initio calculations showed that the cation−π interactions result essentially from a quadrupolar electrostatic interaction.[33,34] The CHARMM force field does not include the description of quadrupoles, but it was

Table 2. Minimum Energy Values Computed with the Different Energy Functions[a]

cation−π pair	ΔE_{HF}	ΔE_{MP2}	$\Delta G_{MP2+solv}$	ΔE_{charmm}	$C^{\mu-true}$ $\Delta W_{dsa1} + \Delta W_{ds2} + \Delta W_{as2}$	$C^{\mu-average}$ $\Delta W_{dsa1} + \Delta W_{ds2} + \Delta W_{da}$	$C^{\mu-true}$ $\Delta W_{dsa1} + \Delta W_{dsa2} + \Delta W_{ds2}$	$C^{\mu-average}$ $\Delta W_{dsa1} + \Delta W_{dsa2} + \Delta W_{ds2}$
F−K	−12.9	−14.1	0.0580	−9.19	−0.767	−0.359	−0.605	−0.020
F−N	−1.05	−2.45	0.723	−2.93	−0.536	−0.069	−0.547	−0.034
F−Q	−1.16	−3.71	−0.321	−5.20	−0.546	−0.088	−0.564	−0.069
F−R	−6.51	−9.14	0.625	−7.96	−0.692	−0.199	−0.593	−0.026
W−K	−18.4	−20.6	−2.20	−12.9	−1.031	−0.353	−0.981	−0.179
W−N	−1.22	−4.33	−0.635	−5.19	−0.540	−0.012	−0.507	−0.069
W−Q	−1.66	−3.72	1.17	−4.97	−0.614	−0.060	−0.622	−0.122
W−R	−11.4	−16.2	−1.63	−12.4	−0.840	−0.169	−0.827	−0.183
Y−K	−13.5	−14.9	−0.420	−16.7	−0.979	−0.328	−0.794	−0.124
Y−N	−1.02	−3.11	0.252	−4.91	−0.558	−0.049	−0.565	−0.122
Y−Q	−1.94	−3.50	0.788	−4.07	−0.509	−0.102	−0.531	−0.036
Y−R	−10.7	−15.7	−2.16	−16.8	−0.833	−0.220	−0.813	−0.158

[a] Lowest energy, in kilocalories per mole, of the 12 cation−π pairs, evaluated with the quantum chemistry approaches and the molecular mechanics force field. In the case of the statistical potentials ($\Delta W_{dsa1} + \Delta W_{ds2} + \Delta W_{as2}$, $\Delta W_{dsa1} + \Delta W_{as2} + \Delta W_{da}$, and $\Delta W_{dsa1} + \Delta W_{dsa2} + \Delta W_{ds2}$), the global minimum of the energy landscape is provided, in kilocalories per mole.

suggested to recover some elements of the cation−π interaction by the presence of partial charges on carbons, nitrogen, and hydrogens.[6,7] Our results show that the cation−π interaction is not only represented by an electrostatic contribution in CHARMM but that the van der Waals part is also important. Indeed, the correlation coefficient with the HF and MP2 energies is only equal to 0.79 and 0.82, as opposed to 0.85 and 0.89, respectively, if the van der Waals interaction energy is not accounted for in ΔE_{charmm}.

We also computed ΔE_{charmm} by ignoring the atoms that are discarded in the quantum chemistry calculations ("simplified side-chain model" in Table 1). In this case, the correlations are slightly lower than those obtained with the complete side-chain model. In the simplified side-chain model, the balance between the electrostatic and the van der Waals energies increases in favor of the former, because mainly van der Waals contacts are removed. It supports the previous observation that the cation−π interaction is modeled in CHARMM not only by the electrostatic interaction between the atoms carrying the charge and the partial charges located on the aromatic ring but also indirectly by neighboring van der Waals contacts.

We tested all the combinations of the database-derived potentials presented in the Methods (eqs 2−8). We constructed the complete energy landscape of the 12 cation−π pairs with all these potentials and located, for each pair and each potential, the global energy minimum. Amino acid side chains were either described by their real geometric center, $C^{\mu-true}$, or by an average centroid, $C^{\mu-average}$. The latter is cruder than the former, but it can easily be used in protein structure prediction methods that deal with simplified residue-level protein representations. The combinations of potentials leading to the largest correlation coefficients with ΔE_{HF} and ΔE_{MP2}, and characterized by a global energy minimum corresponding to a cation−π geometry, are presented in Table 1. The best results are obtained when the inter-residue distances and angles are calculated between the $C^{\mu-true}$'s (Table 1, part B). Some combinations of potentials correlate very well with the MP2 and the HF energies, their correlation coefficients being even larger than those obtained with CHARMM. The minimum energy values of the best performing statistical potentials are supplied in Table 2 and the geometries of the corresponding conformers in Table SI4 of the Supporting Information; the average energies and standard deviations are given in Table SI3 of the Supporting Information. Note that the database-derived potentials are not supposed to correspond to "true" energies but rather to effective energies taking into account the environment of the cation−π pairs. Interestingly, the energies resulting from gas-phase HF and MP2 calculations are of the same order of magnitude as those obtained with the CHARMM force field, while the energies obtained by the statistical potentials are of the same order of magnitude as $\Delta G_{MP2+solv}$ (Table 2). This is consistent with the fact that statistical potentials are effective energies that take into account the screening of cation−π interactions by the solvent.

The potential $\Delta W_{dsa1} + \Delta W_{ds2} + \Delta W_{as2}$ presents a correlation coefficient equal to 0.96 both with MP2 and with HF energies (Table 1). Figure 3 shows the energy landscape of two pairs, F−N and W−R, computed with this potential. $\Delta W_{dsa1} + \Delta W_{ds2} + \Delta W_{as2}$ includes contributions describing the propensity of two residues to be separated by a distance or an angle excluding their individual preferences (ΔW_{ds2} and ΔW_{as2}) and a term taking into account the tendency of one residue to be associated with given distance and angle ranges (ΔW_{dsa1}). Replacing the ΔW_{dsa1} term with a sequence-independent term taking into account the correlation between the distance and the angle bins, ΔW_{da}, induces a slight decrease of the correlation coefficient with the HF energies. However, this change is very small, and their respective performances can be considered as identical. The combination of distance and angular terms, describing the relative position of the charge with respect to the aromatic plane, is necessary to have a good equivalence between the energy minima of the database-derived potentials and the HF or the MP2 energies. Indeed, the largest correlations obtained with a distance or an angular potential taken separately (ΔW_{ds1}, ΔW_{ds2}, ΔW_{as1}, or ΔW_{as2}) are equal to 0.81.

Using $C^{\mu-true}$ to compute the inter-residue distances and angles requires knowledge of the atomic detail of the side chains. This can be viewed as a shortcoming if the potentials must be incorporated in algorithms based on a simplified protein representation, at the residue level. One solution consists of developing algorithms that take into account one side-chain degree of freedom, represented by the position of $C^{\mu-true}$. Another solution consists of modeling the side

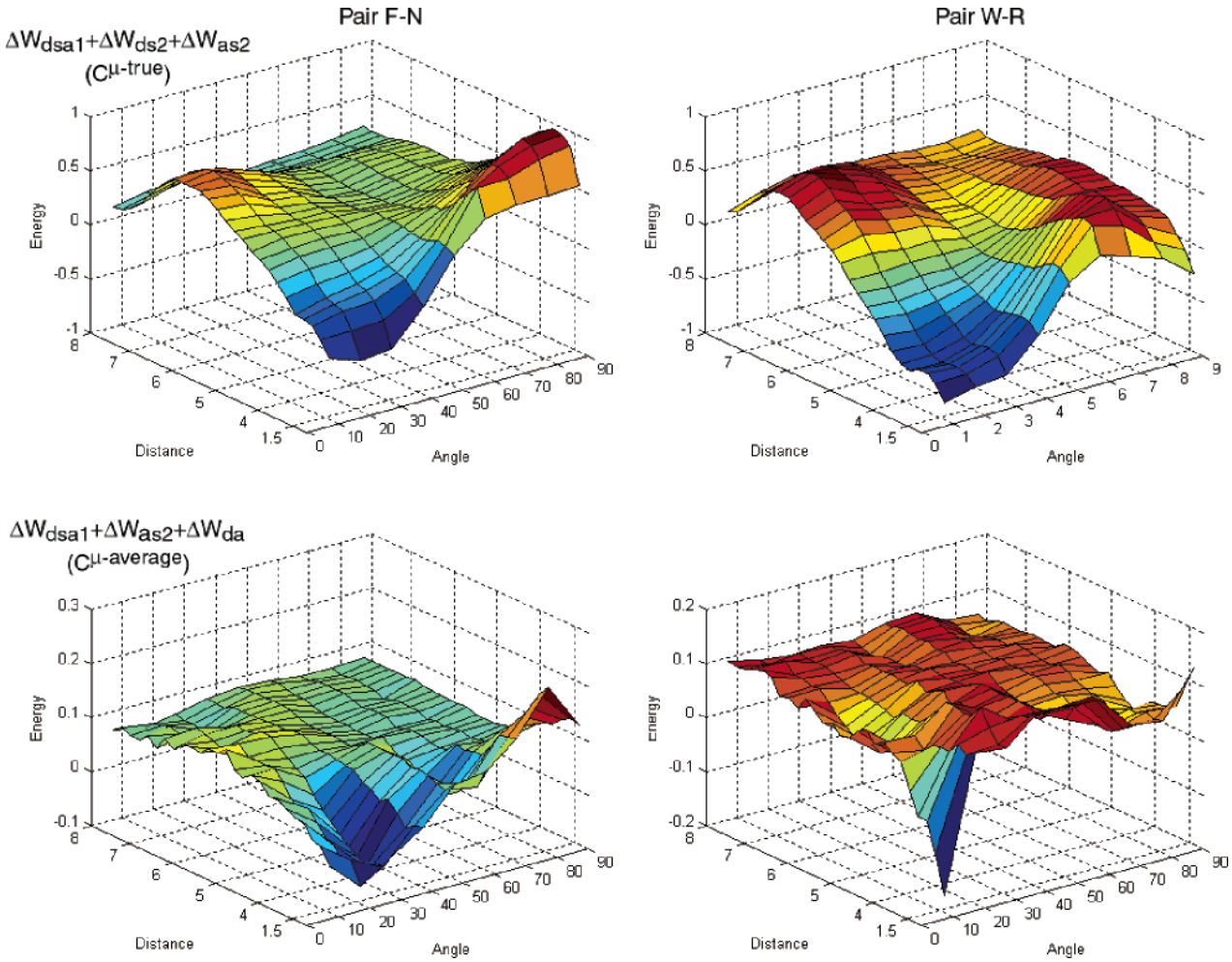

Figure 3. Energy landscape of the F−N and the W−R cation−π pairs. The first line corresponds to the statistical potential $\Delta W_{dsa1} + \Delta W_{ds2} + \Delta W_{as2}$, with the distances computed between the $C^{\mu-true}$'s. The second line corresponds to the statistical potential $\Delta W_{dsa1} + \Delta W_{as2} + \Delta W_{da}$, with the distances computed between the $C^{\mu-average}$'s. The distance is in Angstroms, the angle in degrees, and the energy in kilocalories per mole.

chains by the average centroid, $C^{\mu-average}$. Obviously, lower performances are obtained with potentials computing the distances and the angles between the $C^{\mu-average}$'s (Table 1, part C). Nevertheless, they remain good. Indeed, the combination $\Delta W_{dsa1} + \Delta W_{as2} + \Delta W_{da}$, for instance, presents correlation coefficients of 0.94 and 0.87 with HF and MP2, respectively. This result is similar to, or even better than, that obtained with the CHARMM force field ($r = 0.85$ and 0.89, respectively). Note that the best combination of potentials is partly different according to whether the side chains are characterized by $C^{\mu-true}$ or $C^{\mu-average}$. The two-body distance potential ΔW_{ds2} is included in the best performing combinations when using $C^{\mu-true}$, whereas when $C^{\mu-average}$ is used, it is replaced by the sequence-independent term taking into account the correlation between the distance and the angle bins, ΔW_{da}. Moreover, adding ΔW_{ds2} to the combination $\Delta W_{dsa1} + \Delta W_{as2} + \Delta W_{da}$ provokes a slight decrease of the correlation coefficient (Table 1, part C). Given that $C^{\mu-average}$ represents the relative side-chain positions of both amino acids with less precision than $C^{\mu-true}$, the inter-$C^{\mu-average}$ distances do not seem accurate enough to yield good performances for ΔW_{ds2}. Note also that angular contributions, here ΔW_{as2}, ΔW_{da}, and ΔW_{dsa1}, are essential to getting good performances, as already observed with potentials derived from the $C^{\mu-true}$'s.

Solvent Contribution. We used the IEF−PCM method, which has been shown to be adequate to take into account the solvent contribution in the quantum chemistry calculations (see Methods).[22,35] The average solvent accessibility of the aromatic and charged residues of the 315 cation−π partners is equal to 15% and 22%, respectively. They are, thus, on average, not located at the surface of the protein, and water is not adequate to correctly describe this environment. The protein medium has characteristic dielectric constants ranging from 2 to about 25.[23] Therefore, we chose to perform our quantum chemistry calculations in acetone, a solvent presenting a dielectric constant of 20.7. In the case of the CHARMM force field, three different approaches were used to model the solvent implicitly: the GB model,[25] the EEF1 model,[26] and RDIE (see Methods).

The lowest energies found for the different energy functions are provided in Table 2, and the average energies and their standard deviations are given in Table SI3 of the Supporting Information. The correlation between CHARMM energies and the quantum chemistry energy performed in the presence of a solvent was lower than those in the gas phase (Table 3, part A), as expected, because CHARMM does not include the effects of the solvent. The correlation coefficient remains, nevertheless, significant: 0.75. This good correlation coefficient is not completely surprising, as the

Table 3. Results Obtained with the Quantum Chemistry Calculations Performed in the Presence of the Solvent[a]

CHARMM	MP2 + acetone
Part A	
CHARMM	**0.75**
CHARMM RDIE	*0.70*
CHARMM RDIE4	*0.70*
CHARMM EEF1	0.26
CHARMM GB	0.15
database-derived potentials	**MP2 + acetone**
Part B. $C^{\mu-true}$	
$\Delta W_{dsa1} + \Delta W_{dsa2} + \Delta W_{ds2}$	**0.79**
$\Delta W_{dsa1} + \Delta W_{dsa2} + \Delta W_{ds2} + \Delta W_{as1} + \Delta W_{as2}$	**0.79**
Part C. $C^{\mu-average}$	
$\Delta W_{dsa1} + \Delta W_{dsa2} + \Delta W_{ds2}$	**0.72**
$\Delta W_{dsa1} + \Delta W_{dsa2} + \Delta W_{ds2} + \Delta W_{as1} + \Delta W_{as2}$	**0.71**

[a] Correlation coefficients between the energy minimum of 12 cation−π pairs, computed with quantum chemistry energy calculations at the MP2 level plus a solvation term, and those obtained with the CHARMM force field (part A) or a database-derived potential (parts B and C). Distances and angles are either computed between the true geometric center of the side chains, $C^{\mu-true}$ (part B), or between the average one, $C^{\mu-average}$ (part C). The solvent considered in the quantum chemistry calculations is acetone ($\epsilon = 20.7$). Several solvation models were used in combination with CHARMM. RDIE and RDIE4 correspond to a distance-dependent dielectric function with ϵ equal to 1 and 4, respectively. EEF1 is the effective energy function model. GB is the generalized Born model, which is commonly used to represent the solvent implicitly. We imposed ϵ in GB and EEF1 equal to that of the solvent used to compute the solvation term with the quantum chemistry approach. The correlation coefficients are italicized or bold if the *p* value is lower than or equal to 3% and 1%, respectively.

CHARMM force field is parametrized to model systems in the condensed phase, the parameters set used here being specific for proteins. However, an intriguing result was the decrease of the correlation coefficient when a solvation model was added to the CHARMM energies, this decrease being more marked with EEF1 and GB, the most sophisticated approaches used here.

Database-derived potentials were obtained from frequencies of association between sequence and structure elements in proteins experimentally resolved in the presence of the solvent. The effective energies were derived from a system immersed in water, the presence of the solvent being, therefore, implicitly taken into account, as well as the average properties of the protein medium. Whatever C^{μ} was used, the combinations of potentials leading to the largest correlation coefficient were the same (Table 3, parts A and B): $\Delta W_{dsa1} + \Delta W_{dsa2} + \Delta W_{ds2}$ and $\Delta W_{dsa1} + \Delta W_{dsa2} + \Delta W_{ds2} + \Delta W_{as1} + \Delta W_{as2}$. As expected, the best performances are obtained with potentials using $C^{\mu-true}$. In contrast to the results presented in Table 1, parts B and C, the two-body potential correlating sequence, distance, and angle, ΔW_{dsa2}, is present in both combinations of potentials. It seems, thus, to be important to model the solvent contribution, in conjunction with ΔW_{dsa1}. The angular potential ΔW_{as2} appears in all the combinations, correlating well with the HF and MP2 energies. When the solvent is considered, this potential is associated with its one-body equivalent, ΔW_{as1}. Finally, note that the two-body distance potential ΔW_{ds2} is present in each combination shown in Table 3 parts B and C. This potential describes the particular interactions between two residues, excluding their individual preferences. For instance, it was previously shown that this potential better describes electrostatic interactions than the commonly used distance potential $\Delta W_{ds} = \Delta W_{ds1} + \Delta W_{ds2}$.[13,31]

The comparison of the results obtained with and without the solvent (Tables 1 and 3) shows that the correlation coefficients between the database-derived potentials, ΔW_{DB}, and the quantum chemistry energies including the solvent contribution, $\Delta G_{MP2+solv}$, are markedly smaller (0.79 instead of 0.96). This is quite unexpected as the database-derived potentials contain solvent effects. This result could be explained by the fact that some properties of the solvent, such as the radius of the molecules, play a role in the evaluation of the solvation energy and that the cavitation term directly depends on these properties.[35,36] In particular, it is not clear whether acetone represents the protein environment sufficiently well. Another explanation could be that the use of continuum models induces some approximations which could also be responsible for the drop in correlations.

CONCLUSIONS

Database-derived potentials are widely used in the fields of protein structure prediction and protein design.[9−11] Unlike molecular mechanics or quantum chemistry energy functions, they can deal with simplified models of proteins. They are often suspected to less accurately describe some particular interactions, as a result of their high level of simplification. But this suspicion is, in fact, not always justified. For example, Morozov et al.[37] recently found an outstanding agreement between the energy landscapes of a hydrogen bond statistical potential and those of quantum chemistry calculations; an atomic detail description was used to model protein structures.

In this paper, we focused on cation−π interactions and designed database-derived potentials using a simplified representation of proteins and capturing the essence of these interactions. To assess the performances of our new potentials, we compared them to energies obtained by quantum chemistry and empirical molecular mechanics calculations.

The derivation of potentials from a database of known protein structures involves the search for correlations between sequence and structure elements. The elements considered here are the residue types, the inter-residue distances, and the angles between two side chains (see Figure 2). We extracted the correlations between these elements in view of obtaining the most informative, nonredundant combinations of potentials. The best correlation coefficient between the energies computed by quantum chemistry calculations at the MP2 level and our potentials is equal to 0.96 ($\Delta W_{dsa1} + \Delta W_{ds2} + \Delta W_{as2}$) if the side chains are represented by their geometric centers and to 0.87 ($\Delta W_{dsa1} + \Delta W_{as2} + \Delta W_{da}$) if the atomic details of the side chains are completely ignored. This result is quite remarkable considering the level of simplification used in the potentials. It is, moreover, even better than that obtained with a molecular mechanics force field. We showed that it is necessary to add information about the relative position of amino acid pairs, which is provided by an inter-residue angle, to the usually used distance contribution in order to accurately evaluate the cation−π interactions.

The results presented in this paper suggest several developments. It would be valuable to include our novel

statistical potentials in energy functions used in structure prediction methods. Moreover, these potentials are derived for each type of amino acid pairs and are, thus, able to describe other kinds of interactions that will be analyzed in another study.

ACKNOWLEDGMENT

We are grateful to Martine Prévost for her help with the CHARMM packages. We acknowledge support from the Communauté Française de Belgique through the Action de Recherche Concertée #02/07-289 and from the E. C. through the Concerted Action Qol 2001-3.8.4. M.R. is Research Director at the Belgian National Fund for Scientific Research (FNRS).

Supporting Information Available: (Table SI1) Database of the 315 cation−π partners. The first column corresponds to the PDB code of the protein. Residues 1 and 2 are the one letter code of the amino acids. The PDB number of each partner is also provided, as well as the closest atoms between the aromatic cycle and the (partially) charged group. (Table SI2) Atoms used to define the plane describing the side chain. The first column corresponds to the amino acids' one letter code. A "/" indicates that no plane is defined. (Table SI3) Average energies and their standard deviation (in parentheses), computed for the 12 cation−π pairs with the different energy functions. These averages and standard deviations are calculated on the 315 cation−π partners for ΔE_{HF}, ΔE_{MP2}, $\Delta G_{MP2+solv}$, and ΔE_{charmm}, and on the complete energy landscape for the statistical potentials. (Table SI4) Geometry of the cation−π pairs corresponding to the lowest energy conformers (ΔE_{MP2} and ΔE_{charmm}) and to the minimum of the energy landscape obtained for the best performing statistical potentials, $\Delta W_{dsa1} + \Delta W_{ds2} + \Delta W_{as2}$ and $\Delta W_{dsa1} + \Delta W_{as2} + \Delta W_{da}$. For ΔE_{MP2} and ΔE_{charmm}, the first value is the distance computed between the atom carrying the (partial) positive charge and the closest atom of the aromatic ring. The second value is the angle between the normal to the aromatic plane and the vector linking the center of the aromatic plane to the atom carrying the (partial) positive charge (Figure 2b). For the two statistical potentials, the first value is the distance computed between the C^{α} of each residue. The second value is the angle between the normal to the aromatic plane and the vector linking both C^{α}'s (Figure 2a). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Ma, J.; Dougherty, D. The cation−π interaction. *Chem. Rev.* **1997**, *97*, 1303−1324.
(2) Burley, S. K.; Petsko, G. A. Amino−aromatic interactions in proteins. *FEBS Lett.* **1986**, *203*, 139−143.
(3) Mitchell, J. B.; Nandi, C. L.; McDonald, I. K.; Thornton, J. M.; Price, S. L. Amino/aromatic interactions in proteins: is the evidence stacked against hydrogen bonding? *J. Mol. Biol.* **1994**, *239*, 315−331.
(4) Coleman, D. E.; Berghuis, A. M.; Lee, E.; Linden, M. E.; Gilman, A. G.; Sprang, S. R. Structures of active conformations of Gi alpha 1 and the mechanism of GTP hydrolysis. *Science* **1994**, *265*, 1405−1412.
(5) Biot, C.; Wintjens, R.; Rooman, M. Stair motifs at protein−DNA interfaces: nonadditivity of H-bond, stacking, and cation−π interactions. *J. Am. Chem. Soc.* **2004**, *126*, 6220−6221.
(6) Woolf, T. B.; Grossfield, A.; Pearson, J. G. Indoles at interfaces: calculations of electrostatic effects with desnsity functional and molecular dynamics methods. *Int. J. Quantum Chem.* **1999**, *75*, 197−206.
(7) Minoux, H.; Chipot, C. Cation−π interactions in proteins: can simple models provide an accurate description? *J. Am. Chem. Soc.* **1999**, *121*, 10366−10372.
(8) Wintjens, R.; Lievin, J.; Rooman, M.; Buisine, E. Contribution of cation−π interactions to the stability of protein−DNA complexes. *J. Mol. Biol.* **2000**, *302*, 395−410.
(9) Sippl, M. J. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **1995**, *5*, 229−235.
(10) Moult J. Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.* **1997**, *7*, 194−199.
(11) Lazaridis, T.; Karplus, M. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139−145.
(12) Misura, K. M.; Morozov, A. V.; Baker, D. Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction. *J. Mol. Biol.* **2004**, *342*, 651−664.
(13) Dehouck, Y.; Gilis, D.; Rooman, M. A new generation of statistical potentials for proteins. Submitted for publication.
(14) Wintjens, R. T.; Rooman, M. J.; Wodak, S. J. Automatic classification and analysis of α−α-turn motifs in proteins. *J. Mol. Biol.* **1996**, *255*, 235−253.
(15) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(16) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
(17) Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A* **1978**, *34*, 827−828.
(18) Møller, C.; Plesset, M. S. Note on an approximation treatment for many-electron systems. *Phys. Rev.* **1934**, *46*, 618−622.
(19) Wintjens, R.; Biot, C.; Rooman, M.; Lievin, J. Basis set and electron correlation effects on ab initio calculations of cation−π/H-bond stair motifs. *J. Phys. Chem. A* **2003**, *107*, 6249−6258.
(20) Hobza, P.; Sponer, J. Structure, energetics, and dynamics of the nucleic acid base pairs: nonempirical ab initio calculations. *Chem. Rev.* **1999**, *99*, 3247−3276.
(21) Cancès, E.; Mennucci, B.; Tomasi, J. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032−3041.
(22) Biot, C.; Buisine, E.; Rooman, M. Free-energy calculations of protein−ligand cation−π and amino−π interactions: from vacuum to proteinlike environments. *J. Am. Chem. Soc.* **2003**, *125*, 13988−13994.
(23) Gillès de Pélichy, L.; Eidsness, M. K.; Kurtz, D. M., Jr.; Scott, R. A.; Smith, E. T. Pressure-controlled voltammetry of a redox protein: an experimental approach to probing the internal protein dielectric constant. *Curr. Sep.* **1998**, *17*, 79−82.
(24) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(25) Dominy, B.; Brooks, C. L., III. Development of a generalized Born model parametrization for proteins and nucleic acids. *J. Phys. Chem.* **1999**, *103*, 3765−3773.
(26) Lazaridis, T.; Karplus, M. Effective energy function for proteins in solution. *Proteins* **1999**, *35*, 133−152.
(27) Wang G.; Dunbrack, R. L., Jr. PISCES: a protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589−1591.
(28) Henrick, K.; Thornton, J. M. PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **1998**, *23*, 358−361.
(29) Kocher, J.-P.; Rooman, M. J.; Wodak, S. J. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.* **1994**, *235*, 1598−1613.
(30) Rooman, M.; Gilis, D. Different derivations of knowledge-based potentials and analysis of their robustness and context-dependent predictive power. *Eur. J. Biochem.* **1998**, *254*, 135−143.
(31) Dehouck, Y.; Gilis, D.; Rooman, M. Database-derived potentials dependent on protein size for in silico folding and design. *Biophys. J.* **2004**, *87*, 171−181.

(32) Sippl, M. J. Calculation of conformational ensembles from potentials of mean forces. An approach to the knowledge based prediction of local structures in globular proteins. *J. Mol. Biol.* **1990**, *213*, 859−883.
(33) Dougherty, D. A. Cation−π interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science* **1996**, *271*, 163−168.
(34) Mecozzi, S.; West, A. P., Jr.; Dougherty, D. A. Cation−π interaction in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 10566−10571.
(35) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum mechanical continuum solvation models. *Chem. Rev.* **2005**, *105*, 2999−3093.
(36) Tomasi, J. Thirty years of continuum solvation chemistry: a review, and prospects for the near future. *Theor. Chem. Acc.* **2004**, *112*, 183−204.
(37) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6946−6951.

CI050395B