JCTC Journal of Chemical Theory and Computation

Energy Matrix of Structurally Important Side-Chain/Side-Chain Interactions in Proteins

Karel Berka,[#] Roman A. Laskowski,[‡] Pavel Hobza,[†,#] and Jiří Vondrášek*[,†,§]

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Biomolecules and Complex Molecular Systems, Flemingovo nám. 2, Prague, Czech Republic, Institute of Biotechnology, Academy of Sciences of the Czech Republic, Videnska 1083, 142 00 Prague, Czech Republic, Palacký University, Department of Physical Chemistry, Faculty of Science, tø. 17. listopadu 12, 771 46, Olomouc, Czech Republic, and EMBL Outstation - Hinxton, European Bioinformatics Institute, Welcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

Received January 5, 2010

Abstract: The interactions between amino acid side chains in proteins are generally considered to be the most important stabilizing factor controlling the precise arrangement of the polypeptide chain into a well-defined spatial structure. We used the RI-DFT-D method to calculate the full 20 × 20 matrix of interaction energies between all pairs of amino acid side chains. For each pair, we used a representative 3D conformation extracted from an analysis of known protein structures from Protein Data Bank (PDB). The representative comes from the largest cluster of relative orientations of the two side chains. We find that all of the calculated interaction energies between selected pairs of amino acids are attractive in the gas phase with the exception of side chain pairs having the same total charge. We compared these data with those calculated by the parm03 and OPLS-AA/L force fields to investigate the reliability of simple methods in modeling biomolecules and their behavior. The force fields yield good overall interaction energies for our set but have problems in evaluation of some particular interactions which could be of principal importance for protein stability. We then looked in detail at the 20 side chain interactions involving tryptophan. The histograms of interaction energies showed that the distributions of the interaction energies are neither normal nor Boltzmann-like and that our representative geometries correspond mostly to the minimum energy geometry which is rather poorly populated in the whole pairwise energy distribution. We concluded that cluster representatives obtained by the clusterization algorithm based on geometry criteria cannot be considered as a typical interaction for the whole side chain/side chain interaction distribution. They seem to epitomize the strongest interactions in a protein and are often functionally or structurally important.

Introduction

Proteins are built from 20 natural L-amino acids polymerized into a linear chain of various lengths which, with the exception of the "intrinsically unstructured proteins", fold into a specific and rigid 3D structure either spontaneously or with the help of various factors (chaperones etc.).[1] Anfisen's postulate that protein structure is unambiguously defined by the amino acid sequence is still, to a large extent, valid.[2] The polypeptide chain bears specific and heterogeneous chemical properties given by the different nature of composing amino acids. There is a long history of the efforts to collect and analyze the interactions between amino acid side chains in protein structures — mostly determined experimentally by X-ray crystallography or NMR methods. In the past, Miyazawa and Jernigan[3−5] and others[6−8] attempted to rationalize the character of the contacts between the side chains and to associate it with contact free energy. Such pairwise contact free energies have proven to be useful for scoring the native folds.[9] As the number of solved protein structures has become greater, the distance-dependent and orientation statistical potentials have also been proposed.[9−11] Side-chain/side-chain contacts are characterized geometri-

* Corresponding author tel.: +420 220-410-324, fax: (+420) 220-410-320, e-mail: jiri.vondrasek@uochb.cas.cz.

[†] Academy of Sciences of the Czech Republic.

[‡] European Bioinformatics Institute.

[§] Institute of Biotechnology, Academy of Sciences of the Czech Republic.

[#] Palacký University.

10.1021/ct100007y
Published on Web 05/28/2010

cally and in detail in an online accessible database of side-chain/side-chain interactions created by Laskowski et al.[12]

It is necessary to mention that the predicted free energies calculated from the contact analysis data depend on several approximations, which might not be fully valid for all proteins, as was nicely reviewed by Thomas and Dill.[13] They examined *a priori* potentials based on a simple hydrophobic-polar model. The calculated energies for all of the possible structures in a two-dimensional lattice resulted in the minimal "native" structure, which helped to construct a new potential recursively. They found that the frequencies of the selected pair of amino acids are not independent in terms of the frequencies of the other amino acids in the context of a sequence and that the extracted potential depends quite remarkably on the chain length and the composition.

To be able to evaluate the free energy of a particular amino acid in a pair interaction, one needs computational methods covering both the enthalpy and entropy terms given by the expression for the Gibbs free energy of association. The achievement of this goal can be significantly complicated by two principal difficulties. First is the level of accuracy for the enthalpy term calculation. The empirical potentials usually utilized are not of the required precision especially when the effect of the solvent has to be taken into account. Second, there is no rigorous and reliable theoretical method to evaluate the entropy term at the same level of accuracy as that for the enthalpy term. Most of the methods for the calculation of the entropic contribution are based on the positional variability determined by the NMR technique.[14]

There have been a few attempts to make a comparison of the statistical potential and the *ab initio* calculation of the interaction energy of amino acid side chains. Morozov et al.[15] reported remarkable correspondence between the knowledge-based potential of the hydrogen-bond geometries representing amino acid interactions in proteins and the *ab initio* DFT and MP2 calculations of the hydrogen-bonding energies for model systems. The same authors attempted to evaluate the potential energy surface (PES) for the interaction of aromatic residues at the MP2 and empirical potential levels. The main conclusion of this work is that the interaction is fairly well captured by the empirical potential and "that interactions between cyclic side chains contribute to the geometric distributions observed in protein structures".[16]

Here, we present the results of our study in which we describe and evaluate the interaction energies for all 20 × 20 amino acid side-chain pairs using representative geometries obtained from analysis of known 3D structures of proteins. We use several force fields as well as quantum chemistry methods both in the gas phase and in a protein/water environment. The importance of the obtained energy values for each interacting side chain pair is discussed in the context of the total interaction energy distribution between amino acid side chains.

Methods

Representative Set Selection. To obtain a representative set of amino acid side-chain pairs, we extracted data from a specially updated version of the Atlas of Protein Side-Chain Interactions from October 2006 (http://www.ebi.ac.uk/thornton-srv/databases/sidechains). The Web atlas is based on the printed atlas published in 1992 by Singh and Thornton.[17] It analyzes the interaction geometries of all 20 × 20 amino acid side-chain pairs as derived from a nonhomologous data set of 2548 3D structural models of proteins solved by X-ray crystallography to a resolution of 2.0 Å or better. For each of the 20 × 20 pairs of side chain types, each distance of side chain 2 interacting with side chain 1 is transformed into a common reference frame defined by side chain 1.

The preferred interaction geometries are determined from the local clustering in 3D of the distribution of side chains 2 relative to side chain 1. For each cluster, the most representative side chain 2 is selected, being the side chain which has the minimum total root mean squared distance to all of the other side chains in the cluster. A more detailed description can be found in ref 18. In the work described here, we used the cluster representative from the largest cluster in each of the 20 × 20 distributions. Figure 1 shows top clusters, and their representative side chains, for four example distributions, each involving Trp as side chain 1. Figure 1a and b show the top cluster geometries for Asp and Ser, respectively. Here, the location of side chain 2 is such that it can form a hydrogen bond with the nitrogen of the tryptophan. In Figure 1c and d, the interacting side chains are Leu and Lys. Here, the interactions are hydrophobic in nature, and consequently less specific and less directional.

Geometry Preparation. Each residue was truncated at the Cα atom of the protein backbone, and hydrogen atoms were added using a modified side-chain only force field[18] implemented in the Gromacs molecular dynamics package.[19] It means for example that glycine is approximated by CH_4 and Alanine by C_2H_6 groups. All of the possible positions of the polar hydrogens of the merkaptyl and hydroxyl groups of Cys, Ser, Thr, and Tyr were generated along with two neutral isomers of histidine. Proline was modeled as a cyclic tetrahydropyrrole. This model captures all specific features of proline interactions (pseudoplanarity, cyclic structure) as was shown by Biedermannova et al.[20] The positions of the hydrogens were then optimized in complex geometry for each pair using the SCC-DFTB-D method[21] in the DFTB+ package.[22] The hydrogens in the pairs containing at least one charged residue were optimized separately. The most stable pair determined by means of the benchmark method (see below) was then used for further calculations.

Calculation of the Gas-Phase Interaction Energies. The quantum mechanical energies were calculated using the RI-DFT-D/TPSS|TZVP method.[23] The RI-DFT-D energies were calculated with the Turbomole 5.9 package.[24] This BSSE-free method has proved to be reasonably accurate and computationally efficient on the subset of geometries calculated previously.[18]

We also used two modified force fields parametrized earlier − OPLS-AA/L[25] and parm03.[26] These force fields contain only amino acids truncated at the Cα atom. The residual nonintegral charge is further distributed over added hydrogen atoms attached to the Cα atom. The noncovalent

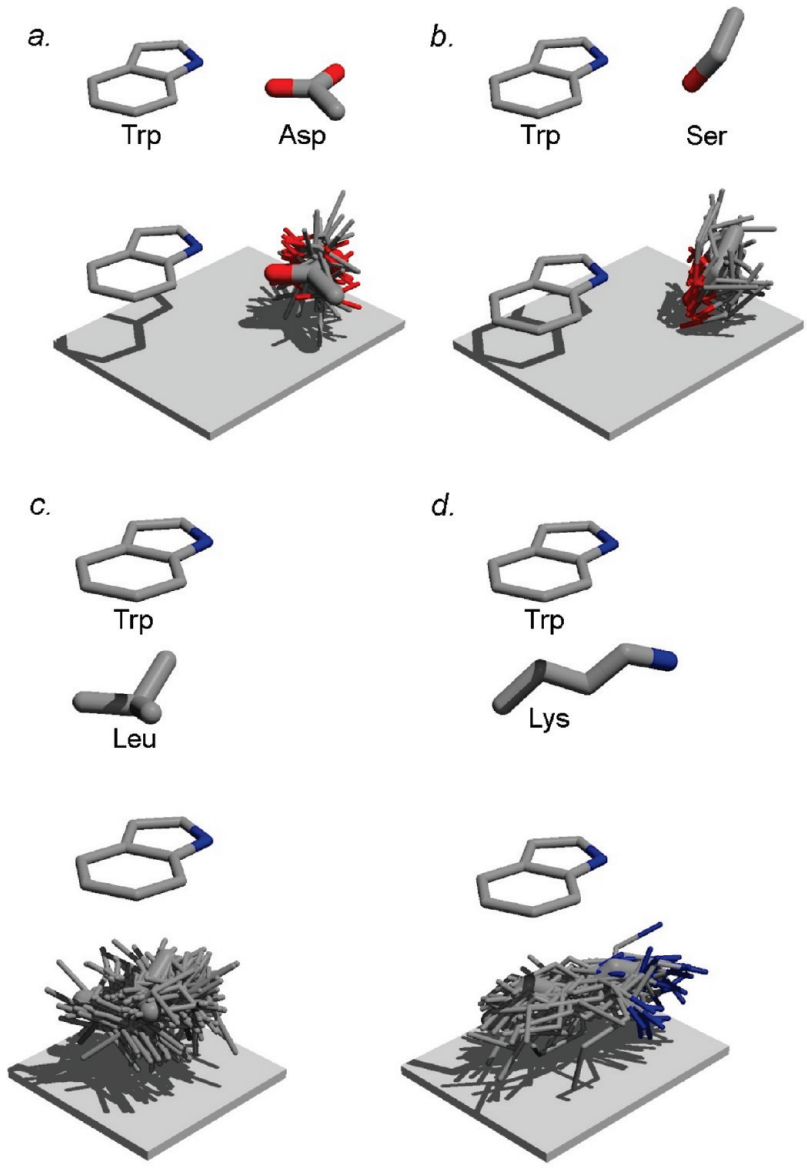

Figure 1. Some examples of side chain interactions in protein 3D structures. All examples involve interactions with tryptophan. The side chains shown are (a) aspartic acid, (b) serine, (c) leucine, and (d) lysine. Each diagram consists of two parts. The lower part shows the largest cluster of the interacting side chains, as extracted from a representative data set of protein structures in the PDB. The "cluster representative" is shown with thicker bonds. This corresponds to the side chain with the lowest total distance to all the other members of the cluster. In the upper part of each figure is shown just the Trp side chain and the cluster representative, both labeled by their three-letter code. The figure was rendered using Raster3D.[35]

interactions were calculated as a sum of the electrostatic and Lennard-Jones terms for the complexes of amino acid fragments forming a particular pair. The force-field calculations were performed with the Gromacs 4.0 package.[27]

Solvent Effect. The effect of an environment was evaluated by the RI-DFT-D method utilizing the COSMO model implemented in the Turbomole package.[28] Two dielectric constants were used to model the effect of a protein/water environment ($\varepsilon = 4, 80$) on the interaction energies.

Results

Benchmark Energy Calculations—Gas Phase Interaction Energies. We have previously shown[18] a correlation between interaction energies for the selected set of side chains evaluated by means of various theoretical methods (CCSD(T)|CBS, DFT-D, RI-DFT-D, MP2, OPLS-AA/L, and parm03 force field). We have found that RI-DFT-D is a reasonable compromise between the accuracy and the speed of the calculations, which supports our choice of this method as the reference. In this paper, we expanded the set to all of the possible combinations of 20 × 20 amino acid side-chain pairs. All of the geometries of the calculated pairs were selected by the cluster analysis described above to represent significantly populated geometry arrangements of interacting amino acids. The reference interaction energies for these pairs calculated by the RI-DFT-D method thus represent a measure of affinity based on the positions of the side chains determined experimentally and stored in the PDB database.[29] The final numbers are presented in Table 1.

Table 1. Gas Phase Interaction Energy Matrix for the Cluster Representatives for All of the 20 × 20 Possible Pairs between Residues within Proteins Calculated with the RI-DFT-D/TPSS|TZVP Method (All energies are in kcal/mol)

DFTD	G	A	V	I	L	F	Y	W	H	P	T	S	N	Q	C	M	K	R	D	E
G	-0.6	-0.7	-0.8	-0.8	-0.9	-1.0	-0.8	-1.6	-0.9	-0.2	-0.9	-1.0	-0.8	-0.8	-1.0	-0.9	-1.8	-0.4	-1.6	-3.8
A	-0.3	-0.2	-1.0	-1.4	-1.3	-1.7	-2.1	-0.7	-1.2	-1.2	-1.2	-1.4	-1.7	-1.5	-0.6	-1.5	-2.4	-3.3	-3.0	-4.6
V	-0.9	-1.5	-1.8	-1.8	-1.3	-1.3	-1.4	-2.1	-0.8	-2.1	-1.1	-1.8	-1.1	-1.5	-0.9	-1.1	-3.5	-3.4	-3.9	-2.9
I	-1.1	-1.5	-1.2	-1.5	-1.7	-3.0	-1.5	-3.0	-1.1	-1.2	-1.2	-1.7	-1.3	-1.6	-0.6	-0.7	-3.8	-3.4	-4.8	-3.3
L	-1.0	-1.0	-1.3	-1.5	-2.0	-2.4	-1.8	-3.9	-1.9	-2.3	-1.4	-1.6	-1.7	-2.3	-1.1	-2.0	-4.9	-4.5	-6.0	-6.4
F	-0.8	-1.4	-1.9	-2.7	-2.3	-2.1	-2.2	-4.6	-2.6	-2.8	-2.5	-2.5	-4.3	-3.0	-1.1	-2.1	-5.7	-9.0	-10.2	-10.2
Y	-0.7	-1.3	-2.5	-2.9	-2.3	-3.3	-3.7	-5.3	-2.8	-4.0	-1.9	-2.9	-3.4	-3.7	-1.3	-2.6	-8.1	-10.4	-29.5	-34.6
W	-1.8	-1.9	-4.5	-2.5	-2.5	-6.0	-5.6	-4.9	-4.5	-3.4	-7.4	-7.0	-5.2	-5.1	-3.7	-4.9	-9.0	-12.6	-27.4	-27.6
H	-0.9	-1.7	-1.7	-3.0	-2.7	-3.0	-2.8	-5.4	-6.1	-2.4	-5.4	-6.1	-8.3	-7.4	-3.0	-1.9	-6.8	-7.7	-27.8	-24.3
P	-1.2	-1.2	-1.9	-1.6	-1.9	-3.3	-1.6	-4.1	-3.4	-1.7	-2.4	-1.7	-0.8	-1.8	-1.1	-1.9	-1.8	-2.9	-6.5	-5.5
T	-0.5	-1.2	-0.2	-1.2	-1.2	-1.0	-3.1	-7.8	-8.0	-0.7	-7.2	-1.7	-2.5	-2.2	-0.8	-1.5	-1.7	-16.9	-12.7	-12.8
S	-0.4	-0.9	-1.8	-1.3	-1.5	-1.4	-2.3	-2.7	-0.3	-2.2	-7.3	-2.9	-6.7	-2.1	-1.1	-2.0	-9.0	-16.0	-13.8	-10.8
N	-0.9	-1.2	-1.3	-0.8	-2.1	-2.8	-3.9	-4.0	-2.6	-1.7	-0.9	-6.6	-7.2	-5.5	-1.8	-2.2	-29.8	-21.4	-25.8	-25.9
Q	-1.1	-1.3	-1.4	-1.7	-1.5	-2.1	-2.1	-4.5	-3.6	-2.7	-2.6	-1.9	-7.1	-9.8	-1.7	-2.3	-6.2	-20.9	-24.3	-25.5
C	-0.6	-0.5	-0.9	-0.7	-1.3	-1.6	-1.3	-3.6	-4.1	-1.1	-0.8	-0.9	-2.2	-2.4	-59.9	-2.4	-5.7	-9.8	-10.6	-8.8
M	-1.2	-0.5	-1.1	-1.4	-2.2	-2.7	-2.4	-2.5	-0.7	-0.9	-1.3	-1.6	-3.8	-3.0	-1.4	-1.9	-6.4	-7.9	-7.0	-11.9
K	-1.9	-2.2	-3.8	-3.7	-3.1	-5.7	-9.5	-6.8	-3.8	-1.3	-2.1	-7.1	-28.9	-28.3	-5.5	-7.3	58.7	55.8	-113.8	-113.7
R	-1.6	-2.8	-3.6	-3.8	-3.6	-7.5	-8.6	-10.6	-6.1	-1.4	-3.5	-15.7	-20.0	-22.8	-5.7	-7.5	51.1	50.7	-115.6	-107.1
D	-1.4	-3.1	-3.3	-5.7	-6.0	-7.1	-40.1	-24.1	-31.6	-8.7	-7.0	-12.0	-27.1	-26.8	-6.7	-4.2	-116.1	-126.5	62.5	50.1
E	-2.1	-2.8	-3.7	-4.1	-4.5	-4.9	-37.2	-27.2	-26.6	-8.2	-12.0	-12.5	-7.2	-26.0	-9.0	-8.6	-109.9	-140.1	51.9	70.4

The first of the important results is that all of the interaction energies for structure representatives presented in Table 1 are attractive with only a few regular exceptions—the pairs containing amino acids with a similar charge. The result thus reflects an important fact regarding the protein's intramolecular stabilization provided by the selective arrangement of interacting amino acids. It must be stressed again that the interaction energy of all of the amino acid pairs was calculated exclusively for the most populated cluster representative geometry, which most probably represents the local distance minimum. The lack of destabilizing contributions is then not so surprising. We can imagine the existence of sterical barriers caused by a tight arrangement of secondary structure elements which could include the studied amino acids. Such an environment may sometimes push the amino acids out of the attractive regime and could result in repulsive behavior of the interacting amino acids. We do not report a single case of such an interaction mode (with the above-mentioned exceptions) for the studied set.

Asymmetry of the Interaction Energy Matrix. The second of the important results which should be properly explained is the asymmetry of the interaction energy matrix. The asymmetry of the matrix is a consequence of the way the clusters were calculated. Figure 2 shows the explanation of the matrix feature. In Figure 2, a and b show two separate interactions between side chains of type S and T: S1 with T1 and S2 with T2. When these interactions are superposed in the frame of reference of the S residues (c), the T residues come close together and might fall in the same cluster. However, when the interactions are superposed on the T residues (d), the S residues are thrown apart and would be unlikely to fall into the same cluster. For the side chain Atlas, this is not a problem, as one is interested in the distribution of side chain B around side chain A, and where the highest concentrations of B are, relative to A (and vice versa). For a symmetrical matrix, however, one would need to calculate full distributions for each pair of amino acids. So, in Figure 2c, one would need to calculate the RMSD between T1 and T2 and to add to it the RMSD between S1 and S2 when the T side chains are superposed (as in d). We still think that the definition used in the Atlas is a fair way to look at the data and accept the asymmetry because we are interested, in principle, about significant structure features which are most probably based on certain geometry preferences between interacting amino acids.

It is worth stressing here again that first we have calculated the pairwise noncovalent interactions between amino acids in the gas phase and that the influence of the environment has not been taken into account. The fact that the system exists in an environment can change the stabilization

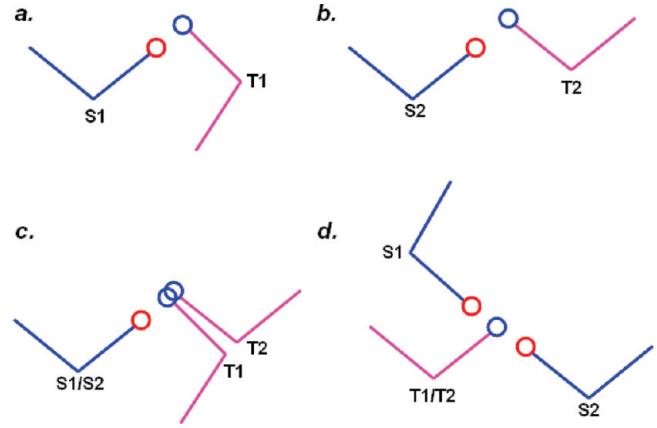

Figure 2. (a and b) Two separate interactions between side chains of type S and T: S1 with T1 and S2 with T2. When these interactions are superposed in the frame of reference of the S residues (c), the T residues come close together and might fall in the same cluster. However, when the interactions are superposed on the T residues (d), the S residues are thrown apart and would be unlikely to fall into the same cluster.

proportions considerably is part of the latter chapter about solvent effects.

We divided the interactions into several groups according to the chemical properties of the compositional amino acids, and in the following paragraphs we shall describe the results separately.

Charge−Charge. The most attractive interactions in the gas phase were obtained for salt bridges − approximately 100 kcal/mol. The interaction energies where Arg is in a pair with a negatively charged carboxylic acid are slightly stronger (up to 140 kcal/mol) in comparison with those containing a Lys residue. This may be the effect of the additional hydrogen bond from the second NH group participating in the interaction or the influence of the electron distribution of the guanidinium group.

Charge−Neutral. The interactions of the charged residues are generally quite strong in the complexes with polar or aromatic residues − around 10−20 kcal/mol for positively charged and 20−30 kcal/mol for negatively charged residues. This difference is more profound for the charged-aromatic pairs. The geometry of the negatively charged-aromatic pairs is different from that of those containing positively charged side chains. The basic amino acids usually interact with aromatics in a stacking-like manner, unlike the acidic residues, which prefer more directional, mostly H-bond interaction. We can find a further difference between Arg and Lys which arises from the fact that Arg mostly stacks above the plane of the ring of the aromatics, indicating clearly a more dispersive character of interaction. Lys, on the other hand, possesses its long aliphatic chain above the ring, so the charged amine group is farther from the ideal contact with the aromatic ring. Both negatively charged residues are oriented in such a way that their carboxylic group is in the plane of the ring with negligible electron contacts with the aromatics. They interact either with the hydroxy group of Tyr or with the amide group of Trp or His. In the case of Glu, Asp−Phe interactions, the carboxylic group is to a certain extent in an interaction with the main chain of the aromatic and is not oriented above the highest electron density on the aromatics.

Polar−Polar and Polar−Aromatics. The third class of interactions is polar−polar and polar−aromatic contacts. Their interaction energy is about 5 kcal/mol. As they are mostly based on the formation of hydrogen bonds in an orientation-dependent manner, the resulting interaction energies fluctuate in the largest range even for the same pairing of amino acids. Good examples are the Thr−Ser pair, where the interaction energy is only −1.7 kcal/mol, and the Ser−Thr pair, which is much stronger, namely at −7.3 kcal/mol. At this point, we have to stress that in both cases the best combination of the rotamers and optimal position of both of the hydroxyl groups was used.

Aromatics−Aromatics. It is well-known that the aromatic−aromatic pairing is abundant in proteins and is also quite homogeneous because of the similar character of the interacting residues. Their interaction energies on average are around 5 kcal/mol. The strongest interaction among aromatic residues comes from pairs containing Trp, mostly due to the aromatic character and size of its indole ring. The Trp is followed by His, which interacts mostly through the hydrogen bond. It should also be noted that the result is dramatically influenced by the selection of an appropriate isomer.

Aliphatics−Others. The largest group of interactions comprises pairs containing aliphatic residues. Their interaction energy with most of the residues is quite small (below 2 kcal/mol) with the exception of the aliphatic−charged and the aliphatic−aromatic pairs, which are stronger. Polar residues cannot create hydrogen bonds and are perceived mainly by dispersion interactions. Proline exhibits special features: it behaves similarly to an aliphatic residue of the same size (Leu, Ile), but its interaction with the charged residues is different.

Sulfurics−Others. A small group can also be derived from sulfur-atom-containing residues, namely, Cys and Met. Their interaction energy is similar to aliphatic residues but with several notable deviations. The biggest difference is the Cys−Cys pair, which is bound covalently and thus cannot be compared to the other cases. However, its dissociation energy for the disulfide bond is about 65 kcal/mol.[30] Because of the better polarizability of the sulfur atom, its interaction with charged residues is approximately twice as strong as in the case of the charged-aliphatic pairs.

Looking at each residue individually in terms of its total interaction energy with all amino acids, we can create a "stability line". It runs from the residue of the highest stabilization potential to the lowest. The energy differences between adjacent residues in the stability line are not constant, so they can be grouped into subclasses. The ">" signifies an important change in the interaction energies:

$$\text{D, E} > \text{R} > \text{K} > \text{N, Q} > \text{W, Y} > \text{H} > \text{S} > \text{T} > \text{F} > \text{M, C} > \text{P, L} > \text{I, V} > \text{A} > \text{G}$$

The strongest stabilization not surprisingly comes from interactions of charged residues even when we take into account the repulsion between amino acids of the same charges. The stability line continues with polar and aromatic residues and ends with aliphatic residues according to their size. It should be mentioned again here that these values are gas phase interaction energies, and hence no effect of an environment has been taken into account.

Parm03 and OPLS-AA/L Force Field Interaction Energies. We evaluated the interaction matrix for the same set of structures with two force fields typically used for the protein study, i.e., Amber parm03 and OPLS-AA/L. We aimed to provide a quantitative comparison between the results obtained by the RI-DFT-D energies and the molecular-mechanical force-field methods. One has to be aware of the difference between the calculation of interaction energies by empirical study and by the quantum chemical approach. In the empirical potential case, the interaction energy is calculated as a sum of the nonbonded interactions between each atom in both residues. The interactions involve a Coulombic term for electrostatics and the Lennard-Jones 6−12 term covering the van der Waals contributions. In the case of RI-DFT-D, the interaction energy is calculated as the difference between the total energy of the complex and the energies of both subsystems.

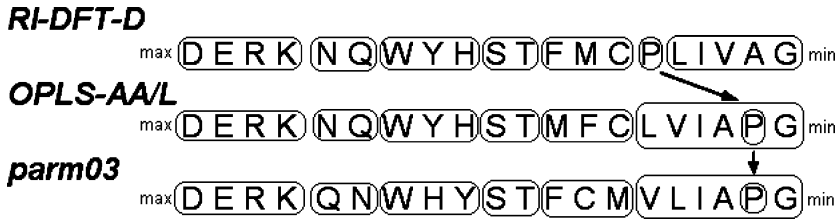

Figure 3. Amino acid families sorted by their summed interaction energy for RI-DFT-D, parm03, and OPLS calculations.

The most notable differences between the force-field and RI-DFT-D energies are substantially weaker interactions per residue provided by empirical potential methods. In contrast with the RI-DFT-D results, we detected additional repulsive behavior of some interacting pairs with the exception of the pairs with the same charges. One of the reasons may be accounted for the Lennard-Jones repulsive term in the force field being generally too steep. This can cause difficulties arising from the fact that the positions of the hydrogens were optimized at the RI-DFT-D level, which consequently shortens the inter-residue distance. This usually leads to an increase of the Coulombic term (repulsive for two hydrogen atoms) and also to the enlargement of the repulsion coming from the Lennard-Jones term in the force-field calculation. Both these aspects can contribute to the overall repulsion calculated energy computed by the force-field method for the pairs which are still attractive in RI-DFT-D. We can also rationalize a reason for a higher number of repulsions coming from the Pro residue from the way we treated this amino acid. The simplification of Pro alters the hybridization state of the N atom from sp^2 (planar) to sp^3 (tetrahedral). This situation improved the interaction energy in quantum chemical methods in contrast with force fields, which lacks proper parameters for such a state.

Last but not least, the partial charges on the amino acids in the force field are generally adjusted for the solvent environment. Depending on the geometry of the interacting amino acid in the structural context of the protein, these interactions are sometimes under/overestimated, which also strengthens the difference in comparison with the quantum chemical interaction calculations in the gas phase.

The previously defined "stability line" can be further subdivided into "families" of amino acids according to their physical-chemical similarities (Figure 3). We define the families as follows: charged residues (DERK), polar residues with peptide-bond motifs (NQ), aromatic residues with at least one polar atom (WYH), hydroxyl-containing polar residues (ST), polarizable residues with electron-rich regions (FMC), unique proline ring (P), and aliphatic residues sorted by their respective size (LIVAG). As is apparent from Figure 3, the amino acids in families behave similarly in all of the methods used.

Both interaction matrices are similar (data not shown), with their correlation coefficients being higher than 0.95. The differences in the results can be seen at the level of the average interaction energies in comparison with the RI-DFT-D values. Both force fields have lower average values of interaction energies (parm03, −3.8 kcal/mol; OPLS, −4.5 kcal/mol; while RI-DFT-D, −6.2 kcal/mol) as well as median values (parm03, −1.6 kcal/mol; OPLS, −1.6 kcal/mol; while RI-DFT-D, −2.5 kcal/mol). Particularly, the median values demonstrate that the force fields have a high level of similarity and that the energies are weaker than those obtained by the RI-DFT-D method. An important detail contributing to the difference between the RI-DFT-D and empirical potential results arises from the fact that force-field parameters in both utilized empirical methods are optimized for molecules in a solvent environment.

The Effect of Environment on the Interaction Energies in the Matrix. While all of the interactions in the gas phase can be calculated explicitly and in principle with reasonable accuracy, most of the interactions of biomolecules and their complexes are realized in a protein or water environment, which makes a precise evaluation of the interaction energy complicated if not impossible because of the heterogeneous conditions around the interacting residues. In order to take the environment roughly into account, we used solvent-implicit models. We used two dielectric constants: $\varepsilon = 4$, mimicking the effect of a protein environment, and $\varepsilon = 80$, for the effect of water. We calculated the interaction energies by the RI-DFT-D method with the COSMO implicit-solvent model.

The results presented in Tables 2 and 3 show that the higher the dielectric constant of the surrounding, the smaller the differences between the interaction energies for all of the interacting pairs of amino acids. The apparent reason is the dielectric screening of the dominant electrostatic interaction. The consequence of this effect is a decrease of the average and the median of the interaction energy. In comparison with the gas phase interaction energy median (−2.5 kcal/mol), the value in a protein-like environment ($\varepsilon = 4$) is −1.4 kcal/mol and in a water-like environment ($\varepsilon = 80$) is only −0.9 kcal/mol.

Charge−Charge. Charged pairs lose most of their interaction energies upon the introduction of the solvent in comparison with other pairs. This is caused by the screening of a substantial part of their interaction energy being dominated by electrostatics. On the basis of the values presented in Tables 1−3, we observed that the like-charged pairs dropped more in their repulsive interaction energy (33%, 4.4% of interaction in gas phase for $\varepsilon = 4$, 80) than the salt-bridge pairs (37%, 7.7% for $\varepsilon = 4$, 80 of the gas-phase values). The repulsion existing in the gas phase can even be surpassed, and the pairs of like-charge residues show an attractive character in a water environment (Arg−Arg). This behavior has recently been reported by Vondrášek el al.[31] on Arg−Arg as a potential stabilizing factor in proteins.

Charge−Neutral. Also, charge−neutral pairs are quite weakened by the presence of a solvent. However, the weakening of the interaction energies is smaller than in the

Table 2. Interaction-Energy Matrix for the Cluster Representatives for All of the 20 × 20 Possible Pairs between Residues Calculated with the RI-DFT-D/TPSS|TZVP Method with the COSMO Model in a Protein-Like Environment ($\varepsilon = 4$) (All energies are in kcal/mol)

	G	A	V	I	L	F	Y	W	H	P	T	S	N	Q	C	M	K	R	D	E
G	-0.5	-0.7	-0.8	-0.8	-0.8	-0.1	-0.5	-1.1	-0.7	-0.2	-0.7	-0.3	0.1	-0.7	-0.8	-0.7	-1.0	0.3	0.1	-0.4
A	-0.3	-0.2	-1.0	-1.4	-1.3	-1.5	-1.7	-0.2	-1.0	-1.1	0.1	-0.6	-0.7	-1.2	-0.2	-1.3	-0.6	-0.8	-0.6	-0.6
V	-0.8	-1.4	-1.7	-1.7	-1.3	-1.0	-1.1	-1.9	-0.4	-2.0	-1.0	-0.5	-0.7	-1.1	-0.4	-0.8	-1.0	-2.0	-1.2	-0.7
I	-1.0	-1.5	-1.1	-1.4	-1.7	-2.7	-1.4	-2.6	-0.8	-1.2	-1.2	-0.8	-0.8	-1.1	-0.1	-0.4	-1.2	-2.1	-0.9	-0.8
L	-1.0	-1.0	-1.3	-1.5	-1.9	-2.0	-1.6	-3.3	-1.5	-1.4	-1.3	-1.4	-1.1	-1.0	-0.6	-1.5	-2.4	-2.7	-1.6	-1.2
F	-0.6	-1.1	-1.6	-2.3	-1.8	-1.6	-1.9	-3.8	-1.8	-2.4	-1.5	-1.5	-2.5	-2.0	-0.4	-1.3	-2.4	-4.3	-2.4	-1.8
Y	-0.4	-0.9	-2.0	-2.3	-1.7	-2.9	-3.2	-4.3	-1.0	-2.7	-1.2	-1.9	-1.8	-2.3	-0.3	-1.9	-4.4	-4.5	-14.9	-20.4
W	-1.3	-1.4	-3.8	-2.3	-2.0	-4.7	-4.4	-4.0	-3.6	-2.8	-5.3	-4.8	-2.9	-3.4	-2.6	-4.1	-4.0	-5.8	-11.8	-13.9
H	-0.6	-1.3	-1.4	-2.5	-2.0	-2.1	-2.0	-3.6	-2.5	-2.0	-3.7	-4.1	-5.1	-4.8	-1.7	-0.8	-3.2	-3.2	-12.9	-11.0
P	-0.9	-1.1	-1.8	-0.8	-1.2	-2.0	0.1	-2.5	-2.0	-1.4	-0.9	-0.4	0.7	-0.8	-0.6	-0.2	-1.5	-1.9	-1.3	-0.2
T	0.3	-0.3	-0.2	-1.1	-1.1	-0.9	-2.3	-5.3	-5.7	-0.7	-5.2	-0.1	-0.1	-0.6	-0.3	-1.2	-0.6	-9.1	-4.9	-5.2
S	0.5	0.1	-1.3	-1.2	-0.8	-1.1	-1.6	-1.9	-0.1	-1.2	-5.0	-1.2	-4.6	0.0	-0.4	-1.2	-3.0	-8.3	-5.6	-3.7
N	-0.2	-0.6	-0.7	-0.3	-1.3	-1.2	-2.2	-1.8	-1.3	-1.0	-0.6	-4.3	-4.7	-3.0	-0.2	-1.1	-14.7	-10.6	-11.7	-11.9
Q	-0.8	-0.9	-1.0	-1.2	-1.0	-1.2	-1.3	-2.9	-1.9	-2.0	0.1	0.6	-4.6	-7.0	-0.1	-1.4	-2.6	-10.4	-10.8	-12.6
C	-0.4	-0.3	-0.3	-0.2	-0.7	-0.8	-0.5	-2.4	-2.6	-0.7	-0.3	1.4	-0.3	-1.3	-56.3	-1.6	-2.3	-3.3	-2.8	-2.7
M	-0.8	-0.4	-0.9	-1.3	-1.8	-2.1	-1.8	-2.3	0.0	-0.5	-0.9	-1.2	-1.9	-1.8	-0.8	-1.6	-2.5	-2.8	-1.0	-4.0
K	-1.0	-1.4	-1.4	-0.8	-2.0	-2.9	-3.9	-3.0	-1.8	0.4	-0.2	-2.0	-13.5	-13.8	-1.2	-3.1	20.8	19.0	-41.1	-42.1
R	-0.6	-1.7	-2.1	-2.3	-1.9	-2.8	-4.8	-4.9	-3.1	0.3	-2.0	-8.2	-10.2	-11.4	-2.7	-3.4	15.2	15.9	-44.5	-38.3
D	0.4	-0.3	-1.1	-1.9	-1.2	-2.6	-24.9	-11.2	-14.5	-2.7	-1.1	-5.1	-12.1	-12.3	-2.2	-1.4	-41.6	-51.8	23.9	16.3
E	0.5	-0.5	-1.2	-1.1	-0.9	-0.8	-22.9	-13.8	-13.0	-3.0	-4.2	-5.1	-2.5	-13.1	-3.3	-2.8	-38.5	-61.0	16.6	30.6

Table 3. Interaction-Energy Matrix for the Cluster Representatives for All of the 20 × 20 Possible Pairs between Residues Calculated with the RI-DFT-D/TPSS|TZVP Method with the COSMO Model in a Water Environment ($\varepsilon = 80$) (All energies are in kcal/mol)

	G	A	V	I	L	F	Y	W	H	P	T	S	N	Q	C	M	K	R	D	E
G	-0.5	-0.6	-0.8	-0.7	-0.8	0.0	-0.3	-0.8	-0.5	-0.1	-0.6	0.1	0.7	-0.6	-0.6	-0.6	-0.7	0.5	0.8	1.0
A	-0.2	-0.2	-0.9	-1.3	-1.2	-1.3	-1.6	0.1	-0.9	-1.1	0.9	-0.2	0.0	-1.0	0.0	-1.2	0.0	0.0	0.3	1.0
V	-0.8	-1.4	-1.7	-1.7	-1.3	-0.9	-1.0	-1.7	-0.1	-2.0	-1.0	0.2	-0.4	-0.9	-0.1	-0.6	-0.2	-1.6	-0.4	0.1
I	-1.0	-1.4	-1.1	-1.4	-1.6	-2.5	-1.3	-2.4	-0.6	-1.1	-1.1	-0.2	-0.5	-0.8	0.3	-0.3	-0.4	-1.7	0.4	0.0
L	-0.9	-1.0	-1.2	-1.4	-1.9	-1.8	-1.5	-2.8	-1.2	-0.8	-1.3	-1.3	-0.8	-0.2	-0.4	-1.3	-1.7	-2.2	0.0	0.9
F	-0.4	-0.9	-1.4	-2.0	-1.6	-1.2	-1.7	-3.3	-1.4	-2.1	-0.9	-0.9	-1.3	-1.4	0.0	-0.8	-1.4	-2.6	0.5	1.8
Y	-0.2	-0.6	-1.7	-1.9	-1.4	-2.6	-3.0	-3.7	0.2	-2.0	-0.7	-1.3	-0.7	-1.4	0.3	-1.5	-3.0	-2.0	-8.8	-14.4
W	-0.9	-1.1	-3.4	-2.1	-1.6	-4.0	-3.7	-3.5	-3.1	-2.3	-4.1	-3.6	-1.3	-2.3	-1.9	-3.6	-2.5	-3.1	-4.8	-7.7
H	-0.3	-1.0	-1.2	-2.2	-1.5	-1.5	-1.5	-2.4	0.0	-1.8	-2.6	-2.8	-3.0	-3.0	-0.8	-0.1	-2.1	-1.4	-5.5	-4.5
P	-0.8	-1.1	-1.8	-0.2	-0.7	-1.1	1.2	-1.5	-1.1	-1.3	0.0	0.4	1.6	-0.2	-0.2	0.9	-1.7	-1.9	0.7	1.9
T	0.7	0.3	-0.2	-1.1	-1.1	-0.8	-1.8	-3.9	-4.4	-0.7	-4.2	0.8	1.3	0.2	0.0	-1.1	-0.5	-5.4	-1.5	-1.7
S	1.0	0.6	-1.0	-1.2	-0.4	-0.9	-1.2	-1.3	0.0	-0.7	-3.8	-0.2	-3.5	1.3	0.0	-0.8	-0.2	-4.7	-1.9	-0.3
N	0.3	-0.2	-0.4	0.0	-0.7	-0.1	-1.1	-0.4	-0.4	-0.7	-0.5	-2.9	-2.9	-1.4	0.9	-0.3	-7.4	-5.1	-4.6	-4.8
Q	-0.5	-0.7	-0.8	-0.9	-0.6	-0.6	-0.9	-2.1	-0.7	-1.6	1.8	2.0	-2.9	-5.2	0.9	-0.7	-1.2	-5.2	-4.0	-6.3
C	-0.3	-0.2	0.0	0.2	-0.4	-0.3	0.1	-1.8	-1.7	-0.4	0.0	2.7	1.0	-0.6	-55.6	-1.1	-0.7	-0.2	1.0	0.1
M	-0.6	-0.2	-0.8	-1.2	-1.6	-1.8	-1.4	-2.2	0.5	-0.2	-0.7	-0.9	-0.8	-0.9	-0.4	-1.3	-0.9	-0.6	1.5	-0.5
K	-0.7	-1.2	-0.6	0.1	-1.7	-2.1	-1.5	-2.2	-1.3	0.8	0.2	0.4	-5.8	-6.7	0.8	-1.2	3.0	2.1	-8.1	-9.6
R	-0.3	-1.3	-1.7	-1.9	-1.4	-1.1	-3.5	-2.8	-1.8	0.9	-1.4	-4.6	-5.3	-5.7	-1.4	-1.6	-1.2	0.1	-12.9	-7.4
D	1.1	0.9	-0.5	-0.5	0.5	-1.1	-18.5	-5.5	-5.9	-0.2	1.5	-2.1	-4.4	-5.1	-0.7	-0.3	-7.7	-18.4	6.3	1.2
E	1.5	0.3	-0.4	-0.1	0.4	0.7	-16.7	-7.8	-6.4	-0.7	-0.7	-1.8	0.1	-6.6	-0.5	-0.4	-6.1	-25.2	0.9	12.8

case of charge−charge pairs (43%, 28% for $\varepsilon = 4$, 80 of the gas-phase values). This fact is in concord with the smaller total interaction energies, not as dominated by electrostatics as is the case of charge−charge pairs.

Polar−Polar or Aromatics. A solvent has a smaller effect on these pairs in comparison with the previous cases. While pairs with a mixed character of polar and aromatic residues are more sensitive to the effect of a water environment (64%, 41% for $\varepsilon = 4$, 80), the polar−polar contacts surprisingly are less affected (65%, 54% for $\varepsilon = 4$, 80).

Aromatics−Aromatics. The decreasing sensitivity of these interacting pairs to the effect of the environment is demonstrated by a moderate decrease of the interaction strength (79%, 68% for $\varepsilon = 4$, 80). The reason for such insensibility is the different nature of their interaction as reported in Berka et al.[32]

Aliphatics−Others. Aliphatic residues are the least sensitive to the effect of the environment. Their interaction energies are almost constant for aliphatic−aliphatic pairs (95%, 91% for $\varepsilon = 4$, 80). Their interactions with polar or

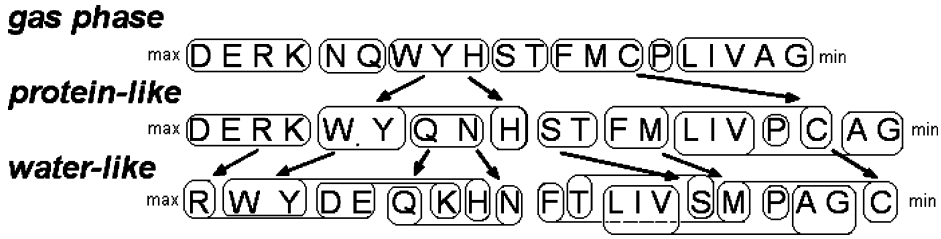

Figure 4. Amino acid families in the environment sorted by their total interaction energy provided by the RI-DFT-D calculations. The dispersion-bound residues are generally shifted upward unlike the electrostatic ones.

aromatic residues are slightly more sensitive (polar: 70%, 59%; aromatic: 79%, 67%), reflecting the different proportionality of the stabilizing forces.

Sulfurics−Others. Sulfur-containing residues act similarly in an environment to the polar residues (62%, 58% for $\varepsilon =$ 4, 80).

The pair interaction energies between the residues are influenced differently by the solvent depending on their characteristics. Charged residues are the most sensitive to the effect of their environment, followed by polar and sulfur-containing residues. Aromatic and aliphatic residues sustain most of their interaction energy despite the significant environment change. This different sensitivity to the environment changes the positions of amino acids and families in the stability line quite significantly, as can be seen in Figure 4.

The strongest effect of the environment is a change of the relative positions of residues in the stability line. The environment promotes the interaction between the residues of an aromatic or aliphatic character (mainly Trp, Tyr, Leu, Ile, and Val). On the other hand, the strength of the interactions involving charged residues is lowered significantly by a water environment, with the only exception being Arg, whose guanidinium group also has a strong dispersion interaction. The polar and sulfuric groups are shifted toward lower stability, whereas the smaller residues of the same kind are moved more (Asn more than Gln, Ser more than Thr, and Cys more than Met). This can be accounted for by the less extensive dispersion interactions.

Interaction Energy Distributions in the Gas Phase. A major question of this study is how the selected cluster representatives are relevant to the overall energy distribution for all of the interacting residues with a particular amino acid, or better said how representative these interactions are. We have shown previously that the interaction energy of the cluster representative is a reasonable approximation of the interaction energy of the whole selected cluster for one particular pair.[18]

The calculation of the 20 × 20 interaction matrix of the cluster representatives shows that some interactions are not symmetrical in terms of their energies. This is more profound for the polar amino acid side chains, namely, Ser−Thr and Thr−Ser. They differ significantly in their interaction energies for cluster representatives (−7.3 vs −1.7 kcal/mol). While the total number of interactions is the same for both pairs, the clustering algorithm apparently provided two different geometries for the cluster representatives. One can expect a symmetry of the interaction energy values if the ensemble of structures is large enough to result in the same geometry for both representatives obtained by the cluster-analysis algorithm.

Our aim was to describe the cluster representative in the context of the overall geometry distribution for the selected pair of amino acids appearing in proteins. To see just how representative it is of the whole distribution required computing all interaction energies for a given side chain/side chain distribution and comparing with the energy of the representative conformation. The only way to achieve this in a reasonable time was to use the parm03 force field. We chose tryptophan (Trp) as our side chain 1 and calculated its interactions with all 20 amino acid side chains. Trp was chosen because of the reasonable level of agreement between the *ab initio* and force-field results for the calculations involving Trp. Moreover, the interactions with this residue do not show any repulsive interaction energy values in any of the methods used.

The results for the gas phase are presented in Figures 5 as histograms of the calculated interaction energies with parm03 force field. Most of the histograms have one peak slightly below the zero value. Only the interactions of Trp with negatively charged residues have clear two-peak distributions. Most of the distributions of interaction energies seem to be limited by zero on one side and by the cluster representative value at the other extreme. To confirm the behavior by a higher level of methodology, we recalculated the distribution for 100 randomly selected pairs for every 20 amino acids interacting with Trp by RI-DFT-D for the optimized structures. As follows from our analysis (see Figure 6; Figure 7 shows those with the OPLS-AA force field), both distributions are very similar, and the energy limit at the zero value is clearly more distinct for the histograms of the RI-DFT-D energies.

One important conclusion can be made on the basis of the obtained results. The cluster representative values are mostly extreme cases of the side-chain/side-chain interactions and cannot serve as a measure of the interaction-energy distribution or its typical value. Particularly in the case of side chains involved in hydrogen bonds with the Trp (eg Asp, Glu, Ser, and Thr), the representatives tend to correspond to low-energy conformations. Where interactions are less directional, as in interactions involving hydrophobic contacts, the representative does not necessarily have a low energy conformation. The results are summarized in Table 4. We cannot relate the data obtained

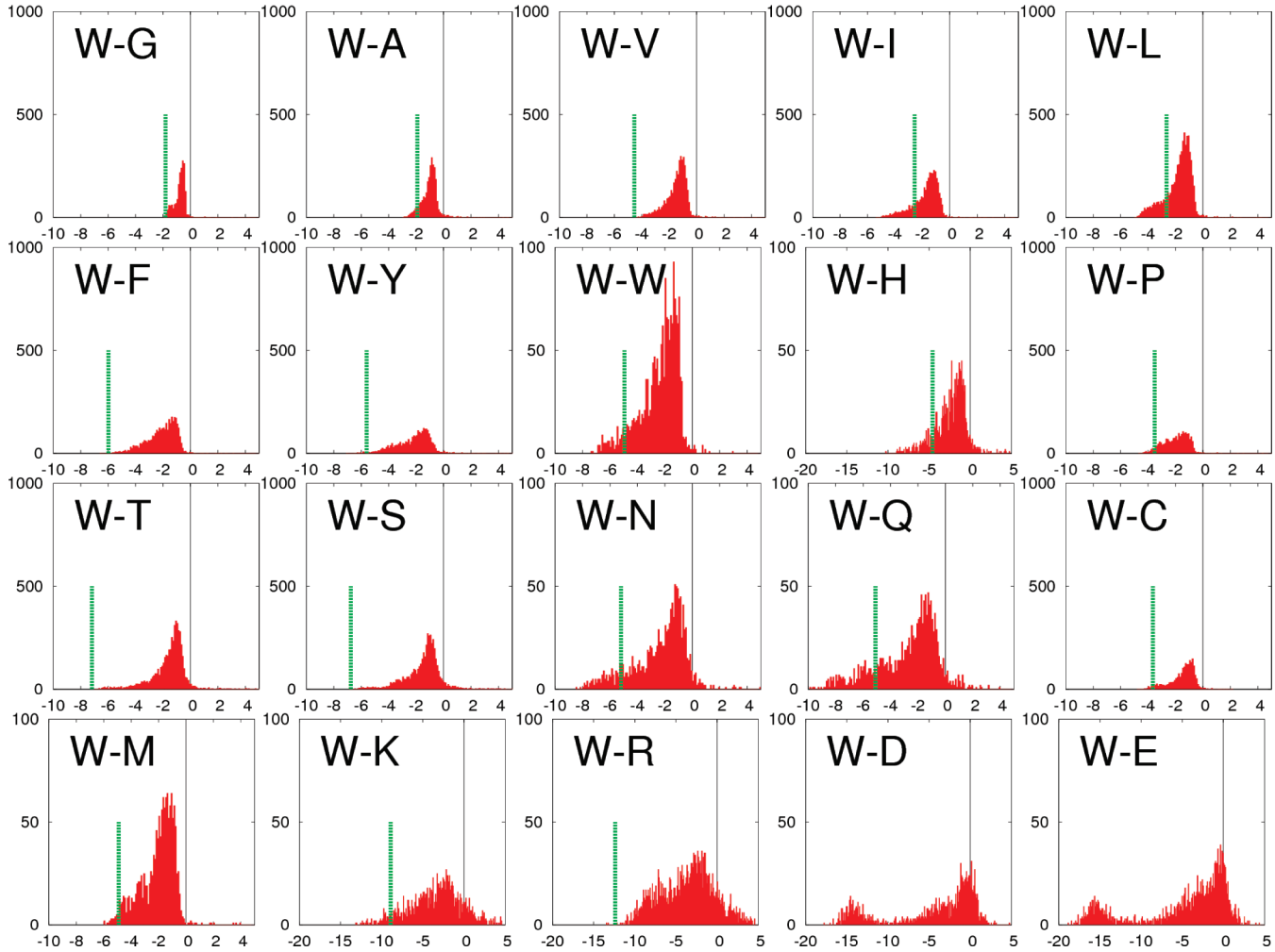

Figure 5. Histograms of the interaction energies of the side chain/side chain interactions of tryptophan with all of the other residues calculated with the parm03 force field in the gas phase. Energies on the *x* axis are in kcal/mol.

for the representative interactions of a particular amino acid to any of the phenomenological matrices published.[4,7,11]

Discussion

It is not a simple task to establish properly the meaning of the calculated interaction energies between the amino acid side chains, especially if we take only one particular interaction as the representative. As described earlier, there are two extreme views of the side-chain interactions in proteins. The first extreme is that their arrangement is completely random and mostly the backbone properties dictate the fold; the second view is that these interactions are the basis of intramolecular stabilization of the fold and their positions are energy tuned. On the basis of our results presented here, we see the protein stabilization and fold in proportion to the specific and nonspecific interactions depends on the structural and sequential contexts of the protein in question.

The complete interaction energy matrix for all of the amino acids in proteins supports the view that the cluster representatives describe the important spatial but mostly local interactions selected by the character of the residue to maximize the interaction strength in a well-defined spatial arrangement. This view is supported by our previous analysis of the cluster representative, which is a maximum of the distribution of the interaction energy for a certain cluster.

Additionally, all of the calculated interaction energies in the matrix were attractive. This is not a trivial finding, even if valid for cluster representatives. The common way of interpreting side-chain/side-chain interactions in proteins is that the resulting interaction is a balance between stabilization and repulsion. Some side chains are displaced in nonfavorable orientations (in extreme cases, they can be repulsive) caused by a much greater influence of the adjacent residues or the secondary structural elements. Our data suggest that this is not the case — at least not for such geometrically exclusive interactions as the calculated set constitutes. A general explanation for protein folding can be attributed to the fact that the sharp character of repulsion does not allow side chains to occupy unfavorable positions and the typical pair geometry in proteins is always adjusted to prevent such an interaction mode.

We are aware of the fact that the benchmark RI-DFT-D values slightly overestimate the interaction energies (by 0.3 kcal/mol on average) for the weakly bound pairs of aliphatic residues such as Ala−Leu as we have proven in a previous paper,[18] and therefore the values are generally higher than

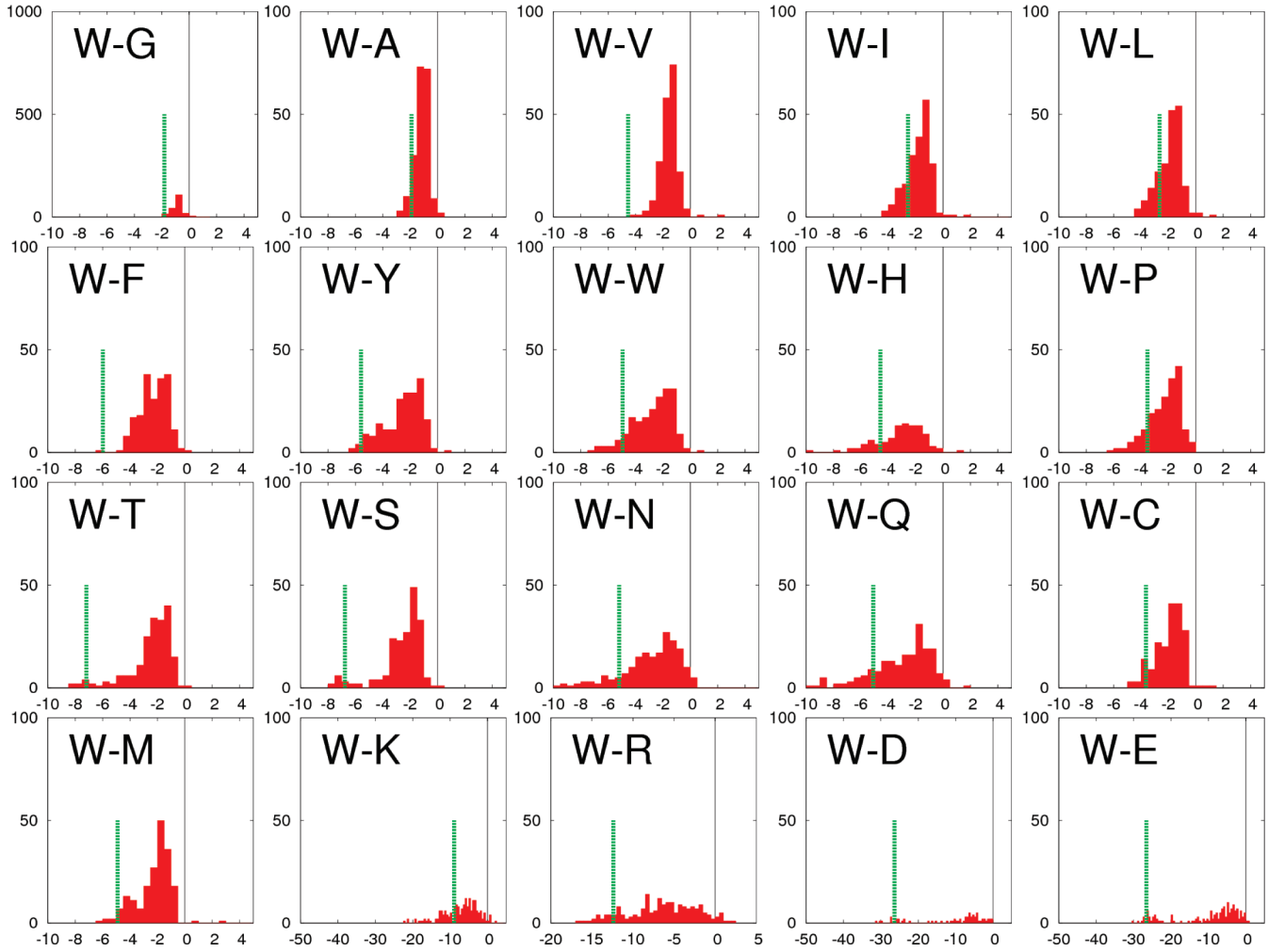

Figure 6. Histograms of the interaction energies of the side chain/side chain interaction of tryptophan with all of the other residues calculated with the RI-DFT-D method in the gas phase. Energies on the *x* axis are in kcal/mol.

those obtained by empirical potentials. Another fact contributing to this difference is that both force fields are parametrized for a solvent in which the interactions are screened by the environment, namely, the partial charges for the parm03 force field have been parametrized for the dielectric continuum model[26] with a dielectric constant equal to 4. Furthermore, the atomic charges for the OPLS-AA/L force field were derived from the parm94 force field. Finally, the non-negligible problem in the utilization of the force field in this study and the interaction energy's accuracy is caused by the addition and optimization of hydrogen atoms. The SCC-DF-TB-D optimization makes them too near the atoms of the other interacting residue. While the optimization of hydrogens had a stabilization effect in the quantum chemical calculations, the opposite is true for the force-field interaction energies.

Although the correlation between the calculated interaction energy matrices is high, especially for the gas phase energies ($r = 0.98$ and 0.95 for OPLS and parm03 in comparison with the RI-DFT-D energies, respectively), the particular differences are not negligible. Fortunately, the interaction energies as a whole are parametrized quite successfully in force fields, but they can vary quite significantly in specific cases.

Unfortunately, some of these specific interactions could be quite important, which is a major problem in the utilization of force fields for the issues addressed by structural biology. One of the reasons for the problematic behavior of the force fields seems to be the repulsive term in the force-field potential form.

Our initial intention was to provide a complete interaction energy matrix for amino acid side-chains and compare it to some extent with the previously published data by Miyazawa and Jernigan and others.[4,7,11,33,34] This comparison is not possible based solely on the results of our calculations for cluster representatives. We have found that the cluster representatives are not statistically significant for the whole ensemble of interactions. And because we limited our analysis to gas-phase interaction energy as the first approximation, we could only attempt to adjust a significance of the representative values by a calculation of the interaction energy distribution for the complete side-chain interactions of tryptophan.

This comparison, i.e., the interaction energies for the representative geometries and the overall distribution of the interaction energies, showed the significance of cluster representative geometries in the context of the protein and investigated the importance of such interactions. Our results

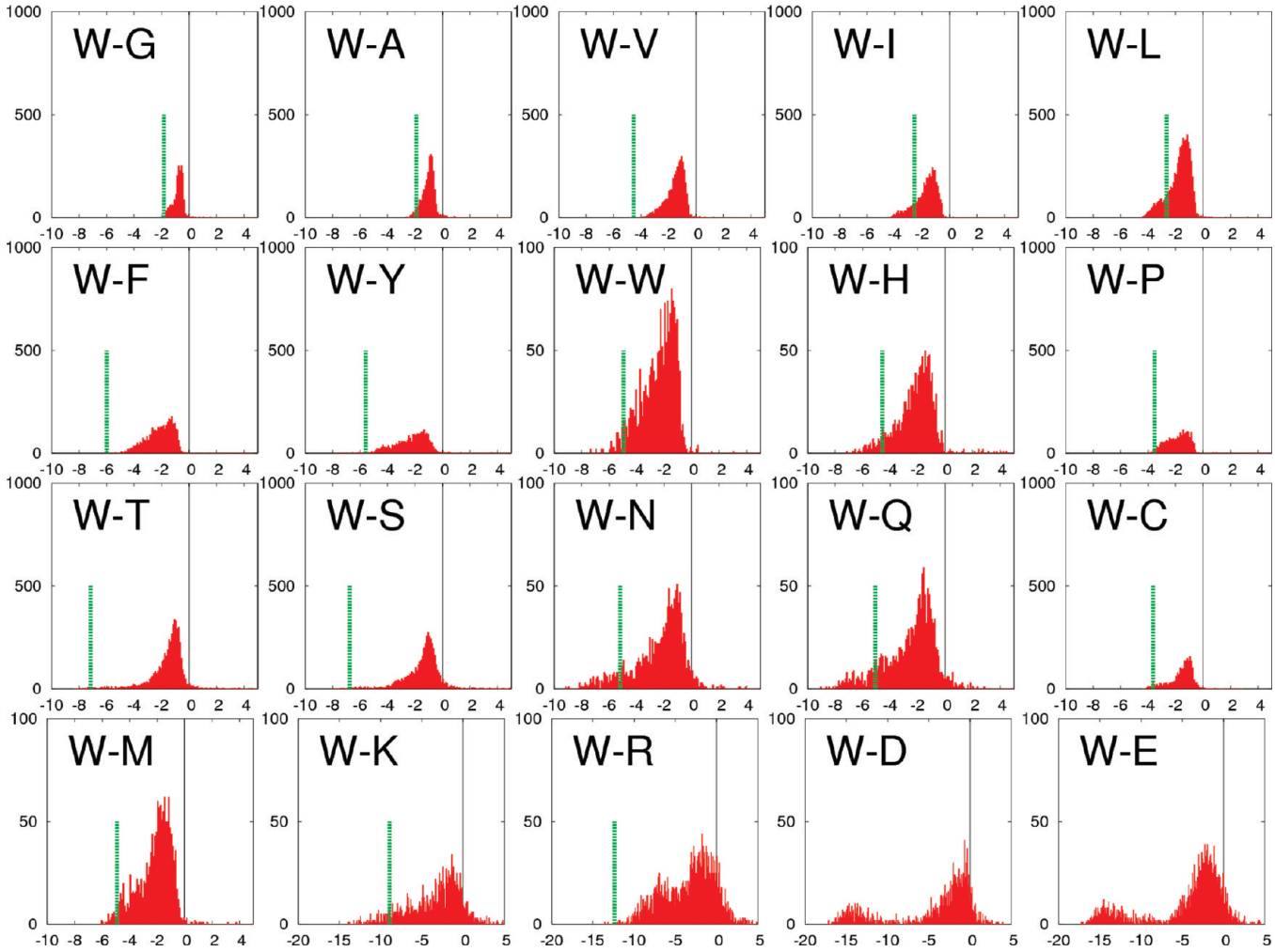

Figure 7. Histograms of all interaction energies of the side chain/side chain interactions of tryptophan with all of the other residues calculated with the OPLS-AA force field in the gas phase. Energies on the *x* axis are in kcal/mol.

Table 4. Comparison between the Interaction Energies for the Cluster Representatives and the Energetically Most Populated Pairs for All of the Pairs of the Trp Residue

system	WG	WA	WV	WI	WL	WF	WY	WW	WH	WP
IE cluster representative	−1.49	−2.08	−4.01	−2.33	−2.27	−4.72	−4.63	−4.26	−4.37	−2.82
most populated IE	−0.55	−0.85	−1.15	−1.15	−1.35	−1.35	−1.45	−1.35	−2.25	−1.45

system	WT	WS	WN	WQ	WC	WM	WK	WR	WD	WE
IE cluster representative	−6.02	−5.23	−4.78	−5.73	−3.20	−5.39	−5.00	−10.78	−14.86	−17.55
most populated IE	−1.05	−1.15	−1.25	−1.25	−0.75	−1.35	−2.15	−2.35	−1.15	−0.35

led to the conclusion that the optimum-energy side-chain interactions are not the most abundant ones in proteins. They are strong enough to be geometrically as well as energetically distinguishable from the mostly random (and mostly attractive) interactions of the majority of side-chain/side-chain pairs. It is therefore plausible to suggest that the interactions represented by cluster representatives are of crucial importance for protein stability or protein function because of their selectivity and strength.

The distributions of the interaction energies also suggest that the approximations lying behind the phenomenological potentials might simply be wrong, as the distributions are not Boltzmann-like. Therefore, the simple calculation of the free energies from the detected contacts is not easily connected to the real energies, as has already been indicated by Thomas and Dill.[13]

Conclusions

We have calculated the matrix of interaction energies by means of the RI-DFT-D method as a benchmark and compared it to the same matrix calculated by the parm03 and OPLS-AA/L force fields in the gas phase while utilizing a simple model of different environments. We have further calculated the distributions of the interaction energies for several pairs to discover the meaning of such interactions.

• All of the interaction energies in the gas phase are attractive with the exception of the ones for pairs with the same charges.

• Force fields generally yield good overall interaction energies for the set, but they can have problems in calculations of specific representative interactions.

• Interaction energies are generally lowered by solvent dielectric screening — the lowest difference between the gas phase and environment goes from aliphatics to aromatics and polars, and the biggest difference can be detected for charged residues.

• The histograms of the interaction energies showed that distributions of interaction energies are neither normal nor Boltzmann-like.

• Geometrically chosen cluster representatives are not representatives for the entire side-chain/side-chain interaction distribution. Most probably, they are representatives of the strongest interactions in a protein, often being functionally or structurally important.

Acknowledgment. This work was supported by Grant No. P208/10/0725 from Grant Agency of the Czech Republic and by grant No. LC512 from the Ministry of Education, Youth and Sports (MSMT) of the Czech Republic. It was also a part of the research projects No. Z40550506 and MSM6198959216. It was also part of Institutional Research Concept No. AV0Z505200701 of the Academy of Sciences of the Czech Republic.

Supporting Information Available: Matrices of interactions calculated in OPLS-AA, parm03 force fields and by the PM6-DH method together with Myiazawa-Jerningan contact energies. This material is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Dyson, H. J.; Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (3), 197–208.

(2) Anfinsen, C. B. Principles That Govern Folding of Protein Chains. *Science* **1973**, *181* (4096), 223–230.

(3) Miyazawa, S.; Jernigan, R. L. A New Substitution Matrix for Protein-Sequence Searches Based on Contact Frequencies in Protein Structures. *Protein Eng.* **1993**, *6* (3), 267–278.

(4) Miyazawa, S.; Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **1996**, *256* (3), 623–644.

(5) Miyazawa, S.; Jernigan, R. L. An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins: Struct., Funct., Genet.* **1999**, *36* (3), 357–369.

(6) Bahar, I.; Kaplan, M.; Jernigan, R. L. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. *Proteins: Struct., Funct., Genet.* **1997**, *29* (3), 292–308.

(7) Betancourt, M. R. Knowledge-based potential for the polypeptide backbone. *J. Phys. Chem. B* **2008**, *112* (16), 5058–5069.

(8) Lu, M.; Dousis, A. D.; Ma, J. OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing. *J. Mol. Biol.* **2008**, *376* (1), 288–301 .

(9) Miyazawa, S.; Jernigan, R. L. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins. *J. Chem. Phys.* **2005**, *122* (2), 4012–4030.

(10) Buchete, N. V.; Straub, J. E.; Thirumalai, D. Orientation-dependent coarse-grained potentials derived by statistical analysis of molecular structural databases. *Polymer* **2004**, *45* (2), 597–608.

(11) Sippl, M. J. Knowledge-Based Potentials for Proteins. *Curr. Opin. Struct. Biol.* **1995**, *5* (2), 229–235.

(12) Laskowski, R. A. http://www.ebi.ac.uk/thornton-srv/databases/sidechains (accessed January 15, 2009).

(13) Thomas, P. D.; Dill, K. A. Statistical potentials extracted from protein structures: How accurate are they. *J. Mol. Biol.* **1996**, *257* (2), 457–469.

(14) Li, D. W.; Bruschweiler, R. A. Dictionary for Protein Side-Chain Entropies from NMR Order Parameters. *J. Am. Chem. Soc.* **2009**, *131* (21), 7226–7232.

(15) 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* (18), 6946–6951.

(16) Morozov, A. V.; Misura, K. M. S.; Tsemekhman, K.; Baker, D. Comparison of quantum mechanics and molecular mechanics dimerization energy landscapes for pairs of ring-containing amino acids in proteins. *J. Phys. Chem. B* **2004**, *108* (24), 8489–8496.

(17) Singh, J.; Thornton, J. M. *Atlas of Protein Side-Chain Interactions*; IRL Press: Oxford, U. K., 1992.

(18) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrasek, J. Representative Amino Acid Side Chain Interactions in Proteins. A Comparison of Highly Accurate Correlated ab Initio Quantum Chemical and Empirical Potential Procedures. *J. Chem. Theory Comput.* **2009**, *5* (4), 982–992.

(19) van der, S. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.

(20) Biedermannova, L.; Riley, K. E.; Berka, K.; Hobza, P.; Vondrasek, J. Another role of proline: stabilization interactions in proteins and protein complexes concerning proline and tryptophane. *Phys. Chem. Chem. Phys.* **2008**, *10* (42), 6350–6359.

(21) Kumar, A.; Elstner, M.; Suhai, S. SCC-DFTB-D study of intercalating carcinogens: Benzo(a)pyrene and its metabolites complexed with the G-C base pair. *Int. J. Quantum Chem.* **2003**, *95* (1), 44–59.

(22) Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a sparse matrix-based implementation of the DFTB method. *J. Phys. Chem. A* **2007**, *111* (26), 5678–5684.

(23) Cerny, J.; Jurecka, P.; Hobza, P.; Valdes, H. Resolution of identity density functional theory augmented with an empirical dispersion term (RI-DFT-D): a promising tool for studying isolated small peptides. *J. Phys. Chem. A* **2007**, *111* (6), 1146–1154.

(24) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. Electronic-Structure Calculations on Workstation Computers - the Program System Turbomole. *Chem. Phys. Lett.* **1989**, *162* (3), 165–169.

(25) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105* (28), 6474–6487.

(26) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24* (16), 1999–2012.

(27) Hess, B.; Kutzner, C.; van der, S. D. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447.

(28) Schafer, A.; Klamt, A.; Sattel, D.; Lohrenz, J. C. W.; Eckert, F. COSMO Implementation in TURBOMOLE: Extension of an efficient quantum chemical code towards liquid systems. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2187−2193.

(29) Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z. K.; Gilliland, G.; Weissig, H.; Westbrook, J. The Protein Data Bank and the challenge of structural genomics. *Nat. Struct. Biol.* **2000**, *7*, 957–959.

(30) Kaur, D.; Sharma, P.; Bharatam, P. V. A comparative study on the nature and strength of O-O, S-S, and Se-Se bond. *THEOCHEM* **2007**, *810* (1−3), 31–37.

(31) Vondrasek, J.; Mason, P. E.; Heyda, J.; Collins, K. D.; Jungwirth, P. The Molecular Origin of Like-Charge Arginine-Arginine Pairing in Water. *J. Phys. Chem. B* **2009**, *113* (27), 9041–9045.

(32) Berka, K.; Hobza, P.; Vondrasek, J. Analysis of Energy Stabilization inside the Hydrophobic Core of Rubredoxin. *Chemphyschem* **2009**, *10* (3), 543–548.

(33) Betancourt, M. R. Empirical model of residue contact probabilities for polypeptides. *J. Chem. Phys.* **2010**, *132* (8), 8613–8621.

(34) Miyazawa, S.; Jernigan, R. L. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins: Struct., Funct., Genet.* **1999**, *34* (1), 49–68.

(35) Merritt, E. A.; Bacon, D. J. Raster3D: Photorealistic molecular graphics. *Macromol. Crystallogr., Part B* **1997**, *277*, 505–524.

CT100007Y