JCTC Journal of Chemical Theory and Computation

Representative Amino Acid Side Chain Interactions in Proteins. A Comparison of Highly Accurate Correlated *ab Initio* Quantum Chemical and Empirical Potential Procedures

Karel Berka,[†,‡] Roman Laskowski,[§] Kevin E. Riley,[||] Pavel Hobza,[†] and Jiří Vondrášek*[,†]

Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic and Center for Complex Molecular Systems and Biomolecules, Flemingovo náměstí 2, Prague 6, 166 10 Czech Republic, Department of Physical and Macromolecular Chemistry, Faculty of Natural Sciences, Charles University in Prague, Hlavova 8, Prague 2, 128 43 Czech Republic, EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K., and Department of Chemistry, P.O. Box 23346, University of Puerto Rico, Rio Piedras, Puerto Rico 00931

Received November 25, 2008

Abstract: Interactions between amino acid side chains play a crucial role both within a folded protein and between the interacting protein molecules. Here we have selected a representative set of 24 of the 400 (20 × 20) possible interacting side chain pairs based on data from Atlas of Protein Side-Chain Interactions. For each pair, we obtained its most favorable interaction geometry from the structural data and computed the interaction energy in the gas phase using several different, commonly used, *ab initio* and force field methods, namely Møller−Plesset perturbation theory (MP2), density functional theory combined with symmetry-adapted perturbation theory (DFT-SAPT), density functional theory empirically augmented with an empirical dispersion term (DFT-D), and empirical potentials using the OPLS-AA/L and Amber03 force fields. All the methods were compared against a reference method taken to be the CCSD(T) level of theory extrapolated to the complete basis set limit. We found a high degree of agreement between the different methods, even though the range of binding energies obtained was extremely large. The most computationally intensive methods yielded the best results. Among the less computationally time-consuming methods, the DFT-D method as well as parm03 force field provided consistently good results when compared to the reference values. We also tested how representative the chosen geometries of the side chains were and investigated the effect on the binding energies of the dielectric constant of the surrounding medium.

Introduction

The building blocks of proteins are the twenty naturally occurring L-amino acids, which are distinguished by their different side chain structures and chemical compositions. The sequence of amino acids defining a given protein determines both its overall 3D structure and, at the local level, the interactions it makes with other molecules when carrying out its biological function. A change to a single amino acid (e.g., due to a point mutation) can either have relatively little effect[1,2] or seriously harm the organism.[3]

The packing of a protein 3D structure into its final, fully folded form is governed by the noncovalent interactions of its side chains. The hydrophobic interactions in the protein's

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

[†] Academy of Sciences of the Czech Republic and Center for Complex Molecular Systems and Biomolecules.

[‡] Charles University in Prague.

[§] European Bioinformatics Institute.

[||] University of Puerto Rico.

10.1021/ct800508v CCC: $40.75
Published on Web 02/25/2009

core[4,5] are of particular importance, as they, along with hydrogen bonds,[6] salt bridges, and disulfide bonds, determine the overall stability and architecture of the protein.

The packing of side chains in a protein is known to be very tight,[7] and two opposing models have been proposed to describe it.[8] The first, the "nuts and bolts in a jar" model, suggests that the main driving force of protein folding is the hydrophobic effect of the surrounding water molecules; consequently, the side chains are crammed together as the protein folds, and their packing is essentially random and the result of entropic factors. In the second, the "jigsaw puzzle" model, energetic (enthalpic) contributions play a more significant role in determining how the side chains come together; consequently, the packing is nonrandom. A large number of studies have analyzed the available 3D structural data in the Protein Data Bank (PDB) and have shown that side chains do indeed have preferred interaction geometries; their packing is not entirely random.[9−12]

Thus, for an accurate energetic picture of protein structure, it is necessary to describe the interactions between amino acid side chains properly. The potential energy landscapes of proteins are most often approximated as a sum of electrostatic charge−charge and Lennard-Jones contributions including exchange-repulsion and dispersion terms. Other energy contributions, such as charge transfer, do not appear to be significant in this schema. There is a reasonable degree of correlation between the molecular mechanics energy landscapes on the one hand and the distributions of amino acid pairs and their geometries observed in protein structures on the other, which suggests that the intrinsic pairwise interaction energies do contribute to packing of side chains in proteins rather than being overwhelmed by the numerous interactions with other atoms within the protein and with the solvent.[13]

The interaction energy functions can generally be divided into two types: an atomic level energy function, whose form is based unambiguously on physical principles, and an energy function based on statistics at the amino acid level, such as the Miyazawa-Jernigan potential.[14,15] The most ideal way to treat these systems is to apply high-level correlated *ab initio* quantum mechanical methods to compute the noncovalent interactions within a protein, incorporating all possible energy contributions. However, these methods are currently limited to small- or medium-sized molecular models. Significant progress has recently been made in overcoming the well-known deficiency of the density functional theory (DFT) approaches in describing dispersion interactions.[16,17] Nevertheless, it is still not possible to make calculations on entire proteins at this level in a reasonable amount of time and at a high (or even medium) level of accuracy. Therefore, a promising approach is to develop improved potential functions for modeling macromolecular interactions involves combining protein structural analysis and quantum mechanical calculations on small molecule models of amino acids.

We have previously obtained a reasonable level of accuracy for determining the interaction energies between amino acid side chains using high-level *ab initio* methods.[18,19] Such characteristics are of utmost importance for the analysis and design of protein structures and can shed light on the enthalpic background of protein stabilization and, to some extent, on the folding process. It must be mentioned here that lower-level theoretical calculations, such as the DFT methods, can incorrectly predict the energetics (and geometries) of these structures. Another fundamental question currently driving research in protein packing is how a single amino acid change in a protein sequence affects the 3D structure.[20] Determining a practical correlation between the two would help move the field of structure prediction and design forward.

Table 1. Computational Methods Used

method	description
CCSD(T)|CBS	reference method—CCSD(T) level of theory, extrapolated to the complete basis set limit (CBS)[27]
RI-MP2|aDZ	MP2 with the aug-cc-pVDZ basis set and resolution of identity approximation
RI-MP2|aTZ	MP2 with the aug-cc-pVTZ basis set and resolution of identity approximation
SCSMI-MP2|TZ	spin-component scaling MP2 perturbation theory, parametrized for molecular interactions (MI) with the cc-pVTZ basis set[32]
DFT-SAPT|aDZ	DFT-symmetry-adapted intermolecular perturbation theory with density fitting with the aug-cc-pVDZ basis set
DFT|TZVP	DFT with the TPSS functional and TZVP basis
DFT-D|TZVP	DFT with the TPSS functional and the TZVP basis set augmented with an empirical dispersion[34]
RI-DFT-D|TZVP	DFT with the TPSS functional and the TZVP basis set augmented with an empirical dispersion and resolution of identity approximation[36]
OPLS	OPLS-AA/L force field[38]
parm03	Duan et al. Amber parm03 force field[39]

Here we explore the intramolecular interaction energies for selected pairs of amino acid side chains. For each pair, we use an empirically determined representative geometry and compare the energies computed using several different *ab initio* and force field methods with the reference method. The interaction geometries are obtained from the Web version of the Atlas of Protein Side-Chain Interactions.[21] The energy calculations are performed using ten different approaches, summarized in Table 1. The estimated CCSD(T)/CBS method is taken as the reference method, whose energies are assumed to be the closest to the "true" energy values and against which all the other methods are compared.

Materials and Methods

1. Representative Set of Amino Acid Side Chain Pairs. 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[21] The Web atlas is based on a printed atlas published in 1992 by Singh and Thornton[22] and analyzes the interaction geometries of all 20 × 20 amino acid side chain pairs as found in experimentally determined 3D structural models of proteins. For each side chain pair, the atlas shows how one side chain is distributed with respect to the other in 3D. The preferred interaction geometries are revealed by clusters in the distributions. The atlas lists the clusters by size and selects a representative side chain pairing for each one.

Table 2. Statistical Data for Selected Pairs Taken from the Updated Version of the Side Chain Atlas[a]

A1	A2	code	$N_{clustered\ contact}$ of $N_{detected\ contacts}$	$p_{A1}p_{A2}$	p_{AA}	$p_{AA}/(p_{A1}p_{A2})$
Leu	Leu	LL	143 of 47638	0.850	3.032	3.57
Val	Leu	VL	107 of 27218	0.660	1.733	2.62
Ile	Leu	IL	82 of 26652	0.518	1.697	3.28
Val	Val	VV	192 of 19723	0.513	1.255	2.45
Ile	Ile	II	112 of 18624	0.315	1.186	3.76
Ala	Leu	AL	159 of 15282	0.771	0.973	1.26
Leu	Tyr	LY	74 of 12030	0.326	0.766	2.35
Phe	Phe	FF	42 of 11127	0.165	0.708	4.30
Leu	Thr	LT	172 of 8233	0.510	0.524	1.03
Lys[b]	Glu[b]	KE	187 of 7755	0.389	0.494	1.27
Arg[b]	Asp[b]	RD	493 of 7391	0.295	0.470	1.60
Leu	Trp	LW	45 of 6487	0.136	0.413	3.04
Leu	Gly	LG	165 of 6368	0.685	0.405	0.59
Tyr	Tyr	YY	51 of 5179	0.125	0.330	2.64
Thr	Thr	TT	238 of 4262	0.307	0.271	0.89
Tyr	Pro	YP	61 of 4149	0.165	0.264	1.60
Thr	Ser	TS	149 of 3132	0.328	0.199	0.61
Asp[b]	His	DH	75 of 2383	0.134	0.152	1.13
Gln	Asn	QN	106 of 2217	0.165	0.141	0.86
Met	Met	MM	19 of 1973	0.034	0.126	3.73
Met	Cys	MC	9 of 641	0.025	0.041	1.65

[a] The part of the total number of detected contacts between selected residue pairs which constitutes the most populated clustered motif, whose representative was used for further energetic analysis. The values of $p_{A1}p_{A2}$ are probabilities in percents that these two residues would be in protein sequences based on the detected numbers of residues within the side chain atlas data set. The value of p_{AA} is the frequency in percents of the observed contact between side chains. Their proportion is thus a measure of preferences between side chains. [b] Charged residues were treated also as neutral.

The atlas is derived using a set of nonhomologous protein chains selected from the structures in the Protein Data Bank (PDB). No two chains have a mutual sequence identity greater than 20%, and the chains are only taken from structures solved by X-ray crystallography to a resolution of 2.0 Å or better. The data in the printed version of the atlas were derived from 62 protein structures, whereas the Web version uses 533. For the current study, we have used the atlas as updated in October 2006, applying 2548 structures.[21]

Interacting side chains are considered to be those having a center-to-center distance between their closest two atoms (excluding backbone atoms) of less than the sum of their van der Waals radii, plus 1 Å to allow for coordinate error. The two amino acids must be at least 4 residues apart in the protein's sequence.

The cluster representatives for a given distribution are determined by considering each side chain in turn. The root-mean-square distance (rmsd) to all other side chains in the distribution is computed using the three atoms that define the side chain's frame of reference. Any side chain with an rmsd of less than 1.5 Å from the selected side chain is considered a "neighbor". The side chain with the largest number of neighbors is taken to be the cluster representative of the largest cluster. This side chain and all its neighbors are then removed from the distribution, and the calculation is repeated to obtain the cluster representative of the second largest cluster, etc.

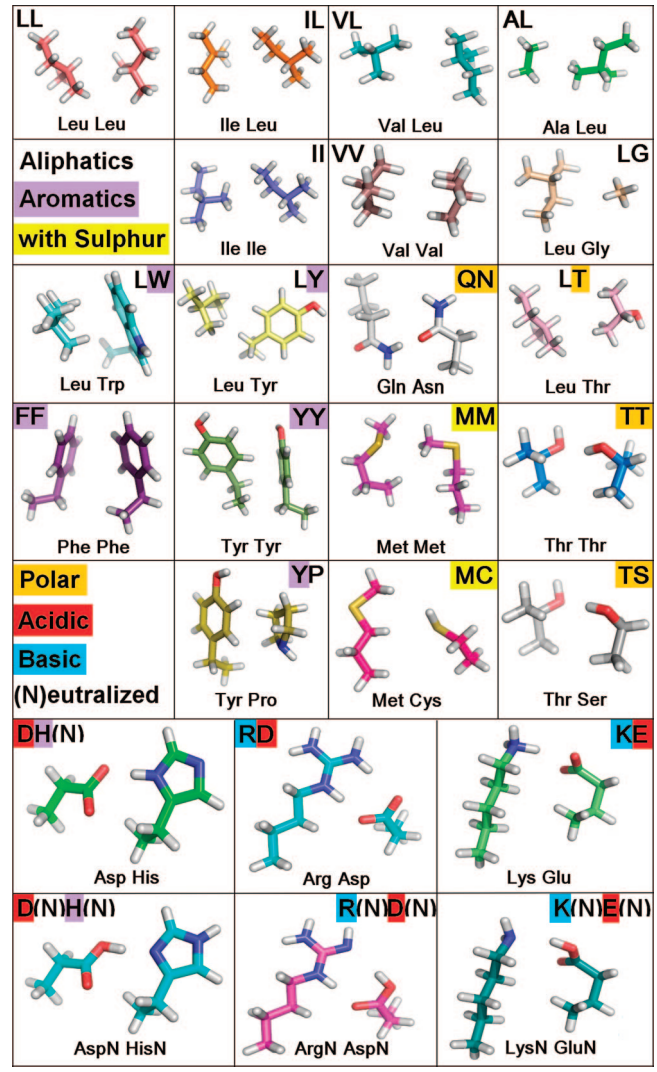


Figure 1. Set geometries of the amino acid residues truncated at the Cα atom and optimized with DFT|TPSS|TZVP, from which the geometries with Cβ fragmentation were derived by the deletion of the Cα methyl group and the insertion of a hydrogen atom in the former methyl direction.

For this study, 24 of the 400 side chain pairs were chosen to be representative of different types of side chain interactions: hydrophobic−hydrophobic, polar−polar, charged−charged, and intermingled interactions (see Table 2 and Figure 1). The side chain pair corresponding to the top cluster representative in each of these 24 distributions was understood to represent that distribution and its geometry used for the various energy calculations described below.

2. System Preparation. As the side chain atlas contains only the heavy atom positions for each cluster representative, the missing hydrogens were added using the Pymol 0.99rc6 package.[23] Two types of model subsystem were defined based on the hydrogenated amino acids: the first contained only the amino acid side chain starting from the C_β atom (C_β model), while the second consisted of the amino acid side chain plus the backbone C_α atom (C_α model). Hydrogen atoms were added at the point of cutting (i.e., either the C_α or C_β atom) in order to complete the valence shell. Proline was modeled as propane or tetrahydropyrrole in the C_β and C_α models, respectively. The positions of hydrogens were

then optimized for each pair at the DFT|TZVP[24] level using the relax procedure in the Turbomole package.[25] The coordinates of the heavy atoms were kept fixed during the entire computational procedure without initial optimization.

3. Reference Interaction Energy in Vacuo. The reference pair stabilization energies were determined in vacuo at the estimated CCSD(T) level of theory and extrapolated to the complete basis set limit (CBS). The CCSD(T)|CBS interaction energy was approximated as[26,27]

$$\Delta E_{CCSD(T)|CBS} = \Delta E_{MP2|CBS} + (\Delta E_{CCSD(T)} - \Delta E_{MP2})_{|small\ basis\ set} \quad (1)$$

The former term, $\Delta E_{MP2|CBS}$, was determined using the Helgaker[28] extrapolation scheme.

The Hartree−Fock and second-order Møller−Plesset (MP2) correlation energies necessary for the extrapolation to the complete basis set limit were determined using systematically improved basis sets; here we have used the aug-cc-pVXZ (X=D, T) basis sets in Turbomole (abbreviated as aXZ). The CCSD(T) term was calculated with a smaller basis set, 6−31G*(0.25, 0.15), in the Molpro 2006 package.[29] The use of this smaller basis set is justified because the accuracy of the quantity (called the CCSD(T) correction term) defined as the difference between the MP2 and CCSD(T) interaction energies (unlike MP2 and CCSD(T) interaction energies themselves) is much less dependent on the size of the basis set and the 6−31G*(0.25, 0.15) basis set has been shown to yield satisfactory values of this difference.[26] The CCSD(T) correction term evaluated for the uracil dimer with this basis set agreed very well[30] with values calculated with the aug-cc-pVDZ and aug-cc-pVTZ basis sets. All the interaction energies were corrected for the basis set superposition error using the counterpoise scheme of Boys and Bernardi.[31] MP2 electronic energies were computed with the resolution of the identity approach (RI-MP2), which has been shown to introduce negligible errors. The frozen core approximation was used systematically throughout the story.

4. Spin-Component Scaling Perturbation Theory SCSMI-MP2. In an attempt to compensate for the overestimation of dispersion contributions generally seen with the MP2 method, the spin-component-scaled MP2 method as parametrized for molecular interactions (SCSMI-MP2) was used.[32] In this method, the parallel and antiparallel spin-contributions of the MP2 correlation energy are empirically scaled with two scaling factors, $c_{PS} = 1.75$ and $c_{AS} = 0.17$. These parameters differ from those of the original SCS-MP2 method described by Grimme:[33]

$$E_{SCSMI\text{-}MP2} = c_{PS}E_{PS} + c_{AS}E_{AS} \quad (2)$$

The parameters were fitted against molecular interaction energies computed at the CCSD(T)|CBS level for the S22 set.[27] The SCSMI-MP2 energies were calculated using the Molpro 2006 package[29] along with the cc-pVTZ basis set (abbreviated as TZ). We prefer to use the SCSMI-MP2 over the original Grimme SCS-MP2 procedure since the former procedure describes not only the stacked interactions but also H-bonding accurately. The original SCS-MP2 procedure works well for stacking interactions, but H-bonded stabilization energies are underestimated.

5. DFT-Based Interaction Energy Augmented with an Empirical Dispersion Term. The energies were also computed using the DFT-D|TZVP method,[34] in which the DFT energies calculated with a TPSS functional in a TZVP basis set are augmented by an empirical dispersion term parametrized against the CCSD(T)|CBS energies in the S22 set.[27] The DFT energies were calculated with the Gaussian03 package.[35]

A faster DFT algorithm using the resolution of the identity approximation and empirically augmented dispersion was also utilized (RI-DFT-D).[36] The RI-DFT-D energies were calculated employing the Turbomole package.[25] This technique provides excellent interaction energies not only for the H-bonded, dispersion-bound and mixed complexes included in the S22 set[34] but also for these noncovalent complexes in general. A strong point of the method is its relatively low computational cost, making it an ideal candidate for calculations on large complexes with hundreds of atoms or even for on-the-fly molecular dynamics simulations.

6. Empirical Force-Field Interaction Energy. All the molecular mechanical force-field calculations were performed using the Gromacs 3.3 package[37] with the built-in OPLS-AA/L (OPLS)[38] and ported parm03[39] force fields. The porting of parm03 was performed according to the method of Sorin and Pande.[40] The amino acid topology and partial charges were changed as follows:

OPLS - The terminal Cα or Cβ methyl group was assigned the same atomic types and partial charges as the other methyl groups.

parm03 - All the original atoms have their original partial charges and the newly added hydrogens on C_α or C_β were assigned to provide the integral charge on the entire residue.

7. SAPT Decomposition of the Interaction Energy in Vacuo. In the DFT-SAPT method,[41] the interaction energy is given as the sum of first- and second-order energies ($E^{(1)}$, $E^{(2)}$) as well as of the δ(HF) term. The first-order energy term contains the electrostatic ($E^{(1)}_{el}$) and exchange-repulsion ($E^{(1)}_{ex}$) contributions, the second-order term includes the induction, exchange-induction, dispersion, and exchange-dispersion contributions. The charge-transfer energy is considered to be part of the induction energy. The δ(HF) energy estimates the contributions from the higher-order energy terms using the Hartree−Fock approximation. In this study, we used the PBE0AC exchange-correlation potential[42] along with the aug-cc-pVDZ basis set (and its corresponding density-fitting basis sets). The PBE0AC functional has been shown to yield accurate first-order as well as induction and dispersion values. The aug-cc-pVDZ set is large enough to provide a reliable estimate of the electrostatic, induction, and exchange components. The dispersion component is underestimated by about 10−20% in this basis set (see ref 41) but should serve well enough for the purpose of comparison. Here we have implemented a gradient-controlled shift procedure required for the asymptotic correction of the exchange-correlation potential, which needs a computed difference (shift) between the vertical ionization potential

and HOMO energy calculated using the same DFT method as used for the DFT-SAPT computation.[42] The DFT-SAPT interaction energy calculated with the aug-cc-pVDZ basis set provides highly accurate interaction energies for DNA base pairs, and if the dispersion energy is augmented by 10−15%, the resulting interaction energies agree fairly well with the CCSD(T)|CBS values.[43] The DFT-SAPT thus provides reasonably accurate stabilization energies as well as their components for various types of noncovalent complexes.

The DFT-SAPT calculation for each pair of selected residues was performed with the density fitting using the Molpro 2006 package.[29] In order to express the obtained DFT-SAPT results in terms of commonly understood physical quantities, the exchange-induction and exchange-dispersion terms were added to the induction and dispersion terms, respectively. The ionization potentials were calculated at the PBE0|TZVP level, while the HOMO values were taken from the aug-cc-pVDZ calculation with the Gaussian03 package.[35]

Results and Discussion

1. Representative Side Chain Pairs. While the most populated types of contacts between the amino acids are mainly those containing leucine, we selected a set of 24 amino acid pairs to best represent the full spectrum of interaction types that occur within proteins, making sure that our set included at least one of each of the 20 natural amino acids (see Figure 1 and Table 2). The most common types of interaction are generally those involving aliphatic-aliphatic contacts, reflecting their tendency to be localized in a protein's hydrophobic core. Thus the largest group within our set comprised aliphatic-aliphatic side chain interactions: LL, IL, VL, AL, II, VV, and LG. The next largest groups consisted of aliphatic-aromatic interactions - LW, LY, and YP - and nonpolar−polar interactions - LT, MC, and MM. An interesting set of contacts are those between aromatic side chains: YY and FF. Polar−polar contacts are represented by the interactions between threonine and serine: TS and TT. A very special type of interaction is that between aromatic and charged residues, DH(N), where the histidine is taken to be neutral, or D(N)H(N), where both residues are taken to be neutral. A typical salt-bridge conformation between charged residues is represented by RD and KE. The final group contains unphysical salt-bridge interactions, where both charged residues are neutralized: R(N)D(N) and K(N)E(N). This situation never happens in solution; however, these complexes would be stable in the gas phase and provide a test case for neutralized charged residues found in force fields.

It is interesting to note how the numbers of expected and observed contacts differ (see Table 2). The expected contact value, $p_{A1}p_{A2}$, is defined as the product of the observed frequencies of all the amino acid side chains detected in the atlas data set. The observed contact value, p_{AA}, is defined as the frequency of the given contact pair divided by the total number of all observed contacts. The observed values differ slightly from the expected values. The ratio between the observed and expected contacts varies between 4.30 for Phe-Phe to 0.59 for Leu-Gly.

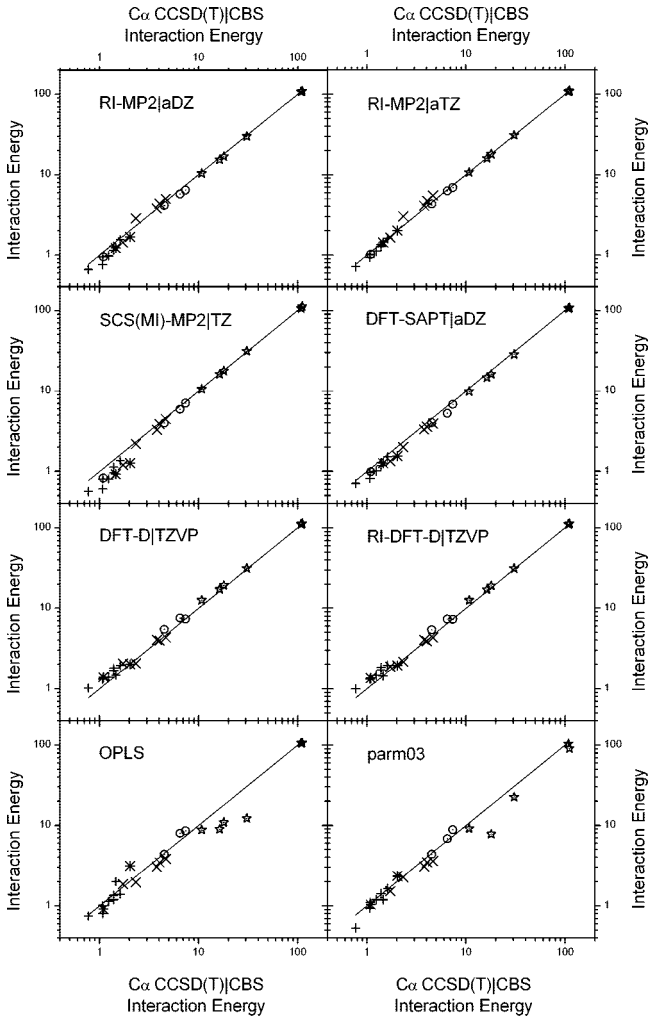
2. Comparison of Binding Energies. To assess the performance of the various energy calculation methods used in this study properly for a range of binding energies spanning 2 orders of magnitude - from the most stable pair, RD, to the least stable, LG - we employed two types of statistics: relative and absolute deviation (see the legend of Table 3). The calculated CCSD(T)|CBS binding energies (see Methods) are taken to be the "true" binding energies of these side chain-side chain interactions as they are, to our knowledge, the most accurate binding energies that can be systematically obtained for all of the complexes described in this work. We should add here that the CCSD(T)/CBS method is the only *ab initio* method that provides accurate interaction energies for different types of noncovalent complexes (keeping in mind that MP2 tends to greatly overbind dispersion bound structures). All the other wave function theory (WFT) as well as DFT procedures which are used in the realm of noncovalent complexes are parametrized, i.e. contain one or more parameters, making it possible to fit the obtained values on the benchmark data. Consequently, using these methods we cannot be certain whether the interaction energies for different types of noncovalent complexes are all reliable.

The CCSD(T)|CBS interaction energies vary over a wide range of values (see column 2 of Table 3). The most stable pairs in the gas phase are the salt bridges between the charged residues RD and KE, having interaction energies of the order of 100 kcal/mol. The second most stable are the DH(N) interactions representing a charged residue interacting with an uncharged residue. Neutralizing the charged residues dramatically lowers the interaction energy, as exemplified by the next group of interactions - D(N)H(N), R(N)D(N) and K(N)E(N) - which have energies between 10 and 18 kcal/mol. Interestingly, neutralization affects charged systems differently: salt bridges R(N)D(N) and K(N)E(N) have only 12% of their former interaction energy, whereas the Asp-His pair's interaction energy decreases to only 58%. The next most stable interaction groups consist of polar contacts, QN, TT, and TS, whose interaction energies are around 5 kcal/mol, and aromatic systems, YY, LW, YP, FF, and LY, whose interaction energies cluster around 4 kcal/mol. The energies of nonpolar pairs containing methionine - MM and MC - are approximately 1.5 kcal/mol, while the final group contains the aliphatic nonpolar pairs (LL, VV, IL, II, LT, VL, VL, AL, and LG), whose binding energies seem to depend mainly on the contact surface between them, favoring more lengthy and more "treelike" structures such as those of leucine or valine.

3. Comparison of Computational Methods. Columns 3−11 of Table 3 show the interaction energies obtained from the nine other computational methods tested here. As can be seen, the methods tend to yield similar absolute values and exhibit a high degree of correlation from the highest to the lowest energy values. This correlation can be more clearly seen in Chart 1. However, there are interesting discrepancies between the methods, which will be discussed next.

For H-bonded complexes, the use of a larger basis set with the RI-MP2 method yields higher-quality stabilization energies while overestimating stabilization results for stacking

Chart 1. Correlations between the Computational Methods[a]



[a] The energies were divided into 5 groups: + - aliphatic-residue-only systems; × - systems with at least one aromatic residue; ✳ - systems containing sulfur; ⊙ - systems with at least one polar residue; and * - systems with at least one charged residue. The systems falling into several groups were arbitrarily added to the more polar group according to the participating amino acids.

complexes. This overestimation is more pronounced when aromatic systems participate in an interaction. Investigating different AA pairs containing at least one aromatic residue (YY, LW, YP, and FF), we found that they exhibit this systematic overbinding and have larger maximal relative errors when larger basis sets are employed. Such behavior when the MP2 method is employed is well-known, with various examples being possible to find in the S22 data set.[19]

The spin component scaled perturbation theory, SCSMI-MP2, performs well on aromatic pair systems but significantly underestimates stabilization energies for loosely bound pairs (interactions whose binding energies are below a threshold of approximately 2 kcal/mol, which is especially true for AL, MC, and MM pairs). This is probably a result of the fitting procedure for this method[32] being carried out in order to obtain minimal errors in the absolute values, and not in terms of relative numbers (i.e., percentagewise), which led to a situation in which contributions from the systems with small binding energies would be underrepresented. Thus the error of the method is relatively modest (0.60 kcal/mol), but larger relative errors occur when stabilization energies are around 1 kcal/mol.

The DFT-SAPT interaction energies are slightly lower than the MP2 ones. Evidently, the DFT-SAPT method does not suffer from an overestimation of the stabilization energies for pairs containing aromatic systems.

The DFT-D method generally overestimates the stabilization energies of these complexes. For polar or charged contacts, the binding energies can be overestimated by up to 2 kcal/mol (approximately 2% of the total interaction energy). At the other extreme, the overestimation of the interaction strength for loosely bound pairs, such as Leu-Gly, is typically around 0.3 kcal/mol.

The errors seen for results obtained using the SCSMI-MP2 are similar in magnitude but opposite in sign to those produced using the DFT-D method. It can also be said that, in terms of absolute energies, the SCSMI-MP2 produces more accurate results, while the DFT-D method yields better results when errors are measured on a relative (i.e., in terms of percentage) scale. It should also be noted here that, as is well-known, the traditional DFT method fails to describe dispersion interactions correctly. Recently introduced modern density functionals, like e.g. the Zhao and Truhlar's MO6 suite of density functionals, do cover the dispersion energy and provide a very good estimate of stabilization energies and geometries for a large spectrum of noncovalent complexes.[44]

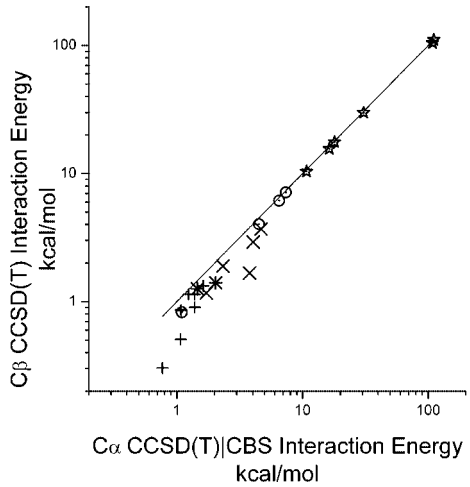
The OPLS-AA/L force field yields results that are slightly unbalanced; several residues are not parametrized for the computation of interaction energies with accuracies similar to those of previous methods. In particular, histidine and methionine exhibit relative errors that are higher than those seen for any other method (DH, DH(N), MM, and MC).

The parm03 force field generally behaves better than the OPLS-AA/L force field. When omitting neutralized charged residues (because of parametrization), the parm03 force field has relative errors comparable to those of the MP2|aDZ method. As the largest errors can be found for the most strongly bound complexes (from RD to QN), the absolute errors of parm03 are significantly higher than those observed for any of the *ab initio* quantum mechanics methods discussed previously.

Surprisingly, both force-field methods were found to perform generally well. Their major weakness concerns the parametrization for the neutralized charged residues (D(N)H(N), R(N)D(N), and K(N)E(N)). It should be noted that the parm03 force field does not even contain a neutralized arginine residue.

To summarize the results, the most accurate method (other than the benchmark CCSD(T)/CBS method) for calculations of interaction energies between amino acid residues in proteins is MP2|aug-cc-pVTZ, which is also the most computationally intensive technique considered here. The less demanding SCSMI-MP2 and DFT-D methods yield similar accuracy with comparable computational expense. The fastest *ab initio* method is RI-DFT-D, which tends to overestimate

Chart 2. Correlation between the Different Fragmentations[a]



[a] The energies were divided into 5 groups: + - aliphatic-residue-only systems; × - systems with at least one aromatic residue; ✷ - systems containing sulfur; ⊙ - systems with at least one polar residue; and * - systems with at least one charged residue. The systems falling into several groups were arbitrarily added only to the more polar group.

interaction energies slightly. The best force-field method is parm03 force field, especially when strongly bound pairs are omitted.

4. Effects of Different Fragmentation (C_α vs C_β Models). The analysis of interaction energies above was based on our C_α model, i.e. where the amino acid side chains were methylated at the C_α position. Are the results markedly different if one uses the C_β model, i.e. residues starting at the C_β atom, instead?

The last column in Table 3 shows the energies obtained from the CCSD(T)|CBS calculations for the C_β model. The average difference between the interaction energies from the two models is approximately 0.3 kcal/mol. Interestingly, this energy loss caused by removing the C_α methyl group is comparable to the interaction energy for the methane dimer, which is around 0.33−0.46 kcal/mol for two methane molecules in close contact.[45] The simple explanation of this fact is that the smaller system - stabilized mainly by dispersion, the smaller interaction energy is. One additional feature of the different fragmentation has to be also taken into account (see Chart 2). In many cases the position of Ca methyl groups improves contact between interacting residues which results in higher stabilization energy than expected based on C_β geometry.

The interaction energy loss varies according to the nature of the interaction. The differences tend to be small for interactions involving charged residues. In the case of RD, for instance, the difference is 0.06 kcal/mol, which is negligible. On the other hand, when at least one aromatic residue, e.g. LW and YY, participates in the interaction the interaction energy, loss due to the demethylation is quite significant. This may be attributable to a difference in the interaction of the system, with the electronic density of the π-system caused by the fragmentation.

Last but not least, an interesting interaction is that between tyrosine and proline (YP). The interaction energy loss here is strongly affected by the opening of the proline ring due to the loss of the main-chain C_α atom. This effect has recently been reported by Biedermannova et al.[46] in interactions between aromatic side chains and proline.

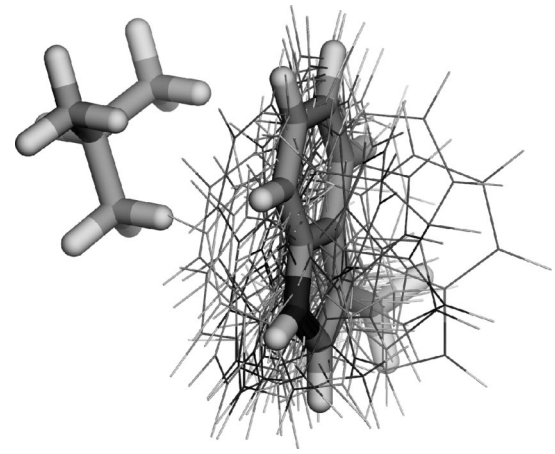


Figure 2. The cluster of LW with a cluster representative.

5. Cluster Representatives. In all the calculations above, the geometry used for each side chain pair is that of the representative of the pair's largest cluster in the Side Chain Atlas. It is natural to ask how valid and meaningful this choice of geometry is. To address this issue we used both the DFT-D and RI-DFT-D methods to compute the interaction energies for all the members of a single cluster and compared these against the energy of that cluster representative. We selected the leucine-tryptophan pair, LW, as it has an intermediate total number of observed contacts (6487) and thus should be more typical than such extreme cases as LL and MC, with 47,638 and 641 contacts, respectively. Moreover, considering that the number of geometries in the cluster is 45 (Figure 2), the number of interaction energies to be computed is sufficiently small for them to be completed in a reasonable time. Finally, to decrease the total calculation time even further, only the C_β fragments were used.

The DFT-D method gave an interaction energy of −2.76 ± 0.550 kcal/mol when averaged over all the Leu-Trp cluster members, which is identical to the energy obtained from the cluster representative using both the DFT-D and RI-DFT-D methods (see Supporting Information Table S1). The RI-DFT-D method yielded a slightly different average cluster value of −2.63 kcal/mol. In this method, however, the median value of the individual energies in the cluster was much closer to the representative's energy than the median value given by DFT-D: −2.75 vs −2.80 kcal/mol, respectively. One can therefore conclude that the geometry of the cluster representative really does approximate to some average energy conformation for the two interacting side chains.

6. Full Optimization of Geometries. To test the relevance of the cluster representative further, we performed a full optimization of their geometry using the C_β model and RI-DFT-D *in vacuo*. In most cases, the geometry changed only negligibly (see Supporting Information Table S2). The differences come mostly for the interactions involving proline, glycine, charged, and polar residues. The largest changes were observed for charged and polar residues,

Table 3. Interaction Energies for Amino Acid Pairs Calculated Using Several Approaches in the Gas Phase[a]

code	CCSD(T) CBS	RI-MP2 aDZ	RI-MP2 aTZ	SCSMI-MP2 TZ	DFT-SAPT aDZ	DFT TZVP	DFT-D TZVP	RI-DFT-D TZVP	OPLS	parm03[b]	CCSD(T) CBS Cβ
RD	−110.80	−109.37	−110.21	−111.71	−107.52	−110.60	−112.93	−112.73	−105.71	−90.37	−110.74
KE	−108.40	−107.36	−107.75	−105.64	−105.78	−108.27	−110.90	−110.86	−106.02	−103.57	−104.67
DH(N)	−30.64	−29.88	−30.91	−31.06	−28.35	−28.83	−31.47	−31.30	−12.20	−22.36	−29.82
D(N)H(N)	−17.97	−16.81	−17.94	−17.68	−16.05	−16.26	−19.29	−19.03	−10.90	−7.80	−17.61
R(N)D(N)	−16.32	−15.29	−15.92	−16.18	−14.68	−14.71	−17.17	−17.01	−8.94		−15.57
K(N)E(N)	−10.76	−10.36	−10.65	−10.50	−9.87	−9.81	−12.60	−12.51	−8.80	−9.11	−10.38
QN	−7.37	−6.41	−6.92	−7.06	−6.83	−5.66	−7.35	−7.31	−8.61	−8.84	−7.14
TT	−6.50	−5.74	−6.28	−5.93	−5.27	−4.81	−7.53	−7.32	−7.96	−6.83	−6.15
YY	−4.66	−4.99	−5.51	−4.49	−3.94	1.35	−4.35	−4.31	−3.84	−3.62	−3.70
TS	−4.50	−4.12	−4.30	−3.99	−4.05	−3.36	−5.47	−5.41	−4.38	−4.40	−4.03
LW	−4.04	−4.38	−4.74	−3.88	−3.58	1.00	−3.97	−3.91	−3.46	−3.46	−2.93
YP	−3.79	−3.78	−4.11	−3.32	−3.34	0.44	−4.06	−4.09	−3.05	−3.09	−1.67
FF	−2.33	−2.85	−3.04	−2.19	−2.01	1.11	−2.07	−2.15	−1.97	−2.26	−1.89
MM	−2.03	−1.67	−2.01	−1.27	−1.56	1.22	−2.01	−1.94	−3.14	−2.35	−1.40
LY	−1.72	−1.43	−1.66	−1.21	−1.34	0.96	−2.07	−1.88	−1.86	−1.52	−1.17
LL	−1.62	−1.54	−1.60	−1.36	−1.52	0.00	−1.93	−1.96	−1.40	−1.66	−1.33
MC	−1.46	−1.22	−1.43	−0.93	−1.27	0.25	−1.48	−1.44	−2.01	−1.20	−1.26
VV	−1.39	−1.14	−1.28	−0.96	−1.18	0.44	−1.79	−1.83	−1.36	−1.43	−0.90
IL	−1.39	−1.28	−1.35	−1.12	−1.29	0.06	−1.68	−1.70	−1.19	−1.41	−1.14
II	−1.24	−0.98	−1.11	−0.80	−1.01	0.62	−1.39	−1.47	−1.13	−1.20	−1.14
LT	−1.09	−0.95	−1.02	−0.83	−0.99	0.02	−1.40	−1.36	−0.91	−1.05	−0.83
VL	−1.08	−0.94	−1.01	−0.81	−0.97	0.11	−1.34	−1.33	−0.81	−1.11	−0.86
AL	−1.07	−0.76	−0.93	−0.60	−0.82	0.71	−1.31	−1.32	−1.00	−0.94	−0.51
LG	−0.77	−0.66	−0.71	−0.56	−0.71	−0.09	−1.02	−1.00	−0.75	−0.53	−0.30
	MRE [%]	*10.96*	*6.52*	*16.05*	*12.01*	*83.61*	*13.04*	*12.64*	*19.54*	*13.55*	*19.68*
	MRX [%]	*28.82*	*−30.62*	*43.57*	*23.69*	*166.28*	*−32.92*	*−31.88*	*60.19*	*56.58*	*60.52*
	MAE	*0.47*	*0.26*	*0.48*	*0.79*	*2.03*	*0.63*	*0.58*	*2.11*	*2.22*	*0.66*
	MAX	*1.43*	*−0.85*	*2.76*	*3.28*	*6.01*	*−2.50*	*−2.45*	*18.44*	*20.43*	*3.73*
	RMS	0.48	0.36	0.60	0.88	1.40	0.73	0.68	4.16	4.78	0.77

[a] All the energies are in kcal/mol. Descriptive statistics: MRE is the unsigned mean relative error, MRX is the signed maximal relative error, MAE is the unsigned mean absolute error, MAX is the signed maximal absolute error, and RMS is the signed root mean square error. [b] Neutral arginine is not defined in the Amber03 force field.

suggesting that the contacts between these residues are shielded by the environment around the charged/polar pair. In the case of Gly and Pro, the interactions tend to involve their main-chain rather than side-chain atoms. The geometries of the nonpolar residue pairs were essentially unchanged upon the optimization, indicating that the stabilization of these pairs is mostly of enthalpic origin.

7. SAPT Decomposition of the Interaction Energies. From the SAPT results shown in Table 4, one can immediately draw a conclusion as to the dominant sources of pair stabilization. From the data, two different groups of side chain interactions may be distinguished: (1) polar and charged residues, which are stabilized mostly by the electrostatic term, and (2) nonpolar residues, which are mostly stabilized by the dispersion term.

However, several specific examples are worth noting: (1) the charged and polar residues in contact are mostly stabilized by the electrostatic term; (2) on the other hand, polar residues in contact with nonpolar residues (LT, MC) are stabilized mostly by the dispersion term. The ratio between the dispersion and electrostatic contributions ranges from 0.05 for charged salt bridges to 9.43 in the case of the Leu-Leu pair. This ratio corresponds to the nonpolarity of the interaction. As expected, the lowest value comes from the salt bridges, while polar contacts have a ratio close to unity and contacts between aromatic residues are lower than those for purely nonpolar aliphatic to aliphatic contacts.

8. Effect of the Solvent. There are many methods for treating amino acid pair interactions within a protein's interior. The most common technique of mimicking the hydrophobic environment in the protein's core is to use implicit solvation models: generalized Born models (GB),[47,48] nonlinear or linear Poisson−Boltzmann models (PB),[49] or a polarizable continuum model (PCM).[50] The most important parameter is usually the dielectric constant of a particular solvent (generally water), which represents the influence of the environment. To model the influence of a protein interior properly is a more complicated task. Several models have been proposed - low dielectric constant ($\varepsilon \sim 2-4$)[51] and high dielectric constant ($\varepsilon \sim 20-40$)[52] as well as more complex models with variable dielectric constants in the interior and at the outer regions of a protein.[53−55]

In this study, we modeled the influence of the protein core environment as solvation by diethyl ether ($\varepsilon = 4.34$), whereas the influence of water was modeled by a dielectric constant $\varepsilon = 80.0$.

We calculated the interaction energies for these two values of dielectric constant using the PCM method (see Table 5 for the results). It appears that the largest influence of the environment is to cause a steep drop of the binding energy for interactions involving charged residues and all the interactions in which the electrostatic contribution is the major term of the interaction. The values obtained for protein environment computations decrease to almost 30% of the gas-phase interaction energy values, whereas, within the water environment, the binding energies are diminished to 2−5% of their gas-phase values. The charged residues are only rarely buried in a protein interior and almost exclusively

Table 4. SAPT Decomposition for Cα in Comparison with the CCSD(T)|CBS Interaction Energies[a]

AA-AA	CCSD(T)	DFT-SAPT	E^1_{pol}	E^1_{exch}	E^2_{ind}	E^2_{disp}	δHF	E^2_{disp}/E^1_{pol}
RD	−110.80	−107.52	**−101.94**	22.28	−14.39	−7.21	−6.25	0.07
KE	−108.40	−105.78	**−96.03**	7.93	−9.99	−4.52	−3.16	0.05
DH(N)	−30.64	−28.35	**−35.96**	35.80	−12.10	−9.24	−6.85	0.26
D(N)H(N)	−17.97	−16.05	**−26.38**	33.71	−8.09	−8.89	−6.40	0.34
R(N)D(N)	−16.32	−14.68	**−19.51**	17.83	−4.04	−6.39	−2.57	0.33
K(N)E(N)	−10.76	−9.87	**−9.52**	6.79	−1.84	−4.20	−1.09	0.44
QN	−7.37	−6.83	**−10.02**	11.23	−2.21	−4.17	−1.66	0.42
TT	−6.50	−5.27	**−9.85**	12.67	−1.76	−4.96	−1.37	0.50
YY	−4.66	−3.94	−3.86	8.93	−0.34	**−7.88**	−0.79	2.04
TS	−4.50	−4.05	**−3.52**	2.92	−0.50	−2.71	−0.25	0.77
LW	−4.04	−3.58	−2.42	6.20	−0.25	**−6.56**	−0.55	2.71
YP	−3.79	−3.34	−2.25	5.24	−0.28	**−5.61**	−0.44	2.49
FF	−2.33	−2.01	−0.65	3.12	−0.13	**−4.08**	−0.26	6.28
MM	−2.03	−1.56	−1.96	5.28	−0.11	**−4.38**	−0.38	2.23
LY	−1.72	−1.34	−1.12	3.80	−0.09	**−3.70**	−0.22	3.30
LL	−1.62	−1.52	−0.21	0.71	−0.01	**−1.98**	−0.03	9.43
MC	−1.46	−1.27	−0.98	2.65	−0.12	**−2.62**	−0.19	2.67
VV	−1.39	−1.18	−0.47	2.01	−0.05	**−2.53**	−0.12	5.38
IL	−1.39	−1.29	−0.25	0.85	−0.01	**−1.83**	−0.04	7.32
II	−1.24	−1.01	−0.56	1.89	−0.02	**−2.23**	−0.09	3.98
LT	−1.09	−0.99	−0.29	0.88	−0.02	**−1.52**	−0.04	5.24
VL	−1.08	−0.97	−0.26	0.89	−0.01	**−1.55**	−0.04	5.96
AL	−1.07	−0.82	−0.66	2.18	−0.02	**−2.21**	−0.10	3.35
LG	−0.77	−0.71	−0.12	0.44	−0.01	**−0.99**	−0.02	8.25

[a] E^1_{pol} is the first-order electrostatics, $E^1_{exch}$is the first-order repulsion, E^2_{ind} is the second-order induction, E^2_{disp} is the second-order dispersion, δ_{HF} is the estimate of higher-order terms and E^2_{disp}/E^1_{pol} is the ratio between the dispersion and electrostatic terms. The most stabilizing terms are boldface.

Table 5. Solvent Effects on the Interaction Energies Calculated by the DFT-D/TPSS|TZVP with PCM in kcal/mol[a]

AA-AA	vacuum	ether		water	
RD	−112.93	−30.70	(27.2%)	−3.23	(2.9%)
KE	−110.90	−33.24	(30.0%)	−7.91	(7.1%)
DH(N)	−31.47	−10.88	(34.6%)	−2.31	(7.3%)
D(N)H(N)	−19.29	−13.86	(71.9%)	−10.45	(54.2%)
R(N)D(N)	−17.17	−7.36	(42.9%)	−2.34	(13.6%)
K(N)E(N)	−12.60	−8.38	(66.5%)	−5.89	(46.7%)
QN	−7.35	−4.45	(60.5%)	−2.55	(34.7%)
TT	−7.53	−5.56	(73.8%)	−4.10	(54.4%)
YY	−4.35	−3.77	(86.7%)	−3.28	(75.4%)
TS	−5.47	−3.18	(58.1%)	−1.59	(29.1%)
LW	−3.97	−3.46	(87.2%)	−3.02	(76.1%)
YP	−4.06	−2.84	(70.0%)	−2.27	(55.9%)
FF	−2.07	−1.55	(74.9%)	−1.26	(60.9%)
MM	−2.01	−1.73	(86.1%)	−1.55	(77.1%)
LY	−2.07	−1.84	(88.9%)	−1.72	(83.1%)
LL	−1.93	−1.87	(96.9%)	−1.85	(95.9%)
MC	−1.48	−1.04	(70.3%)	−0.83	(56.1%)
VV	−1.79	−1.73	(96.6%)	−1.70	(95.0%)
IL	−1.68	−1.64	(97.6%)	−1.62	(96.4%)
II	−1.39	−1.34	(96.4%)	−1.32	(95.0%)
LT	−1.40	−1.32	(94.3%)	−1.28	(91.4%)
VL	−1.34	−1.30	(97.0%)	−1.28	(95.5%)
AL	−1.31	−1.28	(97.7%)	−1.27	(96.9%)
LG	−1.02	−0.98	(96.1%)	−0.96	(94.1%)

[a] The percents in parentheses are relative to the vacuum value.

play a role there as a part of an active site. Their role as a stabilization element in a protein interior is highly improbable.

In terms of the modulation of binding energy strengths with the introduction of solvents, the nonpolar (aliphatic and aromatic) residues behave quite differently than the polar ones. The introduction of neither water nor ether (i.e., protein environment) strongly affects the binding energies for complexes containing these types of amino acid side chains (which is especially true for aliphatic-aliphatic interactions). Generally, solvent effects lower the binding energy of a complex containing nonpolar residues by at most ∼25%. This finding has critical implications in terms of the role of nonpolar interactions in stabilizing a protein.

Conclusion

Here we have calculated the reference binding energies for 24 different pairs of amino acid side chain interactions at the benchmark level of theory (CCSD(T)|CBS). The geometries of the studied structures were derived from X-ray crystal structure data to a resolution of 2.0 Å or better. We expect the resulting interaction energies to be very close to the (still unknown) true interaction energies and to be equally reliable for different types of side chain interactions. One key point concerning the data obtained for these complexes is that each of the interactions was evaluated as attractive. This would not be the case for pairs of similarly charged side chains, and there are no such examples in our set. However, the fact that all the interactions studied here are attractive supports the idea that enthalpic stabilization plays a key role in protein stabilization and that the interactions are nonrandomly distributed within the protein structure. This finding is supported by the geometry optimization of the most populated pairwise interactions, which does not result in any significant changes to the conformations of the interacting side chains taken from the atlas. Allow us to emphasize again that such an essential statement can be made only when using the highly accurate CCSD(T)/CBS procedure. We are certainly aware that all these conclusions concern the stabilization energy and for comparison with experiment it is inevitable to pass to stabilization enthalpy; i.e. to include the zero point vibration energy (ZPVE) term. We determined this term for the weakest pair: the leucine-glycine. Adding the ZPVE (MP2/TZVP] to ΔE we obtained ΔH=−0.3 kcal/

mol. Evidently, the above-mentioned fact that all interaction pairs are attractive remains unchanged even when the ZPVE is taken into account.

In terms of gas-phase binding energies, the strongest interactions were found to be those between two oppositely charged side chains. In general, the strength of a gas-phase interaction is positively correlated with the polarity of the side chains involved in the interaction. Interactions between aromatic side chains are generally stronger than those between aliphatic ones, which has been observed in many previous studies and is chiefly attributable to π-π and CH-π interactions. Amazingly, the range of binding energies observed in this study is extremely large, spanning from -0.77 kcal/mol (LG) to -110.80 kcal/mol (RD). Although polar interactions tend to be the strongest in the gas phase, it is important to keep in mind that these are the types of interactions that are most strongly affected by the introduction of a solvents such as water and ether (which mimics the environment within a protein interior). The introduction of solvent thus dramatically reduces the overall range of binding energies, with interaction energies ranging from -0.83 kcal/mol (MC) to -10.45 kcal/mol (D(N)N(N)) in aqueous solution (as computed using DFT-D).

There are many computational methods that can be used to study protein structure; in this work, we have assessed the performance of several commonly used *ab initio* and force-field techniques in terms of their abilities to produce the accurate binding energies for side chain-side chain complexes (using the estimated CCSD(T)|CBS results as a reference). Not surprisingly, it was found that the computationally intensive MP2|aTZ method yields the most accurate binding energies. The less computationally demanding SCSMI-MP2|TZ and DFT-D|TZVP methods were shown to prove reasonably accurate binding energy results for these complexes, with both techniques yielding better values for strongly bound complexes than for weakly bound ones. The fastest electronic structure method with reasonable accuracy was RI-DFT-D|TZVP. In terms of the force-field methods, the widely used parm03 force field was found to yield the best-balanced interaction energy results.

Acknowledgment. This work was supported by Grant Nos. 203/05/0009, 203/06/1727, and 203/05/H001 from the Czech Science Foundation, Grant No. A400550510 from the Grant Agency of the Academy of Sciences of the Czech Republic, and 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 nos. Z40550506 and MSM6198959216. The authors acknowledge the support from the Praemium Academiae award of the Academy of Sciences of the Czech Republic, awarded to P.H. in 2007. K.R. gratefully acknowledges the support of the NSF EPSCOR program (EPS-0701525).

Supporting Information Available: Interaction energy for the whole cluster of the Leu-Trp pair calculated for Cβ fragmentation (Table S1) and rmsd between the structures before and after full geometry optimization of the pairs (Table S2). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Taverna, D. M.; Goldstein, R. A. *J. Mol. Biol.* **2002**, *315* (3), 479.

(2) Taverna, D. M.; Goldstein, R. A. *Proteins: Struct., Funct., Genet.* **2002**, *46* (1), 105.

(3) McKusick, V. A. *Mendelian Inheritance in Man; A Catalog of Human Genes and Genetic Disorders*; The Johns Hopkins University Press: Baltimore, MD, U.S.A., 1998.

(4) Dill, K. A. *Biochemistry* **1990**, *29* (31), 7133.

(5) Richards, F. M.; Lim, W. A. *Q. Rev. Biophys.* **1993**, *26* (4), 423.

(6) Mcdonald, I. K.; Thornton, J. M. *J. Mol. Biol.* **1994**, *238* (5), 777.

(7) Richards, F. M. *J. Mol. Biol.* **1974**, (82), 1.

(8) Bromberg, S.; Dill, K. A. *Protein Sci.* **1994**, *3* (7), 997.

(9) Banerjee, R.; Sen, M.; Bhattacharya, D.; Saha, P. *J. Mol. Biol.* **2003**, *333* (1), 211.

(10) Chakrabarti, P.; Bhattacharyya, R. *Prog. Biophys. Mol. Biol.* **2007**, *95* (1−3), 83.

(11) Misura, K. M. S.; Morozov, A. V.; Baker, D. *J. Mol. Biol.* **2004**, *342* (2), 651.

(12) Mitchell, J. B. O.; Laskowski, R. A.; Thornton, J. M. *Proteins: Struct., Funct., Bioinf.* **1998**, *29* (3), 370.

(13) Morozov, A. V.; Misura, K. M. S.; Tsemekhman, K.; Baker, D. *J. Phys. Chem. B* **2004**, *108* (24), 8489.

(14) Jernigan, R. L.; Bahar, I. *Curr. Opin. Struct. Biol.* **1996**, *6* (2), 195.

(15) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256* (3), 623.

(16) Andersson, Y.; Hult, E.; Apell, P.; Langreth, D. C.; Lundqvist, B. I. *Solid State Commun.* **1998**, *106* (5), 235.

(17) Rapcewicz, K.; Ashcroft, N. W. *Phys. Rev. B* **1991**, *44* (8), 4032.

(18) Vondrasek, J.; Bendova, L.; Klusak, V.; Hobza, P. *J. Am. Chem. Soc.* **2005**, *127* (8), 2615.

(19) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8* (17), 1985.

(20) Holmes, J. B.; Tsai, J. *J. Mol. Biol.* **2005**, *354* (3), 706.

(21) Laskowski, R. A. M. S.; Thornton, J. M. Atlas of Side-chain Interactions. http://www.ebi.ac.uk/thornton-srv/databases/sidechains. (accessed Oct 31, 2008).

(22) Singh, J.; Thornton, J. M. *Atlas of Protein Side-Chain Interactions*; 1992.

(23) DeLano, W. L. *The PyMOL Molecular Graphics System, 0.99rc6*; DeLano Scientific: Palo Alto, CA, U.S.A., 2002.

(24) Riley, K. E.; Op't Holt, B. T.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2007**, *3* (2), 407.

(25) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. *Chem. Phys. Lett.* **1989**, *162* (3), 165.

(26) Jurecka, P.; Hobza, P. *Chem. Phys. Lett.* **2002**, *365* (1−2), 89.

(27) Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8* (17), 1985.

(28) Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286* (3−4), 243.

(29) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO, version 2006.1, a package of ab initio programs*; 2007. See http://www.molpro.net.

(30) Valdes, H.; Klusak, V.; Pitonak, M.; Exner, O.; Stary, I.; Hobza, P.; Rulisek, L. *J. Comput. Chem.* **2007**, 861.

(31) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19* (4), 553.

(32) Distasio, R. A.; Head-Gordon, M. *Mol. Phys.* **2007**, *105* (8), 1073.

(33) Grimme, S. *J. Chem. Phys.* **2003**, *118* (20), 9095.

(34) Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. R. *J. Comput. Chem.* **2007**, *28* (2), 555.

(35) Frisch, M. J. T. 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.

(36) Cerny, J.; Jurecka, P.; Hobza, P.; Valdes, H. *J. Phys. Chem. A* **2007**, *111* (6), 1146.

(37) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26* (16), 1701.

(38) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105* (28), 6474–6487.

(39) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, *24* (16), 1999.

(40) Sorin, E. J.; Pande, V. S. *Biophys. J.* **2005**, *88* (4), 2472.

(41) Hesselmann, A.; Jansen, G.; Schutz, M. *J. Chem. Phys.* **2005**, *122* (1), 14103.

(42) Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *362* (3−4), 319.

(43) Hesselmann, A.; Jansen, G.; Schutz, M. *J. Am. Chem. Soc.* **2006**, *128* (36), 11730.

(44) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120* (1−3), 215.

(45) Tsuzuki, S.; Uchimaru, T.; Tanabe, K. *Chem. Phys. Lett.* **1998**, *287* (1−2), 202.

(46) Bendova-Biedermannova, L.; Hobza, P.; Vondrasek, J. *Proteins: Struct., Funct., Bioinf.* **2008**, *72* (1), 402.

(47) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104* (15), 3712.

(48) Tsui, V.; Case, D. A. *Biopolymers (Nucleic Acid Sci.)* **2001**, *56*, 257.

(49) Wagoner, J.; Baker, N. A. *J. Comput. Chem.* **2004**, *25* (13), 1623.

(50) Riley, K. E.; Vondrašek, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2007**, *9* (41), 5555.

(51) Gilson, M. K.; Honig, B. H. *Biopolymers* **1911**, *1986*, 2097.

(52) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *J. Mol. Biol.* **1994**, *238* (3), 415.

(53) Archontis, G.; Simonson, T. *J. Phys. Chem. B* **2005**, *109* (47), 22667.

(54) Schutz, C. N.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2001**, *44* (4), 400.

(55) Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; Garcia-Moreno, E. B. *Biophys. J.* **2002**, *82* (6), 3289.

CT800508V