JCTC Journal of Chemical Theory and Computation

Phenylalanyl-Glycyl-Phenylalanine Tripeptide: A Model System for Aromatic−Aromatic Side Chain Interactions in Proteins

H. Valdes,*,† K. Pluhackova,‡ and P. Hobza*,‡,§

Dpto. Química Física y Analítica, Universidad de Oviedo, C/Julián Clavería, 8, 33006 (Oviedo) Asturias, Spain, Center for Biomolecules and Complex Molecular Systems, Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, 16610 Prague 6, Czech Republic, and Department of Physical Chemistry, Palacky University, 771 42 Olomouc, Czech Republic

Received April 10, 2009

Abstract: The performance of a wide range of quantum chemical calculations for the ab initio study of realistic model systems of aromatic−aromatic side chain interactions in proteins (in particular those $\pi-\pi$ interactions occurring between adjacent residues along the protein sequence) is here assessed on the phenylalanyl-glycyl-phenylalanine (FGF) tripeptide. Energies and geometries obtained at different levels of theory are compared with CCSD(T)/CBS benchmark energies and RI-MP2/cc-pVTZ benchmark geometries, respectively. Consequently, a protocol of calculation alternative to the very expensive CCSD(T)/CBS is proposed. In addition to this, the preferred orientation of the Phe aromatic side chains is discussed and compared with previous results on the topic.

Introduction

Proteins are linear heteropolymers composed of amino acids. Most proteins fold into unique three-dimensional structures (tertiary structure or native state) in which they perform their biological functions.[1] A folded protein is a complex structure containing very different types of intra- and interresidual interactions. One of the reasons to be interested in the nature of these intramolecular forces is based on the well-known fact that interactions between side chains largely favor the molecule's acquisition of the folded state. Obviously, then a better understanding of the folding process can be gained by studying both the interactions between adjacent residues covalently bound as well as the interactions between noncovalently bound regions. Noncovalent interactions compile a wide variety of weak intermolecular forces (H-bonds, cation$-\pi$ interactions, etc.) among which the $\pi-\pi$ interactions are known to play a relevant role in the stability of proteins. Indeed, around 64% of aromatic side chains in proteins are likely involved in $\pi-\pi$ intramolecular interactions with neighbor aromatic side chains.

Residues buried in the hydrophobic core of any protein are shielded from solvent, and thus, they attain an environment which may be very similar to that in the gas phase. For that reason, noncovalent interactions occurring in the hydrophobic core of any protein may be studied from a quantum chemical point of view. However, such treatment necessarily requires—for obvious reasons—the definition of a much simpler system than a protein; i.e., quantum chemical studies on noncovalent interactions in proteins are typically restricted to some relevant parts of it. In this respect, we have previously focused on the quantum chemical study of the noncovalent interactions between the peptide backbone and the aromatic side chains in di- and tripeptides.[2−6] In the present work, we concentrate on the study of noncovalent interactions between aromatic side chains of adjacent residues along the protein sequence.

Dealing with the quantum chemical computation of aromatic−aromatic interactions in proteins entails mainly two difficulties, the selection of a good prototype model system

* To whom correspondence should be addressed. E-mail: haydee.valdes@marge.uochb.cas.cz.

† Universidad de Oviedo.

‡ Academy of Sciences of the Czech Republic.

§ Palacky University.

10.1021/ct900174f CCC: $40.75
Published on Web 07/28/2009

and, subsequently, the level of theory to be employed for the study. There are two possible ways to select a prototype system of the $\pi-\pi$ interactions of any protein. Having the crystal structure, one would usually concentrate on the minimal possible interacting parts of interest—aromatic side chains in this particular case—and *remove* everything else. Lacking the crystal geometries, one cannot but optimize ab initio the geometries of the model system. This is, however, a quite critical step, since it is questionable how close the minimum found is to the real geometrical disposition of the aromatic side chains in a protein. In other words, the simplest model of the aromatic−aromatic interactions in proteins one could think of is undoubtedly the benzene dimer. Certainly, phenylalanine residues have benzene-like aromatic side chains that could adopt, at least hypothetically, similar geometrical rearrangements to those shown by the benzene dimer. But, undoubtedly, there are some factors affecting the orientation of the aromatic side chains in proteins that are not present in the benzene dimer model, namely, (a) the likely interaction between the aromatic side chains and the peptide backbone; (b) the ϕ and Ψ preferences of allowed regions in protein structures (Ramachandran plot) that determine the local shape assumed by the protein backbone and, thus, the orientation of the side chains, and, finally, (c) peptide bonds linking the residues have a double character that hinders rotation around its axis—guaranteeing that the α carbons are roughly coplanar—and act upon the geometrical disposition of the side chains. Hence, accurately modeling aromatic−aromatic side chain interactions in proteins requires realistic models including all of the above-mentioned factors.

Aromatic−aromatic interactions have been intensively explored up to now by means of rigorous electronic structure computations. Illustrative examples are, for instance, the elaborated works by Urch et al.,[7] Sherrill et al.,[8] and Diederich et al.,[9] where a detailed bibliography on the topic can be found. So, regarding the level of theory to be employed for the study, it is already a very well-known fact that a definitive treatment of the dispersion energy can only be done by using the CCSD(T) method with large basis sets including multiple polarization and diffuse functions;[8] thus, this should be the method to be employed for any study of such characteristics. However, realistic models of aromatic−aromatic interactions in proteins are incompatible with small-size systems, and obviously, such high-level quantum chemical calculations are prohibitive. Then, a level of theory alternative to the very expensive CCSD(T)/large basis set has to be necessarily found.

Plenty of discussion has been done up to now about the performance of the MP2 and the density functional theory (DFT) methods. On one hand, the major disadvantage of the MP2 method is its overestimation of the dispersion energy—relevant in the aromatic−aromatic interactions—when an extended basis set [or even at the complete basis set (CBS) limit] is applied. Reliable energies (and also geometries) are thus frequently obtained when using medium size basis sets. However, this is evidently due to a compensation of errors and it is impossible to rely on this compensation. On the other hand, DFT lacks a proper description of the dispersion interaction. Consequently, many alternatives have appeared quite recently trying to compensate the deficiencies of both methods.[10−16] Hence the question of which level of theory is more suitable for the study of aromatic−aromatic interactions in protein model systems is topical, and for this reason, we present here a theoretical study on the performance of a vast range of levels of theory in comparison with the CCSD(T)/CBS benchmark data for the phenylalanyl-glycyl-phenylalanine (Phe-Gly-Phe, FGF) tripeptide. We have chosen this system since it fulfills the requirements needed for such kind of study, namely, (a) CCSD(T) single point calculations are affordable, though we are in the upper limit of what is nowadays feasible, and (b) it can be considered a realistic model of $\pi-\pi$ interactions occurring in proteins, and more specifically, unlike previous studies where the prototype systems model the aromatic−aromatic interactions occurring between nonconsecutive residues, the system here studied models those $\pi-\pi$ interactions occurring between adjacent residues along the protein sequence.

Additionally, this work deals with another degree of difficulty in comparison with simpler models, e.g., benzene$\cdots$$NH_3$ dimer[17,18] or benzene dimer,[19] since both H-bond and $\pi-\pi$ interactions coexist in the molecule, which shows the need for a method describing correctly and, more important, simultaneously those interactions. Finally, the importance of the basis set superposition error (BSSE) cannot be forgotten.[20−22] When dealing with the calculation of aromatic−aromatic intermolecular interactions, this error is generally corrected by the counterpoise (CP) procedure.[23] However, the case of isolated systems is far more complicated, because neither a well-established method accounting for this error nor a CP-like procedure has yet been developed. Moreover since small peptides are systems of multiconfigurational character, each particular conformation suffers differently from the intramolecular BSSE, which stresses the importance of choosing a level of theory where this error has been, if not erased, at least minimized.

Ultimately, this work aims to gain a better understanding of the $\pi-\pi$ interactions and by extension their role played in proteins from the information obtained on the pure—without any influence of the solvent—aromatic−aromatic interactions provided by calculations in vacuo on isolated small peptides.

Computational Details

As mentioned already in the Introduction, small peptides are systems of multiconformational character, thus showing a very rich conformational landscape. Since only few conformers are typically experimentally detected,[24−26] we have restricted our benchmark study to a set containing 15 energetic and geometrically different conformers. These structures have been selected according to a strategy of calculation previously proven efficient[5] and they constitute the most stable structures in the potential energy surface of FGF. This set contains those conformers observed experimentally and simultaneously; the number of structures included in it is small enough so that high-level calculations can be carried out for all of them. Different levels of theory have then been tested against CCSD(T)/CBS values. A table containing a time scale for the different computational

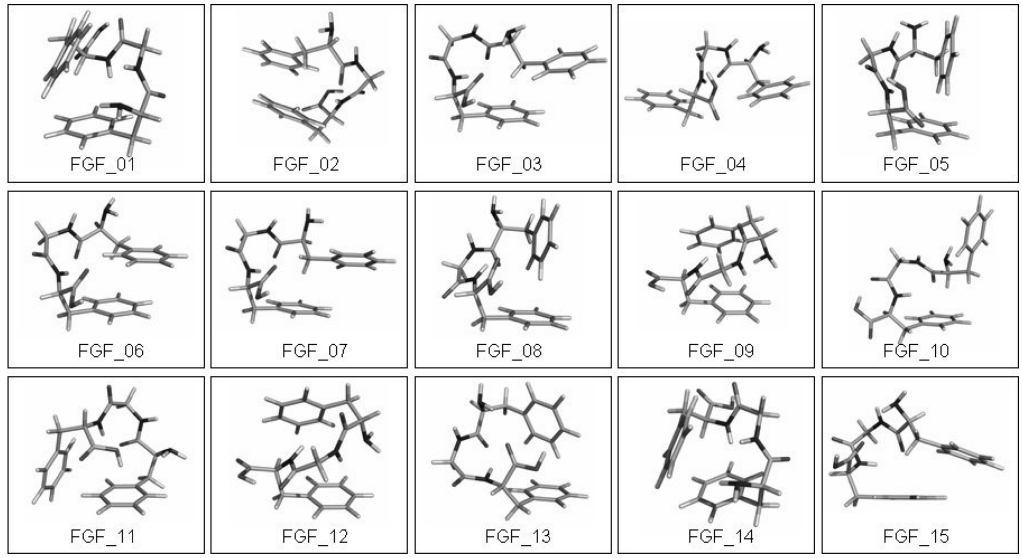

Figure 1. RI-MP2/cc-pVTZ geometries of the 15 most stable structures of FGF tripeptide.

methods employed in the study is included in the Supporting Information (see Table S1).

Empirical Force Field. Single-point energy calculations and geometry optimizations were carried out by means of parm99 empirical force field.[27] B3LYP/cc-pVTZ atomic charges obtained using the restrained electrostatic potential fitting procedure[28] (RESP) have been used for these calculations. Notice that the RESP charges finally used are the average of the RESP charges of six different structures (FGF_02, FGF_03, FGF_04, FGF_05, FGF_10, and FGF_12; see Figure 1).

Tight-Binding Method Extended by an Empirical Dispersion Term (SCC-DF-TB-D). Energy minimizations, single-point energy calculations, and geometry optimizations were obtained by means of the SCC-DF-TB-D[29] method, which includes a term describing the dispersion energy essential for the proper study of peptides containing two aromatic side chains. Additionally, it is a very fast method and particularly interesting for an accurate scanning of the potential energy surface of any small peptide.

Density Functional Theory. Different functionals and basis sets have been tested for geometry optimization, namely, (a) B3LYP[30]/6-31G*,[31] since it is one of the levels of theory commonly used for the study of isolated small peptides; (b) TPSS[32]/6-311++G(3df,3pd)[31] and TPSS-D/6-311++G(3df,3pd), since the latter has been recently proven to perform reasonably well for the study of isolated systems;[3] (c) M06-2X[16]/MIDI!,[33] due to the fact that it belongs to a brand new generation of hybrid meta-generalized-gradient-approximation exchange correlation functionals that include an accurate treatment of medium-range correlation energy that mainly concerns the London dispersion energy; and, finally, (d) M06-L[34]/TZVP[35] since it is much cheaper than M06-2X and consequently, can be combined with larger basis sets without losing computational efficiency.

Additionally, single-point energy calculations on RI-MP2/cc-pVTZ geometries have been also systematically carried out at the following levels of theory: (a) B3LYP/6-311++G(3df,3pd); (b) TPSS/6-311++G(3df,3pd); (c) TPSS-D/6-311++G(3df,3pd); (d) M06-2X/6-311+G(2df,2pd);[31] (e) M06-L/6-311++G(3df,3pd); (f) BH&H[36]/6-311++G-(d,p); and (g) PBE[37]-D/TZVP.

Wave Function Theory (WFT) Calculations. RI-MP2/cc-pVTZ geometries are here considered as the benchmark. CCSD(T) energies where obtained according to the following equation:

$$E_{\mathrm{CBS}}^{\mathrm{CCSD(T)}} = E_{\mathrm{CBS}}^{\mathrm{MP2}} + (E^{\mathrm{CCSD(T)}} - E^{\mathrm{MP2}})\Big|_{\text{6-31G*}} \qquad (1)$$

where MP2/CBS energies were calculated by the extrapolation of the RI-MP2[38,39]/cc-pVTZ[40] and RI-MP2/cc-pVQZ[40] relative energies using the scheme of Helgaker and co-workers.[41] The second term of eq 1 covers the portion of correlation energy beyond the second perturbation order[42] and it has been calculated using a small basis set, as it has been demonstrated that the CCSD(T)-MP2 energy difference depends little on the size of the basis set.[43] Equation 1 has been previously proven as an efficient way to approximate CBS energies for large systems where otherwise such calculations are prohibited.[43] $E_{\mathrm{CBS}}^{\mathrm{MP3}}$ energies were also calculated following eq 1 except for the correlation energy beyond the second order, which is given by the $(E^{\mathrm{MP3}} - E^{\mathrm{MP2}})|_{\text{6-31G*}}$ term. SCS-MP2[14] and SCS(MI)-MP2[15] (with spin-component scaling factors $c_{\mathrm{os}} = 0.40$ and $c_{\mathrm{ss}} = 1.29$) methods have also been tested. For both methods, energies have been extrapolated to the CBS limit using the scheme of Helgaker and co-workers. Additionally, in case of the SCS(MI)-MP2 method, we have also tested another extrapolation scheme suggested by the authors.[15] Since the differences between the results obtained with the different schemes of extrapolation are negligible, for simplicity reasons, we will only show the results obtained using the scheme of Helgaker and co-workers.

Results and Discussion

Figure 1 shows the RI-MP2/cc-pVTZ geometries of the 15 most stable structures in the SCC-DF-TB-D potential energy

Table 1. Root-Mean-Square Deviations (RMSDs) (in Å) between the RI-MP2/cc-pVTZ Benchmark Geometries and Geometries Obtained at Different Levels of Theory[a]

	B3LYP/ 6-31G*	TPSS/ LP_1	TPSS-D/ LP_1	M06-2X/ MIDI!	M06-L/ TZVP	SCC-DF-TB-D	ff99
FGF_01	0.37	0.63	0.14	0.04		0.12	0.21
FGF_02	1.13	1.06	0.82	0.10	0.13	0.08	0.12
FGF_03	0.90	0.88	0.16	0.08	0.15	0.10	0.13
FGF_04	0.30	1.09	0.21	0.14	0.23	0.05	0.12
FGF_05	0.90	0.84	0.29	0.05	0.10	0.05	0.15
FGF_06	2.33	1.09	0.11	0.10	0.14	0.07	0.14
FGF_07	0.78	1.09	0.21	0.08	0.16	0.07	0.32
FGF_08	0.30	0.40	0.12	0.07	0.21	0.09	0.53
FGF_09	0.26	0.51	0.09	0.07	0.06	0.06	0.21
FGF_10	0.62		0.18	0.22	0.39	0.06	0.32
FGF_11	0.45	0.63	0.51	0.14	0.46	0.06	0.13
FGF_12	1.08	0.94	0.31	0.10	0.16	0.07	0.23
FGF_13	0.43	1.66	0.11	0.11	0.13	0.10	0.13
FGF_14	0.41	0.87	0.10	0.10	0.15	0.07	0.16
FGF_15	1.10	0.94	0.17	0.10	0.13	0.08	0.16
average	0.76	0.90	0.24	0.10	0.19	0.08	0.20

[a] Average RMSDs values are also included. LP_1 stands for the 6-311++G(3df,3pd) Pople basis set.

surface of FGF tripeptide. Cartesian coordinates of those structures can be found in the Supporting Information. Additionally, all these coordinates (together with relative energies) can be found in the following web page: www.begdb.com.[44] These geometries have been taken as reference geometries for the assessment of various levels of theory: (a) B3LYP/6-31G*; (b) TPSS/6-311++G(3df,3pd); (c) TPSS-D/6-311++G(3df,3pd); (d) M06-2X/MIDI!; (e) M06-L/TZVP; (f) SCC-DF-TB-D; and (g) geometries obtained using the ff99 force field.

Table 1 collects the root-mean-square deviations (RMSDs) in angstroms obtained for each individual conformer as compared to the benchmark geometries. Average deviations of each particular level of theory are also collected. On the basis of the average deviations shown in Table 1, it is easily inferred that SCC-DF-TB-D and M06-2X/MIDI! show a similar behavior and differ little from the RI-MP2/cc-pVTZ geometries. Indeed, the smaller deviations shown by the SCC-DF-TB-D method suggest that it is highly recommendable for obtaining reliable geometries of aromatic−aromatic protein model systems at an extremely low computational cost. The M06-2X/MIDI! level of theory is equally reliable but certainly more computationally expensive. Additionally, the behavior of the M06-L/TZVP level of theory is slightly worse but still very acceptable and certainly less computational expensive than M06-2X/MIDI!. Also true is that whereas the former two methods rarely show individual conformer deviations larger than 0.2 Å, for the latter, approximately 25% of the individual deviations differ more than 0.2 Å from the benchmark geometries. Notice also that at this level of theory it was not possible to localize (since after 426 steps the convergence criteria was still not satisfied) the FGF_01 structure in the potential energy surface.

Even when on average the ff99 empirical force field performance is comparable to the M06-L/TZVP level of theory, individual geometries obtained using the force field deviate more from the benchmark geometries, as 40% of the conformers show RMSD values higher than 0.2 Å. Consequently, its overall performance is slightly worse than that of the M06-L/TZVP level of theory. Notice, however, that atomic charges of a flexible molecule—like the one here considered—may vary significantly from one conformation to another, obviously affecting the final results. In the present case, the best results were obtained when the RESP charges used were the average of the RESP charges of six structures of different conformations (FGF_02, FGF_03, FGF_04, FGF_05, FGF_10, and FGF_12; see Figure 1).

Geometries obtained at the TPSS-D/6-311++G(3df,3pd) level of theory deviate from the benchmark geometries slightly more than any of the methods previously commented on. However, its overall performance is reasonable, and consequently, geometries obtained at this level should still be reliable. The opposite is true regarding the B3LYP/6-31G* level of theory. Widely used for the study of small peptides, B3LYP/6-31G* gives large RMSD values when compared to the reference geometries, the reason being no other than the poor description of the dispersion interaction given by the functional. A very illustrative example of its different performance in comparison with the TPSS-D functional or the MP2 method can be seen in Figure 2. We cannot forget, though, that the MP2 method suffers from an overestimation of the dispersion energy;[45,46] thus, geometries should be also partially affected by this overestimation, meaning that the comparison between the different methods should be done qualitatively rather than quantitatively. The importance of dispersion in FGF is also supported by the fact that geometries obtained at TPSS/6-311++G(3df,3pd), i.e., removing the dispersion energy from the TPSS functional, gives the largest deviation of all the levels of theory here compared with the RI-MP2/cc-pVTZ values.

Thus, from the analysis of Table 1, one could conclude that any of the levels of theory here discussed—but the B3LYP/6-31G* and TPSS/6-311++G(3df,3pd)—could be in principle used for obtaining reliable molecular structures of aromatic−aromatic protein model systems and its usage should be mostly dependent on the computational resources and time disposal.

Table 2 collects the mean unsigned error (MUE), standard deviation (σ), and maximum unsigned error (Max.) obtained from the comparison between the CCSD(T)/CBS benchmark energies and single-point energies—on the RI-MP2/cc-pVTZ benchmark geometries—obtained at different levels of theory. The statistical values have been calculated as follows. First we have calculated the average energies ($E_{average}$) for each particular method collected in Table 2 as well as the average energy at the CCSD(T)/CBS level of theory. Then, we have obtained the relative energies at all levels of theory subtracting the average energy from the energy of each particular conformer, i.e., $(E_{conformer} - E_{average})^{level\ of\ theory}$ and $(E_{conformer} - E_{average})^{CCSD(T)/CBS}$. Finally, we have calculated the difference $[(E_{conformer} - E_{average})^{CCSD(T)/CBS} - (E_{conformer} - E_{average})^{level\ of\ theory}]$. These last data are those collected in Table 2.

The performance of each method is well-established after the analysis of Table 2. The B3LYP/6-311++G(3df,3pd) level of theory gives the largest errors of all, followed by the empirical force field ff99, and the TPSS/6-311++G-(3df,3pd) level of theory. This is not an unexpected result, though, because dispersion is playing a major role in the

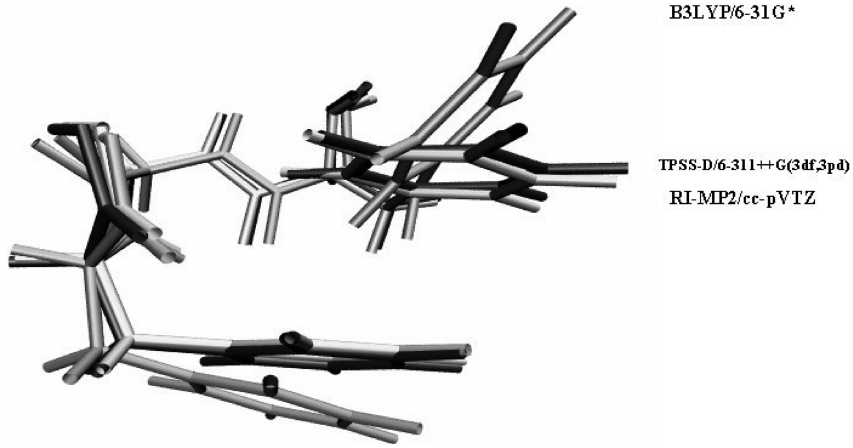

Figure 2. Comparison between the B3LYP/6-31G*, TPSS-D/6-311++G(3df,3pd), and the RI-MP2/cc-pVTZ levels of theory.

Table 2. Mean Unsigned Error (MUE), Standard Deviation (σ), and Maximum Unsigned Error (Max.) Obtained from the Comparison between the CCSD(T)/CBS Benchmark Energies (in kcal/mol) and Single-Point Energies at Different Levels of Theory on RI-MP2/cc-pVTZ Benchmark Geometries

method	MUE	σ	Max.
ff99	1.84	1.55	5.42
SCC-DF-TB-D	0.51	0.39	1.21
B3LYP/6-311++G(3df,3pd)	1.94	1.66	5.80
TPSS/6-311++G(3df,3pd)	1.63	1.72	5.53
TPSS-D/6-311++G(3df,3pd)	0.95	0.51	1.63
PBE-D/TZVP	0.97	0.59	1.91
M06-2X/6-311+G(2df,2p)	0.42	0.38	1.20
M06-L/6-311++G(3df,3pd)	0.39	0.34	0.99
BH&H/6-311++G(d,p)	0.82	0.46	1.51
SCS-MP2/CBS	0.28	0.26	0.86
SCS(MI)-MP2/CBS	0.26	0.28	0.88
MP2/CBS	0.56	0.38	1.37
MP3/CBS	0.40	0.26	0.90

stability of each of the FGF conformers and none of these methods deals properly with dispersion. The next category includes the PBE-D/TZVP, TPSS-D/6-311++G(3df,3pd), and BH&H/6-311++G(d,p) levels of theory with MUEs of approximately 1.0 kcal/mol. It can be seen then that adding dispersion improves the results, which reinforces the former statement about the importance of dispersion in the stability of this system. It also shows that if DFT is to be used for the study of systems of similar characteristics, functionals including dispersion—augmented with dispersion or specifically parametrized to cover dispersion—should be necessarily employed. The SCC-DF-TB-D method, as well as any of the Truhlar's functionals here tested, constitute a significant improvement with respect to any of the levels of theory discussed before. Their statistics are very similar to those given by any of the wave function theory methods. Among the latter, both the SCS(MI)-MP2 and the SCS-MP2 methods show the best performance. This means that the more expensive MP2/CBS and MP3/CBS levels of theory could be avoided, specially, if larger systems were to be calculated.

We have also examined if the global minimum predicted by the different levels of theory here considered is the same. Each individual level of theory except TPSS/6-311++G(3df,3pd), B3LYP/6-311++G(3df,3pd), and f99 predict FGF_01 as the global minimum, in agreement with the benchmark calculations (see Table S2 in the Supporting Information). Notice also that at the BH&H/6-311++G(d,p) level of theory the global minimum predicted is structure FGF_02. However, this structure is only 0.16 kcal/mol more stable than structure FGF_01. Indeed, according to this level of theory, there are four structures within a range of energy of 0.41 kcal/mol and thus it is really difficult to select one as the global minimum. Additionally, in the majority of the levels of theory considered, the 15 structures of the set lie in an interval of energy similar to that of the benchmark calculations (approximately 3.2 kcal/mol). However, this is not the case for the TPSS/6-311++G(3df,3pd), B3LYP/6-311++G(3df,3pd), PBE-D/TZVP, and f99 data. The same holds true for the TPSS-D/6-311++G(3df,3pd) level of theory, where structures lie in a range of 6.00 kcal/mol, suggesting that one should be particularly careful if any selection of the structures is to be done on the basis of the relative energies predicted at this particular level of theory. Also noticeable is that, again at this particular level of theory, structures where an OH···O=C intramolecular H-bond occurs between the carboxylic terminal group and the C=O of the preceding residue (see for instance structure FGF_01) are systematically more stable than structures lacking this particular intramolecular interactions. However, these two families of structures are more interspersed in case of the benchmark calculations. Notice that SCC-DF-TB-D as well as BH&H/6-311++G(d,p) and PBE-D/TZVP show a similar behavior to TPSS-D/6-311++G(3df,3pd) level of theory in this particular respect.

A final comment should be made with respect to our previous paper concerning a similar study, though on a different type of intramolecular interaction within the peptide, i.e., peptide backbone—aromatic side chain interactions.[3] The performance of the methods is consistent for both families of peptides (backbone—aromatic side chain and aromatic—aromatic side chain). Regarding geometries, average RMSD values calculated for both type of systems are almost the same. The only exception is for the B3LYP/6-31G* level of theory, where average RMSD values are larger in the case of the FGF system. Obviously, this is due to the fact that the performance of the B3LYP functional worsens as the aromatic character of the system increases. Regarding

Table 3. Mean Unsigned Error (MUE), Standard Deviation (σ), and Maximum Mean Unsigned Error (Max.) Obtained between the CCSD(T)/CBS Benchmark Energies (in kcal/mol) and Single-Point Energies on the MP2/cc-pVTZ Benchmark Geometries and Geometries Obtained at Each Particular Method

method[a]	MUE	σ	Max.
TPSS/LP_1//RI-MP2/cc-pVTZ	1.63	1.72	5.53
TPSS/LP_1//TPSS/LP_1	1.33	1.00	2.98
TPSS-D/LP_1//RI-MP2/cc-pVTZ	0.95	0.51	1.63
TPSS-D/LP_1//TPSS-D/LP_1	1.12	0.52	1.88
B3LYP/6-31G*//RI-MP2/cc-pVTZ	1.81	1.43	5.35
B3LYP/6-31G*//B3LYP/6-31G*	1.25	1.06	3.19
M06-2X/MIDI!//RI-MP2/cc-pVTZ	0.55	0.44	1.41
M06-2X/MIDI!//M06-2X/MIDI!	0.54	0.38	1.31
M06-L/TZVP//RI-MP2-cc-pVTZ	0.44	0.37	1.28
M06-L/TZVP//M06-L/TZVP	0.44	0.39	1.29
SCC-DF-TB-D//RI-MP2/cc-pVTZ	0.51	0.39	1.21
SCC-DF-TB-D//SCC-DF-TB-D	0.51	0.32	1.05
ff99//RI-MP2/cc-pVTZ	1.84	1.55	5.42
ff99//ff99	1.03	0.70	2.54

[a] LP_1 stands for the 6-311++G(3df,3pd) Pople basis set.

energies, the same trends are observed for both families of peptides. Summarizing, the ff99 force field and standard—not augmented with dispersion or not specifically parametrized to cover dispersion—functionals deviate more from the benchmark energies. Then, DFT improved, i.e., augmented, with dispersion in any possible way performs reasonably well. More specifically, M06-2X and M06-L Truhlar functionals show a better performance than the PBE-D or TPSS-D functionals. Additionally, wave function theory methods show the smallest errors, particularly SCS-MP2 and SCS(MI)-MP2, in comparison with the benchmark data here considered. It should be stressed once again that studying proteins from a quantum chemical point of view necessarily implies restricting, for obvious reasons, the study to those particular areas of interest, e.g. $\pi-\pi$ interactions and peptide backbone−aromatic side chain interactions. However, a global perspective on the topic certainly requires the combination of the conclusions obtained from each individual study. Then, it would be highly unsatisfactory if a certain method worked well, for instance, for peptide backbone−aromatic side chain interactions but failed in the description of $\pi-\pi$ interactions. As we have just seen that the overall performance of the methods is consistent for both $\pi-\pi$ and peptide backbone−aromatic side chain interactions; thus, by choosing the proper method, we can trust that the description of both types of intramolecular interactions is correct.

Table 3 collects the MUE, σ, and Max. obtained from the comparison between single-point energies calculated on the benchmark geometries (RI-MP2/cc-pVTZ) and those geometries included in Table 1 aiming to study the possible influence that the selection of different geometries may have on the final energies. It can be seen that energies calculated using methods covering reasonably well the dispersion energy do not depend much on the geometry chosen for the single-point energy calculations. However, larger differences are found when the TPSS, B3LYP, or ff99 methods are chosen. Interestingly, for these three particular cases, the mean unsigned errors as well as the standard deviations and the maximum error are smaller than those obtained from single-point energies calculated on benchmark geometries, suggesting that a cancellation of error occurs when the former methods are used. Notice that, on one hand, TPSS and B3LYP have the largest geometry deviations compared to RI-MP2/cc-pVTZ geometries, which can have an impact on single-point calculations. On the other hand, ff99 geometries have a low RMSD, but force-field methods are known to be extremely sensitive to changes in the geometry.

Insights into the Preferred Orientation of the Phe Aromatic Side Chains

The FGF tripeptide is a system of multiconformational character where various different types of noncovalent intramolecular interactions are simultaneously present and consequently the quantum chemical method to be used for its study has to deal in a balanced manner with all of these intramolecular forces at once. Thus, examining which method—apart from the very computationally expensive CCSD(T)/CBS—gives the best results at the lowest computational cost is topical and interesting by itself and it constitutes the main scope of this study. However, the study of FGF can also provide interesting biological information. Regarding this point, many different aspects can be analyzed, for instance, the conformational preferences of the peptide backbone or the interaction—via multiple NH$\cdots\pi$ interactions—between the peptide backbone and the aromatic side chains. For the present case, the orientation of the aromatic side chains with respect to each other deserves special attention. Our aim is to shed some light on the *pure*—without the influence of any other factor such as the neighbor residues or hydrophobic effects—intramolecular aromatic−aromatic interactions between the aromatic side chains of residues in the hydrophobic core of globular proteins assuming, as already commented on the introduction, that residues buried in the hydrophobic core attain an environment similar to that in the gas phase. Thus, the conclusions here obtained can be extrapolated up to some point to the behavior of the aromatic side chains of residues within the hydrophobic core of a protein.

Looking at the 15 conformers collected in Figure 1, it is possible to group all these structures into three different categories according to the orientation of the aromatic side chains with respect to each other, namely, (a) *stacked*, those where the aromatic side chains are slipped parallel/parallel to each other, i.e., FGF_02, FGF_06, and FGF_15; (b) *T-shaped*, those where the aromatic side chains are in a T-shaped disposition, i.e., FGF_05, FGF_08, FGF_11, and FGF_13; and (c) *others*, those conformers do not matching any of the criteria mentioned before. According to this geometrical classification, it seems reasonable to conclude then that neither the T-shaped nor the stacked orientations are preferred by the Phe aromatic side chains. Indeed, the number of conformers belonging to both families (three to the first and four to the second) is almost the same. Moreover, according to the data here collected, there is neither a clear preference for the T-shaped family over the stacked or vice versa. The most abundant conformers are those showing a geometrical disposition favoring the maximum number of intramolecular interactions, i.e., those geometries where the

maximum number of H-bonds and NH$\cdots\pi$ and aromatic−aromatic intramolecular interactions are acting together.

We have also calculated the Gibbs relative energies of each particular conformer at $T = 300$ K from TPSS-D/TZVPP ab initio quantum chemical calculations assuming a rigid rotor−harmonic oscillator−ideal gas approximation, since we have already concluded for similar systems that this procedure provides a reliable description of the free energy surface (FES) in a vacuum.[4] According to Gibbs relative energies calculations, these same structures could be ordered as follows (see Table S3 in the Supporting Information): FGF_08 ∼ FGF_07 < FGF_04 < FGF_14 < FGF_06 < FGF_09 < FGF_13 < FGF_12 < FGF_15 < FGF_02 < FGF_03 < FGF_11 < FGF_01 < FGF_05. On the basis of this order, it seems clear that stacked structures are not the most energetically favored. This result may seem in contrast with the work published by Schettino et al.,[47] where it is concluded that, in a hydrophobic environment, such as the protein core, Phe-Phe systems show a slight preference for stacking. However, it should be here explicitly mentioned that these two studies are not straightforwardly comparable, since the prototype systems used by Schettino et al. are simpler than ours. Schettino et al. constructed the models (complexes) for the Phe side chains from the corresponding amino acids by removing the amino and carboxylic groups, and consequently, this work does not take into account the influence of the interactions between the backbone and the aromatic side chains. However, from our study it is clearly inferred that such interactions play a determining role in the final orientation of the aromatic side chains. Indeed, the same conclusion was implicitly obtained by Kollman et al.[19] from a study carried out using benzene and toluene dimers as model systems of aromatic interactions in proteins. Since results obtained with the different model systems were in conflict, Kollman et al. concluded that very simple prototypes hardly model the Phe side chains behavior in proteins. Also interesting is that the structure of the Ac-Phe-Phe-NH_2 system characterized by means of IR/UV double resonance spectroscopy in the gas phase[48] is, as in the case of the FGF peptide, a T-shaped structure.

Summary and Conclusions

Aromatic side chains of proteins often participate in $\pi-\pi$ interactions. Studying $\pi-\pi$ interactions in proteins by means of quantum chemical calculations imposes a restriction on the size of the protein model system to be considered. Too large systems are simply unaffordable from a computational point of view, whereas too small model systems may not be realistic enough and may skip some relevant information, as for instance the geometrical restrictions imposed by the peptide backbone. At the same time, the size of the prototype model influences the level of the quantum chemical calculation. Parallelly, studying model systems of aromatic−aromatic interactions in proteins by quantum chemistry requires the proper treatment of the dispersion energy, which necessarily implies the usage of high-level quantum chemical methods. A satisfactory solution would be then to follow a protocol of calculation that could combine both requirements. In this respect we have shown that geometries optimized at any of the levels of theory here employed—except from B3LYP/6-31G* and TPSS/6-311++G(3df,3pd)—are similar to the RI-MP2/cc-pVTZ geometries (here considered as the benchmark). Particularly, geometries obtained at the SCC-DF-TB-D level of theory are recommended as input geometries for the energy calculations, specially when larger systems are to be calculated.

Energy calculations should never be done using a standard DFT functional that has not been augmented by dispersion or has not been specifically parametrized to cover dispersion energy. These methods also fail in the prediction of the global minimum in the PES. If the size of the system studied is large, then SCC-DF-TB-D or any of the Truhlar functional's here tested should be enough. Otherwise, the final and reliable order of the multiple conformers existing on the potential energy surface of any peptide should be obtained from high-level quantum chemical methods, particularly SCS-MP2 or SCS(MI)-MP2. A necessary condition, which simultaneously deals with the intramolecular basis set superposition error, is the extrapolation of the basis set to the complete basis set limit. Special attention should be paid when selecting the structures according to a specific interval, since TPSS/6-311++G(3df,3pd), B3LYP/6-311++G-(3df,3pd), PBE-D/TZVP, f99 data, and TPSS-D/6-311++G-(3df,3pd) give larger intervals than the remaining methods.

Since the FGF is mainly stabilized by $\pi-\pi$ and H-bond intramolecular interactions, comparing the data obtained at different levels of theory with the benchmark data implicitly tests which method is capable of providing a balanced and accurate treatment of these intramolecular interactions. We have shown that the TPSS-D/6-311++G(3df,3pd), SCC-DF-TB-D, PBE-D/TZVP, and BH&H/6-311++G(d,p), in this order, overstabilize those conformers having an OH$\cdots$O=C intramolecular H-bond.

All the above-mentioned conclusions are in agreement with those obtained after a similar study performed on isolated peptides as model systems of NH$\cdots\pi$ interaction in proteins. This implies that we can combine the results obtained from these two reductionist approaches to obtain a more general overview on the noncovalent interactions occurring in the hydrophobic core of a protein.

FGF is a realistic model system of aromatic−aromatic side chain interactions of adjacent residues along a protein sequence, and consequently, plenty of biological information can be obtained from its study and further used to shed some light on the protein folding process. From the very many structural aspects that can be analyzed, we have focused on the preferred orientation of the aromatic side chains with respect to each other. We have shown that neither the T-shaped nor the stacked orientations are favored. Indeed, for the vast majority of conformers, aromatic side chains adopt a geometrical disposition that favors the maximum number of noncovalent intramolecular interactions.

Benchmark data have been included in the Benchmark Energy & Geometry Database (BEGDB) (http://www.begdb.com/), which aims to provide benchmarks for the testing of many other methods.

Acknowledgment. This work was a part of the research project No. Z40550506 of the Institute of Organic Chemistry

and Biochemistry, Academy of Sciences of the Czech Republic, and it was supported by Grants No. LC512 and MSM6198959216 from the Ministry of Education, Youth and Sports of the Czech Republic. The support of Praemium Academiae, Academy of Sciences of the Czech Republic, awarded to P.H. in 2007 is also acknowledged. H.V. acknowledges the support of the government of Principado de Asturias under the program Plan de Ciencia, Tecnología e Innovación (PCTI) 2006−2009. We thank K. Berka for his help in evaluating the percentage of aromatic side chains that are likely involved in $\pi-\pi$ intramolecular interactions with neighbour aromatic side chains. A portion of the research described in this paper was performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

Supporting Information Available: Cartesian coordinates of all the structures considered in the set, time scale for the different computational methods employed in the study (Table S1), relative energies calculated at different levels of theory (Table S2), and relative Gibbs energies (Table S3). This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Murray, R. F.; Harper, H. W.; Granner, D. K.; Mayes, P. A.; Rodwell, V. W. *In Harper's Illustrated Biochemistry*, 27th ed.; Lange Medical Books/McGraw-Hill: New York, 2006; p 30.

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

(3) Valdes, H.; Pluhackova, K.; Pitonak, M.; Rezac, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2747.

(4) Valdes, H.; Spiwok, V.; Rezac, J.; Reha, D.; Abo-Riziq, A. G.; de Vries, M. S.; Hobza, P. *Chem.—Eur. J.* **2008**, *14*, 4886.

(5) Reha, D.; Valdes, H.; Vondrasek, J.; Hobza, P.; Abo-Riziq, A. G.; Crews, B.; de Vries, M. S. *Chem.—Eur. J.* **2005**, *11* (23), 6803.

(6) Valdes, H.; Reha, D.; Hobza, P. *J. Phys. Chem. B* **2006**, *110* (12), 6385.

(7) Hunter, C. A.; Lawson, K. R.; Perkins, J.; Urch, C. *J. Chem. Soc., Perkin Trans.* **2001**, *2*, 651.

(8) Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10656.

(9) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem., Int. Ed.* **2003**, *42* (11), 1210.

(10) Grimme, S. *J. Comput. Chem.* **2004**, *25* (12), 1463.

(11) Grimme, S. *J. Comput. Chem.* **2006**, *27* (15), 1787.

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

(13) Distasio, R. A., Jr.; Head-Gordon, M. *Mol. Phys.* **2007**, *1058*, 1073.

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

(15) Jung, Y. S.; Lochan, R. C.; Dutoi, A. D.; Head-Gordon, M. *J. Chem. Phys.* **2004**, *121*, 9793.

(16) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.

(17) Rodham, D. A.; Suzuki, S.; Suenram, R. D.; Lovas, F. J.; Dasgupta, S.; Goddard, W. A.; Blake, G. A. *Nature* **1993**, *362*, 735.

(18) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2000**, *122* (46), 11450.

(19) Chipot, C.; Jaffe, R.; Maigret, B.; Pearlman, D. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1996**, *118* (45), 11217.

(20) Holroyd, L. F.; van Mourik, T. *Chem. Phys. Lett.* **2007**, *442* (1−3), 42.

(21) Valdes, H.; Klusak, V.; Pitonak, M.; Exner, O.; Stary, I.; Hobza, P.; Rulisek, L. *J. Comput. Chem.* **2008**, *29*, 861.

(22) van Mourik, T.; Karamertzanis, P. G.; Price, S. L. *J. Phys. Chem. A* **2006**, *110* (1), 8.

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

(24) Bakker, J. M.; Plutzer, C.; Hunig, I.; Haber, T.; Compagnon, I.; von Helden, G.; Meijer, G.; Kleinermanns, K. *Chem. Phys. Chem* **2005**, *6* (1), 120.

(25) Chass, G. A.; Mirasol, R. S.; Setiadi, D. H.; Tang, T. H.; Chin, W.; Mons, M.; Dimicoli, I.; Dognon, J. P.; Viskolcz, B.; Lovas, S.; Penke, B.; Csizmadia, I. G. *J. Phys. Chem. A* **2005**, *109* (24), 5289.

(26) Fricke, H.; Funk, A.; Schrader, T.; Gerhards, M. *Phys. Chem. Chem. Phys.* **2007**, *9* (32), 4592.

(27) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21* (12), 1049.

(28) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97* (40), 10269.

(29) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114* (12), 5149.

(30) Becke, A. D. *Phys. Rev. A* **1988**, *38* (6), 3098.

(31) (a) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650. (b) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P.; von, R. *J. Comput. Chem.* **1983**, *4*, 294. (c) Gill, P. M. W.; Johnson, B. G.; Pople, J. A.; Frisch, M. J. *Chem. Phys. Lett.* **1992**, *197*, 499. (d) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.

(32) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(33) Easton, R. E.; Giesen, D. J.; Welch, A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1996**, *93* (5), 281.

(34) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101/1−18.

(35) Schafer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(36) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37* (2), 785.

(37) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78* (7), 1396.

(38) Eichkorn, K.; Treutler, O.; Ohm, H.; Haser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240* (4), 283.

(39) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97* (1−4), 119.

(40) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96* (9), 6796.

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

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

(43) Pitonak, M.; Janowski, T.; Neogrady, P.; Pulay, P.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1761.

(44) Rezac, J.; Jurecka, P.; Riley, K. E.; Cerny, J.; Valdes, H.; Pluhackova, K.; Berka, K.; Rezac, T.; Pitonak, M.; Vondrasek, J.; Hobza, P. *Collect. Czech. Chem. Commun.* **2008**, *73* (10), 1261.

(45) Beran, G. J. O.; Head-Gordon, M.; Gwaltney, S. R. *J. Chem. Phys.* **2006**, *124*, 114107.

(46) Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Phys. Chem.* **1996**, *100* (48), 18790.

(47) Chelli, R.; Gervasio, F. L.; Procacci, P.; Schettino, V. *J. Am. Chem. Soc.* **2002**, *124*, 6133.

(48) Gloaguen, E.; Valdes, H.; Pagliarulo, F.; Pollet, R.; Tardivel, B.; Hobza, P.; Piuzzi, F.; Mons, M. (Personal communication).

CT900174F