UNIVERZITA KARLOVA

Přírodovědecká fakulta

RNDr. Jiří Vondrášek, CSc.

INTRAMOLECULAR STABILIZATION OF PROTEINS BY MEANS OF COMPUTATIONAL METHODS

INTRAMOLEKULÁRNÍ STABILIZACE PROTEINŮ POHLEDEM VÝPOČETNÍCH METOD

Habilitační práce

Praha, 2018

Prohlášení

Prohlašuji, že jsem tuto habilitační práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu.

V Praze dne 1. května 2018.

Acknowledgements

The proposed results have not been achieved by myself alone, but they are also work of my students and colleagues, who have greatly contributed to the work on the projects represented by the attached peer-reviewed papers published in scientific journals.

I would like to acknowledge the never-ending support from my dear senior colleagues Pavel Hobza, Pavel Jungwirth, Petr Nachtigall, Ota Bludský, Michael Mareš, Vladimír Sklenář, Petter Lyngstaadas, Valerie Daggett and Michael Oliveberg. I am thankful to my junior colleagues Jiří Černý, Tomáš Kubař, Martin Horn, Kvido Stříšovský, Kristýna Boušová, and to all of my students that later became my collaborators: Lada Biedermannová, Karel Berka, Jiří Vymětal, Boris Fačkovec, Jan Heyda, and to all of my current and former students, namely David Jakubec, Martin Sarker and Veronika Jurásková.

My research would not be possible without financial support from the grant agencies Czech Science Foundation, MYES, European Commission, Grant Agency of the CAS – they all deserve my sincere thanks.

I thank all the members of my great family for support and their critical minds, which are the most important elements of any achievement.

Contents

Contents 1

Introduction 2

1 The stability of the hydrophobic core in globular proteins – the origin of the stability 4

1.1 The case of rubredoxin and the analysis of its stabilizing interactions 4

1.2 The motifs of stability – the enrichment of the amino-acid interaction repertoire 12

1.3 Charged amino acids and their role in protein stability 16

2 Comprehensive analysis of amino-acid interactions in proteins 21

2.1 A comparison of ab initio and empirical methods for side-chain interactions 21

2.2 The decomposition of intramolecular interactions 24

3 Practical applications 28

3.1 Amino Acid Interaction web server 28

3.2 Theoretical protein design 30

4 Conclusions 34

References 35

List of publications comprising the habilitation thesis

Appendices: Papers 1–16

Introduction

Proteins are essential molecules of life. They are present in various parts of any organism. They have a large number of functions, which is possible because of the enormous repertoire of their properties dictated by their sequence. As postulated by Afinsen[1], if a protein has a structure, then it is fully defined by the sequence of amino acids and the structure is in the global Gibbs free energy minimum. The classification of proteins is usually based on their cellular localization and structural-chemical properties[2–4]. It is possible to identify about four major protein classes: i) globular and soluble proteins; ii) fibrillar proteins, which are important for tissue structures; iii) membrane proteins; and finally iv) intrinsically disordered proteins. For the sake of clarity, this work will only focus on the globular, one-domain and soluble proteins. Nevertheless, the conclusions that can be drawn from our calculations are rather general, because our approach is based on the physical-chemical principles applicable on any biomolecule.

The structure and function of proteins are not realized via a vast number of completely different protein architectures. Instead, a modular arrangement of some protein subunits, called domains, into a functional molecule is largely utilized in nature[5–7]. The concept of protein domains, which is one of the paradigms of molecular biology, seems to be well understood and heavily utilized during evolution. A particular protein domain can be identified and assigned to specific functional or structural properties. Despite the fact that there is a high but limited number of protein domains, they appear in countless combinations[8,9]. Therefore, the modular organization of proteins is a plausible evolutionary concept explored and proven by bioinformatics methods[6], resulting in specific protein domain databases, such as SMART[10], PFam[11] and InterPro[12]. The advantage of the modular composition of proteins is a much higher number of their functional combinations at low energy cost in comparison with the evolution of a specific new function.

Globular proteins can be characterized by their intramolecular stability. Stabilization energy is defined as the energy needed for protein denaturation – the difference of its native and denatured states – and it is an essential characteristic of any protein. Protein stability is directly related to the process called protein folding, in which the polymeric chain of amino acids realizes its three-dimensional structure[13]. The protein-folding problem has been around for more than 50 years and it is still exciting as well as frustrating[14,15]. Over the years, a large body of knowledge on protein folding has been accumulated, but it is still impossible to predict a native protein structure from amino-acid sequences if there is no homology or the protein is too big, as a result of which the computational approach originating from first principles cannot be applied due to the enormous amount of time necessary for the sampling of the conformational space[16].

Protein function is directly linked to protein flexibility, and any interaction between a protein and another molecule requires the protein to be able to change its conformation[17]. This conformational change may be very small, involving only the rearrangement of a few amino-acid side chains, or it may be large and even may involve the folding of the entire protein. Potentially, a perturbation that changes the flexibility of a protein may interfere with its function. How the function, flexibility and stability are connected is still heavily debated. Not only is a deeper understanding of the interplay between these properties of basic scientific interest, but it will also have implications for protein design and applied protein science.

The most important level of our understanding of protein stability as well as its interactions with other molecules is based on a physically correct description of non-covalent interactions between protein building blocks – amino acids. It is well known that in most of the processes involving a protein, non-covalent interactions play a crucial role and they are realized by interactions of protein backbone atoms as well as by atoms in amino-acid side chains. There is a gap in our ability to accurately evaluate the contributions of enthalpy and entropy to the stabilization Gibbs free energy realized via non-covalent interactions of amino acids composing a protein[1819,20]. On the other hand, the recent development of theoretical chemistry methods as well as the enormous amount of computational resources reflecting advances in computer sciences make it possible to evaluate enthalpy contributions to the total Gibbs free energy at a high level of reliability. Accurate interaction energies can be obtained by complete basis set limit calculations provided that a large portion of correlation energy is covered (e.g. by performing CCSD(T) calculations). The description at the highest theoretical level is still limited to hundreds of atoms. Nevertheless, the use of DFT or semi-empirical methods makes it possible to calculate non-covalent interaction energy for systems of thousands of atoms at a high level of confidence.

The proposed habilitation thesis demonstrates how advanced theoretical chemistry methods could be utilized to establish a solid of level of confidence for less advanced or even semi-empirical and empirical methods in interaction energy evaluation. These quantum chemistry benchmarks are important steps in our realistic descriptions of biomolecular systems and their internal stability realized via non-covalent interactions of their building blocks[21,22].

As demonstrated on the following pages, our understanding of protein stability benefits enormously from such studies.

1. The stability of the hydrophobic core in globular proteins – the origin of the stability

Our interest in understanding protein structure and stability dates back to 1960, when the first high-resolution protein structures were determined by x-ray crystallography. It is a commonly accepted fact that a large contribution to protein stability comes from the hydrophobic residues which are condensed by the effect of entropy in the protein hydrophobic core[23]. This phenomenon also raised fundamental questions regarding the arrangement of hydrophobic residues in the core, a suitable model of their packing and the allowed side-chain conformation within this motif. The role of specific residues in the balance between function, stability and folding rates could be determined. There is a pool of mutational studies shedding light on the role of a particular residue; in addition, a great deal can be learned from a comparison of the sequences of structures having identical folds but low sequence identity[24]. Therefore, the packing of residues in the hydrophobic core appears to be extremely important for the structure, stability, and native-like properties of natural proteins[25–27], and more and more is known not only about principles but also about details at atomistic level[20,28].

The best model systems to study protein stability appear to include proteins with different temperature conditions to reach the maximum stability. There is also a body of evidence to identify what makes proteins exist under physiological conditions and temperature and what makes them behave as mesophilic or hyperthermophilic.[29,30] The following chapters provide a detailed analysis of the non-covalent interactions involved in the packing of the hydrophobic cores of the hyperthermophilic protein rubredoxin[31]. We have used a range of computational methods to identify the stability originating in the close packing of the hydrophobic core residues as well as to find new motifs of interactions at the amino-acid level remarkably contributing to the total stabilization energy.

1.1 The hydrophobic core of rubredoxin and its stabilizing interactions

Rubredoxin is a typical globular one-domain protein, containing a densely packed cluster of interacting residues centered around two phenylalanines (F30 and F49) in the interior of the protein (Figure 1A,B). Since water molecules are not present in the core, water is not directly involved in the core stabilization. The whole cluster was partitioned into two distinct clusters (named after the central residues, F30 and F49) and was further fragmented into well-defined, chemically distinct pairs of neutral amino acids (modeled as methylated aminoacyl residues). The central F30 and F49 phenylalanines thus interact with five (F49, K46, L33, Y13 and Y4) and seven (C39, C6, F30, K46, V5, W37 and Y4) amino acids, respectively. There is one H-bond ascribed to the F30 cluster (a classical CO…HN H-bond in the F30…L33 pair), and another two H-bonds are ascribed to the F49 cluster (a classical CO…HN H-bond in the F49…K46 pair, as well as an unusual CH…π interaction between the methyl group of the capped O terminus of V5 and the π system of the phenylalanine in the F49…V5 pair; cf. Figure 1B,C).

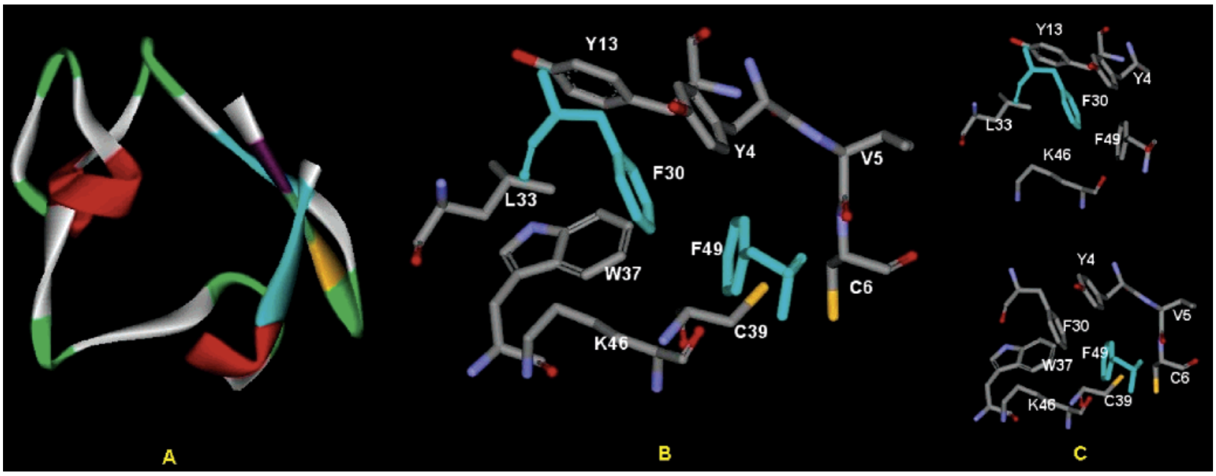


Figure 1. Rubredoxin. (A) Schematic view of the protein; (B) supercluster of F30 and F49; and (C) both subclusters individually.

The total stabilization energy of both clusters was determined as the sum of the pairwise stabilization energies of a central phenylalanine with the amino acids in its neighborhood. These energies were first determined at the frequently used DFT/ B3LYP/6-31G** level. Figure 2 shows that eleven out of twelve DFT pair interaction energies are repulsive and the twelfth one is only slightly attractive. The DFT picture is thus consistent with the expected nature of interactions in a hydrophobic core with low occurrence of hydrogen bonds. All pair interactions are either repulsive or negligible. However, is this conclusion correct? It is evident in Figure 1 that the aromatic rings of the central phenylalanines are in contact with the aromatic and aliphatic side chains of the neighboring amino acids. These contacts should be stabilized by London dispersion energy. Therefore, the calculations should be performed at the highest possible level, excluding the traditional problems of ab-initio quantum chemical calculations, specifically the incompleteness of the AO basis set and the insufficient amount of correlation energy covered.

An inspection of the RIMP2/CBS interaction energies (the lower part of Figure 2 and Table 1) has provided a very surprising picture. All twelve pairs of interaction energies are negative (i.e. stabilizing) and the stabilization energies are relatively high (for six pairs even higher than 4.5 kcal/mol at the CBS limit). What is especially important are the F30…Y4 and F49…V5 pairwise interactions with stabilization energies of about 7 kcal/mol. The first pair is stabilized by the interaction of the two aromatic rings, and the structure corresponds to a parallel-displaced structure of a benzene dimer. The F49…V5 interaction is of a different nature. Due to the fragmentation procedure, the pair contains a CH…π contact instead of the π...π contact present in the real system (the interaction of the π electrons of phenylalanine and a peptide bond). The F30…Y4 and F49…V5 pairs clearly illustrate the stabilization role of the amino-acid aromatic ring and show that strong stabilization (comparable to or even higher than H-bonding) can originate from dispersion attraction without the presence of any classical H-bond.

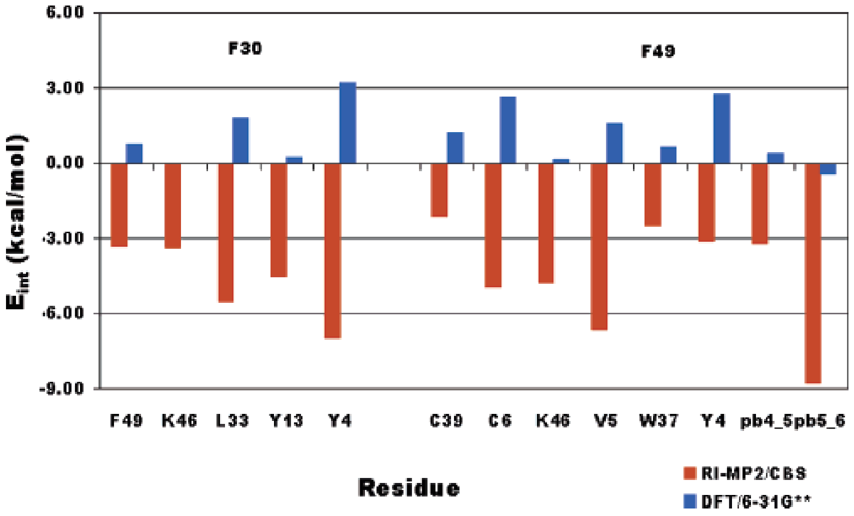


Figure 2. DFT and MP2/CBS interaction energies of F30 and F49 phenylalanines with selected amino acids from the rubredoxin core; the DFT interaction energy of the F30…K46 pair is 0.

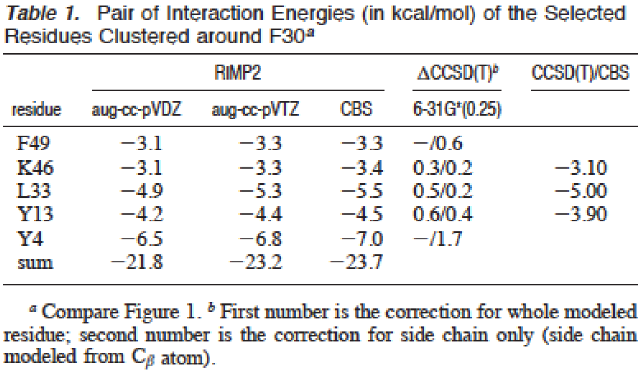


Table 1. Pair of Interaction Energies (in kcal/mol) of the Selected Residues Clustered around F30[a]

	RIMP2			ΔCCSD(T)[b]	CCSD(T)/CBS
residue	aug-cc-pVDZ	aug-cc-pVTZ	CBS	6-31G*(0.25)	
F49	−3.1	−3.3	−3.3	−/0.6	
K46	−3.1	−3.3	−3.4	0.3/0.2	−3.10
L33	−4.9	−5.3	−5.5	0.5/0.2	−5.00
Y13	−4.2	−4.4	−4.5	0.6/0.4	−3.90
Y4	−6.5	−6.8	−7.0	−/1.7	
sum	−21.8	−23.2	−23.7		

[a] Compare Figure 1. [b] First number is the correction for whole modeled residue; second number is the correction for side chain only (side chain modeled from C_β atom).

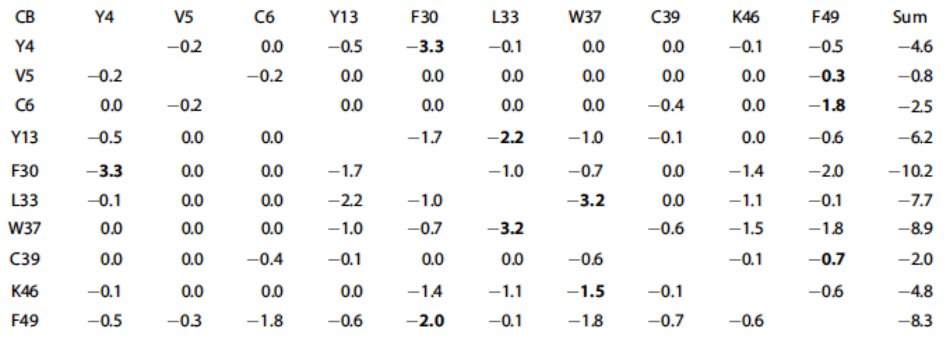
The present results show a complete failure of the DFT calculations, which are not even able to describe the attraction between central phenylalanines and the neighboring amino acids.

The results also fully support the known, but commonly ignored, fact that DFT methods cannot be recommended for simulating systems where London dispersion interactions play a major role and clearly demonstrate further the substantial attraction inside a hydrophobic core. This attraction, originating in London dispersion energy between aromatic rings or between an aromatic ring and an aliphatic chain, is comparable to classical H-bonding. Moreover, residues of aromatic nature can participate in several strong interactions at once, which may be crucial for the role of key residues in the establishment of small-world networks inside a protein[28]. Consequently, the current view on the nature of secondary and tertiary protein structure stabilization and, especially, the origin and nature of protein folding should thus be modified. The hydrophobic nature of a protein core implies that hydrophobic interactions can initiate the folding process. The present results indicate a decisive role of stabilization energy (enthalpy). Possible consequences which means that a significant role during the early stage of protein folding may rather be played by the energy (enthalpy) than hydrophobicity (entropy).

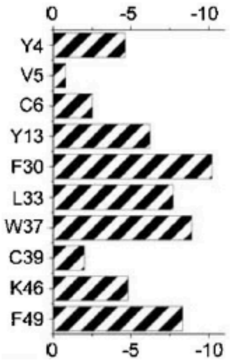
There is a question of how much the identification of the hydrophobic core in rubredoxin and its stabilizing interactions correspond to the balance between different energy terms composing the core. Therefore, we have performed a theoretical analysis of the interaction energy between the amino acids composing the rubredoxin core based on the symmetry-adapted perturbation (SAPT) method. A reliable decomposition of a range of energy terms is feasible only at a reasonable quantum chemical level, and the question of the origin of the stabilizing forces inside the hydrophobic core of the protein rubredoxin could be addressed and these methods make it possible. Different computational procedures allow for the decomposition of the total interaction energy into its energy components. Partial decomposition is also possible through DFT+D techniques[22] or local correlation methods. For the sake of comparison with other stabilizing factors in proteins, we have also performed energy decomposition for the typical hydrogen-bonded structures maintained by the backbone–backbone interaction in Rd.

The total interaction energies and the interaction energies for each pair of interacting side chains determined with the DFT-SAPT method (denoted as CB) are listed in Table 2. The last column is the sum of all the stabilizing contributions for one particular amino acid coming from the interactions of this residue with all the amino-acid residues composing the core. The result shows that the residues with the highest interaction energies are phenylalanine F30 and tryptophan W37. Considerable stabilizing contributions also come from another phenylalanine residue, namely F49, which is also an important side chain in the hydrophobic core of the protein, and the leucine residue in position 33 (L33). It is noteworthy that all contributions (not only the final sums for particular amino acids but also all the interactions maintained by the residue in question) are attractive. The values presented in Table 2 are the upper limits of the gas-phase interaction energy within our model system of the hydrophobic core. The real (gas-phase) stabilization energies will be systematically more negative. The fact that each of the interacting pairs exhibits attraction is not trivial. One would expect the structure of a protein core primarily determined by the entropy-driven hydrophobic effect during hydrophobic collapse not to be energetically at the optimum.

Table 2. The DFT-SAPT interaction-energy matrix for the residue pairs modeled without a backbone (CB) inside the hydrophobic core (Figure 1). All energies are given in kcalmol^{-1}.

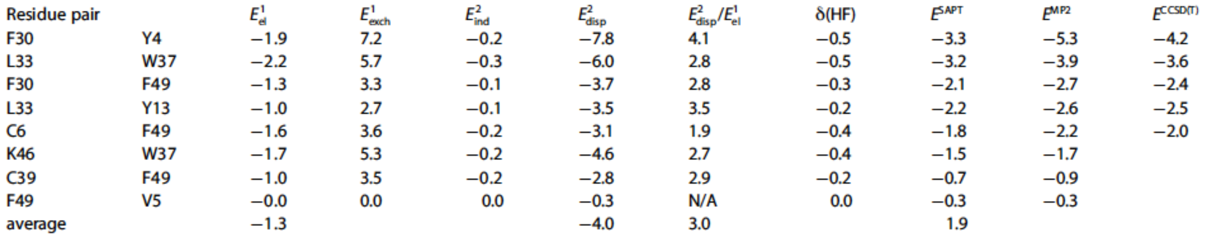


CB	Y4	V5	C6	Y13	F30	L33	W37	C39	K46	F49	Sum
Y4		−0.2	0.0	−0.5	**−3.3**	−0.1	0.0	0.0	−0.1	−0.5	−4.6
V5	−0.2		−0.2	0.0	0.0	0.0	0.0	0.0	0.0	**−0.3**	−0.8
C6	0.0	−0.2		0.0	0.0	0.0	0.0	−0.4	0.0	**−1.8**	−2.5
Y13	−0.5	0.0	0.0		−1.7	**−2.2**	−1.0	−0.1	0.0	−0.6	−6.2
F30	**−3.3**	0.0	0.0	−1.7		−1.0	−0.7	0.0	−1.4	−2.0	−10.2
L33	−0.1	0.0	0.0	−2.2	−1.0		**−3.2**	0.0	−1.1	−0.1	−7.7
W37	0.0	0.0	0.0	−1.0	−0.7	**−3.2**		−0.6	−1.5	−1.8	−8.9
C39	0.0	0.0	−0.4	−0.1	0.0	0.0	−0.6		−0.1	**−0.7**	−2.0
K46	−0.1	0.0	0.0	0.0	−1.4	−1.1	**−1.5**	−0.1		−0.6	−4.8
F49	−0.5	−0.3	−1.8	−0.6	**−2.0**	−0.1	−1.8	−0.7	−0.6		−8.3



The question is, however, what the most stabilizing energy term for these interactions is and whether it is uniform or differs across diverse amino-acid residues. To answer this question, we have performed a DFT-SAPT energy decomposition for each pair of interacting residues. Table 3 presents the DFT-SAPT analysis for the particular pair with the highest stabilization energy of all the pairs of the interacting amino-acid residues. The reference energies determined at the supermolecular MP2/CBS CCSD(T)-corrected level were always larger than those obtained by the DFT-SAPT methods but lower than the MP2/ aug-cc-pVDZ values. This reflects the known fact that MP2 stabilization energies determined with extended basis sets or even at the CBS level are overestimated. This overestimation is removed when passing from the MP2 to the CCSD(T) level.

Table 3. The energy decomposition of the most stabilizing interaction for each residue



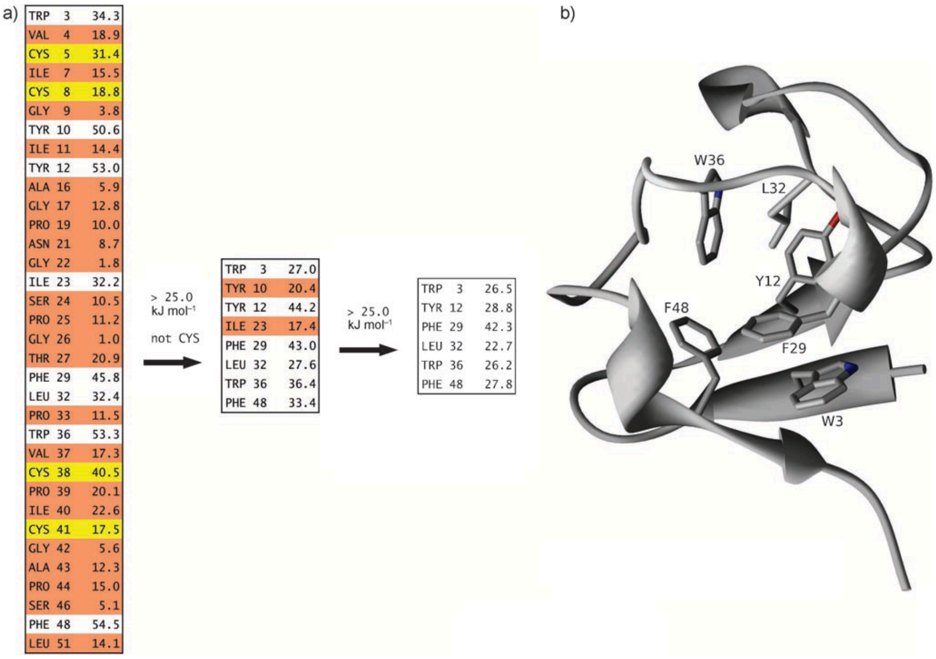
Residue pair		E^1_{el}	E^1_{exch}	E^2_{ind}	E^2_{disp}	E^2_{disp}/E^1_{el}	δ(HF)	E^{SAPT}	E^{MP2}	$E^{CCSD(T)}$
F30	Y4	−1.9	7.2	−0.2	−7.8	4.1	−0.5	−3.3	−5.3	−4.2
L33	W37	−2.2	5.7	−0.3	−6.0	2.8	−0.5	−3.2	−3.9	−3.6
F30	F49	−1.3	3.3	−0.1	−3.7	2.8	−0.3	−2.1	−2.7	−2.4
L33	Y13	−1.0	2.7	−0.1	−3.5	3.5	−0.2	−2.2	−2.6	−2.5
C6	F49	−1.6	3.6	−0.2	−3.1	1.9	−0.4	−1.8	−2.2	−2.0
K46	W37	−1.7	5.3	−0.2	−4.6	2.7	−0.4	−1.5	−1.7	
C39	F49	−1.0	3.5	−0.2	−2.8	2.9	−0.2	−0.7	−0.9	
F49	V5	−0.0	0.0	0.0	−0.3	N/A	0.0	−0.3	−0.3	
average		−1.3			−4.0	3.0		1.9		

All energies are in kcalmol^{-1}. The DFT-SAPT interaction energy E^{SAPT} consists of electrostatic (E^1_{el}), exchange (E^1_{exch}), induction summed with exchange induction (E^2_{ind}), dispersion summed with exchange dispersion (E^2_{disp}), and higher-order term estimations [δ(HF)]. The ratio between E^2_{disp} and E^1_{el} is also shown for the comparison of the relative strength of dispersion and electrostatic stabilization. The SAPT energy is always at the lower limit of the interaction energy given by MP2 in the aug-cc-pVDZ basis set (E^{MP2}) or the benchmark CCSD(T) method ($E^{CCSD(T)}$). Note that the dispersion energy is usually more than twice the electrostatic energy.

The above results show that the structural arrangement of the amino-acid residues forming the hydrophobic core of the protein rubredoxin is mostly maintained by dispersion interactions, unlike hydrogen bonds in secondary- structure elements within the same protein, which are stabilized mostly by electrostatic interactions. Considering all the pairwise interactions within rubredoxin as a whole and taking the environment into account, we have found dispersion energy to be the dominant stabilizing factor in the folded protein structure. It can be expected that the same conclusion is also valid for other proteins of globular nature or protein globular domains.

Further, we addressed fundamental questions as to what the reason for the extreme stability of the protein rubredoxin from *Pyrrococus Furiosus* (Pf Rd) is, how it can be elucidated from a complex set of interatomic interactions and whether it is located in the hydrophobic core of the protein. In order to determine the melting temperature of both wild types as well as a mutant variant of Rd by microcalorimetry measurements, we have combined two approaches: i) computational analysis of the protein and its mutants including the calculation of Gibbs free energy and ii) biophysical experiment.

Theoretical approach was based on the concept of the “interaction-energy matrix” combined with Gibbs free energy calculations by molecular dynamics. The energy matrix helped evaluate an energetic contribution of the hydrophobic core and its most important residues. The interaction-energy matrix was constructed based on pair interaction energy values between all residues within the set. The values were then summed for each row of the matrix to yield the interaction energy of a single amino acid with the others in the set. The resulting list contained three groups of residues (cf. Figure 3): 1) Twenty-two had small total stabilization energy (below 25 kJmol^{-1}; shown in red); these were eliminated from further consideration. The limit of 25 kJmol^{-1} was selected as the strength of an average hydrogen bond; for example, residues which meet the criterion should also possess at least one strong hydrogen bond or more interactions of comparable strength.



a)

TRP	3	34.3
VAL	4	18.9
CYS	5	31.4
ILE	7	15.5
CYS	8	18.8
GLY	9	3.8
TYR	10	50.6
ILE	11	14.4
TYR	12	53.0
ALA	16	5.9
GLY	17	12.8
PRO	19	10.0
ASN	21	8.7
GLY	22	1.8
ILE	23	32.2
SER	24	10.5
PRO	25	11.2
GLY	26	1.0
THR	27	20.9
PHE	29	45.8
LEU	32	32.4
PRO	33	11.5
TRP	36	53.3
VAL	37	17.3
CYS	38	40.5
PRO	39	20.1
ILE	40	22.6
CYS	41	17.5
GLY	42	5.6
ALA	43	12.3
PRO	44	15.0
SER	46	5.1
PHE	48	54.5
LEU	51	14.1

> 25.0 kJ mol^{-1}

not CYS

TRP	3	27.0
TYR	10	20.4
TYR	12	44.2
ILE	23	17.4
PHE	29	43.0
LEU	32	27.6
TRP	36	36.4
PHE	48	33.4

> 25.0 kJ mol^{-1}

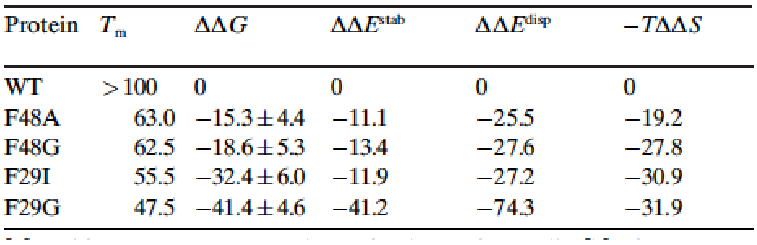
TRP	3	26.5
TYR	12	28.8
PHE	29	42.3
LEU	32	22.7
TRP	36	26.2
PHE	48	27.8

Figure 3. The location of the hydrophobic core (a, left) by the interaction-energy matrix procedure and its position in the protein molecule (b, right).

The change of melting temperature in the mutant relative to the wild type (WT) is obviously related to a change in protein stability, and this phenomenon can be quantified by the change in unfolding Gibbs energy. Because of the small difference in the 3D structure of the Pf Rd mutants indicated by the NMR spectra, we can approach the unfolding Gibbs energy computationally with the molecular dynamics–thermodynamic integration.

The result of the study is depicted in Table 4 and Figure 4. Table 4 shows a comparison of the Tm and computationally determined differences of Gibbs free energy for all constructs.

Table 4. The measured and calculated characteristics of Pf Rd and the mutants (ordered by decreasing thermal stability).



Protein	T_m	$\Delta\Delta G$	$\Delta\Delta E^{stab}$	$\Delta\Delta E^{disp}$	$-T\Delta\Delta S$
WT	>100	0	0	0	0
F48A	63.0	−15.3 ± 4.4	−11.1	−25.5	−19.2
F48G	62.5	−18.6 ± 5.3	−13.4	−27.6	−27.8
F29I	55.5	−32.4 ± 6.0	−11.9	−27.2	−30.9
F29G	47.5	−41.4 ± 4.6	−41.2	−74.3	−31.9

Figure 4 shows the thermal denaturation of Pf Rd and its mutants (a) and the calculated thermodynamic parameters (b).

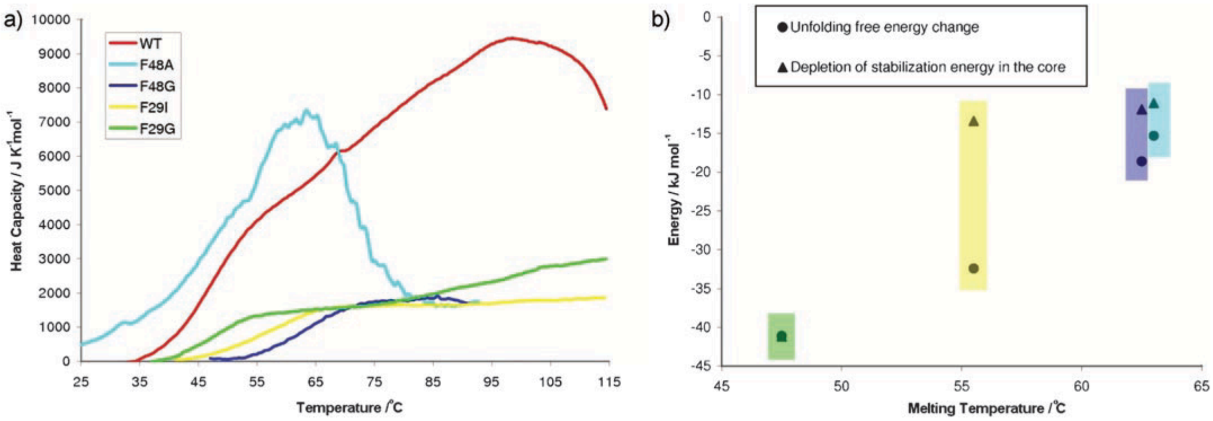


Figure 4. The thermal denaturation of Pf Rd and its mutants (a) and the calculated thermodynamic parameters (b).

The hydrophobic core and its unique spatial arrangement is the part of Pf Rd that notably contributes to its unusual thermal stability. This fact is supported by experimental as well as theoretical results. The relative unfolding Gibbs energy values obtained by the MD method agree with the course of thermal denaturation of the mutant proteins with respect to the WT version. Moreover, it has been possible to trace the overall stability reflected in the relative stabilization energy of the core, which also agrees with the melting temperature of the proteins studied. This results from the weakening of the interactions between the amino-acid side chains composing the hydrophobic core in the mutants relative to WT. Major structural differences between the WT Pf Rd and its mutants are localized in the core through particular side-chain interactions. The overall structure of the molecule is retained, corroborated by the NMR spectroscopic data. We conclude that the high stability content of Pf Rd substantially results from the highly favorable interaction of amino-acid side chains inside the hydrophobic core of the protein, which originates in its entirety in London dispersion interactions. We can speculate that the loss of favorable interactions caused by a mutation inside the core is partially compensated for by a spatial rearrangement of the core, and this occurs at the cost of configurational entropy. Supported by the NMR spectroscopic experiment and the results of calculations, we have assumed that structure and stability can be a reflection of the energy content. This concept can have an impact on various protein-related issues including stability and dynamics.

1.2 The motifs of stability – the enrichment of the amino-acid interaction repertoire

Besides long-recognized forces, for example H-bonds and salt bridges, there are abundant van der Waals interactions, among which the aromatic interactions such as π-π stacking and XH-π H-bonding have also been shown to play an important role for protein structure as well as protein–ligand recognition. The first to point out the importance of aromatic interactions in proteins were Burley and Petsko[32] in a work on interaction between phenylalanine residues. The strength of the stabilization energy in Phe pairs was estimated by gas-phase calculations of benzene and toluene dimer model systems to be 1–2 kcal/mol[32]. It is worth mentioning that π-π stacking is not strictly defined as an interaction of aromatic systems; it is a more general phenomenon that includes interactions of planar systems with delocalized orbitals, such as peptide bonds. Analogically, the XH-π bonding is not limited to aromatic–aromatic interactions, because the aromatic ring can serve as an acceptor for nonaromatic H-bond donors. It has been shown by gas-phase calculations that both electrostatic and dispersion terms are important in the XH-› interaction. The directionality of the interaction is mainly controlled by the electrostatic term; however, the potential energy surface is very flat near the minimum due to the long-range dispersion term.

Therefore, we decided to study one of the elements that we had found in the hydrophobic core of the protein rubredoxin – the interaction between the aromatic ring of phenylalanine and the peptide bond in their stacking arrangement. Aromatic-ring – peptide-bond interactions (modeled as benzene and formamide, N-methylformamide and N-methylacetamide) were studied by means of advanced computational chemistry methods: second-order Möller-Plesset (MP2), coupled-cluster single- and double-excitation model [CCSD(T)], and density functional theory with dispersion (DFT-D). The geometrical preferences of these interactions as well as their interaction energy content, in both parallel and T-shaped arrangements, were investigated.

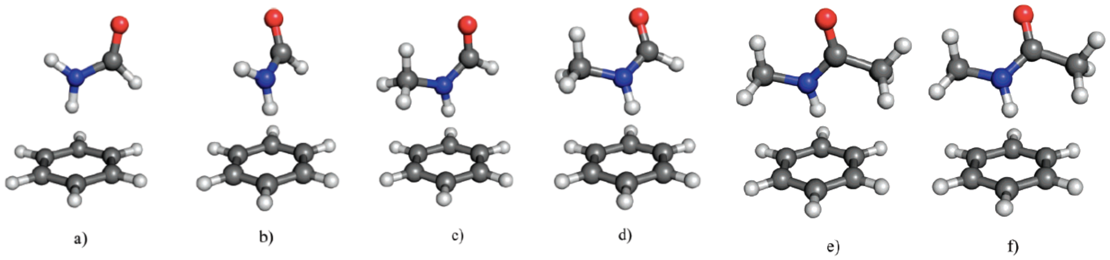
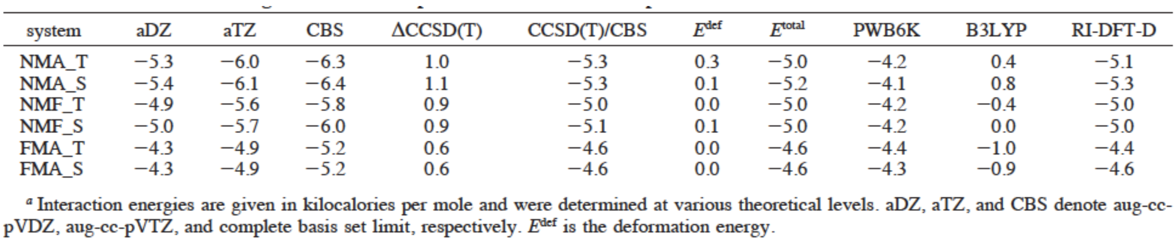


Figure 5. The geometries of optimized complexes: (a) FMA-benzene, starting from a T-shaped arrangement; (b) FMA-benzene, starting from a stacked arrangement; (c) NMF-benzene, starting from a T-shaped arrangement; (d) NMF-benzene, starting from a stacked arrangement; (e) NMA-benzene starting from a T-shaped arrangement; (f) NMA benzene, starting from a stacked arrangement.

Table 5 shows the energy characteristics of the optimized complexes. Columns 2–4 show MP2 interaction energies determined with the aug-cc-pVDZ and aug-cc-pVTZ basis sets and at the CBS limit, respectively. Passing to a larger basis set is connected with a substantial stabilization energy increase (-0.6 kcal/mol), and extrapolation to the CBS limit yields even larger stabilization energies (the highest being 6.4 kcal/mol for NMA_S…benzene).

However, it is known that the MP2/CBS stabilization energies are overestimated and that the CCSD(T) correction term should be included. This term is positive, i.e. of a repulsive character, for all systems – it is similar for both structures of NMA and NMF (-1 kcal/mol) and slightly smaller for both structures of FMA (-0.6 kcal/mol).

Table 5.



system	aDZ	aTZ	CBS	ΔCCSD(T)	CCSD(T)/CBS	E^{def}	E^{total}	PWB6K	B3LYP	RI-DFT-D
NMA_T	−5.3	−6.0	−6.3	1.0	−5.3	0.3	−5.0	−4.2	0.4	−5.1
NMA_S	−5.4	−6.1	−6.4	1.1	−5.3	0.1	−5.2	−4.1	0.8	−5.3
NMF_T	−4.9	−5.6	−5.8	0.9	−5.0	0.0	−5.0	−4.2	−0.4	−5.0
NMF_S	−5.0	−5.7	−6.0	0.9	−5.1	0.1	−5.0	−4.2	0.0	−5.0
FMA_T	−4.3	−4.9	−5.2	0.6	−4.6	0.0	−4.6	−4.4	−1.0	−4.4
FMA_S	−4.3	−4.9	−5.2	0.6	−4.6	0.0	−4.6	−4.3	−0.9	−4.6

[a] Interaction energies are given in kilocalories per mole and were determined at various theoretical levels. aDZ, aTZ, and CBS denote aug-cc-pVDZ, aug-cc-pVTZ, and complete basis set limit, respectively. E^{def} is the deformation energy.

Decomposition analysis of interaction energy in T-shaped, tilted T-shaped and parallel arrangements that exist in proteins was performed. All the arrangements exhibit comparable interaction energies in DFT-SAPT, which is an important observation in light of the fact that H-bonding (expected to be a dominant stabilization feature) exists only in the former two arrangements. Dispersion energy is a major stabilizing term in stacked as well as T-shaped structures; however, they differ significantly in the $E^{2*}_{disp}/E^{1}_{pol}$ ratio.

The stablest arrangements found in the optimizations of all three studied model systems are either T-shaped or tilted T-shaped; the energetic difference between these two arrangements is small, and no substantial barrier exists between these two minima. The interaction energies in both arrangements are large (up to -5.3 kcal/mol in NMA-benzene at the CCSD-(T)/CBS level and comparable to a classic H-bond.) Such arrangements exist in proteins and, consequently, their contribution to protein stabilization should be quite significant. The size of the system brings a large increase of interaction, which is in agreement with the result of our previous work, where the calculated interaction energy in the phenylalanine-NMF complex was -8.2 kcal/mol.

The impressive performance of RI-DFT-D is notable. RIDFT-D/LP stabilization energies agree well with the CCSD-(T)/CBS values, even though they require several orders of magnitude less CPU computational time. The MP2/aug-ccpVDZ interaction energies are also very close to the CCSD-(T)/CBS ones, probably due to the cancellation of errors. It is, however, not advisable to rely on this cancellation.

L-proline it is commonly considered to play a distinctive role in the structure and function of proteins because of the specific character of its side chain. The restraints brought by the cyclic structure of the proline side chain give this residue exceptional conformational rigidity when compared to other amino acids. Upon folding, the residue loses less conformational entropy, which may account, for example, for its higher occurrence in the proteins of thermophilic organisms. Therefore, the stabilizing role of the proline residue for a protein tertiary structure is traditionally believed to be the consequence of its extraordinary rigidity.

For the calculation of aromatic–proline interaction, we have selected two model systems, both taken from the structure of a small protein – Tryptophan cage (Trp-cage, PDB code 1L2Y), namely the pairs of residues Pro17–Trp6 and Pro18–Trp6 (Fig. 6). We have employed two models of the interacting moieties: the “large model” and the “small model”. The large model represents the studied residues as shown in Figure 7; the residue model includes the carbonyl group of the preceding residue in order to take the peptide bond into account (the protein backbone is cut at the C–Ca bond and the resulting fragments capped with hydrogen atoms). This is intended for the determination of the contribution of these polar atoms to the overall stability of a residue–residue complex and allows for a comparison with previous theoretical works on proline interactions, which also included the carboxyl group. In the small model, the system is reduced to the side-chain only, starting from the C_β atom; proline is represented as a pyrrolidine molecule to preserve its cyclic structure.

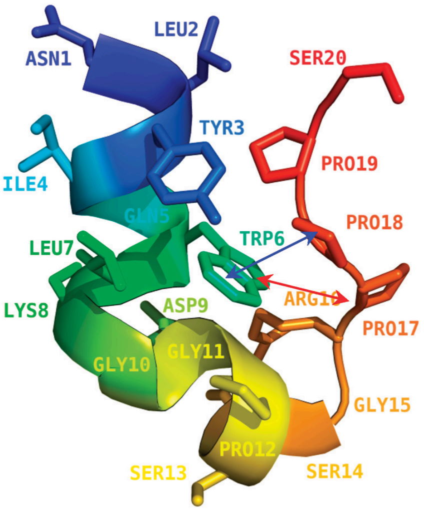


Figure 6. The geometry of the Trp-cage miniprotein. The L-shape arrangement of the interaction between TRP6 and PRO17 is represented by the double arrow red line, whereas the stacked-like arrangement of TRP6 and PRO18 is shown in the blue double arrow line.

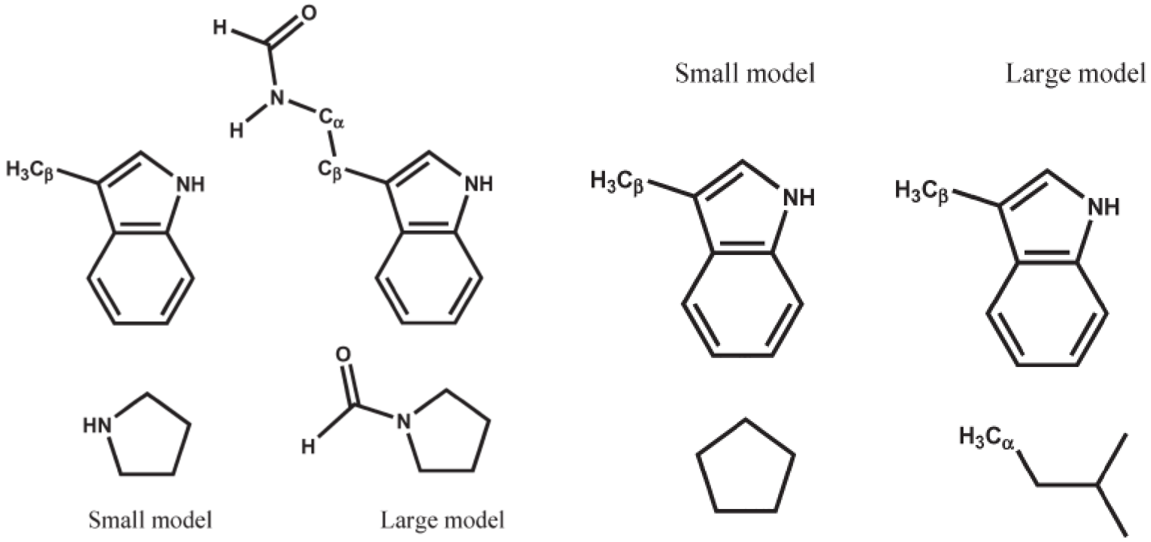


Figure 7. The chemical structure of the molecules used as model systems for the proline–aromatic interaction.

In this work, we have investigated two proline–tryptophan complexes derived from the experimental structure of the Trp–cage miniprotein, one in an “L-shaped” arrangement and the other in a “stacked-like” arrangement. The L-shaped arrangement features H-bond proline and tryptophan residues, whereas in the “stacked-like” arrangements, the residues are in a parallel geometry without any H-bond between them. We have performed correlated MP2 and DFT-D calculations of interaction energies, including benchmark CCSD(T)/CBS calculations for selected complexes, as well as a DFT-SAPT interaction energy decomposition. Our calculations have shown that the L-shaped arrangement is very strongly stabilized and the main source of stabilization is the classical H-bond, as the truncation of the system leads to a dramatic interaction energy decrease (-7.6 vs. -1.1 kcalmol^{-1} for the large and small models, respectively, at DFT-D/TPSS/ TZVP level). The most important result of this study is, however, that the stacked-like arrangement of the tryptophan–proline interaction is also bound very strongly, even without the presence of any classical H-bond, and the truncation of the system does not diminish the interaction energy as profoundly as in the previous case (-6.8 vs. -5.4 kcalmol^{-1} and -8.4 vs. -6.5 kcalmol^{-1} for the large and small models, at DFT-D and MP2 levels, respectively). The fact that the dispersion term in the DFT-D method is responsible for most of the attractive force within this complex indicates that the strong interaction found therein is principally attributable to dispersion forces. However, it should be noted that the electrostatic contribution to this interaction is not negligible and is about half as strong as that of dispersion.

1.3 Charged amino acids and their role in protein stability

Almost pure electrostatic interactions, which originate from the presence of two charged subsystems, are also present in proteins. The best known example is a salt bridge, which is an ion pair of two charged amino-acid side chains. The geometrical definition of a salt bridge from 1983[33,34] requires a distance of 4.0 Å between the charged groups of centroids and the existence of at least one pair of side-chain nitrogen and oxygen atoms within a 4.0-Å distance. This electrostatic interaction element seems to be a key factor in molecular recognition, protein–protein interaction, flexibility and thermostability[35]. It can also play a very important role in the structure and stability of proteins. The strength of the electrostatic attraction between positively and negatively charged subsystems is substantial (by an order of magnitude larger than other contributions) and nearly approaches the strength of a covalent bond. This is true, however, only in the gas phase or in salt crystals. In any other medium, dielectric screening reduces the magnitude of the charge distribution and thus the strength of the electrostatic term. An extreme case is represented by the water phase, where the electrostatic interaction is reduced dramatically by the hydration of both charged partners. This is the case of a salt bridge located at the protein surface, which is thus directly exposed to the water phase. If a salt bridge is partially buried in the protein interior, the situation might be quite different and there is no unambiguous opinion about its strength. Some analyses have shown that the electrostatic term is negligible in this case[36], while others have indicated significant stabilization[37]. This point is of key importance in the case of thermophilic proteins as the thermostability of hyperthermophilic proteins may be related to the abundance of salt bridges.

The aim of the reported study was to investigate the strength of various Glu–Lys ion pairs in the protein rubredoxin as well as their ion-neutral counterparts in which either the Glu is protonated or the Lys is deprotonated. It is known that the salt bridge of the side chains between Lys6 and Glu49 does not stabilize the hyperthermophilic rubredoxin (Pf Rd) variant. In addition to this salt bridge in a wild type and a mutant form of Pf Rd, we have also explored the Glu–Lys interaction localized between residues that are partially buried inside the protein interior. They differ in terms of the distance between the two ionic heads of the side chains. In all of these cases, we considered the dependence of the electronic energy on the dielectric constant of the protein environment as well as of the solvent. For the first time, we have utilized the concept of the total electronic energy and its variant in a continuous protein environment.

As a system for our study, we have chosen again the Pf Rd, of which various ion pairs between the Glu and Lys side chains were selected. The coordinates of all the interacting pairs were obtained from the crystal structures of the hyperthermophilic rubredoxin from Pyrrococus furiosus (pdb code: 1BRF) or its mutants (pdb codes: 1BQ9, 1IU5) and its mesophilic counterpart (pdb code: 1SMM). The amino acids forming salt bridges were excised from the protein and their N termini were set to NH2 and O termini to H–C=O, i.e. not in a zwitterionic form. It has been shown by Strop and Mayo9 that there is a side chain to the side-chain salt bridge between the Lys6 and Glu49 in hyperthermophilic rubredoxin Pf Rd, which has been found not to stabilize the protein.

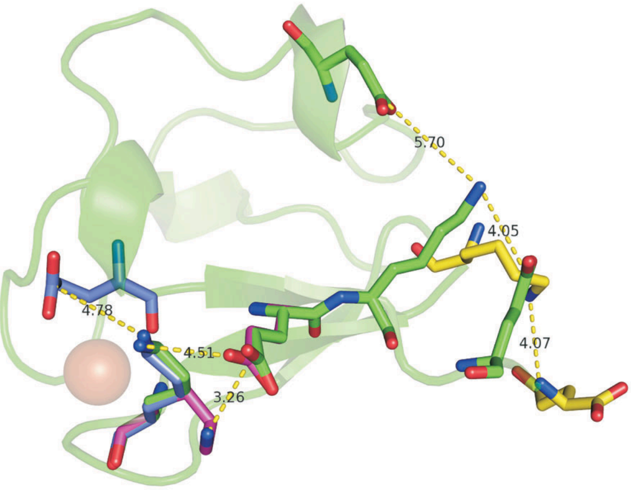


Figure 8. The structure of wild-type rubredoxin from Pyrrococus furiosus (Pf Rd) with salt bridges involved in this study. All other rubredoxin structures (1IU5, 1BQ9, 1SMM) have been aligned to the structure of the wild type (1BRF, green). Salt bridges differ in color (SB1, blue; SB2, violet; SB3, yellow) from those of the wild type (SB4–SB6, green). The distance (in Å) between the COO–carbonyl carbon and NH3 + nitrogen is shown for each salt bridge.

Based on the study performed, we could draw several conclusions.

The CCSD(T) CBS stabilization energies of the Glu-Lys salt bridges determined for the experimental geometries are very large, reaching, and in one case even exceeding, 100 kcal/mol. These values represent new benchmark data for this type of ion-pair amino acids in the gas phase.

The DFT/TPSS/TZVP interaction energies are close to the benchmark data, especially if empirical dispersion energies are included. The dispersion energies themselves are, however, rather small.

The effect of the environment on the electronic energy is of key importance. The protein environment ($\varepsilon = 4$) reduces the stabilization energy of salt bridges by 23–43% and an even larger reduction occurs when the water environment is considered, which sometimes changes large stabilization to destabilization.

The strong stabilization of the Glu–Lys salt bridge is lost upon protonation/deprotonation to an ion-neutral amino-acid pair as a consequence of the altered pH. This effect is independent of the environment. The large difference between the stabilization energies of the ion pairs and ion-neutral pairs as well as the small difference between the corresponding free energies indicate the decisive role of entropy, which should be large for the former pairs and small for the latter pairs.

Some surprises come from the analysis of the crystal structures of proteins, implying that the arrangement of charged residues does not reflect their physical and chemical properties in an expected way. In an aqueous environment, biological macromolecules are subject to a mixture of forces arising from water affinity, cavitation (solvent exclusion), dispersion, and other effects that can outweigh direct electrostatic effects. This leads to a series of "electrostatics-defying" biological structures such as anions bound to anionic protein surfaces, and the bases of DNA, the occurrence of almost completely anionic protein surfaces, and arginine–arginine pairing via positively charged guanidinium (Gdm+) groups within and between protein subunits.

By means of molecular dynamics (MD) simulations of poly-arginine, we have studied the pairing of like-charged side chains. We contrast this behavior to that of poly-lysine, which exhibits a lack of such pairing. Combining MD simulations employing explicit solvent with ab initio calculations of ions in a polarizable continuum model (PCM) of water, we rationalize this effect at a molecular level. Additionally, using analysis of structural databases, we relate the present findings to arginine–arginine interactions in proteins. Despite the fact that these interactions are abundant, their role for protein function is not fully understood yet.

We have performed 50-ns molecular dynamics trajectories of di-arginine and di-lysine and 10-ns trajectories of deca-arginine and deca-lysine in water after 1 ns of equilibration. For each trajectory, 10 000 snapshots have provided the input for consequent analysis.

Ab initio calculations of like-charge pairing were performed employing the polarizable continuum model (PCM) of water and, for comparison and robustness check, also using the COSMO model. For cationic pairs, we employed the cc-pvtz basis, while for the nitrate pair, we utilized the aug-cc-pvdz basis set. Through a comparison to test CCSD(T) calculations for the guanidinium–guanidinium pair, we found the results to be converged within 0.5 kcal/mol at the MP2 level, which was then employed for all ion pairs.

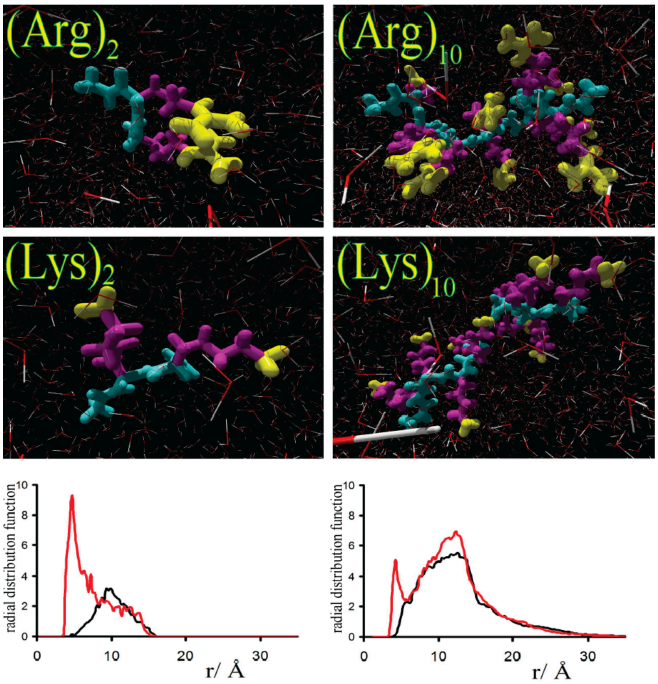
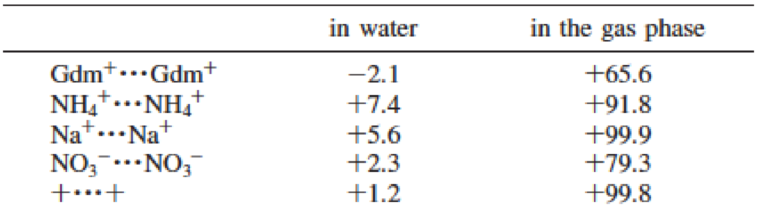


Figure 9. Snapshots from the MD simulation of di- and deca-arginine and lysine (the cationic group in yellow, the side chain in purple, and the backbone in cyan). The lower panels show the radial distribution functions g(r) for the central atom of the cationic group (left for the dimer, right for the decamer; in both cases, the arginine species is shown in red and the lysine species in black).

MD simulations of both $(Arg)_2$ and $(Arg)_{10}$ reveal that the guanidinium groups of the side chains tend to associate. This is apparent both from a visual inspection of the trajectory and from the strong first peak (around 4 Å) of the radial distribution function of the central carbon atoms of the guanidinium groups (Figure 1). In contrast, neither $(Lys)_2$ nor $(Lys)_{10}$ exhibits any direct pairing of ammonium-containing side chains, despite the fact that they are one hydrophobic CH2 group longer (Figure 9).

Table 6. The ab initio energy (in kcal/mol) of like-charged ion pairs optimized at a separation of 3.32 Å in water and in the gas phase (the last line represents a direct application of Coulomb's law).



	in water	in the gas phase
$Gdm^+\cdots Gdm^+$	−2.1	+65.6
$NH_4^+\cdots NH_4^+$	+7.4	+91.8
$Na^+\cdots Na^+$	+5.6	+99.9
$NO_3^-\cdots NO_3^-$	+2.3	+79.3
$+\cdots+$	+1.2	+99.8

Table 6 shows the free energy of association of like-charged ions at a center-of-mass distance of 3.32 Å corresponding to the Gdm+-Gdm+ minimum in PCM water. The free energy minimum of a Gdm+-Gdm+ pair in water amounts to -2.1 kcal/mol at the MP2/cc-pvtz level. This number is stable within 0.5 kcal/mol with respect to further basis set extension and further inclusion of correlation effects at the CCSD(T) level. However, test calculations at the Hartree-Fock level, which lacks correlation effects such as dispersion, bring the free energy of association close to zero.

In summary, we have demonstrated the cationic side-chain association in MD simulations of aqueous oligo-arginines but not oligo-lysines. This effect can be traced to the different behavior of aqueous homoion pairs of Gdm+ and NH4 +, which are the charge carriers of the side chains of Arg and Lys. While the association of Gdm+ ions has been documented in previous calculations[10-12,23], the molecular origin of this effect has not been addressed in detail. The ab initio PCM water calculations presented here both support the previous MD results concerning the energetically favorable formation of the Gdm+-Gdm+ like-charge pair and allow this attraction to be dissected into its individual components. It has been found that a combination of factors results in a favorable Gdm+ like-charge pair but an unfavorable NH4 + ion pairing. Another two factors that bring two Gdm+ ions together are appreciable gains (i) in cavitation (solvent-exclusion) energy and (ii) in dispersion interactions between the two ions upon association. The present results thus provide a molecular rationalization of the electrostatically counterintuitive Arg–Arg pairing, which plays an important role both within and between proteins.

2 Comprehensive analysis of amino-acid interactions in proteins

2.1 A comparison of ab initio and empirical methods for side-chain interactions

Studies of particular protein systems have led to the discovery of unique spatial arrangements of amino acids that still reflect their physical chemical properties. With the existing database of protein structures obtained by high-resolution methods, we can now address even more general questions of statistical preferences for interactions that take place in their 3-dimensional arrangement. This could provide a picture of overall interaction preferences stabilizing the protein structure.

We have selected a representative set of 24 of the 400 (20 °x 20) possible interacting side-chain pairs based on data from the 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 the Møller-Plesset perturbation theory (MP2), the density functional theory combined with symmetry-adapted perturbation theory (DFT-SAPT), the 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. This was to provide benchmarks for different methods, even though the range of binding energies was expected to be extremely large. We could also test how representative the chosen geometries of the side chains were and investigate the effect on the binding energies of the dielectric constant of the surrounding medium.

To obtain a representative set of amino-acid side-chain pairs, we extracted data from the Atlas of Protein Side-Chain Interactions, providing the interaction geometries of all 20 x 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. 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 6 and Figure 9). 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.

Table 6. Statistical data for selected pairs taken from the updated version of the side-chain atlas.

A1	A2	code	$N_{\text{clustered contact}}$ of $N_{\text{detected contacts}}$	$p_{A1}p_{A2}$	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

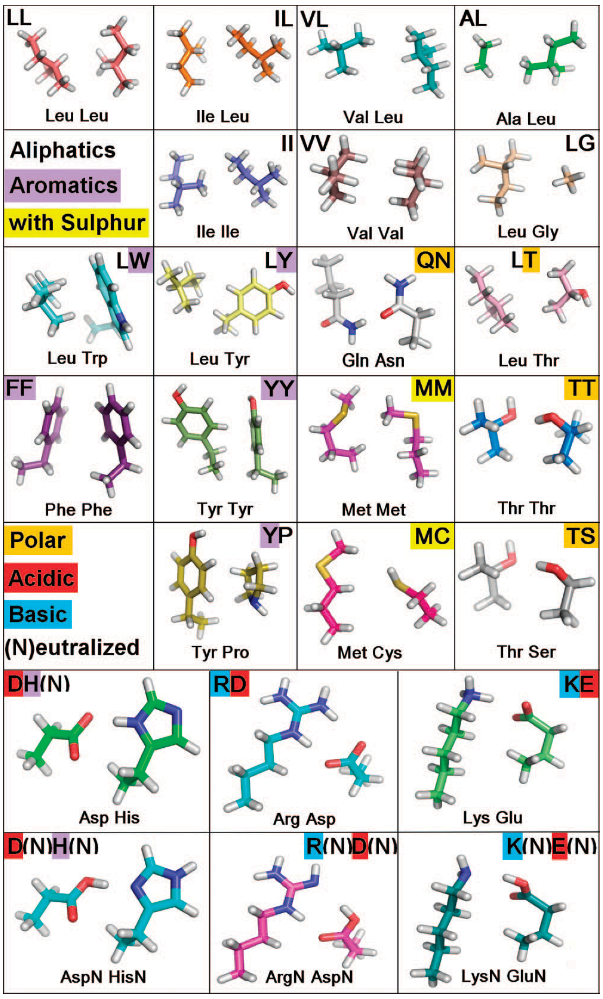
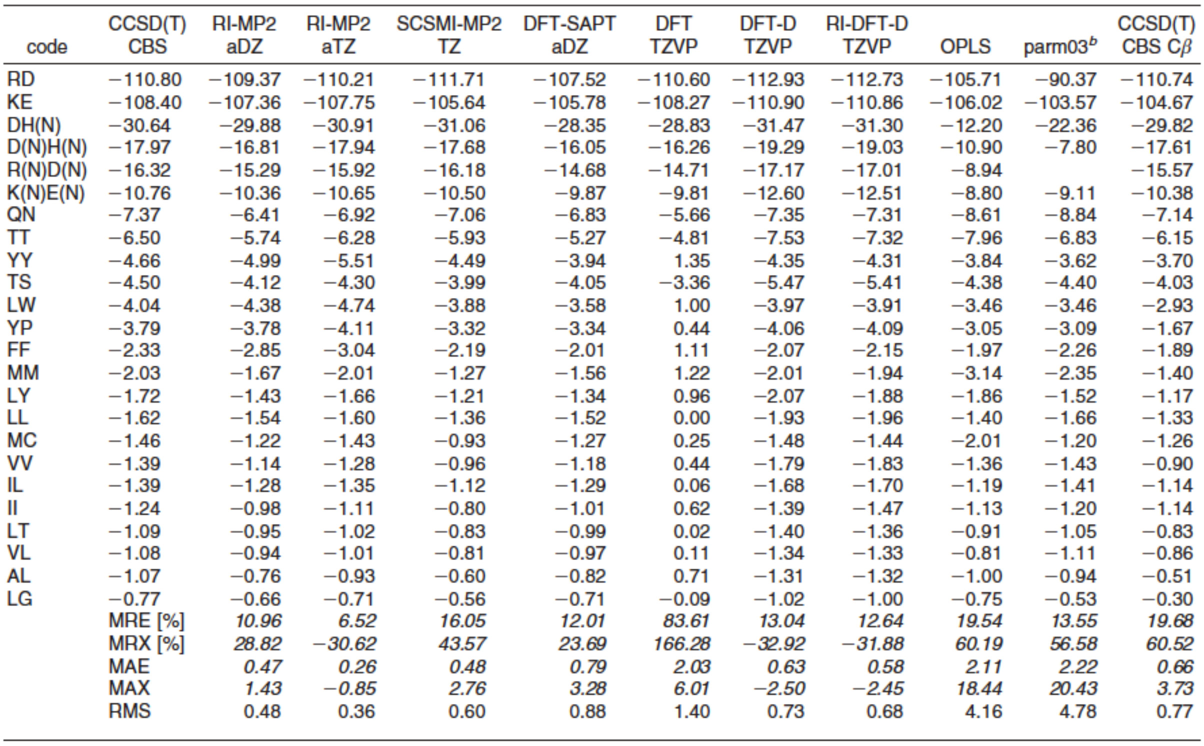


Figure 9. A set geometries of the amino-acid residues truncated at the CR atom and optimized with DFT|TPSS|TZVP, from which the geometries with C_fragmentation were derived by the deletion of the CR methyl group and the insertion of a hydrogen atom in the former methyl direction.

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 performed.

Columns 3–11 of Table 7 show the interaction energies obtained by the nine computational methods tested. 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.

Table 7. The interaction energies for amino-acid pairs calculated using several approaches in the gas phase.



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

To summarize the results, the most accurate method (other than the benchmark CCSD(T)/CBS method) for the 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 interaction energies slightly. The best force-field method is parm03 force field, especially when strongly bound pairs are omitted.

This work presents 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. A 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 non-randomly 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. It should be emphasized 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 that for comparison with experiment, it is inevitable to pass to stabilization enthalpy.

2.2 The decomposition of intramolecular interactions

There have been several attempts to make a comparison between the statistical potential and the ab initio calculation of the interaction energy of amino-acid side chains. Morozov et al.[38,39] have reported remarkable correspondence between the knowledge-based potential of the hydrogen-bond geometries representing amino-acid interactions in proteins and the ab initio DFT and MP2 calculations of the hydrogen-bonding energies for model systems. The same authors have attempted to evaluate the potential energy surface (PES) for the interaction of aromatic residues at MP2 and empirical potential levels. The main conclusion of this work is that the interaction is fairly well captured by the empirical potential and "that interactions between cyclic side chains contribute to the geometric distributions observed in protein structures"[39]. Here, we present the results of our study, in which we describe and evaluate the interaction energies for all 20 x 20 amino-acid side-chain pairs using representative geometries obtained from the analysis of known 3D structures of proteins. We use several force fields as well as quantum chemistry methods both in the gas phase and in a protein/water environment.

Apart from the list of ab initio methods, we also used two modified force fields parametrized earlier – OPLS-AA/L and parm03. These force fields contain only amino acids truncated at the CR atom. The residual nonintegral charge is further distributed over added hydrogen atoms attached to the CR atom. The non-covalent interactions were calculated as a sum of the electrostatic and Lennard-Jones terms for the complexes of amino-acid fragments forming a particular pair. The effect of an environment was evaluated by the RI-DFT-D method utilizing the COSMO model implemented. Two dielectric constants were used to model the effect of a protein/water environment (ε = 4, 80) on the interaction energies.

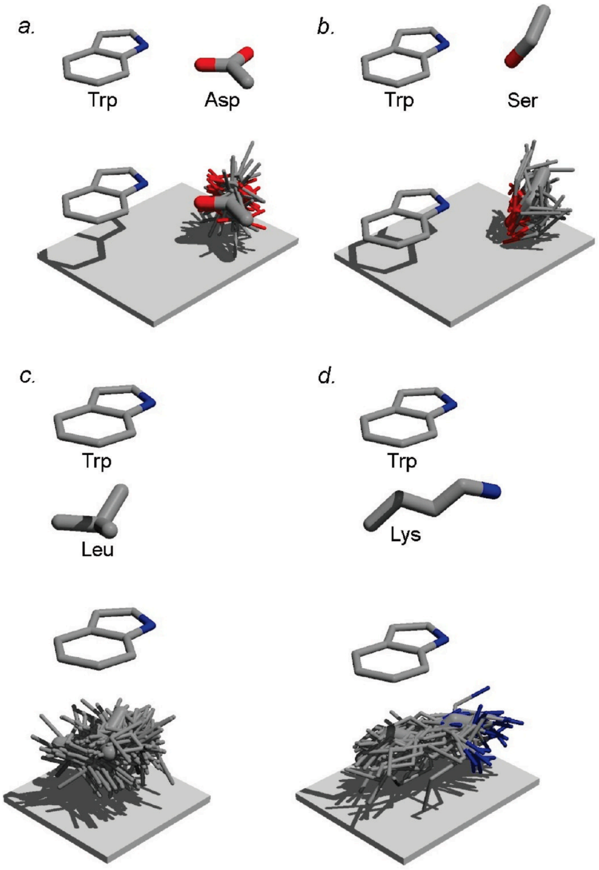


Figure 10. Some examples of side-chain interactions in protein 3D structures. All examples involve interactions with tryptophan. The side chains shown are (a) aspartic acid, (b) serine, (c) leucine and (d) lysine. Each diagram consists of two parts. The lower part shows the largest cluster of the interacting side chains, as extracted from a representative data set of protein structures in the PDB. The "cluster representative" is shown with thicker bonds. This corresponds to the side chain with the lowest total distance to all the other members of the cluster. The upper part of each figure shows only the Trp side chain and the cluster representative, each labeled by its three-letter code. The figure has been rendered using Raster3D.

All of the geometries of the calculated pairs were selected by cluster analysis to represent significantly populated geometry arrangements of interacting amino acids. The reference interaction energies for these pairs calculated by the RI-DFT-D method thus represent a measure of affinity based on the positions of the side chains determined experimentally and stored in the PDB database. The final numbers are presented in Table 8. While all of the interactions in the gas phase can be calculated explicitly and in principle with reasonable accuracy, most of the interactions of biomolecules and their complexes are realized in a protein or water environment, which makes a precise evaluation of the interaction energy complicated, if not impossible, because of the heterogeneous conditions around the interacting residues. In order to take the environment roughly into account, we used solvent-implicit models. We utilized two dielectric constants: ε = 4, mimicking the effect of a protein environment, and ε =

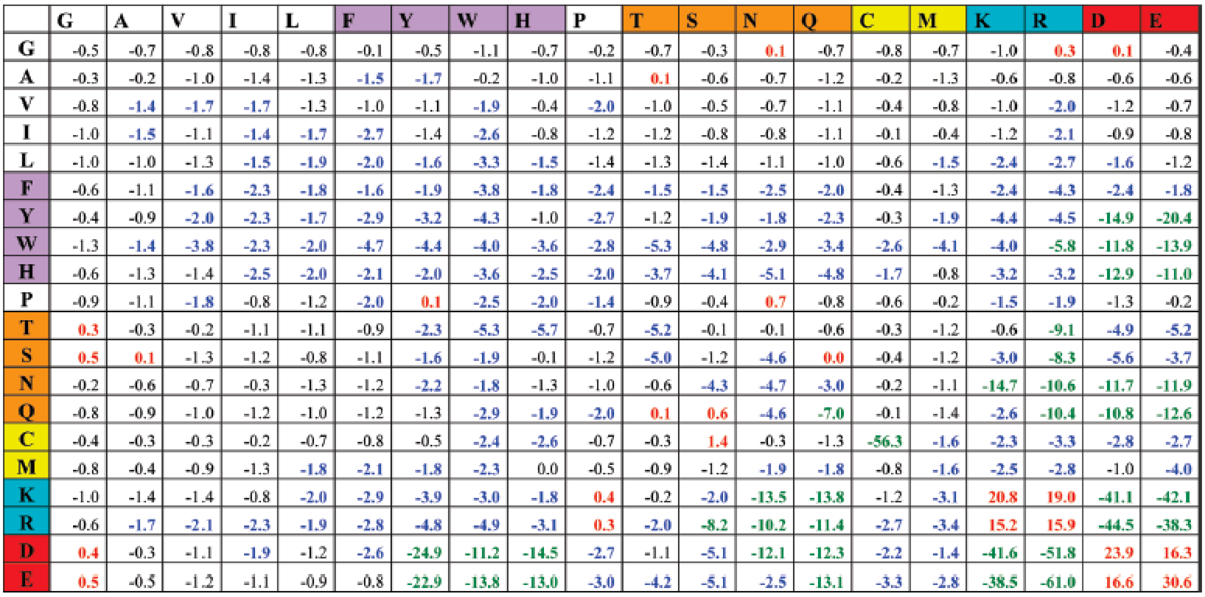
80, for the effect of water. We calculated the interaction energies by the RI-DFT-D method with the COSMO implicit-solvent model.

Table 8. The gas-phase interaction-energy matrix for the cluster representatives for all of the 20 Å~ 20 possible pairs between residues within proteins calculated using the RI-DFT-D/TPSS|TZVP method. All energies are in kcal/mol.

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

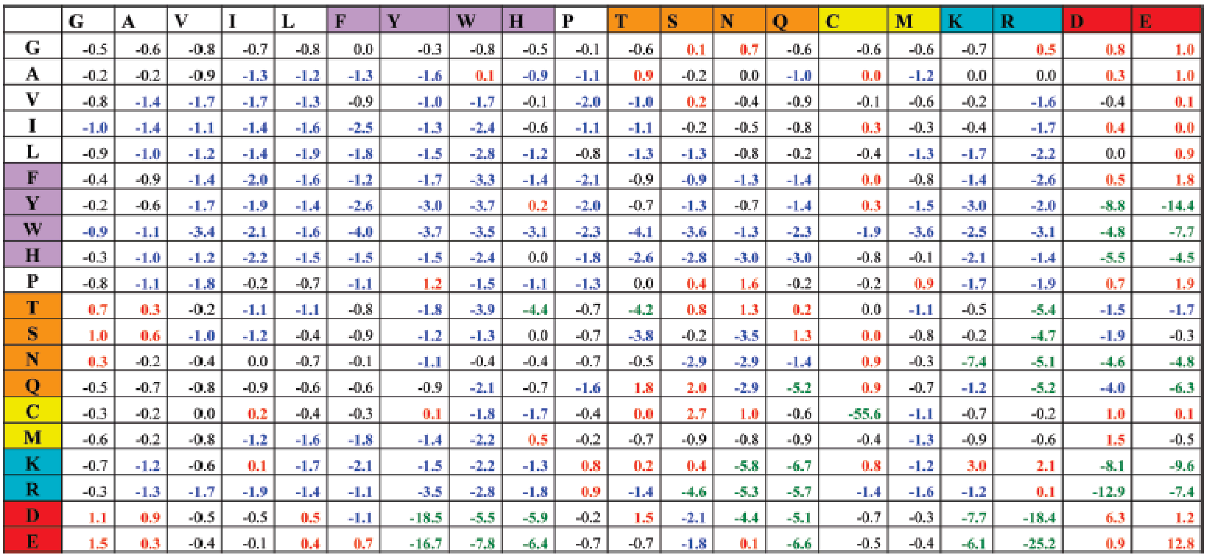
The results presented in Tables 9 and 10 show that the higher the dielectric constant of the surroundings, the smaller the differences between the interaction energies for all of the interacting pairs of amino acids. The apparent reason is the dielectric screening of the dominant electrostatic interaction.

Table 9. The interaction-energy matrix for the cluster representatives for all of the 20 x 20 possible pairs between residues calculated using the RI-DFT-D/TPSS|TZVP method with the COSMO model in a protein-like environment (ε)=4). All energies are in kcal/mol.



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

Table 10. The interaction-energy matrix for the cluster representatives for all of the 20 x 20 possible pairs between residues calculated using the RI-DFT-D/TPSS|TZVP method with the COSMO model in a water environment (ε = 80). All energies are in kcal/mol.



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

Our initial intention was to provide a complete interaction-energy matrix for amino-acid side-chains and compare it to some extent with the data previously published by Miyazawa and Jernigan[40,41]. This comparison is not possible based solely on the results of our calculations for cluster representatives. We have found that the cluster representatives are not statistically significant for the whole ensemble of interactions. Since we limited our analysis to gas-phase interaction energy as the first approximation, we could only attempt to adjust the significance of the representative values by the calculation of the interaction-energy distribution for the complete side-chain interactions of tryptophan. This comparison, i.e. of the interaction energies for the representative geometries and the overall distribution of the interaction energies, showed the significance of cluster representative geometries in the context of the protein and investigated the importance of such interactions. Our results led to the conclusion that the optimum-energy side-chain interactions are not the most abundant ones in proteins. They are strong enough to be geometrically as well as energetically distinguishable from the mostly random (and mostly attractive) interactions of the majority of side-chain/side-chain pairs. It is therefore plausible to suggest that the interactions represented by cluster representatives are of crucial importance for protein stability or protein function because of their selectivity and strength. The distributions of the interaction energies also suggest that the approximations lying behind the phenomenological potentials might simply be wrong, as the distributions are not Boltzmann-like.

3 Practical applications

3.1 Amino Acid Interaction web server

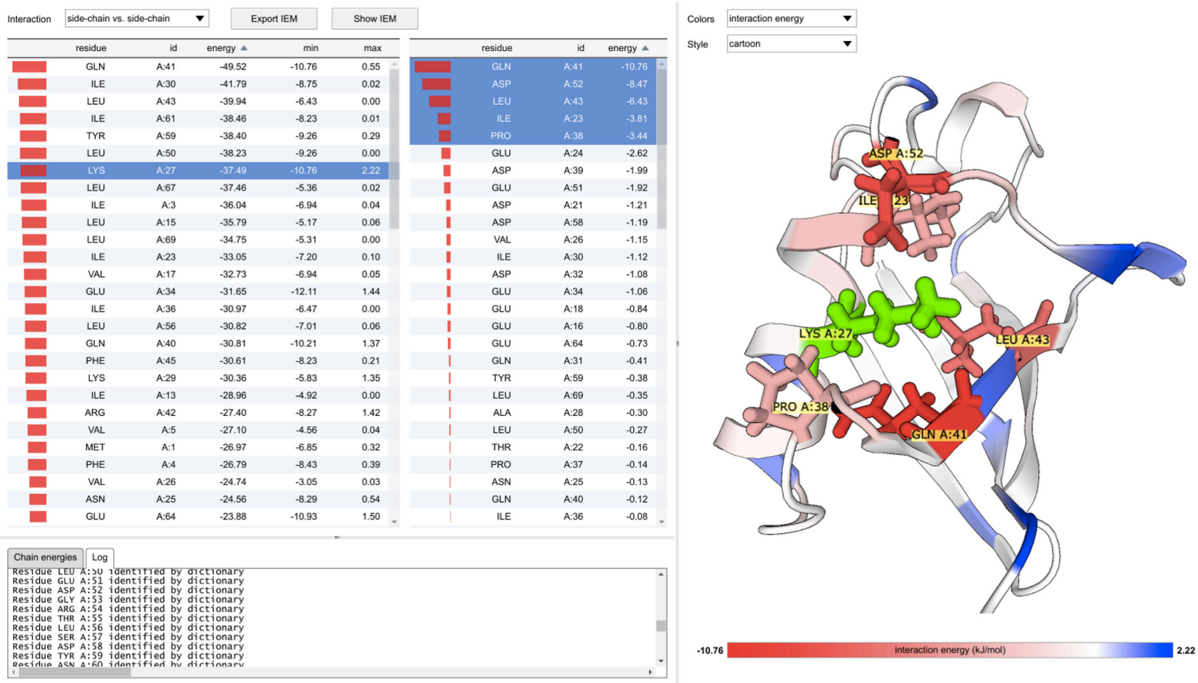
A large body of evidence from mutational studies suggests that amino-acid residues at certain positions in the sequence are more important for the stability and correct formation of a protein fold. Such key residues are usually structurally important and evolutionary conserved across homologous sequences from different organisms. Apart from experimental and alignment-based approaches, an independent, structure-based method has been proposed for the identification of the key residues and their effect on the stability of a protein. This method utilizes the calculation of physically sound interaction energies and evaluates a complete interaction-energy matrix (IEM)[42], involving all pairs of amino acids. The key residues are determined based on the hypothesis that the amino-acid residues with the most stabilizing interactions contribute significantly to the folding enthalpy.

The first calculations and application of the IEM employed precise but time-consuming ab initio methods. Therefore, they were limited to rather small proteins such as Trp-cage or rubredoxin[43,44]. However, it was later demonstrated that the common force fields for biomolecules (OPLS-AA,AMBER parm03) provide acceptable precision for the description of interaction energies. They have made it possible to scale up the computations dramatically and make the calculations feasible for any protein size. The faster calculations thus enabled further application of the IEM, such as an alternative energy-based definition of an amino-acid residue contact or the investigation of protein–protein binding interfaces.

We have proposed the calculation of the IEM as a web service in order to make it easily available for other researchers. The IEM evaluates the pairwise interaction energy (comprising only Lennard-Jones potential and point-charge electrostatics) between well-defined molecular fragments, such as protein and nucleic-acid residues. Although the initial concept of the IEM involved only mutual amino-acid interactions, we have generalized it towards protein–DNA interactions by including support for DNA and RNA residues in our web service.

This service offers the evaluation of the IEM by four common biomolecular force fields, namely OPLS-AA, AMBER parm03, parm99 and charmm36. The supported force fields are commonly used in molecular simulations and represent different parametrization approaches and strategies. These particular force fields have been selected to match the community standards and to reflect our previous work on this topi. Concerning amino-acid residues, all of them provide sounded interaction energies. For calculations of complexes containing nucleic acids, we recommend using the AMBER parm99 or charmm36 force field, whose nucleic-acid parameters are supported in the current version of the IEM service.

The identification of key residues relies on the net interaction energies, which are listed for all residues in the first interactive panel. To guide the eye, the strength of the net interactions is also visualized intuitively by bars next to particular numeric values. For a specific residue selected in the first panel, the decomposition of the net interaction energy is presented in the second panel. Simultaneously, the chosen residues are highlighted in the structure viewer.



Interaction: side-chain vs. side-chain

	residue	id	energy ▲	min	max
🟥	GLN	A:41	-49.52	-10.76	0.55
🟥	ILE	A:30	-41.79	-8.75	0.02
🟥	LEU	A:43	-39.94	-6.43	0.00
🟥	ILE	A:61	-38.46	-8.23	0.01
🟥	TYR	A:59	-38.40	-9.26	0.29
🟥	LEU	A:50	-38.23	-9.26	0.00
🟥	LYS	A:27	-37.49	-10.76	2.22
🟥	LEU	A:67	-37.46	-5.36	0.02
🟥	ILE	A:3	-36.04	-6.94	0.04
🟥	LEU	A:15	-35.79	-5.17	0.06
🟥	LEU	A:69	-34.75	-5.31	0.00
🟥	ILE	A:23	-33.05	-7.20	0.10
🟥	VAL	A:17	-32.73	-6.94	0.05
🟥	GLU	A:34	-31.65	-12.11	1.44
🟥	ILE	A:36	-30.97	-6.47	0.00
🟥	LEU	A:56	-30.82	-7.01	0.06
🟥	GLN	A:40	-30.81	-10.21	1.37
🟥	PHE	A:45	-30.61	-8.23	0.21
🟥	LYS	A:29	-30.36	-5.83	1.35
🟥	ILE	A:13	-28.96	-4.92	0.00
🟥	ARG	A:42	-27.40	-8.27	1.42
🟥	VAL	A:5	-27.10	-4.56	0.04
🟥	MET	A:1	-26.97	-6.85	0.32
🟥	PHE	A:4	-26.79	-8.43	0.39
🟥	VAL	A:26	-24.74	-3.05	0.03
🟥	ASN	A:25	-24.56	-8.29	0.54
🟥	GLU	A:64	-23.88	-10.93	1.50

	residue	id	energy ▲
🟥	GLN	A:41	-10.76
🟥	ASP	A:52	-8.47
🟥	LEU	A:43	-6.43
🟥	ILE	A:23	-3.81
🟥	PRO	A:38	-3.44
🟥	GLU	A:24	-2.62
🟥	ASP	A:39	-1.99
🟥	GLU	A:51	-1.92
🟥	ASP	A:21	-1.21
🟥	ASP	A:58	-1.19
🟥	VAL	A:26	-1.15
🟥	ILE	A:30	-1.12
🟥	ASP	A:32	-1.08
🟥	GLU	A:34	-1.06
🟥	GLU	A:18	-0.84
🟥	GLU	A:16	-0.80
🟥	GLU	A:64	-0.73
🟥	GLN	A:31	-0.41
🟥	TYR	A:59	-0.38
🟥	LEU	A:69	-0.35
🟥	ALA	A:28	-0.30
🟥	LEU	A:50	-0.27
🟥	THR	A:22	-0.16
🟥	PRO	A:37	-0.14
🟥	ASN	A:25	-0.13
🟥	GLN	A:40	-0.12
🟥	ILE	A:36	-0.08

Colors: interaction energy
Style: cartoon

```
Residue LEU A:50 identified by dictionary
Residue GLU A:51 identified by dictionary
Residue ASP A:52 identified by dictionary
Residue GLY A:53 identified by dictionary
Residue ARG A:54 identified by dictionary
Residue THR A:55 identified by dictionary
Residue LEU A:56 identified by dictionary
Residue SER A:57 identified by dictionary
Residue ASP A:58 identified by dictionary
Residue TYR A:59 identified by dictionary
Residue ASN A:60 identified by dictionary
```

Figure 11. The user interface (UI) of the Interaction Energy Matrix Application. The UI provides two interactive tables and an interactive structure viewer. This screenshot captures an analysis of the stabilization role of LYS27 (in 1UBQ). This particular amino-acid residue provides one of the top net interaction energies as found on the left panel, where all net interactions are listed and optionally sorted. The right panel shows the decomposition of the net energy for the selected amino acid. Sorting by energy reveals the strongest interaction partners. The rightmost structure viewer reacts instantly to the actual selection in both panels. In the “interaction-energy” mode, the reference residue is colored green and the others by corresponding interaction energies (the stabilizing interactions in red, the destabilizing interactions and the repulsions in blue). The selected residues with the most stabilizing interactions (ILE23, PRO38, GLN41, LEU43 and ASP52) are additionally highlighted using full-atom representation.

The structure viewer works in two modes. By default, it colors the amino acids based on their net interaction energies, helping the user in finding the key residues. In the alternative mode, the coloring follows the pairwise interaction energies between the chosen reference residue and the others. In this regime, the colors refer directly to a particular row (or column) of the IEM. Furthermore, the web application also offers a decomposition of interaction energies into side-chain–side-chain, backbone–backbone and backbone–side-chain contributions. All analyses and visualizations can be presented for any component of interaction energy.

For medium or bigger proteins, the interaction-energy matrices can be very large. To save the user's internet bandwidth, the application obtains only summary information (net interaction energies) and residue parameters from the server. If necessary, these residue parameters can then be used to calculate a requested subset of pairwise interaction energies on the client's side.

3.2 Theoretical protein design

There is a long history of the de novo computational design of stable and folded peptides that could be utilized in the design of small proteins[16]. There are numerous examples[17–19] of such peptide building blocks, including small β-sheet peptides[20]. Typically, these peptide units are <40 amino acids, allowing us to study in detail the forces and interactions driving protein folding and protein–protein interactions. Recently, there has been great progress in the design of completely new proteins and peptides, as well as those derived from the structures of existing proteins.

Most of the structural and dynamic properties of proteins designed de novo can be explored by computational methods. This description is complementary to experimental characterization and provides some explanation of the protein behavior. State-of-the-art MD simulation techniques can address the dynamics of more complex conformational changes in the protein structure. The development of theoretical methods, together with technical solutions, enables us to successfully simulate the processes of multidomain protein dynamics, interactions and stability. Recombinant-protein expression techniques have helped highlight the advantages of combining different proteins or protein domains. Adding a recombinant fusion partner is one way to change the properties of a target protein.

We have studied chimeric proteins composed of two unrelated types of domains. The first type of domain analyzed includes small artificial or designed mini-proteins able to fold spontaneously into a stable one-domain molecule. On the basis of existing structures and their properties, we selected four small or medium size proteins – of the size of a domain. Their structures, folding, interaction and stability were characterized in detail, giving us a solid background for their utilization in chimeras.

We tested an in silico design strategy using all-atom explicit-solvent molecular dynamics simulations. The well-characterized PDZ3 and SH3 domains of human zonula occludens (ZO-1) (3TSZ), along with five artificial domains and two types of molecular linkers, were selected to construct chimeric two-domain molecules. The influence of the artificial domains on the structure and dynamics of the PDZ3 and SH3 domains was determined using a range of analyses.

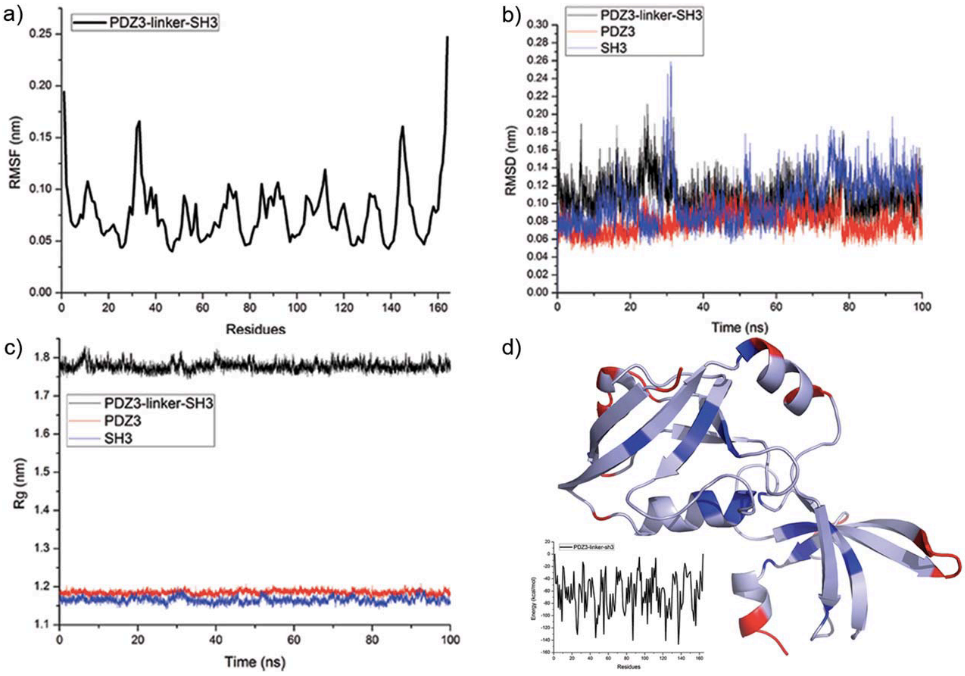


Figure 12. (a) The Ca root-mean-square fluctuation (RMSF) of the PDZ3-linker-SH3 construct. (b) The root-mean-square deviation (RMSD) and (c) the radius of gyration (Rg) calculated for PDZ3, SH3 and the PDZ3-linker-SH3 construct as compared to the starting structure after equilibration. (d) A cartoon representation of PDZ3-linker-SH3 with flexible amino acids (red) and rigid amino acids (blue) determined by interaction-energy matrix (IEM) analysis highlighted. The IEM of the complex is shown in the bottom left corner.

The stability and compactness of the selected structures were calculated by means of Ca-RMSD and radius of gyration (Rg) values (Figure XIII). Steady RMSD values imply that an equilibrium state was reached under the conditions of the simulation. The overall statistical significance of the standard deviation (SD) was considered as the measure of distinct conformational states of each simulated system. We considered the Ca-RMSD and Rg values to be reliable indicators that the structures were stable during the given simulation time and could be considered well-defined for the construction of chimeras with PDZ3 and SH3 proteins.

Structural analyses have confirmed that the most flexible regions in both PDZ3 and SH3 respond similarly to the presence of a modulatory domain and mostly differ in the amplitude of the local residue displacement. The next logical step was to determine whether the flexibility or rigidity of the studied domains is reflected by a change in the interaction-energy profile for particular amino acids in the modulatory domain. For this purpose, we extracted PDZ3 and SH3 for all domain combinations from the MD simulations and colored them according to the spectrum of residual intermolecular energies (Fig XIV). The isolated PDZ3 domain, PDZ3-linker-SH3 chimera and PDZ3-SH3 chimera were considered as benchmarks for further comparison.

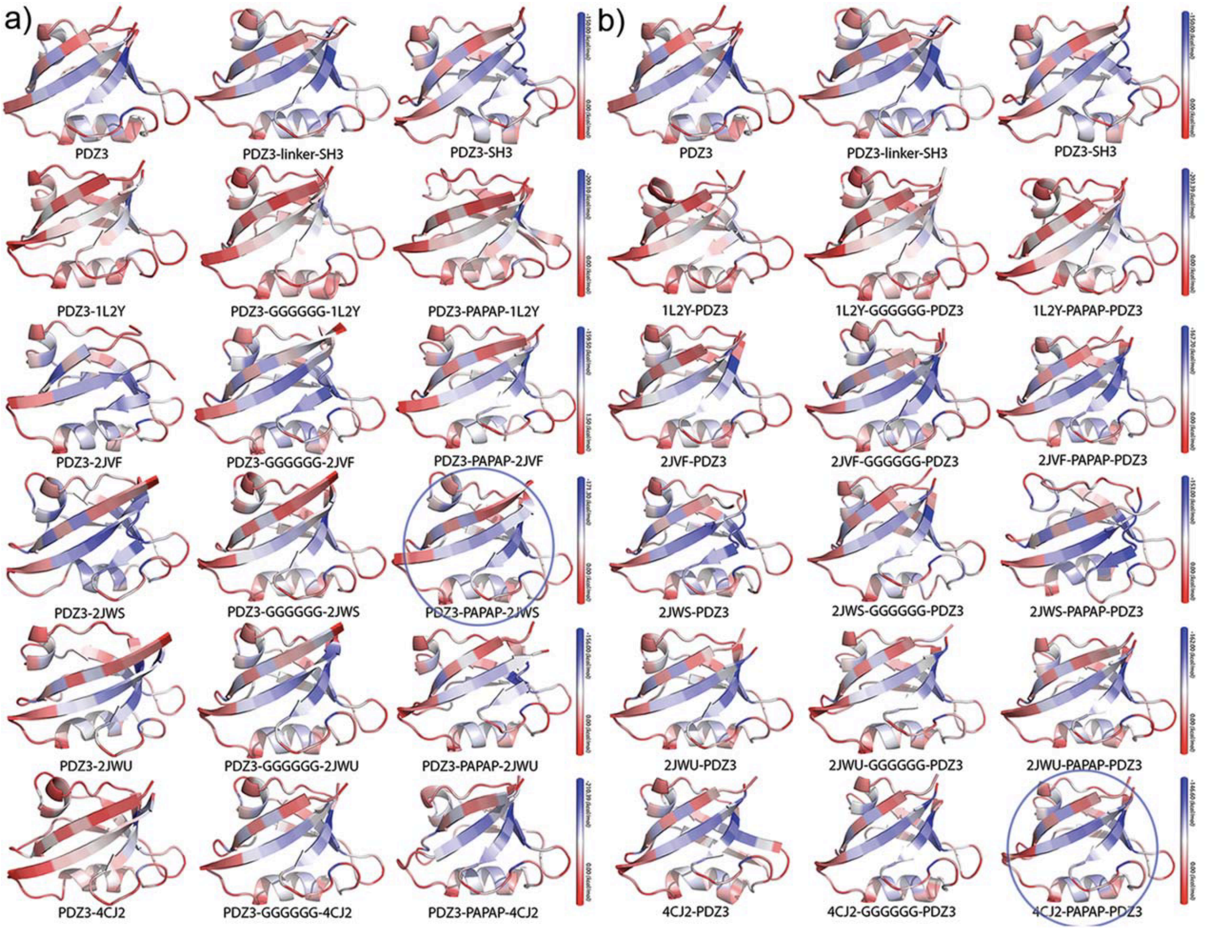


Figure 13. A cartoon representation of the different fusion combination of the PDZ3 domain. A color-coded view of the PDZ3 domain calculated as the IEM for each individual residue of the protein from independent MD simulations of PDZ3 and artificial proteins. (a) A direct combination (NAC terminal fusion). (b) A reverse combination (C-N terminal fusion). The outliers are circled in blue.

A closer inspection of PDZ3 in all constructs (direct and reverse order, no linker, flexible, and rigid linkers) clearly shows the effect of the modulatory domains on the stabilization and dynamics of the PDZ3. Most of the analyzed PDZ3 domains from our constructs have kept the same fold, with a few noticeable exceptions. The very first helix from the N terminus disappears in the chimera construct PDZ3- PAPAP-1L2Y, whereas the reverse combination with the same linker does not show much difference in comparison with the PDZ3 benchmark.

IEM mapping of amino acids was considered to measure stabilization within the local-residue spatial context. To capture trends in local sequence regions, we evaluated the average interaction energy for a residue and its two nearest neighbors [(n21) and (n11)]. Comparisons between these average stabilization energies in the PDZ3 and SH3 domains are shown in Figure 14. Flexible residues are considered to have energy content ranging from 0 to 240 kcal/mol, residues intermediate between flexible and rigid states are considered to have energy content in the range of 240–260 kcal/mol, and rigid residues are those with energy content higher than 260 kcal/mol. We thus evaluated the modulatory effects of the second domain on PDZ3 or SH3 based on these values. The schematic panel (Figure 14(a–d)) portrays the arrangement of residues in the PDZ3 domain and the interaction-energy composition. The stablest (most rigid) residues are represented with squares, intermediates with triangles and the least stable (flexible) residues with crosses. The stablest regions in PDZ3 are centered around F7-K9, D22-F26, S34-A37, E43-L48 and E75-I78, and at least a few of these amino acids or their neighbors appear to participate in the hydrophobic-core formation and protein stabilization.

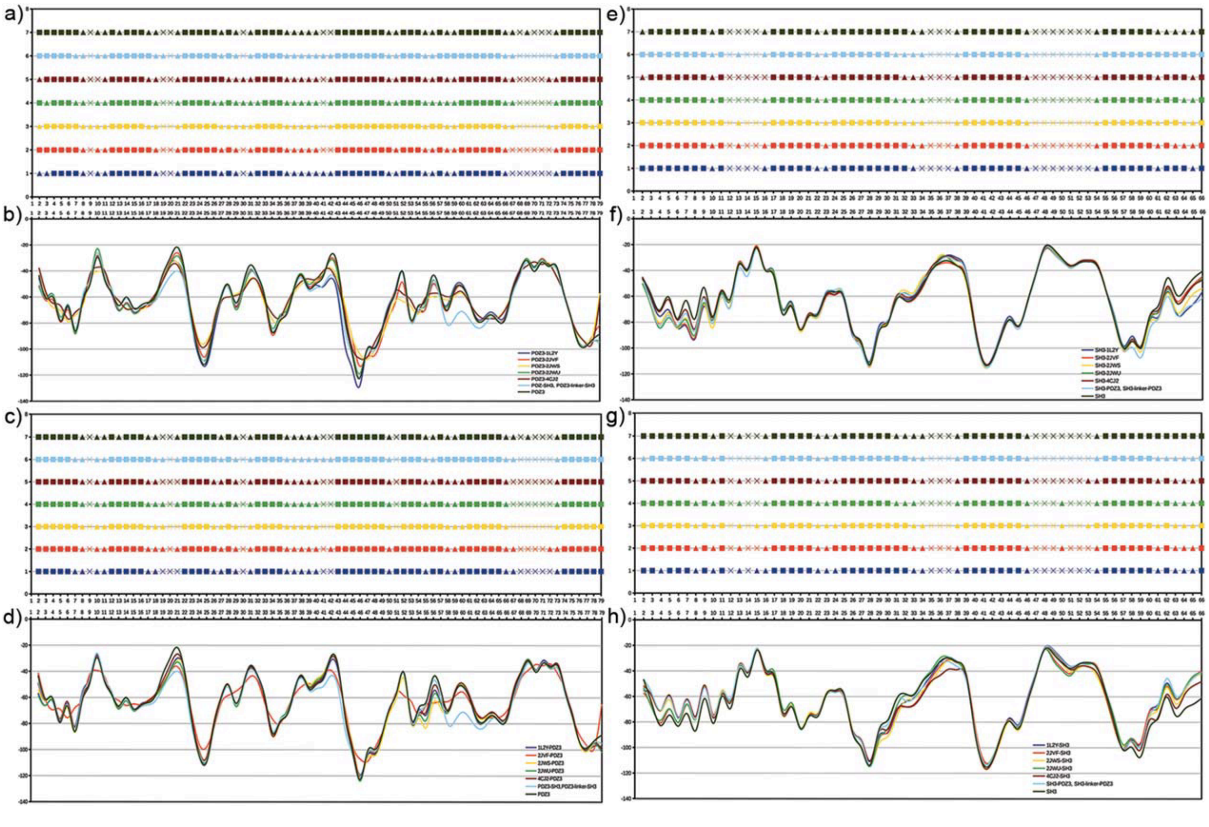


Figure 14. A Schematic representation of the interaction-energy matrices of PDZ3 domains for the direct order of fused domain constructs (panels a and b) and for reverse order (panels c and d). A schematic representation of interaction-energy matrices of SH3 domains for the direct order of fused domain constructs (panels e and f) and for reverse order (panels g and h). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We have shown that both isolated domains and their chimera constructs are reasonably stable and maintain their structure in all MD simulations. The dynamic character of the two-domain constructs and the interactions between the two domains are mainly determined by the character of the linker connecting the domains. Other factors determining chimera behavior include the size and character of the domain structure. We have shown that the localization of flexible regions in the studied domains primarily remains the same in all constructs, but their geometry fluctuations differ in amplitude, depending on the linker, size and surprisingly and unsystematically on the order of the domains. We have found that the reverse order of the PDZ3-SH3 construct exhibits the most distinct behavior, characterized by exceptionally high RMSF values. The direct combination of PDZ3 and SH3 domains is frequent in naturally occurring proteins, but we have not been able to find the reverse combination of SH3 and PDZ in naturally occurring proteins. This could suggest that the order and character of protein-composing domains is not random and that the protein molecules can control allostery by the synergy and character of the composing domains and by their order. We have further shown that the residues determined by the IEM to have the highest energy content correlate well with the stablest residues determined by RMSF values. This clearly shows that the geometry and dynamics of chimeras are products of stabilizing intramolecular interactions, which can be modulated by other domains composing a protein.

4. Conclusions

This habilitation thesis summarizes the results of focused efforts utilizing advanced computational methods to determine important internal stabilizing features in globular proteins. Comprehensive analysis of stabilizing interactions in proteins and their properties in different contexts and environments makes possible to build up a robust statistical framework to understand the importance of particular amino-acid interactions. First of all, the importance of the hydrophobic core for protein stability has clearly been shown to be one of the most important stabilizing elements in proteins. The repertoire of mutual amino-acid arrangements is enormous, but we have been able to determine energetically favorable motifs and the positions contributing the most to the internal protein stability. It has also been quite comforting that the currently used empirical force field describes interaction energies between amino acids at a reasonable level of reliability. On the other hand, we are aware that all of the proposed importance of amino acids in the structural arrangement of a protein would only be hypothetical if not verified by experiment. We have managed to design a mesophilic version of the protein rubredoxin to demonstrate the predictive power of the computational treatment of non- covalent interactions in protein. The applicability of our framework has also been demonstrated on an online version of a web server providing the analysis of intramolecular interactions in proteins for the purposes of an experimentalist. Finally, our approach has been shown to be utilizable in de novo protein design in combination with simulation methods, providing important information on dynamical hot-spots within the protein structure.

References:

(1) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science (80-.).* **1973**, *181* (4096), 223–230.

(2) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; et al. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114* (13), 6589–6631.

(3) Fox, N. K.; Brenner, S. E.; Chandonia, J.-M. SCOPe: Structural Classification of Proteins--Extended, Integrating SCOP and ASTRAL Data and Classification of New Structures. *Nucleic Acids Res.* **2014**, *42* (D1), D304–D309.

(4) Berntsson, R. P.-A.; Smits, S. H. J.; Schmitt, L.; Slotboom, D.-J.; Poolman, B. A Structural Classification of Substrate-Binding Proteins. *FEBS Lett.* **2010**, *584* (12, SI), 2606–2617.

(5) Liberles, D. A.; Teichmann, S. A.; Bahar, I.; Bastolla, U.; Bloom, J.; Bornberg-Bauer, E.; Colwell, L. J.; de Koning, A. P. J.; Dokholyan, N. V; Echave, J.; et al. The Interface of Protein Structure, Protein Biophysics, and Molecular Evolution. *PROTEIN Sci.* **2012**, *21* (6), 769–785.

(6) Bornberg-Bauer, E.; Beaussart, F.; Kummerfeld, S.; Teichmann, S.; Weiner, J. The Evolution of Domain Arrangements in Proteins and Interaction Networks. *Cell. Mol. LIFE Sci.* **2005**, *62* (4), 435–445.

(7) Vogel, C.; Berzuini, C.; Bashton, M.; Gough, J.; Teichmann, S. A. Supra-Domains: Evolutionary Units Larger than Single Protein Domains. *J. Mol. Biol.* **2004**, *336* (3), 809–823.

(8) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: A structural Classification of Proteins Database for For the Investigation of Sequences and

Structures. *J. Mol. Biol. Biol.* **1995**, 536–540.

(9) Vogel, C.; Bashton, M.; Kerrison, N. D.; Chothia, C.; Teichmann, S. A. Structure, Function and Evolution of Multidomain Proteins. *Curr. Opin. Struct. Biol.* **2004**, *14* (2), 208–216.

(10) Letunic, I.; Doerks, T.; Bork, P. SMART: Recent Updates, New Developments and Status in 2015. *Nucleic Acids Res.* **2015**, *43* (D1), D257–D260.

(11) Finn, R. D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; et al. Pfam: The Protein Families Database. *Nucleic Acids Res.* **2014**, *42* (D1), D222–D230.

(12) Mitchell, A.; Chang, H.-Y.; Daugherty, L.; Fraser, M.; Hunter, S.; Lopez, R.; McAnulla, C.; McMenamin, C.; Nuka, G.; Pesseat, S.; et al. The InterPro Protein Families Database: The Classification Resource after 15 Years. *Nucleic Acids Res.* **2015**, *43* (D1), D213–D221.

(13) Thomas, A.; Joris, B.; Brasseur, R. Standardized Evaluation of Protein Stability. *Biochim. Biophys. Acta* **2010**, *1804* (6), 1265–1271.

(14) Shea, J. E.; Brooks, C. L. From Folding Theories to Folding Proteins: A Review and Assessment of Simulation Studies of Protein Folding and Unfolding. *Annu. Rev. Phys. Chem.* **2001**, *52*, 499–535.

(15) Sancho, D. De; Doshi, U.; MunÌƒoz, V. Protein Folding Rates and Stability: How Much Is There beyond Size? *J. Am. ...* **2009**, 2074–2075.

(16) Kloss, E.; Courtemanche, N.; Barrick, D. Repeat-Protein Folding: New Insights into Origins of Cooperativity, Stability, and Topology. *Arch. Biochem. Biophys.* **2008**, *469* (1), 83–99.

(17) Kamerzell, T. I.; Middaugh, C. R. The Complex Inter-Relationships between Protein Flexibility and Stability. **2008**, *97* (9), 3494–3517.

(18) Thomas, P. D.; Dill, K. A. An Iterative Method for Extracting Energy-like Quantities from Protein Structures. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93* (21), 11628–11633.

(19) Lazaridis, T.; Karplus, M. Effective Energy Function for Proteins in Solution. *Proteins* **1999**, *35* (2), 133–152.

(20) Lazaridis, T.; Archontis, G.; Karplus, M. Enthalpic Contribution to Protein Stability: Insights from Atom-Based Calculations and Statistical Mechanics. In *Advances in Protein Chemistry*, Vol. 47; *Advances in Protein Chemistry*; 1995; Vol. 47, pp. 231–306.

(21) Řezáč, J.; Jurečka, P.; Riley, K. E.; Černý, J.; Valdes, H.; Pluháčková, K.; Berka, K.;

Řezáč, T.; Pitoňák, M.; Vondrášek, J.; et al. Quantum Chemical Benchmark Energy and Geometry Database for Molecular Clusters and Complex Molecular Systems (Www.Begdb.Com): A Users Manual and Examples. *Collect. Czechoslov. Chem. Commun.* **2008**, *73* (10).

(22) Rezac, J.; Riley, K. E.; Hobza, P. S66: A Well-Balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **2011**, *7* (8), 2427–2438.

(23) Gaines, J. C.; Clark, A. H.; Regan, L.; O'Hern, C. S. Packing in Protein Cores. *J. Phys. Condens. Matter* **2017**, *29* (29), 293001.

(24) Preissner, R.; Goede, A.; Michalski, E.; Frömmel, C. Inverse Sequence Similarity in Proteins and Its Relation to the Three-Dimensional Fold. *FEBS Lett.* **1997**, *414* (2), 425–429.

(25) Kellis, J. T.; Nyberg, K.; Fersht, A. R. Energetics of Complementary Side-Chain Packing in a Protein Hydrophobic Core. *Biochemistry* **1989**, *28* (11), 4914–4922.

(26) Yue, K.; Dill, K. A. Forces of Tertiary Structural Organization in Globular Proteins. *Proc. Natl. Acad. Sci* **1995**, *92*, 146–150.

(27) Avbelj, F.; Grdadolnik, S. G.; Grdadolnik, J.; Baldwin, R. L. Intrinsic Backbone Preferences Are Fully Present in Blocked Amino Acids. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (5), 1272–1277.

(28) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. Three Key Residues Form a Critical Contact Network in a Protein Folding Transition State. *Nature* **2001**, *409* (6820), 641–645.

(29) Unsworth, L. D.; van der Oost, J.; Koutsopoulos, S. Hyperthermophilic Enzymes − Stability, Activity and Implementation Strategies for High Temperature Applications. *FEBS J.* **2007**, *274* (16), 4044–4056.

(30) Tych, K. M.; Batchelor, M.; Hoffmann, T.; Wilson, M. C.; Hughes, M. L.; Paci, E.; Brockwell, D. J.; Dougan, L. Differential Effects of Hydrophobic Core Packing Residues for Thermodynamic and Mechanical Stability of a Hyperthermophilic Protein. *Langmuir* **2016**, *32* (29), 7392–7402.

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

(32) Burley, S. K.; Petsko, G. A. Amino-Aromatic Interactions in Proteins. *FEBS Lett.* **1986**, *203* (2), 139–143.

(33) Blundell, T.; Barlow, D.; Borkakoti, N.; Thornton, J. Solvent-Induced Distortions and

the Curvature of α-Helices. *Nature* **1983**, *306* (5940), 281–283.

(34) Barlow, D. J.; Thornton, J. M. Ion-Pairs in Proteins. *J. Mol. Biol.* **1983**, *168* (4), 867–885.

(35) Kumar, S.; Nussinov, R. Relationship between Ion Pair Geometries and Electrostatic Strengths in Proteins. *Biophys. J.* **2002**, *83*, 1595–1612.

(36) Kumar, S.; Nussinov, R. Salt Bridge Stability in Monomeric Proteins. *J. Mol. Biol.* **1999**, *293*, 1241–1255.

(37) Pace, C. N. Polar Group Burial Contributes More to Protein Stability than Nonpolar Group Burial. *Biochemistry* **2001**, *40* (2), 310–313.

(38) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. Close Agreement between the Orientation Dependence of Hydrogen Bonds Observed in Protein Structures and Quantum Mechanical Calculations. *Proc. Natl. Acad. Sci.* **2004**, *101* (18), 6946–6951.

(39) Morozov, A. V.; Misura, K. M. S.; Tsemekhman, K.; Baker, D. Comparison of Quantum Mechanics and Molecular Mechanics Dimerization Energy Landscapes for Pairs of Ring-Containing Amino Acids in Proteins. *J. Phys. Chem. B* **2004**, *108* (24), 8489–8496.

(40) Miyazawa, S.; Jernigan, R. L. Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.* **1996**, *256* (3), 623–644.

(41) Miyazawa, S.; Jernigan, R. L. Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules* **1985**, *18* (3), 534–552.

(42) Hobza, P.; Bendova, L. Identifying Stabilizing Key Residues in Proteins Using Interresidue Interaction Energy Matrix. *Proteins* **2008**, No. August 2007, 402–413.

(43) Bendova-Biedermannova, L.; Hobza, P.; Vondrasek, J. Identifying Stabilizing Key Residues in Proteins Using Interresidue Interaction Energy Matrix. *Proteins: Structure, Funct. and Bioinforma.* **2008**, *72* (1), 402–413.

(44) Vondrasek, J.; Kubar, T.; Jenney Jr., F. E.; Adams, M. W. W.; Kozisek, M.; Cerny, J.; Sklenar, V.; Hobza, P. Dispersion Interactions Govern the Strong Thermal Stability of a Protein. *Chem. Eur. J.* **2007**, *13* (32), 9022–9027.

List of publications comprising the habilitation thesis

(1) **Vondrášek, J**., Bendová, L., Klusák, V., and Hobza, P. (2005) Unexpectedly strong energy stabilization inside the hydrophobic core of small protein rubredoxin mediated by aromatic residues: correlated ab initio quantum chemical calculations. *J. Am. Chem. Soc.* ***127***, 2615–9.

(2) Berka, K., Hobza, P. , **Vondrášek, J.** (2009) Analysis of Energy Stabilization inside the Hydrophobic Core of Rubredoxin. *ChemPhysChem.* ***10*** *, 1–7.*

(3) **Vondrášek, J.** Jenney, F. E., Adams, M. W. W., Koz, M., and Hobza, P. (2007) Dispersion Interactions Govern the Strong Thermal Stability of a Protein. *Chemistry A Eur.J.,* ***13****, 9022–9027.*

(4) Bendová, L., Jurečka, P., Hobza, P., **Vondrášek, J.** (2007) Model of Peptide Bond-Aromatic Ring Interaction: Correlated Ab Initio Quantum Chemical Study. *J.Phys.Chem. B,* ***111****,9975-9979*

(5) Biedermannova, L., E. Riley, K., Berka, K., Hobza, P., and **Vondrášek, J**. (2008) Another role of proline: Stabilization interactions in proteins and protein complexes concerning proline and tryptophane. *Phys. Chem. Chem. Phys.* ***10****, 6350-6359*

(6) Řezáč, J., Berka, K., Horinek, D., Hobza, P., and **Vondrášek, J**. (2008) The stabilization energy of the Glu-Lys salt bridge in the protein/water environment: Correlated quantum chemical ab initio, DFT and empirical potential studies. *Collect. Czechoslov. Chem. Commun. 73, 921-936*

(7) Černý, J., Vondrášek, J., Hobza, P. (2009) Loss of Dispersion Energy Changes the Stability and Folding/Unfolding Equilibrium of the Trp-Cage Protein. *J.Phys.Chem.B.* ***113,*** *5567-5660.*

(8) **Vondrášek, J**., Mason, P. E., Heyda, J., Collins, K. D., and Jungwirth, P. (2009) The molecular origin of like-charge arginine-arginine pairing in water. *J. Phys. Chem. B* ***113***, 9041–9045.

(9) Vazdar, M., Vymětal, J., Heyda, J., Vondrášek, J., and Jungwirth, P. (2011) Like-charge guanidinium pairing from molecular dynamics and ab initio calculations. *J. Phys. Chem. A* **115**, 11193–11201.

(10) Bendova-Biedermannova, Lada; Hobza, Pavel; Vondrášek, J. (2008) Identifying stabilizing key residues in proteins using interresidue interaction energy matrix. *Proteins Struct. Funct. Bioinform. 72, 1, 402-413*

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

(12) Berka, K., Laskowski, R. a., Hobza, P., and **Vondrášek, J.** (2010) Energy Matrix of Structurally Important Side-Chain/Side-Chain Interactions in Proteins. *J. Chem. Theory*

Comput. ***6**, 2191-2203.*

(13) Fackovec, B., and **Vondrášek, J**. (2010) Decomposition of Intramolecular Interactions Between Amino-Acids in Globular Proteins-A Consequence for Structural Classes of Proteins and Methods of Their Classification. *Syst. Comput. Biol. – Mol. Cell. Exp. Syst.* 69–82.

(14) Fačkovec, B., and **Vondrášek, J**. (2012) Optimal definition of inter-residual contact in globular proteins based on pairwise interaction energy calculations, its robustness, and applications. *J. Phys. Chem. B* ***116***, 12651–12660.

(15) Galgonek, J., Vymětal, J., Jakubec, D., and **Vondrášek, J**. (2017) Amino Acid Interaction (INTAA) web server. *Nucleic Acids Res.* ***45***, *Web Server Issue*

(16) Kirubakaran, P., Pfeiferová, L., Boušová, K., Bednarova, L., Obšilová, V., and **Vondrášek, J.** (2016) Artificial proteins as allosteric modulators of PDZ3 and SH3 in two-domain constructs: A computational characterization of novel chimeric proteins. *Proteins Struct. Funct. Bioinform.* ***84**, 10, 1358-1374*