Charles University in Prague
Faculty of Science
Department of Physical and Macromolecular Chemistry

Doctoral Thesis

Side-chain Side-chain Interactions in Proteins

RNDr. Karel Berka

Supervisors:
Prof. Ing. Pavel Hobza, DrSc., FRSC
RNDr. Jiří Vondrášek, CSc.

Institute of Organic Chemistry and Biochemistry AS CR
Center for Biomolecules and Complex Molecular Systems

Universita Karlova v Praze
Přírodovědecká fakulta
Katedra fyzikální a makromolekulární chemie

Disertační práce

Interakce mezi vedlejšími řetězci aminokyselin v proteinech

RNDr. Karel Berka

Školitelé:
Prof. Ing. Pavel Hobza, DrSc., FRSC
RNDr. Jiří Vondrášek, CSc.

Ústav organické chemie a biochemie AV ČR
Centrum biomolekul a komplexních molekulárních systémů

I hereby declare that I have written the presented thesis solely by myself and all the literature is properly cited. Neither the thesis nor its parts have been used for obtaining any academic degree.

Prague, 30th November 2009 Karel Berka

Acknowledgement

I would like to thank Prof. Pavel Hobza and Dr. Jiří Vondrášek for their kind guidance and support. I would also like to thank to all members of the Center for Complex Molecular Systems and Biomolecules who were helpful with their important advices and assistance, specifically Dr. Lada Biedermannová, Dr. Jindřich Fanfrlík, Dr. Jan Řezáč and Jiří Vymětal. I am also indebted for the help with the Atlas of Side-Chain Interactions to its author Roman Laskowski. Last but not least, I must acknowledge my debt for the discussions with Dr. Petr Jurečka.

Contents

Preface

Proteins are the most versatile and useful molecules in the cellular arsenal. They are the best catalysts the nature knows. Proteins cover the biggest amount of the cellular functions with range from metabolism and signaling through cell architecture to DNA replication. Variations of their structure and functions are amazing.

And yet, they are built from simple building blocks – amino acids. Each amino acid has many possibilities of interactions with its neighborhood and the sequential context manifested through these possibilities is the main reason for the structure variability.

The experimental investigation of the character and relative strength of interactions between amino acid residues is difficult. On the other hand, theoretical chemistry methods and techniques of are well suited for such task. They can provide useful information about structure, stability and nature of these interactions. The aim of the present thesis is the investigation of interactions between side-chains in the proteins utilizing advanced methods of current theoretical chemistry.

The thesis is based on results of several manuscripts (see bellow). The publications 1 – 3 in the list bellow are dedicated to the accurate *ab initio* calculations of the interaction energies between amino acid residues in the model cases. Subject of publications 4 and 5 is the pairwise interaction energy benchmark provided by the most accurate *ab initio* calculations on interactions between two independent side-chains based on the structural data from Atlas of Protein Side-Chain Interactions. Finally, the paper 6 shows our study on the complete matrix of all possible pairwise side-chain side-chain interactions.

1. Berka, K. et al *ChemPhysChem* **2009,** *10*, 543-548.
2. Biedermannova, L. et al. J. *Phys. Chem. Chem. Phys.* **2008,** *10*, 6350-6359.
3. Řezáč, J. et al. *CCCC* **2008,** *73*, 921-936.
4. Berka, K. et al. *Journal of Chemical Theory and Computation* **2009,** *5*, 982-992.
5. Řezáč, J. et al. *CCCC* **2008,** *73*, 1261-1270.
6. Berka, K. et al. *submitted*

1. Introduction

1.1. Proteins – The Importance of Being Accurate

Protein molecules are the basic machinery and architecture of every cell. Their functions range from catalysis of the chemical reactions, cell signalization and regulation up to the arrangements of the molecular ropes holding the cell together. Therefore, it is extremely important for our understanding of the cellular processes to know how proteins are constructed, stabilized and working. While the protein research spanned over a century and half (1), it brought an enormous knowledge about these cellular tools. However our understanding of all their features is still incomplete.

One of the possible ways towards a better understanding of proteins is based on knowledge of their structure. X-ray diffraction, NMR spectroscopy or electron microscopy provides an atomistic resolution of protein structure. Unfortunately, the structure itself does not explain the protein behavior in its complexity and sometimes even does not provide a clue to all of its functions (2-5). This uncertainty is based on the fact that proteins with the same structural fold can have the different function and that the shape of the ligand binding cavity for the given ligand differs significantly between various proteins (6).

According to the Anfinsen's dogma, the protein spatial structure is defined by its amino acid sequence (7). The "Holy Grail" of the protein research is the knowledge of the rules which are defining the protein spatial structure from the sequence. Amino acid residues in the sequence differ significantly by its physico-chemical properties such as structure, rigidity, polarity, size, and interaction possibilities.

Precise knowledge of the strength and variability of these interactions is thus of a crucial importance. Experimental evaluation of interaction energies is difficult if not impossible at all. In principle, it is possible to obtain the thermodynamical characteristics by the Differential Scanning Calorimetry (DSC) and the Isothermal Titration Calorimetry (ITC). DSC measures the heat capacity and enthalpy changes upon protein thermal denaturation, while ITC measures the enthalpy of ligand binding or enthalpy of the protein assembly. The direct interpretation of these thermodynamical data is not straightforward as they characterize the whole protein together with its environment. These experiments thus cannot be directly used

to measure pair-wise interactions between residues (8). On the other hand, theoretical methods are applicable tools for such task.

The interactions in proteins vary in their nature, strength and directivity. This means that it is necessary to use a theory which provides similar error margins for various interactions. There are several theoretical approaches to obtain the strength of these pair-wise interactions between residues varying in the accuracy and speed. The first approach is the use of knowledge-based potentials while the other possibility is the use of the physical potentials.

1.1.1. Knowledge-based Potentials

The knowledge-based approach uses the known experimental structures for the training of the arbitrary potential, which should cover all the underlying interactions between the residues and with the solvent. The potential is usually constructed from the contact free energies, which are calculated for each pair of residues. The contact energy is based on the quasi-equilibrium between the number of residues in contact and a number of residues separated (9-11). This simple formula was further augmented by additional variables such as the distance between the residues (12-17). Total free-energy of the protein is defined in this approach as a sum of all contact free energies in a given (static) structure.

These pair-wise free-energy potentials have been successful in scoring of the native folds and sequence recognition (11, 18) or in the comparative modeling with SwissMODEL server or within Modeller package (17, 19, 20). Here, the search for the native structure is accomplished by generation of the set of structures similar to the template and then by finding of the structure with a minimum of the free-energy for the specified sequence.

The knowledge-based potentials have also several weaknesses - they greatly depend on the training set (dimensions of the used proteins, as well as amino acid composition) (21-23). The pair-wise potential additivity is also influenced by surrounding residues. The space occupied by other amino acids in a protein strongly limits possible positions for each given pair, which is key factor influencing the statistics of the residual contacts and therefore the free-energy potential based on it.

1.1.2. Physical Potentials

These potentials are constructed from several interaction terms (see later), sometimes augmented with the implicit solvent model. The underlying physical model guarantees the transferability and database independency. Physical potentials can be parameterized to provide directly free-energies or potential energies only.

For free-energy potentials, the search for the native structure is similar to the knowledge-based potentials. The free-energy potential is used to distinguish the native structure from the decoys (24-28). This approach was used successfully to predict the native structures of proteins in the CASP competition by the Rosetta program (29-32).

The potential energy potentials are used for the molecular dynamics simulations of proteins (force field or *ab initio* dynamics). There, a protein is evolving in the used potential. This approach was shown to lead to the native structure for small proteins (33) and it is also used in the Folding@Home distributed computing studies (34). They should properly describe also non-native protein structures. This fact can be used for the further studies of the protein characteristics, such as flexibility, ligand binding, stability, etc. These potentials are the most used ones like Amber set of parm force fields (35-38), OPLS-AA/L (39), CHARMM (40, 41), and MARTINI (42) force fields.

The quality assessment of any potential however needs a proper benchmark. The knowledge-based potentials can be validated by their performance in the new structure prediction as for example in the CASP competition (43). Physical potentials can be further tested by the calculations of the thermodynamical characteristics or in comparison with highly accurate *ab initio* calculations.

1.2. Protein Structure

Proteins are built from 20 amino acids types. Amino acids are connected in the one polypeptide chain as was proposed independently by Hofmeister (44) and Fischer (45) in 1902†. Amino acid residues are organized in the polypeptide chain in a specific sequence ("primary structure"). The chain is built at ribosome during the translation process and it is realized by the formation of a peptide bond between carboxyl and amino group of the neighboring amino acids. Connected amino acid residues interact with each other and with their environment in a number of ways.

The first structural elements ("secondary structure") detected in the protein structure – α-helices and β-sheets found by Pauling and coworkers - are stabilized by hydrogen bonds between the main-chain atoms (46, 47). However, the final shape of the protein ("tertiary structure") is defined not only by the interactions of the main-chain but also by the interactions of residues with the solvent and by the interactions between side-chains.

1.2.1. Characteristics of Amino Acid Residues

Every amino acid residue differs significantly in structure of its side-chain. The side-chains have two main impacts on the structure – firstly, side-chains have different physical-chemical properties, and secondly, their different properties are also influencing the rigidity and secondary structure propensity of the respective main-chain segment. Given that every amino acid has different properties, their roles in protein structures have to reflect this variability (see Figure 1):

- First group of residues are the charged ones. The negatively charged **aspartic** (**D**) and **glutamic** (**E**) **acids** contain carboxyl groups, while the positively charged **arginine** (**R**) and **lysine** (**K**) have guanidine group and ε-amino group, respectively. It is interesting to note, that positively charged residues have groups connected with the main-chain by the long flexible chain, while the negatively charged residues are relatively shorter and more rigid. The side-chain flexibility is important at the protein surface, where the side-chains are exposed into water.

† They proposed the idea in the same day, at the "74th Annual Meeting of the Gesellschaft der deutschen Naturforschen und Ärzte" on September 22, 1902 in Karlsbad (today Karlovy Vary, Czech Republic).

- Polar residues **serine** (**S**) and **threonine** (**T**) contain a hydroxyl group which is capable of the hydrogen bonding. The **asparagine** (**N**) and **glutamine** (**Q**) contain an amidic group, which is capable of the multiple hydrogen bonding. These residues are usually also exposed to the solvent either on the surface or inside the active site.
- Next group consists from residues containing sulphur – **cysteine** (**C**) and **methionine** (**M**). Both of these residues are very unique since they are known to make covalent bonds upon oxidation, while the cysteine bonding is much more known as a disulphide bridges.
- Aromatic residues **phenylalanine** (**F**) and **tyrosine** (**Y**) contain benzene ring, while the tyrosine in addition contains a polar hydroxyl group. **Tryptophane** (**W**) and **histidine** (**H**) contain indole group and imidazole ring, respectively. All of these groups are easily polarizable and they are usually positioned in the central part of the protein.
- Aliphatic residues contain hydrocarbon side-chains, which are with exception of **alanine** (**A**) branched, while the size grows in line **valine** (**V**), **isoleucine** (**I**) and **leucine** (**L**). The aliphatic residues are the most common ones and make the majority of the contacts. The most common amino acid residue is leucine.
- Finally, there are two special residues in proteins. **Glycine** (**G**) is the smallest of the amino acid residues. Due to non existence of its side-chain, it allows the conformational variability of the main-chain unseen in other residues. **Proline** (**P**) has cyclic side-chain avoiding more main-chain conformation possibilities and thus making protein structure more rigid. Both glycine and proline are known to be positioned mainly on the hinges and bends.

The charged and polar residues are usually exposed to the water environment surrounding the protein molecule and they are often referred as “hydrophilic“(water-liking). Aromatic and aliphatic residues are usually found in the central part of the protein avoiding the contact with the bulk water at the surface and due to this fact they are often called “hydrophobic“(water-fearing). The central part of the folded protein is known as “hydrophobic core” due to the presence of hydrophobic residues within the interior of a protein.

The concept of hydrophobicity is the subject of the next chapter.

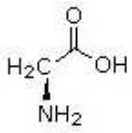
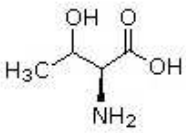
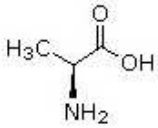
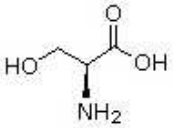
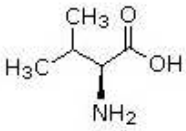
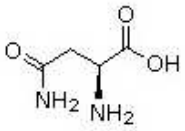
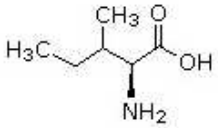
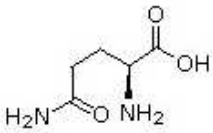
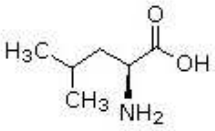
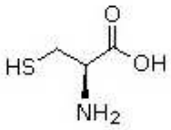
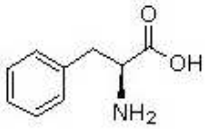
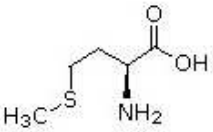
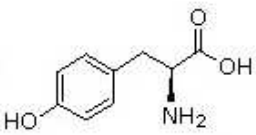
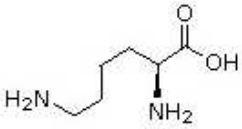
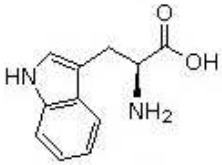
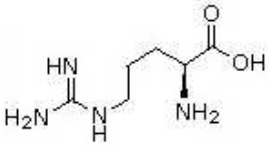
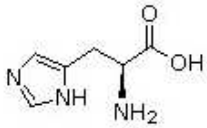
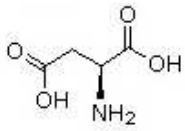
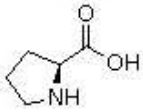
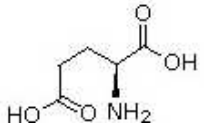
Name	Formula	Abbreviations		Name	Formula	Abbreviations	
Glycine		Gly	G	Threonine		Thr	T
Alanine		Ala	A	Serine		Ser	S
Valine		Val	V	Asparagine		Asn	N
Isoleucine		Ile	I	Glutamine		Gln	Q
Leucine		Leu	L	Cysteine		Cys	C
Phenylalanine		Phe	F	Methionine		Met	M
Tyrosine		Tyr	Y	Lysine		Lys	K
Tryptophan		Trp	W	Arginine		Arg	R
Histidine		His	H	Aspartic Acid		Asp	D
Proline		Pro	P	Glutamic Acid		Glu	E

Figure 1 – Structures and abbreviations of all amino acids.
Amino acid abbreviation backgrounds are colored according to the prevailing character of the residue. Residues are aliphatic (white), aromatic (violet), polar (orange), sulphur-containing (yellow), positively charged (blue), and negatively charged (red). The same coloring scheme is used in the rest of the thesis.

1.2.2. Hydrophobicity

The concept of the hydrophobicity for proteins was established by Kauzmann (48). He suggested that the central part of the protein can be modeled as the mixture of non-soluble hydrocarbons with water, whose possible interactions with water are accompanied with the favorable enthalpy, but with highly unfavorable entropy as water tries to remain in the hydrogen bonding net. He suggested that waters in the vicinity of the hydrophobic molecule forms structured “icebergs” around the solute which retains the interactions (enthalpy) on the cost of their freedom of motion (entropy). Those “iceberg” waters can be freed upon the hydrophobic assembly, which would lead to the increase of entropy during such process.

It was shown later, that the hydrophobicity act differently upon solutes of different sizes (49-51). For the small solutes the hydrophobic effect is mainly of the entropic origin as described above, but for bigger solutes the hydrophobic effect is mainly enthalpic due to the smaller interactions between hydrophobic solute and water than the interactions between water molecules. As a result, water molecules are more mobile near bigger hydrophobic surfaces than in the bulk water (52). Therefore the water between two hydrophobic interfaces is becoming vapor-like and it is readily “dried” and the hydrophobic interfaces can collapse to each other.

On the other side, hydrophilic (water-liking) surface slows waters in its vicinity, where interactions between solute and water are more attractive than between waters (52, 53). The protein surface is however much more diverse than idealized hydrophobic/hydrophilic surface (54, 55), but the hydrophilic surface seems to be prevailing. Upon the contact between two hydrophilic surfaces, the captured water is freed and thus the contact between two hydrophilic surfaces is primarily driven by the entropy like in the Kauzmann’s idea above.

We should note that hydrophobicity (exclusion of the solute from solvent) act indirectly, as it does not exist without the solvent presence. Therefore we should not speak about hydrophobic interactions in protein, as the hydrophobicity is in fact caused by the interactions within the solvent and not within the solute. One of possible ways how to escape this uneasily definable hydrophobicity in a description of side-chain side-chain interactions is a use of cavitation energy instead.

The solvation studies utilize different concept of hydrophobicity definition. The free energy of solvation is defined as a sum of its changes due to the polarization, differences in dispersion and repulsion and lastly by cavitation energy. The hydrophobicity in this concept is the free energy of cavitation, which is the work needed for creation of the cavity inside the solvent in the shape of the solute (56, 57).

$$\Delta G_{solv} = \Delta G_{polarization} + \Delta G_{dispersion} + \Delta G_{repulsion} + \Delta G_{cavitation}, \quad (1)$$

This "physical" definition of hydrophobicity has several advantages over the "biological" definition of the hydrophobic effect. Mainly it does not mix the interactions of the different origin. In other words, interactions between water molecules are stronger in the "biological" definition; however these interactions are mix of electrostatic as well as dispersion between the water molecules, and their change connected with the cavity formation. The "physical" definition also shows the common structural habit of the hydrophobic macromolecules to have as much spherical surface as possible to minimize the cost of the cavity formation.

Last but not least the tertiary structure of protein is defined by the interactions between side-chains. We will focus on them in the next chapters.

1.3. Side-chain Side-chain Noncovalent Interactions

Noncovalent interactions are considerably (by about two orders of magnitude) weaker than the covalent interactions responsible for the formation of a covalent bonds. On the other hand, they still have strong influence on the protein structure – they are numerous, and as they are individually weaker so they can be adjusted in a cooperative way. This cooperativity also allows the protein to overcome bigger structural changes for instance upon ligand binding. There are several types of non-covalent interactions which will be in detail described in the following paragraphs.

1.3.1. Electrostatic Interactions

The electrostatic interactions are the interactions between monopoles and multipoles of the molecules and consist of long-range multipole and short-range overlap parts.

1.3.1.1. Multipole Electrostatic Energy

The multipole electrostatic energy is the energy coming from the two charge clouds (atoms or molecules), due to the Coulomb forces. At larger distances, where the overlap between the charged clouds is negligible, the interaction is usually approximated as a multipole expansion between the charged clouds. For the illustration, the electrostatic energy for the interaction of the point charge with the charged cloud characterized by charge, dipole and quadrupole moments is as follows:

$$E_{el} \approx q_1\left(\frac{q_2}{r}+\frac{\mu_2}{r^2}+\frac{1}{2}\frac{Q_2}{r^3}+\ldots\right), \quad (2)$$

where q_2 is a monopole (total charge), μ_2 is a dipole and Q_2 is a quadrupole of the charged cloud and r is the distance between the centre of the cloud and point charge q_1. For further reading please use the Refs (58, 59).

The electrostatic interactions can be either attractive or repulsive depending on the signs of monopoles and multipoles. As seen from the equation (2), the main electrostatic interaction between the two charged systems would by their charge-charge interactions, while for the charged and polar systems it is charge-dipole term and in the case of charged and aromatics (nonpolar) systems it will be charge-quadrupole interaction, etc. The strength of the electrostatic interaction is usually decreasing in the same line, i.e. the interaction between the two molecules with the nonzero monopoles (charges) is usually the strongest. The distance dependence is also defined by the lowest nonzero multipoles as can be seen from the equation (2), because the interactions between monopoles are the longest.

The multipole electrostatic interaction energy can be classically approximated by two major approaches. Either we can calculate all multipole coefficients per atom in a molecule by the distributed multipole analysis from its wave function (60) or the multipole analysis can be truncated at atomic monopoles – atomic partial charges and the rest of multipoles are fitted onto them. The former method is computationally intensive and also the atomic multipoles can quite vary due to the conformational changes in the molecule. The latter method has advantage in the computational speed as it is rather easy to calculate the electrostatic energy between the partial charges with the Coulomb‘s law (3):

$$E_C = \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}}, \quad (3)$$

where ε_0 is the permittivity of vacuum, q_i is partial charge on atom i, and r_{ij} is interatomic distance.

The use of the atomic partial charges to represent the molecule has also an advantage in a robustness of the method, because the partial charges of the atoms are less affected by the possible conformational changes. There are several ways of calculation of atomic partial charges like Mulliken population analysis (61), or its updated self-consistent version (62), but the best representation of the partial charges is restrained electrostatic potential method (RESP), which fits the electrostatic potential from partial atomic charges onto the electrostatic potential calculated from wavefunction using some quantum mechanical method (63, 64). In this case the resulting atomic charges effectively include atomic dipoles, quadrupoles, and higher multipoles.

1.3.1.2. Overlap Electrostatic Energy

Overlap electrostatic energy (sometimes also called penetration term) is the close-range electrostatic interaction, which is always attractive. This interaction arises from the overlap of the two charge clouds around two point charges with the opposite charge, i.e. two nuclei with the respective electron clouds around them. In the situation when there is no overlap, the charged clouds are repelled with each other and held by their respective nucleus. In case when nuclei are closer, their charged clouds overlap and electrons in these clouds are attracted also to the second nucleus. This is the reason for the attraction coming from this term (60). This energy is at even smaller distances compensated by the exchange-repulsion interaction.

1.3.2. Induction Energy

Induction interaction arises from the adaptation of the molecule to the electrostatic field (*E*) of all its neighbors. The electrostatic field imposes on the molecule induced dipole according to the molecular polarizability (α):

$$\mu_{ind} \approx \alpha \cdot E \quad (4)$$

While the atomic polarizability is isotropic property (see Table 1), the molecular polarizability is a tensor and not a simple sum of the atomic polarizabilities (65).

Table 1 – Atomic polarizabilities taken from Ref (65).

Atom	α [au]	Atom	α [au]
H	2.8	O	5.7
C	8.7	S	16.9
N	6.6	Cl	16.2

The induced dipole interacts with the permanent dipole of the other molecule in a similar way as the normal dipole does, so the pair-wise induction interaction energy is approximately given by:

$$E_{ind} \approx -\frac{\mu_1^2 \alpha_2}{r^6} - \frac{\mu_2^2 \alpha_1}{r^6} \tag{5}$$

Induction interaction energy with the other molecules is (contrary to electrostatic one) always attractive. As the multipole electrostatic energy has its overlap counterpart, the charge-transfer interaction is overlap counterpart of induction. This type of interaction becomes more important in the vicinity of the charged species like ions or coordinated metals like zinc or ferrum in the active site or complexes of electron donor with electron acceptors.

1.3.3. Dispersion Energy

Last type of the noncovalent interaction is dispersion. Dispersion interactions are non-classical effects arising from the correlation between the electron movements. It can be demonstrated classically as the interactions between instantaneous time-dependent and induced dipoles. The pair-wise dispersion energy is thus corresponding to:

$$E_{dis} \approx -\frac{\alpha_2 \langle \mu_1^2 \rangle}{r^6} - \frac{\alpha_1 \langle \mu_2^2 \rangle}{r^6} \tag{6}$$

where ‹μ› is averaged induced dipole.

The dispersion interaction is (like the induction one) always attractive and it is ever present (58). It is the force which is responsible for most of the interactions between the nonpolar species as are aromatic or aliphatic residues.

1.4. Examples of Interaction Types between Side-chains

The types of interaction mentioned above are universal interaction forces. Every residue interact differently according to its charge, multipole moment or polarizability. In the following paragraphs several typical protein specific side-chain side-chain interactions are discussed.

1.4.1. Salt Bridges

Salt bridge is the special case of the multipole electrostatic interactions. It is formed by two oppositely charged residues, which are in vicinity (e.g. nitrogen atom within 4 Å to oxygen atom from the negatively charged residue). One of the typical salt bridges is depicted on Figure 2. The strength of the individual salt bridge is around 100 kcal/mol in the gas phase, which is almost the same value as for the covalent bond between two carbon atoms. Salt bridges can be, however, easily dissociated in the water environment, where the electrostatic interactions with surrounding water molecules are more preferred.

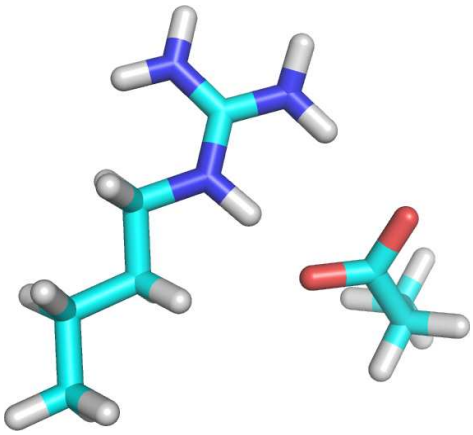


Figure 2 – Example of salt bridge
Interaction between side-chains of arginine (R) and glutamic acid (E).

It has been argued in the literature that salt bridges are the main reason for the protein thermostability (66, 67), as there is higher number of the charged residues on the surface of the hyperthermophilic proteins than on the surface of their mesophilic counterparts. However not all pairs of oppositely charged residues close to each other necessary form the salt bridges (68). The reason is bigger attraction of the water molecules to the charged group which can overcome the salt bridge binding energy.

1.4.2. Hydrogen Bonding Pairs

Hydrogen bonding pairs occur in proteins mostly between main-chain atoms. The side-chains can be a part of the hydrogen bonding in the case of the presence of a heteroatom. The nature and most of the properties of hydrogen bonds can be explained by the electrostatic model. It is a bond between two electronegative atoms operated by the hydrogen atom between them. Hydrogen is covalently bound on the donor atom and it is pointing at the acceptor atom. The hydrogen bond distance is usually below 2.5 Å between acceptor and hydrogen and the donor-hydrogen-acceptor angle is between 90° and 180° with the most frequent angle around 160° (69, 70). The strength of a typical hydrogen bond is around 5 kcal/mol in the gas phase.

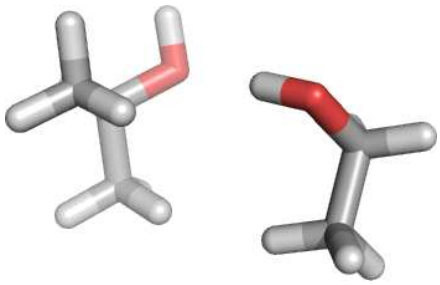


Figure 3 – Example of hydrogen bonding pair.
Interaction between side-chains of polar residues threonine (T) and serine (S) .

The hydrogen bonds are direction-dependent and as thus they can be structure determinants. On the other hand, water molecules also form hydrogen bonds easily and the interaction of the side-chain with water is also entropicaly favorable thanks to the water mobility. To prevent this interaction, the hydrogen bond has to be shielded from the water environment. Examples of such shielding are main-chain hydrogen bonds stabilizing the secondary structure elements. These hydrogen bonds are covered by side-chains exposed to the environment protecting main-chains from the water environment. Bigger flexibility of the side-chains however makes such shielding difficult for the side-chain hydrogen bonds.

1.4.3. Dispersively Bound Pairs

In the case of aliphatic and aromatic residues, the prevailing interaction is the dispersion interaction being the weakest one. Dispersively bound pairs have interaction energies around 1 - 5 kcal/mol. The interactions of the aromatic residues are stronger than those of the aliphatic residues.

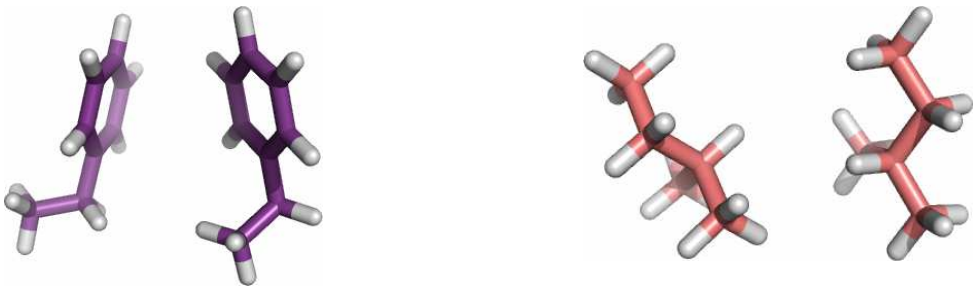


Figure 4 – Examples of dispersively bound pairs.
Interaction between side-chains of aromatic residues phenylalanines (F) (left) and aliphatic residues leucines (L) (right).

Given that dispersion interactions are the weakest ones, it is interesting that dispersively bound pairs are prevailing interactions in the protein hydrophobic core – the most stable structural element of a protein, which is densely packed (71).

1.5. Side-chain Side-chain Covalent Interactions

Besides the non-covalent interactions between side-chains (which are dominant), there are also covalent interactions. The covalent bonds are less numerous than the non-covalent interactions and following expectations, they are also stronger. Further, they do not break as much as the noncovalent interactions do. For a long time the only covalent interactions between side-chains were disulphide bonds. However, another covalent bond has been reported recently between methionine and lysine.

1.5.1. Disulphide Bonds

Disulfide bonds in proteins are formed by oxidation of the thiol (-SH) groups of cysteine residues. The linkage is also called an SS-bond or disulfide bridge and its formation can be seen on Figure 5 and the created residue is called “cystine”. Bond length is about 2.0 Å. The disulphide bond prefers conformations which have dihedral angles approximately 90°; -85.8° for a *left-handed* conformation and 96.8° for a *right-handed* conformation (72).

Disulphide bonds are weaker than the covalent bonds between carbon atoms due to the sulfur size and polarizability. Bond dissociation energy is about 60 kcal/mol (73). On the other hand, they are still considerably stronger than any non-covalent interactions and as such they increase the rigidity of the protein structure. They are more abundant in small proteins. Due to the nature of disulphide bond, it is unstable in a reducing environment like in cytoplasm. They are also rare in hyperthermophilic organisms, probably due to the lesser stability at a higher temperature.

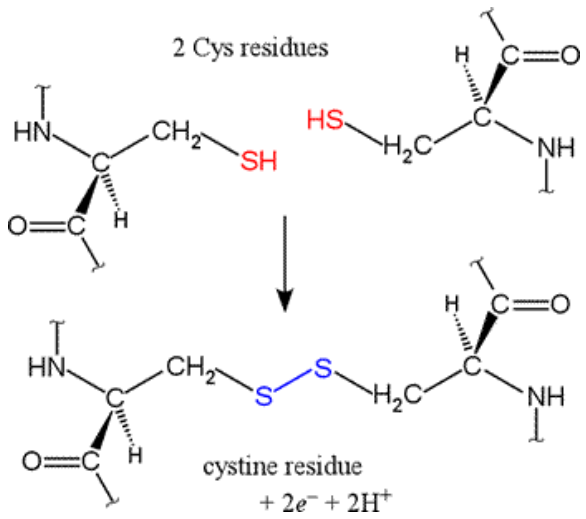


Figure 5 – Formation of the disulphide bond

1.5.2. Methionine-Lysine Bonds

Than et al. reported a novel covalent bonding between methionine and lysine in extracellular matrix in connection between subunits of collagen IV (74) based on the detected electron density between M93 and K211 not explainable by noncovalent interactions (PDB ID: 1LI1). Authors concluded that thioester bond was formed, which means that the bond between carbon and sulphur atoms was formed as shown on Figure 6A.

Recently, this idea was challenged by Vanacore et al (75), who proposed that the bond formed is in fact sulfimine, i.e. that the double covalent bond is formed between sulphur and nitrogen atoms. Mass spectrometry shows that connected peptide fragments lack 2 hydrogen atoms indicating that bond between residues is formed by oxidation of the residues (Figure 6B). However, sulfilimines are usually stabilized by the electron-acceptor group on the nitrogen atom (76), which is not the case in the proposed bonding in proteins. Therefore also cyclic arrangement was proposed as shown on Figure 6C.

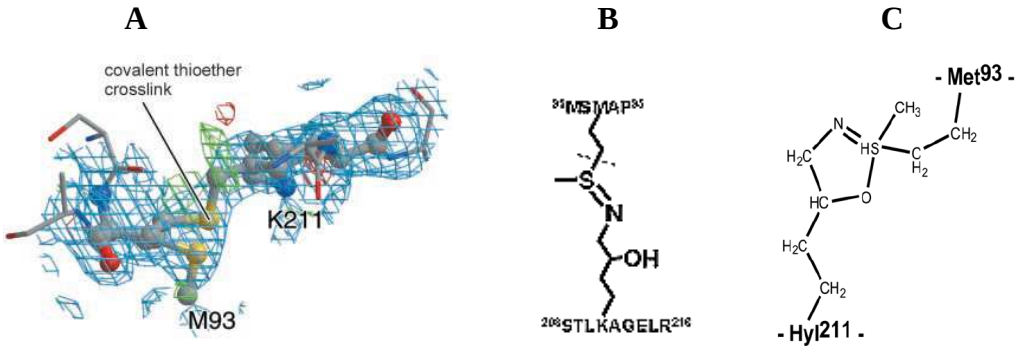


Figure 6 – Binding possibilities between methionine and lysine or hydroxylysine.
A - Thiosester bond proposed by Than et al. (74). B - Sulfilimine bond proposed by Vanacore et al (75). C - Cyclic sulfoximine bond proposed by Vanacore et al. (75).

1.6. Side-Chains Interactions and Protein Stability

The stability of a protein can by defined as a free energy difference between the native structure of protein and its denatured counterparts. The free energy difference has several sources – hydrophobic effect, the rigidity of the peptide bond, conformation preferences of the main-chain and side-chain side-chain interactions.

Each source of the stability has different strength in different stages of the protein folding. According to the folding funnel theory (77), the influence of the hydrophobic effect is the strongest at the beginning of the protein folding upon the hydrophobic collapse, while the last

steps are governed by the selection of the proper side-chain side-chain interactions (78) (see Figure 7).

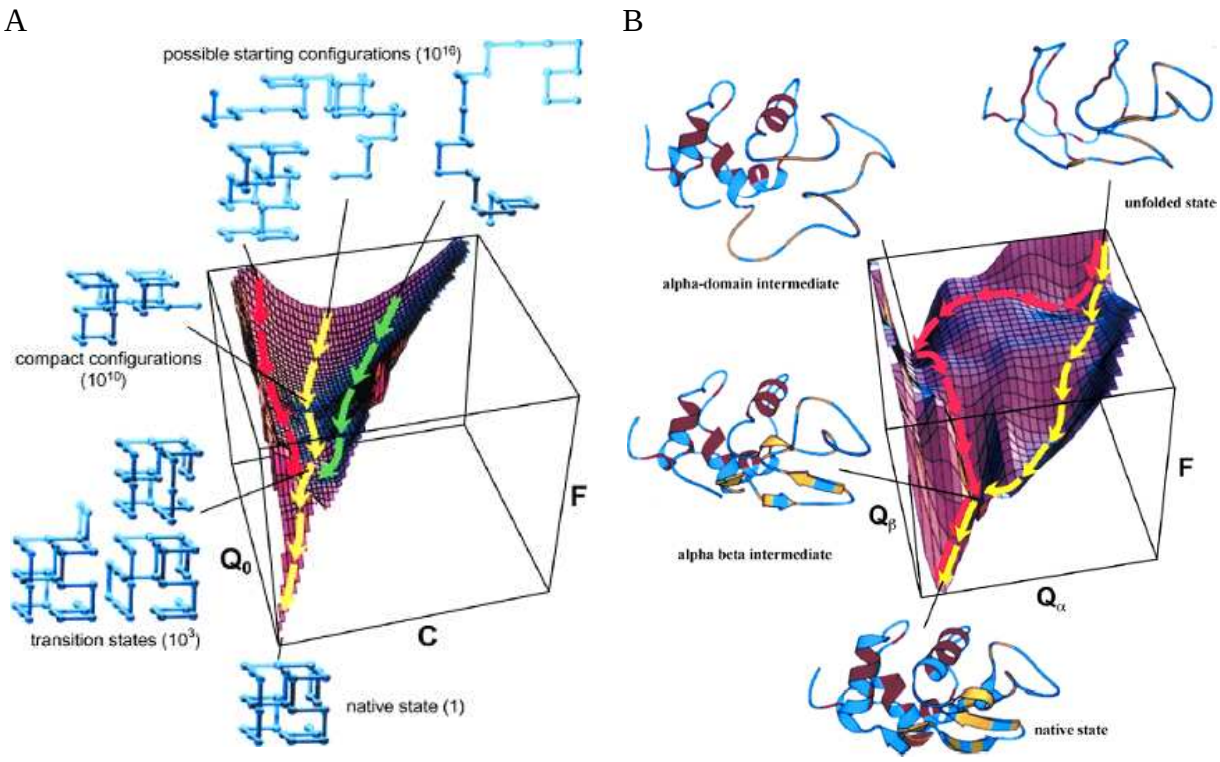


Figure 7 - Folding funnels of lattice model protein (A) and lysozyme (B).
Free energy surface is shown as a function of the total number of contacts and number of native contacts between residues. Adapted from Dobson et al. (78).

This finding is also in concord with the nucleation theory advocated by Shaknovich et al. (79, 80). There the native structure of the protein is formed only after formation of the folding nucleus, which is the transition structure leading to the folding of the native structure of the protein (see Figure 8). The residues from the folding nucleus are those which belong to the hydrophobic core of the protein.

The stability of the protein seems to be also affected by the stability of the hydrophobic core. The mutation of the residues inside the core usually leads to the changes in the stability and/or in the structure of the protein (81, 82). However, other side-chain side-chain interactions were found to be of some importance – such as salt bridges which are more common in the thermophillic counterparts of the mesophilic proteins (66).

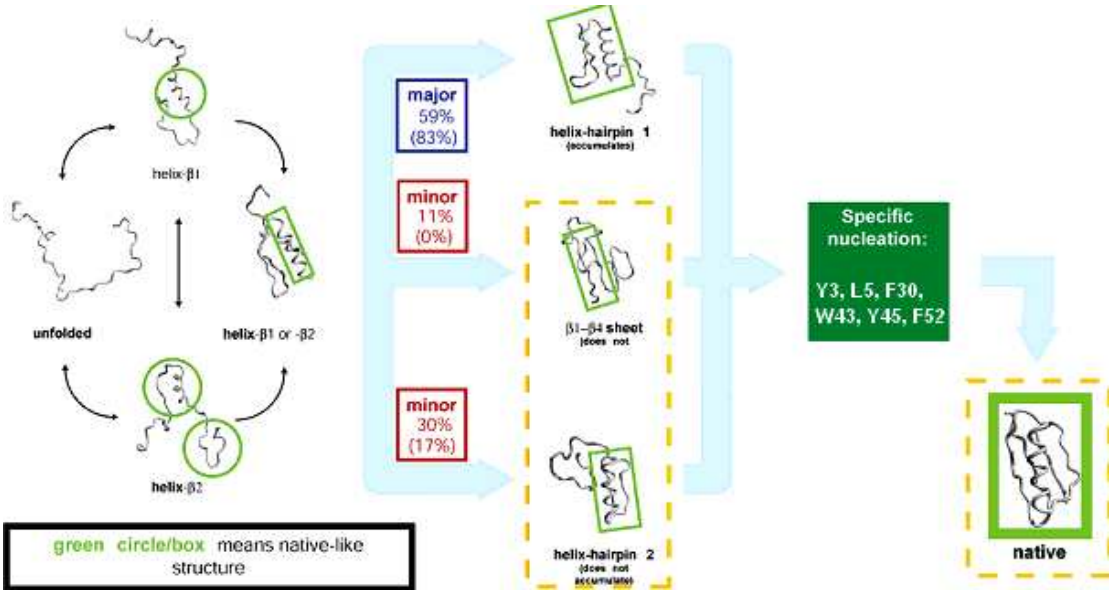


Figure 8 - Mechanism of folding of small protein G (PDB code 1IGD).
Mechanism was derived from all-atom Monte Carlo ensemble folding simulations with the Go potential. Parallel pathways through various helix-hairpin intermediates converge to a common nucleation step that leads to the final folding step. Residues from the nucleation step belongs to the hydrophobic core. Adapted from Shimada, J. et al. (83).

The importance of side-chain side-chain contacts in the latter phase of the protein folding and in the formation of the hydrophobic core were the starting point of this thesis. The work started with the evaluation of the strength of the side-chain side-chain interactions inside the hydrophobic core. It was later enlarged also on the study of the salt bridges and at the end it concluded in the study of all observed side-chain side-chain contacts with the use of the computational methods described in the consequent chapter.

2. Methods

There are two important steps in any computational chemistry workflow and in a protein modeling as well: (a) selection of an appropriate representative model and (b) selection of an applicable computational method providing accurate results in a reasonable time.

2.1. Selection of the Model Geometries

The model system geometry is usually based on coordinates obtained experimentally – either from X-ray structural analysis or NMR experiment. Several geometry models were chosen for the study of different side-chain side-chain interactions.

Studies of interactions inside the hydrophobic core as well as those of salt bridges were based on crystal structures of small protein rubredoxin. Unusually strong interactions of proline with aromatic residues were studied on structures of the Trp-cage protein, and EVH1 and GYF binding domains. The most extensive part of the work was based on geometries from Atlas of Protein Side-Chain Interactions.

2.1.1. Rubredoxin

Rubredoxin from mesophilic organism *Desulfovibrio Vulgaris* (*Df*) was used for the calculations of the side-chain side-chain interactions in the hydrophobic core. *Df* rubredoxin (PDB code 1RB9) is a globular one-domain protein containing a densely packed cluster of residues centered around two phenylalanines (F30 and F49). All residues within a distance of 5 Å around F30 or F49 were cut out from the protein and modeled as a side-chain methylated at the Cβ atom of the side-chain (Figure 9A). Several main-chain hydrogen bonds motifs were selected for comparison of the energy decomposition (Figure 9B). All amino acid residues were treated as neutral.

The same rubredoxin protein family was also selected for the study of the salt bridges strength. Salt bridge coordinates were obtained from structures of hyperthermophilic rubredoxin from *Pyrrococus furiosus* (*Pf*) (PDB code: 1BRF), its mutants (PDB codes: 1BQ9, 1IU5) and its mesophilic counterpart from *Clostridium pasteurianum* (*Cp*) (PDB code: 1SMM). The superimposed salt bridges from the various sources are shown on Figure 10. 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.

A B

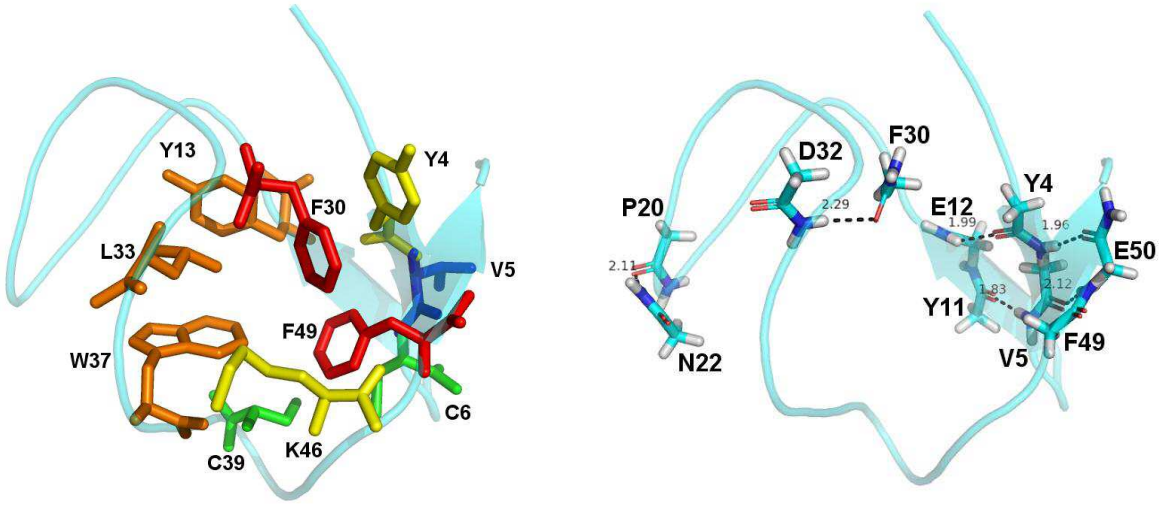


Figure 9 – Visualizations of model systems derived from *Df* rubredoxin (PDB code 1RB9).
(A) Amino acid residues inside the core of. The colors of residues are based on their total interaction energy magnitude. The largest interaction energy (red) is provided by the two phenylalanines F30 and F49, followed by the residues Y13, W37 and L33 (orange), Y4 and K46. The two cysteines C6 and C39 (green) and the valine V5 (blue) provided smallest interaction energy within the studied set.
(B) Hydrogen bonds inside the rubredoxin. Their distances between NH and CO group are shown in Ångströms.

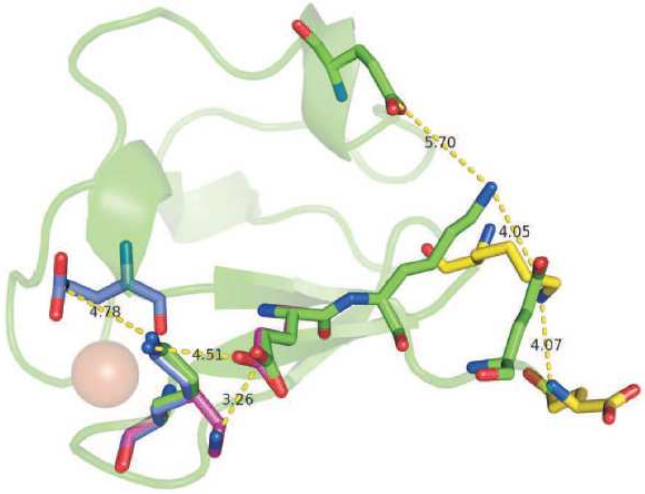


Figure 10 - Structure of rubredoxin from Pyrrococus furiosus (Pf Rd) with salt bridges.
All other rubredoxin structures (1IU5, 1BQ9, 1SMM) were aligned to the structure of wild-type (1BRF, green). Salt bridges differ in color (SB1, blue; SB2, violet; SB3, yellow) from those of the wild-type (SB4–SB6, green). The distance (in Å) between the COO– carbonyl carbon and NH3+ nitrogen is shown for each salt bridge.

2.1.2. Models of Proline Interactions with Tryptophane

For the calculation of proline-tryptophane (PW) interaction, two intramolecular PW motifs were selected from structure of Trp-cage miniprotein (PDB code 1L2Y, see Figure 11).

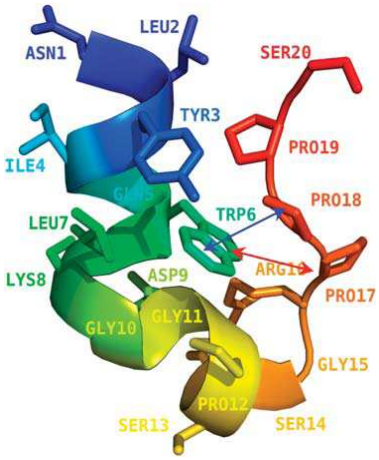


Figure 11 - Structure of Trp-cage miniprotein (PDB code 1L2Y).
The L-shape arrangement of interaction between W6 and P17 is represented by double arrow red line whereas the stacked-like arrangement of W6 and P18 is shown in blue double arrow line.

Two additional models derived from the above mentioned PW motifs were used for evaluation of the proline's nitrogen heteroatom role and the proline's cyclic arrangement role. Replacement of the NH group in proline by CH_2 group resulted in a cyclopentane–tryptophan complex (see Figure 12A). Leucine–tryptophan (LW) complex in stacked-like arrangement was used to evaluate a contribution coming from an acyclic arrangement of the same number of heavy atoms as in cyclopentane. Leucine was modelled only as a side-chain truncated at Cα atom. The structure of the LW complex in stacked-like arrangement was obtained from the Atlas of Protein Side-Chain Interactions (see Figure 12B) (84).

A B

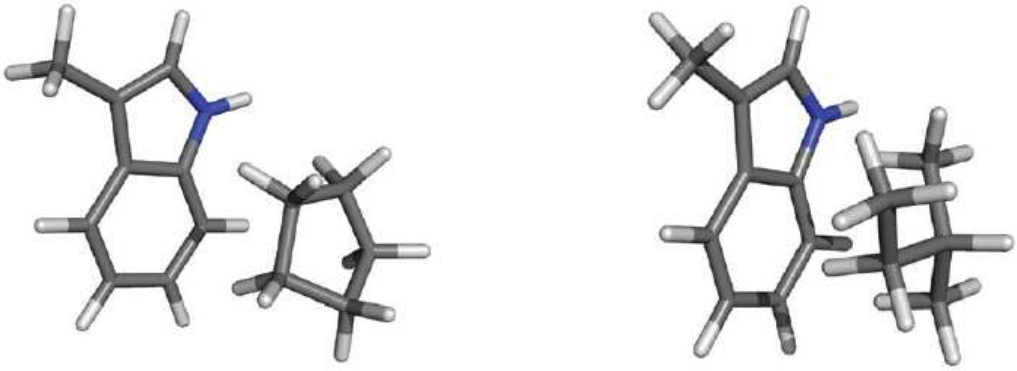


Figure 12 – Structural analogs of the tryptophane...proline.
Tryptophane W6 interacting with cyclopentane (A) based on geometry of W6-P18 pair and WL complex (B) in stacking orientation obtained from the Atlas of Protein Side-Chain Interactions (84).

To further evaluate the strength of intermolecular PW interaction motif found for example in EVH1 and GYF binding domains complexed with proline rich peptides, two X-ray structures (PDB code 1EVH and 1L2Z, see Figure 13) were used. Each proline-tryptophane pair was cut out from the original structure and treated as separate complex.

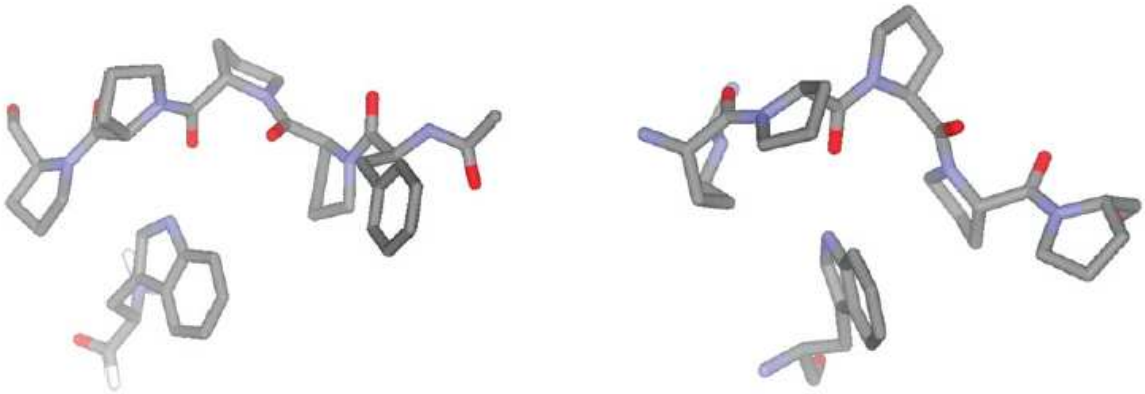


Figure 13 - Models of the intermolecular proline-tryptophane (PW) interactions.
Details of two representative members of protein rich motives (PRM)-binding families with their peptide ligands. Single binding modes known to date for the GYF (left) and one of the binding modes for the EVH1 domains (right).

2.1.3. Atlas of Protein Side-Chain Interactions

The data for the representative set of amino acid side-chains were extracted from a specially updated version of the Atlas of Protein Side-Chain Interactions (84). The online atlas is based on a printed atlas published in 1992 by Singh and Thornton (85) and analyzes the interaction geometries of all 20×20 amino acid side chain pairs as found in experimentally determined 3D protein structures. For each side chain pair, the atlas shows how one side chain is distributed with respect to the other in the space. The preferred interaction geometries are revealed by clusters in the distributions of side-chains around the central residue. The atlas lists the clusters by size and selects a representative side chain pairing for each one.

The atlas is derived using a set of nonhomologous protein chains selected from the structures in the Protein Data Bank (PDB) (86, 87). No two chains have a mutual sequence identity greater than 20%, and the chains are only taken from structures solved by X-ray crystallography to a resolution of 2.0 Å or better. The data in the printed version of the atlas were derived from 62 protein structures, whereas the older online version uses 533 structures (88). For the current study, updated version from October 2006 contained 2548 protein structures total.

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

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

Only a subset of representative structures was used in the first benchmark study. The set covered all important types of side-chain side-chain interactions (Figure 14) and all 20 different amino acid residues. For the later studies we have used either all 20 x 20 representative pairs or even all contacts for selected residues in the Atlas of Protein Side-Chain Interactions dataset.

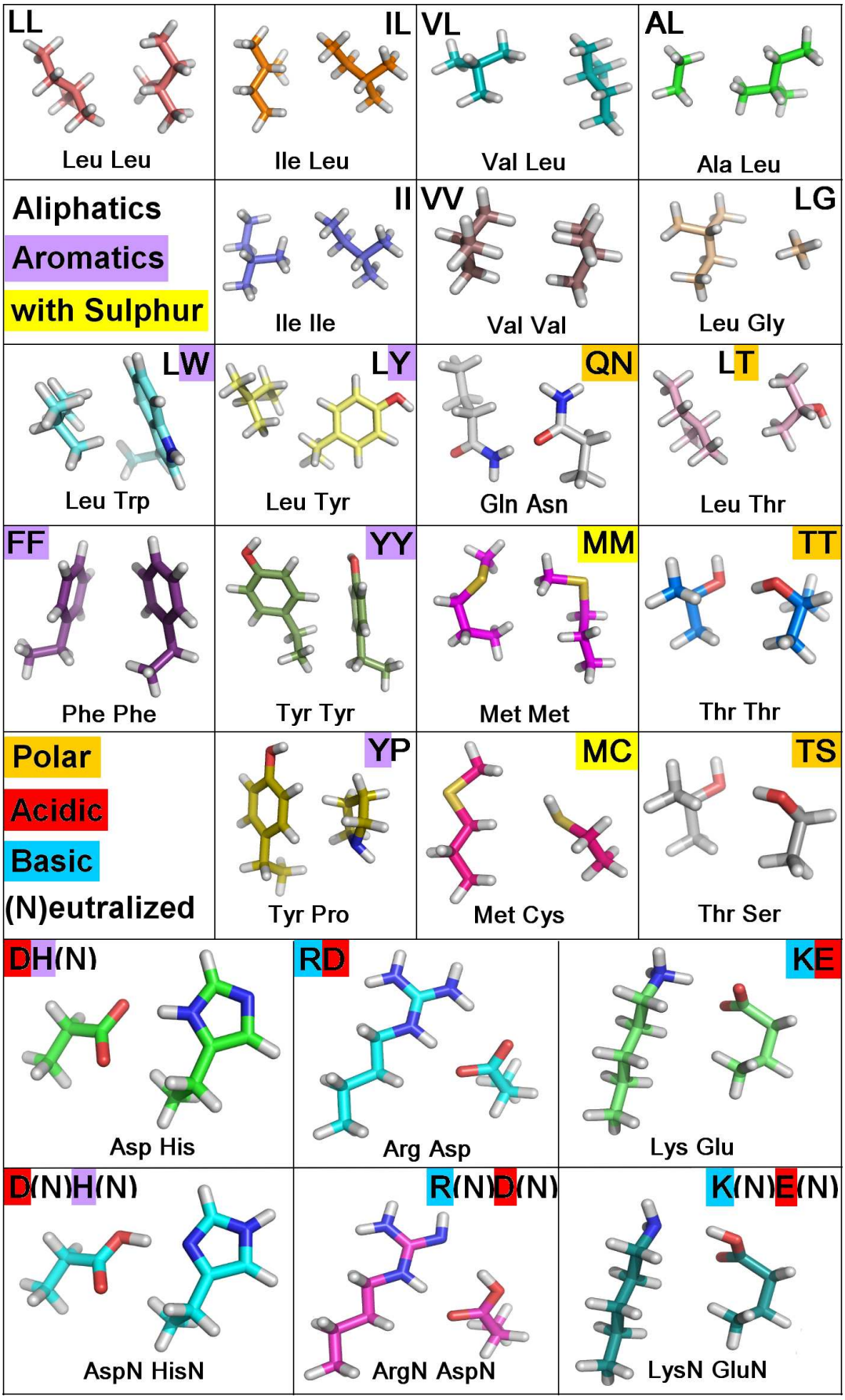


Figure 14 - Geometries of representative set of amino acid side-chain analogs. Side-chains are truncated at Cα atom.

2.2. Selection of Computational Methods

In this thesis, all computational methods are used to evaluate the interaction energy. Interaction energy is defined as the difference between the energy of the complex (E_{AB}) and the sum of energies of the isolated systems ($E_{A,B}$):

$$\Delta E_{\text{int}} = E_{AB} - \left(E_A + E_B\right) \tag{7}$$

The interaction energy cannot be directly assigned to any observable quantity, but it still represents a useful characteristic of the interaction in question. The negative value of the interaction energy is usually referred as the stabilization energy. The calculation of the interaction energy is computationally difficult because the value of the interaction energy is a small number in comparison to the large total energies of the systems. This represents a challenge for the accuracy of the computational method used.

2.2.1. *Ab initio* Methods

Ab initio methods are based on solving the time-independent Schrödinger equation for the system in some representation. Using the Born-Oppenheimer approximation (89), the molecular wave function can be divided into nuclear and electronic parts and be solved separately. With the fixed position of nuclei, the electronic energy can be calculated using the electronic Schrödinger equation (90):

$$\hat{H}_e \chi(r;R) = E_e \chi(r;R), \tag{8}$$

where *r* are all electronic coordinates and *R* are positions of nuclei.

The Schrödinger equation of complex systems cannot be solved analytically and only an approximate numerical solution can be obtained on a wave function expanded in the basis set. The most precise quantum chemical method is the configuration interaction (CI) in an infinite basis set, but this method is extremely expensive and thus cheaper methods are used.

The weakest point of the *ab initio* methods is usually the amount of the correlation energy covered, while this energy is connected with the correlative motions of the electron and it thus have the direct impact on the dispersion interaction defined earlier in the Chapter 1.3.3. Because this interaction is crucial in the determination of the interaction energies, only the methods which are able to calculate the dispersion interaction were used and they are discussed bellow according to their accuracy.

2.2.1.1. Coupled Cluster Method with Complete Basis Set Limit (CCSD(T)|CBS)

A practical route to the complete correlation is the use of the coupled cluster method with variational single and double excitations augmented with perturbative triples (CCSD(T)). This method is based on the description of the wave function as an exponential ansatz (91):

$$\left|\Psi\right\rangle = e^{\hat{T}}\left|\Phi_0\right\rangle, \tag{9}$$

where $|\Phi_0\rangle$ is a Slater determinant usually constructed from Hartree-Fock molecular orbitals . $\hat{T}$ is an excitation operator which, when acting on $|\Phi_0\rangle$, produces a linear combination of excited Slater determinants (singles, doubles, triples and higher orders).

It has been shown that the interaction energies of the stacked as well as hydrogen-bonded model systems calculated at CCSD(T) level were practically identical to those calculated at CCSDT level (92), where all single, double and triple electron excitations are determined iteratively. As the CCSDT energies are close to the full configuration interaction limit (93), this makes CCSD(T) calculations reliable enough to provide benchmark calculations on the interaction energies between biomolecular building blocks such as amino acid side-chains.

For the highly accurate calculation of the interaction energy, the wave function has to be expanded in the basis set as completely as possible to allow the accurate description of the system. The use of the extended basis set is especially important for the calculation of the noncovalent interactions. In order to minimize the error resulting from the usage of the finite basis set, the extrapolation to the complete basis set limit (CBS) can be used thanks to the basis set convergence behavior (94).

It is however impractical to perform CBS extrapolation at the CCSD(T) level as these calculations would be too computationally demanding. Fortunately, the difference between CCSD(T) and MP2 energies show very little basis set dependence unlike these energies themselves (95). Therefore, one can approximate the CCSD(T)|CBS interaction energy ($E_{(CCSD(T)|CBS)}$) as:

$$E_{(CCSD(T)|CBS)} = E_{(MP2|CBS)} + E_{(CCSD(T)|small)} - E_{(MP2|small)}, \tag{10}$$

where $E_{(MP2|CBS)}$ denotes the CBS limit of the interaction energy at the MP2 level and $E_{(method|small)}$ is the interaction energy calculated with the shown method in some smaller basis such as aug-cc-pVDZ or even 6-31G**(0.25,0.15).

Several extrapolation schemes have been proposed for the extrapolation to the MP2 energies to the CBS limit. Since the extrapolation has to be performed systematically, Dunning's correlation consistent polarised basis sets cc-pVxZ (x = D,T,Q,5,6) or their versions augmented with diffuse functions, aug-cc-pVxZ are often utilised (96, 97). The two-point extrapolation scheme suggested by Helgaker et al. (94) is probably the most commonly used today and it is used also in this work.

The described CCSD(T)|CBS procedure solves traditional problems of *ab initio* quantum chemical methods, i.e. the incompleteness of the basis set and insufficient amount of correlation energy incorporated in the computation method.

2.2.1.2. Møller-Plesset Perturbative Treatment (MP2)

Møller-Plesset method (MP) is the post-HF method based on Raleigh-Schrödinger perturbation theory (98). In the MP method, the perturbation represents electron correlation. Since the unperturbed zero-order energy E_{MP0} is Hartree-Fock (HF) energy and the first-order MP energy is zero, at least the second-order MP (MP2) is necessary to improve the HF energy. This leads to the popular MP2 method. The higher-order terms are computationally more demanding. While MP2 method overestimates the correlation energy, more expensive MP3 is underestimating it and only much more expensive MP4 yields good results. Nevertheless, the second–order MP perturbation treatment is widely used, because it is the least expensive wave function method which covers large amounts of the correlation energy.

The evaluation of the two-electron four-centre Coulomb integrals in the Gaussian basis set is a significant component of the overall computational time of many *ab initio* methods such as MP2. The improvement in the computational speed can be obtained with the use of the resolution-of-identity approximation (RI) (also called density fitting (DF)). The basic approach of the RI method is to factorize the four-centre integral into three-centre quantities using a second or "auxiliary" basis set (99). This is formally done by inserting a resolution of identity $1 = \sum_i |i\rangle\langle i|$ into the two-electron integrals:

$$\langle ij|kl\rangle = \sum_t \langle ij|t\rangle\langle t|kl\rangle \tag{11}$$

Owing to the incompleteness of the actual auxiliary basis set, the expansion introduces an error, which should be minimized with use of the bigger basis sets. For example, the RI-MP2 method yields almost identical energies to the exact MP2 method with the time saving being as high as one order of magnitude (100).

Furthermore, interaction energies calculated with small basis sets can be affected by the basis set superposition error (BSSE). This error is the artificial result of the finite basis set size. It is caused by the different number of basis functions used for the description of the wave functions for the complex and for the monomers. This leads to the better description of the complex in comparison with the monomers, yielding artificially increased binding energy. This effect vanishes asymptotically as the complete basis set limit is approached.

The easiest way to eliminate BSSE error is the use of the counterpoise correction (CP) method of Boys and Bernardi (101). In this method, monomer energies are calculated in the basis set of the whole complex by introduction of "ghost orbitals" in the positions of the absent atoms. The interaction energy can be then simply calculated as:

$$\Delta E_{\text{int}} = E_{AB}(AB) - E_A(AB) - E_B(AB) \tag{12}$$

where E_{AB} is the energy of the complex AB, and E_A and E_B are the energies of the monomers. The parentheses denote that the basis set for the whole complex is used in all cases.

2.2.2. Density Functional Theory (DFT)

The density functional methods (DFT) provide a useful alternative to wave-function methods. They are usually computationally less demanding because the many-body wave function of the system is not calculated. Instead, the energy of the system is calculated as a functional of the electronic density. This leads to Kohn-Sham equations, which are solved iteratively like HF equations (102). The energy of the electron gas can be expressed as a functional of the electron density:

$$E[\rho] = T[\rho] + \int V_{ext}(\vec{r})\rho(\vec{r})d\vec{r} + V_H[\rho] + E_{xc}[\rho], \tag{13}$$

where T is a kinetic energy of the electron gas, V_{ext} is an external potential acting on the system, V_H is the Hartree energy and E_{xc} is the exchange-correlation energy, which includes terms accounting for both exchange energy and the electron correlation. The exact exchange-correlation functionals are not known except for the free-electron gas. There are several DFT approximations which differ in evaluation of E_{xc}.

The Generalized Gradient Approximation (GGA) calculates exchange-correlation energy from the electron density and its gradient. Most popular GGA exchange-correlation functionals used now are PBE (103), PW91 (104), and B-LYP (105, 106) functionals. However, the GGA functionals are local and for this reason they are unable to describe the long-range correlation effects, such as dispersion.

For this reason *meta*-GGA functionals have been proposed to contain more semi-local information, such as the kinetic energy density, higher order density gradients or gradients of the Kohn-Sham orbitals. Even though the improvement of the functional form is there, they still have trouble with the description of the nonlocal phenomena. The most widely used meta-GGA functionals are TPSS (107) and M06-L (108).

Another way of improving the DFT performance is to incorporate a fixed amount, typically 20-25%, of the exact Hartree-Fock type exchange to the usual density functional exchange. These functionals are called hybrid exchange-correlation functionals. Although the hybrid functionals usually perform better in the description of long-range interactions, there is still room for development. The most popular and widely used hybrid functional is B3LYP (109, 110); other prevalent functionals from this group are PBE1 (111), PBE0 (112) and M06 suite of functionals. (113).

2.2.2.1. Density Functional Theory with Empirical Dispersion Term (DFT-D)

Even simpler way to improve the performance of the DFT methods in the systems where the nonlocal effects play crucial role is the augmentation of the functional with the empirical dispersion term. The resulting DFT-D method then calculates the energy as a sum of the DFT energy and damped dispersion term. One of the possible forms of the dispersion term is:

$$E_{disp} = f(R)\frac{C_6}{R^6}, \tag{14}$$

where $f(R)$ is a damping function, C_6 is the dispersion coefficient and R is the interatomic distance.

The first succesfull DFT-D method was proposed by Grimme (114), where the damping function scaled the dispersion coefficients, which leads to wrong asymptotic behavior of the dispersion term at very long distances. Another damping function was proposed by Jurečka et al (115), which scales the atomic radii. This form of the dispersion was parameterized on the

CCSD(T)|CBS values on the S22 set (116) so it provide reliable characteristics for the isolated systems as well as for the H-bonded and dispersion-bound complexes (115).

Moreover, the dispersion energy determined by the C_6/R^6 expression agrees surprisingly well with the dispersion contribution calculated with the SAPT method (117, 118). Another advantage of the DFT-D is its favorable computational cost; it can be used for extensive biomolecular systems. The cost of the calculation can be further reduced by applying the resolution-of-identity (RI) approximation.

The combination of TPSS functional (107) and TZVP basis set augmented with empirical dispersion term from Jurečka et al. (115) was used as a RI-DFT-D method in this thesis.

2.2.2.2. Density Functional Tight Binding (DFTB)

The density functional tight-binding scheme (DFTB) (119, 120) is based on a second-order expansion of the Kohn-Sham total energy in DFT with respect to charge density fluctuations and it can be seen as the generalization of the tight-binding method (121). The DFTB method was later modified by a self-consistent redistribution of Mulliken charges (SCC) (122).The SCC charges are used calculate the long-range electrostatic interactions between the point charges at different sites and to include the self-interaction contributions of individual atoms.

The approximate DFT energy functional is given by:

$$E_{total} = \sum_{i}^{occ.} \sum_{\mu\nu} c_{\mu}^{i} c_{\nu}^{j} H_{\mu\nu}[\rho_0] + \frac{1}{2} \sum_{\alpha\beta} \Delta q_{\alpha} \Delta q_{\beta} \gamma_{\alpha\beta} + E_{rep}[\rho_0] \quad (15)$$

where the Hamiltonian matrix elements $H_{\mu\nu}[\rho_0]$ are calculated with GGA functional in a two-center approximation using a minimal basis of atomic-like wave functions. The second term represents long-range Coulomb interactions between point charges at different sites and includes self-interaction contributions of individual atoms. The last term represents the repulsion energy and is approximated as a sum of short-range two-center terms fitted into the ab initio calculations.

The dispersion is calculated in the SCC-DFTB-D scheme by an empirical dispersion term which is added to the SCC-DFTB total energy in the same way as in the DFT-D approach. The diatomic C_6 coefficients are calculated using the Slater–Kirkwood combination rule and

the $1/R^6$ dependence is truncated for small interatomic distances using an appropriate damping function (123).

Due to these extensions, the SCC-DFTB-D method is faster and still quite reliable method for the calculation of biomolecules and due to these characteristic it can be used for the *ab initio* dynamics (124).

2.2.2.3. Symmetry-Adapted Perturbation Theory with DFT (DFT-SAPT)

Another method which is based on the perturbation theory is the Symmetry-Adapted Perturbation Theory (SAPT) (125). In SAPT, the total Hamiltonian for the dimer is partitioned as:

$$\hat{H}_0 = \hat{F} + \hat{V} + \hat{W}\,, \tag{16}$$

where $\hat{F}$ is the sum of the Fock operators for monomers A and B, $\hat{V}$ is the intermolecular interaction operator, and $\hat{W}$ is the sum of the Møller-Plesset fluctuation operators an as such it is intramonomer correlation operator. The interaction energy, E_{int}, is expanded as a perturbative series

$$E_{\text{int}} = \sum_{n=1}^{\infty}\sum_{j=0}^{\infty}\left(E_{pol}^{(nj)} + E_{exch}^{(nj)}\right), \tag{17}$$

where the indices n and j are denoting the orders in the operators $\hat{V}$ and $\hat{W}$, respectively.

SAPT method is thus able not only calculate interaction energies with high accuracy but it also allows the decomposition of the interaction energy into physically meaningful components (electrostatic, exchange, induction and dispersion terms):

$$E_{\text{int}}^{SAPT} = E_{elst}^{(1)} + E_{exch}^{(1)} + E_{ind}^{(2)} + E_{disp}^{(2)} + \delta HF\,, \tag{18}$$

The exponents in equation (18) refer to the perturbation order with respect to the intermolecular operator $\hat{V}$. δHF denotes the estimate for higher-order contributions.

However SAPT method is unfortunately quite expensive with $O(N^7)$ scaling and the calculation of the interaction energies between two bigger systems are beyond the reach of the classical SAPT method. For this reason, the SAPT version with DFT description of the monomers has been introduced (126, 127). This method is called either SAPT(DFT) or DFT-

SAPT, while it is much cheaper with $O(N^6)$ or even $O(N^5)$ scaling with use of density fitting procedure (128).

However, it is necessary to circumvent the common failure of DFT methods to describe correctly the dispersion interaction. This drawback, which occurs due to the wrong long-range behavior of electron densities in commonly used exchange-correlation potentials, can be solved by an asymptotic correction to the exchange-correlation potential. Furthermore, the DFT method is only used for the description of isolated monomers and interaction energies are calculated at a higher level. DFT-SAPT provides similar accuracy to high-level wave function based methods with extrapolation to the complete basis set limit (129).

2.2.3. Semiempirical Methods

Semi-empirical quantum chemical methods are based on *ab initio* methods, but they use additional approximations and parameters from empirical data or from the fit into the higher level *ab initio* calculations. The use of empirical parameters allows some inclusion of electron correlation effects. Semi-empirical calculations are much faster than their *ab initio* counterparts. Their results, however, can be wrong if the molecule being computed is of a different type than the molecules used to parameterize the semiempirical method.

2.2.3.1. Parameterized Model 6 with Corrections (PM6-DH)

Semiempirical method Parameterized Model 6 (PM6) was introduced recently by Stewart (130). It is a method based on the neglect of non-bonded differential overlap (NNDO) improved by the adoption of Viotyuk's core-core diatomic interaction term (131) and Thiel's d-orbital approximation (132). These modifications allowed parameterization of 80 elements and also reduced the error for main group elements (133). However, the PM6 method fails for the description of noncovalent interactions, specifically the dispersion energy and H-bonding.

For this reason, the improvement of noncovalent interactions was done in two directions: (i) the addition of an empirical dispersion energy term that improves the description of complexes controlled by the dispersion energy and (ii) the introduction of an additional electrostatic term that improves the description of hydrogen-bonded complexes. The accuracy of the resulting method, PM6 with corrections for dispersion and hydrogen bonding (PM6-DH), is close to that of correlated ab initio methods on a multiple sets of high-quality benchmark data (134).

2.2.4. Methods Using Empirical Potential

In the most of empirical potentials, the interaction energy can be calculated as a sum over three noncovalent terms – electrostatic, repulsion and dispersion terms between side-chains. The electrostatic interaction energy is calculated as a finite sum over all possible pair-wise electrostatic energies between atoms on both residues using Coulomb's law:

$$E_C = \sum_i^{res1} \sum_j^{res2} \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}}, \tag{19}$$

where ε_0 is the permittivity of vacuum, q_i is the partial charge on atom i, and r_{ij} is an interatomic distance.

Dispersion and repulsion interactions are usually summed up in Lennard-Jones interaction energy term. The interaction energy coming from this term is similarly defined as a sum over all pair-wise Lennard-Jones energies. Lennard-Jones interaction energy can be expressed in two different representations (20):

$$E_{LJ} = \sum_i^{res1} \sum_j^{res2} \left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}} \right) = \sum_i^{res1} \sum_j^{res2} \varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right], \tag{20}$$

where A is repulsion coefficient, B is dispersion coefficient, r_{ij} is interatomic distance, ε is potential depth, and σ is the finite distance, at which the Lennard-Jones potential is zero.

The molecular mechanical empirical potential (force field) also contains bonded terms in the potential, but these terms do not apply in the case of nonbonded interactions. The parameterization of a force field is usually performed by fitting force field parameters to the experimental results and high-level quantum chemical calculations.

The advantage of force field methods is their speed, because they are several orders of magnitude faster than quantum mechanics methods. However their accuracy is highly dependent on their parameterization. For instance, the polarization effects cannot be fully captured with the point charge representation in the normal force field but only with the polarizable force fields, which are on the other hand slower.

Force field methods used in this thesis were modified force fields parm03 and OPLS-AA/L.

2.2.4.1. Optimized Potential for Liquid Simulations (OPLS-AA/L)

OPLS-AA/L (39) force field is a re-evaluated version of OPLS-AA (135) force field. This family of force fields was developed closely with the Amber family of force fields as it used dihedral torsion parameters from the Amber Cornell et al. force field (35), but these were in the later version of OPLS force field reparameterized. Parameters of nonbonded interactions have also been refitted, and the validity of new Coulombic charges and van der Waals parameters were proved through reproducing gas-phase energies of complex formation, heats of vaporization and densities of pure model liquids.

Our modification of OPLS-AA/L force field was only the truncation of the residue at the Cα (or Cβ) atom to provide side-chain atoms only. The terminal Cα (or Cβ) methyl group were assigned the same atomic types and partial charges as the other methyl groups in the OPLS-AA/L force field.

2.2.4.2. Parm03

Duan et al. parm03 (36) force field originates from the family of Cornell et al. parm94 (136) and Wang et al. parm99 (64) force fields and is implemented in the Amber molecular dynamics package (137). This force field was parameterized specifically for the amino acid residues and proteins with fitting of the partial charges by RESP method on the grid calculated with B3LYP/cc-pVTZ//HF/6-31G** method within continuum solvent with an effective dielectric constant of $\varepsilon = 4$.

Our modification of parm03 force field was the truncation of the side-chain topology at the Cα or Cβ atoms. The partial charges and Lennard-Jones parameters of the original atoms were left unchanged. The newly added hydrogens on the truncated atom were assigned such partial charge to provide the integral charge on entire residue.

2.3. Solvent Models

Since the most chemical processes take place in the solvent, our study was extended to the modeling of the influence of the environment on the interaction energy. The solvent can be modeled explicitly but this method considerably increases the computational demands of the calculation. The other way around is to use an implicit, or continuum solvent model (138). In this approach, the solvent is modeled as a bulk medium or a continuum surrounding the studied system.

The solvation models evaluate the solvation free energy ΔG_{solv}, which equals to the free energy of the transfer of the solute molecule from vacuum to the solvent. The ΔG_{solv} can be decomposed into several contributions:

$$\Delta G_{solv} = \Delta G_{polarization} + \Delta G_{dispersion} + \Delta G_{repulsion} + \Delta G_{cavitation}, \quad (21)$$

where the individual terms are electrostatic, dispersion, repulsion and cavitation contributions, respectively.

As this thesis focuses only on the interaction energies between the side-chains, only the change in the interaction energy was studied. The influence of the solvent on the interaction energies between side-chains is mostly given by the change in the electrostatic component. Cavitation term is in this case unusable, as the cavity is not given by the side-chains pair but by the complete protein and thus the cavitation term was omitted from the investigation.

2.3.1. *Ab initio* Solvent Models

Dispersion, repulsion and cavitation terms are often combined in implicit solvent models and they are usually estimated to be linearly proportional to the solvent-accessible surface area of the solute based on the experimentally determined free solvation energies. The electrostatic contribution is very important for charged and polar solutes, due to the polarization of the solvent. It is evaluated through the solvent being modeled as a uniform medium with the dielectric constant - ε. These calculations are usually based on the models derived by Born (139) and Onsager (140).

In the Onsager (140) model, the dipole of the solute induces a dipole in the surrounding medium, which in turn induces an electric field in this cavity (a reaction field). This model can be used in combinations with the *ab initio* calculation. The interactions of the solvent

reaction field with the solute dipole are considered as the perturbation of the Hamiltonian of molecule. The disadvantage of this method, referred to as the self-consistent reaction field (SCRF), is the use of a spherical cavity for the solvent.

A more realistic cavity shape based on the van der Waals radii of the atoms of the solute is used in the polarizable continuum method (PCM) (141). Unlike in the SCRF method, the electrostatic term has to be evaluated numerically. The cavity surface is divided into a large number of surface elements, and a point charge representing the solvent polarization is associated with each element.

The Conductor-like Screening Model (COSMO) (142) of a solvent is a variant of the PCM method, where the cavity is considered to be embedded in a conductor with an infinite dielectric constant. The potential on the surface of the conductor is set to zero, which gives a convenient boundary condition for the determination of the surface charges. Its implementation in the quantum chemistry codes might differ, because while Gaussian version called C-PCM (143) calculates also a cavitation term, Turbomole COSMO version did not calculate the cavitation term at all and focuses only on the electrostatic component (144).

2.3.2. Molecular Modeling Solvent Models

The implicit solvent models are also used in molecular mechanics calculations. The simplest model divides the electrostatic Coulomb term with the dielectric constant. Another simple model uses the dielectric constant linearly dependent on the distance. More realistic implicit solvent models are Poisson-Boltzmann model (PB) and the generalized Born model (GB) are much more complicated.

The former PB approach uses Poisson-Boltzmann equations for the solute in the ionic solvent. These equations are rather complex and they can be solved only slowly numerically (145). The latter GB method uses an approximation of the Poisson-Boltzmann equation. Solute is represented as a set of particles with charges and effective Born radii. The effective Born radius of an atom characterizes its degree of burial inside the solute, which has different dielectric constant than the solvent; qualitatively it can be thought of as the distance of the atom from the molecular surface. Accurate estimation of the effective Born radii is critical for the GB model (146).

3. Aims of the Thesis

In the present thesis we tried to answer following questions concerning side-chain side-chain interactions in proteins.

1. *How strong are interactions inside the hydrophobic core of a protein?*
2. *How strong are other stabilizing interactions in proteins, i.e. in salt bridges?*
3. *What is the reason for the unusually strong interactions of proline with residues of aromatic character?*
4. *Which computational methods have reasonable efficiency and accuracy for interaction energy calculations?*
5. *What can we learn from the energy decomposition by means of SAPT method about interaction energies in proteins?*
6. *How diverse can be side-chain side-chain interactions in proteins?*
7. *How well the generally used force fields describe interaction energies between side-chains?*
8. *How do interaction energies change upon the presence of a solvent?*
9. *How are interaction energies between amino acid side-chains distributed in proteins and what is the meaning of the representative pairs selected in Atlas of Protein Side-Chain Interactions?*

4. Results

4.1. Interactions within the Hydrophobic Core

The hydrophobic core of a typical globular protein consists of tightly arranged residues mostly of the hydrophobic character in the protein interior. The residues in the hydrophobic core are usually better resolved than the rest of the residues in the protein. Such behavior suggests that the hydrophobic core could be stabilized by the forces of the enthalpic nature and not only by the hydrophobic effect. This leads to the question of the source of such forces, because the usual stabilizing interactions such as the hydrogen bonds or salt bridges are usually not present in the hydrophobic core. Recent findings suggest that there are other noncovalent interactions playing the important role in the stabilization (147).

To address a question of stabilizing forces inside the hydrophobic core of a protein, the model based on an arrangement of two amino acids inside the small protein rubredoxin was selected. The core is built around two phenylalanine residues occupying interior of the protein. All interaction energies for the side-chains in a direct contact with either of the residues have been calculated by the DFT-SAPT method, which not only gives the interaction energy with a reasonable accuracy, but it also decomposes the interaction energy into physically valid terms.

The strongest contributions to the overall stabilization of the core come from interactions of aromatic residues F30, F49 and W37, followed by the aliphatic residue L33 (see Figure 15) with the average stabilization energy around 3 kcal/mol per residue. Most of the stabilizing energy originates in the dispersion term and it is about as 2.8 times stronger than the electrostatic energy term. As can be seen from Figure 15, the profiles of the total energy and of the dispersion energy are very similar. This emphasizes that the dispersion is dominant force in the tight arrangement of the hydrophobic core.

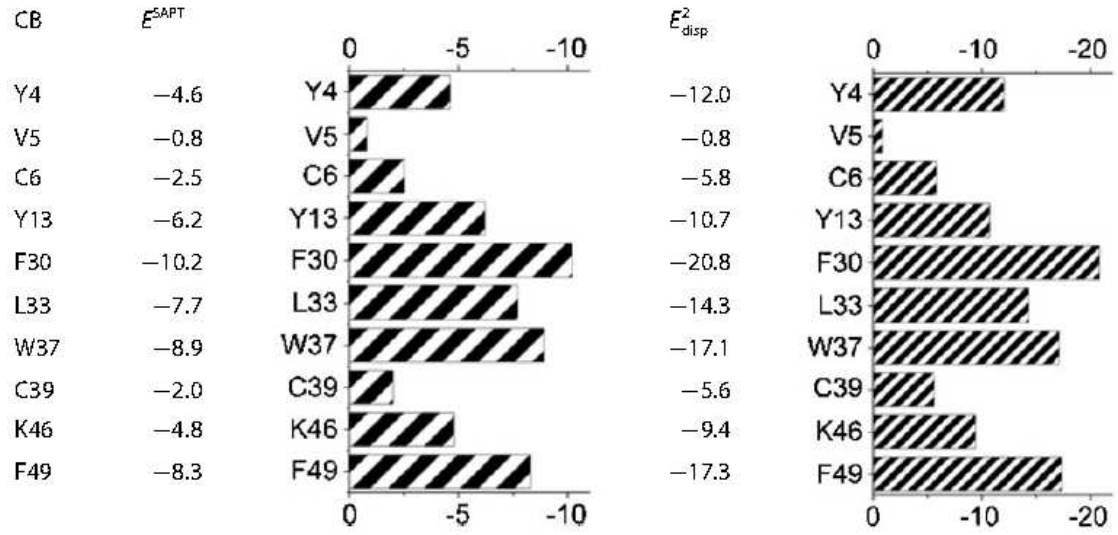


Figure 15 – The profiles of the total interaction energy and of the dispersion energy.
E^{SAPT} – total interaction energy, E^2_{disp} – dispersion energy. Note the similar pattern in both profiles.

Another important fact was that all of the interactions were attractive. One would easily assume due to the fact that the hydrophobic effect is the strongest determinant of the protein compactness, there would be non-ideal or even repulsive interactions present. However, no repulsive interaction was found. This fact ascertains that the structure of the hydrophobic core is finely tuned by the dispersion interactions in such arrangements in which side chains are perfectly complement to each other. For further details see Appendix A.

4.2. Salt Bridges

While the first study was focused on the evaluation of interaction energies between side-chains inside the hydrophobic core of rubredoxin, the second study concentrated on salt bridges. They are thought to provide higher thermostability for rubredoxin family (148) and for thermophilic proteins generally (67). For this reason, six different salt bridges have been selected from the mesophilic as well as thermophilic rubredoxins and their interaction energies were evaluated (Appendix B).

As follows from the calculations, the interaction energies of the salt bridges in the gas phase are well described by the DFT-D approach and their values are around 100 kcal/mol. Similar values can also be obtained by the Cornell et al. force field (64). The interaction energy originated almost exclusively in the ionic interaction between the opposite charges and its strength was almost linearly proportional to the reciprocal distance of the side-chains involved.

These enormous high interaction energy values however substantially weakened when the implicit solvation models was used to model the environment. The salt bridge stabilization energies dropped to about 20% in the protein-like environment whereas salt-bridges introduced to the water environment shown even destabilization. The realistic magnitude of interaction energies for salt bridges is expected to lie between protein and water environments depending on the level of the salt bridge burial from the protein surface. All tested solvent models have similar destabilization behavior.

Not only the change of the environment, but also the change of the pH had a large impact on the stabilization energy of the salt bridge. The neutralization of either cation or anion has lead to the weakening of the interaction energy in all environments.

The role of the charged residues in the protein stabilization is thus probably different from the simple conception of the salt bridges as the stabilizing element. The salt bridge can be significantly destabilized in the water environment depending on the arrangement. However, the role of charged residues would be rather in protection against intermolecular proteins' aggregation and in the shielding of the hydrophobic core from the bulk water.

4.3. Proline Interactions

The thermostability of a protein can be also altered according to the "proline rule". It states that the thermostability of proteins can be increased by the addition of proline (P) amino acid residues at specific positions (149). The reason is that proline restrains movements of the main-chain. Increase of the protein rigidity lowers the conformation entropy. Moreover, it was shown by several studies showed that proline interactions with neighboring residues can be extraordinarily strong (150-152).

To address the question of the importance of such interaction for stabilization, the intramolecular interaction of proline (P) in Trp-cage protein was studied. Trp-cage is a small protein (PDB code: 1L2Y) with the central tryptophane (W) residue and with two tryptophane-proline binding motifs. One interaction motif between residues W6 and P17 is in geometry of an L-shape, while the other motif (W6-P18) has stacking conformation (see Figure 16). According to Bendová-Biedermannová et al. both arrangements are of the comparable interaction energy power (151).

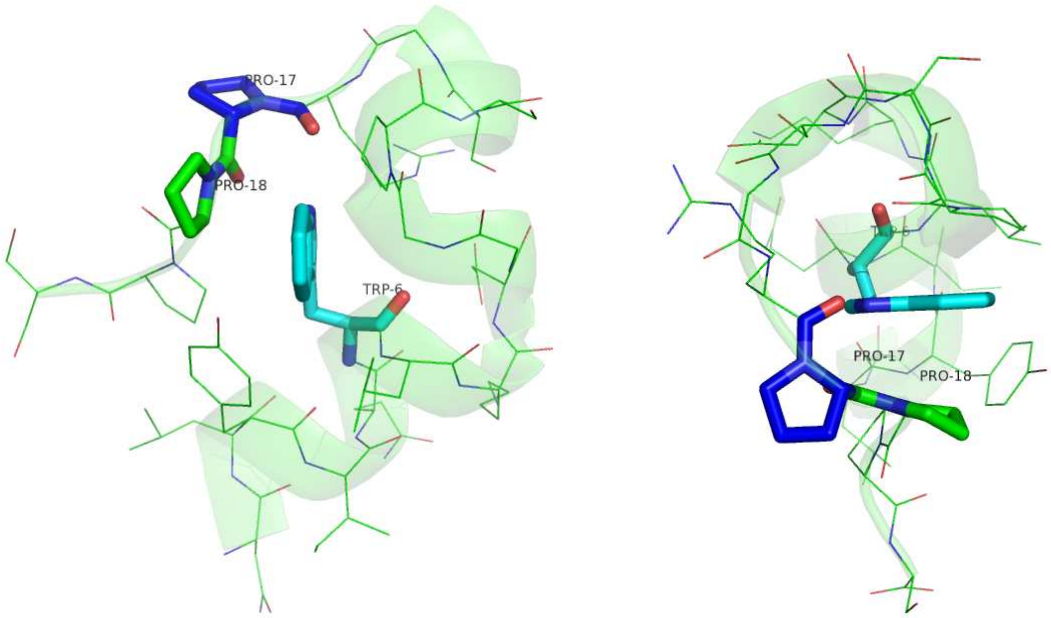


Figure 16 – The structure of the Trp-cage protein with highligthed WP binding motifs. Projections from side (left) and from above (right) are shown. The residues are shown in the large model. L-shaped binding motif W6-P17 is colored in blue and stacked motif W6-P18 in green.

The interaction energies of the binding motifs were calculated at MP2, DFT-D, DFT-SAPT and CCSD(T) levels in two different geometry models (Table 2). The large model is represented by the complete residue including the peptide bond, the small model comprise only side-chain representation of the tryptophane and heterocyclic pyrrolidine representing the proline.

While the interaction energy of the L-shaped binding motif W6-P17 is maintained by the hydrogen bond, the stacked binding motif W6-P18 is apparently stabilized by different mechanism.According to DFT-SAPT calculations, the strongest interaction energy term is clearly the dispersion. This is rather surprising fact because proline is not the aromatic residue and its dispersion interaction should be smaller.

The proline model was modified in our model in order to find the source of the relatively strong interaction energy. The small model of proline was a pyrrolidine containing one heteroatom in the cyclic arrangement. To investigate the role of the heterocyclic atom for the interaction, the nitrogen atom in pyrrolidine was replaced by the carbon. It resulted in cyclopentane molecule. To explore the role of the cyclic arrangement the acyclic leucine (L) was placed instead of the proline with similar pair-wise arrangement with tryptophane. The LW geometry was obtained from Atlas of Protein Side Chain Interactions (84).

Table 2 – Stabilization energies from various methods for all proline complexes.
The L-shaped and stacked-like arrangements of proline complexes in various models of the interaction calculated using CCSD(T), MP2, DFT-D and SAPT methods, and the individual components of the SAPT stabilization energy.

	L-shaped		Stacked-like			
	large model	small model	large model	small model	tryptophane-cyclopentane	tryptophane - leucine
MP2[a]	7.8	0.9	8.4	6.5	5.1	4.1
DFT-D[b]	7.6	1.1	6.8	5.4	4.0	2.9
CCSD(T)[c]	N/A	0.9	N/A	6.0	4.6	3.8
SAPT[d]	7.3	0.8	6.9	5.3	3.7	3.3
E^1_{elst}[e]	9.9	0.2	5.6	3.4	2.0	1.5
E^1_{exch}[e]	-11.2	-0.3	-9.2	-5.6	-5.6	-3.5
E^2_{ind}[e]	3.7	0.1	1.0	0.5	0.3	0.2
E^2_{disp}[e]	4.5	1.2	8.8	6.6	6.8	4.7
δHF[e]	1.8	0.0	0.7	0.4	0.4	0.3

a) MP2/aug-cc-pVDZ. b) DFT-D/TPSS/TZVP. c) CCSD(T)/CBS. d) DFT-SAPT/aug-cc-pVDZ. e) The electrostatic, exchange, induction, dispersion, and higher order contribution terms to the SAPT stabilization energy, respectively. All energies are in kcal/mol.

Both changes of the proline model lowered the stabilization energy (see Table 2). The change of the nitrogen to the carbon atom lowered the stabilization energy by 1.5 kcal/mol. This decrease was decoded by DFT-SAPT analysis as a loss in the electrostatic energy term and it is most likely connected with the change of the molecular dipole. The further modification from the cyclic to acyclic arrangement correlates with the additional loss in interaction energy by approximately 1 kcal/mol. This change can be attributed to the loss in the dispersion contribution as there is one carbon atom missing from the direct contact to tryptophane.

To summarize above described findings - the large interaction energy between proline and tryptophane in the stacked arrangement can be attributed to the favourable electrostatic interaction due to the nitrogen atom and to the facilitation of the close contact due to the cyclic arrangement.

Both binding motifs are intramolecular interactions of proline within the Trp-cage protein. However the unusually strong interactions of the proline were described earlier by Riley et al in their study of the protein-ligand interactions (152). To explore the strong interaction of proline in other intermolecular complexes, we found two representative systems where the polyproline sequence is in the contact with the tryptophane – W28 in the GYF domain and W23 in the EVH1 domain (Figure 13) and both interaction motives are present.

The pair-wise interaction energies of each tryptophane with each proline calculated at DTF-D level showed that the strongest interaction is brought by the L-shape arrangement (around 8 kcal/mol). This interaction was present in both studied complexes. The stacking arrangements of WP pair were only slightly weaker than the L-shape one (around 5 kcal/mol). For further details see Appendix C. Interesting fact is that the binding motives of proline with tryptophane are similar in the intramolecular as well as in the intermolecular case.

4.4. Representative Set of Interactions

The interaction energies between side-chains were studied so far only for specific cases. None of these studies provided a comparison of side-chain side-chain interaction strengths for complete set of 20x20 combinations of interacting residues. The Atlas of Protein Side-Chain Interactions (84) represents a set of 20x20 all possible side-chain side-chain interaction combinations. To calculate all the interaction energies by the benchmark *ab initio* CCSD(T)|CBS method is almost impossible. Therefore a selection of cheaper, but still accurate method is of utmost importance.

The set of 24 side-chain pairs was selected representing typical interactions in proteins (see Figure 14 in Methods section), e.g. aliphatic-aliphatic, aliphatic-aromatic, aromatic-aromatic, polar-polar, aromatic-charged, and charge-charge interactions. The representative set contained all 20 amino acid residues.

The interaction energies for all pairs were calculated in the gas phase by different methods and they were compared with CCSD(T)|CBS benchmark values. For present side-chain side-chain pairs, a high degree of agreement was detected between different methods (see Table 3), even though the range of interaction energies was extremely large – over two orders of magnitude (153).

The RI-DFT-D (154) was found to be the most effective method reasonable level of accuracy. Much cheaper semiempirical methods PM6-DH (134) or SCC-DFTB-D (123) performed noticeably worse, but they still performed better than force field methods parm03 (36) and OPLS-AA/L (39). Both tested force fields showed similar behavior, while the parm03 showed a slightly better accuracy than OPLS-AA/L. For more details see Appendix D or an online database storing benchmark energies and geometries of various non covalent complexes (BEGDB) – www.begdb.com (155). For its description in more detail see Appendix E.

Table 3 - Interaction energies for amino acid pairs calculated with several methods

Code	CCSD(T) CBS	RI-MP2 aDZ	RI-MP2 aTZ	SCS(MI)-MP2 cc-pVTZ	DFT-SAPT aDZ	DFT TPSS TZVP	RIDFT-D TPSS TZVP	OPLS	parm03[a]	SCC-DFTB-D	PM6-DH
RD	**-110.80**	-109.37	-110.21	-111.71	-107.52	-110.60	-112.73	-105.71	-90.37	-105.32	-103.91
KE	**-108.40**	-107.36	-107.75	-105.64	-105.78	-108.27	-110.86	-106.02	-103.57	-104.18	-101.73
DH	**-30.64**	-29.88	-30.91	-31.06	-28.35	-28.83	-31.30	-12.20	-22.36	-26.56	-27.14
DH(N)	**-17.97**	-16.81	-17.94	-17.68	-16.05	-16.26	-19.03	-10.90	-7.80	-9.24	-12.2
RD(N)	**-16.32**	-15.29	-15.92	-16.18	-14.68	-14.71	-17.01	-8.94		-11.12	-17.87
KE(N)	**-10.76**	-10.36	-10.65	-10.50	-9.87	-9.81	-12.51	-8.80	-9.11	-5.66	-10.68
QN	-7.37	-6.41	-6.92	-7.06	-6.83	-5.66	-7.31	-8.61	-8.84	-6.23	-6.89
TT	**-6.50**	-5.74	-6.28	-5.93	-5.27	-4.81	-7.32	-7.96	-6.83	-4.32	-7.47
YY	**-4.66**	-4.99	-5.51	-4.49	-3.94	1.35	-4.31	-3.84	-3.62	-3.84	-5.62
TS	**-4.50**	-4.12	-4.30	-3.99	-4.05	-3.36	-5.41	-4.38	-4.40	-2.6	-5.01
LW	**-4.04**	-4.38	-4.74	-3.88	-3.58	1.00	-3.91	-3.46	-3.46	-3.87	-4.78
YP	**-3.79**	-3.78	-4.11	-3.32	-3.34	0.44	-4.09	-3.05	-3.09	-3.08	-4.7
FF	**-2.33**	-2.85	-3.04	-2.19	-2.01	1.11	-2.15	-1.97	-2.26	-2.14	-2.75
MM	**-2.03**	-1.67	-2.01	-1.27	-1.56	1.22	-1.94	-3.14	-2.35	-2.08	-2.63
LY	**-1.72**	-1.43	-1.66	-1.21	-1.34	0.96	-1.88	-1.86	-1.52	-1.98	-2.87
LL	**-1.62**	-1.54	-1.60	-1.36	-1.52	0.00	-1.96	-1.40	-1.66	-1.47	-1.97
MC	**-1.46**	-1.22	-1.43	-0.93	-1.27	0.25	-1.44	-2.01	-1.20	-1.52	-1.5
VV	**-1.39**	-1.14	-1.28	-0.96	-1.18	0.44	-1.83	-1.36	-1.43	-1.68	-2.17
IL	**-1.39**	-1.28	-1.35	-1.12	-1.29	0.06	-1.70	-1.19	-1.41	-1.3	-1.74
II	**-1.24**	-0.98	-1.11	-0.80	-1.01	0.62	-1.47	-1.13	-1.20	-1.13	-1.74
LT	**-1.09**	-0.95	-1.02	-0.83	-0.99	0.02	-1.36	-0.91	-1.05	-0.97	-1.46
VL	**-1.08**	-0.94	-1.01	-0.81	-0.97	0.11	-1.33	-0.81	-1.11	-1.01	-1.41
AL	**-1.07**	-0.76	-0.93	-0.60	-0.82	0.71	-1.32	-1.00	-0.94	-0.93	-1.61
LG	**-0.77**	-0.66	-0.71	-0.56	-0.71	-0.09	-1.00	-0.75	-0.53	-0.71	-0.98
	MRE [%]	*10.96*	*6.52*	*16.05*	*12.01*	*83.61*	*12.64*	*19.54*	*13.55*	*16.49*	*23.51*
	MRX [%]	*28.82*	*-30.62*	*43.57*	*23.69*	*166.28*	*-31.88*	*60.19*	*56.58*	*48.57*	*32.09*
	MAE	*0.47*	*0.26*	*0.48*	*0.79*	*2.03*	*0.58*	*2.11*	*2.22*	*1.72*	*1.44*
	MAX	*1.43*	*-0.85*	*2.76*	*3.28*	*6.01*	*-2.45*	*18.44*	*20.43*	*8.73*	*6.89*
	RMS	**0.48**	**0.36**	**0.60**	**0.88**	**1.40**	**0.68**	**4.16**	**4.78**	**2.40**	**2.42**

a - All energies are in gas phase and their values are in kcal/mol. Descriptive statistics: MRE is the unsigned mean relative error, MRX is the signed maximal relative error, MAE is the unsigned mean absolute error, MAX is the signed maximal absolute error, and RMS is the signed root mean square error. *b* - Neutral arginine is not defined in the parm03 force field

4.5. Interaction Energy Decomposition

The characteristics discussed above regarding the representative set of side-chain interactions was their total interaction energy. On the other hand, the strength of the interaction is not the only important information one can extract. The physical nature of the interaction is also important characteristic. To meet this requirement, the DFT-SAPT method was used to decompose the interaction energy into the physically valid terms. Certain level of the energy decomposition was used in preceding Chapters 4.1, 4.2 and 4.3, but these decompositions were limited only to a few types of side-chain side-chain interactions.

The DFT-SAPT energy decomposition was also performed for the representative set from the previous chapter (Appendix D), which covers all typical interactions motifs between the side-chains in proteins (153). The results are shown in Table 4.

Table 4 – CCSD(T)|CBS and DFT-SAPT energies for the representative set.

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

[a] E_{pol}^1 is the first-order electrostatics, E_{exch}^1 is the first-order repulsion, E_{ind}^2 is the second-order induction, E_{disp}^2 is the second-order dispersion, δHF is the estimate of higher-order terms and E_{disp}^2/E_{pol}^1 is the ratio between the dispersion and electrostatic terms. The most stabilizing terms are boldface. All energies are in kcal/mol.

The major conclusion of the study is as follows: polar residues interact mostly by the first-order electrostatic interaction, while nonpolar residues interact mostly by the second-order dispersion. Furthermore the ratio between the dispersion and electrostatic terms ranges from 0.05 for the salt bridge to 10 for the aliphatic contact. Moreover, stronger interaction energy is

almost always accompanied by an exchange-repulsion term. Stronger electrostatic energy is also correlated with increase in the induction energy and in higher order interactions. The DFT-SAPT interaction energies are systematically weaker than benchmark CCSD(T)|CBS interaction energies due to the small aug-cc-pVDZ basis set used in this method (156).

4.6. Matrix of Representative Interactions

The knowledge of benchmark values for the representative set of interactions helped to select the reasonably accurate and efficient method (RI-DFT-D) and allowed us to calculate stabilization energies for all 400 (20x20) possible pairs of side chain – side chain interactions (see Appendix F). The results showed (Table 5) that all interaction energies calculated at RI-DFT-D level in full 20x20 matrix are attractive in the gas phase. This can be attributed to the fact that the sharp character of the repulsion does not allow side chains to occupy unfavorable positions and the typical pair geometry in proteins is always adjusted to prevent such interaction mode. The only exceptions are pairs of residues with the same charge (i.e. E-E, D-D, R-R, and K-K). However, these interactions are low populated in the proteins so they cannot change the total attractive character of stabilization contributions of residual contacts.

Differences in interaction energies in Table 5 are enormous. The strongest interactions are those of salt bridges (up to 140 kcal/mol), while the weakest ones are those between small aliphatic residues (around 1 kcal/mol). Repulsion interactions between same charged residues are about a half of size of the attractive ones for the oppositely charged residues (up to -70 kcal/mol). Amino acid residues can be sorted according to their interaction potential as follows: (The“>“ sign shows energy difference of at least 1 kcal/mol)

D, E > R > K > N, Q > W, Y > H > S > T > F > M, C > P, L > I, V > A > G.

The strongest interaction not surprisingly comes from interactions of charged residues even when the repulsion interactions between amino acids of the same charges are included in the total sum of contributions. The line continues by polar and aromatic residues, and it ends with aliphatic residues sorted according their size. Similarly behaving families of residues can be selected (see Figure 17). More detailed description can be found in Appendix F.

Table 5 - Interaction energy matrix for cluster representatives .
Matrix contains all 20x20 possible cluster representative pairs between residues. Calculated with RI-DFT-D/TPSS|TZVP method.
All energies shown in kcal/mol

	G	A	V	I	L	F	Y	W	H	P	T	S	N	Q	C	M	K	R	D	E
G	-0.6	-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

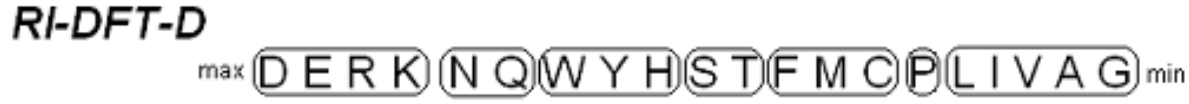


Figure 17 - Amino acid families sorted by their summed interaction energies Calculated with RI-DFT-D/TPSS/TZVP method.

Additional statistical information can be also obtained from Atlas of Protein Side-Chain Interactions representing all side-chain side-chain pair geometries taken from the non-homologous protein structures. Interacting side-chains are considered to be those having a center-to-center distance between their closest two heavy atoms of less than the sum of their van der Waals radii, plus 1 Å to allow for coordinate error. Atlas thus contains all in close range interacting side-chain pairs whose total numbers are shown in Table 6.

It is apparent from Table 6 that all residues have contact mostly with leucine residue The least populated contacts are those with cysteine. This table is also educative in a sense that every important interaction can be distinguished – disulphide bridges or salt bridges are preferred over any other interactions of their respective residues. Another important fact is that the majority of the interactions are between aliphatic and aromatic residues. It also shows some less known facts – for example the likeness which have proline to make contacts with aromatic residues (which is in concord with the unusually strong interaction of the proline with the aromatic residues as was shown in Chapter 4.3) or that the interactions of negatively charged residues with the residues with the hydroxyl group are surprisingly well populated.

Table 6 – Total numbers of the side-chain side-chain contacts
Values taken from Atlas of Protein Side-chain Interactions (84). The color coding shows how much side-chain in row makes contact with side-chain in column – blue color shows the most typical contacts (above 80% percentile) and red color shows the unusual contacts (bellow 20% percentile) of the side-chain in row.

	G	A	V	I	L	F	Y	W	H	P	T	S	N	Q	C	M	K	R	D	E
G	2816	3678	**4697**	**4182**	**6369**	**3871**	3617	*1860*	*1537*	2211	3481	2859	2755	2091	*999*	*1624*	2460	3831	3136	2613
A	3678	6850	**10941**	**9519**	**15282**	**7233**	5144	2680	*1812*	2975	4081	2934	*2304*	2310	*1353*	2857	*2180*	3967	2500	2825
V	4696	10941	**19723**	**17790**	**27218**	**12106**	7632	3869	*2637*	4352	6412	4130	*3003*	3364	*2340*	4572	3963	5190	*3004*	4172
I	4180	9518	**17789**	**18624**	**26652**	**11586**	7310	3672	*2328*	3726	5706	3514	*2744*	3115	*2072*	4498	3816	4636	*2568*	3853
L	6368	15280	**27217**	**26651**	**47638**	**19454**	12030	6487	*3845*	6433	8233	5619	*3999*	5448	*3295*	6993	5696	8423	*4056*	6314
F	3870	7232	**12109**	**11585**	**19454**	**11127**	6660	3676	*2291*	3755	4283	3307	*2468*	2678	*1987*	3865	2935	4135	*2291*	3094
Y	3615	5144	**7633**	**7311**	**12039**	**6658**	5179	2738	*2384*	4149	3568	2952	*2736*	2745	*1287*	*2613*	3724	4613	3531	3980
W	1860	2680	**3870**	**3672**	**6485**	**3676**	2737	1800	*1181*	2135	1773	1571	*1317*	1437	*709*	1412	1425	2303	*1302*	1808
H	1537	1811	**2635**	2329	**3845**	2292	2384	*1179*	1750	1318	1989	1785	1336	1230	*639*	*978*	*1193*	2096	2384	**2583**
P	2211	2975	**4353**	3727	**6438**	**3756**	**4149**	2135	*1318*	2142	2471	1990	1695	1805	*924*	*1486*	*1568*	2955	1754	2382
T	3481	4080	**6411**	**5704**	**8236**	**4283**	3567	*1773*	*1992*	2470	4262	3132	2993	2791	*967*	*1813*	2884	3886	3435	3637
S	2859	2931	**4126**	**3513**	**5614**	**3306**	2951	*1571*	*1789*	1990	3132	2809	2487	2242	*814*	*1245*	2409	3239	3279	3277
N	2753	2302	**3001**	2740	**3996**	2470	2735	*1317*	*1332*	1695	**2991**	2488	2821	2227	*603*	*1052*	2435	**3003**	2844	2861
Q	2087	2307	**3359**	**3105**	**5433**	2668	2737	*1435*	*1224*	1802	2785	2234	2217	2172	*603*	*1154*	2473	**3189**	2375	2591
C	999	1353	**2341**	**2072**	**3296**	1987	1287	709	639	924	967	815	*604*	606	**3490**	641	*597*	860	*560*	*600*
M	1623	2855	**4576**	**4493**	**6993**	**3859**	2613	1411	*979*	1485	1811	1243	*1053*	1154	*641*	1973	1140	1628	*948*	1330
K	2427	2140	**3894**	3758	**5592**	2890	3656	*1407*	*1171*	1523	2834	2365	2395	2428	*587*	*1120*	1616	2232	**5883**	**7755**
R	3814	3955	**5178**	4608	**8397**	4114	4587	2291	*2092*	2942	3862	3230	2986	3185	*856*	*1620*	*2282*	4441	**7391**	**9671**
D	3133	2495	3001	2567	**4050**	2290	**3533**	*1302*	2383	*1753*	3435	3276	2845	2382	*560*	*946*	**5989**	**7416**	2058	1897
E	2607	2808	**4154**	3842	**6285**	3082	3967	*1803*	2576	2369	3623	3260	2840	2594	*596*	*1324*	**7861**	**9678**	*1891*	2475

With the knowledge of total numbers of each residue in the database and their total number of contacts, one can calculate number of total contacts for every residue (see Table 7). The table highlights two main aspects of the residue interaction properties – firstly, the bigger is the residue the higher number of contacts. Secondly, more polar residues have lower number of contacts. They prefer interactions with the surrounding water or ions.

Table 7 – Average number of contacts per residue

AA	<c>	AA	<c>	AA	<c>	AA	<c>
G	***1.7***	**F**	**5.8**	**T**	**2.7**	**M**	**4.9**
A	***2.3***	**Y**	**5.2**	**S**	***1.9***	**K**	***2.1***
V	**4.4**	**W**	**6.3**	**N**	***2.3***	**R**	**3.3**
I	**5.2**	**H**	**3.3**	**Q**	**2.6**	**D**	***2.0***
L	**5.2**	**P**	***2.3***	**C**	**3.9**	**E**	***2.1***

The variability of the strength as well as population of the side-chain side-chain contacts is enormous, and the proper description of all contacts is thus needed. The usual way of the representation of the interaction energies in recent studies is the use of the molecular dynamic simulations with force field potentials. This leads to the question of the force field precision.

4.7. Force Field Accuracy for Side-chain Side-chain Interactions

Force fields were tested firstly against the CCSD(T)|CBS benchmark values on the representative set of the side-chain side-chain interactions (153) (see Chapter 4.4). Their performance was worse than any of *ab initio* methods used, but the correlation with the benchmark values was still high (r = 0.99).

The performance of the force field methods was further tested against interaction energies calculated by RI-DFT-D method for all 20x20 possible pairs of side-chains (
Table 5). The correlation coefficients between the energies calculated with force fields and with more accurate RI-DFT-D method were slightly lower than force field-CCSD(T)|CBS case – above 0.95. This fact is also confired by the good overall preservation of the amino acid families in the stability lines (Figure 18). It can be concluded that force fields are in general quite successful in the description of the interaction energies between side-chains in proteins.

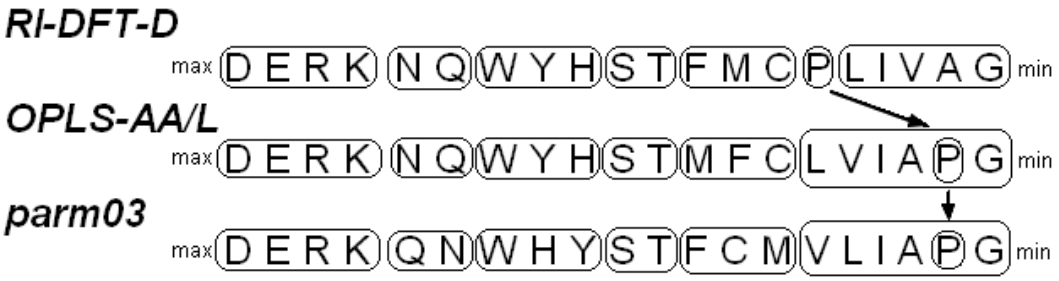


Figure 18 – Changes in amino acid families between different calculation methods.

While the overall performance of the force fields is surprisingly good, their average and median stabilization energies are systematically smaller than those calculated with RI-DFT-D method by about 1 kcal/mol. This has two reasons: (i) firstly, RI-DFT-D method slightly oversized interaction energies for weakly bound pairs of aliphatic interactions (153) and (ii) secondly and more importantly, force field values provide higher numbers of repulsive interaction energies in both force fields applied. The repulsion term seems to be too strong in the Lennard-Jones potential C_{12}/r^{12} term in comparison with *ab initio* methods. More details can be found in Appendix F.

Force fields are reasonably accurate in the estimation of overall stabilization energies within protein. On the other hand, they cannot be used with confidence for reliable evaluation of intermolecular interaction energies between side-chains in all cases.

4.8. Solvent Effect

So far the evaluation of side-chain interactions was focused on gas phase interaction energies between side-chains due to the possibility of comparison of the performance of various methods to highly accurate CCSD(T)|CBS benchmark energies. However, proteins exist in completely different environments. Amino acid residues are surrounded heterogeneously either by other residues or by water molecules. The environment is known to affect the interaction energies at least via the polarization as was shown during the study on the salt bridges in Chapter 4.2.

The change of interaction energies upon introduction of an environment was studied with help of polarizable continuum solvent model (PCM) with two different values of dielectric constants to imitate protein ($\varepsilon = 4$) or water environment ($\varepsilon = 80$). Interaction energies for the representative set are summarized in the Table 8. It shows that interaction energies are weaker

in both environments than those from the gas phase. Interestingly, polar interactions are more weakened than nonpolar ones. Intuitively, interaction energies are stronger in the protein-like environment than in water.

Table 8 – The change of the strength of interaction energies in different environments. Calculated with DFT-D/TPSS/TZVP method with PCM implicit solvent model. All energies shown in kcal/mol.

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

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

Environmental changes of interaction energies have different impact on different side-chains. Therefore the same study was repeated with COSMO solvent model for the complete matrix of 20x20 side-chain pairs discussed earlier to see how interaction energies are influenced by environment on per residue basis. The resulting total interaction energies per residue were sorted into stability lines for each particular environment (see Figure 19).

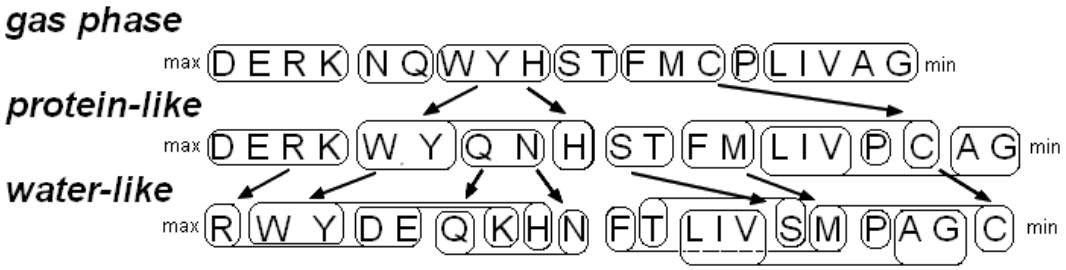


Figure 19 – Changes in amino acid families in the environment.
Residues are sorted by their total interaction energy calculated by RI-DFT-D calculations with COSMO implicit solvent. The dispersively bound residues are generally shifted upward in contrast with the polar ones.

As can be seen in Figure 19, the environment highly promotes interactions between residues of aromatic or aliphatic character (mainly tryptophane, tyrosine, leucine, isoleucine, and valine). On the other hand, the strength of interactions involving charged residues is lowered significantly by the water environment with the only exception of arginine, whose guanidinium group possesses also a strong dispersion interaction. Polar and sulphuric groups are shifted towards the lower stability end, while smaller residues of the same kind are moved more (asparagine more than glutamine, serine more than threonine, and cysteine more than methionine). This can be also accounted to less extensive dispersion interactions whose are unaffected by the environment. The environment changes significantly the order of the stabilization energy in the advantage of aromatic residues.

4.9. The Role of Representative Pairs

There is a question why the cluster representative structures in the Atlas of Protein Side-Chain Interactions are so densely packed that they are repulsive in force field calculations. To answer this question, the leucine-tryptophane pair (LW) was selected as a model system to put obtained characteristic values in larger structural context. This pair was selected due to its size in Atlas. The Atlas of Protein Side-Chain Interactions contains only 6487 LW contacts in total, the most populated cluster hass only 34 structures in and there is 1 structure as a cluster representative.

The interaction energy of the cluster representative pair for LW contacts in Cβ representation is -2.68 kcal/mol. The average interaction energy of the cluster with 34 structures is -2.75 ± 0.55 kcal/mol. This value is comparable with the cluster representative value so it seems that the cluster representative structure provides a reasonable approximation of all structures identified in one cluster. However interaction energies determined for all of 6487 LW contacts provided different average value of -1.60 ± 0.79 kcal/mol (see Figure 20).

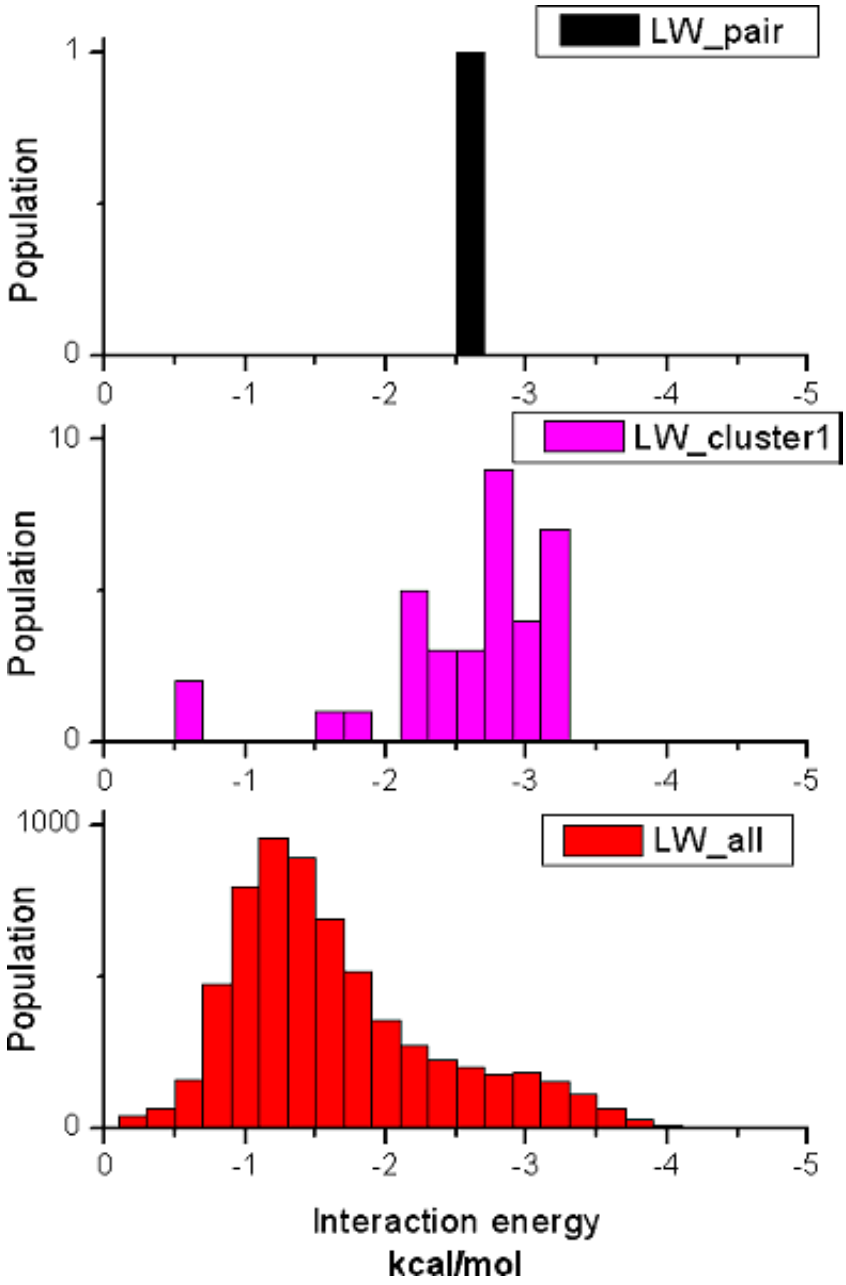


Figure 20 – Histograms of interaction energies for LW pairs calculated by RI-DFT-D.
From above (a) the interaction energy for cluster representative, (b) the histogram of the energies for all geometries from cluster, and (3) the histogram of energies for all leucine-tryptophane pairs from Atlas of Protein Side-Chain Interactions.

The closer look on histograms of interaction energies on Figure 20 showed that the energy of the cluster representative is at the peak of the distribution energies for all structures in cluster. However neither the energy of the representative pair nor energies for the whole cluster are typical enough for the full distribution of interaction energies, which has its peak around -1.3 kcal/mol. The complete distribution of the interaction energies has completely different shape than the distribution of cluster energies. The majority of contacts are significantly weaker than cluster contacts. This leads to the conclusion that cluster geometries can be of some importance.

Since this result needed further verification, the overall distribution of LW pairs was recalculated with the parm03 force field in Cα representation and the observed distribution was similar (see Appendix F). With this assurance, the overall distributions of tryptophane with all other residues were calculated with the parm03 force field. Distributions of interaction energies suggest that the approximations lying behind the phenomenological potentials might simply be wrong, as the distributions are neither normal nor Boltzmann-like. Therefore, the simple calculation of free energies from the detected contacts is not easily connected to the real energies as has already been indicated by Thomas and Dill (21).

When compared with RI-DFT-D interaction energies of the cluster representative pair for respectives of tryptophane-containing pairs, interaction energies of the cluster representative pairs were always stronger than the interaction energies of the most populated interactions (see Figure 21).

Similarly strong interactions as those of the cluster representative pairs were previously found also in the hydrophobic core of rubredoxin (See Chapter 4.1 and Appendix A). There, the strongest interactions between residues in the hydrophobic core (Y4, C6, Y13, F30, L33, W37, and F49). This fact encourages the hypothesis that representative pairs 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. Therefore they should represent structurally or functionally important interactions.

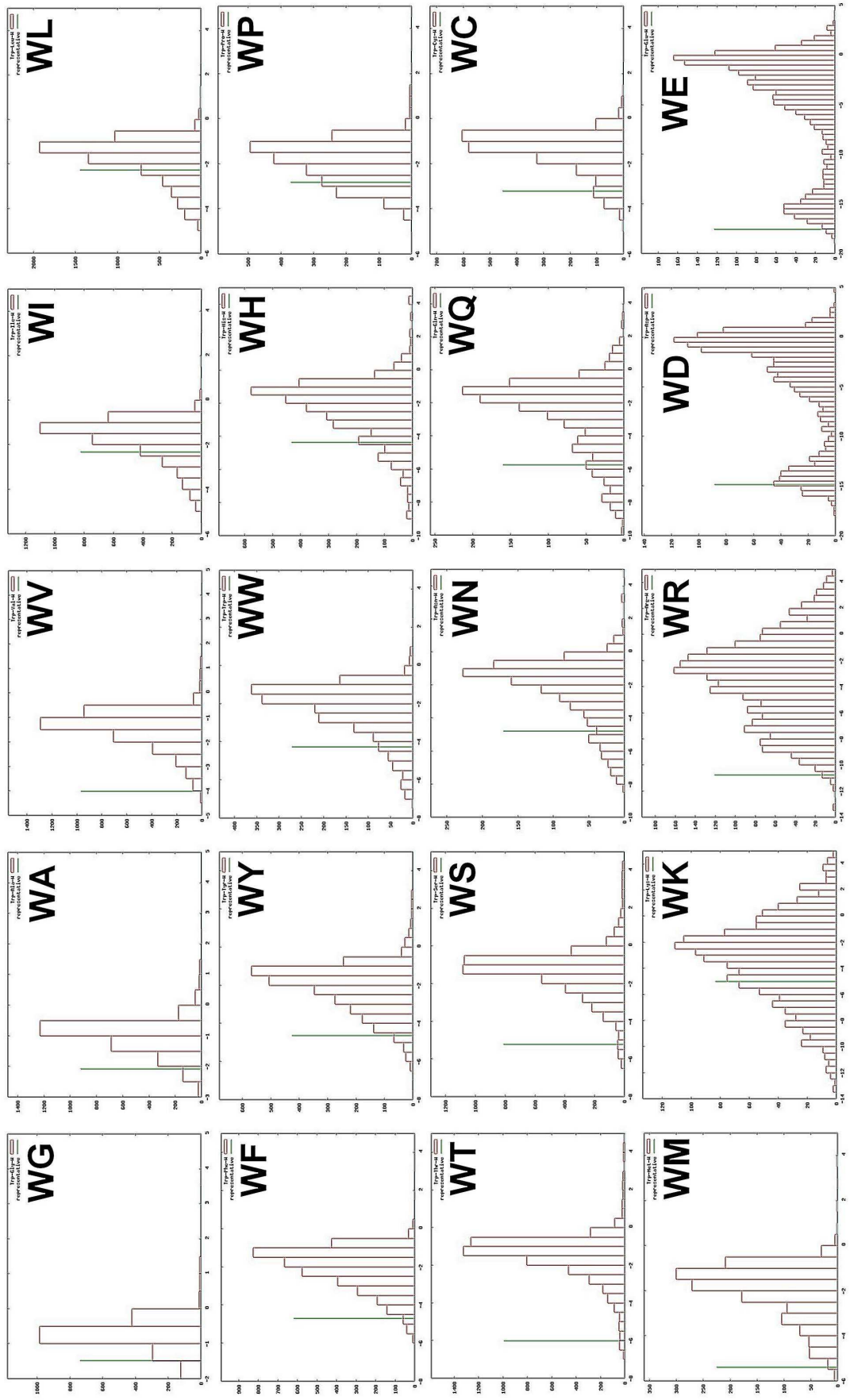


Figure 21 – Histograms of interaction energies of tryptophane with all residues .
The distributions are calculated with parm03 force-field. The green spike corresponds to the RI-DFT-D interaction energy for the cluster representative for the respective pair.

5. Conclusions

The aim of the presented thesis was to investigate the strength of side-chain side-chain interactions in proteins. The interaction energies give us some information about the enthalpic contribution to the overall stabilization of proteins. The results of the thesis can be summarized as follows:

1. The dispersion energy is the main interaction term within the hydrophobic core of rubredoxin. The interaction energies between the residues in the hydrophobic core are also stronger than most of the interactions between the same residues found outside of the hydrophobic core structural context.

2. The strength of the salt bridge interaction is substantially lowered or even negligible upon the presence of protein-like or water environments.

3. Interactions of proline with tryptophane can be as strong as interactions between two aromatic residues mainly for two reasons – the presence of the heteroatom in proline strengthening electrostatic interactions and the cyclic arrangement of the proline residue increasing dispersive contacts.

4. The evaluation of the interaction energies for the side-chain side-chain pairs on benchmark set showed that ab initio method with reasonable accuracy and speed is RI-DFT-D. Much cheaper semiempirical methods PM6-DH or SCC-DFTB-D had worse accuracy, but they were still better than force field methods parm03 and OPLS-AA/L.

   The benchmark data were published in the online database dedicated to the benchmark energies and geometries of various noncovalent complexes www.begdb.com. BEGDB database can be used for testing of other calculation methods.

5. The decomposition of interaction energies showed that polar residues are interacting mostly by the first-order electrostatic interaction, while nonpolar residues are interacting mostly by the second-order dispersion.

6. The variability of the strength of interactions as well as the population of side-chain side-chain contacts is enormous (two orders of magnitude) and as such it poses great demand for the precision of the calculation methods.

7. Force fields provide the rough description of overall interaction energies within protein with reasonable accuracy, but they cannot be used with confidence for specific pairs such as functionally or structurally important pairs.

8. The protein as well as water environment lowers the stabilization energies mostly for the charged and polar side-chains and thus promotes the relative importance of aromatic or aliphatic residues.

9. The distribution of the side-chain side-chain interaction energies is neither normal nor Boltzmann-like. This fact to some extent disrupt the theoretical basis for the statistical potential which assume that the free energy of association of side-chain side-chain pairs in proteins can be gained simply from the numbers of the contacts between respective amino acid residues in the database.

   Representative pairs from the Atlas of Protein Side-Chain Interactions are strong enough to be geometrically as well as energetically distinguishable from the mostly random (and mostly attractive) interactions of the majority of the side-chain side-chain pairs. Therefore they should represent structurally or functionally important interactions.

Programs Used

The hydrogen atoms were added using Pymol 0.99.r6 (157) or with Gromacs 3.3 package (158). Positions of hydrogen atoms were optimized by the TPSS functional using the TZVP basis in Turbomole 5.8 package (159) or by SCC-DFTB-D method in the dftb+ program package (160).

All molecular mechanical force field calculations of the interaction energies were performed using Gromacs 3.3 package (158). The amino acid topology and partial charges have been taken from Sorin and Pande Amberport topologies (161) and they were modified to represent only side-chain analogs truncated at Cα (or Cβ) atoms. In such way, modified version of parm03 (36) and OPLS-AA/L (39) force fields were prepared.

The *ab initio* calculations were calculated with several codes with the common ruby interface called “cuby” created by Dr. Jan Řezáč (162). Most of the *ab initio* calculations were performed with Turbomole package (159) – RI-MP2, RI-DFT-D.

Energy decomposition with DFT-SAPT calculations were performed with the use of two codes – Gaussian 03 (163) was used for the parameterization step of the shift for monomers and the calculation of DFT-SAPT method itself was performed with the use of the Molpro 2006 package (164). Molpro 06 was also used for the calculation of the CCSD(T) method.

Semiempirical calculations were also performed with the cuby framework. The PM6 was calculated with MOPAC2007 (165) and the dispersion and hydrogen bond corrections were added within the ruby code from Jan Řezáč (134). SCC-DFTB-D energies were calculated with dftb+ program package (160).

References

1. Mulder, G. J. (1839) *Journal fur Praktische Chemie* **16**, 129-152.
2. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., & Orengo, C. A. (2000) *Nat Struct Biol* **7**, 991-994.
3. Skolnick, J. & Fetrow, J. S. (2000) *Trends in Biotechnology* **18**, 34-39.
4. Watson, J. D., Laskowski, R. A., & Thornton, J. M. (2005) *Current Opinion in Structural Biology* **15**, 275-284.
5. Whisstock, J. C. & Lesk, A. M. (2003) *Quarterly Reviews of Biophysics* **36**, 307-340.
6. Kahraman, A., Morris, R. J., Laskowski, R. A., & Thornton, J. M. (2007) *Journal of Molecular Biology* **368**, 283-301.
7. Anfinsen, C. B. (1973) *Science* **181**, 223-230.
8. Karshikoff, A. (2006) *Non-covalent Interactions in Proteins* (Imperial College Press, London).
9. Miyazawa, S. & Jernigan, R. L. (1996) *Journal of Molecular Biology* **256**, 623-644.
10. Miyazawa, S. & Jernigan, R. L. (1999) *Proteins-Structure Function and Genetics* **34**, 49-68.
11. Miyazawa, S. & Jernigan, R. L. (1999) *Proteins-Structure Function and Genetics* **36**, 357-369.
12. Sippl, M. J. (1995) *Curr. Opin. Struct. Biol.* **5**, 229-235.
13. Betancourt, M. R. (2008) *J. Phys. Chem. B* **112**, 5058-5069.
14. Bahar, I., Atilgan, A. R., Jernigan, R. L., & Erman, B. (1997) *Proteins-Structure Function and Genetics* **29**, 172-185.
15. John, B. & Sali, A. (2003) *Nucleic Acids Research* **31**, 3982-3992.
16. Melo, F. & Sali, A. (2007) *Protein Science* **16**, 2412-2426.
17. Shen, M. Y. & Sali, A. (2006) *Protein Science* **15**, 2507-2524.
18. Miyazawa, S. & Jernigan, R. L. (2005) *Journal of Chemical Physics* **122**, 024901.
19. Fiser, A. S. & Sali, A. (2003) *Macromolecular Crystallography, Pt D* **374**, 461.
20. Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003) *Nucleic Acids Research* **31**, 3381-3385.
21. Thomas, P. D. & Dill, K. A. (1996) *Journal of Molecular Biology* **257**, 457-469.
22. Vendruscolo, M. & Domany, E. (1998) *Journal of Chemical Physics* **109**, 11101-11108.
23. Vendruscolo, M., Najmanovich, R., & Domany, E. (2000) *Proteins-Structure Function and Genetics* **38**, 134-148.
24. Herges, T. & Wenzel, W. (2004) *Biophysical Journal* **87**, 3100-3109.
25. Verma, A., Schug, A., Lee, K. H., & Wenzel, W. (2006) *Journal of Chemical Physics* **124**.
26. Verma, A. & Wenzel, W. (2009) *Biophysical Journal* **96**, 3483-3494.
27. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004) *Numerical Computer Methods, Pt D* **383**, 66.
28. Baker, D. (2006) *Philosophical Transactions of the Royal Society B-Biological Sciences* **361**, 459-463.
29. Simons, K. T., Bonneau, R., Ruczinski, I., & Baker, D. (1999) *Proteins-Structure Function and Genetics* 171-176.
30. Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. M., & Baker, D. (2001) *Proteins-Structure Function and Genetics* 119-126.
31. Bradley, P., Chivian, D., Meiler, J., Misura, K. M. S., Rohl, C. A., Schief, W. R., Wedemeyer, W. J., Schueler-Furman, O., Murphy, P., Schonbrun, J. *et al.* (2003) *Proteins-Structure Function and Genetics* **53**, 457-468.

32. Bradley, P., Malmstrom, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D. E., Meiler, K., Misura, K. M. S., & Baker, D. (2005) *Proteins-Structure Function and Bioinformatics* **61**, 128-134.
33. Duan, Y. & Kollman, P. A. (1998) *Science* **282**, 740-744.
34. Shirts, M. & Pande, V. S. (2000) *Science* **290**, 1903-1904.
35. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995) *Journal of the American Chemical Society* **117**, 5179-5197.
36. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T. *et al.* (2003) *Journal of Computational Chemistry* **24**, 1999-2012.
37. Cheatham, T. E., Cieplak, P., & Kollman, P. A. (1999) *J Biomol Struct Dyn* **16**, 845-862.
38. Perez, A., Marchan, I., Svozil, D., Sponer, J., Cheatham, T. E., Laughton, C. A., & Orozco, M. (2007) *Biophysical Journal* **92**, 3817-3829.
39. Kaminski, G. A., Friesner, R. A., & Tirado-Rives, J. &. J. W. (2001) *The Journal of Physical Chemistry B* **105**, 6474-6487.
40. Mackerell, A. D., Jr., Banavali, N., & Foloppe, N. (2000) *Biopolymers* **56**, 257-265.
41. Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I. *et al.* (2009) *J. Comput. Chem.*
42. Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., & Marrink, S. J. (2008) *Journal of Chemical Theory and Computation* **4**, 819-834.
43. Moult, J., Fidelis, K., Rost, B., Hubbard, T., & Tramontano, A. (2005) *Proteins-Structure Function and Bioinformatics* **61**, 3-7.
44. Hofmeister, F. (1902) *Ergeb. Physiol.* **1**, 759-802.
45. Fischer, E. (1902) *Chem. Ztg.* **26**, 939.
46. Pauling, L. & Corey, R. B. (1951) *Proceedings of the National Academy of Sciences* **37**, 729-740.
47. Pauling, L., Corey, R. B., & Branson, H. R. (1951) *Proceedings of the National Academy of Sciences* **37**, 235-240.
48. Kauzmann, W. (1959) *Advances in Protein Chemistry* **14**, 1-63.
49. Chandler, D. (2002) *Nature* **417**, 491.
50. Lum, K., Chandler, D., & Weeks, J. D. (1999) *Journal of Physical Chemistry B* **103**, 4570-4577.
51. Ball, P. (2008) *Chemical Reviews* **108**, 74-108.
52. Abseher, R., Schreiber, H., & Steinhauser, O. (1996) *Proteins-Structure Function and Genetics* **25**, 366-378.
53. Giovambattista, N., Rossky, P. J., & Debenedetti, P. G. (2006) *Physical Review E* **73**, 041604.
54. Giovambattista, N., Lopez, C. F., Rossky, P. J., & Debenedetti, P. G. (2008) *Proceedings of the National Academy of Sciences of the United States of America* **105**, 2274-2279.
55. Giovambattista, N., Debenedetti, P. G., & Rossky, P. J. (2009) *Proceedings of the National Academy of Sciences of the United States of America* **106**, 15181-15185.
56. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., & DiNola, A. (1984) *The Journal of Chemical Physics* **81**, 3684-3690.
57. Hofinger, S. & Zerbetto, F. (2005) *Chemical Society Reviews* **34**, 1012-1020.
58. Stone, A. J. (1997) *The theory of intermolecular forces* (Clarendon Press, Oxford).
59. Gray, C. G. & Gubbins, K. E. (1984) *Theory of Molecular Fluids: Fundamentals* (Oxford University Press, Oxford, UK.

60. Stone, A. J. & Alderton, M. (1985) *Molecular Physics* **56**, 1047-1064.
61. Mulliken, R. S. (1955) *Journal of Chemical Physics* **23**, 1833-1840.
62. Elstner, M. (2007) *J. Phys. Chem. A* **111**, 5614-5621.
63. Bayly, C. I., Cieplak, P., Cornell, W. D., & Kollman, P. A. (1993) *J Phys Chem-Us* **97**, 10269-10280.
64. Wang, J., Cieplak, P., & Kollman, P. (2000) *Journal of Computational Chemistry* **21**, 1049-1074.
65. van Duijnen, P. T. & Swart, M. (1998) *Journal of Physical Chemistry A* **102**, 2399-2407.
66. Karshikoff, A. & Ladenstein, R. (2001) *Trends Biochem Sci* **26**, 550-556.
67. Karshikoff, A. & Jelesarov, I. (2008) *Biotechnology & Biotechnological Equipment* **22**, 606-611.
68. Kumar, S. & Nussinov, R. (2002) *Chembiochem* **3**, 604-617.
69. Baker, E. N. & Hubbard, R. E. (1984) *Progress in Biophysics & Molecular Biology* **44**, 97-179.
70. McDonald, I. K. & Thornton, J. M. (1994) *Journal of Molecular Biology* **238**, 777-793.
71. Dill, K. A. (1990) *Biochemistry* **29**, 7133-7155.
72. Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993) *Journal of Applied Crystallography* **26**, 283-291.
73. Kaur, D., Sharma, P., & Bharatam, P. V. (2007) *Journal of Molecular Structure: THEOCHEM* **810**, 31-37.
74. Than, M. E., Henrich, S., Huber, R., Ries, A., Mann, K., Kuhn, K., Timpl, R., Bourenkov, G. P., Bartunik, H. D., & Bode, W. (2002) *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6607-6612.
75. Vanacore, R., Ham, A. J. L., Voehler, M., Sanders, C. R., Conrads, T. P., Veenstra, T. D., Sharpless, K. B., Dawson, P. E., & Hudson, B. G. (2009) *Science* **325**, 1230-1234.
76. Gilchrist, T. L. & Moody, C. J. (1977) *Chemical Reviews* **77**, 409-435.
77. Leopold, P. E., Montal, M., & Onuchic, J. N. (1992) *Proceedings of the National Academy of Sciences of the United States of America* **89**, 8721-8725.
78. Dobson, C. M. (2004) *Seminars in Cell & Developmental Biology* **15**, 3-16.
79. Shakhnovich, E. (2006) *Chemical Reviews* **106**, 1559-1588.
80. Abkevich, V. I., Gutin, A. M., & Shakhnovich, E. I. (1994) *Biochemistry* **33**, 10026-10036.
81. Vlassi, M., Cesareni, G., & Kokkinidis, M. (1999) *J Mol Biol* **285**, 817-827.
82. Vondrasek, J., Bendova, L., Klusak, V., & Hobza, P. (2005) *Journal of the American Chemical Society* **127**, 2615-2619.
83. Shimada, J. & Shakhnovich, E. I. (2002) *Proceedings of the National Academy of Sciences of the United States of America* **99**, 11175-11180.
84. Laskowski, R. & Thornton, J. M. (2008) (http://www.ebi.ac.uk/thornton-srv/databases/sidechains, accessed Oct 31, 2008).
85. Singh, J. & Thornton, J. M. (1992) *Atlas of Protein Side-Chain Interactions* (IRL press, Oxford).
86. Berman, H., Henrick, K., Nakamura, H., & Markley, J. L. (2007) *Nucleic Acids Res.* **35**, D301-D303.
87. Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., & Westbrook, J. (2000) *Nat. Struct. Biol.* **7 Suppl**, 957-959.
88. Laskowski, R., Luscombe, N., & Moes, S. (1999) (http://www.biochem.ucl.ac.uk/bsm/sidechains, accessed Oct 31, 2009).
89. Born, M. & Oppenheimer, R. (1927) *Annalen der Physik* **84**, 457-484.

90. Schrödinger, E. (1926) *Physical Review* **28**, 1049-1070.
91. Čížek, J. (1966) *J. Chem. Phys.* **45**, 4256.
92. Pittner, J. & Hobza, P. (2004) *Chemical Physics Letters* **390**, 496-499.
93. Burda, J. V., Zahradnik, R., Hobza, P., & Urban, M. (1996) *Molecular Physics* **89**, 425-432.
94. Halkier, A., Helgaker, T., Jorgensen, P., Klopper, W., Koch, H., Olsen, J., & Wilson, A. K. (1998) *Chemical Physics Letters* **286**, 243-252.
95. Jurecka, P. & Hobza, P. (2002) *Chemical Physics Letters* **365**, 89-94.
96. Dunning, T. H. (1989) *Journal of Chemical Physics* **90**, 1007-1023.
97. Woon, D. E. & Dunning, T. H. (1995) *Journal of Chemical Physics* **103**, 4572-4585.
98. Moller, C. & Plesset, M. S. (1934) *Physical Review* **46**, 618-622.
99. Klopper, W. (2004) *Journal of Chemical Physics* **120**, 10890-10895.
100. Jurečka, P., Nachtigall, P., & Hobza, P. (2001) *Physical Chemistry Chemical Physics* **3**, 4578-4582.
101. Boys, S. F. & Bernardi, F. (1970) *Molecular Physics* **19**, 553.
102. Kohn, W. & Sham, L. J. (1965) *Physical Review* **140**, 1133.
103. Perdew, J. P., Burke, K., & Ernzerhof, M. (1996) *Physical Review Letters* **77**, 3865-3868.
104. Perdew, J. P., Chevary, J. A., Vosko, S. H., Jackson, K. A., Pederson, M. R., Singh, D. J., & Fiolhais, C. (1992) *Physical Review B* **46**, 6671-6687.
105. Becke, A. D. (1988) *Physical Review A* **38**, 3098-3100.
106. Lee, C. T., Yang, W. T., & Parr, R. G. (1988) *Physical Review B* **37**, 785-789.
107. Tao, J. M., Perdew, J. P., Staroverov, V. N., & Scuseria, G. E. (2003) *Physical Review Letters* **91**, 146401.
108. Zhao, Y. & Truhlar, D. G. (2008) *Accounts of Chemical Research* **41**, 157-167.
109. Becke, A. D. (1993) *Journal of Chemical Physics* **98**, 5648-5652.
110. Stephens, P. J., Devlin, F. J., Chabalowski, C. F., & Frisch, M. J. (1994) *J Phys Chem-Us* **98**, 11623-11627.
111. Perdew, J. P. & Wang, Y. (1992) *Physical Review B* **45**, 13244-13249.
112. Burke, K., Ernzerhof, M., & Perdew, J. P. (1997) *Chemical Physics Letters* **265**, 115-120.
113. Zhao, Y. & Truhlar, D. G. (2008) *Theor Chem Acc* **120**, 215-241.
114. Grimme, S. (2006) *Journal of Computational Chemistry* **27**, 1787-1799.
115. Jurečka, P., Cerny, J., , H. P., & Salahub, R. (2007) *J. Comp. Chem.* **28**, 555-569.
116. Jurečka, P., Sponer, J., Cerny, J., & Hobza, P. (2006) *Physical Chemistry Chemical Physics* **8**, 1985-1993.
117. Sedlak, R., Jurečka, P., & Hobza, P. (2007) *Journal of Chemical Physics* **127**.
118. Kolar, M., Berka, K., Jurečka, P., & Hobza, P. (2009) *submitted*.
119. Porezag, D., Frauenheim, T., Kohler, T., Seifert, G., & Kaschner, R. (1995) *Physical Review B* **51**, 12947-12957.
120. Seifert, G., Porezag, D., & Frauenheim, T. (1996) *International Journal of Quantum Chemistry* **58**, 185-192.
121. Harrison, W. A. (1986) *Physical Review B* **34**, 2787-2793.
122. Elstner, M., Porezag, D., Jungnickel, G., Elsner, J., Haugk, M., Frauenheim, T., Suhai, S., & Seifert, G. (1998) *Physical Review B* **58**, 7260-7268.
123. Elstner, M., Hobza, P., Frauenheim, T., Suhai, S., & Kaxiras, E. (2001) *Journal of Chemical Physics* **114**.
124. Seabra, G. D., Walker, R. C., Elstner, M., Case, D. A., & Roitberg, A. E. (2007) *Journal of Physical Chemistry A* **111**, 5655-5664.

125. Rybak, S., Jeziorski, B., & Szalewicz, K. (1991) *Journal of Chemical Physics* **95**, 6576-6601.
126. Misquitta, A. J. & Szalewicz, K. (2005) *Journal of Chemical Physics* **122**, 214109 .
127. Hesselmann, A. & Jansen, G. (2003) *Physical Chemistry Chemical Physics* **5**, 5010-5014.
128. Hesselmann, A., Jansen, G., & Schutz, M. (2005) *Journal of Chemical Physics* **122**, 14103.
129. Podeszwa, R. & Szalewicz, K. (2005) *Chemical Physics Letters* **412**, 488-493.
130. Stewart, J. J. P. (2007) *J Mol Model* **13**, 1173-1213.
131. Voityuk, A. A. & Rosch, N. (2000) *Journal of Physical Chemistry A* **104**, 4089-4094.
132. Thiel, W. & Voityuk, A. A. (1996) *J Phys Chem-Us* **100**, 616-626.
133. Stewart, J. J. P. (2008) *J Mol Model* **14**, 499-535.
134. Rezac, J., Fanfrlik, J., Salahub, D., & Hobza, P. (2009) *Journal of Chemical Theory and Computation* **5**, 1749-1760.
135. Jorgensen, W. L., Maxwell, D. S., & TiradoRives, J. (1996) *Journal of the American Chemical Society* **118**, 11225-11236.
136. Cieplak, P., Cornell, W. D., Bayly, C., & Kollman, P. A. (1995) *Journal of Computational Chemistry* **16**, 1357-1377.
137. Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C., Wang, J., Duke, R. E., Luo, R., Crowley, M., Walker, R. C., Zhang, W. *et al.* (2008) San Francisco: University of California).
138. Cramer, C. J. & Truhlar, D. G. (1992) *Science* **256**, 213-217.
139. Born, M. (1920) *Zeitschrift für Physik A Hadrons and Nuclei* **1**, 45-48.
140. Onsager, L. (1936) *J Am Chem Soc* **58**, 1486-1493.
141. Miertus, S., Scrocco, E., & Tomasi, J. (1981) *Chemical Physics* **55**, 117-129.
142. Klamt, A. & Schuurmann, G. (1993) *J Chem Soc Perk T 2* 799-805.
143. Barone, V. & Cossi, M. (1998) *Journal of Physical Chemistry A* **102**, 1995-2001.
144. Schaefer, A., Klamt, A., Sattel, D., Lohrenz, K., & Eckert, F. (2000) *Phys. Chem. Chem. Phys.* **2**, 2187-2193.
145. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., & Mccammon, J. A. (2001) *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10037-10041.
146. Onufriev, A., Case, D. A., & Bashford, D. (2002) *Journal of Computational Chemistry* **23**, 1297-1304.
147. Vondrasek, J., Bendova, L., Klusak, V., & Hobza, P. (2005) *J. Am. Chem. Soc.* **127**, 2615-2619.
148. Richie, K. A., Teng, Q., Elkin, C. J., & Kurtz, D. M., Jr. (1996) *Protein Sci.* **5**, 883-894.
149. Watanabe, K., Masuda, T., Ohashi, H., Mihara, H., & Suzuki, Y. (1994) *Eur J Biochem* **226**, 277-283.
150. Morozov, A. V., Misura, K. M. S., Tsemekhman, K., & Baker, D. (2004) *Journal of Physical Chemistry B* **108**.
151. Bendova-Biedermannova, L., Hobza, P., & Vondrasek, J. (2008) *Proteins-Structure Function and Bioinformatics* **72**, 402-413.
152. Riley, K. E., Cui, G. L., & Merz, K. M. (2007) *Journal of Physical Chemistry B* **111**, 5700-5707.
153. Berka, K., Laskowski, R., Riley, E., Hobza, P., & Vondrášek, J. (2009) *Journal of Chemical Theory and Computation* **5**, 982-992.
154. Černý, J., Jurečka, P., Hobza, P., & Valdes, H. (2007) *J. Phys. Chem. A* **111**, 1146-1154.

155. Rezac, J., Jurečka, P., Riley, K. E., Cerny, J., Valdes, H., Pluhackova, K., Berka, K., Rezac, T., Pitonak, M., Vondrasek, J. *et al.* (2008) *Collection of Czechoslovak Chemical Communications* **73**, 1261-1270.
156. Riley, K. E. & Hobza, P. (2008) *Journal of Chemical Theory and Computation* **4**, 232-242.
157. DeLano, W. L. (2002) (DeLano Scientific, Palo Alto, CA, USA., http://www.pymol.org, accessed Jun 31, 2007).
158. Lindahl, E., Hess, B., & van der Spoel, D. (2001) *J Mol Model* **7**, 306-317.
159. Ahlrichs, R., Bar, M., Haser, M., Horn, H., & Kolmel, C. (1989) *Chemical Physics Letters* **162**, 165-169.
160. Aradi, B., Hourahine, B., & Frauenheim, T. (2007) *J. Phys. Chem. A* **111**, 5678-5684.
161. Sorin, E. J. (2005) *Biophysical Journal* **88**, 2472-2493.
162. Rezac, J. (2009) (www.cuby.rezacovi.cz, accessed Oct 31, 2009).
163. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Cheeseman, J. R., Montgomery Jr., J. A., Vreven, T., Kudin, K. N. *et al.* (2004) (Gaussian, Inc., Wallingford CT, USA, http://www.gaussian.com accessed Jun 6, 2005).
164. Werner, H. J., Knowles, P. J., Lindh, R., Manby, F. R., Schtz, M., Celani, P., Korona, T., Rauhut, G., Amos, R. D., Bernhardsson, A. *et al.* (2007) http://www.molpro.net accessed Dec 8, 2007).
165. Stewart, J. J. P. (2007) (Stewart Computational Chemistry, Colorado Springs, CO, USA, http://OpenMOPAC.net accessed Jul 10, 2008).

List of Abbreviations

All amino acid residues are in the text thoroughly cited in their one or three letter abbreviations. All energies noted in this work are in kcal/mol.

BSSE – Basis Set Superposition Error
BEGDB – Benchmark Energy and Geometry Database
CASP – Critical Assessment of Techniques for Protein Structure Prediction
CBS – Complete Basis Set
CCSD(T) – Coupled Clusters with Single, Double and Perturbative Triple Excitations
CCSDT – Coupled Clusters with Single, Double and Triple Excitations
COSMO – Conductor-like Screening Model
CI – Configuration Interaction
CP – Counterpoise Correction
DFT – Density Functional Theory
DFT-D – Density Functional Theory with Empirical Dispersion
DFT-SAPT – Density Functional Theory with Symmetry-Adapted Perturbation Theory
DSC – Differential Scanning Calorimetry
GB – Generalized Born Solvent Model
GGA – Generalized Gradient Approximation
HF – Hartree-Fock Method
ITC – Injection Titration Calorimetry
MD – Molecular Dynamics
MM – Molecular Mechanics
MP – Møller-Plesset Method
MP2 – Second-order Møller-Plesset Method
NMR – Nuclear Magnetic Resonance
OPLS-AA/L – Optimized Potential for Liquid Simulations Optimized for Amino Acids
Parm03 – Amber parm03 Force Field
PB – Poisson-Boltzmann Implicit Solvent Model
PCM – Polarizable Continuum Solvent Model
PM6 – Parameterized Model 6
PM6-DH – Parameterized Model 6 with Dispersion and Hydrogen Bond Corrections
PDB – Protein Databank
QM – Quantum Mechanics
RESP – Restrained Electrostatic Potential Method
RI – Resolution-of-Identity Approximation
rmsd – Root Mean Square Distance
SAPT – Symmetry-Adapted Perturbation Theory
SCC-DFTB-D – Self-Consistent Charge Density Functional Tight Binding Method with Empirical Dispersion
SCRF – Self-Consistent Reaction Field Method

Table of Figures

Tables

Appendix

A. Berka, K., Hobza, P., Vondrášek, J.: Analysis of energy stabilization inside the hydrophobic core of rubredoxin. *ChemPhysChem* **2009,** 10 (3), 543-548.

B. Řezáč, J., Berka, K., Horinek, D., Hobza, P., Vondrášek, J.: The stabilization energy of the glu-lys salt bridge in the protein/water environment: Correlated quantum chemical ab initio, dft and empirical potential studies. *Collection of Czechoslovak Chemical Communications* **2008**, 73 (6), 921-936.

C. Biedermannová, L., Riley, K. E., Berka, K., Hobza, P., Vondrášek, J.: Another role of proline: stabilization interactions in proteins and protein complexes concerning proline and tryptophane. *Physical Chemistry Chemical Physics* **2008**, 10 (42), 6350-6359.

D. Berka, K., Laskowski, R., Riley, K. E., Hobza, P., Vondrášek, J.: Representative amino acid side chain interactions in proteins. a comparison of highly accurate correlated ab initio quantum chemical and empirical potential procedures. *Journal of Chemical Theory and Computation* **2009,** 5 (4), 982-992.

E. Řezáč, J., Jurečka, P., Riley, K. E., Černý, J., Valdes, H., Pluháčková, K., Berka, K., Řezáč, T., Pitoňák, M., Vondrášek, J., Hobza, P.: Quantum chemical benchmark energy and geometry database for molecular clusters and complex molecular systems (www.begdb.com): A users manual and examples. *Collection of Czechoslovak Chemical Communications* **2008**, 73 (10), 1261-1270.

F. Berka, K., Laskowski, R., Hobza, P., Vondrášek, J.: The Matrices of Side-Chain–Side-Chain Interactions in Proteins *Submitted* **2009**