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

Side-chain Side-chain Interactions in Proteins

Doctoral Thesis Abstract

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

Praha 2009

Introduction

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.

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?*
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?*

Methods

The methodical part of this work consisted from two parts: (a) selection of an appropriate representative model of side-chain side-chain interactions and (b) selection of an applicable computational method providing interaction energies.

Selection of model

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 tryptophane 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.

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. 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 and all 20 different amino acid residues. We have also used either all 20 x 20 representative pairs or even all contacts for selected residues in the Atlas of Protein Side-Chain Interactions dataset..

Selection of computational method

We have utilized several *ab initio* or semiempirical as well as empirical force field methods to test their accuracy and speed for the calculations of the side-chain side-chain interactions. As a benchmark method was used the most accurate CCSD(T)|CBS method.

The *ab initio* calculations were calculated with several codes with the common ruby interface called “cuby” created by Dr. Jan Řezáč. Most of the *ab initio* calculations were performed with Turbomole 5.8 package – RI-MP2, RI-DFT-D. Energy decomposition with DFT-SAPT calculations were performed with the use of two codes – Gaussian 03 and Molpro 2006 package. Molpro 2006 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 and the dispersion and hydrogen bond corrections were added within the ruby code from Jan Řezáč. SCC-DFTB-D energies were calculated with dftb+ program package.

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

Results

All interaction energies for the side-chain contacts within the hydrophobic core of rubredoxin were calculated by DFT-SAPT method decomposing 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. Most of the stabilizing energy originates in the dispersion term. Even 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.

Salt bridges are thought to provide higher thermostability for thermophilic proteins. For this reason, six different salt bridges have been selected from the mesophilic as well as thermophilic rubredoxins and their interaction energies were evaluated. The major conclusion is that the strength of the salt bridge interaction is substantially lowered upon the presence of protein-like or water environments or with the change of the pH.

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. One of the reasons can be unusually strong interactions between proline and aromatic residues. 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.

Because all previous studies were focused only on a partial selection of side-chain side-chain interactions in proteins, the set of 24 side-chain pairs was selected representing typical interactions in proteins. 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 selected side-chain pairs, a high degree of agreement was detected between different methods, even though the range of interaction energies was extremely large – over two orders of magnitude. The RI-

DFT-D was found to be the most effective method reasonable level of accuracy. Much cheaper semiempirical methods PM6-DH or SCC-DFTB-D performed noticeably worse, but they still performed better than force field methods parm03 and OPLS-AA/L.

The energy decomposition of the interaction energies for the set with DFT-SAPT method showed that polar residues interact mostly by the first-order electrostatic interaction, while nonpolar residues interact mostly by the second-order dispersion.

The knowledge of benchmark values for the representative set of interactions allowed us to calculate stabilization energies for all 20×20 possible pairs of side chain – side chain interactions with selected RI-DFT-D method. The results showed that most of interaction energies calculated at RI-DFT-D level are attractive in the gas phase. The variability of the strength as well as population of the side-chain side-chain contacts is enormous.

Force fields methods are the most used methods for the simulations of proteins. Fortunately they 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.

The change of interaction energies for the set or the complete matrix of side-chain side-chain interactions upon introduction of an environment was studied with the help of PCM or COSMO solvent models with two different values of dielectric constants to imitate protein-like ($\varepsilon = 4$) or water environment ($\varepsilon = 80$). The environment highly promotes interactions between residues of aromatic or aliphatic character.

The leucine-tryptophane pair (LW) was selected as a model system to put characteristic values of interaction energies in larger structural context. 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 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.

Conclusions

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 interactions between the same residues found elsewhere.
2. The strength of the salt bridge interaction is substantially lowered or even negligible upon the presence of environment.
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 interaction energies for side-chain pairs on benchmark set showed that 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 www.begdb.com.
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 as well as the population of side-chain interactions is enormous and it poses a 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. Representative pairs from 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.